QATAR UNIVERSITY

COLLEGE OF ARTS AND SCIENCES

EXPONENTIAL MODEL FOR BREAST CANCER PARTLY INTERVAL CENSORED DATA VIA MULTIPLE IMPUTATION

BY

SALMAN UMER

A Thesis Submitted to

the College of Arts and Sciences

in Partial Fulfillment of the Requirements for the Degree of

Masters of Science in Applied Statistics

January 2021

COMMITTEE PAGE

The members of the Committee approve the Thesis of

Salman Umer defended on 29/11/2020.

_____

Dr. Faiz Ahmed Elfaki
Thesis/Dissertation Supervisor

_____

Dr. Reza Pakyari
Co-Supervisor

Approved:

_____

Ibrahim AlKaabi, Dean, College of Arts and Sciences

ABSTRACT

UMER, SALMAN, Masters: January : 2021, Applied Statistics

Title: Exponential model for breast cancer partly interval censored data via multiple imputation

Supervisor of Thesis: Dr Faiz Ahmed Mohamed Elfaki.

The estimation problem for interval-censored data has been investigated by several authors. The application of conventional methods to interval censored data that has been considered by Lindsey and Ryan (1998) showed misleading results when they tended to underestimate the standard errors of the estimated parameters.

In this thesis, we apply the likelihoods in the exponential model in order to estimate the parameters and function of survival when multiple imputation and left imputation methods are used for partly interval censored data. We pay particular attention to the performance of our model. In particular, we present the Likelihood Ratio Test (LRT) with their p-value.

We undertake a simulation study with different percentage of exact observations (0%, 25%, 50%, and 75%) in order to quantify and analyze the relative performances of maximum likelihood estimation for exponential model. The numerical evidence suggests that the estimates from multiple imputation are more accurate. We apply the proposed method to a real breast cancer data.

DEDICATION

I would like to dedicate this work to my family. To my parents and siblings, who have always encouraged me to pursue my goals. To my colleagues for their support throughout the course period.

ACKNOWLEDGMENTS

Firstly, I would like to thank my supervisor Dr. Faiz Elfaki for taking me under his supervision and for his continuous unwavering support and guidance throughout the research period. Furthermore, I would like to thank the committee member Dr. Reza Pakyari for his valuable comments and support during my research period.

I would like to express my gratitude towards the members of Mathematics, Statistics and Physics Department, Qatar University for their support and assistance during my studies.

I would like to express my heartfelt gratitude to Dr. Mufeedh of Al Amal Hospital for providing us with the data and supporting us from the beginning until the end of the thesis period.

Moreover, my family and friends for all their support and encouragement throughout this journey.

I would like to thank my colleagues Faraz, Raneem, Amal and Zainab for their support and encouragement. Special thanks to my managers at work Sheikha Amal Al-Thani and Mr. Martin Poveda, for their encouragement, continuous help and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

CHAPTER OVERVIEW

In this chapter we introduce, the survival analysis, censored data and exponential distribution model. The chapter also, discusses the background of the research, the problem statement, the objective and the scope of the research.

1.1 Background

In this modern era, data analysis is gaining high importance due to its application capabilities in all areas where data is involved. Analyzing data can provide different alternative results, useful information and precise conclusion which helps in making decisions. Researchers use statistical methods to deal with data. Survival analysis is one such method which is in high demand in the global market.

Analyzing time-to-event type of data is called survival analysis. The time to event type data, shows the span of the variable of interest from the origin time until the end point time. Due to its high focus on death and failures of components, survival analysis has huge importance (Singh and Totawattage, 2013).

Survival analysis has many applications, mainly including but not limited to; engineering, medical and education. Analyzing the application of survival analysis in education field, Singer and Willett (1993) released the paper on duration and the timing of events based on discrete-time survival analysis. The paper finds that the discrete-time survival analysis provides a framework which helps in analyzing a type of event occurring data that is used in educational field. Furthermore, the discrete-time approach helps in examining the hazard function shape, where in the shape is ignored due to the shift in parameters associated with the covariates.

Researchers use survival analysis on the school data for retrieving valuable information regarding the student characteristics. Plank et al. (2008) studies on the surviving of students in high school. The objective of the paper was to combine core academic courses and Career & Technical Education (CTE) that influence the likelihood of leaving school. The study uses the method of cox regression model to use the hazard model for student dropouts. The hazard model indicates the significant curvilinear association between the CTE and the risk of reducing the youths in the school when they were 14.

Currently, survival study is a wide branch which includes studies of retrospective correlative, prospective cohort, and retrospective cohort along with clinical trials. The analysis is connected with censoring on time to event modeling data. In the study of survival analysis, there are subjects in the study who cannot complete the survival time due to censoring. The below section provides detailed view on censoring.

1.2 Censored data

Taking into consideration the fact that the survival data are not normally distributed, they usually contain incomplete information, called censored subjects. Censoring of the subjects can be on left or right. It is necessary to include censored subjects in the statistical analysis.

Censoring occurs when there is prevalence of incomplete information about the survival time of some individuals in the study. It is an important factor in the study of survival analysis. At the end of the observation period, we know that not all patients for example experience the event (death) in this case the actual survival times are unknown for some patients. General reasons for censoring include: the subject was lost in the middle of

the study period or the subject dies or the subject didn't experience the event before the end of the study. Furthermore, patients discontinue from the clinical trial because of the side effects the treatment causes on them or a patient is removed from the study if there is evidence of the treatment having no effect on them (Williams & Lagakos, 1977).

The most common type of data censoring is the right-censoring. Here, the event time is not observed when it is larger than the right-censoring time. In clinical trials, almost always, an analysis needs to be done before everyone is dead, otherwise some people might be lost to follow-up. Also, the subject is right censored when the subject leaves the study before the event occurs (Kleinbaum and Klein, 2005). Right censoring is commonly used in many of the studies. The study has a left censored subject when the subject has failed before the study. The subject is interval censored when the event of interest occurs within an interval of time. It is common in clinical trials, where the event of interest occurs in between visits. Interval-censored data results from incomplete observations of random variable, when the values cannot be observed exactly but instead lie in between intervals. When a study consists of exact data and interval censored data, then we have Partly Interval Censored (PIC) (Kim, 2003).

In this study, exponential distribution model will be used based on the partly interval censored data. In the next section, exponential distribution model will be introduced.

1.3 Exponential distribution

In survival analysis, exponential distribution is one of the most simple, convenient and useful distribution for analyzing and modeling. P.V. Sukhatme in his paper in 1937 mentioned exponential distribution can be used as an alternative for normal distribution

where the form of variation is removed. Furthermore, he established that the exponential spacings are exponentially distributed. Davis (1952) used exponential distribution to discuss the analysis of failure data and compared the analysis with the normal distribution.

Teisser (1934), Weibull (1939) and Steffensen (1930) demonstrated applications of the exponential distribution on biological, engineering and actuarial problems. D.J. Davis (1952) used exponential distribution to discuss the analysis of failure data and compared the analysis with the normal distribution.

The probability density function (pdf) of the exponential distribution is given as:

$$f(t) = \lambda e^{-\lambda t}$$

where $t > 0$, $\lambda > 0$, λ is the parameter that is widely used in applied statistics, and the pdf of the exponential distribution is shown in Figure 1.1.

The exponential distribution in survival analysis can be characterized by its survival function survival $S(t)$ and the hazard function $\lambda(t)$. The function of survival is the probability that an individual survives at least time t so that;

$$S(t) = 1 - F(t) = 1 - \int_0^t f(u)du = e^{-\lambda t}$$

where $F(t)$ is cumulative distribution function, and the hazard function is given as;

$$\lambda(t) = \frac{f(t)}{S(t)} = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda$$

In the next chapters, we will derive the model under survivorship and the curve of the survival probability function will be estimated using censoring data based on exponential distribution model.

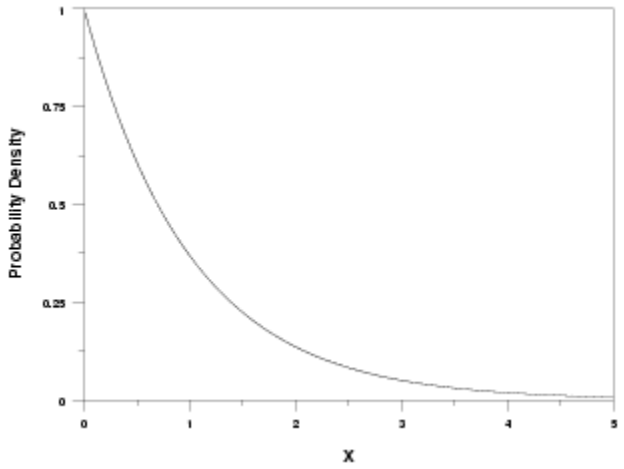

Figure 1.1: Pdf function of exponential distribution

1.4 Problem statement

When it comes to statistical field, in contrast to other topics, there are limited number of researchers who have worked on PIC. For example, Huang (1999) discussed asymptotic properties of the nonparametric MLE of a distribution function based on PIC data. Peto and Peto, (1972) published his work on PIC data, where an exact observation was seen as an interval-censored observation with a short interval. Zhao. et al., (2008)

discussed PIC failure time data based on log-rank test. Kim (2003) explained the Cox model for PIC data.

However, there are very few studies that focus on the partly interval censored data using exponential distribution and even fewer being applied to breast cancer related application. Alharpy and Ibrahim (2014) discusses the comparison problem in the presence of partly interval-censored (PIC) failure time data for two reliability functions. The study furthermore aims to use the parametric test under non-proportional hazard in the presence of PIC data using multiple imputation technique.

In this research, the focus will be given on partly interval censored data for survival analysis and a model will be applied which will be used for the breast cancer data. Moreover, the procedure will be simplified by using exponential distribution based on multiple imputation techniques.

1.5 Objective

The study focuses to predict the survival ability of the breast cancer patients in the hospital for partly interval censored data based on exponential model with different imputation techniques. Furthermore, MLE method will be used to estimate the parameters in the study. The main objectives in the study are:

- To modify breast cancer PIC data based on multiple imputation techniques via exponential model.
- To compare the parameter estimate obtained from our model with existing model.
- Evaluate the performance and operating characteristics of the parameter estimating methods through conducting extensive simulation studies.

1.6 Scope of the research

The research study is focused on breast cancer partly interval censored data based on the exponential model to predict the survival-ability of the breast cancer patients in the Amal Hospital. This model is described in chapter 3 and the MLE is going to be used to estimate the parameters in the model. Multiple imputation techniques will be used to modify the real data set into PIC data.

Chapter 2 includes literature review of survival analysis, partly interval censoring and exponential distribution model. Chapter 3 includes exponential distribution model based on survival analysis and derivation of maximum likelihood estimator for parameters. Chapter 3 includes likelihood ratio test as well. Towards the end of chapter 3, breast cancer data will be explained along with the use of multiple imputation techniques to modify the data.

Chapter 4 shows exponential model that is suitable for the modified data sets after implementing the process of multiple imputation techniques. Later, the chapter shows the breast cancer data, real data set and simulated data. To conclude, chapter 5 provides an overall idea of the entire study and provides few suggestions for future works on this topic for the researchers.

CHAPTER 2: LITERATURE REVIEW

CHAPTER OVERVIEW

In this chapter, we discuss and review some of the existing literature related with survival analysis which has major applications in areas like medicine, education and engineering. Furthermore, focus will be given on partly interval censoring in this paper. Finally, some additional existing literature related to exponential distribution model will be provided.

2.1 Survival analysis

Kartsonaki, (2016) describes survival analysis as the time-to-event analysis of data, which describes the length of time between the origin and endpoint of the event. The general term used is survival time, even though it will be applied to the time 'survived' from complete remission to relapse or from diagnosis to death (Clark et. al., 2003). It is described as a significant advancement in the field of mathematical statistics towards the end period of 20^{th} century (Ma and Krings, 2008).

Survival analysis is used in many areas: education, engineering, medical and other sectors. This method has huge importance in medical studies. It has such a huge importance in the statistical study because it is connected with failure and death of objects Singh and Totawattage (2013). Brenner and Hakulinen (2002) mentions using the Finnish Cancer Registry for providing an empirical assessment of survival analysis. They use the technique to compare the 10-year actual relative survival curves of patients diagnosed with one of the 15 most common forms of cancer in 1983 to 1987 against the most up-to-date relative survival curves that might have been obtained using traditional or period analysis. Likewise, Weissberg et al. (1984) in his study mentions an example of medical application

that deals with the survival analysis theory in determining life expectancy of 379 patients with chronic hepatitis B from the time of first contact. The estimated 5-year survival study showed that the usual cause of death was liver failure and its sequel.

There has been studies that have made use of the survival analysis techniques from the education sector. Kaminski & Geisler (2012) used survival analysis to track a total of 2966 individual assistant professors hired in science and engineering from 1990 at 14 United States universities from time of hire to time of their departure using publicly available bulletins and catalogs. Similarly, the study by Willet and Singer (1991), investigates student retention in undergraduate engineering program using large longitudinal database of 100,179 engineering students from nine universities which spans for around 19 years.

Giolo (2004) used interval censored data for discussing the nonparametric method for estimation of the survival function. He used Turnbull's algorithm to estimate the survival function. He observed that analysis based on this type of censored data which applies this method to standard time. Due to this, Giolo asks analysts to be careful when they use new methods for analyzing interval censored data.

Plank et al. (2008) studies on the surviving of students in high school. The study was focused to combine Career and Technical Education (CTE) along with core academic courses that plays a role in the likelihood of leaving school. In the study, they used cox regression model, which is the most common method of estimation for using hazard model for dropout of youths. The study observed that the hazard model indicates significant curvilinear association between the CTE to academic courses.

Bhandari & Boutros (2016) in their paper discusses that treatment of cancer is

converging more towards personalized and that biomarkers are used to refine treatment decisions. Tumor mRNA data abundance data for example is commonly used as a biomarker to predict survival of a patient. Survival analysis presents the challenge of not knowing whether mRNA abundance information gives different results if it is analyzed in a continuous or discrete manner. Thus, they analysed 1988 primary breast tumor transcriptome. The result showed an average of 60% of all genes were showing difference between the discrete and continuous cox proportional hazards model. Furthermore, hybrid models outperformed models using a single type of information.

In public health research or biomedical, it is common to collect both longitudinal and survival time for a subject, along with the subject's risk factors and characteristics. Investigators are interested in finding variables that are important for predicting survival time which correlates with the same subject. Choi et al. (2015) simultaneously model the survival time with a stratified cox proportional hazards model and the longitudinal categorical outcomes with a generalized linear mixed model. The study uses data from the Carolina Head and Neck Cancer Study (CHANCE) to compare the results based on the simultaneous analysis and the analysis conducted using the generalized linear mixed model with cox proportional hazards model.

Chen et al. (2019) discusses the exposure of statin on improving the survival rate in several cancers. The study focuses on evaluating the association between statins and patients with lung cancer patients. The study collected thirteen studies with data from a sample of 99,297 meeting the inclusion criteria. The study showed that statin exposure was associated with improved overall survival, cancer-specific survival and recurrence-free survival. Meanwhile, subgroup analysis shows that users of statin after being diagnosed

with lung cancer had more survival benefit for overall survival than those before diagnosis. Hence the study finds strong association between statin exposure and improved survival of patients with lung cancer.

Most of the work done in survival analysis are based on Cox model (Cox, 1972) which is commonly used in medical research for investigating the association between predictor variables and survival time of the patients. Cox (1972) proposed a proportional hazard model for the analysis of censored survival data that allows the inclusion of covariates. The hazard rate is;

$$h(t, z) = h_0(t) \exp(\beta' z) \qquad (1)$$

where $h_0(t)$ is the underlying hazard function, z represent covariates and β is the unknown regression coefficient. Cox (1972, 1975) obtained estimates of β and asymptotic covariance matrix using a partial likelihood argument. Breslow (1974) proposed an estimate for the underlying hazards rate assuming that the hazard rate was constant between death times.

The Cox model focuses on simultaneously exploring the effects of different variables on the survival. Additionally, it is based on modelling approach to the analysis of survival data. The Cox model is a famous statistical technique used in the studies and process of analyzing survival data. The model allows its users to isolate the treatment effects from the other variable's effect. The model could improve the effect of treatment estimate by narrowing the confidence interval.

Even though, the above reviewed studies have topics similar to my research study due to its focus on survival analysis. There are huge differences mainly on the methods and

type of censoring used. In this research, we will focus on Exponential model based on partly interval censoring.

