The University of New Mexico

Department of Mathematics & Statistics

An Exploration of Mortality Trends Using the SEER Data Set

Author:
Katherine Freeland

Advisor:
Dr. Michael Sonksen

Abstract

Using the SEER data set, a massive publicly available cancer database, we wish to explore mortality trends. The data set captures a wide scope of demographic information, as well as treatment and diagnostic information that we believe will enhance our understanding of mortality prediction. We will compare mortality trends, using both a logistic regression model and a mixed effects model, to the United Health Foundation's State Health Rankings. We believe a relationship exists between state mortality rates and general state health. Also, a Cox proportional hazards model will be used to better predict mortality using the variables collected in the SEER data set. The relationship between time until death and location of treatment will also be explored. We conclude that although mortality can be explained by a hierarchy of subpopulations (by state), this does not correlate directly to the State Health Rankings.

Contents

1 Introduction

Using data collected by the Surveillance Epidemiology and End Results (SEER) program, a massive publicly available cancer database, we wish to explore several questions concerning breast cancer mortality trends. The data set captures a wide scope of demographic information, as well as treatment and diagnostic information that we believe will enhance our understanding of treatment effects and aid in mortality prediction.

We suspect more impoverished areas, with less access to quality health care, have greater breast cancer mortality. Using both a logistic regression model and a mixed effects model, we will explore cancer mortality by state and compare our findings to the 2009 (the last year included in the most recent data set) America's Health Rankings, sponsored by the United Health Foundation. These rankings compare states using indicators such as health insurance, childhood povery, premature death, etc to determine a general 'healthfulness' score. The subset of SEER data used contains data from the following states: California, Connecticut, Georgia, Hawaii, Iowa, Michigan, New Mexico, Utah, and Washington.

The logistic regression model was selected because it examines the relationship between a binary outcome ('1' = death, '0' = alive) and predictor variables that can be both categorical and numeric. In doing so, it generates a model that predicts the odds of occurrence. This model is commonly used in the public health field for mortality prediction. After the logistic regression model is established, information about the location of treatment (by state) will be added as a random effect. A random effect establishes a hierarchy of different populations, in this case by treatment location, and allows analysis of different populations based on that hierarchy.

The immense amount of information contained in the SEER data set has the ability to aid in the prediction of time until death. We will use a Cox proportional hazards model, which examines the time it takes for an event to occur (in this case death) based on a set of predictors (commonly called covariates). The SEER data set includes current patients and therefore we cannot always observe when (and if) the patient has died of cancer. In other words, we have right-censored data. The Cox model is especially useful when dealing with this problem. The model evaluates the varying impact of the covariates on predicting survival time and allows us to predict, given a patient's information, an estimated time until death from cancer.

All models, calculations, and graphics will be generated using SAS statistical software.

2 Data Description

Beginning in 1973, the Surveillance Epidemiology and End Results (SEER) program at the National Cancer Institute began to collect data on cancer cases throughout the country. As collection continued, the data set expanded to include many more geographic areas and additional demographic information. The current data set includes information collected between 1973 and 2009. Presently, the data has registries in 18 geographic areas with records from over 2 million patients. Registries associated with the SEER program collect data on primary tumor site, tumor morphology and stage at diagnosis, first course of treatment, and follow-up for vital status for patients with many types of cancer. The SEER data is unique in its comprehensive collection of information about the stage of cancer at diagnosis as well as patient survival. SEER registries currently cover approximately 28% of the US population.

The SEER data set has 9 different anatomical areas: breast, colon and rectum, other digestive, female genital, lymphoma and leukemia, male genital, respiratory, urinary, and all other areas. Within each of these subsets there are 134 variables capturing various socio-demographic and cancer specific information as well as SEER data recodes. Some of these 134 variables are cancer specific and thus do not apply to all of the 9 anatomical areas. In the breast cancer subset there are approximately 650,000 observations. For the purposes of this research a subset of this data was extracted, keeping all patients diagnosed between 2004 and 2009. This time frame was selected to create consistency among cancer stage information. In 2004 information regarding the American Joint Committee on Cancer 'stage group' was added to the SEER data. This narrower window also allows for greater consistency among treatment options available to patients. This subset of the data contains 89,694 individuals.

Approximately 75%, if not more, of time spent on this project was spent on data cleaning and preparation. As mentioned early, the breast cancer data subset contains variables that are not applicable to this particular cancer or contain only 'unknown' entries. A basic understanding of each variable, provided in part by the SEER Research Data Record Description, was required to remove the unnecessary variables while retaining the most important and influential. Another large part of the data cleaning process was consolidating the groups for certain variables. For the race/ethnicity variable, the SEER registry provides 29 different race options and an unknown category. This level of specificity was not needed for this study and thus this variable was consolidated into 5 groups: White, Black, Asian, other, and unknown.

Several other variables in the data subset underwent a similar consolidation process for the sake of simplicity and computational convenience.

When reading in the SEER data, all variables are categorical. However, variables such as age, survival time, number of primary tumors, size of tumor, etc. are much better understood as numeric variables. To facilitate this change, unknown entries, such as '99' or '989', needed to be recoded as 'not applicable' or deleted for anaylsis in SAS. In the case of survival time, a YYMM time needed to be converted into a single number representing both the number of years and months of survival after diagnosis. A combination of backward model selection and prior knowledge was used to reduce the remaining variables down to 18 that either the literature felt were important mortality predictors or were shown to be statistically significant. These 18 variables were used in all three models. They can be seen in the in Table 1.

