Survival Analysis of Young Leukemia Patients

Authored by: Theren Williams[1], Zachary Smith[2], Drew Seewald[3]
University of Michigan Dearborn, Department of Mathematics and Statistics

Faculty Advisors: Dr. Keshav P. Pokhrel[4], Dr. Taysseer Sharaf[5]

Abstract

With cancer as a leading cause of death in the United States, the study of its related data is imperative due to the potential patient benefits. This paper examines the Surveillance, Epidemiology, and End Results program (SEER) research data of reported cancer diagnoses from 1973-2014 for the incidence of leukemia in young (0-19 years) patients in the United States. The aim is to identify variables, such as prior cancers and treatment, with a unique impact on survival time and five-year survival probabilities using visualizations and different machine learning techniques. This goal culminated in building multiple models to predict the patient's hazard. The two most insightful models constructed were both neural networks. One network used discrete survival time as a covariate to predict one conditional hazard per patient. The prediction rate is nearly 95% for testing datasets. The other network built hazards for discrete time intervals without survival time as a covariate and predicted with lower accuracy, but captured variable effects from initial testing better.

Acknowledgment

Special thanks to UM-Dearborn ORSP and the UM-Dearborn Scholar's Grant
We would also like to thank the SIURO editor and the two anonymous referees.

[1] Undergraduate student, therenw@umich.edu
[2] Business Intelligence and Systems Coordinator, Ann Arbor Area Community Foundation, smithzac@umich.edu
[3] Graduate student, University of Wisconsin Stevens Point, dseew009@uwsp.edu
[4] Assistant Professor of Statistics, kpokhrel@umich.edu
[5] Assistant Professor of Statistics tsharaf@umich.edu

Introduction

Cancers are the source of nearly a quarter of all deaths in the United States and among them; leukemia is one of the leading causes of death for children. Our study aims to isolate features with a unique impact on survival, in addition to more accurately predicting the likelihood of death of leukemia patients from age 0 to 19, for whom leukemia is the most commonly occurring cancer type [1]. According to the Leukemia & Lymphoma Society (LLS) [2], the 5-year survival rates for leukemia patients have increased from 34% in the mid-1970s to approximately 63% from 2006 to 2012. The primary treatment options for leukemia patients are chemotherapy, radiation, or a combination of the two. However, since leukemia is a cancer of the blood cells, radiation is not typically recommended as often as chemotherapy since there is rarely a cancer cluster at which they can direct the beam radiation.

Furthermore, due to the lack of distinct clusters, analysis into leukemia patients is made more difficult due to atypical staging. The vast majority of cancers are staged based on tumor size, and if they have spread from the place of origin, both of which are qualities that are absent from instances of leukemia. According to Cancer Treatment Centers of America [3], doctors often stage leukemia by various blood cell counts or the number of leukemia cells that get built up in a patient's organs, such as the liver. There is further complexity in staging based on the leukemia subtype, often calling for dramatically different techniques to be used in examining the condition and making the call for a stage. The National Cancer Institute Surveillance, Epidemiology, and End Results program (SEER) points out that across the board there is no standard staging method for leukemia [4], resulting in a large vacancy of a typical and commonly interpretable factor toward a patient’s risk.

Data

Making further use of the information available from SEER for 1973 to 2014, we extracted the data used in this study from the program’s database. This data is also the same data that the LLS used to find their 5-year survival rates. It is expected to see that the time of diagnosis will have a significant impact on the 5-year survival rates and the patient hazards to be discussed later on. We restricted the data to only the desired ages, 0-19 years, and removed patients diagnosed before 1975. The restriction on diagnosis year was made to simplify the task of discretizing time-periods into five-year blocks in the discrete modeling section of the research. Second, all exact duplicate rows were removed leaving only 102 duplicated patient ID's left in the data. Of the duplicated patients there was some discrepancy in treatment. Some patients had multiple rows, each with a different combination of treatment variables. Those rows were simplified to include the complete treatment information to represent treatments received for that particular cancer. We designated any further duplicated patient ID's as a new instance of cancer, with different age, diagnosis year, sequence number and unique survival time. These rows were left in the final data to assess the impact of each independent cancer, as we are not examining patients but the instance of cancer. With this in mind, due to the possible dependencies that could be present due to a repeated host patient for multiple cancers, we performed sensitivity analysis on our final models with their omission, discussed in the results section. These filtrations left us with 18421 unique observations of cancers.

Furthermore, we simplified radiation treatment into only two levels including received or did not receive/unknown if received. While this is not a typical leveling strategy, the SEER leveling method for chemotherapy is in terms of treatment vs. none/unknown. We extended this strategy to radiation where there is no specification between no treatment and unknown, so the decision was made to follow the simplified coding method SEER employed for chemotherapy on radiation as well. Year of diagnosis was divided into five year periods, for us to create eight time-periods. We marked all patients that were still alive at the end of the study as censored, as well as those patients who lost contact and did not follow up during the time of the study. The final split put the original data into three different formats: The standard format with all 18421 rows for data exploration and continuous modeling, a long-format for the discrete extension of Cox Proportional Hazard modeling (Cox PH), and a wide-format for discrete neural network modeling.

For the long-format, each row had multiple repetitions, one for each of the discrete time-periods that the cancer was present in the study. For example, if a patient died in the fourth period, their records would be repeated four times with an indicator noting that they died in that last period, as well as indicators for which time-period that row represents.

The wide-format contains eight additional columns representing the patient's status in the first through last time-period. If a patient was alive in a given time-period, that entry was marked 0, and if they died or were deceased in a time-period, the entry was marked 1. Patients who lost contact had their vital status marked with the hazard for each given time-period they were absent from the study. For example, if a patient were lost to follow-up in the fifth time-period, then the sixth through eighth periods would be marked each with their respective hazard. For neural network models, all factor variables took the form of n-1 dummy variables, where n is the number of factor levels.

Initial Investigations

Figures 1a and 1b show the geometry for survival time, split across censored and uncensored patients. This triangular pattern present in survival time is due to the structure of the study. Patients had their survival time recorded regardless of their event, and in a living patient's case, represents the time in months from their diagnosis to the end of the study. Since the study cuts off all data collection in 2014, despite bringing new patients in up to then, the hypotenuses of these triangular areas represent the maximum survival time patients could have at a given diagnosis time. We also observed numerous censored cases where patients lost contact and were last classified as alive but have less than the maximum survival time for their diagnosis year. The trend for deceased patients is what we expect in most survival data, most dense in the realm of shorter survival times.

In general, inference on cancer data is derived from a core list of different features. Some of the common factors are surgical procedures and tumor size, among others. In the case of leukemia, these features are not particularly relevant or available.

