Comparison of the Mean Survival Analysis with the Restricted Mean Survival Time in Clinical Trial Study

Fangya Tan
Harrisburg University

Abstract

In many clinical trial studies, the time to event duration or survival time is one the most important primary endpoint to measure the efficacy of the treatments. Conventionally, it is industry standard to apply log-rank test to Survival Analysis to estimate the median of the survival time and use the Kaplan Meier method to graph the results. The data structure must pass the Cox Proportional Hazard Ratio Test to apply survival analysis. In order to avoid the misleading or biased results while the data failed Cox PH model assumption, we use the Restricted Mean Survive Time (RMST) as an alternative method to calculate the survival tim. In this paper, I used the Haberman's Survival dataset to perform Cox Proportional Hazard Ratio Test by Age Group (<= 60, or >60) and by Positive Axillary Nodes Group (<5 , or >= 5). Then I perform Survival Analysis and Restrict Mean Survival analysis to explore the relationship between the variable, Survival Time Length and the variable Age Group and Positive Axillary Nodes Group. Finally, I compared and interpreted the statistical descriptive results from Survival Analysis and RMST and discussed the underlying reasons for similarities and possible future works.

Keywords: Survival Analysis, Restricted Mean Survival Analysis, Kaplan Meier, Cox PH model, Haberman Survival Dataset

Literature Review

Survival analysis is a set of statistical methods to study the duration of data, how long until an event occurs. For example, how long patients will survive after being diagnosed with cancer? How long will patient recover from tuberculous after injecting the medication? Although the median of survival time is the most critical measure for survival analysis, depends on the data structure, there are advantages and constraints for various survival analysis methods, such as Mean survival analysis, Restricted Mean Survival Time (RMST), Kaplan Meier Survival Curve, and Cox Proportional Hazard Regression. Since applying survival analysis to inappropriate datasets would lead to bias mean survival estimation and interpretations, we would like to investigate what the best method is to perform based on the data frame to optimize the estimation result.

We begin with a question what the mathematical concept for the distribution of is survive times in survival analysis. Fox, J. (2008) focused on demonstrating the survival modeling examines the relationship between survival and one or more predictors, which examine entails the specification of a linear-like model for the log hazard. Furthermore, the author justified how Cox Proportional Hazards(PH) Model fits the concept by showing the baseline hazard function a(t)=logh0(t) unspecified, then Cox model is: $hi(t) = h0(t)\mathrm{exp}(b1xi1 + b2xi2 + \cdots + bkxik)$, which is the linear predictor $ni = b1xi1 + b2xi2 \ldots..bkxik$. The fancy point about this model is that regardless of the info of a specified baseline hazard, the Cox model can still be estimated by the method of partial likelihood as long as the data satisfies the proportional assumptions.

Beyond the basic mathematical concept, we are also interested in the importance of survival analysis in real world applications. Sedgwick,P.(2011) presented the median survival time as a primary measurement in the investigation of the effects of palliative chemotherapy in patients with locally advance or metastatic colorectal cancer. This study conducted seven trials with a common study period of 24 months on a total of 866 patients. The outcome measures the length of time until death after starting treatment. The study result demonstrated the survival time to be about 8 months in the supportive care group and 11.7 months in the chemotherapy group. The median survival time for each treatment group is the length of time corresponding to the probability of 0.5. Based on the median survival time from the survival analysis, we can conclude that there is a significant improvement of life expansion in the chemotherapy group from supportive care group with loss of generality

Clark,TG. (2003)'s paper further demonstrated the substantial of survival analysis in oncology studies by showing the main outcome is the assessment of survival time, the duration from finding to an event of interest. The major challenge with survival data is that the data structure is generally not Normally Distributed. It is skewed and comprises typically many early events and relatively few late ones. It is usually caused by some individuals have not had the event of interest, leading to their true time to event to be unknown. This paper introduced three general methods to perform survival analysis: Kaplan-Meier (KM) plots, log-rank test, and Cox PH regression. KM method is best for visualization purpose. It can show the survival curves and is generally used for directly describing the survival experience of a study cohort. The log rank test may be used to test for differences between survival curves for groups, such as study drug and placebo. We can check the p-value to compare the results for these groups. If p-value <0.05, this represents the study drug group to have a significant difference with placebo group. The Cox PH function is used to predict the instantaneous potential of having an event at a specific time, given survival up to that time. In consideration of the strong relationship between Cox PH model and survival analysis, Clark,TG. (2003) discussed in detail how to apply various statistical models to survival function and Cox PH model to analyze and summarize the survival data. How to estimate the effect of one or more factors that may

predict survival. If we define the survival time as the time between a fixed starting point (e.g. diagnosis of cancer) and a terminating event (e.g death), and focus on finding a reasonable approach to estimate the proportion of individuals to remain alive and disease-free at the end of the follow-up period. The authors provided Cox-PH model and

Accelerated Failure Time(AFT) model and concluded that if the given data fits the proportional hazard assumption the Cox model offers greater flexibility. Otherwise, if we know the direct estimation of baseline hazard function, then AFT is a better approach.

Benitez-parejo,N(2011) and Goel, M (2010) also stated that survival time, time elapse between a given exposure and the outcome of a certain event to occur, is an important term in clinical trials data. These articles described that the analysis of survival time study from Kaplan-Meier estimation or Cox regression models is more applicable than Accelerated Failure Time(AFT) model in reality. However, there are limitation to KM and Cox PH model. When we perform Kaplan Meier to measure the period of subjects living after treatment, Goel, M (2010) argued that study subjects could affect the estimated mean survival time intentionally or unintentionally. For instance, some subjects are uncooperative to remain in the study while other subjects never experience the event or death before the end of study or have moved out of the area and cannot continue to contribute data in the follow-up period. Nonetheless, KM estimate is the simplest way of computing survival over time despite all these difficulties associated with subjects or situations. KM can provide the survival probabilities for two group subjects as well as their statistical difference in survivals. KM plot is famous for being visually friendly. The survival curves clearly illustrate the censor, death information, survival probability and survival length information. In addition, we can customize the graph to show

different treatment groups with corresponding colors. Benitez-parejo,N(2011) provided the detailed example how the KM estimator and Cox PH Regression are calculated and validated the result with the software R.

As the previous paragraph suggested the Kaplan-Meier curve estimation is one of the most common and simple technique dealing with studies involving survival times. Rich,J.(2014)

explained how Kaplan-Meier curves are generated and analyzed, and discussed in detail how KM estimates in the contest of “survival” before the event of interest. The survival probability for each period is obtained by the fraction of objects at risk over surviving objects. Object at risk is defined as objects have not experienced an interest of event, e.g. patients have not died in cancer studies. Survival probability at any point is calculated as the product of preceding probabilities. The author also described how to interpret the result generated by log rank test for chi-square to compare the full curves of each group. If P-value > 0.05, it represents the two curves not to be statistically significantly different. As the researchers mentioned frequently, if we are only looking at the outcome from the KM curves respectively, it may easily lead to biased result. The author suggested that a better way is to compare the result of hazard ratio with KM curves, in which case the result will be more promising.