2.2 Partly interval censoring

When a study consists of exact data and interval censored data, then we have Partly Interval Censored (PIC) (Kim, 2003). PIC data can be found in different fields such as medical studies and reliability studies (Odell et al., (1992); and Lu and Meeker, (1993); Elfaki et al. (2012); Alharpy and Ibrahim (2014); Zyoud et al. (2016); Yousif et al. (2016); Saeed (2018) and Saeed and Elfaki (2020).

Kim (2003) in his study researched on partly interval censored data for Cox model using the method MLE. The study uses generalized profile information procedure and missing information principle to estimate variance-covariance matrix. The simulation studies states variance for bias and samples of size, both the methods work. He illustrated his method using an application in Denmark on a diabetes data. Zhao et al. (2008) used generalized log-rank test to study partly interval censored data as discussed by Peto and Peto (1972). The researchers used a set of real diabetes and simulation study data to evaluate their method.

Gao et al. (2017) explains efficient semi parametric estimation using partly interval censored data on the Accelerated Failure time model. The study generalizes the Buckley-James estimator for right-censored data to partly interval censored data. Then, a score for regression parameters was estimated using a one- step estimator. Furthermore, extensive simulation studies are conducted to analyze the performance of the estimators and to apply the methods to an AIDS study data.

Zyoud et al. (2016) discusses the partly interval-censored data to estimate the survival function using nonparametric analysis. Methods like right-point, left- point and mid-point are included under simple imputation. While, mean imputation, conditional, median, random and multiple imputation are included under probability-based imputation methods. R program was used for estimating the survival function. They concluded that mean, random and median imputation techniques were better than other imputation methods available.

Elfaki et al. (2013) discusses estimation of parameters using semi parametric Cox's proportional hazards regression for partly interval censored based on EM algorithm. Weighting Technique (WT) model and Censoring Complete (CC) model were two competing risks models used in the study. They studied the effect of covariates on the development of complications to investigate the association between the anti D in Rhesus time and treatment which is being applied to a time data set which arises from anti D in Rhesus D negative women in Sudan who are pregnant. The study concluded that there are no significant differences for the covariates since the negative group was showing a higher risk with the onset of anti D rhesus.

Wu et al. (2019) analyzes partly interval censored data using cox regression using semi parametric sieve MLE method. To analyze the data, they use the semi parametric spline-based sieve MLE approach and included the non-mixture cox regression. Furthermore, for both nonparametric and parametric parts of the sieve estimator they have used the theory of modern empirical process. Confirming that the sieve estimator is consistent they simulated the data and concluded that the sieve MLE is acceptable. At the end, the proposed method is applied on spontaneous abortion study.

Yousif et al. (2016) uses Bayesian method for partly interval-censored data to estimate the regression coefficients using Cox model. The study verified the model using the simulated data. Finally, they showed that the developed model is well applicable and performs well with the use of simulated data.

2.3 Exponential distribution model

Exponential distribution is seen as a life-time distribution with constant hazard rate. In general, exponential distributions are commonly used in stochastic processes and formation of models of lifetime distributions. Sukhatme (1937) indicated that the exponential distribution could be used in cases where form of variation is removed from the normal, which will be a suitable alternative to the normal distribution.

Davis (1952) used the exponential distribution to discuss the analysis of failure data and compared the analysis with the normal distribution analysis.

Friedman (1982) mentioned that there is a similarity between the likelihood function for the piecewise exponential model and the likelihood function for a log-linear model for frequency data. Prentice (1973) used exponential model for survival data. While, Holford (1980) along with Laird and Oliver (1981), used Iterative Proportional Fitting to obtain the MLE of the piecewise exponential model, so that the techniques used to analyze the frequency data can be used for survival analysis.

Staplin et al. (2015) discusses sensitivity analysis method on piecewise exponential survival model. This method helps in analyzing the sensitivity of the results of standard survival models to minute amounts of dependence between time to censoring and time to failure variables. The paper uses more flexible models for the marginal distributions by having simple computation methods. The study finds that the sensitivity analysis acts good

in all the situations, apart from the condition when the data have a large proportion of censoring.

Jung et al. (2018) states that a cancer clinical trial is designed usually to assume an exponential distribution for a time-to-event outcome such as overall survival. The study assumes overall survival and survival post-progression to be exponentially distributed. The study focuses on deriving a sample size calculation formula for comparing overall survival between two treatment arms using log-rank test following a gamma or hypo-exponential distribution.

Mazucheli et al. (2013) discusses estimates for the parameters which are seen in non-mixture lifetime and long-term mixture models which are applied to analyze the survival data. The study considers the case in which the lifetime data have a two-parameter exponential distribution. Bayesian and classical procedures are used to avail confidence intervals of the parameters. Furthermore, the general survival model where

Lee et al. (2007) states to use a bivariate exponential model with Bayesian analysis of paired survival data. But, the bivariate exponential model has limitation of having lack of a closed form likelihood function. They introduced a latent variable that removes the difficulty in the Bayesian computation. Furthermore, predictive Bayesian p-value is used for model checking procedure.

Chakhoyan et al. (2018) discusses on quantifying prognostic values and changes of MRI measurements obtained using stretched exponential models, mono-exponential and diffusion kurtosis imaging models before and after chemo radiation. The study concludes that advanced diffusion models take more acquisition time even though there is increased tissue complexity after chemo radiation.

Jeong et al. (2017) discusses on differentiating and diagnosing the pathological grades and subtypes of uterine cervical carcinoma using the metrics derived from stretched exponential model (SEM), bi-exponential model (BEM) and mono-exponential model (MEM). The study concluded that diffusion parameters from stretched exponential model and bi-exponential model offers better information in cervical carcinoma diagnosis along with predicting subtypes of tumor and grades.

CHAPTER 3: METHODOLOGY

CHAPTER OVERVIEW

This chapter presents the estimation of the parameters of the exponential distribution with the help of maximum likelihood estimator under general and censored data. The chapter also discusses the likelihood ratio test and the real data set is described which is used in this study and the process to treat the data with survival time. At the end of the chapter, multiple imputation technique will be presented.

The models in survival analysis are extended to include regression variables, called factors or exploratory variables (covariates), measured on the individuals in the study. These types of models are used to investigate the rule of these exploratory variables in modifying the response and to take into account the confounding factors in the estimate or exposure effect or treatment. One of these models are exponential distribution in survival analysis.

3.1 Exponential model

Exponential distribution is often used to model survival data because of its potential to be a simple distribution to characterize the data. Even though, as mentioned previously, exponential distribution is a simple method used in analysis of survival data, it has the capacity to be used as the underlying process that leads to censored values. Moreover, the hazard function is constant with respect to the time. The distribution has importance in reliability and survival analysis applications. The distribution is utilized for modeling the behavior of units with failure rate that is constant. With the recent developments in scientific world, exponential distribution has a huge important role in the survival analysis study.

Suppose that failure times as defined early as T from exponential distribution and the failure rate depend on the k exploratory variables $z_i = (z_{1i}, z_{2i}, ..., z_{ki})$ on the ith individual which is characterized by the set of data (t_i, δ_i, z_i) as in (1). For given T and model $E(T)$ in terms of z_i, it implies modeling λ given that $E(T) = 1/\lambda$, where λ is the hazard rate for $\lambda > 0$. Marubini and Valsecchi (1995) suggested the exponential model as;

$$\lambda_i^{-1} = \beta_0 + \beta' z_i = \mu_i \qquad (2)$$

where β_0 indicates the intercept and also represents the regression coefficient of z_{0i} taken to be identically 1 for all individual. Then the linear predictor which includes $k+1$ regression is given as;

$$\mu_i = \beta_0 z_{0i} + \beta_1 z_{1i} + ... + \beta_k z_{ki} \qquad (3)$$

In the next section, the exponential distribution will be estimated by using maximum likelihood estimator which is a common method used for estimation.

3.2 Maximum likelihood estimators

Maximum likelihood estimation is a famous technique for deriving estimators with wide range of applications. Maximum likelihood, also known as the maximum likelihood method, is the method of finding the value of one parameter or more for a given statistic which makes the known likelihood distribution become maximum. The advantages for this method are plenty. One of the advantages is that it presents a consistent approach towards parameter estimation problems. This shows that the maximum likelihood estimates can be developed for a large variety of estimation situations. Other benefits include the maximum likelihood estimators having desirable mathematical and optimality properties, specifically in cases where the sample size increases and the minimum variance will be unbiased estimator.

In order to write the likelihood function in the presence of censoring it is convenient to use the set of data (t_i, δ_i, z_i) that is presented in (1) to identify the outcome of each individual. The likelihood of an individual who fail at $t_i(\delta_i = 1)$ is the probability without survival function. However, for an individual whose survival time is censored at $t_i(\delta_i = 0)$ then the contribution to the likelihood is given by the probability of survival function $S(t_i)$. The general likelihood function in the presence of censoring is;

$$L(\lambda;t) = \prod_{i=1}^{n} [f(t_i)]^{\delta_i} [S(t_i)]^{1-\delta_i} \qquad (4)$$

where $f(t_i) = \lambda \exp(-\lambda t_i)$ if $\delta_i = 1$ is probability density function, $S(t_i) = \exp(-\lambda t_i)$ if $\delta_i = 0$ is survival function and λ is hazard rate.

The likelihood function equation (4) became;

$$L(\lambda;t) = \prod_{i=1}^{n} [\lambda \exp(-\lambda t_i)]^{\delta_i} [\exp(-\lambda t_i)]^{1-\delta_i} \qquad (5)$$

The log-likelihood is given as;

$$\log[L(\lambda;t)] = \sum_{i=1}^{k} [\delta_i (\log \lambda - \lambda t_i) + (1-\delta_i)(-\lambda t_i)] \qquad (6)$$

The log-likelihood (6) is rewritten accounting for the fact that each individual i is related to a different λ_i . By substituting (2) into (6) for λ , we obtain the log-likelihood of the sample as;

$$\log[L(\beta_0, \beta; t, z)] = \sum_{i=1}^{n} [\delta_i (\log \mu_i^{-1} - \mu_i^{-1} t_i) + (1-\delta_i)(-\mu_i^{-1} t_i)] \qquad (7)$$

Maximizing equation (7) to estimate the k+1 parameters of β_0 and β . However, Glasser (1967) defined λ_i as;

$$\lambda_i = \lambda_0 \exp(\beta' z_i) = \exp(\mu_i) \qquad (8)$$

To fit the model (8) to our data with one independent variable z having values 0 and 1; where 0 denotes the 'with treatment' and 1 represents 'without treatment'. Using equation (5), the log-likelihood of the sample is;

$$\log[L(\beta_0,\beta)] = \sum_{i=1}^{n}[\delta_i(\beta_0+\beta' z_i) - \exp(\beta_0+\beta' z_i)t_i] \tag{9}$$

The first derivatives with respect to β_0 and β is given as;

$$\begin{aligned}\frac{\partial}{\partial\beta_0}\log[L(\beta_0,\beta)] &= \sum_{i=1}^{n}[\delta_i - \exp(\beta_0+\beta' z_i)t_i] \\ &= \sum_{i=1}^{n_1}[\delta_i - \exp(\beta_0)t_i] + \sum_{i=n_1+1}^{n}[\delta_i - \exp(\beta_0+\beta' z_i)t_i] \\ &= d_1 + d_2 - \exp(\beta_0)\sum_{i=1}^{n_1}[t_i - \exp(\beta_0+\beta')\sum_{i=n_1+1}^{n} t_i] \end{aligned} \tag{10}$$

$$\begin{aligned}\frac{\partial}{\partial\beta_1}\log[L(\beta_0,\beta)] &= \sum_{i=1}^{n}[\delta_i z_i - \exp(\beta_0+\beta' z_i)t_i z_i] \\ &= \sum_{i=n_1+1}^{n}[\delta_i - \exp(\beta_0+\beta_1)t_i] = d_2 - \exp(\beta_0+\beta')\sum_{i=n_1+1}^{n} t_i \end{aligned} \tag{11}$$

where d_1 and d_2 are observed deaths due to breast cancer, additionally n_1 and $n-n_1$ indicate the number of patients under treatment and without treatment respectively.
Newton Rapson will be used to solve the above system of equations (10) and (11) when it is equal to zero for β_0 and β respectively.

For the case of Partly Interval Censored (PIC), we assume that the exact failure time for n_1 participants that will be observed, and the failure time for interval censored will be the remaining subject that is; $n-n_1=n_2$. As mentioned earlier in this thesis by exact failure times we mean any patient that has the event of interest during the inspection times or the hospital examination of the patient's condition. In this case the likelihood is given as;

$$L(\lambda;t)=\prod_{i=1}^{n_1}\{f(t)\}^{\delta_i}\prod_{i=n_1+1}^{n}\{S(t)\}^{1-\delta_i}$$

$$=\prod_{i=1}^{n_1}[\lambda\exp(-\lambda t_i)]^{\delta_i}\prod_{i=n_1+1}^{n}[\exp(-\lambda t_i)]^{1-\delta_i} \qquad (12)$$

The log-likelihood is given as;

$$\log[L(\lambda;t)]=\sum_{i=1}^{n_1}[\delta_i(\log\lambda-\lambda t_i)+(1-\delta_i)\sum_{i=n_1+1}^{n}[(-\lambda t_i)] \qquad (13)$$

Following equations (7) and (8), equation (13) becomes;

$$\log[L(\beta_0,\beta')]=\sum_{i=1}^{n_1}[\delta_i(\beta_0+\beta' z_i)-\exp(\beta_0+\beta' z_i)t_i)+(1-\delta_i)\sum_{i=n_1+1}^{n}[-(\beta_0+\beta' z_i)t_i] \qquad (14)$$

The normal equations for deriving the MLEs of equation (14) become;

$$\frac{\partial}{\partial\beta_0}\log[L(\beta_0,\beta')]=\sum_{i=1}^{n_1}[\delta_i-\exp(\beta_0+\beta' z_i)t_i]+(1-\delta_i)\sum_{i=n_1+1}^{n}[-t_i]$$

$$=d_3+d_4-\exp(\beta_0)\sum_{i=1}^{n_1}[t_i]-\exp(\beta_0)\sum_{i=n_1+1}^{n}[t_i]-\exp(\beta')\sum_{i=n_1+1}^{n}[z_i t_i] \qquad (15)$$

$$\frac{\partial}{\partial\beta_1}\log[L(\beta_0,\beta')]=\sum_{i=1}^{n_1}[\delta_i z_i-\exp(\beta_0+\beta' z_i)t_i z_i]+(1-\delta_i)\sum_{i=n_1+1}^{n}[z_i t_i]$$

$$=d_1+d_2-\exp(\beta_0+\beta')\sum_{i=1}^{n_1}t_i+\sum_{i=n_1+1}^{n}t_i \qquad (16)$$

where d_1 and d_2 are described in equation (11).

After setting the above equations equal to zero to maximize the function β_0 and β, later we can solve the system by Newton Rapson method.

The Likelihood Ratio Test (LRT) will be presented in the next section to investigate the performance of exponential model.

3.3 Likelihood ratio test

Likelihood Ratio Test (LRT) is a hypothesis test that helps to choose the best model from two nested models. In this research project, the likelihood ratio test will be used to perform tests of hypotheses about parameters that have been estimated by MLE in two situations. One of the test statistics is for testing whether all parameters in the distribution are equal to certain values and the other test statistics is for testing whether some of the parameters in the distribution are equal to certain values. To test a subset of parameter in a distribution, let $\beta^* = (\beta_1, \beta_2, ..., \beta_k)$ denote all the parameters in a parametric distribution. Then the hypothesis will be

$$H_0 : \beta^* = 0 \qquad (17)$$

Let $\hat{\beta}$ be MLE of $\beta^* = (\beta_1, \beta_2, ..., \beta_k)$ the MLE of β_1 given $\beta_1 = \beta_0$. Under H_0 the statistic test has chi-square distribution with degrees of freedom equal to the dimension of β_2 or the number of parameters in β_2. Then the likelihood ratio test statistic is given as;

$$X_L = 2\left[l(\hat{\beta}) - l(\hat{\beta}_1(\beta_0), \beta_0)\right] \qquad (18)$$

If the number of parameters in β_1 is equal to q, for a given significant level α. Then H_0 is rejected if $X_L > X^2_{q,\alpha}$ (Lee and Wang, 2003).

3.4 Multiple Imputation

Imputation techniques are often used to transform the problem of analyzing the data. In this study, the data will be modified based on the imputation technique to right censored, interval censored and partly interval censored data. The imputation process is simple with different methods to be used on the data. Two different types of imputation techniques are simple imputation and multiple imputation. In this study, multiple

imputation technique will be used.