Table 1: Variables Selected for Models

Variable Names	Number of Levels
Marital Status	6
Laterality	5
Grade	5
AJCC Stage Group	11
SEER Summary Stage	7
Surgical Procedure of Other Site	7
ER Status	4
Race	5
Surgery/Radiation Therapies	5
Site Specific Factor 2	4
Method of Radiation Therapy	7
State of Treatment	9
Age	Numeric
Year of Diagnosis	Numeric
Number of Primary Tumors	Numeric
Tumor Size (in mm)	Numeric
Regional Lymph Nodes Examined	Numeric
Positive Regional Lymph Nodes	Numeric

Once the steps described, as well additional ones not mentioned, were completed, the models were able to be fit and give the researcher a clearer picture of mortality trends and survival time prediction.

3 Literature Review

The SEER data set is used by many researchers to study cancer trends in the United States. In the case of breast cancer, much of this research focuses on the correlation between race/ethnicity and incidence or mortality. The SEER data set is ideal for tasks such as these as it is highly representative of the US population and is considered to be one of the most comprehensive information sources for cancer research. In addition to being used to look at cancer trends in the United States, it is also employed regularly to study model development for predicting mortality and survival. Because of the high quality of the data and the myriad of variables captured, it is a powerful tool that is readily available to researchers.

3.1 The SEER Data Set In Use

Each year, the American Cancer Society (ACS) estimates the number of new cancer cases and deaths and makes predictions regarding incidence, mortality and survival for the coming year. Along with data from the Centers for Disease Control and Prevention, the National Center for Health Statistics, and the North American Association of Central Cancer Registries, the ACS uses the SEER data to make these estimations. Although their report is not cancer specific, due to the high prevalence of breast cancer this particular cancer is discussed at length. The methods of analysis are not discussed in the report, but regional variation in cancer incidence was examined. They concluded that "cancers that can be detected by screening or other testing practices", such as breast cancer, vary by state due to "differences in the use of screening tests or detection practices in addition to differences in disease occurrence" [10] .

ACS also provided more specific estimates of breast cancer cases and deaths in an article by Smigal, et al [11]. Employing the same data sets as the annual Cancer Statistics report, they looked at the incidence rates across states. Comparing White and African-American women, they looked at the percent of patients over 40 years of age, percent of patients without health insurance, average incident rates, mortality rates, etc. They found that breast cancer mortality rates are higher in the Northeast compared to other areas of the country. However, there has been a decrease in geographic variation in recent years due to worsening mortality, especially in the South.

Lacey, et al. examined geographic variation in mortality rates (not cancer specific) among white women in the United States using the SEER data set. They found that rates were considerably higher across the Northeast and

lower in the South. They believe this variation is the result of reproductive characteristics and sociodemographic factors. An interest was expressed in the further study of the correlation between dietary or environmental risk factors and geographic mortality variation [8].

The SEER data set (supplemented by data from additional sources) was also used by Jemal, et al. to examine trends in breast cancer incidence rates by age and tumor characteristics. Information on stage at diagnosis, tumor size (in three categories), and ER or PR status was used to construct a joinpoint regression model (fitting a series of joined straight lines on a log scale). Only women over age 40 were included in the analysis [7].

3.2 Methods for Analyzing the SEER Data

The SEER data set has been used to fit many different types of models. Popular in the literature are logistic regression, Cox proportional hazards models, artificial neural networks, and decision trees. It is important to note that extensive data cleaning and preparation were crucial steps in all model building procedures and much of the early stages of research are spent on this task.

3.2.1 Logistic Regression

Logistic regression is commonly used to predict mortality in public health literature because it is able to predict a discrete response variable (death from cancer vs. no death from cancer). Unlike linear regression, logistic regression generates a model that predicts the odds of occurrence. The question of mortality is a two-class problem and therefore, odds greater than 50% are assigned to the class designated "1" and those less than 50% are designated 0 [5].

Delen, et al. used the SEER data set to assess the accuracy of breast cancer predictions generated by two different data mining algorithms (artificial neural networks and decision trees) and logistic regression models. A 10-fold cross-validation method was used to compare the performance of these three models. The logistic regression model generated in this study was of particular interest to this reader. An earlier version of the SEER data set used in this study and contained only 54 input variables (versus the 88 contained in the most current version). After an extensive data cleaning and preparation process 17 variables (16 predictors and 1 response) were selected and 202,932 records were used. The 11 categorical predictor variables used were race, marital status, primary site code, histology, behavior, grade, ex-

tension of disease, lymph node involvement, radiation, stage of cancer, site specific surgery code. The 5 quantitative variables were age, tumor size, number of positive nodes, number of nodes, and number of primaries. Of these variables grade, stage of cancer, radiation, and number of primaries were the most important prognostic factors, respectively. Although the regression model performed the worst, compared to the two data mining algorithms it still achieved a classification accuracy of 0.8920, with a sensitivity of 0.9017 and a specificity of 0.8786 [5].

For his master thesis, Wang used various statistical methods, including logistic regression to examine to effect of various risk and prognosis factors on breast cancer survival **??**. The SEER data set was used. In SAS, a table of odds ratios was generated for the various categories in each factor. Using this information, comparisons within each factor were made. Patients with the lowest 5-year survival probability included those over 65 years of age, African-Americans, unmarried (single) patients, those who did not receive radiation or surgery, etc. Adjustments were made for some risk and prognosis factors but this did not alter the groups identified as having the lowest 5-year survival probability.

Bradley, et al. used a subset of the SEER data set (examining only data from the Metropolitan Detroit Cancer Surveillance System, one of the SEER registries) to examine the effect of race, socioeconomic status, and treatment on breast cancer survival. Additional data was used to supplement the SEER data subset and provide more sociodemographic information. Odds ratios were examined and logistic regression was carried out to estimate the odds of death from cancer. The study concluded that mortality among African-American women was not statistically different from their white counterparts when age, socioeconomic status, and insurance coverage were controlled [3].