The natural of survival data structure skewed from normal distribution is one of the most trivial and essential reasons for KM estimation might lead to biased results. We generally obtain the estimated survival time by calculating the area under the Kaplan-Meier curve, whereas some patients may never experiences an interest of event, which leads to an underestimation of the mean survival time. Under this circumstance, we would like to investigate what the confidence interval for the KM survival estimation is, or whether there are adjustments in the model to reduce the tolerance. Zhong, M.(2009), suggested an approach for modifying the nonparametric estimate mean survival time by setting the largest observation to an event time if it is censored. The study simulated data from seven distributions: exponential, normal, uniform, log-normal, gamma, log-logistic, and Weibull to compare the results of the estimates to the known true values. By conducting this simulation, the authors found a bias of the modified KM mean estimator. Furthermore, it increases with the proportion of censoring. For example, log-logistic, log normal, exponential increased the quickest. The other distributions are relatively unbiased until around 60% censoring. From these findings, the authors concluded that the mean estimator depends heavily on the nature of the distribution. We must be very careful while using the estimate mean from KM with 30% or more censoring data.

Theoretically, for right censored survival data, it is easy to estimate mean survival time using KM method when the objects of the censoring time contains the objects of survival time. We have no doubt about the advantages and power of Kaplan Meier curve estimation in survival analysis. Its application is ubiquitous in clinical trial studies. Yet, we would like to explore other statistical models for more accurate estimations in the scenarios of sophisticated data structures. Ying, d. (2015) showed in practice that it is an easily violated the assumption that all objects are contained in censoring time. Because the follow-up of a study is most likely within a finite window. When the object survives beyond or not censored within the final cut-off time then the estimated time might lead to a biased result. The Authors found an alternative solution by showing that the mean survival time is still estimable from a linear model when object of some covariate with nonzero coefficient is unbounded regardless of the length of follow-up. The paper shows that, when the conditions of the linear model are stated correctly, the reasonable mean square prediction errors outperform the Cox model in the condition of heavy censoring and short follow-up time.

To find a more robust solution to estimate the survival time for incomplete dataset, Langova, K. (2008) focused on their research in developing the method to mine a maximum of relevant information from studies finished at a time when the collected data has not been completed. As we mentioned before, the study of the survival time or length of life of cancer patients. Because generally most studies on survival analysis are terminated before the observed event(end-point), such as death, occurs for all subjects. The author compared the mean and median using survival function and Cox PH model and concluded that Cox PH model is a better approach. Because there is a danger of statistical misinterpretation of data when directly applying standard statistical procedures or Kaplan Meier Curve with incomplete data. The trivial result is mean or median of respecting data would underestimate the reality.

After reviewing and comparing these common statistical models to analyze survival data, Go, T. (1997), showed the parametric methods to have less constrains for the data. Whereas if the assumptions for nonparametric methods hold, the resulting estimates have smaller standard errors and are easier to interpret. Consequently, we are more confident with the estimations. Nonparametric techniques include the Kaplan-Meier method for estimating the survival function and the Cox proportional hazards model to identify risk factors and to obtain adjusted risk ratios. When the assumption of proportional hazards is not tenable, the authors suggested a solution to stratify the data to fit a model with different baseline functions in each stratum. From this paper, we can conclude that the Cox proportional hazards model is the most robust

method for estimating mean survival time in survival analysis.

Zhu, X.(2017) and his colleagues' research in 112 studies further demonstrated the popularity of Cox PH model usage. They illustrated that Cox models is one of the most popular regressions in survival analysis out of a variety of statistical forms. Cox PH model is extremely powerful because it only assumes the proportional hazards ratio for different groups while no restrictions on the baseline hazard rate. If the data satisfy this condition, it can analyze the outcome of a lifetime and estimate the mean survival time with independent variables and variables. However, violating or neglecting assumptions of Proportional hazard for the Cox model, or lack of supplicated statistical software to check the PH hypothesis might lead to potentially inaccurate, biased, or even erroneous estimations and conclusions.

As powerful as the Cox PH model to estimate the survival times, all the research papers showed that it is crucial for the dataset to meet the proportional hazard ratio requirement when we use it. However, with the development of clinical trials, there are more and more therapies that are novel to have distinct mechanisms compared to conventional treatments. Many of them have departed from the proportional hazard assumption with a time-to-event end point. In those situations, the restricted mean survival time (RMST) is an alternative solution which is robust and is a clinically interpretable summary measure that does not rely on PH assumption. The survival probability at a specific time point, say t, does not transparently capture the temporal profile of this endpoint up to t. Geographically, RMST is the area under the survival curve up to t. Zhao, L.(2014), illustrated the use of RMST through two clinical trials, one from oncology and the other one from cardiology. Based on the findings, the authors are able to achieve the conclusion that the RMST method provides some clinically meaningful interpretations for the difference between two treatment groups in a time scale under an equivalence setting.

In addition, Huang,b.(2018) demonstrated RMST to be the best alternative solution by conducting extensive simulations to evaluate the RMST-based inference against the PH-based inference. While PH assumption is satisfied, the RMST has a similar result as the Kaplan-Meier survival curves. In non-PH scenarios, the RMST-based test has better performance than the log-rank test. This paper concluded by selecting the minimum and maximum time point for RMST will produce a better result than truncating time in clinical settings where there is suspicion of substantial violation of the PH assumption.

To support the argument, RMST is an improvement method for regular Mean survival analysis. Calkins, K.(2018) applied restricted mean survival time (RMST) to estimate the time- to-event analysis on antiretroviral therapy for HIV-infected persons who are injected drugs(PWID) and persons who do not receive the treatment. The article showed the advantage of RMST in contrast to regular mean survival time by showing mean survival curves going to zero during the observation time. Due to the data collection process, some persons are still alive during data collecting. Therefore, “restricted” component of mean survival avoids extrapolating the integration beyond the last observed time point by calculating the difference in the area under two survival curves A1(t)-A2(t). Although few epidemiological studies utilize RMST, the article argued the following advantages of applying the RMST: It 1) does not require the assumption of proportional hazards, 2) can summarize the difference in survival when survival curves initially diverge and later converge, and 3) provides information about absolute risk.

