QATAR UNIVERSITY

COLLEGE OF ARTS AND SCIENCES

LOG-LOGISTIC COX MODEL FOR BREAST CANCER PARTLY INTERVAL

CENSORED DATA

BY

ALAA AHMED EL-SALEM

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
Alaa Ahmed El-Salem defended on 15/11/2020.

_____
Dr. Faiz Ahmed Elfaki
Thesis/Dissertation Supervisor

_____
Dr. Ayman Suleiman Bakleezi
Committee Member

Approved:

_____
Ibrahim AlKaabi, Dean, College of Arts and Sciences

ABSTRACT

EL-SALEM, ALAA, AHMED Masters: January: 2021, Applied Statistics

Title: Log-logistic Cox Model for Breast Cancer Partly Interval Censored Data

Supervisor of Thesis: Faiz Ahmed Mohamed Elfaki

The research in this study is concerned with implementing techniques in data which include censored observations for the evaluation of survival analysis. Analysis of survival research has numerous distinctions in the areas of health, architecture, finance, science, and other fields and it is recognized as failure time analysis. Partly Interval Censoring (PIC) is one of the censoring strategies used in the survival analysis, which may help with several forms of data, especially the incomplete ones. Log-logistic distribution is perhaps the most widely employed lifetime delivery in durability applications. We use the log-logistic Cox model in this thesis focused on adjusted medical with PIC data, as well as simulation data based on PIC.

We find that our model is effective and flexible for breast cancer PIC data and simulated data. From the analysis of our real medical data and simulation data for this specific case, we may infer that our suggested distribution better represents the complexity of the model in terms of the importance of predictions of the scale and the shape parameters. Survival distribution feature plots against failure periods are used to analyze the predicted trends of survival for the two kinds of failures.

DEDICATION

This dissertation is lovingly dedicated to my parents, Dr. Ahmed El-Salem and Hayat El-Mawed, and to my sister Dr. Samah El-Salem and my brother Dr. Bilal El-Salem. Their blessings, affection and kindness have guided me throughout my life. They have encouraged me, empowered me and led me to the correct path of my ambitions with passion and dedication. They showed me the value of endurance, determination, and gratitude. Without their efforts, this work would not have been made possible.

It is also a huge pleasure for me to devote this thesis to all my lecturers at the College of Arts and Sciences at Qatar University. Their dedication and passion for their area of knowledge have made me a successful undergraduate student, without their motivation, this dissertation would not have been completed.

ACKNOWLEDGMENTS

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER OVERVIEW

The introduction and Background of the survival study, censored data, and log-logistic distribution model are presented in this chapter. The chapter also presents the statement of the problem, the purpose and the study scope.

1.1 Introduction and Background

The goal of data analysis is to find out useful information to make appropriate decisions and suggest conclusions in various studies and researches. It is needed in all fields such as engineering, medical science, business, social science, etc. There are many methods for analyzing data like, the study of events that happen during the time. One of the methods of data analysis is called survival analysis, to which much attention was given in the last few decades.

Survival analysis or lifetime failure time analysis was described as one of the most significant research approaches to statistics in the past centuries (Sam and Krongs, 2008). In fact, Singh and Totawattage (2013) noted that this is a significant statistical priority, since it includes the failures of death and as well as the tools. The survival analysis was described by Kleinbaum and Klein (2005) as the statistical method that assesses the outcomes and the timely consequence before an occurrence might occur.

There are several applications of survival analysis for example in; medical science, engineering, education, economic and other areas. Usually the approach to survival research has been commonly utilized in both electronics and biomedical applications. As stated by Xian Liu (2012), one of the examples of engineering applications dealing with the method of survival analysis is the reliability of a mechanical or an electrical system on the existence of research. The scientist is implementing this technique to monitor the life span of the products and material in order to estimate the reliability of the product.

Patient / participant chance can survive this is the key period estimate or calculate a survival function in survival analysis. Demographers and social scientists are involved in the length of any human life. Lawless, (2003) defined the length of a marriage as lifetime; a marriage can end because of annulment, divorce, or death. Another illustration from the educational field, as stated by Eagle and Barnes (2014) used the survival method analysis to measure time before an incident might happen and compensate for teachers attrition.

Emmert-Streib and Dehmer (2019) described survival analysis in general, as the set of statistical research techniques for data analysis in which time is a variable outcome. The time refers to the one preceding a given occurrence. For example, an event can include mortality, heart attack, drug use, divorce or parole violation. Through such varied cases, it is evident that the survival approach can be extended to other situations in various fields. In fact, survival approach is used in many fields; such as genetics, medical science, engineering, marketing, and social or behavioral sciences. The extensive usage of this approach in various areas led to the development of many synonyms.

Yin et al., (2012) defined the variable which measures the survival time from a starting time to a specific endpoint of interest. For every research there are some subjects could not complete a survival time due to censorship and some time we don't know exactly how long they survived. Within the following section the censored data would be introduced.

1.2 Censored Data

Survival analysis is included in all areas of life; it plays an important role in the medical field. It depends on the time until an event occurs, such as time to die, time to be sick, time to get a job, etc. We might have censorship, that happens when one or

more than one of the patients is excluded during the observation period, and that's because some patients may leave the study before the end of the observation period (loss of follow-up), or the study ended and the event did not yet occur or because the individual has been withdrawn from the study for some reasons.

Right censorship, left censorship or interval censorship are the general types of censorship. Kartsonaki (2016) discussed that the easiest and most common type of censorship handled by the survival analysis is the right censorship. The latter can occur when an individual is followed-up from the begging of the study and he/she has not get the event of interest. Right censored may occur when an individual is excluded from the study before the study ended, or because of the event didn't occur and the study already came to an end.

Prinjaet et al., (2010) defined the left censored if the patient had been on risk for a period of time before starting the study. For example, consider the event is the age at which children are able to learn the alphabet at school. There may be some children who are able to recite the alphabet before starting school; these subjects are left censored (Islam 2016).

Censorship of intervals usually reflected an incomplete data structure (Zhang2009). Through interval censorship, we say that a random interest variable is only known to lie into interval. In survival analysis, the random variable is the time to some event, such as death, a recurrence of the disease. There is one common example of medical or health studies that require regular follow-up.

Lu and Tsai (2009) used lifetime distribution for log-logistic based on interval censored sampling plans. Dugueet al., (2016) discussed that the exact moment at which the event occurred is unknown in the censored data frames, although there was a time period during which the event occurred. In such cases the dataset includes two

variables: T1 and T2. The time period between T1 and T2 is the time frame during which the event occurs .

Partly interval censored (PIC) contains accurate data and censored intervals. This means that all subjects with equal failure times are measured, but the survival time of interest for the remaining subjects is found to belong only to an interval (Kim, 2003). PIC data, which deals with an exact observation as a very short interval-censored observation Peto and Peto (1972). However, in this study, the log-logistic Cox model will be used based on PIC data. In the next section the log-logistic model will be introduced.

1.3 Log-logistic Distribution Model

The log-logistic distribution is an alternative model of Weibull distribution. The log-logistics is one of the parametric models used in survival time to which the hazard rate may be increasing, or decreasing or it may increase and then decrease (hump-shaped). Also, the distribution has a fairly flexible functional form. The log-logistic can be a mixture of Gompertz distributions and gamma distributions mixture variable with mean and variance equal to one.

Several researchers used log-logistic regression in their study based on survival analysis as; Abbott (1985) used log-logistic regression when the event time was grouped into interval. In their study, they found that log-logistic regression is useful compared to Cox based and this conclusion is based on censored data.

In particular, the log-logistic distribution is useful for modeling unimodal (i.e., non-monotone) hazard functions. Let T has a log-logistic distribution then;

$$Y = \log T = \alpha + \sigma W$$

Where W has a standard logistic distribution, with probability density (pdf) as;

$$f(w) = \frac{e^w}{(1+e^w)^2}$$

The cumulative distribution function as;

$$F(w) = \frac{e^w}{1+e^w}$$

The survival functions as;

$$S(w) = 1 - F(w) = 1 - \frac{e^w}{1+e^w} = \frac{1}{1+e^w}$$

By changing the variables into T (since W has the extreme value distribution, $\alpha = -\log \lambda$ and $p = 1/\sigma$), we find that the pdf, survivor and hazard based on the log-logistic distribution respectively given as;

$$f(t) = \frac{\lambda p (\lambda t)^{p-1}}{[1+(\lambda t)^p]^2} \qquad (1.1)$$

$$S(t) = \frac{1}{1+(\lambda t)^p} \qquad (1.2)$$

$$h(t) = \frac{\lambda p (\lambda t)^{p-1}}{1+(\lambda t)^p}, \quad p > 0; \ \lambda > 0 \qquad (1.3)$$

where $t > 0$ is the distribution support, while $p > 0$ and $\lambda > 0$ are the parameters, λ is reparametrized in terms of predictor variables and regression parameters, the shape parameters p is held fixed. It is easy to check that the log-logistics distribution's hazardous function is monotonous, decreasing in the case of $p \le 1$, and unimodal in the case of $p > 1$ as shown in Figure 1.

Loglogistic hazard function

Loglogistic cumulative hazard function

Loglogistic density function

Loglogistic survivor function
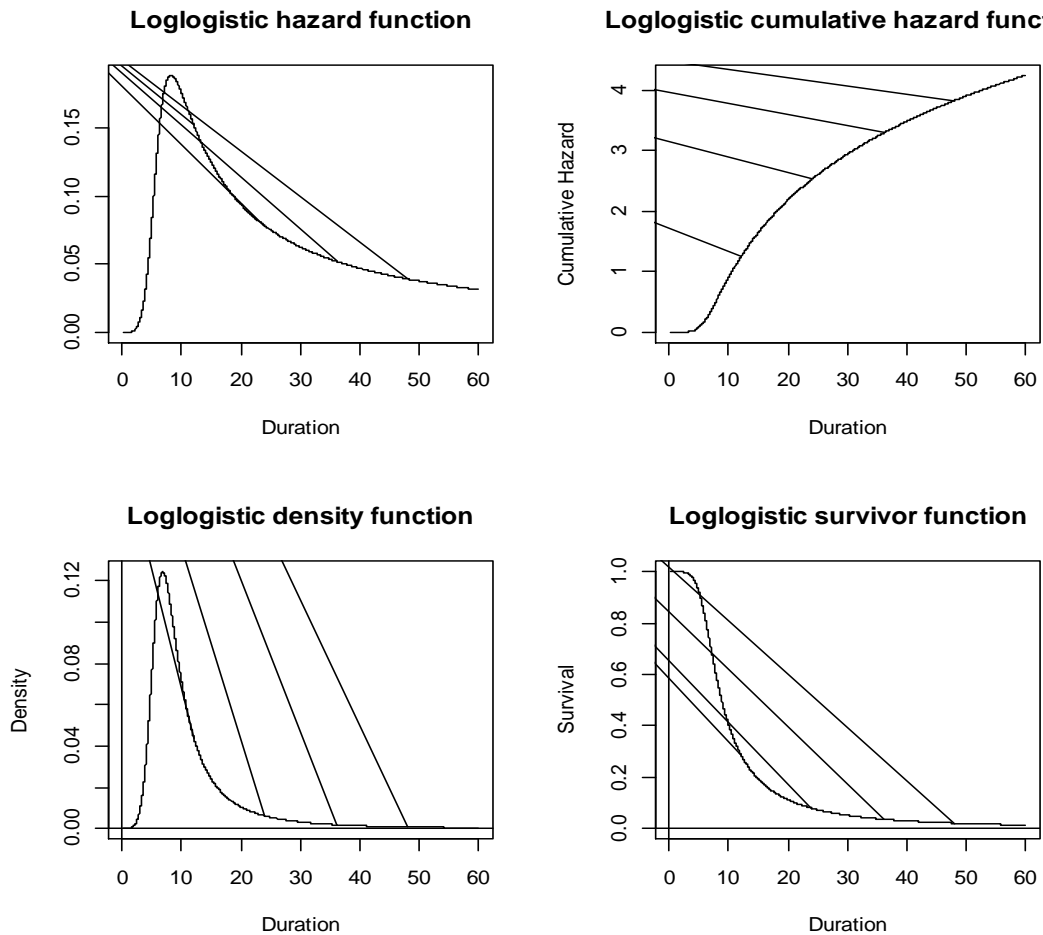
Figure 1: The graphs for hazard, cumulative hazard, density and survival functions Based on log-logistics distribution.

The log-logistic distribution is commonly used to characterize the course of a disease in which mortality reaches a peak after certain specified time duration, then gradually declines (Bennett 1983). For example, the log-logistic model can be used to explain the lifetimes of patients with breast cancer (peak mortality of patients with breast cancer occurs after around three years (Langlands et al., 1979).

Bayaga (2010)found that the dependent variable should be binary so the relationship between independent variables and dependent variables is based on binary log-logistic regression. However, it can be as a problem in survival analysis after the

event time's group into interval. They showed how binary logistic regression can be viewed as a special case of survival analysis. Also the model obtained from survival analysis with logistic distribution is the same as logit and probit models. The distribution mostly used in the survival analysis and the failure-time data is called Weibull distribution. Bennett (1983) stated that the logistic distribution is an alternative useful way in modeling the survival and failure time data to Weibull distribution.

Tang et al., (2017) Extended log-logistic regression to transformed fractional responses from censored survival data. By developing a median test for censored survival data that can be used to measure a group impact that can be adjusted for multiple confounding factors. A quasi-likelihood inferential procedure to construct the test statistics is adopted. Simulation studies show the probability of empirical form I error and the powers for the modified two sample median test are rational. The approach is demonstrated with a dataset on breast cancer. Under certain conditions a parametric model can produce estimates that are more efficient than nonparametric.

Gupta et al., (1999) analyzed lung cancer data based on Log-Logistic model via Maximum Likelihood Estimator (MLE). Chatterjee and Chatterjee (2010) applied binary logistic regression for survival analysis. Cithoet et al. (2012) analyzed the skin test reactivity based on ordinal logistic regression.

Khan and Khosa (2016) proposed a model that has similar properties to log-logistic and Weibull distribution. Senaviratn and Cooray (2019) used logistic regression model for diagnosing multi-collinearity. In this study, we will consider PIC in log-logistic Cox model based on simple imputation techniques which are used to simplify the procedure. We are also, looking to apply our model which will be significantly used in medical data.

1.4 Problem Statement

Among several studies of logistic regression in survival analysis, Abbott (1985)used the mentioned model when the event time is grouped into intervals based on Framingham Heart study. Bayaga (2010) showed the relationship between independent and dependent variables based binary logistic model. Bennett (1983) showed that logistic distribution is an alternative useful way in modeling the survival and failure time data to Weibull distribution. Tang et.al.(2017) extended logistic regression to transformed fractional responses from censored survival data. Khan and Khosa (2016) compared the generalization of log-logistic distribution with different models based on several data sets.

One of our interests will be modelling breast PIC data and predicting future observations. The aim is to estimate the parameters of the flexible log-logistic Cox model based on MLE method via imputation techniques.

1.5 Objective

In this study, we would like to focus on the prediction of patients' survivability in the hospital based on the log-logistic Cox model in the present of medical PIC data. Maximum likelihood estimator will be used to estimate the parameters in the model. Moreover, the main objectives of this study are:

- To modify the log-logistic Cox model to render its suitable with cancer PIC data based on imputation techniques.

- To estimate the survival function when the data is PIC.

- To investigate the performance of our model using the likelihood ratio test.

- To use real data and simulation data.

1.6 Research Questions and/or Working Hypothesis

- Is the modified log-logistic suitable for PIC data?

- Is the simple imputation technique suitable for PIC data?

1.7 Organization of Thesis

       The research study is limited to partially use interval censored data from breast cancer based on the Cox model to predict Hamad Medical Corporation's survival-ability of patients. This model is defined in Chapter III, and the MLE is used to estimate the model parameters. It will also use simple imputation techniques to modify the real data set into PIC data. Chapter 2 presents the literature review of the survival analysis, Log-logistic distribution, partial interval censorship and Cox model. The Cox model will be presented in Chapter 3, based on survival analysis and maximum likelihood estimator derivation for parameters. Also, the techniques of likelihood ratio research and imputations will be discussed in Chapter 3. At the end of Chapter III, data will be presented on the real set of Breast Cancer data and the process to be viewed as survival time. Additionally, the simple imputation techniques used to alter educational data to be right, interval and PIC. Cox model which is appropriate for our updated data sets based on simple methods of imputation will be described in Chapter IV. At the end of this chapter, an example will be provided based on our secondary medical data model, real data set and simulated data. Finally, the results drawn in previous chapters are outlined in Chapter V and recommendations for future study are given.

CHAPERT OVERVIEW

Throughout this segment, we will examine some current survival research based literature that has been applied throughout different fields such as medicine, engineering, economics …etc. Also some current similar literature for log-logistic distribution, and we will concentrate on one form of censoring data used in this thesis which is PIC. Eventually, any relevant literature referring to Cox model.

2.1 Survival Analysis

Giolo (2004) implemented the non-approach for estimating the survival function using censored interval data. Giolo used the Turnbull method to achieve the survival prediction parameters based on R program. It has been found that analyzing this form of censored data and then applying this approach to normal time to event data will lead to inferences that are worthless. Hence, Giolo recommends researchers to be more careful by using the new approaches to analyze censored data at intervals.

Singh and Totawattage (2013) discussed survival research techniques through data on time failure intervals. Five separate parametric and non-parametric approaches were used. The methods used in their analysis were estimator Kaplan-Meier, process Turnbull, survival curve log spline, scale Weibull and piecewise exponential model to predict survival function parameters. To demonstrate the methodology of their analysis they used different data sets, which included AIDS, Hemophilia, and Breast Cancer. From their data, they showed that the parametric approach can be more satisfying in efficiency, particularly when the parametric estimation of the log-normal family or Weibull is chosen because it provides a broad range of distributive types. So, they recommend using a piecewise constant model of danger to enable additional robust simulation with weak parametric assumptions.

Singer and Willett (1993) studied the period and timing of events from the perspective of the application for education focused on a discrete-time survival study. Also, they found that the discrete-time survival model offers a readily accessible method for evaluating data on a form of case incident that is often obtained in educational studies. Interpretation of the parameters of the discrete-time hazard model that can be conveniently implemented basis on a conventional logistic regression study. Furthermore, the discrete-time method allows the analysis of the hazard function form which is opposite to the Cox regression model, where the hazard function form is neglected in favor of calculating only the change parameters correlated with covariates under the proportionality principle.

Plank et al., (2008) Provides the survival of a high school student. Their research goal was to combine career and technical education (CTE) with core academic courses that affect the probability of school leaving. Throughout their research, they used one of the more popular models of calculating the danger rate for youth dropout, which is the Cox Regression model. It was found that the hazard model suggests that there is a strong curvilinear correlation between the CTE and the academic course taking ratio and the possibility of lowering it for younger and14-year-olds after reaching the ninth grade.

Additionally, Eagle and Barnes (2014) used survival analysis to calculate the function of incident results and compensate for the depletion of participants using the process of Cox regression relative hazards and the Accelerated Failure Time (AFT). We are also shown that the time data obtained from smart tutors are appropriate for survival research and useful when a sample experiences turnover of participants.

Weybright et al., (2017) defined the possibility of male and female adolescents dropping out of high school based on an approach to survival research. Based on

secondary empirical evidence, they analyze the effect of predictors of drug and leisure behaviors when accounting for demographics. To estimate the survival and hazard functions they used Kaplan-Meier (KM). They also used SAS PROC PHREG to approximate the parameters of the Cox Regression model with a separate time-survival study focused on variables of age, drug usage, and recreational activity to forecast dropout.

Even though the studies reviewed are similar to our research, in the sense that they also focus on survival analysis, however, in many ways they are different from this research. For example, they are different as they used different types of censored data and methods as well. In this research, we will rather focus on the Cox log-logistic model based on PIC data.

2.2 Log-logistic Distribution

There are some studies that used log-logistic distribution in their research in survival analysis. Throughout survival analysis, logistic regression may be used where the time of the occurrence is divided into intervals. Abbott (1985) suggested that logistic regression in the study of survival evidence represents a viable solution for the proportional hazard model.

Binary logistic regression is used to forecast the relationship between independent variables and dependent variables; binary will be the dependent variable. However, it may be a concern in the study of survival after the community of incident periods in intervals. Devlina et al., (2010) explored how to interpret conditional logistic regression as a specific survival study. The concept derived from a logistic regression survival analysis is often the same as the logit and probit methods.

The distribution that used mostly in the analysis of survival and failure time data is the Weibull distribution. Steve (1983) provided that logistic distribution is an alternative

useful way in modeling the survival and failure time data to Weibull distribution. The model is fitted on GLIM and an example is given of its use with lung cancer survival data.

Mayfield (1961 and 1975) used the Mayfield method to estimate the survival rates of nests. The logistic regression of Mayfield incorporates the strengths of two commonly used approaches into one approach. Kirsten (2004) discussed how Mayfield logistic regression widely looks relevant as nests are located at different stages of the nesting process and are of concern to numerous explanatory variables affecting nest survival. This expands the conventional estimator of Mayfield to include individual covariates in a system for evaluating logistic-regression.

We will focus on the Cox log-logistic model based on PIC data in this thesis, the next section we will discuss the PIC data.

2.3 Partly Interval Censoring

Some researchers used PIC data in their research among them; Kim (2003) examined data slightly filtered by intervals utilizing the Cox model via MLE. In their analysis, they used two methods to estimate the regression parameter MLE variance–covariance matrix which generalized missing information theory and generalized profile information method. The simulation experiments show that both approaches function well with moderate-size samples and bias in terms of variance. Furthermore, the researcher explained this approach using a Diabetes data program in Denmark.

Zhao et al., (2008) used generalized log-rank research to analyze PIC failure results, as discussed by Peto and Peto (1972).They used a collection of actual data from a diabetes research and simulation tests to test the process.

Alharpy and Ibrahim (2014) used PIC to several statistical test procedures that have been suggested to aid in solving the comparison problem. The suggested test

procedures included Finkelstein (1986) who came up with a score test to compare numerous survival functions under Cox model. Moreover, Pan (2000) used multiple imputation techniques, which depend on the approximate Bayesian bootstrap to compare between two interval-censored samples with proportional and non-proportional hazards. In addition to that, Lim and Sun (2003) suggested nonparametric two sample tests for no proportional as well as proportional hazards.

Research is still ongoing with regards to PIC, however with inadequate number of studies. Peto and Peto (1972) explained that PIC data are dealing with the same observations as the extremely short interval-censored observations. Huang (1999) extracted the asymptotic properties of the nonparametric estimation for the distribution function by the usage of PIC data.

Guure et al., (2012) used PIC data that focused on Weibull distribution model with several estimates of parameter methods that were MLE, Least Square (LS) estimators of one variable over another variable to evaluate the survival estimator of these methods to estimate the parameters and to prove that the parameter bias of the estimators is valid values. Their used of MSE, bias in comparing their respective approaches dependent on simulation analysis. They noticed the MLE was best for calculating the parameter of the scale. On the other hand, the least square on the first variable was more accurate on estimating the form parameter with fairly limited samples, but the optimal approach was the least square for larger samples for the other side.

Elfaki et al., (2012) used the Weibull parametric proportional hazard model, based on the AIDS research application's based on Expectation Maximization (EM) algorithm for PIC results. Partially filtered periods may also be used in their work mainly to measure the survivability of the failure rate. They studied the HIV / AIDS

diagnosis of hemophiliacs at two Sudanese hospitals. It was also shown in their analysis that there are no variations in actual data collection between the two therapies. Furthermore, for two tests the Likelihood Ratio Test (LRT) and Score Test (ST) based on PIC are provided.

Alharpy and Ibrahim (2013) utilized parametric Weibull distribution. They found that the LRT is easier to check the parametric for PIC under Weibull distribution than the performance method. Elfaki et al., (2013) suggested a rival risk model focused on EM algorithm to approximate the parameters of the Cox's proportional hazards regression for PIC. They used two contrasting risk models which are the Complete Censoring (CC) model and the Weighting Technique (WT) model. They examined the possible correlation between treatments and anti-D in Rhesus time to search the impact of covariates on the occurrence of complications added to a collection of time results from anti-D in Sudan's Rhesus D negative pregnant women.

Yousif et al., (2016) approximated the coefficients of regression for PIC data using Bayesian approach. Simulation studied is used in order to check the model concept and it showed that the method was performed well as if it was simple to use.

Zyoud et al., (2016) used nonparametric analysis via imputation approaches to approximate the survival function based on PIC. They used several imputation techniques that included mean imputation, median imputation, conditional mode, multiple imputation and random imputation. Their proposal to approximate the survival function was introduced using R tools. They found the natural, standard, and median imputations are stronger relative to other strategies of imputation.

