11-5-2010

Statistical Analysis and Modeling of Breast Cancer and Lung Cancer

Chunling Cong
University of South Florida

Statistical Analysis and Modeling of

Breast Cancer and Lung Cancer

by

Chunling Cong

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Mathematics and Statistics
College of Arts and Sciences
University of South Florida

Major Professor: Chris Tsokos, Ph.D.
Gangaram Ladde, Ph.D.
Kandethody M. Ramachandran, Ph.D.
Wonkuk Kim, Ph.D.
Marcus McWaters, Ph.D.

Date of Approval:
November 5, 2010

Keywords: decision tree, survival analysis, accelerated failure model,
Cox proportional hazard model, Kaplan-Meier

Dedication

I would like to dedicate my dissertation to my parents, Weihua Cong and Lanzi Peng and my sister Lian Cong for their unconditional support and trust in me which enables me to be confident in myself.

I would also like to dedicate my dissertation to my advisor Dr. Chris Tsokos. In the past three years, his continuous support and encouragement inspired me to never giving up the dream of becoming not only a good statistician, but also a better person. He has also given me lots of opportunities to achieve whatever goal I may have. I have been truly enjoying the wonderful mentoring relationship with him. He made the whole Ph.D. process a journey full of endeavors, learning experience and feeling of accomplishments.

Acknowledgements

First of all, I would like to thank my dissertation committee for their great efforts in refining my research design and writing. Dr. Chris Tsokos has been given my help and suggestions throughout the whole process and my growing interest in research to a good extent owes to the helpful inputs from Dr. Wonkuk Kim. Dr. Gangaram Ladde has been always very kind and supportive in every step of the way. Dr. Kandethody M. Ramachandran certainly also contributed a lot to my pleasant and rewarding dissertation process.

I am truly grateful for the internship opportunity from American Cancer Society and valuable inputs provided by Dr. James Kepner which makes the current research applicable to real world problems. At last, I would like to extend my appreciation to all the members in the cancer research team, their hard work and cooperation made my statistical analysis possible.

Table of Contents

List of Tables

List of Figures

Abstract

The objective of the present study is to investigate various problems associate with breast cancer and lung cancer patients. In this study, we compare the effectiveness of breast cancer treatments using decision tree analysis and come to the conclusion that although certain treatment shows overall effectiveness over the others, physicians or doctors should discretionally give different treatment to breast cancer patients based on their characteristics. Reoccurrence time of breast caner patients who receive different treatments are compared in an overall sense, histology type is also taken into consideration. To further understand the relation between relapse time and other variables, statistical models are applied to identify the attribute variables and predict the relapse time. Of equal importance, the transition between different breast cancer stages are analyzed through Markov Chain which not only gives the transition probability between stages for specific treatment but also provide guidance on breast cancer treatment based on stating information.

Sensitivity analysis is conducted on breast cancer doubling time which involves two commonly used assumptions: spherical tumor and exponential growth of

tumor and the analysis reveals that variation from those assumptions could cause very different statistical behavior of breast cancer doubling time.

In lung cancer study, we investigate the mortality time of lung cancer patients from several different perspectives: gender, cigarettes per day and duration of smoking. Statistical model is also used to predict the mortality time of lung cancer patients.

Chapter 1

Introduction

1.1 Breast Cancer and Lung Cancer

Cancer is a class of diseases when a cell or group of cells display uncontrolled growth, invasion and sometimes spread to other locations in the body via lymph or blood (metastasis). It causes about 13% of all human deaths in 2007 with a total of 7.6 million affecting people at all ages. Although there are many causes of cancer, 90-95% of cancer is caused due to lifestyle and environmental factors and 5-10% are due to genetics.

1.1.1 Breast cancer

According to an authoritative source of information on cancer incidence and survival in the United States: the Surveillance, Epidemiology and End Results (SEER) Program of the National Cancer Institute (NCI) collects and publishes cancer incidence and survival information from around 28 percent of the US population, it is estimated that 207,090 women will be diagnosed with and 39,840 women will die of cancer of breast in year 2010. From the statistics based on 2003 to 2007, the median age at diagnosis for cancer of the breast was 61 years old, and the incidence rate was 122.9 per 100,000 women per year. From the

same data source, the median age at death for breast cancer was 68 years old with a death rate of 24.0 per 100,000 women per year. The overall 5-year survival for 1999-2006 was as high as 89.0%. Based on rates from 2005-2007, 12.15% of women born today will be diagnosed with cancer of the breast at some time during their lifetime. In other words, 1 in 8 women will be diagnosed with breast cancer during their lifetime.

1.1.2 Lung Cancer

For lung cancer, it is estimated that 222,520 men and women will be diagnosed with and 157,300 men and women will die of lung and bronchus cancer in 2010. Based on NCI's SEER Cancer Statistics Reivew, the incidence rate was 62.5 per 100,000 men and women per year and the median age at death for lung and bronchus cancer was 72 years old, based on the cases diagnosed in 2003-2007 from 17 SEER geographic areas. The death rate was 52.5 per 100,000 men and women per year. Overall 5-year survival rate was 15.8% based on data from 1999-2006 with highest survival rate 18.6% for white women and lowest rate 11.3% for black men. As to the lifetime risk, 6.95% men and women born today will be diagnosed with lung and bronchus cancer at some point during their lifetime. In other words, 1 in 14 men and women will be diagnosed with lung or bronchus cancer during their lifetime.

Of all cancer incidences among women, breast cancer comprises 10.4% worldwide and it is the most common type of non-skin cancer in women and the fifth most common cause of cancer death. The primary epidemiologic and risk

factors identified are sex, age, lack of childbearing or breastfeeding, higher hormones level, race and economic status.

The most common cause of cancer-related death in men and women is lung cancer, responsible for 1.3 million deaths worldwide annually. It is a disease of uncontrolled cell growth in tissues in lung.

Due to the high incidence rate and death rate cause by breast cancer and lung cancer, significant amount of statistical analysis has been done on causes of cancer, treatment effectiveness, transition between cancer stages, prediction of reoccurrence time and survival time.

1.2 Decision tree

Recently, the decision tree analysis plays a very significant role in the analysis and modeling of various types of medical data, especially in cancer research. In addition, decision tree analysis has been extensively used in areas in the financial world, for example, loan approval, portfolio management, health & risk assessment, insurance claim evaluation, supply chain management, etc. It is also widely applied in fields such as engineering, forensic examination and biotechnology. The objective of present study is to review the theory behind decision tree analysis and to illustrate its usefulness by applying the subject area to various applications. Furthermore, statistical software information is given to assist scientists in applying decision tree analysis.

1.2.1 Introduction to Decision Tree

A decision tree as a visual and analytical decision support tool is a hierarchical tree structure. Inductive machine learning algorithms are used to learn the decision function stored in the data of the form $(X,Y)=(X_1,X_2,X_3...X_k,Y)$ that maps some sets of attributes $(X_1,X_2,X_3...X_k,Y)$ to the conclusion about some target variable Y, and then the target variable Y can be classified or predicted as necessary. The attributes could be any type of variables and based on the type of the outcomes that we are interested in, a decision tree can be called classification tree in descriptive manner if the outcome is discrete or regression tree in a predictive manner if there are continuous outcomes.

The theory of a decision tree has the following main parts: a “root” node is the starting point of the tree; branches connect nodes showing the flow from question to answer. Nodes that have child nodes are called “interior” nodes. “leaf” or “terminal” nodes are nodes that do not have child nodes and represent a possible value of target variable given the variables represented by the path from the root. The following graphs are two examples of decision trees of 320 breast cancer patients who received the medical treatment of tamoxifen and radiation and 321 patients who received tamoxifen alone respectively. The target variable is relapse time, and the attributes are age, hgb, hist, nodediss, hrlevel, pathsize (will be explained in detail in section 3). As can be seen Figure 1.1 and 1.2, not all attributes are used to split the nodes. The next section explains the

mathematical algorithms of how to construct a decision tree including how an attribute and the value of attribute are chosen to split a given node.

There are several advantages of decision tree over other classification theory tools that make decision tree popular besides its simplicity and interpretability. The approach is supervised learning that given a training data that consist of input and output, we can induce a decision tree even with little hard data; it performs well with large data in a short time, and other statistical or mathematical techniques can be easily incorporated in it.

1.2.2 Theory behind Decision Tree Analysis

The basic idea of decision tree analysis is to spit the given source data set into subsets by recursive portioning of the parent node into child nodes based on the homogeneity of within-node instances or separation of between-node instances with respect to target variables. For each node, attributes are examined and the splitter is chosen to be the attribute such that after dividing the nodes into two child nodes according to the value of the attribute variable, the target variable is differentiated to the best using algorithm. Because of this, we need to be able to distinguish between important attributes, and attributes which contribute little to overall decision process. This process is repeated on each child node in a recursive manner until splitting is either non-feasible or all certain pre-specified stopping rules are satisfied.

Classification & Regression Tree is a decision tree algorithm (L. Breiman, 1984) is a non-parametric probability distribution free technique to construct binary classification or regression trees as shown in Figure 2.1. Splitting points – attribute variables and values of chosen variables – are chosen based on Gini impurity and Gini gain are given by:

$$i(t) = 1 - \sum_{i=1}^{m} f(t,i)^2 = \sum_{i \neq i} f(t,i)\, f(t,j)$$

$$\Delta i(s,t) = i(t) - P_L \cdot i(t_L) - P_R \cdot i(t_R),$$

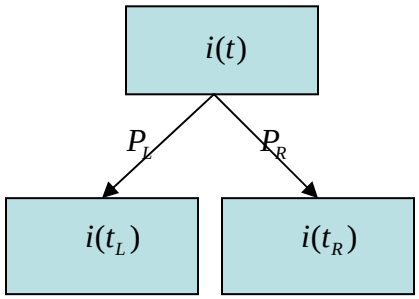

Figure 1.1. CART

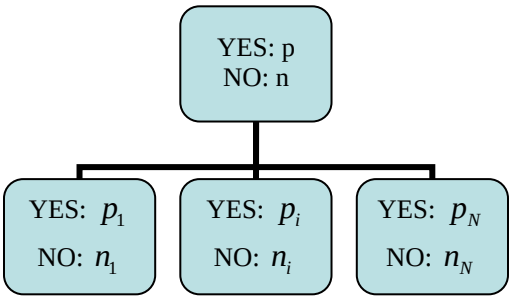

Figure 1.2 ID.3

where $f(t,i)$ is the probability of getting i in node t, and the target variable takes values in {1,2,3…m}. P_L is the proportion of cases in node t divided to the left child node and P_L is the proportion of cases in t sent to the right child node. If the target variable is continuous, the split criterion is used with the Least Squares Deviation (LSD) as impurity measure. If there is no Gini gain or the preset stopping rule are satisfied, the splitting process stops.

CHAID (Chi-Squared Automatic Interaction Detection) classification technique introduced by Kass (1980) for nominal predictors and extended by Magidson (1993) to ordinal predictors is another effective approach for nominal or ordinal target variable. CHAID exhausts all possible pairs of categories of the target variable and merge each pair until there is no statistically significant difference within the pair using Chi-square test.

ID.3 (Iterative Dichotomiser 3) developed by Ross Quinlan (1986) is a classification tree used the concept of information entropy first brought in a publication by Claude Shannon and Warren Weaver (1949) . This provides a method to measure the number of bits each attribute can provide, and the attribute that yields the most information gain becomes the most important attribute and it should go at the top of the tree. Repeat this procedure until all instances in the node are in the same category.

As shown in Figure 2.2, It works in the following manner. Suppose there are only two outcomes “Yes” and “No” in the root node T of target variable. Let p and n denotes the number of “positive records and negative records, respectively. The initial information entropy is given by:

$$I(p,n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n},$$

If attribute X with values $\{x_1, x_2, \ldots, x_N\}$ is chosen to be the split predictor and partition the initial node into $\{T_1, T_2, \ldots, T_N\}$, and p_i and n_i denotes the number of

positive records and negative records in the child node i. Then the expected information $EI(X)$ and information gain $G(X)$ are given by,

$$EI(X) = \sum_{i=1}^{N} \frac{p_i + n_i}{p + n} \cdot I(p_{i,} n_i),$$

Information gain $G(X) = I(p,n) - EI(X)$.

In 1993, Ross Quinlan made several improvements to ID.3 and extended it to C4.5**.** Unlike ID.3 which deals with discrete attributes, C4.5 handles both continuous and discrete attributes by creating a threshold to split the attribute into two groups, those above the threshold and those that are up to and including the threshold. C4.5 also deals with records that have unknown attribute values. C4.5 algorithm used normalized information gain or gain ratio as a modified splitting criterion of information gain which is the ratio of information gain divided by the information due to the split of a node on the basis of the value of a specific attribute. The reason of this modification is that the information gain tends to favor attributes that have a large number of values.

The best approach in selecting the attribute for a specific node is to choose the one that maximize the given ratio. Stopping rule of C4.5 needs to be pre-specified and it initiated a pruning procedures by replacing branches that do not help with leaf nodes after they are created to decrease overall tree size and the estimated error of the tree. A rule set can be derived from the decision tree constructed by writing a rule for each path from the root node to the leaf node. After C4.5, Quinlan created C5.0 as an extended commercial version of C4.5

featuring a number of improvements including smaller decision trees, weighting different attributes and misclassification types, reducing noise, speed and memory efficiency, support for boosting which gives the trees more accuracy.

As a binary-split algorithm, like CART, QUEST (Quick, Unbiased, Efficient, Statistical Tee) proposed by Loh and Shih in 1997 is a classification algorithm dealing with either categorical or continuous predictor X. Pearson's chi-square test is applied to target variable Y and predictor X's independence if X is a categorical predictor. Otherwise, if X is continuous, ANOVA F test is performed to test if all the difference classes of Y have the same mean of X. In both cases, p-values are calculated and compared to a Bonferroni adjusted threshold to determine if further Levene's F-statistics test needs to be performed to determine if the predictor should be chosen as the split predictor for the node.

Overfitting occurs in large tree models where the model fits noise in the data, such as including some attributes that are irrelevant to the decision-making process. If such a model is applied to data other than the training set, the model may not perform well. There are generally two ways to reduce overfitting: stop growing when data is split not statistically significant, or grow full tree, and then post prune. For example, if Gain of the best attribute at a node is below a threshold, stop and make this node a leaf rather than generating children nodes.

1.2.3 Survival Tree and Random Forest

A decision tree is of great importance in classification and modeling of health-related data and in many situations the data is censored due to various reasons one of which is that some patients left before the end of the period of study. Due to the incompleteness of the data, a special portioning and pruning algorithm should be used to construct a survival tree. Gordon and Olshen (1985) gave the first adaption of CART algorithm in censored data using Wasserstein metrics to measure distances between Kaplan-Meier curves and certain point masses . After that, Segal (1988) extended regression-tree methodology to right-censored target variables by replacing the splitting rules with between-node separation rules based on the Tarone-Ware or Harrington-Fleming classes of two-sample statistics and a new pruning algorithm was also devised, and truncation and time-dependent covariates were included in the method proposed by Bacchetti and Segal (1995). Davis and Anderson (1989) used likelihood-ratio test to split nodes under parametric exponential distribution or within-node constant hazard assumptions. LeBlanc and Crowley used martingale residuals for splitting rule assuming a proportional hazards model and also developed an corresponding efficient pruning algorithm, and the model was extended to time-dependent case assuming the survival times are piecewise exponential by Huang, Chen and Soong (1998). Both Davis and Leblanc algorithms are based on a definition of a within-node homogeneity measure, unlike Segal's algorithm which tried to maximize between-node separation. Su and Fan extended the CART algorithm to multivariate survival data by introducing a gamma distributed frailty to account

for the dependence among survival times based on likelihood ratio test as the splitting function. In addition, this method was extended to competing risks based on proportional hazards for subdistribution of competing risks and deviance was used to grow a tree proposed by Ibrahim, Kudus, Daud and Bakar .

Random forest is an ensemble classifier first developed by Leo Breiman and Adele Cutler in 2001. Random forest has more accuracy than the single-tree model, and handles a very large number of input variables. Besides, it provides a experimental way to detect variable interactions, etc. Instead of using all training data, a random sample of N observations with replacement is chosen to build a tree. In the tree building process, for each node, a random subset of the predictor variables is considered as possible splitters for each node, a predictor excluded from one split is allowed to be used as splitters in the same tree. Repeat the above procedure until a large number of trees are constructed. The average of the predicted value in regression trees are computed as the predicted value and the most frequently predicted category in the classification trees are considered to be the predicted category.

1.3 Survival Analysis

Survival analysis is widely used in areas that deal with biological organism and failure of mechanical systems. It is a branch of statistical analysis that are commonly seen in engineering, economics or sociology when modeling time to event data, such as death of a cancer patient, failure of a equipment. The difference of survival analysis is that it deals with censoring. Censoring is a form

of missing data problem which is common seen in those above mentioned areas. If it is known only that the time of an event is after some date, it is called right censoring. This often happens when the individual are lost to follow up or when the study ends after a certain period. Similarly, if the time of event is know to be less than a certain time; the data is called left censoring. The object of primary interest is to investigate the time of event or the probability of the occurrence of an event after certain time, which is also called a survival probability. Time-to-event data are increasingly common in health research, particularly in longitudinal or cohort studies where the onset of certain health outcomes is observed. The timing of event onset, in addition to the outcome event (e.g. cure of disease, development of a symptom, death), provides important information about disease progression or treatment effects. Furthermore, the outcome may not be observed on every study subject because of limitations in the study design. For example, a study may terminate before a subject develops the symptom of interest. This characteristic of incomplete observation is called censoring, must be considered in evaluating the study.

A survival function measures the probability of nonevent after certain time which defined as

$$S(t) = \Pr(T > t),$$

where t is some time, and T is a random variable denoting the time of an event. According to definition, a survival function is always between 0 and 1. It must be non-increasing and approaches 0 as time goes to infinitely.

The corresponding lifetime distribution function is defined as

$$F(t) = \Pr(T \le t) = 1 - S(t).$$

The hazard function is defined as the event rate at time t conditional on survival until time t or later.

$$\lambda(t)dt = \Pr(t \le T \le t + dt \mid T \ge t) = \frac{f(t)dt}{S(t)} = -\frac{S'(t)dt}{S(t)}.$$

1.3.1 Kaplan – Meier Estimator

Nonparametric analysis is used to analyze data without assuming an underlying distribution which avoids potentially larges errors brought about by making incorrect assumptions about the underlying distribution. However, nonparametric analysis usually generated much wider confidence bounds than those calculated via parametric analysis and predictions outside the range of observations are not possible. Kaplan – Meier (KM) estimator, also called product limit estimator is commonly used to get the survival function of lifetime data. A plot of Kaplan - Meier estimate of the survival function is a series steps of declining magnitude. When the sample size is large enough with respect to the population, Kaplan-Meier estimator approaches the true survival function for the population.

Let $S(t)$ be the probability that an individual will not have reoccurrence of an event after time t. For a sample of size n, denote the observed times until death

of n sample members as $t_1 \le t_2 \le t_3 \le ... \le t_n$. Then the nonparametric Kaplan-Meier estimator of the survival function is estimated by:

$$\hat{s(t)} = \prod_{t_i \le t} \frac{n_i - d_i}{n_i}$$

Kaplan-Meier estimator (also known as the product limit estimator) estimates the survival function from survival related data. In many medical researches, it is used to measure the portion of patients living for a certain amount of time after treatment. Kaplan-Meier is useful when we have censored data, and it is equivalent to the empirical distribution when no truncation or censoring occurs.

1.3.2 Accelerated Failure Time model

When covariates are considered, we assume that the relapse time has an explicit relationship with the covariates. Furthermore, when a parametric model is considered, we assume that the relapse time follows a given theoretical probability distribution and has an explicit relationship with the covariates.

Let T denote a continuous non-negative random variable representing the survival time (relapse time in this case), with probability density function (pdf) $f(t)$ and cumulative distribution (cdf) $F(t) = \Pr(T < t)$. We will focus on the survival function $S(t) = \Pr(T > t)$, the probability of being alive at t (reoccurrence free in this case. In this model, we start from a random variable W with a standard distribution in $(-\infty, +\infty)$ and generate a family of survival distributions by introducing location and scale parameters to relate to the relapse time as follows:

$$\log T = Y = \alpha + \sigma W \,, \qquad (1)$$

where α and σ are the location and scale parameters, respectively.

Adding covariates into the location parameter in equation (1) we have

$$Y = \log T = x'\beta + \sigma W \,,$$

where the error term W has a suitable probability distribution, e.g. extreme value, normal or logistic. This transformation leads to the Weibull, log-normal and log-logistic models for T. This type of statistical models are also called accelerated failure time (AFT) model.

1.3.3 Cox Proportional Hazard Model

An alternative approach to modeling survival data is to Cox Proportional Hazard (Cox - PH) model which assumes that the effect of the covariates is to increase or decrease the hazard function by a proportionate amount at all durations. Thus,

$$\lambda(t, x) = \lambda_0(t) e^{x'\beta}$$

or

$$\ln \frac{\lambda(t, x)}{\lambda_0(t)} = x'\beta \,,$$

where $\lambda_0(t)$ is the baseline hazard function or the hazard for an individual with covariate values 0, and $e^{x'\beta}$ is the relative risk associated with the covariate values x. Subsequently, for the survival functions

$$S(t,x) = S_0(t)^{e^{x'\beta}} .$$

Hence the survival function for covariates x is the baseline survivor raised to a power.

Parameter estimates in the Cox-PH model are obtained by maximizing the partial likelihood as opposed to the likelihood. The partial likelihood is given by

$$L(\beta) = \prod_{Y_i \text{ uncensored}} \frac{\exp(x_i'\beta)}{\Sigma_{Y_j \geq Y_i} \exp(x_j'\beta)} .$$

The log partial likelihood is given by

$$l(\beta) = \log L(\beta) = \sum_{Y_i \text{ uncensored}} \left\{ x_i'\beta - \log[\sum_{Y_j \geq Y_i} \exp(x_j'\beta)] \right\} .$$

In application of the Cox-PH model, we also included the interactions of the attributable variables.

Chapter 2

Parametric and Nonparametric Analysis of Breast Cancer Patients with Decision Tree Analysis

2.1 Background and Data

Extensive literature and studies can be found related to whether radiation shows a benefit to breast cancer patients with respect to relapse time. It is clear that radiation makes a difference in recurrence for some women. However, the potential side effect of heart damage from breast radiation makes it desirable to avoid radiotherapy unless it is absolutely necessary. Therefore, it is of great importance to identify the patients who could potentially benefit from radiation and those who would be put at higher risk for receiving radiation treatment. The aim of the present research is to perform parametric, nonparametric, and decision tree analysis to answer the above question. Our parametric and nonparametric analysis confirms the overall advantage of combined radiation and tamoxifen (RT+Tam) over tamoxifen (Tam) alone in reducing the probability of relapse; however, after utilizing decision tree analysis in conjunction with survival analysis of relapse time of breast cancer patients, we have concluded under some conditions, giving both treatments to patients without considering the clinicopathological characteristics could be negatively effective or catastrophic.

Between December 1992 and June 2000, a total of 769 women were enrolled and randomized, of which 386 received combined radiation and tamoxifen (RT+Tam), and the rest, 383, received tamoxifen-alone (Tam). The last follow up was conducted in the summer of 2002. Only those 641 patients enrolled at the Princess Margaret Hospital are included: 320 and 321 in RT+Tam and Tam treatment group, respectively.

This censored data consists of 77 uncensored observations and 564 censored observations as shown in Figure 2.1. The censored observations are mostly due to two reasons: (1) the breast cancer patient emigrated out of the study area; (2) the individual survived (did not experience occurrence) past the end of the study period. Due to the fact that nearly 90% of the data are censored observations, we take into consideration two datasets, 77 uncensored dataset, and 641 censored dataset for later analysis.

