5-2015

Summary of survival analysis with SAS procedures.

Derek Duane Childers 1990-
University of Louisville

SUMMARY OF SURVIVAL ANALYSIS WITH SAS PROCEDURES

By

Derek Duane Childers

B.S., University of Louisville, 2013

A Thesis

Submitted to the Faculty of the

School of Public Health of the University of Louisville

in Partial Fulfillment of the Requirements

for the Degree of

Master of Science in Biostatistics: Decision Science

School of Public Health

University of Louisville

Louisville, Kentucky

May 2015

SUMMARY OF SURVIVAL ANALYSIS WITH SAS PROCEDURES

By

Derek Duane Childers

B.S., University of Louisville, 2013

A Thesis Approved on

April 13, 2015

By the following Thesis Committee:

_____

Shesh Rai

_____

Guy Brock

_____

Dongfeng Wu

_____

Frank Groves

DEDICATION

This thesis is dedicated to my parents

Mr. Duane Eddie Childers

And

Mrs. Cynthia Fern Childers

who have provided me background and knowledge to not only

be successful in life, but to enjoy every second of it!

ABSTRACT

SUMMARY OF SURVIVAL ANALYSIS

WITH SAS PROCEDURES

Derek D. Childers

April 13, 2015

The research conducted for this thesis was performed to summarize some of the most commonly used survival analysis techniques as well as to create one macro that will provide the solutions for these techniques. Some of the techniques that this thesis focuses on are survival and hazard functions, mean and median survival times, life table, log rank test, proportional hazards/model building, and competing risk. To further analyze these survival analysis techniques I will use the Bone Marrow Transplantation for Leukemia dataset. This trial consists of either acute myelocytic leukemia (AML 99 patients) or acute lymphoblastic leukemia (ALL 38 patients). There are two risk level for AML, low risk first readmission (54 patients) and high risk second readmission or untreated first relapse (15 patients) or second or greater relapse or never in remission (30 patients). Any further details of this study can be found in (Copelan, 1991).

TABLE OF CONTENTS

PAGE

LIST OF TABLES

TABLE PAGE

LIST OF FIGURES

FIGURES PAGE

CHAPTER I

INTRODUCTION

When someone says ‘survival analysis’, our initial thought refers to time-to-event data where the event is death. However, there are many different event’s that could be used for survival analysis, such as sickness, hospitalization, transplantation, marriage, or even something as insignificant as buying a new car (Despa). We see time to event data in many different fields of study, such as engineering, public health, epidemiology, biology and medicine to only say a few. Now that being said, for our purposes, we will be focusing on the clinical side of survival analysis which is truly the soul of this topic.

The definition of survival analysis could be written as a number of procedures to analyze data where our outcome variable is the time it took for a specific event to occur. This time to event variable can be measured in days, weeks, years, etc. We can have certain observations that would be considered censored. A censored variable can be assigned when a participant drops out of the study or they never obtain the event of interest. For example, if the study population is cancer patients who are in remission and the event we are interested in is relapse; if a patient never acquires a relapse by the end of the study then that patient would be right censored since they never relapsed. There are other ways to be censored as well but usually it would depend on the study.

To perform survival analysis most biostatisticians will use SAS, or Statistical Analysis Software. SAS is a collection of programs that can be used to manipulate data or analyze

the data. This data can be from Excel, a database, regular text file or many other formats. We can also run SAS on many different computing platforms such as UNIX, Linux or the average PC (Cody, 2011). However, this software isn't just used by biostatisticians, it is used by millions of people all around the world and over 134 countries with approximately 60,000 sites (Delwiche, 2012).

One of the most powerful capabilities of SAS is being able to create a macro, which is a program that allows the user to use very few lines of code and run many different analytical procedures. This is accomplished by writing a program, not visible to the user, which has macro definitions and variables (Stroupe). Then we can call this code through the macro function and we specify 'options' that the programmer pre-defined and based on which options we have chosen we get different results. I have created a macro with survival analysis in mind. I have done this since there are many commonly used procedures in survival analysis and it would be beneficial to have one macro that does many different survival analysis techniques.

Every time my macro is used by a user it will print out a life table that has the following variables: time, number left, number of events, survival estimate, survival standard error, cumulative hazard estimate, and cumulative hazard standard error. The reason I have decided to let this specific table be printed no matter what options you are wanting is because the survival life table is always a good way to make sure the data looks like what the user thinks; plus it is always a very nice presenting tool to give to a customer. The options that a user can choose from are showing the survival or cumulative hazard plots, performing a log rank test, calculating the mean or median survival times, two competing risk procedures (cause specific hazard and cumulative

incidence), printing the cumulative incidence plot, and creating a proportional hazards model which has its own small list of options.

The layout of my report is first giving all the necessary background and methodology for each of survival analysis techniques I focused on while creating my macro, this is chapter 2. For all of these methods I have focused on continuous random variables and nonparametric analysis only. In chapter 3 I will go through all the details of my macro and explain what each step does and how to use my macro efficiently. Chapter 4 will be a brief results review, showing that my macro does indeed perform as expected. Then I will end my report in chapter 5 with a discussion and conclusion.