Kim,D.(2017)’s research paper on Cardiovascular therapeutics is another case utilizing the RMST with its dominant feature, which does not need to meet the Cox PH model. It is well- known that cardiovascular therapeutic areas have been criticized for being underpenetrated considering older adults with frailty and multiple chronic conditions. How to interpret the treatment effect with limited life expectancy is one of the most important challenges in older adults’ cardiovascular studies. In 2010, restricted mean survival time (RMST) was proposed to measure the treatment effect that offers advantages in design, analysis, and interpretation over the conventional measures. The authors explain how different measures of treatment effect are interpreted by evidence-based communication using 5-year follow-up data from the placement of Aortic Transcatheter Aortic Transcatheter Valves A, B trials, and conventional medications. In this study, the hazard for treatment A in the first 30 days was greater than patients in treatment B, hence the hazard rate is not proportional. Researchers reached a 95% Confidence Interval for that the median survival time of treatment TAVR was 19 months longer than medications by performing survival analysis using the RMST method.

In classical regression analysis of quantitative, outcomes focused on mean value regression models for survival data are often specified from the hazard function. While RMST is a well-established method to estimate the median of survival time, in a clinical industry, few studies have utilized it. Andersen, P.(2004), conducted Monte Carlo simulations for pseudo-observations and two real data sets for regression analysis of mean survival time and related quantity. The restricted mean survival time is reviewed and compared to a method

based on pseudo-observations. From the result, the authors concluded for two real data set, existing regression methods for analysis of the mean is a good approach. Moreover, RMST is theoretically a superior model for calculating mean survival time based on the pseudo-observations. The advantage of this RMST method is based on a simple non-parametric estimator for the mean obtained as the integrated KM estimator.

RMST as an expansion of conventional survival analysis with Cox PH model, researchers and scientists have never stopped seeking for more advanced techniques and statistical models to achieve a higher accuracy in estimate mean survival time, as well as developing various processes to avoid the violation of proportional hazard ratio assumption. Fernandez, T. (2016), introduced a semi-parametric Bayesian model for survival analysis. The model is centered on a parametric baseline hazard and uses a Gaussian process to model variations away from it nonparametrically, along with dependence on covariates. The special feature of this framework is that it does not impose unnecessary constraints in the hazard rate or in the survival function. The report is based on the experimental results on synthetic and real data, showing that the model performs better than the competing model such as Cox proportional hazards, ANOVA-DDP, and random survival forests. Further work consists in increasing a method to choose the initial parameter to avoid sensitivity problems in the beginning.

Likewise, Dehbhi, H. (2017) presented two contemporary mechanisms to calculate the estimate mean, life expectancy difference (LED), and Life expectancy ratio(LER). The authors are aware of the fact that when survival curves cross over of separate only after a considerable time, then the Cox model assumption is violated which HR result might lead to a biased result, although hazard ratio (HR) is the most frequently used measurement for time-to-event outcomes. LED and LER models can be used for evaluating the treatment effect in the situation when HR changes over time. LED, the life expectancy difference, represents the difference between mean survival times in intervention and control arms. LER, Life expectancy ratio, expresses the ratio of intervention and control arms survival times. Both measures can be interpreted as absolute and relative gains or losses in life expectancy due to an event of interest. The unique advantage for those methods is that they can be estimated for any survival curves regardless of the shapes.

Dehbhi, H. (2017) also provided a detailed application of LED in a randomized study of 370 patients to compare the effect of rituximab and standard chemotherapy with untreated mantle cell lymphoma. When performing the Grambsch-Therneau test for Non-proportional hazards to the treatment groups, the result was statistically significant (p=0.025). Since the data failed the Cox PH test, when applying the regular mean survival analysis, it shows the difference in median survival to be 7.5 months for two groups, whereas LED demonstrates a difference of 17.7 months in two treatment groups. From the result of this study, we can conclude that LED is a much more appropriate method than regular mean survival in this situation.

Up to this point, we have been focusing on comparing two methods at the same time, such as the KM curve vs. Cox PH model, RMST vs. regular mean survival, or LED vs. regular mean survival. Kasza, J. (2014) reviewed five different techniques for the analysis of survival data in respiratory studies, such as Kaplan-Meier survival plot, the Cox PH model, parametric PH models, accelerated failure time models(AFT), and Restricted mean survival time. The outcome of interest is the time to death for lung cancer patients or the time to recovery for pneumonia patients. The authors used 137 patients in Veterans Administration lunch cancel trial data through the research, they concluded that although the Kaplan-Meir approach is the most straightforward approach, Cox Proportional hazards models are the most popular choice for modeling survival data. RMST is a novel measurement to simulate mean survival time.

Methods

I conducted a Survival Analysis on the survival time length from the breast cancer patients who survived in the operation in the Haberman’s Survival Data Set (Lim.T, 1999). My research analyzed the whether the continuous variable, Survival Time length, is affected by the categorical variable, Age Group, or Positive Axillary Nodes Group. If so, what is the association degree between Age Group and Survival Time length, and Positive Axillary Nodes Group between Survival Time Length. I also applied the Restricted Mean Analysis to check if the findings are consistent with Survival Analysis.

My Survival Analysis on the Haberman’s Survival Data is a mixed analysis. It investigated the relationship between the continuous variable survival time length from the breast cancer and the patients operating Age Group or the patients Positive Axillary Nodes.

Participants

The participants are 306 survival patients who had undergone surgery of breast cancer in Haberman's Survival Data Set from Kaggle (Lim.T, 1999). The dataset contains cases from a study that was conducted between 1958 and 1970 at the University of Chicago's Billings Hospital on breast cancer patients' survival status. This original dataset included four variables. It includes the continuous variable, the age, in unit of years, of patients performed the operation and the year of operation. It also includes the continuous variable, the number of positive axillary nodes detected. In addition, it includes the categorical variable survival status, alive or dead. I created three new variables in order to perform my survival analysis, Survival Time Length, Age Group and Positive Axillary Nodes Group. There is no information about patients' ethnicity and sex.

Due to the nature of breast cancer, I can assume with a high confidence interval the ratio of female and male participants would be extremely high. To avoid any potential biased result leading by this un-equal sex rate participants, I will not consider sex as a factor of survival length. I will categorize the participants by age and number of positive axillary nodes detected. Moreover, the fundamental concept of Survival analysis is investigating the association between patients' death or live status with other variables, therefore I would not do any imputation if there are missing values in the dataset.

The inclusion criteria for participates in the Haberman Breast Cancer Study included the patient must have undergone surgery for breast in University of Chicago's Billings Hospital and the age is between 30 to 83 years old.

Procedure

I acquired the Haberman's Survival Data Set from Kaggle (Lim.T, 1999). The dataset is originally collected from a study at the University of Chicago's Billings Hospital on the survival of patients who had undergone surgery for breast cancer between 1958 and 1970. It includes four variables for 306 patients between 30 to 83 years old when the operation was performed, Age of patient at the time of operation, Patient's year of operation, Number of positive axillary nodes detected and Survival Status. There is no missing value in this dataset.