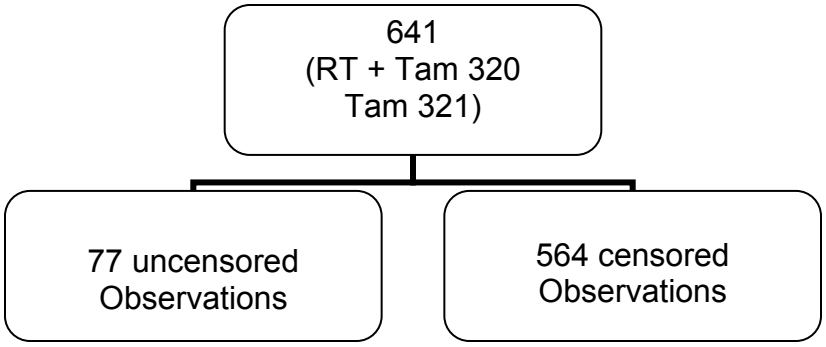

Figure 2.1 Patient Treatment Data

In the original data, three relapse events are recorded: local relapse, axillary relapse and distant relapse. The original dataset was used to analyze competing

risks (also called multiple causes of death) including relapse, second malignancy, and other causes of death. Since in the present study we are interested in the relapse time regardless of the reoccurrence type, minimum time of the three types of relapse is chosen for analysis purpose, and the values of censoring indicator variable are adjusted accordingly. Variables assessed at the time of randomization are: pathsize(size of tumor in cm) ; hist(Histology: DUC=Ductal, LOB=Lobular, MED= Medullar, MIX=Mixed, OTH=Other); hrlevel(Hormone receptor level: NEG=Negative, POS=Positive); hgb(Hemoglobin g/l); nodediss(Whether axillary node dissection was done: Y=Yes, N=No); age(Age at diagnosis in years). All these attributable variables will be used in the modeling of breast cancer in a separate study where various statistical models are used to identify the significant prognostic factors in the relapse of breast cancer, as well as the interactions between the variables and ranking of significant individual attributable variables and interactions.

2.2 Nonparametric Analysis

Kaplan-Meier estimator (also known as the product limit estimator) estimates the survival function from survival related data. In many medical researches, it is used to measure the portion of patients living for a certain amount of time after treatment. Kaplan-Meier is useful when we have censored data, and it is equivalent to the empirical distribution when no truncation or censoring occurs.

Let $S(t)$ be the probability that an individual will not have reoccurrence of breast cancer after time t. For a sample of size n, denote the observed times until death

of n pateints as $t_1 \le t_2 \le t_3 \le \ldots \le t_n$, the nonparametric Kaplan-Meier estimator of the survival function is estimated by:

$$\hat{s(t)} = \prod_{t_i \le t} \frac{n_i - d_i}{n_i}$$

Kaplan-Meier estimates of the survival curves of relapse time for the two treatment groups are shown in Figure 2.2.

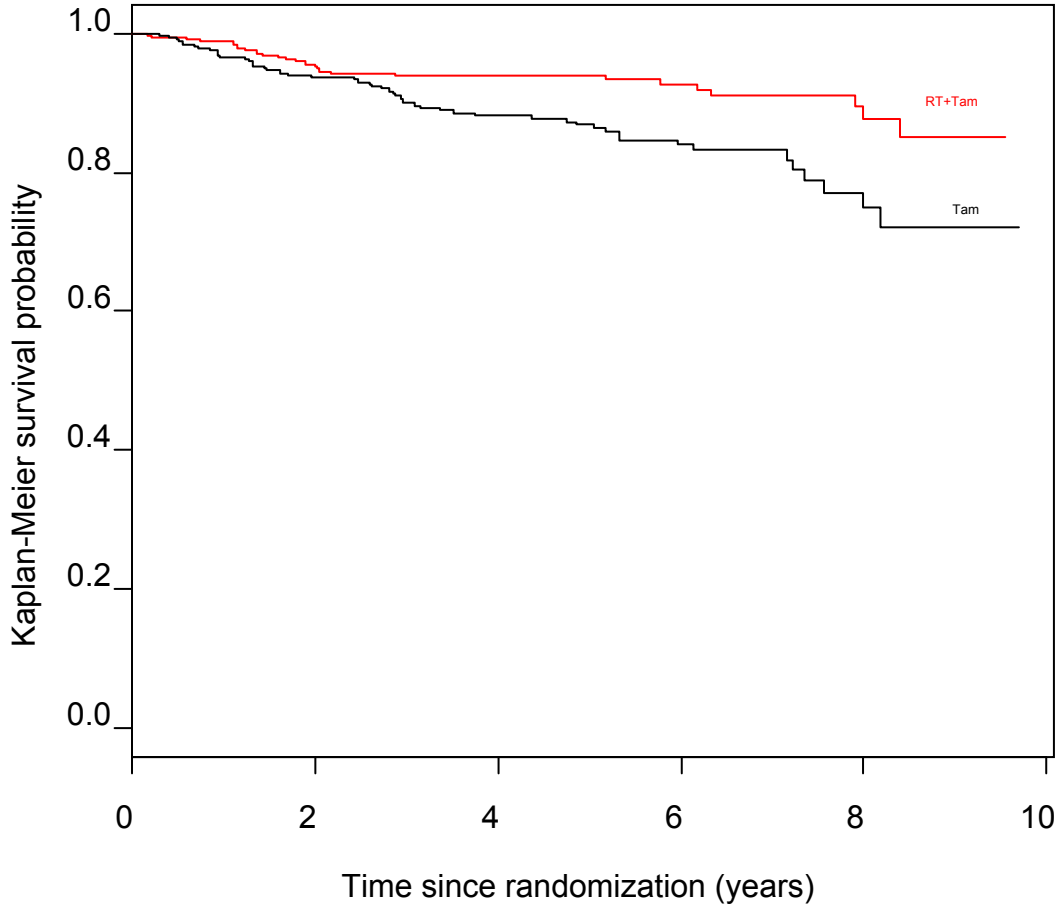

Figure 2.2 Survival Curves of Two Treatment Groups

Kaplan-Meier is a nonparametric procedure for estimating the survival curve; however, it is not commonly used to compare the true mean effectiveness of the

two treatments. In the present study, we perform actual nonparametric analysis utilizing Wilcoxon rank sum test and Peto & Peto modification of the Gehan-Wilcoxon test. We proceed in nonparametric direction for comparison purpose with the results obtained using parametric analysis. Utilizing the two different nonparametric tests, we found the information in Table 2.1, which shows that the combination of the two treatments (RT+Tam) is more effective than using the single treatment (Tam) which is consistent with Figure 2.1.

Table 2.1 Test the Difference of Means of Two Treatments

	Chi-Square	Degrees of freedom	P-value
Log-rank	9.8	1	0.0017
Peto & Peto modification of the Gehan-Wilcoxon	9.6	1	0.00197

2.3 Parametric Analysis

First, censored dataset which consists of 641 patients are analyzed, and the characteristic of the behavior of relapse time in RT+Tam arm is investigated through goodness of fit tests. The best probability distribution is the lognormal distribution, with corresponding maximum likelihood estimator (MLE) of the

following form $\hat{\mu}$ =5.148, $\hat{\sigma}$ =2.47 (as shown in Table 2.2). A graphical presentation of the cumulative distribution function (CDF) is shown by Figure 2.3 where Kaplan-Meier curve and its 95% confidence band, as well as CDF of the fitted lognormal distribution are plotted.

Table 2.2 Estimators and log-likelihood of Lognormal Distribution

	$\hat{\mu}$	$\hat{\sigma}$	Log-likelihood
Totality	4.101	2.04	-367
RT+Tam	5.148	2.47	-134.4
Tam	3.491	1.79	-227.3

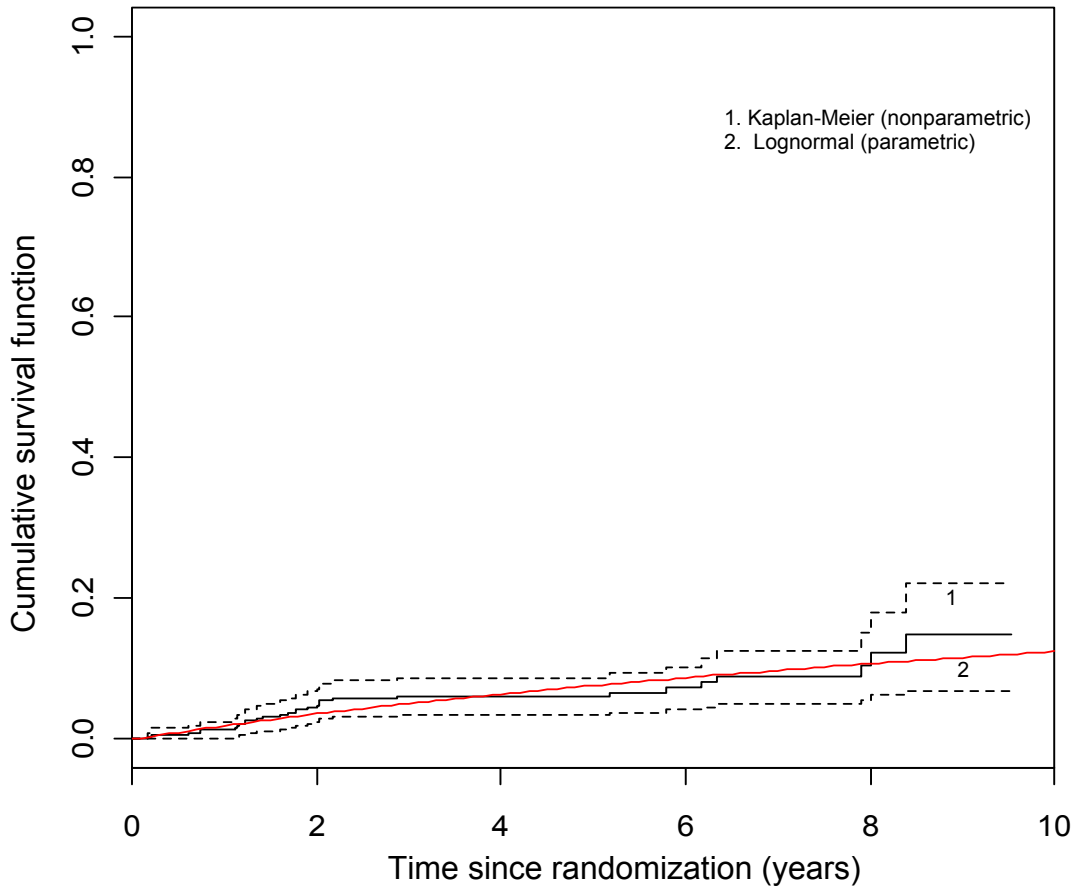

Figure 2.3 Fitted Lognormal CDF Curve for RT+Tam

As can be seen from the above graph, lognormal probability distribution seems to be a good fit for the relapse time of breast cancer patients in RT+Tam, and the survival curve from the lognormal probability distribution with estimated parameters is very close to the Kaplan-Meier survival curve and it is within the 95% confidence band constructed from Kaplan-Meier survival curve.

Similarly, we perform a parametric analysis for patients in Tam arm. It has been proven through goodness-of-fit test that the subject data follows a lognormal distribution as well, with MLE of $\hat{\mu}$ =3.419, $\hat{\sigma}$ =1.79 (as shown in Table 2.2). Therefore, the final estimated form of the lognormal probability distribution is given in Table 2.2 and a graphical form of the cumulative distribution function is given in Figure 2.4.

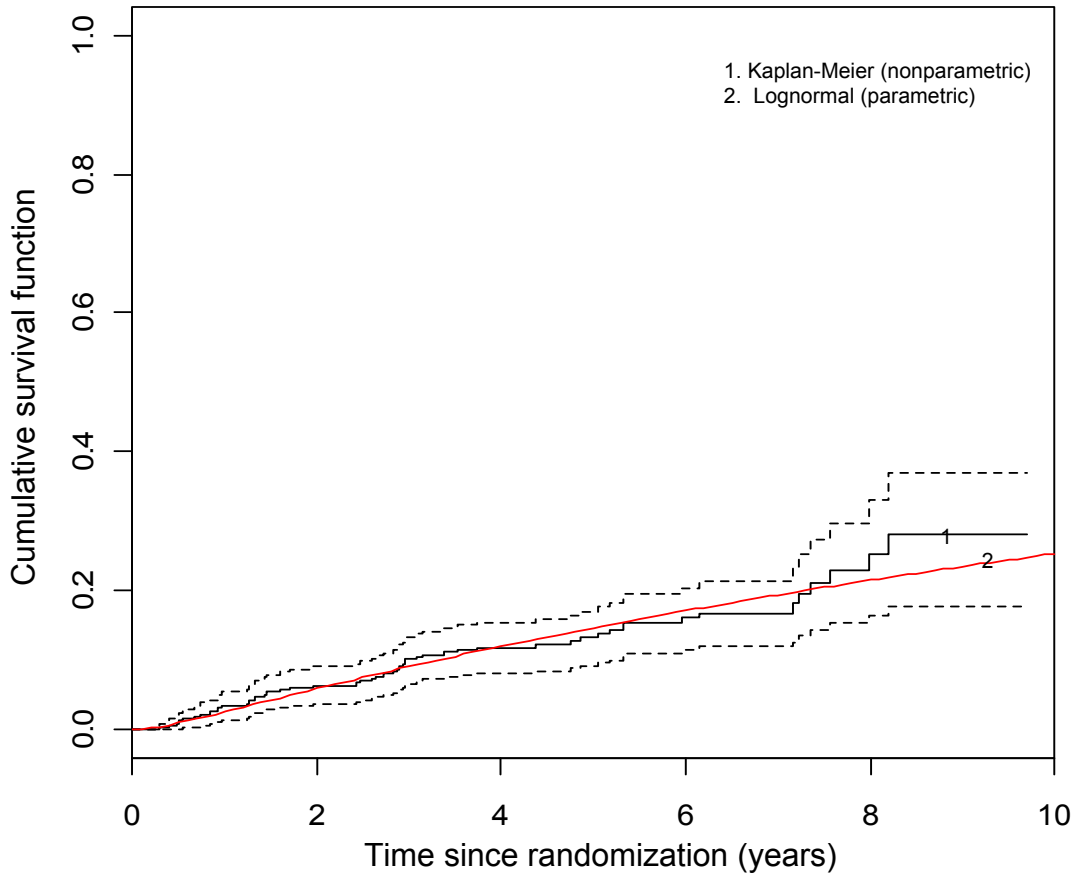

Figure 2.4 Fitted lognormal survival curve for Tam

Since relapse time in RT+Tam and Tam arm both follow lognormal probability distribution, the log-likelihood ratio test can be applied to test hypothesis

$$H_0: \mu_1 = \mu_2 = \mu \text{ vs. } H_1: \mu_1 \neq \mu_2$$

The log-likelihood ratio test statistic is given by

$$T=-2[l(\mu, \mu)-l(\mu_1, \mu_2)]=10.6$$

with one degree of freedom, and from the Chi-square distribution table, p-value is between 0.05 and 0.001. Thus, there is significant difference between the locations of the two treatment groups, which is consistent with the conclusion using nonparametric tests.

While for the uncensored dataset of the 77 breast cancer patients, of which 26 are treated with RT+Tam and 51 with Tam alone, in order to perform goodness of fit test to identify the PDF of the 26 patients, we employ bootstrapping technique to increase the sample size of the RT+Tam arm. Through goodness of fit tests including Kolmogorov-Smirnov, Anderson-Darling and Chi-Square tests, the best fit for RT+Tam is log-logistic probability distribution while the best for Tam arm is general Pareto probability distribution. Considering the difference in probability distributions of the two groups, further analysis or tests are not conducted to check the mean difference in relapse time. Since consistent results were obtained using nonparametric and parametric tests with regard to the censored dataset, we only considered the censored dataset for the subsequent analysis. However, as we will see in the following discussion, after applying

decision tree analysis to partition the subject data as a function of the tumor size, age of patient and haemoglobin, the findings of the two treatments give contradictory results which could be quite misleading in the treatment of breast cancer patients as the nonparametric and parametric analysis indicates.

2.4 Decision Tree Analysis

The clinicopathological characters of breast cancer patients are heterogeneous. Consequently, the survival times are different in subgroups of patients. Decision tree analysis is used to homogenize the data by separating the data into different subgroups on the basis of similarity of their response to treatment. The general goal of such applications is to identify prognostic factors that are predictive of survival outcome and time to an event of interest (relapse time in this study). For example, a tree-based decision analysis enables the natural identification of prognostic groups among patients, using information available regarding several clinicopathologic variables. Such groupings are important because patients treated with RT+Tam and Tam present considerable heterogeneity in terms of relapse time, and the groupings allow physicians to make early yet prudent decisions regarding adjuvant combination therapies.

The concept of exponential decision tree analysis [5] is to reduce the impurity within nodes by splitting based on covariates using a specified loss function. Assuming the hazard rate within a given node is constant, $h(y) = \lambda_j$ for all y in group j, and then the survival function within each node is an exponential function. The split point is selected so that the loss among the possible binary

splits defined by the covariates are minimized. The loss function for a node t is given by

$$R(t) = -\hat{L}(t) = D_t - D_t \log(D_t / Y_T),$$

Where $D_t = \sum_i d_i$ is the number of complete observations at the node and $Y_t = \sum_i y_i$ is the total observed time.

Considering our main focus here is to compare the two treatments instead of analyzing each treatment alone, the maximum tree depth is set to be 3 with complexity parameter 0.02. The trees of RT+Tam and Tam are shown in Figure 2.5 and Figure 2.6 shown below.

RT+Tam arm is divided into 4 groups denoted by RT1,RT2,RT3,RT4 from the left to the right; Tam arm is divided into 4 groups denoted by T1,T2,T3,T4 from the left to the right. To further investigate the survival curves of a subgroup from different treatment arms, Kaplan-Meier survival curves are plotted in Figure 2.7.

Using decision tree analysis we conclude that giving radiation to a patient whose tumor size exceeds 3.05cm would be catastrophic as has been shown in Figure 5.3 since patients in RT1 are most likely to relapse. Furthermore, treatment Tam is more effective than treatment RT+Tam with respect to relapse time has also been shown by the survival curves of T2 and RT2.

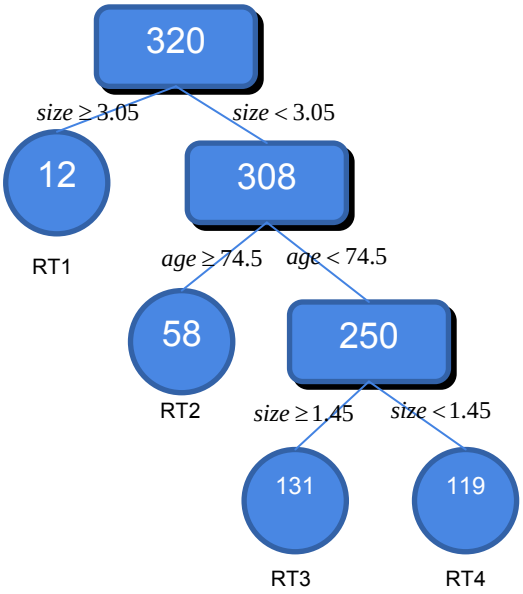

Figure 2.5 Radiation +Tamoxifen

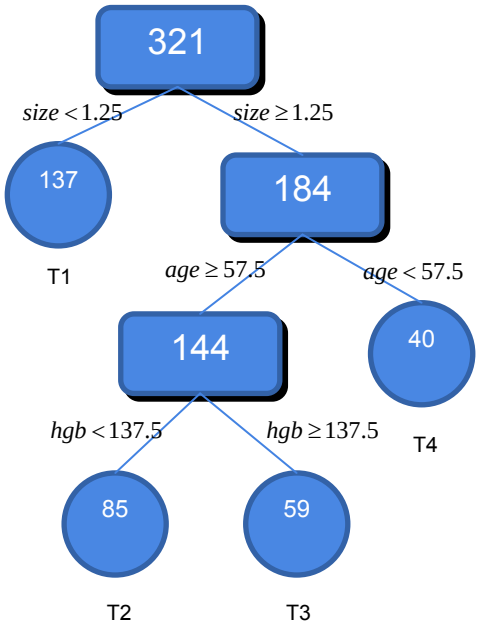

Figure 2.6 Tamoxifen

In addition, we can conclude that by using decision tree analysis and the corresponding survival analysis, we can group the breast cancer patients into three clusterings that identify the effectiveness of treatment RT+Tam versus treatment Tam. For example, the survival curve of RT3 is very close to that of T1, which suggests that for patients whose age is under 74.5 and have tumor size between 1.45cm and 3.05 cm, RT+Tam shows no advantage over Tam. Thus, it would be desirable for this patient not to consider receiving radiation.

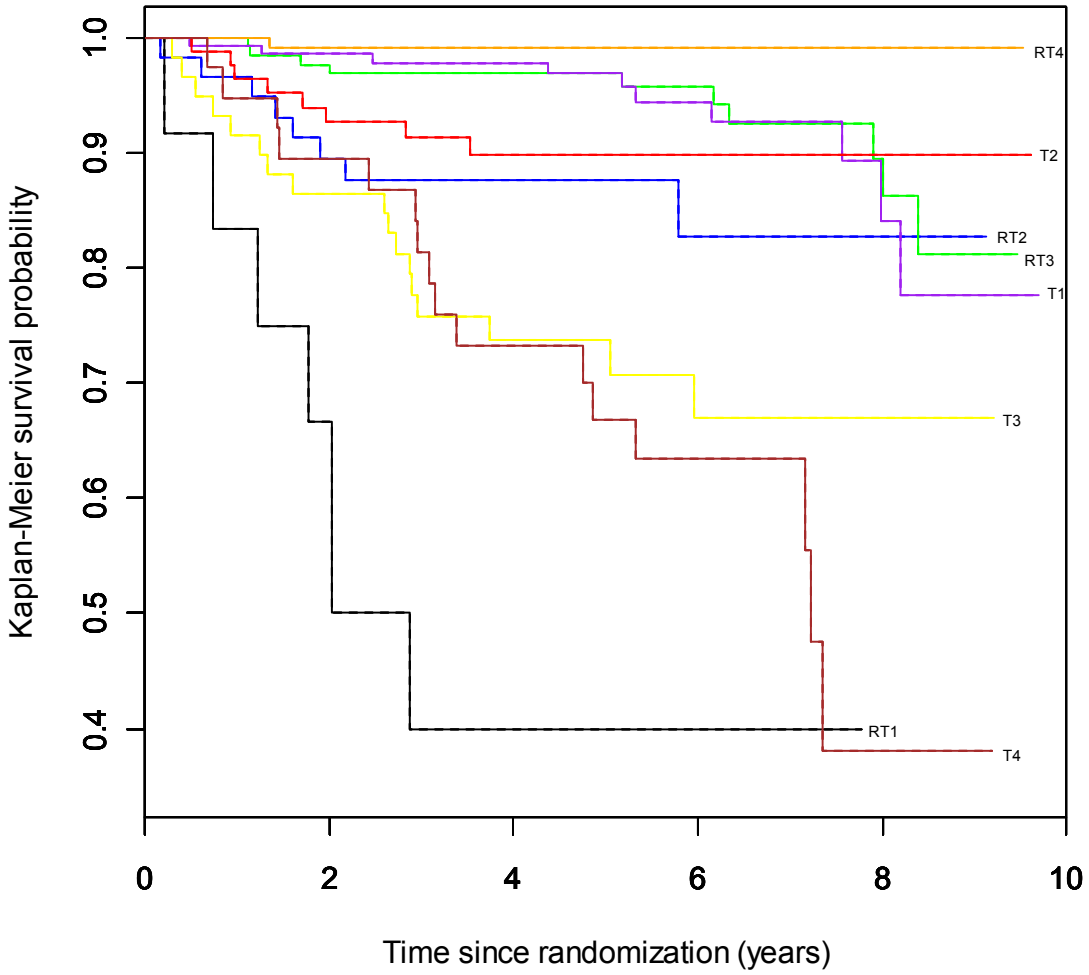

Figure 2.7 Survival Curves for Different Subgroups

We summarize below when RT+Tam and Tam are almost equally effective

(1)RT4, T2, RT3, RT2, T1

(2)T3, T4

(3)RT1

Thus, our findings are important in guiding the physicians to recommend tamoxifen alone without radiation rather than a combined treatment of tamoxifen and radiation when they are equally effective to breast cancer patients with certain size of tumor, age and hemoglobin level.