Figure 1: Survival time groupings by event

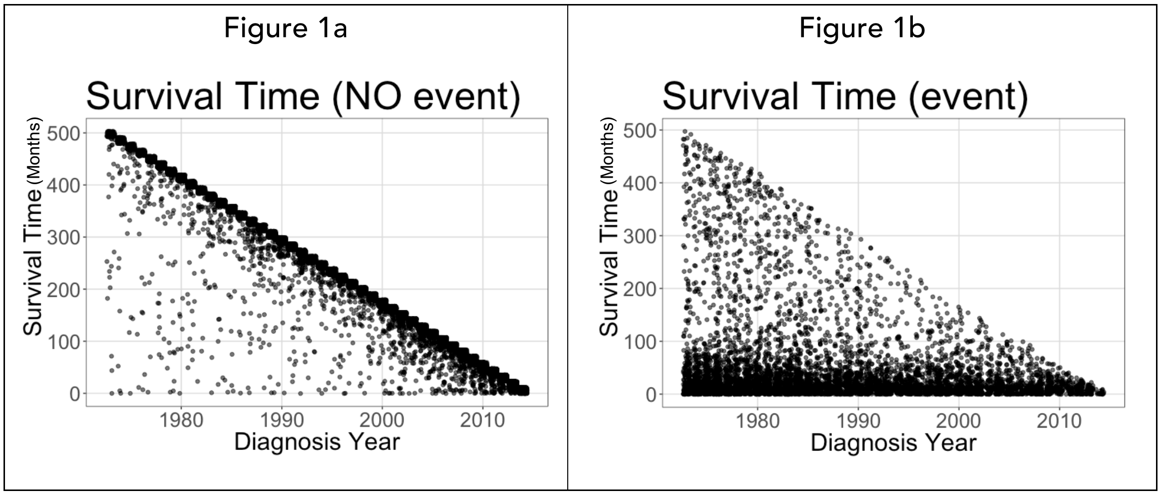


Figure 1a

Figure 1b

Additionally, some of the variables were rarely tracked, being marked as unknown or the entries were left blank. Many of these unknowns were coupled with the variables that had inconsistent tracking across all time-periods. Not all data was recorded for the entire length of the study. Some variables only had records in the earlier years, and others were introduced within the most recent years. The decision was made to remove these variables as the particular aim of the study required that our covariates be traceable from the beginning of the study to the end. In some cases, variables were re-coded to be more accurate to different recording conventions.

Of the complete features that remained, we examined variables such as age, sex, treatment, sequence, among others (Full variable table found in the appendix.) First, looking into treatment, we notice that radiation shows a surprising level of survival, as well as longer survival times. Additionally, we noted that the patients were mainly in the earliest diagnosis periods and among the oldest patients. Figure 2a shows this interaction. We note that there is an evident cluster of the patients with only radiation administered, in the earlier time-periods at older ages, that seemed to survive longer. Since leukemia is blood based, it typically is not as affected by radiation alone, so this cluster was worth further inspection. Through examining the cluster, we noticed that the majority of these patients had a primary site identifying that they were

Figure 2: Survival Time and Treatment Effects With and Without Hodgkin’s Disease Patients