My research design started by creating two new categorical variables and one continuous variable: Age Group, Positive Axillary Nodes Group and Survival Time Length from the Operation. Then Finally, I will apply the Restricted Mean Analysis by adding a limit to the survival time to check the I can achieve the same result.

To create the Age Group variable, I split the age into two age groups: patients are smaller than 60 years old and patients greater than 60 years old. I also split the Positive Axillary Nodes into two groups: smaller or equal to 10 nodes or greater than 10 nodes. For the continuous variable Survival Time Length, I use the year of the end of study 1970 minus the year of operation for the patients are still alive. Then I will perform two individual Proportional Hazard Ratio Assumption for a Cox regression model in R respect to Age Group and Axillary Nodes Group to test if the data structure meets the proportional hazard ratio assumption. After confirming the data structure is suitable for Survival Analysis, I will apply Survival fit and Survival Plot packages in R(3.3.4, R cores Team, 2018) to present the Survival Analysis summary and Kaplan Meier Curve respectively. If the P-value from the Survival Analysis summary is smaller than 0.05, I conclude the Age Group or Positive Axillary Nodes Group have a significant influence with the patients' survival length depending on which group I tested

Measures

I will start the data processing process by running a statistical summary on the variable, Age of Operation. Based on the Median of age summary, I will divide the participants into two groups, Age Group 1 <=60, and Age Group 2> 60. There are 229 patients in Age Group 1 and 77 patients in Age Group 2. In addition, I will perform another summary, Number of positive axillary nodes detected. Then categorize patients into two groups based on the number of positive axillary nodes <=5. There are 236 patients have positive axillary nodes smaller or equal to 5 and 70 patients have positive axillary nodes >5. Then I will calculate the variable of Participants' Survival Year by using the end of study year (1970) – Operation Year +1. I will run the proportionality test by the Cox PH model twice of the patients by Age Group, and Node group, and check they pass the proportionality test.

Analysis

I use R (3.3.4, R cores Team, 2018) and R-packages *A Package for Survival Analysis in S* (Therneau. T,2015) for my analysis. I will use the Kaplan Meier Method to graph the survival time of the breast cancer survival patients by Age Group and Node Group. If I can observe one group's median survival length is out perform than the other visually, then I will continue investigating the dataset by applying the Mean survival analysis and calculate the P-value from Log-Rank test. If the P-value is smaller than 0.05 then I conclude the Operation Age/Number of Positive Node has clinical significance in survival time of breast cancer. I will perform the same procedure for Restricted Mean Survival analysis and compare the results.

Result

In initial purpose of this research paper was conducting a survival analysis to investigate the relationship between the survival time length and the variable Age Group and Positive Axillary Nodes Group in Haberman's Survival Dataset. Then I applied the Restricted Mean Survival analysis to check if my findings are consistent. In this section I would share my findings in two parts, survival analysis and restricted mean survival analysis respect to Age Group, and survival analysis and restricted mean survival analysis respect to Positive Axillary Nodes Group.

In regards the Survival Analysis and Restricted Means Survival Analysis, I choose 1980 minus the operation year for survival patients greater than 5 years, and 1971 minus the operation year for the survival patients' survival status less than 5 years for the time limit of Survival Analysis. Meanwhile, I choose 1985 minus the operation year for survival patients greater than 5 years, and 1973 minus the operation year for the survival patients' survival status less than 5 years for the time limit of Restricted Mean Survival Analysis.

Case I: Survival Time respect to Age Group in Survival Analysis and RMST

In order to investigate the relationship between survival time after the breast cancer operation year is associated with age, I divided the age into two groups: Greater than 60 or Smaller to equal to 60 years old when the participates had undergone the breast operation. I used Coxph function in R Survival Package to check my first hypothesis is satisfied, the data structure proportionally distributed. Moreover, From Table 1 we can see the Hazard Ratio, Exp(coef) = 1.098 means there is no effects of survival length after operation between Age Groups. The P-value=0.709 also suggested there is no significant difference between Age smaller or equal to 60 years old or Age greater than 60 years older. To further support the statistical analysis provided the readers two group has no significant difference in survival time, the Kaplan Meier survival plot shows the survival time is about 74% for Age Group <= 60 years old, and about 70% for Age Group >60. Although from Table 1 and Graph 1 demonstrated there are no visible effect for Survival time length and Age Group, I apply the Restricted Mean Survival Analysis, the P-value= 0.726 and Hazard Ratio = 1.092 , which means the results are consistent with my previous finding. Table2 and Graph2 shows the Restricted Mean Survival Analysis result and Kaplan-Meier Plots

Case II: Survival Time respect to Positive Node Group in Survival Analysis and RMST

To investigate the relationship between survival time after the breast cancer operation year is associated with Node groups, I divided the patients have less than 5 nodes to Node Low Group, and patients have more than 5 nodes to Node Hight Group. The summary of Coxph function in R Survival Package in Table 3 to show me the data structure is proportionally distributed which satisfy the survival analysis hypothesis. Moreover, the statistical P-value= 2.52e-07, I conclude the variable Node Group has highly statistically significance. The Regression Coefficients, Low=-1.16, the negative sign indicates that the Node Low Group have lower risk of death than Node High Group. The Hazard Ratio, Exp(coef) = 0.31 means being in Low Node Group reduces the hazard by a factor of 0.31. The overall 95% confidence interval of all the statistical results is (0.20 , 0.487). Finally, the Kaplan Meier Curve showed (graph 2) the consistency with the statistical results. Table 3, the Survival probability for patients in Node Low Group is about 80%, whereas the probably is about 50% for patients in Node High Group. Table 4 and Graph 4 shows the result from Restricted Means Survival Analysis, we can see P-value=2.98e-07 and Hazard Ratio= 0.3158; this indicates the Positive Nodes Group is statistical significant for Survival time and if a patient in Positive Nodes Low Group then the risk of death will decrease, which is consistent with our findings.

Discussion

In this research paper, I have worked on Haberman's Survival Data Set to compare the statistical results from Haberman's Survival dataset in the Survival analysis and Restricted Mean Survival analysis, and explored the relationship between the patients' survival time length with the Age Group and Positive Axillary Nodes Group. I have concluded not only from the descriptive statistical outputs from the Cox PH Regression Model but also the visualization of data from the Kaplan-Meier Survival Probability Plots, the Survival time for patients' undergone a breast cancer is highly associated with the Positive Nodes (P-value= 2.52e-07), but not affected by the age of operation. More specifically, if a patient has less than 5 Positive Axillary Node when she or he undergone the breast surgery, then the 5-year death risk of her/him would decrease a factor by 0.31 compared to the patient has 5 or more Positive Axillary Node. In addition, if the data structure passes the Cox PH model assumption then the Survival Analysis and Restricted Mean Survival Analysis generated very similar statistical outputs.