To create an effective model for predicting a diagnosis of breast cancer in women (separated into premenopausal and postmenopausal populations), Barlow, et al also used logistic regression [2]. Using data from a variety of sources (not including the SEER program) a model was constructed with a minimal number of predictors and no interaction terms. The most statistically significant predictors for both populations were breast density and age. The four variables include in the model for premenopausal women were age, breast density, number of first-degree relatives with breast cancer, and prior breast procedure. In postmenopausal women all of those variables were included with the addition of race, BMI, age at birth of first child, current hormone therapy use, surgical menopause, and previous mammographic outcome. These models were constructed with 75% of the data and validated

with the remained 25%.

3.2.2 Cox Proportional Hazards Model

The Cox proportional hazard model relates the time of an event, in this case death, to a number of explanatory variables (called covariates). The model assesses the impact of these covariates in predicting survival time, the effect of a unit increase in a covariate related to the hazard rate. This model assumes the covariates are multiplicatively related to the hazard rate.

Tai, et al. , wished to explore the relationship of age and mortality in TI-2 breast cancer using a Cox proportional hazards model and the SEER data. Including variables such as region, year of diagnosis, race, marital status at diagnosis, histology, grade, ER/PR status, tumor size, number of nodes examined and involved, and treatment (surgery vs. radiation), Tai et al. generated hazard ratios, using both the original data and transformed data. A proportional hazards check for age was done, and suggested that mortality increases linearly (for the log hazard ratio) with each year of age and at age 50, the risk increases quasi-quadraticaly. The study concluded the relationship between age and mortality is biphasic and that young women experience a higher risk of death then their older counterparts and risk of death from cancer does not decrease with increasing age [12].

To model the effect of tumor size in early breast cancer, Verschraegen, et al. used a similar log-hazard analysis on the SEER data set [13]. They found that patients with both node-negative and node-positive breast cancer have a linear increase of mortality with tumor size until the tumor reaches 30mm. Beyond 30mm (up to 50mm), mortality plateaus. Hazard ratios of untransformed covariates, hazard ratios of gompertzian transform size, and change of crude death rate were examined to arrive at this conclusion. It was historically thought that axillary lymph node involvement was the single most important predictor in overall survival; however this finding questions nodal involvement as a predictor of death or disease-free survival.

In addition to the logistic regression carried out by Wang, a Cox proportional hazards model was used to check the effect of various risk and prognostic factors on overall breast cancer survival using the SEER data [1]. A reference category in each factor was first selected to be compared to other levels within that same factor. Wang (2012) concluded that age, race, tumor size, lymph node involvement and/or distant metastasis, negative tumor markers, and poorly differentiated/undifferentiated tumors all greatly affected mortality risk.

3.2.3 Artificial Neural Networks

Artificial neural networks (ANN) model extremely complex non-linear functions. These models, inspired by biological neural networks, consist of interconnected groups (which receive one or more inputs and generate an output) and adapt its structure as it learns. Models such as these are used to find patterns in large data sets. The studies examined here used a popular ANN structure called multi-layer perceptron with back-propagation.

In both his 2004 (breast cancer) and 2009 (prostate cancer) studies of data mining methods to predict cancer survivability, Delen et al examined artificial neural networks using the SEER data set [5] [4]. These models were evaluated using measures of classification accuracy, sensitivity, and specificity. In the case of breast cancer, the ANN model performed best in all three areas. The ANN model for prostate cancer came in second to the support vector machine model. The sensitivity analysis on the ANN, although criticized, provides insight about which variables are used and how much they contribute to the dependent variable. In the case of Delens' study on breast cancer, grade, number of primaries, stage of cancer, radiation, and number of lymph nodes were the top five contributors.

3.2.4 Decision Trees

Decisions trees are classification algorithms that separate observations, creating a tree like structure with the goal of improving prediction accuracy. A variety of mathematical algorithms are used (i.e. Chi-squared test) to identify a threshold for a variable that splits the incoming observations into two or more subgroups. This process is repeated until all significant 'leaves are identified. Individual nodes from a tree may be used as predictor variables in logistic regression [6]. Alternatively, combinations of trees (random forests) may be used as prediction models.

Decision trees proved most accurate, sensitive, and specific in Delens 2004 study of data mining methods to predict breast cancer survivability [5].

3.2.5 Crude Cumulative Probabilities of Death

Crude cumulative probabilities of death can also be used to measure patient prognosis. This probability represents the chance a patient will die from his/her cancer given other causes of death. This measure assures mortality patterns experienced within a group of cancer patients who are also affected by non-cancer related causes of death.

Schairer, et al used the theory of competing risks to calculate crude cumulative probabilities of death from breast cancer to assess the burden of mortality by age and race. Using the variables stage, race, age, tumor size, and ER status collected by the SEER program, cumulative weighted differences were used to compare white and African-American subpopulation. They concluded that patients less than 50 years of age of both races were most likely to die from their cancer if the disease was localized. At older ages, the probability of death from other causes (not cancer) became increasingly important. Overall, the probability of death from breast cancer was statistically significantly greater in African-American patients than in their white counterparts at all ages [9].

3.3 Discussion

Breast cancer trends and their relation to certain factors are a much studied topic, greatly aided by the wealth of information found in data sets like SEER. Previously conducted studies have found mortality and incidence trend variation by geographic location, which supports the findings of the research presented here. There is also support for the use of logistic regression model in the current literature, as well as the specific predictor variables chosen by the stepwise selection procedure conducted in SAS. The studies presented here represent a small fraction of the available research regarding breast cancer mortality trends and model development. These specific articles were included because of their relevance to the research being conducted and their availability to the researcher.

