QATAR UNIVERSITY

COLLEGE OF ARTS AND SCIENCES

ACCELERATED FAILURE TIME FOR WEIBULL DISTRIBUTION BASED PARTLY INTERVAL CENSORED DATA

BY

IBRAHIM ALI ELFAKY

A Project 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 Project of

Ibrahim Ali Elfaky defended on 26/11/2020.

_____

Dr. Faiz Ahmed Elfaki
Thesis/Dissertation Supervisor

_____

Dr. Reza Pakyari
Co-Supervisor

ABSTRACT

ELFAKY, IBRAHIM, ALI., Masters: January : 2021, Applied Statistics

Title: Accelerated failure Time for Weibull Distribution Based Partly Interval Censored Data

Supervisor of Project: Faiz Ahmed Mohamed Elfaki.

In this project, the performance of maximum likelihood estimators of the parameters of Accelerated Failure Time (AFT) regression model based on Weibull distribution with simple imputations methods under Partly-Interval Censored (PIC) data is studied and compared with semiparametric Cox model. From a real data set, the results indicate that the AFT with Weibull distribution is comparable with Cox model under PIC breast cancer data. This result suggests that the parameters of the model are stable and the treatments are significant for breast cancer patients.

Hence, maximum likelihood estimation is an appropriate method for estimating the parameters of our Weibull AFT based on simple imputation method under PIC data. In the simulation study, using the AIC and LRT with their p-values, the results show that our model is fit well and flexible under PIC data especially for exact observation. This finding has led us to deduce the fact that the AFT with Weibull distribution can be useful for modeling PIC data.

DEDICATION

I would like to dedicate this work to my family. To my parents, to my friends who have always encouraged me to pursue my goals with patience. The completion of this work would have never been possible without them.

ACKNOWLEDGMENTS

First, I would like to express my heartfelt gratitude to my supervisor, Dr. Faiz Elfaki, for giving me an opportunity to work under his supervision and for his academic guidance, support and encouragement during my research work. We also thank the members of the co-supervisor Dr. Reza Pakyari for their assistance, valuable comments and support. I would also like to thank my co-workers for all their support and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

In this chapter, will presented the survival study, censored, accelerated failure time model. The chapter also presents the statement of the problem, the purpose and the study scope.

1.1 Introduction and Background

Survival analysis is a branch of statistics that is defined as a set of methods and techniques for analyzing the duration of time until the occurrence of one or more events of interest. In other words, survival analysis is studying the analysis of data by a set of techniques until the event has happened and the result is the time taken for it to happen. There is common *parametric* model in survival analysis underline the exponential, Weibull, Gompers, Lognormal and Log logistic distributions, etc. The Accelerated Failure Time (AFT) and Proportional Hazards Regression Model (PHRM) are derived from a Log logistic (Wei 1992). There are important *nonparametric* methods in analysis of survival such as: Kaplan Meier estimator, PHRM and log-rank test. The Kaplan Meier estimator is use to evaluation the survival probabilities for example the function of time. While PHRM is the most common approach for modeling survival data.

The purpose of this project is to study one of the important models in industrial fields and clinical trials that AFT model which is a flexible model that provides an alternative to the commonly used PHRM (Wei 1992; Saikia and Barman 2017). It is used to study the reliability of industrial products. In the past a PHRM model was used for statistical modeling of survival data. The AFT model is also known as the log-location scale model given by Lawless (1982). One of the assumptions of it is that the effect of covariates acts multiplicatively (proportionally) with respect to the survival time.

1.2 Accelerated Failure Time Model

It's a huge challenge when estimating the regression parameters of the AFT model in the presence of PIC since the actual failure times are not observed directly and their exact values are unknown. In the literature suggested that the parameter estimating methods are mainly the basis of the estimated distribution function of censored observations through the censored intervals and assumed the examination times that to be laid between censored failure time. However, if the actual failure times are not available, parametric method for estimating the parameters is very appealing to researchers since it does not involve complicated statistical assumptions about the distribution of the variables in the model and the error terms of the model.

Several researches used AFT model in their studies such as; Pike (1966) studied the AFT model based on the carcinogenesis data. Nelson and Hann (1972) used AFT to determine the relationship between failure time data and temperature. Kalbfleisch and Prentice (2002), introduced the semi- parametric class of survival model, which was the class of log-linear models for time. In AFT model, the covariate effects act multiplicatively on survival time. Both the PHRM and AFT models are regression models.

Brown and Wang (2007) propsed weighted rank estimators that involoes smoothing method based on the inference procedure for semiparametric AFT models. While Johnson and Strawderman (2009) extended the method of Brown and Wang (2007) for semiparametric AFT model based on Newton Raphson algorithm and other common numerical methods.

Zheng and Yu (2011) and Zhou (2005) used semiparmetric AFT models to analyzing multivarite failure time data. Also, Wang and Fu (2011) used Gehan weight function based multivariate estimating procedure for modeling clusters and censored

failure time data. Jin (2016) used semiparmetric AFT model for right censored data. Shankar et al (2012) discussed the application of acute liver failure patients for AFT models based on the maximum likelihood estimation method. They found that when the distribution of survival time is known the best alternative model is the AFT model. Vallinayagam et al (2014) proposed the parametric regression models based on loglogistic and Weibull distributions for breast cancer data. In addtional to that Saikia1 and Barman (2017) reivewed the AFT models compared to others models based on Akaike Information

Criterion (AIC) and Baysian Information Criteria (BIC), and the found that AFT is a good alternative to PHRM. Liu and Lim (2018) predicted time to health events based on Weibull AFT regression mode. However, in this research project we will use parametric AFT model based on Weibull distribution for cancer PIC data.

Liu and Lim (2018) introduced that the Weibull Accelerated Failure Time model for a subject $i = 1, 2, ..., n$, which takes the form

$$\log T_i = \theta' z_i + \sigma \varepsilon_i \qquad (1.1)$$

where random variables $T_1, T_2, ..., T_n$ denote the failure times of n independent subjects, $z_1, z_2, ..., z_n$ denote their associated $p \times 1$ vector of covariates, θ is a $p \times 1$ vector of unknown regression parameters, σ is a scale parameter and ε_i 's are the error terms of the model, which are independent and identically distributed according to a Gumbel.

The functions for the extreme value distribution of the probability density (pdf), cumulative (cdf), survival (S(t)) and hazard (h(t)) based on AFT model given below in equations (1.1) to (1.5). However, the graphs of respective functions for the pdf, cdf, survival and hazard based on AFT model are given in Figures 1 to 4.

$$f_{\varepsilon}(x) = e^{z} e^{-e^{z}} \tag{1.2}$$

$$F_{\varepsilon}(x) = 1 - e^{-e^{z}} \tag{1.3}$$

$$S_{\varepsilon}(x) = 1 - F_{\varepsilon}(x) = 1 - (1 - e^{-e^{z}}) = e^{-e^{z}} \tag{1.4}$$

$$h_{\varepsilon}(x) = \frac{f_{\varepsilon}(x)}{S_{\varepsilon}(x)} = \frac{e^{z}\, e^{-e^{z}}}{e^{-e^{z}}} = e^{z} \tag{1.5}$$

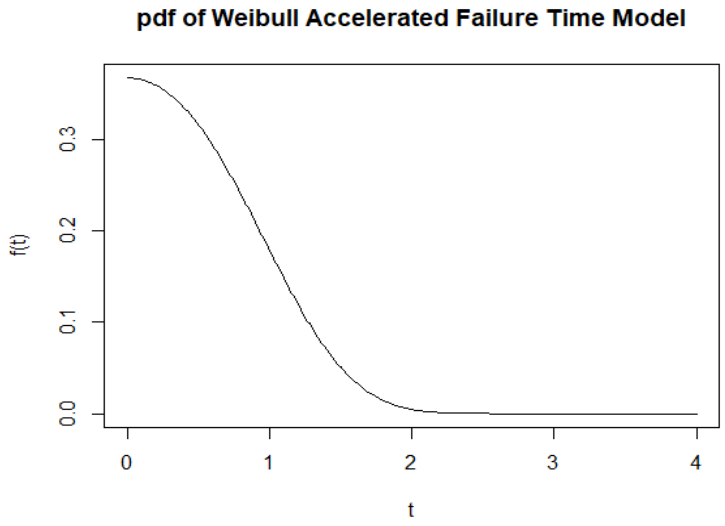

Figure 1: PDF of Weibull Accelerated Failure Time Model

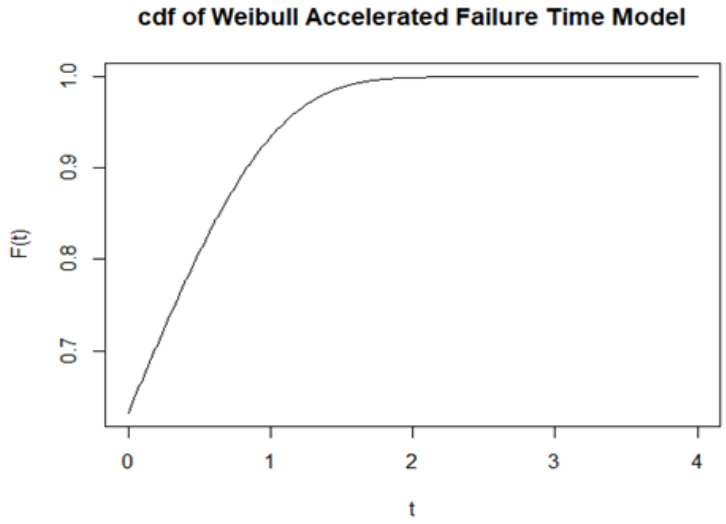

Figure 2: CDF of Weibull Accelerated Failure Time Model

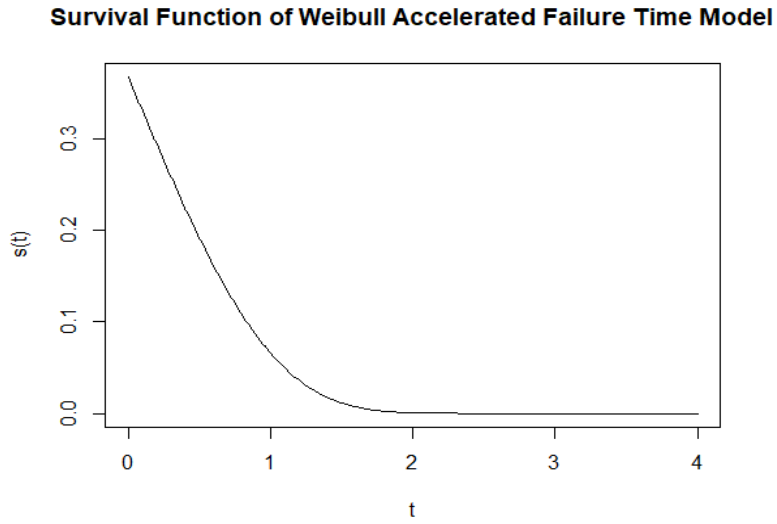

Figure 3: Survival Function of Weibull Accelerated Failure Time Model

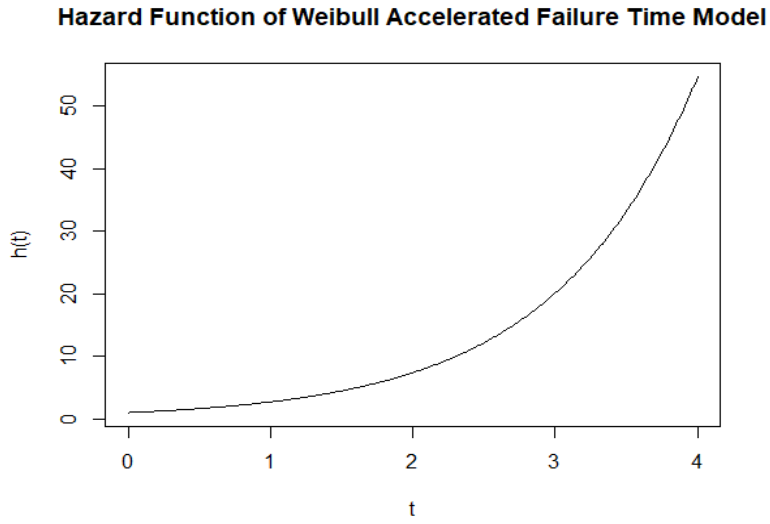

Figure 4: Hazard Function of Weibull Accelerated Failure Time Model

1.3 Censored Data

Survival data is any type of data that contains survival time, failure time, lifetime, time- to-event, time-to-respond, and characteristics of the experiment subject that are related to the response variable. Several examples of survival data in different filed such as clinical trials or biological studies (Lee and Wang 2003), and in industrial (Lawless 2011).

Censoring is a special feature for analysis of survival data and an important issue in survival analysis. Time-to-event is described as censored observation if the failure time of the subject is not observed during the study period and subject does not experience failure at the end of the study. An example of this in clinical trials when a patient is withdrawn from the experiment or is lost to follow-up in the experiment. The censored data is an important information, but it is not respective the population.

There are three main types of censoring which are right, left, and interval. Interval censoring is the combination of both right and left censoring as stated by Giolo (2004). According to the Allison (2011), this sort of censoring is likely to occur in

survival analysis when observations are made at infrequent intervals and it is difficult to get retrospective information on the exact timing of events.

We focus in our study on one type of censoring witch is the interval censoring, according to Kim (2003) PIC data consist of interval censored data also exact data . Which means that while some subject of interest lays between an interval, others exactly observed.

Partly censored has been used in several researches such as Kim (2003), Zhao.et al. (2008) used for his PIC failure data generalized log-rank test and as Peto and Peto (1972) discussed the PIC data and treated the observation for censored interval with very short interval. Elfaki et al., (2012) in the study of parametric Cox's model also used of Proportional Hazard Weibull Model in the study for PIC of the application for the AIDS study, in order to estimate the survivability of failure rate partly interval censored will be mainly used In this research.

Guure, et al., (2012) used and defined PIC is one type of censoring that occurred when units which are put to test begin at different times and the test is terminated before all unit are said to have failure, and where it is observed that there is some intertwine among censoring times and that failure times. In their study, their scopes of PIC are consisting of failure, right censoring and interval censoring observations by using Weibull distribution model with several estimating of parameter methods. Failure situation is occurred when an item stops functioning before the end of the observation study. In this study we used AFT model based on PIC data.

Interval-censored data consist of adjacent inspection times that surround an unknown failure time. We seek to determine the best estimator for the Weibull scale

parameter using interval-censored survival data. Consideration is given to the classical maximum likelihood and Bayesian estimation under squared error loss with interval censoring using non-informative prior and a proposed generalization of non-informative prior. The study is based on simulation and comparisons are made using mean squared error and absolute bias. We find that the proposed generalized non-informative prior is the preferred estimator of the scale parameter.

1.4 Problem Statement

The Accelerated Failure Time (AFT) model is considered to be semiparametric when the probability distribution of the error terms is unspecified. Even though there are many semiparametric or parametric methods that have been significant theoretical advances for estimating the parameters of the model in presence of censored data, those methods are not common in applications because they are computationally complicated or not completely reliable. This problem include the lack of reliable methods for modeling partly interval censored data, computational challenges that the estimating procedure and the problems with estimating the covariance matrix of parameter estimators.

In the literature, the most of proposed parametric methods concentrated on estimating procedures on AFT model based on right censored, left censored and few on interval censored. Therefore, making inference in the presence on partly interval censored data requires modifications of the methods that are proposed for right censored or left censored. Clearly, such methods will be very challenging for estimating the parameter based on covariance matrix and conducting hypothesis testing or constructing confidence intervals for the estimator’s parameters.

1.5 Objective

The main objectives of this research are:

- Modified the procedures for estimating the regression parameters of parametric accelerated failure time models based on partly interval censored.
- Compare the AFT with nonparametric estimation of the distribution of the model error terms such as Kaplan-Meier method.
- Evaluate the performance and operating characteristics of the parameter estimating methods through simulation studies.
- Illustrate the parameter estimating methods in applications through real data sets.

1.6 Scope of the Project

The research study is limited to use the partly interval censored data based on the AFT model to predict of the survival-ability of medical data in Qatar. This model is described in chapter III and the MLE will be used to estimate the parameters in the model. Simple imputation techniques will also be used to modify the real data set into PIC data.

The literature review of the survival analysis, partly interval censoring and Accelerated Failure Time model are presented in Chapter 2. The Accelerated Failure Time model based on survival analysis and derivation of maximum likelihood estimator for parameters will be described in chapter 3. At the end of chapter III the simple imputation techniques that used to modify the education data to be right, interval and PIC.

CHAPTER 2: LITERATURE REVIEW

CHAPERT OVREVIW

In this chapter, we will discuss some existing related literature for Accelerated Failure Time (AFT) which have been applied in various areas such as medicine and manufacture. Then, we will focus on one type of censoring data that is used in this project which is a partly interval censoring (PIC).

2.1 Accelerated Failure Time Model

A lot of researchers studied the accelerated failure time (AFT) model in their research areas, such as medical and industrial fields and other fields. Shankar et al., (2012) discussed the application of AFT models in acute Liver Failure Patients survival in India. First, the terms of uses of the model was mentioned and how to estimate the effect of covariate. They used two exploratory methods to identify the appropriate survival model, the first method was based on the shape of the baseline hazard function and through a plot of appropriately transformed survival function with the log of survival time. Also, they used the maximum likelihood estimation (MLE) method to fit the model and found that AFT models would be the best alternative if the distribution of survival time is known.

Saikia1 and Barman (2017) discussed the AFT models in terms of uses and how to estimate this model including technical development and past researches in case of survival analysis. The last idea was how to check the appropriate distribution by Akaike's information criterion (AIC), Bayesian information criterion (BIC) and Cox-Snell residual. They found out that the AFT is a good alternative for a cox PH model as it was used to study the reliability of industrial products.

Vallinayagam et al., (2014) studied the parametric regression models for breast cancer survival data processing based on AFT model. The aim of it was comparing the

quality of parametric models using records from German breast cancer. The analytical method is likelihood ratio test. Following that in their studied they found that the Weibull model was the perfect parametric mannequin match for forecasting survival following both HIV and AIDS diagnosis, leading to the conclusion of how the log-logistic AFT model is ideally suited for randomized placebo-controlled trial to avoid Tuberculosis in Uganda, adults diagnosed with HIV proposed the parametric regression models based on log-logistic and Weibull distributions for breast cancer data.

In another report, Newby (1988) addressed the AFT models for reliability data analysis. This model is a decent alternative for the proportional hazards model (PHM) to evaluate certain forms of reliability data since the PHM model is not ideal for certain circumstances. They examined the option between accelerated failure time models, PHM and describes the techniques used in the analysis of data from the fatigue crack growth experiment.

In another study, Odell et al., (1992) discussed the two ways to estimate the Weibull AFT model for interval censored data, which are the MLE and the midpoint estimate. Simulation study was used in their studied and indicating that for relatively large samples there are many instances when the MLE is superior to the midpoint estimate.

In another study, Liu and Lim (2018) debated the use of Weibull AFT regression model to estimate time for wellness activities. They identified a statistical technique to accomplish that. This tactic of prediction is popular in studies on engineering reliability but seldomly practiced for medical forecasts, such as survival time. An explanation of how to make predictions using the provided released data set, the high likelihood is the perfect mean of measuring.

2.2 Partly Interval Censored

As you can see in this chapter, many studies have used a method of distribution based on the PIC records. That is because this method is flexible when being used to estimate the parameter. The purpose of our study is to apply the AFT for Weibull distribution which is suitable for education PIC data. Also, to investigate the performance of AFT model and to ascertain its effectiveness by using suitable methods.

Several Researchers used PIC in their studies such as; Kim (2003) who utilized two techniques to appraise the MLE variance-covariance matrix of the regression parameter that summed up missing data rule and the profile data strategy based on an application to diabetes information PIC data in Denmark. In the simulation part a two techniques function was used regarding the variance for moderate samples of size and the bias. In addition, Zhao et al., (2008) considered the PIC failure data based on the log-rank test which has been studied by Peto and Peto (1972). They assessed the technique utilizing a lot of real data from a diabetes study and simulation data.

Furthermore, Guure et al., (2012) applied PIC based on Weibull distribution model via MLE and Least Square (LS) methods for estimators of one variable on other variables to decide the endurance estimator among these techniques for assessing the parameters and to show that the estimators of the parameters is a bias. They applied MSE bias to look for the difference between two techniques that depend on the reproduction study. They showed that the MLE was better for evaluating the scale parameter. Then again, the least square for the first factor was increasingly dependable for evaluating the shape parameter with generally little examples, yet with bigger examples, least square on other variables was a good strategy.

To add more, Elfaki et al., (2012) examined the parametric proportional hazard Weibull model based on the Expectation-Maximization (EM) algorithm for PIC data

with an application of AIDS study. In addition, the PIC would be used primarily in heir work to evaluate the survivability of failure rate. They investigated the treatments of the mentioned data for hemophiliacs in two hospitals in Sudan. In their study they showed that there are no differences between the two treatments in the real data set. Moreover, AL harpy and Ibrahim (2013) used parametric Weibull distribution for two testing that the Likelihood Ratio Test (LRT) and Score Test (ST) based on PIC data. They observed that the LRT is better than the score test to examine the parametric for the partly interval censored under Weibull distribution.

