USING SURVIVAL ANALYSIS TO INVESTIGATE BREAST CANCER IN THE KURDISTAN REGION OF IRAQ

CITY, UNIVERSITY OF LONDON

By

Mahdi Saber Raza

Supervised by

Professor Mark Broom

Thesis submitted to City, University of London

for the degree of Doctor of Philosophy

City, University of London

School of Mathematics, Computer Science and Engineering

Department of Mathematics

October 2016

Table of Contents

List of tables

List of tables in Appendix A1

Appendix **Page**

List of tables in Appendix A2

Appendix **Page**

List of figures

Figure Page

List of figures in Appendix 3

Acknowledgements

I am grateful to my supervisor, Professor Mark Broom, for his guidance and advice from the start of my research to the completion of this thesis and among the staff and students at the School of Mathematics, Computer Science and Engineering, City University London, all those who taught me or helped me during my studies, most particularly, Professor Martin Newby for his encouragement at the beginning of my study.

Thanks also to my family-my wife Tara for her support and tolerance throughout completion of this work. Finally, I would like to dedicate this thesis to the memory of my mother Hajiah Fatima. My father and mother in-law for their inspiration; no-one could have appreciated my endeavours better.

Declaration

I hereby declare that this thesis is my work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

Mahdi Saber Raza

October 2016

Abstract

The objective of this thesis is to carry out a survival analysis for patients with breast cancer. Using data from the Nanakaly and Hewa hospitals in the cities of Erbil and Suleimaniah, respectively, cases where there is hidden censoring on survival time were investigated. The aim of this study was to identify the main risk factors and quantify the overall risk for breast cancer. We developed a new Markov chain-based method for generating survival curves and hazard functions. In particular we adjusted the Kaplan Meier analysis to find a survival curve with hidden censoring of the data, and also estimated a survival function from the biased one obtained directly from the data by generating new models in two cases; with and without censoring. To ensure the validity of the suggested model we considered different simulation techniques applied to the Nanakaly data. Because of the availability of a good survival function, we chose to work with a German data set. As a result we conclude that our model performs well in many circumstances, and its predictions, even when less accurate, are always an improvement on considering the apparent survival curves from the unadjusted data.

For the data from Nanakaly hospital, the only variable we had to consider was age at diagnosis and the survival results showed that this was a significant variable. With far more detailed reports available for Hewa hospital, we were able to identify estrogen abundance, smoking habits and tumour grade, as having a statistically significant impact on the incidence of breast cancer. On the other hand, when analysing the Nanakaly and Hewa data for comparison with German data, in all three cases the survival curve is greater among younger patients. The suggested models may be verified using cross validation or by using new data.

Finally, we note that it would be preferable to have accurate data to applying our methods to imperfect data. Therefore we established both a general and a specific flowchart to collect the data in the future. Encouraging the implementation of the recommended procedures might serve to obtain the data needed to develop a more comprehensive understanding of breast cancer in Kurdistan.

1 CHAPTER 1: Introduction

1.1 Introduction

In general, cancer is a type of disease which makes the cells divide, grow and multiply uncontrollably. The cancer is named after the part of the body which it starts from. Breast Cancer means the unregulated growth of the cells which arise in the breast tissues and its multiplication and spread. Infected cells which divide and multiply rapidly may form a mass of extra tissues. These mass tissues are called tumors. The tumours are either cancerous (malignant) or (benign). The malignant tumors multiply and invade the intact tissues of the body (Carol 2005).

Most kinds of breast cancer start from the inner lining of milk ducts and therefore are known as ductal carcinomas, whereas so called lobular carcinomas appear in the lobules. When breast cancer starts to spread outside the breast, the cancer cells reach the lymph nodes under the armpit. In this case, the cancer starts to spread to all body lymph nodes (Carol 2005).

The age-adjusted method had been used by Althuis et al. (2005) from 1973-77 to 1993-97. Based on their study the incidence of breast cancer appears to be increasing world-wide, as it rose by 30-40% from the 1970s to the 1990s in most countries. According to information collected by the American cancer society program of the National Cancer Institute (NCI) it is estimated that 246,660 women will be diagnosed with, and 40,450 women will die of, cancer of the breast in 2016 in the USA (Rebecca et al. 2016).

In Western countries there is plenty of information about breast cancer which shows an increase in breast cancer cases but declining mortality. This may be due to the increasingly effective treatments and early detection (Rennert 2006).

Information about non-Western countries is limited and there are problems with accuracy, however, breast cancer may still be on the rise (Rennert 2006). The cultural tendency in the West of late marriages and limited childbearing results in fewer children being raised and breastfed, which is actually an activity protective against breast cancers forming after menopause (Rennert 2006). Although the studies are limited, Egypt is believed to have the highest rate of breast cancer in the Middle East, with the largest increase in women aged 30 to 60 (McCredie et al., 1994; Rudat et al., 2013).

In Iran and most of Middle Eastern countries, age is considered a main risk factor for breast cancer. For example a study in Iran found that women who were never married and women whose family members have suffered from breast cancer have a significant increase in the risk of breast cancer (Montazeri et al., 2003).

The age distribution of breast cancer patients in Iraq was quite similar to that in Iran, Egypt, Jordan, and among Israeli Arabs (Ebrahimi et al., 2002: Montazeri et al., 2003). In these countries, age at diagnosis had an average and median of less than 50 with a quarter of them being less than 41 years of age. In comparison to the United States, the median is 61 years of age, 65% are older than 55, and 10.6% less than 44 years of age (Ries et al., 2008).

A study by Hussein and Aziz (2009) suggests that decreased child birth could possibly increase the risk in Kurdish women of older age because patients more than 50 years of age who had never been pregnant or given birth to fewer children in comparison to the controls had breast cancer. The prevalence of age specific breast cancers for Jordanian and Israeli Arab women is around 25 to 50% less than that in the United States up until the age of 40 to 44 years (Rennert 2006).

Traditionally, Middle-Eastern cultures have had large families which may imply that increased childbearing protects younger women against those breast cancers that develop when they are older. In comparison to the controls, Kurdish patients with a family history of breast cancer were more likely to suffer from it according to (Othman et al., 2011). Almost 20% of the Kurdish breast cancer patients have a family history of breast cancer but this high figure is offset by lower rate of 7.1% reported by the control which is much lower than Western standards of 17% (Majid et al., 2012). There were near equal age distributions between patients and controls ruling out the possibility that there may have been a large proportion of pre-menopausal women with breast cancer.

1.2 Problem statement

Cancer incidence is increasing all over the world (Curado et al., 2007). Despite new advances in cancer research, the ethiology of many types of cancer is still unknown. Some independent reports from different cities of Iraq had showed an increased incidence of different types of cancer; (Al-Humadi 2009) showed increase in risk of colon cancer in Iraq of 25% to 50% during a 30 year period (1965-1994). (Habib et al., 2007) demonstrate overall increase in incidence cancer rate as compared to previously reported figures in Basrah. Additionally, their study showed a high cancer incidence rate especially among females compared to neighbouring countries, with an age standardized rate (ASR) of 123.4 and ASR of 114.3 for both males and females respectively (Habib et al., 2007). The Kurdistan Region as a part of

Iraq has been exposed to many environmental and epidemiological changes that predispose it to an increase in the risk of cancer in this region. On one hand, there is a shift toward the western-style of living and dietary habits, on the other hand there are the effects of chemical hazard of the Iraqi/Iranian War for 8 years from (1980-1988) and use of chemical weapons against Kurdish people from the Central Government for example the chemical bombardment of Halabja City in Kurdistan on 1988 (Salih 1995). In Kurdistan, North of Iraq, no research has been undertaken to identify the incidence rate of cancer and to highlight the increased risk of malignancy in this region.

There are numerous studies on Survival Analysis in different countries, on various types of diseases. However, there are hardly any Survival Analysis studies performed on breast cancer by Iraqi researchers in general and researchers in the Kurdistan Region in particular. Because of the importance of this for both society and the individuals, this work considers this subject, in order to bring it to the attention of the policy makers in the Ministry of Health and providing the doctors with ways and tools to control the incidence of cases of breast cancer in the region.

1.3 Research question and significance of the study

These questions are as follows:

1. What are the main risk factors in breast cancer in the region?
2. What are the factors that have a big influence in breast cancer’s increase in the region?
3. How can we construct an applicable survival function model with limited or problematic data?
4. How can we collect better data in the future?

The main goal of this study is to inform the Kurdish Government and Ministry of Health on the treatment of breast cancer, as the Government is the largest provider of health care in the region.

In this study we construct a new model based on data from the Nanakaly and Hewa hospitals, which can be used to compensate for problems with the existing data. This analysis was then helpful with analysing this data.

From the available data we generated two new models, presuming censoring and disregarding it for Nanakaly and Hewa data. For the first case, we estimate only the number of observations while in the second one we additionally estimate the number of deaths. These models might serve to advance healthcare infrastructure in various respects, specifically the collection of data. This model can also be useful to scientists and public servants from all sectors in identifying any possible gap in their datasets.

1.4 Research aims and objectives

The project is strongly focused on the Kurdistan Region of Iraq. The aim is to find whether there has been an increase in the incidence of breast cancer in Kurdistan, to quantify the risk, and to identify the main risk factors for breast cancer in Kurdistan. Thus the research will directly benefit the community.

Objectives:

1. Carry out an extensive literature review regarding breast cancer in Kurdistan and more generally, and on survival analysis methodology as applied to breast cancer.
2. Carry out work in preparation for obtaining the data from Kurdistan (see objective 3): in particular obtaining ethical approval, using SPSS for data analysis, choosing the appropriate survival analysis methodology (see chapter 4 for more detail).
3. Collect the breast cancer data from two main hospitals in Kurdistan; Nanakaly and Hewa, both of which specialize in all types of cancer.
4. Carry out a systematic analysis of the data obtained in objective 3 by using SPSS program package version 22.
5. Develop new Markov models to deal with problems related to censoring and recorded deaths with the data.
6. Develop a methodology for collecting better data in the future.

1.5 Methodology

Classical survival analysis is applied to this data, including Cox regression to determine the significant risk factors, the Kaplan-Meier method to find the survival curve for the chosen significant variable and log-rank tests to compare within each specific variable. These methods are described in detail in chapter four. The type of data supplied by each hospital required us to develop a new Markov chain based model in order to properly carry out the above analysis, and in particular to find the survival function as described in chapter five.

In this study we encountered some problems while attempting to apply classical survival analysis, due to the limitations of the data collected. The reports from Nanakaly hospital featured instances of individuals being lost to the study, the implications of which for the statistical analysis are described in detail in chapter six. For the data set from Hewa hospital, in addition to the problems of lost individuals, the definitive times of death were not recorded. We discuss the shortcomings of the data in detail in chapter six, specifically their effects on the validity of the models. These limitations of the data are to some extent expected because most of the reports are recorded on paper rather than electronically in a database.

1.6 Structure of the thesis

This thesis is comprised of eight chapters: Chapter One contains an introduction, problem statement, research question and significance of the study, research aims and objectives, methodology, scope and limitation of the study and finally the structure of the thesis. Chapter Two deals with a review and critical analysis of selected literature relevant to the study topic which will help determine the current state of research in the areas of breast cancer. Chapter Three discusses the theoretical concepts of Survival Analysis. Chapter Four explains the basic features of the data.

Chapter Five discusses the Markov-chain model for breast cancer and it includes applications of survival analysis for data from Nanakaly and Hewa Kurdish Hospitals, Markov models without and with censoring for Nanakaly and Hewa data and simulations for the Nanakaly data. This work has been published in Raza and Broom (2016)

Chapter Six explains survival analysis for breast cancer; survival analysis for Nanakaly, Hewa and German data, the connection between German and Nanakaly data, the connection between German and Hewa data, the connection between German, Nanakaly and Hewa data and finally the connections between unadjusted and adjusted data for Nanakaly and Hewa data.

Chapter Seven discusses a proposed data collection methodology for Kurdish hospitals, including general procedures to collect the data, itemising the data required for survival analysis, the feasibility of the plan and presentation to users.

Finally Chapter Eight presents conclusions and future works.

The following, Figure 1.1, presents a brief summary of the chapters involved in this study:

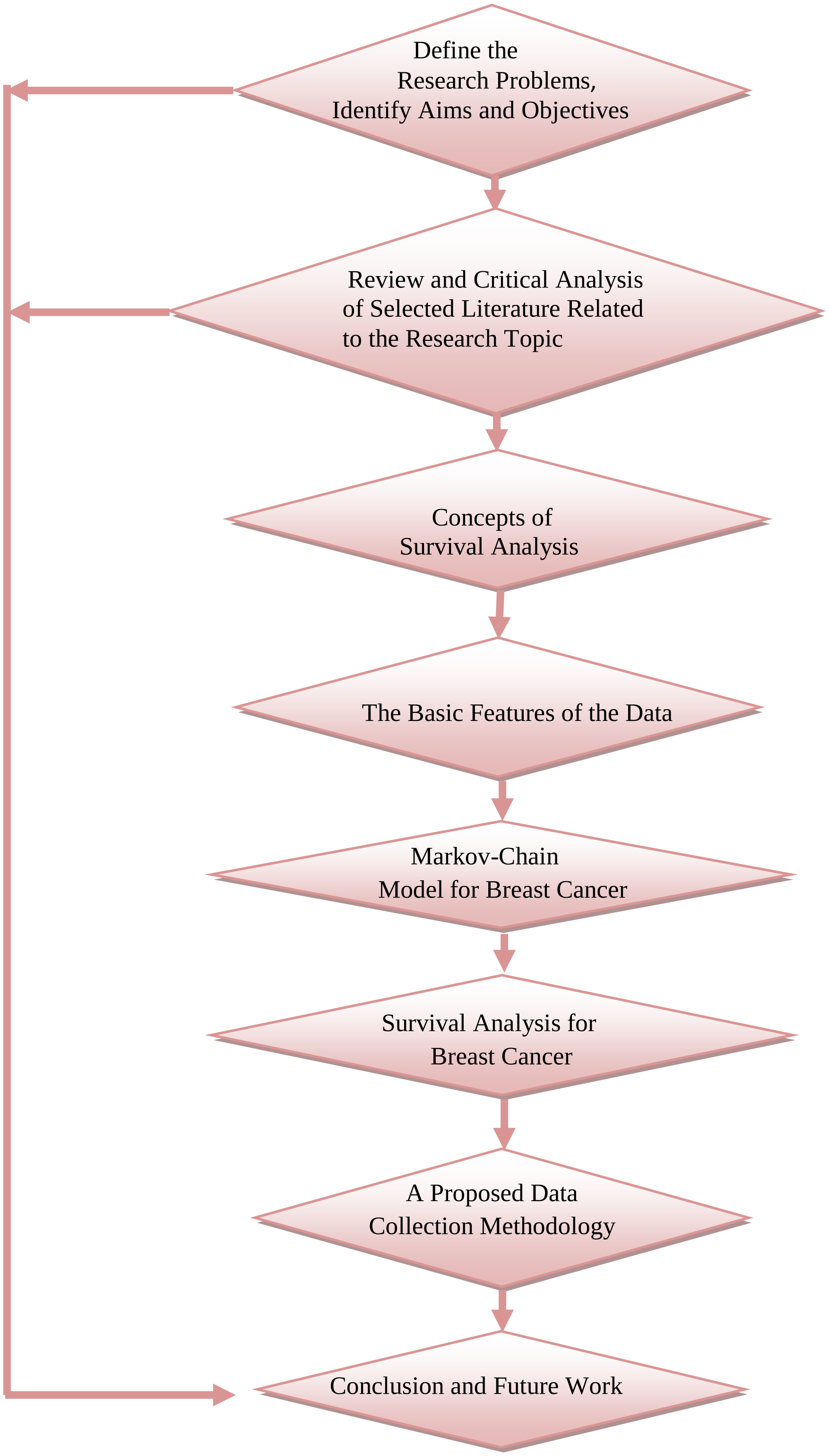


Figure 1-1 Structure of the thesis

2 CHAPTER 2: Literature review

2.1 Introduction

Breast cancer is the most frequent cancer diagnosed in women and the second most common in all humans (Darendeliler and Ağaoğlu 2003; Ozmen 2008). Nearly 25 million are estimated to be living with cancer (Kamangar et al., 2006), and it is a leading cause of death worldwide (WHO 2008). Although its prevalence varies in different societies, it is accepted that one out of 8-10 women in Western society is likely to develop breast cancer during her lifespan. Over half a million women are estimated to have died in 2004 alone due to breast cancer, and while this disease is often regarded as a disease of the West, almost 70% of all breast cancer deaths actually occurs in developing countries (WHO 2008). However, the prevalence of the disease is notably higher in North American and European countries than the rest of the world (Stewart and Kleihues 2003).

Belgium and North America head the table with age standardized rates of occurrence as high as 99.4 per 100,000 women but the rates of breast cancer occurrence vary greatly, with Eastern Europe, South America, Southern Africa, and western Asia showing moderate, but rising incidence rates. The UK currently has the 11th highest breast cancer rate with 89.1 of every 100,000 women every year expected to develop breast cancer (NHS Choices. 2011). Breast cancer is also the most commonly diagnosed of all cancers among Malaysian women, accounting for 16.5% of all cancer cases registered in 2006 (Omar et al., 2006). The lowest numbers are to be found in most African countries but here numbers are also on the increase.

Breast cancer survival rates vary in a similar way, ranging from 80% or over in North America, Sweden and Japan to around 60% in middle-income countries and below 40% in low-income countries (Coleman et al., 2007: 2008). The latter can be explained not only by the lack of early detection programmes in poorer countries, resulting in a much higher number of women diagnosed only in the later stages of the disease, but also by the lack of sufficient diagnosis and treatment facilities available in under-developed countries, especially in Africa.

In developed countries it is the second most deadly disease, and in developing countries it is one of three leading causes of death (Parkin et al., 2005; WHO 2004). Globally breast cancer accounts for approximately 23% of all female cancers according to recent research (Parkin et al., 2005).

The 5-year survival rate of people suffering from breast cancer is 90.3% overall. A study carried out by the World Health Organisation (WHO) in 1990 reported that there were 796,000 breast cancer cases and 482,000 censored due to the disease in 1990. A more recent study of the same kind carried out by the International Agency on Cancer Research (IARC) in 2002 reported 1,152,000 new cases and 741,000 survivals. With all phases of the disease considered, the five-year survival rate has been reported as 73% in developed countries and 53% in developing countries. Early diagnosis via mammography scans and better treatment in developed countries may explain this significant difference.

2.2 The history of breast cancer

Breasts are made up of fat, connective tissue and thousands of tiny glands known as lobules which produce the milk. To allow breastfeeding in women who have given birth, milk is delivered to the nipple through tiny tubes called ducts. Bodies are made up of billions of tiny cells which grow and multiply in an orderly way under normal conditions. New cells are created only when and where they are needed. This process goes wrong and cells begin to grow and multiply in an uncontrollable way in cancer. Breast cancer usually shows as a lump or thickening in the breast tissue (although most breast lumps are not cancerous). If the lump can be detected at an early stage then treatment is usually successful in preventing spreading to nearby body parts (Ananya Mandal 2013).

Even since ancient times, cancers have been known to human beings and indeed have been mentioned in almost every period of history. Early physicians were able to record the visible symptoms of the disease, especially in the later stages when the lumps progress to tumours. This is even more evident in the case of breast cancer because, unlike other internal cancers, these lumps most often become apparent as visible tumours in the breast. The ancient Egyptians were the first to record the disease more than 3,500 years ago. Both the Edwin Smith and George Ebers papyri describe the condition with a good degree of accuracy. One of the descriptions refers to bulging tumours of the breast that have no cure.

In 460 B.C., Hippocrates, known as the father of Western Medicine, described breast cancer as a humeral disease. His belief was that the body consisted of four humors - blood, phlegm, yellow bile, and black bile and his idea that cancer was caused by an excess of black bile was probably based on observations that in appearance, breast cancer was black, since hard tumours were seen to burst if left untreated and would produce a black fluid. He named the cancer *karkinos*, a Greek word for "crab," because the tumours seemed to have tentacles like the legs of a crab.

In A.D.200 Galen described cancer also suggesting the cause was excessive black bile but theorized that some tumours were more dangerous than others. Medications prescribed were opium, castor oil, licorice, sulphur, and salves for medicinal therapy. During this time of history, breast cancer was regarded as a disease that affected the whole body and so surgery was not considered. Galen's theories on breast cancer held success even into the late 17th century when in 1680, a French physician called Francois de la Boe Sylvius realized the humoral theory of cancer had to be reconsidered. He hypothesized that cancer did not stem from too much black bile but came from a chemical process affecting lymphatic fluids, changing them from acidic to acrid, as he described. In the 1730s, a Paris physician Claude-Deshais Gendron also rejected Galen's theory and postulated that cancer developed when nerve and glandular tissue mixed with lymph vessels and formed hard masses only curable by removal.

Bernardino Ramazzini hypothesized in 1713 that the high frequency of breast cancer in nuns was due to lack of sexual intercourse. He said that without regular sexual activity the reproductive organs of women, including the breast, may fall into decay and develop cancers for this very reason. Friedrich Hoffman of Prussia said that women who had regular sex still developed cancer claiming that it must be due to the fact that they were practicing energetic sex, which his theory said could lead to lymphatic blockage.

There have been plenty of other theories about the causes of breast cancer over the years, including Giovanni Morgagni who blamed curdled milk, Johanes de Gorter who cited pus-filled inflammations in the breast, Claude-Nicolas Le Cat from Rouen who held depressive mental disorders responsible, Lorenz Heister who mentioned childlessness, others blaming a inactive lifestyle to name but a handful.

It was 1757 when Henri Le Dran, a leading French physician of his time, recommended that surgical removal of the tumour could help treat breast cancer as long as the infected lymph nodes of the armpits were also removed. Claude-Nicolas Le Cat similarly argued that surgical therapy was the only method to combat this type of cancer. By the mid-nineteenth century, surgery was available as an option for curing breast cancer and what's more, the development of antiseptics, anesthesia and blood transfusion during this time also made survival after such a surgery more likely.

This prescribed course of treatment lasted well into the twentieth century and led to the establishment of the radical mastectomy or extensive removal of the breast as the main approach to dealing with breast cancer. William Halstead of New York, in particular, made radical breast surgery the norm for the next 100 years. He developed a form of mastectomy that involved the removal of breast, axillary nodes (nodes in the armpits), and both chest muscles in a single *en bloc* procedure as the only certain means to prevent the spread of the cancer if these were removed one-by-one.

Radical mastectomy was the most common form of treatment for the first four decades of the twentieth century but though this radical mastectomy helped women survive longer, especially if performed in the early stages of the disease, many women rejected it as an option since it left them severely disfigured. In addition, there were related problems caused such as a deformed chest wall, lymphedemae or swellings in the arm due to the lymph node removal, not to mention not inconsiderable pain.

In 1895, a Scottish surgeon called George Beatson discovered that removing the ovaries from one of his patients led to a reduction in the size of her breast tumour. This he surmised was due to the fact that estrogen from the ovaries could be shown to help in the growth of the tumour and their removal thus helped reduce the size of breast tumours. As this link became more established, many more surgeons began removing both ovaries as a treatment for breast cancers.

Next it was found that estrogen was still being produced in women without ovaries by the adrenal glands and so in the 1950s Charles Huggins performed an adrenalectomy, in that way removing the woman’s adrenal gland so as to deprive the tumour of any further source of estrogen. But Rolf Lefft and Herbert Olivecrona also initiated the removal of the pituitary gland as it became known as another site of estrogen production. Again in the 1950s, the systemic theory of cancer started to become widespread. George Crile first suggested that cancer was not a local disease but instead one that spreads throughout the body. Bernard Fisher also recognized a cancer’s capability of metastasizing.

Fisher publicized his results in 1976 of using breast-conserving surgery followed by radiation or chemotherapy. He proved that his methods of treatment were just as effective, and so preferable to, radical mastectomy, which was the method still most widely applied in the fight against cancer at that time. However, with the arrival of modern medicine less than 20 years later, less than 10 percent of breast cancer-afflicted women would have to undergo a partial or full mastectomy. The 1990s also saw the development of more innovative therapies for breast cancer including hormone treatments, surgeries and biological therapies. The X-ray of the breast (Mammography) was also a significant development for the early detection of the cancers and now scientists have even isolated the genes that actually cause breast cancer: in particular BRCA 1, BRCA2 and ATM.

While in earlier years individuals were ashamed to be a victim of breast tumours meaning that detection and diagnosis was rare and made the mention of breast cancers in any literature beyond those of a medical nature relatively uncommon, recent developments dating back three or four decades or so have seen the involvement of more and more women in movements that actively bring the disease out into the open thus breaking old taboos. A case in point is the 1990s' symbol of breast cancer, the pink ribbon, which brought about a revolution against this particular form of cancer, together with several high–profile women like Angelina Jolie going public with their personal experiences of this disease.

2.3 Risk factors of breast cancer

Breast cancer is a complex disease and no single cause can be isolated but research has identified a number of risk factors linked to increased likelihood of this appearing. Reviewed studies by Hider and Nicholas (1999) add to the information in the Pullon and MacLeod (1996) report and generally endorse the results reported there where family history is confirmed as a top-of-the-list high risk factor, as is childhood treatment for cancer. As expected, increasing age is also associated with increasing risk.

However, new research has been done to change our depth of knowledge on lesser risk factors. Studies are still being published which investigate the effects of nulliparity (the non-bearing of children), and age at first and last birth. These results are consistent with the work published by Pullon and MacLeod (1996). A number of studies on diet, alcohol, BMI, smoking and exercise have been published in the period covered by this review. Nevertheless, however small these risk factors may be, they are modifiable ones of relevance to all women. A number of additional factors to those discussed in (Pullon and MacLeod 1996) should be noted, such as the factor of previous abortions (seemingly a no or very small risk factor), prenatal environment (including related factors such as hormone levels), environmental toxins, electromagnetic radiation, stress and occupation-linked factors.

Thus while it is true to say that a number of risk factors for breast cancer have been reasonably well documented, for the majority of women with breast cancer, it is not always possible to identify risk factors (IARC 2008; Lacey et al. 2009). A family history of breast cancer increases the risk by a factor of two or three. Particular mutations like BRCA1, BRCA2 and TP53, result in much higher risks of breast cancer. However, these are rare and account for a small portion.

Prolonged exposure to endogenous estrogens related to the reproductive cycle, such as early menarche (onset of menstruation), late menopause, and a later age of childbirth for the first time are also among the most important risk factors for breast cancer to have been identified. Exogenous hormones also suggest a higher risk thereby oral contraceptive and hormone replacement therapy users are at a higher risk. On the other hand, breastfeeding has been shown to have an inhibitive effect on the likelihood of developing the disease (IARC 2008, Lacey et al., 2009).

Just how various modifiable risk factors, excluding reproductive factors, contribute to overall breast cancer incidence was calculated by Danaei et al., (2005). They conclude that 21% of all breast cancer deaths worldwide can be blamed on alcohol use, being overweight or obese, and lack of physical exercise. High-income countries comprise a higher proportion (27%) of global figures with obesity being identified as an important factor. In low and middle-income countries, these particular risk factors fall to 18%, and lack of exercise was the most important factor (10%).

The differences in breast cancer incidence in developed and developing countries can also partly be explained by dietary differences combined with age at first childbirth, nulliparity rates, and duration of breastfeeding (Peto 2001). The conclusion is that the increasing adoption of western life-styles in lower and middle-income countries is directly related to the increase of breast cancer.

2.4 Breast cancer research in developed countries

Breast cancer incidence rates vary greatly around the world. It is lower in less-developed countries and greatest in the more-developed countries (Hussein and Aziz, 2009). It is the most commonly diagnosed cancer among women. More than 1.1 million women globally are newly diagnosed yearly. It accounts for at least 1.6% of worldwide female deaths annually (Lan et al., 2013).

Annual incidence rates per 100,000 women (figures age-standardized) in the following regions have been recorded as follows: Eastern Asia, 25.3; South-Eastern Asia, 31; South Central Asia, 24; Western Asia, 32.5; sub-Saharan Africa, 22; North Africa, 32.7; South Africa, 38.1; Western Africa, 31.8; Middle Africa, 21.3; North America, 76.7; Central America, 26; South America, 44.3; Eastern and Central Europe, 45.3; Western Europe, 89.9; Northern Europe, 84; Southern Europe, 68.9; Oceania, 74 and in Australia / New Zealand 85.5 (Stewart and Kleihues 2003 and Jemal et al. 2011). In the UK it affects about 48,000 women annually where 8 out of 10 are over 50 (NHS 2012)**.**

A group of 4,764 women from Piedmont, Italy, diagnosed with breast cancer between 1979-81 was studied by Boffetta et al. (1993). It was followed by a study on mortality until 1986 or 1987. It established that there were better survival rates for women aged 40-49 at time of diagnosis. Between one and four years after diagnosis mortality peaked, and was lowest between five and seven years. Women aged 80 and over had the lowest survival rates. Prognosis for single women was worse than married women. Alternatively, Yang et al., (1998) showed that family history has more significant impact in causing breast cancer in the USA using a proportional hazard model, logistic regression and line of best fit.

In a study by Tokunaga, et.al. (1987) relating to mortality rates in breast cancer cases associated with radiation, it was indicated that in first four decades of life exposure of female breast tissue to radiation, especially in early childhood, can cause breast cancer to develop later in life. Moreover, the length of time that tumour promoters such as endogenous hormones operate following exposure plays an important role in the future development of radiation-induced breast cancer. Similarly, Little and Boice (1999) compared breast cancer occurring among Japanese women exposed briefly to atomic bomb radiation and among Massachusetts women exposed repeatedly over an extended period of time to medical radiation as part of tuberculosis therapy. Results showed that the excess relative risk of breast cancer incidence in the Japanese atomic bomb survivors was significantly higher than that in the Massachusetts fluoroscopy patients.

In a detailed study, Little and Boice (1999) reported the best estimate of the ratio between the excess relative risk coefficients for the Massachusetts and Japanese cohorts to be 2.11 (95% CI 1.05, 4.95). However, the higher risk can be explained by the lower (baseline) risk among all Japanese women in comparison to all Massachusetts women, thus the excess absolute risks are statistically the same (two-sided $P = 0.32$). A significant finding in their study was an early appearance of radiation-induced breast cancer among Japanese atomic bomb survivors but not in the Massachusetts sample. Differences in the patterns of risk were seen (two-sided $P = 0.04$) over the period of time following exposure between two groups who had been exposed to radiation in childhood. Overall there is no difference between the Massachusetts and Japanese data sets in terms of age and time distribution of risk of radiation-induced breast cancer is concerned and their data also provided minimum evidence for a prognosis of reduced breast cancer risk after fractionated irradiation (Little and Boice, 1999).

Studies on the effect of gene mutations on breast cancer survival are contradictory. No difference was found by Lee et al (1999). However, Johannsson et al. (1998) found evidence that it may be worse for some changes, specifically BRCA1 carriers. Watson et al. (1998) confirmed this for a specific situation only. Rubin et al., (1996) found higher probability of survival of BRCA1 carriers. Edwin et al. (2000) found the relationship between BRCA1 and BRCA2 carrier status and survival after breast cancer by using the Cox proportional hazards model. The survival rate of non-irradiated mutation carriers (0.990) is higher than that of non-carries, more specifically 0.844 for BRCA1 and 0.924 for BRCA2.

Robson et al. (2004) studied the prognostic significance of germline change in BRCA1 and BRCA2 in women with breast cancer. To address the lack of convincing data available in this area of research a combined analysis was performed. It was found that BRCA1 mutations are associated with reduced survival in Ashkenazi women receiving breast-conserving treatment for invasive breast cancer. The poor prognosis related to germline BRCA1 change can be removed by more chemotherapy. Observations after 10 years of follow up show that the risk of metachronous ipsilateral disease does not amplify in either BRCA1 or BRCA2.

Konecny et al., (2003) studied the correlation between HER-2 and hormone receptor expression. They analyzed HER-2~~/neu~~, estrogen receptor, and progesterone receptor as continuous variables in breast cancer cell lines in two cohorts of primary breast cancer patients. It was shown that reduced ER/PR expression may be one explanation of relative resistance of HER-2 to endocrine therapy.

A retrospective analysis of records from Surveillance, Epidemiology and End Results (SEER), and Medicare claims was considered by Goodwin et al. (2004) to investigate the effect of previous history of depression on diagnosis, treatment and survival of older women with breast cancer. It was shown that there is a higher risk of death associated with prior diagnosis of depression for women receiving treatment. It was found that women with a relatively recent diagnosis of depression are at greater risk than others of receiving treatment and have worse survival rates after breast cancer diagnosis, and differences in treatment cannot explain worse survival rates.

In a study by Aggarwal et al. (2008) the association of symptoms of depression in postmenopausal women with breast or colorectal cancer and both screening rates and stage of cancer was investigated. They found that among a self-motivated and healthy cohort of women, depressive symptoms reported by the patient herself could be associated with lower rates of screening mammography but this was not so in the cases of colorectal cancer screening. No association was found as regards stage of cancer at diagnosis.

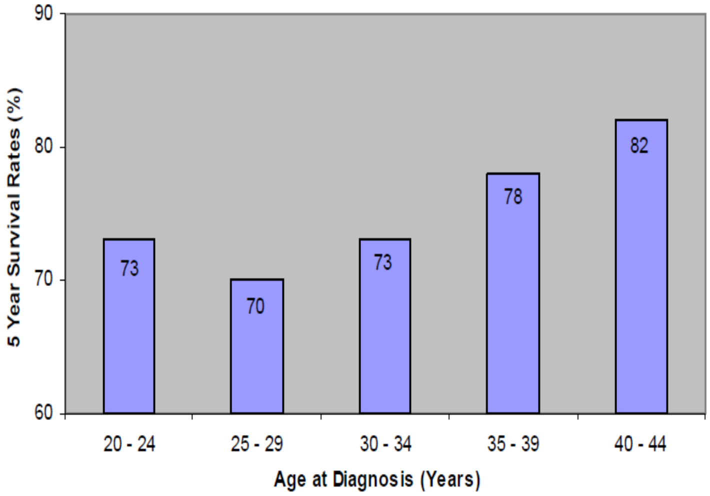
Research has indicated that the risk increases by 30% - 50% in obese women. Obesity has been shown to be associated with worse health regardless of their menopausal status. There is an increased risk of 33% in obese women in comparison to women with acceptable BMI, (Rudat et al., 2013). Denmark-Wahnefried et al. (2005) also reported that cancer survivors are more likely to be obese and subject to more ongoing diseases than the general population. Ogle et al. (2000) similarly reported 68.7% of comorbidity (the presence of an additional disorder) among 15,626 cancer survivors. Chlebowski et al. (2002) stated that comorbid conditions are related to reduced survival and increased mortality. The association of both obesity and lack of physical activity with cancer recurrence and overall survival in cancer survivors is well documented (Holmes et al., 2005; Meyerhardt et al., 2006; Patnaik et. al., 2011; Pierce et al., 2007). For these reasons, it can be seen that switching to a health-promoting lifestyle is advisable for cancer survivors to defend against subsequent diseases and improve prognosis and survival in addition to bettering their overall health. There are therefore many researchers who recommend this (Blanchard et al. 2003; Patterson et al. 2003; Satia et al., 2004). However, many studies also report difficulties in adjusting lifestyle (Denmark- Ahnefried et al. 2005; Irwin et al., 2005). Irwin et al., (2005) reported an example of this where 68% of women gained weight after being diagnosed.

Despite the awareness in cancer survivors, few health behaviour differences have been reported between them and controls (Bellizzi et al. 2005; Caan et al., 2005; Coups and Ostroff, 2005). This highlights the need for intervention to promote healthy lifestyles based on the evidence of higher rates of comorbidity within cancer survivors and that unhealthy lifestyles increase the risk of other cancers. However, data providing evidence when this would be most beneficial is lacking (Denmark-Wahnefried and Jones 2008).

In another study from Denmark (Olsen et at., 2012) examined the possible connection between a common major life-changing event such as the loss of a partner, and the repetition of breast cancer and all-cause mortality using Cox regression analyses. It was concluded that women who were widows before diagnosis or in the years immediately afterwards did not have a significantly higher risk of suffering a return or even dying than women who had not been through a similar shock. In other words, the results did not support the fear that a stressful life event like the loss of a partner may adversely affects prognosis in breast cancer.

Desreux et al. (2011) remarked on the fact that Belgium has the highest breast cancer incidence of all European countries, with 9,697 new cases in 2008 and 106/100,000 women affected per year. The explanation they give for this high incidence is an accumulation of lifestyle risk factors, and the impact of screening and registration of cases. The relative bearing of each of these factors on statistics cannot be clearly ascertained due to a dearth of relevant powerful statistical studies. The rate in Belgium is just above the European mean for breast cancer mortality (19.4/100.000 women per year) with an all stages 15-year survival rate of 75%. Their article investigates the causes of this high national incidence and reasons for the current decrease of cancer incidence recorded in western countries, while reviewing both familiar and less known risk factors of breast cancers.

The American Cancer Society reports that around 250,000 breast cancer cases will be diagnosed in the U.S. per year, and of these, almost 10 percent will affect women under the age of 45. While this percentage may sound relatively insignificant, in comparison to the total number of women diagnosed annually, it is a noteworthy ratio particularly when compared to other cancers. In women under 40 breast cancer is the leading cause of cancer deaths (Ries 2007). In younger women the disease tends to be more aggressive and diagnosed in its later stages. Cancer statistics that have been published by the N.C.I. (National Cancer Institute) state, in fact, that the 5-year relative survival rate is lowest in women below 40 who are diagnosed with breast cancer (82 percent) compared to women diagnosed at ages 40 and older (89 percent). Numbers thus separated by age, emphasize the difference in survival rates in women under 40 (Ries 2007 see Figure 2.1).



Source: Bleyer A, O'Leary M, Barr R, Ries LAG (eds): *Cancer Epidemiology in Older Adolescents and Young Adults 15 to 29 Years of Age, Including SEER Incidence and Survival: 1975-2000*. National Cancer Institute, NIH Pub. No. 06-5767. Bethesda, MD 2006. Available online at http://seer.cancer.gov/publications/aya/

Figure 2-1 Breast cancer survival rate in women by age

Tumour registry data for Surveillance, Epidemiology, and End Results was linked with data on Medicare claims by (Gilligan et al. 2007). To find a relationship between number of breast cancer operations performed in a hospital (hospital volume) and all-cause mortality Cox's proportional hazard survival analysis, logistic regression and linear and quadratic components was also used. The study found moderate reductions in both all-cause mortality and breast cancer–specific mortality for women who were treated in hospitals with annual volumes of over 40 operations carried out on Medicare breast cancer patients. Even though careful control for possible confounders such as patient characteristics and tumour prognostic characteristics were made this was still observed. For the duration of standard 5-year median follow-up time, the effect of hospital volume remained measurable.

Pritchard et al. (2006) examined the hypothesis that the gene HER2 for epidermal growth factor receptor type 2, and the overexpression of its production in breast-cancer cells, could be related to responsiveness to anthracycline chemotherapy. Kaplan–Meier estimates of the survival rates were made using the presence or absence of increase of HER2 (as referring to FISH and PCR results) or overexpression of HER2 (as referring to immunohistochemical analysis results) or using the followed treatment regime and comparing it with the aid of a log-rank test. Moreover, the Cox's proportional-hazards model with a single covariate produced the hazard ratios for relapse or death with associated 95% confidence intervals in order to contrast the groups. The Cox model was used with treatment, intensification status and their interaction as covariates, to assess the interrelation between treatment type and growth status. Using the Cox model, multivariable analyses were use and adjusted for a number of variables: age (below 50, over 50), number of positive lymph nodes (<3 - ≥4): estrogen-receptor level (≥10 -. <10 fmol/mg (femtomole /per milligram), type of surgery (total or partial mastectomy), tumour size according to the tumour node metastasis staging system (T1, T2, or T3). A 95% associated confidence intervals and a kappa statistics were used to measure the agreement betwwen the four assays of HER2. In conclusion, the authors claimed a clear association of HER2 growth (or its over expression) in breast cancer cells with clinical responsiveness to anthracycline-type chemotherapy. It was shown that there was a greater benefit from CEF than CMF as adjuvant chemotherapy.

Dunnwald et al. (2007) looked at hormone receptor ER/PR status (positive or negative) and the relative risk of mortality according to demographic or clinical variations. They examined data from 11 population-based cancer registries taking part in the Surveillance, Epidemiology, and End Results program and included in their study 155,175 women from the years 1990 to 2001, who were over 30 years old and had a primary diagnosis of invasive breast carcinoma. The goal of their study was to determine relations between joint hormone receptor status and breast cancer mortality risk and the Cox proportional hazards model was implemented to compare results within categories divided by diagnosis year, diagnosis age, ethnicity, histologic tumour type, stage at time of study, size and grade, and axillary lymph node status. Results showed that in comparison to women with ER+/PR+ tumours, women with ER+/PR-, ER-/PR+, or ER-/PR- tumours experienced higher risks of mortality, irrespective to a great extent of the various demographic and clinical tumour characteristics assessed. The higher relative mortality risks noted among joint ER/PR negative patients with low-grade or small tumours raises the question of whether there may be an effective role for adjuvant chemotherapy in this group of patients.

Miecznikowski et al. (2010) state that one in eight (12%) women in the United States will develop breast cancer in their lifetime and even though advancements in treatment options, with regards to both surgery and chemotherapy, breast cancer is still the cancer with the second highest number of deaths in women. They investigated five data sets concerning breast cancer by means of gene set analysis with cancers being categorized into subsets according to a scoring system that was based on their genetic pathway activity. Their comparative survival study used the Cox proportional hazards regression model to discover significant variables correlated with risk with reported *P* values based on the sample estimate obtained from the Wald statistical test. Multivariate survival and univariate analysis were performed to select the clinical variables and/or their interactions significant for each separate dataset.