Wu et al. (2017) suggested the semi-parametric based on MLE approach for evaluating censored data in PIC. They used the non-mixed Cox cure rate model and the semi-parametric spline via sieve MLE to evaluate these findings. Then, a strong sieve

estimator was noted. Simulation studied is illustrated to the efficiency of the proposed system, and found that the proposed method based on MLE was satisfying.

Noora & Elfaki studied (2020) used parametric Weibull model based on simulation PIC via education data sets. They found that the Weibull model based on simulation PIC data are much more suitable compared with interval data.

As can be seen in this section, several experiments have used the PIC data dependent distribution process since its approach is flexible as used for parameter estimation. Our analysis attempts to adapt the Cox model via log-logistic distribution, which deals with data on medical PIC data. In addition to that, we are looking to examine the efficiency of the model and to check its efficacy using the suitable method.

2.4. Cox Model

Cox model is very popular in survival analysis, due to the two reasons: (a) the probability distribution of the survival times does not need any assumption, and (b) the data can be fitted well in the model regardless of the parametric model that is used (Khan and Khosa 2016). However, the model is flexible to the extent that it can be used as nonparametric, semi-parametric and fully parametric (Kalbfleisch and Prentice 2002; Lawless 2002). Therefore, for the parametric the examination of the appropriateness of the chosen distribution becomes a requirement.

Hashemian et al. (2017) studied the affection of the survived patients with colorectal cancer based on the log-logistic model and non-parametric Cox model. They found that both models had almost similarly results; but, owing to the advantages of parametric models, log-logistic may be supplemented with Cox model in surveying the recovery period of patients with colorectal cancer based on Cox Snell residuals. Since the Log-logistic and Cox models used regression coefficients remained the same, given several researchers ' preference to use Cox model in survival analysis experiments,

parametric models will forecast the long-term likelihood of the goal occurrence and have a good view of the survival time and hazard function.

Bender et al. (2005) described methods for producing survival functions from simulation experiments based on the Cox proportional hazards. They showed that the exponential, Weibull, and Gompertz distributions can be used to produce appropriate survival functions via simulation studies. The partial likelihood in the classic Cox model does not depend on the baseline hazard, the option of the distribution of the produced survival times in the Cox model via simulation studies is not given much consideration. In such situations, the distribution of the produced survival times should represent the condition of the data being considered in order to obtain sufficient simulation results.

Khan and Ababneh (2016) used survival distribution and Kaplan-Meier for survival function as descriptive approaches to approximate a sample life table distribution of survival times. There are also many regression models to approximate the relationship of continuous variables to survival times which are commonly included in the Weibull regression and the Cox models. The model of Cox refers to a broader range of distributions which is a semi-parametric concept whereas the model of Weibull regression is a completely parametric model. When a data set is applies to the Weibull model then the Cox model could be used as an alternative model for the same data set. In their results showed that the Cox model are less reliable and less effective than those for the Weibull model. It also seems the tests for the Cox model are less efficient than those for the Weibull regression model. In fact, the parametric models can provide advantages over the Cox model because of the complexity of the process.

Rezaei et al., (2014) used generalized linear models based on ordinary least square for censored data. Cox (1972) introduced a modern approach that would be helpful in situations that require censored data. Several researchers continued utilizing this approach without mentioning models of reference hazards. These approaches are defined and compared to a modern hazard proportional model. In their study also they introduced the extended exponential geometric (EEG) of Cox model. They used different models that is the semi-parametric Cox model, the model of linear regression with and without log translation and generalized linear models, e.g. for which they used of purely positive values for the event. Survival theory explores and models the period required for incidence of the incidents. The distribution of survival times is main focuses of the survival study.

There are well-known ways of calculating the distributions of unconditional life. Many important models of survival explore the interaction between survival and one or more predictors, typically referred to as covariance in the literature on survival research. However, in this thesis we will use the log-logistic Cox model based on partly interval censored data.

CHAPTER OVERVIEW

This chapter presents the estimation of the parameters of log-logistic distribution based on Cox model using maximum likelihood estimator and under PIC. The chapter also, presents the likelihood ratio test. At the end of the chapter, the simple imputation technique will be presented.

3.1 Introduction

The Cox PHRM is a technique of regression that allows estimation of survival times (or hazard functions) for the collection of exploratory variables of the model that was first developed by Cox (1972). The Cox model enables estimation of the survival function. The association between survival and a set of possible risk factors is widely used in medical research. The Cox model is used in randomized clinical trials to assess the efficacy of new treatments or interventions on survival or on the incidence of an unusual case. The Cox proportional hazard model may be called a Kaplan-Meier generalization (or product limit) estimator of a survival curve that describes both discrete and continuous risk factors. Compared to other regression methods, the benefits of Cox model for survival data lie in its consistency as well as the interpretability of its parameter estimates.

As mentioned in one of the previous chapter of this thesis, the parametric model can be fitted with the data well regardless of the model that is used. The PHRM is considered the most common part in the survival analysis due the following reasons; the first one is the survival times distribution does not require any assumption (example when the model is a semiparametric), and the second reason is that it fits the data well usually when the model is parametric. Following that, a fully parametric PHRM is required for distribution assumption (Kalbfleisch 2020, Lawless 2003). Hjort (1992)

found that more efficient estimates of parametric models than the certain condition of the Cox model. In additional to that Collett (2003) found that there are small standard errors of estimation in the parametric model with valid assumption of the distribution than would be in the absence of the distribution assumption. The most closed and common parametric under PHRM is the Weibull model. In this thesis, we will use parametric PHRM based on log-logistic distribution.

Cox (1972) introduced the PH models with covariates $z = (z_1, z_2, ..., z_p)'$ as;

$$h(t, z) = h_0(t; k) \exp(z^T \theta) \tag{3.1}$$

where the baseline hazard $h_0(t; \alpha)$ is defined by as the a vector of parameters α, and the regression coefficients by vector $\theta = (\theta_1, \theta_2, ..., \theta_p)'$.

Khan and Khosa (2016) described the hazard function for a non-negative random variable T of log-logistic distribution as:

$$h(t; k) = \frac{\alpha \rho (\rho t)^{\alpha-1}}{1 + (\lambda t)^\alpha} \tag{3.2}$$

Here $t > 0$, $k = (\alpha, \gamma, \rho)'$ and $\alpha, \lambda \& \rho$ are parameters that have the conditions $\alpha > 0, \lambda > 0$ and $\rho > 0$.

Lawless (2002) reduced (3.2) to the hazard function when λ depends on ρ through $\lambda = \rho$ and $\lambda = \rho \eta^{-1/\alpha}$ with $\eta > 0$.

Respectively the function of survivor and the function of probability density based on distribution of log-logistic given as;

$$S(t; k) = \left(1 + (\lambda t)^\alpha\right)^{-(\frac{\rho}{\lambda})^\alpha} \tag{3.3}$$

$$f(t; k) = \frac{\alpha \rho (\rho t)^{\alpha-1}}{[1 + (\lambda t)^\alpha]^{(\frac{\rho}{\lambda})^\alpha + 1}} \tag{3.4}$$

The parametric log-logistic Cox model can be obtained when $h_0(t; k)$ is full specify in

20

(3.1). Therefore, if we substituted (3.2) into (3.1) then (3.1) became as;

$$h(t;k) = \frac{\alpha e^{z'\theta/\alpha} (e^{z'\theta/\alpha} t)^{\alpha-1}}{1+(\lambda t)^{\alpha}} \qquad (3.5)$$

Thus, the Cox model based on log-logistic is closed under hazards proportionality. However, the semi parametric of Cox PH model when we left arbitrary of the baseline hazard function in (3.1) that can be denoted by $h_0(t)$.

3.2 Estimations the Model Parameters

Let the censored random sample consisted of the data (t_i, δ_i, z_i), $i = 1, 2, ..., n$, where z_i is defined in (3.1) and the failure or censoring time represent by t_i according to $\delta_i = 1$ or $\delta_i = 0$. The likelihood function for is given as;

$$L(k,\theta) = \prod_{i=1}^{n} [f(t_i, k, \theta)]^{\delta_i} [S(t_i, k, \theta)]^{1-\delta_i} \qquad (3.6)$$

By substitute equation (3.3) and (3.4) into (3.6) equation (3.6) become;

$$L(k,\theta) = \prod_{i=1}^{n} \left[\frac{\alpha \rho (\rho t)^{\alpha-1}}{[1+(\lambda t)^{\alpha}]^{(\frac{\rho}{\lambda})^{\alpha}+1}} \right]^{\delta_i} \left[\left(1+(\lambda t)^{\alpha}\right)^{-(\frac{\rho}{\lambda})^{\alpha}} \right]^{1-\delta_i} \qquad (3.7)$$

Following Khan and Khosa (2016), assume $a_i = (\lambda t_i)^{\alpha}$, $b_i = e^{z_i'\theta}$, $m = \sum_{i=1}^{n} \delta_i$, and $\beta = (k', \theta')'$ then the function of log likelihood of (3.7) for model (3.5) given as;

$$
\begin{aligned}
L(\beta) = & m \log \alpha + m\alpha \log \rho + (\alpha-1) \sum_{i=1}^{n} \delta_i \log t_i - \sum_{i=1}^{n} \delta_i \log(a_i+1) \\
& + \sum_{i=1}^{n} \delta_i \log b_i - (\frac{\rho}{\gamma})^{\alpha} \sum_{i=1}^{n} \delta_i \log(a_i+1)
\end{aligned}
\qquad (3.8)
$$

The method of MLE will be used to estimate the parameters $(\theta, \alpha, \rho, \lambda)$.

Solving the parameters that maximize equation (3.8) will yield the parameters for the PHRM model with log-logistics, which are obtained by solving the following partial derivatives

$$\frac{\partial L(\beta)}{\partial \alpha} = 0, \ \frac{\partial L(\beta)}{\partial \rho} = 0, \ \frac{\partial L(\beta)}{\partial \lambda} = 0, \ \frac{\partial L(\beta)}{\partial \theta_i} = 0$$

The n-th derivative for α, ρ, γ and θ of the equation (3.8) are

$$\frac{\partial L(\beta)}{\partial \alpha} = \frac{m}{\alpha} + m \log \rho + \sum_{i=1}^{n} \delta_i \log t_i - \frac{1}{\alpha} \sum_{i=1}^{n} \delta_i a_i c_i - \left(\frac{\rho}{\lambda}\right)^{\alpha} \left(\frac{1}{\alpha}\right) \sum_{i=1}^{n} \delta_i a_i c_i$$
$$- \left(\frac{\rho}{\lambda}\right)^{\alpha} \log\left(\frac{\rho}{\lambda}\right) \sum_{i=1}^{n} b_i \log(a_i + 1) \tag{3.9}$$

Since $a_i = (\lambda t_i)^{\alpha}$, then; $\log a_i = \alpha \log(\lambda t_i) \Rightarrow \log(\lambda t_i) = \frac{\log a_i}{\alpha}$ and

$(\lambda t_i)^{\alpha} \log(\lambda t_i) = \frac{a_i \log a_i}{\alpha}$.

Therefore, $\qquad \frac{\partial a_i}{\partial \alpha} = \frac{\partial}{\partial \alpha}(\lambda t_i)^{\alpha} = (\lambda t_i)^{\alpha} \log(\lambda t_i) = \frac{a_i \log a_i}{\alpha} \tag{3.10}$

Based on (3.10) we can obtain;

$$\frac{\partial \log a_i}{\partial \alpha} = \frac{1}{a_i} \frac{\partial}{\partial \alpha}(a_i) = \frac{1}{a_i}\left(\frac{a_i \log a_i}{\alpha}\right) = \frac{\log a_i}{\alpha}$$

$$(3.11)$$

$$\frac{\partial \log(1 + a_i)}{\partial \alpha} = \frac{1}{a_i + 1} \frac{\partial}{\partial \alpha}(a_i + 1) = \frac{1}{a_i + 1} \frac{\partial}{\partial \alpha}(a_i) = \frac{1}{a_i + 1} \frac{a_i \log a_i}{\alpha}$$
$$= \frac{a_i \log a_i}{\alpha(a_i + 1)} = \frac{a_i c_i}{\alpha} \tag{3.12}$$

Where $c_i = \dfrac{\log a_i}{(a_i + 1)}$.

$$\frac{\partial[a_i \log(a_i + 1)]}{\partial \alpha} = a_i \frac{\partial}{\partial \alpha} \log(a_i + 1) + \log(a_i + 1) \frac{\partial}{\partial \alpha}(a_i) \tag{3.13}$$

By substitute (3.10) and (3.11) into (3.13), equation (3.13) became;

$$\frac{\partial[a_i \log(a_i)]}{\partial \alpha} = a_i \frac{\log a_i}{\alpha} + \log(a_i) \frac{a_i \log a_i}{\alpha} = \frac{a_i \log a_i}{\alpha}[1 + \log a_i] \tag{3.14}$$

In equation (3.12) we have $c_i = \dfrac{\log a_i}{(a_i + 1)}$, then;

$$\frac{\partial}{\partial \alpha}[c_i] = \frac{\partial}{\partial \alpha}[\frac{\log a_i}{(a_i+1)}] = \frac{(a_i+1)[\frac{\partial}{\partial \alpha}\log a_i] - \log a_i[\frac{\partial}{\partial \alpha}(a_i+1)]}{(a_i+1)^2} \qquad (3.15)$$

By substitute (3.10) and (3.11) into (3.15), equation (3.15) will be;

$$\frac{\partial}{\partial \alpha}[c_i] = \frac{\partial}{\partial \alpha}[\frac{\log a_i}{(a_i+1)}] = \frac{(a_i+1)[\frac{\log a_i}{\alpha}] - \log a_i[\frac{a_i \log a_i}{\alpha}]}{(a_i+1)^2} = \frac{\log a_i}{\alpha(a_i+1)}\left(\frac{(a_i+1)}{(a_i+1)} - \frac{a_i \log a_i}{(a_i+1)}\right)$$

$$= \frac{\log a_i}{\alpha(a_i+1)}\left(1 - \frac{a_i \log a_i}{(a_i+1)}\right) = \frac{c_i(1-a_ic_i)}{\alpha} \qquad (3.16)$$

Also,

$$\frac{\partial}{\partial \alpha}[a_ic_i] = a_i\frac{\partial}{\partial \alpha}[c_i] + c_i\frac{\partial}{\partial \alpha}[a_i] = a_i\frac{c_i(1-a_ic_i)}{\alpha} + c_i\frac{a_i \log a_i}{\alpha} = \frac{c_ia_i(1-_i a_ic_i+\log a_i)}{\alpha}$$

$$= \frac{c_ia_i(1-a_i[\frac{\log a_i}{a_i+1}]+\log a_i)}{\alpha} = \frac{c_ia_i(\frac{1+a_i-a_i \log a_i+\log a_i+a_i \log a_i}{a_i+1}]}{\alpha} = \frac{c_ia_i(1+\frac{\log a_i}{a_i+1}]}{\alpha} = \frac{c_ia_i(1+c_i]}{\alpha}$$

$$(3.17)$$

By assuming $d_i = \frac{a_i}{(a_i+1)}$, then;

$$\frac{\partial}{\partial k}[d_i] = \frac{\partial}{\partial k}[\frac{b_i}{(1+b_i)}] = \frac{(a_i+1)[\frac{\partial}{\partial \alpha}a_i] - a_i[\frac{\partial}{\partial \alpha}(a_i+1)]}{(a_i+1)^2} = \frac{(a_i+1)[\frac{a_i \log a_i}{\alpha}] - a_i[\frac{a_i \log a_i}{\alpha}]}{(a_i+1)^2}$$

$$= \frac{a_i \log a_i}{\alpha(a_i+1)}[1 - \frac{a_i}{(a_i+1)}] = \frac{a_i \log a_i}{\alpha(a_i+1)} * \frac{1}{a_i+1} = \frac{\log a_i}{\alpha(a_i+1)} * \frac{a_i}{a_i+1} = \frac{c_id_i}{\alpha} \quad (3.18)$$

$$\frac{\partial}{\partial \alpha}[\log(1-d_i)] = \frac{\partial}{\partial \alpha}[\log(1-\frac{a_i}{(a_i+1)}] = \frac{\partial}{\partial \alpha}[\log(\frac{1}{(a_i+1)})] = -\frac{\partial}{\partial \alpha}[\log(a_i+1)]$$

$$= -\frac{1}{a_i+1}\frac{\partial}{\partial \alpha}[(a_i+1)] = -\frac{1}{a_i+1}\frac{\partial}{\partial \alpha}[(a_i)] = -\frac{1}{a_i+1}\frac{a_i \log a_i}{\alpha} = -\frac{a_ic_i}{\alpha} (3.19)$$

Since in equation (3.7) we have $a_i = (\lambda t_i)^\alpha$, then;

$$\frac{\partial}{\partial \lambda}[a_i] = \frac{\partial}{\partial \lambda}[(\lambda t_i)^\alpha] = \alpha\lambda^{\alpha-1}t_i^\alpha = \alpha\lambda^{-1}(\lambda^\alpha t_i^\alpha) = \frac{\alpha}{\lambda}(\lambda t_i)^\alpha = \frac{\alpha}{\lambda}a_i \qquad (3.20)$$

As we assuming in (3.18) that is $d_i = \frac{a_i}{(a_i+1)}$, and based on (3.20) we have;

$$\frac{\partial}{\partial\lambda}[d_i] = \frac{\partial}{\partial\lambda}[\tfrac{a_i}{(a_i+1)}] = \frac{(a_i+1)[\frac{\partial}{\partial\lambda}a_i] - a_i[\frac{\partial}{\partial\lambda}(a_i+1)]}{(a_i+1)^2}$$

$$= \frac{(a_i+1)[\frac{\alpha}{\lambda}a_i] - a_i[\frac{\alpha}{\lambda}(a_i)]}{(a_i+1)^2} = \frac{\alpha}{\lambda}\frac{a_i}{(a_i+1)}[\frac{(a_i+1)-a_i}{(a_i+1)}] = \frac{\alpha}{\lambda}d_i(1-d_i) \quad (3.21)$$

Similarly, to (3.19) we can also have;

$$\frac{\partial}{\partial\lambda}[\log(1-d_i)] = -\frac{\partial}{\partial\alpha}[\log(1+b_i)] = -\frac{1}{(a_i+1)}\frac{\partial}{\partial\alpha}[(a_i+1)]$$

$$= -\frac{1}{(a_i+1)}\frac{\alpha}{\lambda}a_i = -\frac{\alpha}{\lambda}(\frac{a_i}{a_i+1}) = -\frac{\alpha}{\lambda}d_i \qquad (3.22)$$

Now, similarly to (3.9), we can find the first and second derivatives for α, ρ, λ and

θ using equation (3.8) and (3.10) to (3.22) as follows:

$$\frac{\partial L(\beta)}{\partial\rho} = \frac{m\alpha}{\rho} - \left(\frac{\rho}{\lambda}\right)^\alpha\left(\frac{\alpha}{\rho}\right)\sum_{i=1}^{n}b_i\log(a_i+1) \qquad (3.23)$$

$$\frac{\partial L(\beta)}{\partial\lambda} = -\left(\frac{\alpha}{\lambda}\right)\sum_{i=1}^{n}\delta_i d_i - \left(\frac{\rho}{\lambda}\right)^\alpha\left(\frac{\alpha}{\lambda}\right)\sum_{i=1}^{n}b_i d_i - \left(\frac{\rho}{\lambda}\right)^\alpha\left(\frac{\alpha}{\lambda}\right)\sum_{i=1}^{n}b_i\log(1-d_i) \quad (3.24)$$

$$\frac{\partial L(\beta)}{\partial\theta_j} = \sum_{i=1}^{n}\delta_i z_{ij} - \left(\frac{\rho}{\lambda}\right)^\alpha\sum_{i=1}^{n}b_i\log(a_i+1)z_{ij} \text{ ; For } j=1,2,...,p \quad (3.25)$$

The second derivatives given as;

$$\frac{\partial^2 L(\beta)}{\partial\rho^2} = \frac{\partial}{\partial\rho}[\frac{\partial L(\beta)}{\partial\rho}] = \frac{\partial}{\partial\rho}[\frac{m\alpha}{\rho} - \left(\frac{\rho}{\lambda}\right)^\alpha\left(\frac{\alpha}{\rho}\right)\sum_{i=1}^{n}b_i\log(a_i+1)]$$

$$= -\frac{m\alpha}{\rho^2} - \frac{\alpha(\alpha-1)}{\rho^2}\left(\frac{\rho}{\lambda}\right)^\alpha\sum_{i=1}^{n}b_i\log(a_i+1)] \qquad (3.26)$$

$$\frac{\partial^2 L(\beta)}{\partial\theta_j^2} = \frac{\partial}{\partial\theta_j}[\frac{\partial L(\beta)}{\partial\theta_j}] = \frac{\partial}{\partial\theta_j}[\sum_{i=1}^{n}\delta_i z_{ij} - \left(\frac{\rho}{\lambda}\right)^\alpha\sum_{i=1}^{n}b_i\log(a_i+1)z_{ij}]$$

$$= -\left(\frac{\rho}{\lambda}\right)^\alpha\sum_{i=1}^{n}b_i\log(a_i+1)z_{ij}^2 \qquad (3.27)$$

Based on (3.10), then the second derivative of α is given by;

$$\frac{\partial}{\partial\alpha}(\frac{\partial L(\beta)}{\partial\alpha}) = \frac{\partial}{\partial\alpha}[\frac{m}{\alpha} + m\log\rho + \sum_{i=1}^{n}\delta_i\log t_i - \frac{1}{\alpha}\sum_{i=1}^{n}\delta_i a_i c_i - \left(\frac{\rho}{\lambda}\right)^{\alpha}\left(\frac{1}{\alpha}\right)\sum_{i=1}^{n}\delta_i a_i c_i$$

$$-\left(\frac{\rho}{\lambda}\right)^{\alpha}\log\left(\frac{\rho}{\lambda}\right)\sum_{i=1}^{n}b_i\log(a_i+1)]$$

$$= -\frac{m}{\alpha^2} + \frac{1}{\alpha^2}\sum_{i=1}^{n}\delta_i a_i c_i - \frac{1}{\alpha^2}\sum_{i=1}^{n}\delta_i a_i c_i(1+c_i) - \left(\frac{\rho}{\lambda}\right)^{\alpha}\log\left(\frac{\rho}{\lambda}\right)\left(\frac{1}{\alpha}\right)\sum_{i=1}^{n}\delta_i a_i c_i$$

$$+\left(\frac{\rho}{\lambda}\right)^{\alpha}(\frac{1}{\alpha^2})\sum_{i=1}^{n}\delta_i a_i c_i - \left(\frac{\rho}{\lambda}\right)^{\alpha}(\frac{1}{\alpha^2})\sum_{i=1}^{n}\delta_i a_i c_i(1+c_i)$$

$$-\left(\frac{\rho}{\lambda}\right)^{\alpha}\log\left(\frac{\rho}{\lambda}\right)(\frac{1}{\alpha})\sum_{i=1}^{n}\delta_i a_i c_i - \left(\frac{\rho}{\lambda}\right)^{\alpha}[\log\left(\frac{\rho}{\lambda}\right)]^2\sum_{i=1}^{n}b_i\log(a_i+1)$$

$$= -\frac{m}{\alpha^2} - \frac{1}{\alpha^2}\sum_{i=1}^{n}\delta_i a_i c_i^2 - \left(\frac{\rho}{\lambda}\right)^{\alpha}(\frac{1}{\alpha^2})\sum_{i=1}^{n}\delta_i a_i c_i^2 - \left(\frac{2}{\alpha}\right)\left(\frac{\rho}{\lambda}\right)^{\alpha}\log\left(\frac{\rho}{\lambda}\right)\sum_{i=1}^{n}\delta_i a_i c_i$$

$$-\left(\frac{\rho}{\lambda}\right)^{\alpha}[\log\left(\frac{\rho}{\lambda}\right)]^2\sum_{i=1}^{n}b_i\log(a_i+1) \tag{3.28}$$

The second derivative of λ based on (3.24) is given by;

$$\frac{\partial}{\partial\lambda}[\frac{\partial L(\beta)}{\partial\lambda}] = \frac{\partial}{\partial\lambda}[-\left(\frac{\alpha}{\lambda}\right)\sum_{i=1}^{n}\delta_i d_i - \left(\frac{\rho}{\lambda}\right)^{\alpha}\left(\frac{\alpha}{\lambda}\right)\sum_{i=1}^{n}b_i d_i - \left(\frac{\rho}{\lambda}\right)^{\alpha}\left(\frac{\alpha}{\lambda}\right)\sum_{i=1}^{n}b_i\log(1-d_i)]$$

$$= \left(\frac{\alpha}{\lambda^2}\right)\sum_{i=1}^{n}\delta_i d_i - \left(\frac{\alpha}{\lambda}\right)^2\sum_{i=1}^{n}\delta_i d_i(1-d_i) + \alpha\rho^{\alpha}\left(\frac{\alpha+1}{\lambda^{\alpha+2}}\right)\sum_{i=1}^{n}b_i d_i$$

$$-\left(\frac{\alpha}{\lambda}\right)^2\left(\frac{\rho}{\lambda}\right)^{\alpha}\sum_{i=1}^{n}b_i d_i(1-d_i) + \alpha\rho^{\alpha}\left(\frac{\alpha+1}{\lambda^{\alpha+2}}\right)\sum_{i=1}^{n}b_i\log(1-d_i) + \left(\frac{\alpha}{\lambda}\right)^2\left(\frac{\rho}{\lambda}\right)^{\alpha}\sum_{i=1}^{n}b_i d_i$$

$$= \left(\frac{\alpha}{\lambda^2}\right)\sum_{i=1}^{n}\delta_i d_i - \left(\frac{\alpha}{\lambda}\right)^2\sum_{i=1}^{n}\delta_i d_i(1-d_i)$$

$$+\left(\frac{\alpha(\alpha+1)}{\lambda^{\alpha+2}}\right)\left(\frac{\rho}{\lambda}\right)^{\alpha}\sum_{i=1}^{n}b_i[d_i + \log(1-d_i)] + \left(\frac{\alpha}{\lambda}\right)^2\left(\frac{\rho}{\lambda}\right)^{\alpha}\sum_{i=1}^{n}b_i d_i^2 \tag{3.29}$$

The inverse of the information matrix, evaluated at $\hat{\rho}, \hat{\lambda}, \hat{\alpha}$ and $\hat{\theta}$, that is $I^{-1}(\rho)$, $I^{-1}(\lambda), I^{-1}(\alpha)$ and $I^{-1}(\theta)$, is the estimated covariance matrix of $\hat{\beta}$ and $\hat{\theta}$ respectively. Numerical methods such as Newton Rapson will be used to solve the system of the equations.