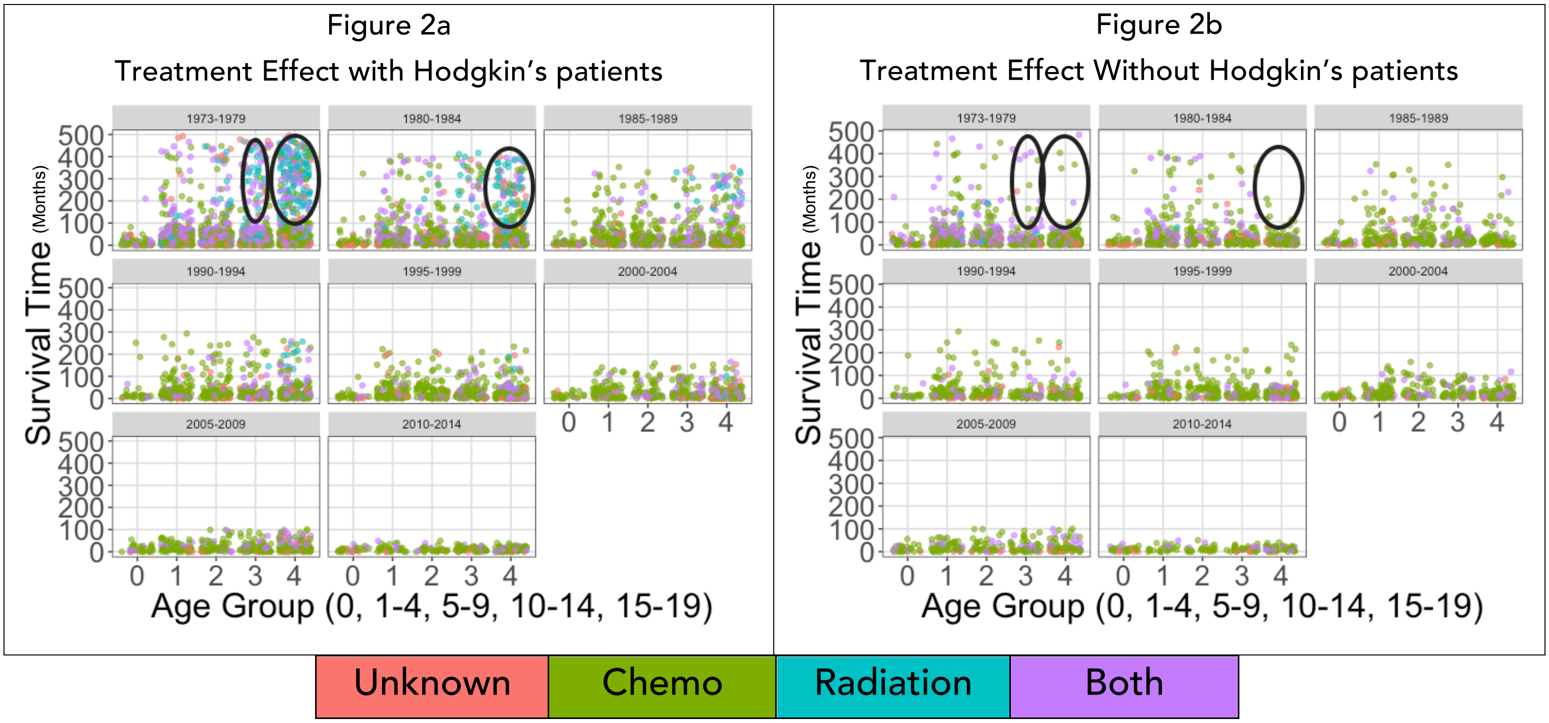


Figure 2a

Treatment Effect with Hodgkin’s patients

Figure 2b

Treatment Effect Without Hodgkin’s patients diagnosed with Hodgkin’s disease. Since Hodgkin’s disease has a primary site of the lymph nodes, radiation has a more substantial impact on it, and therefore patients show significantly better survival rates than other patients. Re-examining the initial plot, omitting the Hodgkin’s patients, shows a more typical relationship in Figure 2b where the patients who survived had either chemotherapy or both chemotherapy and radiation. Because patients with Hodgkin's disease were a moderately sized group of patients present and there was a clear impact on treatment by these individuals, we repeated our hypothesis testing, discussed momentarily, with their exclusion to determine how sensitive the other covariates were to this diagnosis discrepancy. Upon repeating these tests, we encountered similar results, suggesting that the only difference between these patients could be found in the treatment clusters of figure 2a. Under this reasoning, these patients remained in the study and their unique information was to influence the models as the effect of radiation on a patients survival. Thus, radiation-only treatment would act as our factor variable, in essence, representing an intercept term specific to this diagnosis type.

While performing our tests we allowed each patient to be represented and their general information to vary as to capture the overall relationship between each of the variables discussed. Thus, the tests were carried out under the natural data structure, without variables held constant, to identify specific conditional interactions between multiple covariates. All tests were carried out using a standard 0.05 level of significance, determined beforehand.

After performing the Fligner-Policello for the difference between two sample medians and z-test for comparison of two proportions, we observed that there was no difference between the survival times and survival rates of male and female patients. Due to this, we excluded sex from any further examination.

Patients were assigned one of five groups based on age at diagnosis. One group contains only infants and the remaining four are structured in five-year intervals. Kaplan-Meier plots [13] showed the infants as the group with the lowest survival, with the 15-19 year-olds showing the highest survival. Again, through use of the Fligner-Policello test, we observed that, like the survival probabilities, infants had the shortest survival with 15-19 year-olds surviving the longest, with the remaining three groups showing insignificant differences from one another at the 0.05 level.

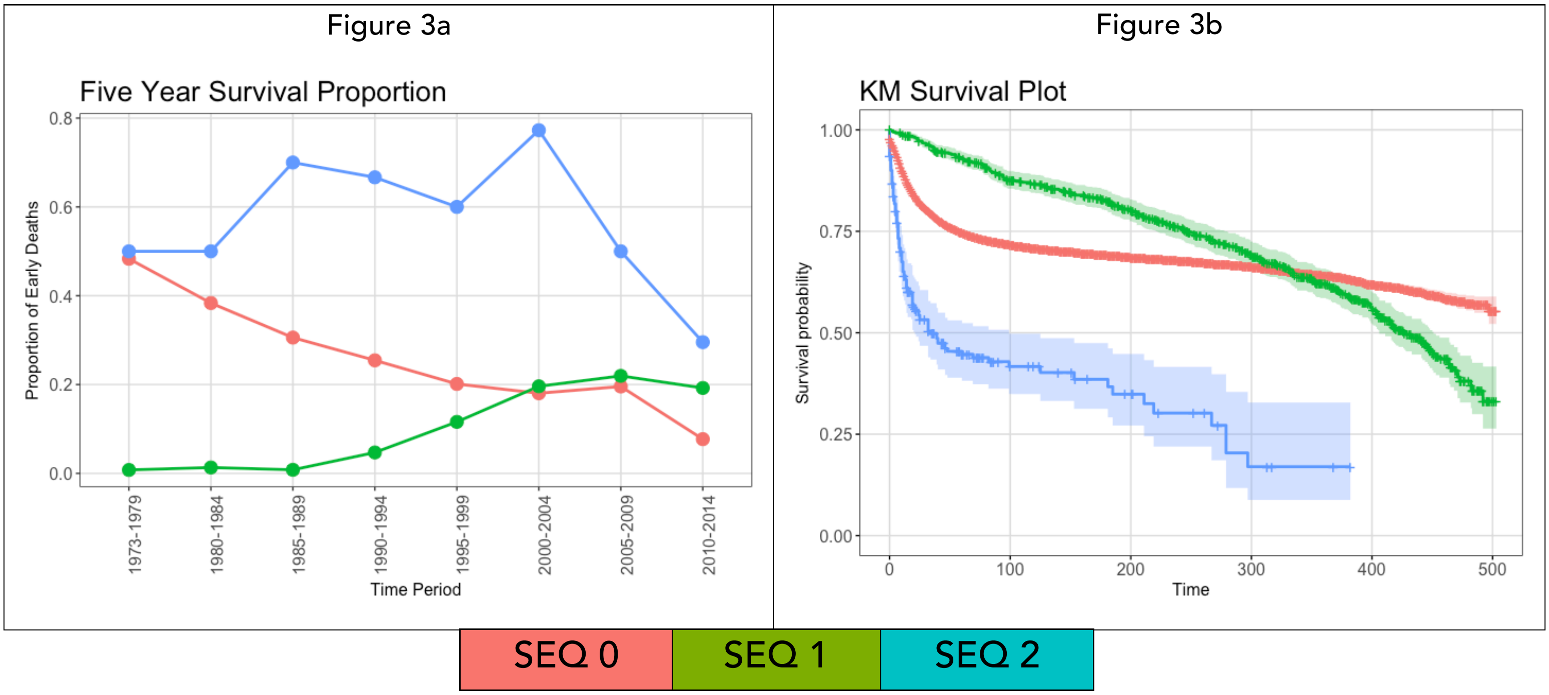
One of the additional factors often used in the medical field to determine the severity of leukemia is the grade of cancer, typically denoted by the cell type [5]. The primary cells for classification are B and T cells so we examined them with a particular interest. Typically, T cell leukemia is more aggressive and usually results in deadlier cancers. Hypothesis testing showed that B and T cell types are different among patients' survival. Kruskal-Wallis tests for several independent samples showed that there was no significant difference between the cell types and median survival time. T cells did, however, show shorter but insignificant survival times, which fell in line with current understandings of leukemia. Further z-testing for comparison of proportion did provide more evidence for the inclusion of cell type in our work. We revealed that early death proportions for T cell leukemia were significantly higher than those patients with B cell type leukemia, until leveling off near five years after diagnosis.

Time-period of diagnosis depicts expected trends for survival time. Even disregarding the improvements in treatment that have come with time, the nature of the study lent itself towards higher survival rates and shorter overall survival times for the more recent periods. Since SEER recorded final survival time and status of all patients in 2014, as diagnoses approached 2014, patients were given less time in the study to experience the event. Thus, maximum survival time decreases, and we observe a reduced number of events. Additionally, the change across each time-period was relatively consistent and fit the expectations built around the study's structure.