A study from Jerusalem, Israel (al-Quds University) by Ora et al. (2004) looked at the incidence of cancer among women with and without a history of pre-eclampsia. It was determined that cancer developed in 91 women who had suffered pre-eclampsia and in 2204 who had not (hazard ratio 1.27, 95% confidence interval 1.03 to 1.57). Risk of site-specific cancers was greater than before, particularly in cases of the stomach, ovary Epithelium, breast, and lung or larynx. These particular incidences of cancer increase in women with one child at study entry who had a pre-eclampsia history. Increased overall risk of cancer and incidence at several sites is correlated with a history of pre-eclampsia, which may be explained by environmental and genetic factors.

To test further the differences in survival among breast cancer subgroups, univariate Cox regression was performed to estimate the hazard ratios for basal-like breast cancer instead of luminal A, and for HER2+/ER− breast cancer instead of luminal A. Powerful calculations were performed using a computer program developed by Dupont and Plummer. These computations concluded that the estimation was at least 70%-80% or excellent (>80%) for the majority of comparisons. Statistical analysis was performed with the here of Reliability Centered Maintenance. This population-based study concluded that there is a higher frequency of Basal-like breast tumours among pre-menopausal African American patients than in postmenopausal African American and non-African American patients. The poor prognosis of young African American women with breast cancer could be due to a higher frequency of basal-like breast tumors and a lower prevalence of luminal a tumours (Carey et al., 2006).

In another study, (Gennari et al., 2008), compared different treatments: two dose levels of epirubicin plus cyclophosphamide vs cyclophosphamide, methotrexate, and 5-fluorouracil (CMF). Inverse variance weighting was used to pool the hazard ratios for the two epirubicin-based arms versus CMF. Log hazard ratios for overall survival and disease-free status were pooled across the studies, and both in relation to respective HER2 status overall by inverse variance weighting. In each study, formal tests for treatment by HER2 status interaction were implemented and test results were compared with the results of interaction tests reported by the individual studies. Publication bias was assessed by way of visual evaluation of direct plots mapped for study size versus treatment effect and also with the Egger regression asymmetry test. Sensitivity analyses were also done to determine if the strength of interaction between HER2 status and efficacy of adjuvant anthracycline treatment was in any way related to the method of HER2 assessment used, the type of anthracycline-based regimen implemented, or proportion of patients assessed for HER2 status. According to their results, in early stages of breast cancer, HER2 status is a predictor of responsiveness to adjuvant anthracycline therapy. The lack of observable effect of anthracyclines in HER2$^-$ negative disease implies that such patients might be spared the unnecessary toxic effects of this class of agents.

El Fatemi et al. (2013) carried out a case study of a patient with a rapidly growing nodule in the right breast. They confirmed that factors affecting survival are early stage diagnosis, conservative surgery, radiotherapy and combined modality treatment. Analysis showed that the node status is the best single predictor of survival. Improved molecular techniques together with classic histological diagnosis is necessary due to the breast lymphomas being rare and the problems associated with diagnosing it. In another study by Biggar et al. (2013) on Danish women with breast cancer, it was assumed that the use of digoxin might affect tumour characteristics and cause an increased relapse risk. It was concluded that breast cancers arising in digoxin-using women presented better prognostic features, overall breast cancer relapse risk in digoxin users showed no significant increase after adjustment for markers, however, in the first year subsequent to diagnosis recurrence hazards for ER positive tumours were higher.

Simsek (2000) examined major results for three main breast cancer treatment types in North Carolina. It showed that those treated with both Breast Conserving Surgery (BCS) and radiations have survival rates similar to patients who had done a mastectomy. The results imply that survival rates are similar for all three standard treatments and that BCS and radiation (for stages I and IIA) can safely substitute for mastectomy.

Psychological response to breast cancer was studied by Watson et al. (1999) in 578 women. Hazard ratios for psychological response were obtained using Cox regression. A high helplessness/ hopelessness score was found to have a moderately detrimental impact for 5 year event-free survival. Thus, a high score for depression was linked to a more significantly reduced chance of survival. In contrast, a high "fighting spirit" score in those women had no significant effect on prognosis. The authors do point out, however, that these conclusions are based on a small number of patients and should, for that reason, be interpreted with caution.

Renard et al. (2010) noted the incidence rate of breast cancers in Belgian women was as high as 152.7 per 100,000 in 2003 and estimated the effects of HRT (hormone replacement therapy) on incidences of breast cancer between 1999 and 2005 in women in Flanders between 50 and 69 years of age. The proportion of women aged 50-69 years and using HRT in Flanders was seen to have increased since 1992, peaking in 2001 at 20%, and from then on decreasing to 8% in 2008. In parallel, the diagnoses of breast cancer per 100,000 women aged 50-69 years in the same region rose from 332.8 in 1999 to 407.9 in 2003, then in 2005 numbers fell to 366.0. The number of HRT attributed breast cancers peaked at 11% in 2001, decreasing thereafter. Since participation in mammography screening by 50-69 year old Flemish women was still on the rise in 2003 and never surpassed 62%, they attribute the noticeable decrease in breast cancer incidence to the decrease in HRT use rather than to the level of screening carried out (Renard et al. 2010).

Holleczek and Brenner (2012) on the other hand, looked at the most recent 5-year relative survival rates in women with breast cancer and compared them with preceding trends in the U.S.A. and Germany. Life tables were calculated for intervals of 5 years from 1990 using period analysis to derive the 5-year relative survival and previous survival trends according to age and stage. Poisson regression models were also used for relative survival that fitted modelling of the logarithm of excess deaths with a linear predictor of follow-up year, age group and calendar period. Age standardized relative survival has progressively become better both in Germany and the U.S.A since 1993 to 83% and 88% respectively. Relative survival of localized cancer was over 97% in both countries, and 79% or 83%, respectively, for more advanced stages of breast cancer between the years 2005–08 (Holleczek and Brenner 2012). Metastasized disease prognosis is reported to have remained poor generally, with real improvement only to be found in younger patients. Patients being diagnosed with localized breast cancer was proportionally consistently higher in the U.S.A. When adjusted for stage, the differences in relative survival rates between the countries diminished over time and now effectively cease to exist.

2.5 Breast cancer research in developing countries

It has recently been estimated that worldwide there are over 25 million people living with cancer within five years of diagnosis (International Union against Cancer, 2010). In developed countries, breast cancer is an established health priority whereas in middle-income countries there is still insufficient attention paid. In developing countries recent evidence has shown breast cancer to be a leading cause of death and disability among women (Mathers et al., 2006). Although research has produced many new treatment options, most are prohibitively expensive. Therefore, modern breast cancer should be considered a significant challenge for health system funding.

Porter reported in 2008 that the risks of both breast cancer and death due to it are without a doubt rising worldwide. Most rises occur in low- and middle-income countries such that they account for 45% of more than a million new cases of breast cancer diagnosed each year, and more than 55% of breast-cancer–related deaths (Porter, 2008).

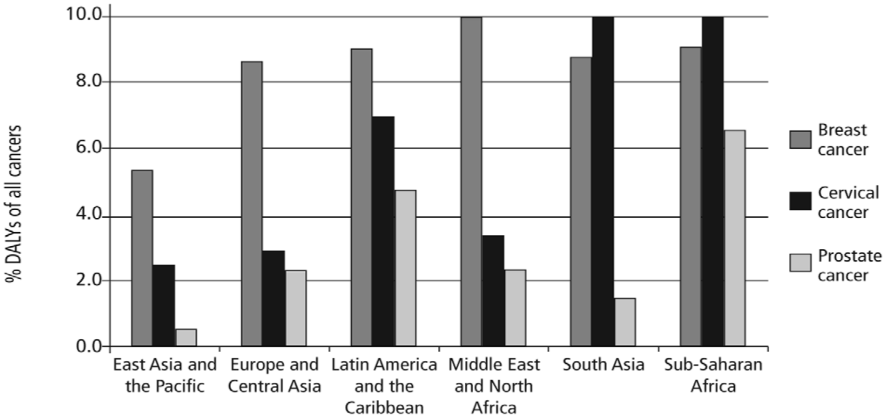


Figure 2-2 Percentage of DALYs lost from breast, cervical and prostate cancer as a proportion of all cancers, by world bank region, Source: Brown, M. L., Goldie, S. J., Draisma, G., Harford, J. and Lipscomb, J. (2006). Chapter 29. Health service interventions for cancer control in developing countries.

Recent analysis of mortality and morbidity trends by Brown et al. (2006) show the full extent of the disease in developing countries. Breast cancer surpasses cervical and prostate cancer in all regions of the developing world apart from sub-Saharan Africa and South Asia when using DALYs (Disability Adjusted Life Years) lost to cancer. It is the top cause of lost DALYs due to cancer at 9% in Latin America and the Caribbean. The differences are even more marked in other regions. In the Middle East and North Africa, as well as Europe and Central Asia, three to four times more DALYs are attributed to breast cancer than to cervical or prostate cancers, and in East Asia and the Pacific twice as many (Figure 2.2).

According to the National Cancer Center (2009) the five-year relative survival rate of all cancers was reported at 52.2% in 2005 for all cancers and is likely to continuously increase in the near future. Health professionals have begun to focus on other long-term health issues as well because the number of cancer survivors continues to rise. The risk of contracting a secondary cancer or other diseases such as heart disease and diabetes, is higher in cancer survivors, than the general population. They also more frequently present physical and psychological symptoms than healthy people (Bower et al., 2006; Helgeson and Tomich, 2005; National Cancer Center, 2009). This is illustrated by a Korean study that reported breast cancer survivors did worse than controls in terms of their performance in physical, emotional and social functioning (National Cancer Center, 2009).

About 1 out of 100 people in Korea can be found living with cancer within five years of diagnosis (National Cancer Information Center, 2009). This is due to improving survival rates of cancer sufferers as well as increasing incidence rates. Estimations are that new cases will increase from 152,600 in 2007 to 235,100 in 2015 which is 50% in eight years (National Cancer Center, 2009). Cases in Vietnam have also been increasing steadily over the last decade from, roughly 13.8 per 100,000 women in 2000 to 28.1% in 2010 (Lan et al., 2013). One way to encourage healthy lifestyles is to understand the behaviour influencing variables and therefore create effective involvement allowing survivors to be proactive and pursue this health-benefiting lifestyle. This was examined by Yi and Kim (2013) who studied the relationships among the internal health locus of control, depression, social support, and health-promoting behaviours in Korean breast cancer survivors with a view to identifying the factors that influence health-promoting behaviour. They used a predictive design and data was collected by means of questionnaires from a sample of 258 breast cancer survivors in Korea from a single year (2007). Apart from those receiving chemotherapy, significant differences in health-promoting lifestyle based on demographic and illness-related characteristics were not found. The internal health locus of control, depression, and social support correlated significantly with a health-promoting lifestyle on the other hand. Using stepwise multiple regression analysis, it was determined that social support, depression, and chemotherapy accounted for almost 40% of the variance in health-promoting lifestyle. The level of social support was found to most affect a health-promoting lifestyle, followed by that of chemotherapy and depression. The study results conclusively demonstrate the value of social support and the importance of depression in explaining the incidence of health-promoting lifestyle among Korean breast cancer survivors (Yi and Kim, 2013).

In Saudi Arabia, Rudat et al., (2013) considered the fact that obesity is increasing in a number of low-income and middle-income countries. Indeed to date more than 1.3 billion people globally are believed to be overweight or obese. It is commonly recognized to be a risk factor linked to cardiovascular disease, metabolic syndrome, certain types of diabetes and cancers including breast cancer. It has been estimated that the risk of developing postmenopausal breast cancer rises by as much as 30% - 50% in overweight or obese women. Furthermore, it has been directly linked with a poorer prognosis, regardless of menopausal status of the patient. A recent systematic review of both breast cancer-specific death and death from all causes indicated an increased risk of 33% in obesity as compared to lean women. This is especially pertinent in Saudi Arabia where breast cancer is the most common cancer, accounting for over 25% of all newly diagnosed cancers in women and for almost 14% in both genders combined (National Cancer Institute, 2012 and H. Al-Eid, 2012).

Huo et al. (2009) in Nigeria investigated the fact that black women experience a disproportionately heavy burden of aggressive breast cancer as compared to white women, since reasons for this phenomenon were to-date either unknown or understudied. This was the first study to determine the distribution of molecular subtypes of invasive breast tumours in indigenous black women in West Africa. The overall conclusion is that there is an urgent need for research into the ethiology and treatment of the aggressive molecular subtypes disproportionately affecting young African women worldwide in order to close the gap in the difference across populations.

The study by Rosmawati (2010) was the first to note low survival rates were related to later stages in Malaysian women of all ages and origin even though breast self-examination (BSE) can reliably be used for early detection. In addition, a questionnaire was formulated for systematic random sampling to be applied in a cross-sectional study and information was compiled through a guided interview. The results were assessed to determine the knowledge, attitude to and practice of BSE among women at least 15 years old and above. The average age of the 86 respondents was 40.5 years (SD: 15.51) of which a majority (80%) are educated to secondary or tertiary educational level. The total scores tabulated were as follows: 16.9 (total mean percent: 60.4%) for knowledge: 37.1 (77.3%) for attitude and 9.56 (34.1%) for practice. The respondents scored 38.4%, 73.3% and 7.0% for respondents for knowledge, attitude and practice respectively. The population studied was seen to have poor knowledge of the disease and there was a wide gap between attitude and practice. The factors related to poor practices were being unaware of the correct BSE method, not having knowledge of signs of cancer and not having support from parents, husband or friends. Thus measures that would have a significant positive impact on BSE among young Malaysian women are improved breast cancer awareness programs and health care workers recognising religious and social hurdles that include spouse, family and community.

There will be over 50.000 breast cancer cases by 2012 in Turkey based on recent documentations of breast cancer incidence rates (Özmen, 2008). They point out that while breast cancer is becoming more prevalent, it is characterized by a slow growth rate and early diagnosis can achieve positive treatment outcomes. Early diagnosis and early treatment of breast cancer is the most valuable in increasing life expectancy, reducing mortality rate, improving quality of life, and reducing physical pain and psychosocial problems in women (İğci and Asoğlu, 2003 and Özkan et al., 2010) describe basic attitudes to health issues and the BSE practices of Turkish female nursing and midwifery students and evaluate the benefit of educating their mothers, sisters, and other female relatives in BSE. They employed descriptive statistics and determined that better knowledge about breast cancer and BSE continual training programs should be planned for nursing and midwifery students. With improved access to facts, belief and attitudes would be enhanced and medical motivation with BSE should also increase accordingly.

Mandana et al., (2002) conducted a case-control study in Tehran, Iran between April 1997 and April 1998, in which 249 control women and 286 suffering from breast cancer were interviewed. With a multivariate and logistic regression method of analysis to derive Ors (odd ratio) and 95% CIs (confidence intervals), they examined the relationship between reproductive status and other risk factors in breast cancer occurrence in Iranian women. The results showed that family history and marital status could be associated with the incidence of breast cancer in Iranian women. It was unexpected that there was a lack of significant correlation between breast cancer and the other variables studied and the authors acknowledge that this may be explained by the limited capability of the study to estimate these risk variables.

Another study relating to breast cancer in Iran was done by Montazeri et al. (2003) who examined the extent of delay in patients seeking medical advice among Iranian breast cancer patients. A group of 190 diagnosed breast cancer patients were interviewed and subsequently completed questionnaires either at a university hospital or a breast clinic. Information regarding the time-lapse from first recognition of symptoms to first medical consultation was collected, which was used to calculate degree of patient delay. Their study findings confirm that patient delay is an important health problem, and must be reduced by educating women who are at higher risk. Examining the extent of patient delay and associated factors is only the first step. The next step is to establish interventions to reduce delays in seeking help and improve outcomes in all breast cancer cases.

Rezaianzadeh et al. (2009) set out to determine risk factors and breast cancer survival as related to socio-demographic and pathologic factors in Southern Iran, for which they noted there had been no previous study conducted. The main purpose of their study was to examine the effect of a wide range of variables on breast cancer survival, and for that reason the only outcome considered was that of survival. All 44 variables recorded at the cancer registry (from inauguration 01/01/2000 up until 31/12/2005) were used. These explanatory variables divide naturally into three groups: socioeconomic or demographic, clinical/pathological factors, and distant metastases. The association between survival, socio-demographic and pathological factors, and distant metastasis at diagnosis was investigated, and Cox regression was used to assess treatment options. The results demonstrate that the survival rate was relatively poor and this can be attributed to late stage diagnosis of the disease (patient delay). They assume that this was because of cultural inhibitions, low level of awareness, lack of access to screening programs and subsequent late access to treatment.

Harirchi et al. (2011) stated that breast cancer is the most regularly occurring cancer (estimated as 23% of all cancers) and fatal form of tumour among women that accounts for 16% of deaths. According to the Iranian Centre for the Prevention and Control of Disease (Ministry of Health and Medical Education, 2000, Iran), it is the most prevalent cancer among Iranian women that accounts for 21.4% of all tumours. In Europe and the USA it is estimated that the incidence rate is 8-10%. The lowest rates are found in some Asian countries (roughly 1%). In Iran the incidence rate was 6.7/100,000 in 2002, much lower than other countries (Rezaianzadeh et al., 2009). Later, it ranked as the number one tumour in Iranian women, accounting for 24.4% of all tumours and an age standardized incidence rate (ASR) of 23.65 in the year 2006. However, due to the lack of studies describing the clinicopathologic features, stages, and age distributions of breast cancer, the prediction of present and future patterns is difficult as is carrying out appropriate defensive and therapeutic actions to decrease its effect (Harirchi et al., 2011). It appears that most studies agree that the survival rate in Iranian breast cancer patients is lower than Europe and the USA (Vahdaninia et al. 2004).

By comparing data from other cities and countries, Ziaei et al.(2013) investigated the survival rates in Tabriz (Northwest Iran). The sample consisted of 271 breast cancer patients who visited a university clinic between 1997 and 2008. The survival rates for one, three, five, seven and ten years were taken. The sample had a lower survival rate compared to western countries, in particular, the survival rate of around 60% is significantly lower than those in European countries and the United States (e.g. 64% in Oman, 65% in Greece, 71% in Germany, 78% in Belgium, 84% in the United Kingdom and 89% in the United States (Mousavi et al., 2011). However, a larger sample size is required to conduct a better survival analysis, particularly for those under 40 years of age.

In twin studies in Tunisia, Hsairi et al. (2002) estimated the national incidence of main cancerous sites for the period 1993-1997. They first determined the relationship between cancer incidence and life expectancy at birth, Evo, in certain countries, and then calculated the level of regional incidence with data from local registers and compared them by using similarity of average (Evo) level as a basis. Their results indicated the significance of tobacco control, screening for breast cancer and cervix-uterine cancer, as components fundamental in reducing the number of cancers.

Abdelkrim et al. (2010) studied ER, PR Her-2 (estrogen receptor, progesterone receptor and the human epidermal growth factor receptor-2 respectively) according to which breast cancer can be categorised into four molecular subtypes (luminal A, luminal B, Her-2, and basal-like) and the possible correlations between these subtypes and clinico-pathological features. Univariate and multivariate analyses were used to analyse the data from their pathology department and used in the study. Their analysis showed that the Her-2 and basal-like subcategories could be associated with factors such as tumour size associated with a poor prognosis. Because the luminal A subtype was the commonest found in Tunisian women (just over 50% of all patients), they suggested that this showed that breast cancer in this country had no aggressive phenotype.

Maher et al. (2006) set out to evaluate the prevalence of ER, PR and HER2/neu among Jordanian women who have breast cancer of ductal and lobular types by retrospectively analyzing data of 267 cases from June 2003 to June 2004 at the King Hussein Cancer Center Hospital. To evaluate the two hormone receptors and HER2/neu, they used the standard Immunohistochemical test (IHC) and further evaluation of HER2/neu was carried out by the Fluorescence In Situ Hybridization (FISH) test in certain cases. Despite the limited extent of the sample group, results revealed that when compared to white American females, Jordanian breast cancer patients have lower hormone receptor positivity rates, which is more similar to results seen in black Americans and Chinese women in the States. This study may help provide further insight into breast cancer ethiology among different ethnic populations.

Breast cancer is the most widespread tumour diagnosed in Saudi Arabia. It amounts to 26% of all newly diagnosed women. With an age-standardized rate of 21.6/100,000 of the female population, the incidence is considerably lower than in the United States (124/100,000) (National Cancer Institute, 2012). About a quarter are younger than 40 years of age while the mean age at diagnosis is 47 (Al-Eid, 2012). The higher percentage of younger victims may partially be due to the demographics, since 50% of the females are younger than 20 years of age. By contrast, in the United States only about 7% of breast cancer patients are below 40, and the median age at diagnosis is 61 years of age (Rudat et al., 2013).

After diagnosis of cancer, weight gain has been frequently reported. Ethnic differences in the pattern of weight gain after the onset of breast cancer treatment are possible because of the differences in the epidemiology of breast cancer and obesity when comparing Asian and Western regions. There is no available data for Malaysian women on weight changes before and after adjuvant (of therapy) treatment with breast cancer (Yaw et al. 2010).

2.6 Methods used by researches in breast cancer

The Bayesian approach is used in the analysis of survival breast cancer data. Kalbfleisch (1978) modelled the cumulative base-line hazard by a gamma process within a proportional hazards setting. Clayton (1991) and Gelfand and Mallick (1995) considered extensions to multivariate survival data. Other examples include Dey et al. (1998) and Ibrahim et al. (2001). Many, although not all, such models applying Bayesian methods, concentrate on proportional hazard models. Gamerman (1991) for example, developed a Bayesian survival model with time-dependent effects, making use of a correlated prior process that Leonard (1978) had first introduced, where the true underlying course is estimated by a piecewise constant process. Evolution from interval to interval was plotted in a non-parametrical way, although some other conjugate assumptions had to be imposed so as to obtain an estimation using the linear Bayesian methods. Similar models were also produced by others, for example, (Arjas and Gasbarra, 1994, Arjas and Heikkinen, 1997, Sargent, 1997 and Sinha et al., 1999). These developments principally concentrated on parametric evolution, using Monte Carlo Markov chain (MCMC) methods for estimation. Arjas and Gasbarra (1994) also developed a Bayesian alternative to the frequentist non-parametric approach, by way of modelling the hazard by a piecewise constant process, relating parameters between intervals (using the gamma distribution) and treating the amount of smoothing as a fixed value. In addition, the times at which the values of the dynamic parameter changes were modelled as a random process.

Mostert et al. (1998) also investigated Bayesian analysis of survival data using the Linex Loss and Rayleigh model. Ahmed et al. (2005) investigated robust weighted likelihood estimation in the specific case of the Weibull distribution. Among others, Wang and Li (2005) used estimators for survival function with known censoring times. Saleem and Aslam (2008b) again worked on type I right-censored data with fixed censoring time. On the other hand, Abu-Taleb et al. (2007) examined exponentially distributed survival times with an exponentially distributed censoring time. Some interesting expressions were presented on computational aspects and Ali et al. (2005) in turn, presented Bayes estimation of exponential parameters. Subsequently, Saleem and Aslam (2008a, 2008b) produced a study using ordinary type I right-censored data for Bayesian analysis of a Rayleigh mixture. Rayleigh survival times that assumed random censoring time was considered by Saleem and Aslam (2009). Whereas Saleem et al. (2010) made use of ordinary type I right-censored data for Bayesian analysis of power function mixtures. This paper extended the work carried out by Abu-Taleb et al. (2007) adding algebraic expressions, numerical results and a simulation study used to compare and investigate the properties of the estimators. A posterior predictive distribution was derived and the equations required for the construction of predictive intervals were presented. The HPD (Highest Posterior Density) interval was formulated analytically and numerically in Saleem and Raza (2011). Maximum likelihood estimators and Bayes estimators assuming Squared Root Inverted Gamm (SRIG) were used by Saleem and Aslam (2009).

In Miecznikowski et al. (2010), model fitting for each gene expression profile was calculated using each gene individually, and then combined with ER status and tumour size, then with best model resulting from minimizing the AIC (Akaike Information Criterion), and finally minimizing the BIC (Bayesian Information Criterion). Statistical significance was ascertained for individual genes by means of controlling the FDR (false discovery rate) for testing multiple genes at 0.2 using the Benjamini and Hochberg scheme for the p-values obtained from log-rank tests in each gene model.

Hemming and Shaw (2002) made use of Bayesian methodology and Monte Carlo Markov chain (MCMC) estimation methods. For some of the predictive indicators it was indicated that the estimated effects evolve with increasing follow-up time. In general, those predictive indicators which were regarded as representative of the most hazardous groups had a declining effect.

Cox (1972, 1975) proposed a partial likelihood function for the estimation of parameters as well as for the discrete time logistic model. If m failure times are tied at time t and n individuals are at risk just prior to t then the partial likelihood contribution involves a summation over all possible subsets of size m from the n at risk. With large data sets calculations like this are no longer feasible. The 'marginal' likelihood of Kalbfleisch and Prentice (1973) is simpler because the contribution at time t involves only the m! Permutations among the m individuals with failure time t. Categorical data regression models were considered for data analysis. Cox (1972) proposed a model for discrete survival time which is closely related to the Mantel-Haenszel approach (e.g. Mantel, 1966) to survival analysis.

Clark et al. (2003) published a series of four articles where the basic concepts of survival analysis were introduced and explained. The first article presented basic concepts, including how to develop and interpret survival curves and how to assess, quantify and test survival differences between two or more groups of patients with the aid of Cox's proportional hazard, Kaplan-Meier log-rank test and accelerated failure time. The other papers dealt with multivariate analysis and the last one introduced some more advanced concepts in the form of brief questions and answers.

An alternative approach to regression modelling involves stratification of nuisance variables. Inferences on primary factors then proceeded within strata, and summary inferences were made by combining test statistics and estimators across strata (see e.g. Mantel 1966; Hankey and Myers 1971; Godwin and Brown 1975). What became apparent were the constraints of such a procedure in terms of sample sizes. The stratification approach is conceptually simple and provided a solution to the possibility of high order interaction effects of nuisance variables on survival time.

Tabatabai et al. (2012) studied the survival of breast cancer patients in the Netherlands through the exploration of the role of a metastasis variable in conjunction with both clinical and gene expression variables. The Netherlands Cancer Institute provided data for a hyper tabastic model, which uses a two parameter probability distribution, for an in-depth analysis of 295 breast cancer patients. In comparison to Cox regression model assumptions, the increased accuracy was complemented by their subsequent ability to analyze the time course of the disease by means of progression hazard and survival curves, which were described in detail. Deciles were also computed for survival and probability of survival to a given time. Their primary aim was the introduction of a variable representing the existence of metastasis and its effects on other gene expression and clinical variables.

In the Czech Republic, Horov´a et al. (2007) built a model to study the survival rate of cancer patients. A parametric form of hazard function was used based on a model of cancer cell population dynamics (Kozusko and Bajzer, 2003) that is dependent on several parameters. A detailed method is outlined to estimate such parameters. For survival data the nonparametric methods seem more suitable, which include methods of kernel estimation of hazard functions, though there was a serious difficulty on deciding a smoothing parameter. An alternative method for the bandwidth selection was proposed.

Parmar et al. (1998) summarized studies addressing similar matter using Meta-analyses. They extracted estimates of these statistics using various methods which included the log hazard ratio. This improved the accuracy and reliability of their meta-analyses. For the sample size of 2152 black women and 25968 white women in the USA, Prentice and Gloeckler (1978) suggested the grouped data version of the proportional hazards model function. A generalization of the log-rank test for the comparison of survival curves as given for testing the hypothesis of a zero regression coefficient. Its application to breast cancer data from the National Cancer Institute indicated that race differences in breast cancer survival times is because of differences in the phase of disease and demographic features when diagnosed. This method is free from the length of the survival time grouping.

Using the West Midlands Cancer Intelligence Unit, Hemming and Shaw (2002) studied time-dependent effects of prognostic indicators on breast cancer survival times. While noting that the greater part of analysis of the cancer registry data uses the semi-parametric proportional hazards Cox model, estimation methods used were Monte Carlo Markov chain and Bayesian methodology. This model is similar to that developed by Sinha et al. (1991) and Gamerman (1999) amongst others. It was shown that the estimated effects change with more follow-up time (Hemming and Shaw, 2002).

In Melbourne, Australia, (Baglietto et al., 2005) evaluated the effect of vitamin B9, the dietary folate, on the relation between alcohol consumption and breast cancer risk. Hazard ratios were estimated using Cox regression with age being used as the time metric. Polynomial relations were compared with log hazard rate using fraction polynomials for alcohol consumption and folate intake. Interaction between alcohol consumption and dietary folate intake was compared in non-nested models with the Akaike information criterion. It showed that sufficient dietary intake of folate may defend against increased risk of breast cancer that is related to alcohol consumption.

The applications of parametric survival models were extended by Abadi et al. (2012) to include cases where the AFT (accelerated failure time) assumption is not satisfied, and looked at parametric and semi-parametric models according to different proportional hazards (PH) and AFT assumptions. In the specific 1990–1999 study that they used, 12,531 women diagnosed with breast cancer in British Columbia, Canada, were categorized into eight groups according to age and disease stage. It was assumed that each group had different AFT and PH criterias. The saturated generalized gamma (GG) distribution was fitted for parametric models, and then compared with the straight AFT model. A likelihood ratio statistic was applied to compare both models to the simpler forms which included Weibull and lognormal. Cox's PH model or the stratified Cox model was fitted for semi-parametric models according to the PH assumption and Schoenfeld residuals were used to test them. The GG family was compared to the log-logistic model by means of the Akaike information criterion and the Baysian information criterion. When PH and AFT assumptions were satisfied, semi-parametric and parametric models both gave valid descriptions of breast cancer patient survival. In the case that the AFT condition held when the PH assumption failed, the parametric models were more reliable than the stratified Cox model. When neither AFT nor PH assumptions were met, the log normal distribution provided a reasonable fit. Both parametric and semi-parametric models were found appropriate when both PH and AFT assumptions were satisfied. When the PH assumption is not satisfied, then parametric models should be considered, whether the AFT assumption is met or not (Abadi et al., 2012).

Jeong (2006) pointed out that an approach that is based on the Kaplan-Meier estimator (Kaplan and Meier, 1958) may overestimate the proportion of occurrence of local or regional events (Korn and Dorey, 1992; Pepe and Mori, 1993; Gay-nor et al., 1993; Lin, 1997) whereas the cumulative incidence function (Kalbfleisch and Prentice, 1980) affords correct estimates for the cumulative probability of local or regional recurrences in the presence of other competing events without the assumption of interdependence of event times. Gray (1988) used a nonparametric inference procedure to compare the cumulative incidence estimates of different samples. However, a semi-parametric regression model on the cumulative incidence function was also examined by Fine and Gray (1999). To assess the full parameterization of the cumulative incidence, Benichou and Gail (1990) used piecewise exponential or simple exponential distributions.

In survival analysis the simple exponential distribution and existing Weibull distribution families may prove to be too restrictive. Also, a conclusive estimate may not be possible because the piecewise exponential distribution involves too many parameters. Due to this, a new distribution family to parameterize the cumulative incidence function was proposed by Jeong (2006) that included the three-parameter generalized Weibull distribution (Mudholkar et al., 1996; Jeong et al., 2003). To evaluate the increased efficiency for the parametric estimates as a function of time, a comparison was made between the mean-square errors of the parametric estimates of the cumulative incidence function to the same quantities of the nonparametric estimates. To compare the cumulative incidence functions at a particular time, a simple parametric two-sample test statistic was created that can be applied to data sets on breast cancer treatment from phase III National Surgical Adjuvant Breast and Bowel Project clinical trials.

Gray (1992) investigate applying fixed knot splines to model the progress of breast cancer tumours, and then the parameters were estimated using the penalized partial likelihood. These methods give a useful understanding of how prognosis diverges as a function of continuous covariates and shows how the covariate effect changes with respect to follow-up time. This may be achieved by using penalized likelihood. For example Verweij and van Houwelingen (1995) used a finite sequence for the dynamic effects. In contrast, Hastie and Tibshirani (1986) used cubic splines (piece-wise polynomial functions) with knots at unique failure times, and Gray (1992) used both quadratic and piecewise constant splines, but with fewer knots. (Grambsch and Therneau 1994) showed how plots of rescaled Schoenfield residuals against time can be used to both show the extent of non-proportionality and to test a null hypothesis of proportional hazards.

In Islamabad, Pakistan, Saleem and Aslam (2009) looked at lifetime data analysis and considered the suitability of Rayleigh distribution survival times to derive maximum likelihood in conjunction with Bayes estimates for unknown parameters. The Rayleigh distribution is particularly useful in the analysis of lifetimes of objects that age with time, which may be modeled by a monotonically increasing hazard function. Lifetimes with constant hazard rates are studied using the exponential distribution, given its memoryless property. In the same area there have been a number of breast cancer studies initiated by Raqab and Ahsanullah (2001) who focused on ordered generalized exponential distribution (GED).

Carey et al. (2006) in the U.S.A analyzed population-based distributions and clinical associations for breast cancer subtypes. Data was presented stratified by four different patient groups. Differences between breast cancer subtypes were investigated regarding clinico-pathologic characteristics using one-way analysis of variance (ANOVA) for age, and χ^2 tests for the remaining variables. The Fisher test was used when expected cell counts were less than 5.

From Switzerland, Spitale et al. (2008) did a large European population-based study which investigated prevalence, clinicopathologic features and overall survival (OS) of molecular subtypes of cancers. These molecular subtypes were defined by IHC (immunohistochemical) markers and were evaluated using (ANOVA one-way analysis of variance) for patient age and size (diameter) of tumour. The Bonferroni method was used to calculate pair-wise differences in the molecular subtypes so as to control the overall significance level by adjusting the p value. The χ^2 test assessed the relationship between the different subtypes and main clinicopathologic characteristics believed to be of prognostic importance, for example histologic grade (well\moderate vs poorly differentiated), menopausal state according to age, laterality (right vs left), multifocal or not, metastasis status at time of diagnosis etc. The Fisher exact test was used for expected cell counts under five using the Monte Carlo method. For verification of correct IHC subtype definition, the distribution of cases according to HER2 expression, PR status and ER status were made available.

Strati et al. (2013), compare different analytical methodologies for circulating tumour cell (CTC) discovery and molecular characterization and conclude that standardization of investigative procedures is urgently needed and essential before such methodologies are implemented in clinical practice. For statistical analysis, the chi-square test was used to evaluate concordance in early breast cancer, while in a smaller group presenting verified metastasis, the Fisher exact test was preferred. The Kappa test was utilized in all cases in order to calculate the degree of agreement between the three different CTC molecular methods. Resulting data indicated the importance of the heterogeneity of circulating tumour cells (CTC) to detect different analytical procedures.

Sinha et al. (1999) looked at the class of models in which both the log-base-line hazard and the covariate effects are modelled by piecewise constant functions, with parametric distributions controlling evolution. Such models do not lose much in the way of flexibility compared with piecewise linear or other spline models, but do gain much in terms of ease of communication with health professionals and clinicians, and further avoid approximation methods, by providing the flexibility to estimate model parameters with the aid of Monte Carlo Markov chain (MCMC) methods. Furthermore, in contrast to all the aforementioned approaches, a simple reparameterization must be introduced to improve the convergence of the Gibbs sampler, thus giving a set of parameters that is to a much lesser extent highly correlated.

Arjas and Heikkinen (1997) subsequently generalized this work to include spatial intensities. Sargent (1997) in turn combined a partial likelihood approach with dynamic effects modelled by a piecewise constant process and included the smoothing parameter as a hyperparameter. He then noticed slow convergence within the Gibbs sampler (Gilks et al., 1996) due to high correlation of neighbouring parameters. A new parametric survival model was applied to cancer prevention studies by Sinha et al. (2002). The model was formulated along the lines of a stochastic modelling of the occurrence of tumours throughout two stages, namely the initiation of an undetected tumour and development of the tumour to a cancer that is detectable.

Siddiqui et al. (2001) studied the survival rate of patients with Metastatic breast cancer (MBC) from a particular care institution in Pakistan. Univariate (descriptive) statistics and median survival time were used. Important factors identified causing MBC to progress were emotional energy and financial resources, race and ethnic origin.

2.7 Breast cancer research in Iraq and Kurdistan region

Therapeutic response and prognosis of breast cancer can be predicted according to hormone receptor (HR) and HER2 expression. Breast cancer is diagnosed at a comparatively young age in the Middle-East and Arabic women show a low occurrence of HR positive tumours. In Iraq breast cancer is the most widespread cancer and is the most important cancer-related mortality among women (Majid et al., 2009: Al Tamimi et al., 2010). Arabic women in Saudi Arabia and Jordan have a high percentage of ER- and PR- breast cancers which arise mostly before menopause (Sughayer et al., 2006: Al Tamimi et al., 2010). In the Middle-East breast cancer is typically a relatively aggressive form with a negative diagnosis for individual patients (Sughayer et al., 2006; Al Tamimi et al., 2010). Even though the Kurds are ethnically different from the Arab population in Southern Iraq, the incidence rates for age specific breast cancer in Kurdish females was documented to be comparable to that of Egypt and Jordan (Majid et al., 2009; Rennert., 2006). Studies in the Middle-East on HR and HER2 are not general and their conclusions differ (Dey et al., 2010; al-Alwan et al., 2000).

Breast cancer subtypes are indicators of both the options of and possible reactions to treatment. Nichols (2010) refers to epidemiological studies in Iraq that considers the risk factors of the cancers, in particular the uniqueness and behaviours of cancer in patients present in different geographic at regions. Even though 90% of women who examine themselves identify a lump, only 32% seek medical advice within a month's time. As a result, almost half presented at an advanced stage of breast cancer. ER positive tumours accounted for 65% of the cases and PR positive tumours were noted in 45% of the total. The statistical analysis is not mentioned and demographic and clinico-pathological presentations have yet to appear. There is no data available for the incidence rate of breast cancer in the region of Suleimani, Southern Iraq prior to 2006.

Hussaion and Aziz (2009) were interested in collating statistical data on the incidence rate of breast cancer in Suleimaniyah in 2006 as well as collecting demographical data, and detecting some risk factors associated with breast cancer. Results showed that the incidence rate in females is 10.1 per 100000 adults for the specific year 2006. A hypothesis test was carried out using a Chi-Square test at 95% confidence level to investigate the difference between the study individuals and a control group; the study involved 61 Kurds (60 females and 1 male) . The group displaying highest occurrence was between the ages of 45 and 54. Most of the females were housewives, married, fertile, living in city or urban areas, with no family histories of breast carcinoma, nonsmokers, had breast fed, with a BMI index above normal, and mammography or U/S screening test for breast mass was done for only a small minority prior to diagnosis. Results revealed invasive ductal carcinoma of breast is the most frequent tumour in the cohort. Of the patients, 11 (18.0%) of them had been exposed to chemical weapons in 1988.

Majid et al. (2009) studied breast cancer incidence in the Kurdish region of northern Iraq, investigating age specific cancer rates in the province of Suleimaniyah. The risks associated both with reproductive history and family history of breast cancer were evaluated in an age-matched case-control study. The relationships between clinical stage and patient age were studied as well to investigate whether this is a way to evaluate tumour development in younger and older women. The objective was to compare incidence rates, severity, and risk linked to reproductive and family history with published research from other Middle-Eastern countries and United States. Groups were compared by t-tests if data met normality assumptions and equal variance tests and by Mann-Whitney rank-sum tests if not. Kolmogorov-Smirnov tests were employed to resolve whether data were normally distributed. Conditional logistic regression was used for age grouped patients and controls so as to analyze the observed risk of breast cancer with respect to marital status and number of children. The connection between patient age and tumour stage was tested through linear regression and differences between groups for categorical values by chi-square tests. Values of $P < 0.05$ were considered statistically significant in all tests.

In Kurdish Iraq breast cancer is mainly a disease of pre-menopausal women with several pregnancies according to Majid et al. (2009). Incidence rates for younger patients were comparable to Western statistics. However, they were higher than most Middle-Eastern countries which noticeably declined with age, unlike in the West. The genetic breast cancer risk for both older and younger women was within the general population risk seen in Western countries. Delays in diagnosis were unrelated to patient age and led to clinical stages being more advanced. Better preventative projects were recommended since screening programs for breast cancer in the Kurdish region of Iraq were established.

To identify possible risks of cancer, Othman et al. (2011) studied cancer incidence in this same region for which data was provided by cancer registries from 9 hospitals located in three Kurdistan cities. Information was examined to verify that it was not duplicated, place of residence was correct and to check for other possible errors. The overall total of registered cases in 2007, 2008 and 2009 were 1444, 2081, 2356 respectively, 49% of cases were males and 51% females. A direct adjustment method was used for computing age-standardized rate (ASR) and was found to be 89.83/100 000 among males and 83.93/100 000 among females. Among the three Kurdish Governorates, there were considerable differences in incidence rates of the different types of cancer according to the results. In addition to this, there was an indication of increased risk of cancer in these regions. Among male cases hematological malignancies (blood cancer: leukemia) were the most widespread (21.13% of all cancer in males) and the second most widespread in female (18.8% of all female cancers), only breast cancer was higher.

Majid et al. (2012) investigated the expression of hormones HR and HER2 among Kurdish and Arabic women. The Suleimaniyah Directorate of Health recorded 514 Suleimaniyah Kurdish women, 227 Kurdish women of other Governorates, and 83 Arabic women with a primary diagnosis of breast cancer between 2008 and 2010. Of these, the breast cancers of 432 women were tested using an immunohistochemistry test (IHC) for estrogen and progesterone receptors (ER and PR) and HER2 and age specific and age standardized incidence rates were calculated for Suleimaniyah Kurds. These results were compared with Egypt and with United States (US) for which SEER data had been used. For proportional distribution of patients among different groups analysis Chi square tests were used. Dunn’s variance on ranks analysis (this test helps analyse the specific sample pairs for stochastic dominance: Kruskal-Wallis one way anova) was used to compare the age of individuals in the three populations. Logistic regression was used to compare the relationships between HER2 status (dependent variable) and age, tumour grade, and ER status (independent variables). For all statistical procedures $P<0.05$ was considered significant. Lower age standardized and age specific breast cancer incidence rates were found in Kurdish women in comparison to US rates. However, the proportional HR and HER2 expression for both Kurds and Arabs was comparable to American Caucasian females. The vast majority of breast cancers are ER+/HER2- and responsive to anti-estrogen therapy. However the comparison is not to be considered entirely dependable due to the nature of society and other environmental factors.