2.5 Conclusion

The objective of this chapter is to perform parametric, nonparametric, and decision tree analysis to evaluate two treatments that are being used for breast cancer patients. Our study is based on utilizing real data which was initially used in “Tamoxifen with or without breast irradiation in women of 50 years of age or older with early breast cancer” [9], and the data is supplied to us by N.A. Ibrahim “Decision tree for competing risks survival probability in breast cancer study” [2]. We agree upon certain aspects of our findings with the published results. However, in this manuscript, we focus on relapse time of breast cancer patients instead of survival time and parametric analysis instead of semi-parametric decision tree analysis is applied to provide more precise recommendations of effectiveness of the two treatments with respect to reoccurrence of breast cancer.

Although overall parametric and nonparametric comparisons of RT+Tam and Tam arms show that the combination of radiation and tamoxifen is more effective than tamoxifen alone with regard to the relapse time of a breast cancer patient, a decision tree analysis for censored data reveals that the heterogeneity of clinicopathological characteristics lead to important difference between

subgroups of the two treatment groups, thus affecting the decision making process in choosing suitable treatment for breast cancer patients.

Chapter 3

Statistical Modeling of Breast Cancer Relapse Time with Different Treatments

3.1 Background and Data

In the current chapter, same data is used to predict the relapse time of breast cancer patients with different treatments. The proposed statistical models in the present study are constructed for patients in RT+Tam Group and Tam group, respectively. Information concerning potential prognostic factors (attributable variables) are pathsize (size of tumor in cm) ; hist(Histology: DUC=Ductal, LOB=Lobular, MED= Medullar, MIX=Mixed, OTH=Other); hrlevel (Hormone receptor level: NEG=Negative, POS=Positive); hgb (Hemoglobin g/l); nodediss (Whether axillary node dissection was done: Y=Yes, N=No); age (Age of the patient in years). The dependent variable or response variable is the relapse time (in years) of a given patient.

One important question that we will address is that which of these attributable variables are significantly contributing to the response variable - the relapse time. In addition, identify all possible contributing to relapse time.

3.2 AFT Model and Cox- PH Model

As mentioned, for censored data, the most commonly used methods is survival analysis. To model the relapse time of breast cancer patients, accelerated failure time model and cox proportional hazard model are applied. The major objective of applying these models is to identify which of the six attributable variables are significant contributing to the relapse time of breast cancer patients receiving different treatments. The six explanatory variables used in the models are pathsize (size of tumor in cm) ; hist(Histology: DUC=Ductal, LOB=Lobular, MED= Medullar, MIX=Mixed, OTH=Other); hrlevel (Hormone receptor level: NEG=Negative, POS=Positive); hgb (Hemoglobin g/l); nodediss (Whether axillary node dissection was done: Y=Yes, N=No); age (Age of the patient in years).

The most commonly used AFT models such as exponential, Weibull and log-normal AFT models and Cox-PH model are applied. After running the model including all covariates and interactions between covariates, number of parameters that drive the attributable variables are reduced using stepwise regression based on Arkariki Information Critria (AIC) is a measure of the goodness of fit of an estimated statistical model. It is trades off the complexity of an estimated model against how well the model fits the data. It is given by $AIC = -2\log(likelihood) + 2(p+k)$, where p is the number of parameter, and k is the number of parameters in the distribution. Statistical models with lower AIC are preferred. Table 3.1 given below shows the covariates and interactions in the related statistical models chosen using the AIC as well as their corresponding p-

values for the breast cancer patients that were treated with both radiation and tamoxifen. * indicates the variables is statistically significant.

Table 3.1 Factors in Parametric Regression Models for RT+Tam

RT+Tam	lognormal	exponential	Weibull	Cox-PH
AIC	237.64	234.49	235.97	245.8
age	0.002*	0.008*	0.011*	0.01*
pathsize	0.01*	0.0002*	0.0002*	0.00086*
nodediss	0.021*	0.009*	0.012*	0.012*
hrlevel	0.027*	0.010*	0.008*	0.016*
age:nodediss	0.037*	0.022*	0.026*	0.028*
nodediss:hrlevel	0.009*	0.0005*	0.0008*	0.00067*
pathsize:hrlevel	0.078	0.060	0.041*	0.099

As can be seen from the table, age, pathsize, nodediss , hrlevel, and the interactions between age and nodediss, and interaction between nodediss and hrlevel are significant with respect to relapse time of breast cancer patients who received radiation and tamoxifen. The interaction of pathsize and hrlevel proves to be significant only in Weibull AFT model.

Table 3.2 given below address the same aspects as table 3.1, for breast cancer patients that were treated with tamoxifen only.

Table 3.2 Factors in Parametric Regression Models for Tam

Tam	lognormal	exponential	Weibull	Cox-PH
AIC	439.34	439.55	440.76	525.89
age	0.343	0.294	0.287	0.32
hgb	0.037*	0.645	0.630	0.68
pathsize	0.339	0.316	0.300	0.33
nodediss	0.025*	0.017*	0.020*	0.018*
hrlevel	0.006*	0.002*	0.003*	0.002*
age:pathsize	0.143	0.112	0.106	0.120
age:nodediss	0.038*	0.006*	0.007*	0.0065*
hgb:nodediss	0.054	0.077	0.079	0.075
age:hgb	NA	0.131	0.128	0.150

For patients who received tamoxifen only, only nodediss, hrlevel as single attributable variables are significant with respect to relapse time in this group. It is worth noticing that although age itself is not significantly contributing to relapse time, the interaction between age and nodediss is significant. hgb is found to be significant only in lognormal AFT model.

Comparing the results from the two treatment groups, for each group at significance level of 0.05, the three AFT models give almost the same results. Significant prognostic factors for relapse time of breast cancer patients who

received combined treatment of radiation and tamoxifen are age, pathsize, nodediss, hrlevel, age:nodediss, nodediss:hrlevel which appears statistically significant in all lognormal, exponential and Weibull regression models. Only in Weibull regression model pathsize: hrlevel shows significant contribution to the model. For patients who are in Tam arm, all three models show nodediss, hrlevel and age:nodediss are significant contributing, only in lognormal regression model hgb shows significance.

Furthermore, significant prognostic factors identified using Cox-PH model confirm our conclusion. There are six significantly contributing variables two of which are interactions for RT+Tam arm and three significantly contributing variables one of which is interaction for Tam arm.

Next the predicted survival curves of the three AFT models and Cox-PH model for each arm are compared to Kaplan-Meier survival curve along with 95% confidence band to determine the best predicting model for relapse time and the results will be shown and discussed.

3.3 Kaplan-Meier VS. Parametric Survival Analysis

From the above four models, we identified the significant attributable variables and interactions between them that contributes to the relapse time of breast cancer patients in two different treatment groups. To investigate which model gives the best fit of the relapse time of breast cancer patients in those two groups, graphical presentation would be a useful tool. In this study, Kaplan-Meier curve as a commonly used nonparametric survival curve and its 95% confidence limits are plotted against the survival curves obtained from the four models discussed above to see which model gives the closest curve to Kaplan-Meier survival curve.

Using the breast cancer data for patients from RT+Tam arm, the Kaplan-Meier curves along with its 95% confidence limits against the lognormal AFT model are plotted in Figure 3.1 below.

As can be seen from Figure 3.1, for the second year, third year and around the sixth year , the survival curve from lognormal AFT model runs out of the 95% confidence band of Kaplan-Meier curve.

For exponential AFT model, the same graphical representation is given in Fiugre 3.2 below. Form this graph, the survival curve estimated from the exponential AFT model is off the 95% confidence band from year 1 to year 4, and from year 5 to year 6.

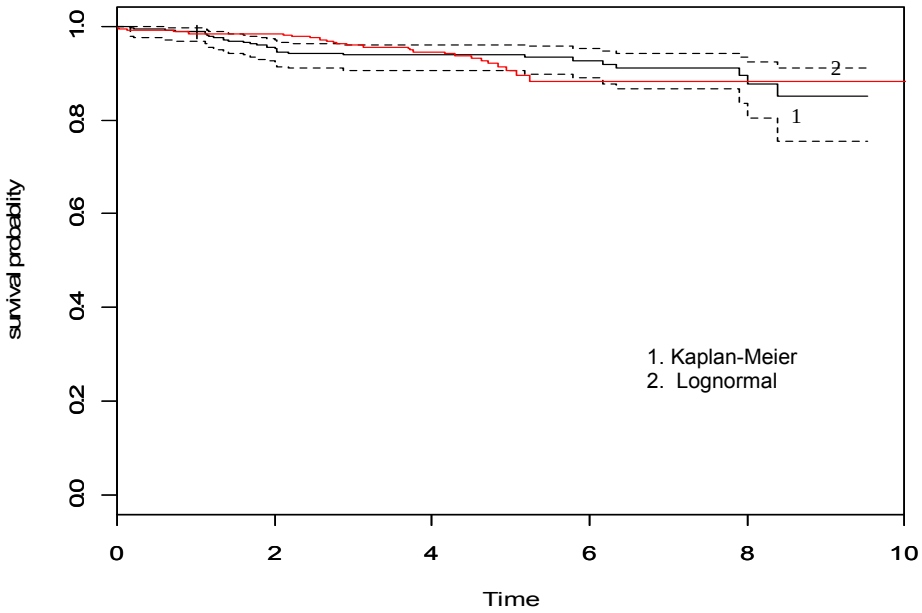

Figure 3.1 Lognormal VS. Kaplan-Meier for RT+Tam

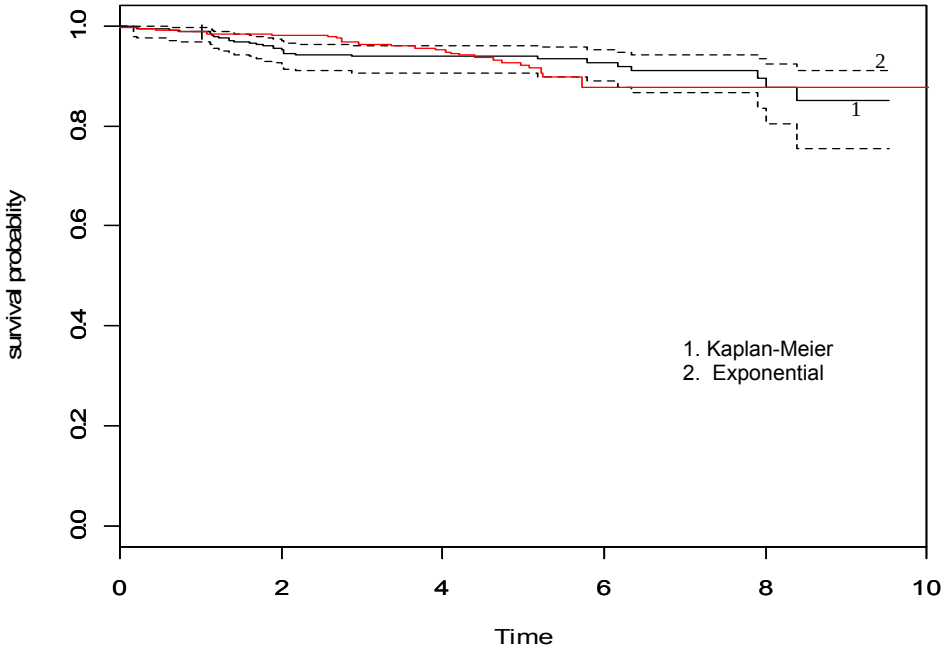

Figure 3.2 Exponential VS. Kaplan-Meier for RT+Tam

Figure 3.3 shows the graph of survival curve obtained from the Weibull AFT model, it deviates from the 95% confidence band of the Kaplan-Meier in a similar pattern as the survival curve of the exponential AFT model.

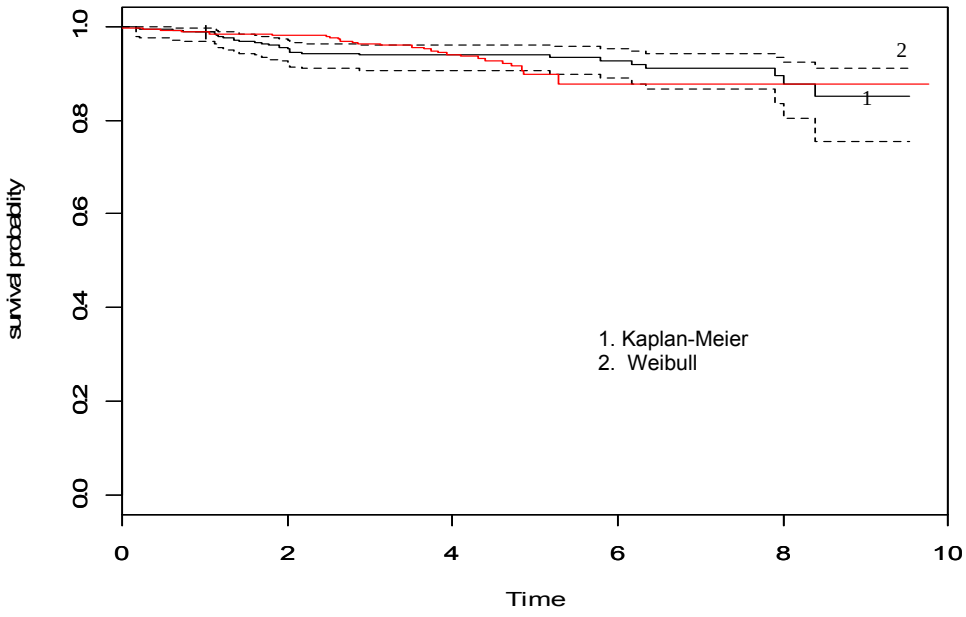

Figure 3.3 Weibull VS. Kaplan-Meier for RT+Tam

However, in Figure 3.4 which shows the survival curve obtained from the Cox-PH model, it is clear that most of the time, the survival curve lies within the 95% confidence band of Kaplan-Meier curve.

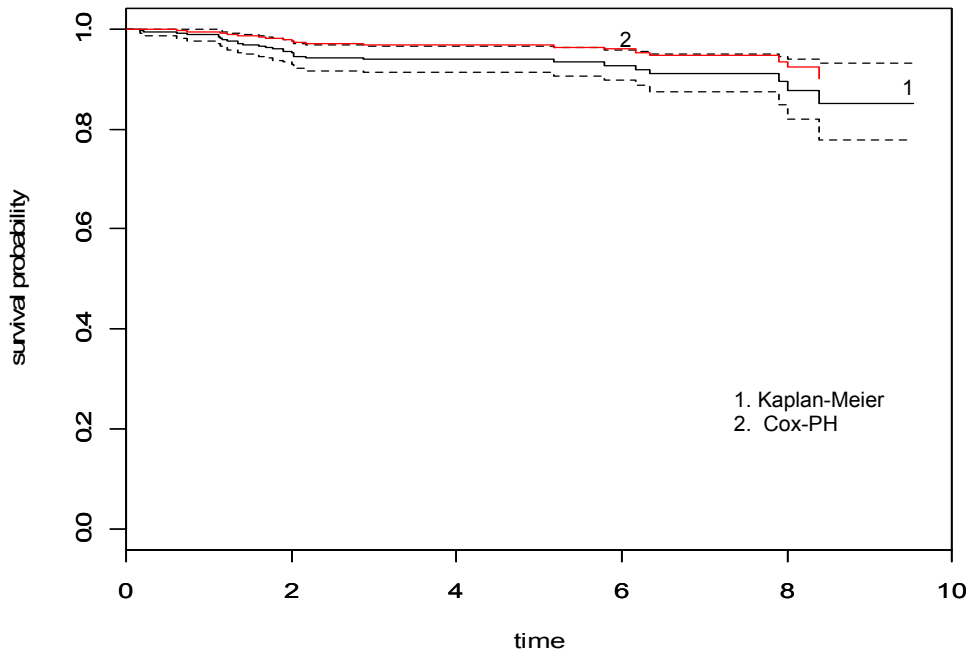

Figure 3.4 Cox-PH VS. Kaplan-Meier for RT+Tam

Thus, we can conclude from the above analysis that Cox-PH model with interactions gives a better prediction of relapse possibility of breast cancer patients in RT+Tam arm comparing to the three AFT models.

Similarly, we proceed to perform a survival analysis of the relapse time for the patients who are treated with tamoxifen only. Figures 3.5, 3.6 and 3.7 shows the survival curves obtained from lognormal, exponential and Weibull AFT models. It is clear that those survival curves fall out of the 95% confidence limits of the Kaplan-Meier curve most of the time. However, in Figure 3.8 which shows the survival curve obtained from the Cox-PH model with interactions, we can see the survival curve lies within the 95% confidence band. Therefore, we can conclude

that for patients who received tamoxifen only, Cox-PH model with interactions gives a more precise prediction of the relapse time than AFT model.

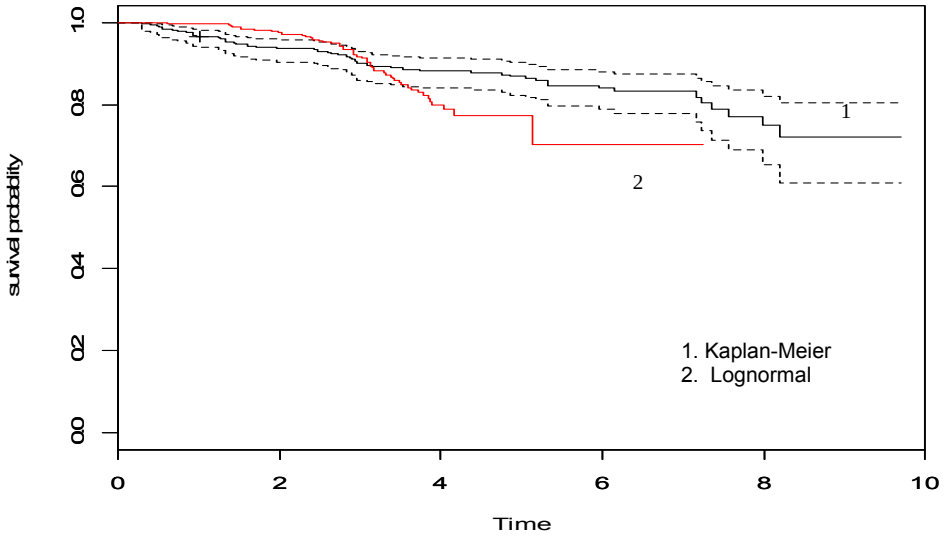

Figure 3.5 Lognormal VS. Kaplan-Meier for Tam

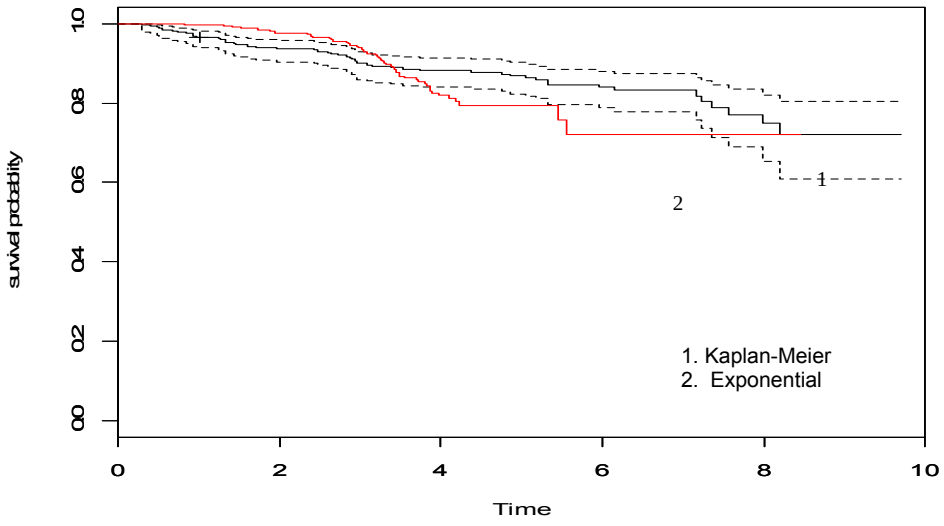

Figure 3.6 Exponential VS. Kaplan-Meier for Tam

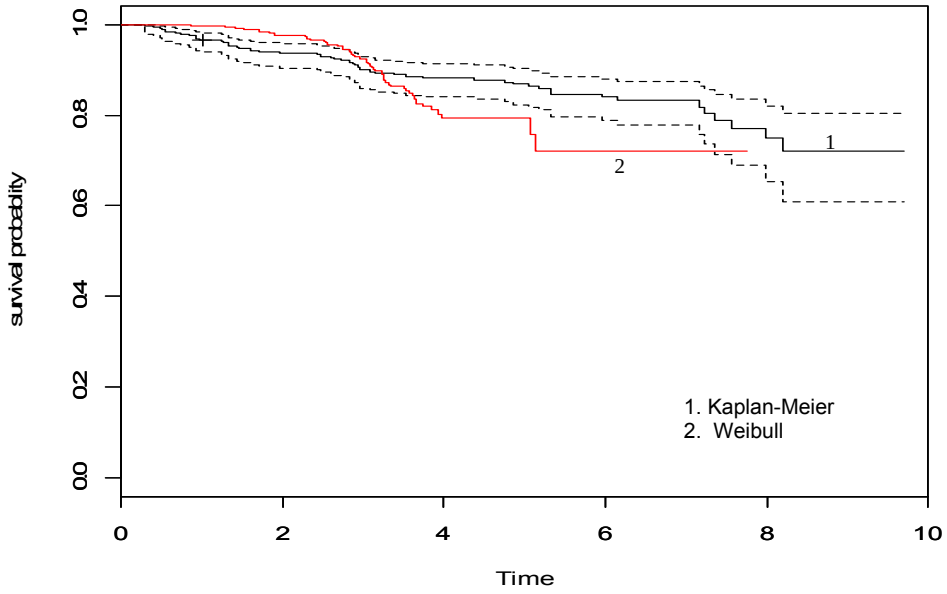

Figure 3.7 Weibull VS. Kaplan-Meier for Tam

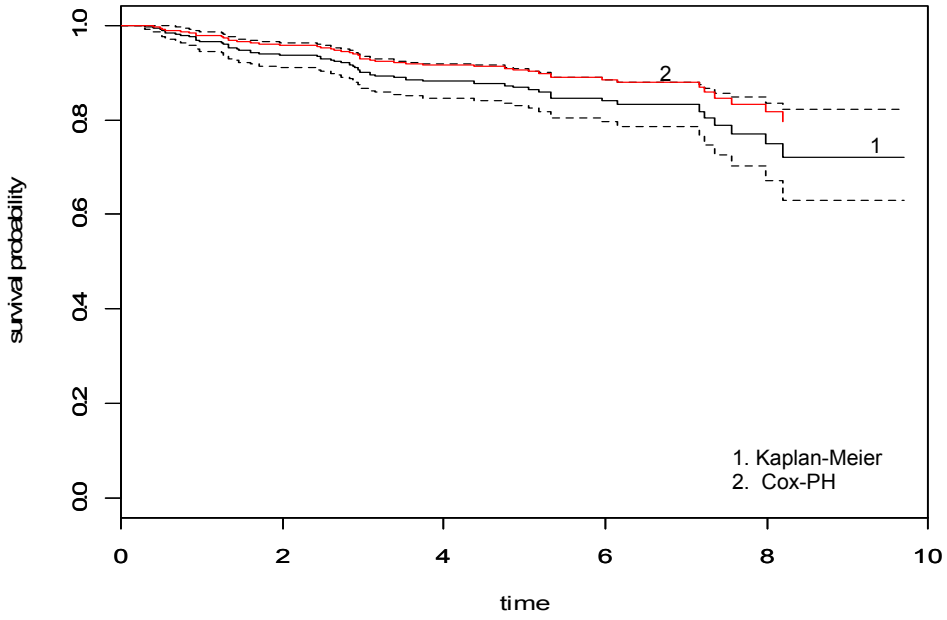

Figure 3.8 Cox-PH VS. Kaplan-Meier for Tam

Since Cox-PH model gives better prediction of relapse possibility than AFT models for both groups, we recommend Cox-PH model to approximate the probability of having

2-year, 5-year, and 8-year reoccurrence-free and the results are shown in Table 3.3 below.