4 Methods and Models

To predict mortality we used both a logistic regression model as well as a mixed effect model. A logistic regression model was chosen because of its ability to predict a discrete response variable and its prevalence in existing public health literature. A mixed effect model was used because we believe variation in mortality can be explained by examining state subpopulations. To explore time until death a Cox proportional hazards model was used. The relationship between location and time until death was also examined using this model. Each model includes 18 predictor variables and 1 response variable. See table 1 for a list of predictor variables. No interaction terms were used in any of the models generated.

4.1 Logistic Regression

A logistic regression model has a binary response variable, taking on the values 1 and 0. Y is a Bernoulli random variable with the parameter $E(Y) = P$. This can be written:

$$Y_i = \begin{cases} 1 & \text{patient died of cancer} \\ 0 & \text{else} \end{cases}$$

$$Y_i \overset{iid}{\sim} Bern(P_i)$$

$$logit(P_i) = \log_e \left(\frac{P_i}{1 - P_i} \right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}$$

$$\text{for } i = 1, 2, \cdots, n \text{ and } p = \text{number of parameters}$$

The coefficients can be interpreted as the estimated odds, $\frac{P_i}{1-P_i}$, multiplied by $exp(\beta_i)$ for any unit increase in X, while holding all other variables are held constant. To fit this model, `proc logistic` in SAS was used. This model was assessed using cross validation. This cross validation was done by fitting the model with the complete data set and then using predicted probabilities to conduct a receiver operating characteristic (ROC) analysis. Using the predicted probabilities to fit a new model simulates the process of fitting the model ignoring a single observation and then estimating that predicted probability using the fitted model. We can be assured of the models predictive abilities by using the ROC contrast test which tests whether the fitted model is better than a uninformative model applied to the validated data set. This test will be explored in further detail in the "Results" section.

4.2 Mixed Effects Model

Mixed effects models contain both fixed and random effects. In this case, location of treatment was used as the random effect. This allows us to make inferences about each state in comparison to the other states. The mixed effect model is of the following form (note: τ_i is the random effect):

$$logit(P_i) = (\beta_0 + \tau_{state_i}) + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}$$

$$\text{where} \sum_{i=}^{n} (\tau_{state_i}) = 0$$

To fit this model `proc glimmix` in SAS was used, specifying the random intercept to be location of treatment (STCOUNTY). A covariance test of the random effects was conducted to test whether a model **without** random location effects fits the data as well as the model **with** random location effects. The results of this covariance test will be examined later. An analysis of the intercept for each state can also help us identify the relationship between states in regards to breast cancer mortality, the goal of this project. It is important to note that using a random intercept model only allows inference on populations, not on individuals.

4.3 Cox Proportional Hazards Model

The Cox proportional hazards model relates the amount of time that passes before in event (in this case death) to certain predictor variables, often called covariates. This model assumes the effect of the predictors is the same at all times and that the covariates act additively on the time until death. $h(t)$ is is called the baseline hazard. This model is given by:

$$h(t|X) = h(t)exp(X_1\beta_1 + \cdots + X_p\beta_p)$$

In this instance, we are more interested in the parameter estimates than the shape of the hazard. The parameter estimates help us identify which predictors have the greatest effect on time until death. This model was fitted in SAS using `proc.phreg`. With this data set, 96.45% of the data is censored - meaning that 96.45% of individuals in the data set have not yet died (either as a result of their cancer or from other causes). The Cox proportional hazards model is useful when dealing with censored data such as this. To examine the survival distribution function estimates by state, `proc.lifetest` was used. It generated a graph which displayed the odds of survival over time by state. This also helped address the relationship between breast cancer mortality and location of treatment.

5 Results

5.1 Logistic Regression Results

Although the parameter estimates are of some interest with the logistic regression model, we will first discuss the ROC analysis to show the predictive accuracy of the model. The area under the ROC curve (AUC) is commonly used to assess the predictive accuracy of a logistic regression model. The

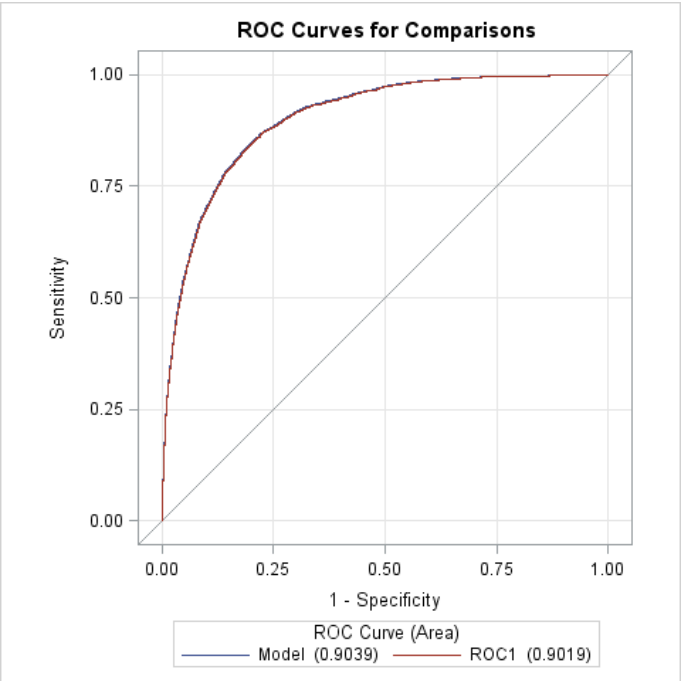

Figure 1: ROC Comparison

AUC for the fitted model applied to the SEER data set is 0.9039. When applied to the predicted probabilities validation set, the AUC is 0.9019. The graph below (see Figure 1) shows both curves (the one generated by the SEER data set and the one generated by the cross-validation set). These curves lie almost exactly on top of each other. This is a good indication of strong predictive ability and robustness of a model.