Multiple imputation technique was being used from the early 1970s, and since then gained popularity through the decades. The technique is simulation-based for managing missing data. It is important that there is enough compatibility between the imputation model and the analysis model or the imputation model has more generality than the analysis model. Analyzing a multi-imputed data set is seen as simulating under imputation models, the predictive distributions of desired summary statistics. There are three tasks that are needed to create the imputations: imputation task, estimation task and the modeling task. The estimation task and imputation task are technical in nature. The modeling task would require an additional development of tools that are suitable for relating respondents and non-respondents.

Jakobsen et al. (2017) mentions the different types of multiple imputation methods. The following are presented according to their increasing degrees of complexity:

1. Single value regression analysis
2. Monotonic imputation
3. Chained equations or the Markov chain Monte Carlo (MCMC) method

The datasets can be analysed by possibly any type of method that would be appropriate once the MI is created and if the data were complete. A possible example would be performing logistic or linear regression procedures using any possible standard statistical package available. From the result, for each imputed dataset any model will have to fit m times, and the results will vary across these available datasets based on the missing-data uncertainty.

To obtain a set of estimated standard errors and coefficients, the estimates and

standard errors has to be stored from each of the m imputed datasets, and using the rules provided by Rubin (1987) to combine the results.

Rubin's rule is stated as follows. Let $\hat{Q}$ be denoted as an estimate of a population quantity of interest and *U* is its estimated variance. For an example, $\hat{Q}$ can be denoted as an estimated regression coefficient and *U* represented as its squared standard error. We have m equally plausible estimates $\hat{Q}_1, \hat{Q}_2, ..., \hat{Q}_m$.

After performing the analysis on each imputed dataset and their respective corresponding variances $U_1, U_2, \ldots, U_m$. The MI estimate or overall estimate is provided by the following

$$\bar{Q} = \frac{1}{m}\sum_{i=1}^{m} \hat{Q}_i$$

There are two components for the total variance estimation that takes in to account the variability across datasets and within each dataset. The within-imputation variance is as below

$$\bar{U} = \frac{1}{m}\sum_{i=1}^{m} U_i$$

An approximate 95% confidence interval can be obtained using $\bar{Q} \pm 2\sqrt{T}$. Generally, it is wise to calculate the intervals using approximation as below

$$\bar{Q} \pm t_{df}\sqrt{T}$$

where t_{df} denotes the quantile of student's t-distribution with degrees of freedom

$$df = (m-1)\left(1+\frac{m\bar{U}}{(m+1)B}\right)^2$$

p-values for testing the null hypothesis Q = 0 can be derived by comparing the ratio $Q/\sqrt{T}$

to the same t-distribution

The total variance reduces to the sum of the two-variance components with an infinite number of imputations ($m \to \infty$). Furthermore, the confidence interval is based on a normal distribution ($df \to \infty$). The degrees of freedom are affected by the relative sizes of B and $\bar{U}$. When $\bar{U}$ dominates B , the degrees of freedom tends to approach infinity. On the other hand, when B dominates $\bar{U}$, the degrees of freedom are close to the minimum value of $m-1$. If the computed value of df is large, it suggests that little will be gained from a larger m . On the other hand, a smaller computed value of df , which approximates to less than 10 suggests greater efficiency which means narrower intervals and more accurate estimates.

Rubin (1987) also presents an estimate of the fraction of missing information about the population quantity Q:

$$\gamma = \frac{r + 2/(df + 3)}{r + 1}$$

where

$$r = \frac{(1 + m^{-1})B}{\bar{U}}$$

shows the relative increase in the variance based on the nonresponse. Both of the quantities are useful diagnostics to reveal how strongly the estimation of Q may influence the missing data. The below procedures are followed for the above method:

1. To impute the exact observations by taking a suitable model to use that incorporates random variation.
2. To repeat the first step 3-5 times.

3. To perform the desired analysis by using standard and complete data methods on each data set.
4. To obtain a single point estimate, the average values of the estimates on parameter across the imputed value samples.
5. To average the squared standard errors of the imputed value estimates to calculate the standard errors.

CHAPTER 4: RESULTS AND DISCUSSION

4.1 Breast cancer data

One of the common cancer diseases in Qatar is Breast Cancer. According to Qatar Cancer Register, the risk of development of breast cancer among women is 56 per 100,000. It accounts to 31% of the entire women's cancer cases in Qatar. There have been rapid changes in providing multidisciplinary care for the cancer patients through treatment modalities like surgery, chemotherapy and radiation therapy.

National Center for Cancer Care & Research (Al Amal Hospital), part of Hamad Medical Corporation provided the breast cancer data. In the study all the patients were under observation in Al Amal Hospital. The data contain 24 variables collected from the period of 1/2/2016 to 19/01/2020. The data provides details on 1008 patients of which, 557 patients are treated by chemotherapy, 770 treated by surgery, 555 treated by hormone and 533 treated by radiotherapy. The study focuses on comparing cosmetic effects of each treatment separately on women with early breast cancer. The event of interest in the study was the time to the first occurrence of breast retraction. It is noted that the sample has interaction and some patients have undergone surgery treatment and additionally, later they undergo through other treatments. Therefore, we consider a dummy variable for our analysis. The upcoming section will discuss the simulation study used on the data. R package software was used for analyzing the breast cancer data and conducting simulation.

The coefficient, confidence interval (CI) and standard error (SE) obtained by the exponential model based on chemotherapy, hormone and radiotherapy are very similar to the respective values obtained by the Cox model, as shown in Table 4.1. Figure 4.1, 4.2, and 4.3 shows the survival function obtained by our exponential model is comparable to

the results obtained by the Cox model. This result indicates that the validity of our model is comparable with the Cox model.

Table 4.1**:** Estimation obtained by Exponential model and Cox model based on breast cancer data.

Treatment	Method	Coefficient	CI of 95%	SE
Hormone	Exponential	-0.279289	(-0.4132, -0.14542)	0.0641902
	Cox Model	-0.410110	(-0.5602, -0.2065)	0.0683113
RT	Exponential	-0.160000	(-0.2763, -0.01252)	0.0710976
	Cox Model	-0.270290	(-0.28435, -0.0056)	0.0645300
Chemotherapy	Exponential	-0.390898	(-0.2248,0.02742)	0.0671416
	Cox Model	-0.356040	(-0.4845, -0.03471)	0.0659520

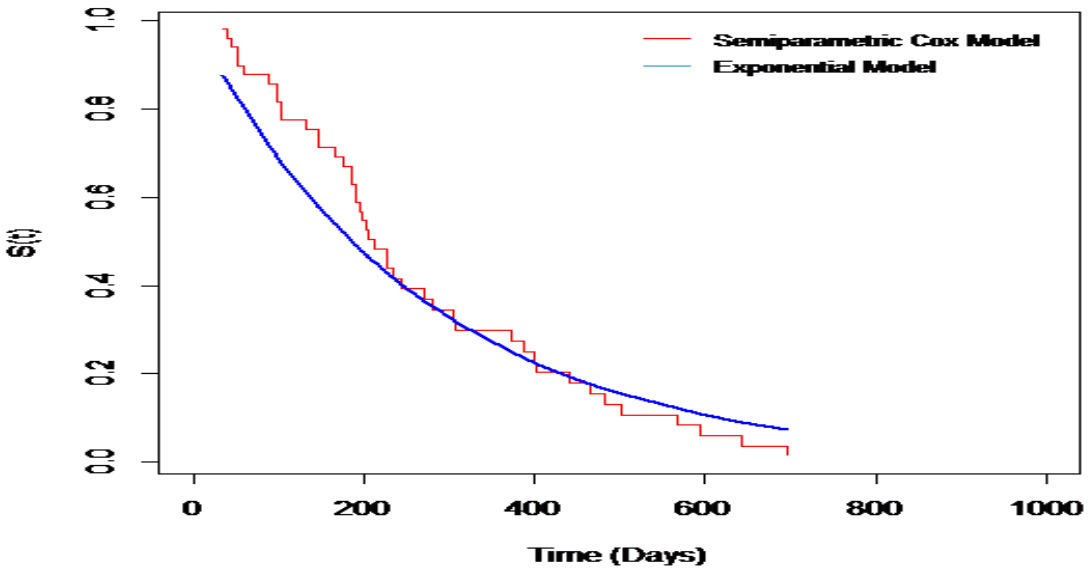

Figure 4.1: Function of survival obtained by exponential model and Cox model for chemotherapy treatment

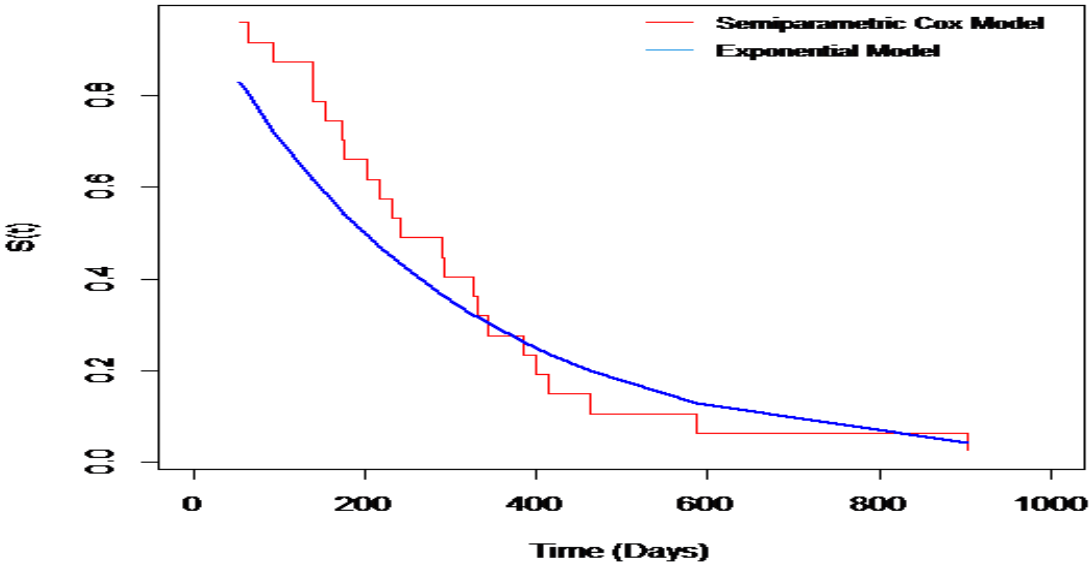

Figure 4.2: Function of survival obtained by exponential model and Cox model for hormone treatment.

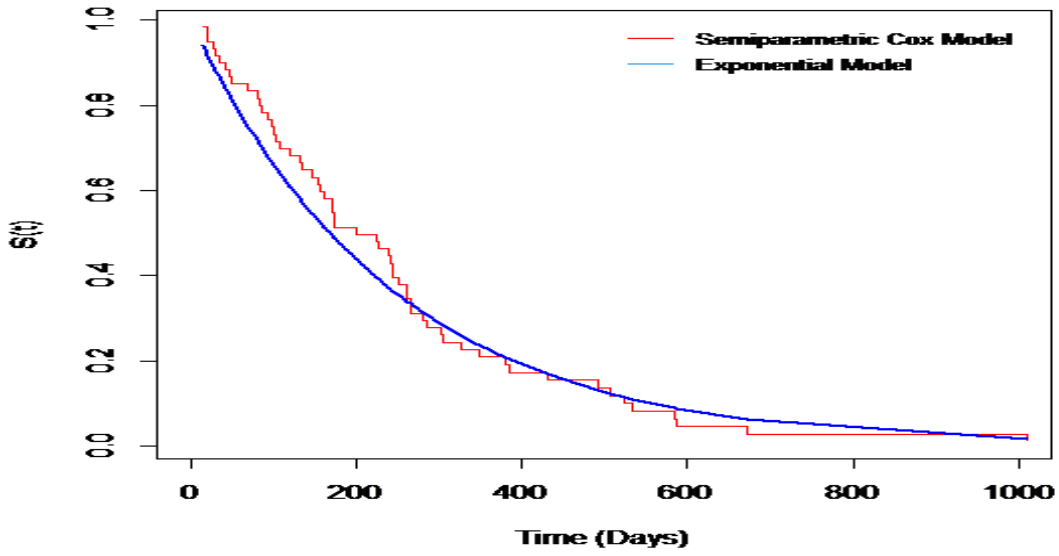

Figure 4.3: Function of survival obtained by exponential model and Cox model for RT treatment

Figure 4.1, show that the survival chance of a breast cancer patient obtained by exponential model which is comparable to the one obtained by cox model. Furthermore, the patient who undergoes the chemotherapy treatment has greater chance of survival by having an increased likelihood of living greater number of days. Therefore, the probability of living increases with the patient undergoing chemotherapy treatment and the LRT - 7654.09 (5.16e-07) along with p-value shows that the treatment is not significant.

The breast cancer patient's chance of survival increases based on hormone treatment as provided in figure 4.2. Likewise, the patient receiving hormone treatment has greater survival chance since they have an increased likelihood of living greater number of days compared to the patient not receiving treatments as shown shows that for the hormone treatment there is significant effect.

The breast cancer patient's chance of survival increases based on surgery treatment as provided in figure 4.3. Likewise, the patient receiving RT treatment has greater survival chance since they have an increased likelihood of living greater number of days compared to the patient not receiving treatments.

Figure 4.4 compares the three treatment methods of chemotherapy, radiotherapy and hormone testing using exponential model. The graph shows that there is no significant difference between the three treatments. However, the graph shows slightly different results in the treatments for example the hormone and radiotherapy has longer survival rate compared to the other treatment.

From the above, on the basis of the survival rates, we recommend radiotherapy followed by hormone and chemotherapy treatment.

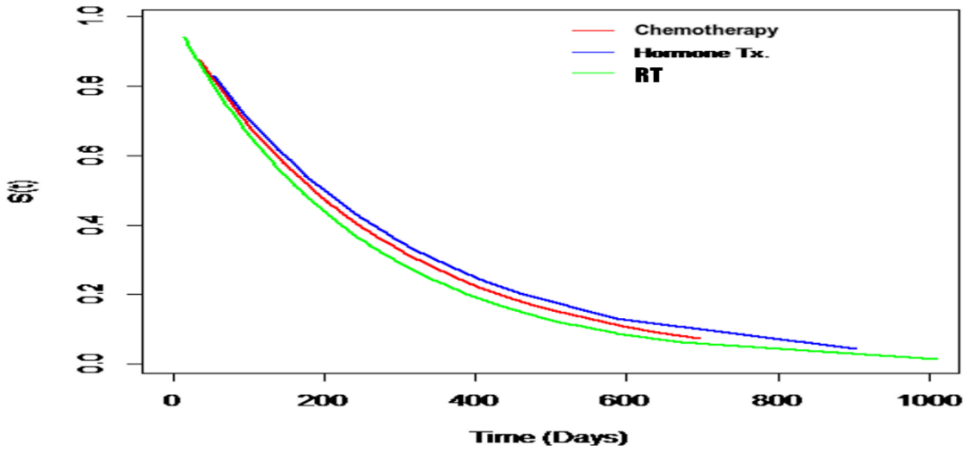

Figure 4.4: Comparing the three treatments obtained by exponential model

4.2 Simulated Data

Simulation is a method to model random events, in a way that simulated outcomes are closely related to real-world outcomes. Using the simulated outcomes, researchers understand and gain more insight on the real world. The process allows to consider the properties of methods as bias. The simulation technique is an extremely valuable tool for statistical research, to compare alternative methods and to evaluate new methods. Generally, it may be difficult to obtain analytic results. As opposed to more general analytic results which covers many scenarios; simulation studies obtain empirical results on the performance of statistical methods. Furthermore, in many situations mathematical treatment does not work as other techniques might be expensive, time-consuming or difficult to analyze. In these cases, simulation could approximate real-world results, but requires lesser time, effort and money than other approaches.

A simulation study is carried out to examine the influence of the exponential model and furthermore to compare the covariates in the data set of breast cancer data.

Normal distribution is used to generate the simulation data since we find that the normal is more reasonable based on real data compared to some other distributions such as Weibull and log-logistic as the Akaike's information criterion (AIC) was found to 14610.14 a normal distribution, 15059.36 for log-logistic and 14672.15for Weibull. Furthermore, the sample used in the study is 20000 times for each treatment.

To generate the data for treatment, a mean and standard deviation of -0.27926 and 0.06419 for hormone (for example) are used on the basis of percentage of exact observation- 0%, 25%, 50% and 75% for the partly interval censored (PIC) data. Moreover, we obtained the function of survival for each simulation data for the two groups of each treatment that are based on the exact observation compared to the estimation by imputation methods which are mean, left point and midpoint.