Shabila et al. (2012) investigated primary care providers’ perspectives as to the foremost concerns in the provision of primary care services and potential opportunities to expand the present system. Discussions were held and scripts fully transcribed and translated and analysed qualitatively by content analysis, followed by a thematic analysis. To improve the system it was suggested to include the application of a family approach and ensure effective planning and monitoring. The qualitative study was based on participants separated into four focus groups involving 40 primary care providers from 12 primary health care centers in the Erbil Governorate in the Iraqi Kurdistan region between July and October 2010. To guide discussions, there was a list of topics that included questions on both positive aspects of existing problems with the current primary care system as well as prioritizing needs for improvement.

Throughout reviewing the literature on breast cancer in different countries, so far no specific studies have been carried out that explicitly consider survival analysis for breast cancer in the Kurdistan region of Iraq. This thesis addresses this important topic.

2.8 Summary

Breast cancer is the most common type of cancer in women in both developed and developing countries. The incidence of breast cancer is increasing in developing countries due to increased life expectancy, increased urbanization and wider adoption of western lifestyles. Although some risk reduction might be achieved with prevention, these strategies cannot eliminate the majority of breast cancers that develop especially in low and middle-income countries where breast cancer is diagnosed in very late stages. Early detection is therefore required to improve the outcome of breast cancer and survival remains the cornerstone of breast cancer control.

The World Health Organization promotes breast cancer control within the context of national cancer control programs integrated with non-communicable disease control and prevention. At present, WHO together with support from the Komen Foundation is conducting a 5-year breast cancer cost-effectiveness study in 10 low and middle-income countries. The project includes a program-costing tool to assess affordability. It is hoped that the results of this project will contribute to providing evidence for the formulation of adequate breast cancer policies in less developed countries (WHO, 2013).

3 CHAPTER 3: The survival analysis concept

3.1 Introduction

Survival analysis is primarily concerned with modelling and analysing time-to-event data, which are generally referred to as “failures.” Some examples are time until an electrical component fails, time to first recurrence of a tumour (i.e., length of remission) after initial treatment (Tableman and Kim 2004).

It is possible that a “failure” time will not be observed due to deliberate design or random censoring. In this study this would occur if a patient is still alive at the end of a clinical trial period or has moved away. The primary reason for developing specialized models and procedures for failure time data is brought on by the necessity of obtaining methods of analysis that accommodate censoring. Survival analysis can then be thought of as a collection of statistical procedures that accommodate time-to-event censored data. Previously, incomplete data were treated as missing data and omitted. This loss of information introduced bias in estimated quantities. The procedures discussed here avoid bias and are more powerful as they utilize the partial information available on a subject or item (Tableman and Kim, 2004). Survival analysis is the study of the occurrence and timing of events. Covariates are studied to determine their effect on survival duration. Censoring and time-dependent covariates (time-varying explanatory variables) are unique to survival analysis (Cox and Oakes, 1984).

Leung. et al. (1997) highlighted three common methods of survival analysis: the life-table method, the Kaplan Meier method, and the Cox proportional hazards method. Survival curves are used in the preliminary examination of data and visual inspection tells us whether there are obvious differences between the two groups. In general, survival analysis is used to follow-up patients under treatment by various experimental therapies, evaluate survival after diagnosis with specific diseases, summarize and evaluate mortality in different groups (Cleves et al., 2002).

3.2 The concepts of survival analysis

Miller (1980) considered a random variable T > 0, which can be thought as the lifetime or the survival time of a patient. If T has a density function *f(t)* and distribution function F(t) then the survival function of T (see e.g. Crowder, 2012), is defined as the following:

$$S(t) = 1 - F(t) = P\{T > t\}. \qquad (3.1)$$

The hazard rate or hazard function is

$$h(t) = \frac{f(t)}{1 - F(t)} \qquad (3.2)$$

i.e.

$$h(t)dt = P\{t < T < t + dt \,|\, T > t\} \qquad (3.3)$$

$=P$ {death in the interval (t, t+dt) given survival past time t}.

Integrating $h(t)$

$$\int_0^t h(u)du = \int_0^t \frac{f(u)}{1 - F(u)} du = -\log(1 - F(u))\,|_0^t$$

$$= -\log(1 - F(t)) = -\log S(t),$$

which leads to the important expression

$$S(t) = e^{-\int_0^t h(u)du}. \qquad (3.4)$$

Notice that F(+∞) = 1 (i.e., S(+∞) =0) iff $\int_0^\infty h(u)du = \infty$.

Continuity will be assumed but concepts and formulae can be modified to include jumps in the density function when it is important (Miller, 1980).

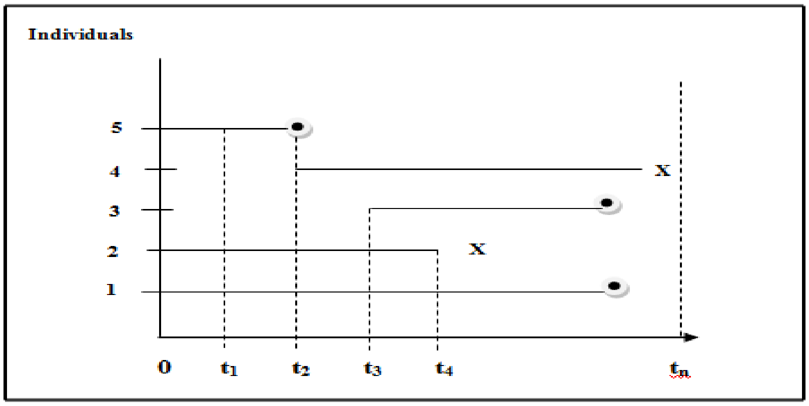


Figure 3-1 Illustration of the survival data where (.) is a censored observation and (X) is an event (death), Source: Originated by the researcher based on Hosmer and Lemeshow (1999)

The illustration of survival data in Figure 3.1 shows several features which are typically encountered in the analysis of survival data:

- Staggered entry: Individuals enter the study at different times, e.g. individual 3 enters the study at time t3 and the 5th individual was censored between time 0 and t_2 and died from t_2 onwards.

- Not all individuals have had the event (death) when the study ends.

- Other individuals drop out or get lost in the middle of the study.

The last two relate to "censoring" of the failure time events (Cox and Oakes, 1984).

3.3 The problem of censoring

According to Leung et al. (1997) censoring occurs when an individual is not followed up until occurrence of the event of interest. There is loss of information due to this incomplete observation because their different experience would lead to bias in the study. It is caused by failure to follow-up, withdrawal from the study, study termination when subjects had different dates of enrolment, or death due to a competing risk. They contribute to the analysis until the time of censoring. It is assumed by Leung et al., (1997) that censored subjects would have had the same rates of outcomes as those not censored at that time if they had been followed beyond the point in time at which they were censored. Existence of similar censoring patterns between different treatment groups suggests that the censoring assumptions hold (Leung et al., 1997). The situation is further complicated by effects we summarize as "hidden censoring": Due to inconsistency in follow-up checks, multiple transfers between hospitals and the general difficulty of contacting and keeping track of patients without comprehensive records, censoring may in some cases go unnoticed; when censoring is noticed we shall (see chapter 5) refer to it as “overt” censoring.

Kleinbaum and Klein (2005) consider three types of censored observations: right-censored, left-censored, and interval-censored. In order to analyze such data John and Melvin (1997) defined C_i as an element of the set $\{1,0\}$, where

1 means that the ith data point is not censored (death),
0 means that the ith data point is censored.

The likelihood function is a statistical methodology, which provides a way of estimating the unknown parameters of a probability distribution (or density) based on a given data sample by way of maximum likelihood estimation. The likelihood function of an estimator based on censored data looks like the usual likelihood function. However, the information given by censored data has to be added, (John and Melvin, 1997; Lawless, 2003). For right-censored data, we have:

$$L = \prod_{i=0}^{t-1} [f(t_i)]^{C_i} [S(t_i+)]^{1-C_i} \qquad (3.5)$$

where $S(t_i) = P_r(T > t) = 1 - F(t)$ is the survival function and $f(t_i)$ is the probability density function, t_i + is the time at the end of the interval and t_i - is the time at the start of the interval. The same can be done for left-censored data:

$$L = \prod_{i=0}^{t-1} [f(t_i)]^{C_i} [S(t_i-)]^{1-C_i} \, . \qquad (3.6)$$

For interval censored data, this becomes:

$$L = \prod_{i=0}^{t-1} [f(t_i)]^{C_i} [S(t_i-) - S(t_i+)]^{1-C_i} \, . \qquad (3.7)$$

An example of each one of these is shown in Figure 3.2 which illustrates three forms of censoring; Right-censored: suppose a subject is lost to follow-up after 10 years of observation and the time of event (death) is not observed because it happened after the 10th year (i.e., t > 10). A subject is Left-censored if the event (death) happens before the 10th year but the exact time is unknown, hence the subject is left-censored at 10 years (i.e., t < 10). Lastly, Interval-censoring represents the subject having the event (death) with exact time unknown but occurring between the 8th and 10th year. This subject is interval censored (i.e., 8 < t < 10).

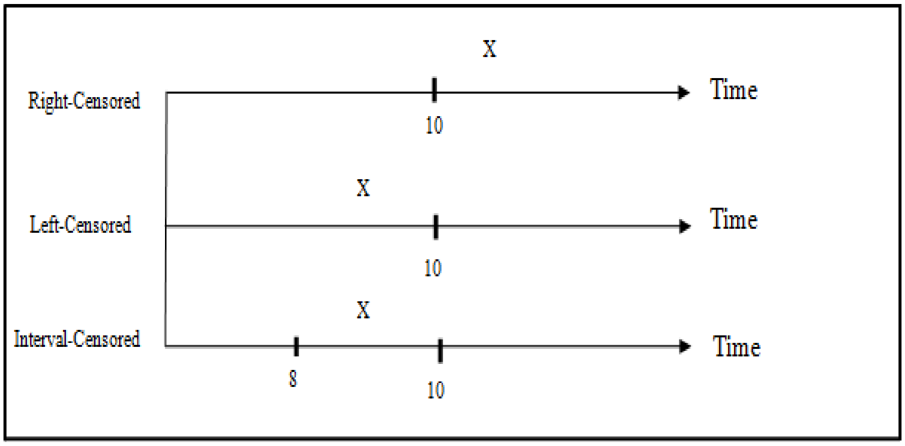


Figure 3-2 Forms of censoring data: the right, left and interval censoring

3.4 Survival analysis techniques

Techniques used for dealing with censored data can be broadly classified into non-parametric (Kaplan Meier, product limit method and Life tables), parametric (exponential, log-logistic, Gompertz and Weibull methods) and semi-parametric (Cox-proportional hazards method). The latter two can also be applied as regression-based models.

3.4.1 Non parametric survival analysis

Non-parametric analysis estimates probabilities associated with dependent variables without making assumptions about shape of the distribution (Cleves et. al., 2002).

3.4.1.1 Kaplan Meier (Product limit methods)

Let time be partitioned into a fixed sequence of intervals T_0, T_1, T_2, …, T_K. These intervals are almost always, but not necessarily, of equal lengths. The survival function of the Kaplan-Meier method is formed as follows:

$$h(t) = \frac{d_t}{n_t} \qquad (3.8)$$

$$s(t) = 1 - h(t) \qquad (3.9)$$

$$S(t+1) = \prod_{k=0}^{t-1} s(k) \qquad (3.10)$$

where:

c_t is the number of censored (withdrawing) observations at time point t,

d_t is the number of deaths at time point t,

$n_{t+1} = n_t - d_t - c_t$ (for censored data),

$n_{t+1} = n_t - d_t$ (for uncensored data, $c_t = 0$),

n_t is the number of individuals (entering) at risk,

d_t / n_t represents the probability of dying at time t+1 conditional to being at risk (alive) at time t.

The censored individuals are excluded from the denominator of ‘at risk’ individuals at the point when they are censored, however, they are included at each preceding point.

The Kaplan-Meier method computes the probability of dying at a certain point in time conditional on survival up to that point. Meier (1958) and Crowder (2012) utilize the information of censored individuals up to the point where the patient is censored in order to maximize the use of the information available from the study sample.

Tables 3.1 and 3.2 illustrate a theoretical dataset and computation of survival probability using the Kaplan Meier estimator, respectively. Note in Table 3.1 columns 6-8 represent a rewarding of the data in columns 1-5.

Table 3-1 Theoretical data to illustrate survival analysis

1	2	3	4	5=3-2	6	7	8
Patient number (nt)	**Time of operation (in week)**	**Time observation ended (in week)**	**Reason observation ended (death=1, censoring=0)**	**Time under observation**	**Ordered patient number (nt)**	**Time under observation**	**Reason observation ended (death=1, censoring=0)**
1	0	120	0	120	6	30	1
2	0	68	1	68	12	30	1
3	0	40	1	40	5	30	0
4	4	120	0	116	10	35	1
5	5	35	0	30	3	40	1
6	10	40	1	30	9	40	0
7	20	120	0	100	11	50	1
8	44	115	1	71	2	68	1
9	50	90	1	40	8	71	1
10	63	98	1	35	7	100	0
11	70	120	1	50	4	116	0
12	80	110	1	30	1	120	0

Table 3-2 Theoretical illustration of estimating probability of survival, *S(t+1)*, using the Kaplan Meier estimator

Time period (t)	Time weeks	Patient number (n_t)	Death (d_t)	Censored (c_t)	$1-\frac{d_t}{n_t}$	$S(t+1)$
0	0	12	0	0	1	1
1	30	12	2	1	0.83	0.83
2	35	9	1	0	0.89	0.74
3	40	8	1	1	0.88	0.65
4	50	6	1	0	0.83	0.54
5	68	5	1	0	0.80	0.43
6	71	4	1	3	0.75	0.32

A large number of individuals censored at a single point of time affect the shape of the survival curve leading to sudden spurious large jumps or large flat sections. Machin et al., (2006) stated other factors leading to such spurious jumps such as an extremely low number of individuals at risk especially toward the end of the study or pre-arranged clinic visit schedule. The reliability of the different portions of the survival curve is dependent on the number of individuals at risk at that stage as shown in Prinja et al. (2010). The majority of studies are likely to have some individuals for which the outcome event is not recorded. This can be due to limited resources to carry forward the study until outcomes are recorded for each and every study individual. A measure of the maturity of the data then shows the quality of data in terms of the frequency of individuals for which the outcome event (death) is recorded. A simple measure of this is the average (median) follow-up period and a more robust graphical technique involves constructing a survival curve by reversing censoring (Machin et al. 2006).

3.4.1.2 The life tables method

A life table subdivides the period of observation into shorter time intervals. All people who fall in that interval are used to calculate the probability of the event occurring in that interval. These probabilities are then used to estimate the overall probability of the event occurring at different time points. As per Lawless (2003), the classical method of estimating life table, $S_T(t)$, and the actuarial method used in epidemiology and actuarial science are discussed below. The life table estimates are calculated by counting the number of events and censored observations that occur in the time intervals (T_t, T_{t+1}) for *t=0,1,2,...,k-1*, where $t_0 = 0$, and n_t is the number of units entering the time intervals (T_t, T_{t+1}). We now needed to adjust for censoring.

Individuals lost during time interval (T_t, T_{t+1}) are assumed to be at risk for half the interval on average.

The effective sample size N_t of the number of individuals during the interval exposed to risk is then defined as

$$N_t = n_t - \frac{c_t}{2} \qquad (3.11)$$

where:-

n_t is the number of individuals entering in the study during time interval (T_t, T_{t+1}),

c_t is the number of withdrawing observations during time interval (T_t, T_{t+1}),

d_t is the number of deaths during time interval (T_t, T_{t+1}).

The probability of failure q_t conditional on the proportion of deaths between the time interval (T_t, T_{t+1}) is

$$q_t = \frac{d_t}{N_t} \text{ and so } p_t = 1 - q_t \qquad (3.12)$$

where p_t is the proportion surviving in the time interval (T_t, T_{t+1}).

The hazard rate, a life table estimator which is evaluated at the midpoint of the interval, is

$$h\,(t) = \frac{N_t q_t}{(T_{t+1} - T_t)N_t p_t + ((T_{t+1} - T_t)/2)(N_t - N_t p_t)}, \qquad (3.13)$$

which then leads to

$$h(t) = \frac{2q_t}{(T_{t+1} - T_t)(1 + p_t)} \quad . \qquad (3.14)$$

The estimated survival function $S\,(t)$, which is another life table estimate, is as follows:

$$S\,(t) = \prod_{k=0}^{t-1}\left[1 - \frac{d_k}{N_k}\right] = \prod_{k=0}^{t-1} p_k\,. \qquad (3.15)$$

The equation used to derive survival probability at time *t* is the same as the life table rule above**.**

The term “life tables” refers to any of a number of statistical tools used to assess the probability of an event occurring in a given time interval and its dependence on additional factors. The expression originated in actuarial science, where the probability of a person dying in the next year is modelled as a function of age, smoking and drinking habits, previous illnesses etc. Using a nonparametric method. Generally speaking, these approaches produce a table with rows labeled by time intervals, columns labeled by the variables whose effect is being studied and cells containing the probability (measured or predicted) of the event in question.

By contrast, Kaplan-Meier survival analysis (KMSA) is a technique of (again, non parametrically) associating a survival or, complementarily, hazard function to a given event history dataset. For instance, measurements of the incubation period of a virus may be cast into the form of a function which associates the cumulated probability of outbreak in a specimen to the time elapsed since infection, which in turn can be plotted. KMSA is most commonly used when the impact of factors other than time are considered insignificant.

3.4.2 Log rank test method

The log-rank test is a large sample chi-squared test which uses as its test criterion a statistic that provides an overall comparison of the Kaplan Meier curves. This log-rank statistic makes use of observed versus expected cell counts over categories of outcomes. The categories for the log-rank statistics are defined by each of the ordered failure times for the entire set of data being analysed. It is expressed as follows:

$$e_{1j} = (\frac{n_{1j}}{n_{1j} + n_{2j}}) * (m_{1j} + m_{2j}), \qquad (3.16a)$$

$$e_{2j} = (\frac{n_{2j}}{n_{1j} + n_{2j}}) * (m_{1j} + m_{2j}), \qquad (3.16b)$$

where:

e_{1j} is the expected number of individual events in group one,

e_{2j} is the expected number of individual events in group two,

n_{1j} is the number at risk in group one,

n_{2j} is the number at risk in group two,

m_{1j} is the number of failures in group one,

m_{2j} is the number of failures in group two.

Here the data is divided into J categories, labelled j= 1, 2, ..., J.

Further the difference between the observed and expected number of failures in each group is given by

$$O_i - E_i = \sum_{j=1}^{J} (m_{ij} - e_{ij}) \ , \qquad (3.17)$$

with variance

$$Var(O_i - E_i) = \sum_{j=1}^{J} \frac{n_{1j} n_{2j} (m_{1j} + m_{2j})(n_{1j} + n_{2j} - m_{1j} - m_{2j})}{(n_{1j} + n_{2j})^2 (n_{1j} + n_{2j} - 1)} \ , \ i = 1, 2. \qquad (3.18)$$

Here:

$E_1 = \sum_{j=1}^{J} e_{1j}$ is the expected number of all events in the first group,

$E_2 = \sum_{j=1}^{J} e_{2j}$ is the expected number of all events in the second group,

O_1 is the number of observations in the first group,
O_2 is the number of observations in the second group,
and J is end time of the study.

Basically, the log-rank test is a hypothesis test with the null hypothesis: there is no difference between the two survival curves. With this hypothesis the log-rank statistic is approximately chi-square with one degree of freedom. The p-value for the log-rank test is determined from tables of the chi-square distribution. The alternative hypothesis is simply that there is a difference between the two survival curves.

Based on the equations 3.16a, 3.16b, 3.17 and 3.18 then the log-rank test method gives:

$$\text{Log-rank test statistic} = \frac{(O_1 - E_1)^2}{Var(O_1 - E_1)} + \frac{(O_2 - E_2)^2}{Var(O_2 - E_2)}. \qquad (3.19)$$

The log-rank test is used to test whether the difference in survival times between two groups is statistically significant or not, but it does not test the effect of the other independent variables.

3.4.3 Parametric survival analysis:

The parametric approach derives estimates of failure time statistics while accounting for the presence of censoring in the data as in the non-parametric approach. The main difference is to derive estimates using a parametric model, which make specific assumptions about the distribution of failure times by assuming a particular functional form for the hazard rate. This functional form can specify the hazard rate as a function of time. Alternatively, it can incorporate covariate information so that the hazard rate is specified as a function of time and specific covariates. Failure time is then related to a set of covariates thus leading to a regression approach (Machin et al., 2006).

Parametric methods of survival analysis assume distribution of hazard rates as a function of time with the assumption of independent censoring. The hazard rate is defined as an instantaneous probability of dying in the next short interval conditional upon having survived until time t. The functional form of the event times for parametric methods is constructed using various statistical distributions. Commonly used statistical distributions are: Exponential, Weibull, Log-Logistic and Gompertz. These distributions are useful for survival analysis data. While the exponential distribution is used to model processes with a constant hazard rate, the other three are more flexible two parameter distributions which allow for the modelling of a wide variety of shapes. They are possible candidate because their parameters have positive values.

Table 3-3 Example probability density functions, in each case valid on the region [0,∞)

Distribution function	**Probability density function** $f\,(t)$
Exponential	$\lambda \exp(-\lambda t)$ where $\lambda > 0$
Weibull	$\delta\lambda\,(\lambda t)^{\delta-1} \exp[-(\lambda t)^{\delta}]$ where $\lambda > 0, \delta \geq 1$
Log-Logistic	$\delta\lambda^{\delta} t^{\delta-1}[1+(t\lambda)^{\delta}]^{-2}$ where $\lambda > 0, \delta \geq 1$
Gompertz	$\delta\lambda\, e^{\delta t}\, e^{\lambda} \exp(-\lambda e^{\delta t})$ where $\delta, \lambda > 0$

They generally involve two parameters: the scale (δ) and shape (λ) parameters, where:

(λ) The shape is generally assumed to be constant across individuals,

(δ) The scale parameter is estimated by using a regression model (Crowder, 2012).

Theoretically this is the same as in a linear regression model but the Normal distribution is replaced by the exponential distribution. It is implemented in a regression framework, with estimates found by maximizing the likelihood of the data for patients observed to have an event (death) at time t. The likelihood contribution is represented by $p_r\,(T = t) = f\,(t)$ which is the density function at time t and p_r represent a probability. For patients censored at time t, their likelihood is represented by $p = p_r\,(T > t) = S\,(t)$.

Generally the functions which characterize parametric distributions are as follows:

- Density Function: $f\,(t) = p_r\,(T = t)$

- Cumulative Incidence: $F\,(t) = p_r\,(T \leq t)$

- Survival Distribution: $S(t) = p_r(T > t)$

- Hazard Function: $h(t) = \dfrac{f(t)}{S(t)}$, see also Lambert and Royston (2009).

3.4.4 Semi parametric survival analysis method

Semi-parametric regression models are used to describe survival time in a comparative sense, for example, if we are interested in how a new treatment affects survival compared to an old treatment. A functional form for the hazard over time need not be specified. The commonly adopted approach is the Cox-proportional hazards model.

Cox-proportional hazard does not assume any functional form of the distribution of hazard rate but assumes that the hazard functions of any two individuals are proportional with the ratio being determined by the covariates. If one is unsure of the functional form of the hazard function then adopting a semi-parametric approach would be the preferred alternative rather than imposing specific parametric assumptions (Cox, 1972: Breslow, 1972). The basic Cox PH hazard function regression model is formulated as follows:

$$h_T(t,x) = h_0(t)\exp(x'\beta) = h_0(t)\exp(\sum_{i=1}^{p}\beta_i x_i). \qquad (3.20)$$

where x_i represents the value of the ith variable, β represents the coefficient parameter measure of the risk of variable i and $h_0(t)$ is the baseline hazard function.

By letting $\phi(x) = \exp(\sum_{i=1}^{p}\beta_i x_i)$,

we obtain

$$h_T(t,x) = h_0(t)\phi(x). \qquad (3.21)$$

The formula for the Cox PH survival function is expressed as follows:

$$S_T(t,x) = [S_0(t)]^{\phi(x)}. \qquad (3.22)$$

Here, $S_0(t) = \exp(-H_0(t))$, and $H_0(t) = \int_0^t h_0(u)du$

where,

$S_0(t)$ is the baseline survival function,

and

$H_0(t)$ is the baseline cumulative hazard rate,

By taking logs in equation (3.22),

$$\ln S_T(t,x) = \phi(x) * \ln \left[S_0(t)\right]. \qquad (3.23)$$

Since $0 \le S_T(t,x) \le 1$, $\ln S_T(t,x)$ and $\ln S_0(t)$ are negative. Thus equation (3.23) should be multiplied by (-1) and taking the logarithm again for it as follows (Newby, 2010) yields

$$\ln[-\ln S_T(t,x)] = \ln[\phi(x)] + \ln[-\ln[S_0(t)]]. \qquad (3.24)$$

3.5 Advantages and disadvantages of survival analysis

Survival analysis accounts for both censored observations and time to event because the t-test and linear regression can be used to compare the mean time to event between two groups. Logistic regression can be used to compare proportions of events whilst ignoring the time. The non-parametric method uses the smallest number of assumptions but can only compare a limited number of groups. It cannot deal with continuous variables nor control for other variables. The parametric technique deals with both discrete and continuous explanatory variables and allows for a large number of explanatory variables. However, assumptions on time dependence and how the explanatory variables influence the risk of death need to be made.

The semi-parametric technique requires the last assumption only. However, the estimated parameters will be less precise and hypotheses about time dependence can no longer be tested (Cleves et al., 2002).

4 CHAPTER 4: Basic features of the data

4.1 Introduction

Building on the previous one, this chapter deals with the analytical part, methods and procedures of the study. To determine the breast cancer incidence rate among women in Iraqi Kurdistan and to identify the factors contributing to it, survival analysis and univariate statistics were used. Subsequently we assembled, classified and tabulated the data and codified the variables. Descriptive and survival analyses were employed to fit the data with the help of the Statistical Package for Social Sciences (SPSS) version 22, Statgraphics version 16 and Wolfram Mathematica 10 software packages.

In order to plan and perform a data analysis of this type we need to specify three things: the data type needed, the data collection method to be used, and the data processing mechanism. The questions we want to answer are as follows:

1. What are the factors that have a high impact on breast cancer in the region?
2. How can we produce an appropriate survival functions for the data?

This chapter studies the concerns regarding the selection of research methods in order to answer these questions, methodology for the analysis of quantitative indicators, and justifications of the tools used.

4.1.1 The required data

We will be dealing with descriptive statistics, in particular, demographic data which was obtained from official sources in the Kurdistan region. The demographic data includes the patients' general information. Since the data was obtained from official sources it is of better quality than data obtained from other sources such as interview, public news paper and survey. We will focus on right censored data which we can divide into two different groups:

1. Social data which includes age, marital status, education, occupation, religion, ethnicity, weight, height, body mass index (BMI), residency, smoking, family income, family history, alcohol, menstrual cycles, number of children, the first pregnancy age, breast feeding, moderate levels of exercises and obesity.
2. Information on hormones: ER, PR, tumour size, tumour grade and lymph nodes.

4.1.2 Data collection procedures

This research uses secondary data based on laboratory investigation including hematology, biochemical study and/or radiology. This data is supplied by two main hospitals in the Kurdistan Region of Iraq, which are: Hewa Hospital in Suleimaniaha, and Nanakaly Hospital for Leukemia in Erbil (from 1991 until 1 June 2014). The data was collected from existing databases with no access to the names of the individuals and therefore there will be no direct involvement of participants in the study.

Prior to the actual start of the study, approval by the Research Ethics Committees at City University had to be obtained by means of the standard application process. In addition, in order to facilitate the actual data collection in Kurdistan hospitals, official permission by the Kurdish Ministry of Higher Education and Scientific Research was required. Following preliminary approval by the Post Graduate Research Office at City University and the Kurdistan Regional Government Representative to the United Kingdom, the ministry issued a formal letter of concession to Professor Mark Broom and myself in May 2013. This in turn enabled us to propose our work to the Kurdish Ministry of Health, which, under the condition of confidentiality, instructed the hospital administration to provide the required datasets.

The actual process of data acquisition in the Hewa and Nankaly hospitals in Suleimaniah and Erbil, respectively, took place in the summer of 2013. The respective Departments of Statistics provided the data in the form of extensive Excel sheets, which for further processing were later transferred to SPSS.

As was specified in the confidentiality agreements with the Ministry of Health and City University, the data was only stored and made available electronically and will be destroyed after the end of the study; access was limited to Professor Broom and myself.

In summary, this is the first analysis of this kind carried out for this region and the only one based on this specific dataset. However, given that the collection of the relevant data is a simple process and the quality of the data may be improved further as outlined in the final remarks of this dissertation, we are hopeful that this will not remain the only study of this type and the systematic analysis of breast cancer data will prove beneficial for public health in Northern Iraq.

4.1.3 The data processing

The data collected is analysed using a generic framework that best suits this data. The framework involves processing data through three stages:

1. Univariate statistics.

2. Adjusting the basic Markov chains model for both Nanakaly and Hewa data with and without censoring.

3. Classical survival analysis which includes the semi-parametric method; Cox regression, and non-parametric methods such as Kaplan-Meier and log-rank test model estimates of the survivor function.

4.2 Univariate statistics

In our initial analysis, we look at descriptive statistics which are used to describe the basic features of the data. We provide simple summaries of the data and its measurements. We stall look at the Nanakaly and Hewa data in turn.

4.2.1 Univariate statistics for Nanakaly data

To begin with, a table of descriptive statistics is obtained to give a general idea about the variables used. Here we have two variables, age of patients and survival time for them.

Table 4-1 Univariate statistics for Nanakaly data

	Age (years)	Survival Time (days)
N	713	713
Missing	0	0
Mean	48.96	862.62
Std. Error of Mean	.435	26.626
Median	48.00	624.00
Std. Deviation	11.608	710.969
Variance	134.74	505477.03
Range	71	2601
Minimum	18	1
Maximum	89	2602

In Table 4.1 age and survival time are non categorical variables, and simply represent the age in years of the patients and the survival time of the patients in days. Where N is the number of patients.

4.2.2 Univariate statistics for Hewa data

There are two types of variables in Table 4.2, categorical and non categorical variables. The non categorical variables include age, weight, height, BMI, estrogen and progesterone. They represent the status of the patients, for instance the age variable is the age in years of the patients, weight is measured in kilograms (kg) while height is measured in meters (m). The BMI is the standard measurement for the patients based on the formula weigh/(height)^2. Estrogen and progesterone are the main female hormones which are measured by pg/ml (picograms/ milligram) and ng/ml (nanogram/milligram). One nanogram of progesterone is 1000 picograms. This section is an exploration of the data and that without additional information on the general population, no information can be extracted about the risk factors associated with breast cancer from the data.

We discuss below each of the categorical variables (marital status, religion, occupation, income, menopause, hormone, tumour grade, exercise, smoking, drinking alcohol, family history and breastfeeding) which may have an effect on the likelihood of contracting breast cancer. The marital status is the patient's situation with regard to whether she is single, married, divorced or widowed which are categorized by 1,2,3, and 4 respectively. The religion variable refer to a patient's belief in and worship of God or gods and they are classified as Muslim (1), Christian (2), other (3). While occupation defines a patient's regular work or profession; job or principal activity and this includes 12 categories; housewife (1), business manager (2), doctor (3), educator (high school graduate, self employed) (4), lawyer (5), police officer (6), retired (7), student (8), teacher (9), university teacher (10), worker (11) and other (12). The family income variable is the monetary payment received for goods or services, or from other sources, as rents or investments and they are labelled as very good (1), good (2), medium (3) and poor (4). The possibility of contracting breast cancer is affected by the number of menstrual periods of women (Abuelghar et al., 2013). Those who have had less menstrual cycles because they started menstruating late or stopped menstruating at an early age or because of pregnancy, have a slightly lower risk of breast cancer. Women are categorized as either post or pre menopausal with "Yes" meaning post menopausal category (1) and "No" meaning pre menopausal category (2) as referred to in Table 4.4. The hormone category means that either the patients have hormone problems related to breast cancer or not. The hormone problem is identified when the clinical practice tests the hormone levels for the patients.

A hormone balance of cortisol, DHEA, estrogen, progesterone and testosterone are essential to good health for women of all ages especially two main hormones estrogen and progesterone connected to breast cancer. There are three types of tests that can be used to determine hormone levels. There is a Saliva test, Serum or blood test and Follicle-stimulating hormone (FISH) test. FISH is the most common test and it is frequently used to determine the hormone status of premenopausal women who may complain of hot flushes, mood changes or other symptoms. When a hormone imbalance is detected early and steps are taken to correct it, symptoms can be relieved and progression to disease states may be prevented. Here, in Table 4.5 the "Yes" category (1) refer to the patients who have an imbalance in estrogen and progesterone hormone and the "No" category (2) means that the patients do not have a hormone imbalance.

The two primary female sex hormones are estrogen and progesterone. The activity of the receptors associated to these hormones is strongly linked to the growth of typical breast cancer cells and plays a role in many cases. Cancer cells respond to these hormones through the estrogen receptors ER and progesterone receptors PR. ER and PR are cells receiving these hormones circulating in the blood. The tumour is tested for these receptors in a test called a hormone receptor assay. If a cancer does not have these receptors, it is referred to as hormone receptor negative, in particular, ER negative and/ or PR negative. On the other hand, if the cancer has these receptors then it is referred to as hormone receptor positive, that is ER positive and/ or PR positive. These receptors are important because cancer cells that are ER or PR positive often stop growing when drugs that either block the effect of ER and PR or decrease the body’s levels of ER are taken. These drugs lower the chance of the cancer recurring, thereby improving the chances of living longer. They form part of the treatment in patients whose breast cancer is ER or PR positive. However, these hormone active drugs are not effective when the cancer does not contain these receptors. All types of breast cancer should be tested for hormone receptors except for lobular carcinoma in situ because it is a very advanced stage of cancer, with the cancer cells spread almost all over the breast. Women should ask their doctor for these test results so that they can determine whether hormone treatment active drugs can form part of their treatment. The "Yes" category in Table 4.5 indicates that 39.7% of the patients have hormone problems; while "No" means that 60.3% of the patients do not have this problem (note all patients are tested for hormone problems).

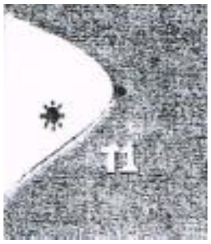
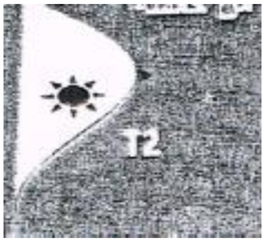
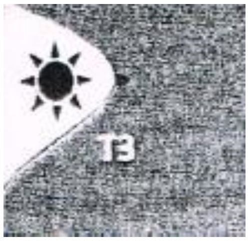
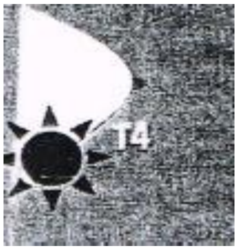
Tumour size describes the size of the original tumour and is measured in millimetres (mm) (see Table 4.3). For example, if the diameter of the tumour is smaller than 2 cm then the tumour size will be classified as T1, which means that it is in the initial stage out of the four grades from Table 4.3. Note that these T classifications are not used in our data. Lymph nodes are small glands located through system. They act as filters, removing waste fluid from the body American community survey (ACS 2006). Lymph nodes can be measured along the short or long axis (Hoang et al., 2013). In this study the lymph node values of 1 to 44 refers to the number of cancerous nodes, with median equal to 4. It is medically graded by numbers from N0 to N3, i.e. for N0 the cancer is not separated under the breast tissue, N1 represents 1 to 3 lymph nodes under the arm, N2 denotes that the cancer are separated to 4 to 9 lymph nodes under the arm finally N3 refers to 10 or more lymph nodes existing under the arm or including extra lymph nodes around the breast. See Table 4.3 for more details, although again the N classification is not used in our data.

Table 4-2 Univariate statistics for Hewa data

	Age	Survival Time	Progesterone Receptors	Estrogen Receptors	Menopause	Hormone	Tumour Size (mm)	Tumour Grade (cm)	Lymph Nodes	Marital Status	Exercise
N	1163	1163	1163	1163	1163	1163	1163	1163	1163	1163	1163
Missing	0	0	0	0	0	0	0	0	0	0	0
Mean	49.37	762.92	122.3	86.1	1.52	1.6	32.31	1.77	7.77	*	1.67
Std. Er. of Mean	.340	21.634	5.292	2.86	.015	.01	.440	.019	.233	*	.014
Median	48.00	704.00	65.00	63.0	2.00	2.0	28.00	2.00	4.00	*	2.00
Std. Deviation	11.59	737.78	180.5	97.6	.500	.49	15.00	.632	7.951	*	.470
Variance	134.4	544328	32574	9534	.250	.24	225.0	.400	63.216	*	.221
Range	72	6770	2396	851	1	1	91	2	43	*	1
Minimum	20	3	0	0	1	1	9	1	1	*	1
Maximum	92	6773	2396	851	2	2	100	3	44	*	2

	Weight	Height	BMI	Family History	Religion	Smoking	Drinking Alcohol	Occupation	Income	Breast Feeding	
N	1163	1163	1163	1163	1163	1163	1163	1163	1163	1163	
Missing	0	0	0	0	0	0	0	0	0	0	
Mean	71.3	168	32.15	1.89	*	1.83	1.85	*	*	1.68	
Std. Er. of Mean	.870	3.05	3.200	.009	*	.011	.010	*	*	.014	
Median	68.0	160	26.04	2.00	*	2.00	2.00	*	*	2.00	
Std. Deviation	29.7	104	109.1	.310	*	.374	.354	*	*	.466	
Variance	881	1086	11908	.096	*	.140	.126	*	*	217	
Range	560	1635	2800	1	*	1	1	*	*	1	
Minimum	40	1.5	17.8	1	*	1	1	*	*	1	
Maximum	150	1.65	55.9	2	*	2	2	*	*	2	

Table 4-3 Illustrates information for tumour size and lymph nodes

Tumour Size T	The diameter of the tumour is less than 2 cm 	The diameter of the tumour is more than 2 cm and less than 5 cm 	The diameter of the tumour is more than 5 cm 	The tumour penetrates the skin or the rib cage 
Lymph Nodes N	**N 0** There are no cancerous cells in the lymph nodes	**N 1** The cancer spreads to non-adjacent lymph nodes under the armpit from the same side of the infected breast	**N 2** The cancer spreads to adjacent lymph nodes under the armpit from the same side of the infected breast or metastasis of breast internal lymph nodes	**N 3** The cancer spreads to lymph nodes under or above the clavicle or breast internal lymph nodes and under the armpit.

Cited from http://www.thebestoncologist.com/Arabic/breast_cancer_stages.html

Generally there are three stages for tumour grade based on 3 cm measurements; small which is about 3 cm, medium is greater than 3 cm but it is not spread in the chest and large tumour grad is greater than 3 cm and it is the advanced stage of the cancer cells. These grades depend on the number of the combined tumour sizes. Pathologists look at breast cancer tissue under a microscope to determine the grade of the tumour, which depends upon how much it looks like normal breast tissue. Cancers that closely resemble normal breast tissue tends to grow and spread more slowly, and get a smaller grade. In general, it indicates a cancer that is less likely to spread, and a larger grade indicates a cancer that is more likely to spread. Tumour grade is based on the arrangement of the cells in relation to each other, whether they form tubules (small tumour about 1cm or less), how closely they resemble normal breast cells (nuclear grade), and how many of the cancer cells are in the process of dividing (mitotic count). A small tumour grade cancer may also be called “well differentiated” because it more closely resembles normal breast cells. Similarly a large tumour grade may also be called “poorly differentiated” since the cells have lost their resemblance to normal breast cells. In Table 4.6 the small tumour grade is categorized by (1), the medium tumour grade by category (2) and category (3) refers to the large tumour grade.

Physical activity in the form of exercise reduces breast cancer hazard with "Yes" indicating that the patient is doing exercises such as running, swimming, or any other sports at least three times a week and "No" that they are not (Calle, et al., 2003). In Table 4.8, the "Yes" category indicates that 33% of patients are exercising. Meanwhile, 67% of patients are not doing the mentioned activities. Unfortunately the numbers of patients who are doing exercise are less than those not doing it.

Family history plays a significant role in causing breast cancer. The possibility of infection with breast cancer increases between 1.5-3 times in those with a first degree relative (mother, sister or daughter) who suffer from breast cancer, and there are mutations in genes which lead to an increase of the possibility of the disease. Here disease often appears in the patients at a younger age. These are reflected by category (1) "Yes" that they have a family history of breast cancer and the "No" category (2) that they do not, as shown in Table 4.9. The risk rates for patients who smoke more than three cigarettes per day on average are more than for non smokers, the "Yes" category (1) in Table 4.13 refers to the fact that 16.9% of patients smoke at or above this level while the "No" category (2) that 83.1% of patients do not.

Furthermore, the risk for patients who drink alcohol generally more than three times per week is higher than for the patients that do not drink. Table 4.14 shows that 14.7% of patients drink alcohol as represented by category (1) "Yes", and 85.3% of them do not drink alcohol, represented by category (2) "No". Finally, breastfeeding may slightly decrease risk, and the possibility of breast cancer during the breastfeeding period is lowered. Table 4.15 illustrated categories (1) "Yes" meaning that the patients are breastfeeding and the "No" category (2) that they are not. Table 4.3 shows the general information about 21 variables included in the Hewa data.