3.3 Likelihood Ratio Test

A likelihood ratio may be conceived of as a combination of two statistical representations of the evidence being studied. That model has a likelihood density for the measurements and several unknown parameters that can be calculated from the results. In a broad variety of specific cases, the density is the normal multivariate distribution, and the parameters are the results under various circumstances, along with the error variance. Despite of the restrictions on the state implies, the two models vary. For example, a model in which two ways of situation vary can be contrasted with a model in which the measures are similar. The match of each model and the observations may be estimated by measuring the data likelihood, provided the best estimates of the model parameters: the more probable the results are, the better the result. In this case, the better parameter estimates are those that maximize the likelihood of results and are referred to as maximum-likelihood estimates. The ratio of two such probabilities is the maximum likelihood ratio; it gives an overview of the relative correlation of the two equations to the results observed.

Based on the details mentioned above and for a large sample size, the chi-square is an approximated distribution of the LRT. The degrees of freedom of this distribution is equal to the difference in the number of coefficients in the two models. This test is defined as:

$$LR = -2[L_{subset} - L_{full}] = -2\ln\left(\frac{l_{subset}}{l_{full}}\right) \qquad (3.30)$$

The -2 in LR equation adjusts the test in a way that the chi-square distribution can be used to approximate the distribution of the test.

3.4 Imputation

Imputation approaches are sometimes used to transform the problem of analyzing data analysis. In this thesis, we will modify the data based on imputation technique to be PIC or interval censored data.

The motivation behind that is that the imputation process is very simple and there are numerous methods to deal with the data. There are different imputations methods such as simple and multiple imputations methods. In this study we will use a simple imputation techniques.

3.4.1 Simple Imputation

The Simple imputation technique is one of the most common techniques used to treat the missing data. Because the simple imputation is conjectural and appealing , it is often used in the simple cases of observations. As mentioned by Zyoud, et al. (2016), the simple imputation methods such as;

1. The right limit of the interval R_i which represents right point.

2. The left limit of the interval L_i that represents the left point.

3. The midpoint of the interval $(R_i + L_i)/2$ for which represents the midpoint.

CHAPTER 4: RESULTS AND ANALYSIS

4.1 Brest Cancer Data

Breast cancer is Qatar's most common disease, as 31% of women have breast cancer. The risk of breast cancer development among women in the over all population is 56 per 100,000 (source: Qatar Cancer Register). However, breast cancer can affect both women and men. It is important to remember that many breast cancer cases can be treatable if patients are been diagnosed and checked up early, they will recover.

The proposed methods in this thesis were applied to the data set taken from Hamad Medical Corporation (HMC). The data was collected from 2/1/2016 to 1/19/2020 and it contains 24 variables. The First case was on 2/1/2016 and the last case was on 1/19/2020.The data consist of 1008 patients, 770 treated by surgery, 557 patients treated by chemotherapy, 555 treated by the hormone and 533 treated by radiotherapy (RT). In this study, the patients were observed at Al-mal clinic visits, the event of interest was the time of the first occurrence of breast retraction, the actual dates of the event were recorded exactly if available otherwise the interval of events were noted. The main objective here is to compare the cosmetic effects of each treatment alone on women with early breast cancer. 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 in this section. In later section, the simulation study will be highlighted.

As mentioned in chapter 3 that the log-logistic model can be used based on the Cox model by allowing α (scale parameter) to differ between the treatments, or more generally by introducing covariates that affects α but not β (shape parameter) by modeling $\log(\alpha)$ as a linear function of covariates.

Table 1: Log-logistic Model for Breast Cancer Data

Treatment	Parameter	Coefficient	CI of 95%	SE
	Shape	3.751545	(3.55575, 3.958124)	0.102600
	Scale	656.4018	(619.453, 695.5546)	19.40320
Hormone	Coefficient	0.313580	(0.25079, 0.376375)	0.032039
RT	Coefficient	0.142124	(0.07885, 0.205398)	0.032283
Chemotherapy	Coefficient	0.061392	(0.02986, 0.119798)	0.029800

The Likelihood Ratio Test (LRT) with p-value (-7254.742; 0.000045), together with test whether the model coefficients/parameters as a group equal to zero. In this case, larger LRT values and small p-values indicate a greater confidence in rejecting the null hypothesis. Therefore, based on LRT we can conclude that the model is fit well and the coefficient, shape and scale parameters cannot be equal zero as shown in Table 1. Also, since the values of standard error (SE) are less than 0.05, therefore they are statistically significant in nature for all the treatments that are; hormones, radiotherapy (RT) and chemotherapy and all they affect survival.

This means that the chances of survival depend on RT, hormones and chemotherapy with the impacts as shown in the result of data points in the Table 1 or the Figures 2, 3 and 4 also accurately provides the difference between the impacts of different treatments.

From Figure 2, we can conclude that the chances of survival of a breast cancer patient increases or we can say that the likelihood of living for more number of days of the patient increases after radiotherapy. As per the expectations the likelihood of living more after surgery increases first and then decreases as the value of chances of survival decreases but the results shows that it is still always better to have radiotherapy.

Based on the value of coefficient and SE of hormone it accurately shows that there is a significant correlation between both failure times but while comparing among the three different treatment methods it can be classified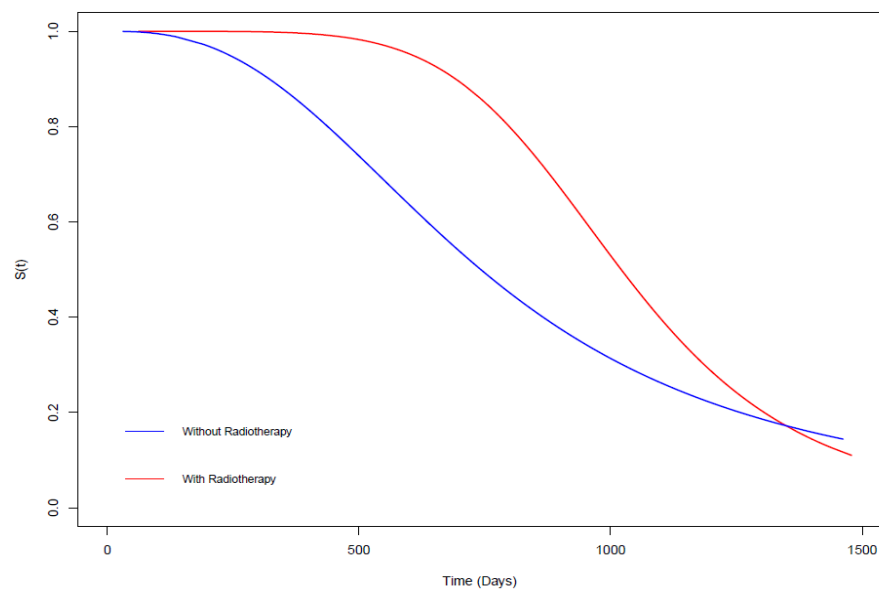 as the second lowest. The chances of survival of breast cancer patient increases based on hormone treatment as shown in Figure 3 or we can say that the likelihood of living for more days increases after hormone. As per the expectations the likelihood of living more after hormone increases first and then decreases even though the chances of survival decreases, but the results shows that it is still always better to have hormone.

Similarly the value of LRT and it's p-value for RT treatment shows that there is a significant correlation between both the data sets but while comparing among the four different treatment methods it is second highest.
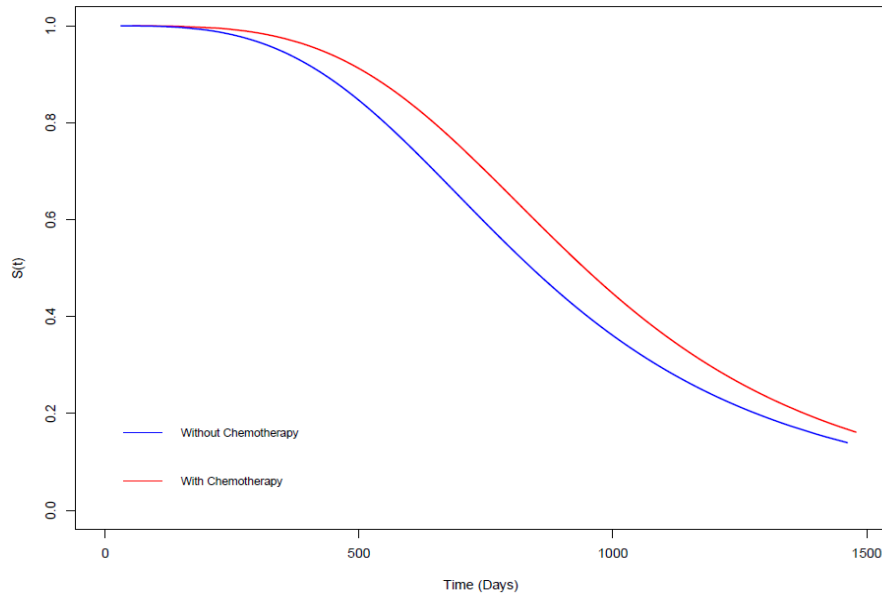

Figure 2: The function of **s**urvival estimated by log-logistic model based on radiotherapy treatment
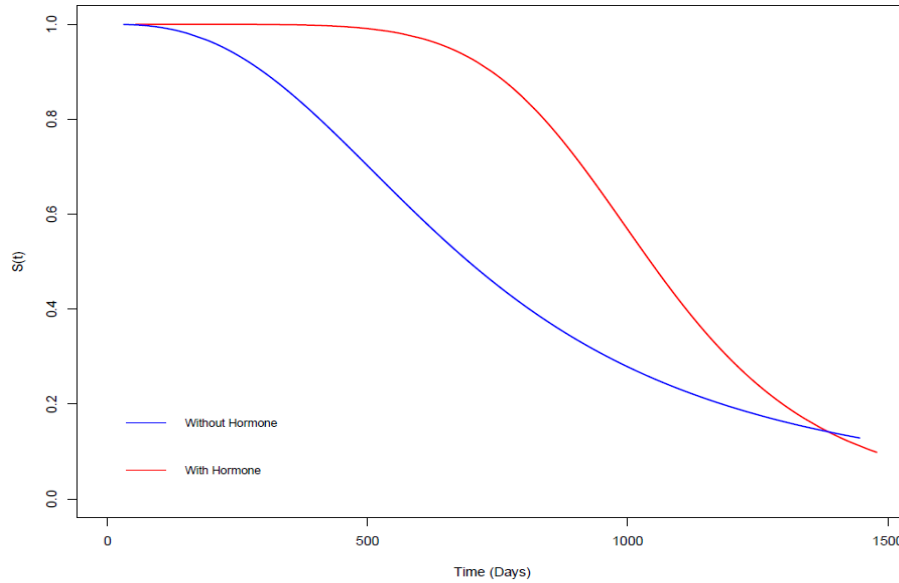
Figure 3: The function of survival estimated by log-logistic model based on hormone treatment.

Based on Figure 3, we can conclude that the chances of survival of a breast cancer patient increases or we can say that the likelihood of the patient to live for more days increases after hormone. As per the expectations the likelihood of living more after surgery increases first and then decreases later. Even the value of chances of survival decrease but the results show that it is still always better to have hormone treatment.

As per the data, hormone treatment has the highest shape parameter i.e. maximum dispersion which directly states that the chances of survival increases significantly when compared with a patient who haven't gone through the hormonal treatment while fighting with breast cancer. However, based on SE it shows that there is a significant correlation between both data sets but while comparing among the three different treatment methods it shows the highest.

Figure 4: The function of survival estimated by log-logistic model based on chemotherapy treatment.

Similarly based on Figure 4, we can conclude that the chances of survival of a breast cancer patient increases or we can say that the likelihood of living for more number of days of the patient increases after surgery. As per the expectations the likelihood of living more after surgery increases first and then decreases but the results shows that it is still always better to have chemotherapy.

As per the data chemotherapy has the lowest value of shape parameter i.e. lowest dispersion which directly states that the chances of survival doesn't increase significantly when compared with a patient who has not gone through the Chemotherapy Treatment while fighting with breast cancer. In additional to that and based on the value of LRT and p-value we can conclude that there is no significant correlation between both the data sets but while comparing among the four different treatment methods it is lowest among four of them.

In summary, the result in this section suggested that it useful to use all three treatment that is the hormone as consider highly used then followed by RT and lastly chemotherapy for breast cancer patients.

4.2   Simulated Data

Simulation studies are computational tests involving the production of results through pseudo-random sampling from established probability distributions. They are indispensable resources for statistical analysis, especially in the appraisal of modern approaches and the evaluation of alternative approaches. Simulation experiments are commonly used in the pages of statistics in medicine; however our experience is that some statisticians lack the knowledge required to perform a simulation test with conviction, whereas others are over-confident and thus neglect to think carefully about the design and the recorded outcomes.

Simulation experiments are used to produce scientific data on the efficiency of statistical methods under some situations, as opposed to more common analytical (algebraic) tests, which can include other situations. It is not often feasible or complicated to achieve empirical tests. Simulation experiments are self-contained where procedures make incorrect decisions or results become messy so they may determine the durability of procedures under these circumstances. It is not necessarily true for observational findings, as the conclusions will only be replicated where the details are extracted from a particular model.

We carried out a simulation study based on the real breast cancer data (that highlighted in this thesis) to examine the influence of the log-logistic model and to compare the covariates in the data sets. The Normal distribution is used to generate the simulation data since we find that the normal one is more reasonable as it is based on real data (graphics of the data are relatively similar to curves of normal) compared to

some other distributions such as Weibull and log-logistic as displayed in Figures 5, 6 and 7. Moreover, the Akaike's information criterion (AIC) was found to be 14610.14 a normal distribution, 15059.36 for log-logistic and 14672.15 for Weibull for which indicated and confirmed the mentioned results in the Figures (5, 6 &7). In addition to that, sample we used is 20000 times for each treatments.

Mean and standard deviation of 0.3135801 & 0.032038 and 0.3418023 & 0.03492 are used to generate the data for treatment with and without hormone respectively, based on (0%, 25%, 50%, and 75%) as percentage of exact observation for the partly interval censored (PIC) data.

In each simulation data we obtained the function of survival for the two groups of each treatment that are based on the exact observation compared to the one estimated by imputations methods that is; mean, midpoint and left point via our model.

The estimates approach in Figures 8, 9, 10, 11, 12, 13, 14, 15 and 16, shows that the estimate of the results based on exact observation, is similar compared with the one obtained by means and midpoint and left point. However, the group who used surgery treatment are more likely to survive longer compared to those without surgery, suggesting that our mean and midpoint methods provide an acceptable approximate estimation especially when more exact observations (25%, 50% and 75%) are used in the data compared to the less exact observations (0%) as displayed in Figures 17, 18 and 19 (left point).

Table 2 showed the results obtained by our model based on mean point imputation for surgery treatment with different percentages of exact and interval censored data. It showed significant concerning with respect to LRT and their p-value. These results indicate that for more exact observation in the data the results are better (as high value of AIC=286714.4 when 100% exact compare to AIC= 286904.3 for 0%

exact). Moreover, the chances of survival increase significantly when patient use surgery treatment compared with a patient who haven't gone through the surgery treatment while fighting with breast cancer.
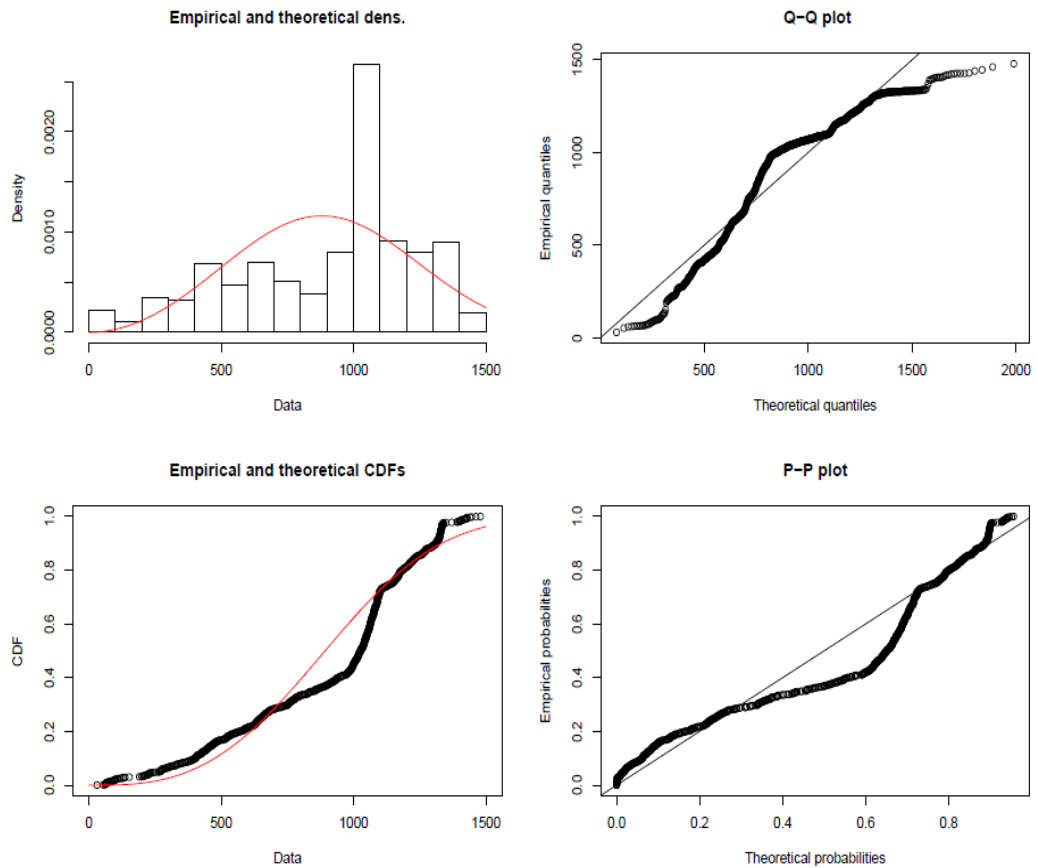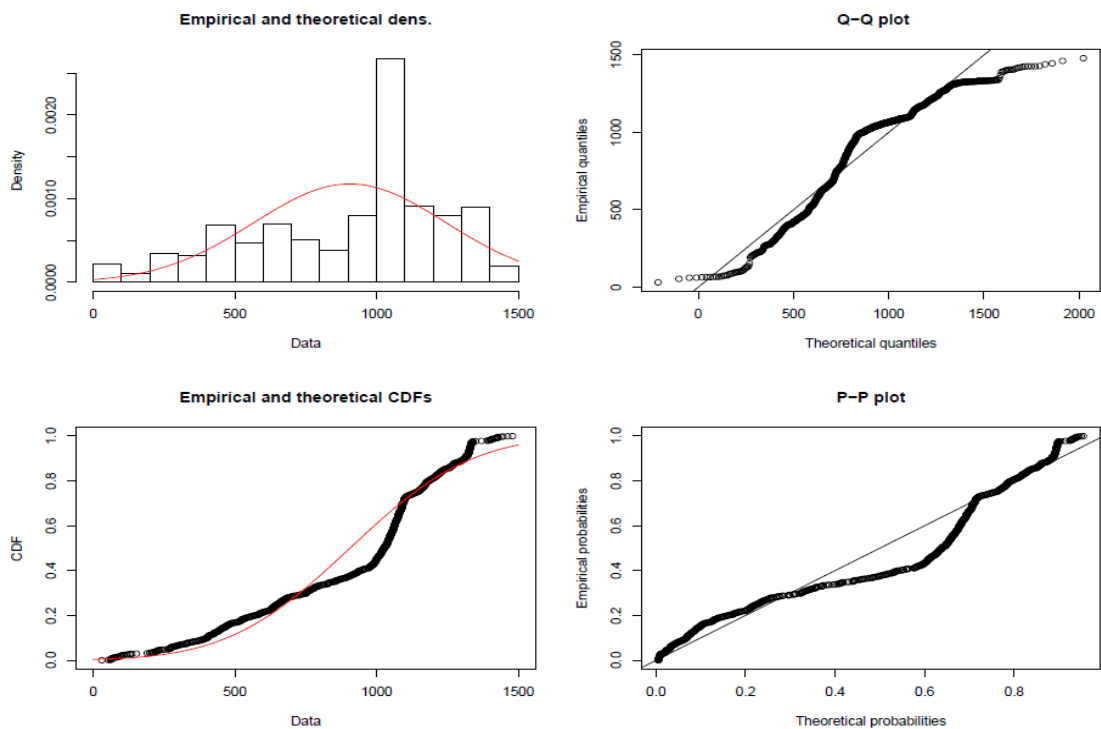

Figure 5: The empirical quintiles, density function, empirical probabilities and cumulative density function obtained by Weibull Distribution.

Table 2: Result from simulation data for surgery based mean point via Log-logistic Model.