Table 3.3 Reoccurrence-free Probability

		2-year	5-year	8-year
RT+Tam	K-M	0.98	0.95	0.93
	Cox-PH	0.98	0.97	0.95
Tam	K-M	0.94	0.88	0.76
	Cox-PH	0.97	0..92	0.84

Although there is consistency on indentifying significant prognostic factors for reoccurrence of breast cancer, it can be seen from the above six graphs, regression model might not be a good choice for predicting purpose. Cox-PH models with interactions show more efficiency over regression models with respect to predicting power. So it would be advisable to use Cox-PH model with interactions to predict the relapse time of a breast cancer patient given all the information of the attributable variables. And as can be seen form the reoccurrence-free table, patients with combined treatments have higher possibility of free of reoccurrence of cancer than those with single treatment.

3.4 Cure Rate Statistical Model

3.4.1 Model introduction

Any clinical trial consists of a heterogeneous group of patients that can be divided into two groups. Those who respond favorably to the treatment and become insusceptible to the disease are called cured. The others that do not respond to the treatment remain uncured. It would be of interest to determine the proportion of cured patients and study the causes for the failure of the treatment or reoccurrence of the disease. Unlike the above mentioned survival parametric regression model and semi-parametric Cox-PH model with interactions that assume each patient is susceptible to failure of treatment or reoccurrence, cure rate statistical models are survival models consisting of cured and uncured fractions. These models are being widely used in analyzing cancer data from clinical trials. The first model to estimate cure fraction was developed by Boag (1949) which is called mixture model or standard cure rate model. Let π denote the proportion of cured patients and 1-π is the proportion of uncured patients, then the survival function for the group is given by

$$S(t) = \pi + (1-\pi)S_u(t),$$

where $S_u(t)$ is the survival function of the uncured group.

It follows that the density function is given by

$$f(t) = (1-\pi) f_u(t) .$$

For uncured patients, we assume that the failure time or relapse time T follows a classical probability distribution, and also we can add the effect of covariates into the model using the parametric survival regression models that we studied in the previous section. For cure rate π , it can either be assumed constant or dependent on covariates by a logistic model, that is,

$$\log(\frac{\pi}{1-\pi}) = \exp(x'\beta) .$$

Thus, covariates may be used either in cure rate or in the failure time probability distribution of the uncured patients. These different conditions will be considered in developing the modeling process.

Estimates of parameters in the model can be obtained by maximizing the overall likelihood function given by

$$L = \Pi_{i=1}^{n} \left\{(1-\pi_i) f_u(t_i)\right\}^{\sigma_i} \left\{\pi_i + (1-\pi_i) S_u(t_i)\right\}^{1-\sigma_i} ,$$

where t_i is the observed relapse time, and σ_i is the censoring indicator with σ_i =1 if t_i is uncensored and σ_i =0, otherwise.

3.4.2 Model results for the breast cancer data

For the survival regression part, Weibull, lognormal (Lnormal), Gamma, generalized

log-logistic (GLL), log-logistic(Llogistic), generalized F(GF), extended generalized gamma(EGG) and Rayleigh parametric regression are used. The following cases encompass the above statistical analysis as set forth.

Case 1: No covariates in π and $S_u(t)$: when both cure rate and survival curve of uncured groups are independent of covariates. But we get very different cure rates using different distributions which suggest the model is very sensitive to the underlying distribution of the failure time of uncured patients.

Case 2: No covariates in π, six single covariates in $S_u(t)$: when we consider covariates in survival function of uncured group, table 3.4 shows there is some kind of consistency of cure rate among different distribution assumptions.

Table 3.4 Cure Rate of Uncured Patients

	RT+Tam		Tam	
	Likelihood	Cure rate	Likelihood	Cure rate
Weibull	-98.9810	0.1000	-171.2401	0.0748
Lnormal	-96.6371	0.0057	-171.4864	0.0748
Gamma	-95.9291	0.0064	-171.6280	0.449
GLL	-96.6900	0.1000	-171.3320	0.0748

Llogistic	-100.6583	0.1000	-171.6338	0.0748
GF	-96.2180	0.002152	-171.9223	0.0748
EGG	-95.6891	0.0038	-167.6022	0.5582
Rayleigh	-117.5059	0.1000	-204.8281	0.0748

Case 3: Six single covariates in π, six single covariates in $S_u(t)$: when we consider covariates in both cure rate and survival function of uncured group. Although we add six covariates into cure rate, there is not much improvement in the likelihood and sometimes the likelihood is even lower, which suggests cure rate might not be dependent on those covariates; instead, we can consider it as a constant.

Case 4: No covariates in π, six single covariates and their interactions in $S_u(t)$: Since there is no significant difference in maximum likelihood between case 2 and case 3, it shows including covariates does not improve the model much. Thus, in the following analysis, we consider cure rate as a constant, i.e. independent of those covariates. Table 3.5 shows uniformity of cure rates using different parametric regression models.

Table 3.5 Cure Rate with Interactions of Uncured Patients

	RT+Tam		Tam	
	Likelihood	Cure rate	Likelihood	Cure rate
Weibull	-100.5568	0.1000	-166.6240	0.0748
Lnormal	-109.4816	0.1000	-167.8399	0.0748
Gamma	-88.9750	0.1000	-167.7902	0.0748
GLL	-89.3220	0.1000	-164.8092	0.0748
Llogistic	-101.7427	0.1000	-165.9777	0.0748
GF	-89.0793	0.1000	-164.6293	0.0748
EGG	-90.5768	0.1000	-163.8507	0.0748
Rayleigh	-95.4851	0.1000	-186.6157	0.0748

After computing the AIC of the above models for each group, the smallest one for RT+Tam is Gamma, the smallest one for Tam is EGG. Hence, we choose mixture cure model with Gamma regression for uncured RT+Tam group and with EGG regression for uncured Tam group. For patients who received radiation and tamoxifen, 10% of them will be cured of breast cancer and not be subject to reoccurrence. However, for those who received tamoxifen alone, only 7.48% will be cured of breast cancer which suggests that giving radiation to breast cancer patients who take tamoxifen could possibly decrease the probability of reoccurrence of breast cancer.

3.4.3 Conclusions

By applying AFT and Cox-PH models, the significant factors and interactions that contribute to relapse time of a breast cancer patient receiving different treatments are identified and AFT and Cox-PH gives consistent results. With respect to predicting survival curve, Cox-PH model gives better fit than AFT models. Thus, given information of covariates of a given breast cancer patient, Cox-PH model with interactions can be applied to determine the time before reoccurrence of breast cancer.

From a different perspective, cure rate model takes into consideration the fact that some part of the patients are cured and will never experience reoccurrence. It is found that cure rates for RT+Tam and Tam groups both are independent of the covariates and are different. For RT+Tam group, the cure rate is 0.1 which is higher than that of Tam group which is 0.0748. Thus, using the cure rate statistical model we conclude that patients received combined treatment of radiation and tamoxifen are more likely to be cured of breast cancer and less susceptible to reoccurrence of breast cancer than those who received single treatment.

Chapter 4

Markov Modeling of Stages of Breast Cancer Patients

4.1 Background

Markov chain model was first produced by Andrey Markov (1906) theoretically and has been applied in various fields such as physics, queueing theory, internet application, economics, finance, and social sciences. As an efficient way of describing a process in which an individual moves through a series of states in continuous time, it has also been extensively used in health filed where the progression of a certain disease are of great importance to both patients and doctors. In this chapter, the main objective is to investigate the progression of breast cancer patients in three different stages who took different treatments, of which the first group of patients received combined treatments of tamoxifen and radiation, and the other group of patients received tamoxifen alone. The three stages that we are interested in are: alive with no relapse, alive with relapse, and death as shown in Figure 1. Even though breast cancer patients who have reoccurrence may be treated and recovered from breast cancer and become alive with no relapse, due to the fact that the data does not include any observations of that process, we consider the second state- alive with relapse- as

those patients who ever had relapse and are still alive, no matter they are recovered from breast cancer or not.

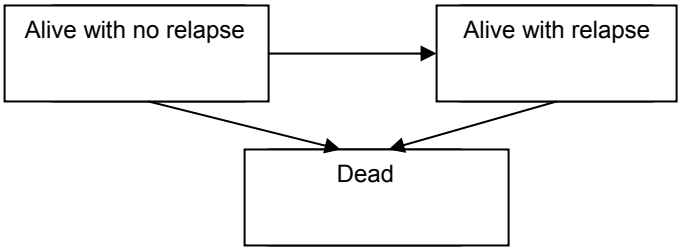

Figure.4.1 Three Stages of Breast Cancer

4.2 Markov Chain Model

Markov chain is a model for a finite or infinite random process sequence $X = \{X_1, X_{2,\ldots,X_N}\}$ Unlike i.i.d model which assumes the independency of a sequence of events Xi's, Markov model takes into account the dependencies between Xi's.

Suppose a random process $X = \{X_t\}_{t\geq 1} = \{X_1, X_2, \ldots\}$ of random variables taking values in a discrete set of states $S = \{1, 2, 3, \ldots, s\}$ and X_t represents the state of the process of an individual at time t. In this study, there are three states: alive without relapse, alive with relapse and death and the arrows in Figure.1 show which transitions are possible between states. Consider a realization of the entire history of the process up to and including time t is $\{X_t = x_t, X_{t-1} = x_{t-1}, \ldots, X_1 = x_1\}$ where $x_t, x_{t-1}, \ldots, x_1$ is a sequence of states at different times. A random process is called a Markov Chain if the conditional probabilities

between the outcomes at different times satisfy the Markov property, that is, the conditional probability of an event one step into future conditioned on the entire past of the process is equal to the conditional probability of the future event given just the present. In other words, the one-step future state depends only on the current state:

$$P(X_{t+1} = x_{t+1} \mid X_t = x_t, X_{t-1} = x_{t-1}, ..., X_1 = x_1) = P(X_{t+1} = x_{t+1} \mid X_t = x_t),$$

For every sequence $x_1, ..., x_t, x_{t+1}$ of elements of S and every $t \geq 1$.

The transition probability from state i to state j at time t and transition intensity is defined as

$$p_{ij}(t) = p(X_{t+1} = j \mid X_t = i),$$

$$q_{ij}(t) = \lim_{h \to 0} \frac{P(X(t+h) = j \mid X(t) = i)}{h} .$$

If the transition probabilities do not depend on time, $p_{ij}(t)$ can simply be written as p_{ij}, then the Markov chain is called time-homogeneous. A transition probability matrix $P(t)$ consists of all the transition probabilities between states in the matrix form

$$P(t) = \begin{Bmatrix} p_{11}(t) & p_{12}(t) & \dots & p_{1s}(t) \\ p_{21}(t) & p_{22}(t) & \dots & p_{2s}(t) \\ \dots & \dots & \dots & \dots \\ p_{s1}(t) & p_{s2}(t) & \dots & p_{ss}(t) \end{Bmatrix},$$

where probabilities in each row add up to 1.

The transition probability matrix can be calculated by taking the matrix exponential of the scaled transition intensity matrix $P(t) = Exp(tQ)$ where

$$Q = \begin{Bmatrix} q_{11} & q_{12} & \cdots & q_{1s} \\ q_{21} & q_{22} & \cdots & q_{2s} \\ \cdots & \cdots & \cdots & \cdots \\ q_{s1} & q_{s2} & \cdots & q_{ss} \end{Bmatrix}$$

The exponential of a matrix A is defined by

$$Exp(A) = 1 + A^2 / 2! + A^3 / 3! + \ldots$$

where each summand in the series is matrix products.

Next, the intensity matrix and transition probabilities matrix can be obtained by maximizing the likelihood $L(Q)$. Consider an individual consist of a series of times $(t_1, t_2, ..., t_n)$ and corresponding states $(x_1, x_2, ..., x_n)$. More specifically, we consider a pair of successive states observed to be i and j at time t_i and t_j. Three scenarios are considered as follows.

(1) If the information of the individual are obtained at arbitrary observation times and the exact time of the transition of stages is unknown, the contribution to the likelihood from this pair of states is

$$L_{ij} = p_{ij}(t_j - t_i)$$

(2) If the exact times of transitions between different states are recorded and there is no transition between the observation times, the contribution to the likelihood from this pair of states is

$$L_{ij} = p_{ij}(t_j - t_i)q_{ij}$$

(3) If the time of death is known or $j = death$ but the state on the previous instant before death is unknown which we denote by k (k could be any possible state between state i and death), the contribution to the likelihood from this pair of states is

$$L_{ij} = \sum_{k \neq j} p_{ik}(t_j - t_i)q_{kj}$$

After the likelihood function $L(Q)$ is constructed, the estimated intensity and transition probabilities would be the ones that maximize $L(Q)$.

4.3 Breast Cancer Markov Chain Results

The breast cancer patients are divided into two groups RT+Tam and Tam based on the different treatment they received. For those patients who received combined treatments, 26 patients experienced relapse, 13 patients died without reoccurrence of breast cancer during the entire period of study, 14 died after reoccurrence of breast cancer. For the patients in Tam group, 51 patients

experienced relapse, 10 died without reoccurrence of breast cancer, 13 died after reoccurrence of breast cancer.

After running the Markov model, the transition intensity matrixes for both groups are obtained as shown in the tables below.

Table.4.1 Transition Intensity Matrix of RT+Tam

	State 1	State 2	State3
State 1	-0.02301	0.01957	0.0034
State 2	0	-0.3074	0.3074
State 3	0	0	0

Table.4.2 Transition Intensity Matrix of Tam

	State 1	State 2	State3
State 1	-0.03917	0.03528	0.003889
State 2	0	-0.08533	0.08533
State 3	0	0	0

As we can observe from the two tables, patients who received single treatment are have a higher transition intensity form State 1 to State 2, thus they are more likely to have breast cancer reoccurrence. For those patients who died without relapse, there is not much significant difference between the two treatments as illustrated by the intensity form State 1 to State 3. Combined treatment is also

more effective than single treatment with respect to the possibility of death without relapse as can be seen from the transition intensity from State 1 to State 3. However, for those who already experienced relapse of breast cancer, patients who received combined treatments are more likely to die than those who received single treatment. In other words, combined treatment should be chosen over single treatment to avoid reoccurrence, but for those patients who already had breast cancer relapse, it would be advisable to choose single treatment to extend the time from reoccurrence to death.

The following two graphs give a clearer view of the effectiveness of the two treatments with respect to the survival probability. The two graphs show the survival curves of the patients who had reoccurrence and who had no reoccurrence in each treatment group.

Comparing the above two graphs, it is clear that for patients who are in State 2 (ever experienced reoccurrence of breast cancer), the curve for RT+Tam group in Figure 2 has a much faster decreasing slope than the curve for Tam group in Figure 3 which suggests combined treatment of tamoxifen and radiation is not as effective as radiation alone for patients who had reoccurrence.

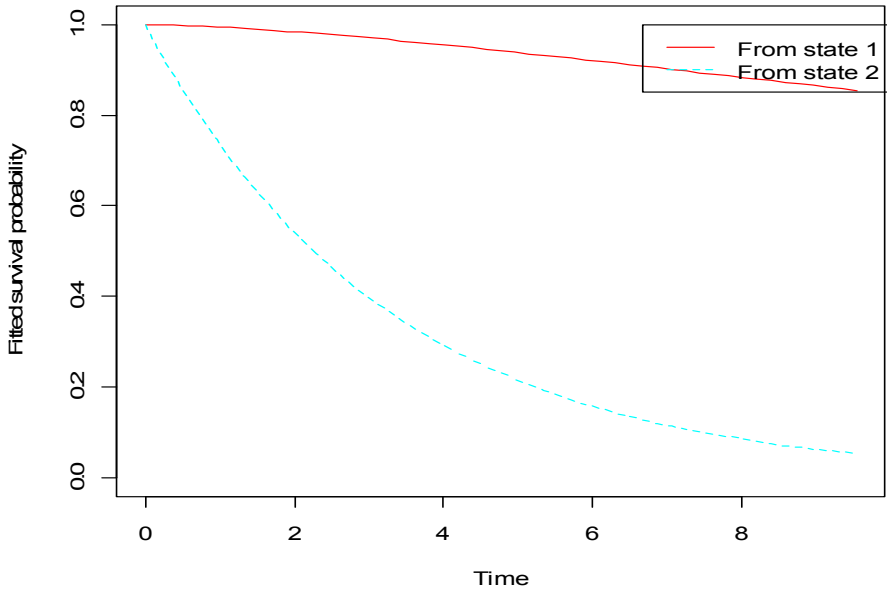

Figure.4.2 Survival Curves of Patients in RT+Tam

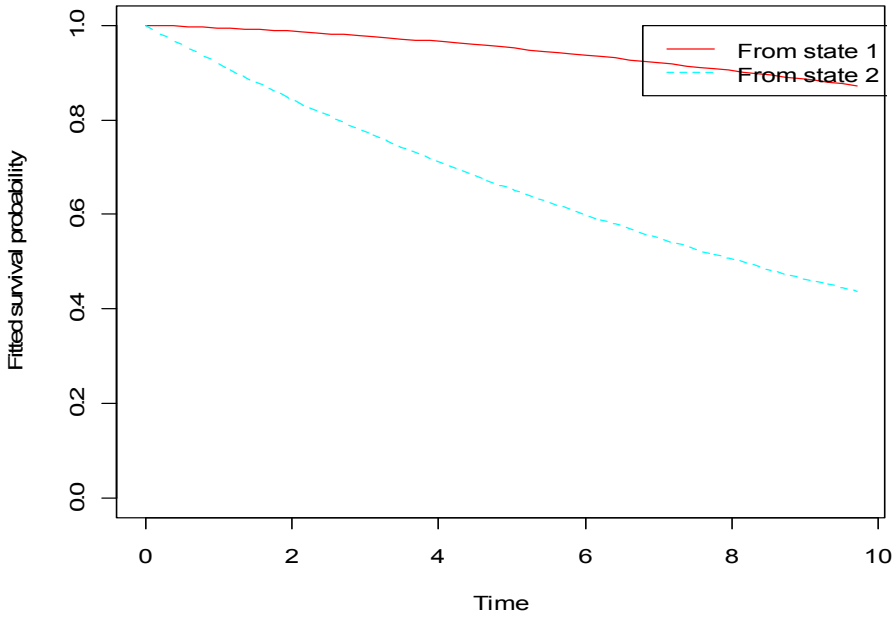

Figure.4.3 Survival Curves of Patients in Tam

From the above analysis, Markov chain model gives us recommendations on which treatment to choose for breast cancer patients with respect to relapse and survival time. Moreover, it provides patients with very important information on the exact time or possibilities of reoccurrence and death. Estimated mean sojourn times in each transient state for patients who received combined treatment are 43.46 and 3.25 in State 1 and State 2, respectively. Estimated mean sojourn times for patients who received single treatment are 25.53 and 11.72 in State 1 and State 2. This further confirms our conclusion that patients with combined treatment will stay in State 1 longer than those with single treatment, however, for patients who had relapse of breast cancer, patients with single treatment with stay alive longer than those with combined treatment.

Another major concern of this study is to provide transition probability matrix at different times so that given specific time period, we will be able to tell the probability that a given state will transit to another state. Tables 3 to Table 10 give 2-year, 4-year, 5-year and 10-year transition probability matrixes of patients in RT+Tam and Tam groups.

Table 4.3 2-year Transition Probability for RT+ Tam

	Stage 1	Stage 2	Stage 3
Stage 1	0.9550	0.0285	0.0165
Stage 2	0	0.5408	0.4592
Stage 3	0	0	0

Table 4.4 2-year Transition Probability for Tam

	Stage 1	Stage 2	Stage 3
Stage 1	0.9247	0.0623	0.0130
Stage 2	0	0.8431	0.1569
Stage 3	0	0	0

Table 4.5 4-year Transition Probability for RT+Tam

	Stage 1	Stage 2	Stage 3
Stage 1	0.9121	0.0426	0.0453
Stage 2	0	0.2925	0.7075
Stage 3	0	0	0

Table 4.6 4-year Transition Probability for Tam

	Stage 1	Stage 2	Stage 3
Stage 1	0.8550	0.1102	0.0348
Stage 2	0	0.7108	0.2892
Stage 3	0	0	0

Table 4.7 5-year Transition Probability for RT+Tam

	Stage 1	Stage 2	Stage 3
Stage 1	0.8913	0.0466	0.0621
Stage 2	0	0.2151	0.7849
Stage 3	0	0	0

Table 4.8 5-year Transition Probability for Tam

	Stage 1	Stage 2	Stage 3
Stage 1	0.8221	0.1295	0.0484
Stage 2	0	0.6527	0.3473
Stage 3	0	0	0

Table 4.9 10-year Transition Probability for RT+Tam

	Stage 1	Stage 2	Stage 3
Stage 1	0.7945	0.0515	0.1540
Stage 2	0	0.0463	0.9537
Stage 3	0	0	

Table 4.10 10-year Transition Probability for Tam

	Stage 1	Stage 2	Stage 3
Stage 1	0.6759	0.1910	0.1331
Stage 2	0	0.4260	0.5740
Stage 3	0	0	0

4.4 Conclusion

Through Markov chain modeling of the three stages of breast cancer patients: alive with no relapse, alive with relapse, and death, it shows that combined treatment of tamoxifen and radiation is more effective than single treatment of tamoxifen in preventing the reoccurrence of breast cancer. However, for patients who had relapse of breast cancer, single treatment of tamoxifen proves to be more effective than combined treatment with respect the survival probability. Transition probabilities between different stages during 2 years, 4 years, 5 years and 10 years are also constructed for predicting purpose.

Chapter 5

Statistical Comparison of Breast Cancer Patients with Different Histology Types

5.1 Background and Data

The relapse time and survival time of breast cancer patients could be different between groups who receive different treatment. Another important factor that should be taken into consideration is the histology type of breast cancer patients. In order to see the effect of histology type of breast cancer patients on survival time, relapse time, the previous mentioned data is divided into the following several subgroups shown in the Figure 5.1 based on the histology type and treatment they received, those 641 breast cancer patients can be divided into the following several subgroups shown in Figure 1 for later analysis. It can be noticed that the majority of the breast cancers are ductal (397) or mixed(174), only a small number are lobular (31), medullar(5), mucinous (16) or others(18).