In another study, Elfaki et al., (2013) proposed the semiparametric PHM for PIC as a competing risk model based on EM algorithm for estimating the parameters. They used two contrasting risk models, the Censoring Complete (CC) model and the Weighting Technique (WT) model. Their studied the potential correlation between medication and the anti D in Rhesus time to study the impact of covariates on the development of complications, which have been applied to a collection of anti D time data for negative pregnant women in Rhesus D in Sudan. The study concluded that the covariates do not have a major difference, as the negative party had a higher chance of anti D rhesus initiation after infection. Estimating the survival function using R software. He observed that the random, mean, and median imputations are better compared with the other imputation techniques.

Wu et al., (2017) implemented the system of semi-parametric sieve MLE for PIC data analysis using Cox Regression. In their work, the non-mixed Cox Regression cure rate design was considered, and the semi-parametric spline-based sieve maximum probability strategy was adopted to analyze these results. They also illustrated the methods using conventional empiric process hypothesis for both the parametric and non-parametric aspects of the sieve estimator. Next, they observed that the sieve

estimator was consistent and simulated the data to show the performance for the proposed method. Later on, they observed that the suggested sieve MLE was satisfactory and agreed applying the introduced method on spontaneous abortion studies and have it applied successfully.

Zyoud et al., (2016) used non-parametric imputation-based research processes to approximate the survival function to analyze the partly interval-censored outcomes. Simple imputation includes imputation at the top, left and midpoint. On the other hand, chance-dependent modes of imputation include mean imputation, median imputation, conditional mode, multiple imputation, and random imputation. Their proposal to approximate the survival function was made using R devices. They found that natural, standard, and median imputations are stronger relative compared to other imputation strategies.

In conclusion, most of the previous researches studied the AFT model including its interval censored. Accordingly, this is the reason why in our present research we are trying to adopt this approach by using the same model only with partly interval censored.

CHAPTER 3: METHODOLOGY

CHAPTER OVERVIEW

This chapter shows the estimation of the regression parameters of Accelerated Failure Time model and estimation of parameters for the Weibull AFT model with covariate using maximum likelihood estimator with Partly Interval Censored data. The chapter also, observe the maximum likelihood estimator and, Newton's Raphson method and describes the real data in this study and the process to analyses the data to survival time data.

3.1 Introduction to Accelerated Failure Time model

The Accelerated Failure Time model is one of the continuous probability distributions. Model of AFT has appeared to be a time-failure model that can be utilized to evaluate time to event information. It is designed to test the impact of covariate on accelerate or decelerate time of survival. The covariate impact can be multiplied on the AFT time scale, while on the hazard scale it can be multiplied in the hazard models of proportional. This model is a letdown time model but can be used sometimes for the assessment of time to date of occasion. The model tries to check the effect of covariate to decelerate or accelerate a for time of endurance. The effect of covariate is able to multiply on AFT model time scale, however it is multiplicative in models of relative hazard scale. The model is shall be adapted by applying the most extreme probability technique of estimation by utilizing Newton Raphson iterative strategy.

The density likelihood role of AFT model for the exceptional value appropriation is

$$f_{\varepsilon}(x) = \mathrm{e}^{x}\mathrm{e}^{-\mathrm{e}^{x}} \qquad (3.1)$$

The probability density function for are independent and identically distributed of the AFT model based on Weibull distribution.

The role of survival for the extra ordinary value of AFT model distribution is

$$S_{\varepsilon}(x)=P\{X>x\}$$

$$S_{\varepsilon}(z)=1-P\{X\leq x\}$$

$$S_{\varepsilon}(x) = 1- F\varepsilon(x)$$

$$S_{\varepsilon}(x) =1- (1 - e^{-e^{x}})$$

$$S_{\varepsilon}(x)= e^{-e^{x}} \qquad (3.2)$$

The hazard function for the extreme value distribution of AFT model is

$$h_{\varepsilon}(x) = \frac{f_{\varepsilon}(x)}{S_{\varepsilon}(x)} = \frac{e^{x}e^{-e^{x}}}{e^{-e^{x}}}$$

$$h_{\varepsilon}(x) = e^{x}e^{-e^{x}}e^{e^{x}} = e^{x} \qquad e^{x} > 0 \qquad (3.3)$$

where ε and x zare defined earlier in equation (1.1).

3.2 Research Methods

It's a huge challenge when estimating the regression parameters of the AFT model in the presence of PIC since the actual failure times are not observed directly and their exact values are unknown. In the literature suggested that the parameter estimating methods are mainly the basis of the estimated distribution function of censored observations through the censored intervals and assumed the examination times that to be laid between censored failure time. However, if the actual failure times are not available, parametric method for estimating the parameters is very appealing to researchers since it does not involve complicated statistical assumptions about the

distribution of the variables in the model and the error terms of the model.

Let random variables $T_1, T_2, \dots, T_n$denote the failure times of n independent subjects and let $z_1, z_2, \dots, z_n$denote their associated $p \times 1$vector of covariates. The AFT Weibull model proposed by Liu and Lim (2018) for $i = 1,2, \dots, n$ takes the form

$$log\, T_i = \theta' z_i + \sigma \varepsilon_i \qquad (3.4)$$

Where θ is a $p \times 1$vector of unknown regression parameters, σ is a scale parameter and ε_i's are the error terms of the model, which are both independent and identically distributed. parameter of AFT Weibull model can be estimated by using maximum likelihood estimator method. The probability role of the n perception log (t) is given by Lim and Liu (2018). and the formula is

$$L(\theta, \sigma; z_i) = \prod_{i=1}^{n} [f_Z\,(z_i)]^{\delta_i} [S_Z(z_i)]^{1-\delta_i} \qquad (3.5)$$

where

$S_Z(z_i)$ is the survival function for the ith individual at the time ti respectively.

$f_Z(z_i)$ is the probability density function of Z i.e $log\, T_i$,

$$\delta_i = \begin{cases} 1 & if\ an\ event\ has\ occurred \\ 0 & if\ the\ event\ has\ not\ occurred \end{cases}$$

δ_i is the event indicator for the ith subject with $\delta_i = 1$if an event has occurred

$\delta_i = 0$ if the event has not occurred.

Presently the work of survival for the ith individual (during that point) related p×1vector of covariates which is equivalent to the work of the survival for the outrageous worth distribution $\{\varepsilon i\}$.

$$S_Z(z_i) = S_{\varepsilon i}(z_i) \qquad (3.6)$$

where

$$z_i = \frac{\log T - \mu - \beta 1 X1 - \cdots \beta pXp}{\sigma} \qquad (3.7)$$

S_Z is a survival for the ith individual, $S_{\varepsilon i}$ is a survival for the outrageous worth distribution and β_1,……..β_p are the regression coefficients of interest, z_i is called the ith individual at the time t_i (which is associated vector of covariates).

3.3 Maximum Likelihood Estimators

The strategy for maximum likelihood is the most famous procedure for inferring estimators and has a lot of application. Maximum likelihood estimation (MLE)is the parameter esteem for which the watched test is no doubt comparative. There are a lot of benefit to use it. One of them is favorable position of utilizing it is available a predictable way to deal with parameter estimation issues. This implies most extreme probability evaluations can be created for an enormous assortment of estimation circumstances. Another preferred position of the most maximum likelihood estimators is having alluring numerical and optimality properties, particularly when the example size expanded the base change will be fair-minded estimators.

According to the formal given by Lim and Liu (2018), the first step of maximum likelihood estimation is the parameter esteem for L, which is the logarithm of function L.

$$\log L = \sum_{i=1}^{n} \sigma \frac{\mathrm{logT} - \mu - \beta 1 \mathrm{X1} - \cdots \beta \mathrm{pXp}}{\sigma}^{-\delta_i} \mathrm{f}_{\varepsilon \mathrm{i}} \left(\frac{\mathrm{logT} - \mu - \beta 1 \mathrm{X1} - \cdots \beta \mathrm{pXp}}{\sigma} \right)^{\delta_i} \ldots$$

$$\ldots \mathrm{s}_{\varepsilon \mathrm{i}} \left(\frac{\mathrm{logT} - \mu - \beta 1 \mathrm{X1} - \cdots \beta \mathrm{pXp}}{\sigma} \right)^{1-\delta \mathrm{i}}$$

$$\log L$$

$$= \sum_{i=1}^{n} -\sigma \log(\sigma \frac{\log T - \mu - \beta 1 X1 - \cdots \beta pXp}{\sigma}) + \delta_i \log f_{\varepsilon i}\left(\frac{\log T - \mu - \beta 1 X1 - \cdots \beta pXp}{\sigma}\right)$$

$$+ (1 - \delta i) \log s_{\varepsilon i}(\frac{\log T - \mu - \beta 1 X1 - \cdots \beta pXp}{\sigma})$$

The MLE of μ, and $\beta 1$,……..βp unknown parameters are found by maximizing the log-likelihood function using Newton Raphson procedure.

σ is a scale parameter

ε is the random disturbance terms

3.4 Estimation of parameters for the Weibull AFT model without covariates

The probability density function of Weibull AFT model is

$$F(y: \mu, \sigma) = \frac{1}{\sigma} e^{(\frac{(y-\mu)}{\sigma})} e^{(-\frac{e(y-\mu)}{\sigma})} \quad where \ (-\infty < y < \infty) \qquad (3.8)$$

If the εi follows the extreme value distribution than T_i follows the Weibull distribution The survival function and density function of the extreme value distribution are respectively.

The survival function Weibull AFT model is :

$$S(z_i) = e^{(-e(z_i))} \qquad (3.9)$$

Take a Log of survival function we will have:

$$\log(S_Z(z_i)) = -e(z_i)$$

Take a differentiating of the log of survival function respective to z_i to fine the PDF.

$$\frac{d \log(s(z_i))}{dz_i} = -e^{(z_i)}$$

The probability density function of Weibull AFT model as follow:

$$F_{\epsilon i}(z_i) = e^{(z_i)} e^{(-e(z_i))} \qquad (3.10)$$

Take a log of PDF and diffrentating respective to z_i to help find the estimator.

$$\log\left(F_{\epsilon i}(z_i)\right) = z_i - e^{(z_i)}$$

The differentiating respective to z_i to help to find the estimator the and σ .

$$\frac{d\log\left(F_{\epsilon i}(z_i)\right)}{dz_i} = 1 - e^{(z_i)} \qquad (3.11)$$

Consider the Weibull Accelerated Failure Time Model is AFT models are fitted by using maximum likelihood estimation (MLE) method

Now

$$S_i(t_i) = S_{\varepsilon i}(z_i)$$

$$l = \frac{1}{\sigma}\prod_{i=1}^{n}\left[(f_Z(z_i))^{\delta_i}\left(S_Z(z_i)\right)^{1-\delta_i}\right]$$

The t_i are the failure times of i independent subjects

Where $(f_Z(z_i))$ and $\left(S_Z(z_i)\right)$ are the density function and the survival function for the ith individual at the time zi respectively. δi is the event indicator for the ith individual. Where,

$$z_i = logT - \mu - \beta_1 x_1 - \cdots \beta_p xp\sigma$$

The likelihood function for Weibull ACCELERATED FAILURE TIME MODEL is given by

$$\log L = \sum_{i=1}^{n} \delta_i \, log\left(\frac{1}{\sigma}\right) + \delta_i \log[F_{\epsilon i}\,(z_i)) + (1-\delta_i)\log\left(s_{\epsilon i}(z_i)\right)$$

If the εi follows the extreme value distribution then Ti follows the Weibull distribution.

$$\log l = -\,r log\,\sigma + \delta_i \sum_{i=1}^{n} \log[F_{\epsilon i}\,(z_i)) + (1-\delta_i)\sum_{i=1}^{n}\log\left(s_{\epsilon i}(z_i)\right)$$

$$\log l = -\,r log\,\sigma + \delta_i \sum_{i=1}^{n} \log\left(e^{(z_i)}e^{(-\mathrm{e}(z_i))}\right) + (1-\delta_i)\sum_{i=1}^{n}\log\left(e^{(-\mathrm{e}(z_i))}\right)$$

Then take the log-likelihood function, we can have;

$$log\ L = \sum_{i=1}^{n} \delta_i\ log\left(\frac{1}{\sigma}\right) + \delta_i\ log[F_{\epsilon i}\ (z_i)) + (1 - \delta_i)\ log\big(s_{\epsilon i}(z_i)\big)$$

$$log\ l = -\ rlog\ \sigma + \delta_i \sum_{i=1}^{n} log[F_{\epsilon i}\ (z_i)) + (1 - \delta_i) \sum_{i=1}^{n} log\big(s_{\epsilon i}(z_i)\big)$$

The log-likelihood function is

$$log\ l = -\ rlog\ \sigma + \delta_i \sum_{i=1}^{n} log\big(e^{z_i} e(-e^{z_i})\big) + (1 - \delta_i) \sum_{i=1}^{n} log\big(e^{-e^{z_i}}\big)$$

$$log\ l = -\ rlog\ \sigma + \delta_i \left[\sum_{i=1}^{n} l(log(e^{z_i})) + \sum_{i=1}^{n} log\big(e^{-e^{z_i}}\big)\right]$$

$$+ \sum_{i=1}^{n} log\big(e^{-e^{z_i}}\big) - \delta_i \sum_{i=1}^{n} log\big(e^{-e^{z_i}}\big)$$

$$log\ l = -\ rlog\ \sigma + \delta_i \left[\sum_{i=1}^{n} z_i + \sum_{i=1}^{n} -e^{z_i}\right] + \sum_{i=1}^{n} -e^{z_i} + \delta_i \sum_{i=1}^{n} e^{z_i}$$

$$log\ l = -\ rlog\ \sigma + \delta_i \sum_{i=1}^{n} z_i - \delta_i \sum_{i=1}^{n} e^{z_i} - \sum_{i=1}^{n} e^{z_i} + \delta_i \sum_{i=1}^{n} e^{z_i}$$

$$log\ l = -\ rlog\ \sigma + \delta_i \sum_{i=1}^{n} z_i - \sum_{i=1}^{n} e^{z_i} \qquad (3.12)$$

Diffrentating of the log of survival function respective to $z_i$to help to get the estimator.

$$\frac{d\log l}{dz_i} = 0 + \delta_i \sum_{i=1}^{n} (1 - e^{z_i}) + (1 + \delta_i) \sum_{i=1}^{n} -e^{z_i}$$

$$\frac{d\log l}{dz_i} = \delta_i \sum_{i=1}^{n} 1 - \sum_{i=1}^{n} e^{z_i} = \delta_i \sum_{i=1}^{n} (1 - e^{z_i}) + (1 - \delta_i) \sum_{i=1}^{n} -e^{z_i}$$

$$\frac{d\log l}{dz_i} = \delta_i \sum_{i=1}^{n} 1 - \delta_i \sum_{i=1}^{n} e^{z_i} + \sum_{i=1}^{n} -e^{z_i} + \delta_i \sum_{i=1}^{n} e^{z_i}$$

$$\log L = r log\, \sigma + \textstyle\sum_{i=1}^{n} \delta_i \frac{\boldsymbol{y-\mu}}{\boldsymbol{\sigma}} + \sum_{i=1}^{n} -e^{\frac{\boldsymbol{y-\mu}}{\boldsymbol{\sigma}}} \qquad (3.13)$$

Differentiating this equation (13) with respect of parameters μ and $\boldsymbol{\sigma}$ one can get the parameters values of Weibull AFT mode.

In this way the parameters of the AFT mode can be estimate

$$\log l = r log\, \sigma + \sum_{i=1}^{n} \delta_i \frac{\boldsymbol{y-\mu}}{\boldsymbol{\sigma}} + -\delta_i \, \mathrm{e}^{\frac{\boldsymbol{y-\mu}}{\boldsymbol{\sigma}}}$$

$$\log L = r log\, \sigma + \textstyle\sum_{i=1}^{n} \delta_i \frac{\boldsymbol{y-\mu}}{\boldsymbol{\sigma}} + \sum_{i=1}^{n} -\mathrm{e}^{\frac{\boldsymbol{y-\mu}}{\boldsymbol{\sigma}}} \qquad (3.14)$$

Now the Differentiating of this equation with respect of parameters μ is

$$\frac{d\log l}{d\boldsymbol{\mu}} = 0 - 1 \sum_{i=1}^{n} \frac{\delta_i}{\boldsymbol{\sigma}} - (-\frac{1}{\sigma})\mathrm{e}^{\frac{\boldsymbol{y-\mu}}{\boldsymbol{\sigma}}}$$

$$\frac{d\log l}{d\boldsymbol{\mu}} = -\frac{1}{\sigma}(\sum_{i=1}^{n} \delta_i - \sum_{i=1}^{n} \mathrm{e}^{\frac{\boldsymbol{y-\mu}}{\boldsymbol{\sigma}}})$$

Then, make the equation equal zero .

$$\frac{d\log l}{d\boldsymbol{\mu}} = 0$$

$$\textstyle\sum_{i=1}^{n} (\delta_i - e^{\frac{\boldsymbol{y-\mu}}{\boldsymbol{\sigma}}}) = 0 \qquad (3.15)$$

The Differentiating of this equation with respect of parameters σ is

$$\frac{d\log l}{d\sigma} = \frac{r}{\sigma} - \sum_{i=1}^{n} \delta_i(-\frac{y-\mu}{\sigma}) + \sum_{i=1}^{n} -(-\frac{y-\mu}{\sigma})\mathrm{e}^{\frac{\boldsymbol{y-\mu}}{\boldsymbol{\sigma}}}$$

$$\frac{d\log l}{d\sigma} = \frac{r}{\sigma} + \sum_{i=1}^{n} \delta_i(\frac{y-\mu}{\sigma}) + \sum_{i=1}^{n} (\frac{y-\mu}{\sigma})\mathrm{e}^{\frac{\boldsymbol{y-\mu}}{\boldsymbol{\sigma}}}$$

$$\frac{r}{\sigma} + \sum_{i=1}^{n} \left(\frac{\boldsymbol{y-\mu}}{\boldsymbol{\sigma}}\right)\mathrm{e}^{\frac{\boldsymbol{y-\mu}}{\boldsymbol{\sigma}}} - \sum_{i=1}^{n} \delta_i \left(\frac{\boldsymbol{y-\mu}}{\boldsymbol{\sigma}}\right) = 0$$

Make the Differentiating respect of parameters σ equal zero.

$$\frac{r}{\sigma} + \sum_{i=1}^{n} \left[\frac{\boldsymbol{y-\mu}}{\boldsymbol{\sigma}}\left(e^{\frac{\boldsymbol{y-\mu}}{\boldsymbol{\sigma}}} - \delta_i\right)\right]^1 = 0 \qquad (3.16)$$

$$\log l = r log\, \sigma + \sum_{i=1}^{n} \delta_i \frac{y-\mu}{\sigma} + -\delta_i\, e^{\frac{\boldsymbol{y-\mu}}{\boldsymbol{\sigma}}}$$

The equation for differentiating this equation with respect of parameters μ equal zero is

$$\sum_{i=1}^{n} (\delta_i - e^{\frac{\boldsymbol{y-\mu}}{\boldsymbol{\sigma}}}) = 0$$

The equation for differentiating this equation with respect of parameters $\boldsymbol{\sigma}$ equal zero is

$$\frac{r}{\sigma} + \sum_{i=1}^{n} \left[\frac{\boldsymbol{y-\mu}}{\boldsymbol{\sigma}}\left(e^{\frac{\boldsymbol{y-\mu}}{\boldsymbol{\sigma}}} - \delta_i\right)\right]^1 = 0 \qquad (3.17)$$

At this point, we can solve without censored data because δ_i=1 by log function, however the system of the above equations we can be solve using Newton-Raphson method.

3.5 Imputation Approaches

Imputation approaches are once in a while utilized to change the data analysis issue. In this research project, the imputation-based data will be adjusted to be as partly-interval-censored (PIC) and interval-censored data. The purpose of this procedure is basic and there are various methods to manage the information and impute the exact observation for PIC data. There are two differing imputation kinds: numerous imputation and simple imputation. The basic-imputation strategy will be utilized in this study.

3.5.1 Simple Imputation

Basic or simple imputation is one of the most frequently used methods to handle the missing information. And since the basic type of imputation is approximate and is engaging regularly, it has been used in simple cases of perceptions. The simple imputation techniques as referenced by Zyoud, et al. (2016), have three key types which are;

1. The right limit of the interval R_i which is the right point on the right side.

2.The left limit of the interval L_i representing the left point on the lift side.

3. The midpoint of the interval, $\frac{R+L}{2}$.

CHAPTER 4: RESULTS AND ANALYSIS

CHAPERT OVREVIW

We will represent the execution of the strategies studied in the prior parts using two informational collections, the first one is breast cancer data, and the second one will be simulated data. All calculations were computed using R software.

4.1 Breast Cancer Data

One of the most common disease in Qatar is breast cancer with a count of 31% cases of women's cancer. According to Qatar Cancer Register the danger of being diagnosed with this type of cancer among ladies in the populace is 56 for each 100,000. Nonetheless, both genders men and women can be affected by this ailment. It is essential to recollect that numerous patients can recover when their tumor is analyzed and treated early.

This data set, which was taken from Hamad Medical Corporation (HMC), was applied to the proposed method in the project This medical data that is breast cancer data set is taken from Hamad Medical Corporation (HMC) in Qatar, was applied to the proposed method in the project. The data was compiled from 1/2/2016 to 1/9/2020 including 24 variables. The first case was on 1/2/2016 and the last case was on 1/19/2020. The data included 1008 patients, 770 patients were treated with surgery, 557 with chemotherapy, 555 patients treated with hormones and 553 patients were treated with radiotherapy (TR). There are overlap between all treatments, which are 237 patients. Therefore, dummy variables accordingly will be use in our analysis in this section.