The ROC contrast test results give a chi-squared test statistic of 161.1830 with a corresponding p-value of $< .0001$ (see Table 2). This test indicates the fitted model is better than an uninformative model applied to the validation data set.

Table 2: ROC Contrast Test Results

Contrast	DF	Chi-Square	P-value
Reference=Model	1	161.1830	$< .0001$

A full table of the parameter estimates is available in the appendix. Here we will examine the parameter estimates for the different states (Table 3). Individual states with the largest positive parameter estimates (i.e. New Mexico and Utah) have a higher likelihood of dying from cancer than individuals from states with large negative parameter estimates (i.e. Connecticut and California). The 2009 United Health Foundation rankings are listed along the states. There does not seem to be a pattern between the state's health ranking and a greater likelihood of dying from this cancer.

Table 3: Logistic Regression Parameter Estimates

State Health Ranking (2009)	State	Parameter Estimate
31	New Mexico	0.2056
2	Utah	0.1796
4	Hawaii	0.0430
30	Michigan	0.0243
15	Iowa	0.0134
11	Washington	0
43	Georgia	-0.0091
23	California	-0.1210
7	Connecticut	-0.2110

Although no pattern is apparent in this model, we will continue to examine the mixed effects model and the cox proportional hazards model.

5.2 Mixed Effect Model Results

First we will examine the covariance test (Table 4) of the random effects to test whether the inclusion of location of treatment random effects improves the model. The -2 Residual Log Pseudo-Likelihood is 678,682 with a p-value of $< .0001$. This small p-value indicates that the model **with** random location of treatment effects fits the data better than a model without random location of treatment effects. This indicates the cancer mortality can be better understood based on a hierarchy of populations, in this case, by state.

Again, we will examine the estimates for the location of treatment variable (Table 5). However, this time these numbers represent the random intercept associated with the model for each state. (Note: all other variables included in the model are fixed)

Table 4: Test of Covariance Parameters

Label	DF	-2 Log Res Log P-Like	Chi-Square	P-value
No random effect	1	678682	14.52	$< .0001$

Table 5: Solutions for Random Effects

State Health Ranking (2009)	State	Parameter Estimate
31	New Mexico	0.1446
2	Utah	0.1354
30	Michigan	0.0287
4	Hawaii	0.0281
15	Iowa	0.0169
43	Georgia	0.0015
7	Connecticut	-0.0929
11	Washington	-0.0968
23	California	-0.1654

Although the states in this table is not in the exact same order as the table generated by the logistic regression procedure, they are similar. The four states with the largest positive parameter estimates are the same and the four states with the largest negative parameter estimates are the same, with Iowa in the middle in both cases. Again, there does not appear to be a correlation between state health ranking and cancer mortality trends. However, some consistency between models is seen.

5.3 Cox Proportional Hazards Model Results

Before we look at the Cox proportional hazards model, we will look at the graph of probability of survival vs. time since diagnosis (in years). In this graph (Figure 2) it appears the states with the most negative slopes, indicating lower chances of survival as time increases, are New Mexico, Utah, Georgia, and Michigan. Once again, Iowa appears to be in the middle (between the lower four and the upper four). The states with the least negative slopes are Connecticut, California, Hawaii, and Washington. It is important to notice that the lowest survival probability is 0.88. This is an indication of the progress made in treatment technology and the greatly improved odds of breast cancer survival.

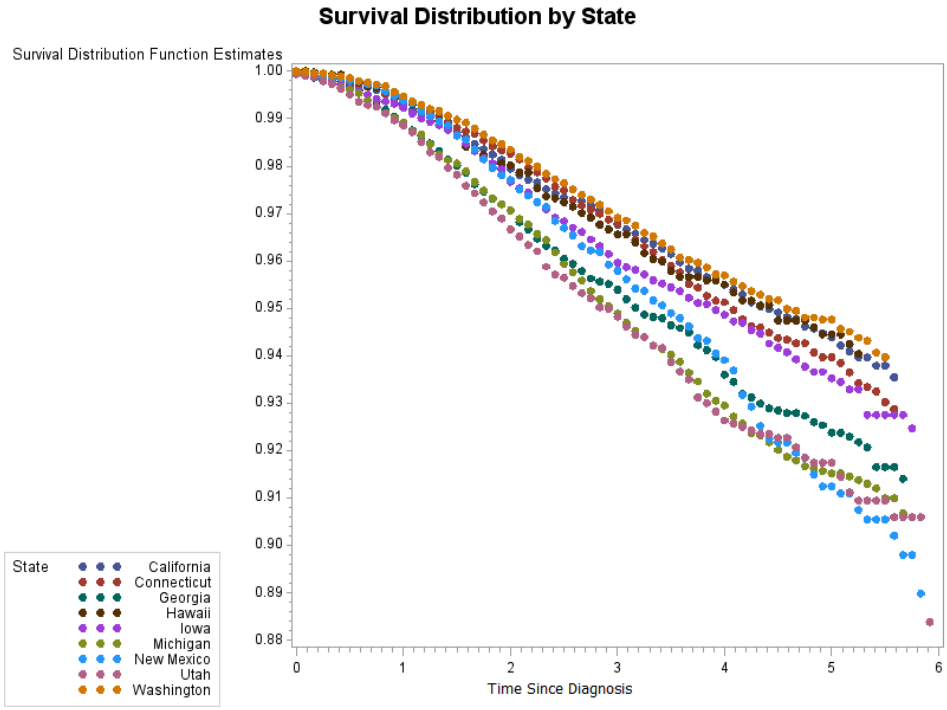

Figure 2: Probability of Survival vs. Time Since Diagnosis

Now we will look at the parameter estimates from the Cox proportional hazards model as well as the hazard ratios (Table 6). When looking at the parameter estimates smaller (or negative) values are 'better', meaning the hazard of death is less. Although there still does not appear to be a trend between health rankings by state and hazard of death we see a similar grouping of the lowest four states and the highest four states (with Iowa, again, in the middle).