Table 4-4 Frequency table for menopause

		Frequency	Percent
Menopause	Yes	558	48.0
	No	605	52.0
	Total	1163	100.0

Table 4.4, shows Menopause Status for Hewa Hospital patients. Out of 1163 women with breast cancer, 48% are menopausal while 52% of them are not. In general, the age of menopause in the Kurdistan Region is around 50-52 years old.

Table 4-5 Frequency table for hormone

		Frequency	Percent
Hormone	Yes	462	39.7
	No	701	60.3
	Total	1163	100.0

Table 4.5 illustrates that 39.7% of women are suffering from a hormone-related anomaly, whilst 60.3% are not. It is a clinically established fact that, in the case of receptor positive breast cancer cells, hormonal therapy can be employed to counter the effect of hormones on the cells' growth and overall functioning.

Table 4-6 Frequency table for tumour grade

		Frequency	Percent
Tumour Grade	Small(I)	397	34.1
	Middle(II)	637	54.8
	Large(III)	129	11.1
	Total	1163	100.0

Table 4.6 is the frequency table for the tumour grade; an indicator of how quickly a tumour is likely to grow and spread in the breast. It shows that 54.8% of patients have the intermediate degree tumour grade. Cancer cells do not look like normal cell and they grow faster than normal. Even though 11.1% are diagnosed late, their treatment is easier than the middle grade because the cancer cells can be eradicated. With regards to the small tumour grade, which represents 34.15% of patients in Hewa hospital, diagnosis at the early stages is better when controlling the cancer because they are usually growing more slowly.

Table 4.7, shows that out of 1163 patients 50.7% are married, 5% are single and 39.2% are widows.

Table 4-7 Frequency table for marital status

		Frequency	Percent
Marital Status	Single	58	5.0
	Married	590	50.7
	Divorced	59	5.1
	Widow	456	39.2
	Total	1163	100.0

Table 4-8 Frequency table for exercise

		Frequency	Percent
Exercise	Yes	384	33.0
	No	779	67.0
	Total	1163	100.0

Women whose close blood relatives have breast cancer have a higher risk of this disease. Table 4.9 illustrates the frequency table for patients who have a family history of breast cancer. The risk is doubled if the women have a first-degree relative (mother, sister or daughter). Having one first-degree relative (mother, sister, or daughter) or two second degree relatives with breast cancer increases the risk approximately threefold. Women also have increased risk of breast cancer if their father or brother has breast cancer but the exact effect on the risk is not known.

Table 4-9 Frequency table for family history

		Frequency	Percent
Family History	Yes	125	10.7
	No	1038	89.3
	Total	1163	100.0

Overall more than 10% of women have a family member that has breast cancer, but a majority of the women (over 89%) do not have a family history of breast cancer.

Table 4-10 Frequency table for occupation

		Frequency	Percent
Occupation	Housewife	933	80.2
	Business	2	.2
	Doctor	3	.3
	Educator (Self Employed)	3	.3
	Lawyer	1	.1
	Officer	102	8.8
	Retired	24	2.1
	Student	1	.1
	Teacher	89	7.7
	University Teacher	1	.1
	Worker	3	.3
	Other	1	.1
	Total	1163	100.0

Table 4.11 is the exploration of the family income level of breast cancer patients. It shows that 57.2% are middle income women, whilst this figure is only 6% for those with a very good income. This reflects the larger number of middle income women in the general population, and is not an indicator of risk.

Table 4-11 Frequency table for family income

		Frequency	Percent
Income	Very Good	70	6.0
	Good	295	25.4
	Medium	665	57.2
	Poor	133	11.4
	Total	1163	100.0

Table 4.12 shows that 85.7% of patients in Hewa hospital were Muslims and 10.7% were Christian, while the remaining patients 3.5% were from other religions.

Table 4-12 Frequency table for religion

		Frequency	Percent
Religion	Muslim	997	85.7
	Christian	125	10.7
	Others	41	3.5
	Total	1163	100.0

Table 4-13 Frequency table for smoking

		Frequency	Percent
Smoking	Yes	196	16.9
	No	967	83.1
	Total	1163	100.0

Table 4-14 Frequency table for drinking alcohol

		Frequency	Percent
Drinking Alcohol	Yes	171	14.7
	No	992	85.3
	Total	1163	100.0

Table 4-15 Frequency table for breastfeeding

		Frequency	Percent
Breastfeeding	Yes	371	31.9
	No	792	68.1
	Total	1163	100.0

Table 4.13, 4.14 and 4.15 reflect the frequency percentage for three variables smoking, drinking alcohol and breast feeding respectively.

4.3 Timing data

In general, the process of collecting data in the health sector or any other sector in a developing country such as Iraq is not easy, because there is no accurate database system. The most dependable data are available in the official records but not obtainable readily. One inevitably has to refer to numerous government agencies to obtain relevant information from official sources such as the Ministry of Health, especially for information regarding the time of death.

The nature of this study requires the collection of primary data in two main hospitals in the Region. The Hewa data includes general information about the breast cancer patients, including their age, religion, tumour size, tumour grade, lymph nodes, exercise, the educational level, family history, breast feeding, smoking, drinking alcohol, occupation, progesterone, estrogen, menopause, marital status and income. In addition there are three times involved in this data, time of admission, time of diagnosis and time of death.

In studying survival analysis, it is necessary to have all of the relevant information about time of diagnosis and the time of death. The time of admission refers to the first time when the patient visits the hospital and the staff of the hospital administration register general information about her. The diagnosis time means the time when the doctor diagnoses the patient and refers them to the laboratory to make necessary required tests based on the symptoms that they are suffering from. Finally there is the actual time of death of the patient.

In initial analysis later in this study we use *z* as an intermediate measurement for the rate of real death (which is not given) from the time of diagnosis. This serves to account for the missing reports on times of death of a number of patients who did not return for the follow up appointments, a phenomenon which is indicative of a larger problem with the patients' reaction toward this particular diagnosis which may be due to poor health education and general lack of awareness of the importance of keeping detailed and complete hospital records. The reason may include a general fear of disease or death and the hope of receiving better treatment elsewhere. Hospital record consistencies and general compliance appears to be correlated to economic status and doctor-patient interaction, and also the apparent termination of a patient's follow up treatment may simply be due to bad bookkeeping. In particular it is clear that actual death records are almost entirely absent in the data that was obtained.

That is the basic problem leading to not having the real time of death, which is reflected in the results of the survival curve when we applied the SPSS program at the first step, because we used the time of admission instead of the time of death. Clearly there are differences between the time of admission and the time of death, and this leads to significant problems with the analysis. That is why we extend the analysis of the survival curve using a Markov process. This is the most natural and simple extension to the model to try to deal with the absence to the actual time of death. Here when we applied the Markov process we considered two modifications which are without censoring and with censoring.

We used Markov processes because of some problems regarding the applications of the survival function curve due to the lack of knowledge of the time of deaths and hidden censoring. Nevertheless there are some serious limitations to this model. For example if individuals do not get censored at constant rate, the censoring time will follow another distribution rather than an exponential distribution and we will obtain a different picture to that obtained from a Markov process.

Potentially more significant problems result from the lack of knowledge of the times of death. Deaths are assumed to follow a Markov process from the time of diagnosis category (see Section 4.5.1) and there are two main source of error. Firstly the rate z of this process is unknown and had to be estimated and we have thus considered a range of values. Secondly, again this may not be a Markov process, which would also affect the shape of the survival function.

Finally the following table is general information and descriptions for medicine treatment.

Table 4-16 General information about specific medical terms and descriptions

Medical terms	Descriptions
Mammography	A type of medical examination used for early detection and diagnosis of breast lumps.
BSE	Breast self examination.
Lymphatic	A breast network of blood that brings in nourishment and remove waste products.
Mastectomy	Remove entire breast including the nipple.
Metastasis	A tumour which has spread beyond its original domain.
BRCA1 & BRCA2	Breast cancer gene (1 and 2) are two genes which are linked to breast cancer risk.
Protein 53 (P53)	The Li-Fraumeni Syndrome is a genetic disease which is caused by a mutation of the P53 gene and whose symptoms include the occurrence of soft tissue disease caused at a young age.
Estrogen (ER) and Progesterone Receptors(PR)	A high rate of activity of the receptors associated to the female sex hormones estrogen and progesterone in breast cancer cells may stimulate tumour growth.
Tamoxifen	Anti Estrogen drug used most often.
CK56	Cytokeratin 5/6 is an indicator commonly used replacement immunohistochemical for tumours with the basal-like gene expression profile.
Test (Ki-67)	The Ki-67 test measures the speed at which a tumour grows.
BMI	Body mass index.
Immunohistochemical (IHC)	The IHC test is used to determine the HER2-receptor protein in a tumour.
Letrozole	A type of medication.
HER2	Human epidermal growth factor receptor in the tumours of the breasts fall into two categories: The HER2- positive type exhibits multiple HER2-genes and receptor overexpression while the HER-negative type shows no anomalies in HER2-gene expression. Tumours of the first type generally grow faster and more aggressively.
Ductal Carcinoma In Situ (DCIS)	DCIS is a type of cancer which does not spread beyond the milk ducts and is therefore called Non-Invasive.
Lobular Carcinoma In Situ (LCIS)	LCIS is a type of tumour characterized by regions of uncontrolled growth of lobular tissue. While it is non-invasive and is generally not considered cancer itself, it indicates increased risk of developing invasive breast cancer.
Invasive Ductal Carcinoma (IDC)	A tumour of the milk ducts which has advanced into the surrounding tissue is called IDC.
CTC	Circulating tumour cell.
Anthracycline	Type of chemotherapy.
Fluorescence In Situ Hybridization (FISH)	The FISH test is designed to detect additional copies of the HER2 gene in cells.
EVO	The time of birth.
HRT	Hormone receptor therapy.

Medical terms	Descriptions
Hormone Receptor (HR)	Hormone receptor are proteins in the membrane which facilitate the signal transfer into the cell.
S-Phase Fraction	The S-Phase Fraction of a given cell sample is defined as the percentage of cells which currently undergo DNA replication. For tumours of the breast anything above 10% is considered a high S-Phase Fraction value.
PCR and LH	(Female hormone profile).
Cortisol	A stress hormone produced in the outer layers of the adrenal gland.
DHEA (Dehydroepiandrosterone)	A hormone created in the adrenal gland which is an intermediate stage in the synthesis of sex hormones such as estrogen and testosterone.
Luminal A & Luminal B Tumour	A hormone receptor positive tumour is said to be of the Luminal type. One further distinguishes between Luminal A tumours which are HER2-negative and Luminal B ones which are HER2-positive.
Basal-like Breast Cancer	The basal-like breast cancer type is both hormone receptor and HER2-negative (and therefore sometimes called triple-negative).
Digoxin	A drug which by inhibiting sodium-potassium pumps in the cell membrane reduces atrial flutter.
BCL-2	Beta-cell lymphoma leukemia 2; it is a mitochondrial protein known to inhibit apoptosis triggered by chemotherapy and radiation therapy.
DALYs	Disability adjusted life years.
ASRs	Age standardized rates.
NHS	National health service

5 CHAPTER 5: Markov chain models for breast cancer

5.1 Introduction and application of survival analysis to Nanakaly Kurdish data

There are well-established survival analysis methodologies for data sets which are complete, with accurate information on censoring as discussed in chapter 3. But what if they are not complete? In this chapter we consider how to analyse cases where "hidden censoring" occurs, where individuals have left the study but the hospital is unaware of this. We develop a new Markov chain-based methodology for generating survival curves and hazard functions, and demonstrate this using our breast cancer datasets from the Kurdistan region of Iraq.

This section studies the status of breast cancer in two main hospitals in the Kurdistan Region. The work in this section has been published in Raza and Broom (2016). Firstly we try to determine the survival time for breast cancer patients in the Nanakaly data (see Figure 5.2) in general, where the Kaplan-Meier curve for the original data (713 patients) has been used,.

Here the vertical axis represents the number of individuals and the horizontal axis is survival time.

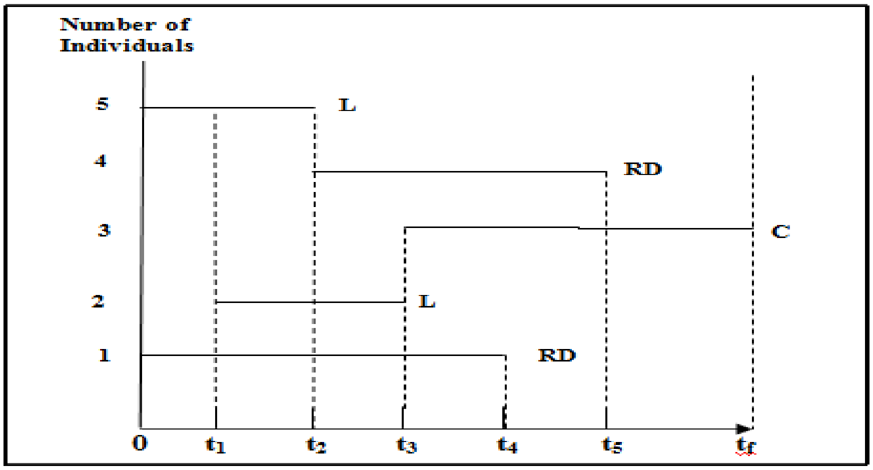


Figure 5-1 The plot including censoring for the Kurdish data from Nanakaly hospital

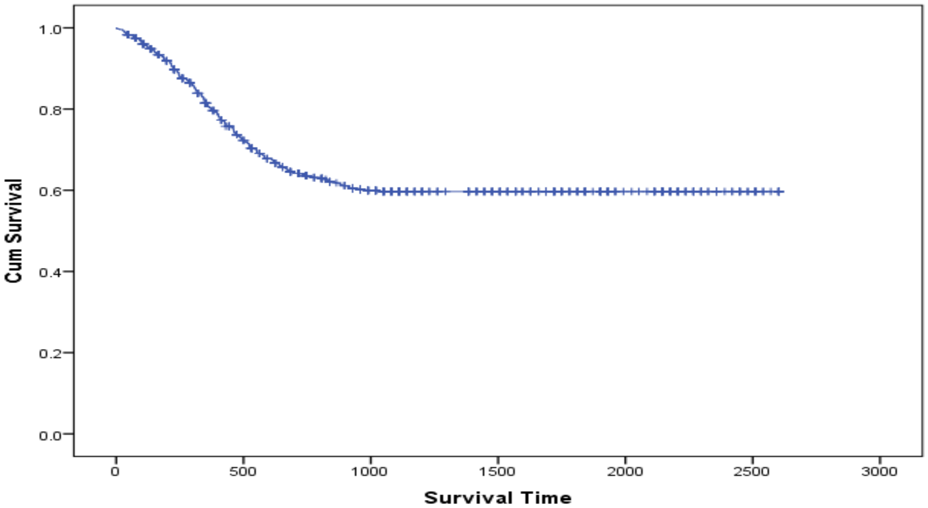


Figure 5-2 The original survival curve for the Kurdish data from Nanakaly hospital

Detailed times of death were provided, with censoring at the end of the study period on 1st June 2014, where C represents of end of period censoring, (RD) is recorded death, and (L) is hidden censoring. i.e. individuals unknowingly lost to the study, (see Figure 5.1 and the explanation below). Analysing the above data using SPSS, provided the Kaplan-Meier survival curve in Figure 5.2. This is clearly not a realistic survival curve. The problem with the survival curve from Figure 5.2 is that we calculated it on the assumption that all individuals other than those who died (or were censored by reaching the end of the study period) were still active in the study, but in fact individuals often did not return to the hospital after initial treatment, and there are no clear records of when the deaths of these individuals occur, or of which individuals these are. Thus there is some secret censoring that we do not have knowledge about. In other words, whilst the values of d_t are accurate, the values of n_t are not, where d_t and n_t are as defined in section (3.4.1.1) and we are (after some time, greatly) overestimating them.

The survival function flattening out to effectively a horizontal line, indicates a hazard rate tending to zero, suggesting some problems with the data. This shows a need for adjustment, which we carry out using a Markov chain model. The properties of Markov chains fit well with the study. For instance both cases (Nanakaly and Hewa) can be modelling assuming that they depend continuously on time with constant transition rate represented by:

$$P' = \frac{dP}{dt} = PQ \qquad (5.1)$$

for a given transition matrix $Q = (q_{is})$, where q_{is} is the rate of flow from ($i \rightarrow s$), which is a |P| * |P| matrix of transition rates if it fulfils the following two conditions:

a) Q has no negative off-diagonal entries, i.e. $q_{is} \geq 0$ for all $i \neq s$.

b) Q has row sums equal to zero, or $\sum_s q_{is} = 0$ for all i.

$$Q = \begin{pmatrix} q_{11} & q_{12} & \cdots & q_{1s} & \cdots & q_{1n} \\ q_{21} & q_{22} & \cdots & q_{2s} & \cdots & q_{2n} \\ . & . & \cdots & . & \cdots & . \\ q_{i1} & q_{i2} & \cdots & q_{is} & \cdots & q_{in} \\ . & . & \cdots & . & \cdots & . \\ q_{n1} & q_{n2} & \cdots & q_{ns} & \cdots & q_{nn} \end{pmatrix}$$

The solution of 5.1 identifies the following equation for *P*:

$$P = P_0 e^{tQ} \qquad (5.2)$$

subject to initial conditions $P_I(0) = P_0$.

5.1.1 Markov model without censoring (Nanakaly data)

We shall first introduce a Markov model without overt censoring. In our data the only observed censoring was caused by the end of the study period, although as patients were being recruited all the time during the period, the censoring time could be small, and such censoring could occur for any time less than 2602 days, the time from the earliest record considered to the end of the study period.

The following Figure 5.3 represents the Markov survival model with no censoring.

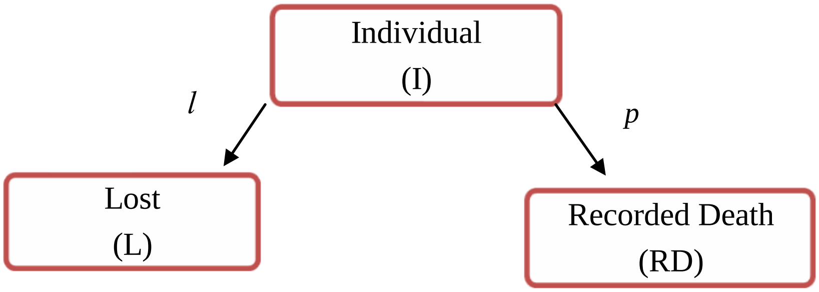


Figure 5-3 The markov survival model with no censoring

Consider a population of individuals in three categories; either at risk I, died RD or who have left the study (without our knowledge), which we shall call “lost” L. Individuals simply move from state I to the other two states at constant rates l to L and p to RD. We thus have a population as described by Figure 5.3. Denoting the proportion of individuals in states I, L and RD at time t by $P_I(t)$, $P_L(t)$ and $P_{RD}(t)$ respectively, we have

$\Omega = [I, L, RD]$,

so Ω is a 1*3 vector of Ω_i terms representing the state i at time t. Hence, states 1, 2, 3 representing I, L and RD respectively, and then $P = \left[P_I(t) \quad P_L(t) \quad P_{RD}(t)\right]$.

The transition rate matrix Q, which consists of all transition rates between states, is represented as:

$$Q = \begin{pmatrix} -(l+p) & l & p \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Equation 5.1 then becomes:

$$\begin{pmatrix} \frac{d}{dt}P_I(t) & \frac{d}{dt}P_L(t) & \frac{d}{dt}P_{RD}(t) \end{pmatrix} = \begin{pmatrix} P_I(t) & P_L(t) & P_{RD}(t) \end{pmatrix} \begin{pmatrix} -(l+p) & l & p \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \Rightarrow$$

$$\begin{pmatrix} \frac{d}{dt}P_I(t) & \frac{d}{dt}P_L(t) & \frac{d}{dt}P_{RD}(t) \end{pmatrix} = \begin{pmatrix} -(l+p)\ P_I(t) & l\,P_I(t) & p\,P_I(t) \end{pmatrix}.$$

The transitions from state I is represented as follows: flow in is equal to zero and flow out is equal to $(l+p)\,P_I(t)$, giving

$$\frac{d}{dt}P_I(t) = 0 - (l+p)\,P_I(t) \Rightarrow P_I(t) = k_1\, e^{-(l+p)t} = e^{-(l+p)t} \qquad (5.3)$$

where k_1 is a constant, which must equal 1 due to the fact that $P_I(0) = 1$, (since at time zero all individuals are in category I).

The transition rate from state I to state L is represented as follows:

The flow in is equal to $l\,P_I(t)$ and flow out is equal to zero, then using the expression from equation 5.3 for $P_I(t)$ we get $\frac{d}{dt}P_L(t) = l\,P_I(t) - 0 \Rightarrow$

$$P_L(t) = \frac{-l}{l+p} e^{-(l+p)t} + k_2. \qquad (5.4)$$

To find the value of k_2, consider time zero; i.e.

$t = 0 \Rightarrow P_L(0) = 0$. Thus

$$\frac{-l}{l+p} e^{-(l+p)*0} + k_2 = 0 \Rightarrow k_2 = \frac{l}{l+p}. \qquad (5.5)$$

Then substituting equation 5.5 into equation 5.4 we get equation 5.6 below:

$$P_L(t) = \frac{l}{l+p}(1 - e^{-(l+p)t}). \qquad (5.6)$$

Furthermore, the rate of moving from states I to state RD is as follows:

The flow in is equal to $p\,P_I(t)$ and flow out is equal to zero. Then substituting equation 5.3 into equation 5.7, we obtain

$$\frac{d}{dt} P_{RD}(t) = p\,P_I(t) - 0 \Rightarrow$$

$$P_{RD}(t) = \frac{-p}{l+p} e^{-(l+p)t} + k_3. \qquad (5.7)$$

The same procedure as above is repeated to find the value of k_3 using time zero; i.e.

$t = 0 \Rightarrow P_{RD}(0) = 0$, i.e.

$$\frac{-l}{l+p} e^{-(l+p)*0} + k_3 = 0 \Rightarrow k_3 = \frac{p}{l+p}. \qquad (5.8)$$

Now substituting equation 5.8 into equation 5.7 we will have the following equation 5.9:

$$P_{RD}(t) = \frac{p}{l+p}(1 - e^{-(l+p)t}). \qquad (5.9)$$

Table 5.1 is the summary of all the above mentioned steps:

Table 5-1 Summary of all state transitions for model I Nanakaly data

State	In flow	Out flow	Equation for state probabilities	Probabilities values
I	0	$(l+p)\,P_I(t)$	$\frac{d}{dt}P_I(t) = 0 - (l+p)\,P_I(t)$	$P_I(t) = e^{-(l+p)t}$
L	$l\,P_I(t)$	0	$\frac{d}{dt}P_L(t) = l\,P_I(t) - 0$	$P_L(t) = \frac{l}{l+p}(1 - e^{-(l+p)t})$
RD	$p\,P_I(t)$	0	$\frac{d}{dt}P_{RD}(t) = p\,P_I(t) - 0$	$P_{RD}(t) = \frac{p}{l+p}(1 - e^{-(l+p)t})$

We denote by *l/p* the ratio of probabilities for an individual to be lost to the study or die, respectively, to account for the right-censoring in this population caused by the end of the study period. Thus we can use the number of recorded deaths to estimate the number of lost individuals provided we can estimate *l*/p, which we alternatively denote by α . Suppose that, as in the original survival plot, we consider the data without realizing that the category L exists. We can see from equations 5.6 and 5.9 that

$$\frac{P_L(t)}{P_{RD}(t)}=\frac{\frac{l}{l+p}(1-e^{-(l+p)t})}{\frac{p}{l+p}(1-e^{-(l+p)t})}=\frac{l}{p}=\alpha. \qquad (5.10)$$

We will choose a time τ sufficiently large that (essentially) all of the important events have occurred. As previously noted (see Figure 5.2), the numbers of deaths are very few after *t* = 1000; thus we consider $\tau = 1000$, Based on the original data we can estimate the total number of lost individuals by $\hat{l} = n_{1000}$ (where n_{1000} is the number of remaining individuals at time 1000), because after this time we have a very small number of deaths indicating a small number of remaining individuals in total and the total of recorded deaths is given as $\hat{p}=\sum_{t=0}^{1000} d_t$. This yields

$$\hat{\alpha}=\frac{\hat{l}}{\hat{p}}=\frac{n_\tau}{\sum_{t=0}^{n} d_t}=\frac{n_{1000}}{\sum_{t=0}^{1000} d_t}\ . \qquad (5.11)$$

For every recorded death, we have on average α lost individuals, so we lose α extra individuals for each death.

Denoting;

$$o=\frac{p}{l+p} \qquad (5.12)$$

we obtain

$$(l+p)o=p \Rightarrow \frac{l}{p}=\frac{1}{o}-1\cdot \qquad (5.13)$$

Using equation 5.12 we get

$$\hat{o}=\frac{\sum_{t=0}^{n} d_t}{n_\tau+\sum_{t=0}^{n} d_t} \qquad (5.14)$$

where $\hat{o}$ is the proportion of observed deaths in the study and an estimate of o from the data. Let $\tilde{n}_t$ represent the estimated remaining population size at time *t*; if *t* = 0 then

$\tilde{n}_1 = n_1$; because at that time we have neither censored nor dead individuals.

Then $\tilde{n}_t$ can be expressed as:

$$n_{t+1} = n_t - d_t - c_t \ , \qquad (5.15)$$

$$\tilde{n}_t = n_t - (\frac{1}{\hat{o}} - 1)\sum_{i=1}^{t-1} d_i \, , \qquad (5.16)$$

where c_t represents the number of censored patients. By substituting equation 5.15 into equation 5.16, considering equation 5.16 at times t and t+1, we get

$$\tilde{n}_{t+1} = n_t - d_t - c_t - (\frac{1}{\hat{o}} - 1)\sum_{i=1}^{t} d_i \ \Rightarrow$$

$$\tilde{n}_{t+1} = n_t - (\frac{1}{\hat{o}} - 1)\sum_{i=1}^{t-1} d_i - (\frac{1}{\hat{o}} - 1)d_t - d_t - c_t \Rightarrow$$

$$\tilde{n}_{t+1} = \tilde{n}_t - \frac{d_t}{\hat{o}} - c_t \, . \qquad (5.17)$$

Then we adjust our estimates of the hazard and survival functions (and denote these using the subscript a) from chapter 3 equations 3.8, 3.9 and 3.10 to take account our estimates of the true number of individuals at risk to get

$$\hat{h}_a(t) = \frac{d_t}{\tilde{n}_t} \ , \qquad (5.18)$$

$$\hat{s}_a(t) = 1 - \hat{h}_a(t) \ , \qquad (5.19)$$

$$\hat{S}_a(t) = \prod_{k=0}^{t-1} \hat{s}(k) \ \cdot \qquad (5.20)$$

For our data for $\tau = 1000$, $\sum_{t=0}^{i-1} d_t$ = 240 and n_τ =232, which gives $\hat{o}$ = 0.50850 as the proportion of observed deaths before the end of the period, and from equation 5.11 give $\hat{\alpha} = 0.96670$

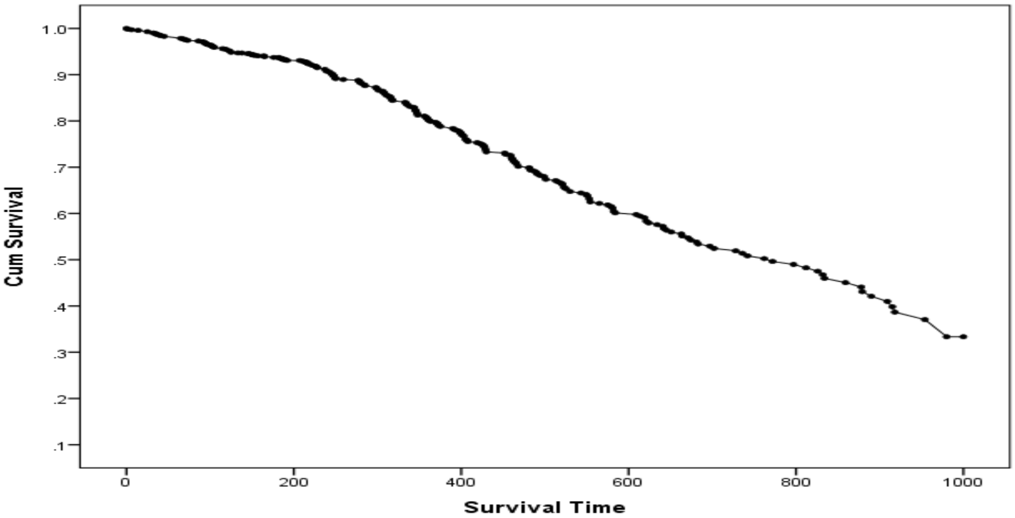


Figure 5-4 Adjusted survival curve for the Nanakaly data using the method without censoring

Our method applied without censoring gives the adjusted survival curve in Figure 5.4. This figure says that the cumulative survival probability up to 245 days is 0.902 and up to 315 days is 0.852; between these periods there were 20 deaths and 24 censored patients. However, between 363 to 481 days, which has cumulative survival probability 0.799 and 0.699 at the start and end of the interval respectively, 44 patients died and 23 patients survived. Finally, the values of the cumulative survival functions at 772 and 1000 days were 0.496 and 0.334 respectively.

Figures, 5.5 and 5.6 illustrate the adjusted hazard function (without censoring) and their smoothing based on a five days average for the Nanakaly data. The spikes in the first figure are due to the discrete nature of the hazard function, taking distinct values at every point. Averaging over a period of time as in equation 5.18 yields a scaled hazard function and a smoothed graph as seen in the second figure.

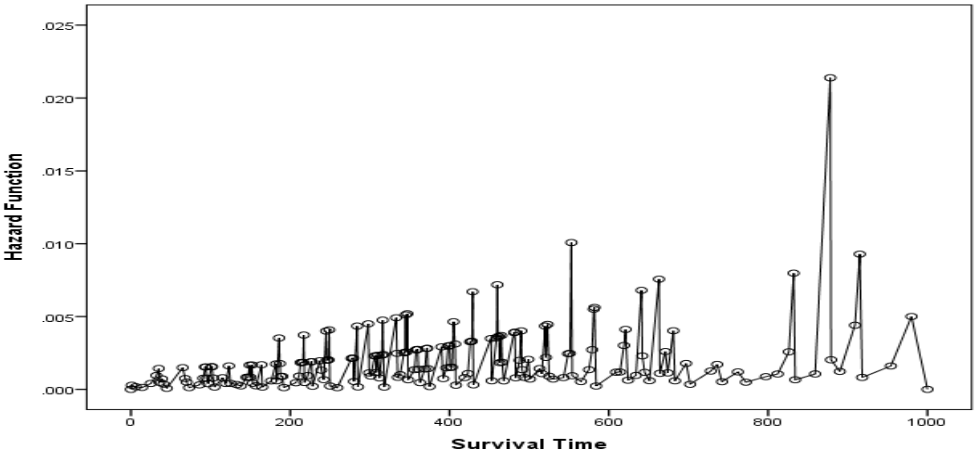


Figure 5-5 Adjusted hazard function curve for the Nanakaly data using the method without censoring

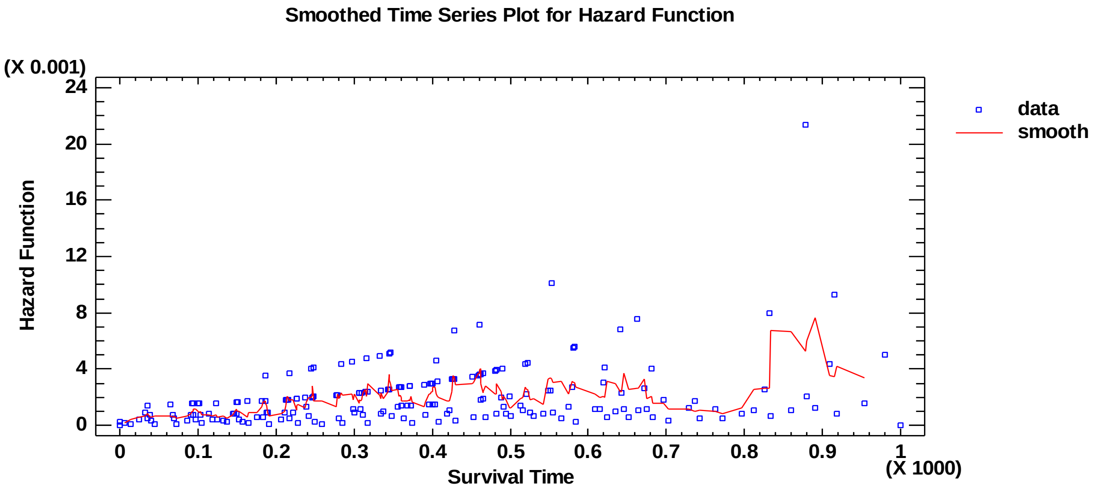


Figure 5-6 Adjusted smoothed hazard function curve for the Nanakaly data using the method without censoring

5.1.2 Markov model with censoring (Nanakaly data)

More generally we would like to allow for observed censoring as well as hidden censoring within our model. Thus we now add an extra “censored” category C to our model, where individuals move from I to C at rate q. Importantly, individuals also move from the lost category L to C at the same rate q. This is clearly appropriate for our dataset, since the only overt censoring is due to the end of the study, and thus any individual will reach this at the same time, whether in category I or L. We thus now have a population as described by Figure 5.7. We note that for individuals censored because we know that they have dropped out of the study prior to the end time, it would seem reasonable to assume that these and the “lost” individuals would be entirely separate, and so that the transition rate *q* from state L to state C would be absent.

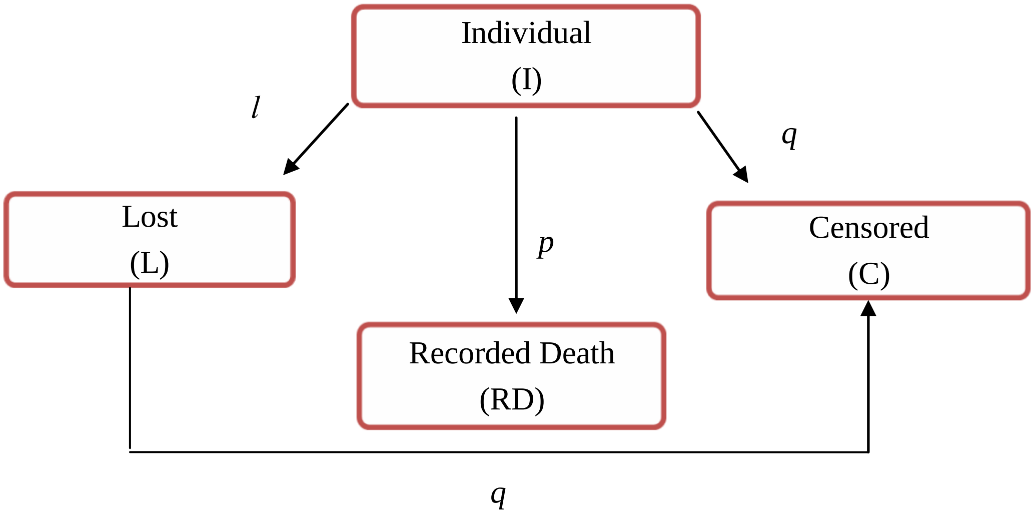


Figure 5-7 The markov survival model with censoring

The transition rate matrix Q, which consists of all the transition rates between states, is represented as follows:

$$Q = \begin{pmatrix} -(l+p+q) & l & q & p \\ 0 & -q & q & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Equation 5.1 now becomes

$$\begin{pmatrix} \frac{d}{dt}P_I(t) & \frac{d}{dt}P_L(t) & \frac{d}{dt}P_C(t) & \frac{d}{dt}P_{RD}(t) \end{pmatrix} = \begin{pmatrix} P_I(t) & P_L(t) & P_C(t) & P_{RD}(t) \end{pmatrix} \begin{pmatrix} -(l+p+q) & l & q & p \\ 0 & -q & q & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \Rightarrow$$

$$\begin{pmatrix} \frac{d}{dt}P_I(t) & \frac{d}{dt}P_L(t) & \frac{d}{dt}P_C(t) & \frac{d}{dt}P_{RD}(t) \end{pmatrix} = \begin{pmatrix} -(l+p+q)\,P_I(t) & l\,P_I(t) - q\,P_L(t) & q\,P_I(t) + q\,P_L(t) & p\,P_I\,(t) \end{pmatrix}\cdot$$

The transitions from state I is represented as follows: flow in is equal to zero and flow out is equal to $(l+p+q)\,P_I$, giving

$$\frac{d}{dt}P_I(t) = 0 - (l+p+q)\,P_I(t) \Rightarrow P_I(t) = k_1\,e^{-(l+p+q)t} = e^{-(l+p+q)t} \qquad (5.21)$$

where k_1 is a constant, which again must equal 1 due to $P_I(0) = 1$, since at time zero all individuals are in category I.

The transitions from state I to state L are represented as follows:
The flow in is equal to $l\,P_I(t)$ and flow out is equal to $q\,P_L(t)$, then using the expression from equation 5.21 for $P_I(t)$ we get

$$\frac{d}{dt}P_L(t) = l\,P_I(t) - q\,P_L\,(t) \Rightarrow \frac{d}{dt}P_L(t) + q\,P_L(t) = l\,P_I\,(t)\cdot$$

Multiplying both side of this equation by e^{qt} we get

$$(\frac{d}{dt}P_L(t) + q\,P_L(t))e^{qt} = l\,P_I(t)\,e^{qt}\cdot$$

Noting that following the product rule,

$$\frac{d}{dt}(\,f(t)\,g(t)) = f(t)\frac{d}{dt}g(t) + g(t)\frac{d}{dt}\,f(t)\;,$$

setting $f(t) = e^{qt}$ and $g(t) = P_L(t)$ we obtain

$$\frac{d}{dt}(e^{qt}\ P_L(t)) = e^{qt}\ \frac{d}{dt}P_L(t) + P_L(t)\frac{d}{dt}e^{qt} \Rightarrow$$

$$\frac{d}{dt}(e^{qt}\ P_L(t)) = (\frac{d}{dt}P_L(t) + q\,P_L(t))e^{qt}\ \cdot$$

Using the Integrating factor $e^{\int qdt} = e^{qt}$ we have

$$\frac{d}{dt}(e^{qt}\ P_L(t)) = l\,P_I(t)\,e^{qt}\ = l\,e^{-(l+p)t}\ \Rightarrow$$

$$e^{qt}\ P_L(t) = \int l\,e^{-(l+p)t}\ dt = k_2 - \frac{l}{l+p}\ e^{-(l+p)t}. \qquad (5.22)$$

To find the value of k_2 consider time zero; i.e.

$t = 0 \Rightarrow P_L(0) = 0$, so that

$$k_2 - \frac{l}{l+p}e^{-(l+p)0} = 0 \Rightarrow$$

$$k_2 = \frac{l}{l+p}\ . \qquad (5.23)$$

Then substituting equation 5.23 into equation 5.22 we get equation 5.24 below:

$$P_L(t) = e^{-qt}\ \frac{l}{l+p}(1 - e^{-(l+p)t}). \qquad (5.24)$$

Furthermore, the rate of moving from state I to state RD is as follows:

The flow in is equal to $p\,P_I(t)$ and flow out is equal to zero. Then substituting equation 5.21 into equation 5.26, we obtain

$$\frac{d}{dt}P_{RD}(t) = p\ \ P_I(t) - 0\ \Rightarrow \qquad (5.25)$$

$$P_{RD}(t) = \frac{-\,p}{l+p+q}e^{-(l+p+q)t} + k_3\ \cdot \qquad (5.26)$$

The same procedure above is repeated to find the value of k_3 by considering time 0;

$$t = 0 \Rightarrow P_{RD}(0) = 0 \Rightarrow \frac{-\,p}{l+p+q}e^{-(l+p+q)0} + k_3\ = 0 \Rightarrow$$

$$k_3\ = \frac{p}{l+p+q}. \qquad (5.27)$$

Now substituting equation 5.27 into equation 5.26 we will have the following equation 5.28:

$$P_{RD}(t) = \frac{p}{l+p+q}(1 - e^{-(l+p+q)t}). \qquad (5.28)$$

The rate of transition from state I to state C is as follows:

The value of flow in is equal to $q\,(P_I\,(t)+P_L\,(t))$ and flow out is equal to zero. Thus

$$\frac{d}{dt}P_C\,(t)=q(P_I\,(t)+P_L\,(t))-0\,.$$

In general; $P_I\,(t)+P_L\,(t)+P_{RD}(t)+P_C\,(t)=1$, and so

$$P_C\,(t)=1-P_L\,(t)-P_I\,(t)-P_{RD}(t). \qquad (5.29)$$

Then substituting equations 5.21, 5.24 and 5.28 into equation 5.29 we get:

$$P_C\,(t)=1-e^{-qt}\,\frac{l}{l+p}(1-e^{-(l+p)t})-e^{-(l+p+q)t}\;-\frac{p}{l+p+q}(1-e^{-(l+p+q)t})\,.$$

Rearranging the above equation we get equation 5.30, below:

$$P_C\,(t)=\frac{l+q}{l+p+q}(1-e^{-(l+p+q)t})-e^{-qt}\,\frac{l}{l+p}(1-e^{-(l+p)t})\;. \qquad (5.30)$$

Note that equation 5.24 is equivalent to

$$P_C\,(t)=-P_L\,(t)+\frac{q+l}{l+p+q}(1-e^{-(l+p+q)t}). \qquad (5.31)$$

Table 5.2 is the summary of the second model for all of the above mentioned steps.