Exact	Parameter	Estimate	CI of 95%	SE	LRT*(P-value)
0%	Shape	3.720395	(3.6763, 3.765)	0.02262	-143354.2 (2e-16)
	Scale	730.2488	(723.44,737.12)	3.48957	
	Coefficient	0.237332	(0.2246,0.2501)	0.00651	
25%	Shape	3.717644	(3.6736, 3.762)	0.02260	-143369.8(2e-16)
	Scale	730.2005	(723.39,737.08)	3.49123	
	Coefficient	0.237427	(0.2247,0.2502)	0.00651	
50%	Shape	3.711552	(3.6676,3.7561)	0.02257	-143408.8(2e-16)
	Scale	730.4620	(723.64,737.34)	3.49568	
	Coefficient	0.237076	(0.2243,0.2498)	0.00652	
75%	Shape	3.707289	(3.6633,3.7517)	0.02255	-143432.6 (2e-16)
	Scale	730.5007	(723.68,737.39)	3.49868	
	Coefficient	0.237042	(0.2242,0.2498)	0.00652	
100%	Shape	3.703924	(3.6600, 3.748)	0.02253	-143449.1(2e-16)
	Scale	730.2192	(723.39,737.11)	3.49956	
	Coefficient	0.237508	(0.2247,0.2503)	0.00653	

Figure 6: The empirical quintiles, density function, empirical probabilities and cumulative density function obtained by Normal Distribution.
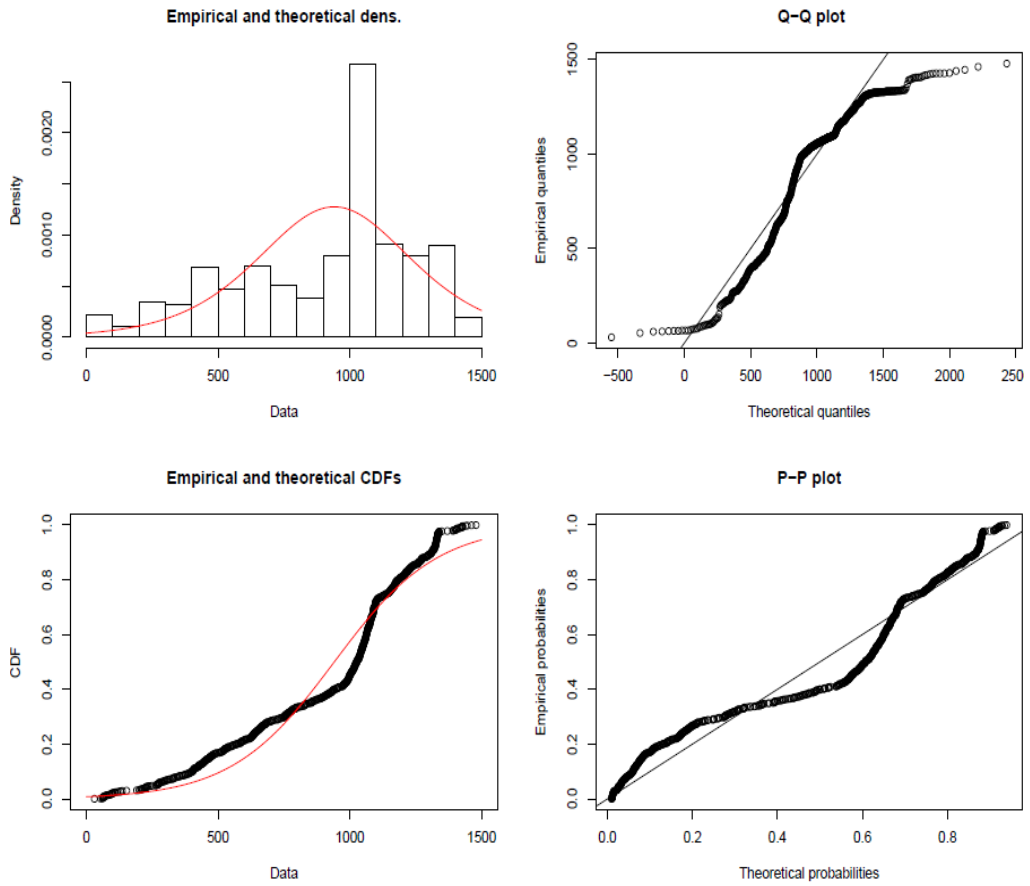
Figure 7: The empirical quantiles, density function, cumulative density function and Empirical probabilities obtained by Log-logistic Distribution.
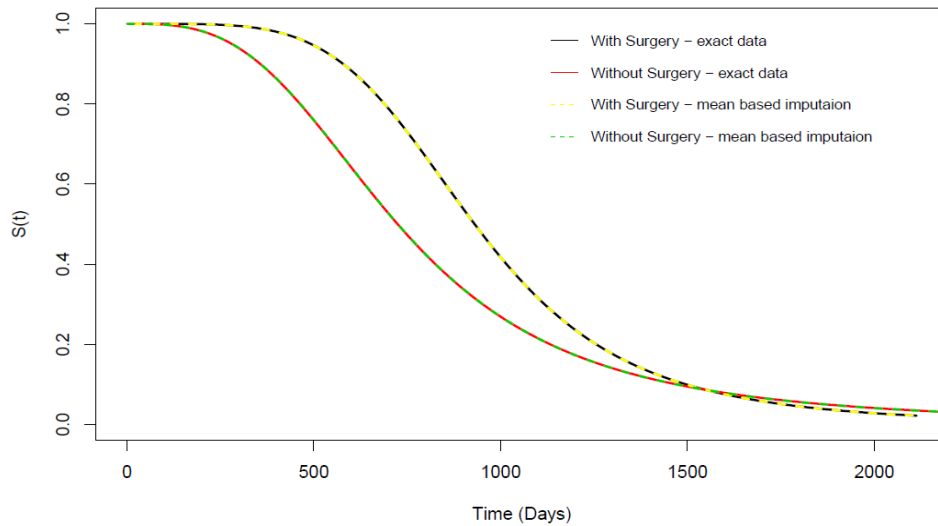
Figure 8: The function of survival estimated by mean imputation for 75% exact data based on surgery treatment.
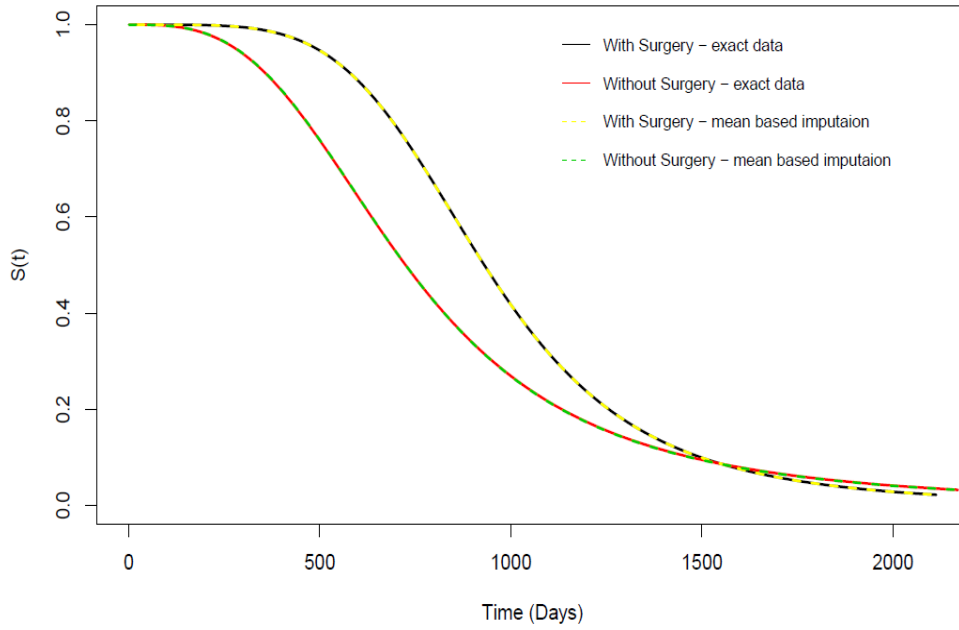
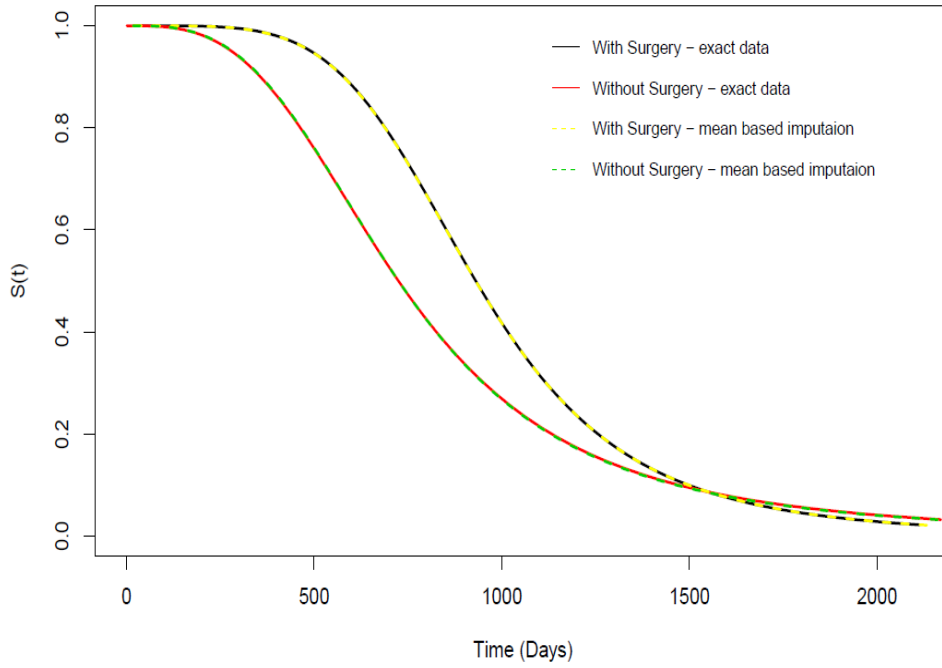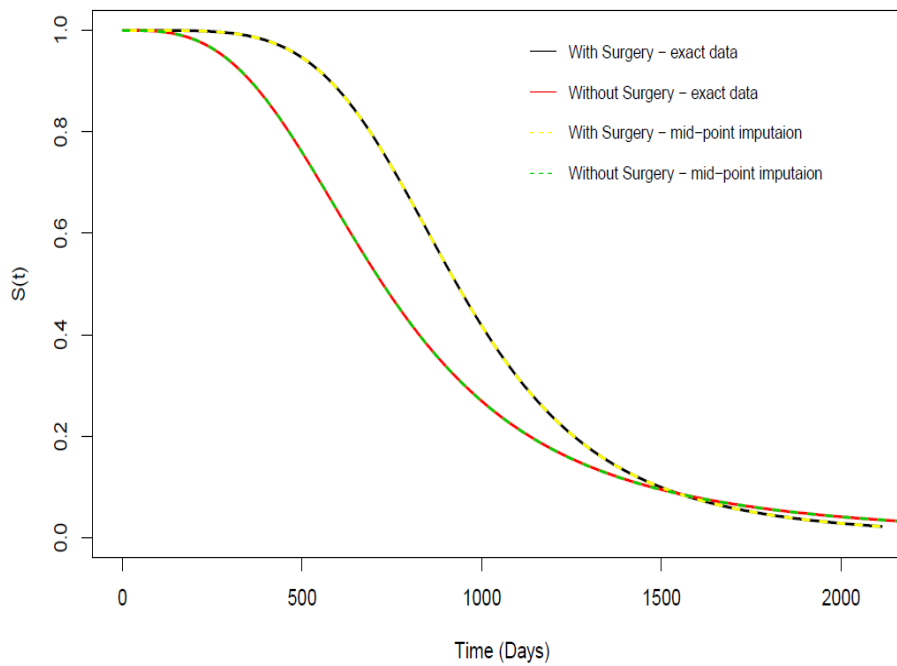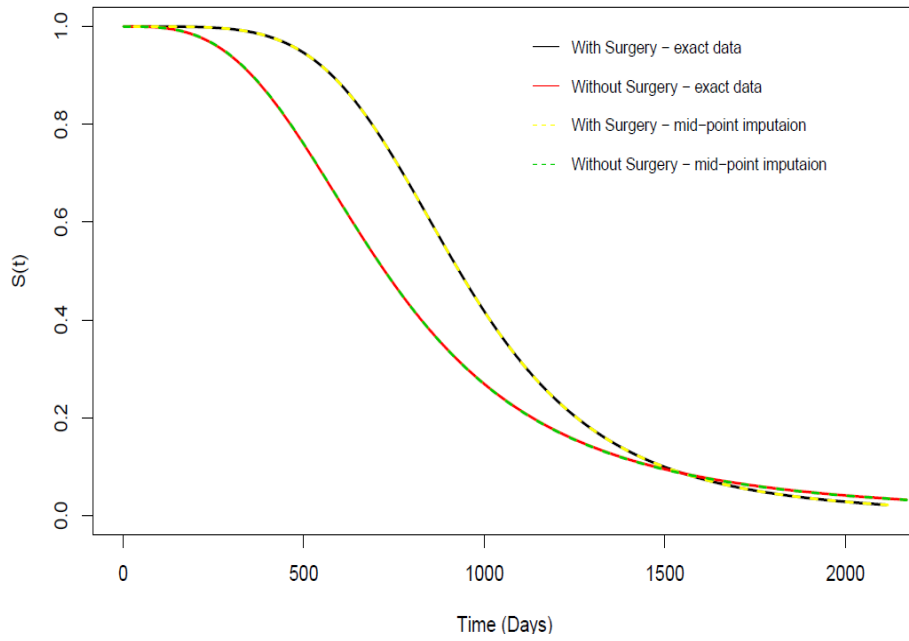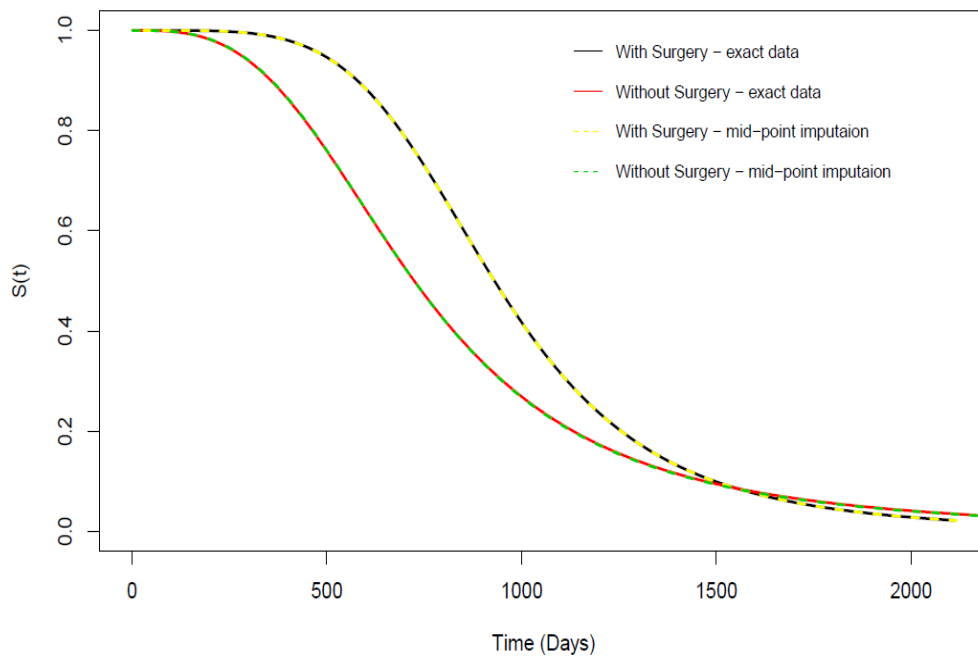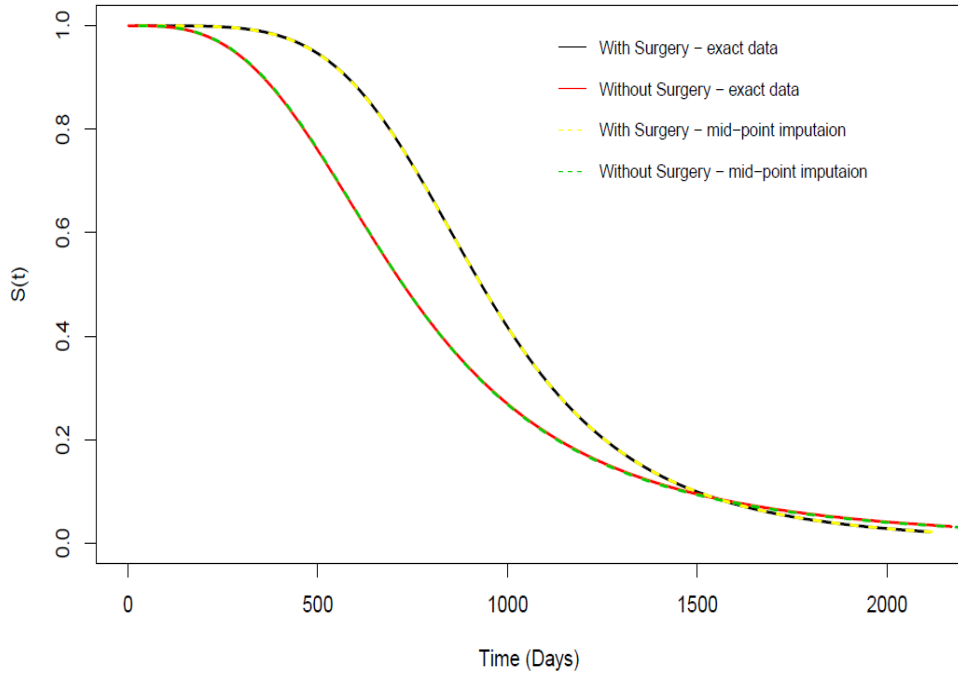Figure 9: Thefunction of survival estimated by mean imputation for 50% exact data based on surgery treatment.
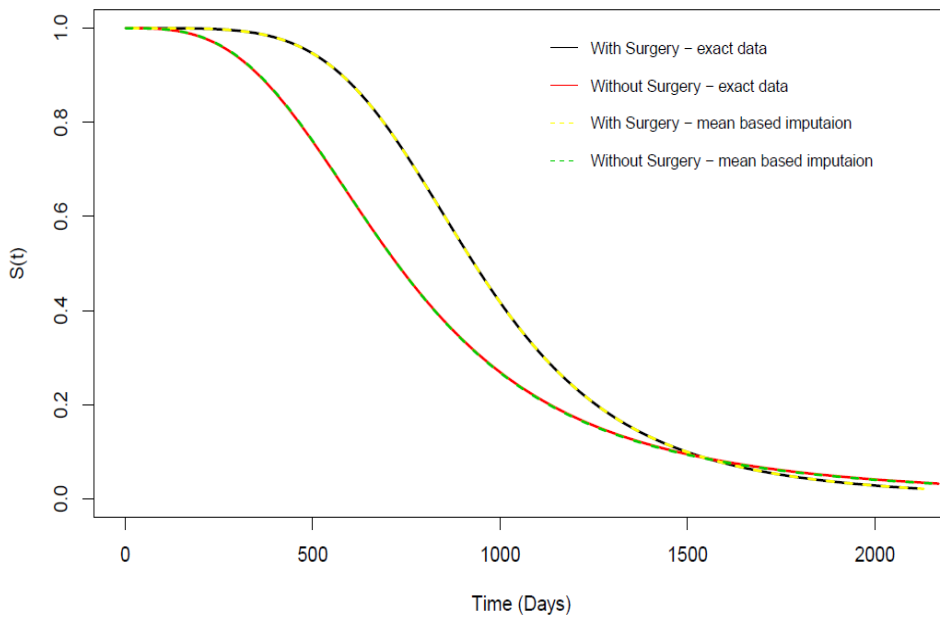
Figure 10: The function of survival estimated by mean imputation for 25% exact data based on surgery treatment.

Figure 11: The function of survival estimated by mean imputation for 0% exact data based on surgery treatment.

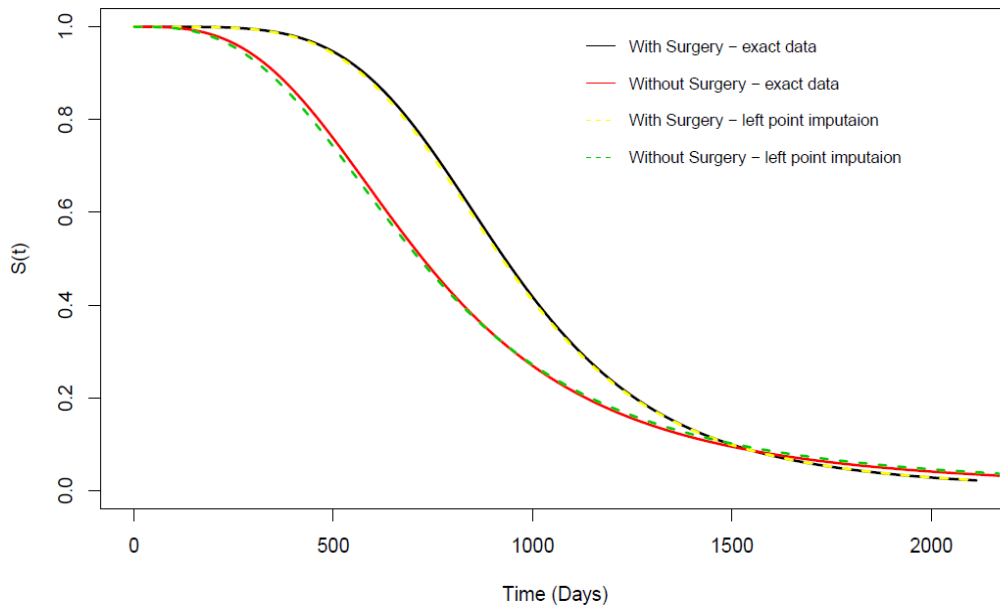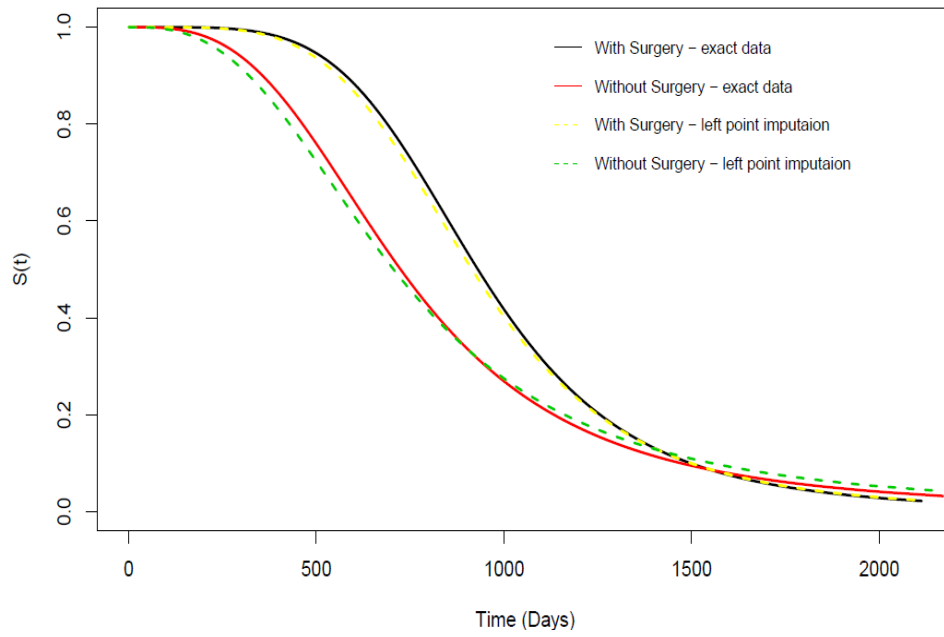Figure 12:Thefunction of survival estimated by midpoint imputation for 75% exact data based on surgery treatment.