CHAPTER 2

METHODS

2.1 Survival Function and Plot

The most common method to describing any time to event data is the survival function. The survival function is defined as the probability that an individual with survive beyond a certain time period:

$$S(x) = \Pr(X > x)$$

If our random variable X is continuous then we can say that the survival function is strictly decreasing. Which makes sense, if we say X is time then as time goes on more people tend to experience the event (for example death) which in turn drops our overall survival rate. The survival function is related to other functions as well. When we look at the cumulative hazard function $F(x)$, $S(x) = 1-F(x)$, I will further discuss the cumulative hazard function in the next section. We can also see that when provided with the probability density function, $f(x)$, we can find the survival function by integration:

$$S(x) = \int_x^\infty f(t)dt$$

After we have calculated our survival function, we commonly want a graphical representation of our survival function which is denoted as the survival curve. These

survival curves are non-increasing and monotone functions which equal one at zero and zero as the time goes to infinity (Klein, 2003).

2.2 Cumulative Hazard Function and Plot

In order to further discuss the cumulative hazard function we first need to review the basic hazard function. The hazard function or hazard rate is defined by (Klein, 2003):

$$h(x) = \lim_{\Delta x \to \infty} \frac{P[x \leq X < x + \Delta x \,|X \geq x]}{\Delta x}$$

When our X is a continuous random variable we can say that h(x) = f(x) / S(x). Our quantity we focused on, the cumulative hazard function, is defined below:

$$H(x) = \int_0^x h(t)dt$$

One could say that it's is the probability of failure at time x given survival until that time x (Easyfit).

2.3 Mean Survival

After the collection of some time-to-event data, the question of how long does someone has to live can come up regularly. The mean residual life for an individual at x is this measure of expected remaining lifetime or the average remaining survival given that they have survived past time x. It is defined as (Klein, 2003):

$$mrl(x) = E(X - x \mid X > x) = \frac{\int_x^\infty (t-x) f(t) dt}{S(x)} = \frac{\int_x^\infty S(t) dt}{S(x)}$$

This leads to our mean survival time

$$\mu = E(X) = \int_0^\infty t f(t) dt = \int_0^\infty S(t) dt$$

which is the average survival rate (Tsiatis).

2.4 Median Survival

In analysis, we can sometimes find cases where the median is more beneficial than the average. These cases could be when our data is skewed one way or another as an example. The p^{th} quantile for survival is the smallest x_p that satisfies:

$$S(x_p) \leq 1 - p;\ x_p = \inf\{t: S(t) \leq 1 - p\}$$

However, we have an X that is a continuous random variable, so $S(x_p) = 1 - p$. This would lead to the outcome for the median lifetime (50^{th} percentile) or $x_{0.5}$, where $S(x_{0.5}) = 0.5$ (Klein, 2003).

2.5 Interrelationships

When our lifetime variable X is continues, we can have many different interrelationships between the different methods that I've discussed thus far. Below you can see these relationships and they are important because sometimes we cannot do a certain method so we must have ways around that block to still get what we need for our analysis (Klein, 2003).

$$S(x) = \int_x^{\infty} f(t)dt$$

$$= \exp\left[-\int_0^x h(t)dt\right] = \exp[-H(x)] = \frac{mrl(0)}{mrl(x)}\exp[-\int_0^x \frac{dt}{mrl(t)}]$$

$$f(x) = -\frac{d}{dx}S(x)$$

$$= h(x)S(x) = (\frac{d}{dx}mrl(x) + 1)(\frac{mrl(0)}{mrl(x)^2})\exp[-\int_0^x \frac{dt}{mrl(t)}]$$

$$h(x) = -\frac{d}{dx}\ln[S(x)]$$

$$= \frac{f(x)}{S(x)} = (\frac{d}{dx}mrl(x) + 1)/mrl(x)$$

$$mrl(x) = \frac{\int_x^{\infty} S(t)dt}{S(x)} = \frac{\int_x^{\infty}(\mu - x)f(t)dt}{S(x)}$$

2.6 Life Table

A survival life table is a way to numerically show as our time increases and events keep occurring, the survival rate will decrease. We will use the Kaplan-Meier or Product Limit estimation process do estimate the methods. The first column of the life table is sometimes the event time (t_i) while the second column is the number of individuals who have experience the event, otherwise known as the number of events (d_i). The third column for us will be the number of individuals left or number at risk (Y_i). The fourth and fifth columns are related to the survival estimate. We will first calculate the survival rate at that time with the following formula:

$$\hat{S}(t) = \prod_{t_i} (1 - \frac{d_i}{Y_i})$$

Then we will find the standard error for the survival estimate via:

$$\sqrt{\hat{S}(t_{j-1})^2 (\sum_{i=1}^{j-1} \frac{d_i}{Y_i(Y_i - d_i)})}$$

The last two columns are for the cumulative hazard rate and its standard error. The cumulative hazard rate and its standard error can be calculated by using this two equations respectively (Klein, 2003):

$$\sum_{i:t_i<t} \frac{d_i}{Y_i}, \sqrt{\sum_{i:t_i<t} \frac{d_i}{Y_i^2}}.$$

2.7 Log Rank Test

When our study consists of two or more groups, such as a treatment group and a placebo group, we would always like to test the hypothesis of no difference between the group's survival rates. We can think of this as testing whether the two survival curves are identical or not (Survival Analysis). Moreover,

$$H_0: no\ statistially\ significant\ difference$$

$$H_1: statistically\ significant\ difference$$

The survival curves will be estimated by the Kaplan-Meier process for each group and then we use the log rank test to compare the group's survivals. There are a few variations of the log rank test statistic. However, due to using SAS for my analysis the test statistic method is already pre-defined.