Table 5-2 Summary of all state transitions for model II Nanakaly data

State	In flow	Out flow	Equations for state probabilities	Probabilities values
I	0	$(l+p+q)\;P_I(t)$	$\frac{d}{dt}P_I(t)=0-(l+p+q)P_I(t)$	$P_I\,(t)=e^{-(l+p+q)t}$
L	$l\,P_I\,(t)$	$q\,P_L\,(t)$	$\frac{d}{dt}P_L(t)=l\,P_I(t)-q\,P_L(t)$	$P_L\,(t)=e^{-qt}\,\frac{l}{l+p}(1-e^{-(l+p)t})$
C	$q(P_I\,(t)+P_L\,(t))$	0	$\frac{d}{dt}P_C\,(t)=q(P_I\,(t)+P_L\,(t))-0$	$P_C(t)=\frac{l+q}{l+p+q}(1-e^{-(l+p+q)t})-e^{-qt}\,\frac{l}{l+p}(1-e^{-(l+p)t})$
RD	$p\,P_I\,(t)$	0	$\frac{d}{dt}P_{RD}(t)=p\,P_I\,(t)-0$	$P_{RD}(t)=\frac{p}{l+p+q}(1-e^{-(l+p+q)t})$

Note that in our data we cannot observe which individuals are in state I and L separately, only their sum. We only observed in the study whether an individual had died, been censored or neither. The real death rate is equal to *p* while the apparent (i.e. the observed) death rate equals the following:

$$h(t)=\frac{P_I\,(t)}{P_I\,(t)+P_L\,(t)}\;\;p\Rightarrow$$

$$h(t) = \frac{p\,e^{-(l+p+q)t}}{e^{-(l+p+q)t} + \frac{l}{l+p}\,e^{-qt} - \frac{l}{l+p}\,e^{-(l+p+q)t}} = \frac{p\ e^{-(l+p+q)t}}{\frac{p}{p+l}\,e^{-(l+p+q)t} + \frac{l}{l+p}\,e^{-qt}} \ . \qquad (5.32)$$

Multiplying the top and bottom of the above equation by $(p+l)\ e^{(l+p+q)t}$, we get equation 5.33 below:

$$h(t) = \frac{(l+p)\ p}{p+l\ e^{(l+p)t}} \ . \qquad (5.33)$$

We first consider the apparent survival function, given by equation 5.34:

$$S(t) = e^{-\int_0^t h(u)du} \ . \qquad (5.34)$$

Integrating the adjusted hazard function in equation 5.34, we get;

$$\int_0^t h(u)du = \int_0^t \frac{(l+p)\,p}{p+l*e^{(l+p)u}}\,du \ , \qquad (5.35)$$

Using the substitution $v = e^{(l+p)u}$ we obtain

$$\int_0^t h(u)du = \int_1^{e^{(l+p)t}} \frac{p}{v}\ \frac{1}{p+l*v}dv \Rightarrow$$

$$[\int_1^{e^{(l+p)t}} (\frac{1}{v} - \frac{l}{p+l*v})\,dv] = [\ln v - \ln\,(p+l*v)]_1^{e^{(l+p)t}} \ \Rightarrow$$

$$[[\ln\,(\frac{v}{p+l*v})]_1^{e^{(l+p)t}}\ ] \Rightarrow \int_0^t h(u)\,du = \ln\,(\frac{e^{(l+p)t}}{p+l*e^{(l+p)t}}) - \ln\,(\frac{1}{p+l}) \Rightarrow$$

$$\int_0^t h(u)du = \exp[-\ln\,(\frac{(p+l)e^{(l+p)t}}{p+l*e^{(l+p)t}})]. \qquad (5.36)$$

Substituting equation 5.36 into equation 5.34 and rearranging we get the apparent survival function, equation 5.37 below:

$$S(t) = \frac{l+p\,e^{-(l+p)t}}{p+l}. \qquad (5.37)$$

When time t is equal to zero then the partial survival function equals 1 (a necessary condition for survival functions). On the other hand, when time t equals infinity then the apparent survival function takes the value $S(\infty) = l/(l+p)$. Recall that $\alpha = l/p$ from Section 5.1.1. For this case we do not use the previous estimate of ($\hat{\alpha}$) because there is overt censoring in this population caused by the end of the study period. This creates a potentially significant problem, because even the “lost” individuals are censored in this way, and so without adjustment the number of individuals at risk can be underestimated due to double counting (effectively the same individual being lost and then censored can be removed twice). This in turn leads to a lower estimate of $\hat{\alpha}$ than would otherwise be the case (in the alternative model below we shall see a different, higher, estimate of $\tilde{\alpha}$). In general when overt censoring occurs, $\hat{\alpha}$ will be smaller than $\tilde{\alpha}$ because the first model neglects the influence of censoring in the estimation procedure.

We will use the following method to find $\tilde{\alpha}$. Multiplying top and bottom in equation 5.37 by *p* and substituting *l*/p by α we obtain

$$S(t) = \frac{\alpha + (e^{-pt})^{1+\alpha}}{1+\alpha} = \frac{\alpha + (S_c(t))^{1+\alpha}}{1+\alpha}, \qquad (5.38)$$

where $S_c(t) = e^{-pt}$ represents the real survival function (since the hazard rate for all I individuals is *p*). We can then find the true survival function as a function of the apparent survival function,

$$S_c(t) = [(1+\tilde{\alpha})S(t) - \tilde{\alpha}]^{\frac{1}{1+\tilde{\alpha}}}. \qquad (5.39)$$

Figure 5.2 shows the apparent function $S(t)$ for our data. From this we obtain $\tilde{\alpha}/(1+\tilde{\alpha}) = 0.59688 \Rightarrow \tilde{\alpha} = 1.4807$. Thus using equation 5.39 we obtained $S_c(t)$ in Figure 5.8, below.

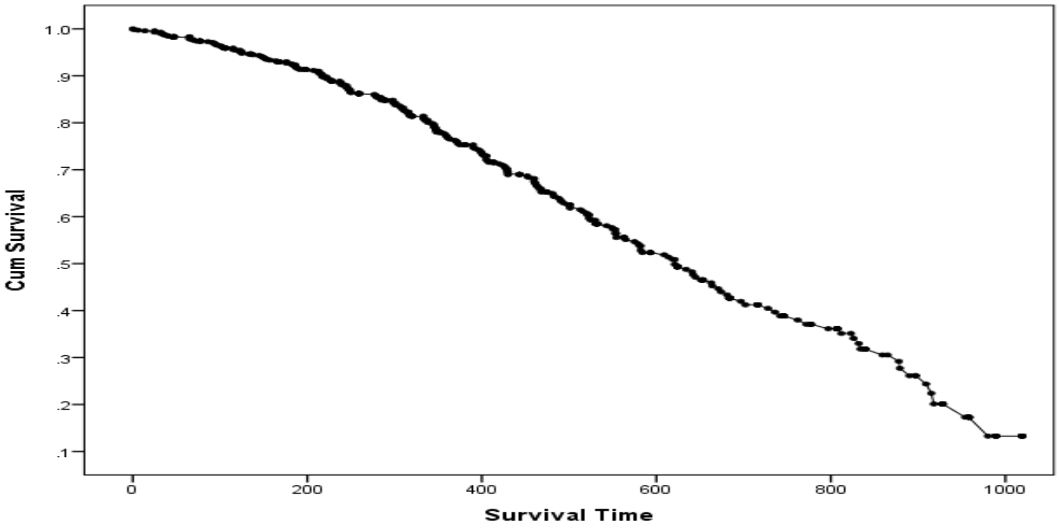


Figure 5-8 Adjusted survival curve for the Nanakaly data using the method with censoring

We can see that in Figures 5-4 and 5-8 that the two alternative survival curves generated by our methods now resemble classical survival curves for example that of the German data, which can be obtained from Hosmer et al. (2008) and which we discuss in chapter 6.

Comparing the two curves from Figures 5.4 and 5.8, we see that initially the two curves are roughly the same, but for later times, the curve in 5.4 is clearly above that in 5.8. We should also note that our methods are likely not to be very accurate near the end of the curves, i.e. when the last of their recorded deaths occur. Thus in the case of the Nanakaly data, the curves beyond about 1000 days are likely to be inaccurate; indeed they would be reliable for a considerably shorter time e.g. 500 days. We should also note that the above methodology might be applied to advance healthcare communication in various respects, specifically the collection of data, or for patients.

Adjusted hazard functions (with censoring) and their smoothing also based on a five day average for the Nanakaly data, are shown in Figures 5.9 and 5.10. As mentioned above, the sharp spikes in the first figure can be remedied by introducing a scaled hazard function to smoothen the plot.

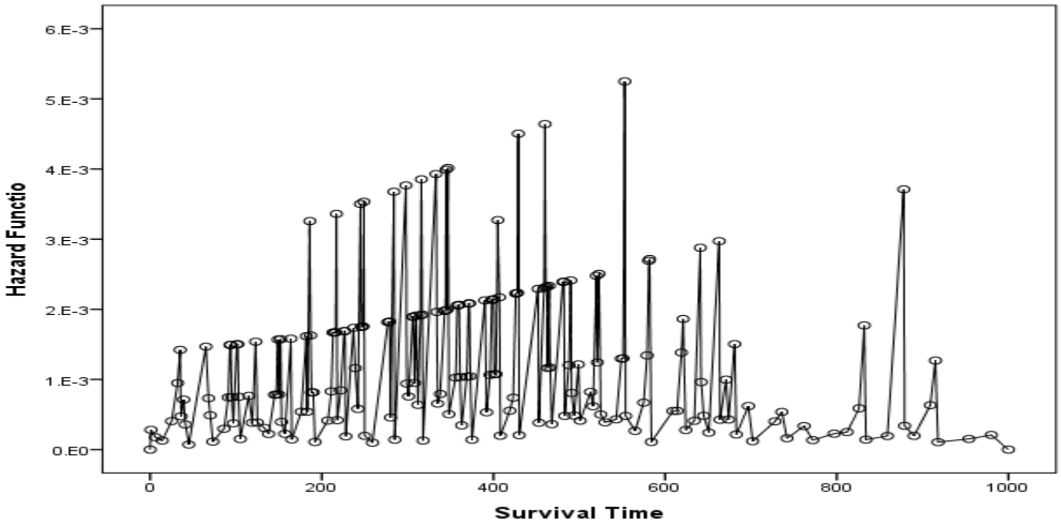


Figure 5-9 Adjusted hazard function curve for the Nanakaly data using the method with censoring

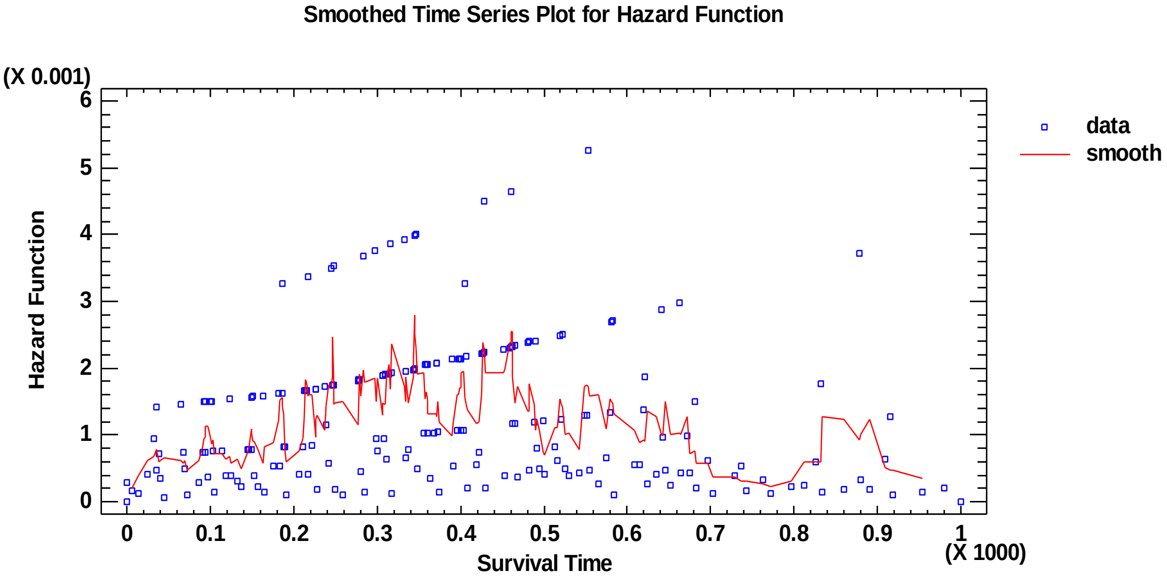


Figure 5-10 Adjusted smoothed hazard function curve for the Nanakaly data using the method with censoring

5.2 Application of survival analysis to Hewa Kurdish hospital data

In general to find a good model for Hewa Hospital data we began by plotting the survival curve using the Kaplan Meier method and it shows that, the curve is not reliable when compared to the Kaplan Meier curve for the German data model.

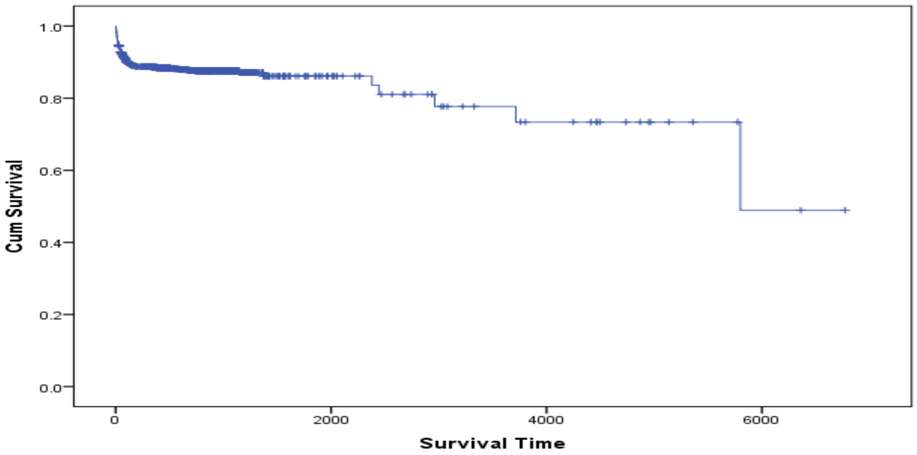


Figure 5-11 Survival curve including censoring for the Kurdish data from Hewa hospital

Figure 5.11 shows that the probability of death appears very small after 700 days as the curve flatters out around this time. As in the Nanakaly data, this is likely because only a small number of patients remain in the study after a this time. Since we see a small number of patients remaining for several thousand days, the data after this period was removed as shown in Figure 5.12.

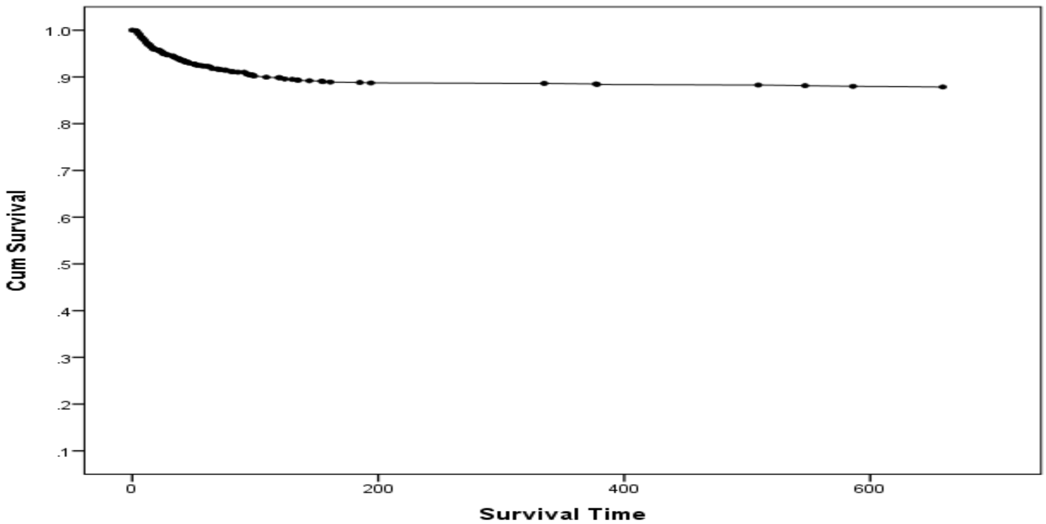


Figure 5-12 Survival curve including censoring for the Kurdish data from Hewa hospital data for 700 days

5.2.1 Markov model without censoring for Hewa data

A major issue for the Hewa data is the lack of recording of the true times of death as we discussed in Section 4.3. We addressed this problem through estimating these numbers by constructing two new models; each with and without censoring using Markov chains and estimating the number at risk $\tilde{n}$ and deaths $\tilde{d}$.

Figure 5.13, shows the first model for the Hewa data; here we have four stages, three of them are the same as for the first model of the Nanakaly data plus on extra stage from recorded death (where we use the admission time as a proxy for recorded death) to death z. Here the same conditions and protocol will be required for the continuous-time Markov chain as applies in the Nanakaly data Model I, but for the different sample space;

$\Omega = [I, L, RD, D]$.

The Markov Chain structure for Model I in the Hewa data is represented in Figure 5.13 below:

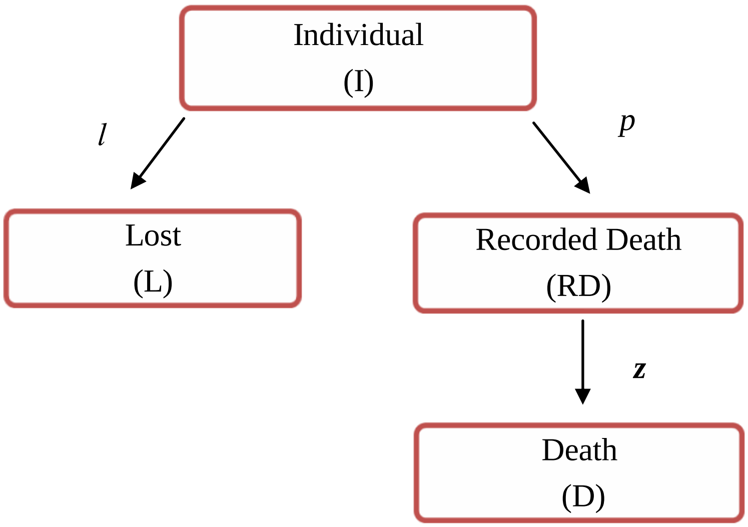


Figure 5-13 The markov survival model without censoring

The state probabilities at time t, are represented by the following below:

$P = \left(P_I(t) \quad P_L(t) \quad P_{RD}(t) \quad P_D(t)\right).$

The transition rate matrix Q is given by:

$$Q = \begin{pmatrix} -(l+p) & l & p & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -z & z \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

The derivations from equation 5.1 are represented by:

$$\left(\frac{d}{dt}P_I(t) \quad \frac{d}{dt}P_L(t) \quad \frac{d}{dt}P_{RD}(t) \quad \frac{d}{dt}P_D(t)\right) = \left(P_I(t) \quad P_L(t) \quad P_{RD}(t) \quad P_D(t)\right) \begin{pmatrix} -(l+p) & l & p & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -z & z \\ 0 & 0 & 0 & 0 \end{pmatrix} \Rightarrow$$

$$\left(\frac{d}{dt}P_I(t) \quad \frac{d}{dt}P_L(t) \quad \frac{d}{dt}P_{RD}(t) \quad \frac{d}{dt}P_D(t)\right) = \left(-(l+p)P_I(t) \quad l\,P_I(t) \quad p\,P_I(t) - z\,P_{RD}(t) \quad z\,P_{RD}(t)\right).$$

The transition rate out of state I is the same as in Model I for the Nanakaly data, i.e. as in equation 5.3.

For state I we thus have the same equation and initial conditions and so the same solution as for the Nanakaly model I, i.e. $P_I(t) = e^{-(l+p)t}$. Similarly the equation, initial condition and solution for L are identical to before, i.e. $P_L(t) = \frac{l}{l+p}(1 - e^{-(l+p)t})$.

However, for the Hewa model we add the rate of transition from state RD to state D of patients, given by z.

For state RD the flow in equals $p\,P_I(t)$ and flow out equals $z\,P_{RD}(t)$.

$$\frac{d}{dt}P_{RD}(t) = p\,P_I(t) - z\,P_{RD}(t) \Rightarrow$$

$$\frac{d}{dt}P_{RD}(t) + z\,P_{RD}(t) = p\,P_I(t),$$

then multiplying both sides by the integrating factor e^{zt} we get

$$(\frac{d}{dt}P_{RD}(t) + z\,P_{RD}(t))e^{zt} = p\,P_I(t)e^{zt} \Rightarrow$$

$$\frac{d}{dt}(e^{zt}\,P_{RD}(t)) = p\,P_I(t)e^{zt}$$

$$= p\,e^{zt}\,e^{-(l+p)t} \Rightarrow e^{zt}\,P_{RD}(t) = \int P\,e^{(z-l-p)t}dt \Rightarrow$$

$$e^{zt}\,P_{RD}(t) = \frac{p}{z-l-p}\,e^{(z-l-p)t} + k_5\,. \qquad (5.40)$$

Dividing both sides of equation 5.40 by e^{zt} we get the following equation 5.41.

$$P_{RD}(t) = k_5\,e^{-zt} + \frac{p}{z-l-p}\,e^{-(l+p)t} \qquad (5.41)$$

To find the value of constant k_5 we use the fact that at time zero $P_{RD}(0) = 0$ and so

$$k_5 + \frac{p}{z-l-p} = 0 \Rightarrow k_5 = \frac{p}{l+p-z}. \qquad (5.42)$$

Then substituting equation 5.42 in equation 5.41 we get equation 5.43:

$$P_{RD}(t) = \frac{p}{l+p-z}[e^{-zt} - e^{-(l+p)t}]. \qquad (5.43)$$

As already mentioned the transition rate from state RD to state D is z, and so the flow in to state D is equal to $z\,P_{RD}(t)$ and the flow out is equal to zero. Thus

$$\frac{d}{dt}P_D(t) = z\,P_{RD}(t) - 0\,.$$

By substituting equation 5.43 into the above formula we get

$\frac{d}{dt}P_D(t) = z\frac{p}{l+p-z}[e^{-zt} - e^{-(l+p)t}]$. If we let $w = \frac{p}{l+p-z}$ and then integrate we obtain equation 5.44 below:

$$P_D(t) = -we^{-zt} + \frac{zw}{l+p}e^{-(l+p)t} + k_6 \, . \qquad (5.44)$$

When $t = 0 \Rightarrow P_D(0) = 0$ then

$$k_6 = w - \frac{zw}{l+p}. \qquad (5.45)$$

Substituting equation 5.45 into equation 5.44 we get

$$P_D(t) = \frac{p}{l+p-z}[(1-e^{-zt}) - \frac{z}{(l+p)}(1-e^{-(l+p)t})]. \qquad (5.46)$$

The following Table 5.3 is the summary of all the above mentioned solutions:

Table 5-3 Summary of all state transitions for model I Hewa data

State	In flow	Out flow	Equations for state probabilities	Probabilities values
I	0	$(l+p)P_I(t)$	$\frac{d}{dt}P_I(t) = 0 - (l+p)P_I(t)$	$P_I(t) = e^{-(l+p)t}$
L	$l\,P_I(t)$	0	$\frac{d}{dt}P_L(t) = l\,P_I(t) - 0$	$P_L(t) = \frac{l}{l+p}(1-e^{-(l+p)t})$
RD	$p\,P_I(t)$	$z\,P_{RD}(t)$	$\frac{d}{dt}P_{RD}(t) = p\,P_I(t) - z\frac{p}{l+p-z}(e^{-zt} - e^{-(l+p)t})$	$P_{RD}(t) = \frac{p}{l+p-z}[e^{-zt} - e^{-(l+p)t}]$
D	$z\,P_{RD}(t)$	0	$\frac{d}{dt}P_D(t) = z\frac{p}{l+p-z}(e^{-zt} - e^{-(l+p)t}) - 0$	$P_D(t) = \frac{p}{l+p-z}[(1-e^{-zt}) - \frac{z}{(l+p)}(1-e^{-(l+p)t})]$

The results in Table 5.3 are helpful in setting up the second Hewa model (now including censoring), but these should be applied using the same mathematical formalism as for the first model (which does not account for censoring). Note that $P_I(t)$ and $P_L(t)$ are the same as for the first Nankaly model.

We can derive, from the Markov process (see Figure 5.13), estimates of the number of individuals moving from state I to RD and then D. In the calculation below we shall use the following terms:

$\tilde{d}_t$ is the estimated number of real deaths, while d_t is the number of recorded deaths, and x_t is the probability of death of an individual in the RD category, all within the *t th* time interval starting at T_t and ending at T_{t+1}.

In time interval T_t to T_{t+1}, d_t individuals move from I to RD, which are assumed to happen at the start of the interval. Individuals in RD can then move to D, at the start of the subsequent time interval, which they do with probability x_t .Thus the number of individuals remaining in state RD at the end of the time interval is $\sum_{i=1}^{t} d_i \prod_{j=i}^{i-1}(1-x_j)$ and, so to estimate the number of deaths for Hewa patients, the following equations have been used:

$$\tilde{d}_1 = 0 \;\; , \;\; \tilde{d}_2 = x_1 d_1 \;\; , \;\; \tilde{d}_3 = x_2 d_2 + x_2(1-x_1)d_1 \qquad (5.47a)$$

and

$$\tilde{d}_{t+1} = x_t \sum_{i=1}^{t} d_i \prod_{j=i}^{t-1}(1-x_j) \; , \qquad (5.47b)$$

where

$$x_t = 1-\exp(-z(T_{t+1}-T_t)) \, . \qquad (5.48)$$

Factoring out $\tilde{d}_t$ in equation (5.47b) one finds the recursive expression

$$\tilde{d}_{t+1} = x_t [d_t + \frac{1-x_{t-1}}{x_{t-1}} \tilde{d}_t]. \qquad (5.49)$$

z is the rate of death (recorded death to death) as shown in Figure 5.19. Thus the number of individuals dying in the *t th* period is the sum of all the probabilities of the death of individuals whose deaths were recorded before this. To estimate the number at risk ($\tilde{n}_{t+1}$) we use a similar method to ~~in~~ the first Nanakaly model and so need the value of $\hat{o}$ as in the following equation:

$$\hat{o} = \frac{\sum_{t=0}^{n_\tau} \tilde{d}_t}{n_\tau + \sum_{t=0}^{n_\tau} \tilde{d}_t} \; . \qquad (5.50)$$

The equation of the total number of patients n_τ is

$$n_\tau = \text{Number of individuals remaining in the study } + \sum_{t=0}^{n_\tau} d_t - \sum_{t=0}^{n_\tau} \tilde{d}_t \, . \qquad (5.51)$$

Recall that due to only a small number of events occurring later in the study, we cut off the data at time 659 just after one of the death events thus letting $\tau = 659$ $n_\tau = 569.63887$ and $\sum_{t=0}^{n_\tau} d_t = 137$. Using $z = 0.005$ we obtain $\sum_{t=0}^{n_\tau} \tilde{d}_t = 127.36113$ and $\hat{o} = 0.18273$. Thus

$$\tilde{n}_{t+1} = \tilde{n}_t - \frac{\tilde{d}_t}{\hat{o}} - c_t \; , \qquad (5.52)$$

The adjusted hazard function $\hat{h}_a(t)$ and the adjusted survival function $\hat{S}_a(t)$ are given by:

$$\hat{h}_a(t) = \frac{\tilde{d}_t}{\tilde{n}_t} \, , \qquad (5.53)$$

$$\hat{s}_a(t) = 1 - \hat{h}_a(t), \qquad (5.54)$$

$$\hat{S}_a(t) = \prod_{k=0}^{t-1} \hat{s}(k). \qquad (5.55)$$

Figure 5.14 shows the survival function, the Kaplan-Meier curve, based on the estimated data, supposing that z =0.005.

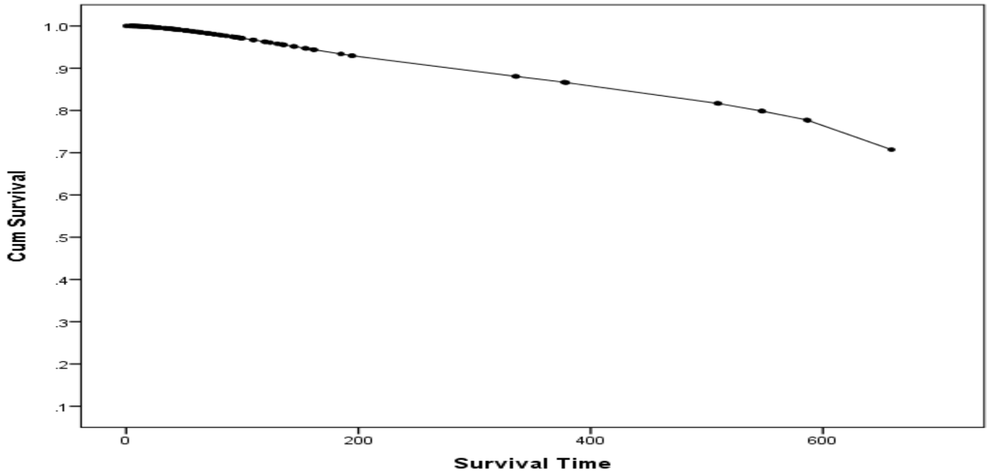


Figure 5-14 Adjusted survival curve for the Hewa data using the method without censoring (z=0.005)

From the Hewa data, the value obtained for $S(\infty)$ (the limiting value of the survival curve) is 0.87607, which gives an estimate of $\tilde{\alpha}$ = 7.07029 individuals lost per death event. Our method applied without censoring gives the adjusted survival curve in Figure 5.14. This figure says that the probability of surviving up to 50 days is 0.990 and up to 100 days is 0.971, during these periods (i.e. up to 100 days) there were 113 death and 51 censored patients. Finally, up to 700 days, there is probability of survival 0.707, where an additional 20 patients died and 383 patients were censored.

Figure 5.15 is the survival curve when z =0.05, which shows a different shape of survival curve for the same period compared to Figure 5.14 when z =0.005. For the first period of 50 days the survival probability is 0.941 and for the second period up to 100 days it is 0.889, while for the period up to 700 days the probability is 0.744.

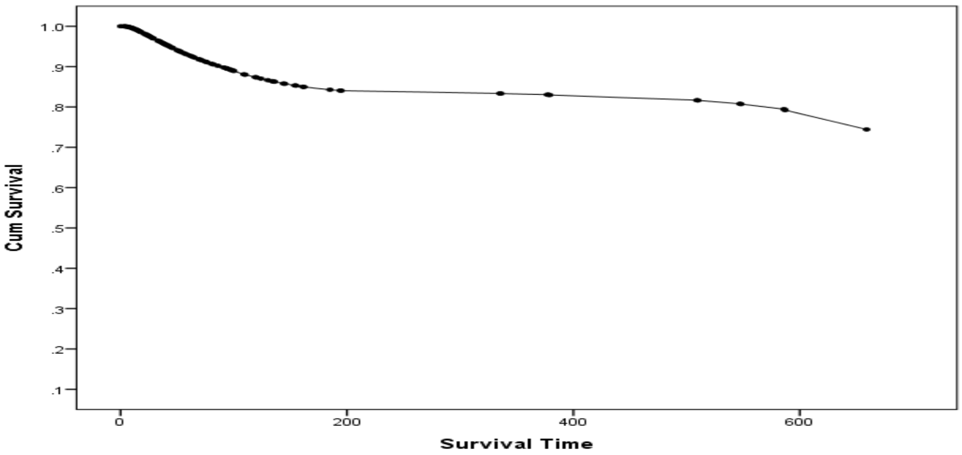


Figure 5-15 Adjusted survival curve for the Hewa data using the method without censoring (z=0.05)

The following Figure 5.16 when $z = 0.0005$ shows different survival probabilities again for the same period respectively. For up to 50 and up to 100 days the survival probabilities are 0.999 and 0.997 respectively, while the survival probability between up to 700 days is 0.929.

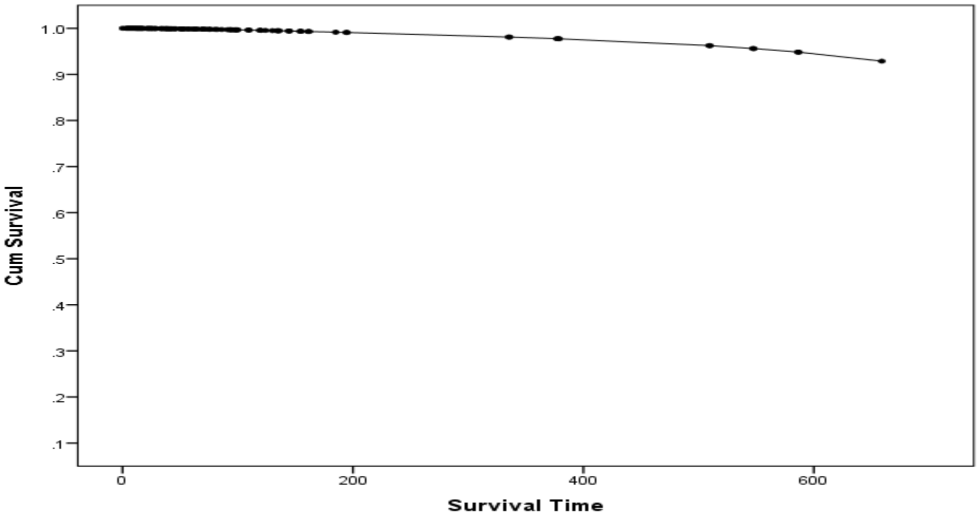


Figure 5-16 Adjusted survival curve for the Hewa data using the method without censoring (z=0.0005)

Since *z* depends upon an intuitive estimate using little evidence, we considered *z* to be 10 times bigger and 10 times smaller than 0.005. As above, we took the times of death to be given by the Markov process starting at the time of diagnosis which we discussed in Section 4.3.

Figures 5.17 and 5.18 demonstrate the adjusted hazard function (without censoring) and the smoothed hazard function based on a five days average for the Hewa data. Again, the spikes in Figure 5.17 may be removed by passing from the original to an appropriately scaled hazard function and thereby smoothing the graph, the scale of this smoothing is 5 days.

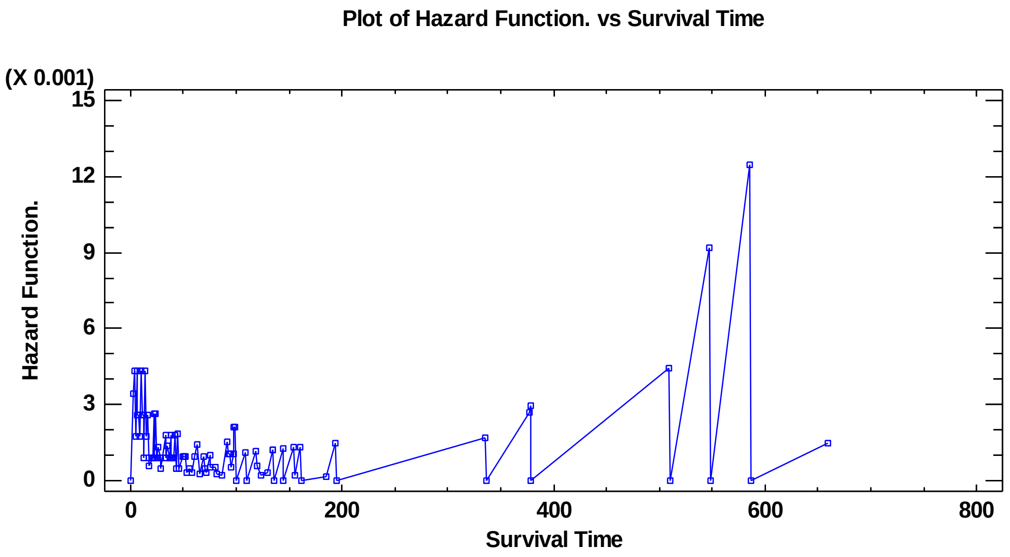


Figure 5-17 Adjusted hazard function curve for the Hewa data using the method without censoring

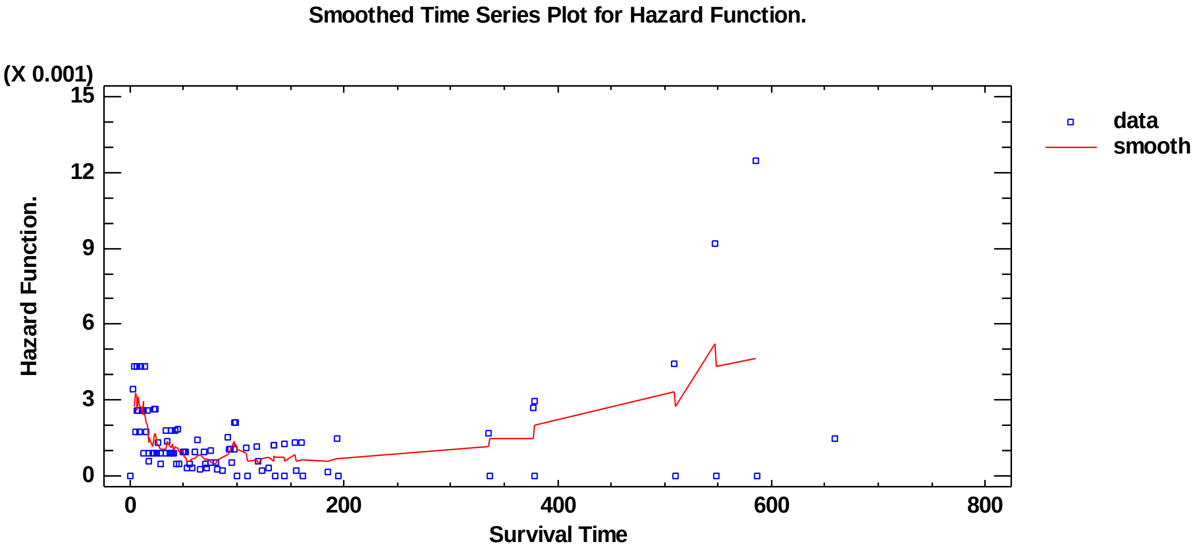


Figure 5-18 Adjusted smoothed hazard function curve for the Hewa data using the method without censoring

5.2.2 Markov model with censoring for Hewa data

The second Markov Chain structure model for the Hewa data is represented in Figure 5.19 below:

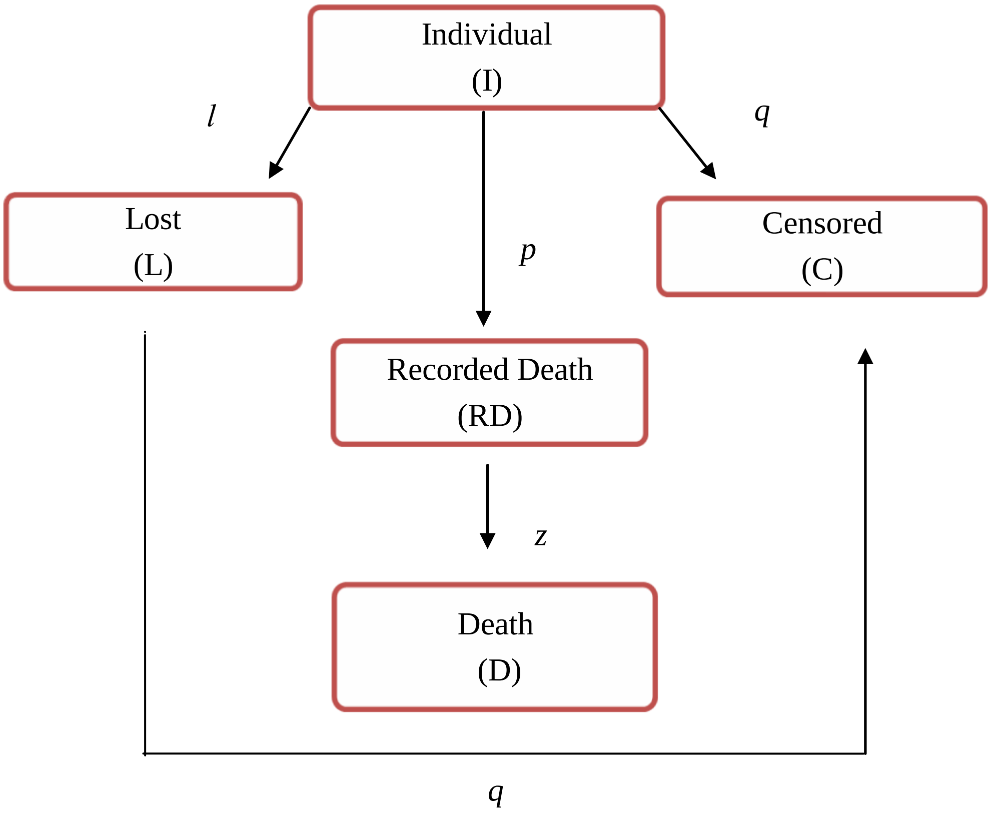


Figure 5-19 The markov survival model with censoring

The probabilities at time t of being in state I, L,C, RD and D respectively are represented by the following vector:

$$P = \left(P_I(t) \quad P_L(t) \quad P_C(t) \quad P_{RD}(t) \quad P_D(t) \right).$$

The transition rate matrix Q is given by

$$Q = \begin{pmatrix} -(l+q+p) & l & q & p & 0 \\ 0 & -q & q & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -z & z \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

The derivatives from equation 5.1 are represented by:

$$\left(\frac{d}{dt}P_I(t) \quad \frac{d}{dt}P_L(t) \quad \frac{d}{dt}P_C(t) \quad \frac{d}{dt}P_{RD}(t) \quad \frac{d}{dt}P_D(t) \right) = \left(P_I(t) \quad P_L(t) \quad P_C(t) \quad P_{RD}(t) \quad P_D(t) \right) \begin{pmatrix} -(l+q+p) & l & q & p & 0 \\ 0 & -q & q & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -z & z \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \Rightarrow$$

$$\left(\frac{d}{dt}P_I(t) \quad \frac{d}{dt}P_L(t) \quad \frac{d}{dt}P_C(t) \quad \frac{d}{dt}P_{RD}(t) \quad \frac{d}{dt}P_D(t)\right) = (-(l+q+p)P_I(t) \quad l\,P_I(t)-q\,P_L(t) \quad q\,(P_I(t)+P_L(t)) \quad p\,P_I(t)-z\,P_{RD}(t) \quad z\,P_{RD}(t)) \quad .$$

The transition rates out of the I state are the same as in the second model for the Nanakaly data, but for the Hewa model we add the rate of transition from state I to state RD (as opposed to state D) given by *p*. Thus the equation, initial condition and probability for state I are the same as before, i.e.

$$P_I(t) = e^{-(l+p)t}.$$

Similarly the equation, initial condition and probability for states L and C are identical to in the previous model and so

$$P_L(t) = \frac{l}{l+p}(1-e^{-(l+p)t}) \ ,$$

$$P_C(t) = \frac{l+q}{l+p+q}(1-e^{-(l+p+q)t}) - e^{-qt}\,\frac{l}{l+p}(1-e^{-(l+p)t}) .$$

For the state RD the flow in equals $p\,P_I(t)$ and the flow out is $z\,P_{RD}(t)$. Thus

$$\frac{d}{dt}P_{RD}(t) = p\,P_I(t) - z\,P_{RD}(t) \Rightarrow$$

$$\frac{d}{dt}P_{RD}(t) + z\,P_{RD}(t) = p\,P_I(t) .$$

Thus as in the previous Hewa model (see equation 5.43) we get

$$\frac{d}{dt}(e^{zt}\,P_{RD}(t)) = e^{zt}\,\frac{d}{dt}P_{RD}(t) + P_{RD}(t)\,ze^{zt} = p\,P_I(t)\,e^{zt} = p\,e^{zt}\,e^{-(l+q+p)t} \Rightarrow$$

$$e^{zt}\,P_{RD}(t) = \int p\,e^{(z-l-q-p)t}dt = \frac{p}{z-l-q-p}\,e^{(z-l-q-p)t} + k_5 . \qquad (5.56)$$

Dividing both side of equation 5.56 by e^{zt} we get

$$P_{RD}(t) = k_5\,e^{-zt} + \frac{P}{z-l-q-p}e^{-(l+q+p)t} . \qquad (5.57)$$

To find the value of constant k_5 we consider time zero, so that

$$k_5 + \frac{P}{z-l-q-p} = 0 \Rightarrow k_5 = \frac{p}{l+q+p-z}. \qquad (5.58)$$

Then substituting equation 5.57 into equation 5.58 we get equation 5.59:

$$P_{RD}(t) = \frac{p}{l+q+p-z}[e^{-zt} - e^{-(l+q+p)t}]. \qquad (5.59)$$

Now the last transition from state RD to state D has rate z. The value of flow in is equal to $z\,P_{RD}(t)$ and flow out is equal to zero. Thus

$$\frac{d}{dt}(P_D) = z\,P_{RD} - 0.$$

By substituting equation 5.59 into the above formula we get

$\frac{d}{dt}P_D(t) = z\frac{p}{l+q+p-z}[e^{-zt} - e^{-(l+q+p)t}]$. If we let $w = \frac{p}{l+q+p-z}$ and then integrate here we obtain equation 5.60 below:

$$P_D(t) = -we^{-zt} + \frac{zw}{l+q+p}e^{-(l+q+p)t} + k_6\,. \qquad (5.60)$$

When $t=0$ then $P_D(0)=0$ and thus

$$k_6 = w - \frac{zw}{l+q+p}. \qquad (5.61)$$

Substituting equation 5.61 into equation 5.60 we get

$$P_D(t) = \frac{p}{l+q+p-z}[(1-e^{-zt}) - \frac{z}{(l+q+p)}(1-e^{-(l+q+p)t})]. \qquad (5.62)$$

The following Table 5.4 is the summary of all the above mentioned steps:

Table 5-4 Summary of all state transitions for model II Hewa data

State	In flow	Out flow	Equations for state probabilities	Probability values
I	0	$(l+p+q)P_I(t)$	$\frac{d}{dt}P_I(t) = 0 - (l+p+q)P_I(t)$	$P_I(t) = e^{-(l+p)t}$
L	$l\,P_I(t)$	0	$\frac{d}{dt}P_L(t) = l\,P_I(t) - 0$	$P_L(t) = \frac{l}{l+p}(1-e^{-(l+p)t})$
C	$q\,P_I(t)$	0	$\frac{d}{dt}P_C(t) = q\,P_I(t) - 0$	$P_C(t) = \frac{l+q}{l+p+q}(1-e^{-(l+p+q)t}) - e^{-qt}\frac{l}{l+p}(1-e^{-(l+p)t})$
RD	$p\,P_I(t)$	$z\,P_{RD}(t)$	$\frac{d}{dt}P_{RD}(t) = p\,P_I(t) - z\frac{p}{l+q+p-z}(e^{-zt} - e^{-(l+q+p)t})$	$P_{RD}(t) = \frac{p}{l+q+p-z}[e^{-zt} - e^{-(l+q+p)t}]$
D	$z\,P_{RD}(t)$	0	$\frac{d}{dt}P_D(t) = z\frac{p}{l+q+p-z}(e^{-zt} - e^{-(l+q+p)t}) - 0$	$P_D(t) = \frac{p}{l+q+p-z}[(1-e^{-zt}) - \frac{z}{(l+q+p)}(1-e^{-(l+q+p)t})]$

Using equations 5.21 and 5.59 we will estimate the second model for Hewa data to determine the survival curve in the Hospital. We consider the estimated hazard function $\hat{h}_c(t)$, represented below:

$$\hat{h}_c(t) = \frac{z\,P_{RD}(t)}{P_{RD}(t) + P_I(t)} \ . \qquad (5.63)$$

This is the ratio of the rate of deaths in the population among the risk individuals (in categories RD and I) and the number of at risk individuals.