Although restricted by the sample data size, from this study it is noted with a great concern that the number Positive Axillary Nodes has crucial impact on the 5-year survival rate of Breast Cancer while undergo a surgery. We observed the 5 year Overall Survival Rate for patients have less than 5 Positive Axillary Nodes breast cancer about 80%, therefore surgery is a relative efficient choice for those patients; whereas patients have more positive Axillary Nodes would take more risk for surgery because the 5 year ORR dropped to 50%. As the rapid development of novice clinical trials drugs in recent years, there are many new cancer medicines designed special for breast cancer patients have too many positive nodes or have Triple Negative breast cancer.

Comparing to previous research Zhu, X.(2017) and his colleagues showed the popularity and Cox PH model and if the data neglecting assumptions of Proportionality would cause potential biased estimations. Zhu, X.(2017) only emphasized the PH problem, whereas my research provided a solution, if the data does not meet the PH assumption, then we can use Restricted Mean Survival Analysis to estimate the survival time. In addition, Zhao, L.(2014) illustrated RMST provides some clinically meaningful interpretations in cardiology. In my research I demonstrated why RMST is an alternative solution for survival analysis, because the statistical outputs are extremely similar from these two methods.

My original intention was to compare the survival analysis and Restricted Mean Survival Analysis on an ongoing clinical data trial or a dataset does not satisfy the Cox Proportional Hazard Ratio assumption; I would expect a more visible improvement from the statistical results and Kaplan Meier Survival Plot in these cases. However, the Haberman’s Survival Dataset is completely collected by 1990 and fits the Cox Model, so there is no ongoing data that might leads to an underestimation of median survive time nor unproportionable data structure cause potential biased statistical results. Therefore, we did not see much difference of the statistical outputs in Survival analysis and Restricted Mean Survival Analysis. In other words, Survival analysis and RMST generated extremely close statistical outputs, which showed our hypotheses, when the survival time between the control group and placebo group are proportional, then mean survival analysis is a more accurate method to estimate survival time than RMST. Otherwise, RMST is the best alternative to estimate the survival time.

Among the results, I found the Age group did not have a statistical significance in survival length very interesting; it is intuitive that a patient would higher risk of death when he/she undergone a cancer surgery in older age than younger. This counter-intuitive finding might due to the fact that there is no accurate survival length information from the Haberman's Survival dataset, I only considered the end of study in this study. In future researches, we can focus on ongoing cancer study datasets that provides patients detailed survival information.

In summary of this research, I applied Cox PH model on Haberman's Survival Data Set and compared the descriptive statistical results for Survival analysis and Restricted Mean Survival analysis. From the both survival technique results and Kaplan Meier plots I concluded the patients' survival time is highly associated with the Number of Positive Nodes. I also reached the general conclusion, when the survival time between the control group and placebo group are proportional, then mean survival analysis and RMST generate very similar outputs. Otherwise, RMST is the best alternative to estimate the survival time.

References

Andersen, P. , Hansen, M. , & Klein, J. , (2004) Liftime Data Analysis,

*Regression analysis of Restricted Mean Survival Time Based on Pseudo-Obsevations.*10,335-350

Benitez-parejo,N. Rodriguez del Augila, M.M, & Perez-Vicente,S.(2011). Allergol Immunopathol(MAdr).

Survival analysis and Cox regression

Calkins, K. , Canan.C, Moore, R. , Lesko.C & Lau,B. (2018) BMC Medical Research Methodology

*An application of restricted mean survival time in a competing risks setting: comparing time to ART initiation by injection drug use.*18-27

Clark, TG. , Brandurn, MJ., Love, SB., & Altman, DG. (2003), British Journal of Cancer.

Survival Analysis Part I: Basic concepts and first analyses, 89(2): 232-238

Clark, TG. , Brandurn, MJ., Love, SB., & Altman, DG. (2003), British Journal of Cancer.

Survival Analysis Part I: Basic concepts and first analyses, 89(3): 431-436

Dehbhi, H. , Royston, P., & Hackshaw.A. (2017), Research Methods & Reporting,

Life expectancy difference and life expectancy ratio: two measures of treatment effects in randomized trials with non-proportional hazards, 357:j2250

Fernandez,T., Rivera,N., & Teh,Y., (2016), Cornell University Library Archive, *Gaussian Processes for Survival Analysis,* arXiv:1611.00817

Fox, J. (2008) , Cox proportional-Hazards Regression for Survival Data,

Appendix to An R and S-Plus Companion to Applied Regression, p1-13

Go, T., & Lee, E., (1997), Annual Review of Public Health,

*Survival Analysis in Public Health Research,*18(1): 105-34

Goel, M., Khanna, P., & Kishore.P(2010), International Journal of Ayurved Research,

Understanding survival analysis: Kaplan-Meier estimate, 1(4):274-278

Huang, .b, & Kuan, pf. (2018), Pharm Stat.

Comparison of the restricted mean survival time with the hazard ratio in superiority trials with a time-to-event end point. 17(3):202-213

Kasza, J., Wraith, D., Lamb, K., & Wolfe, R.(2014), Respirology.

*Survival analysis of time-to-event data in respiratory health research studies,*19(4): 483-492

Kim, D., Uno H, &Wei, L., (2017), Jama Cardiology,

Restricted Mean Survival Time as a Measure to Interpret Clinical Trial Results.

2(11):1179-1180.

Langova, K. (2008), Biomed Pap Med Fac Uni Palacky Olomouc Czech Repub.

Survival Analysis for Clinical studies, 152(2): 303-307

Lim, T., 1999, *Haberman's Survival Data Set*, Retrieved from

https://www.kaggle.com/gilsousa/habermans-survival-data

set

R cores Team, (2017). A language and environment for statistical computing.

Rich,J. , Neely, J., Paniello, R., Voelker, C., Nussenbaum, B., & Wang,E., (2014)

Otolaryngol Head Neck Surg,

A practical guide to understanding Kaplan-Meier Curves. 143(3):331-

336 Sedgwick, P. , & Joekes, K., (2011), British Medical Journal,

Survival(time to event) data: median survival times. 343: d4890

Therneau. T, (2015), _*A Package for Survival Analysis in S*_. version

2.38 Ying, d, & Nan, b. (2015). Scand Stat Theory Appl.

Estimating mean survival time: when is it possible? 42(2): 397-413

Zhao, L., Claggett, B., Tian, L. , Uno H., Pfeffer, M., Solomon.S, Trippa.L, Wei, L.(2014),