2.8 Competing Risk

In a study there could be many reasons a subject could experience a problem and fail, these problems are competing risks. An example of a competing risk is when a patient dies from another cause than what we are focusing on in our study. In order to look into the parameters for competing risks we will focus on the latent failure time approach. We will let X_i, i=1,…,K be the possible impalpable time to occurrence of the i^{th} competing risk. We will observe the time at which a individual experiences any cause which makes them fail, $T = \min(X_1,\ldots,X_p)$ and an indicator variable denoted by δ which will show which of K causes made the individuals experience the failure. The standard competing risk parameter is the cause-specific hazard rate, this is shown by:

$$h_i(t) = \lim_{\Delta t \to 0} \frac{P[t \leq T < t + \Delta t, \delta = i \,|T \geq t]}{\Delta t}$$

$$= \lim_{\Delta t \to 0} \frac{P[t \leq X_i < t + \Delta t, \delta = i \,|X_i \geq t, j = 1, \ldots, K]}{\Delta t}$$

$h_i(t)$ is telling us the rate at which the participants who have not experienced a competing risk yet are going to experience the i^{th} risk. We now can find the overall hazard rate by using the equation below (Klein, 2003):

$$h_T(t) = \sum_{i=1}^{K} h_i(t).$$

We are often interested in the probability of death from a specific cause in the real world where all other risks are acting on the patient. This probability can be expressed as the cumulative incidence function (Klein, 2003):

$$F_i(t) = P[T \leq t, \delta = i]$$

The cumulative incidence function can also be referred to as the sub-distribution function which follows from the following definitions:

$$F_i(t) = \int_0^t S(t)h_k(t)dt = \int_0^t S(t)dH_k(t)dt, k = 1, \ldots, K$$

where $H_k(t)$ is the cause specific cumulative hazard rate function (So).

The cause specific hazard function and cumulative incidence sub-distribution hazard function will usually not yield the same conclusion due to the fact they have different risk sets. Cause specific risk set will go down each point in time at which a different cause produces a failure. Sub-distribution risk set is when an individual fails from a competing risk which allows us to classify them as experiencing an event at time infinity. Furthermore, we can say that this treats them as being censored at the last follow up time (Brock). Whenever competing risk are present, we can seek out Gray's Test for comparing two sub-distribution functions. This is comparable to the log rank test discussed earlier when competing risk are not present for that analysis.

2.9 Cox Proportional Hazards

There are methods in which we can assess several risk factors at the same time, similar to multiple regression. One of the more popular regression methods is the Cox proportional hazards regression which will relate different risk factors to the survival time. Our measure of effect is the hazard rate or risk of failure, given that the individual has not experienced the event up to a specified time. Cox proportional hazards is a semi-parametric model, it will explain the effect of the explanatory variables on the hazard rates. The survival time has an assumption that it follows its own hazard function which is denoted by $\lambda_i(t)$ and can be shown as:

$$\lambda(t; Z_i) = \lambda_o(t)\exp(Z'_i\beta)$$

where $\lambda_o(t)$ is a baseline hazard function, Z is a vector of the explanatory variables and β is the vector of unknown parameters. This β is also assumed to be the same for each individual (SAS/STAT).

Proportional hazards can also see cases where ties are present, ties can happen if the survival times are calculated from a continuous time model and are grouped. For continuous data there are three methods: exact, Breslow, and Efron. The exact method will calculate the exact conditional probability under the model where the set of observed ties happen before other larger values. Breslow and Efron are quite close in their calculations and are approximations of the exact method. Both of these methods will use their own partial likelihoods. Breslow's partial likelihood is defined as:

$$L(\beta) = \prod_{i=1}^{D} \frac{\exp(\beta^t s_i)}{[\sum_{j \in R_i} \exp(\beta^t Z_j)]^{d_i}}$$

and we can express Efron's partial likelihood as:

$$L(\beta) = \prod_{i=1}^{D} \frac{\exp(\beta^t s_i)}{\prod_{j=1}^{d_i} [\sum_{k \in R_i} \exp(\beta^t Z_k) - \frac{j-1}{d_i} \sum_{k \in D_i} \exp(\beta^t Z_k)]}.$$

We can also identify some important prognostic factors by variable selection. We can use forward selection, backward elimination, stepwise selection, or best subset selection. The best subset selectin is based on the likelihood score statistic which identifies a specific number of models that contain a few variables which are considered the best until a single model contains all of the explanatory variables. If we choose the forward selection process then the parameter estimates for effects are first calculated. Then we would compute the score statistic for each of effects not included in the model. These statistics are the chi-square statistic based off of the score test for testing the null that the related effect is not in the model. If the largest of these are significant then they are added to the model and never removed, this process is repeated until there are no remaining effects. The backward elimination process takes all the parameters in the beginning for the complete model. Then it will get the Wald statistic for each effect, the Wald statistic is the chi-square of the Wald test for testing the null of correspondence with the null. Next, the smallest of these statistics are looked at to see if they are significant or not; if they are not significant then it will be removed. This process is repeated until no other effects meets the removal criteria. The stepwise procedure is relatable to forward selection but they differ due to the stepwise procedure possibly

removing effects that are already in the model. Effects are entered into and removed from the model in a way where the forward selection can be followed by some backward elimination steps. This procedure will remove the effect if no further effect can be added or if it's just been entered and is the only effect to be removed in the next backward elimination step.

CHAPTER 3

MACRO

The foresight of this macro was to have a program that did some of the bare essentials in survival analysis, and have those techniques in one location for convenient access. My macro will, in a short explanation, will do the procedures described in chapter 2 of this report. The macro is created in a way that not every procedure is ran every time, because realistically we do not always need to run every procedure that I talked about thus far. Furthermore, I have created where you can specify which procedure you would like to run on your data and I will further explain this now. The only procedure that will always print out is the life table because this survival tool is a very useful table and can be useful even when not expected.