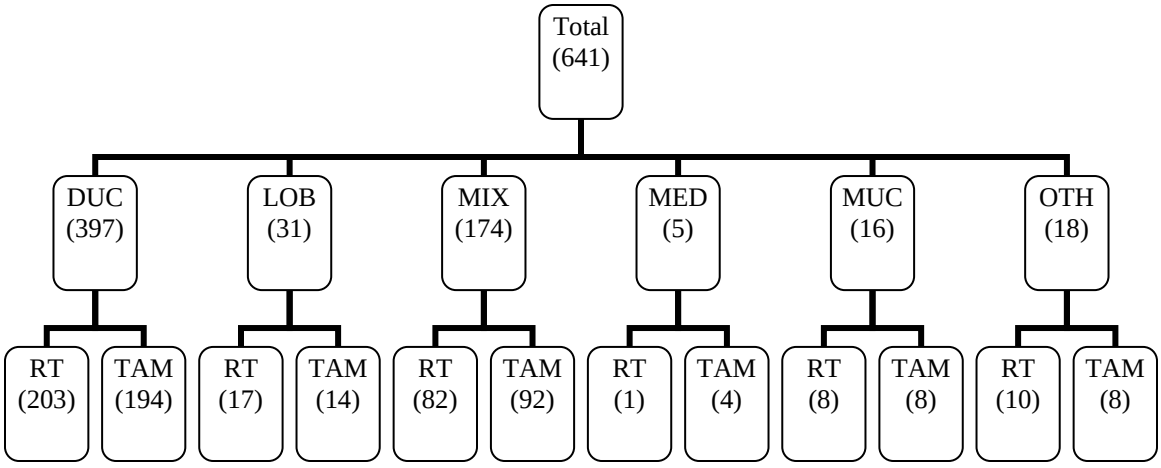

Figure 5.1.Breast Cancer Patients Grouped by Histological Types and Treatments

Information concerning potential prognostic factors (attributable variables) are pathsize (size of tumor in cm); hist (Histology: DUC=Ductal, LOB=Lobular, MED= Medullar, MIX=Mixed, MUC=mucinous, OTH=Other); hrlevel (Hormone receptor level: NEG=Negative, POS=Positive); hgb (Hemoglobin g/l); nodediss (Whether axillary node dissection was done: Y=Yes, N=No); age (Age of the patient in years). The response variables we are interested in are survival time and relapse time of a given patient.

In the next several sections, related research of breast cancer based on the same dataset is first reviewed, then we proceed to addresses the following questions: is there a difference of survival curves between different histological types; is there a difference of relapse time between different histological types; do patients of different cancer types react differently to treatments with respect to survival time and relapse time?

Despite the usefulness of the previous work done on the dataset, it does not take into consideration of the possible different behavior of different histological breast cancer types. For example, patients with different cancer type would react differently to the same treatments, and also there are potential significant differences among various cancer types with respect to survival time and relapse time. In this study, we divide the dataset into several subgroups based on the histology of the tumors as shown previously, and confine our study to the major two breast cancer types: ductal (DUC) and mixed (MIX) to address the following questions:

1. Is there significant difference for survival time among different histological breast cancer types?
2. Is there significant difference for relapse time among different histological breast cancer types?
3. Do patients with different histological breast cancer types react the same way to treatment with respect to survival time and relapse time?

5.2 Comparison of Survival Time and Relapse Time

It is of importance to see if the survival curves of patients in different cancer types are the same. Thus Kaplan-Meier [8] curves are plotted for each of the three major breast cancer types.

Kaplan-Meier estimates of the survival curves of relapse time for the two treatment groups are shown in Figure 5.2.

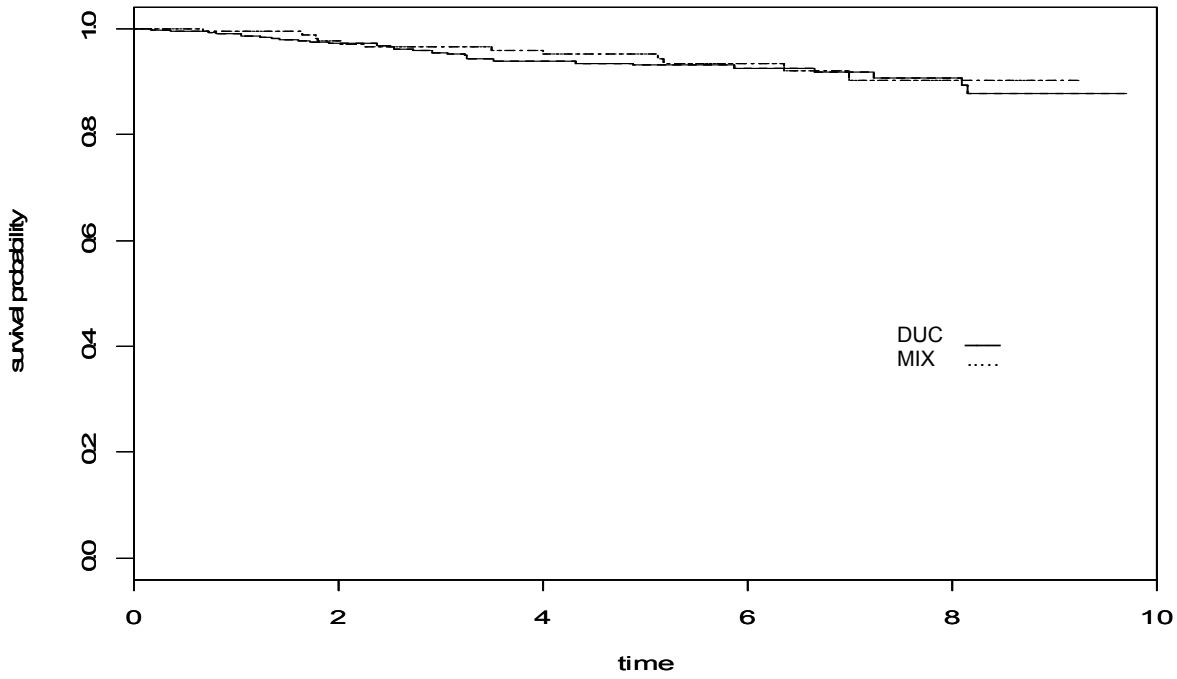

Figure 5.2. Kaplan-Meier Curves of Survival Time in DUC and MIX

As seen from the graph, the two curves almost overlap showing there is not much difference for survival time of the two breast cancer types. To verify that, Log-rank test is applied and p-value of 0.693 showing that there is no significant difference between survival curves of ductal breast cancer patients and mixed breast cancer patients. This suggests that there is homogeneity of survival time with respect to breast cancer types, so when analysis is conducted on survival time of breast cancer patients, there is no need to separate data into subgroups based on histology type.

Similar analysis is conducted for relapse time and the Kaplan-Meier survival curves are shown in Figure 5.3 below.

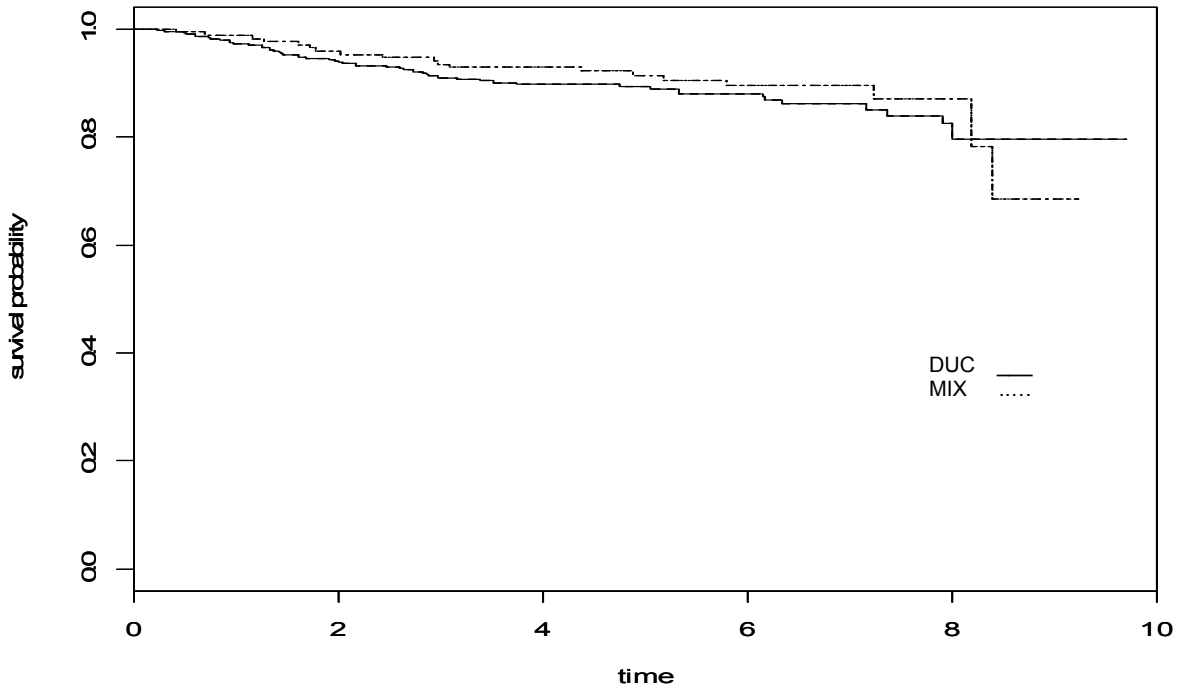

Figure 5.3. Kaplan-Meier Curves of Relapse Time in DUC and MIX

Furthermore, p-value 0.516 of Log-rank test indicates that there is no significant difference of relapse curve between Ductal and Mixed breast cancer patients.

5.3 Treatment Effectiveness of Different Histology

From the previous analysis we find that histology type does not affect the survival and reoccurrence behavior of breast cancer patients. Therefore, we proceed to investigate the treatment effects in different histology types. In another words, we are interested in if combined treatment and single treatment affect survival and relapse time in the same pattern for breast cancer patients with different histological types.

First, the survival curves of survival time and relapse time of patients of combined treatment group (RT+Tam) and single treatment group (Tam) are compared to see the overall effectiveness of the two treatments. For survival time and relapse time, the Kaplan-Meier curves are shown in Figure 5.4 and 5.5 below. And the p-values of the Log-rank test are 0.379 and 0.00192 for survival time and relapse time, respectively. Under significance level of 0.05, it can be concluded that there is no significant difference for survival time between the two treatments. However, combined treatment seems to be more effective than single treatment with respect to relapse time of breast cancer patients.

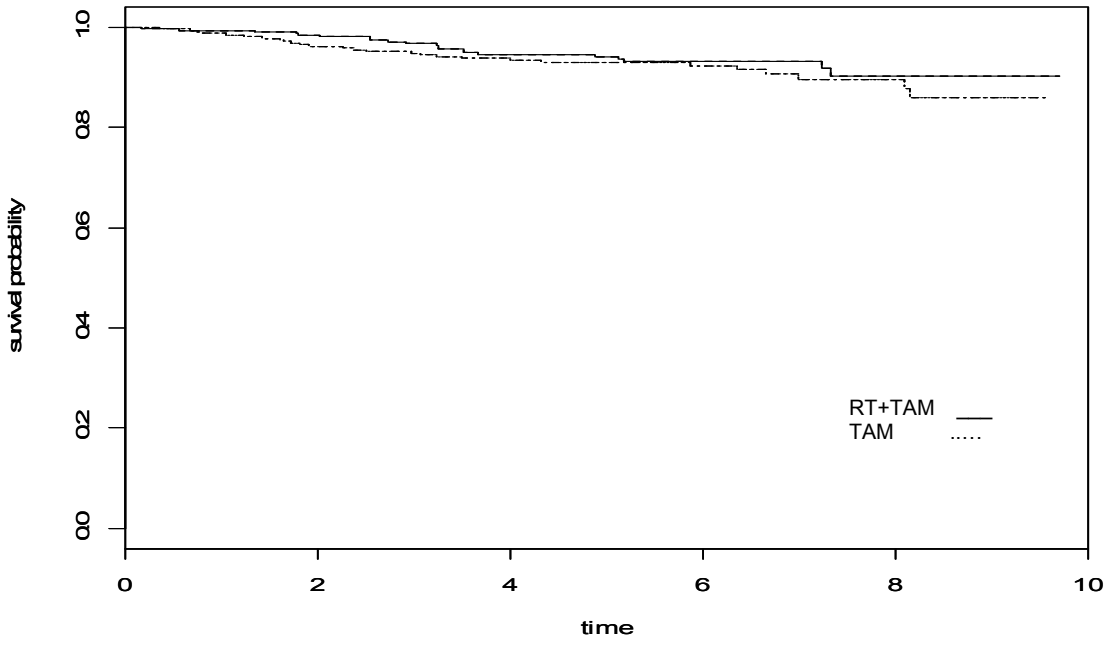

Figure 5.4. Kaplan-Meier Curves of Survival Time in RT+TAM and TAM

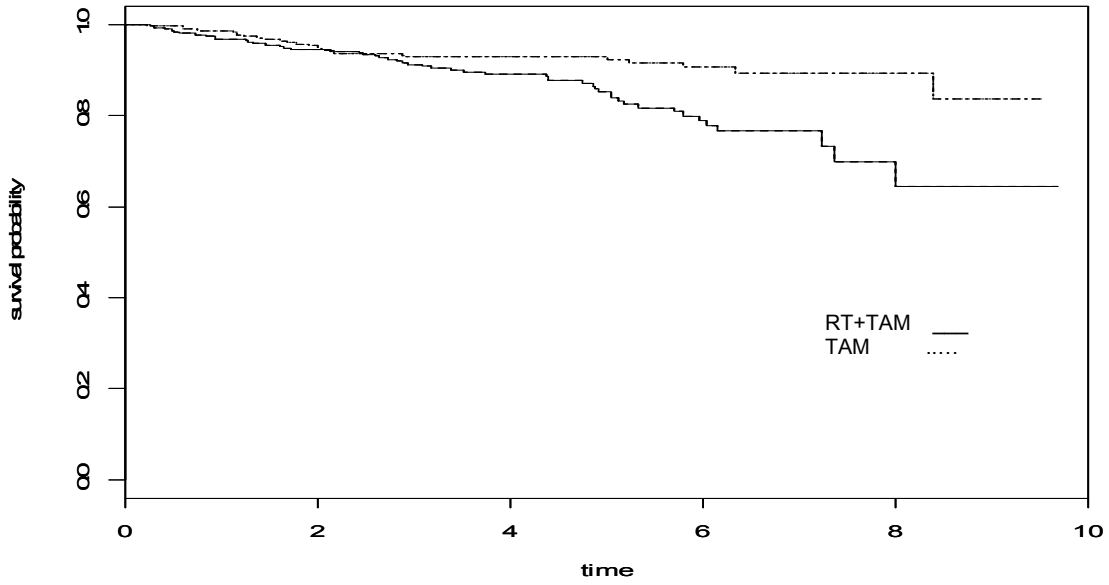

Figure 5.5. K-M Curves of Relapse Time of RT+TAM and TAM

Same analysis is conducted to the patients groups determined by various histology types. As mentioned above, we confine our analysis to the major two histological types: ductal and mixed because the other histology types do not have enough number of observations for statistical analysis. After running Log-rank test for survival time and relapse time with respect to two different treatments of each histology type, the p-values are obtained and listed in Table 5.1.

Table 5.1. Log-rank Test for Survival Time and Relapse Time

Histology Type	DUC		MIX	
Survival/Relapse	Survival Time	Relapse Time	Survival Time	Relapse Time
P-value	0.217	0.0114	0.708	0.0256

As can be observed from Table 5.1, for patients in both DUC and MIX group, survival time of patients who received combined treatment does not significantly differ from those who received single treatment. However, there is significant difference for relapse time between different treatment groups within both DUC and MIX cancer types. This result is consistent with the result obtained from the complete data that consists of all histology types. Thus, breast cancer type does not affect the choice of treatment with respect to survival time and relapse time.

5.4 Conclusion

Previous research on real breast cancer data is reviewed and the question of homogeneity among different breast cancer histology types is brought into consideration. By dividing the data based on histology type, Kaplan-Meier curve and Log-rank test are performed to test the homogeneity of survival time, relapse time, and treatment effect between the two major histology types: ductal and mixed. Results show there is no significant difference for survival time and relapse time between this two histology types of breast cancer, and treatment

effect is the same between the two breast cancer types as well. Thus, there are no significant treatment effects with respect to survival time for DUC, MIX, and the totality data, and combined treatment is more effective than single treatment with respect to relapse time for DUC, MIX and the totality data. This finding provide useful information for statistical analysis and modeling of breast cancer data in the way that all the observations from different histology types can be analyzed as a combined dataset because of the homogeneity among different histology types instead of splitting them into subgroups, and could effectively reduce the time and effort spent on modeling of breast cancer data.

Chapter 6

Sensitivity Analysis of Breast Cancer Doubling Time

6.1 Background

Growth rate of breast cancer tumors is a critical aspect to understanding the natural history of breast cancer and doubling time (DT) is most commonly used to indicate how fast tumors grow. To be more specific, doubling time is the time it takes for a tumor to double in size. With respect to that objective, we need to identify the geometrical behavior of the tumor so that we can obtain an estimate of its volume, we need to identify the mathematical behavior of the growth of the tumor and once we have determined the doubling times, we need to obtain the best possible fit of a probability distribution that characterizes their behavior. Peer P. G. M. et. al in an article concerning age dependent growth rate of primary breast cancer assumed one of four possible geometrical formulas to calculate the volume of the tumor. If we consider any of the others that are also commonly used, the overall results will be different. They also assumed exponential growth of the tumor when in fact actual data reveals that the tumor growth is decaying exponentially up to the age of 48, then it follows a quadratic growth between the age 49 and 68, and then follows a exponential growth up to age 100. Thus, assuming exponential growth for all ages will lead to misleading decisions.

Furthermore, in the subject paper, they assumed that the doubling time follows a two parameter lognormal probability distribution. However, our goodness-of-fit statistical testing shows that the lognormal is not the best probability distribution. The methodology and results of the subject article lead to a sequence of other publications.

Most recently, Green, L. (2009), in a conference presentation, “Age Dependent Screening” used Peer’s et al. results in their research on the subject matter, that is exponential growth of the tumor, the same geometrical volume formula and the lognormal probability distribution of the doubling time. Thus, the results that followed are subject to the above comments.

However, it is rarely possible for medical doctors to obtain the exact doubling time of a given breast cancer patient since there is no record of two mammograms of which the volume of larger tumor is exactly twice as large as the other of one given patient. Thus, we need to assume the shape of the tumor to estimate the volume of the tumor, and pre-specify the tumor growth model to estimate the doubling time. As a result of different volume and growth model assumptions, the probability behavior of doubling time can be significantly different.

In the present study we shall consider the four commonly used formulas to measure the volume of the tumor, in conjunction with negative exponential, linear and quadratic growth behavior of the tumor as a function of age. For each case we will calculate the doubling time and proceed to identify the appropriate

probability distribution for parametric analysis. Thus, having these various scenarios a physician can make the optimal decision concerning his patients. In other words, the following questions are addressed:

1. What is the mathematical growth behavior of breast tumor as a function of age?

2. What are the time (age) intervals that have the same analytical form of the average tumor growth?

3. What is the best mathematical expression that best characterize the behavior of the average tumor size for specific time (age) intervals?

4. Can we use these analytical characterizations of the average tumor size to predict or estimate the rate of tumor growth as a function of age?

5. Are the four different commonly used volumes to determine doubling time robust with respect to the analytical form of the growth of the tumor?

6. Do the resulting doubling time of the twelve possible configurations of volume and analytical form of the average size of tumor results in the same probability distribution?

7. Can we justify using the standard lognormal probability distribution for any of the four volumes or different analytical growth of the average tumor size?

8. How are the current findings compared with the commonly used standard lognormal probability distribution to characterize the probabilistic behavior of the doubling times?

6.2 Breast Cancer Data

The present data was first used by Heuser et.al. (1979) where 108 women underwent screening for breast cancer at the Breast Cancer Detection and Demonstration Project conducted at the University of Louisville. All of these 108 women received mammography as their screening method among which 45 were diagnosed on the initial mammography; thus, there is no previous mammography record. However, for the remaining 64 women, 32 had two or more mammograms based on which we conduct the subject study on tumor growth. For each of these 32 patients, the mammograms were displayed in series, and measurements of the major axis (a) and minor axis (b) are obtained from the medio-lateral views as shown in Figure 6.1 and the details are listed below in Table 6.1. It is clear that there is no growth of breast tumor in 4 patients.

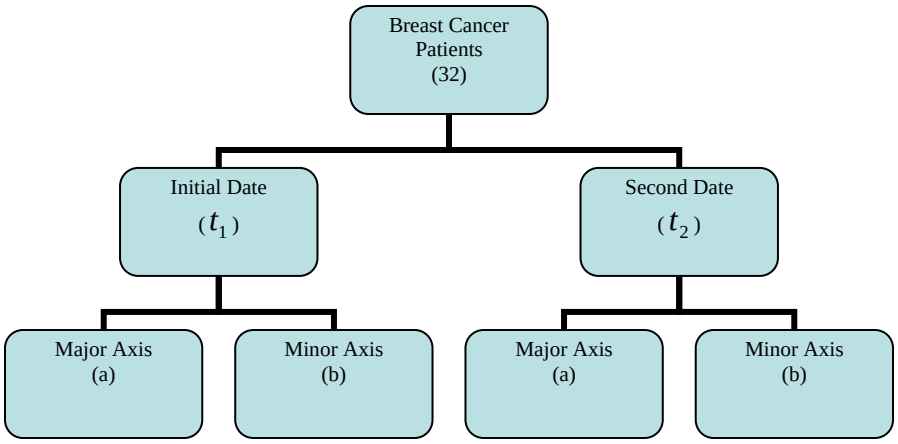

Figure 6.1 Breast Cancer Mammogram Data

Table 6.1 Date and Dimensions of Tumor Observations

Patient ID	Initial Date	Initial Dimension (mm)	Second Date	Second Dimension (mm)
1	04/25/75	12*11	01/06/76	30*14
2	10/15/74	22*17	01/15/75	27*20
3	01/30/74	10*8	01/28/75	22*15
4	10/14/74	7*5	07/10/75	16*8
5	04/12/76	30*20	04/28/77	70*28
6	08/29/73	6*3	02/11/75	20*5
7	11/08/73	20*7	11/21/74	30*12
8	06/04/75	13*10	06/18/76	18*16
9	01/02/75	30*15	04/01/75	40*15
10	01/24/74	6*5	01/10/75	18*5

11	07/17/75	15*15	07/21/76	26*22
12	02/24/76	5*4	07/22/76	8*4
13	03/17/75	8*6	03/09/76	12*8
14	10/09/75	11*8	09/22/76	12*12
15	11/18/74	13*11	11/07/75	20*13
16	01/07/75	18*15	05/29/75	18*17
17	06/23/75	20*18	04/02/75	25*20
18	02/11/75	12*11	01/09/76	20*15
19	10/24/74	18*7	06/23/76	19*9
20	03/23/75	70*70	02/24/76	90*80
21	10/24/74	22*12	10/09/75	24*14
22	03/23/75	18*10	04/01/77	21*11
23	02/20/75	10*6	02/24/76	13*6
24	10/08/75	25*17	04/01/76	25*17
25	12/06/74	5*3	03/07/75	5*3
26	06/26/75	6*5	12/10/75	6*5
27	02/01/75	8*6	08/01/75	9*6
28	06/03/75	18*13	06/04/76	19*13
29	08/03/74	10*8	08/10/75	11*8
30	02/06/75	6*6	02/23/76	7*6
31	01/30/74	38*27	01/28/75	44*26
32	07/15/74	20*12	08/01/75	20*12

The above table identifies the patient data of first and second mammograms along with the major & minor axis.

6.3 Geometrical Formulas of Tumor Volume

In the present study we will investigate the volume of the breast tumor that is commonly used in the subject matter.

Spherical Shape:

$$V = \frac{4}{3} \cdot \pi \cdot r^3;$$

where *r* is the radius calculated from the major axis, $r = a/2$ with *a* being the major axis. This formula is commonly used for purpose of simplicity as in Hesuer's paper and Green's presentation.

Averaged Spherical Shape:

$$V = \frac{4}{3} \cdot \pi \cdot r^3.$$

The radius *r* is the average of the major and minor axis, $r = \frac{2a + b}{6}$, where *a* is the major axis and *b* is the minor axis. This formula assumes the same shape of tumor but takes into consideration the two measurements instead of focusing on

only one dimension of the tumor. This form of the volume was also used in Hesure's paper.