Initial testing on the number of prior diagnoses for each patient, referred to as sequence/sequence number, left us with consistent, yet surprising trends. Fligner-Policello testing revealed that the number of prior cancers (counts 0-2) did show a significant difference in the survival time of the patients who experienced the event. In examining each pair, we found that each sequence level was different from both others. Sequence two showed the lowest median survival time (8 months), followed by no prior cancers (19 months) and one prior cancer (200 months). The survival time of patients with one prior cancer is unexpectedly higher than zero and two prior cancers.

Further investigation was conducted to verify these results. Figure 3a and 3b show the proportion of patients that died within the first five years after receiving their diagnosis (conditional on when they were diagnosed) and Kaplan-Meier probabilities over time (in months) respectively. Both figures illustrate behavior similar to that shown in hypothesis testing. These plots are divided across time-periods to help see the changes over time. However, in general, a sequence of one still shows to be beneficial

Figure 3: Sequence Effects on Patient Survival



in survival time across all periods and only showing five-year survival for zero prior cancers surpassing it in the most recent 15 years. Finally, Kaplan-Meier estimates express that having no previous cancers only benefits patients after they have already survived for 30+ years. The sequence did not show a relationship that simply explained itself, as was the case with treatment and Hodgkin's patients. We hypothesize that these patients, with the surprising survival times, despite the event taking place, could have potentially been caught in an earlier stage (made difficult by no standard staging method). Alternatively, these patients could all share a particular feature which is not present in our given data. These trends were also present when removing patients that had multiple instances of leukemia in the data, thus showing that the patterns were not confounding effects caused by potential dependencies between those particular cancers.

In summary, we selected age group, grade, sequence, chemotherapy, radiation treatment, and time-period diagnosed as the covariates for modeling. This decision was made based on the variables' trends through initial testing, their completeness, and usage as standard metrics. Key hypothesis testing results are summarized in tabular form in the appendix.

Methodology

Let T be a random variable representing survival time of interest, the time until a given event takes place. We note that this random variable can be discrete or continuous based on how the sample space is defined.

Often, it is not practical to discuss the probabilities of survival, but the functional form of survival which is easily derived and can help represent a patient's risk. The probability of having not experienced the event at time *t* can be defined as

$$S(t) = P(T > t). \tag{1}$$

The survival of an individual in any given time interval *t* can be found by taking the product of the conditional probabilities for each previous interval where the risk set changed. This can then be written as

$$S(t) = \prod\nolimits_{t(i)} \frac{n_i - d_i}{n_i}, \tag{2}$$

where *t(i)* represents the set of all time intervals that individuals left the risk set, n_i representing the size of the risk set at interval t_i and d_i representing the number of events in interval t_i.

The risk, or hazard, for an individual experiencing the event at any instant, can be defined as

$$h(t) = \lim_{\Delta t \to 0} \frac{P[t \leq T \leq t + \Delta t | T \geq t\}}{\Delta t}. \tag{3}$$

From (3), it can be derived that *h(t)* can be more succinctly written as

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}\{\ln(S(t))\}, \tag{4}$$

where $f(t)$ is the PDF, $P(T = t)$.

One of the most common proportional hazard models, the Cox proportional hazard model [9], is written as

$$h(t, x, \beta) = h_0(t) * r(x, \beta), \tag{5}$$

where $r(x, \beta)$ is the exponential parameterization of the hazard function

$$r(x, \beta) = e^{(\beta_0 + \sum_{i=1}^{n} \beta_i x_i)}, \tag{6}$$

and $h_0(t)$ is referred to as "baseline hazard," the rate at which hazard changes over time. Parameters of this type of regression procedure are estimated by maximizing the log-likelihood function. However, Cox used the "partial likelihood" (6) method, whereby we maximize

$$\ell_p(\beta) = \left(\prod_{i=1}^{n} \frac{e^{x_i \beta}}{\sum_{j \in R_{(t_i)}} e^{x_j \beta_j}} \right)^{c_i}, \tag{7}$$

where $R_{(t_i)}$ is the total number of individuals in the risk set at time t_i and c_i is the censorship term, one for the event and zero for non-events. The Cox model calls for three major assumptions to hold for the method to be valid. 1) The hazard ratio for any two individuals must be constant with respect to time, 2) there must be a limited number of non-event cases, and 3) each time of event is unique. In other words no more than one event happened at a given instant, though there are multiple common methods of compensating if this assumption fails. While the method is widely used and offers great interpretability, the rigidity of the model assumptions makes it difficult to use on instances of imperfect data.

By re-examining the structure of our random variable for survival time and the interpretation of our event, we can extend out to alternate methods. By breaking our survival time into a discrete random variable across the range of survival times in a given data set, we can use logistic regression to estimate an individual's hazard using discrete survival time as a conditional covariate [10]. We also briefly utilized support vector machines (SVM) [11] to classify event status. The SVM method aims to maximize the distance separating events from censored cases. To reduce classification errors, two SVM's were chained together. The goal of the first SVM in the chain is to use all covariates and predict a penalty term, how likely it is for a patient's event status to be incorrect. The second vector takes all covariates and the new penalty term and uses these terms to predict the final patient vital status.

Our final method involved the use of artificial neural networks to provide a nonlinear alternative to Cox PH "not constrained to strong assumptions on the effect of the covariates" [12]. We benefit from the use of neural networks as they learn off patients' actual vital status and thus learn how to predict for both events and censored cases. Neural networks treat each variable as an input node and each response as an output node. In the case of a categorical response, each category level is represented as a distinct output node. Between these two layers can be an arbitrary number of "hidden layers" with an arbitrary number of nodes in each layer. Each node in each layer connects to each node in the subsequent layer, and those connections all hold a weight, or importance, to the model structure. The model is trained and goes through the learning process by multiplying each input by each connection weight (which is repeated layer to layer) and feeding forward through the network. At the hidden layer, the hyperbolic tangent (tanh) function

$$tanh(a) = \frac{e^{a} - e^{-a}}{e^{a} + e^{-a}}, \tag{8}$$

is applied for simplicity as it reduces inner network values onto the range of [-1,1]. Upon reaching the output layer, the weighted values have the sigmoid function

$$sigmoid(z) = \frac{1}{1 + e^{-z}}, \tag{9}$$

which restricts the range onto [0,1] and has an interpretation akin to a hazard probability. The network then applies the cross-entropy error function (similar to the use of MSE in regression)

$$CE(\hat{y}) = (-y * \ln(\hat{y})) + \big((1 - y) * \ln(1 - \hat{y})\big), \tag{10}$$

where $\hat{y}$ is the predicted value and y is actual patient vital status. The network then uses the gradient of (10) with respect to the connection weights to minimize the error and replace the weight which will reduce the total network bias the most. The network then repeats the process with the updated weights until the errors converge. The outputs from this final iteration of the model can then be properly compared to the actual vital status to determine model predictive effectiveness.