The 'mandatory' options when using this macro are *dat, status, grp, time,* and *indicator*. *dat* is the location of the data in SAS you will be using for your analysis; your data has to be uploaded into SAS as a SAS dataset to work properly. *status* represents the value that is used as your disease free indicator, normally this value is 0 (or 1) but it could be different depending on how the data was collected. The *grp* variable is used to specify the strata variable and *indicator* is the variable in your data which lists which observation has experienced the event or not.

The 'optional' choices for this macro are *survplot, cumhazplot, logrank_mean, median, med_maxtime, med_bytime, comp_risk, risk_test, cum_plot, prop_haz,*

prop_haz_bygroup, *cov_list*, *ties,* and *selection*. The *survplot* and *cumhazplot* have 'yes' or 'no' values for whether you want to see those printed or not. *survplot* is the survival plot while *cumhazplot* is the cumulative hazard plot. *logrank_mean* will calculate the log rank test to compare multiple survival curves and will automatically produce the mean survival time estimate, this also has a 'yes' or 'no' value. When you specify *median*=yes in the program you will have to make sure you also include two other parameters with *median.* They are *med_maxtime* and *med_bytime* and those values are for the interval in which we will calculate the median survival for. *med_maxtime* is the maximum value and *med_bytime* is what value we want to increase by. For the competitive risk code (when *comp_risk* = yes, if =no then no need for these options) you will first need to specify *risk_test* as either cause-spec or cum_inc. These represent cause specific hazards and cumulative incidence respectively. For when you choose cum_inc for *risk_test* you may want to also see the cumulative incidence plot, to do this use the option *cum_plot*=yes which actually uses a SAS internal macro %cumincid to get the graph; if you do not wish to see the cumulative incidence plot use *cum_plot*=no. For the cox proportional hazards section, first specify *prop_haz* as yes or no. If yes, there are four other options you must specify as well. At the start you will need to specify if you wish to proceed by group or not (*prop_haz_bygroup*=yes or no), for either case you will still have to give these other three parameter values: *cov_list*, *ties,* and *selection*. *cov_list* is a list of the covariates/effects you want to test your model with, for example if you want to test the age, gender, and race your code would be *covlist=age gender race*. Next you will give which method you wish to use for *ties*: discrete, Breslow, Efron, or exact. For the *selection* variable you will choose either none, forward, backward, or stepwise. Each

of these methods listed above were discussed in detail in chapter 2, so if you have questions on which to choose please review chapter 2.

CHAPTER 4

BRIEF RESULTS

For the analysis I conducted we will look at the disease free survival time which is the time to event or end of study with our indicator variable being disease free (1-dead/relapsed and 0-alive/disease free) as well. I will also stratify by group which is either ALL-1, AML low-2, or AML high-3. To further review the macro itself please see the appendix section of this report. Below you see the life table for this analysis:

t2	num_event	num_left	surv	stderr	h	stderrh
1	1	38	0.97368	0.025967	0.02632	0.02632
55	1	37	0.94737	0.036224	0.05334	0.03772
74	1	36	0.92105	0.043744	0.08112	0.04685
86	1	35	0.89474	0.049784	0.10969	0.05487
104	1	34	0.86842	0.054836	0.13910	0.06226
107	1	33	0.84211	0.059153	0.16941	0.06924
109	1	32	0.81579	0.062886	0.20066	0.07597
110	1	31	0.78947	0.066135	0.23291	0.08253
122	2	30	0.73684	0.071434	0.29958	0.09505
129	1	28	0.71053	0.073570	0.33530	0.10153
172	1	27	0.68421	0.075405	0.37233	0.10808
192	1	26	0.65789	0.076960	0.41079	0.11472
194	1	25	0.63158	0.078252	0.45079	0.12149
230	1	23	0.60412	0.079522	0.49427	0.12904
276	1	22	0.57666	0.080509	0.53973	0.13681
332	1	21	0.54920	0.081223	0.58735	0.14486
383	1	20	0.52174	0.081672	0.63735	0.15325
418	1	19	0.49428	0.081860	0.68998	0.16203
466	1	18	0.46682	0.081788	0.74553	0.17129
487	1	17	0.43936	0.081457	0.80436	0.18111
526	1	16	0.41190	0.080862	0.86686	0.19159
609	1	14	0.38248	0.080260	0.93829	0.20447
662	1	13	0.35306	0.079296	1.01521	0.21846
2081	0	1	0.35306	0.079296	1.01521	0.21846

Table1

Next we will see our disease free Kaplan-Meier survival plot and next the cumulative hazard plot:

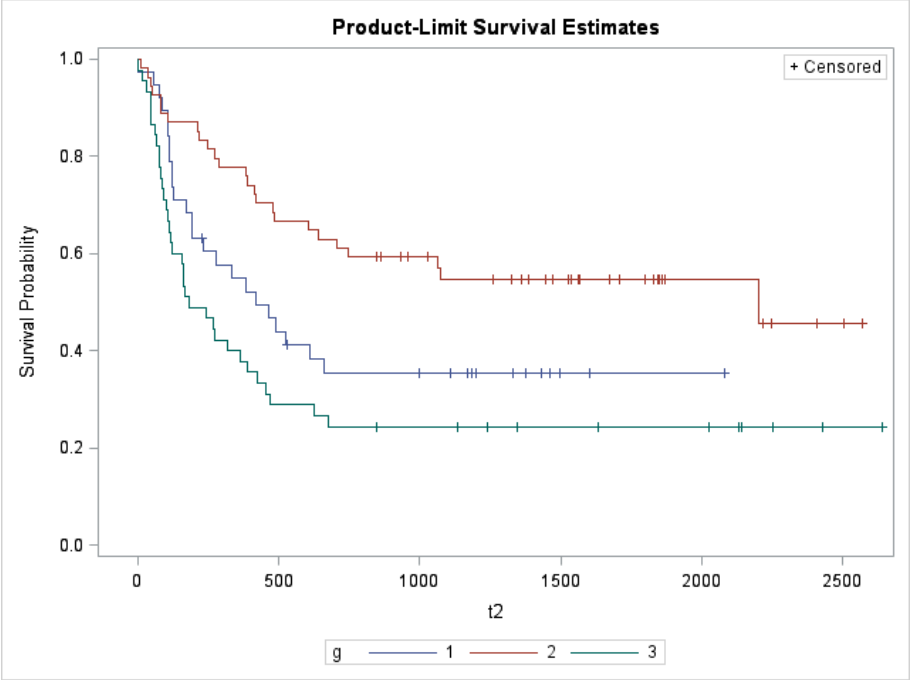

Figure1

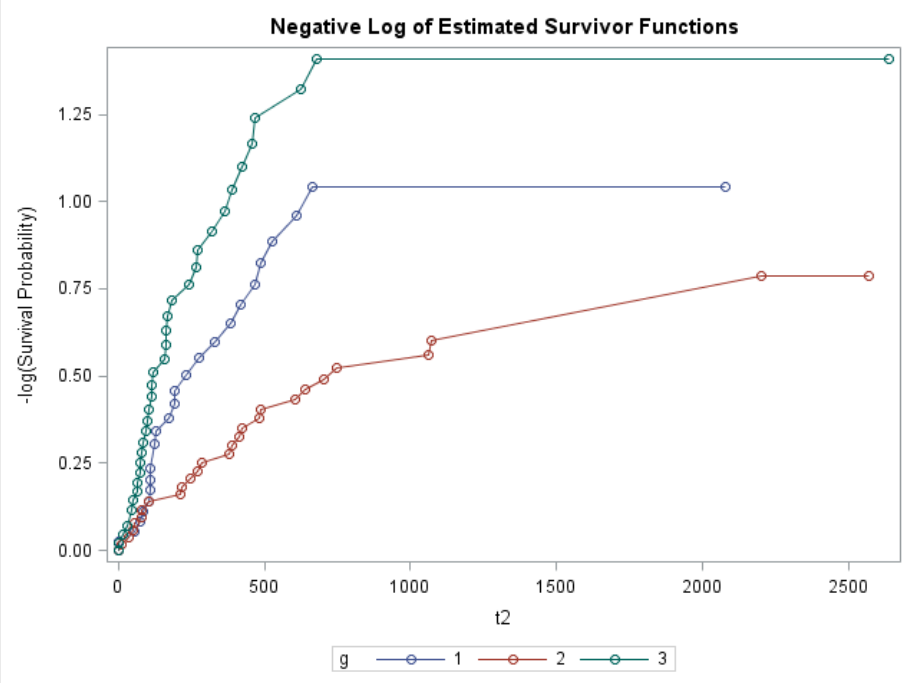

Figure2

We can see from our survival plot that AML low has the best chance for survival and then ALL is second followed by AML high, but are these visual difference's significant? That is why we will now perform a log rank test, to compare the survival curves.

Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	13.8037	2	0.0010
Wilcoxon	16.2407	2	0.0003
-2Log(LR)	19.5313	2	<.0001

Table2

So from the results in the table above we can see from our Log-Rank test that our p-value is p=0.001 which is less than the normal cut off of 0.05. This indicates that there is indeed a significant difference between the survival curves.

The mean survival times estimated by SAS are 398.24 for ALL with a standard error of 41.11, 1382.46 for AML low with a standard error of 129.65, and 312.47 for AML high with a standard error of 38.69. Below you will see the first few observations which include the median survival rates:

Interval		Number Failed	Number Censored	Effective Sample Size	Conditional Probability of Failure	Conditional Probability Standard Error	Survival	Failure	Survival Standard Error	Median Residual Lifetime	Median Standard Error	Evaluated at the Midpoint of the Interval			
[Lower,	Upper)											PDF	PDF Standard Error	Hazard	Hazard Standard Error
0	50	6	0	45.0	0.1333	0.0507	1.0000	0	0	195.0	33.5410	0.00267	0.00101	0.002857	0.001163
50	100	7	0	39.0	0.1795	0.0615	0.8667	0.1333	0.0507	237.5	78.0625	0.00311	0.00108	0.003944	0.001483
100	150	5	0	32.0	0.1563	0.0642	0.7111	0.2889	0.0676	300.0	70.7107	0.00222	0.000937	0.00339	0.001511
150	200	5	0	27.0	0.1852	0.0748	0.6000	0.4000	0.0730	337.5	64.9519	0.00222	0.000937	0.004082	0.001816
200	250	1	0	22.0	0.0455	0.0444	0.4889	0.5111	0.0745	500.0	117.3	0.000444	0.000439	0.00093	0.00093

Table3

For the cause-specific competitive risk we will look at the table at the top of the next page which gives the hazard ratios for each group versus the other:

Cause-Specific Hazards: Hazard Ratios for g			
Description	Point Estimate	95% Wald Confidence Limits	
g 2 vs 3	0.384	0.228	0.646
g 3 vs 2	2.603	1.548	4.379
g 2 vs 1	0.563	0.321	0.989
g 1 vs 2	1.776	1.011	3.118
g 3 vs 1	1.466	0.868	2.476
g 1 vs 3	0.682	0.404	1.152

Table4

So when a point estimate is greater than 1, we can say that the first group has a higher chance of experiencing an event. For example, g 3 vs 2 has a point estimate of 2.603, then would be that AML high (group 3) has a 1.603 times higher chance of experiencing death/relapse than AML low (group 2) has. For cumulative incidence we will also look at the hazard ratios, which can be seen below and we notice that the point estimates don't much differ from cause-specific but there are some slight differences within the confidence limits.