Oblate Spheroid:

$$V = \frac{4}{3}\pi \cdot (\frac{a}{2})^2 \cdot (\frac{b}{2}) \ .$$

This formula assumes different shapes of volume as the two above. However, it gives more emphasis on the major axis. This formula is also mentioned in Hesure's paper.

Averaged Oblate Spheroid:

$$V = \frac{4}{3}\pi \cdot \frac{1}{2}a \cdot \frac{1}{2}b \cdot \frac{1}{2} \cdot (\frac{1}{2}a + \frac{1}{2}b) .$$

This formula uses the average of major and minor axis's, thus gives equal weight to the two measurements as used in Peer, et. al (1993) publication .

We shall study each of these volumes separately with respect to different growth rate of the tumor and the resulting probability distribution of the doubling time.

Figure 6.2 below shows a scattered diagram of the average breast tumor size as a function of the age of the cancer patients.

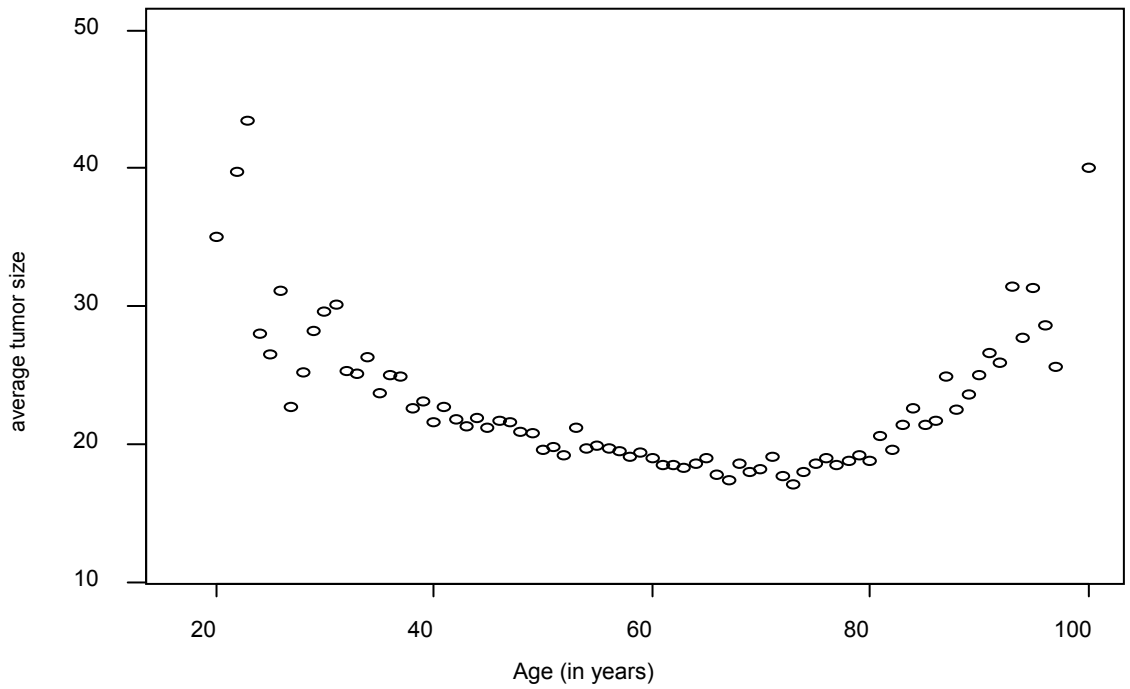

Figure 6.2 Average Tumor Size VS. Age

It is clear from the above scattered diagram that the mathematical configuration over all the ages is different. Thus, we shall group the average tumor size into three age groups. First group starts from 17 to 48, the second group from 49 to 78, and the last one from 79 years old and on. The first group clearly shows an exponential decay, and then flat linear or quadratic decrease and then exponential increase. Thus, we shall approximately partition the average tumor sizes into three groups as mentioned above and proceed to identify the best mathematical fitting function of the observed data in each of the three regions. We begin by graphing respectively the three age intervals, the first being from age 17 to 48. Figure 6.3 gives a better diagram of the data in that interval.

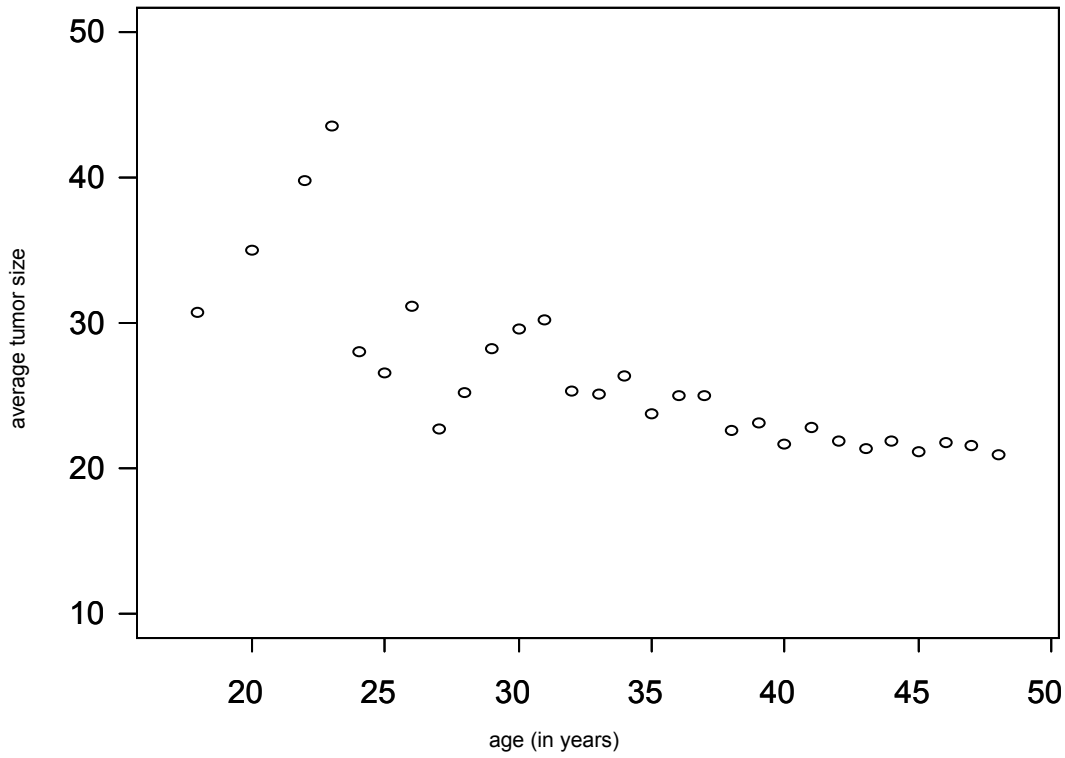

Figure 6.3 Average Tumor Size with Age 17-48

The best mathematical function that describes breast tumor size as a function of age in this interval is given by

$$f(t) = 42 \cdot e^{-0.015254 \cdot t},\ 17 \le t \le 48.$$

A residual analysis supports that $f(t)$ gives a good fit of the breast tumor growth in the given interval. Furthermore, we can differentiate the above equation with respect to t, and evaluate it in a specific age of interest to identify an estimate of the rate of growth of the size of the breast tumor.

The scattered diagram of the breast tumor size between the patients age between 49 and 78 is shown below by Figure 6.4.

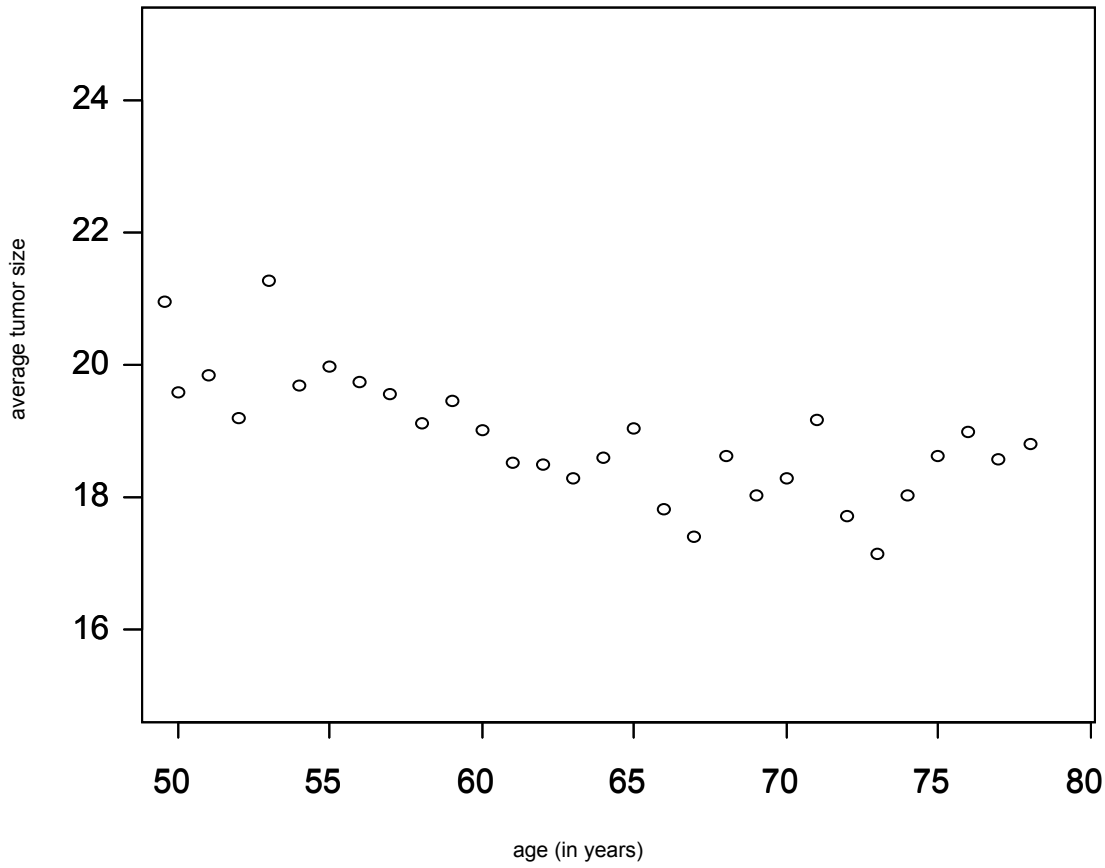

Figure 6.4 Average Tumor Size with Age 49-78

Thus, visually it could be approximated by either a linear or quadratic function and therefore, we will include both in our statistical analysis. If we assume it is linear, the best estimate is given by

$$f(t) = 22.67865 - 0.05925 \cdot t\,,\ 49 \le t \le 78\,.$$

The quadratic function form of the average breast tumor size in the second interval is given by

$$f(t) = 45.014257 - 0.764531 \cdot t + 0.00547 \cdot t^2,\ 49 \le t \le 78.$$

Residual analysis reveals that the analytical quadratic function gives a better fit than the linear function. However, since some medical scientist use the linear function, we will include it in the present analysis.

Finally, the scattered diagram for breast cancer patients older than 78 years old is given by Figure 6.5 below.

It is clear that the breast tumor size slowly increases exponentially after the age of about 78 years old. The best mathematical form that fits the tumor growth behavior in this age interval is given by

$$f(t) = 1.208443 \cdot e^{0.033947 \cdot t},\ t > 78.$$

Residual analysis that we performed supports the quality of the fit for the given function in the third age interval.

Again, once we have identified the analytical function form of the size of the breast tumor, we can differentiate it with respect to time (age) and evaluate at a specific age to determine the change of the tumor size.

Now we will proceed to obtain the doubling time for each volume and growth rate we have identified.

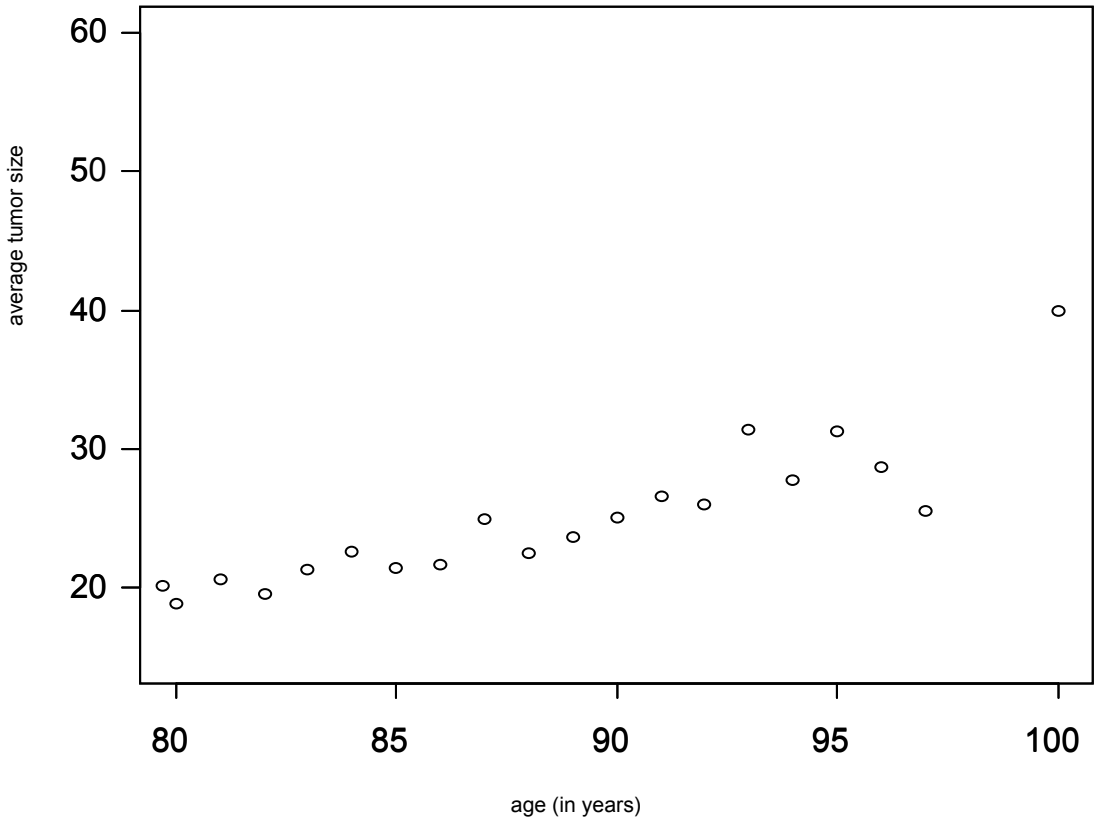

Figure 6.5 Average Tumor Size with Age above 78

6.4 Analytical Calculation of Doubling Time

After having identified the best possible analytical form of the average size of breast tumor as a function of age, we can proceed to calculate the doubling time of tumors. Since we only have two observations of each patient, for the growth function, we have to be specific so that there is only one parameter in the function. First, we have to choose the suitable geometric volume of the tumor and obtain the corresponding initial volume and second observation of volume.

After that we will see the calculation of the doubling time of breast tumor based on different growth assumptions.

Under linear growth, the following relation between two observations from a given individual can be obtained:

$$V_1 - V_o = \alpha t .$$

Thus,

$$\alpha = \frac{V_1 - V_0}{t},$$

and

$$DT = \frac{2V_0 - V_0}{\alpha} = \frac{V_0}{\alpha} = \frac{V_0}{\frac{V_1 - V_0}{t}} = \frac{V_0 t}{V_1 - V_0} \quad .$$

Estimates of α and DT in the above equations will be obtained using actual data.

Quadratic growth model takes into consideration the quadratic growth with respect to time. As mentioned above, since there are only two observations of each patient, only one parameter is involved in the quadratic function.

$$V_1 - V_0 = \alpha t^2 .$$

Thus,

$$\alpha = \frac{V_1 - V_0}{t^2},$$

and

$$DT = \sqrt{\frac{2V_0 - V_0}{\alpha}} = \sqrt{\frac{V_0}{\alpha}} = \sqrt{\frac{V_0}{\frac{V_1 - V_0}{t}}} = \sqrt{\frac{V_0 t}{V_1 - V_0}} \quad .$$

Under exponential growth, the following growth model can be obtained:

$$V = V_0 \cdot \exp(\beta t),$$

Where V_o is the initial tumor size, and β and doubling time DT can be obtained as

$$\beta = \frac{\ln V_1 - \ln V_o}{t} \quad ,$$

and

$$DT = \frac{\ln 2}{\beta} \quad .$$

The above two expressions are the true analytical forms and we will use actual data to obtain the approximate estimates.

Thus, we will use the derived analytical forms of DT for the four different volumes to obtain the actual doubling times.

If we use the finding that the breast tumors follow linear growth, we can calculate the doubling times under each of the four different volumes: spherical; averaged spherical; oblate spherical; averaged oblate spherical. The results of the calculated doubling times are shown in Table 6.2. Because the last observation has no change in major axis, thus there is no change in spherical volume which causes the doubling time not applicable.

Table 6.2 Doubling Time under Linear Growth

Spherical	**Average Spherical**	**Oblate Spherical**	**Averaged Oblate Spherical**
17.50427	64.57835	83.94580	50.32479
108.42457	129.10155	131.68303	124.32018
37.62438	51.85714	53.90255	48.53482
24.58487	47.44559	55.44759	42.60181
32.55380	78.84934	106.62313	70.52073
14.73484	36.30769	64.29148	36.79299
159.15789	125.73555	110.91018	126.00000
229.67263	153.33034	149.33495	192.76923
64.94595	151.37008	267.00000	167.03911
13.50000	80.51613	175.50000	141.35294
87.93395	133.86104	135.60104	66.56897
48.12661	172.38178	248.33333	120.61721

150.73684	205.27523	214.80000	131.47059
1170.07305	254.67568	239.93750	327.08969
134.02344	278.69026	308.15827	236.00000
201.44262	523.18977	499.21875	702.90000
100.28571	340.55765	353.45455	297.79592
2078.39533	166.94660	173.40157	148.05788
335.89119	670.56684	491.34247	703.84615
1173.42569	550.07197	556.45494	481.98513
1258.58268	787.28970	721.87500	828.48101
308.27068	1648.27603	1797.57085	1585.71429
427.05991	789.45389	1230.00000	678.62069
2084.07400	1148.29500	1448.00000	881.39130
1123.86707	5262.16597	6606.00000	4095.72000
649.70079	3103.11913	3720.00000	2308.96552
657.11718	2169.25414	2292.00000	1447.57895
NA	2665.43211	4924.49853	1807.95146

Similarly, for the quadratic growth of the breast tumors, we can calculate the doubling times under each of the four different volume formulas: spherical; averaged spherical; oblate spherical; averaged oblate spherical. The results are given in Table 6.3 below.

Table 6.3 Doubling Time under Quadratic Growth

Spherical	**Averaged Spherical**	**Oblate Spherical**	**Average Oblate Spherical**
66.94097	128.5771	146.5951	113.5039
99.87522	108.9832	110.0674	106.946
116.866	137.2011	139.8808	132.7333
81.32238	112.9728	122.1286	107.0509
111.3687	173.3251	201.5525	163.9158
88.45451	138.8502	184.7668	139.7751
245.2788	218.0093	204.7536	218.2384
295.4244	241.3825	238.2169	251.9422
76.02756	116.0687	154.1525	112.1624
68.83676	168.1106	248.1945	152.8585
180.3762	222.5502	223.9919	211.2543
84.68096	160.265	192.3582	139.9611
232.3011	271.0877	277.3056	262.7002
639.027	298.1305	289.3755	337.8673
217.8171	314.0961	330.2848	289.0398
196.6647	272.5673	266.25	315.9301
191.0602	255.709	260.5058	239.1167
872.177	246.5128	251.2333	232.1488

356.3241	495.4064	424.0653	507.5507
640.858	455.9904	458.6284	426.8376
965.0654	524.9299	502.6492	538.4871
337.2712	1104.411	1153.344	1083.249
278.0249	539.73	673.6987	500.4109
874.56	455.8963	511.9453	399.4144
646.5899	1389.682	1557.049	1226.022
498.1824	1074.412	1176.367	926.7876
488.399	910.3049	935.7051	743.623
NA	983.6421	1337.009	810.115

Finally, as we have shown that in the third age interval, for patients older than 78 years of age, the breast tumor growth follows the exponential function , we can calculate the doubling times under each of the four different volume formulas: spherical; averaged spherical; oblate spherical; averaged oblate spherical.

Table 6.4 Doubling Time under Exponential Growth

Spherical	**Averaged Spherical**	**Oblate Spheroid**	**Average Oblate Spheroid**
64.55217	110.74799	126.8724	98.24536
103.7944	118.52572	120.358	115.1284
106.3733	121	122.9973	117.70724

75.18303	98.26084	105.5403	93.70481
103.89462	149.7647	173.7151	142.23461
101.90185	133.89533	165.3744	134.50317
215.39842	188.78535	176.6209	189
269.79901	211.30055	208.1474	222.0329
71.47948	133.39783	214.4385	126.32282
73.81878	144.91768	221.4563	132.49819
155.41991	193.48532	194.8772	182.79169
73.2469	165.79857	219.7407	136.30841
204.00168	245.83176	252.9968	236.37022
926.73079	280.29596	269.4032	333.1652
189.86678	299.2805	320.7961	267.79066
198.80216	409.89326	393.1944	534.91821
198.80216	297.65657	306.7415	267.46283
1564.0522	218.07114	223.0551	203.33541
347.51902	582.46856	455.7098	605.88872
929.38618	500.91682	505.442	452.51952
1109.1505	659.58923	613.6952	688.45059
324.957	1383.1473	1487.7378	1339.2671
355.0586	666.93933	974.871	1339.2671
1568.3256	857.13751	1065.1758	671.71212
901.79507	3773.2189	4704.9767	2964.3103
572.56148	2277.4113	2705.3852	1726.1698

572.09291	1632.4261	1717.6844	1130.6136
NA	1970.6666	3537.7173	1375.1483

6.5 Probability Distribution of Doubling Time

Now that we have calculated the doubling time for each configurations of tumor growth behavior and the different volume of the tumor as shown in Table 6.2, 6.3 and 6.4, we will proceed to identify the best possible probability distribution that characterizes the probabilistic behavior of the doubling time. That is, we want to statistically using general goodness-of-fit test to identify the best possible probability distribution that characterizes the behavior of the doubling time. To find the best probability distribution, we utilize three different types of goodness-of-fit tests including Kolmogorov-Smirnov (1971), Anderson-Darling (1952) and Chi-Square (1954) tests. Once we have identified the best possible probability distribution, we shall give the basic and useful statistics of each of the scenarios for proper and relevant interpretation for comparison purpose along with confidence limits of the true doubling time.

Use the above test criteria for goodness-of-fit, we have found the following probability distribution of each growth function.

Case 1. Linear Growth and Four Different Volumes

(a) Linear Growth and Spherical Volume: Using the data for doubling time as given by Table 6.2 and the goodness-of-fit test that we mentioned above, the best probability distribution is Fatigue Life with the following probability density function (p.d.f).

$$f(x) = \frac{\sqrt{x/174.79} + \sqrt{174.79/x}}{2 \cdot 1.8276 \cdot x} \cdot \Phi\left(\frac{1}{1.8276}\left(\sqrt{\frac{x}{174.79}} - \sqrt{\frac{174.79}{x}}\right)\right).$$

(b) Linear Growth and Average Spherical Volume: The three-parameter lognormal is the best fit distribution whose p.d.f. is given by

$$f(x) = \frac{\exp\left(-\frac{1}{2}\left(\frac{\ln(x - 33.052) - 5.4105}{1.7292}\right)^2\right)}{(x - 33.052) \cdot 1.7292 \cdot \sqrt{2\pi}}.$$

(c) Linear Growth and Oblate Spherical Volume: The four-parameter Pearson probability distribution is the best fit with p.d.f.

$$f(x) = \frac{((x - 26.447)/18.82)^{5.2477}}{18.82 \cdot B(6.2477, 0.76608)(1 + (x - 26.447)/18.82)^{7.01378}}.$$

(d) Linear Growth and Average Oblate Spherical: The three-parameter lognormal is the best with p.d.f.

$$f(x) = \frac{\exp\left(-\frac{1}{2}\left(\frac{\ln(x - 34.391) - 5.2388}{1.7891}\right)^2\right)}{(x - 34.391) \cdot 1.7891 \cdot \sqrt{2\pi}}.$$

In Table 6.5 below, we have calculated the basic statistics and 95% confidence limits of the true doubling time along with the same estimates for the commonly used lognormal probability distribution for comparison purpose. As can be seen that there are significant differences of the basic statistics and confidence limits between the justified p.d.f.s and the one that is not statistically acceptable.