Figure 13: The function of survival estimated by midpoint imputation for 50% exact data based on surgery treatment.
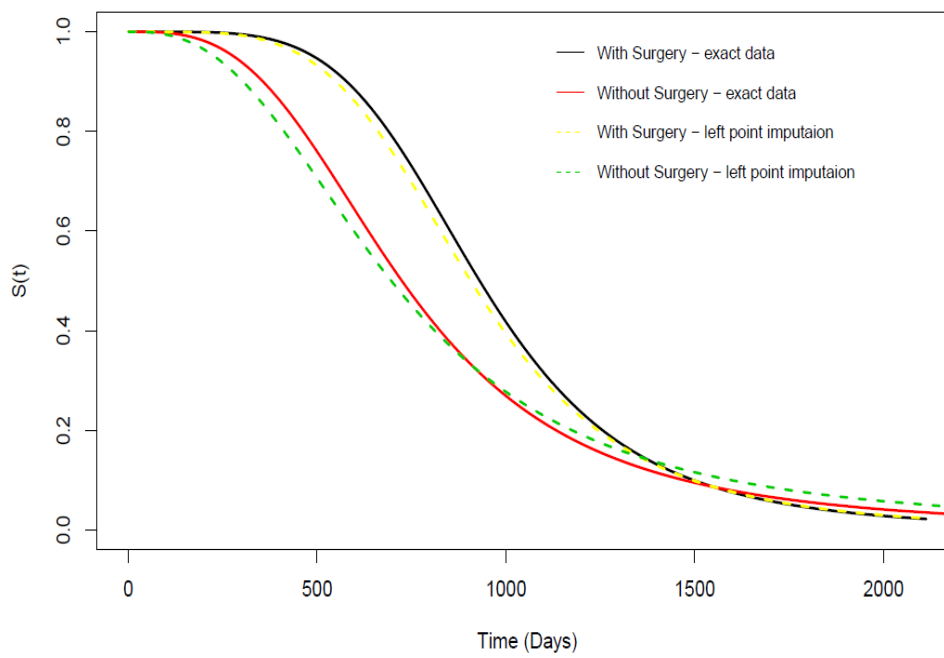

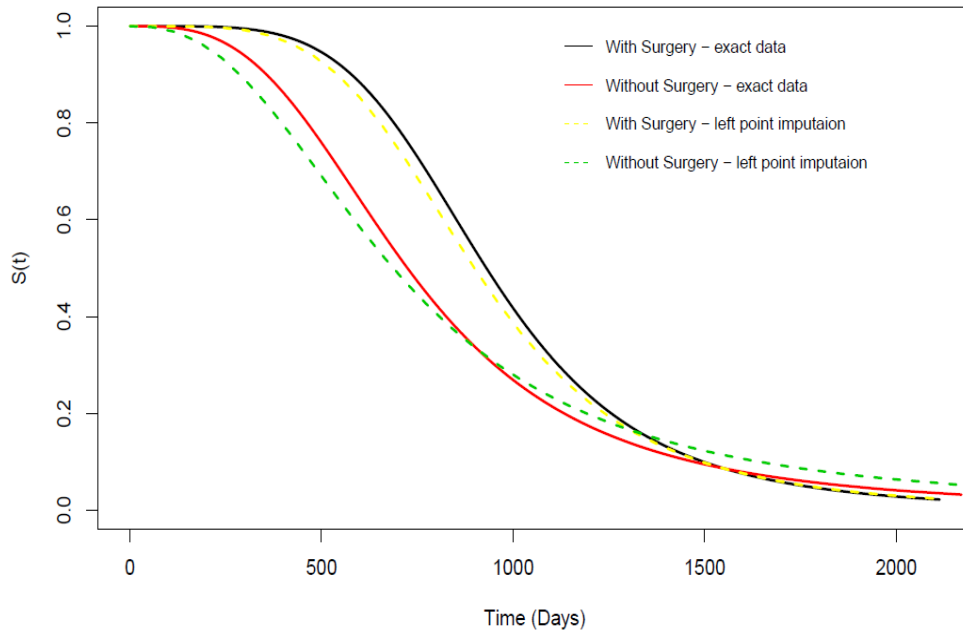Figure 14: The function of survival estimated by midpoint imputation for 25% exact data based on surgery treatment.

Figure 15:The function of survival estimated by midpoint imputation for 0% exact data based on surgery treatment.

Figure 16: The function of survival estimated by left imputation for 75% exact data based on surgery treatment.

Figure 17: The function of survival estimated by left imputation for 50% exact data based on surgery treatment

Figure 18: The function of survival estimated by left imputation for 25% exact data based on surgery treatment

Figure 19: The function of survival estimated by left imputation for 0% exact data based on surgery treatment.

To generate the data for the radiotherapy covariate (with RT treatment and without RT treatment) mean and standard deviations used as 0.2421249 & 0.0321832 and 0.11227867 & 0.025425 respectively. Exact observation with different percentage that is 0%, 25%, 50%, and 75% in PIC data are also used and compared.

Figures 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30 and 31 show the results obtained based on our model mentioned in the early chapter based on different three imputations that are; mean point, mid-point and left point, respectively. The figures are acceptable since it showed similar result between the exact observation and the one obtained by our imputations methods mentioned except the one obtained by left point with exact observation of 0% and 25%.

However, significant results are shown based on the midpoint imputation with respect to LRT and their P-value which indicate that the null hypothesis is rejected (H_0: there is no different between patient who use RT and not use RT treatment) as shown

in Table 3.Clearlyour Cox model via log-logistic distribution with imputations method is easily implemented to partly-interval censored breast cancer data.
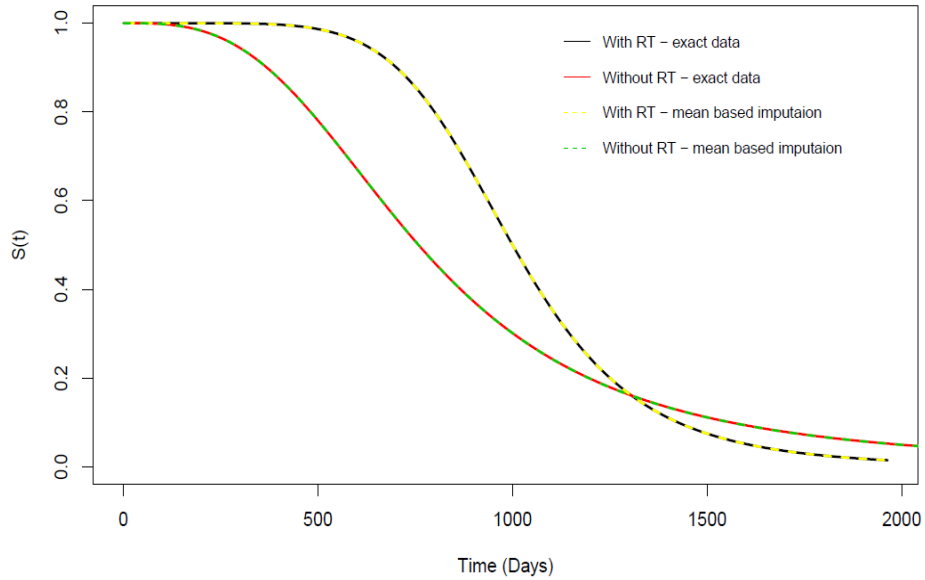

Figure 20: The function of survival estimated by mean imputation for 75% exact data based on RT treatment
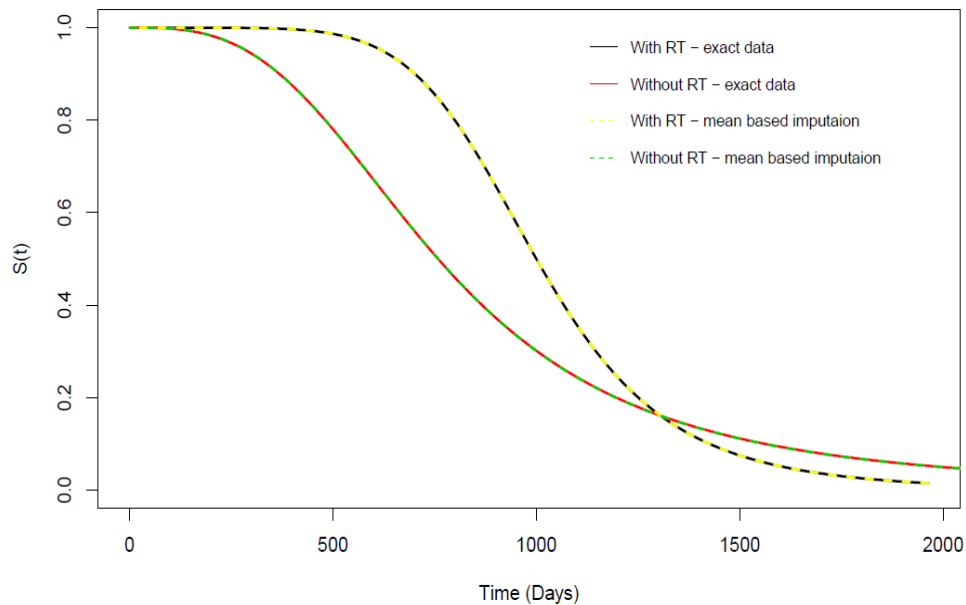

Figure 21: The function of survival estimated by mean imputation for 50% exact data based on RT treatment.
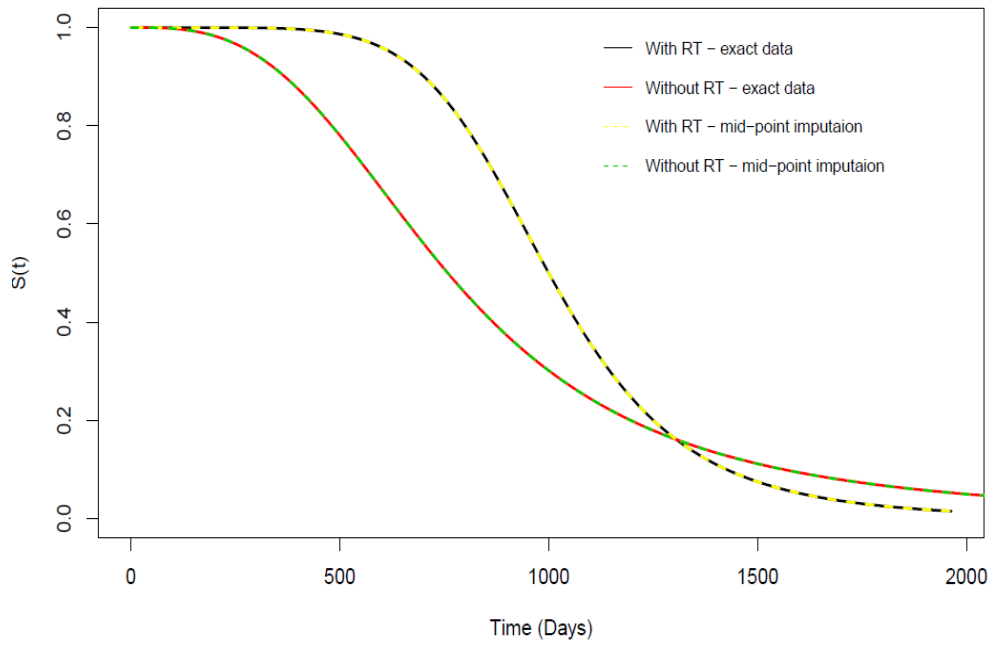
Figure 22: The function of survival estimated by mean imputation for 25% exact data based on RT treatment.
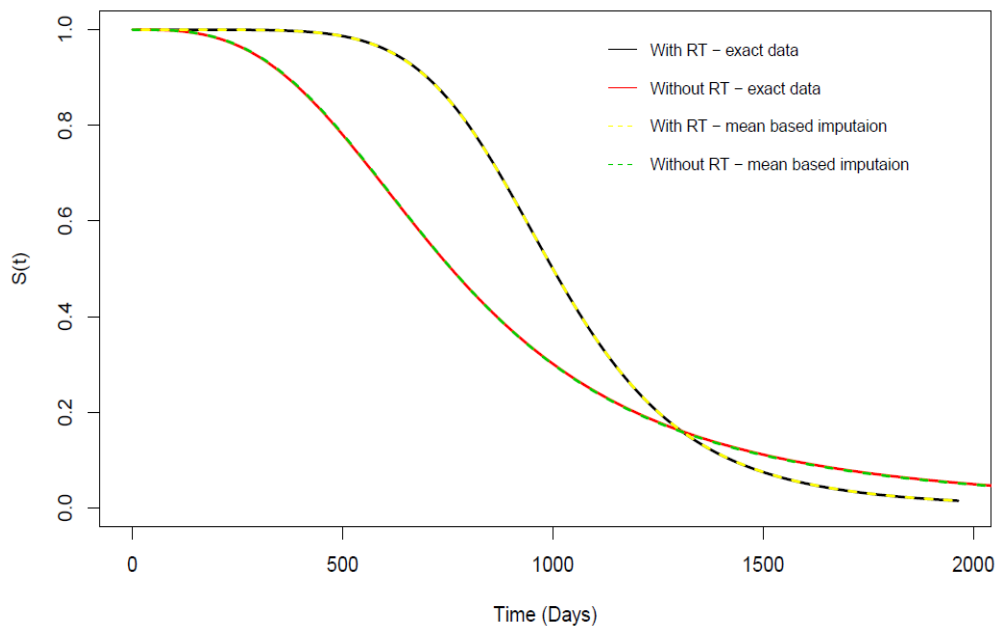

Figure 23: The function of survival estimated by mean imputation for 0% exact data based on RT treatment
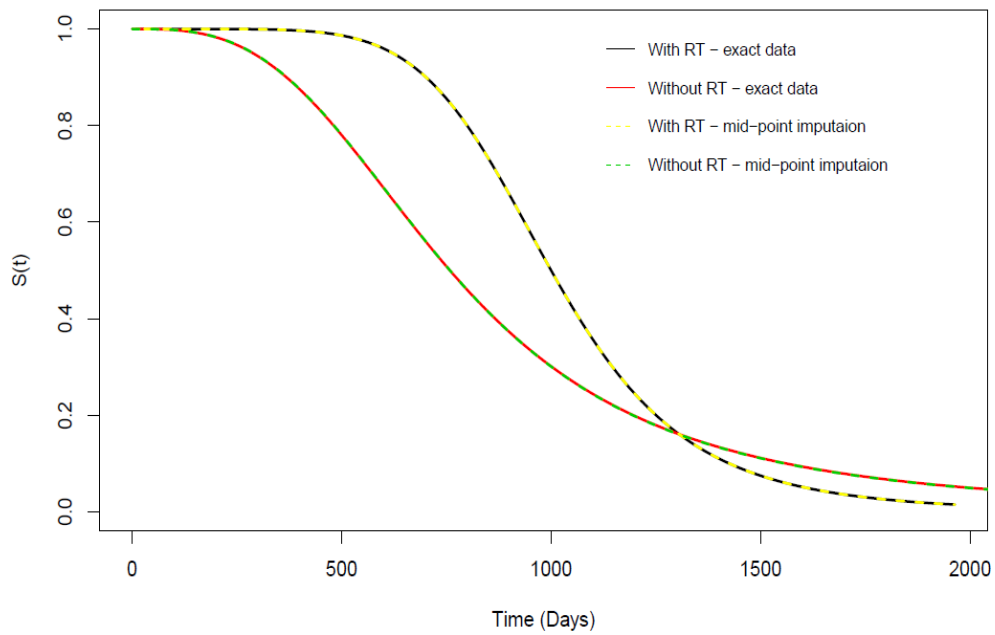
Figure 24: The function of survival estimated by midpoint imputation for 75% exact data based on RT treatment.
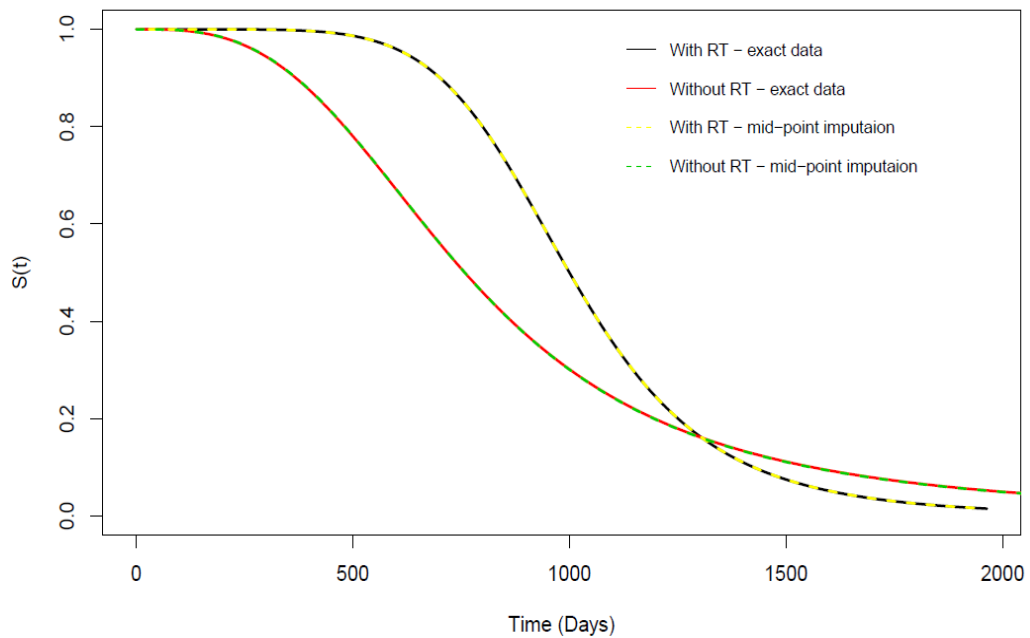

Figure 25: The function of survival estimated by midpoint imputation for 50% exact data based on RT treatment.
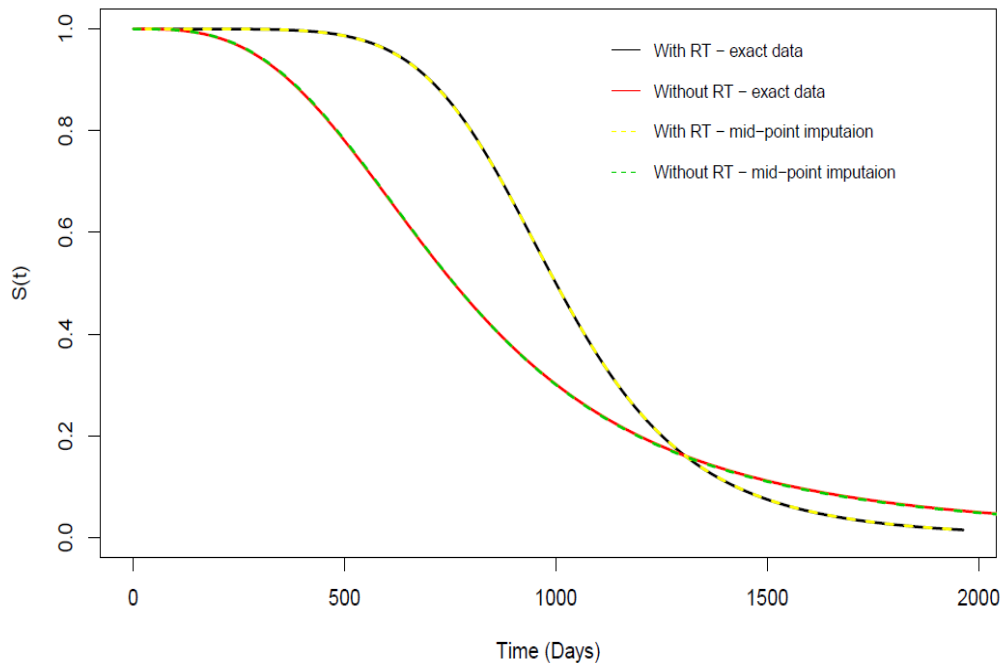
Figure 26: The function of survival estimated by midpoint imputation for 25% exact
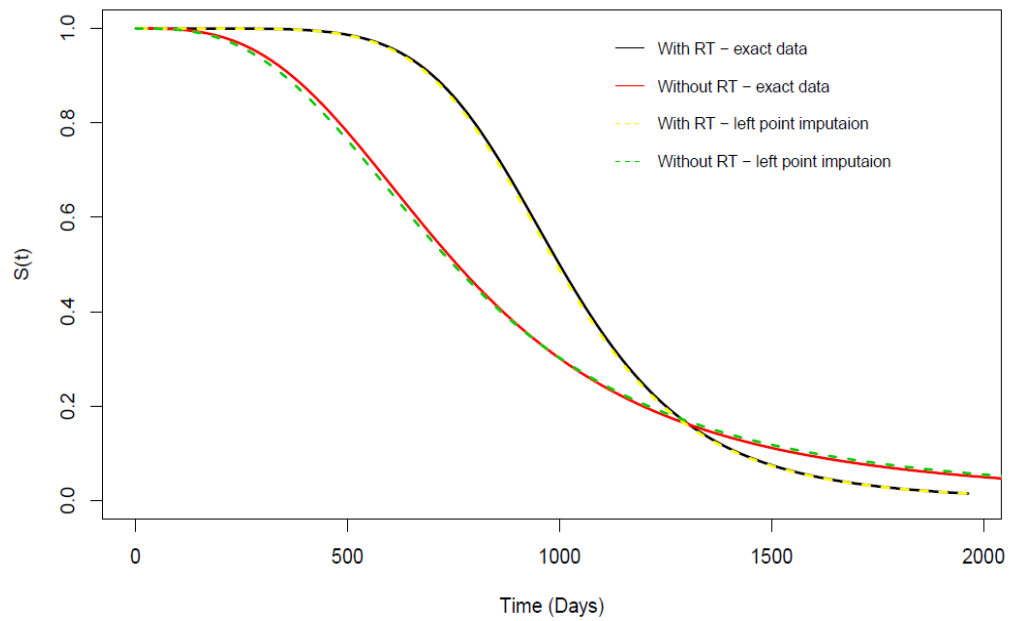
data based on RT treatment

Figure 27: The function of survival estimated by midpoint imputation for 0% exact

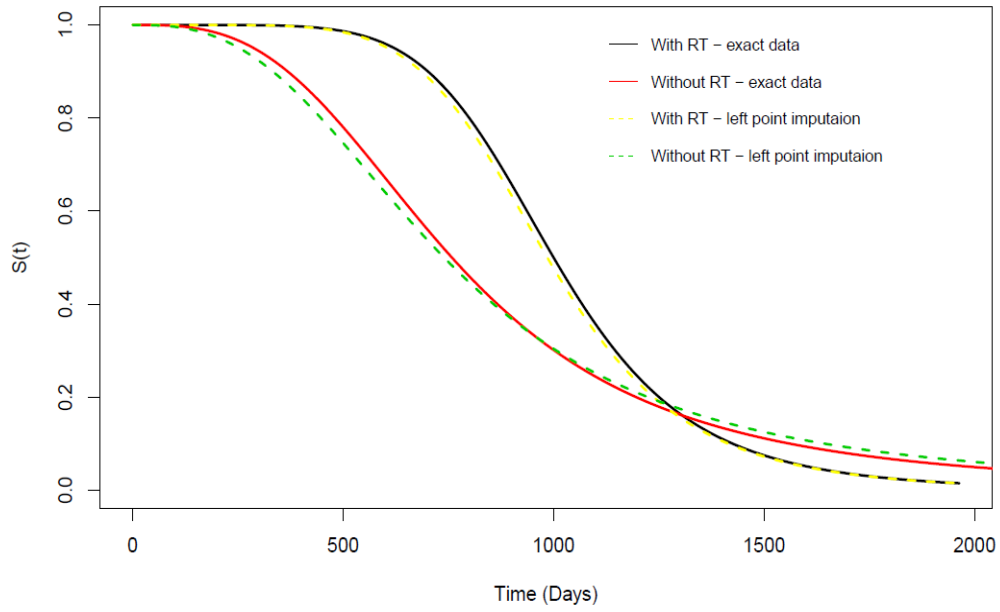Data based on RT treatment.