Cumulative Incidence Hazards: Hazard Ratios for g			
Description	Point Estimate	95% Wald Confidence Limits	
g 2 vs 3	0.384	0.227	0.649
g 3 vs 2	2.603	1.542	4.396
g 2 vs 1	0.563	0.328	0.967
g 1 vs 2	1.776	1.034	3.048
g 3 vs 1	1.466	0.861	2.495
g 1 vs 3	0.682	0.401	1.161

Table5

Be that as it may, the confidence intervals are based off of Greenwoods formula which can be less efficient when compared to other methods such as Wilson's method, Agresti-Coull's method or other methods similar to these (Yuan). The macro that was created for this analysis does not report the confidence intervals using these methods.

At the top of the next page we can also look at the cumulative incidence function graph:

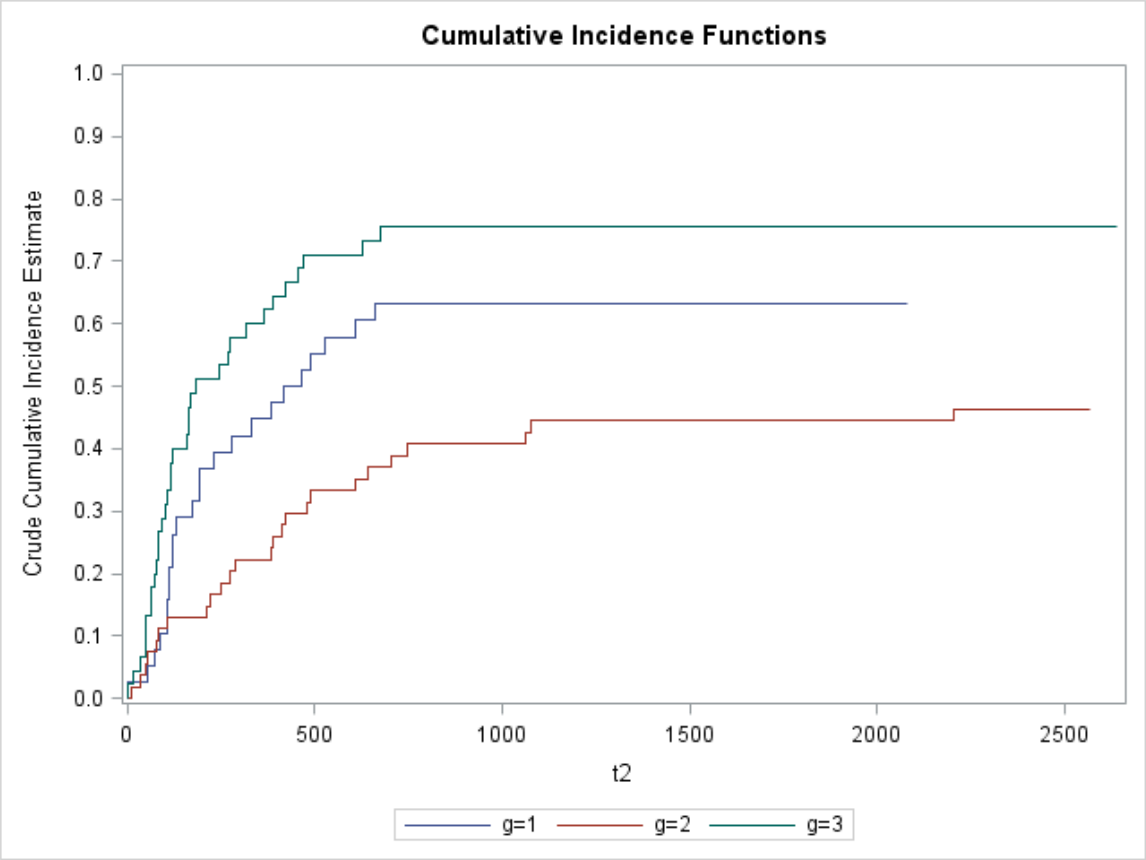

Figure3

For the proportional hazards model I am using the stepwise procedure for selection and the Breslow method for ties. My covariate/effect list will consist of patient's age, sex, CMV status (0-negative, 1-positive) and which hospital they were located in (1-Ohio State, 2-Alfred, 3-St. Vincent, 4-Hahnemann). When I run the analysis to choose the best model I get that the only significant effect is the patient's age.

CHAPTER 5

CONCLUSION

In this report we observed a macro that will provide the user with a life table, Kaplan-Meier survival plot, cumulative hazard plot, log rank test, mean survival estimates, median survival estimates, cause-specific hazard ratios, cumulative incidence hazard ratios and plot, and a cox proportional hazard model. The methodology for each of these techniques were also discussed in detail to provide clarity for what the program is doing.