Results

After exploring different parametric and classification models, we were left with two neural network models using discrete survival time. These models are arguably the most robust of those explored, with the most reliable predictions.

The initial starting point was Cox PH as it is the standard method for working with time-to-event data. Several assumptions (most of which our data severely failed) and poor prediction accuracy on our data led us to find possible alternatives to the Cox model. We shifted focus to discrete methods with logistic regression to predict the probability that an individual would experience the event in a five-year period. Though offering stronger results than the Cox model, the estimates of the logistic regression model were still weaker than desired. We then explored machine learning methods, even though we lose some interpretability. Labeled SVM strings offered stronger predictions than logistic regression but lacked a level of customizability offered by other machine learning methods, such as neural networks.

All the models confirmed that the covariates employed were influential and demonstrated an impact on survival, falling in line with initial work and making it clear that it was appropriate to update these models and search for better prediction accuracy. Neural networks allowed for an amalgamation of the benefits of these models without many of the shortcomings. We used two discrete methods to be sure that we captured the temporal relationship observed in the data, as well as the impact development in the medical field had on patient's survival over time. The overall structure of the hidden layer for each of the two models was determined by training with cross-validation and the comparison of total model errors. Each model used the ten-fold cross-validation method. The data was split into 70% training set and 30% validation set. The training data was then divided into ten equal folds, then the models were trained on each nine fold combination and tested on the remaining fold. Optimal models were then trained on the entire training set and validated with the unseen data to check for overfitting.

The critical difference between the two discrete models is how they include time and their output layer structure. The first model, referred to as Model 1, took the patients' recorded survival time and discretized it based on five-year intervals, up to 40 years. The range of 40 years was decided based on the maximum length of the study, and five-year intervals offered a clear cut-off point which could also serve as outputs in the second model. Since a patient can only survive into time interval t if they did not already experience their event in a time-period less than t, the results from the model are strictly conditional on the patients having survived to that point. Since each individual only has one hazard based on their covariates and conditional survival time, only one output node is necessary for this model.

The second model, Model 2, predicts a hazard for each of the time-periods rather than a single value per cancer instance. Limiting to eight intervals of five year spans made this method reasonable by using a standard length to capture the significant medical development over time without being exceptionally difficult computationally. To predict eight separate hazards, we took each patient's survival time and converted it into actual vital status for each time-period. If a patient's survival time depicts a presence of an event within 10-15 years, then their outputs would be recorded 0 for periods [0, 5) and for [5, 10) with each subsequent time-period being marked 1 as having experienced the event in these periods. These values then formed the responses for our eight outputs on which we trained the model. Figure 4a and 4b present the final network diagrams of Model 1 and Model 2, showing the structure of each component layer.

To restate, while the initial models yielded a relatively low accuracy on our data, the variables from these models were used as the basis for the neural networks that comprise our final models. Model 1 was built using age group, sequence, treatment, grade, and survival time to predict the probability that a patient will experience the event. The sequence variable was scaled between 0 and 1 to help reduce the computational power needed to train the network. The final form of Model 1 had a single hidden layer with six nodes and a decay value of 0.075. These values were found by training with all possible combinations of 1-10 nodes and decay terms 0.025, 0.05, 0.075, and 0.1. We trained each model on these combinations and then validated with ten-fold cross-validation. The final decision on parameters to use is made based on the set of parameters that has the lowest cross-entropy. This process can be visualized in

Figure 4: Model 1 and Model 2 Network Diagrams

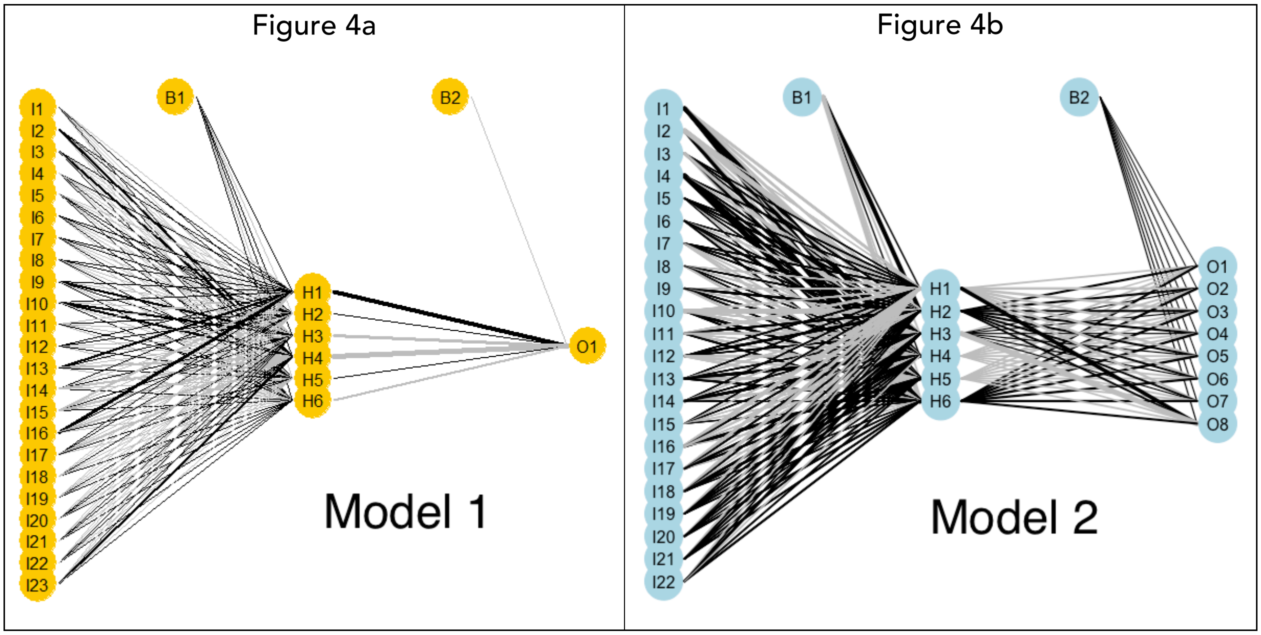


Figure 4a

Figure 4b

Appendix Figure 2. The cross-entropy given by Model 1 is 472.88. Due to the fact the model outputs a single probability for each patient, a cutoff point is set to determine whether or not the patient is predicted to have experienced the event. Tuning the cutoff point from .6-.8 to increase the model's accuracy was unnecessary because each output value is extremely close to either 0 or 1. Figure 5 depicts a contingency table using validation data where Model 1 achieves a correct classification rate of 0.847. Note that time-period diagnosed is removed from Model 1 because of high collinearity between time-period diagnosed and survival time.