Firstly we use the real hazard function to find the real survival function as shown below:

$$\hat{h}_c(t) = \frac{z\,P_{RD}(t)}{P_{RD}(t) + P_I(t)} = \frac{z\,\dfrac{p}{l+p+q-z}\left[e^{-zt} - e^{-(l+p+q)t}\right]}{\dfrac{p}{l+p+q-z}\left[e^{-zt} - e^{-(l+p+q)t}\right] + e^{-(l+p+q)t}} . \qquad (5.64)$$

After multiplying top and bottom by $(l+p+q-z)\,e^{(l+p+q)t}$ the real hazard function takes the following form:

$$\hat{h}_c(t) = \frac{z\,p\left[e^{(l+p+q-z)t} - 1\right]}{l+q-z+p\,e^{(l+p+q-z)t}} . \qquad (5.65)$$

The real survival function is given by:

$$\hat{S}_c(t) = e^{-\int_0^t h_c(u)\,du}$$

where, using equation 5.65,

$$\int_0^t h_c(u)du = \int_0^t \frac{z\,p\left[e^{(l+p+q-z)u} - 1\right]}{l+q-z+p\,e^{(l+p+q-z)u}}du \, .$$

Letting $v = e^{(l+p+q-z)u}$ we have

$$\frac{dv}{du} = (l+p+q-z)\,e^{(l+p+q-z)u} \Rightarrow$$

$$\int_0^t h_c(u)du = \int_1^{e^{(l+p+q-z)t}} \frac{z\,p[v-1]}{(l+q-z+pv)}\,\frac{1}{v\,(l+p+q-z)}\,dv . \qquad (5.66)$$

Using partial fractions we have

$$\frac{z\,p[v-1]}{(l+q-z+pv)\,v} = \frac{A}{v} + \frac{B}{l+q-z+pv} \qquad (5.67)$$

$$\Rightarrow \frac{(l+q-z+pv)A + Bv}{v(l+q-z+pv)} = \frac{z\,p[v-1]}{(l+q-z+pv)v}$$

which yields

$$A = \frac{-z\,p}{l+q-z} \qquad (5.68)$$

and

$$B = \left(\frac{l+q-z+p}{l+q-z}\right) z\, p. \qquad (5.69)$$

Substituting A and B into equation 5.67 and equating it into equation 5.66, we get the following equation:

$$\int_{1}^{e^{(l+p+q-z)t}} \frac{z\,p[v-1]}{(l+q-z+pv)\,v\,(l+p+q-z)}dv = \int_{1}^{e^{(l+p+q-z)t}} \frac{1}{l+p+q-z}\left[\frac{-z\,p}{(l+q-z)\,v} + \frac{(l+p+q-z)\,z\,p}{(l+q-z)\,(l+q-z+pv)}\right]dv$$

$$= \frac{z\,p}{(l+p+q-z)\,(l+q-z)}\left[\int_{1}^{e^{(l+p+q-z)t}} -\frac{1}{v}dv + (l+p+q-z)\int_{1}^{e^{(l+p+q-z)t}} \frac{1}{l+q-z+pv}dv\right] =$$

$$\frac{z\,p}{(l+p+q-z)\,(l+q-z)}\left[-(l+p+q-z)t + \frac{(l+p+q-z)}{p}\Big(\ln(l+q-z+p\,e^{(l+p+q-z)t}) - \ln(l+q-z+p)\Big)\right].$$

To simplify the above equation we set $h = l+p+q-z$ and $w = e^{(l+p+q-z)t}$. This gives

$$\int_0^t h_c(u)du = \frac{z\,p}{h(h-p)}\left[-ht + \frac{h}{p}\big(\ln(h-p+pw) - \ln h\big)\right] =$$

$$-\frac{z}{h-p}pt + \frac{z}{h-p}[\ln(h-p+pw) - \ln h] \Rightarrow$$

$$-\int_0^t h_c(u)du = \frac{z}{h-p}pt + \frac{z}{h-p}\left[\ln\frac{h}{h-p+pw}\right]. \qquad (5.70)$$

Using equation 5.70 the real survival function $\hat{S}_c(t)$ takes the following form:

$$\hat{S}_c(t) = \exp\left[-\ln\left(\frac{p\,e^{(l+p+q-z)t} + l+q-z}{(l+p+q-z)\,e^{pt}}\right)^{\frac{z}{l+q-z}}\right] = \left[\frac{(l+p+q-z)\,e^{pt}}{p\,e^{(l+p+q-z)t} + l+q-z}\right]^{\frac{z}{l+q-z}}. \qquad (5.71)$$

5.2.3 Estimating the p, α, l , q and z values ;

We estimate the rate of recorded death p using the unadjusted survival function $S(t)$ as in the following equation

$$\hat{p} = \frac{1}{t}\ln\left(\frac{S(0)}{S(t)}\right) \qquad (5.72)$$

where;

p is the rate of recorded death, t is the time of death, $S(0)$ is the survival probability at time zero, ($S(0)=1$), $S(t)$ is the survival probability of an individual at time t , i.e. the probability of not entering category RD.

At low t, $S(t)\simeq e^{-pt}$, so that $\dfrac{S(0)}{S(t)}\simeq e^{pt}$ and $\dfrac{1}{t}\ln(\dfrac{S(0)}{S(t)})\simeq p$, leading to equation 5.72.

We tested different values of time t for equation 5.72. In principle the lower the t value, the better, except that for very small t, there is little data to use to estimate the value of p. There was sufficient data at time t =10, where there had been 29 recorded deaths. Here $S(t=10)=0.9790$ and so we obtain the following estimate of p,

$$\hat{p}=\frac{1}{10}\ln(\frac{1}{0.9790})\simeq 0.00212.$$

To estimate α we apply the same methods as in Section 5.1.2 using equation 5.38 as repeated below:

$$S(t)=\frac{\alpha+(e^{-pt})^{1+\alpha}}{1+\alpha}\ \Rightarrow S(\infty)=\frac{\alpha}{1+\alpha},$$

where $S(\infty)$ is the limiting apparent survival probability for the data. In practice, in contrast to the Nanakaly data, a small number of individuals remained in the study indefinitely, and so we had to choose a practical cut-off value. We selected the value associated with time t=750**,** which led to an estimated value of $\tilde{\alpha}=7.07029$. We selected this value because it was close to the equivalent cut off value from Figure 5.12 and it gives us an estimated survival curve very close to the survival curve from the real data except for when there are few individuals left in the study. We see this in Figures 5.20 and 5.21.

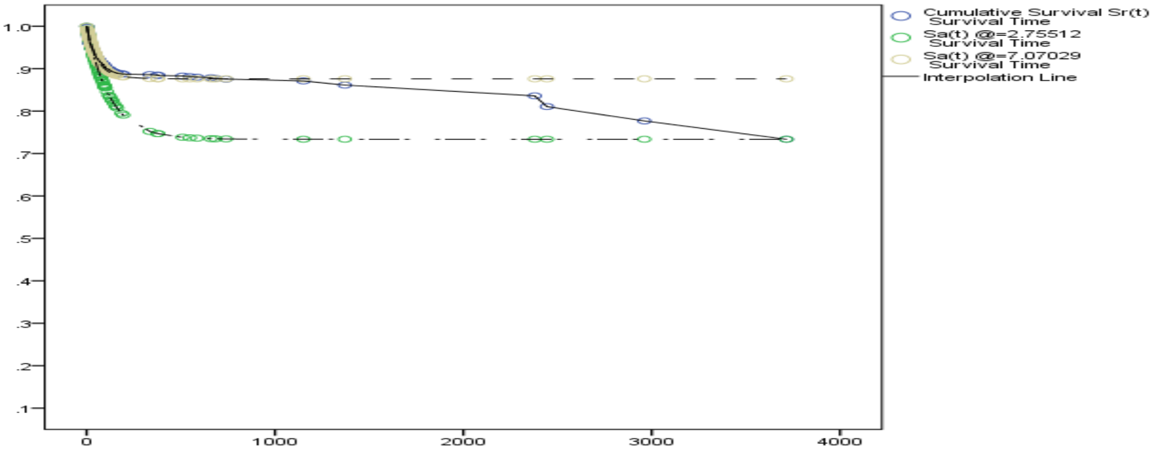


Figure 5-20 Kaplan Meier method for apparent and estimated survival functions at time 3715 days

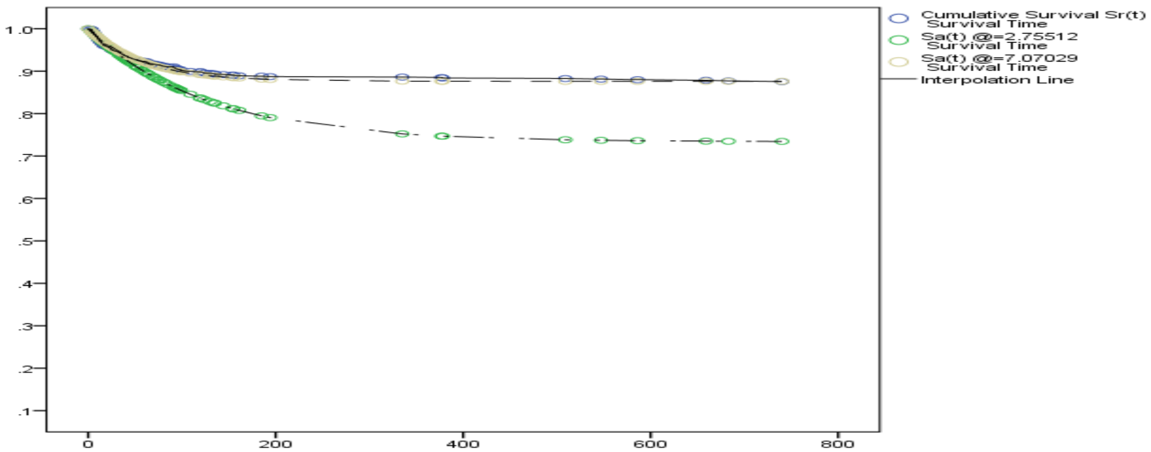


Figure 5-21 Kaplan Meier method for apparent and estimated survival function at time 700 days

To find the rate of loss of individuals, denoted by l, we use the definition of α ,

$\alpha = \frac{l}{p} \Rightarrow l = \alpha\, p$. Thus we have

$\hat{l} = 0.01499$.

The rate of censoring individuals q at time t can be estimated using equation 5.30, repeated below.

$$P_c(t) = \frac{l+q}{l+p+q}(1-\exp^{-(l+p+q)t}) - e^{-qt}\frac{l}{l+p}(1-e^{-(l+p)t} \ ,$$

where $P_c(t)$ is the proportion of censored individuals at time t.

After testing different value of t we conclude that a small value of t has not enough censored individuals. In practice censoring here is not a homogeneous Markov process (as implicit in the model); note for the models in Section 5.1 we saw that this was not important for that model, but here it is. A large number of censored individuals between t=700 to t=750 also made these values unreliable. A sensible choice is t= 1000 because the variations in $\hat{q}$ are larger below t= 1000. Then the estimated q value is $\hat{q} = 0.00160$. The discussion above is illustrated by Table 5.5.

Table 5-5 The estimated rate of censored individuals ($\hat{q}$) at different time (t)

t	200	500	700	750	875	**1000**	2000
$\hat{q}$	0.00093	0.00078	0.00085	0.00211	0.00182	**0.00160**	0.00159

The last component we need to estimate is the rate of recorded death to death z . For this case we depend on the previous z value of the first model without censoring, which was chosen to be equal to 0.005. We see this in Figure 5.22.

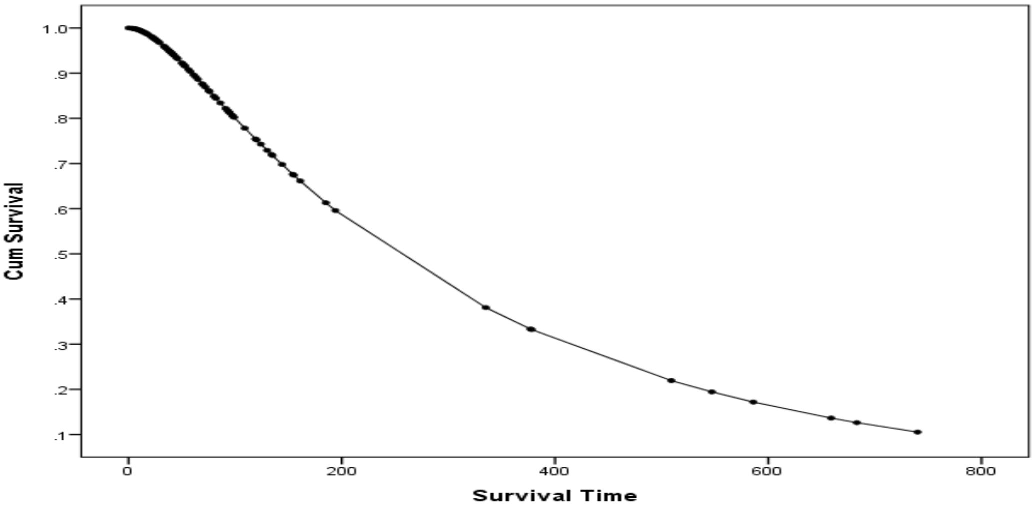


Figure 5-22 Adjusted survival curve for the Hewa data using the method with censoring (z=0.005)

Below we show some other figures for survival with the same estimated *p*, *l* and *q* with different *z* values. For the same reason as mentioned in the first model above, the following Figures 5.23 and 5.24 use *z* equal to 0.05 and 0.0005.

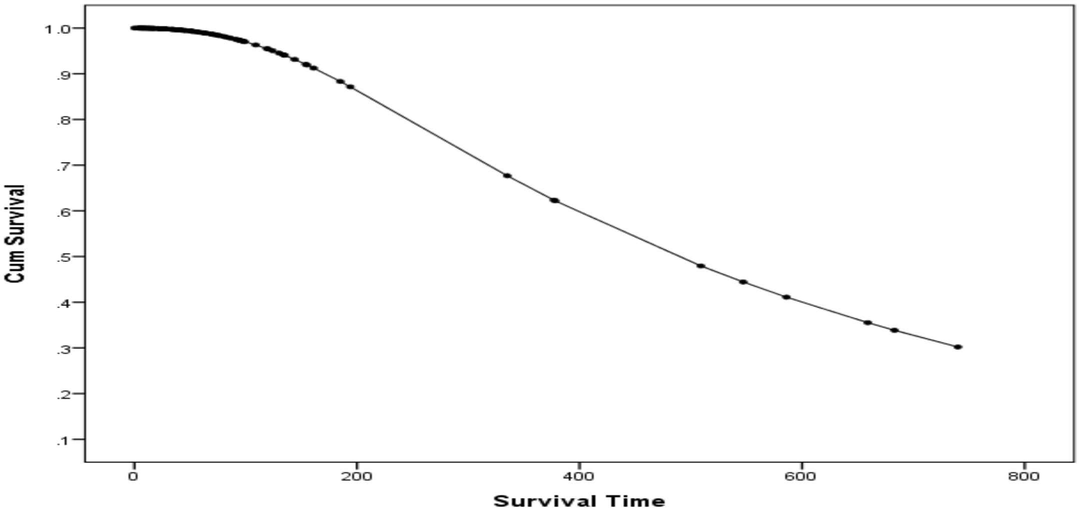


Figure 5-23 Adjusted survival curve for the Hewa data using the method with censoring (z =0.05)

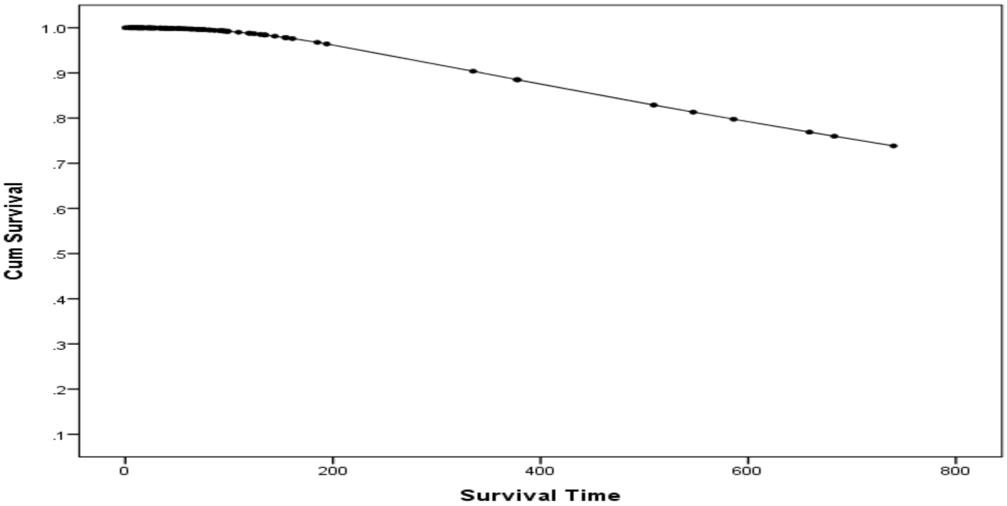


Figure 5-24 Adjusted survival curve for the Hewa data using the method with censoring (z=0.0005)

For the second model we used the Markov Chain to estimate the rate of recorded deaths, the rate of censoring, the rate of losing individuals and the rate of death (p, q, l and z). The survival curves arising from the two models are considerably different (see Figures 5.14 and 5.22), partly because of non-homogeneous censoring in the data. Observe that while the survival curve corresponding to z = 0.005 appears to be approaching zero, setting z = 0.05 yields a curve decreasing more slowly. Finally, for z = 0.0005, we obtain an even slower drop of the survival function, which does not appear to be approaching zero in the observed timeframe, casting doubt on the reliability of this estimate. To sum up, the time of death for which no data is available unlike for the time of diagnosis, has to be modelled by a Markov chain, leading to further uncertainties (see Section 4.3).

6 CHAPTER 6: Survival analysis for the breast cancer data

6.1 Survival analysis for the Nanakaly data

The next step is to determine major factors which impact breast cancer among women in the Kurdistan Region of Iraq. This is performed by finding the survival curves for the selected unadjusted variables, beginning with the use of the Kaplan Meier method and comparison of the variables by using tests including the log-rank test. For Nanakaly hospital, we have only one variable which is age. Applying Cox regression gives the results as show in Table 6.1. Age is highly statistically significant, with a p-value under 0.005.

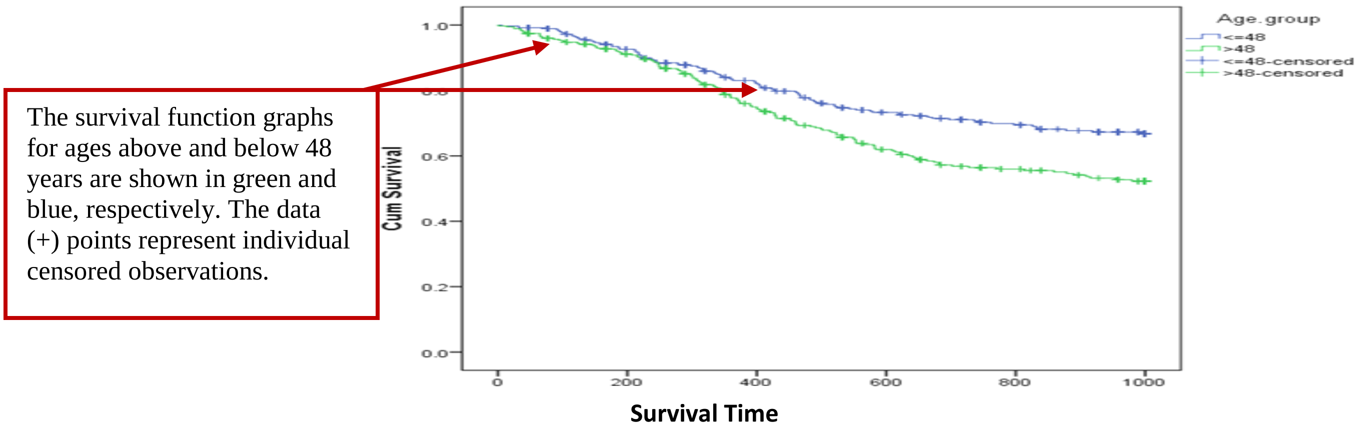
Table 6-1 Significant variable in the Cox regression model for Nanakaly data

	B	SE	Wald	Df	*p*. value	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
Age	.032	.005	33.258	1	.001	1.032	1.021	1.043

Based on the above table, the hazard ratio, $\phi(x) = \exp(\sum_{i=1}^{p} \beta_i x_i)$, can be written as the following equation:

$\phi(x) = \exp(0.032 \text{ Age})$.

The positive value of the hazard ratio indicates that there is a greater risk with higher age. For example $\exp(0.032 * 21.66) = 2$ means that an age difference of 22 years doubles the risk. To determine which age group has a longer survival time, the age variable between the individuals in the group is compared by dividing them into two groups based on their median, 48. The chi-square value when comparing two age groups, less and equal to the median and greater than it, is equal to 11.483 with p-value 0.001. Figure 6.1 and Figure 6.2 show the Kaplan Meier survival curve and hazard function curve for the patients for the different age groups. It is clear that the patients who are aged less than and equal to their median, 48, years old have more chance to survive from breast cancer disease than those of age greater than 48 years old.



The survival function graphs for ages above and below 48 years are shown in green and blue, respectively. The data (+) points represent individual censored observations.

Figure 6-1 Survival function curves (Kaplan Meier method) for the age variable (Nanakaly data)

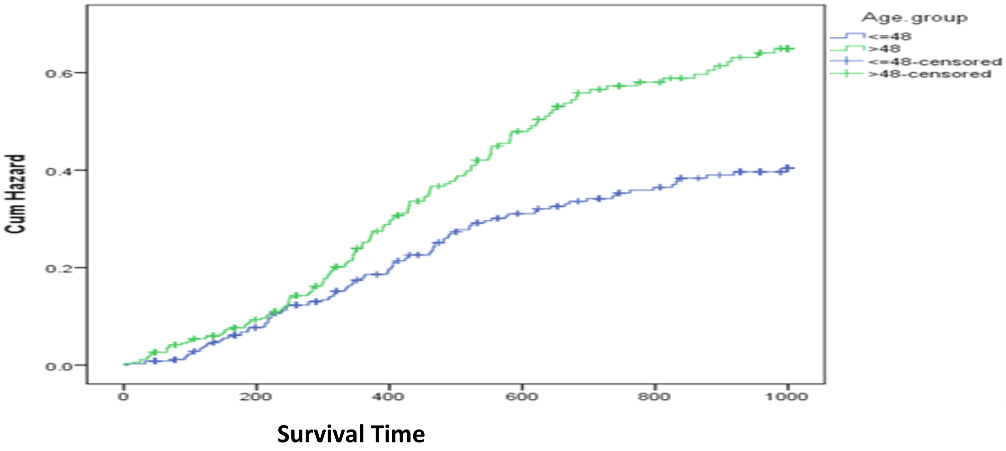


Figure 6-2 Hazard function curves for the age variable (Nanakaly data)

The survival function up to 200 days for the both age classes are equal to 0.9, while up to 400 days the survival curve for the age class greater than 48 years is 0.7 and less than equal to 48 it is 0.8. Correspondingly, in Figure 6.2, which shows the cumulative hazard function curve (h(t)= 1- cumulative survival functions), the case where up to 200 days is 0.1 for both age classes, while after 400 days the cumulative hazard for age class less than or equal to 48 is increased to 0.2, but for age class greater than 48 it is 0.3.

6.2 Survival analysis for Hewa data

As for the Nanakaly Hospital data, we use the SPSS program package to carry out the survival analysis for the unadjusted Hewa data. Then Cox-regression is used to determine the significant variables among the 20 used in the study. Based on the largest p-value, we delete sequentially the variables repeating the process until we get all significant variables with *p*-value less than 0.05. This is shown in the following Tables 6.2 and 6.3:

Table 6-2 All variables in the Cox regression model for Hewa data

	B	SE	Wald	df	*p*.value	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
Age	.001	.013	.007	1	**.935**	1.001	.976	1.027
Prog.recp	-.001	.001	3.442	1	.064	.999	.997	1.000
Estr.rec	-.002	.001	3.099	1	.078	.998	.995	1.000
Menopause	.264	.274	.931	1	.335	1.303	.761	2.229
Hormone	-.422	.238	3.130	1	.077	.656	.411	1.047
Tumour size	-.004	.006	.486	1	.486	.996	.985	1.007
Lymph nodes	-.012	.013	.912	1	.340	.988	.964	1.013
Religion.co			2.749	2	.253			
Religion.co(1)	.911	.817	1.244	1	.265	2.487	.502	12.331
Religion.co(2)	-.353	.425	.689	1	.407	.703	.305	1.617
Smoking.co	.486	.295	2.701	1	.100	1.625	.911	2.899
Drinking.co	1.183	.795	2.218	1	.136	3.266	.688	15.502
Weight	.002	.002	1.030	1	.310	1.002	.998	1.006
Height	-.009	.012	.499	1	.480	.991	.968	1.016
BMI	-.001	.002	.151	1	.698	.999	.995	1.003
Family.His.co	-.325	.312	1.080	1	.299	.723	.392	1.333
Occupa.co			5.550	11	.902			
Occupa.co(1)	6.461	65.274	.010	1	.921	639.891	.000	2331
Occupa.co(2)	.586	78.756	.000	1	.994	1.797	.000	1957
Occupa.co(3)	.333	76.014	.000	1	.997	1.396	.000	7044
Occupa.co(4)	8.071	65.282	.015	1	.902	3201.189	.000	1185
Occupa.co(5)	-.056	102.523	.000	1	1.000	.945	.000	1751
Occupa.co(6)	6.790	65.274	.011	1	.917	888.600	.000	3238
Occupa.co(7)	6.590	65.277	.010	1	.920	727.510	.000	2663
Occupa.co(8)	.188	91.636	.000	1	.998	1.206	.000	1208
Occupa.co(9)	6.030	65.275	.009	1	.926	415.647	.000	1516
Occupa.co(10)	.002	89.998	.000	1	1.000	1.002	.000	4047
Occupa.co(11)	7.139	65.282	.012	1	.913	1260.748	.000	4667
Family Income.co			4.527	3	.210			
Income.co(1)	-.687	.500	1.889	1	.169	.503	.189	1.340
Income.co(2)	-.390	.314	1.535	1	.215	.677	.366	1.254
Income.co(3)	-.515	.255	4.081	1	.043	.597	.362	.985
Martial.st.co			5.696	3	.127			
Martial.st.co(1)	.021	.491	.002	1	.965	1.021	.390	2.676
Martial.st.co(2)	.557	.239	5.449	1	.020	1.745	1.093	2.786
Martial.st.co(3)	.073	.452	.026	1	.872	1.075	.443	2.608
Exercise.co	-.065	.198	.109	1	.741	.937	.635	1.382
Breast.Fee.co	-.196	.266	.540	1	.462	.822	.488	1.386
Tumour grade			7.530	2	.023			
Tumour grade1	-.637	.309	4.259	1	.039	.529	.289	.968
Tumour grade2	-.078	.264	.088	1	.767	.925	.551	1.551

By applying Cox-regression we see that Age is the variable with the largest p-value. We thus remove it and repeat the analysis without the age code variable, getting Table 6.3. We can see from these table the occupation categories will be the next to be removed.

Table 6-3 Variables in the Cox regression model for Hewa data

	B	SE	Wald	df	*p*.value	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
Progesterone Receptor	-.001	.001	3.500	1	.061	.999	.997	1.000
Estrogen Receptor	-.002	.001	3.129	1	.077	.998	.996	1.000
Menopause	.254	.244	1.088	1	.297	1.289	.800	2.079
Hormone	-.426	.232	3.366	1	.067	.653	.414	1.030
Tumour Size	-.004	.006	.481	1	.488	.996	.985	1.007
Lymph Nodes	-.012	.012	.966	1	.326	.988	.964	1.012
Religion co.			2.761	2	.251			
Religion co. (1)	.910	.817	1.241	1	.265	2.484	.501	12.312
Religion co. (2)	-.348	.421	.684	1	.408	.706	.310	1.610
Smoking co.	.488	.294	2.746	1	.098	1.628	.915	2.899
Drinking co.	1.180	.793	2.211	1	.137	3.253	.687	15.399
Weight	.002	.002	1.027	1	.311	1.002	.998	1.006
Height	-.009	.012	.497	1	.481	.991	.968	1.016
BMI	-.001	.002	.151	1	.698	.999	.995	1.003
Family History co.	-.325	.312	1.083	1	.298	.723	.392	1.333
Occupation co.			5.568	11	**.901**			
Occupation co. (1)	6.467	65.275	.010	1	.921	643.772	.000	2350
Occupation co. (2)	.596	78.760	.000	1	.994	1.814	.000	1992
Occupation co. (3)	.338	76.022	.000	1	.996	1.402	.000	7194
Occupation co. (4)	8.065	65.283	.015	1	.902	3181.584	.000	1180
Occupation co. (5)	-.052	102.525	.000	1	1.000	.950	.000	1764
Occupation co. (6)	6.794	65.275	.011	1	.917	892.089	.000	3258
Occupation co. (7)	6.601	65.278	.010	1	.919	736.143	.000	2700
Occupation co. (8)	.181	91.637	.000	1	.998	1.199	.000	1204
Occupation co. (9)	6.034	65.276	.009	1	.926	417.488	.000	1526
Occupation co. (10)	.004	89.999	.000	1	1.000	1.004	.000	4067
Occupation co. (11)	7.145	65.283	.012	1	.913	1267.830	.000	4703
Family Income co.			4.537	3	.209			
Income co. (1)	-.687	.500	1.891	1	.169	.503	.189	1.339
Income co. (2)	-.390	.314	1.539	1	.215	.677	.365	1.254
Income co. (3)	-.516	.255	4.092	1	.043	.597	.362	.984
Marital Status co.			5.694	3	.127			
Marital Status co. (1)	.010	.472	.000	1	.983	1.010	.400	2.550
Marital Status co. (2)	.556	.238	5.447	1	.020	1.743	1.093	2.780
Marital Status (3)	.066	.445	.022	1	.882	1.068	.447	2.553
Exercise co.	-.065	.198	.107	1	.743	.937	.635	1.382
Breast Feeding co.	-.201	.259	.602	1	.438	.818	.493	1.358
Tumour Grade			7.675	2	.022			
Tumour Grade (1)	-.635	.307	4.277	1	.039	.530	.291	.967
Tumour Grade (2)	-.078	.264	.088	1	.767	.925	.551	1.551

The above steps are continued as illustrated in Appendix A Tables A1.1 to A1.16, until we get Table 6.4. which shows that there are three significant variables with p-value under 0.05.

p-value, hazard rates and CI for each variable

Table 6-4 Significant variables in the Cox regression model for Hewa data

	B	SE	Wald	df	*p*.value	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
Estrogen Receptor	-.003	.001	5.339	1	.021	.997	. 995	1.000
Smoking Code	.540	.197	7.490	1	.006	1.716	1.166	2.527
Tumour Grade			9.000	2	.011			
Tumour Grade(1)	-.578	.294	3.869	1	.049	.561	.315	.998
Tumour Grade(2)	.022	.258	.007	1	.933	1.022	.616	1.696

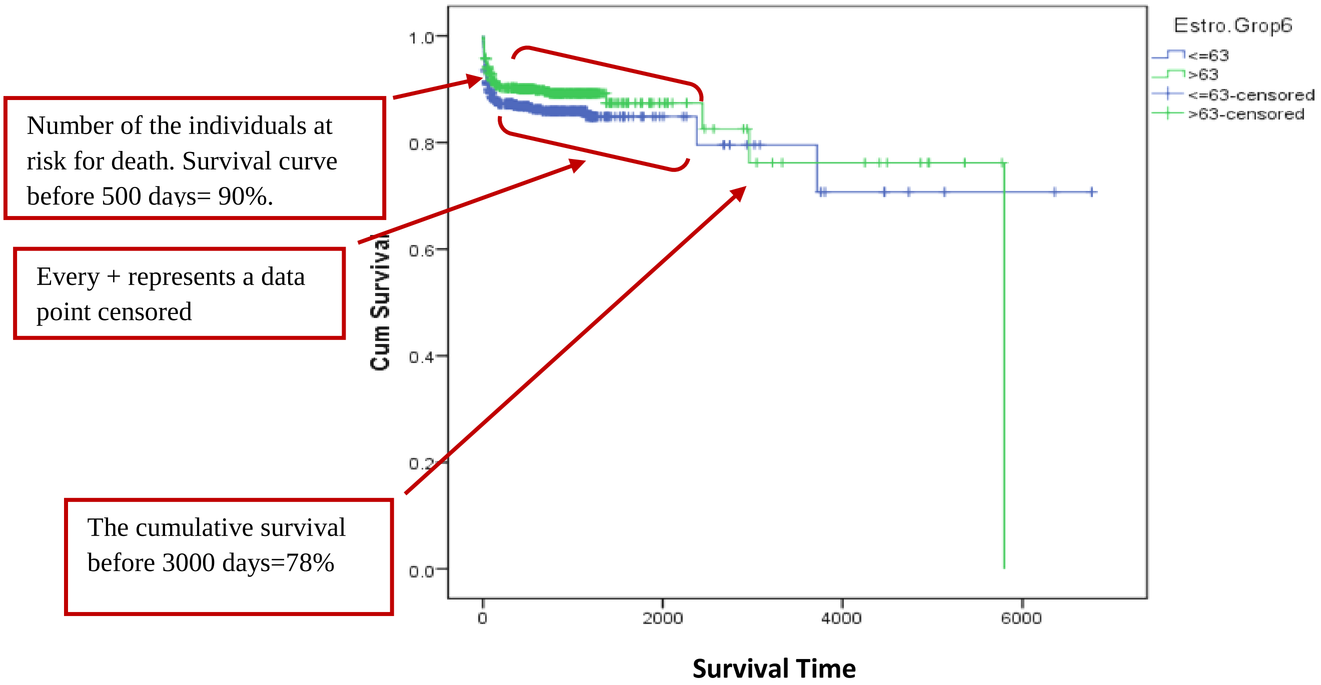
In the above table the hazard ratio, $\phi(x) = \exp(\sum_{i=1}^{p} \beta_i x_i)$, is represented as the following equation:

$$\phi(x) = \exp[(-0.003\, estogen\ receptor) + 0.540\, smoking + (-0.578\, tumour\ grade(1)) +$$
$$0.022\, tumour\ grade(2)] \cdot \qquad (6.1)$$

The estrogen is the main female hormone because it plays an important role in women’s menstrual cycle, sexual development, pregnancy, and childbirth. It can also cause cancer to grow. Here the value of estrogen receptor equals 0.997 and since it is less than one its influence appears to lower the risk of breast cancer. The hormonal therapy may help to slow or stop the growth of hormone receptor positive breast cancer by lowering the body’s estrogen levels or blocking the effects of estrogen.

There are two groups of estrogen receptors based on the median value such that in group 1 the estrogen receptor is less than or equal to 63, whilst for group 2 it is greater than 63. The difference between these two groups is not significant since the value of Chi-Square for the estrogen receptor equals 2.768 and the p-value equals to 0.096.

This conclusion is supported by the Kaplan-Meier survival curve as shown in Figure 6.3 where the survival curve for the group 2 is very close to that of group 1.


Number of the individuals at risk for death. Survival curve before 500 days= 90%.

Every + represents a data point censored

The cumulative survival before 3000 days=78%

Figure 6-3 Survival function curve (Kaplan Meier method) for estrogen receptor Hewa data

Figure 6.4 the classical cumulative hazard curve showing how it increases with ~~difference~~ time, while the estrogen receptor for the age less than or equal 63 which is group 1 is more risky than group 2.