I used the bone marrow transplant data for leukemia patients in order to show my program working with ease and simplicity. We conducted a log rank test that showed us how the three groups (ALL, AML low and AML high) differ when comparing their survival rates. We also seen this fact in the survival plot since AML low had the best visual survival appearance compared to ALL being in the middle and AML high which being the worst. Then we further investigated this fact by looking at the cause-specific and cumulative incidence competing risk hazard ratios and we seen that the AML low group always had the best chance of survival while the AML high group always had the worst chance. Lastly we conducted a small model building exercise with some of the basic covariates like age and sex and by using a stepwise selection procedure and Breslow's method for ties we concluded that a model with just the age effect was the best model with the four effects that we chose.

We have seen how SAS is a very powerful and robust statistical software and this this fact cannot be denied since it is being used by statisticians all around the world. I found that in my idea of what I wanted to do for this paper, SAS was the best option for me to get a perfect working program to satisfy all of my survival analysis needs. Moreover, by using SAS I was able to create a macro that could potentially be used for more than the procedures I created, this macro could become 'the' survival analysis package that every statistician wants in their tool bag when tackling a survival problem.

REFERENCES

Copelan, E. A. Biggs, J. C., Thompson, J. M., Crilley, P., Szer, J., Klein, J. P., … Kapoor, N. (1991). *Treatment for Acute Myelocytic Leukemia with Allogeneic Bone Marrow Transplantation Following Preparation with Bu/Cy.* Blood, 78, 838-843.

Despa, S. (n.d.). *What is Survival Analysis?* (StatNews #78) Retrieved from Cornell University website: http://www.cscu.cornell.edu/news/statnews/stnews78.pdf

Cody, R. P., & SAS Institute. (2011). *SAS statistics by example*. Cary, N.C: SAS Pub.

Delwiche, L. D., & Slaughter, S. J. (2012). *The little SAS book: A primer, fifth edition*. Cary, N.C: SAS Institute.

Stroupe, J. (n.d.). *Nine Steps to Get Started using SAS Macros* (56-28). Retrieved from SAS Institute, Chicago, IL website: http://www2.sas.com/proceedings/sugi28/056-28.pdf

Klein, J. P., & Moeschberger, M. L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data.* New York: Springer.

EasyFit - Cumulative Hazard Function. (n.d.). Retrieved from http://www.mathwave.com/help/easyfit/html/analyses/graphs/cumulative_hazard.html

Tsiatis, A., & Zhang, D. (n.d.). *Analysis of Survival Data* [PDF]. Retrieved from http://www4.stat.ncsu.edu/~dzhang2/st745/chap1.pdf

Survival Analysis. (n.d.). Retrieved from http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Survival/BS704_Survival_print.html

So, Y., Lin, G., & Johnston, G. (n.d.). *Using the PHREG Procedure to Analyze Competing-Risks Data*. Retrieved from SAS Institute Inc. website: http://support.sas.com/rnd/app/stat/papers/2014/competingrisk2014.pdf

Brock, Guy N Barnes, Christopher Ramirez, Julio A Myers, John. (2011). *How to handle mortality when investigating length of hospital stay and time to clinical stability*. (BioMed Central Ltd.) BioMed Central Ltd.

SAS/STAT(R) 9.22 User's Guide. (n.d.). Retrieved from http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_phreg_sect001.htm

Yuan, X., & Rai, S. (2011). *Confidence Intervals for Survival Probabilities: A Comparison Study, Communications in Statistics - Simulation and Computation*(40:7, 978-991). Taylor & Francis.