The Berkeley Ecteronic Press,

On the Restricted Mean Survival Time Curve Survival Analysis, paper 187

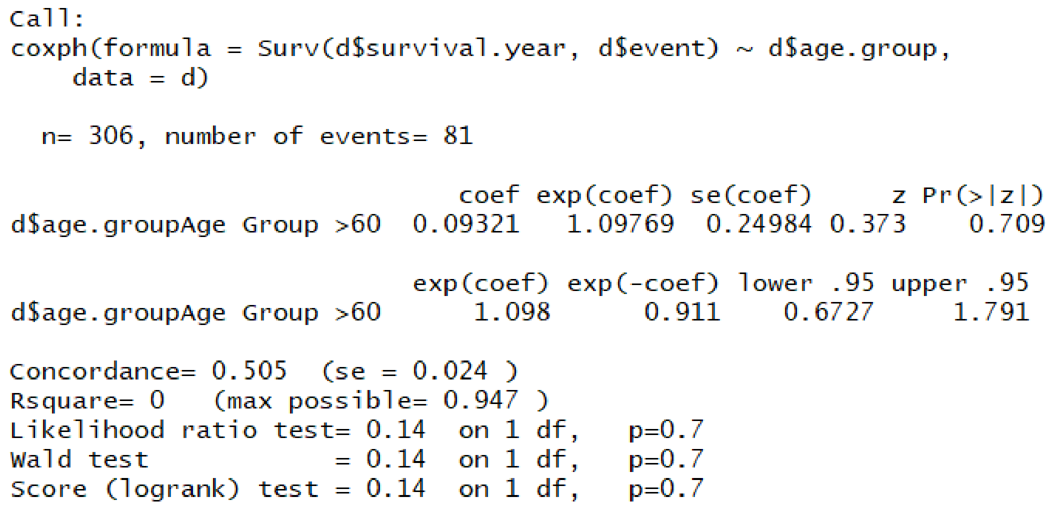
Zhong, M., & Hess, K. (2009), The Berkeley Electronic Press,

Mean Survival Time from Right Censored Data, paper 66

Zhu, X. Zhou.X , Zhang.Y , Sun.X, Liu.H & Zhang,Y (2017), Medicine(Baltimore).

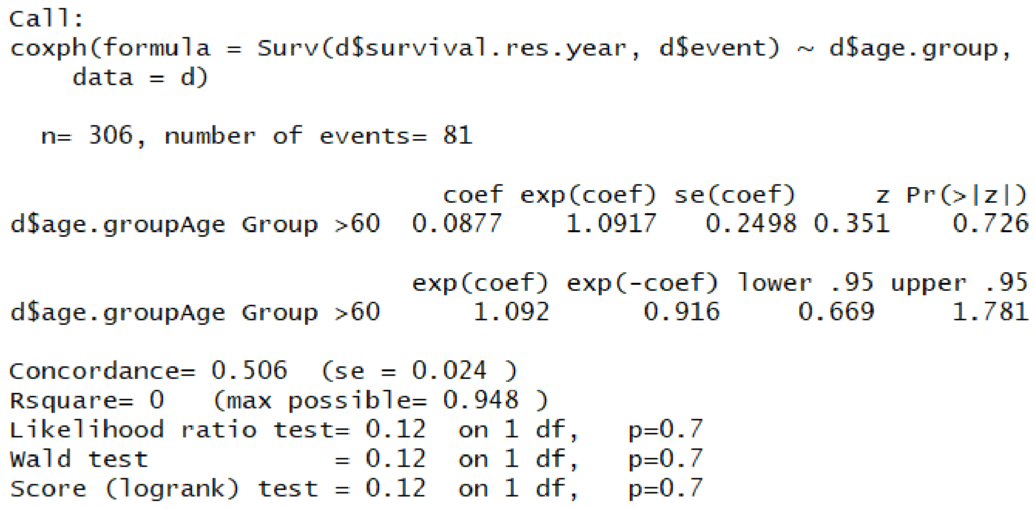
Report and methodological quality of survival analysis, 95(50) e9204

Table 1: Survival Analysis respect to Age Group



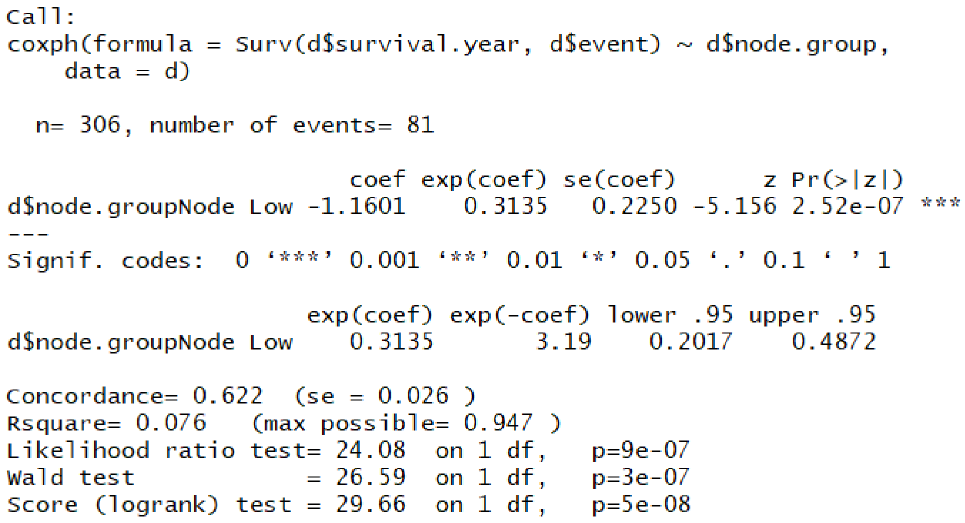
```
Call:
coxph(formula = Surv(d$survival.year, d$event) ~ d$age.group,
    data = d)

  n= 306, number of events= 81 

                             coef exp(coef) se(coef)     z Pr(>|z|)
d$age.groupAge Group >60  0.09321   1.09769  0.24984 0.373    0.709

                         exp(coef) exp(-coef) lower .95 upper .95
d$age.groupAge Group >60     1.098      0.911    0.6727     1.791

Concordance= 0.505  (se = 0.024 )
Rsquare= 0   (max possible= 0.947 )
Likelihood ratio test= 0.14  on 1 df,   p=0.7
Wald test            = 0.14  on 1 df,   p=0.7
Score (logrank) test = 0.14  on 1 df,   p=0.7
```