It is to be noted that 0% exact observation is defined by 100% interval censored. Similarly, 25%, 50% and 75% exact observation indicate that it has 75%, 50% and 25% interval censored respectively.

Table 4.2: Results from surgery obtained by exponential model with MI imputation based on simulation data

% Exact	Parameter	Estimate	CI of 95%	SE	LRT* (P-value
0%	Coefficient	-0.21757	(-0.245682,-0.18946)	1.434e-02	-151186.6(2e-16)
	rate	0.001274	(0.0012487, 0.00130)	1.292e-05	
25%	Coefficient	-0.21765	(-0.245758,-0.18954)	1.434e-02	-151186.3(2e-16)
	rate	0.001274	(0.001248, 0.00130)	1.292e-05	
50%	Coefficient	-0.21729	(-0.24539, -0.18918)	1.434e-02	-151190.5(2e-16)
	rate	0.001273	(0.001248, 0.00130)	1.292e-05	
75%	Coefficient	-0.21718	(-0.245286,-0.18907)	1.434e-02	-151191.5(2e-16)
	Scale	0.001273	(0.001248, 0.00130)	1.292e-05	
100%	Coefficient	-0.28938	(-0.317523,-0.26124)	1.436e-02	-151718.9(2e-16)
	rate	0.001264	(0.001239, 0.00129)	1.283e-05	

Table 4.2 shows the results obtained by exponential model based on multiple imputation with different percentages of exact for the PIC data on the basis of with and without surgery treatment. The results show that surgery is significant with respect to their p-value and LRT (based on α=0.05).

Figures 4.5, 4.6, 4.7 and 4.8 shows the results obtained on the estimation function of the survival obtained by the exponential model via the multiple imputation technique with different exact observations as 0%, 25%, 50% and 75% as mentioned before in this chapter. The estimated survival function obtained from exact observation with 0%, 25%, 50% and 75% are similar to the one obtained by multiple imputation technique. Furthermore, Table 4.2 shows significant results by using the multiple imputation with respect to the values of LRT and their subsequent p-value.

Table 4.2 shows the results obtained through our exponential model using multiple imputation for surgery treatment via different percentages of interval censored data and exact data. It shows significant results with respect to LRT and their p-value. The results depict that the more the exact observations in the data, the better the results. Furthermore, the breast cancer patient using surgery treatment has a greater chance of survival compared to the patient who have not received surgery treatment. Additionally, the null hypothesis test (Ho: there is no difference between the patient who underwent surgery treatment and the patient without surgery treatment) is rejected.

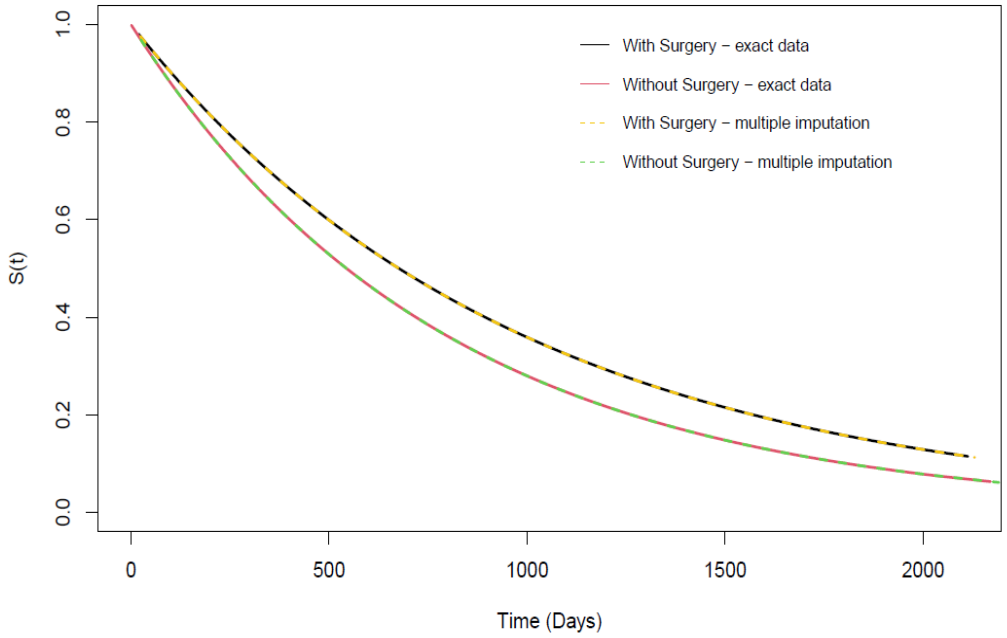

Figure 4.5: Function of survival obtained by multiple imputation for 0% exact data based on surgery treatment

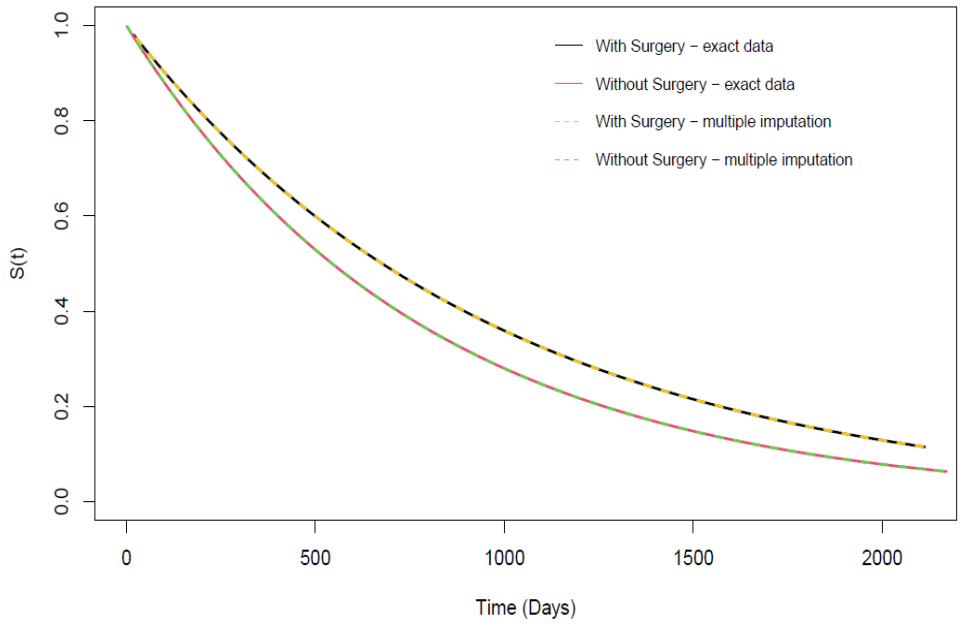

Figure 4.6: Function of survival obtained by multiple imputation for 25% exact data based on surgery treatment.

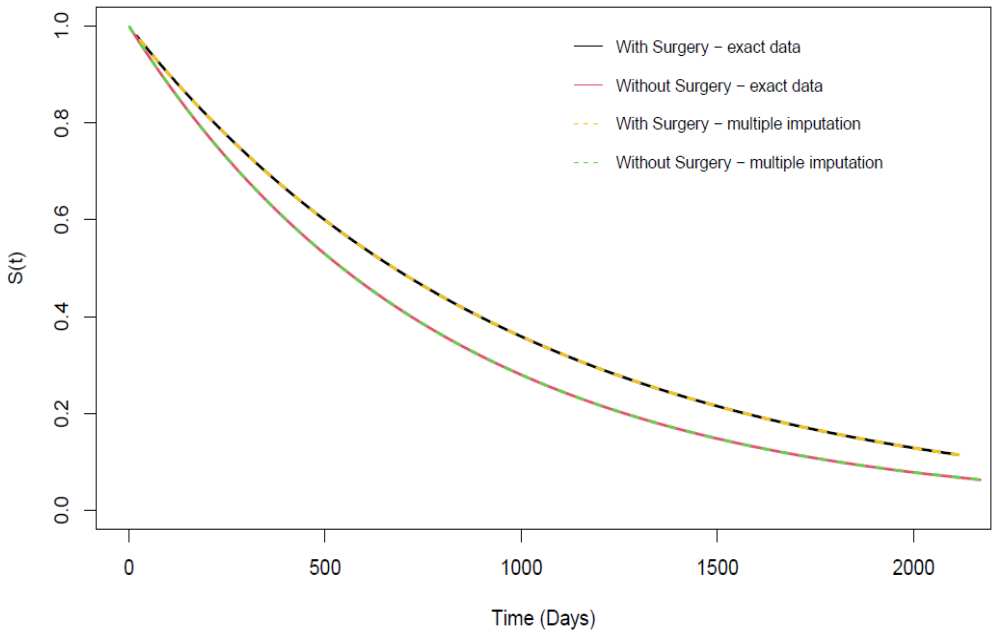

Figure 4.7: Function of survival obtained by multiple imputation for 50% exact data based on surgery treatment.

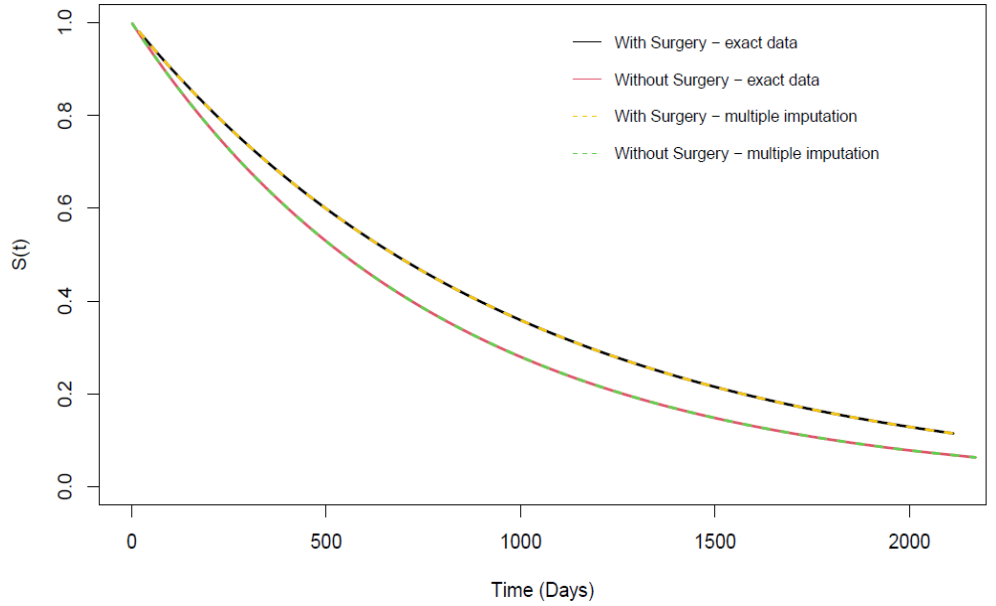

Figure 4.8: Function of survival obtained by multiple imputation for 75% exact data based on surgery treatment.

Table 4.3: Results from RT obtained by exponential model with MI imputation based on simulation data

% Exact	Parameter	Estimate	CI of 95%	SE	LRT* (P-value
0%	Coefficient	-0.23562	(-0.26377, -0.20748)	1.436e-02	-151863.6(2e-16)
	rate	0.001214	(0.001190, 0.00124)	1.232e-05	
25%	Coefficient	-0.23577	(-0.26392, -0.20762)	1.436e-02	-151864.5(2e-16)
	rate	0.001214	(0.001190, 0.001239)	1.232e-05	
50%	Coefficient	-0.23577	(-0.26392, -0.20762)	1.436e-02	-151865.3(2e-16)
	rate	0.001214	(0.001190, 0.00124)	1.232e-05	
75%	Coefficient	-0.23587	(-0.264018, -0.2077)	1.436e-02	-151863.7(2e-16)
	Scale	0.001214	(0.001190, 0.00124)	1.232e-05	
100%	Coefficient	-0.28938	(-0.317523,-0.26124)	1.436e-02	-151718.2(2e-16)
	rate	0.001264	(0.001239, 0.001289)	1.283e-05	

LRT*: Likelihood Ratio Test

Table 4.3 explains the results obtained using our exponential model via multiple imputation by utilizing four different percentage values of exact for the PIC data based on the RT treatment and without RT treatment. It is depicted that RT treatment is significant on the basis of its LRT and p-value.

The results obtained using the estimation of survival function on the figures 4.9, 4.10, 4.11 and 4.12 are obtained using exponential model via the multiple imputation technique with alternative exact observations of 0%, 25%, 50% and 75% which were discussed previously in this chapter. The estimated function of survival acquired with the exact observations of 0%, 25%, 50% and 75% and the curve obtained by multiple imputation technique are alike. Additionally, with respect to the values of LRT and their respective p-values the above table 4.3 shows significant results by using multiple imputation.

For the RT treatment significant results are attained as denoted by the LRT and its

p-value. It shows that with the presence of more exact data, the results are better. Moreover, the patient receiving RT treatment has a greater likelihood of survival chance in comparison to the patient who have not received RT treatment. In conclusion, the null hypothesis of no difference between the patient who completed RT treatment and the patient who did not receive RT treatment is rejected.

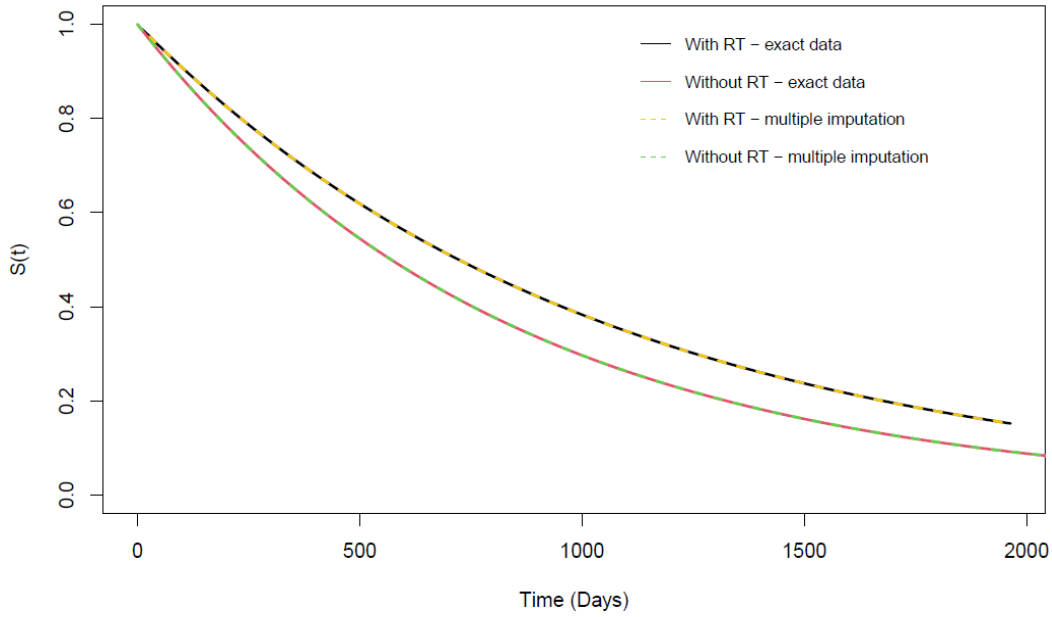

Figure 4.9: Function of survival obtained by multiple imputation for 0% exact data based on RT treatment.

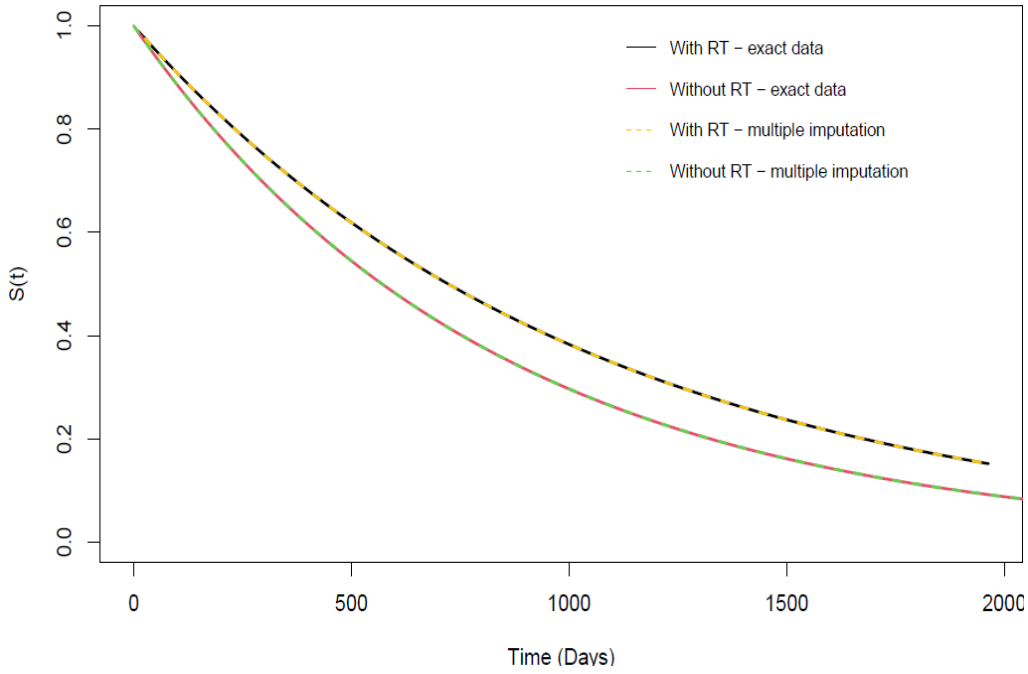

Figure 4.10: Function of survival obtained by multiple imputation for 25% exact data based on RT treatment.