Figure 28: The function of survival estimated by left point imputation for 75% exact data based on RT treatment.
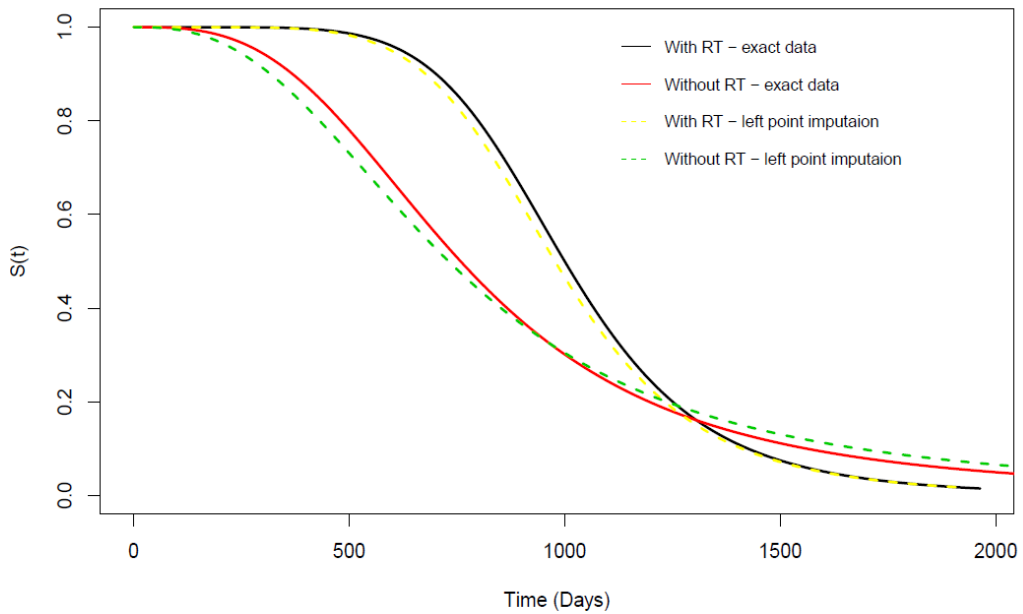

Figure 29: The function of survival estimated by left point imputation for 50% exact data based on RT treatment.
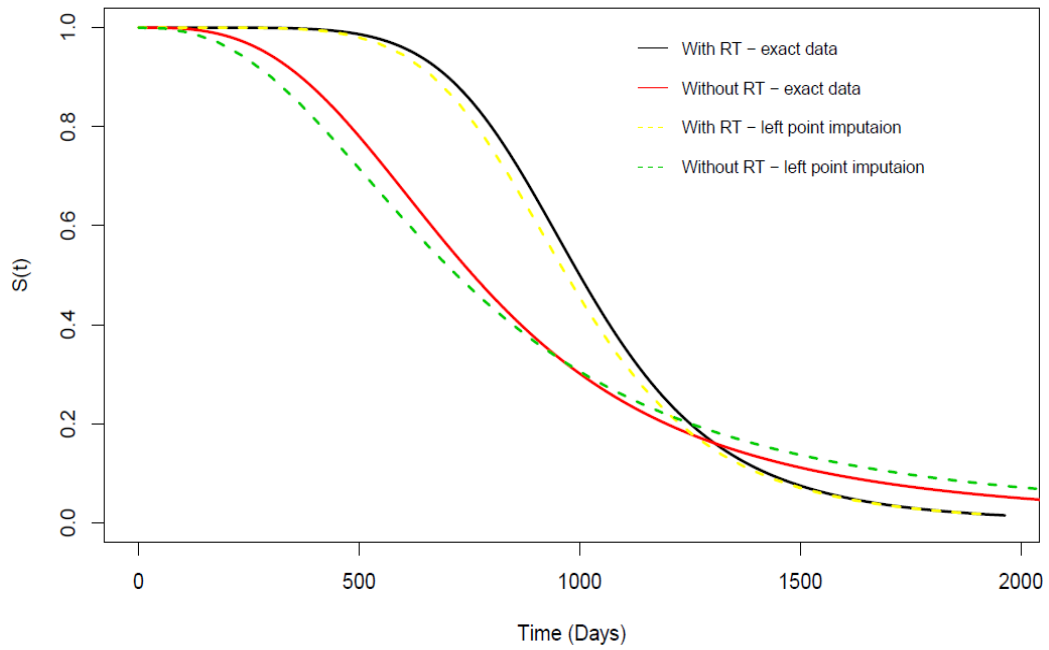
Figure 30: The function of survival estimated by left point imputation for 25% exact data based on RT treatment
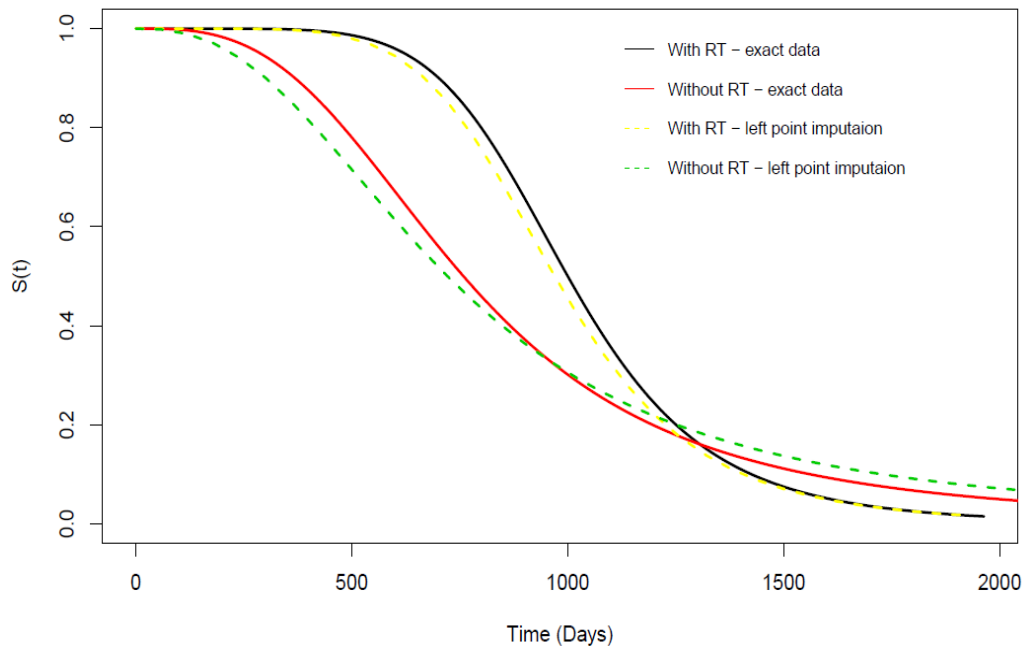

Figure 31: The function of survival estimated by left point imputation for 0% exact data based on RT treatment.

Table 3: Result from simulation data for Radiotherapy based midpoint via Log-logistic Model.

Exact	Parameter	Estimate	CI of 95%	SE	LRT*(P-value)
0%	Shape	4.074947	(4.0264, 4.124)	0.02491	-142665.5 (2e-16)
	Scale	769.0776	(762.33,775.88)	3.45678	
	Coefficient	0.258039	(0.2464,0.2697)	0.00596	
25%	Shape	4.064404	(4.0159, 4.113)	0.02485	-142720.1(2e-16)
	Scale	769.0731	(762.31, 775.9)	3.46350	
	Coefficient	0.258181	(0.2465,0.2699)	0.00597	
50%	Shape	4.064382	(4.0159, 4.113)	0.02485	-142721.6(2e-16)
	Scale	769.0425	(762.28, 775.86)	3.46309	
	Coefficient	0.258267	(0.2466,0.26996)	0.00597	
75%	Shape	4.058851	(4.0104, 4.108)	0.02482	-142748.7 (2e-16)
	Scale	768.9808	(762.22,775.80)	3.46589	
	Coefficient	0.258240	(0.2465, .2699)	0.00597	
100%	Shape	4.057459	(4.0091, 4.106)	0.02482	-142755.7 (2e-16)
	Scale	768.8760	(762.11,775.70)	3.46531	
	Coefficient	0.258317	(0.2466,0.2700)	0.00598	

For the hormone treatment (with and without hormone treatment), the data we generate with mean and standard deviation as 1022.85 & 246.25 and 762.64 & 379.78 via different exact observations of PIC that is with; 0%, 25%, 50%, 75% and 100%.

Figures 32, 33, 34, 35, 36, 37, 38, 39,40, 41, 42 and 43 displayed the result of the estimated the function of survival obtained by our model and imputation techniques that is; mean point, midpoint and left point. These figures look almost similar in case of the one obtained by mean and midpoint especially for exact observation that more than 25%, but little difference compared with one obtained by left point in case of exact observation that more less than 50%. However, the patient treated with hormone treatment have a long survival compared to those without hormone treatment as shown clearly in the figures above as well in Table 4 concerning LRT and p-value. In additional to that the null hypothesis (H0: there is no different between patient who use hormone and not use hormone treatment) is rejected.
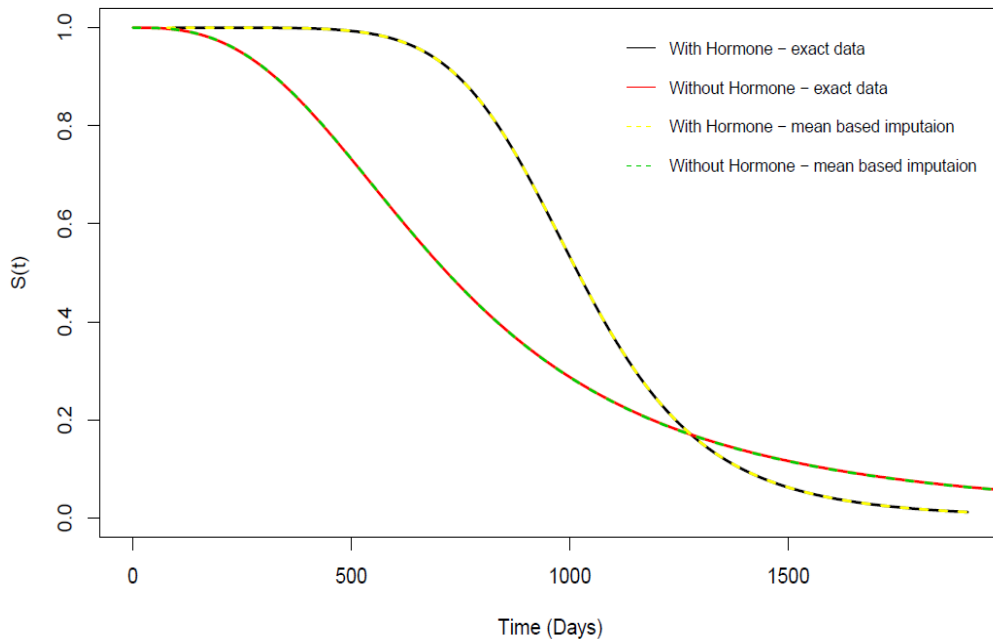
Figure 32: The function of survival estimated by mean imputation for 75% exact data based on hormone treatment

Figure 33: The function of survival estimated by mean imputation for 50% exact data based on hormone treatment.

Figure 34: The function of survival estimated by mean imputation for 25% exact data based on hormone treatment.
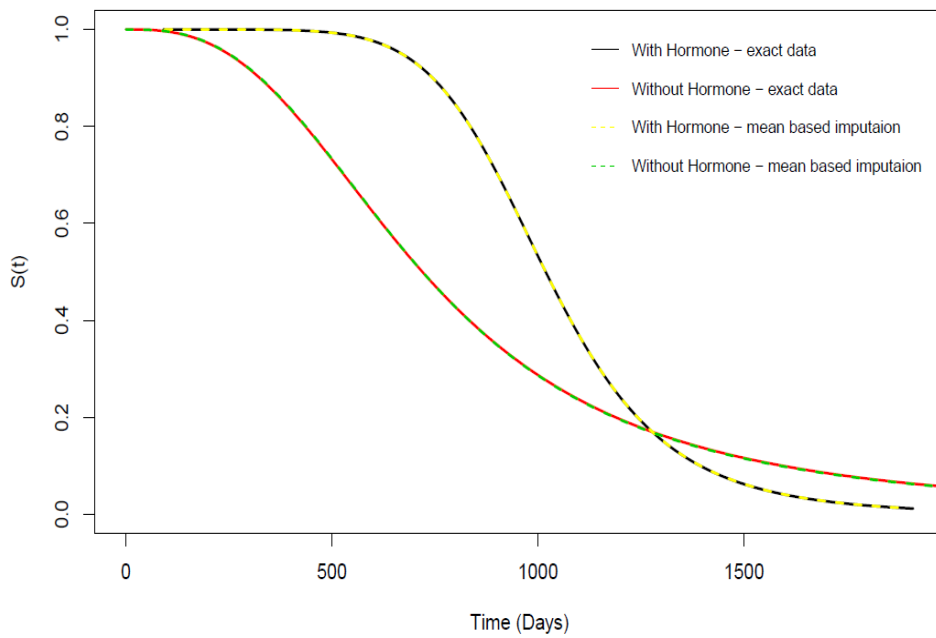

Figure 35: The function of survival estimated by mean imputation for 50% exact data based on hormone treatment.
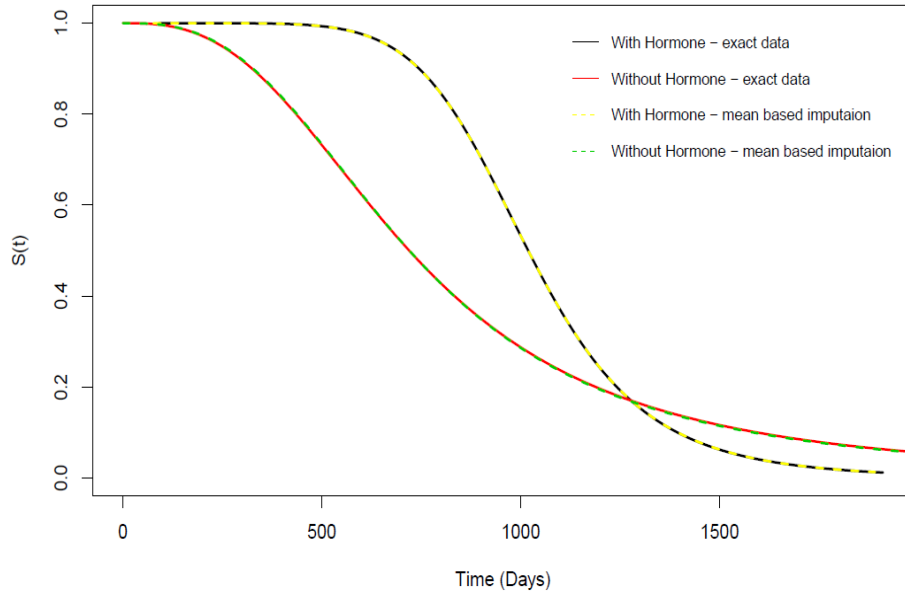
Figure 36: The function of survival estimated by midpoint imputation for 75% exact data based on hormone treatment.
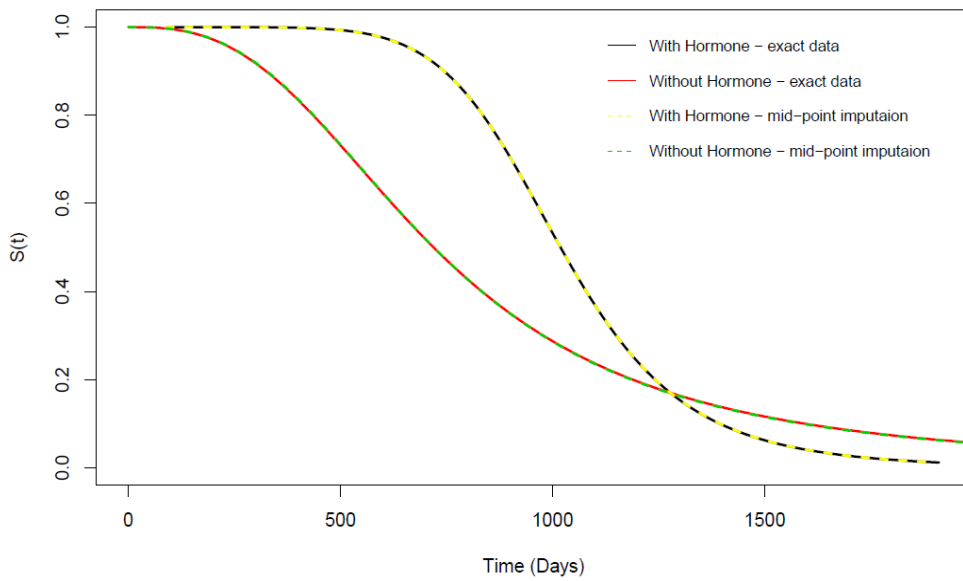
Figure 37: The function of survival estimated by midpoint imputation for 50% exact data based on hormone treatment.
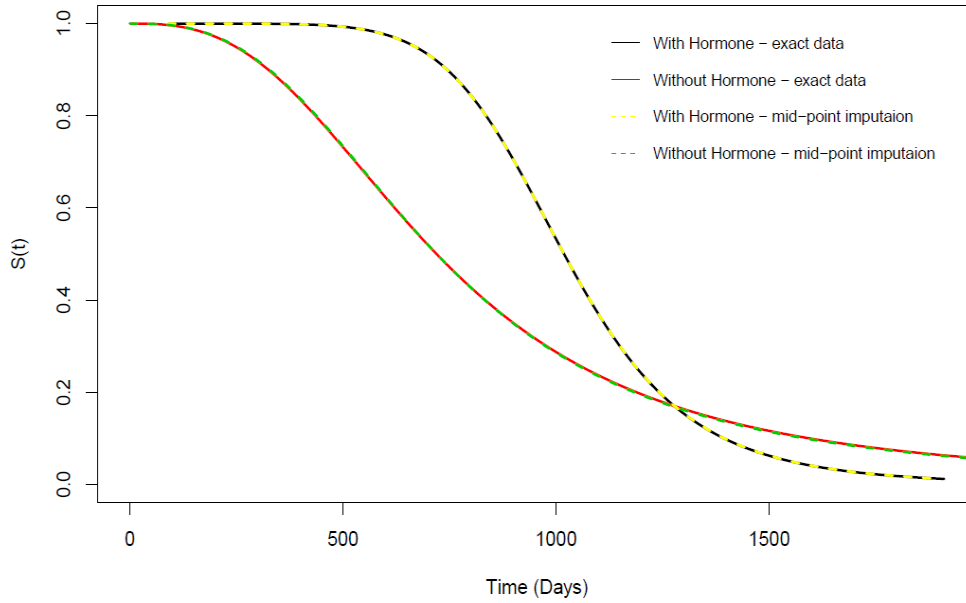
Figure 38: The function of survival estimated by midpoint imputation for 25% exact data based on hormone treatment.

Figure 39: The function of survival estimated by midpoint imputation for 0% exact data based on hormone treatment.

Figure 40: The function of survival estimated by left point imputation for 75% exact data based on hormone treatment.
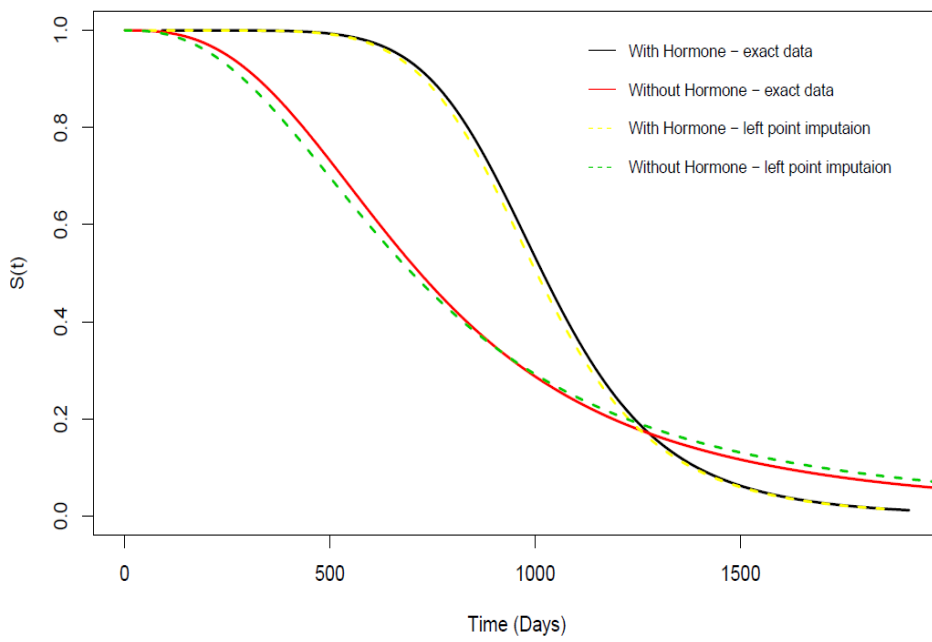
Figure 41: The function of survival estimated by left point imputation for 50% exact data based on hormone treatment.
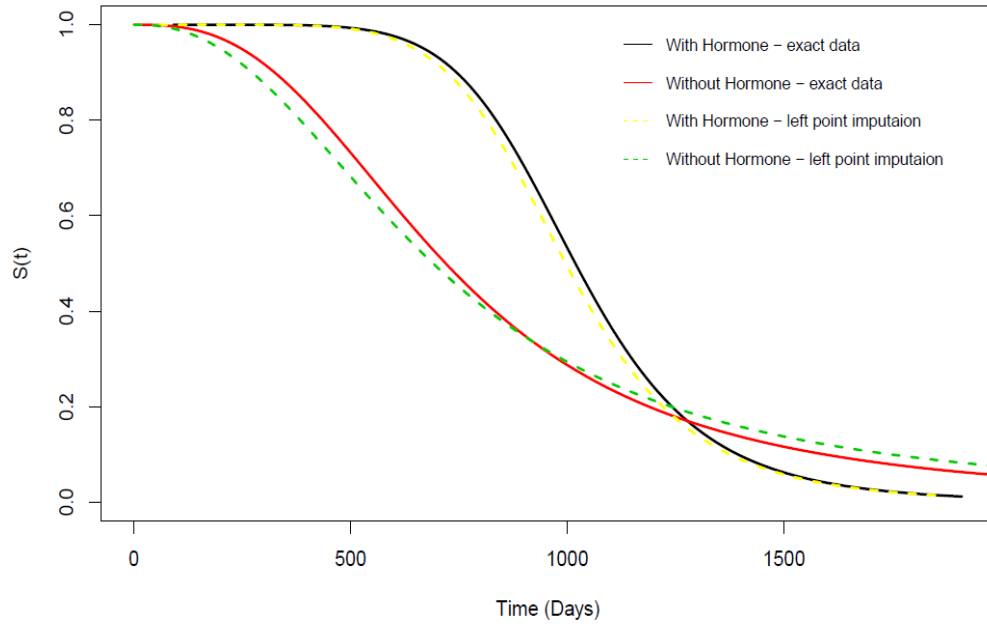
Figure 42: The function of survival estimated by left point imputation for 25% exact data based on hormone treatment.
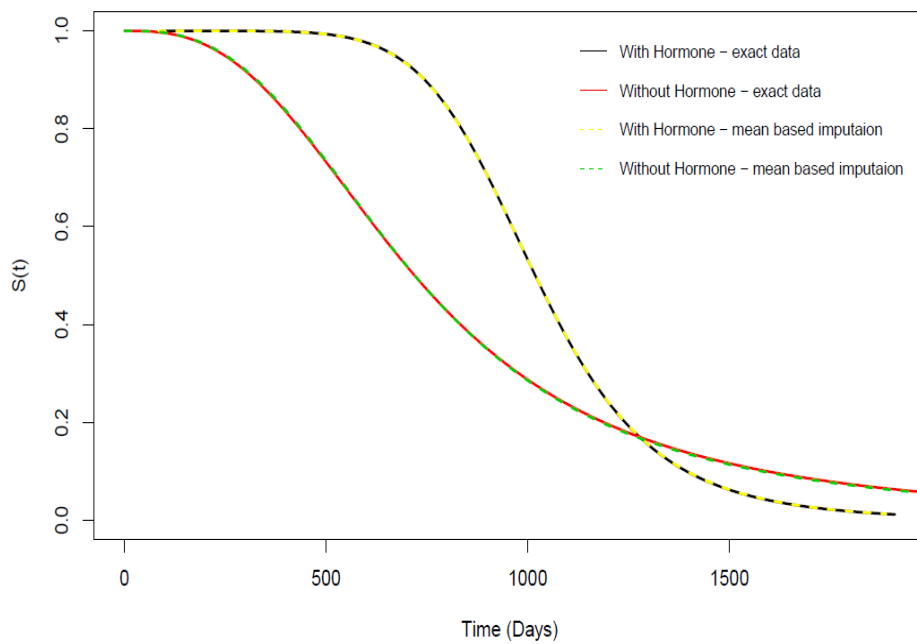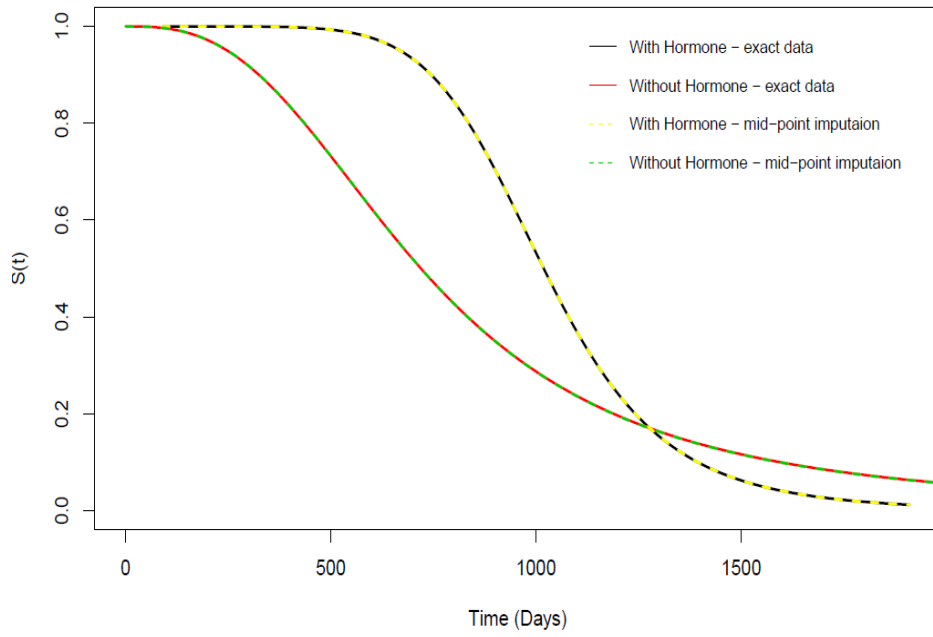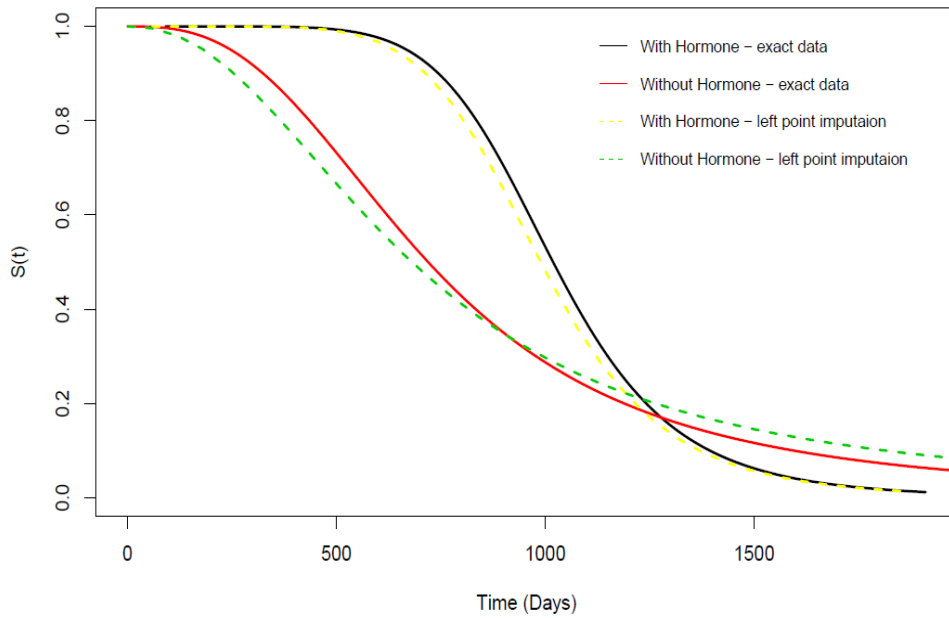