Table 2: Restricted Mean Survival Analysis respect to Age Group



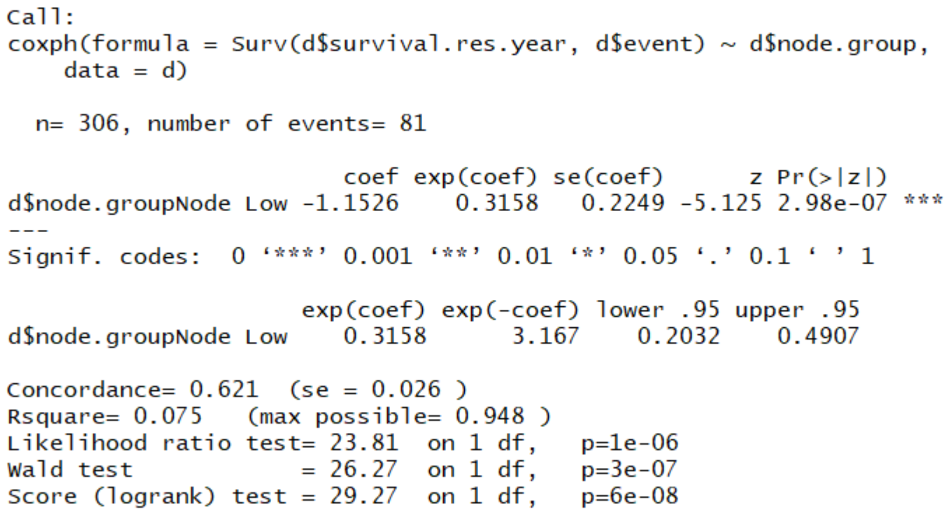
```
Call:
coxph(formula = Surv(d$survival.res.year, d$event) ~ d$age.group,
    data = d)

  n= 306, number of events= 81 

                            coef exp(coef) se(coef)     z Pr(>|z|)
d$age.groupAge Group >60  0.0877    1.0917   0.2498 0.351    0.726

                         exp(coef) exp(-coef) lower .95 upper .95
d$age.groupAge Group >60     1.092      0.916     0.669     1.781

Concordance= 0.506  (se = 0.024 )
Rsquare= 0   (max possible= 0.948 )
Likelihood ratio test= 0.12  on 1 df,   p=0.7
Wald test            = 0.12  on 1 df,   p=0.7
Score (logrank) test = 0.12  on 1 df,   p=0.7
```

Table 3: Survival Analysis Respect to Node Group



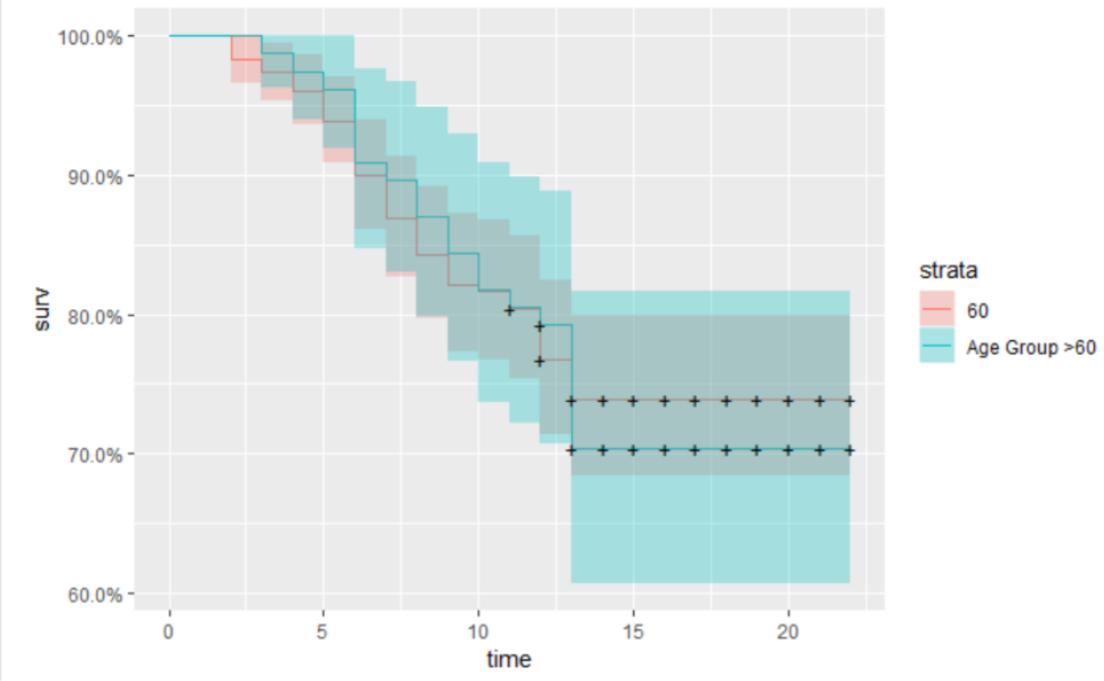
```
Call:
coxph(formula = Surv(d$survival.year, d$event) ~ d$node.group,
    data = d)

  n= 306, number of events= 81

                           coef exp(coef) se(coef)      z Pr(>|z|)
d$node.groupNode Low -1.1601    0.3135   0.2250 -5.156 2.52e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                     exp(coef) exp(-coef) lower .95 upper .95
d$node.groupNode Low    0.3135       3.19    0.2017    0.4872

Concordance= 0.622  (se = 0.026 )
Rsquare= 0.076   (max possible= 0.947 )
Likelihood ratio test= 24.08  on 1 df,   p=9e-07
Wald test            = 26.59  on 1 df,   p=3e-07
Score (logrank) test = 29.66  on 1 df,   p=5e-08
```