Figure 5: Neural Network Contingency Tables

Figure 5						
	Model 1			Model 2		
	Actual			Actual		
		Alive	Dead		Alive	Dead
Predicted	Alive	8613	1423	Alive	30438	9777
	Dead	571	2288	Dead	1161	2832

Model 2 consisted of the majority of variables from Model 1 with slight alterations. Age group, sequence, treatment, and grade are used by with the addition of discrete diagnosis time-period. After cross-validation, the final form of Model 2 used a single hidden layer with six nodes and a decay value of 0.05. We determined these values by finding the network with the minimum value for the cross-entropy error. While performing ten-fold cross-validation, we observed six nodes in the hidden layer and a decay value of 0.05 showed the minimum average error from models built from all combinations of 4-15 nodes and decay terms 0, 0.05, 0.075, and 0.1. When applying our validation set, the model produced a total cross-entropy of 23442.66 with a mean and standard deviation of 0.53 and 0.23 respectively. After considering the eight outputs, we observed cross-entropy of approximately 2930.00 per output. Model 2's classifications used a cutoff hazard of 0.6, with any value equal or greater than 0.6 representing an event. Unlike Model 1, Model 2 varied in accuracy across different cutoffs, and thus 0.6 was used as it shad the highest accuracy. In summary, Model 2 demonstrated a classification rate of 0.753. Assuming each of the eight outputs contributes proportionately to the errors, each output accounts for approximately 3% of the total 24% error.

To better understand the particular variable effects of sequence and treatment we created sets of evaluation data, two for each model, where we evaluated across each level of the two respective variables and held all other variables fixed. To avoid particular conditions impacting the evaluation, treatment was held constant for sequence as no treatment, and sequence was held constant for treatment at a level of no prior cancers. Additionally, the grade variable was held constant at B-cells as they are typically less aggressive. Age groups were held constant at the median level, and time-period diagnosed, in Model 2, was held constant at the fifth time-period (early 2000's) of diagnosis as this is often a popular reference point in the study of cancer. The evaluation data was fed through each model to simulate the relative variable effects, and in the case of Model 1, the data was sent through conditional on surviving into each of the time-periods to obtain eight hazards, comparable to Model 2's outputs.

Figures 6-9 provide a visualization of our evaluated data for variable effects. Figures 6 and 7 both show treatment having an overall and consistently low hazard. Figure 6 shows a substantial decrease in the hazard of the patients who survived five years or longer as compared to the very slight positive trend of figure 7. Figure 8 and 9 show sequence elicits a much higher variance between the variable levels than treatment. Figure 8 shows another negative correlation with hazard from Model 1 and Figure 9 shows a slight positive correlation in sequence for Model 2. Additionally, Figure 7 and 9 show the unique relationships from initial testing: a sequence of one demonstrating the best survival among sequence levels, as well as radiation showing the best survival among treatments. Both of these trends are absent from Figures 6 and 8.

To account for the potential dependencies due to repeated patients, individuals present with multiple instances of leukemia, sensitivity analysis was conducted on Model 1 and two. The optimal models were once again trained and tested on the data with the omission of the patient IDs for those with multiple cases. Recalculation of the overall cross-entropy error and classification rate on Model 1 revealed only a 2.8% decrease in model error and a 1% decrease in correct classifications. Recalculation of error and classification for Model 2 showed a 3% decrease in the overall model error coupled with a 1% increase in the classification rate. We concluded that these changes in error and classification did not represent a substantial model sensitivity to patients with repeated instances of leukemia. We also note that the removal of these patients

Figure 6-9: Model Comparison Plots for Treatment and Sequence for Each Time Period