Figure 43: The function of survival estimated by left point imputation for 0% exact data based on hormone treatment.
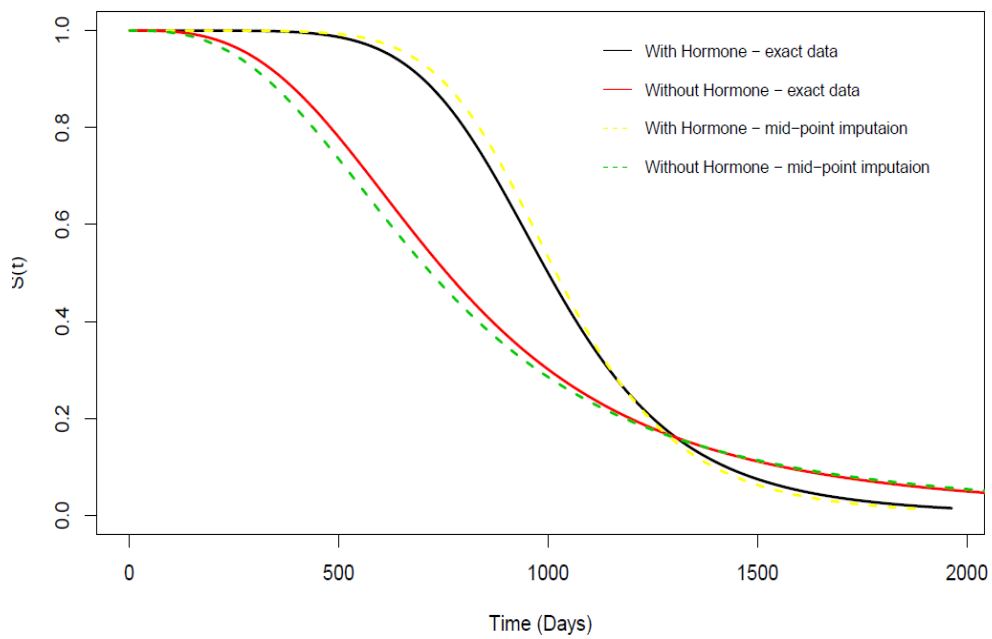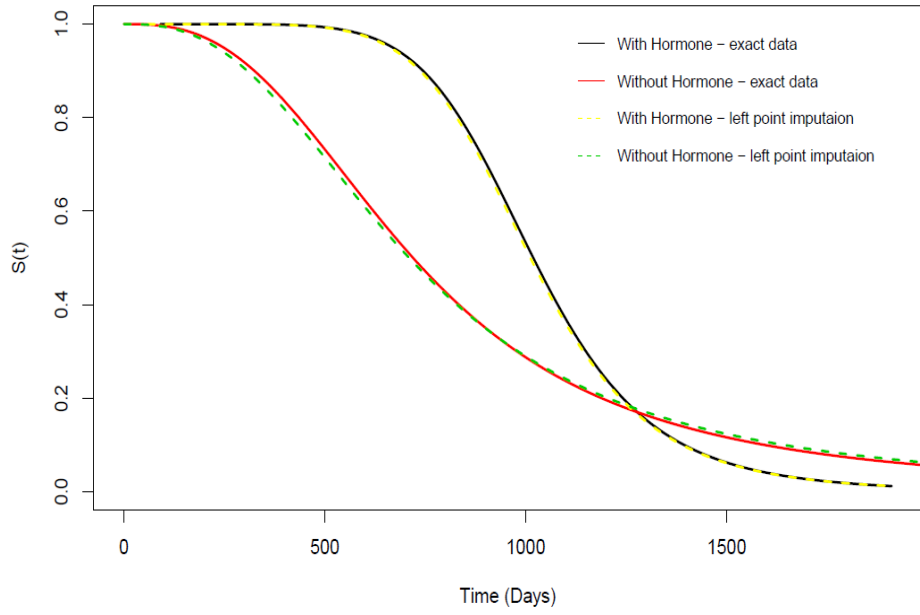
Table 4: Simulation data for hormone based mean point via Log-logistic Model.

Exact	Parameter	Estimate	CI of 95%	SE	LRT*(P-value)
0%	Shape	3.95712	(3.9097,4.0052)	0.02437	-143114.6 (2e-16)
	Scale	734.632	(727.90,741.43)	3.45239	
	Coefficient	0.32244	(0.3104,0.3345)	0.00613	
25%	Shape	3.94721	(3.8998,3.9951)	0.02431	-143161.9 (2e-16)
	Scale	734.346	(727.60,741.15)	3.45737	
	Coefficient	0.32279	(0.3108,0.3348)	0.00614	
50%	Shape	3.93673	(3.8895,3.9846)	0.02425	-143216.3 (2e-16)
	Scale	734.391	(727.63,741.21)	3.46420	
	Coefficient	0.32268	(0.3106,0.3347)	0.00616	
75%	Shape	3.92707	(3.8799,3.9748)	0.02420	-143266.8(2e-16)
	Scale	734.239	(727.47,741.07)	3.46891	
	Coefficient	0.32292	(0.3108,0.3350)	0.00617	
100%	Shape	3.92137	(3.8742,3.9690)	0.02417	-143295.9(2e-16)
	Scale	734.165	(727.39,741.00)	3.47155	
	Coefficient	0.32290	(0.3108,0.3350)	0.00617	

Table 5: Result from simulation data for chemotherapy based midpoint via Log-logistic Model.

Exact	Parameter	Estimate	CI of 95%	SE	LRT*(P-value)
0%	Shape	4.037694	(3.9901, 4.0858)	0.02444	-142962.6 (2e-16)
	Scale	828.8237	(821.86, 835.84)	3.56489	
	Coefficient	0.114750	(0.1030, 0.1265)	0.00601	
25%	Shape	4.035990	(3.9884,4.0842)	0.02443	-142972.4(2e-16)
	Scale	828.8043	(821.85, 835.82)	3.56538	
	Coefficient	0.114723	(0.1029,0.1265)	0.00601	
50%	Shape	4.037423	(3.9898, 4.0856)	0.02444	-142966.5(2e-16)
	Scale	828.8207	(821.86, 835.84)	3.56439	
	Coefficient	0.114801	(0.1030,0.1266)	0.00601	
75%	Shape	4.032079	(3.9845, 4.0802)	0.02441	-142995.1(2e-16)
	Scale	828.7023	(821.74, 835.72)	3.56720	
	Coefficient	0.115005	(0.1032, 0.1268)	0.00601	
100%	Shape	4.032791	(3.9852, 4.0809)	0.02442	-142993.8(2e-16)
	Scale	828.6519	(821.69, 835.67)	3.56594	
	Coefficient	0.115118	(0.1033, 0.1269)	0.00601	

To generate the data for the chemotherapy (with chemotherapy treatment and without chemotherapy treatment) the mean and standard deviation is used with a value of 948.28& 338.29 and 853.59 & 332.65 respectively. Exact observations with a different percentage of for PIC data are also used that is 0%, 25%, 50%, 75% and 100%.Figures 44, 45, 46, 47, 48, 49, 50, 51, 52 and 53 displayed the results obtained based on our model by three imputation methods mentioned earlier in this chapter. The figures are almost the same for the result obtained except for the one obtained by left point with exact 0% and 25% (Figures 55 and 56). However, significant results are obtained via midpoint imputation that has been shown concerning LRT and their P-value (Table 5).

Table 5 showed the results obtained by our model based on midpoint imputation for chemotherapy treatment with different percentages of exact and interval censored data. It showed significant with respect LRT and their p-value. These results indicate that for more exact observation in the data the result are better (as lower value of AIC=285931.2 when 100% exact compare to AIC= 285993.6 for 0% exact). Moreover, the chances of survival increases significantly when patient use surgery treatment compared with a patient who haven't gone through the surgery treatment while fighting with breast cancer

Figure 44: The function of survival estimated by mean imputation for 75% exact data based on chemotherapy treatment.

Figure 42: The function of survival estimated by mean imputation for 50% exact data based on chemotherapy treatment.
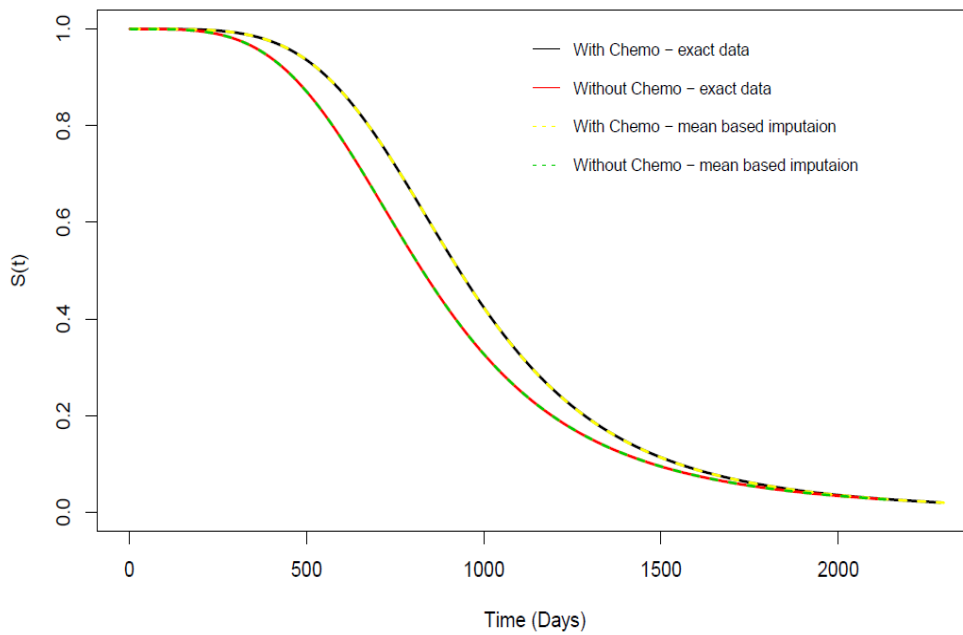
Figure 46: The function of survival estimated by mean imputation for 25% exact data based on chemotherapy treatment.
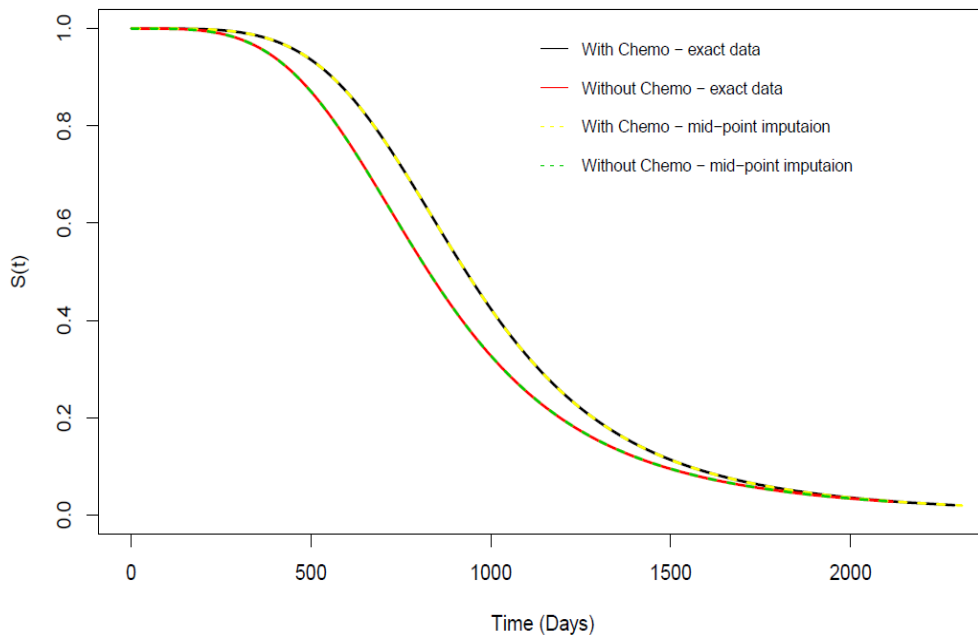
Figure 47: The function of survival estimated by mean imputation for 0% exact data based on chemotherapy treatment.
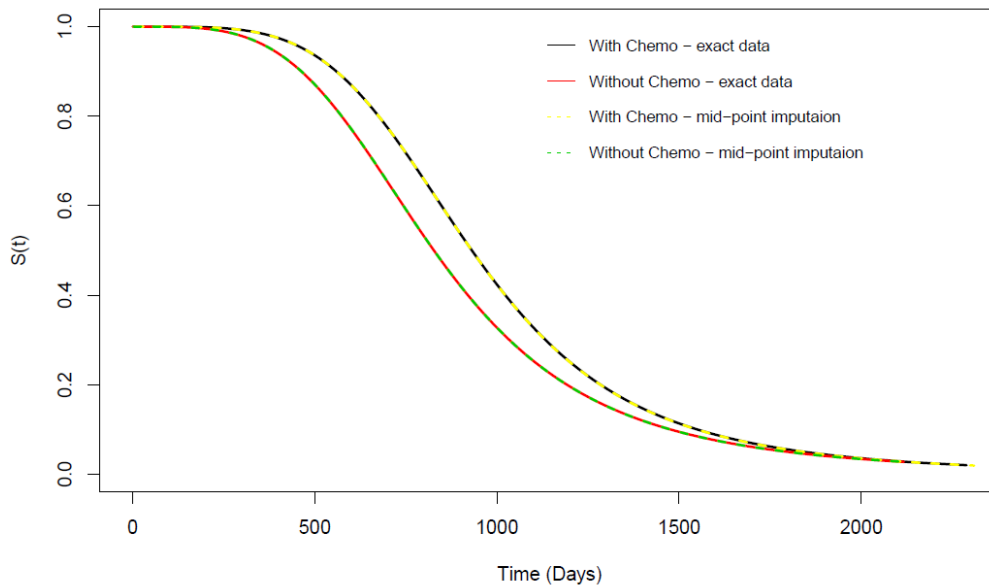
Figure 48: The function of survival estimated by midpoint imputation for 75% exact data based on chemotherapy treatment.
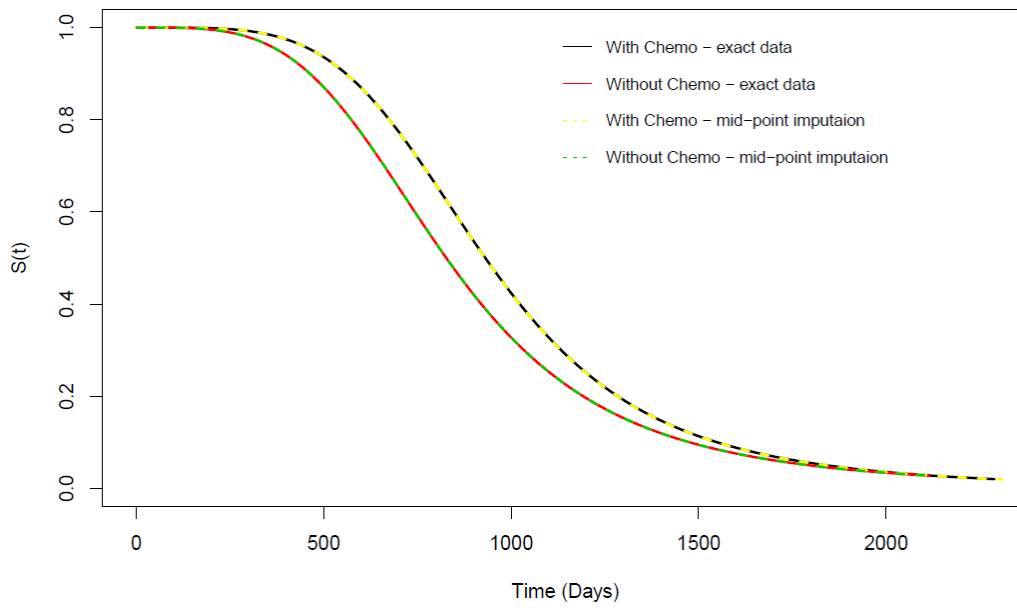
Figure 49: The function of survival estimated by midpoint imputation for 50% exact data based on chemotherapy treatment.
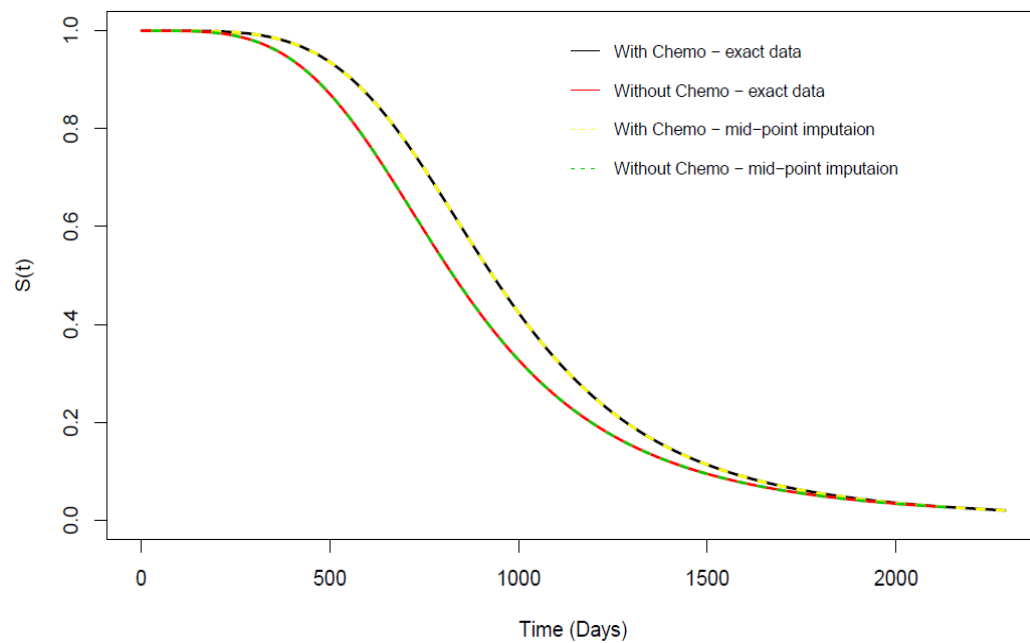
Figure 50: The function of survival estimated by midpoint imputation for 25% exact data based on chemotherapy treatment.
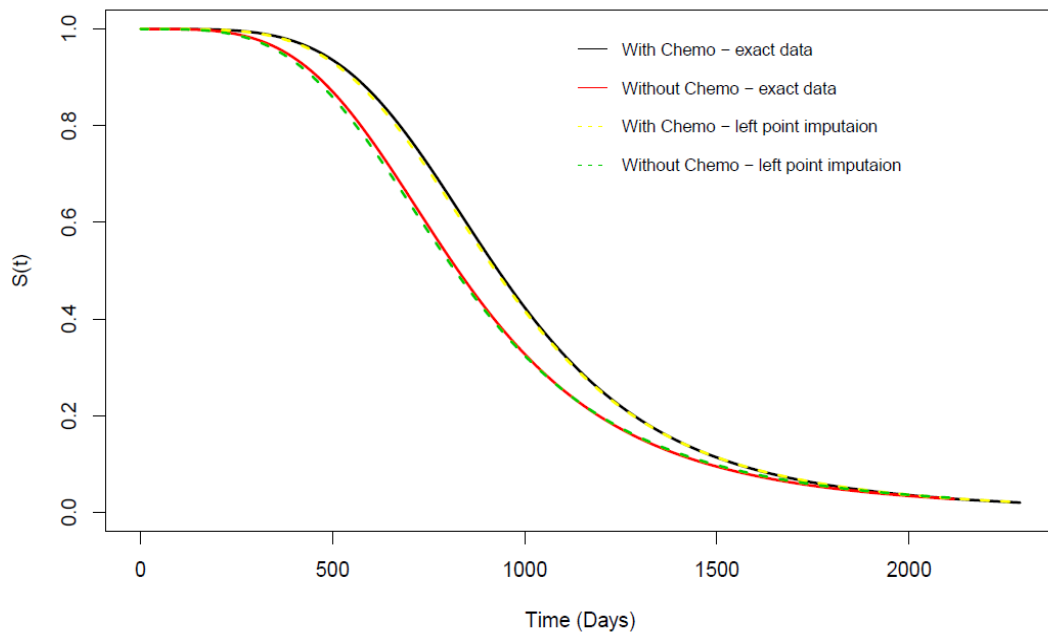
Figure 51: The function of survival estimated by midpoint imputation for 0% exact data based on chemotherapy treatment.
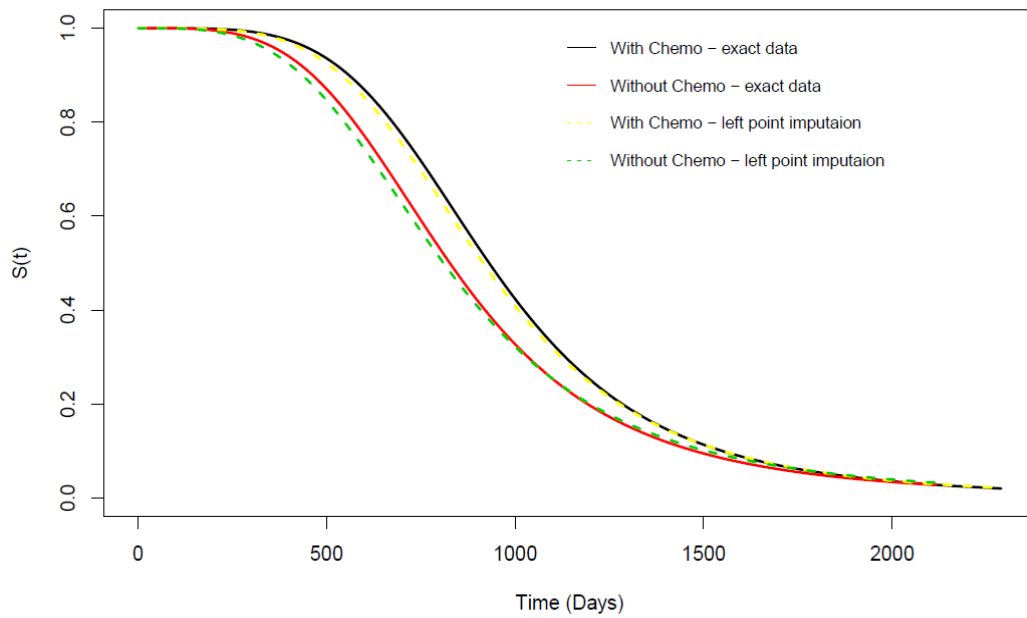
Figure 52: The function of survival estimated by left point imputation for 75% exact data based on chemotherapy treatment.
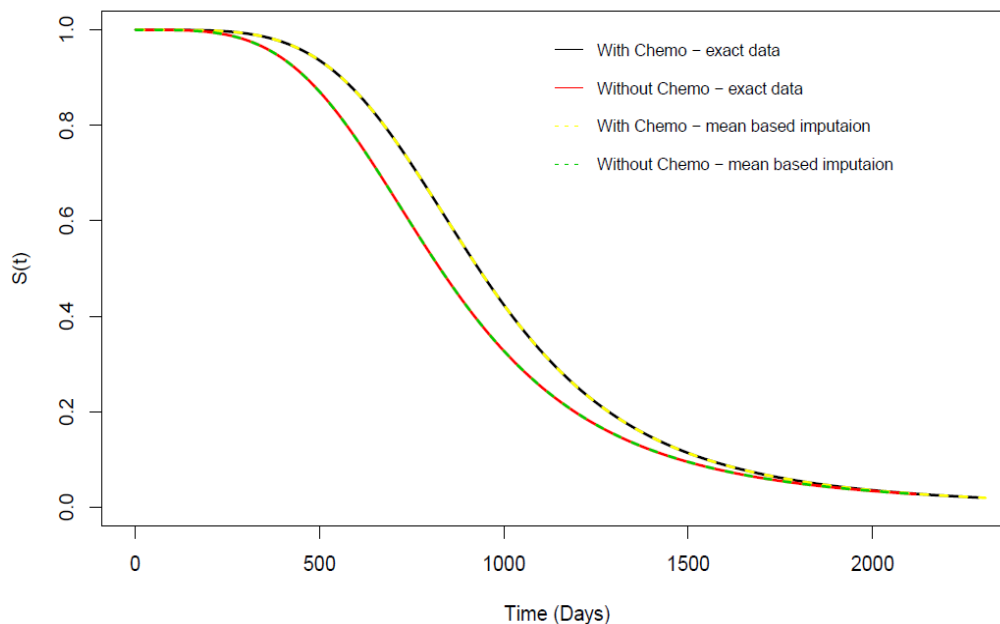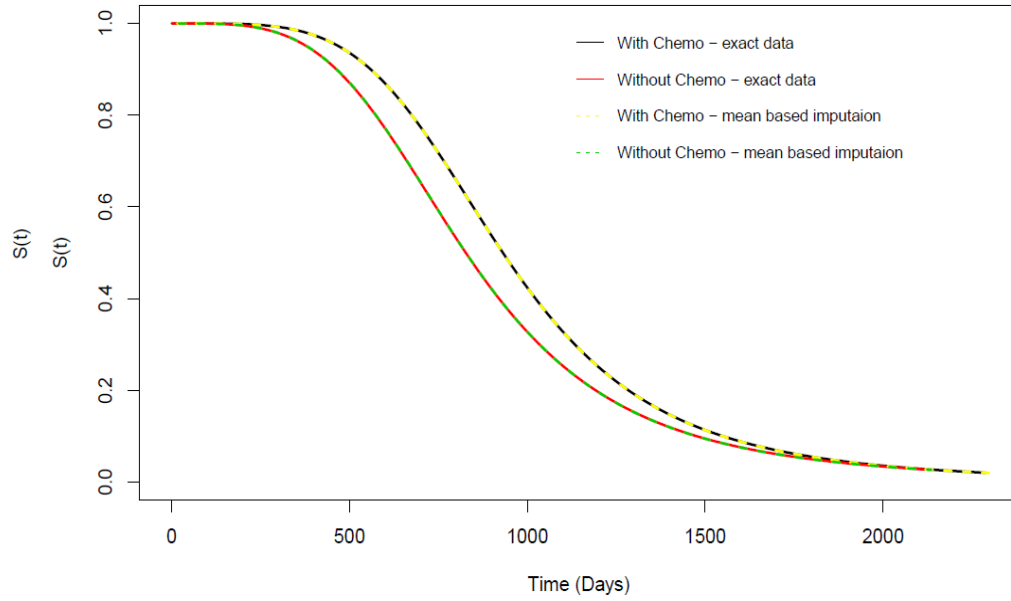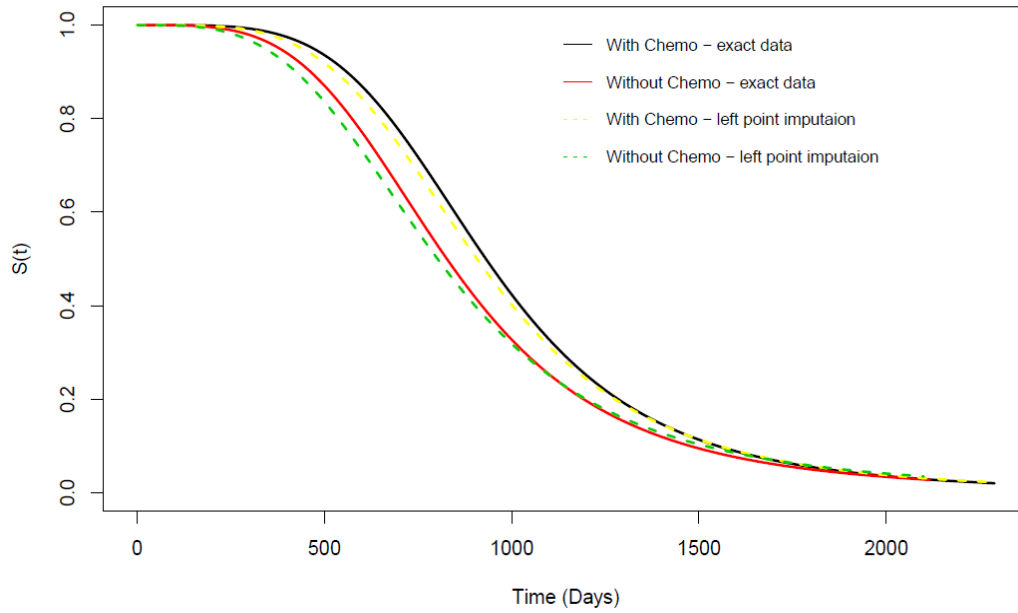
Figure 53: The function of survival estimated by left point imputation for 50% exact data based on chemotherapy treatment.
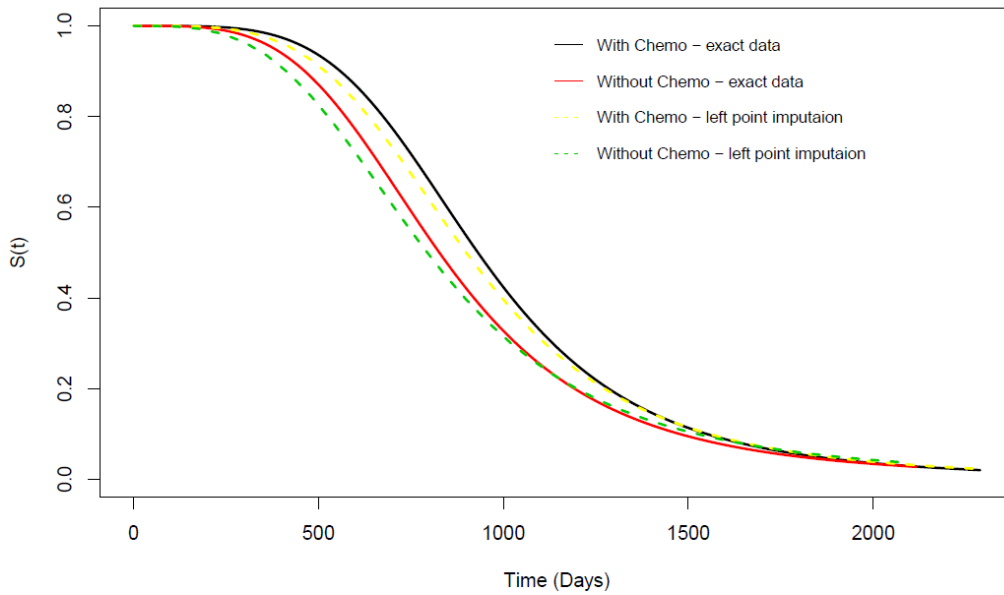
Figure 54: The function of survival estimated by left point imputation for 25% exact data based on chemotherapy treatment.
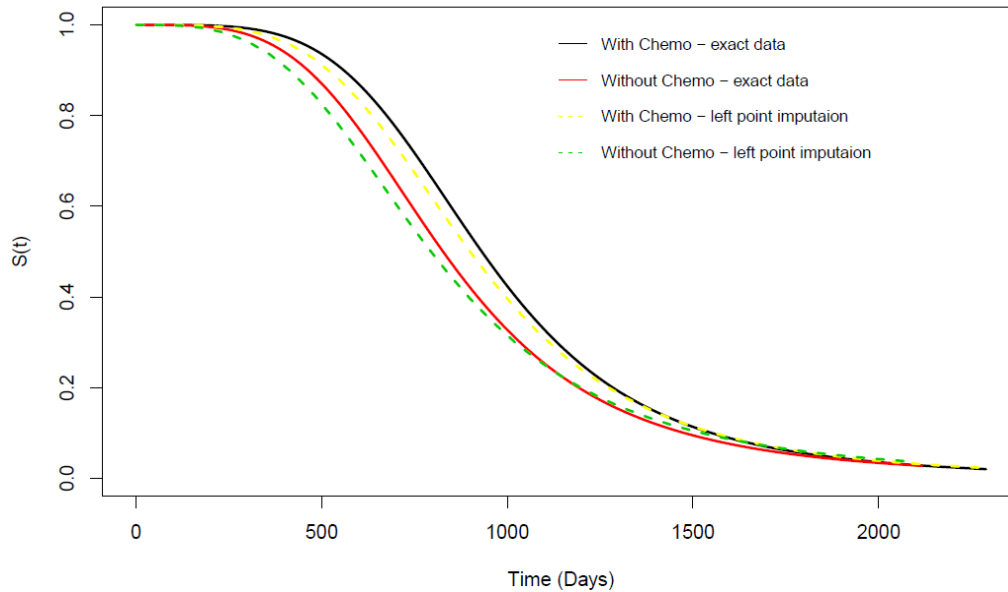
Figure 55: The function of survival estimated by left point imputation for 0% exact data based on chemotherapy treatment.