Table 4: Restricted Mean Analysis to Node Group

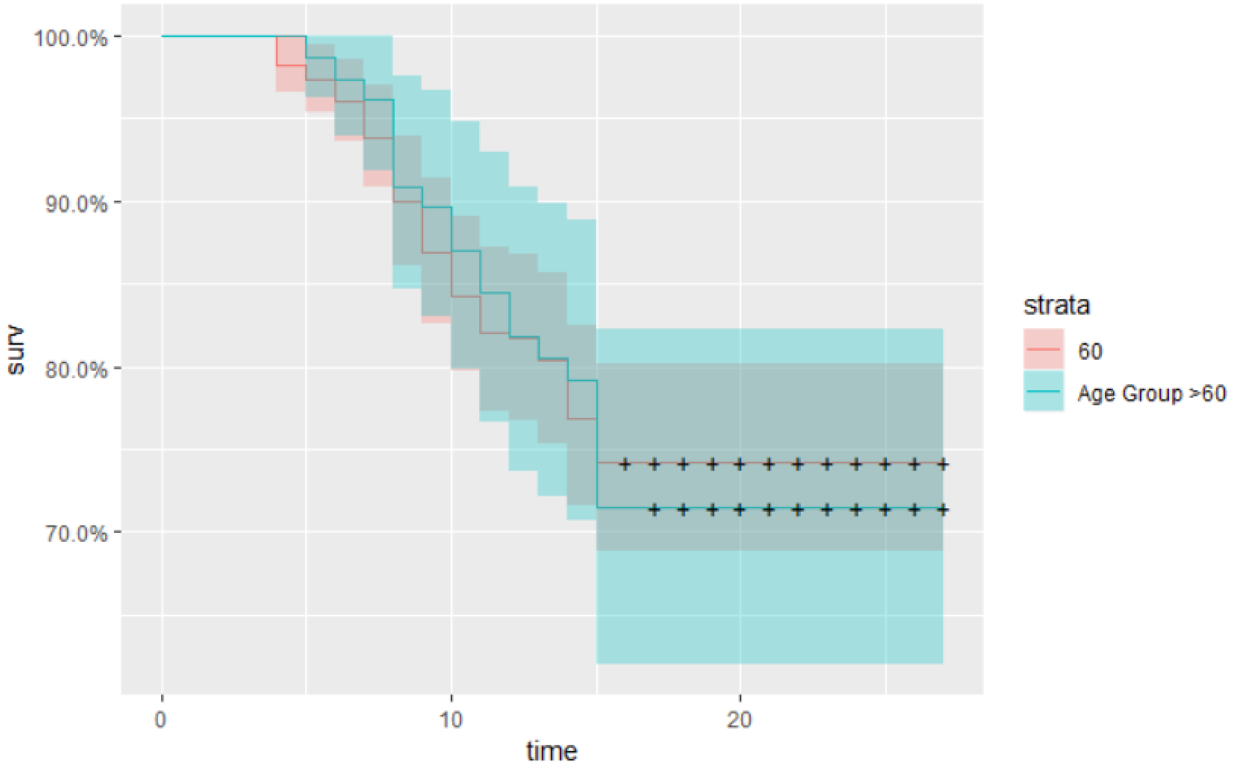


```
Call:
coxph(formula = Surv(d$survival.res.year, d$event) ~ d$node.group,
    data = d)

  n= 306, number of events= 81

                           coef exp(coef) se(coef)      z Pr(>|z|)
d$node.groupNode Low -1.1526    0.3158   0.2249 -5.125 2.98e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                     exp(coef) exp(-coef) lower .95 upper .95
d$node.groupNode Low    0.3158      3.167    0.2032    0.4907

Concordance= 0.621  (se = 0.026 )
Rsquare= 0.075   (max possible= 0.948 )
Likelihood ratio test= 23.81  on 1 df,   p=1e-06
Wald test            = 26.27  on 1 df,   p=3e-07
Score (logrank) test = 29.27  on 1 df,   p=6e-08
```

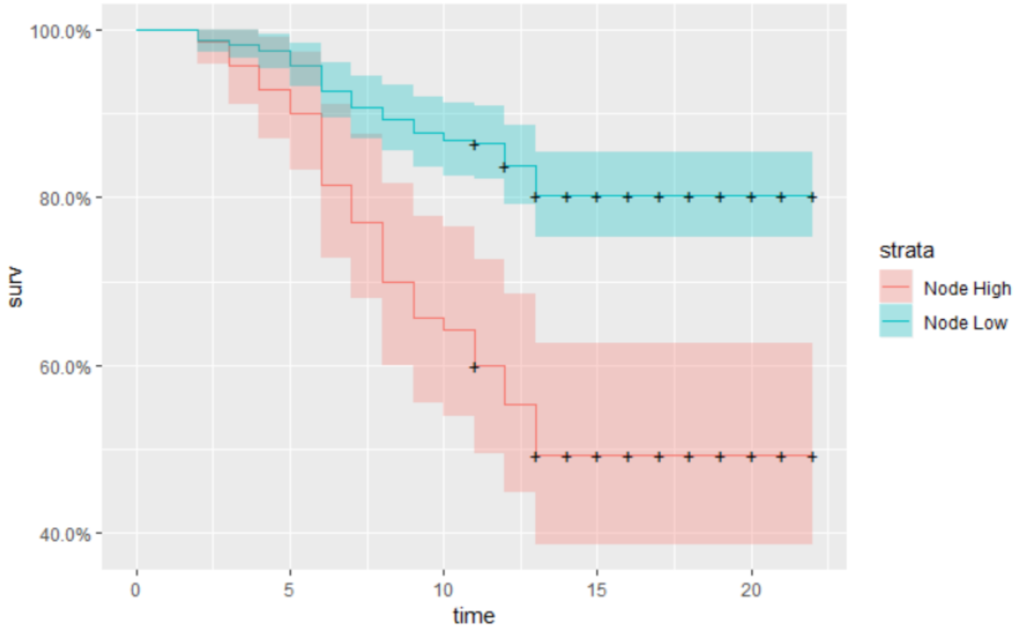
Graph1: Survival Analysis respect to Age Group



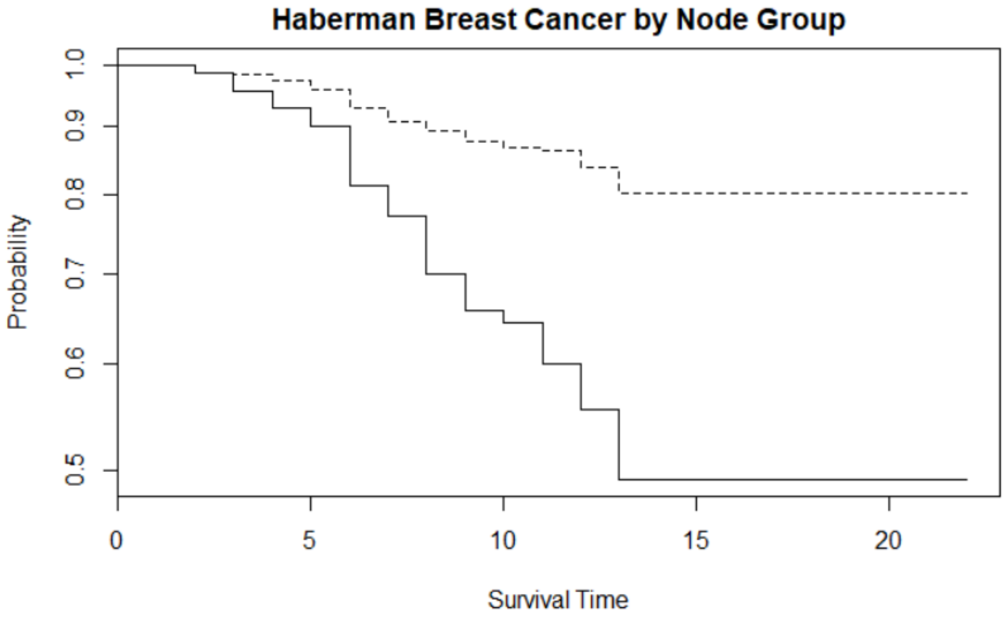
Graph 2: Restricted Mean Survival Analysis respect to Age Group



Graph 3: Survival Analysis Respect to Node Group



Graph 4: Restricted Mean Survival Analysis Respect to Node Group


Haberman Breast Cancer by Node Group
Probability
Survival Time