Another very interesting number to consider is the hazard ratio. In this model the baseline group is Washington state. Consequently, the hazard of death from breast cancer for individuals from Connecticut is 2.8% lower ($.972 - 1 = -.028$) than individuals from Washington. At the other extreme, the hazard of death from breast cancer for individuals from Utah is 36.8% **higher** ($1.35 - 1 = .35$) than individuals from Washington. All other comparisons between other states and Washington can be made in a similar fashion. To compare a state to another state (other than Washington) the two hazard ratios can be divided. For example, individuals from Utah are 1.01% ($1.368 \div 1.35$) more likely to die from cancer than individuals from New Mexico.

Table 6: Proportional Hazards Model Parameter Estimates & Hazard Ratios

State Health Ranking (2009)	State	Parameter Estimate	Hazard Ratios
2	Utah	0.3133	1.368
31	New Mexico	0.3013	1.35
4	Hawaii	0.2246	1.25
30	Michigan	0.1653	1.18
15	Iowa	0.1648	1.18
43	Georgia	0.1527	1.17
23	California	0.0145	1.02
11	Washington	-	-
7	Connecticut	-0.0289	.972

6 Discussion

As seen in the results section above, there was not a direct relationship between a state's health ranking and breast cancer mortality. This is not suprising since the State Health Care Rankings are a compositie measure of many factors. However, these models are useful both for predictive purposes and the identification of other trends in cancer mortality across the country.

6.1 Conclusions

For this project we wished to explore the hypothesis that cancer mortality was related to the United Health Foundation's State Health Rankings using the SEER data set. We used a subset of the breast cancer data set, with all patients diagnosed between 2004 and 2009. After extensive data cleaning, three different models were fitted to gain a better understanding of mortality trends. We suspected states with lower healthfulness scores would indicate higher cancer mortality rates. However, after examining the three different models, we concluded that this was not the case. Although, we did not find the trends we expected, all three models are useful in their own right.

The logistic regression model has a classification accuracy of 90.39%. This means that using a subset of 18 variables in the SEER data set, we can correctly predict cancer mortality 9 out of 10 times. Furthermore, an examination of the parameter estimates provides a better understanding of which variables contribute most to breast cancer mortality.

The covariance test conducted with the mixed effects model concluded

that the inclusion of a location of treatment random effect better explained the data than a model without it. This allows us to further conclude that the variation in mortality is related to state subpopulations. Although these subpopulations were not related in the same way as described by the United Health Foundation's State Health Rankings, they may be related to other socioeconomic or demographic characteristics.

Lastly, the Cox proportional hazards model attempted to systematically predict years of survival after diagnosis given a subset of 18 variables found in the SEER data set. Although, probability of survival decreases as time increases, the negative slope is small. This is most likely the result of improved screening and early detection (i.e. more frequent mammograms) and technological advances in treatment. The results of this analysis were very encouraging in that regard.

There is some variation in the mortality estimates for each state in each of these three models. Although this data set is large, not all states are represented equally. We hypothesize that some of the variation between models may be due to unequal sampling.

Although massive data sets such as the SEER data set and various modeling techniques are capable of finding patterns and relationships, it is important to remember that the results have little meaning without the cooperation and support of the medical community. The conclusion of analyses such as this need the review of the medical community before it can direct public health efforts on a national scale. A cooperation between the statistics community and the medical community has the opportunity to discover many new trends in health across the nation and large and comprehensive data sets, such as the SEER data, are able to help.

6.2 Future Work

It would be interesting in future work to look at other possible information to explain the variation in mortality by state, whether this be an alternative health ranking, economic information, or cultural and social information. Also, much research has been done with regards to race/ethnicity and its relation to cancer mortality. Fitting a mixed effects model, with race as the random intercept, might be a powerful tool in quantifying this relationship and helping direct public health resources appropriately.

Lastly, this project also set out to explore the relationship between radiation therapy and cancer reoccurrence. Due to time constraints this question was not able to be addressed. Using propensity scores and the SEER data set, it may be possible to look for epidemiological evidence to support research

that identifies a link between exposure to radiation and cancer reoccurrence.

This project has done much to enhance the researcher's understanding of statistical modeling as well as expose her to both the power and the problems associated with massive data sets. She hopes to continue work in both health related statistics and big data methodology.

7 References

[1] Effect of risk and prognosis factors on breast cancer survival: Study of a large dataset with a long term follow-up. Master's thesis, Georgia State University, 2012.

[2] William E Barlow, Emily White, Rachel Ballard-Barbash, Pamela M Vacek, Linda Titus-Ernstoff, Patricia A Carney, Jeffrey A Tice, Diana SM Buist, Berta M Geller, Robert Rosenberg, et al. Prospective breast cancer risk prediction model for women undergoing screening mammography. *Journal of the National Cancer Institute*, 98(17):1204–1214, 2006.

[3] Cathy J Bradley, Charles W Given, and Caralee Roberts. Race, socioeconomic status, and breast cancer treatment and survival. *Journal of the National Cancer Institute*, 94(7):490–496, 2002.