.In this project, patients were seen at Al-Amal center visits, the occasion of premium was the opportunity to first event of breast withdrawal, if the dates of the occasion were available, they get recorded precisely and if not in any case the time

period were note. The fundamental goal is to evaluate the cosmetic impacts of every treatment on females with early breast cancer.

In the next section, the result of this data set will be observed dependent on the analysis of the data as interval censored and PIC. In the later section the simulation study will be considered.

As mentioned in chapter 3 that our model can be used based on AFT 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(α) as a linear function of covariates. The data was collected from 2/1/2016 to 1/19/2020 and it contains 24 variables.

Table 1: Results from breast cancer data obtanied by AFT model

Treatment	**Parameter**	**Coefficient**	**CI of 95%**	**SE**
	Shape	3.0087668	(2.851099, 3.175153)	0.0826283
	Scale	847.88573	(813.0309, 884.2347)	18.159250
Hormone	Coefficient	0.1552978	(0.111023, 0.199573)	0.0225896
RT	Coefficient	0.0801384	(0.034276, 0.126001)	0.0233999
Chemotherapy	Coefficient	0.0759168	(0.031263, 0.120569)	0.0227825

Likelihood Ratio Test (LRT) and the p-value are use in Table 1, the two test whether the model coefficients/ parameters as a group equivalent to zero. For this situation, the bigger value of LRT qualities and smaller value p-values is best for confidence in rejecting the null hypothesis. Therefore, we can conclude from LRT (715.44, 9.08e-05) that the model is fit and the coefficient, shape and scale boundaries can't be equivalent to zero as seen in Table 1. Likewise, the estimations of standard error are less than 0.05 for all treatments, so they are factually critical in nature for all the treatments such as

radiotherapy (RT), hormones and chemotherapy. This implies the odds of endurance on each treatment with the effects as appeared in the result of data points in the Table 1. In the next Figures additionally precisely gives the contrast between the effects of various treatments.

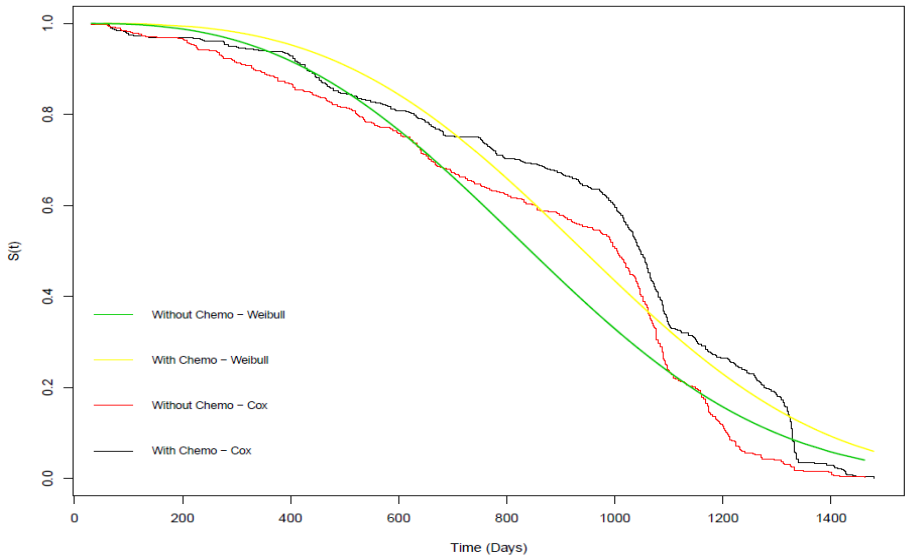

Figure 5: Estimated of the survival function based on chemotherapy treatment via AFT model.

According to Figure 5, we can conclude that the chances of survival of a breast cancer patient with AFT are almost similar to the one obtained by Cox model. In addition, the likelihood of living for a greater number of days of the patient increases with Chemotherapy. According to predictions, the probability of living longer with chemotherapy increases and decreases for interval of time as the importance of the chances of survival decreases but the result show that chemotherapy is still better. The value of LRT for chemotherapy treatment shows that there is a significant correlation

between both the data sets but while comparing among the four different treatment methods it is lowest.

The chances of survival of breast cancer patient increases based on hormone treatment as shown in Figure 6. In addition, the likelihood of living for a greater number of days of the patient increases with hormone. According to predictions, the probability of living longer with hormone increases and decreases for interval of time as the value of chances of survival decreases but the results shows that it is still always better to have hormone.

As showed in table 1, the value of LRT with their p-value for hormone treatment shows that there is a significant result that indicate we reject the null hypothesis that the coefficient of hormone treatment equal to 0 which is in line with result in figure 6.

Similarly, the estimate of survival function of a breast cancer patient based on RT treatment via AFT model are almost similar to the one obtained by Cox model as shown in Figure 7. The likelihood of living for a greater number of days of the patient increases after the patient used RT treatment. According to predictions, the probability of living longer with RT increases and decreases for interval of time as the value of chances of survival decreases but the results shows that it is still always better to have RT than without RT treatment.

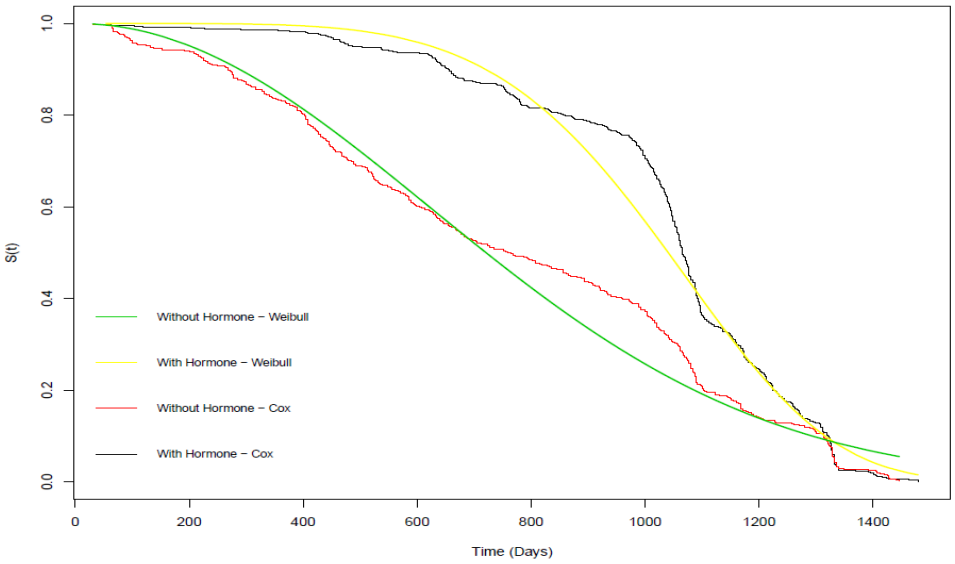

Figure 6: Eestimated of the survival function based on hormone treament via AFT model.

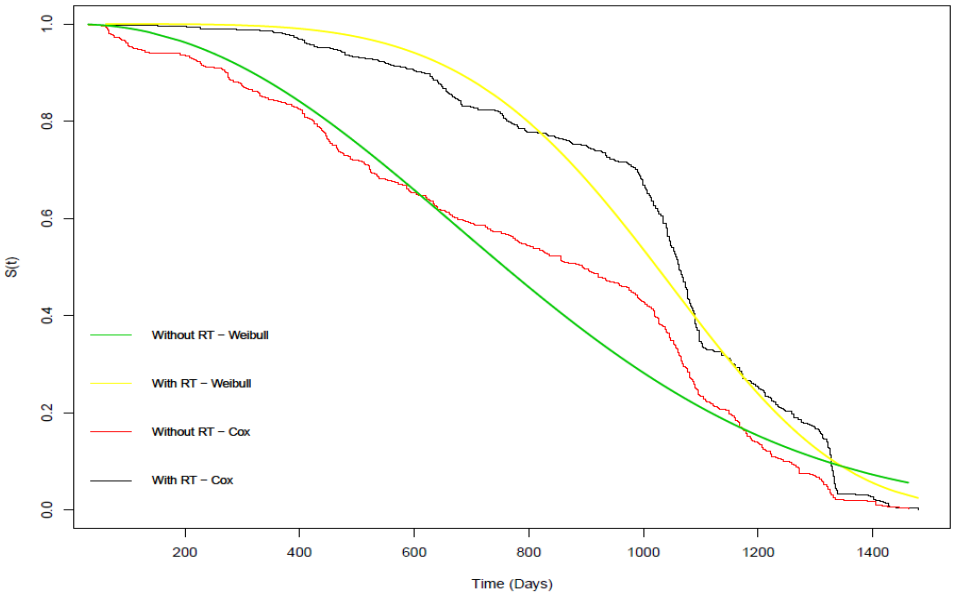

Figure 7: Eestimated of the survival function of survival based on RT treament via AFT model.

For the breast cancer patients that has been treated by the chemotherapy they have the most minimal estimation of shape boundary for example most lowest dispersion which legitimately expresses that the odds of endurance doesn't increments altogether when contrasted and a patient who haven't experienced the chemotherapy Treatment while fighting with breast cancer. The next observation is the value of LRT with their p-value for chemotherapy treatment shows that there is a no significant correlation between both the data sets also while comparing among the four different treatment methods it is lowest among four of them.

In conclusion, the outcome in this segment recommended that it valuable to utilize each of the four treatment that is the hormone as consider profoundly utilized at that point followed by RT and surgery and last of all chemotherapy for breast cancer patients based on LRT with p-value as displayed in Table 1.

4.2 Simulated Data

Simulation studies are computational tests including results creation through pseudo-random sampling from built up likelihood appropriations. These studies are an irreplaceable statistical analysis asset, particularly in the evaluation of current tactics and the assessment of other methodologies. In medicine, the simulation experiments are known to be used in the pages of statistics, yet the analysts sometimes in our experience do not have the enough information required to play out a simulation test with conviction, while others are arrogant and in this way disregard to think carefully about recording consequences and the project. Under certain circumstances, simulation tests are utilized to deliver logical information on the statistical strategies productivity, instead of algebraic tests which is a more common analytical that can incorporate different circumstances.

It is not easy to achieve empirical tests for a lot of people. This problem however

is solved by our simulation experiments. They are independent where messy systems settle on inaccurate choices or results so they could decide the sturdiness of strategies under these conditions. It is not true for observational discoveries, as the outcomes may be duplicated where the aspects are removed from a specific model.

A simulation study was done dependent on the true information of breast malignant featured in this hypothesis to look at the impact of the AFT via Weibull model and to assess the covariates in the informational indexes.

As we discover that the normal distribution is sensible for the real data (designs of the information are moderately close to curves of ordinary) we utilized it to produce the simulation data contrast with Weibull and log-normal distributions (that not showed in this project). Furthermore, the Akaike's data measure known as (AIC) was proven to be 14610.14 for ordinary dissemination, 15059.36 for log-normal and 14672.15 for Weibull.

Adding up to that sample, we used 20000 times for each treatment. Using mean and standard deviation of 0.0759168 & 0.0227825 for chemotherapy treatment, in light of (0%, 25%, half, and 75%) as levels of accurate perception for the partly interval censored (PIC) data. We acquired the measurements of endurance in every simulation data for the two gatherings of every treatment that depends on the precise observation contrasted with the one evaluated by imputation strategies that is; mean, midpoint and left point through our model.

In each simulation data we obtained the function of survival for the two groups (with treatment and without treatment) of each treatment that based on the exact observation compared to the one estimated by imputations methods that is; median, random and right point via our model.

Table 2: Results from chemotherpay obtanied by AFT model with median imputaion based on simulation data

% of Exact	Parameter	Estimate	CI of 95%	SE	LRT* (P-value
0%	Coefficient	0.0940416	(0.0844, 0.1037)	0.004943	-141283.5 (2e-10)
	Shape	2.9032264	(2.8712, 2.9357)	0.081959	
	Scale	972.71857	(965.92, 979.56)	3.479775	
25%	Coefficient	0.0942438	(0.0846, 0.1039)	0.004944	-141292.9 (2e-16)
	Shape	2.9027543	(2.8707, 2.9352)	0.016458	
	Scale	972.23215	(965.44, 979.08)	3.478308	
50%	Coefficient	0.0941029	(0.0844, 0.1038)	0.004942	-141279.6(2e-16)
	Shape	2.9042715	(2.8722, 2.9367)	0.016461	
	Scale	972.53363	(965.74, 979.37)	3.477762	
75%	Coefficient	0.0942234	(0.0845, 0.1039)	0.004945	-141295.7(2e-16)
	Shape	2.9020972	(2.8700, 2.9345)	0.016454	
	Scale	972.30079	(965.51, 979.14)	3.479340	
100%	Coefficient	0.0942438	(0.0846, 0.1039)	0.004944	-141292.9(2e-16)
	Shape	2.9027543	(2.8707, 2.9352)	0.016458	
	Scale	972.23215	(965.44, 979.07)	3.478308	

LRT*: Likelihood Ratio Test

Table 2 showed the outcomes obtained by AFT model based on median point imputation for chemotherapy treatment with different percentages of exact and interval censored data. It showed significant with respect LRT and their p-value. This results indicate that for more exact observation in the data the result are better based on standard error and LRT. Moreover, the chances of survival increases significantly when patient used chemotherapy treatment compared with a patient who have not gone through the chemotherapy treatment while fighting with breast cancer.

The estimates approach in Figures 8 to 19 which is shown that the estimate of the outcomes based on exact observation and compared with the one obtained by median,

random and right point. However, the group who used chemotherapy treatment are more survival compare to those without chemotherapy, suggesting that our median, random methods are provides an acceptable approximation estimation especially when more exact observations (25%, 50% and 75%) are used in the data compare to the less exact observations (0%) as displayed in Figures 8, 9 and 10 (right point).

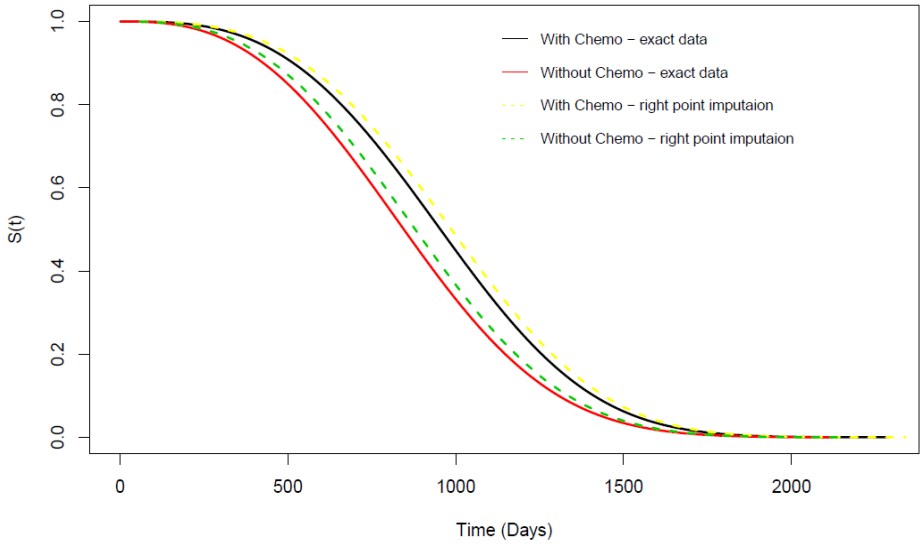

Figure 8: Eestimated of the function of survival based on chemotherpy with 0% exact observation from right point imputaion via AFT model.

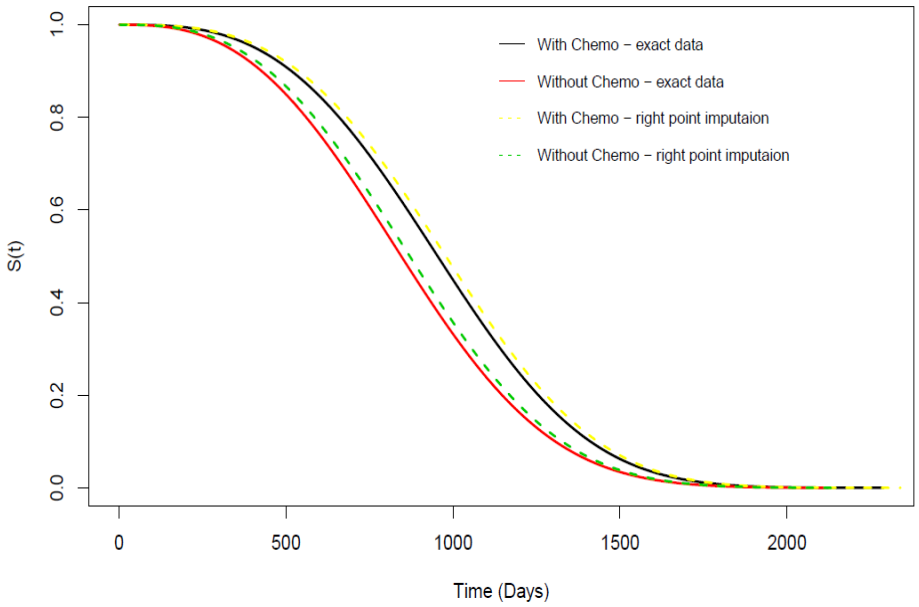

Figure 9: Eestimated of the function of survival based on chemotherpy with 25% exact observation from right point imputaion via AFT model.

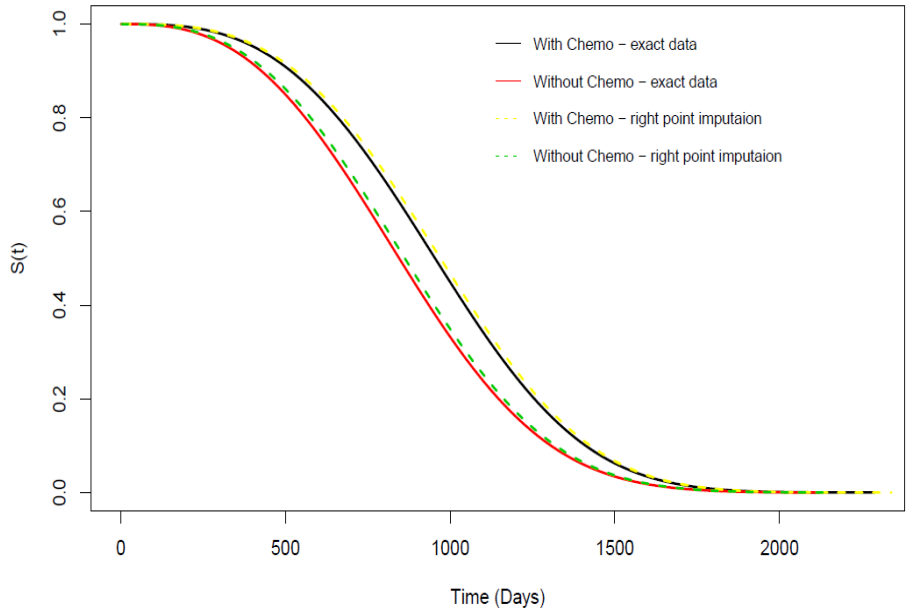

Figure 10: Eestimated of the function of survival based on chemotherpy with 50% exact observation from right point imputaion via AFT model.

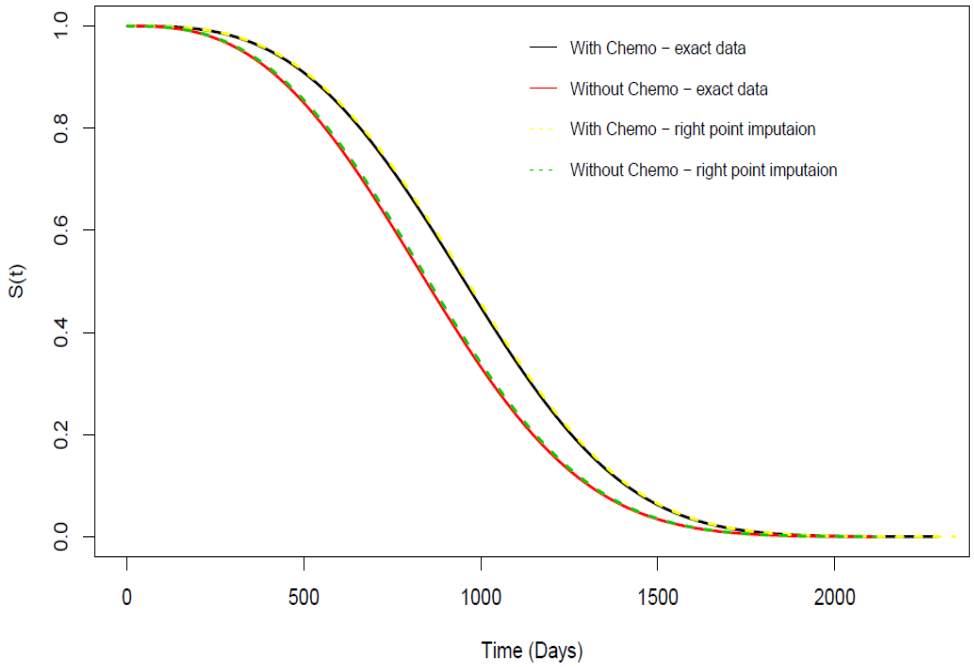

Figure 11: Eestimated of the function of survival based on chemotherpy with 75% exact observation from right point imputaion via AFT model.