Table 6.5 Distribution of Doubling Time under Linear Growth

Volume	**Distribution**	**Mean**	**S.D**	**95% lower limit**	**95% upper limit**
Spherical	Fatigue Life	466.69	726.68	11.84	2580.4
	Lognormal	573.99	1741.8	9.0575	3563.2
Averaged Spherical	Lognormal (3P)	1030.8	4336.1	40.601	6664.9
	Lognormal	769.1	1747.2	22.048	4354.5
Oblate Spherical	Pearson 6 (4P)	969.97	2378.3	54.229	15768.0
	Lognormal	898.33	2028.5	26.075	5074.4
Averaged Oblate Spherical	Lognormal (3P)	963.54	4509.4	40.016	6284.8
	Lognormal	655.05	1387.4	21.682	3607.4

Case 2. Quadratic Growth and Four Different Volumes

We shall use the same statistical criteria to identify the best possible p.d.f. that characterize the doubling time data given in Table 6.3.

(a) Quadratic Growth and Spherical Volume: Using the data for doubling time and the goodness-of-fit test that we mentioned above, the best probability distribution is Johnson SB with the following probability density function (p.d.f).

$$f(x) = \frac{0.50062}{1034\sqrt{2\pi}x(1-x)} \exp(-\frac{1}{2}(0.84509 + 0.50062 \ln(\frac{x}{1-x}))^2), \text{ where } 61.303 \le x \le 1095.303$$

(b) Quadratic Growth and Averaged Spherical Volume: The three-parameter Frechet is the best fit distribution whose probability density function (p.d.f.) is given by

$$f(x) = \frac{1.1887}{138.76}(\frac{138.76}{x - 60.814})^{2.1887} \exp(-(\frac{138.76}{x - 60.814})^{1.1887}) .$$

(c) Quadratic Growth and Oblate Spherical: The Burr probability distribution is the best fit with p.d.f.

$$f(x) = \frac{5.7215 \cdot 0.22469 \cdot (\frac{x}{160.19})^{4.7215}}{160.19(1 + (\frac{x}{160.19})^{5.7215})^{1.22469}} .$$

(d) Quadratic Growth and Average Oblate Spherical: The p.d.f. of the best possible probability distribution Frechet is given by

$$f(x) = \frac{1.5838}{196.1}\left(\frac{196.1}{x}\right)^{2.58381} \exp\left(-\left(\frac{196.1}{x}\right)^{1.5838}\right).$$

The basic statistics along with the corresponding 95% confidence limits are given in Table 6.6 below.

Table 6.6 Distribution of Doubling Time under Quadratic Growth

Volume	**Distribution**	**Mean**	**S.D**	**95% lower limit**	**95% upper limit**
Spherical	Johnson SB	331.47	274.66	65.101	994.63
	Lognormal	336.95	345.11	44.756	1238.0
Averaged Spherical	Frechet (3P)	873.98	475.31	107.09	3118.6
	Lognormal	403.76	350.54	70.173	1324.7
Oblate Spherical	Burr	712.34	477.81	110.47	2823.9
	Lognormal	435.77	374.89	76.813	1420.7
Averaged Oblate Spherical	Frechet	473.19	422.42	86.007	1997.8
	Lognormal	377.7	316.05	69.464	1207.9

It can be observed that, under linear growth assumption as shown in the previous section, Fatigue Life is the best fitting probability distribution. However, when quadratic growth function is considered, the best-fitting distribution changes to Johnson SB distribution, and the statistical properties such as mean and 95% confidence bands change as well. This suggests doubling time is sensitive to the growth assumption chosen and thus careful research would be done on the shape of breast cancer before assuming any probability distribution for the doubling time.

Furthermore, since doubling time is used widely as an indication of how fast tumors grow, mean and 95% confidence bands give both doctors and patients useful information on the progression of a breast cancer tumor. However, these statistical properties vary significantly from distribution to distribution. Not only tumor shape affects the doubling time distribution, tumor growth function also affects the way tumor volume is calculated and subsequently the doubling times calculated from tumor volumes as illustrated both in the table above and the analysis below.

Case 3. Exponential Growth and Four Different Volumes

Utilizing the same approach, we calculate doubling times of the 28 breast cancer patients under four different geometric shape scenarios: spherical, averaged spherical, oblate spherical, averaged oblate spherical as shown below.

(a) Exponential Growth and Spherical Volume: Using the data for doubling time and the goodness-of-fit test that we mentioned above, the best probability distribution is the three-parameter lognormal with the following p.d.f.

$$f(x) = \frac{\exp(-\frac{1}{2}(\frac{\ln(x-62.612)-4.8435}{1.7292})^2)}{(x-62.612)\cdot 1.7292\cdot\sqrt{2\pi}}.$$

(b) Exponential Growth and Averaged Spherical Volume: The three-parameter lognormal is the best fit distribution whose probability density function (p.d.f.) is given by

$$f(x) = \frac{\exp(-\frac{1}{2}(\frac{\ln(x-95.239)-5.1591}{1.6535})^2)}{(x-95.239)\cdot 1.6535\cdot\sqrt{2\pi}}.$$

(c) Exponential Growth and Oblate Spherical: The three-parameter Frechet probability distribution is the best fit with p.d.f.

$$f(x) = \frac{0.87589}{135.21}(\frac{135.21}{x-80.287})^{1.875841}\exp(-(\frac{135.21}{x-80.287})^{0.87584}) \quad .$$

(d) Quadratic Growth and Average Oblate Spherical: The p.d.f. of the best possible probability distribution, three-parameter Fatigue Life is given by

$$f(x) = \frac{\sqrt{(x-84.284)/179.88}+\sqrt{179.88/(x-84.284)}}{2\cdot 1.8229\cdot(x-84.284)}\cdot\Phi(\frac{1}{1.8229}(\sqrt{\frac{x-84.284}{179.88}}-\sqrt{\frac{179.88}{x-84.284}}))$$

Their 95% confidence limits of the true doubling time along with those of the commonly used Lognormal probability distribution of each group of doubling times are given in Table 6.7.

Table 6.7 Distribution of Doubling Time under Quadratic Growth

Volume	**Distribution**	**Mean**	**S.D**	**95% lower limit**	**95% upper limit**
Spherical	Lognormal (3P)	628.61	2459.9	66.894	3824.6
	Lognormal	419.36	551.35	35.631	1809.0
Averaged Spherical	Logrnnomal (3P)	778.0	2950.5	102.05	4542.0
	Lognormal	588.58	793.94	47.653	2578.4
Oblate Spherical	Frechet	770.88	1529.0	110.75	9073.2
	Lognormal	685.92	983.2	49.467	3113.7
Averaged Oblate Spherical	Fatigue Life (3P)	563.05	744.44	96.523	2728.1
	Lognormal	543.93	699.01	48.221	2313.9

From the above analysis, it is clear that the probability behavior of doubling time varies significantly with respect to the tumor volume, and under all circumstances, the popular lognormal probability distribution is far from the best fitting probability distribution that characterizes the statistical behavior of doubling time. Besides that, there is no consistency between the mean and 95% confidence limits constructed from the best-fitting distributions.

Thus, it is of great importance to investigate the shape of the tumor as well as the tumor growth pattern before any statistical analysis of the doubling time are calculated from the volume and tumor growth assumptions vary significantly from scenario to scenario.

6.6 Conclusion

As a result of the present study, we can conclude that

1. The analytical growth behavior of the average breast cancer tumor size is not exponential for all ages as commonly assumed. For example, we have found that age between 17 and 48, the growth is exponentially decaying. For ages between 48 and 78, the tumor growth is best characterized by a quadratic analytical function. However, not quite as good as the linear behavior. And exponential growth function best fits the average tumor size for patient older than 78 years.

2. There are four commonly used formulas for determining the volume of breast tumor, and using these four configurations of the volume results in different

doubling times. Thus, the results are sensitive with respect to the choice of the specific volume we use.

3. Calculation of the doubling time of each volume and different mathematical growth results in twelve cases with different probability distribution that characterize the probabilistic behavior of the doubling time. Table. 6.8 gives a summary of the actual probability distribution that one should use along with the types of growth rate the volume.

Table 6.8 Summary of Results

	Spherical	Averaged Spherical	Oblate Spheroid	Averaged Oblate Spheroid
Linear	Three-parameter lognormal	Three-parameter lognormal	Three-parameter Frechet	Three-parameter Fatigue Life
Quadratic	Fatigue Life	Three-parameter lognormal	Six-parameter Pearson	Three-parameter lognormal
Exponential	Johnson SB	Three-parameter Frechet	Burr	Frechet

4. Our findings clearly show that the commonly used exponential growth of breast tumor will lead to incorrect decisions.

5. The commonly used standard lognormal probability distribution to characterize the behavior of the doubling time is not correct and will lead to wrong decisions.

6. One should be very careful in selecting one of the four volumes for a given situation because all are sensitive with respect to growth rate an age of the patients.

Chapter 7

Statistical Modeling of Lung Cancer Mortality Time

7.1 Background and Data

Lung cancer is a disease of uncontrolled cell growth in tissues of the lung and one of the deadliest common cancers in both men and women. Annually, 1.3 million deaths are caused by lung cancer worldwide. It is more common in older adults than in people under age 45. It is known that cigarette smoking is the leading cause of lung cancer, which means the risk of getting lung cancer is strongly associated with the number of cigarettes smoked per day and the time when one starts and quits smoking. Secondhand smoke contributes to lung cancer as well and there is a chance that people who have never smoked will get lung cancer.

The data that were first collected in 1982 and the mortality follow-up in the dataset is complete through 2006. It encompasses 1.2 million subjects in 50 states. Only data from those who got lung cancer are included in this study, whether from smoking or non-smoking. For ex-smokers, the total number of lung cancer patients is 5,316, of which 1,523 are females and 3,793 are males. For

non-smokers, the total number of lung cancer patients is 2,010, of which 1,386 are females and 624 are males.

Although there are many other causes associated with lung cancer such as air pollution, radon gas, asbestos, family history of lung cancer, radiation therapy to the lungs, and exposure to cancer-causing chemicals, we confined our interest in smoking only due to the lack of data pertaining to these other causes. The four variables of interest are the number of cigarettes per day (CPD), the age at which an individual started smoking (t_s), the age at which an individual quit smoking (t_q), smoking duration (in years), $t_q - t_s$ (DUR), and mortality time (t_m). The following diagram gives a clearer view of what the data looks like.

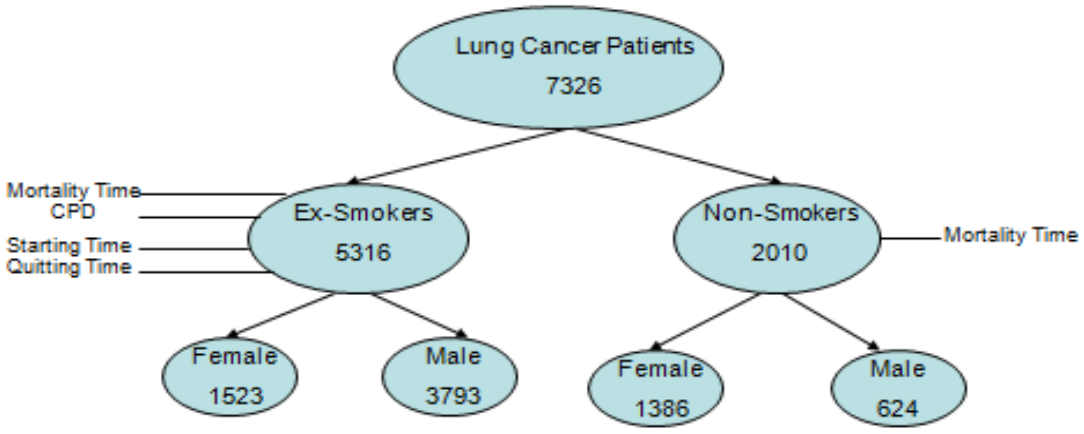

Figure 7.1. Lung Cancer Data

The objective of this study is to address the following questions related to some of the most important entities in lung cancer: cigarettes per day (CPD), time the patient started smoking (t_s), time the patient quit smoking (t_q), duration of

smoking which is defined as the difference between the two above mentioned times (DUR), and, most importantly, the mortality time (t_m):

(1) What is the probabilistic nature of mortality time in ex-smoker lung cancer patients and non-smoker lung cancer patients, for female, male, and the totality of female and male patients?

(2) Is there significant difference of mortality time between ex-smoker and non-smoker patients?

(3) For ex-smokers, are there any differences with respect to the key variables such as mortality time, CPD, and duration of smoking between female and male patients?

(4) For non-smokers, can we notice a difference in mortality time between female and male patients?

(5) Can we accurately predict mortality time given information on CPD, starting time and quitting time for a specific lung cancer patient who smokes?

7.2 Results of Parametric Analysis

Before modeling mortality time as a function of CPD, DUR, t_s , t_q , basic parametric analysis should be performed to understand its probabilistic behavior. More than 40 different classical distributions are fit to the data and three goodness-of-fit tests, Kolmogorov-Smirnov. Anderson-Darling and Chi-Square are conducted for the mortality time of lung cancer patients for female ex-

smokers, male ex-smokers, all ex-smokers, female non-smokers, male non-smokers, all non-smokers, respectively. The parameters of the three distributions ranked the highest by goodness-of-fit are listed and the 90% and 95% confidence intervals are constructed. The results appear in the following tables where the first table shows the mean and variance of the best fitting distribution and Table 7.2 shows the 90% and 95% confidence bands of the true mean of the estimated distributions.

As can be seen from the tables, the best fitting distribution is always Johnson SB followed by Beta distribution and three-parameter Weibull distribution. With this finding, we can find the mean and variance and construct the 90% and 95% confidence intervals for mortality time. An interesting thing worth noting is that no matter which distribution is chosen, Beta, Johnson SB, or three-parameter Weibull, their 90% and 95% confidence intervals are very close. Although Johnson SB appears to be the best fit for both female and male ex-smokers and theoretically a likelihood ratio test could be applied to test the difference of means of mortality time in these two groups, the parametric comparison is not used here due to the extremely complicated calculation. Furthermore, it appears that there are no significant differences between the means and variances of mortality times for females and males. However, it is observed that ex-smoker lung cancer patients have decreased mortality compared to non-smoker lung cancer patients. These parametric results lead us to compare the key variables in the different groups, between female and males, or between non-smokers and ex-smokers using non-parametric methods as described next.

Table 7.1 Mean and Standard Deviation of Fitted Distributions

	Johnson SB	Beta	Three - parameter Weibulll
Female ex-smokers	NA	73.995 (8.9577)	74.007 (8.9365)
Male ex-smokers	NA	74.543 (8.1875)	74.542 (8.2119)
Ex-smokers	NA	74.384 (8.4155)	74.387 (8.428)
Female non-smokers	NA	76.117 (10.213)	76.148 (10.165)
Male non-smokers	NA	76.011 (9.6368)	76.015 (9.6551)
Non-smokers	NA	76.085 (10.041)	76.103 (10.022)

Table 7.2 Confidence Interval of the True Mean

	Johnson SB	Beta	Three - parameter Weibulll
Female ex-smokers	(59.9, 87.622) (55.419, 90.07)	(58.586, 88.016) (55.364, 90.327)	(58.456, 87.916) (55.371, 90.168)
Male ex-smokers	(60.527, 87.493) (57.827, 89.55)	(60.557, 87.52) (57.849, 89.591)	(60.304, 87.386) (57.519, 89.482)
Ex-smokers	(59.9, 87.622) (57.061, 89.701)	(59.98, 87.659) (57.057, 89.85)	(59.751, 87.541) (56.871, 89.68)
Female non-smokers	(58.145, 91.777) (54.682, 93.933)	(58.304, 91.886) (54.794, 94.142)	(58.277, 91.742) (54.582, 94.205)
Male non-smokers	(58.888, 90.419) (55.014, 92.541)	(58.87, 90.391) (55.045, 92.434)	(58.671, 90.298) (54.727, 92.425)
Non-smokers	(58.367, 91.373) (54.761, 93.541)	(58.461, 91.428) (54.811, 93.643)	(58.354, 91.302) (54.567, 93.657)

7.3 Results of Nonparametric Comparison

After finding the mortality time for both ex-smokers and non-smokers, the next question is whether there is significant difference of mortality time between ex-smokers and non-smokers, between female and male groups. We are also interested in the impact of the number of cigarettes smoked per day and duration of smoking on female and male smokers (for ex-smokers only since they are all zeros for non-smokers). The Wilcoxon Rank Sum two - sample test by Wilcoxin (1945) was performed to detect location differences. The results are shown in Table 3. For all these tests of hypothesis, we first set the null hypothesis to be two-sided, if p-value is large enough; we fail to reject the null hypothesis. However, if p-value is small which suggests the rejection of the null hypothesis, we proceed to test the one-sided hypothesis.

Table 7.3 Wilcoxon Two-Sample Test Result

H_o	$\bar{t}_{m(ex)} \geq \bar{t}_{m(non)}$	$\bar{t}_{m(female-ex)} = \bar{t}_{m(male-ex)}$	$\bar{t}_{m(female-non)} = \bar{t}_{m(male-non)}$	$\overline{CPD}_{male} \leq \overline{CPD}_{female}$	$\overline{DUR}_{male} \leq \overline{DUR}_{female}$
p-value	0.0018	0.1180	0.8106	<0.0001	0.0001
Conclusion	Reject	Accept	Accept	Reject	Reject

(1) Mortality time between ex-smokers and non-smokers

Mortality time of ex-smokers and nonsmokers are compared using the Wilcoxon two-sample test. Under hypothesis that $\bar{t}_{m(ex-smokers)} \geq \bar{t}_{m(non-smokers)}$, the p-value is 0.0018. Thus, using a significance level of 0.05, we reject the null hypothesis and conclude that non-smoker lung cancer patients have longer mortality time than that of ex-smokers.

(2) Ex-Smokers mortality time between female and male

There is no significant difference between the female and male smokers with respect to the death time from lung cancer. For two-sided hypothesis, the p-value is 0.1180 and the p-value for one-sided hypothesis is 0.0590 which is still higher than 0.05. Thus, using a significance level of 0.05, death time of female ex-smokers is not significantly different from that of male ex-smokers (p = 0.1180).

(3) Non-Smokers mortality time between female and male

As can be seen from the two-sided p-value 0.8106 and one-sided p-value of 0.4053, there is insufficient evidence to conclude that there is a difference between female and male non-smoker lung cancer patients. Thus, no difference of mortality time can be found between female and male lung cancer patients, both in ex-smokers and non-smokers, which is consistent with the conclusion from parametric analysis.

(4) Ex-smokers CPD

As can be seen from the p-value, which is less than 0.0001 under the hypothesis that $C\bar{P}D_{male} \leq C\bar{P}D_{female}$, there is strong evidence that males tend to have more cigarettes per day than females.

(5) Ex-smokers DUR

Similarly, the p-value of 0.0001 under null hypothesis that $D\bar{U}R_{male} \leq D\bar{U}R_{female}$ suggests that smoking duration for male smokers exceeds the smoking duration of female smokers.

In summarizing the above analyses, the following conclusions are obtained:

(1) There is no significant difference in mean mortality time for females and males for both ex-smokers and non-smokers. Ex-smokers tend to have a shorter mortality time than non-smokers.

(2) For CPD, mean CPD of males is larger than that of females.

(3) For DUR, males have longer duration of smoking than females.

7.4 Results of Modeling of mortality time

After finding the probabilistic behavior of mortality time and comparison of key entities with respect to race and smoking status, we proceed to investigate the relation between mortality time and other attributable variables such as CPD,

time an individual started smoking (t_s), and time an individual quit smoking (t_q), where multiple regression models is most commonly used tool. First, for the female ex-smokers, multiple regression models were run and the backward selection method is used to eliminate any variables that do not significantly contribute. However, after multiple regression is applied using mortality time as the response variable and CPD , t_s , t_q , and the second-order interaction between them as well as the quadratic terms, the R-square (0.2249) of the full model is pretty small which indicates multiple regression model is not a good choice here. The same procedure was applied to male ex-smokers, where the R-square was only 0.1301.

Although multiple regression models do not perform well, they give us some guidance on which variables are not important and can be eliminated in the modeling process later. We then proceeded to utilize the survival regression model, also called the accelerated failure time (AFT) model, which assumes certain distribution of the response variables.

(1) AFT model

When covariates are considered, we assume that the relapse time has an explicit relationship with the covariates. Furthermore, when a parametric model is considered, we assume that the relapse time follows a given theoretical probability distribution and has an explicit relationship with the covariates.

Females

Survival regression models were run using statistical software including exponential, generalized gamma, loglogistic, lognormal, logistic, normal, and Weibull distributions. Their log likelihoods were: -1532, 1215, 1189, 1194, -5326, -5309, 1185. Thus, the generalized gamma was determined to prove the best fit and backward elimination was used to eliminate the unimportant variables. In the final model, the variables left are CPD, t_s , interaction between t_s and t_q , interaction between CPD and t_q, and quadratic terms of CPD and t_q.

All terms in the model are significant and they are ranked according to their significance. The quadratic term of quitting time ranks first followed by CPD, starting time, interaction between starting time and quitting time, quadratic term of CPD, and interaction between CPD and quitting time.

The following percentage plot Figure 7.2 is obtained for the final model.

After the estimations of the parameters in the model are obtained, the value of log(T) can be predicted by plugging the parameters into the equation, and thus mortality time T can be calculated by simply taking natural exponentials. The mean and standard deviation of the difference between predicted mortality time and observed mortality time are 0.1378148 and 7.911363, respectively. However, the mean and variance of the difference between predicted log(T) and observed log(T) (residual) are only 0.008175027 and 0.1108183, respectively. Figure 7.3 shows the survival curves constructed by predicted mortality time and observed mortality time.