[4] Dursun Delen. Analysis of cancer data: a data mining approach. *Expert Systems*, 26(1):100–112, 2009.

[5] Dursun Delen, Glenn Walker, Amit Kadam, et al. Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial intelligence in medicine*, 34(2):113–128, 2004.

[6] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[7] Ahmedin Jemal, Elizabeth Ward, and Michael Thun. Recent trends in breast cancer incidence rates by age and tumor characteristics among us women. *Breast Cancer Research*, 9(3):R28, 2007.

[8] James V Lacey, Susan S Devesa, and Louise A Brinton. Recent trends in breast cancer incidence and mortality. *Environmental and molecular mutagenesis*, 39(2-3):82–88, 2002.

[9] Catherine Schairer, Pamela J Mink, Leslie Carroll, and Susan S Devesa. Probabilities of death from breast cancer and other causes among female breast cancer patients. *Journal of the National Cancer Institute*, 96(17):1311–1321, 2004.

[10] Rebecca Siegel, Deepa Naishadham, and Ahmedin Jemal. Cancer statistics, 2012. *CA: A Cancer Journal for Clinicians*, 62(1):10–29, 2012.

[11] Carol Smigal, Ahmedin Jemal, Elizabeth Ward, Vilma Cokkinides, Robert Smith, Holly L Howe, and Michael Thun. Trends in breast cancer by race and ethnicity: update 2006. *CA: a cancer journal for clinicians*, 56(3):168–183, 2006.

[12] Patricia Tai, Gábor Cserni, Jan Van De Steene, Georges Vlastos, Mia Voordeckers, Melanie Royce, Sang-Joon Lee, Vincent Vinh-Hung, and Guy Storme. Modeling the effect of age in t1-2 breast cancer using the seer database. *BMC cancer*, 5(1):130, 2005.

[13] Claire Verschraegen, Vincent Vinh-Hung, Gábor Cserni, Richard Gordon, Melanie E Royce, Georges Vlastos, Patricia Tai, and Guy Storme. Modeling the effect of tumor size in early breast cancer. *Annals of Surgery*, 241(2):309, 2005.

8 Appendix

Logistic Regression Parameter Estimates

Table 7: Logistic Regression Parameter Estimates

Variable	Parameter Estimate
Intercept	1103.2
Marital Status	
Single (never married)	0.0653
Married	-0.1195
Separated	-0.0526
Divorced	0.1103
Widowed	0.1051
Laterality	
Right: Origin of Primary	-0.3185
Left: Origin of Primary	-0.3204
Only one side involved, not specified	-0.2228
Bilateral Involvement	0.369
Grade	
Grade 1: well differentiated	-0.7130
Grade 2: moderately differentiated	-0.0967
Grade 3: poorly differentiated	0.5598
Grade 4: undifferentiated	0.2312
Stage Group	
Stage 0	-2.1078
Stage I	-1.5775
Stage IIA	-0.6642
Stage IIB	-0.1204
Stage III NOS	0.7982
Stage IIIA	0.2841
Stage IIIB	0.3334
Stage IIIC	0.4179
Stage IV	1.2751
NA	1.4670

Table 7: Logistic Regression Parameter Estimates

Variable	Parameter Estimate
SEER Summary Stage	
In Situ	-1.3378
Localized	-0.0598
Regional, direct extension	0.3357
Regional, lymph nodes only	0.1446
Regional, extension and nodes	0.7944
Distant	0.8479
Surgery at Other Sites	
None, diagnosed at autopsy	-0.3389
Nonprimary surgery performed	0.1991
Nonprimary surgery to other regional sites	0.0652
Nonprimary surgery to distant nodes	-0.2264
Nonprimary surgery to distant site	0.4861
Combination of the above	1.0798
ER Status	
Positive	-0.4381
Negative	0.1804
Borderline	0.4172
Race	
White	0.00887
Black	0.4578
Asian	-0.2965
Radiation Sequence w/ surgery	
NA	-0.0245
Radiation before surgery	0.2960
Radiation after surgery	-0.2772
Radiation before and after surgery	-0.4125
Site Specific Factor 2 (PRA)	
Positive/elevated	-0.2738
Negative/normal	0.2915
Borderline	0.660
Radiation	
None, diagnosed at autopsy	0.1695
Beam Radiation	0.0405
Beam Radiation and	
Radioactive implants or radioisotopes	-0.5112

Table 7: Logistic Regression Parameter Estimates

Variable	Parameter Estimate
Radiation, not specified	0.2797
Recommended, unknown if given	-0.0553
State	
California	-0.1210
Connecticut	-0.2111
Georgia	-.00910
Hawaii	0.0430
Iowa	0.0134
Michigan	0.0243
New Mexico	0.2056
Utah	0.1796
Numeric	
Age	0.0208
Year of Diagnosis	-0.5517
Nodes examined and removed	-0.0275
Nodes found to contain metastases	0.0712
Size of tumor (mm)	0.0012
Number of primary tumors	0.0714

Mixed Effects Model Parameter Estimates

Table 8: Mixed Effects Model Parameter Estimates

Variable	Parameter Estimate
Intercept	1100.91
Marital Status	
Single (never married)	0.1649
Married	-0.01829
Separated	0.04242
Divorced	0.2104
Widowed	0.2048
Unknown	0
Laterality	
Right: Origin of Primary	-0.8415
Left: Origin of Primary	-0.8434
Only one side involved, not specified	-0.7685
Bilateral Involvement	-0.1695
Unknown	0
Grade	
Grade 1: well differentiated	-0.7302
Grade 2: moderately differentiated	-0.1148
Grade 3: poorly differentiated	0.5401
Grade 4: undifferentiated	0.2084
Unknown	0
Stage Group	
Stage 0	-1.9494
Stage I	-1.4214
Stage IIA	-0.5604
Stage IIB	-0.0369
Stage III NOS	0.9606
Stage IIIA	0.4416
Stage IIIB	0.4945
Stage IIIC	0.6283
Stage IV	1.4280
NA	1.6194
Unknown	0
SEER Summary Stage	
In Situ	-0.6185