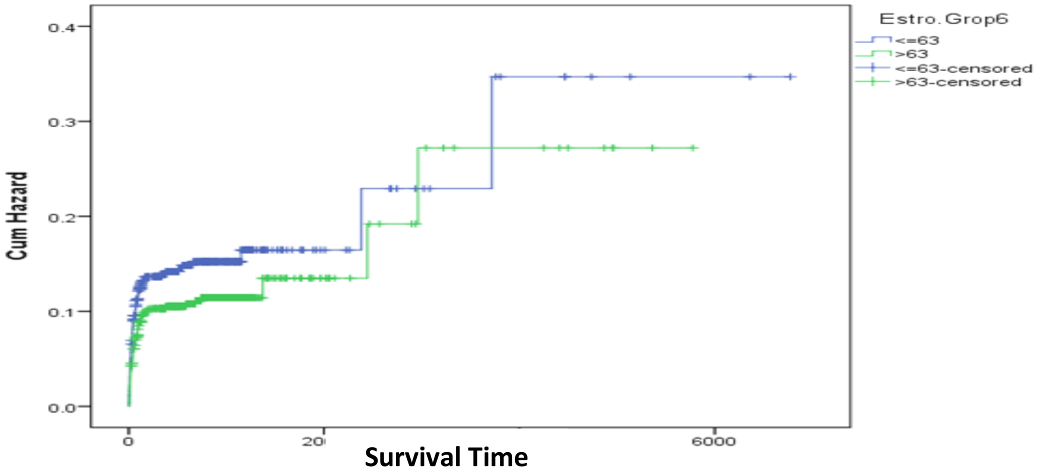


Figure 6-4 Hazard function for estrogen receptor Hewa data

Using the log-rank test, smoking is statistically significant because the Chi-Square is equal to 5.368 and giving a p-value less than 0.005. The survival curve illustrates that the non smokers have better opportunity to live longer than smokers, and the accoutres for smokers is more than for non smokers. See, Figure 6.5 and 6.6 below.

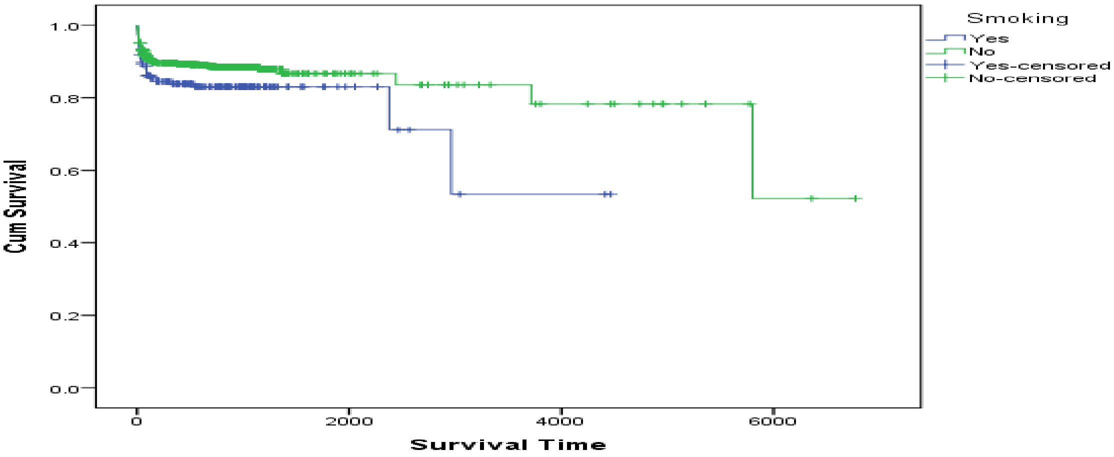


Figure 6-5 Survival function curve (Kaplan-Meier method) for smoking (Hewa data)

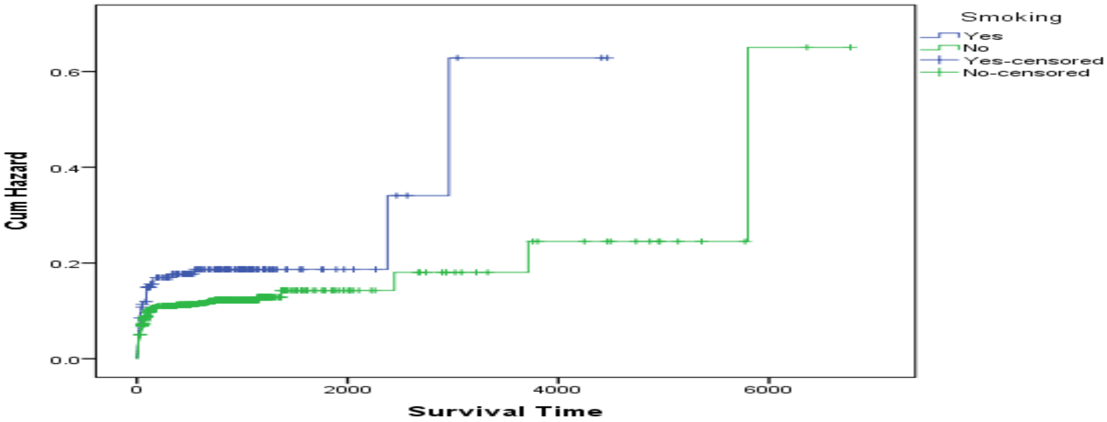


Figure 6-6 Hazard function for smoking (Hewa data)

In Table 6.4 above, tumour grade 1 has 0.561 the risk of tumour grade 3 so the death rate of tumour grade 1 is 0.561 times the death rate of tumour grade 3. Whilst for tumour grade 2 death rate is marginally larger than tumour grade 3 because its death rate is equal to 1.022.

Table 6.5, states the results of the log-rank test for the last significant variable in the study, which is the tumour grade. It shows that the difference between individuals where small and medium are stronger than that of small and large.

Table 6-5 log-rank test for the tumour grade variable for the Hewa data

Tumour Grade		Small(I)		Medium(II)		Large(III)	
		Chi-Square	*p*. value	Chi-Square	*p*. value	Chi-Square	*p*. value
Log Rank (Mantel Cox)	Small(I)			9.404	.002	3.467	.063
	Medium(II)	9.404	.002			.069	.793
	Large(III)	3.467	.063	.069	.793		

Figure 6.7 the survival curves, demonstrate that the patients with a small tumour grade have a better survival rate than those with medium and large tumours, while for the tumour grade medium the cumulative survival function is higher than that for the tumour grade large. Similarly in Figure 6.8, the cumulative hazard functions for the tumour grade medium and large are higher than tumour grade small.

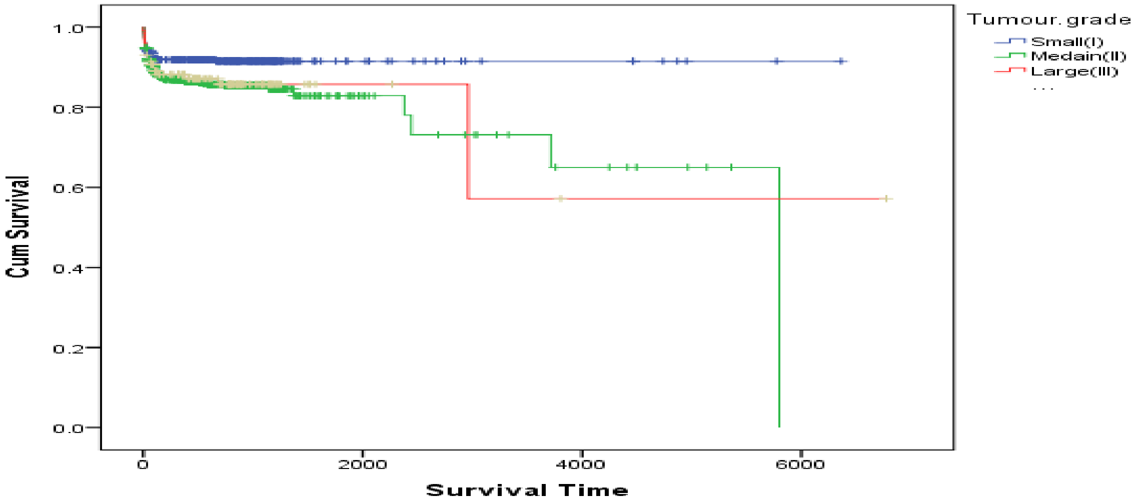


Figure 6-7 Survival function curve (Kaplan-Meier method) for tumour grade (Hewa data)

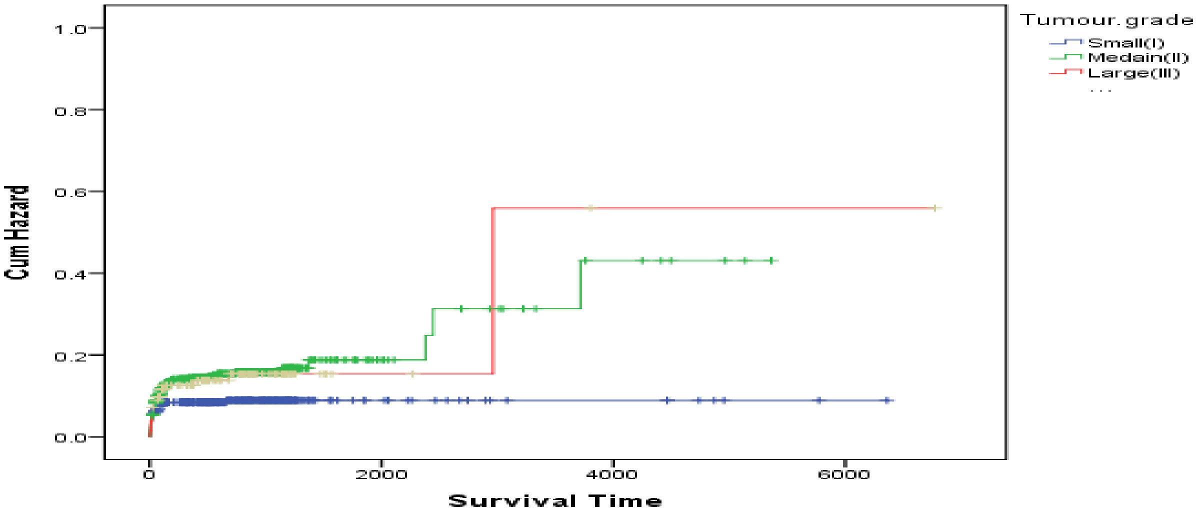


Figure 6-8 Hazard function curves for tumour grade (Hewa data)

6.3 Survival analysis for the German data

Following the analysis on breast cancer among women in the Kurdistan region of Iraq, we would like to assume survival analysis for the German data then compared our results.

First of all, the Cox regression equation is used to determine the significant variables among all eight variables used in the study. Similarly the analysis for the Kurdish data, Hewa and Nanakaly, we depend on the largest p-value and delete the variables each time and repeat the process until we get the significant variables with p-value less than 0.05. This is shown in the following Tables 6.6 and 6.7:

Table 6-6 All variables in the Cox regression model for German data

	B	SE	Wald	df	*p*.value	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
Age	.007	.012	.317	1	.573	1.007	.983	1.031
Menopause	-.091	.253	.130	1	**.718**	.913	.556	1.498
Hormone	.269	.169	2.543	1	.111	1.308	.940	1.821
Tumour Size	.013	.005	7.460	1	.006	1.013	1.004	1.023
Tumour Grade			8.501	2	.014			
Tumour Grade(1)	-1.127	.442	6.501	1	.011	.324	.136	.770
Tumour Grade(2)	-.352	.169	4.343	1	.037	.703	.505	.979
Lymph Nodes	.052	.010	30.239	1	.000	1.054	1.034	1.074
Progesterone Receptor	-.005	.001	20.188	1	.000	.995	.992	.997
Estrogen Receptor	.000	.001	.216	1	.642	1.000	.999	1.001

We see that Menopause has the highest p-value, and so remove it. By applying the Cox regression method again without the menopause variable, we get Table 6.8:

Table 6-7 Variables in the Cox regression model for German data

	B	SE	Wald	df	*p*.value	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
Age	.010	.008	1.663	1	.197	1.010	.995	1.026
Hormone	.264	.168	2.464	1	.116	1.302	.937	1.810
Tumour Size	.013	.005	7.410	1	.006	1.013	1.004	1.023
Tumour Grade			8.524	2	.014			
Tumour Grade(1)	-1.131	.442	6.550	1	.010	.323	.136	.767
Tumour Grade(2)	-.351	.169	4.319	1	.038	.704	.506	.980
Lymph Nodes	.052	.010	30.070	1	.000	1.054	1.034	1.074
Progesterone Receptor	-.005	.001	20.560	1	.000	.995	.992	.997
Estrogen Receptor	.000	.001	.224	1	**.636**	1.000	.999	1.001

The above steps are continued as shown in Tables A2.1 to A2.3 in the Appendix A2 until we get Table 6.8, which shows that there are four significant variables with p-values under 0.05.

Table 6-8 Significant variables in the Cox regression model for German data

	B	SE	Wald	df	*p*.value	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
Tumour Size	.013	.005	8.130	1	.004	1.014	1.004	1.023
Lymph Nodes	.051	.009	29.193	1	.000	1.053	1.033	1.072
Progesterone Receptor	-.005	.001	21.566	1	.000	.995	.992	.997
Tumour Grade			8.958	2	.011			
Tumour Grade (1)	-1.150	.441	6.798	1	.009	.317	.133	.752
Tumour Grade (2)	-.359	.167	4.605	1	.032	.698	.503	.969

In the above table the hazard ratio, $\phi(x) = \exp(\sum_{i=1}^{p} \beta_i x_i)$, is represented as the following equation:

$$\phi(x) = \exp\,[(0.013\, Tumour\ Size) + (0.051\, Lymph\ Nodes) + (-0.005\, \Pr ogetrone \operatorname{Re} ceptor) + (-1.150\, Tumour\ Grade(1)) + (-0.359\, Tumour\ Grade(2))].$$

The positive value 0.013 means that larger tumours carry greater risk, while the negative value (-0.005) means that higher progesterone receptor levels mean lower risk. In addition the highly negative value (-1.15) of tumour grade 1, means that a small tumour is a much lower risk than large grade tumours and (-0.359) of tumour grade 2, means that an intermediate tumour is a significantly lower risk than large grade tumours. Figure 6.9 shows the survival curve for the German data.

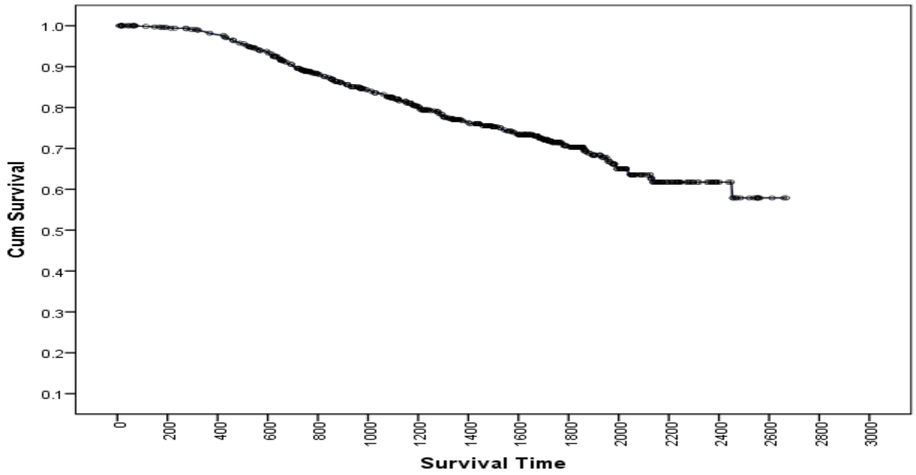


Figure 6-9 Cumulative survival function curve for German data

We compare between two groups of tumour size, divided by their median 25. The Chi-Square value from using the log rank test is equal to 12.089 which indicates a significant difference between the groups as expected. We denote these as groups 1 and 2, group 1 is for tumour size which is less or equal to 25 and group 2 is for tumour size greater than 25. Here we see that there is a significance relationship between the two groups. This conclusion was supported by Figure 6.10 where the survival curve for group 1 is better than for group 2.

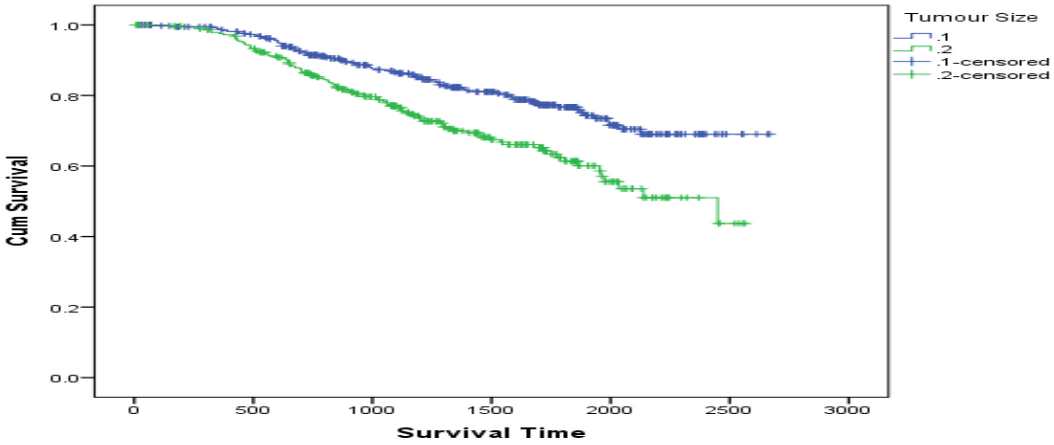


Figure 6-10 Survival function for tumour size (German data)

Naturally, Figure 6.11 shows that the corresponding cumulative hazard function for group 2 is higher than for group 1.

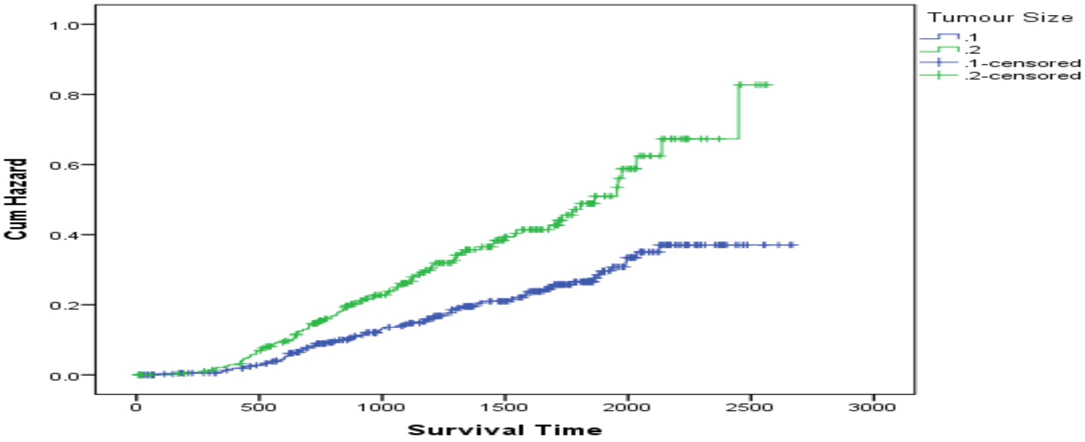


Figure 6-11Hazard function for the tumour size (German data)

Here, the individuals are divided into two groups divided by their median lymph node value 3. Group 1 represents the number of Lymph nodes less or equal to 3 and group 2 if it is greater than 3. There is a significant difference between the two Lymph Nodes categories of individuals, as the value of Chi-Square for the Lymph Nodes equals to 50.649 giving a p-value near 0.001. The cumulative survival curve and cumulative hazard functions for group 1 and 2, as shown below, indicate that group 1 has better survival rates than group 2.

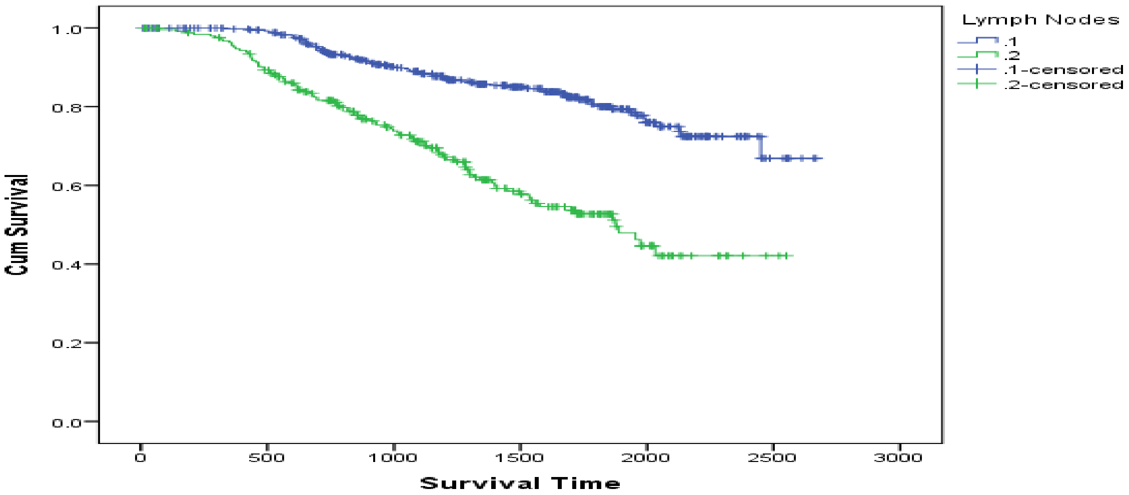


Figure 6-12 Survival function for the lymph nodes (German data)

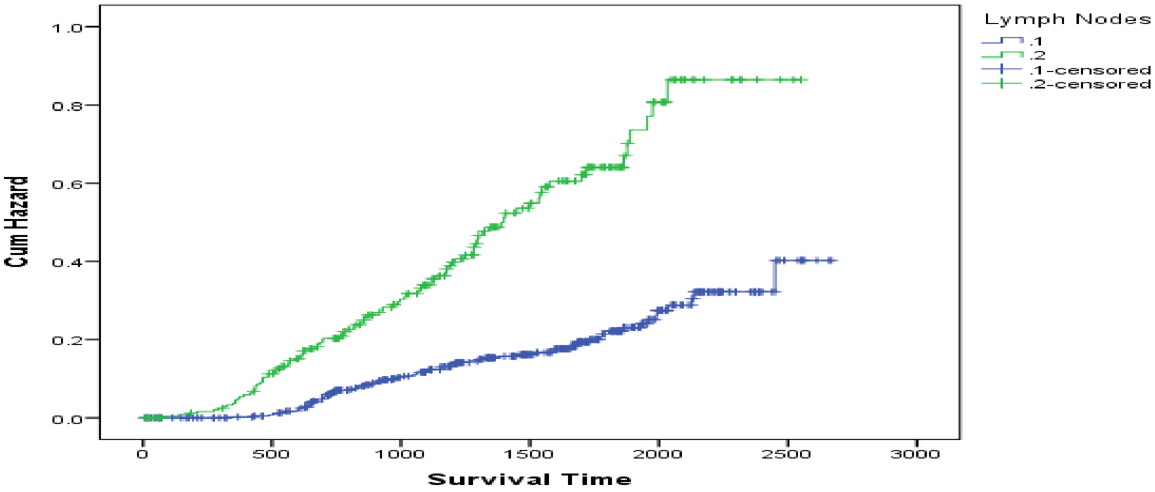


Figure 6-13 Hazard function for the lymph nodes (German data)

Figure 6.14, represents the cumulative survival curve for individuals in two progesterone receptor categories 1 and 2, chosen so that group 1 represents a level under the median value of 33, and group 2 represents a level above this value, and we see that survival for group 2 is greater than for group 1. The Chi-Square value of progesterone receptors equals 36.241; this result shows that the group 1 and 2 are also statistically significant based on their p-value of under 0.001.

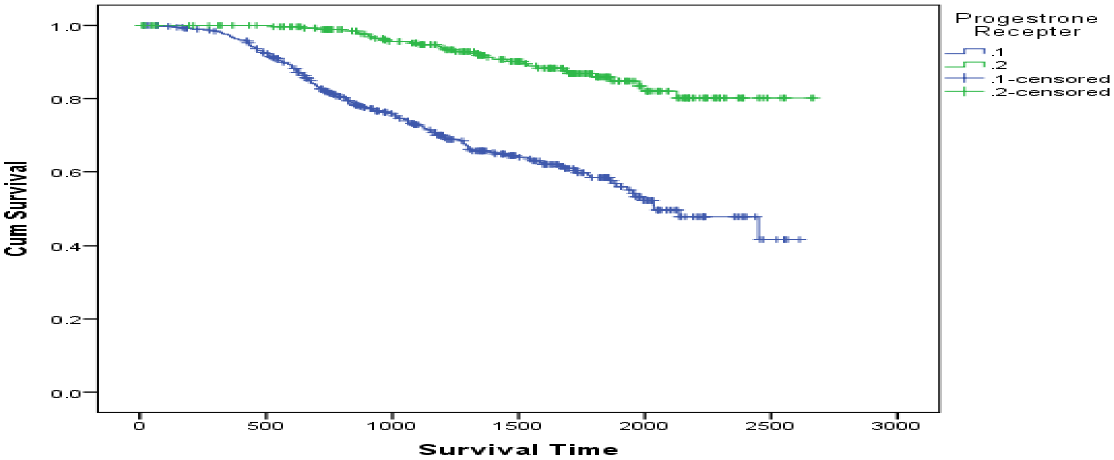


Figure 6-14 Survival function for the progesterone receptor (German data)

Naturally, the hazard for group 2 is less than for group 1 as shown in Figure 6.15 below:

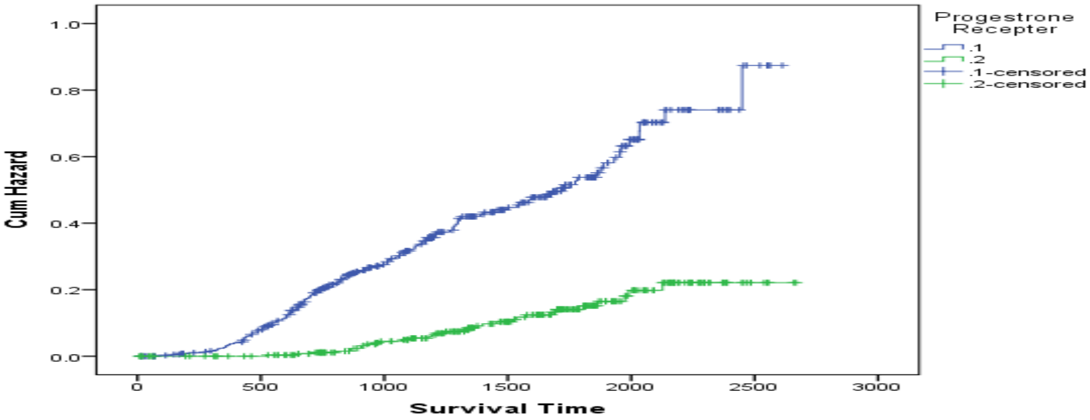


Figure 6-15 function for the progesterone receptor (German data)

Table 6.9, illustrates the log-rank test for the tumour grade. The results show the relation sheep between its individuals; the difference between small and large are stronger than that of small and medium.

Table 6-9 Log-rank test for the tumour grade variable for German data

Tumour Grade		Small		Medium		Large	
		Chi-Square	*p*.value	Chi-Square	*p*.value	Chi-Square	*p*.value
Log Rank (Mantel-Cox)	Small			10.060	.002	23.286	.001
	Medium	10.060	.002			14.813	.001
	Large	23.286	.001	14.813	.001		

The survival curves in Figure 6.16, demonstrates that survival for individuals with small tumours is better than for those with medium and large tumours and the survival curve for intermediate tumours is higher than that for the largest. See also Figure 6.17 for the corresponding hazard functions.

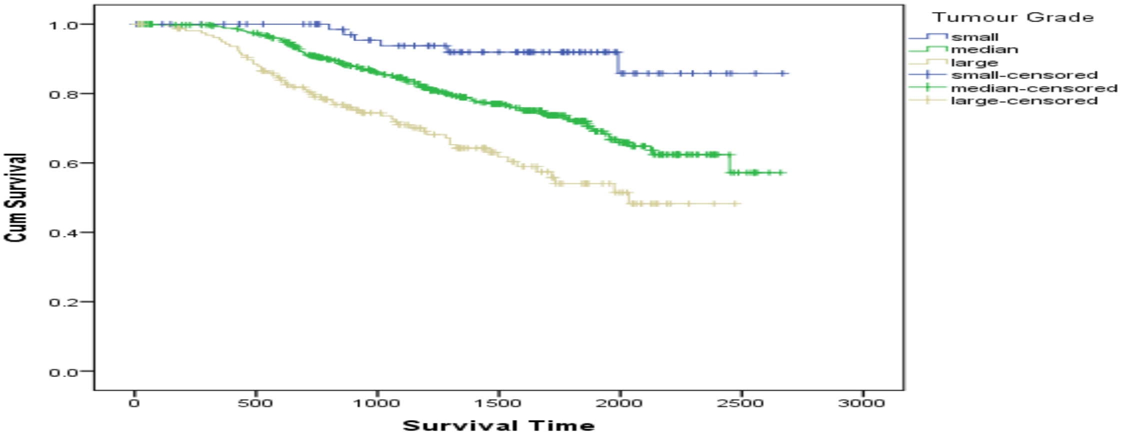


Figure 6-16 Survival function for the tumour grade (German data)

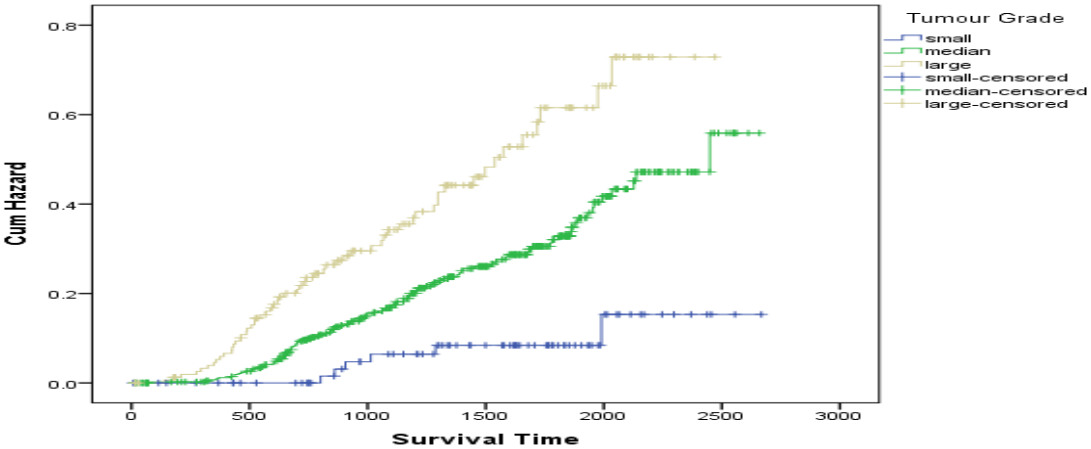


Figure 6-17 Hazard function for the tumour grade (German data)

6.3.1 Simulations

In this section we consider simulations to investigate the validity of our modeling procedure. We consider the example German data from Figure 6.9, as we have an accurate survival function for this because of the accurate data. For each simulation, we chose a distribution and simulated each individual from the German data being "lost" following this distribution. Thus if death happens before the individual is lost, we observe the death, but if the individual is lost first we assume that they are still in the study, and do not observe their death, if it occurs. This thus replicates what happens in the Nanakaly data, and the situation that we are modelling. The models that we have considered are Markov with constant rate of lost individuals, which would yield an exponentially distributed time of loss. We considered various values of this distribution.

One set of simulations considered a mean loss time of 2000 days. Given the length of the German study, this accounted for quite a significant loss of data. This is shown in Figure 6.18 where the apparent survival probability after 2000 days has only fallen to approximately 0.764 instead of the true value of just over 0.612 as a result. The survival curves generated for our two models with and without censoring are shown in Figures 6.19 and 6.20. The values of cumulative survival function up to 500 days are equal to 0.967 and 0.961 and again up to 1000 days are equal to 0.874 and 0.848 respectively. Finally, after 2000 days their cumulative survival rates for Figures 6.19 and 6.20 are 0.711 and 0.560 respectively. We can see that in both cases, the models significantly correct the survival function from the apparent survival function shown in Figure 6.18. The first model for this group gives a somewhat conservative correction, which is higher than the true survival function in Figure 6.9. i.e. we lose less individuals for the second model. As explained in Section 5.1.1, this is because of the double counting of lost and censored individuals.

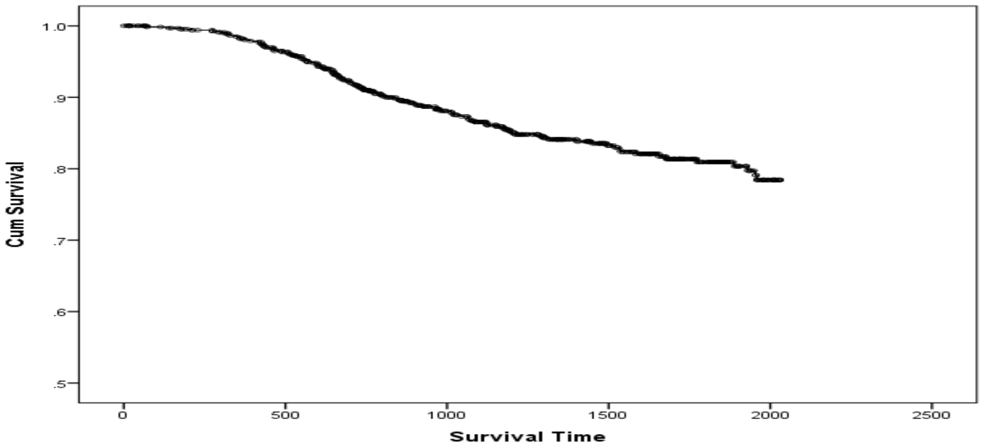


Figure 6-18 First simulation: the survival curve for a sample simulation of loss from the German data, where loss of individuals occurs following an exponential time with mean 2000 days

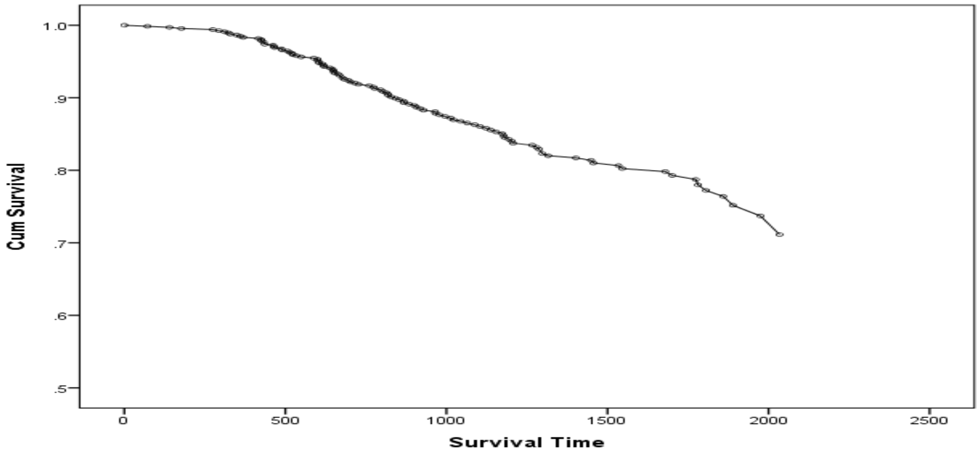


Figure 6-19 First simulation: an adjusted survival curve for the German data with simulated loss following an exponential distribution with mean 2000 days, using the method without censoring from Section 5.1.1

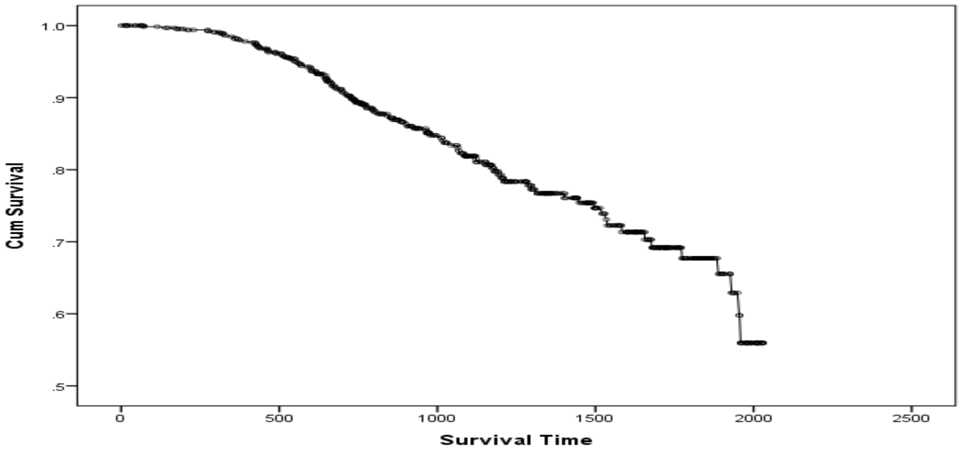


Figure 6-20 First simulation: an adjusted survival curve for the German data with simulated loss following an exponential distribution with mean 2000 using the method with censoring from Section 5.1.2

The second simulation with mean equal to 1000 is shown in Figure 6.21 where the apparent survival probability after 2000 days has only fallen to approximately 0.875 instead of the true value of just over 0.612 as a result. The survival curves for model one without censoring and the second model with censoring are shown in Figures 6.22 and 6.23. Furthermore, the survival rate after 2000 days is 0.756 and 0.709 respectively.

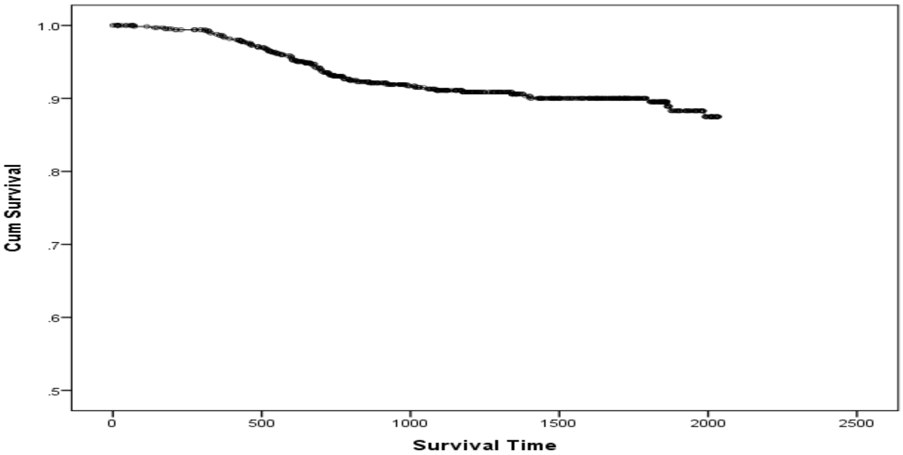


Figure 6-21Second simulation: the survival curve for a sample simulation of loss from the German data, where loss of individuals occurs following an exponential time with mean 1000 days

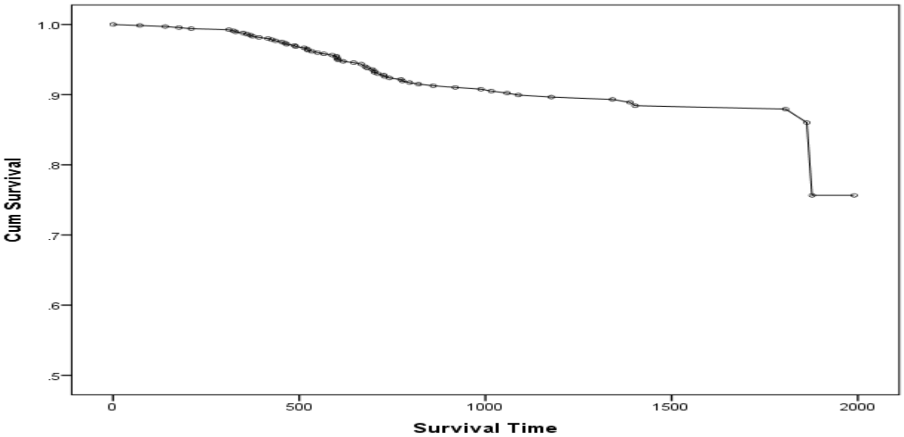


Figure 6-22 Second simulation: an adjusted survival curve for the German data with simulated loss following an exponential distribution with mean 1000 days, using the method without censoring from Section 5.1.1

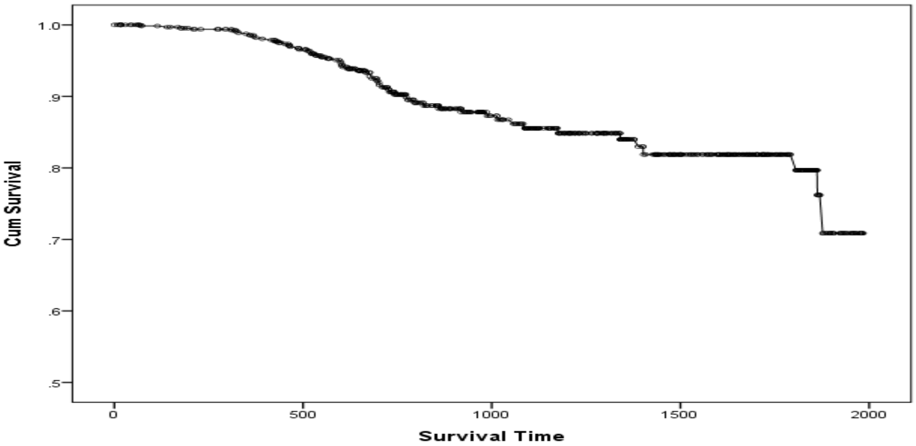


Figure 6-23 Second simulation: an adjusted survival curve for the German data with simulated loss following an exponential distribution with mean 1000 using the method with censoring from Section 5.1.2

An example set of simulations with mean 500 is shown in Figure 6.24. After 2000 days the apparent survival probability is approximately equal to 0.939 instead of the actual value of 0.612. The survival curves generated for both models are shown in Figures 6.25 and 6.26 and after 1500 days instead of 2000 days because there are a few numbers of individuals at risk toward the end of the study as we discussed in section 3.4.1.1, and the survival rates are 0.915 and 0.839 respectively.