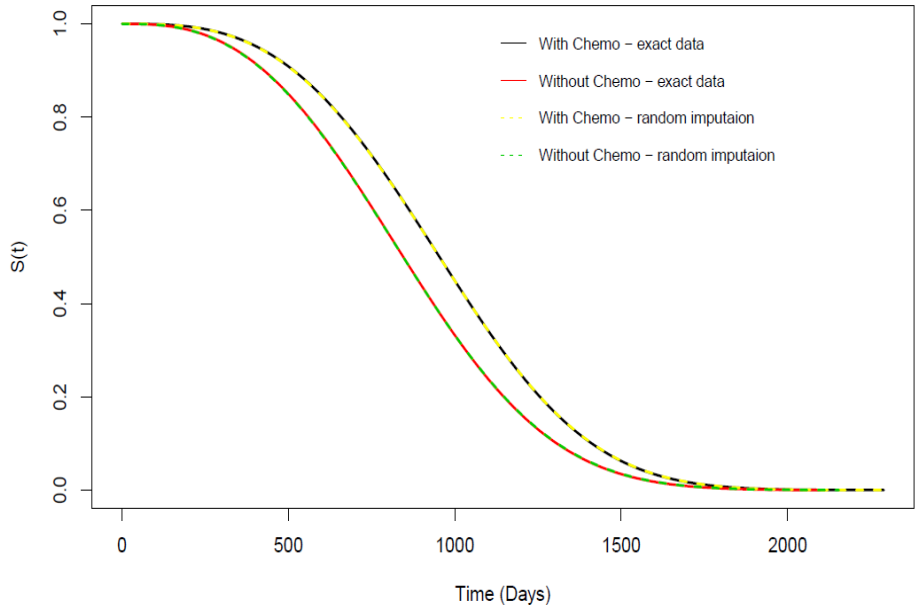

Figure 12: Eestimated of the function of survival based on chemotherpy with 0% exact observation from random imputaion via AFT model.

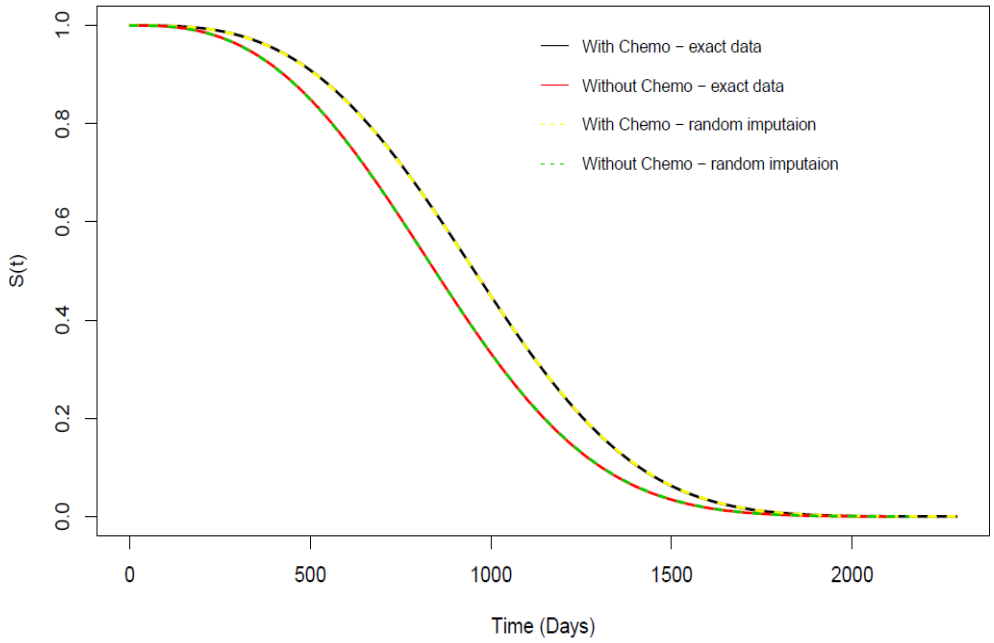

Figure 13: Eestimated of the function of survival based on chemotherpy with 25% exact observation from random imputaion via AFT model.

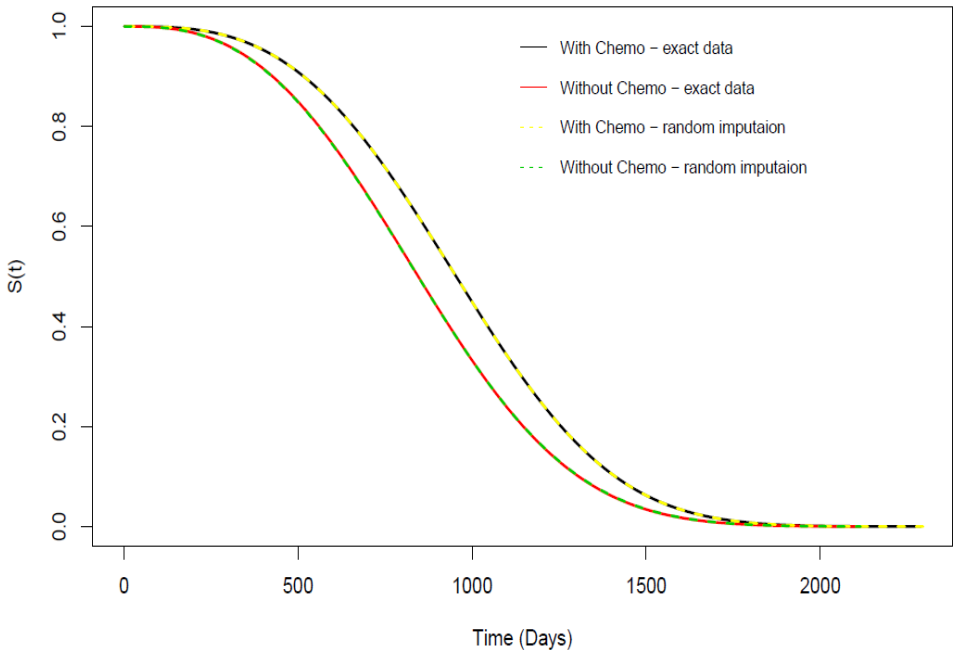

Figure 14: Eestimated of the function of survival based on chemotherpy with 50% exact observation from random imputaion via AFT model.

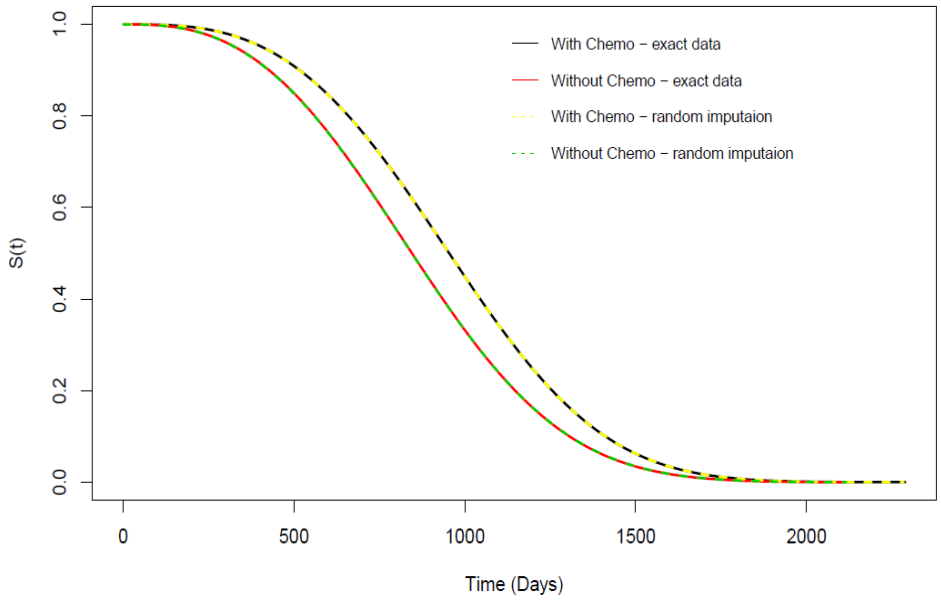

Figure 15: Eestimated of the function of survival based on chemotherpy with 75% exact observation from random imputaion via AFT model.

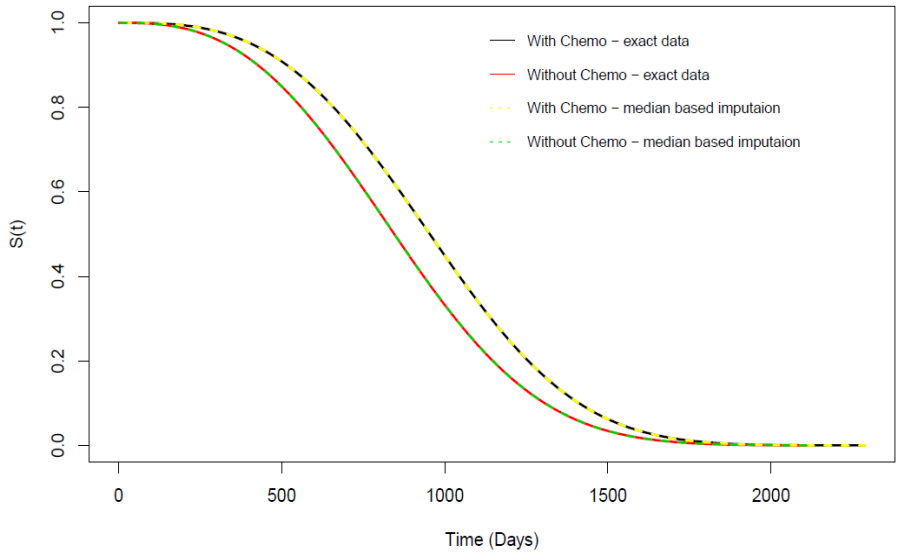

Figure 16: Eestimated of the function of survival based on chemotherpy with 0% exact observation from median imputaion via AFT model.

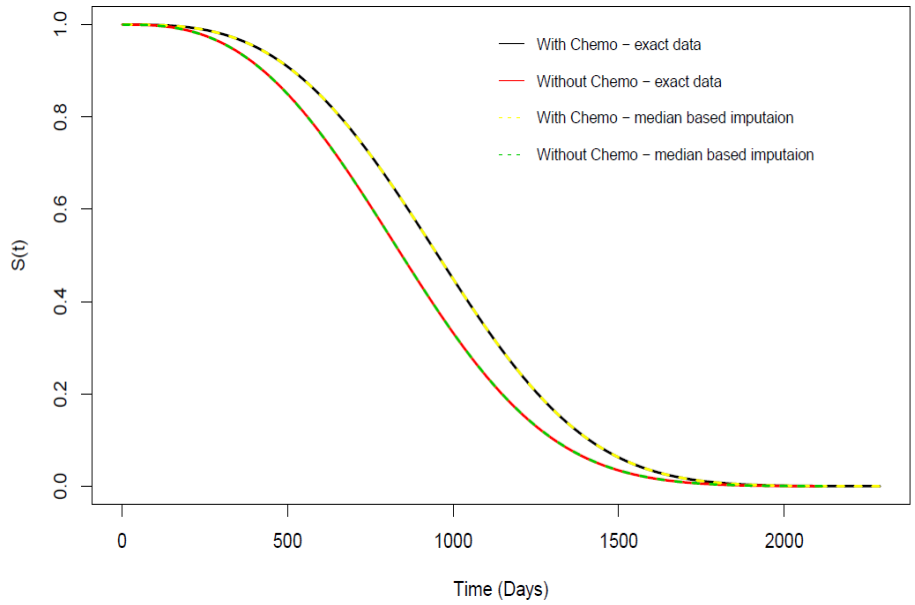

Figure 17: Eestimated of the function of survival based on chemotherpy with 25% exact observation from median imputaion via AFT model.

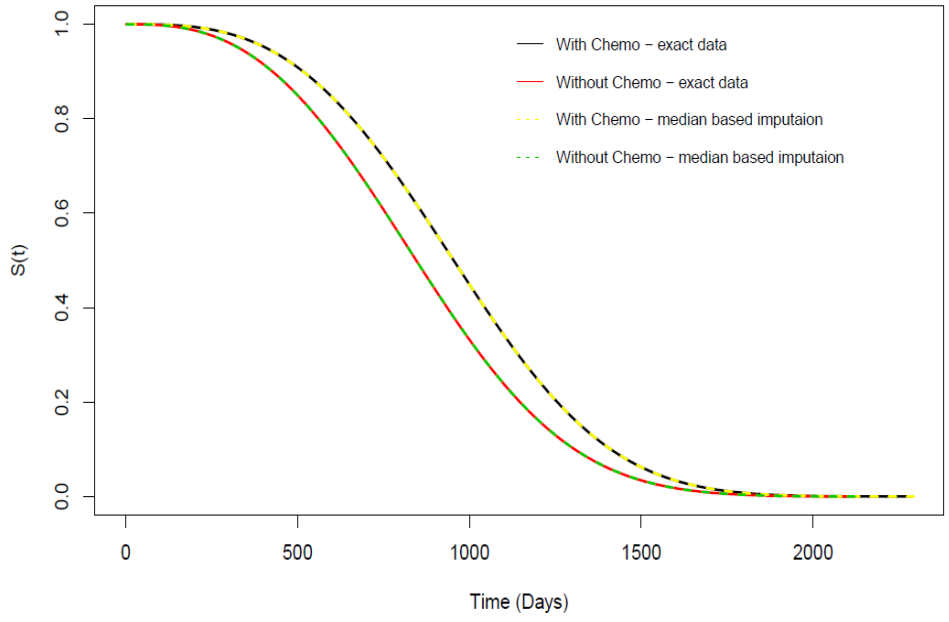

Figure 18: Eestimated of the function of survival based on chemotherpy with 50% exact observation from median imputaion via AFT model.

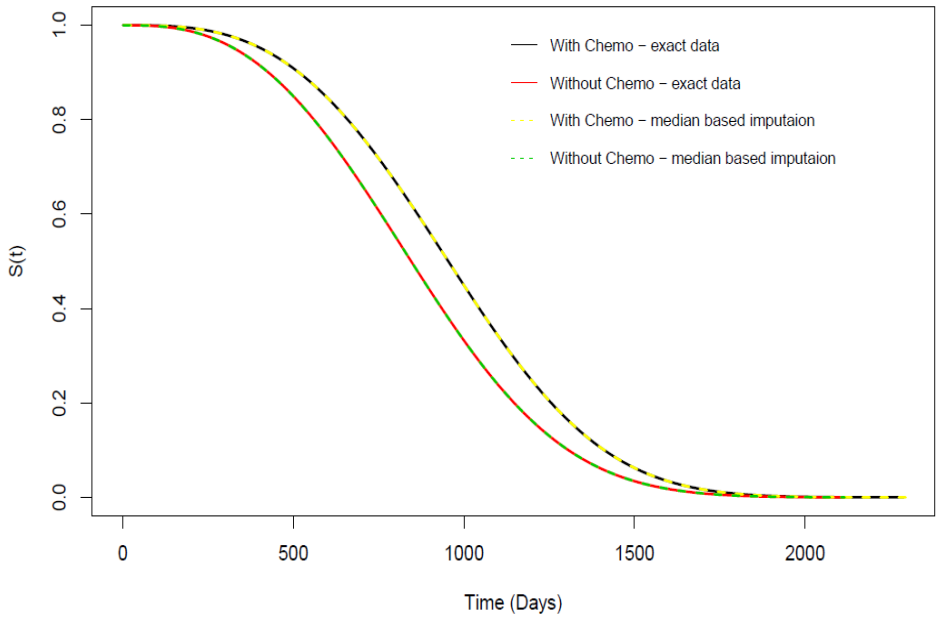

Figure 19: Eestimated of the function of survival based on chemotherpy with 75% exact observation from median imputaion via AFT model.

Table 3: Results from hormone obtanied by AFT model with random imputaion based on simulation data

% of Exact	**Parameter**	**Estimate**	**CI of 95%**	**SE**	**LRT* (P-value**
0%	Coefficient	0.167267	(0.1574, 0.17714)	0.005038	-141194.2 (2e-10)
	Shape	2.880782	(2.8483, 2.9136)	0.016654	
	Scale	920.6089	(913.99, 927.28)	3.389513	
25%	Coefficient	0.1675895	(0.1577, 0.1775)	0.005046	-141223.4 (2e-16)
	Shape	2.8760510	(2.8436, 2.9088)	0.016633	
	Scale	920.15048	(913.52, 926.83)	3.392943	
50%	Coefficient	0.1674553	(0.1576, 0.1774)	0.005055	-141259.6(2e-16)
	Shape	2.8709014	(2.8385, 2.9036)	0.016609	
	Scale	920.11481	(913.48, 926.80)	3.398423	
75%	Coefficient	0.1678243	(0.1579, 0.1778)	0.005063	-141291(2e-16)
	Shape	2.8662033	(2.8339, 2.8989)	0.016587	
	Scale	919.69367	(913.05, 926.39)	3.401836	
100%	Coefficient	0.1679411	(0.1580, 0.1779)	0.005068	-141307.9(2e-16)
	Shape	2.8636293	(2.8313, 2.8963)	0.016577	
	Scale	919.42314	(912.78, 926.12)	3.403494	

LRT*: Likelihood Ratio Test

Table 3 showed the results for the hormone treatment based on random imputation with and without hormone treatment through different exact observations of PIC with 0%, 25%, 50%, 75% and 100%. The patient treated with hormone treatment have a long survival compare to those without hormone while fighting with breast cancer as indicated in Table 3 and Figures 20 to 31.

Figures 20 to 31 showed the result of the estimation function of survival obtained by our model and imputation techniques namely; right point, median and random point. . The figures are acceptable since it showed similar result between exact observation and the one obtained by our imputations methods mentioned except the one obtained by right point with exact observation of 0%. However, significant results are showed based on the median and random imputation with respect to LRT and their P-value which indicate that the null hypothesis is reject (H_0: there is no different between use hormone treatment and not use hormone treatment) as shown in Table 3.

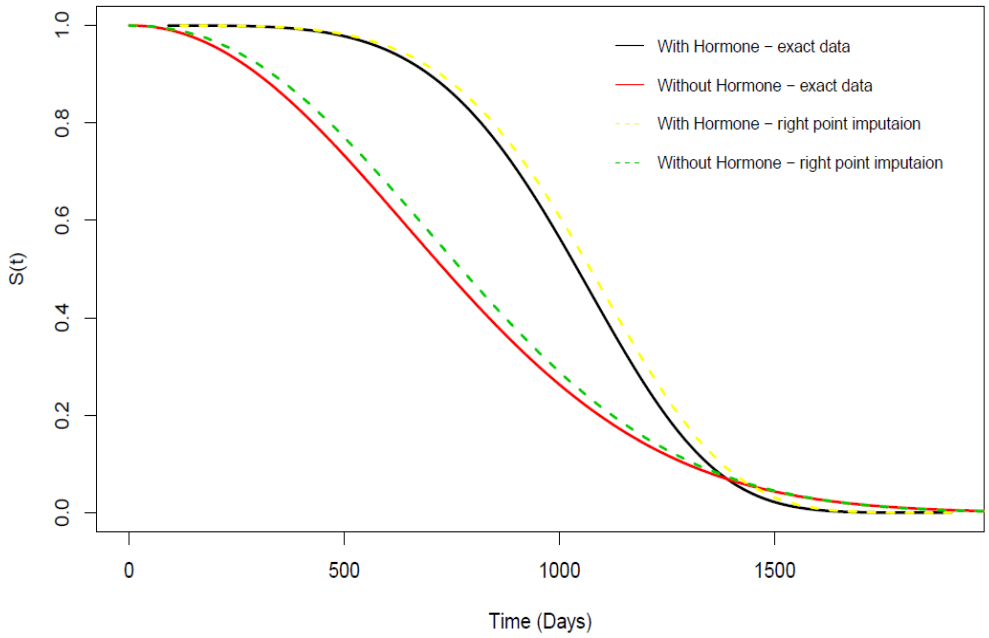

Figure 20: Eestimated of the function of survival based on hormone with 0% exact observation from right point imputation via AFT model.

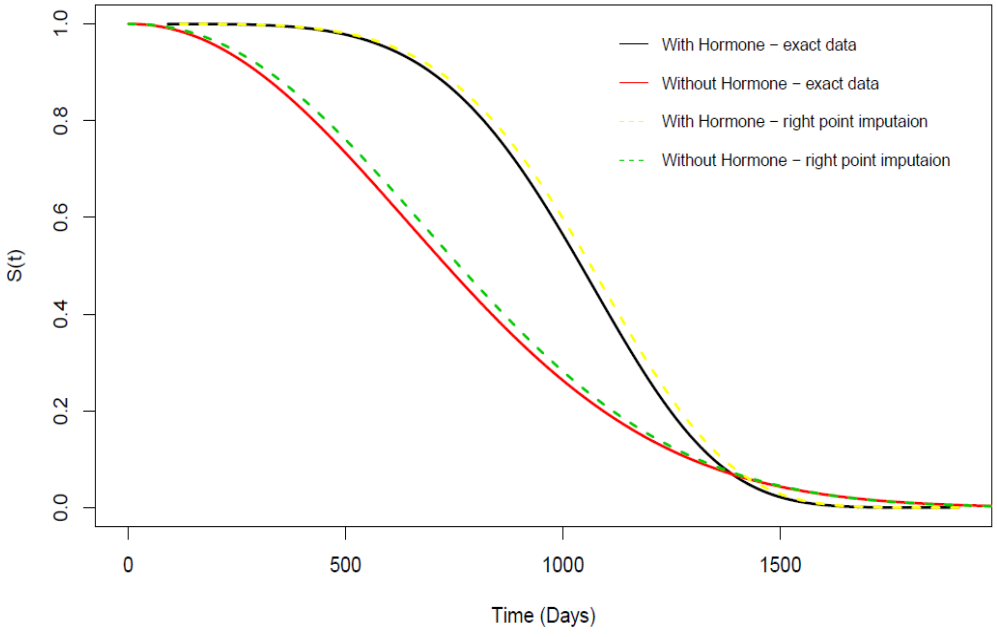

Figure 21: Eestimated of the function of survival based on hormone with 25% exact observation from right point imputation via AFT model.