Table 8: Mixed Effects Model Parameter Estimates

Variable	Parameter Estimate
Localized	0.6646
Regional, direct extension	1.0616
Regional, lymph nodes only	0.8680
Regional, extension and nodes	1.5165
Distant	1.5759
Unknown	0
Surgery at Other Sites	
None, diagnosed at autopsy	0.9061
Nonprimary surgery performed	1.4435
Nonprimary surgery to other regional sites	1.3049
Nonprimary surgery to distant nodes	1.0105
Nonprimary surgery to distant site	1.7366
Combination of the above	2.3293
Unknown	0
ER Status	
Positive	-0.2829
Negative	0.3371
Unknown	0
Borderline	0.5657
Race	
White	0.1690
Black	0.6123
Asian	-0.1376
Unknown	0
Radiation Sequence w/ surgery	
NA	-0.4544
Radiation before surgery	-0.1105
Radiation after surgery	-0.6939
Radiation before and after surgery	-0.8288
Unknown	0
Site Specific Factor 2 (PRA)	
Positive/elevated	-0.1994
Negative/normal	0.3645
Borderline	0.1454
Unknown	0

Table 8: Mixed Effects Model Parameter Estimates

Variable	Parameter Estimate
Radiation	
None, diagnosed at autopsy	0.1420
Beam Radiation	-0.0029
Beam Radiation and Radioactive implants or radioisotopes	0.0432
Radiation, not specified	.2558
Recommended, unknown if given	-0.8736
Unknown	0
Numeric	
Age	0.0208
Year of Diagnosis	-0.5513
Nodes examined and removed	-0.0272
Nodes found to contain metastases	0.0712
Size of tumor (mm)	0.0001
Number of primary tumors	0.0697

Cox Proportional Hazards Model Parameter Estimates

Table 9: Cox Proportional Hazards Parameter Estimates

Variable	Parameter Estimate	Hazard Ratio
Marital Status		
Single (never married)	0.1616	1.175
Married	-0.0187	0.982
Separated	0.0529	1.054
Divorced	0.1590	1.172
Widowed	0.2098	1.233
Laterality		
Right: Origin of Primary	-0.8560	0.425
Left: Origin of Primary	-0.8599	0.423
Only one side involved, not specified	-0.0304	0.970
Bilateral Involvement	-0.9505	0.387
Grade		
Grade 1: well differentiated	-0.7848	0.456
Grade 2: moderately differentiated	-0.1737	0.841
Grade 3: poorly differentiated	0.4273	1.533
Grade 4: undifferentiated	.1101	1.116
Stage Group		
Stage 0	-1.9806	0.138
Stage I	-1.4622	0.232
Stage IIA	-0.55525	0.574
Stage IIB	-0.0087	0.991
Stage III NOS	0.8037	2.234
Stage IIIA	0.4117	1.509
Stage IIIB	0.4434	1.558
Stage IIIC	0.6091	1.839
Stage IV	1.8730	3.278
NA	1.33065	3.784
SEER Summary Stage		
In Situ	-0.4699	0.625
Localized	0.8190	0.313
Regional, direct extension	1.1691	3.187
Regional, lymph nodes only	1.0303	2.802
Regional, extension and nodes	1.6255	5.081

Table 9: Cox Proportional Hazards Parameter Estimates

Variable	Parameter Estimate	Hazard Ratio
Distant	1.6925	5.443
Surgery at Other Sites		
None, diagnosed at autopsy	0.5137	1.671
Nonprimary surgery performed	0.9291	2.532
Nonprimary surgery to other regional sites	0.7498	2.117
Nonprimary surgery to distant nodes	0.5013	1.651
Nonprimary surgery to distant site	1.2505	3.492
Combination of the above	2.3299	10.277
ER Status		
Positive	-0.3385	0.713
Negative	0.2397	1.271
Borderline	0.4741	1.607
Race		
White	0.1232	1.131
Black	0.5452	1.725
Asian	-0.1394	0.870
Radiation Sequence w/ surgery		
NA	0.1075	1.113
Radiation before surgery	0.00531	1.005
Radiation after surgery	-0.4117	0.663
Radiation before and after surgery	-0.5554	0.574
Site Specific Factor 2 (PRA)		
Positive/elevated	-0.1889	0.828
Negative/normal	0.3186	1.375
Borderline	0.1121	1.119
Radiation		
None, diagnosed at autopsy	0.0915	1.096
Beam Radiation	0.0834	1.087
Beam Radiation and Radioactive implants or radioisotopes	-0.6654	0.514
Radiation, not specified	0.3333	1.396
Recommended, unknown if given	-0.2664	0.766

Table 9: Cox Proportional Hazards Parameter Estimates

Variable	Parameter Estimate	Hazard Ratio
State		
California	0.0145	1.015
Connecticut	-0.0289	0.972
Georgia	0.1527	1.165
Hawaii	0.2246	1.252
Iowa	0.1648	1.179
Michigan	0.1653	1.180
New Mexico	0.3013	1.352
Utah	0.3133	1.368
Numeric		
Age	0.0203	1.021
Year of Diagnosis	-0.0519	0.949
Nodes examined and removed	-0.0283	0.972
Nodes found to contain metastases	0.05713	1.059
Size of tumor (mm)	0.00006	1.000
Number of primary tumors	-0.0055	0.994