We can conclude that our model which is Cox with log-logistic via simple imputations methods fits very well the PIC breast cancer data as well as the simulation data with different percentages of exact observation. We highlight in the data and in the simulation data two different failure rates for each treatment that is; with surgery& without surgery, with RT treatment & without RT, with hormone treatment & without hormone and with chemotherapy & without chemotherapy. The simulation study runs for 20000 times based on normal distribution. The results showed that patients who are treated with surgery, RT, hormone and chemotherapy have better opportunities to survive longer compared to those without treatment as shown in Figures 2 to 56and Tables 2 to 5.

Surgery is done to enhance the quality of living rather than to cure the disease itself, for example, to alleviate discomfort induced by a tumor that is crushing the nerve or bone, or to extract a tumor that covers the intestine. Radiation therapy may be used to cure virtually every stage of breast cancer.

Radiation therapy is an effective approach used to reduce the risk of breast cancer arising after surgery. In addition to that, it is commonly used to treat cancer-induced complications that have been spread to other sections of the body. Hormone medication after surgery, radiation or chemotherapy have been proved effective in reducing the likelihood of recurrence of breast cancer in women's bodies at an early-stage hormone-sensitive breast cancer. It also significantly reduces the chance of breast cancer development and advancement for people with hormone-sensitive tumors. Breast cancer chemotherapy is sometimes used in contrast to other treatments, such as surgery, radiation treatment or hormone therapy. It can affect the chances of success of a cure, minimize the possibility of recurrence of cancer, minimize the symptoms of cancer, or help people with cancer to survive longer with higher life quality.

CHAPTER 5: CONCLUSION AND SUGGESTION FOR FUTURE RESEARCH

CHAPTER OVERVIEW

There will be two parts to this chapter. The first part is about a conclusion that summarizes the results obtained in previous chapters. The second part will deal with possible suggestions for future research.

5.1 Conclusion

In this study, we used the log-logistic Cox model via simple imputation techniques to simplify the procedure for partly interval censored (PIC) data. Log-logistic distribution model has been mostly applied in medical experiments. The estimated of survival function was obtained using the maximum likelihood estimator and under PIC. Comparisons were made with existing one under the Weibull and lognormal distribution for generating the simulation data. The first step of this study is to look for real data set to confirm that our model is useful. Based on the result extracted from this data set, we found that our model fits well and it is easy to be implemented with respect to the survival functions estimated (Figures 2, 3, 4 and 5), the value of LRT and their p-value and AIC for four treatments that is; RT, surgery, hormone and chemotherapy.

The medical data set used to implement our methods was collected from the Al-amal hospital in Qatar. The data was followed up the cancer patients from 2/1/2016 until 1/19/2020. This study was implemented to compare cosmetic effects of each treatment alone for women in an early breast cancer stage and the event of interest was the time of first occurrence of breast retraction and the patients were observed at clinic visits, where the actual dates of the event were exactly recorded.

Then, we modified the data set to be PIC data and interval data for research needs.

Overall, the results concluded from medical data have shown that, from survival curves for the two failure times for hormone (treated with hormone & without hormone), RT (treated with RT and without RT), surgery (treated with surgery and without surgery) and chemotherapy (treated with chemotherapy and without chemotherapy) are reasonable. In additional to that the breast cancer patients treated by hormone, RT, surgery and chemotherapy respectively, have more chances to survive longer compared with those who didn't receive these treatments as shown by Figures 2 to 5 and LRT (Table 1). This result indicates how effective the methods of treatment are in increasing the survival chances for longer period compared with other treatments.

In present research, we carried out simulation study based on the real breast cancer data. The data generated for 20000 times from the each treatments mentioned earlier in this thesis with different percentage of the exact observations that 0% (interval), 25%, 50%, 75% and 100% via simple imputations methods (mean, midpoint, and left point) to achieve the partly interval data. The shape, scale and treatment coefficient using our Cox model with log-logistic based on MLE has been estimated. It has been witnessed that the our model with different imputations methods fits the data well especially when the data is partly-interval censored. Although the left point imputation in the simulation study is not performing well especially when we have less exact observation in the data. But the consistency, basic study and comparisons of present estimations and improved parameters (when left point is used) estimation will be conducted in future research.

In a study conducted by researchers such as Kim (2003), Alharpy and Ibrahim (2013), Zyoud et al. (2016) based on PIC data, it was found that when there is more exact observations in the data are more likely to be correct and effective, the results become more stable. Similarly, we have obtained the same conclusion as theirs, hence, our

methods is strongly stable and flexible for PIC breast cancer compares to interval data.

5.2 Suggestion for Future Research

This study on breast cancer has utilized tools such as p-value, likelihood ratio test as well as the Cox model to help us arrive at the significance of the statistical decision. Once we can utilize these, we can make an educated analysis of the test.

There are a few scopes for further research.

1. The biasedness of the sample needs to be checked to ensure there is no sampling error in this case.
2. More research is need when left point imputation are used.

REFERENCES

Abbott, R. D. (1985). "Logistic Regression In Survival Analysis." *American Journal of Epidemiology*, *121*(3):465–471. doi: 10.1093/oxfordjournals.aje.a114019.

Alharphy and Ibrahim (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.

Bayaga, A (2010). "Multinomial Logistic Regression: Usage and Application in Risk Analysis." Journal of Applied Quantitative Methods. Vol 5(2)288-297.

Bennett, S. (1983). Log-Logistic Regression Models for Survival Data. *Applied Statistics*, *32*(2):165-171. https://doi.org/10.2307/2347295

Bender, R., Augustin, T., & Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine*, *24*(11), 1713–1723. doi: 10.1002/sim.2059

Christiana Kartsonaki (2016) Survival analysis. *Science direct.* doi:https://doi.org/10.1016/j.mpdhp.2016.06.005

Citho. D, Milewska. A, Wasilewska. J, and Kaezmarski M. (2012). "Ordinal logistic regression for the analysis of skin test reactivity to common aeroallergens." *Studies in Logic, Grammar and Rhetoric.* Vol 29(42): 87-98.

Chatterjee, D., & Chatterjee, A. (2010). Binary Logistic Regression Using Survival Analysis. *SSRN Electronic Journal.* doi:10.2139/ssrn.1672759

Dugue, A. E., Pulido, M., Chabaud, S., Belin, L., & Gal, J. (2016). How to Deal with Interval-Censored Data Practically while Assessing the Progression-Free Survival: A Step-by-Step Guide Using SAS and R Software. *Clinical Cancer Research*, *22*(23), 5629–5635. doi: 10.1158/1078-0432.ccr-16-1017

Eagle and Barnes (2014), Survival Analysis on Duration Data in Intelligent Tutors,

Intelligent Tutoring Systems, 8474, pp 178-187

Emmert-Streib, F., & Dehmer, M. (2019). A Machine Learning Perspective on Personalized Medicine: An Atomized, Comprehensive Knowledge Base with Ontology for Pattern Recognition. *Machine Learning and Knowledge Extraction*, *1*(1), 149–156. doi: 10.3390/make1010009

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

Elfaki, Abobakar, Azram and Usman (2013), Survival Model for Partly Interval-Censored Data with Application to Anti D in Rhesus D Negative Studies, International Scholarly and Scientific Research & Innovation, 7, pp 347-350.

Eagle and Barnes (2014), Survival Analysis on Duration Data in Intelligent Tutors, *Intelligent Tutoring Systems, 8474*, pp 178-187

Freund, J. E., Miller, I., Miller, M., & Freund, J. E. (1999). *John E. Freunds mathematical statistics*. Upper Saddle River, NJ: Prentice Hall.

Finkelstein, D. M. (1986). A Proportional Hazards Model for Interval-Censored Failure Time Data. *Biometrics*, *42*(4), 845. doi: 10.2307/2530698

Gupta, R. C., Akman, O., & Lvin, S. (1999). A Study of Log-Logistic Model in SurvivalAnalysis.*BiometricalJournal*,*41*(4),431–443.doi:10.1002/(sici)1521-4036(199907)41:4<431:aidbimj431>3.0.co;2-u.

Giolo (2004), Turnbull's Nonparametric Estimator for Interval-Censored Data. Department of Statistics, Federal University of Paraná, researchgate.net.

Guure, Ibrahim and Adam (2012), On partly censored data with the Weibull distribution. *Science Asia*, 39S: 75–79.

Harrell, F. R. A. N. K. (2016). *Regression Modeling Strategies: with applications to linear models, logistic and ordinal ... regression, and survival analysis*. Place of publication not identified: SPRINGER.

Hashemian, Garshasbi, & Pourhoseingholi (2017). A comparative study of COX and Log-Logistic (with and without Frailty) models comparison study of colorectal cancer survival estimation. *Research gate*. doi: 10.4314/jmbs.v6i2.2

Huang. (1999). Asymptotic properties of nonparametric estimation based on partly interval-censored data. *StatisticaSinica*.

Hazler, K. R. (2004). Mayfield Logistic Regression: A Practical Approach For Analysis Of Nest Survival. *The Auk*, *121*(3), 707. doi: 10.1642/0004-8038(2004)121[0707:mlrapa]2.0.co;2

Kalbfleisch, J. D., & Prentice, R. L. (2003). The statistical analysis of failure time data. New York: J. Wiley.

Khan, M. H., & Ababneh, F. (2016). Efficiency Of Weibull Regression Model Over Cox Regression Model: A Simulation Study. JP Journal of Biostatistics, 12(2), 169-178. doi:10.17654/jpjbdec2015_169_178

Khan, S. A., &Khosa, S. K. (2016). "Generalized log-logistic proportional hazard model with applications in survival analysis."*Journal of Statistical Distributions and Applications*, 3(16):1-18. doi: 10.1186/s40488-016-0054-z

Kim (2003), Maximum likelihood estimation for the proportional hazards model with partly interval-censored data. J R. Statist. Soc., Series B65, pp 489-502.

Kleinbaum, D. G., & Klein, M. (2005). Survival Analysis. *Statistics for Biology and Health*. doi:10.1007/0-387-29150-4

Lawless (2003), Statistical Models and Methods for Lifetime Data. John Wiley & Sons, Canada.

Liu, X. (2012). Survival Analysis. doi:10.1002/9781118307656

Lu, W., & Tsai, T.-R. (2009). Interval censored sampling plans for the log-logistic lifetime distribution. *Journal of Applied Statistics*, *36*(5), 521–536. doi: 10.1080/02664760802554180

Langlands, A. O., Pocock, S. J., Kerr, G. R., & Gore, S. M. (1979). Long-term survival of patients with breast cancer: a study of the curability of the disease. *Bmj*, *2*(6200), 1247–1251. doi: 10.1136/bmj.2.6200.1247

Lawless, J. F. (2003). *Statistical models and methods for lifetime data*. New York: John Wiley & Sons.

Meng, L. (2016). Method for Computation of the Fisher Information Matrix in the Expectation–Maximization Algorithm.

Morris, T. P., White, I. R., &Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine,38*(11), 2074-2102. doi:10.1002/sim.8086

Abraham, A. (2009). *Data mining*. Berlin: Springer.

Peto R. and Peto, J. (1972), Asymptotically Efficient Rank Invariant Test Procedures.” JR. Statist. Soc., Series A135, pp. 187-220.

Pan, W. (2000). A Multiple Imputation Approach to Cox Regression with Interval-Censored Data. *Biometrics*, *56*(1), 199–203. doi: 10.1111/j.0006-341x.2000.00199.x

Prinja, S., Gupta, N., & Verma, R. (2010). Censoring in clinical trials: Review of survival analysis techniques. *Indian Journal of Community Medicine*, *35*(2), 217. doi: 10.4103/0970-0218.66859

Plank, DeLuca and Estacion (2008), High School Dropout and the Role of Career and Technical Education A Survival Analysis of Surviving High School, *Sociology*

of Education, 81, pp 345-370.

Sam and Krongs (2008), Survival Analysis Approach to Reliability, Survivability and Prognostics and Health Management (PHM). IEEE Aerospace Conference.

Rezaei, S., Hashami, S., & Najjar, L. (2014). Extended Exponential Geometric Proportional Hazard Model. *Annals of Data Science*, *1*(2), 173–189. doi: 10.1007/s40745-014-0013-9.

Singh, & Totawattage. (2013). The Statistical Analysis of Interval-Censored Failure Time Data with Applications. *Open Journal of Statistics* .doi: 10.4236/ojs.2013.32017.

Saeed, N. M., &Elfaki, F. A. M. (2020). Parametric Weibull Model Based on Imputations Techniques for Partly Interval Censored Data. *Austrian Journal of Statistics*, *49*(3), 30–37. doi: 10.17713/ajs.v49i3.1027

Senaviratna, N. A., & Cooray, T. M. (2019). Diagnosing Multicollinearity of Logistic Regression Model. *Asian Journal of Probability and Statistics,* 1-9. doi:10.9734/ajpas/2019/v5i230132

Singer and Willett (1993), It's About Time: Using Discrete-Time Survival Analysis to Study Duration and the Timing of Events. Journal of Educational Statistics, 18(2), pp 155-195.

Saeed, N. M., & Elfaki, Faiz. (2020). Parametric Weibull Model Based on Imputations Techniques for Partly Interval Censored Data. *Austrian Journal of Statistics,49*(3), 30-37. doi:10.17713/ajs.v49i3.1027

Tang. S, Jeong. J, and Song. C (2017). "Fractional Logistic Regression for Censored Survival Data." *Journal of Statistical Research*, Vol 51(2):101-114.

Yuan Wu, Christina D.Chambers, Ronghui Xu (2017), Semiparametric Sieve Maximum Likelihood Estimation Under Cure Model with Partly Interval

Censored and Left Truncated Data for Application to Spontaneous Abortion Data. 1708.06838, https://arxiv.org/abs/1708.06838.

Weybright, Caldwell, Xie, Wegner and Smith, (2017), Predicting secondary school dropout among South African adolescents: A survival analysis approach. South African Journal of Education, 37, pp 1-11.

Yousif. Y, Elfaki. Faiz, 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.

Yin, Huang, Li and Zhou (2013), A Survival Modeling Approach to Biomedical Search Result Diversification Using Wikipedia, IEEE Transactions on Knowledge and Data Engineering, 25, pp 1201-1212.

Zyoud. A, Elfaki. Faiz and Hrairi. M. (2016), Non-parametric estimate based in imputations techniques for interval and partly interval censored data, Science International (Lahore), 28, pp 879-884.

Zhao, Zhao, Sun and Kim (2008), Generalized Log-Rank Tests for Partly Interval-Censored Failure Time Data. *Biometrical Journal, 50 (3), pp 375-385.*