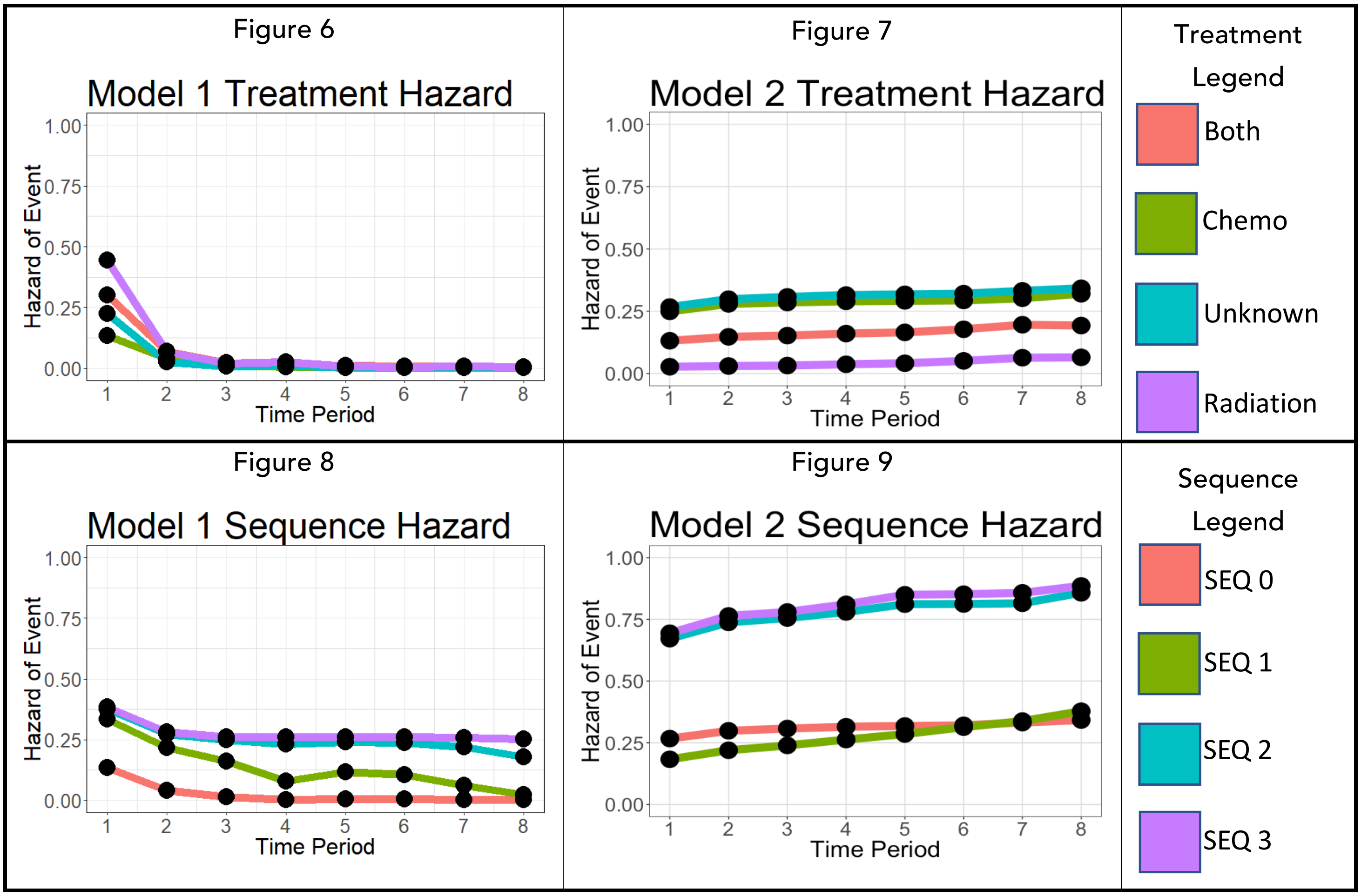


Figure 6

Figure 7

Figure 8

Figure 9

is not the only factor contributing to the changes. Despite using the same random assignment method of patients into train and test groups, the change of sample size does result in individuals being assigned to a set they may not have been present in for the first iteration. Therefore, due to minimal error and classification change, as well as making up only a small portion of the data, we do not find sufficient reason for the omission of multiple occurrence patients from the modeling.

Discussion and Conclusions

Looking into the outcome of both models regarding the overall model error and classification percentage, Model 1 shows itself to be preferable with almost a 10% higher classification rate and nearly 2% of the overall error. Our results are different from those in [6], regarding which network resulted in less error. In "Two Artificial Neural Network Approaches For Modeling Discrete Survival Time of Censored Data" [6], they uncovered that the second network structure resulted in a lower error than that of the first. Model 2 offers better capture of variable trends, as well as identifying factors that were expected to influence hazard. A case in point, this model captures the effect of radiation due to the Hodgkins patients within the data. Even though it does not perfectly capture the intricacies between individual levels, Model 2 captures the features identified as particularly critical. These outcomes create a different comparison of the two models: one is more accurate and holds less error while the other can identify trends and variable levels that were determined to be highly relevant.

While capturing the effects of contributing variables is the most robust feature of Model 2, Model 1 identifies the overall importance of a patient's risk based on how long they have survived up until time t, in addition to prediction accuracy. The influence on conditional survival time is seen in Figure 6 involving treatment. For all levels, once a patient enters their second time-period (years 5-10) of survival post diagnosis they are what some may consider "cured" [7], having their risk of death drop off considerably. This shift is far less visible in Figure 8. The overall negative trend in hazard does show that eventually a patient's risk of experiencing the event will stabilize and fall mostly to a constant level. However, Model 2 shows that the hazard increases as time increases, which was expected based on the model structure. To continue to predict a patient's vital status (aiming to avoid going from a dead prediction in one time-period to an alive prediction in the next period) for each subsequent time-period, the model predicts slightly higher hazards as time progresses for the evaluation data.

There is no obvious strategy *regarding* conventional methods of choosing preferred models based on two different areas of success. Since the models' success are not solely based on quantifiable measures, selecting a model without a well-defined method is less than optimal. Each model provides an essential part of the story of modeling, and thus, with the current scope of our research, we lack sufficient reason to select only one and instead claim that each is necessary to gather the most reliable insights about childhood leukemia. Model 1 creates the opportunity to evaluate a patient through follow up appointments, stressing the importance of surviving their first few years, while Model 2 can give instantaneous expectations of a patient's hazard into the future.

Future work aims to identify additional features which could potentially increase variability among covariates in our model. Lack of variability and reliance on categorical data does put a limit on how different each patient is from another, and thus how robust the model can be. Furthermore, we aim to examine interactions between the covariates and identify key variable level groupings that are particularly notable.

References

[1] *Cancers that Develop in Children*, (n.d.), https://www.cancer.org/cancer/cancer-in-children/types-of-childhood-cancers.html.

[2] *Facts and Statistics*, Leukemia & Lymphoma Society, https://www.lls.org/http%3A/llsorg.prod.acquia-sites.com/facts-and-statistics/facts-and-statistics-overview/facts-and-statistics, Accessed 20 Apr. 2018.

[3] *Stages Of Leukemia: What Are the Stages? | CTCA*, CancerCenter.com, 1 Jan. 1AD, www.cancercenter.com/leukemia/stages/, Accessed 20 Aug. 2018.

[4] *Cancer Stat Facts: Leukemia*. Surveillance, Epidemiology, and End Results (SEER) Program, https://seer.cancer.gov/statfacts/html/leuks.html. Accessed 20 Aug 2018.

[5] *How Is Acute Lymphocytic Leukemia Classified*?, https://www.cancer.org/cancer/acute-lymphocytic-leukemia/detection-diagnosis-staging/how-classified.html, (2016, February 18). Accessed January 19, 2018.

[6] T. Sharaf and C. P. Tsokos, *Two Artificial Neural Network Approaches For Modeling Discrete Survival Time of Censored Data*, Adv. Artif. Intel., 2015, Article ID 270165.

[7] *Childhood Leukemia Survival Rates*, https://www.cancer.org/cancer/leukemia-in-children/detection-diagnosis-staging/survival-rates.html, (2016, February 3), Accessed October 2018.

[8] Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Research Data (1973-2014), National Cancer Institute, DCCPS, Surveillance Research Program, released April 2017, based on the November 2016 submission.

[9] D. R. Cox, *Regression Models and Life-Tables,* J. R. Stat. Soc. Ser. B Stat. Methodol. Vol. 34, No. 2. (1972), pp. 187-220.

[10] T. Sharaf, and C. P. Tsokos, (2014), *Predicting Survival Time of Localized Melanoma Patients Using Discrete Survival Time Method,* J. Appl. Stat. Methodol.: Vol. 13: Iss. 1 , Article 9. DOI: 10.22237/jmasm/1398917280.

[11] Y.-J. Lee, O. L. Mangasarian, and W. H. Wolberg, (2001), *Survival Time Classification of Breast Cancer Patients* (technical report): ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/01-03.ps.

[12] E. M. Biganzoli, F. Ambrogi, and P. Boracchi, (2009), *Partial logistic artificial neural networks (PLANN) for flexible modeling of censored survival data*, in 2009 International Joint Conference on Neural Networks, doi:10.1109/ijcnn.2009.5178824.