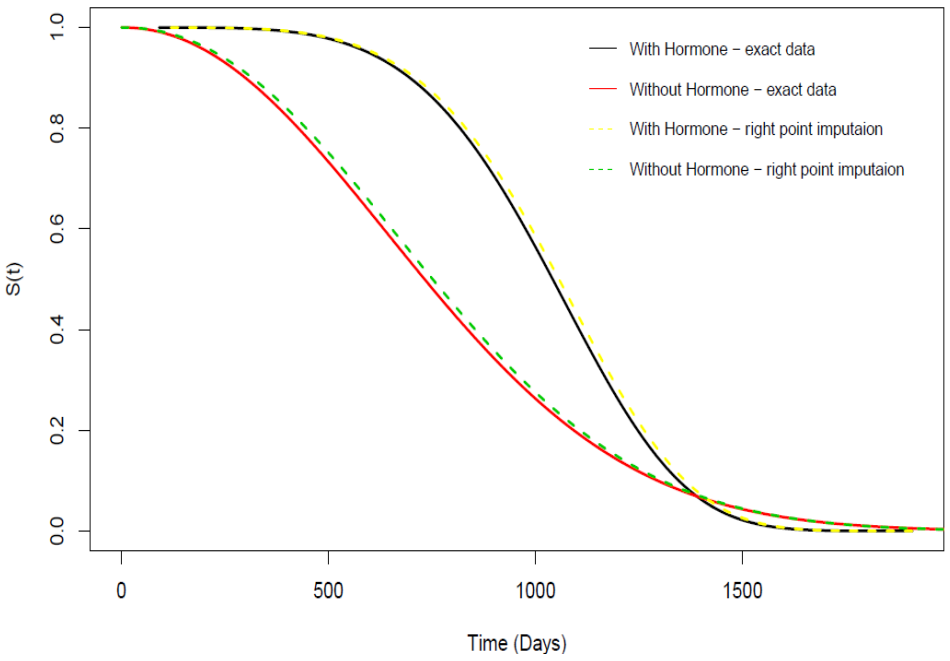

Figure 22: Eestimated of the function of survival based on hormone with 50% exact observation from right point imputation via AFT model.

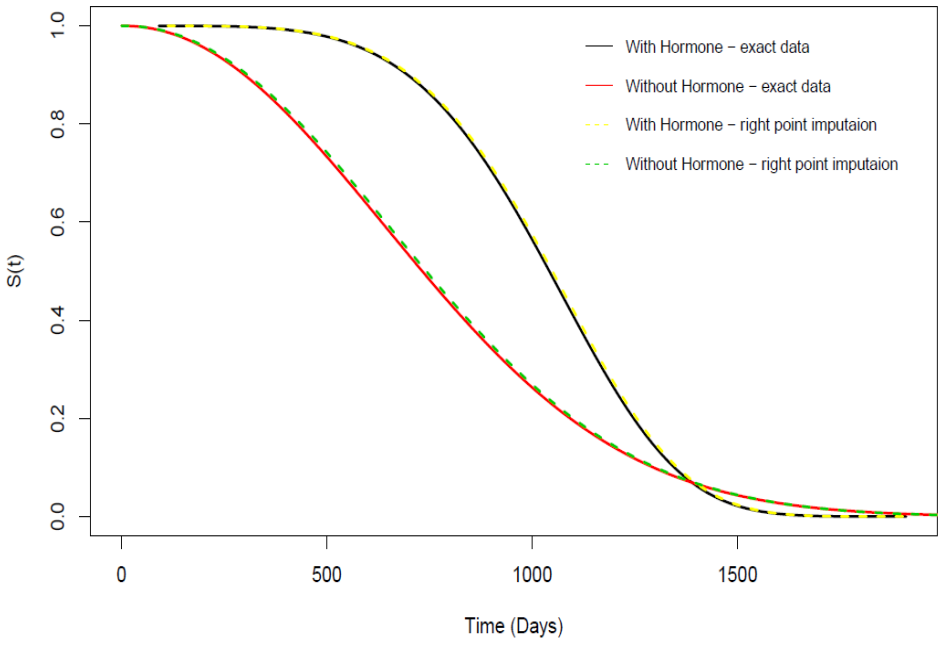

Figure 23: Eestimated of the function of survival based on hormone with 75% exact observation from right point imputation via AFT model.

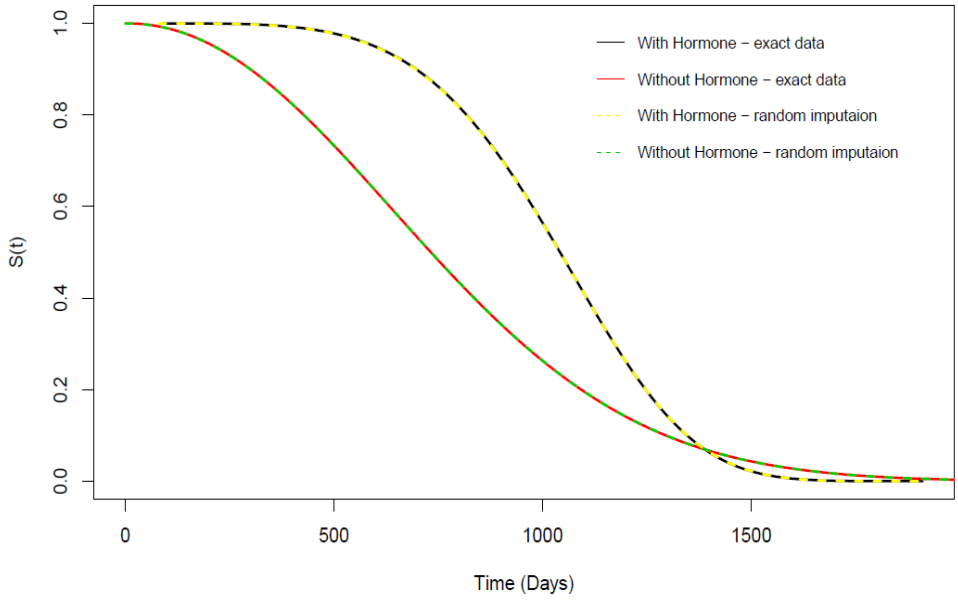

Figure 24: Eestimated of the function of survival based on hormone with 0% exact observation from random point imputation via AFT model.

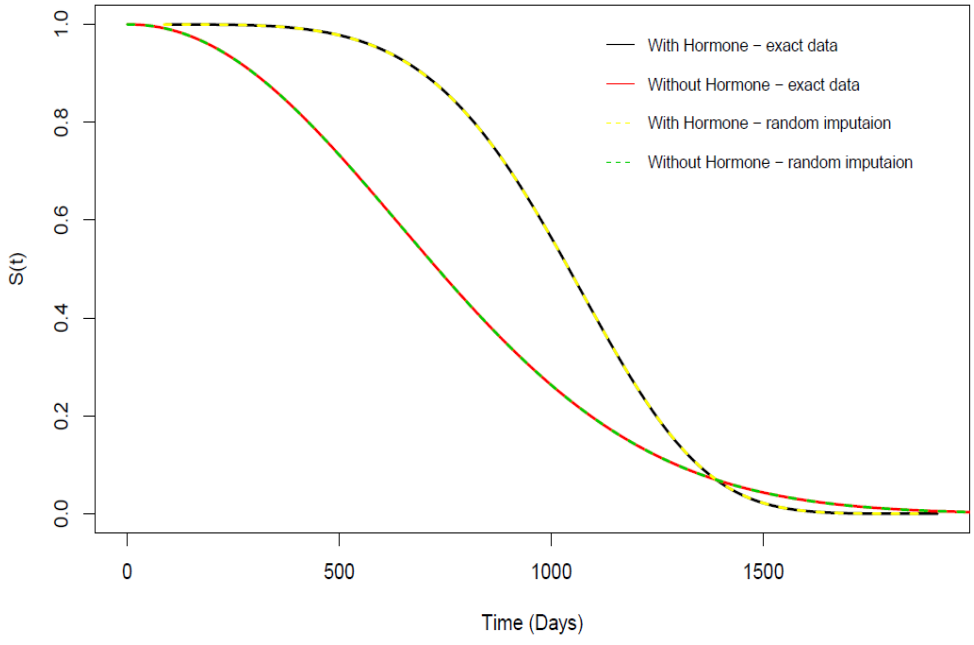

Figure 25: Eestimated of the function of survival based on hormone with 25% exact observation from random point imputation via AFT model.

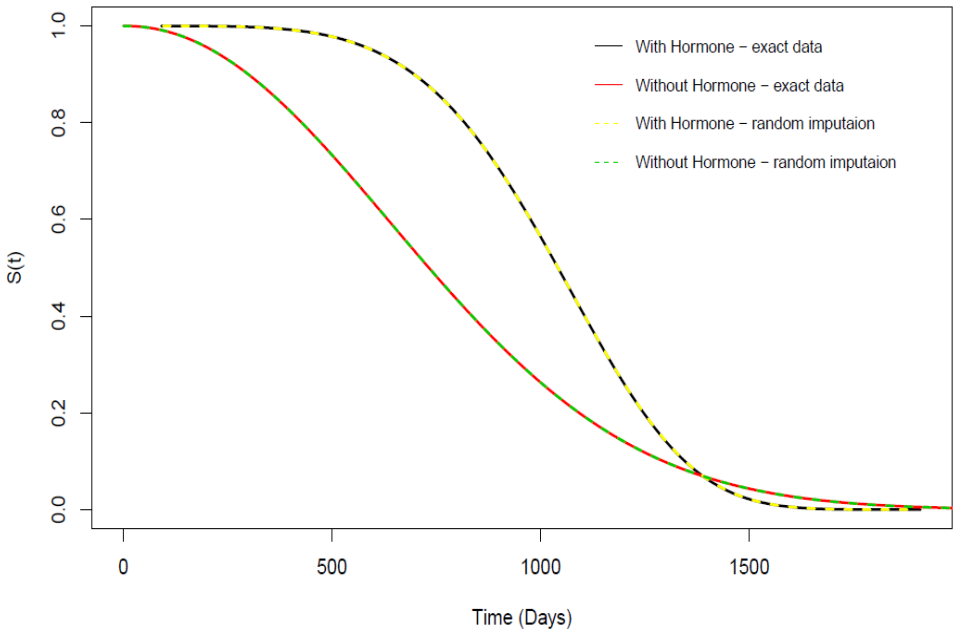

Figure 26: Eestimated of the function of survival based on hormone with 50% exact observation from random point imputation via AFT model.

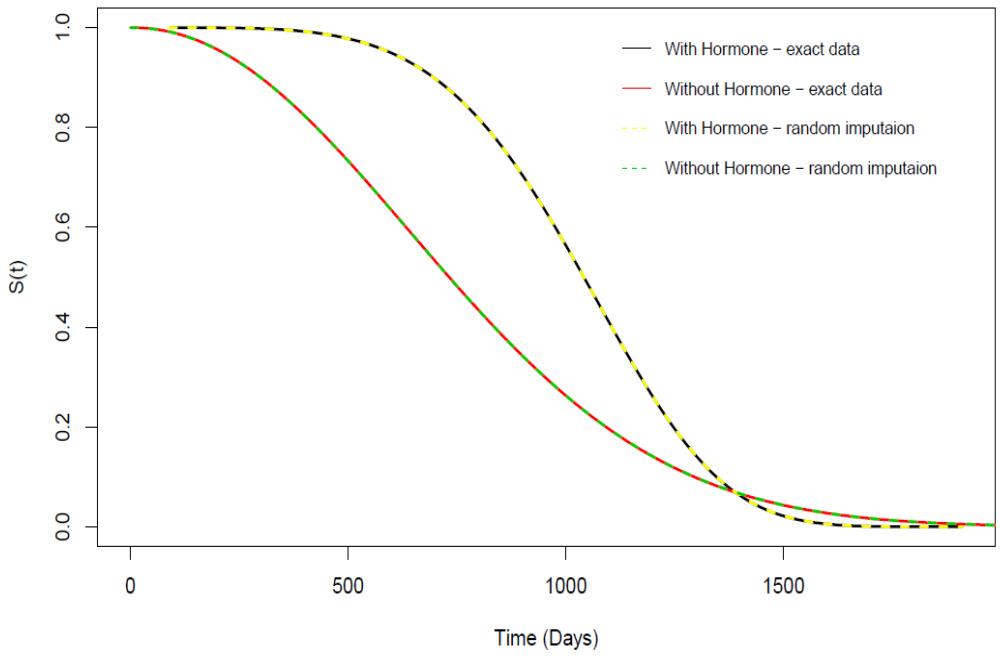

Figure 27: Eestimated of the function of survival based on hormone with 75% exact observation from random point imputation via AFT model.

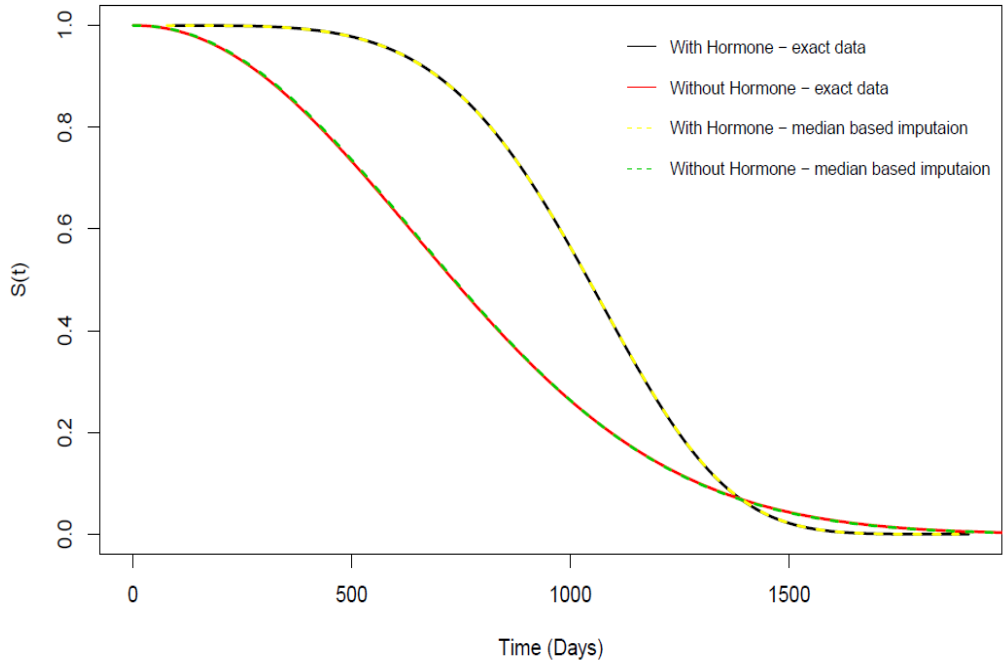

Figure 28: Eestimated of the function of survival based on hormone with 0% exact observation from median point imputation via AFT model.

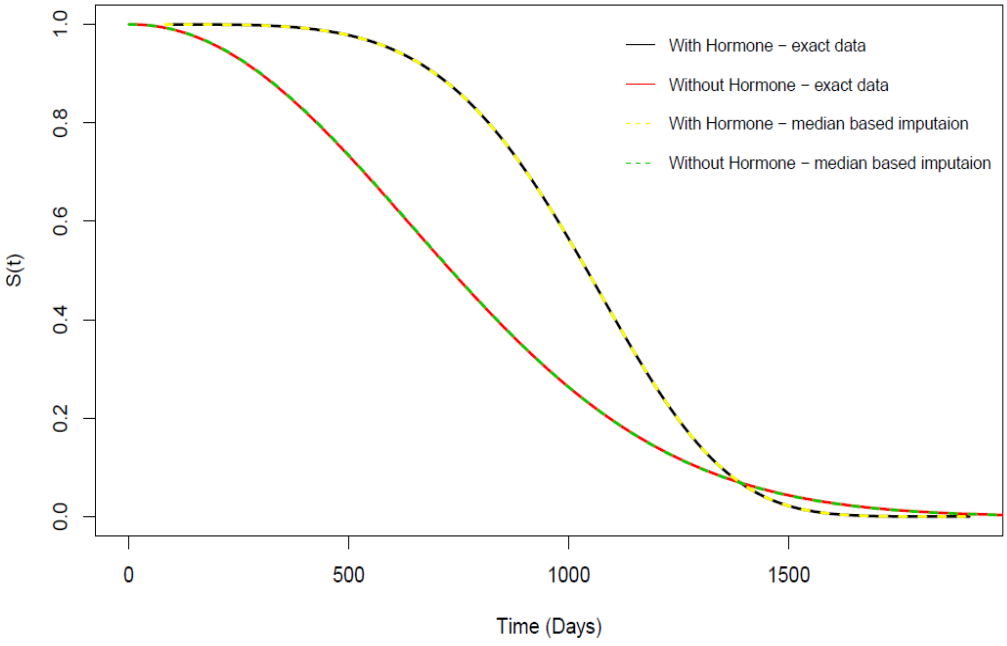

Figure 29: Eestimated of the function of survival based on hormone with 25% exact observation from median point imputation via AFT model.

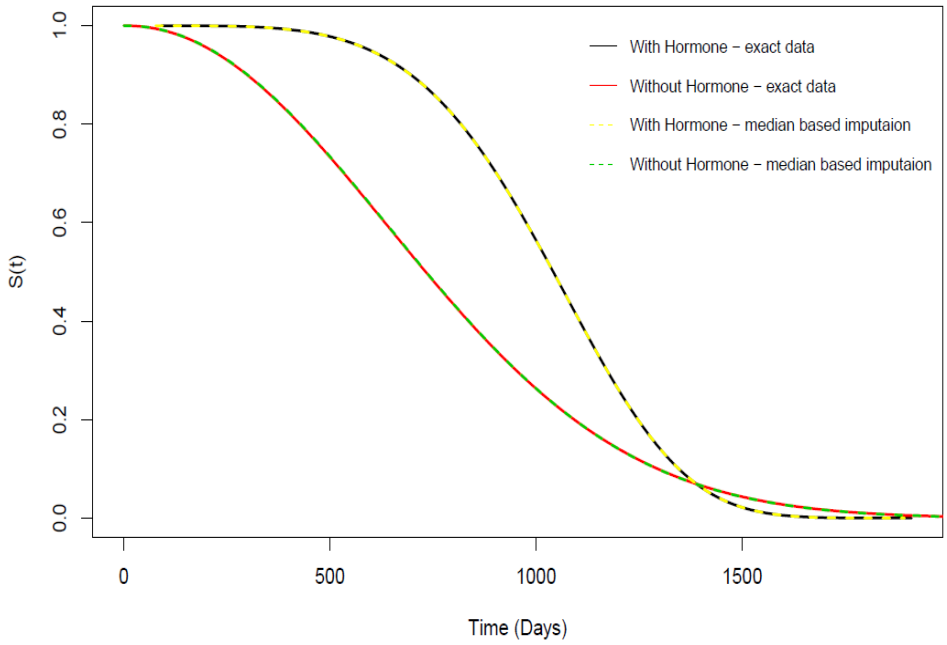

Figure 30: Eestimated of the function of survival based on hormone with 50% exact observation from median point imputation via AFT model.

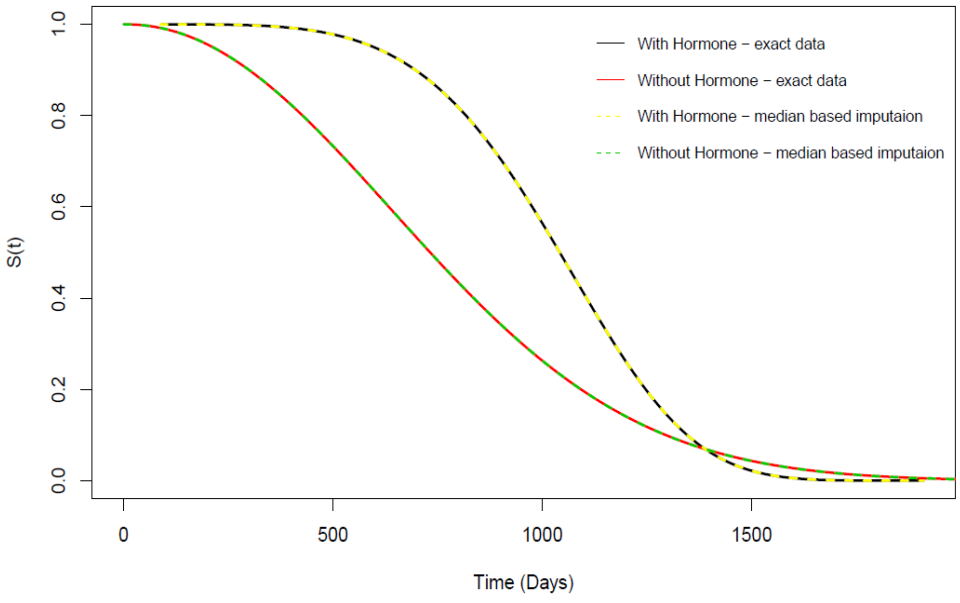

Figure 31: Eestimated of the function of survival based on hormone with 75% exact observation from median point imputation via AFT model.