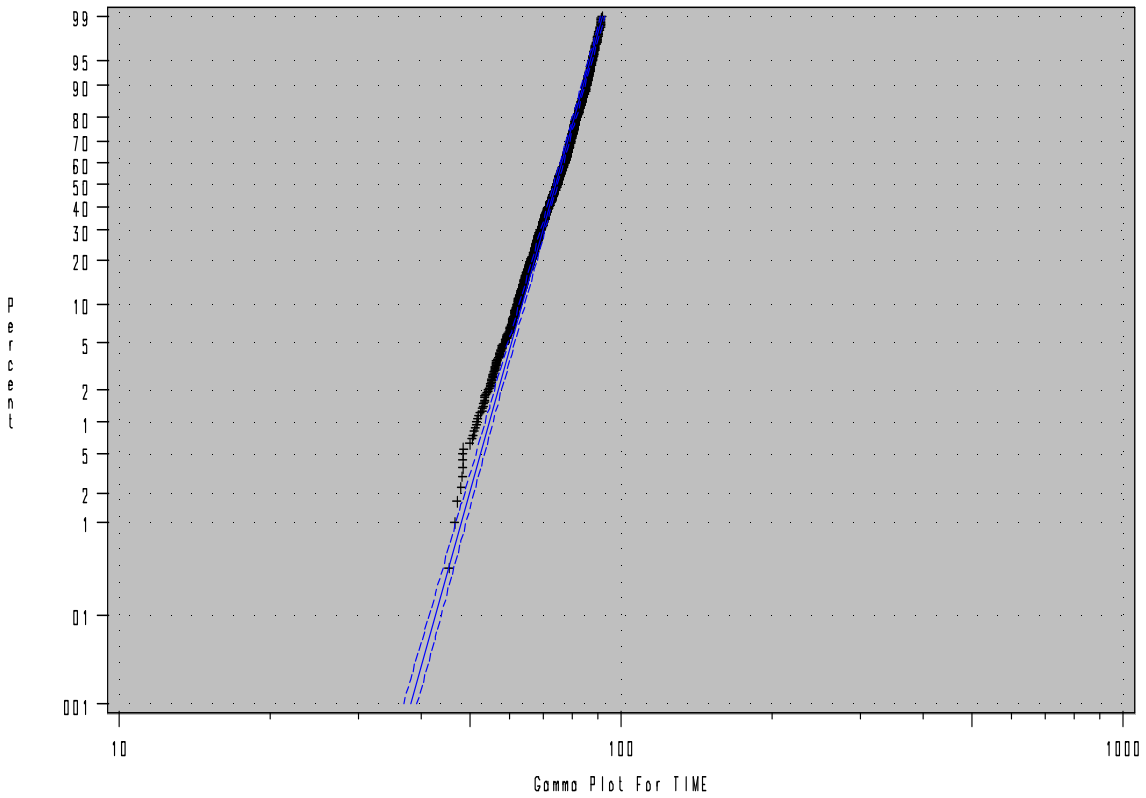

Figure 7.2 Percentage Plot of Female Ex-Smokers

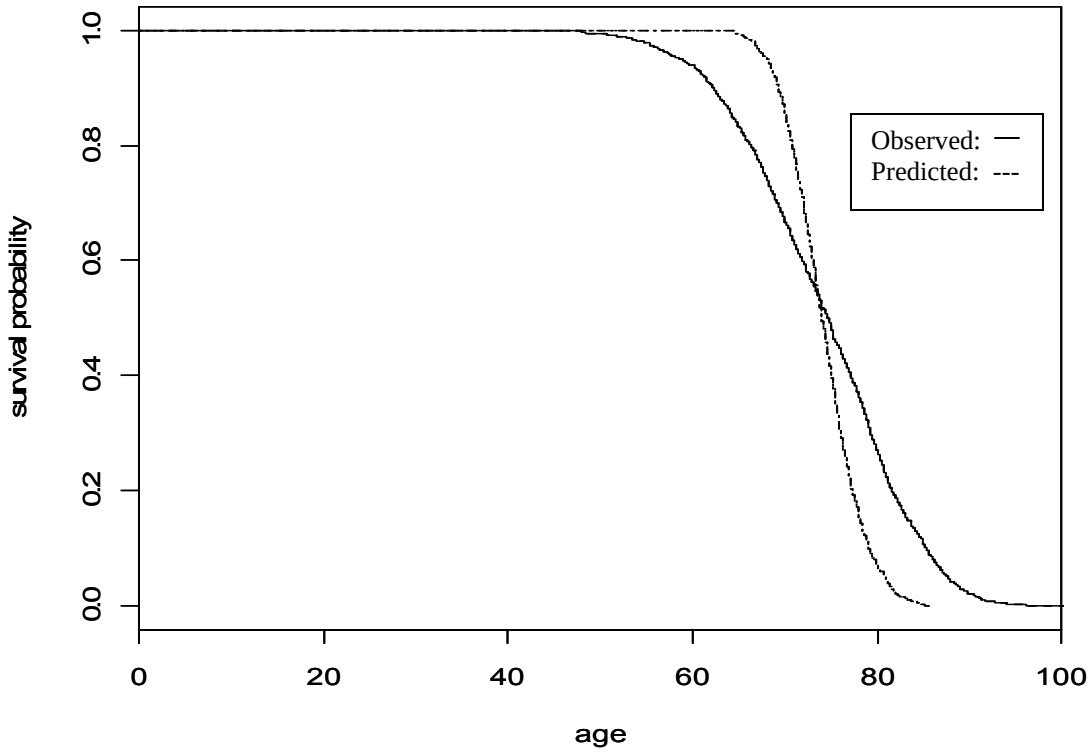

Figure 7.3 Predicted Survival Curve of Female Smokers

After elimination of the insignificant variables, all the variables left in the models are significant and the loglikelihood is not much changed. As can be observed from the Figure 7.2, all the data falls within the 95% confidence interval of the estimated percentage except in the left tail which suggests the model is fairly accurate.

Males

The same procedure is followed for male ex-smoker lung cancer patients. Survival regression models were run including exponential, gamma, loglogistic,

lognormal, logistic, normal, and Weibull. And their log likelihoods are as follows: -3814, 3209, 3128, 3160, -13152, -13093, 3121. Thus, the three parameter gamma is chosen to be the best fitting distribution and backward elimination is used to eliminate interactions between CPD and t_s, t_s and t_q, and the quadratic term of t_s. In the final model, the variables left are CPD, t_s, t_q, interaction between CPD and t_q, and quadratic terms of CPD and t_q.

All the terms are significant, and quadratic term of quitting time ranks first followed by quitting time, CPD, starting time, interaction between CPD and quitting time, and quadratic term of CPD comes last.

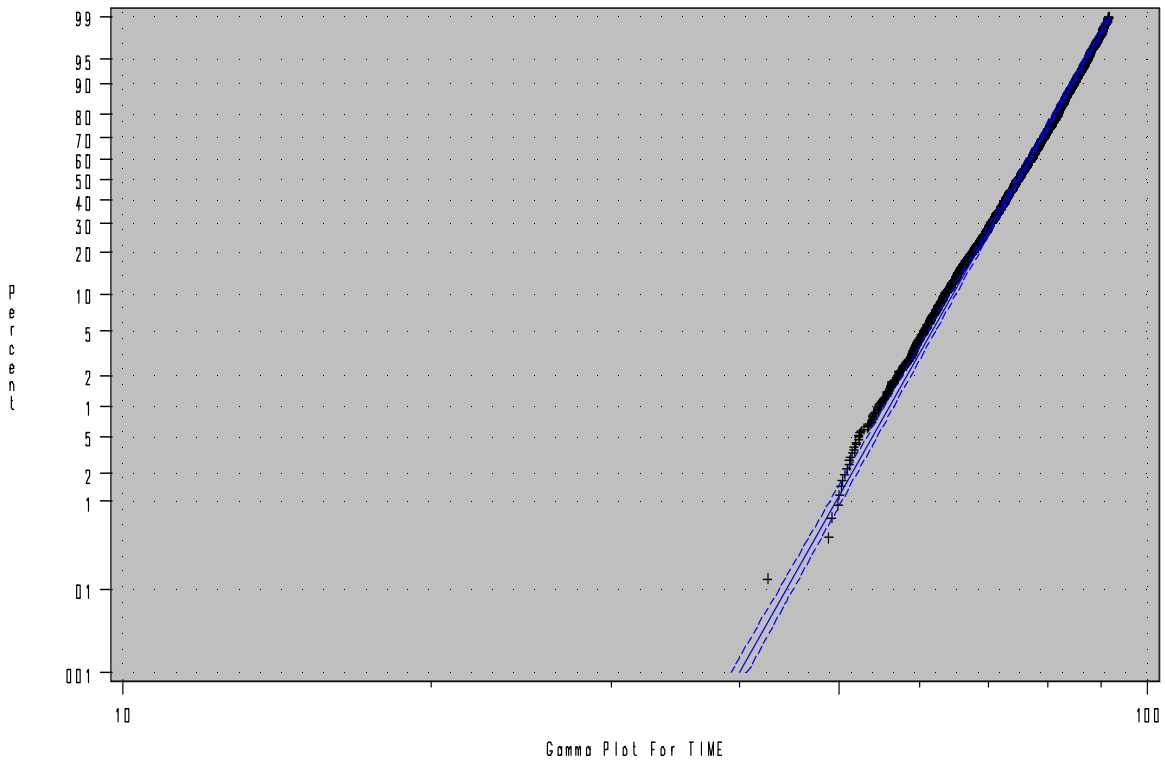

Figure 7.4 Percentage Plot of Male Ex-Smokers

Similarly, after plugging the estimated parameters into the model and obtaining the value of log(T) , mortality time T can be easily calculated by taking natural exponentials. The mean and standard deviation of the difference between predicted mortality time and observed mortality time are 0.1439845 and 7.651358, respectively. However, since the model is constructed using log(T) as the response variable, the mean and variance of the difference between predicted log(T) and observed log(T) are only 0.007539421 and 0.1054347, respectively, which indicates the predictive power of the model. Figure 5 below shows the survival curves constructed by predicted mortality time and observed mortality time.

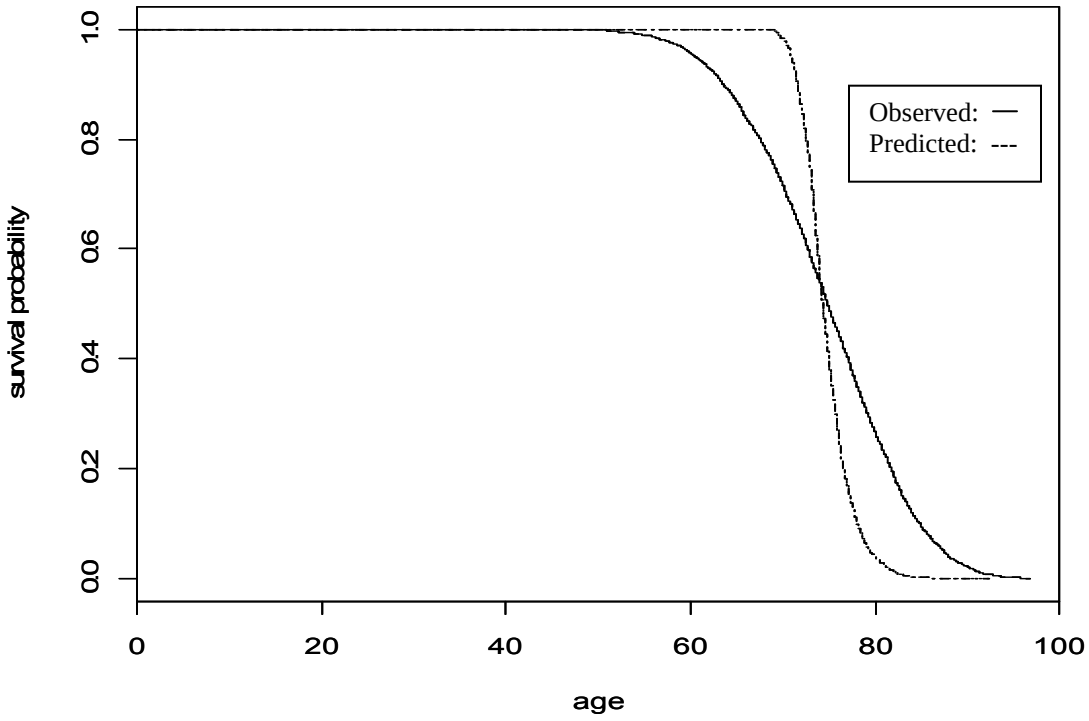

Figure 7.5 Predicted Survival Curve of Male Smokers

As can be observed from Figure 7.4, the percentage calculated from real data falls well within the 95% confidence interval constructed from the mode. This suggests the model is fairly accurate.

7.5 Discussion

The current chapter performs parametric and nonparametric analysis to address some very important questions concerning lung cancer utilizing real lung cancer data: What is the probabilistic nature of mortality time in ex-smoker lung cancer patients and non-smoker lung cancer patients, for female, male, and the totality of female and male patients? Is there significant difference of mortality time between ex-smoker and non-smoker patients? For ex-smokers, are there any differences with respect to the key variables such as mortality time, CPD, and duration of smoking between female and male patients? For non-smokers, can we notice a difference in mortality time between female and male patients? Can we accurately predict mortality time given information on CPD, starting time and quitting time for a specific lung cancer patient who smokes? Thus best fitting probability distributions are identified and their parameters are estimated. Mean mortality times are compared between non-smokers and ex-smokers, female non-smokers and male non-smokers, and female ex-smokers and male non-smokers. Important entities related to lung cancer mortality time, such as cigarettes per day (CPD), and duration of smoking (DUR), are compared between female and male ex-smoker lung cancer patients. Finally, a model is

developed to predict the mortality time of ex-smokers with a high degree of accuracy.

By using parametric analysis, distributions of mortality time for both female and male ex-smokers and non-smokers are found. Ninety percent and 95% confidence intervals are constructed which provide basic information on the probabilistic behavior of mortality time. Using nonparametric methods, we found that there is no significant difference in mean mortality time for females and males for both ex-smokers and non-smokers. Ex-smokers tend to have a shorter mortality time than non-smokers; mean CPD of males is larger than that of females; males have longer duration of smoking than females. Lastly, an accelerated failure time model is constructed for female and male lung cancer patients, respectively, so that given information on cigarettes per day, time started smoking, and time quit smoking of a specific smoker, mortality time can be predicted.

Chapter 8

Conclusion and Future Research

8.1 Conclusions

We applied various statistical approaches to modeling and predicting the survival time, relapse time of breast cancer patients, mortality time of lung cancer patients. We also utilized parametric and nonparametric comparisons including decision tree techniques to investigate the effectiveness of breast cancer treatments and showed the combined treatment of Tamoxifen and radiation is not always more effective than Tamoxifen only and different treatment should be given to patients with heterogeneous prognostics factors. Markov Chain also confirmed that different treatment should be given based on the stages of breast caner patients and the transition probability are calculated between stages of patients with different treatments.

We also used parametric analysis to show the sensitivity of breast tumor doubling time with different volume and growth assumptions. The results showed the probabilistic behavior of doubling time is very sensitive to the choice of volume and growth assumptions of tumors and lognormal probabilistic distribution is not the best choice to characterize tumor doubling time.

Nonparametric comparisons are conducted to analyze the mortality time between different genders, smoking status and mortality time is modeled for prediction purpose.

8.2 Future Research

Future validation of the models may be conducted using cross validation or by using new data. Other datasets such as SEER could provide more relevant information on breast cancer such as surgery, chemo theory, number of lymph nodes involved, ect, thus providing more compressive understanding of breast cancer. With the increased number of variables and need to identified attributable variables, considering the fact that accelerated failure model is extensively used in survival analysis and current statistical software does not provide variable selection in the accelerated failure time regression model, statistical package should be developed to satisfy such need. Currently a stepwise selection SAS macro based on p-value in accelerated failure model SAS procedure as shown in Figure 8.1 has been written and will be implemented the analysis in the future.

Stepwise Selection Algorithm:

1. Run univariate regression with each variable; select the one with the lowest p-value into the initial model

2. Run the existing model, if any term in the model has p-value > exit tolerance, remove the variable with largest p-value and go to 3.

3. Run the model with each remaining variable added to the existing model one at a time, if any term in the model has p-value < entrance tolerance, add the one with smallest p-value and repeat 2 and 3.

4. Stop until all the terms in the model have p-value < exit tolerance, and all the remaining variables have p-value > entrance tolerance if added to the existing model one at a time. That is, no variable can be eliminated from the existing model, and no variable needs to be added to the existing model.

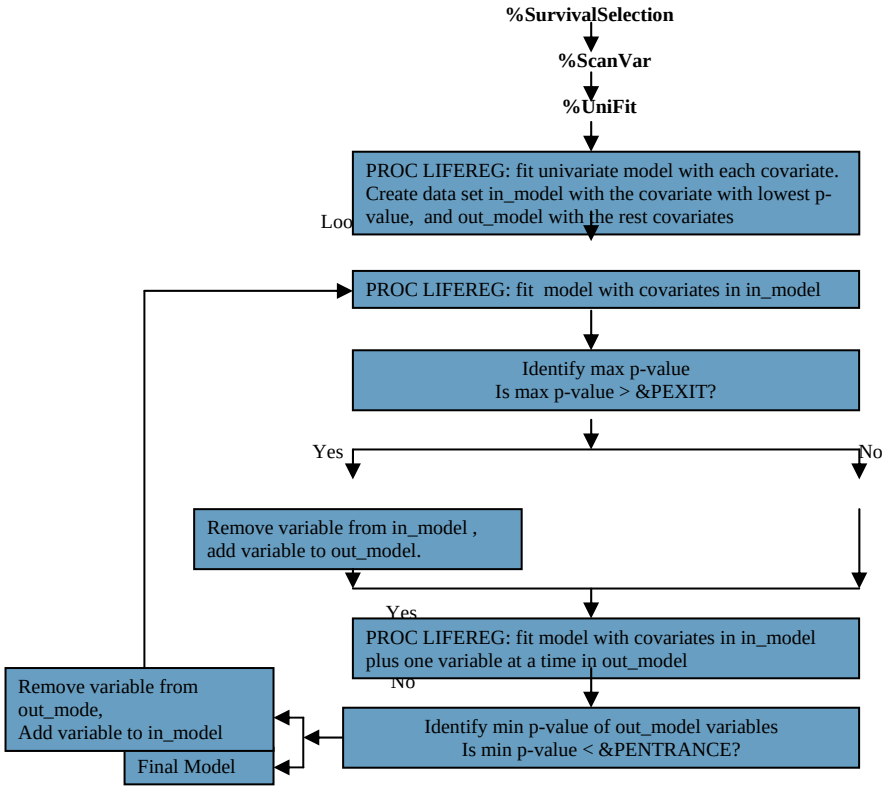

Figure 8.1 Stepwise Variable Selection Macro

Last but not least, the methods used in the current study could be implemented in the study of other types of cancer in providing important information on

treatment of cancer patients, transition of cancer stages and prediction of reoccurrence and survival time.

References

A. W. Fyles, D.R. McCready, et al. (2004) ”Tamoxifen with or without breast irradiation in women 50 years of age or older with early breast cancer”, New England Journal of Medicine 351, pp. 963-970.

A.A. Markov. (1971) "Extension of the limit theorems of probability theory to a sum of variables connected in a chain". reprinted in Appendix B of: R. Howard. Dynamic Probabilistic Systems, volume 1: Markov Chains. John Wiley and Sons.

Akaike H. (1974) A new look as the statistical model identification. IEEE Trans Automatic Control, 1974, 19: 716-23.

Anderson, T. W.; Darling, D. A. (1952) "Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes". *Annals of Mathematical Statistics,* , **23**: 193–212..

Bacchetti, P. and Segal M. R. (1995): Survival trees with time-dependentcovariates: application to estimating changes in the incubation period of AIDS Lifetime Data Analysis Vol. 1, number1.

Boag JW. (1949) Maximum likelihood estimates of the proportion of patients cured by cancer therapy. J Roy Stat Soc B; 11: 15-44.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984): Classification and Regression Trees, New York; Chapman and Hall

Breiman, Leo (2001): Random Forests, Machine learning, **45** (1): 5–32.

Bryant, A; Cerfolio RJ. (2007) "Differences in epidemiology, histology, and survival between cigarette smokers and never-smokers who develop non-small cell lung cancer". *Chest*, **132** (1): 198–192.

Chernoff, H.; Lehmann E.L. (1954) "The use of maximum likelihood estimates in χ^2 tests for goodness-of-fit". *The Annals of Mathematical Statistics*, **25**: 579–586.

Chin-Shang Li, Jeremy M.G. Taylor (2002) A semi-parametric accelerated failure time cure model. Statistics in Medicine; 21: 3235-3247.

Claude Shannon and Warren Weaver(1949) publication ”model of communication ” .

Davis, R. and Anderson, J. (1989): Exponential survival trees, Statistics in Medicine 8, pp 947-962.

F. Gao, A. K. Manatunga, and S. Chen (2004), “Identification of prognostic factors with multivariate survival data”, Computational Statistics and Data Analysis 45, pp. 813-824

Fabien Corbiere, Pierre Joly (2007). A SAS macro for parametric and semiparametric mixture models. Computer Methods and Programs in Biomedicine; 85(2): 173-80.

Farewell, V.T. (1982) The use of mixture models for the analysis of survival data with long-term survivors. Biometrics; 38: 1041-1046.

Gordon, L. and Olshen, R. A. (1985): Tree-structured survival analysis. Cancer Treatment Reports 69, 1065-1069.

Green L. (2009) “ Age Dependent Screening”, SIAM Conference Mathematics for Industry: Challenges and Frontiers, San Francisco, California.

Harrington, D. P. and Fleming, T. R. (1982): A class of rank test procedures for censored survival data. Biometrika 69, 553-566.

Heuser, L. et. al. (1979). “Growth Rates of Primary Breast Cancers”. *Cancer* **43**: 1888-1894.

Jones, D.R, Powles, R.L., Machin, D. and Sylvester, R.J.(1981) On estimating the proportion of cured patients in clinical studies. Biometrie-Praximetrie,; 21: 1-11.

Kaplan, E.L.; Meier, Paul. (1958): Nonparametric estimation from incomplete observations, J. Am. Stat. Assoc. 53, 457-481.

Lebalanc, M.; Crowlry, L. (1992): Relative risk trees for censored survival data, Biometrics. v48. 411-425.

LeBlanc, M., Crowley, J. (1993). Survival trees by goodness of split. Journal of the American Statistical Association 88, 457–467

Loh, W. Y. and Shih, Y. S. (1997): Split selection methods for classification trees. Statistica Sinica, Vol. 7, p. 815 - 840.

M.T. Uddin, M.N. Islam and Q.I.U. Ibrahim (2006), An analytical approach on cure rate estimation based on uncensored data. Journal of Applied Sciences; 6(3): 548-552.

MAGIDSON, J. (1993): The use of the new ordinal algorithm in CHAID to targetprofitable segments, The Journal of Database Marketing, 1, 29–48.

Mann, H. B., & Whitney, D. R. (1947) "On a test of whether one of two random variables is stochastically larger than the other". *Annals of Mathematical Statistics, 18*, 50–60.

N.A Ibrahim, et al. (2008): Decision tree for competing risks survival probability in breast cancer study, International Journal of Biomedical Sciences Volume 3 Number 1.

Orbe J, Ferreira E, Nunez-Anton V. (2002) Comparing proportional hazards and accelerated failure time models for survival analysis. Statist med; 21: 3493-510.

Peer P. G. M. et al. (1993). “Age Dependent Growth Rate of Primary Breast Cancer”. *Cancer*, **71**: 3547- 51.

Plackett, R.L.(1983) "Karl Pearson and the Chi-Squared Test". *International Statistical Review* , **51** (1): 59–72.

Quinlan, J. R. (1986): Induction of Decision Trees. Machine. Learning 1, 1, 81-106.

Segal M. R. (1988): Regression trees for censored data, Biometrics 44, pp.35-47.

Thun, MJ; Hannan LM, Adams-Campbell LL et al. (2008), "Lung Cancer Occurrence in Never-Smokers: An Analysis of 13 Cohorts and 22 Cancer Registry Studies". *PLoS Medicine*,**5** (9): e185.

Wilcoxon, F. (1945), "Individual comparisons by ranking methods". *Biometrics Bulletin,* 1, 80–83.

Yamaguchi K. (1992), Accelerated failure-time regression model with a regression model of surviving fraction: an application to the analysis of “permanent employment” in Japan. Journal of the American Statistical Association; 83:222-230.

Yingwei Peng, Keith B.G. Dear and J.W. Denham (1998), A generalized F mixture model for cure rate estimation. Statistics in Medicine; 17: 813-830.

About the Author

Chunling Cong got her undergraduate degree at Nanjing University of Technology in China and came to the United States for higher education at University of South Florida at fall, 2005. She got her master degree in Mathematics at the end of 2006 and continued pursuing her Ph. D degree with a concentration in Statistics. She got a summer internship at Statistical Evaluation Center in American Cancer Society in 2009. Beside Statistical Analysis, she developed great interest in actuarial science and worked in Travelers Companies, Inc as a summer intern which led to a permanent position as a senior consultant in Claim Research Department in 2010.