[13] E. L. Kaplan, and P. Meier, (1958), *Nonparametric Estimation from Incomplete Observations*, J. Amer. Statist. Assoc., Vol. 53(282), pp. 457-481. doi:10.1080/01621459.1958.10501452.

Appendix

Hypothesis testing results			
Variable	***Test***	***Parameter***	***P-Value***
Sex(male/female)	Fligner-Policello for of two med.	Median srv. time	0.40
Chemo(Yes/No, unknown)	Fligner-Policello for of two med.	Median srv. time	0.003
Age group (5 levels)	Kruskal-Wallis for of multi. med.	Median srv. time	2.9e-8
Age Groups (1,2)	Wilcox Rank Sum for two med.	Median srv. time	0.76
Age Groups (2,3)	Wilcox Rank Sum for two med.	Median srv. time	0.92
Age Groups (0,1)	Fligner-Policello for of two med.	Median srv. time	4.8e-10
Age Groups (0,2)	Fligner-Policello for of two med.	Median srv. time	3.4e-8
Age Groups (0,3)	Fligner-Policello for of two med.	Median srv. time	8e-9
Age Groups (0,4)	Fligner-Policello for of two med.	Median srv. time	1.6
Age Groups (1,3)	Fligner-Policello for of two med.	Median srv. time	0.56
Age Groups (1,4)	Fligner-Policello for of two med.	Median srv. time	0.08
Age Groups (2,4)	Fligner-Policello for of two med.	Median srv. time	0.05
Age Groups (3,4)	Fligner-Policello for of two med.	Median srv. time	0.02
Chemo Rad interactions	Kruskal-Wallis for of multi. med.	Median srv. time	2e-16
Grade (T, B, Null, Ndet)	Kruskal-Wallis for of multi. med.	Median srv. time	0.657
Sequence (0,1,2)	Kruskal-Wallis for of multi. med.	Median srv. time	2e-16
Sequence (0,1)	Fligner-Policello for of two med.	Median srv. time	0.00
Sequence (0,2)	Fligner-Policello for of two med.	Median srv. time	6e-8
Sequence (1,2)	Fligner-Policello for of two med.	Median srv. time	3e-184
Grade prop (T, B)	Two prop test for early deaths	Early death prop	0.03
Sequence prop (0,1)	Two prop test for early deaths	Early death prop	2e-16
Sequence prop (1,2)	Two prop test for early deaths	Early death prop	2e-16

Figure 3b by Time Bin (Appendix Figure 1)

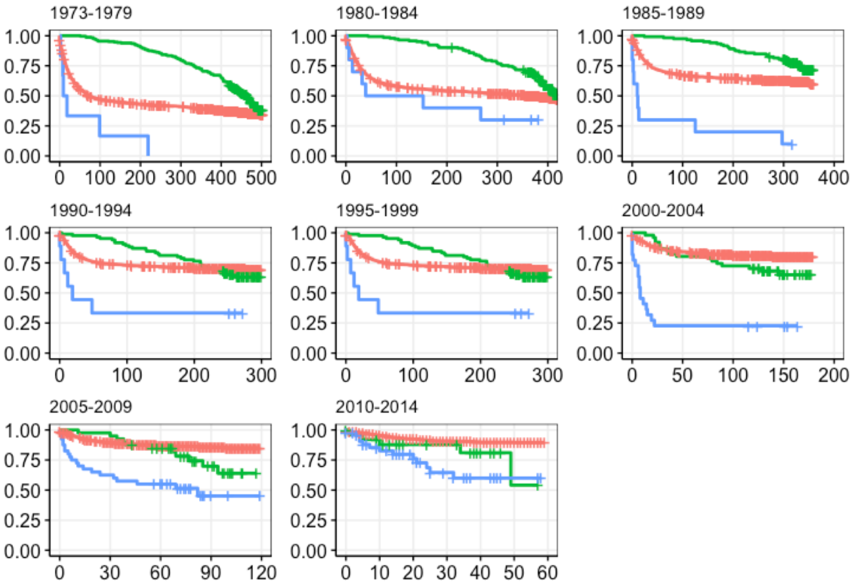
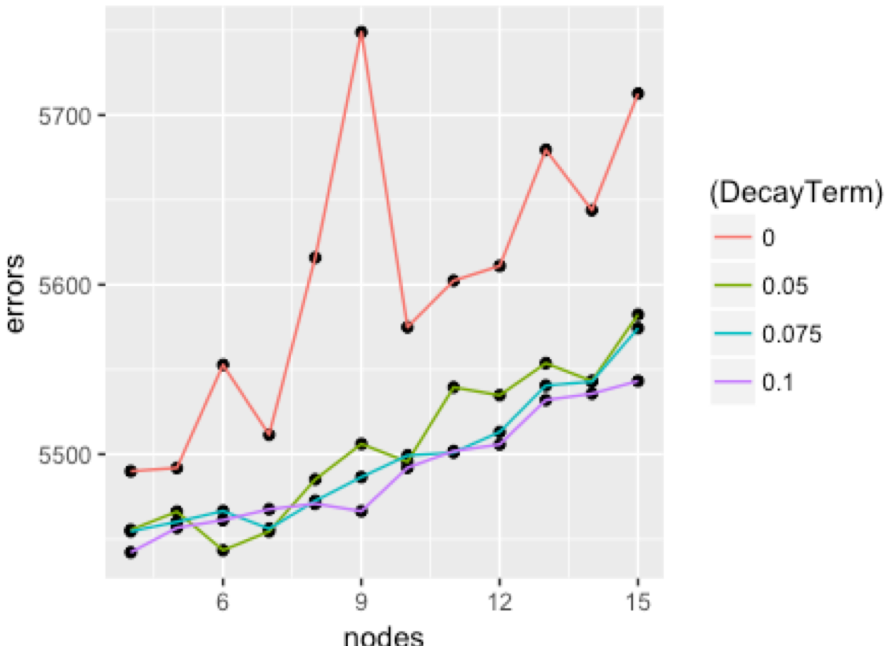


Table of Variables Examined

Variable name	*Model Inclusion*
Age Group	Yes
Chemotherapy Treatment	Yes
Derived Stage	No
Ethnicity	No
Gender	No
Geographic Location (state and county)	No
Grade	Yes
Primary Site	No
Radiation Treatment	Yes
Sequence	Yes
Survival Time	Yes
Time Period Diagnosed	Yes
Year Diagnosed	No

Ten-Fold Cross-Validation Errors (Appendix Figure 2)


errors
nodes
(DecayTerm)