Table 4: Results from Radiotherapy (RT) obtanied by AFT model with median imputaion based on simulation data

% of Exact	Parameter	Estimate	CI of 95%	SE	LRT* (P-value
0%	Coefficient	0.134629	(0.1250, 0.1443)	0.004918	-140894.5(2e-10)
	Shape	2.941726	(2.9089, 2.9750)	0.016868	
	Scale	947.51813	(940.88, 954.21)	3.400072	
25%	Coefficient	0.1347957	(0.1251, 0.1445)	0.004925	-140926.9 (2e-16)
	Shape	2.9374182	(2.9046, 2.9706)	0.016850	
	Scale	947.29828	(940.65, 953.99)	3.403819	
50%	Coefficient	0.1348801	(0.1252, 0.1445)	0.004925	-1410927.6(2e-16)
	Shape	2.9373304	(2.9045, 2.9705)	0.016849	
	Scale	947.23029	(940.58, 953.93)	3.403646	
75%	Coefficient	0.1351449	(0.1255, 0.1448)	0.004928	-140942.2(2e-16)
	Shape	2.9352125	(2.9024, 2.9684)	0.016841	
	Scale	946.89045	(940.24, 953.59)	3.404387	
100%	Coefficient	0.1353154	(0.1257, 0.1450)	0.004930	-1410946.3(2e-16)
	Shape	2.9343472	(2.9015, 2.9675)	0168368	
	Scale	946.66885	(940.02, 953.37)	3.404489	

LRT*: Likelihood Ratio Test

Table 4 showed the results for the Radiotherapy (RT) treatment based on random imputation with and without RT treatment through different exact observations of PIC with 0%, 25%, 50%, 75% and 100%. These results indicate that when we have more exact observation the result is more acceptable as AIC in 100% is 281795.1 compare with 0% is 281898.7.

Figures 32 to 43 showed the result of the estimation function of survival obtained by our model and imputation techniques namely; right point, median and random point.

These Figures look almost similar in case of the one obtained by median and

random point especially for exact observation that more than 25%, but little difference compared with one obtained by right point in case of exact observation that more less than 25%. However, the chances of survival increase significantly when patient use RT treatment compared with a patient who have not gone through the RT treatment as shown clearly in this figures mentioned above as well in Table 4 with respect to LRT and p-vale. In additional to that the null hypothesis (H0: there is no different between patient who use RT and not use RT treatment) is rejected.

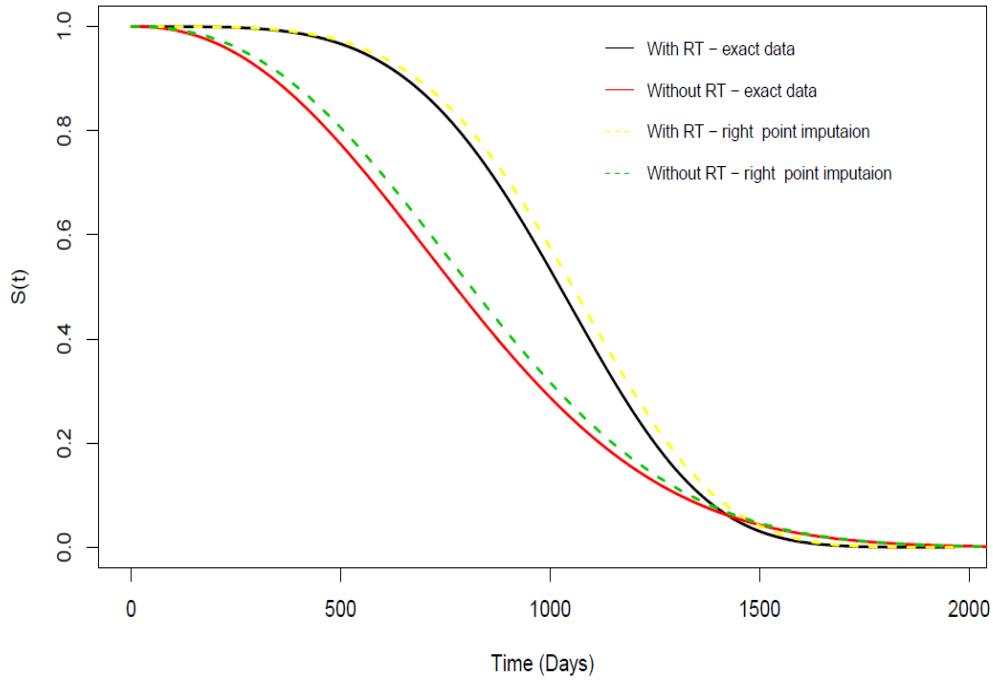

Figure 32: Eestimated of the function of survival based on radiotherapy with 0% exact observation from right point imputation via AFT model.

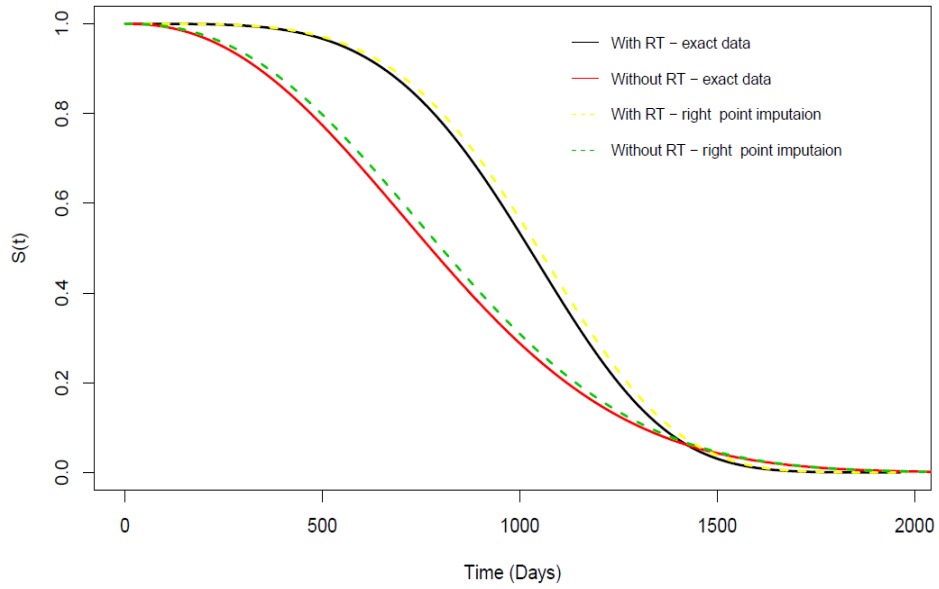

Figure 33: Eestimated of the function of survival based on radiotherapy with 25% exact observation from right point imputation via AFT model.

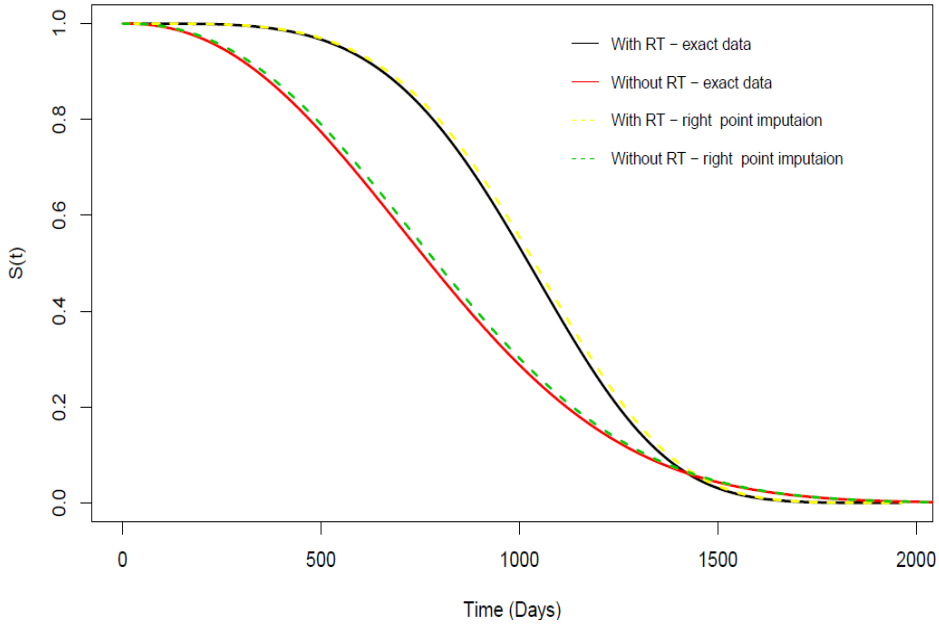

Figure 34: Eestimated of the function of survival based on radiotherapy with 50% exact observation from right point imputation via AFT model.

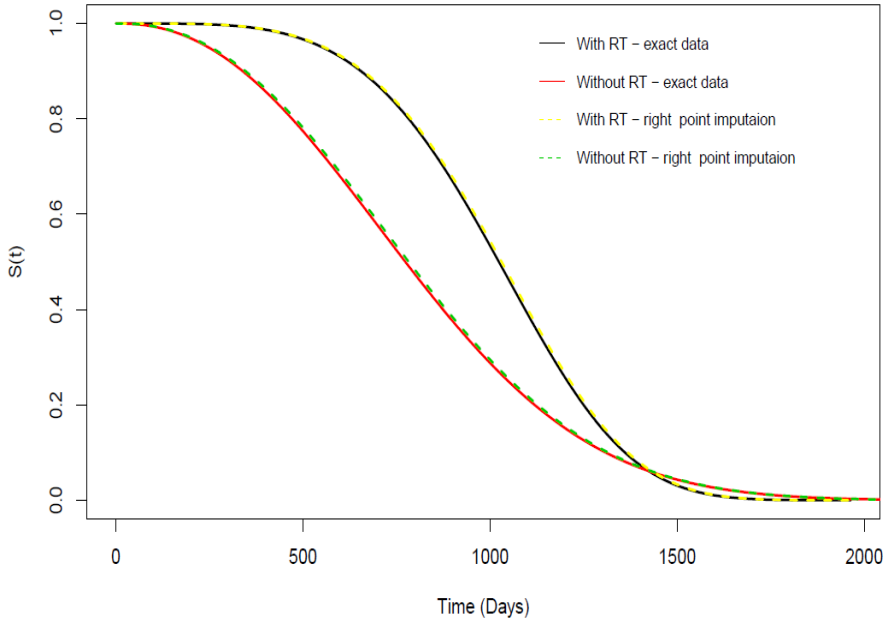

Figure 35: Eestimated of the function of survival based on radiotherapy with 75% exact observation from right point imputation via AFT model.

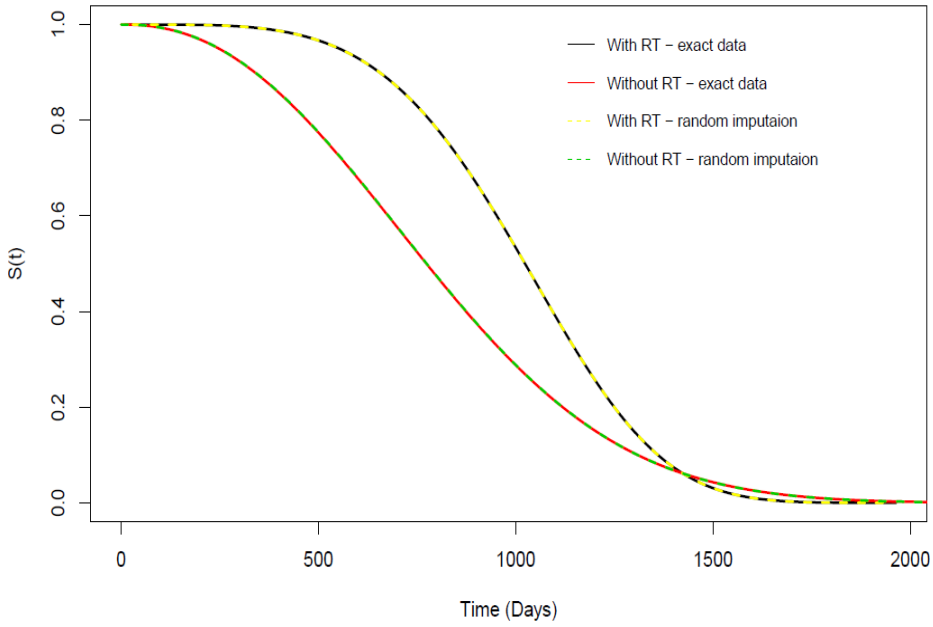

Figure 36: Eestimated of the function of survival based on radiotherapy with 0% exact observation from random point imputation via AFT model.

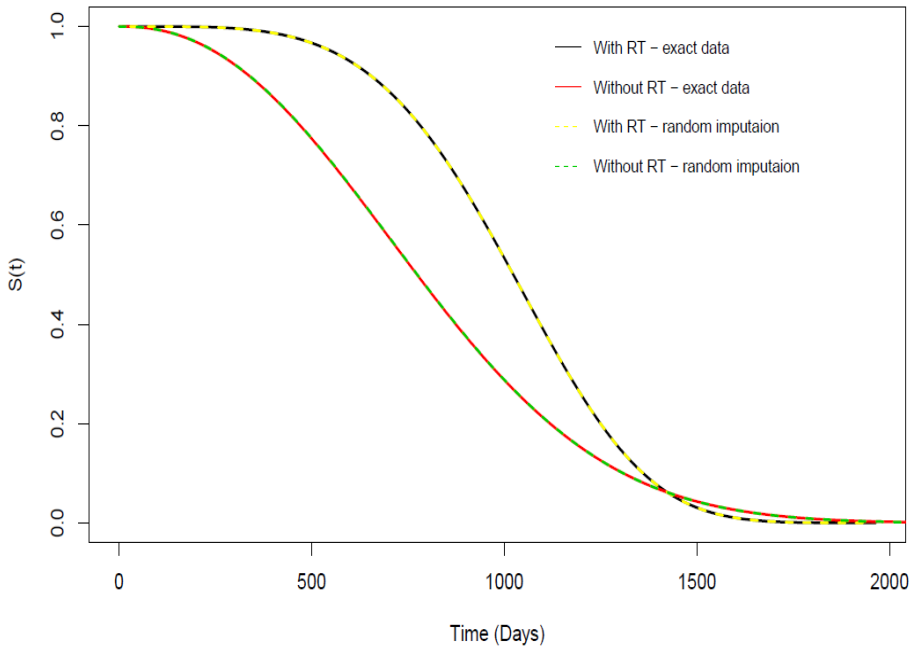

Figure 37: Eestimated of the function of survival based on radiotherapy with 25% exact observation from random point imputation via AFT model.

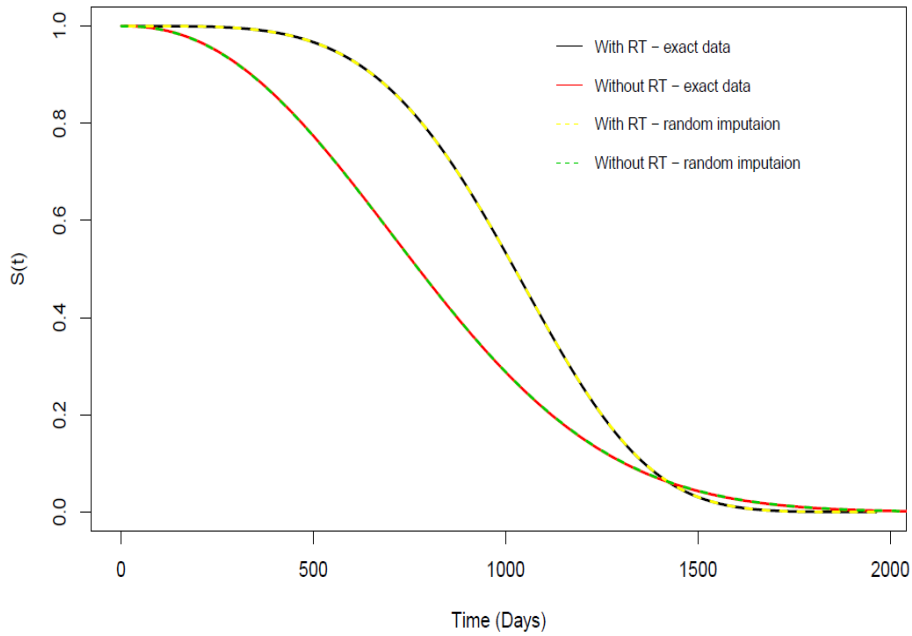

Figure 38: Eestimated of the function of survival based on radiotherapy with 50% exact observation from random point imputation via AFT model.

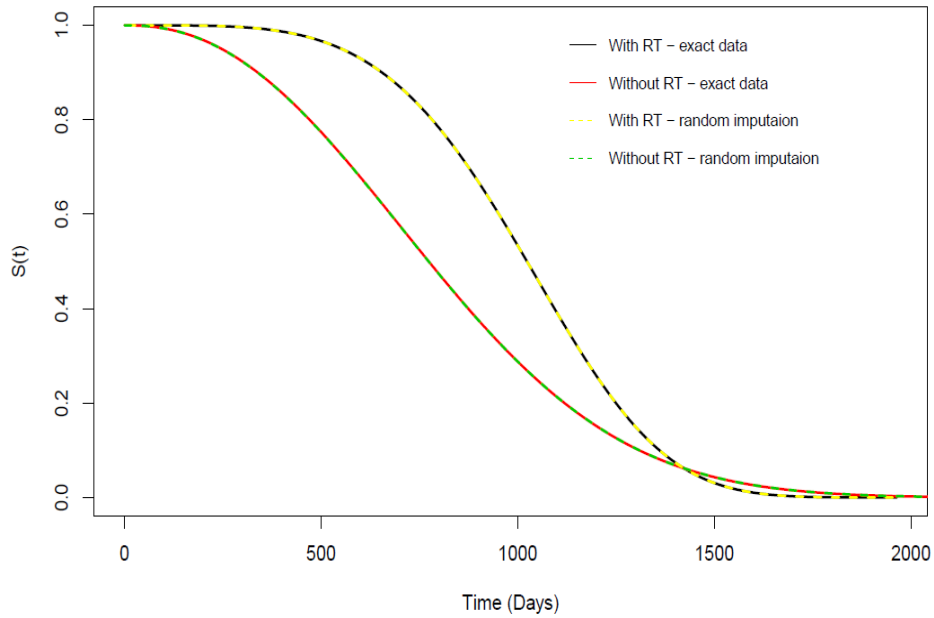

Figure 39: Eestimated of the function of survival based on radiotherapy with 75% exact observation from random point imputation via AFT model.

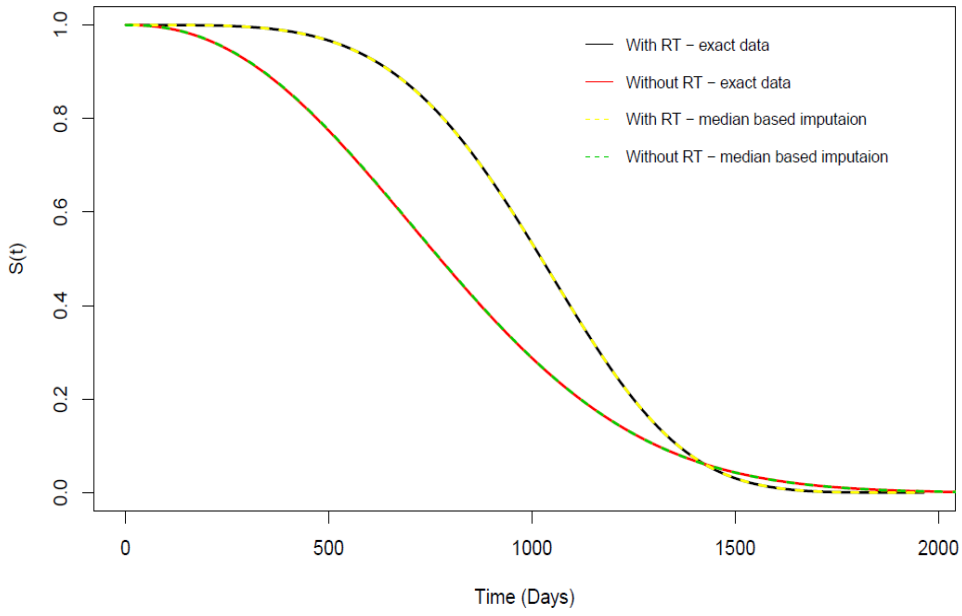

Figure 40: Eestimated of the function of survival based on radiotherapy with 0% exact observation from median point imputation via AFT model.