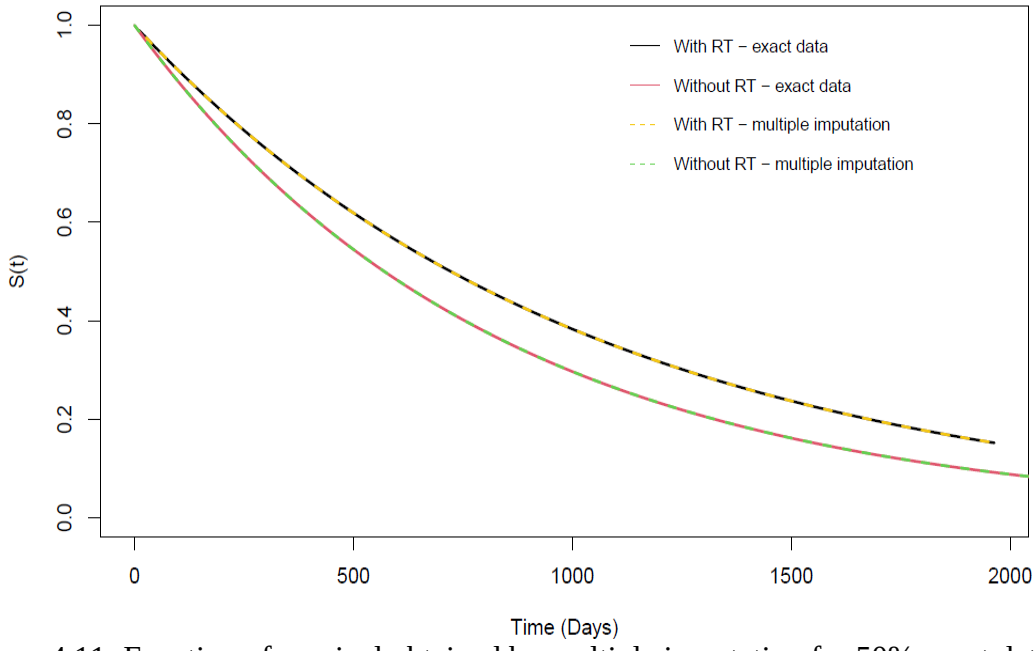

Figure 4.11: Function of survival obtained by multiple imputation for 50% exact data based on RT treatment.

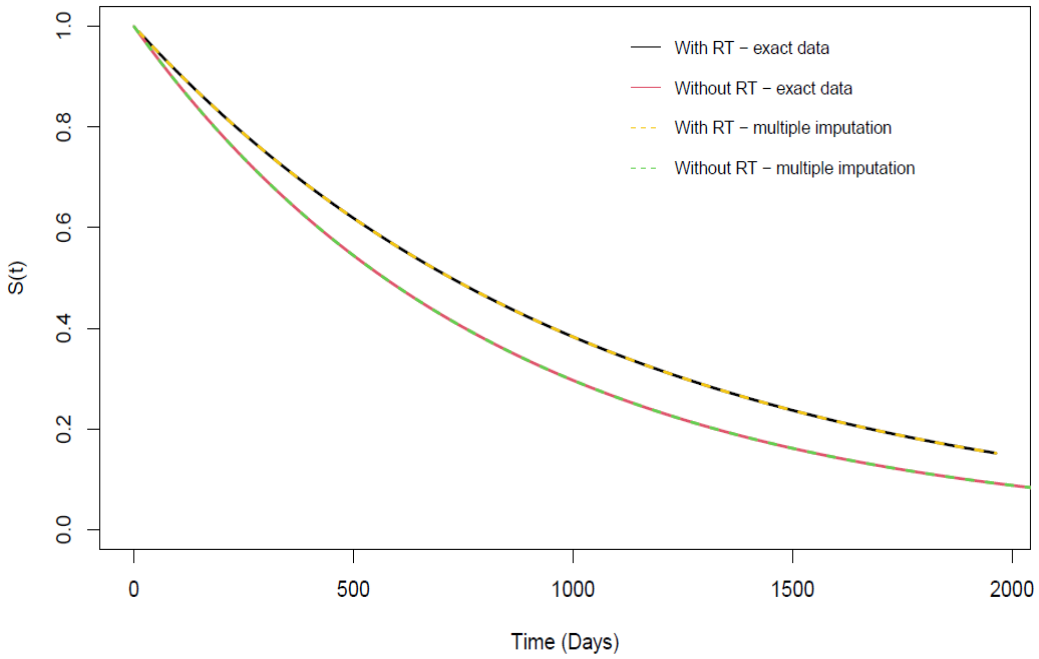

Figure 4.12: Function of survival obtained by multiple imputation for 75% exact data based on RT treatment.

Table 4.4: Results from Hormone obtained by exponential model with MI imputation based on simulation data

% Exact	Parameter	Estimate	CI of 95%	SE	LRT* (P-value
0%	Coefficient	-0.289336	(-0.317476, -0.2612)	1.436e-02	-151719.2 (2e-16)
	rate	0.0012637	(0.001239, 0.00129)	1.283e-05	
25%	Coefficient	-0.289523	(-0.317662,-0.26138)	1.436e-02	-151719.2(2e-16)
	rate	0.0012640	(0.001239, 0.00129)	1.283e-05	
50%	Coefficient	-0.289259	(-0.317398, -0.2611)	1.436e-02	-151720.8(2e-16)
	rate	0.001264	(0.001238, 0.00129)	1.283e-05	
75%	Coefficient	-0.289423	(-0.317563,-0.26129)	1.436e-02	-151720(2e-16)
	Scale	0.0012639	(0.001239, 0.00129)	1.283e-05	
100%	Coefficient	-0.289383	(-0.317523, -0.2612)	1.436e-02	-151722 (2e-16)
	Rate	0.0012640	(0.001239, 0.00129)	1.283e-05	

LRT*: Likelihood Ratio Test

Table 4.4 shows the results obtained by exponential model based on multiple imputation with four different exact percentages for the PIC data on the basis of the following type of failures that is with and without hormone treatment. It is concluded that the results show that the hormone is significant with respect to the p-value and LRT.

Figures 4.13, 4.14, 4.15 and 4.16 displays the results obtained on the estimation function of the survival attained with the exponential model via the multiple imputation technique with the previously discussed exact observations of 0%, 25%, 50% and 75%. The curves obtained by multiple imputation technique and the estimated survival function using exact observations of 0%, 25%, 50% and 75% are closely alike. Furthermore, with the use of multiple imputation, table 4.4 demonstrates significant result on basis of the LRT and its p-value.

The results prescribe significant results on the basis of LRT and its p-value. As mentioned before, this means the results are better as more exact observations in the data is translated to better results. In addition, the results show that the breast cancer patient using hormone treatment has a greater survival likelihood in comparison to the patient who did not receive the treatment. Finally, the null hypothesis test of no difference between the patient who received and did not receive the hormone treatment is rejected.

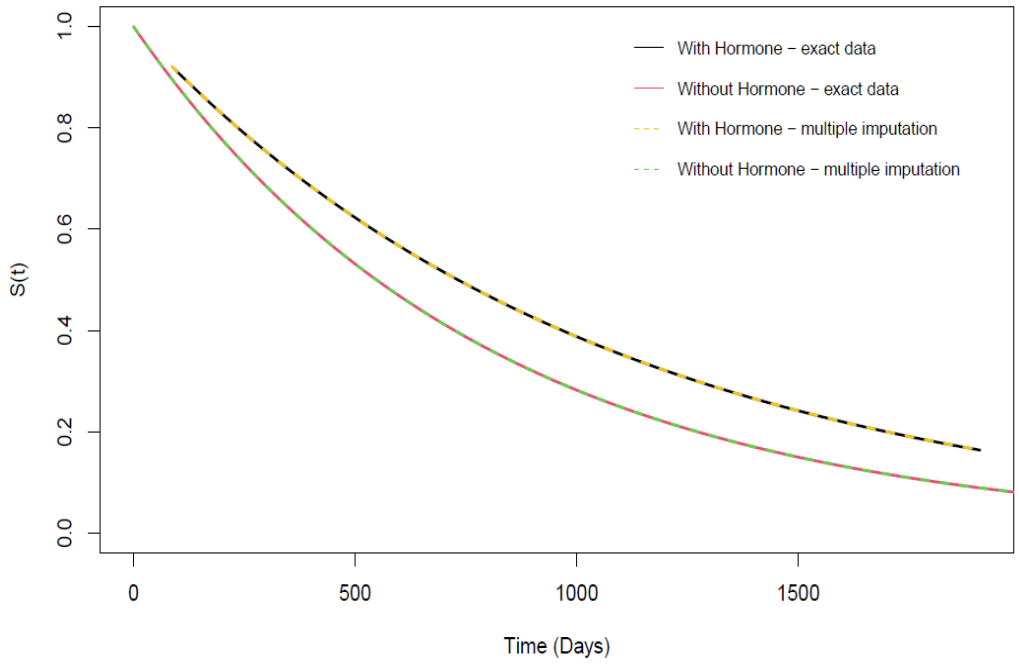

Figure 4.13: Function of survival obtained by multiple imputation for 0% exact data based on Hormone treatment.

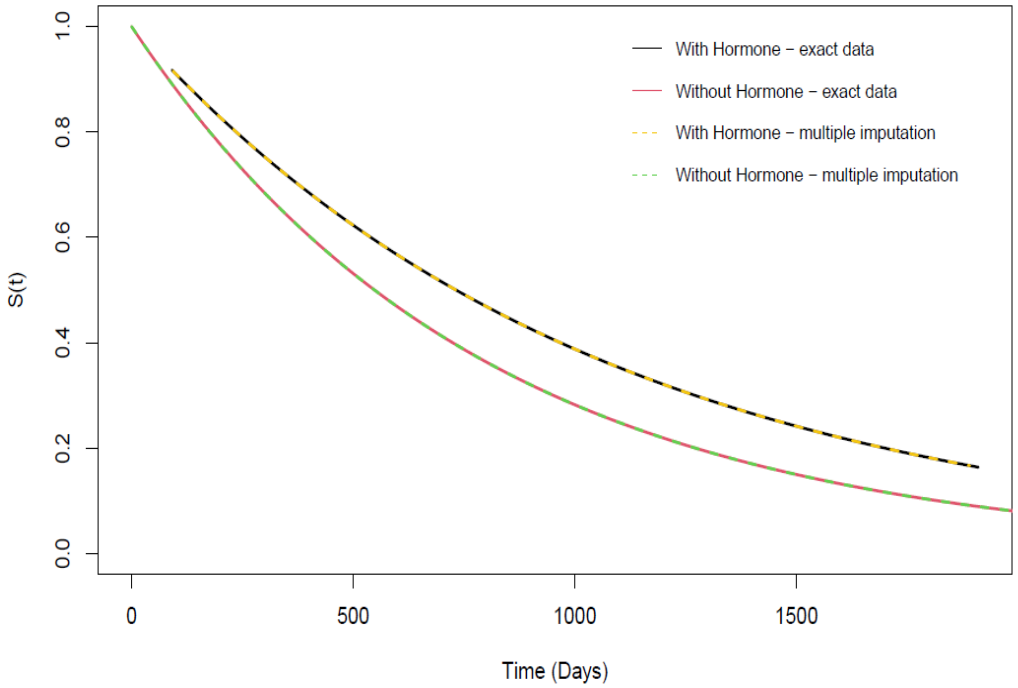

Figure 4.14: Function of survival obtained by multiple imputation for 25% exact data based on Hormone treatment.

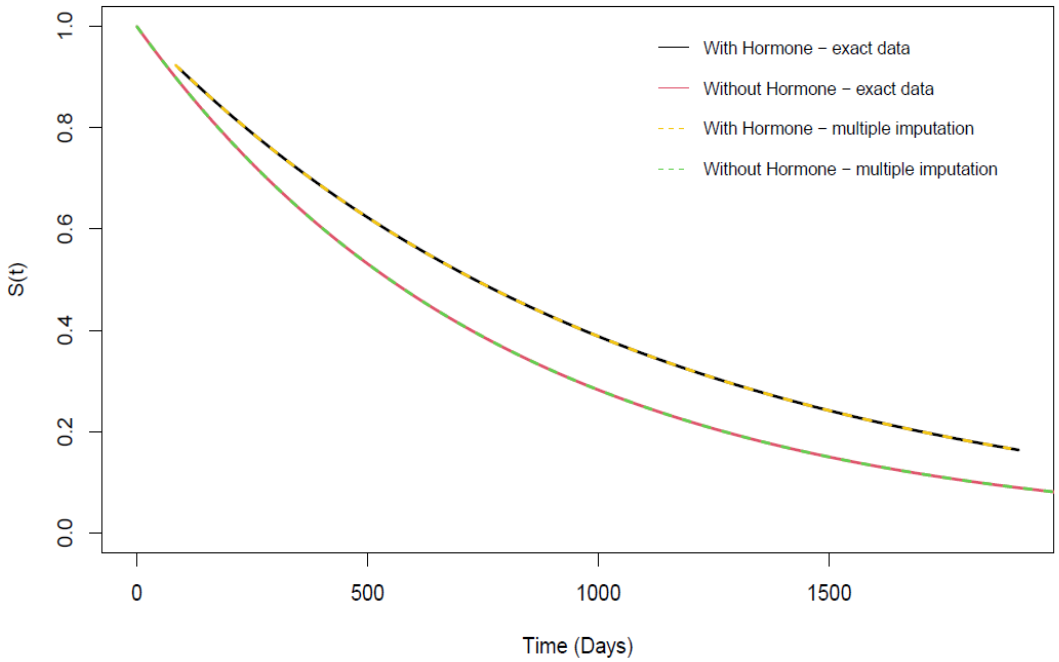

Figure 4.15: Function of survival obtained by multiple imputation for 50% exact data based on Hormone treatment.

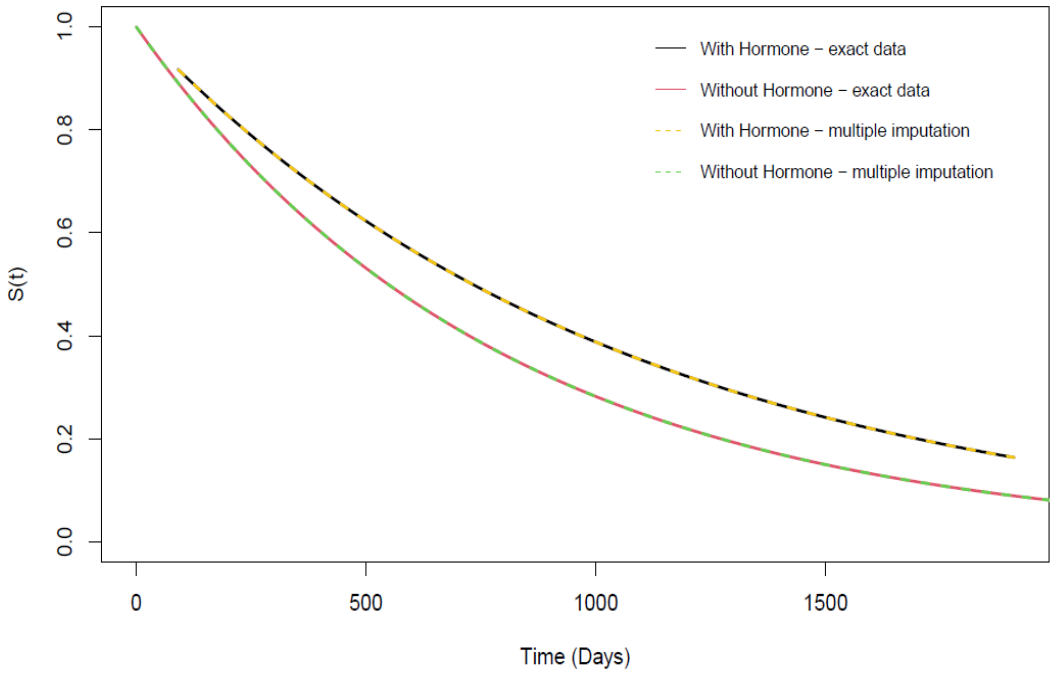

Figure 4.16: Function of survival obtained by multiple imputation for 75% exact data based on Hormone treatment.