APPENDICES

```
/*will use %include
'C:\Users\DCMC\Desktop\survival_analysis_overview_macro.sas' so I don't
have to open this exact file up every time to
            run a new analysis*/

%macro surv_all(dat= , status= ,grp= ,time= ,indicator= ,survplot=,
cumhazplot=, logrank=, mean=, median=, med_maxtime=, med_bytime=,
                                    comp_risk=, risk_test=, cum_plot=,
prop_haz=, prop_haz_bygroup=, cov_list=, ties=, selection=);

%let varlist=&time num_event num_left surv stderr h stderrh;

/*Below is to create my table to use for my estimates of the survival
and cum hazards*/
proc sql noprint;
  create table raw_data as
  select  &time, sum(&indicator) as num_event, count(&time) as
sub_total
  from &dat
  where &grp = 1
  group by &time;
quit;

/*Below is to Estimate the Survival Rate and the Cum Hazard Rate*/
data table1;
  set raw_data nobs = all;
  total1 = lag(sub_total);
  retain num_left 38;
   if _n_ > 1 then do;
   num_left = - total1 + num_left; end;
   retain surv 1 delta  h sigma2 0;
   surv = surv*(1-num_event/num_left);
   delta= delta + num_event/(num_left*(num_left-num_event));
   var =(surv**2)*delta;
   stderr = var**.5;
   h = h + num_event / num_left;
   sigma2 = sigma2 + num_event /(num_left)**2;
   stderrh = sigma2**.5;
   keep &varlist;
   if num_event~=0 or _n_ = all;
run;

proc print data = table1 noobs;
  var &varlist;
run;

*Here I am setting all the options in which my macro can run;
```

```
%if &survplot=no %then %do ;
%end;

%if &survplot=yes %then %do;
      %survplot;
%end;

%if &cumhazplot=no %then %do ;
%end;

%if &cumhazplot=yes %then %do;
      %cumhazplot;
%end;

%if &logrank=no %then %do ;
%end;

%if &logrank=yes %then %do;
      %logrank;
%end;

%if &mean=no %then %do ;
%end;

%if &mean=yes %then %do;
      %mean;
%end;

%if &median=no %then %do ;
%end;

%if &median=yes %then %do;
      %median;
%end;

%if &comp_risk=no %then %do ;
%end;

%if &comp_risk=yes %then %do;
      %comp_risk;
%end;

%if &cum_plot=no %then %do ;
%end;

%if &cum_plot=yes %then %do;
      %cum_plot;
%end;

%if &prop_haz=no %then %do ;
%end;

%if &prop_haz=yes %then %do;
      %prop_haz;
%end;

%mend;
```

```
%macro survplot;

/*Below is to Plot the Survival Rate*/
proc lifetest data = &dat plots = (s) graphics;
  time &time*&indicator(&status);
  strata &grp;
  title "Survival Plot";
  symbol v= none;
run;

%mend;

%macro cumhazplot;

/*Below is to Plot the Cum Hazard Rate*/
ods graphics on;

proc lifetest data=&dat plots=(ls) notable;
time &time*&indicator(&status);
strata &grp;
title "Cumulative Hazard Plot";
run;

ods graphics off;

%mend;

%macro logrank;
/*Log rank test*/
proc lifetest data = &dat;
  time &time*&indicator(&status);
  strata &grp;
run;
%mend;

%macro mean;
/*mean survival estimates*/
proc lifetest data = &dat;
  time &time*&indicator(&status);
  strata &grp;
run;
%mend;

%macro median;
/*median survival estimates*/
proc lifetest data = &dat method=lt intervals=(0 to &med_maxtime by
&med_bytime);
  time &time*&indicator(&status);
  strata &grp;
run;
%mend;

%macro comp_risk;
/*competing risk - cause specific hazard or cumulative incidence*/

%if &risk_test=cause_spec %then %do;
```

```
proc phreg data=&dat;
class &grp / order=internal ref=first param=glm;
model &time*&indicator(&status) = &grp;
hazardratio 'Cause-Specific Hazards' &grp / diff=pairwise;
run;
%end;

%if &risk_test=cum_inc %then %do;
proc phreg data=&dat;
class &grp / order=internal ref=first param=glm;
model &time*&indicator(&status) = &grp/ eventcode=1;
hazardratio 'Cumulative Incidence Hazards' &grp / diff=pairwise;
run;
%end;

%mend;

%macro cum_plot;

*here I am using a internal macro to plot the cumulative incidence
function;
%cumincid(data=&dat,time=&time, strata=&grp, status=&indicator,
event=1, compete=0, options=plotcl);

%mend;

%macro prop_haz;

*below is model with no by statement;
%if &prop_haz_bygroup=no %then %do;
proc phreg data = &dat;
  model &time*&indicator(&status) =  &cov_list / ties=&ties
selection=&selection include=1 details risklimits=both type3(ALL)
itprint;
run;
%end;

*here you will model with the by statement;
%if &prop_haz_bygroup=yes %then %do;
proc phreg data = &dat;
  model &time*&indicator(&status) =  &cov_list / ties=&ties
selection=&selection include=1 details risklimits=both type3(ALL)
itprint;
  by &grp;
run;
%end;

%mend;

options symbolgen;

      *survival plot practice;
%surv_all(dat= thesis.bmt,status=0, grp= g,time= t2,indicator= d3,
survplot=yes, cumhazplot=no, logrank=no, mean=no, median=no,
comp_risk=no,
      cum_plot=no, prop_haz=no);
```

```
      *cumulative hazards plot practice;
%surv_all(dat= thesis.bmt,status=0, grp= g,time= t2,indicator= d3,
survplot=no, cumhazplot=yes, logrank=no, mean=no, median=no,
comp_risk=no,
      cum_plot=no, prop_haz=no);

      *log rank test and mean survival practice;
%surv_all(dat= thesis.bmt,status=0, grp= g,time= t2,indicator= d3,
survplot=no, cumhazplot=no, logrank=yes, mean=yes, median=no,
comp_risk=no,
      cum_plot=no, prop_haz=no);

      *median survival practice;
%surv_all(dat= thesis.bmt,status=0, grp= g,time= t2,indicator= d3,
survplot=no, cumhazplot=no, logrank=no, mean=no, median=yes,
med_maxtime= 700,
      med_bytime= 50, comp_risk=no, cum_plot=no, prop_haz=no);

      *cause-specific practice;
%surv_all(dat= thesis.bmt,status=0, grp= g,time= t2,indicator= d3,
survplot=no, cumhazplot=no, logrank=no, mean=no, median=no,
comp_risk=yes,
      risk_test=cause_spec, cum_plot=no, prop_haz=no);

      *cumulative incidence practice;
%surv_all(dat= thesis.bmt,status=0, grp= g,time= t2,indicator= d3,
survplot=no, cumhazplot=no, logrank=no, mean=no, median=no,
comp_risk=yes,
      risk_test=cum_inc, cum_plot=yes, prop_haz=no);

      *proportional hazards practice;
%surv_all(dat= thesis.bmt,status=0, grp= g,time= t2,indicator= d3,
survplot=no, cumhazplot=no, logrank=no, mean=no, median=no,
comp_risk=no,
      cum_plot=no, prop_haz=yes, prop_haz_bygroup=yes, cov_list=z1 z3
z5 z9, ties=breslow, selection=stepwise);
```

CURRICULUM VITA

NAME: Derek Duane Childers

ADDRESS: Department of Biostatistics
485 East Gray Street
University of Louisville
Louisville, KY 40202

DOB: Lexington, KY – October 11, 1990

EDUCATION & TRAINING: B.S., Mathetmatics
University of Louisville
2009-2013

PUBLICATIONS: Co-Author for the cohort titled “Comparing Autonomic and Behavioral Changes in Individuals with Autism Spectrum Disorders Entering Black versus White Sensory Room” by Alok R Amraotkar, MD, MPH, MHA