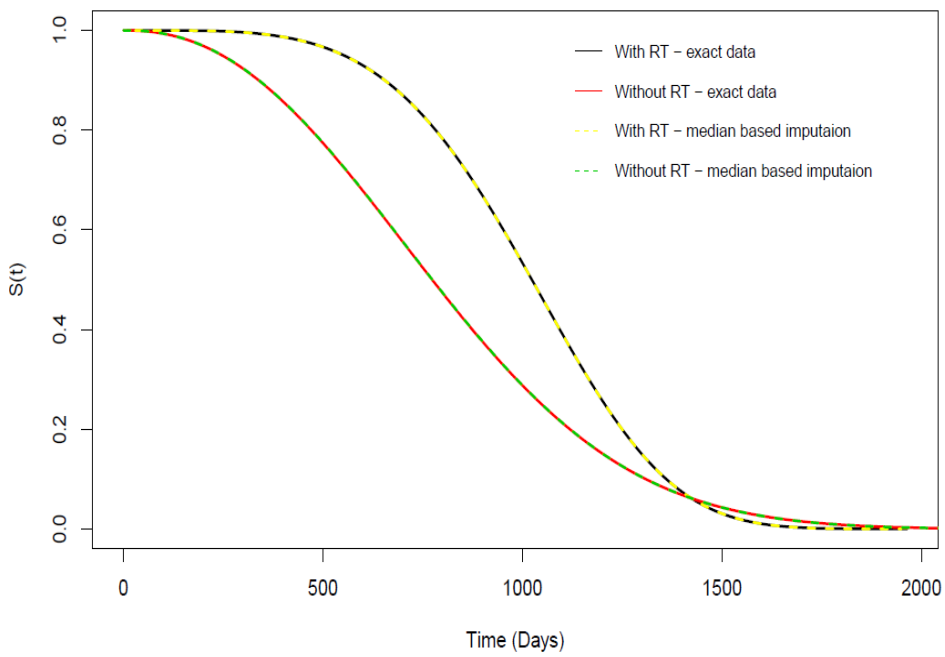

Figure 41: Eestimated of the function of survival based on radiotherapy with 25% exact observation from median point imputation via AFT model.

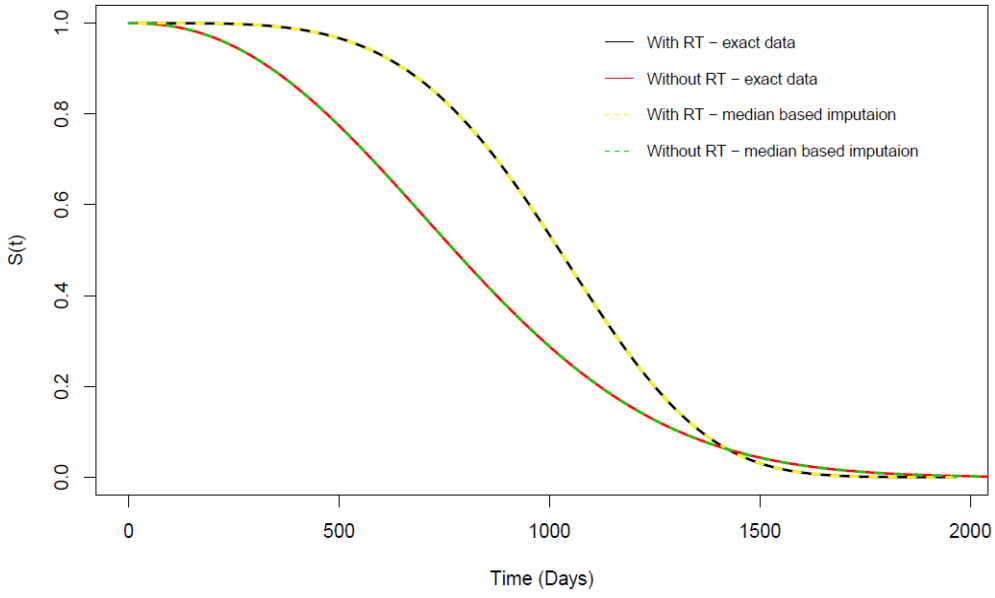

Figure 42: Eestimated of the function of survival based on radiotherapy with 50% exact observation from median point imputation via AFT model.

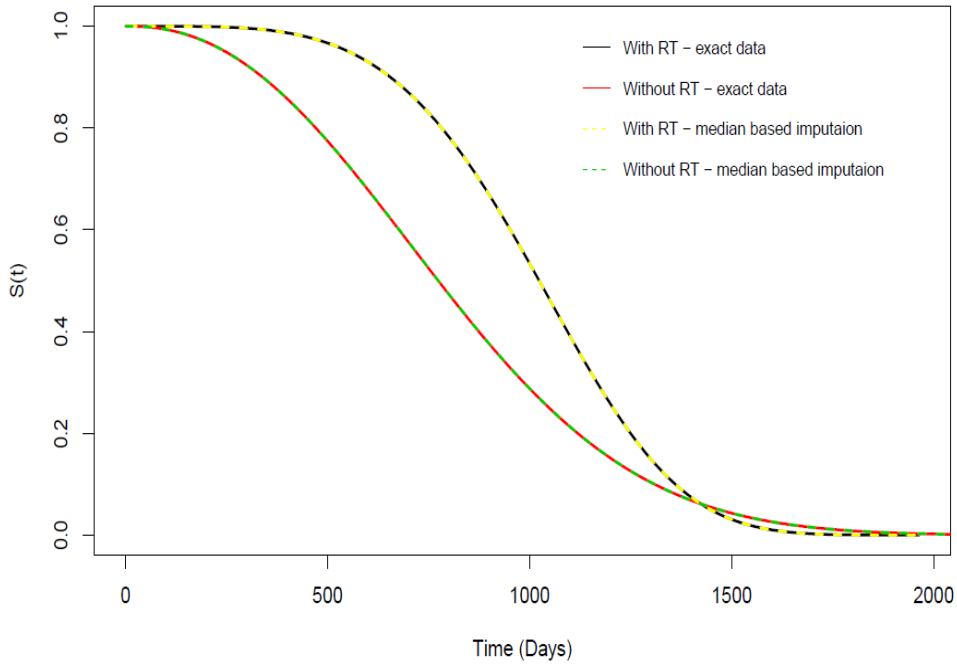

Figure 43: Eestimated of the function of survival based on radiotherapy with 75% exact observation from median point imputation via AFT model.

Table 5: Results from surgery obtanied by AFT model with random imputaion based on simulation data

% of Exact	**Parameter**	**Estimate**	**CI of 95%**	**SE**	**LRT* (P-value**
0%	Coefficient	0.164277	(0.1539, 0.1747)	0.00531	-141489.3(2e-10)
	Shape	2.704938	(2.6748, 2.7354)	0.01545	
	Scale	887.0654	(880.37, 893.81)	3.42914	
25%	Coefficient	0.164408	(0.1540, 0.1748)	0.00531	-141519.9 (2e-16)
	Shape	2.703855	(2.6738, 2.7343)	0.01545	
	Scale	886.9470	(880.25, 893.70)	3.42988	
50%	Coefficient	0.164222	(0.1538, 0.1747)	0.00532	-141519.9(2e-16)
	Shape	2.701681	(2.6716, 2.7321)	0.01544	
	Scale	886.9837	(880.28, 893.74)	3.43237	
75%	Coefficient	0.164177	(0.1538, 0.1746)	0.00532	-141534.9(2e-16)
	Shape	2.699769	(2.6697, 2.7302)	0.01543	
	Scale	886.9672	(880.26, 893.72)	3.43458	
100%	Coefficient	0.164611	(0.1542, 0.1751)	0.00533	-141544 (2e-16)
	Shape	2.698304	(2.6682, 2.7287)	0.01543	
	Scale	886.5889	(879.88, 893.35)	3.43483	

LRT*: Likelihood Ratio Test

Table 5 showed the outcomes obtained by AFT model based on random imputation for surgery treatment with different percentages of exact of PIC data. It showed significant with respect LRT and their p-value.

Figures 44 to 55 showed the result of the estimation function of survival obtained by our model and imputation techniques namely; right point, median and random point. The figures are almost the same for the result obtained except for the one obtained by right point with exact 0% and 25% (Figures 44 and 45). However, significant results are obtained via random imputation that have showed with respect to LRT and their P-value (Table 5).

Table 5 showed the results obtained by our model based on random imputation for surgery treatment with different percentages of exact and interval censored data. This results indicate that for more exact observation in the data the result are better (as high value of AIC=282984.6 when 100% exact compare to AIC= 283094 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

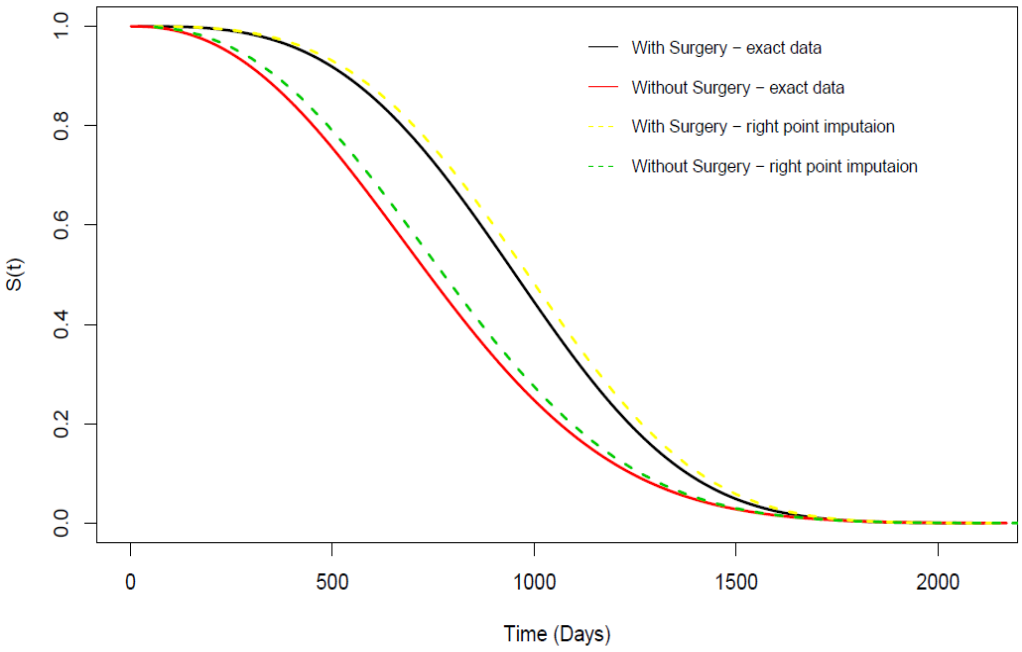

Figure 44: Eestimated of the function of survival based on surgery with 0% exact observation from right point imputation via AFT model.

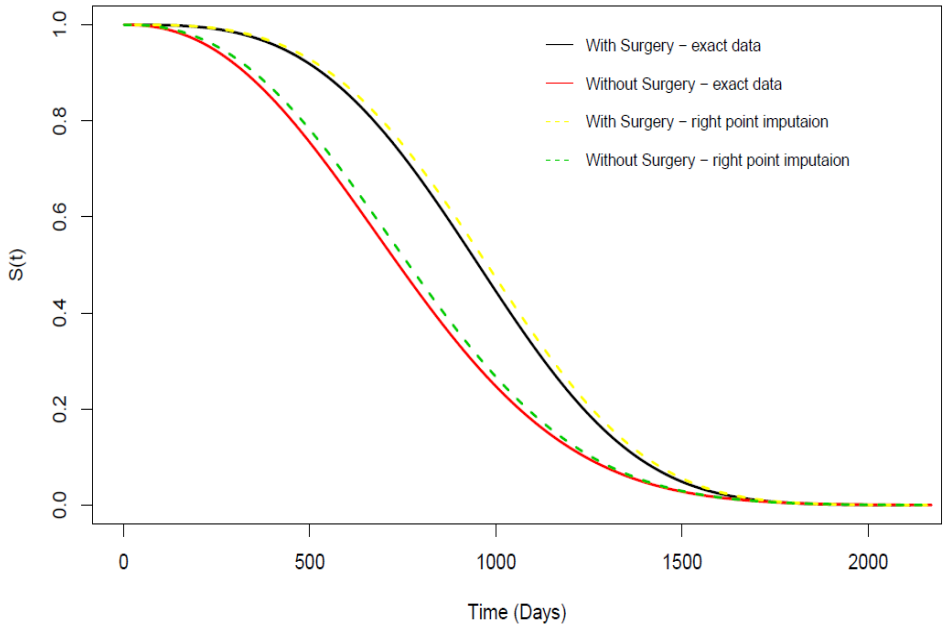

Figure 85: Eestimated of the function of survival based on surgery with 25% exact observation from right point imputation via AFT model.

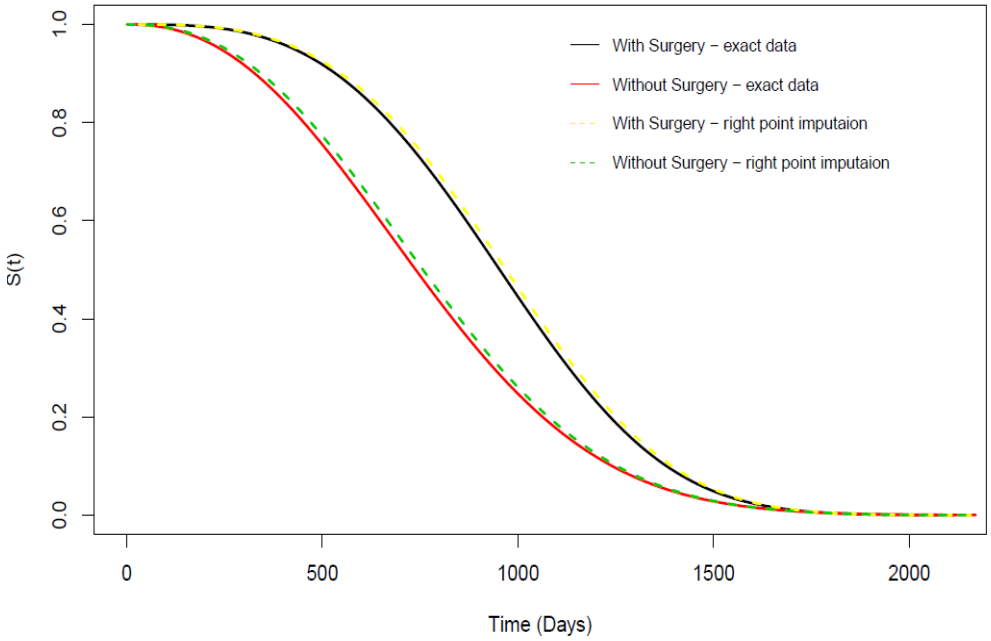

Figure 46: Eestimated of the function of survival based on surgery with 50% exact observation from right point imputation via AFT model.

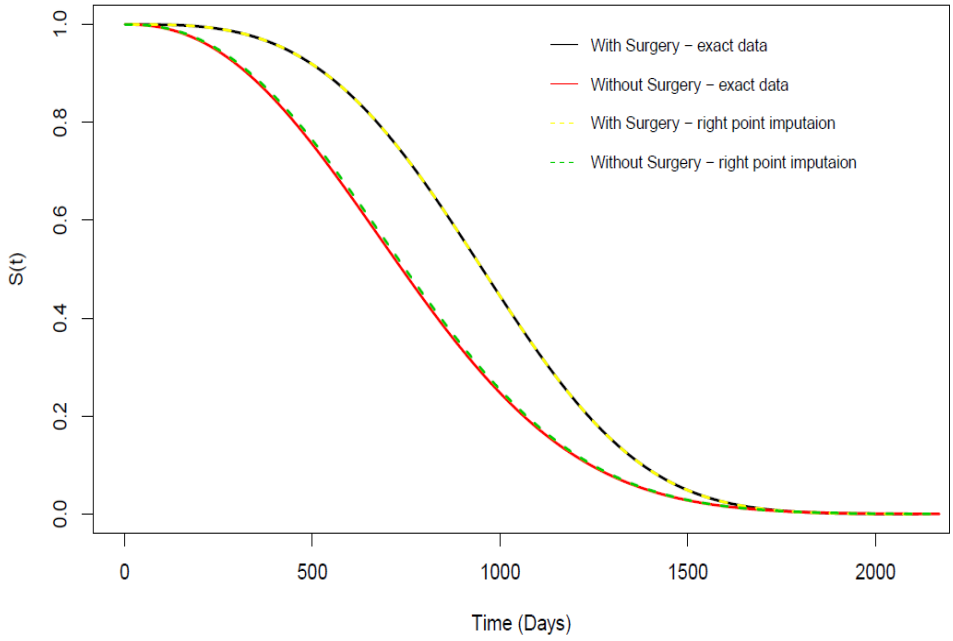

Figure 47: Eestimated of the function of survival based on surgery with 75% exact observation from right point imputation via AFT model.

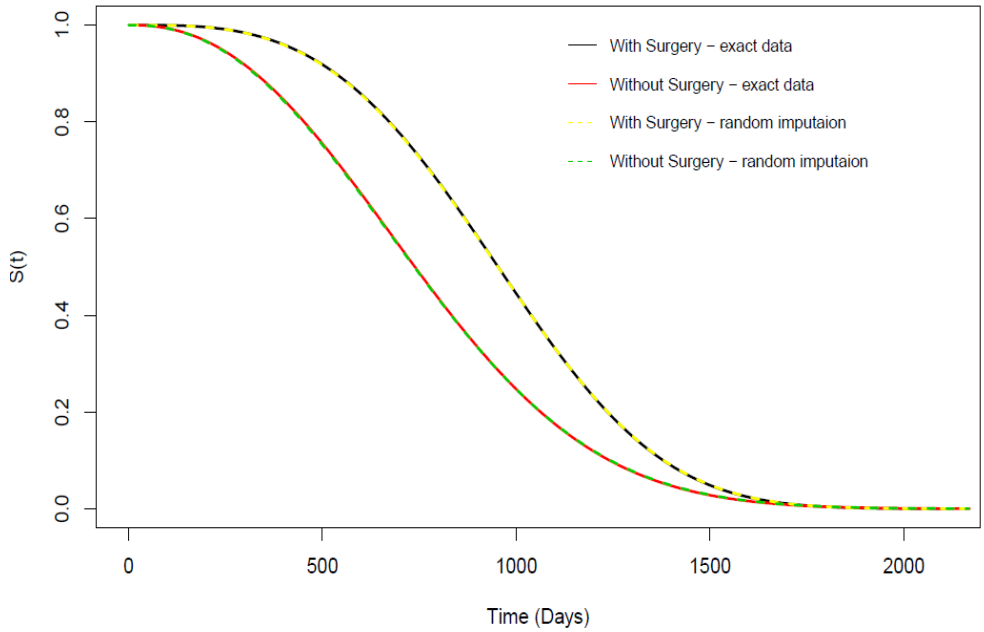

Figure 48: Eestimated of the function of survival based on surgery with 0% exact observation from random point imputation via AFT model.

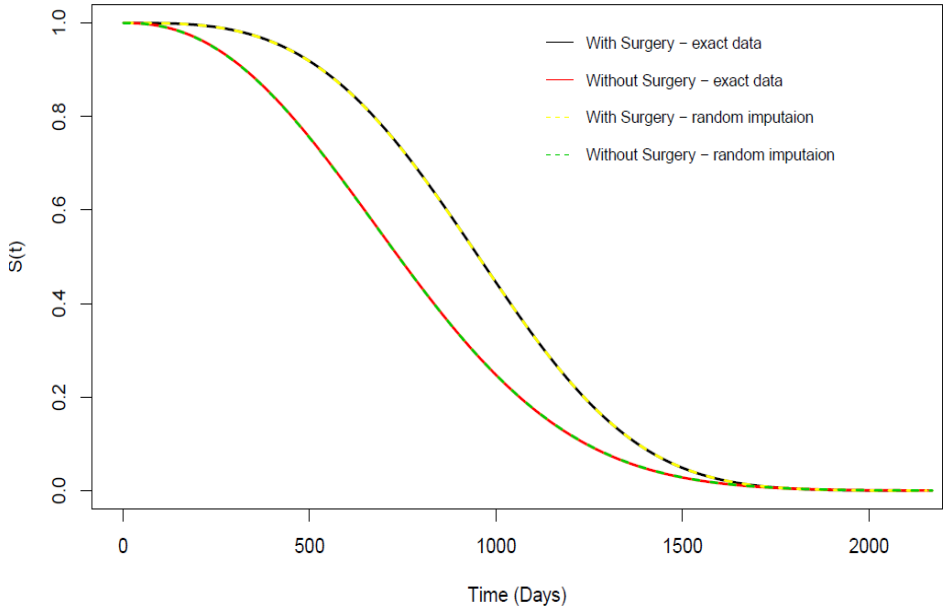

Figure 49: Eestimated of the function of survival based on surgery with 25% exact observation from random point imputation via AFT model.