Table 4.5: Results from Chemotherapy obtained by exponential model with MI imputation based on simulation data

% Exact	Parameter	Estimate	CI of 95%	SE	LRT* (P-value
0%	Coefficient	-0.11309	(-0.14121, -0.08496)	1.435e-02	-152199.5(3.2e-15)
	rate	0.001139	(0.001116, 0.00116)	1.154e-05	
25%	Coefficient	-0.11319	(-0.141312,-0.08507)	1.435e-02	-152198.1(3.1e-15)
	rate	0.001139	(0.00112, 0.00116)	1.154e-05	
50%	Coefficient	-0.11322	(-0.141345,-0.08510)	1.435e-02	-152198.2(3.1e-15)
	rate	0.001139	(0.001116, 0.00116)	1.154e-05	
75%	Coefficient	-0.11336	(-0.141485, -0.0852)	1.435e-02	-152197.7(2.8e-15)
	Scale	0.001139	(0.001116, 0.00116)	1.155e-05	
100%	Coefficient	-0.11342	(-0.141539,-0.0853)	1.435e-02	-152197.9(2.8e-15)
	rate	0.001139	(0.001116, 0.00116)	1.154e-05	

LRT*: Likelihood Ratio Test

Table 4.5 shows the results obtained by exponential model based on multiple imputation on the two failure times, with and without chemotherapy treatment via different percentages of exact for the PIC data. With respect to the LRT and its p-value, it is concluded that chemotherapy is significant.

Figures 4.17, 4.18, 4.19 and 4.20 with the exact observations of 0%, 25%, 50% and 75% shows the results obtained on the estimation function of the survival are attained by the exponential model and the multiple imputation technique. The curves using exact observation of 0%, 25%, 50% and 75% obtained by both the multiple imputation technique and the estimated survival function are alike. Furthermore, LRT and p-value in table 4.5 demonstrates that the results are significant.

LRT and its p-value in Table 4.5 demonstrate that it shows significant results. Furthermore, the results suggest that the patient undergoing chemotherapy has a higher survival chance when compared to the patient receiving no chemotherapy treatment. As

discussed before, the above table also shows that the exact observations being higher is translated to better results. Moreover, the null hypothesis test of our study showing no difference between patient undergoing chemotherapy and no chemotherapy treatment is rejected.

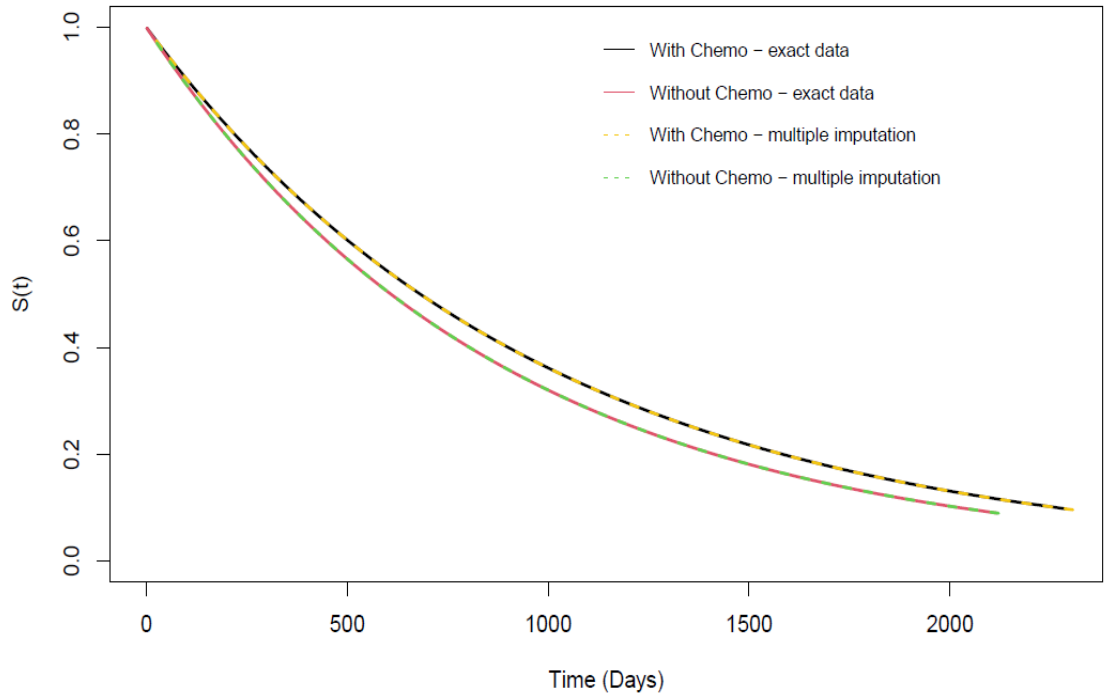

Figure 4.17: Function of survival obtained by multiple imputation for 0% exact data based on Chemotherapy treatment.

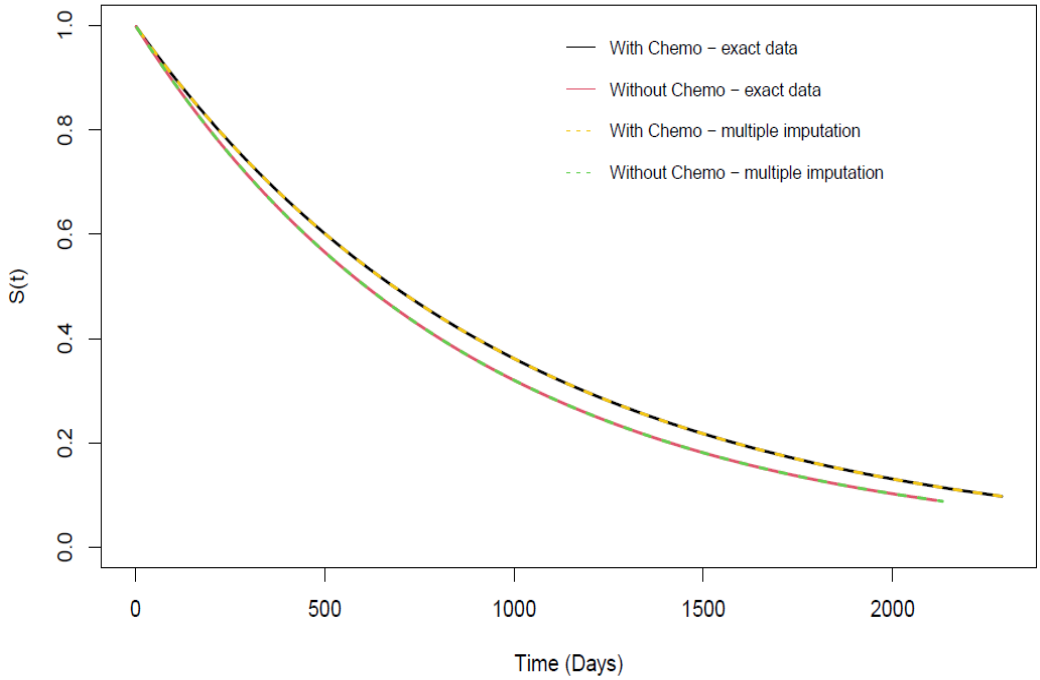

Figure 4.18: Function of survival obtained by multiple imputation for 25% exact data based on Chemotherapy treatment.

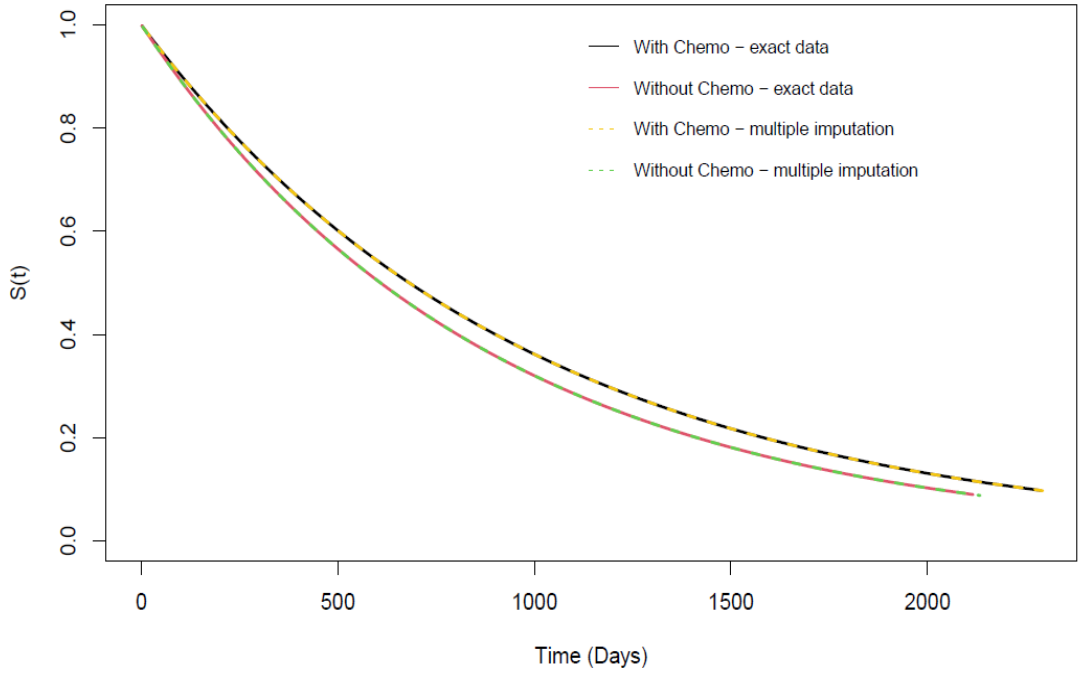

Figure 4.19: Function of survival obtained by multiple imputation for 50% exact data based on Chemotherapy treatment.

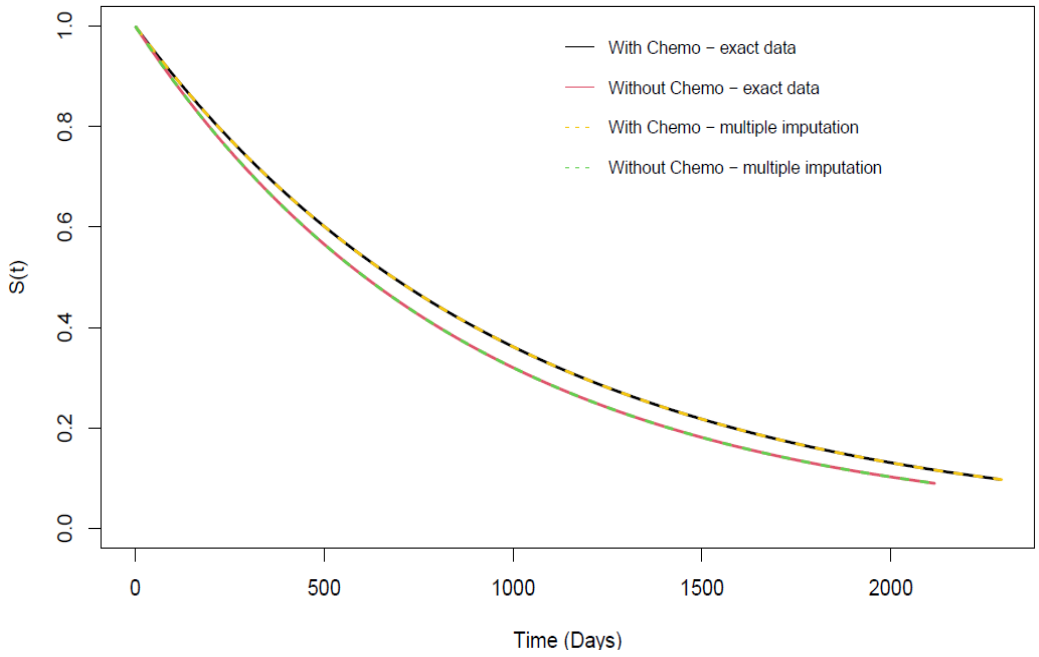

Figure 4.20: Function of survival obtained by multiple imputation for 75% exact data based on Chemotherapy treatment

A simple approach to dealing with missing data is simple imputation. The process involves obtaining a single estimated value for the missing observation, thereby applying augmented data set by enabling standard statistical methods.

The simple imputation includes left imputation, right imputation and mid-point imputation. For a better clarity on our results obtained using multiple imputation, we will compare the results found using the multiple imputation method against the results achieved from left imputation via the following analysis.

The left imputation technique is used in this thesis for the purpose of comparison**.** The tables (table 4.5 as example) define λ as the rate which is defined as the likelihood that an item will survive to a certain point in time on the basis of its survival to an earlier. Also, we defined a 95% confident that rate will fail between the two endpoints of the interval. However, when the confidence interval is shorter, it indicates the estimation is better.

Tables 4.6, 4.7, 4.8 and 4.9 along with figures 4.21 to 4.28 are obtained by our model via left imputation from the simulation data set with 0% and 25% exact observation. Furthermore, the mentioned Tables and Figures demonstrate how the different treatments of chemotherapy, surgery, RT and hormone treatment are different compared to the one obtained by multiple and left imputation using 0% and 25% exact data respectively. It shows that the multiple imputation is better than the left imputation since there is a slight difference between the results obtained via exact observation.

Table 4.6: Results from Chemotherapy obtained by exponential model with Left imputation based on simulation data

% Exact	Parameter	Estimate	CI of 95%	SE	LRT* (P-value
0%	Coefficient	-0.11682	(-0.144940,-0.08869)	1.435e-02	-151552.7(3.9e-16)
	rate	0.00118	(0.001156, 0.00120)	1.196e-05	
25%	Coefficient	-0.1159	(-0.144008,-0.0878)	1.434e-02	-151715.3(6.8e-16)
	rate	0.00117	(0.001146, 0.0012)	1.185e-05	
50%	Coefficient	-0.11504	(-0.14316, -0.08692)	1.435e-02	-151877.9(1.1e-15)
	rate	0.00116	(0.001136, 0.00118)	1.435e-02	
75%	Coefficient	-0.11426	(-0.14238, -0.08613)	1.435e-02	-152037.7(1.7e-15)
	Scale	0.00115	(0.001126, 0.00117)	1.165e-05	
100%	Coefficient	-0.11342	(-0.141539, -0.0853)	1.435e-02	-152197.9 (2e-15)
	rate	0.001139	(0.001239, 0.0013)	1.155e-05	

LRT*: Likelihood Ratio Test

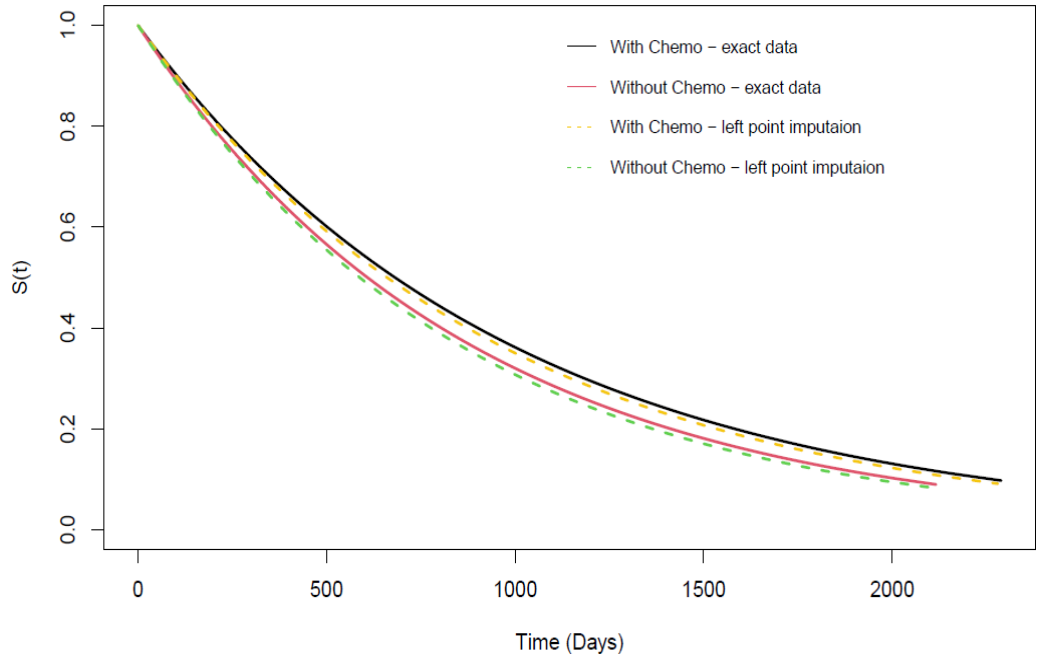

Figure 4.21: Function of survival obtained by left imputation for 0% exact data based on Chemotherapy treatment