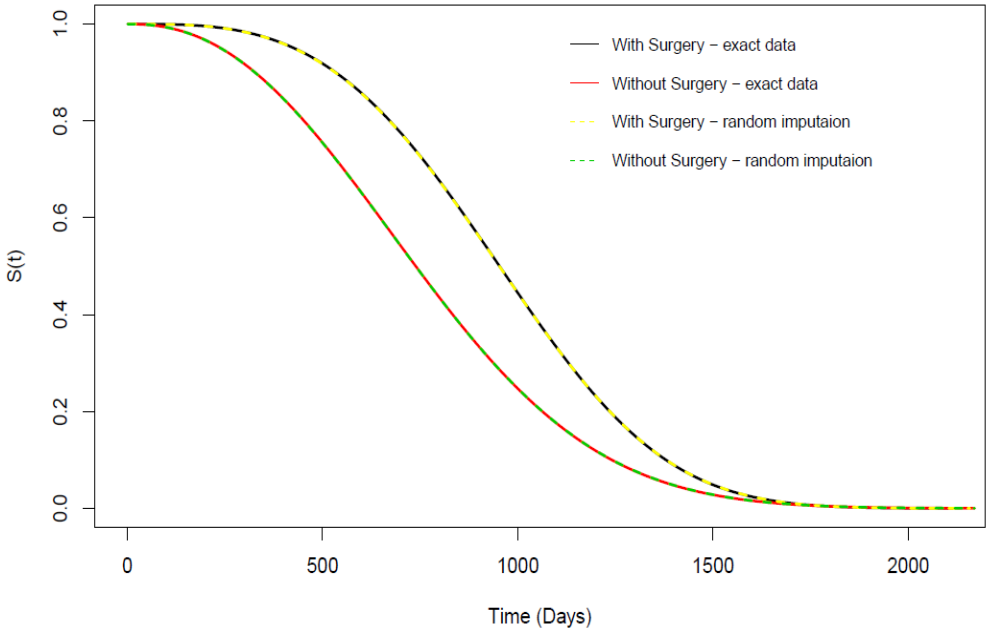

Figure 50: Eestimated of the function of survival based on surgery with 50% exact observation from random point imputation via AFT model.

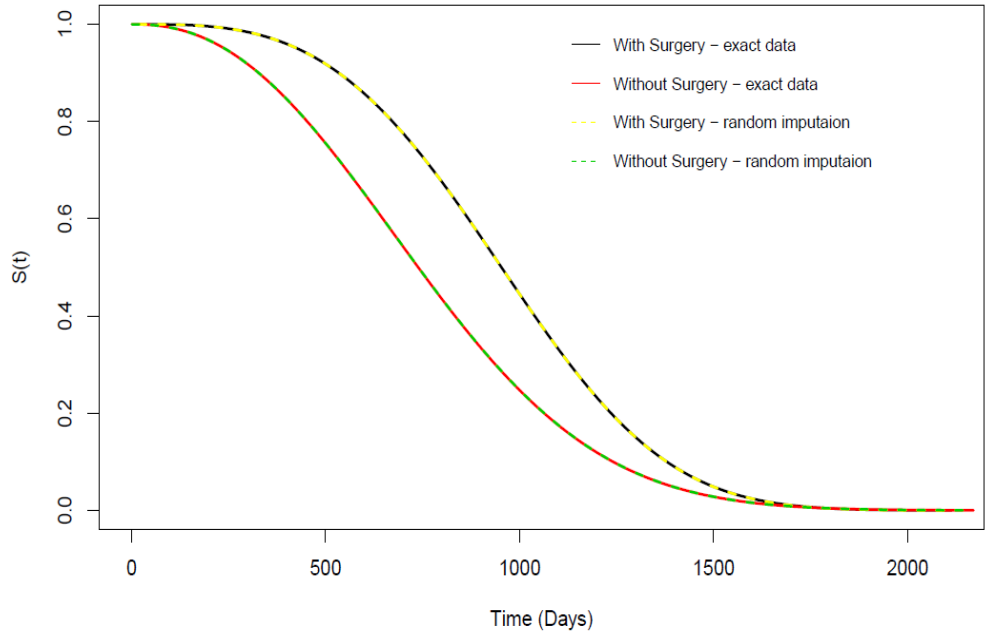

Figure 51: Eestimated of the function of survival based on surgery with 75% exact observation from random point imputation via AFT model.

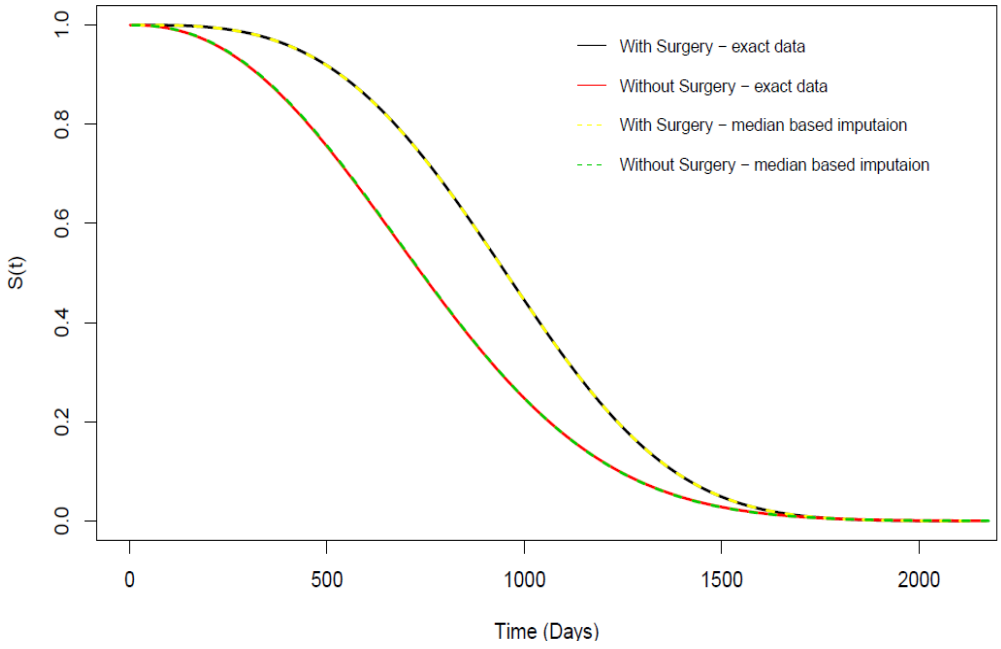

Figure 52: Eestimated of the function of survival based on surgery with 0% exact observation from median point imputation via AFT model.

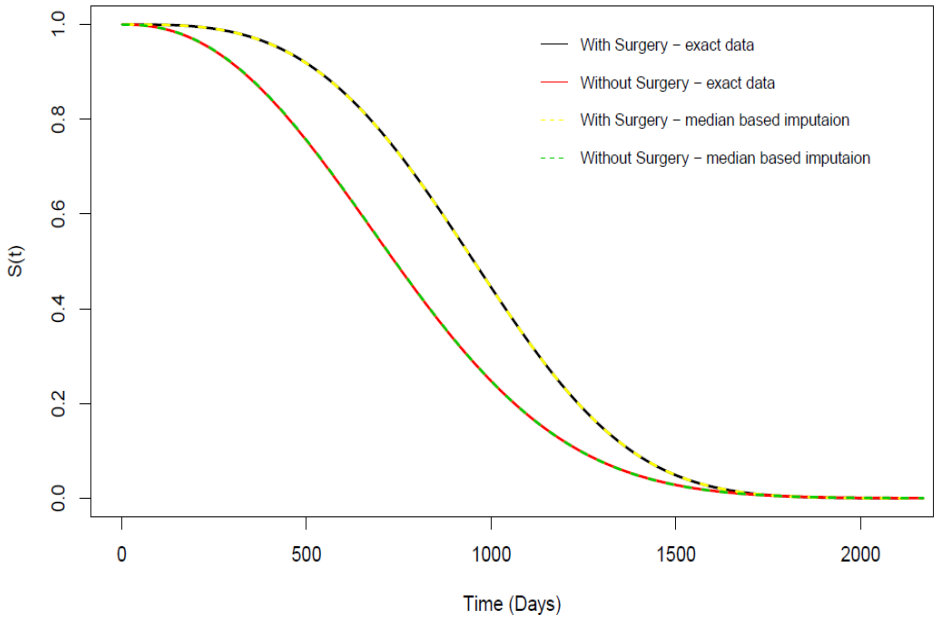

Figure 53: Eestimated of the function of survival based on surgery with 25% exact observation from median point imputation via AFT model.

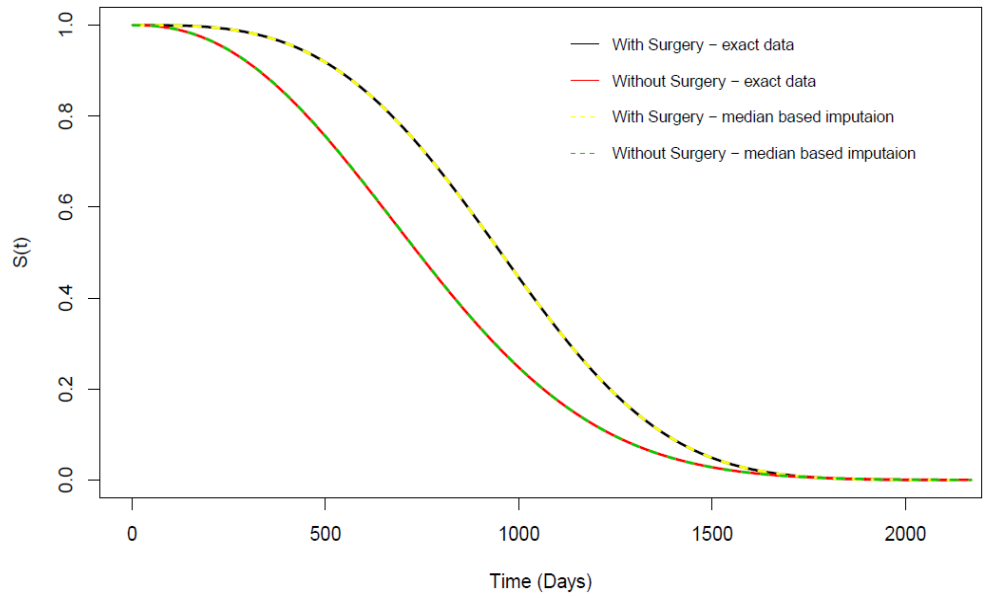

Figure 54: Eestimated of the function of survival based on surgery with 50% exact observation from median point imputation via AFT model.

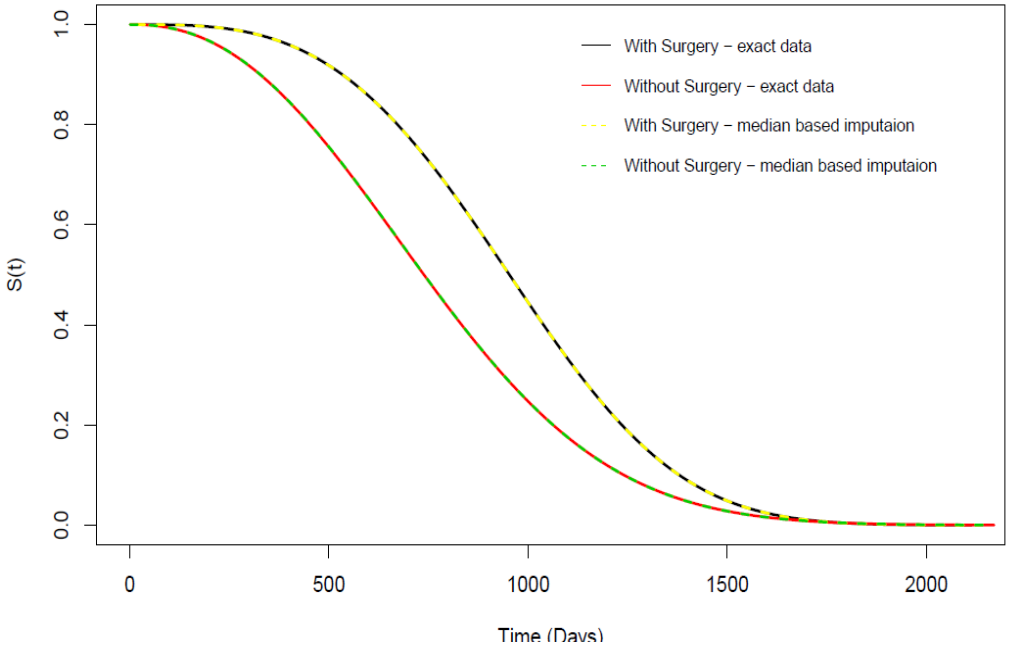

Figure 55: Eestimated of the function of survival based on surgery with 75% exact observation from median point imputation via AFT model.

We can conclude that our model via simple imputations methods fit very well the PIC breast cancer data as well as the simulation data with different percentage of exact observation in the data. We consider in the data and in the simulation data two different failure rates for each treatment that; treated with surgery& without surgery, treated RT& without RT, treated with hormone & without hormone and treated with chemotherapy & without chemotherapy.

The simulation study runs for 20000 times based on normal distribution. It showed that patients treated with; surgery, hormone, chemotherapy and RT have long survival compared to those without treatment as shown in Figures 8 to 55 & Tables 2 to 5.

Surgery is done to improve the nature of living instead of treating the illness itself, for instance, to lighten inconvenience instigated by a tumor that is pounding the nerve or bone, or to separate a tumor that covers the digestive system. Radiation treatment might be utilized to treatment for all intents and purposes each phase of bosom disease. The hormone medication after medical procedure has been found to decrease the probability of repeat of bosom disease growth in ladies with beginning phase hormone-touchy bosom disease. It likewise essentially diminishes the opportunity of bosom disease improvement and headway in individuals with hormone-delicate tumors. Chemotherapy for breast cancer is utilized as opposed to different therapies, for example, medical procedure, radiation therapy or hormone treatment.

Chemotherapy for breast cancer growth can influence the odds of a fix, limit the chance of repeat of disease, limit the indications of malignancy, or help individuals with malignant growth to endure longer with higher caliber of life. Radiation treatment is a viable way to deal with decrease the danger of breast cancer arising after medical procedure. Moreover, it is usually used to treat cancer-induced complications that have spread to different parts of the body.

CHAPTER 5: CONCLUSION AND SUGGESTION FOR FUTURE RESEARCH

CHAPTER OVERVIEW

The conclusion that summarizes the results obtained in the previous chapters will be provided in the first section of this chapter. Suggestions for Future research are discussed in the second section of this chapter.

5.1 Conclusion

In this study, we used AFT model based on Weibull distribution via simple imputation technique to simplify the procedure for partly interval censored (PIC) data. Weibull distribution model have been mostly applied in medical application. In this research project, we used it for medical data that is breast cancer data from Hamad Medical Corporation (HMC) in Qatar.

The estimated survival function was attained utilizing maximum likelihood estimator with PIC. Correlations were made with existing one under the Accelerated Failure Time model via Weibull distribution with covariate. The initial step of this examination is to search for real data set to support our model from all around fit to utilize. In view of the outcome this data set index that our model fit well and simple to execute as for the survival functions estimated (Tables 2, 3, 4 and 5), the estimation of LRT and their p-value and AIC for four medicines that is medical procedure, hormone, surgery, chemotherapy and RT.

The medical data set used to implement our methods was taken from Hamad Medical Corporation (HMC) in Qatar. The data was compiled from 1/2/2016 to 1/9/2020 including 24 variables foe cancer patients. This examination was executed to think about the corrective impacts of every therapy alone on women with early breast cancer growth and the occasion of intrigue was the opportunity to first event of bosom withdrawal and the patients were seen at facility visits, where the real dates of the occasion were recorded precisely if accessible. At that point, we altered the

informational index to be PIC information and span information for research need.

Generally, the outcome from clinical information has demonstrated that, from survival curves for the two failure times for medical procedure (treated with medical procedure and without medical procedure) as example for hormone (treated with hormone and without hormone), chemotherapy (treated with chemotherapy and without chemotherapy), RT (treated with RT and without RT) are sensible. Moreover, to that the bosom disease patients what their identity was dealt with surgery, hormone, chemotherapy and RT individually, have a more extended endurance contrast and the patients who didn't approached by these treatments regarding Figures 8 to 55 and LRT (Table 1 to 5). This outcome shows that the techniques for treatment regarding impacts on survival chances for a longer period in the event of surgery compare with other treatments.

In present exploration, we did simulation study dependent on the real breast cancer data. The date rated for 20000 times from the every medicines referenced earlier in this project with various level of the exact observations that 0% (interval), 25%, half, 75% and 100% through simple imputations techniques (right point, median and random point) to accomplished PIC data. The shape, scale and treatment coefficient utilizing AFT model dependent on MLE has been assessed. It has been seen that our model with various attributions techniques fits the data well particularly when the data is PIC. Although the right point right in the reproduction study isn't performing great particularly when we have less exact observation in the data. However, the consistency, fundamental investigation and examinations of present assessments and improved boundaries (when right point is utilized) assessment will be led in future exploration.

In the end, the result of this project is similar with other results found by other researchers such as Kim (2003), AL harpy and Ibrahim (2013), Zyoud et al. (2016)

based on PIC data. They found that there is more exact observations in data the model well be fit. The simulation study indicate that our methods is flexible and better for PIC breast cancer than interval data.

5.2 Suggestion for Future Research

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

1. More research is need when right point imputation is used.
2. Looking into the performance of our model through multiple imputation techniques such as EM algorithm.

REFERENCES

Allison, P. D. (2010). *Survival analysis using SAS: a practical guide*. SAS Institute.

Brown B. M, Wang Y. G. (2007). Induced smoothing for rank regression with censored survival times. *Statist Med.* Vol 26:828–836.

Eagle, M., & Barnes, T. (2014). Survival analysis on duration data in intelligent tutors. *International Conference on Intelligent Tutoring Systems* (pp. 178-187). Springer, Cham.

Elfaki, F. A. M., Azram, M. and Usman, M., (2012). Parametric Cox ' s Model for Partly Interval -Censored Data with Application to AIDS Studies. 2(5), pp.352–354.

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

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

Guure, C. B., Ibrahim, N. A., & Adam, M. B. (2012). On partly censored data with the Weibull distribution. *ARPN Journal of Engineering and Applied Science*, 1329-1334.

Johnson L. M, Strawderman R. L. (2009). Induced smoothing for the semiparametric accelerated failure time model: asymptotics and extensions to clustered data. *Biometrika*. 96(3):577–590.

Kalbfleisch J. D, Prentice R. L.(2002) *The Statistical Analysis of Failure Time Data.* 2nd Ed. Hoboken, New Jersey: Wiley-Interscience.

Khanal, S. P., Sreenivas, V., & Acharya, S. K. (2014). Accelerated failure time models:

an application in the survival of acute liver failure patients in India. *Int J Sci Res*, 3, 161-66.

Kim, J. S. (2003). Maximum likelihood estimation for the proportional hazards model with partly interval-censored data. *Journal of the Royal Statistical Society*: Series B (Statistical Methodology), 65(2), 489-502.

Lawless, J. F.(1982). *Statistical models and methods for lifetime data snalysis*, New York: Wiley & Sons.

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

Lee, E. T., & Wang, J. (2003). *Statistical methods for survival data analysis* (Vol. 476). John Wiley & Sons.

Liu, E. and Lim, K (2018). *Using the Weibull accelerated failure time regression model to predict time to health events*. Cold Spring Harbor Laboratory https://doi.org/10.1101/362186

Ma, Z., & Krings, A. W. (2008). Survival analysis approach to reliability, survivability and prognostics and health management (PHM). *IEEE Aerospace Conference* (pp. 1-20). DOI:10.1109/AERO.2008.4526634.

Nelson, W.B. and Hahn, G.J.(1972). Linear Estimation of a Regression Relationship from Censored data: 1,Simple Methods and Their Applications (With Discussion), *Technometrics,14,247-276.*

Newby M. (1988). Accelerated failure time models for reliability data analysis. *Reliability Engineering & System Safety*, Vol 20(3): 187-197.

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

Pike, M. C. (1966). A method of analysis of a certain class of experiments in

carcinogenesis, *Biometrics*, 22:142-61.

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

Saikia, R., & Barman, M. P. (2017). A Review on Accelerated Failure Time Models. *International Journal of Statistics and Systems*, 12(2), 311-322.

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

Vallinayagam, V., Prathap, S., and Venkatesan, P. (2014) Parametric Regression Models in the Analysis of Breast Cancer Survival Data, *International Journal of Science and Technology*, 3(3) 2049-7318.

Wei, L.J. (1992) The Accelerated Failure Time Mode: A Useful Alternative to the Cox Regression Model in Survival Analysis, *Statistics in Medicine*, vol 11:1871-1879.

Weybright, E. H., Caldwell, L. L., Xie, H., Wegner, L., & Smith, E. A. (2017). Predicting secondary school dropout among South African adolescents: A survival analysis approach. *South African Journal of Education*, 37(2).

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

Yousif, Y., Elfaki, F. A. M., & Hrairi, M. (2016). Analysis of Partly Interval-Censored Data under Competing Risks Framework. *IJCSNS*, 16(8), 25.

Zhao, X., Zhao, Q., Sun, J., & Kim, J. S. (2008). Generalized log-rank tests for partly

interval-censored failure time data. Biometrical Journal: *Journal of Mathematical Methods in Biosciences,* 50(3), 375-385.

Zhou, Y. (2005). International technology diffusion: the case of robotics. Academy of International Business Annual Conference, Québec City, Canada, July 9-12, 2005.

Zyoud, A., Elfaki, F. A. M., & Hrairi, M. (2016). Nonparametric extimate based on imputations techniques for interval and partly interval censored data. *Science International,* 28(2).