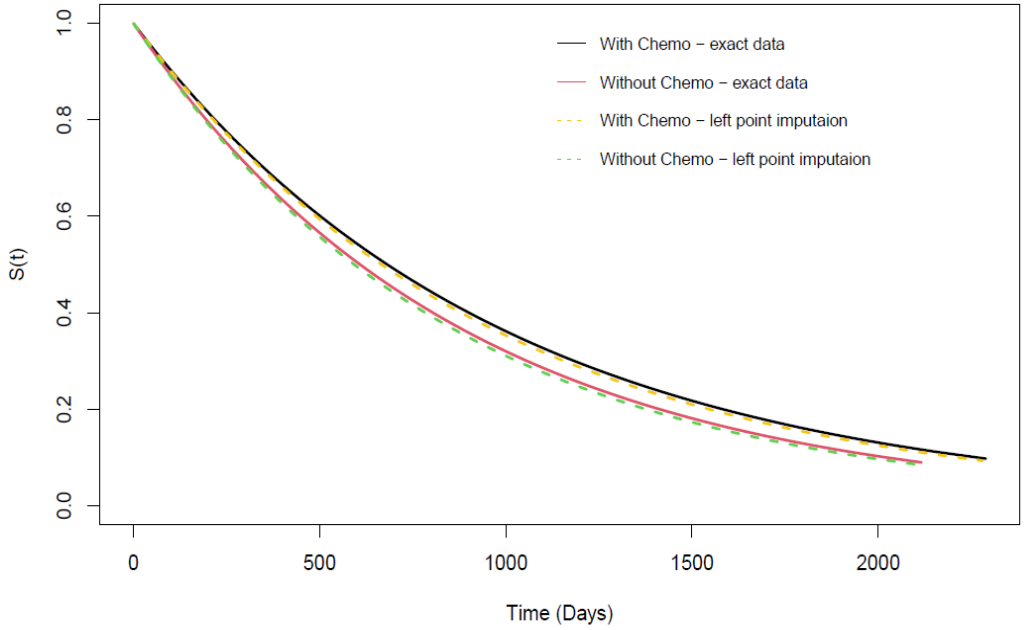

Figure 4.22: Function of survival obtained by left imputation for 25% exact data based on Chemotherapy treatment

Table 4.7: Results from Hormone obtained by exponential model with Left imputation based on simulation data

% Exact	Parameter	Estimate	CI of 95%	SE	LRT* (P-value
0%	Coefficient	-0.29925	(-0.32739, -0.27111)	1.435e-02	-151059.9 (2e-16)
	rate	0.00131	(0.001288, 0.00134)	1.334e-05	
25%	Coefficient	-0.29690	(-0.325035,-0.26876)	1.435e-02	-151224.6 (2e-16)
	rate	0.001302	(0.001276, 0.00133)	1.321e-05	
50%	Coefficient	-0.29412	(-0.32226, -0.26599)	1.436e-02	-151392.7(2e-16)
	rate	0.00129	(0.001263, 0.00131)	1.308e-05	
75%	Coefficient	-0.29183	(-0.319970,-0.26369)	1.435e-02	-151557.6 (2e-16)
	Scale	0.00128	(0.001251, 0.00130)	1.295e-05	
100%	Coefficient	-0.2894	(-0.317523, -0.2612)	1.435e-02	-151718.9 (2e-16)
	rate	0.00126	(0.001239, 0.00129)	1.283e-05	

LRT*: Likelihood Ratio Test

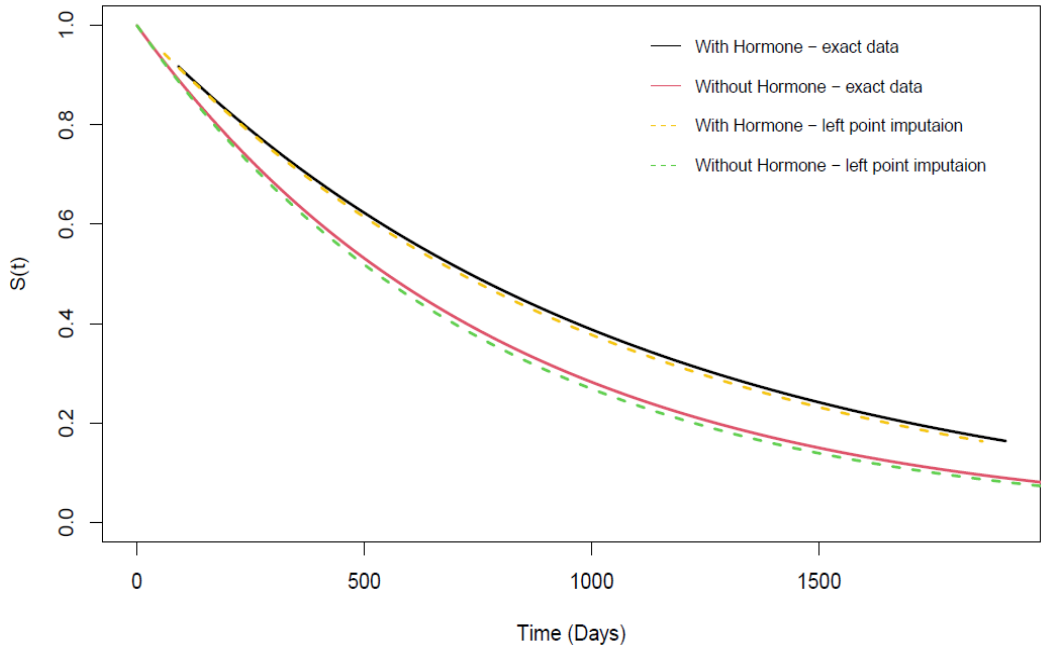

Figure 4.23: Function of survival obtained by left imputation for 0% exact data based on Hormone treatment

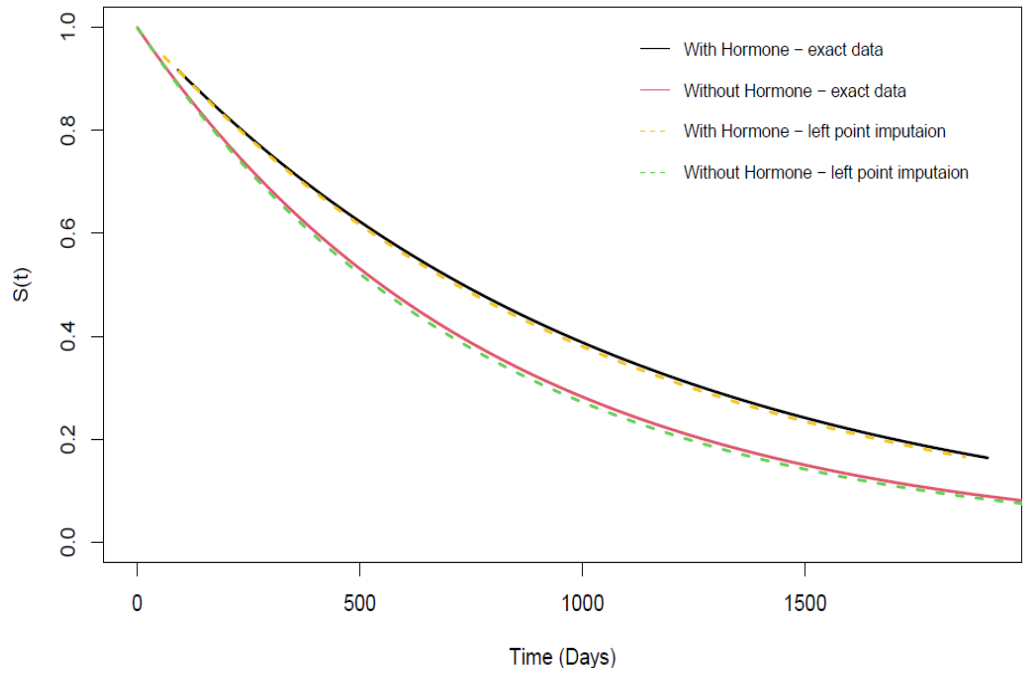

Figure 4.24: Function of survival obtained by left imputation for 25% exact data based on Hormone treatment

Table 4.8: Results from RT obtained by exponential model with Left imputation based on simulation data

% Exact	Parameter	Estimate	CI of 95%	SE	LRT* (P-value
0%	Coefficient	-0.2434	(-0.27154, -0.21524)	1.436e-02	-151211.9 (2e-16)
	rate	0.00126	(0.001235, 0.00129)	1.279e-05	
25%	Coefficient	-0.2416	(-0.26970, -0.21340)	1.436e-02	-151377.7 (2e-16)
	rate	0.00125	(0.001224, 0.00127)	1.267e-05	
50%	Coefficient	-0.2396	(-0.26779, -0.21149)	0.0143621	-151542.1 (2e-16)
	rate	0.00124	(0.001213, 0.00126)	0.0000125	
75%	Coefficient	-0.2378	(-0.26594, -0.20964)	1.436e-02	-151702.7(2e-16)
	Scale	0.00123	(0.001201, 0.0013)	1.243e-05	
100%	Coefficient	-0.2360	(-0.26415,-0.20785)	1.436e-02	
	rate	0.00121	(0.00119, 0.00124)	1.232e-05	-151862.2(2e-16)

LRT*: Likelihood Ratio Test

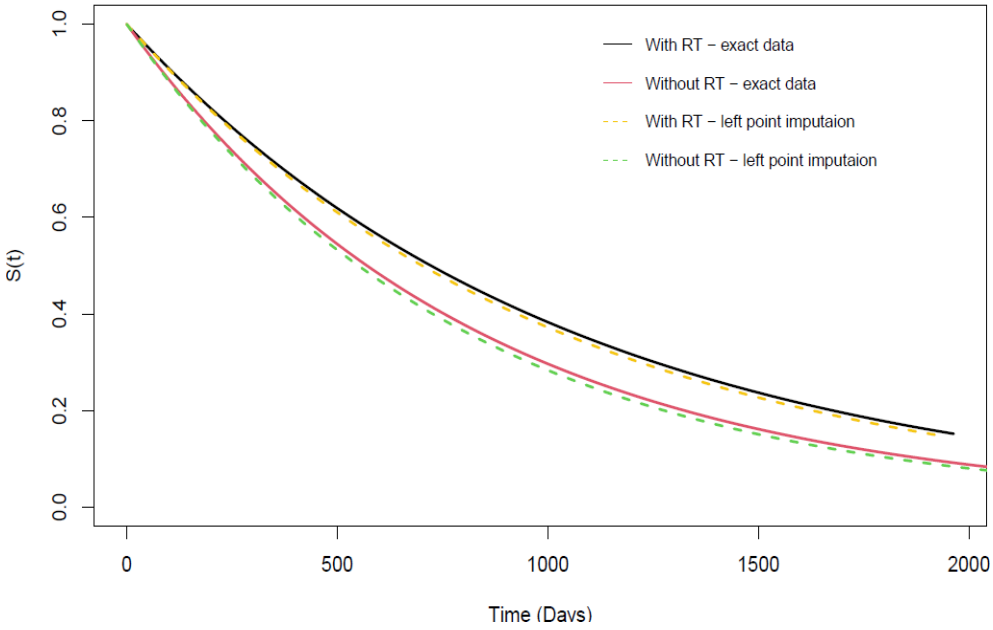

Figure 4.25: Function of survival obtained by left imputation for 0% exact data based on RT treatment

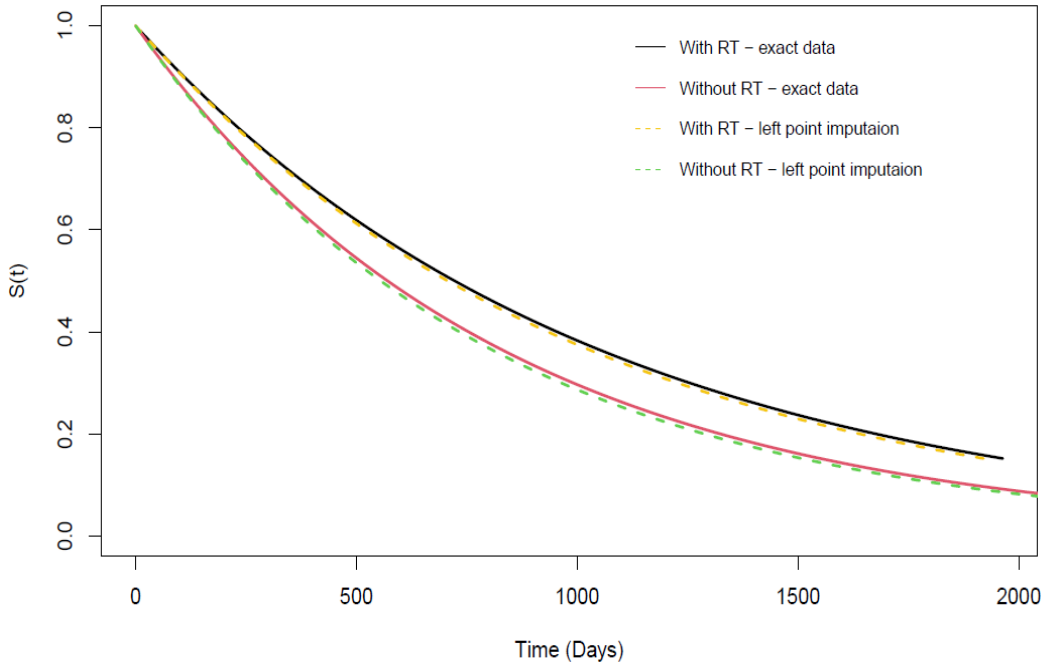

Figure 4.26: Function of survival obtained by left imputation for 25% exact data based on RT treatment.

Table 4.9: Results from surgery obtained by exponential model with Left imputation based on simulation data

% Exact	Parameter	Estimate	CI of 95%	SE	LRT* (P-value
0%	Coefficient	-0.2254	(-0.25354, -0.19732)	1.434e-02	-150495.6 (2e-16)
	rate	0.00133	(0.001299, 0.001351)	1.344e-05	
25%	Coefficient	-0.2235	(-0.251596,-0.19538)	1.434e-02	-150671.3 (2e-16)
	rate	0.00131	(0.001286, 0.00134)	1.331e-05	
50%	Coefficient	-0.2212	(-0.24929, -0.19307)	1.434e-02	-150848.1 (2e-16)
	rate	0.00130	(0.001273, 0.00132)	1.317e-05	
75%	Coefficient	-0.2191	(-0.247207,-0.19099)	1.434e-02	-150495.6 (2e-16)
	Scale	0.00129	(0.001260, 0.00131)	1.304e-05	
100%	Coefficient	-0.2176	(-0.245697,-0.18948)	1.434e-02	-151189.6 (2e-16)
	rate	0.00127	(0.001248, 0.00130)	1.292e-05	

LRT*: Likelihood Ratio Test

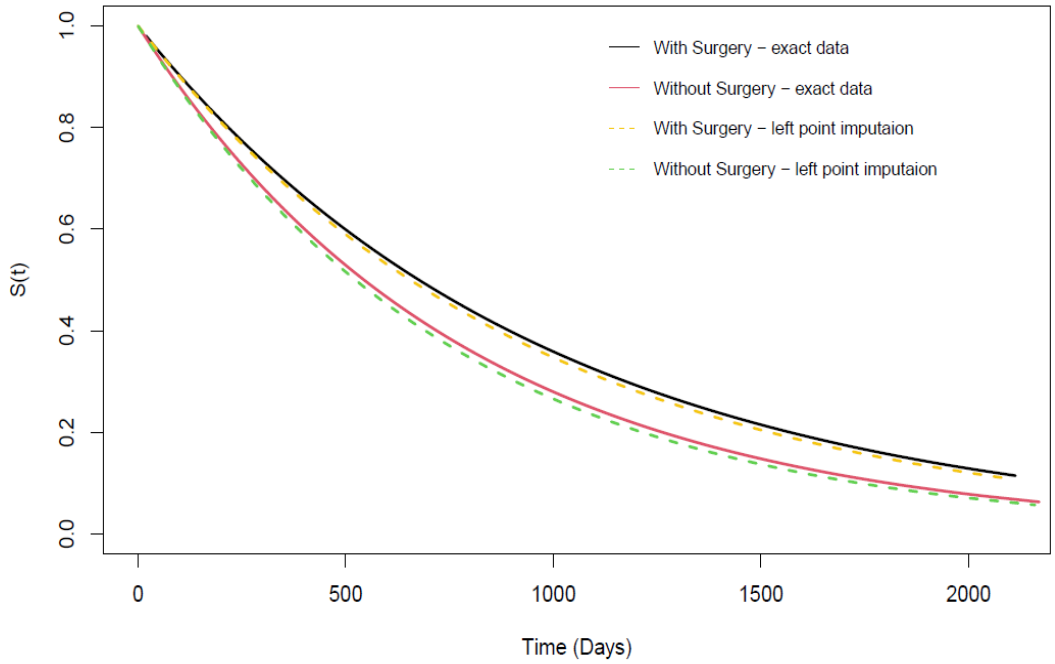

Figure 4.27: Function of survival obtained by left imputation for 0% exact data based on Surgery treatment.

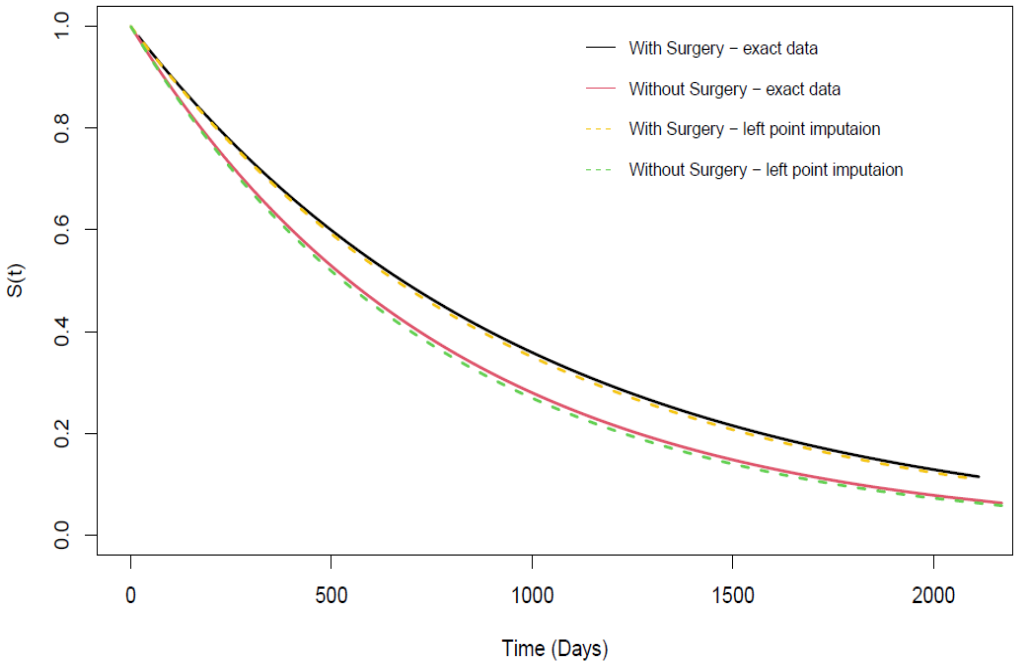

Figure 4.28: Function of survival obtained by left imputation for 25% exact data based on Surgery treatment.

CHAPTER 5: CONCLUSION AND SUGGESTION FOR FUTURE RESEARCH

CHAPTER OVERVIEW

The following chapter discusses two sections. Firstly, the conclusion summarizes the result obtained in the previous chapters. Additionally, the second section provides suggestions for future studies.

5.1 Conclusion

In the above study, we use exponential model based on multiple imputation technique which helps for simplifying the procedure of partly interval censored data. Exponential model has been applied in many different sectors and subjects. In this study, the model is used for medical data. The estimated survival function was derived using the maximum likelihood estimation. The medical data using exponential model was compared with Cox model.

To setup our data as partly interval censored, we consider two months as the interval and we impute the exact observation based on multiple imputation technique that mentioned in earlier chapter. The result from the breast cancer data shows that the survival curves obtained using exponential model lies closer to the survival curves obtained through Cox model. The results are similar in all cases of treatments of chemotherapy, radiotherapy and hormone. Furthermore, the models fit well and was flexible to use for the real data as well for the simulation data with different percentages of exact observations. The coefficient and standard error of the two models were close to each other for the four treatments. The graphs show that there is no significant difference when comparing the four treatments. Furthermore, the graphs show that the surgery treatment has longer

survival rate compared to the other available treatments.

Simple imputation is standardized and is an easy tool for interval data with a shorter period. For the medical data in our study we use multiple imputation which gives a more reliable and better result, since medical data has more missing observations.

The simulation was conducted based on the breast cancer data. The sample was used 20,000 times for each treatment. In addition to that to setup the data as partly interval censored, the interval of two-month period is given and the exact value will be generated via multiple imputation technique and left imputation technique. The function of survival was obtained for each simulation data for each treatment of the two groups (with treatment and without treatment). It is noted that the simulation results obtained from medical data and the results from the simulation data are similar to each other. Furthermore, the medical breast cancer data is suitable for the partly interval censored.

Overall, the result of this study using survival curves has shown that the four treatments provide the patients a longer survival rate than them not receiving any treatment. It is observed that the estimated parameters for our model are almost similar by using different types of exact observations for the PIC data. On the contrary, we note that there is a slight difference in the estimates of parameters between different exact observations for the PIC data.

The model was implemented based on medical data as well as simulation data via LRT. We conclude that both results from the simulation and real data are suitable for partly interval censored via MI and left point imputations.

Finally, the result observed that when the observations in the data are more exact, the model is a better fit, which is a similar inference other researchers like Alharpy and Ibrahim (2013) and Zyoud et al. (2016) have reached in their respective published papers. The simulation study strongly supports the concept that if the data is partly interval censored then the exponential model is a suitable option and it has the potential for being applied to many areas such as medical, education, engineering, and others.

5.2 Suggestions for future research

This research focuses only on the treatments in the data sets; future research can be extended to study the properties of other parameters in the model as well as different factors in the data such as age, gender, level of education, etc.

Moreover, in this study we use the MI and the results are compared to left imputation; future research in this topic can be conducted by comparing MI with EM algorithm which will provide more accurate results. Additionally, inclusion of more factors of the data sets in the analysis ensures more accurate outcomes.

REFERENCES

Alharphy, A. M. and Ibrahim, N. A (2013), Parametric Test for Partly Interval-censored Failure Time Data under Weibull Distribution via Multiple Imputation. *Journal of Applied Science*, 13(4), pp 621-626.

Alharpy, A. M. and Ibrahim, N. A (2014). A Two-Sample Parametric Test for Partly Interval-Censored Data with Non proportional Hazard. *Mathematical Problems in Engineering*, Vol. 2014, Article ID 702847, 10 pages.

Bhandari, V., & Boutros, P. C. (2016). Comparing continuous and discrete analysis of breast cancer survival information. *Genomics*, *108*(2), 78-83.

Brenner, H., and Hakulinen, T. (2002). Up-to-date long-term survival curves of patients with cancer by period analysis. *Journal of Clinical Oncology*, *20*(3), 826-832.
censored data. *Statistics in Medicine*. Vol (17):219-238.

Buckley, J., and James, I. (1979). Linear regression with censored data. Biometrika, 66(3), 429-436.

Chakhoyan, A., Woodworth, D. C., Harris, R. J., Lai, A., Nghiemphu, P. L., Liau, L. M., and Ellingson, B. M. (2018). Mono-exponential, diffusion kurtosis and stretched exponential diffusion MR imaging response to chemoradiation in newly diagnosed glioblastoma. *Journal of neuro-oncology*, *139*(3), 651-659.

Chen, Y., Li, X., Zhang, R., Xia, Y., Shao, Z., and Mei, Z. (2019). Effects of statin exposure and lung cancer survival: A meta-analysis of observational studies. *Pharmacological Research*, *141*, 357-365.

Choi, J., Cai, J., Zeng, D., and Olshan, A. F. (2015). Joint analysis of survival time and longitudinal categorical outcomes. *Statistics in biosciences*, *7*(1), 19-47.

Clark, T. G., Bradburn, M. J., Love, S. B., and Altman, D. G. (2003). Survival analysis part I: basic concepts and first analysis. *British journal of cancer*, *89*(2), 232-238.

Cox, D. R (1972). Regression models and life tables (with discussion). J R Statist Soc B 34: 187–220

Davis, D. J. F. (1952). An analysis of some failure data. *Journal of the American Statistical Association*, *47*(258), 113-150.

Elfaki, F. A. M., Abobakar, A., Azram, M., and Usman, M. (2013). Survival model for partly interval-censored data with application to anti d in rhesus d negative studies. *International Journal of Bioengineering and Life Sciences*, *7*(5), 347-350.

Elfaki, F. A., Azram, M., and Usman, M. (2012). Parametric Cox's Model for Partly Interval-Censored Data with Application to AIDS Studies. *International Journal of Applied Physics and Mathematics*, *2*(5), 352-354.

Friedman, M. (1982). Piecewise exponential models for survival data with covariates. *The Annals of Statistics*, *10*(1), 101-113.

Gao, F., Zeng, D., and Lin, D. Y. (2017). Semiparametric estimation of the accelerated failure time model with partly interval-censored data. *Biometrics*, *73*(4), 1161-1168.

Giolo, S. R. (2004). Turnbull's nonparametric estimator for interval-censored data. *Department of Statistics, Federal University of Paraná*, 1-10.

Glasser, M. (1967). Exponential survival with covariance. *Journal of the American Statistical Association.* Vol 62:561-568.

Holford, T. (1980). The Analysis of Rates and of Survivorship Using Log-Linear Models. *Biometrics, 36*(2), 299-305. doi:10.2307/2529982

Huang, J. (1999). Asymptotic properties of nonparametric estimation based on partly interval-censored data. *Statistica Sinica,* 501-519.

Steffensen, J. F. (1930). Infantile mortality from an actuarial point of view. *Scandinavian Actuarial Journal, 1930*(2), 272-286.

Teissier, G. (1934). *Dysharmonies et discontinuités dans la Croissance* (No. 591.13 T4).

Jakobsen, J. C., Gluud, C., Wetterslev, J., and Winkel, P. (2017). When and how should multiple imputation be used for handling missing data in randomised clinical trials–a practical guide with flowcharts. *BMC medical research methodology*, *17*(1), 162.

Jeong, H., Hong, S. R., Chae, S. W., Jin, S. Y., Yoon, H. K., Lee, J., ... and Jung, J. J. (2017). Comparison of unsatisfactory samples from conventional smear versus liquid-based cytology in uterine cervical cancer screening test. Journal of pathology and translational medicine, 51(3), 314.

Jung, I., Ko, H. J., Rha, S. Y., and Nam, C. M. (2018). Reconsideration of sample size and power calculation for overall survival in cancer clinical trials. *Contemporary clinical trials communications*, *12*, 90-91.

Kaminski, D., and Geisler, C. (2012). Survival analysis of faculty retention in science and engineering by gender. *science*, *335*(6070), 864-866.

Kartsonaki, C. (2016). Survival analysis. *Diagnostic Histopathology* 22 (7), 263–270.

Kim. J. S. (2003). Maximum Likelihood Estimation for the Proportional Hazards Model with Party Interval-Censored Data. *J. R. Statist. Soc.* Series B 65: 489-502.

Kleinbaum, D. G., and Klein, M. (2005). Competing risks survival analysis. Survival Analysis: A self-learning text, 391-461.

Laird, N., and Olivier, D. (1981). Covariance analysis of censored survival data using log-linear analysis techniques. *Journal of the American Statistical Association*, *76*(374), 231-240.

Lee, J., Kim, J., and Jung, S. H. (2007). Bayesian analysis of paired survival data using a bivariate exponential distribution. *Lifetime Data Analysis*, *13*(1), 119-137.

Lindsey, J. C. and Ryan, L. M. (1998). Tutorial in biostatistics methods for interval censored data. *Statistics in Medicine*. Vol (17):219-238.

Lu, C. J., and Meeker, W. O. (1993). Using degradation measures to estimate a time-to-failure distribution. *Technometrics*, *35*(2), 161-174.

Marubini E., and Valsecchi M.G. (1995). *Analysing Survival Data from Clinical Trials and Observational Studies*, 267–292, John Wiley: London

Mazucheli, J., Ghitany, M. E., and Louzada, F. (2013). Power Lindley distribution: Different methods of estimations and their applications to survival times data. *Journal of Applied Statistical Science*, *21*(2), 135.

Odell, P. M., Anderson, K. M., and D'Agostino, R. B. (1992). Maximum likelihood estimation for interval-censored data using a Weibull-based accelerated failure time model. *Biometrics*, 951-959.

Peto, R., and Peto, J. (1972). Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society: Series A (General)*, *135*(2), 185-198.

Plank, S. B., DeLuca, S., and Estacion, A. (2008). High school dropout and the role of career and technical education: A survival analysis of surviving high school. *Sociology of Education*, *81*(4), 345-370.

Prentice, R. L. (1973). Exponential survivals with censoring and explanatory variables. *Biometrika* 60, 279-84.

Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. New York" John Wiley and Sons.

Saeed, N. M, and Elfaki F. A. M. (2020). Parametric Weibull Based on Imputations Techniques for Partly Interval Censored Data. *Austrian Journal of Statistics*. Vol Vol 49: 30-37.

Saeed, N. M. (2018). Weibull Distribution Based on Education Partly Interval Censored Data. *Master Project*. Qatar University.

Ma, Z., and Krings, A. W. (2008). Survival analysis approach to reliability, survivability and prognostics and health management (PHM). In *2008 IEEE Aerospace Conference* (pp. 1-20). IEEE.

Singer, J. D., and Willett, J. B. (1993). It's about time: Using discrete-time survival analysis to study duration and the timing of events. *Journal of educational statistics*, *18*(2), 155-195.

Singer, J. D., and Willett, J. B. (1991). Modelling the days of our lives: Using survival analysis when designing and analyzing longitudinal studies of duration and the timing of events. *Psychological Bulletin*, 110(2), 268–290.

Singh, R. S., and Totawattage, D. P. (2013). The statistical analysis of interval-censored failure time data with applications.

Staplin, N. D., Kimber, A. C., Collett, D., and Roderick, P. J. (2015). Dependent censoring in piecewise exponential survival models. *Statistical Methods in Medical Research, 24*(3), 325-341.

Sukhatme, P. V. (1937). Test of significance for samples of the x2-population with two degrees of freedom. *Annals of Eugenics, 8*(1), 52-56.

Weibull, W. (1939). *A statistical theory of the strength of materials*. Generalstabens Litografiska Anstalts Förlag, Stockholm.

Weissberg, J. I., Andres, L. L., Smith, C. I., Weick, S., Nichols, J. E., Garcia, G., Robinson, W.S., and Gregory, P. B. (1984). Survival in chronic hepatitis B: an analysis of 379 patients. *Annals of internal medicine, 101*(5), 613-616.

Williams, J. S., and Lagakos, S. W. (1977). Models for censored survival analysis: Constant-sum and variable-sum models. *Biometrika, 64*(2), 215-224.

Wu, Y., Chambers, C. D., and Xu, R. (2019). Semiparametric sieve maximum likelihood estimation under cure model with partly interval censored and left truncated data for application to spontaneous abortion. *Lifetime data analysis, 25*(3), 507-528.

Yousif Y. A, Elfaki F. A. M., and Hrairi M., (2016). Analysis of Partly Interval-Censored Data under Competing Risks Framework, *International Journal of Computer Science and Network Security, 16, pp* 25-28.

Zhao, X., Zhao, Q., Sun, J., and Kim, J. S. (2008). Generalized log-rank tests for partly interval-censored failure time data. *Biometrical Journal: Journal of Mathematical Methods in Biosciences, 50*(3), 375-385.

Zyoud. A., Elfaki F. A. M., and Hrairi M. (2016). Nonparametric Estimate Based on Imputations Techniques for Interval and Partly Interval Censored Data. *Science International (Lahore) Journal*. Vol 28(2):879-884.