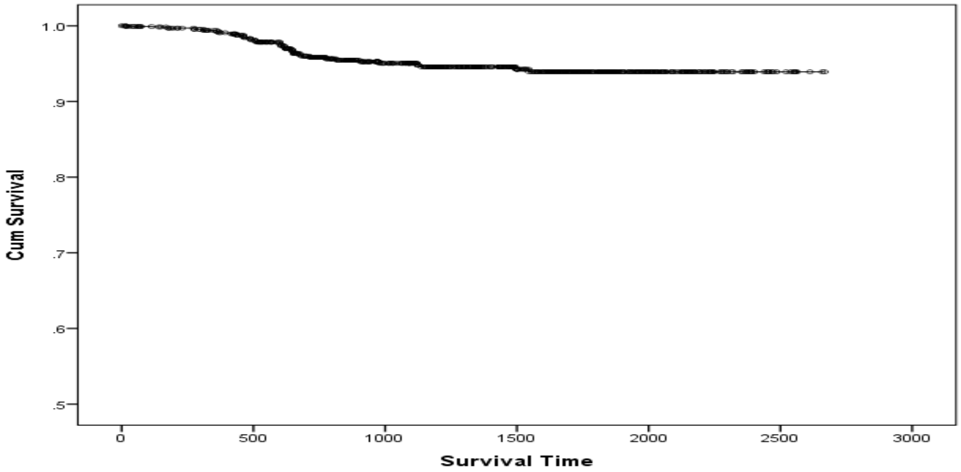


Figure 6-24 Third simulation: the survival curve for a sample simulation of loss from the German data, where loss of individuals occurs following an exponential time with mean 500 days

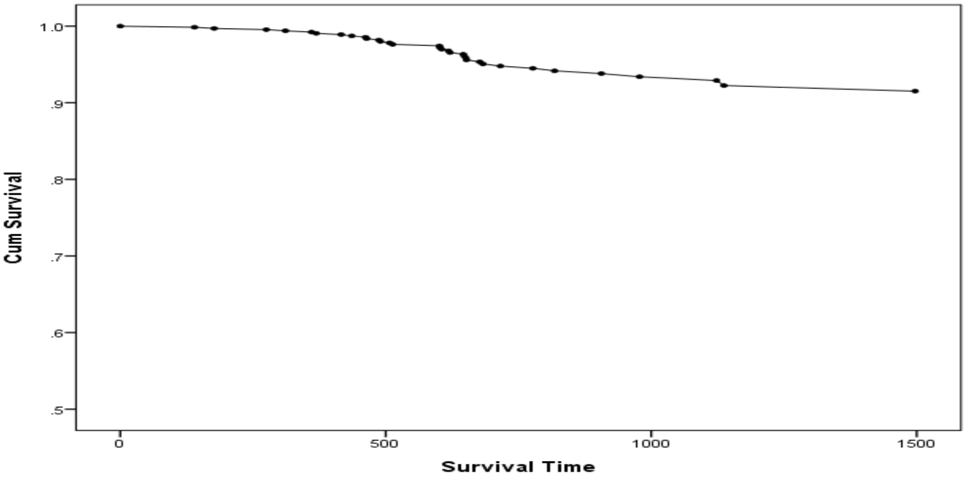


Figure 6-25 Third simulation: an adjusted survival curve for the German data with simulated loss following an exponential distribution with mean 500 days, using the method without censoring from Section 5.1.1

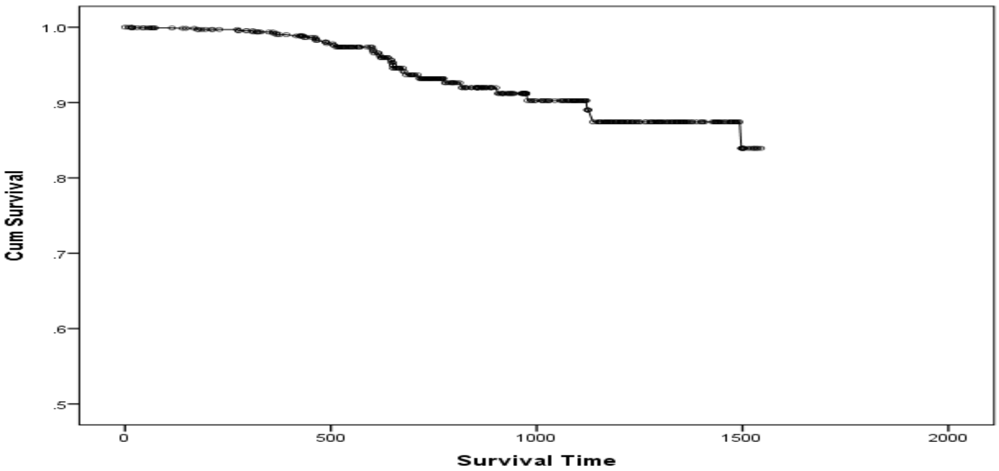


Figure 6-26 Third simulation: an adjusted survival curve for the German data with simulated loss following an exponential distribution with mean 500 days, using the method with censoring from Section 5.1.2

As shown in Figure 6.27, after 2000 days the apparent survival probability is equal to 0.854 instead of the actual value of 0.612. The survival curves generated for both models are shown in Figures 6.28 and 6.28 and after 1500 days instead of 2000 days (again because of the small number of remaining individuals as we mentioned in section 3.4.1.1), the survival rates are 0.827 and 0.625 respectively.

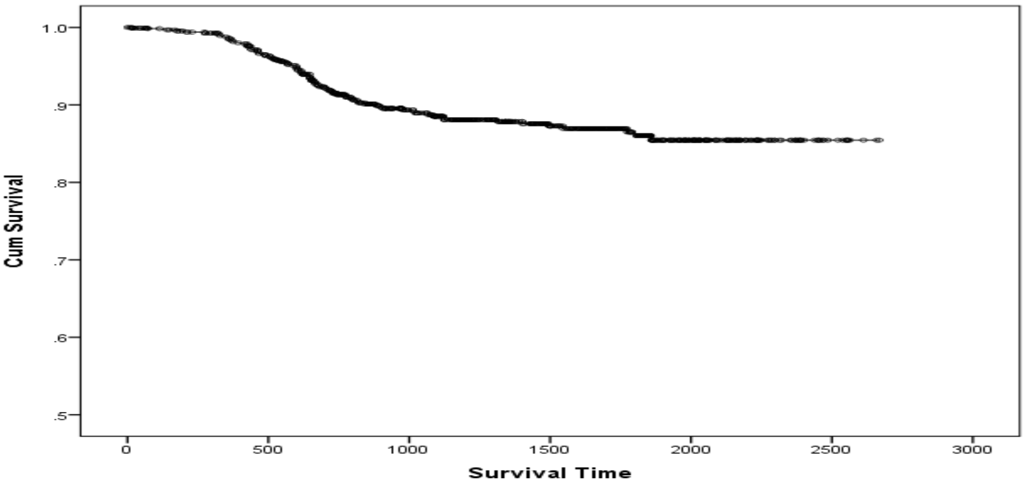


Figure 6-27 Fourth simulation: the survival curve for a sample simulation of loss from the German data, where loss of individuals occurs following Gamma (3,θ) time with mean 1000 days

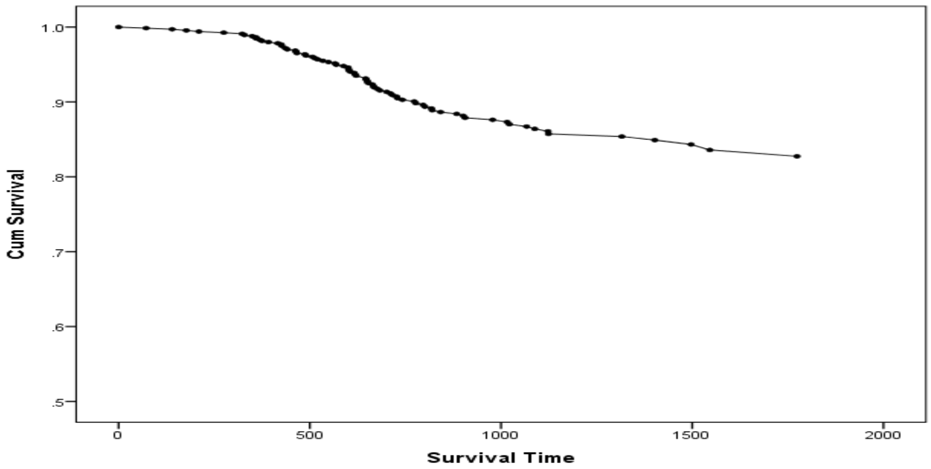


Figure 6-28 Fourth simulation: an adjusted survival curve for the German data with simulated loss following Gamma (3,θ) distribution with mean 1000 days, using the method without censoring from Section 5.1.1

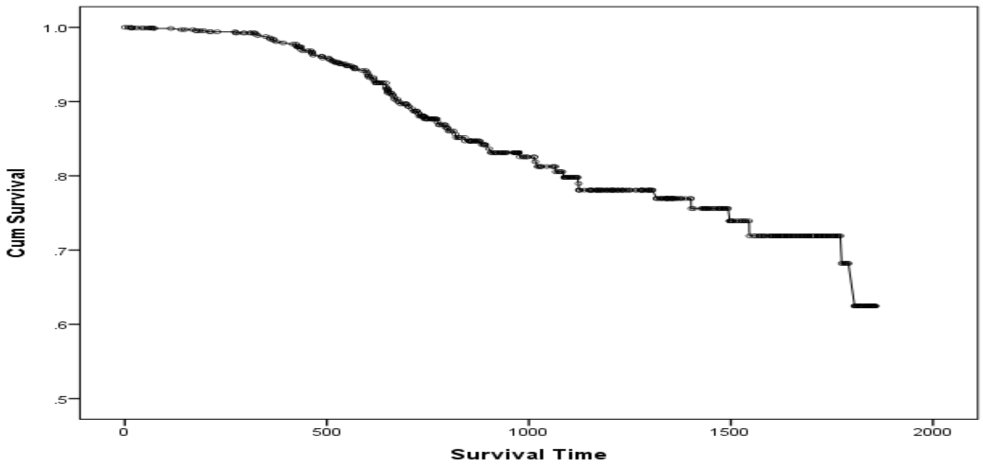


Figure 6-29 Fourth simulation: an adjusted survival curve for the German data with simulated loss following Gamma (3,θ) distribution with mean 1000 days, using the method with censoring from Section 5.1.1

We also considered the non-exponential distribution, e.g. gamma (3, 3/1000). When this led to a large number of lost individuals (for sufficiently high means this did not, and thus as above the corrections were not large and were accurate) we would expect our model to perform worse in such circumstances, as this would indicate that the underlying Markov assumption was not correct. This was indeed the case, although the models still corrected the false apparent curves to significant effect, and as for exponential distributions with small means, the effect was generally to produce slightly conservative survival functions, which overestimated the true survival curve.

Thus we see that our first simulation models perform well in many circumstances, and even when less accurate, are always an improvement on considering the apparent survival curves from the unadjusted data.

In the appendix A3 we consider sets of simulations to match the four above; simulating German data for the first and second models (with and without censoring) with different means 2000, 1000, 500 as well as Gamma distribution (3, 3/1000) with mean 1000 days. For each case we did ten simulations; the aim was to verify that the single simulations that we give above are truly representative and that results do not vary greatly from simulation to simulation. In general the cumulative survival functions were consistent for the unadjusted and adjusted first and second model data. The same situations were repeated when we took different simulations as shown in Appendix A3.

In Figure 6.18 for the unadjusted German data for the exponential distribution with a mean of 2000, the cumulative survival function curve after 2000 days is equal to 0.784 whereas in Figures A3.1, A3.4, A3.7, A3.10, A3.13, A3.16, A3.19, A3.22, A3.25 and A3.28 with the same distribution, the mean cumulative survival curve after 2000 days is 0.756 and the standard error is 0.021. This indicates that the results are very close to the observations as in Figure 6.18. When comparing the cumulative survival function of the first model in Figure 6.19, with 10 simulations for the same period, Figures A3.2, A3.5, A3.8, A3.11, A3.14, A3.17, A3.20, A3.23, A3.26 and A3.29, they both have a similar value of the cumulative survival function of 0.711 and 0.683, on average, respectively with the standard deviation 0.032. On the other hand, the adjusted cumulative survival function for the second model, Figure 6.20, is 0.560 in comparison to the simulations, Figures A3.3, A3.6, A3.9, A3.12, A3.15, A3.18, A3.21, A3.24, A3.27 and A3.30, that, on average, have value 0.502. This gives a difference of 0.058. Here the value of the standard deviation is 0.055, so that the results in this case are a little less reliable.

We now calculate the cumulative survival up to 2000 days for a mean of 1000. For the unadjusted model, Figure 6.21 and simulations, Figures A3.31, A3.34, A3.37, A3.40, A3.43, A3.46, A3.49, A3.52, A3.55 and A3.58, we have 0.875 and, on average, 0.843, respectively, which constitutes a deviation slightly above the standard error of 0.029. Meanwhile, for the first adjusted model, Figure 6.22 we have a value of 0.756 whereas simulations, Figures A3.32, A3.35, A3.38, A3.41, A3.44, A3.47, A3.50, A3.53, A3.56 and A3.59 give an average of 0.777 indicating a slight difference between them within the bounds of the standard deviation of 0.067. For the second adjusted model Figure 6.23 it is 0.709 and the simulations, Figures A3.33, A3.36, A3.39, A3.42, A3.45, A3.48, A3.51, A3.54, A3.57 and A3.60 average to 0.654 with a standard deviation of 0.035.

The cumulative survival function for the unadjusted model, Figure 6.24 at 1500 days (instead of 2000 days because there are few patients at risk) for a mean of 500 is 0.939 which is in agreement with the simulation average of 0.933 Figures A3.61, A3.64, A3.67, A3.70, A3.73, A3.76, A3.79, A3.82, A385 and A3.88, and the standard deviation of 0.006. On the other hand, the cumulative survival function for the adjusted first model over the same period in Figure 6.25 is 0.915 which exceeds the mean from the 10 simulations, Figures A3.62, A3.65, A3.68, A3.71, A3.74, A3.77, A3.80, A3.83, A3.86 and A3.89 of 0.871 and standard deviation 0.032. Again, when comparing the cumulative survival curve of the second model adjusted for the same period, Figure 6.26 with 10 simulations in Figures A3.63, A3.66, A3.69, A3.72, A3.75, A3.78, A3.81, A3.84, A3.87 and A3.90, they both give similar values of 0.839 and 0.821, respectively, with a standard deviation 0.025.

To check the accuracy of the suggested model we fitted the gamma distribution with parameters 3 and 1000/3, and so a mean of 1000, to our data (see Figure 6.27) and compare it to the respective simulations (see Figures A3.91, A3.94, A3.97, A3.100, A3.103, A3.106, A3.109, A3.112, A3.115 and A3.118). As a result, the cumulative survival for up to 2000 days is equal to 0.854 which agrees with the average value of the ten simulations of 0.842 which had a standard deviation of 0.009. From Figure 6.28 the cumulative survival function for the adjusted first model up to 1500 days is equal to 0.739, for the simulations (Figures A3.92, A3.95, A3.98, A3.101, A3.104, A3.107, A3.110, A3.113, A3.116 and A3.119), on average we find 0.746 with a standard deviation of 0.054, which again shows agreement.

Finally we compare the cumulative survival function for the second adjusted model with ten simulations (Figures A3.93, A3.96, A3.99, A3.102, A3.105, A3.108, A3.111, A3.114, A3.117 and A3.120). Figure 6.29 shows that after 1500 days the cumulative survival curve is equal to 0.625, which overestimates the simulated value of 0.590. The associated standard deviation from the ten simulations is 0.029. In general the results show good agreement. In summary we can conclude that the results from the simulation procedures are consistent from simulation to simulation.

6.3.2 Connections between German and Nanakaly data for survival analysis

For the purpose of comparison, we do additional analysis of the data after finding the survival analysis and apply it on both Kurdish Hospitals; Hewa and Nanakaly, and German patients. We follow the same procedures after combining the data from the three hospitals. This will be done firstly by comparing Nanakaly and German patients then Hewa and German and finally comparing Nanakaly, Hewa and German patients all together. Combining German patients with Nanakaly patients is done by combining the common variables in both of them. There are 686 patients in the German data and 713 patients in Nanakaly hospital.

Figures 6.30 illustrate the cumulative survival functions for the German and unadjusted Nanakaly data.

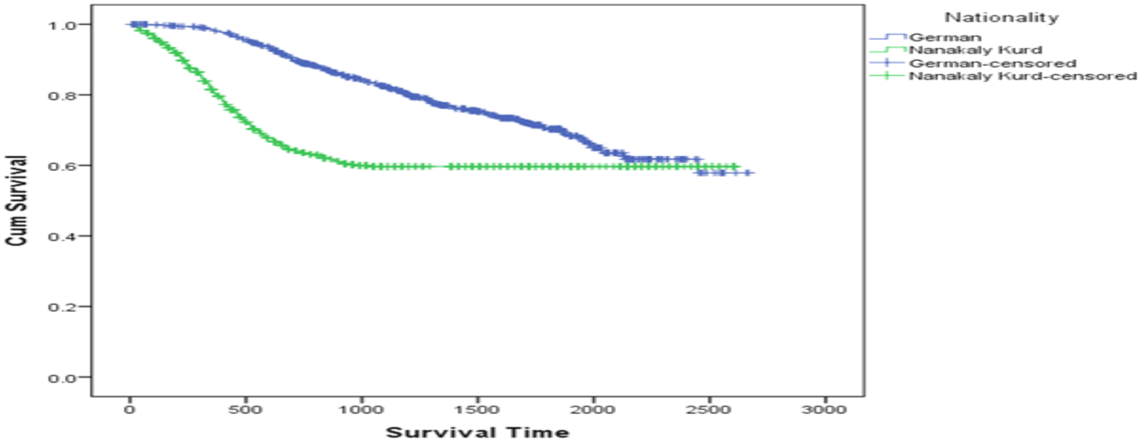


Figure 6-30 Cumulative survival function curves for German and Nanakaly data: the Nanakaly curve is for the unadjusted data

When comparing the blue line from Figure 6.30, the cumulative survival function for the German data and Figure 6.31, the adjusted cumulative survival curve for the Nanakaly data from the first model as shown in section 5.1.1, it is clear that the cumulative survival curve for the German data is higher than that for the Nanakaly data, largely due to the superior health care system. For German data the value of the cumulative survival function up to 100 days inclusive is equal to 0.999 while for the adjusted Nanakaly data it is equal to 0.965. Specifically in the German data there were 15 censored individuals and 1 death during the aforementioned period whereas for the adjusted Nanakaly data there are 20 censored individuals and 24 deaths. The cumulative survival function on day 500 is 0.956 and the adjusted Nanakaly counterpart is 0.679. The number of censoring events and deaths of patients in Germany from day 101 to day 500 inclusive is 21 and 28 patients while in the adjusted Nanakaly data for the same period there are 104 censored individuals and 152 deaths. For the last section, the cumulative survival function on day 1000 is 0.841 for German and 0.333 for the adjusted Nanakaly data. There were further 105 censored and 70 deaths and 109 censored and 63 deaths in the adjusted Nanakaly data from day 501 to day 1000.

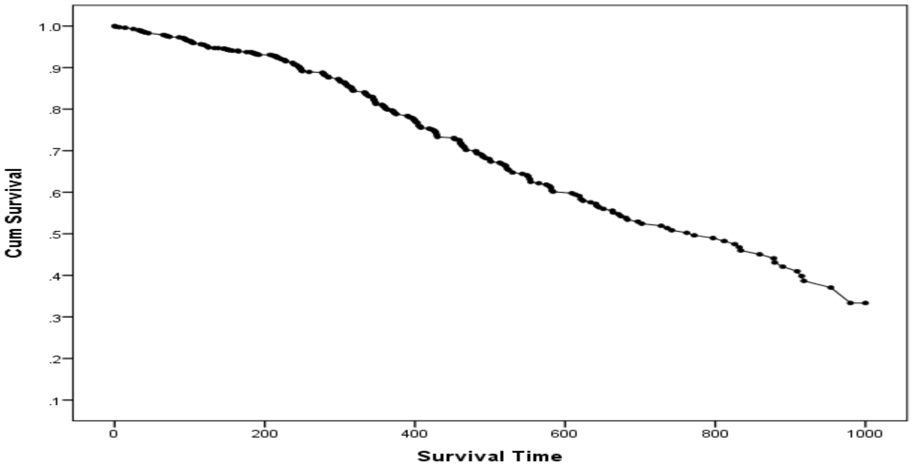


Figure 6-31Cumulative survival function curve for the adjusted Nanakaly data from the first model

Figure 6.32 shows the cumulative survival function for the unadjusted Nanakaly data for the age classes less than or equal to 48 years old and greater than 48 years old. We split the data at 48 based on the median age of all patients.

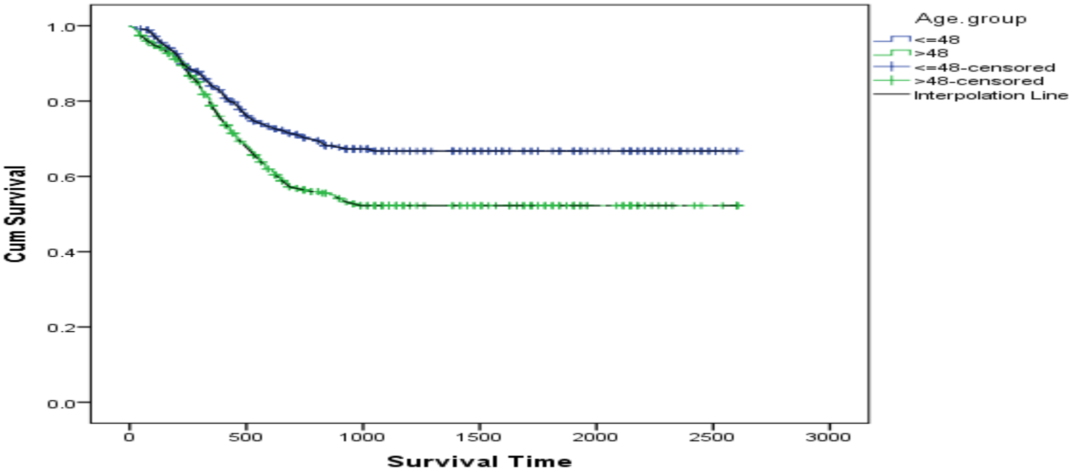


Figure 6-32 Cumulative survival function for age less than equal and greater than 48 years for the unadjusted Nanakaly data

The following Figures 6.33 and 6.34 illustrate the cumulative survival functions for the age class less than or equal to 48 years old and greater than 48 years old for the patients in the adjusted Nanakaly and German data respectively. For the age class of less than or equal to 48 years old, the difference between German and adjusted Nanakaly cumulative survival functions is 0.074. On day 600, the German survival function is higher than the adjusted

Nanakaly function by 0.297 for the same age class. Finally, up to 800 days, the cumulative survival function for the German data exceeds its adjusted Nanakaly counterpart by 0.333. Whilst when comparing the age class greater than 48 years old, for the same time points, the difference between German data and the adjusted Nanakaly data is 0.088 0.409 and 0.497 respectively, with the Germans on the higher side.

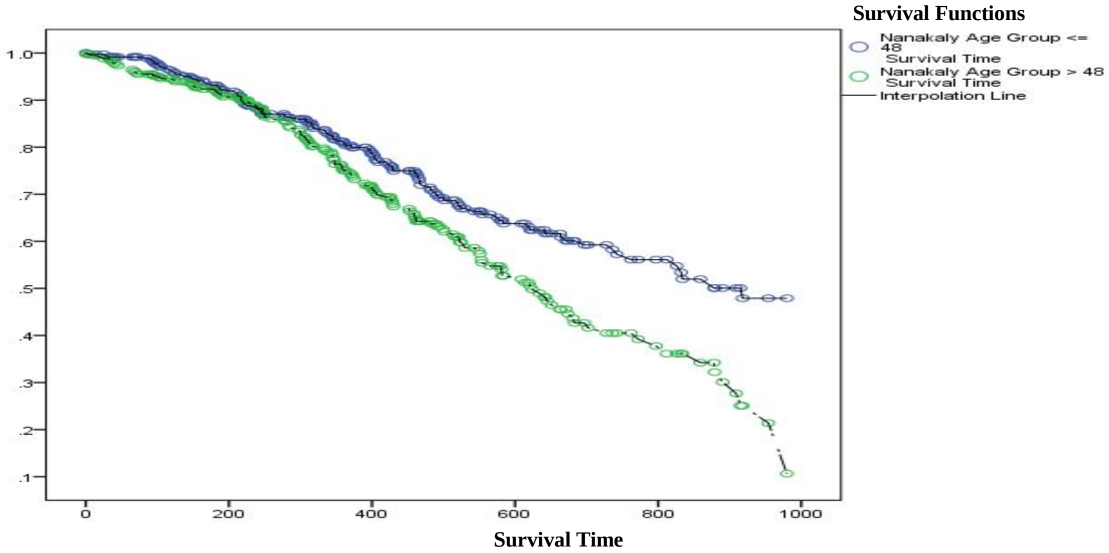


Figure 6-33 Cumulative survival function for age less than or equal to and greater than 48 years for the adjusted Nanakaly data

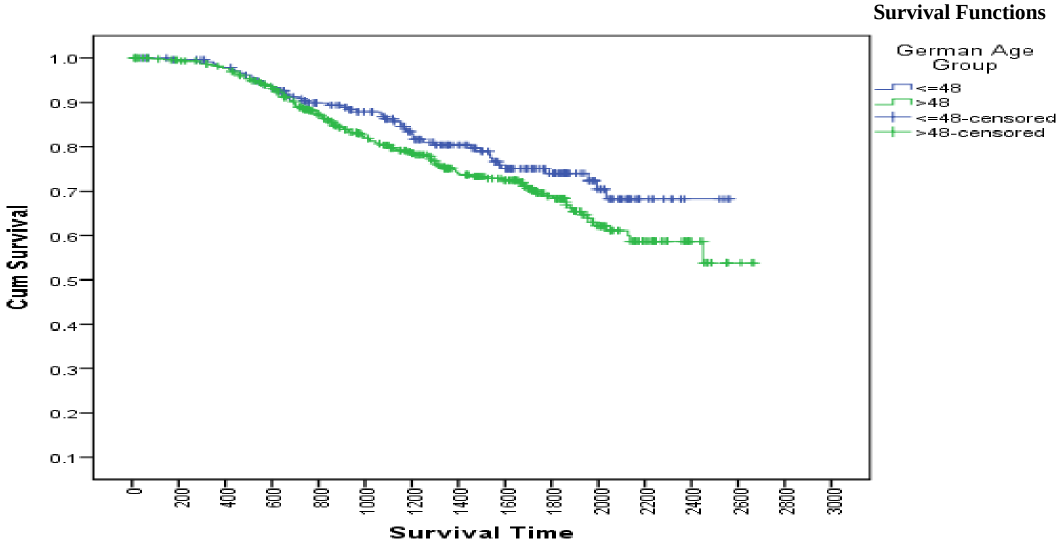


Figure 6-34 Cumulative survival function for age less than or equal to and greater than 48 years for the German data

6.3.3 Connections between German and Hewa data for survival analysis

Now we compare the German and Hewa data, which will be done by selecting the common variables in both of them. There are 686 and 1163 patients in the German and Hewa data sets, respectively. Figure 6.35 shows the cumulative survival functions for the German and the unadjusted Hewa data.

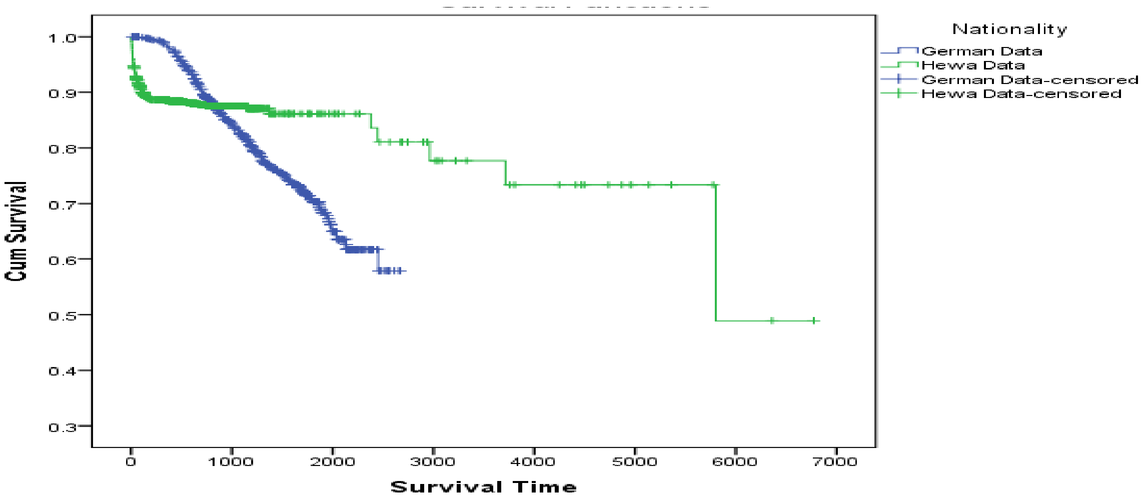


Figure 6-35 Cumulative survival function curve for German and Hewa data: the Hewa curve is for the unadjusted data

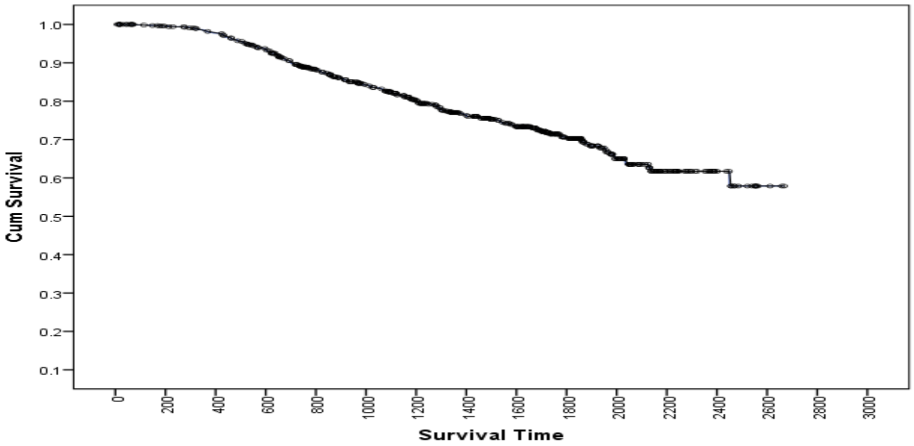


Figure 6-36 Cumulative survival function curve for German data (this is a copy of figure 6.30)

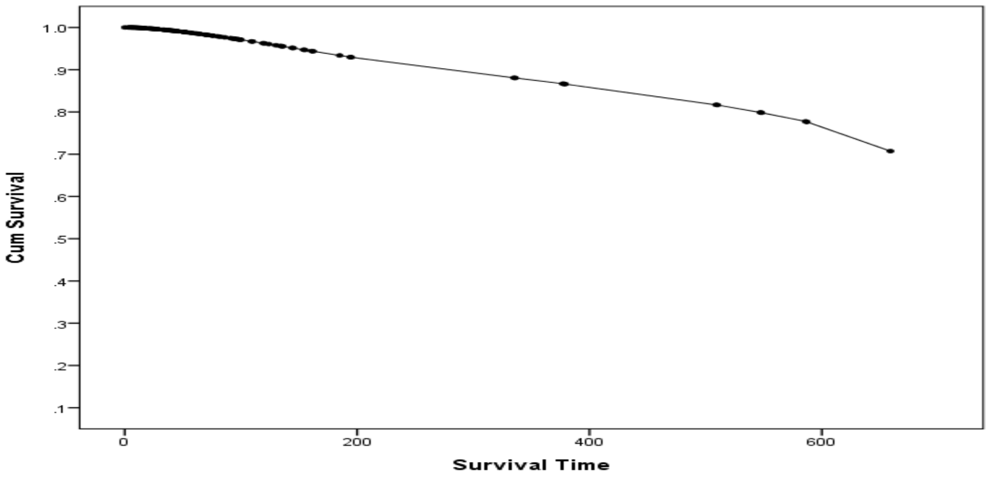


Figure 6-37 Cumulative survival function curve for the adjusted Hewa data from the first model of Section 5.2.1

In general the cumulative survival function for the Hewa data is lower than for the German data as illustrated in Figures 6.36 and 6.37. The value of the cumulative survival function up to 100 days for the German data is equal to 0.999 while for the adjusted Hewa data it is equal to 0.971. At day 500 the cumulative survival function value is 0.956 for the German data and 0.866 for the adjusted Hewa data. Lastly the cumulative survival rate at day 700 is 0.902 for the German data and 0.707 for the adjusted Hewa data. The first adjusted survival curve model for the Hewa data is discussed in Section 5.2.1.

The following Figures 6.38 and 6.39 are the cumulative survival functions for tumour grade in the German and unadjusted Hewa data. There are three grades for both data sets small; medium and large.

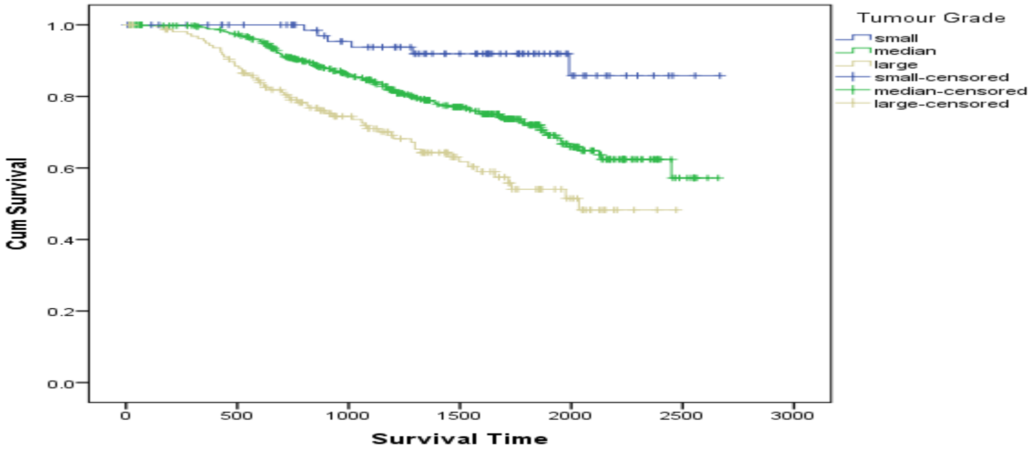


Figure 6-38 Cumulative survival function for the tumour grade variable for the German data

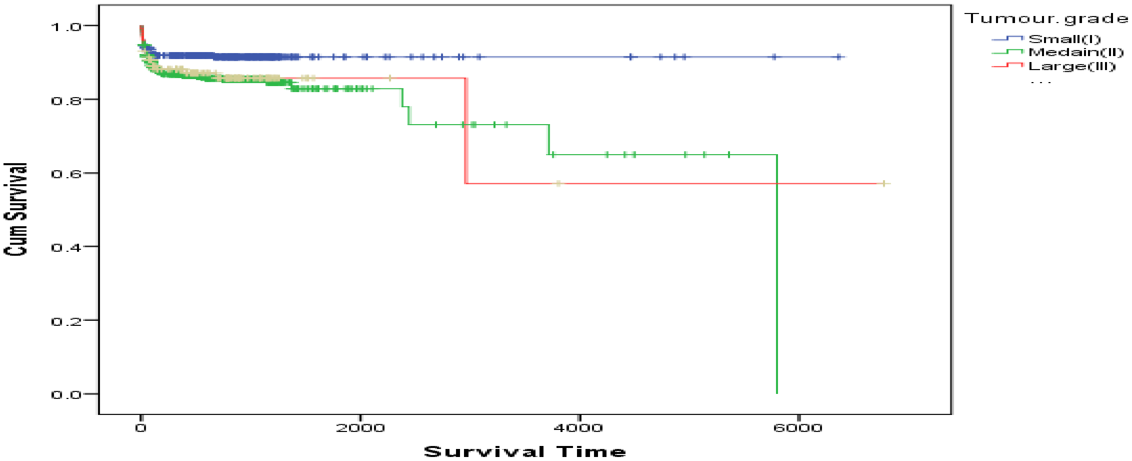


Figure 6-39 Cumulative survival function for tumour grade for the unadjusted Hewa data

The cumulative survival functions for the small tumour grade in the German and Hewa data are shown in Figures 6.40 and 6.41 respectively. In the German figure the cumulative survival probabilities up to 50, 100 and 700 days are to 0.999, 0.999 and 0.986 respectively. For the same period for the adjusted Hewa data these are 0.990, 0.975 and 0.803.

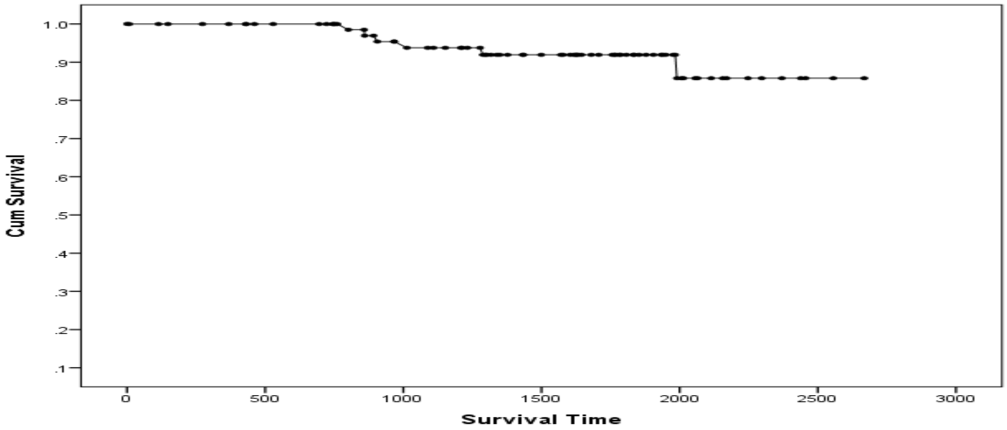


Figure 6-40 Cumulative survival function for the small tumour grade in the German data

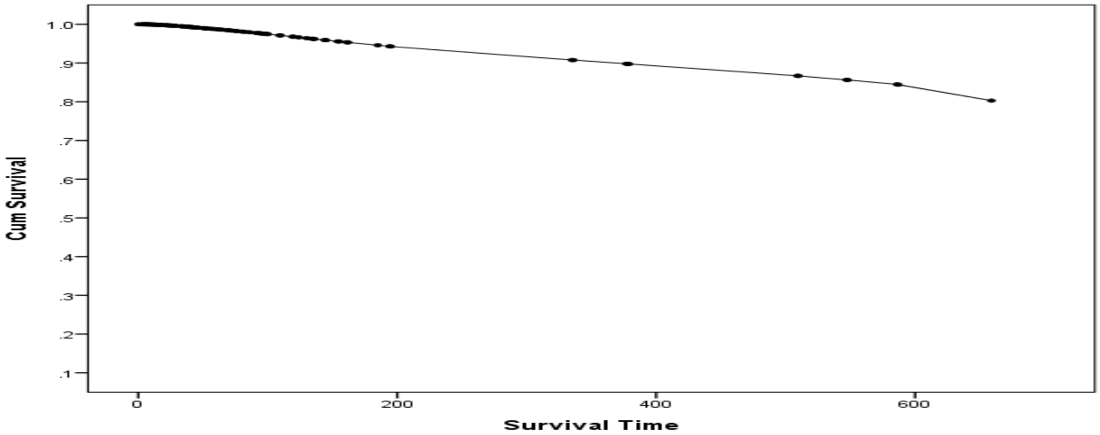


Figure 6-41Cumulative survival function for the small tumour grade in the adjusted Hewa data

The following Figures 6.42 and 6.43 show the cumulative survival functions for the medium tumour grade for the German and adjusted Hewa data. In the German figure the cumulative survival probabilities up to 50, 100 and 700 days are 0.999, 0.998 and 0.916 respectively. For the same period for the adjusted Hewa data these are 0.989, 0.969 and 0.657.

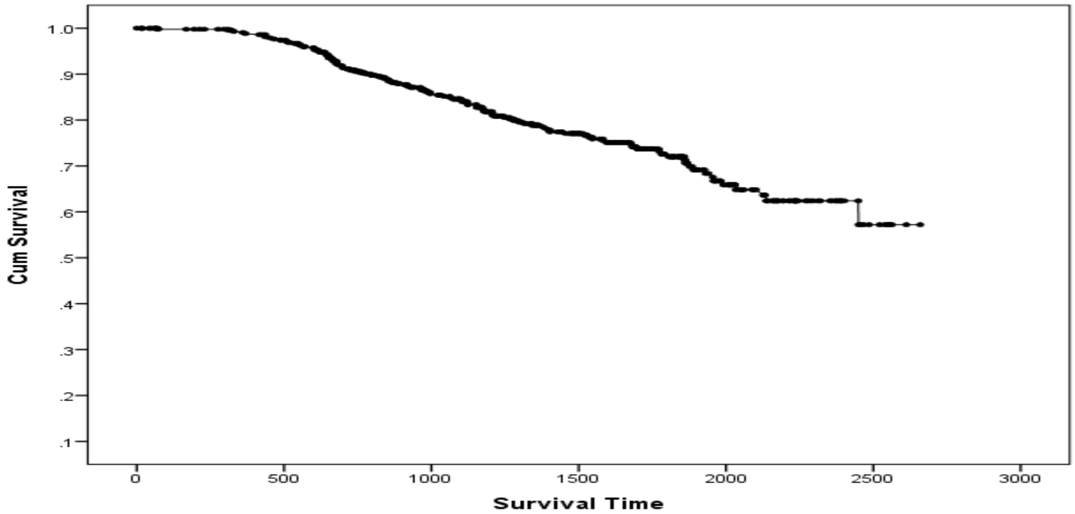


Figure 6-42 Cumulative survival function for the medium tumour grade in the German data

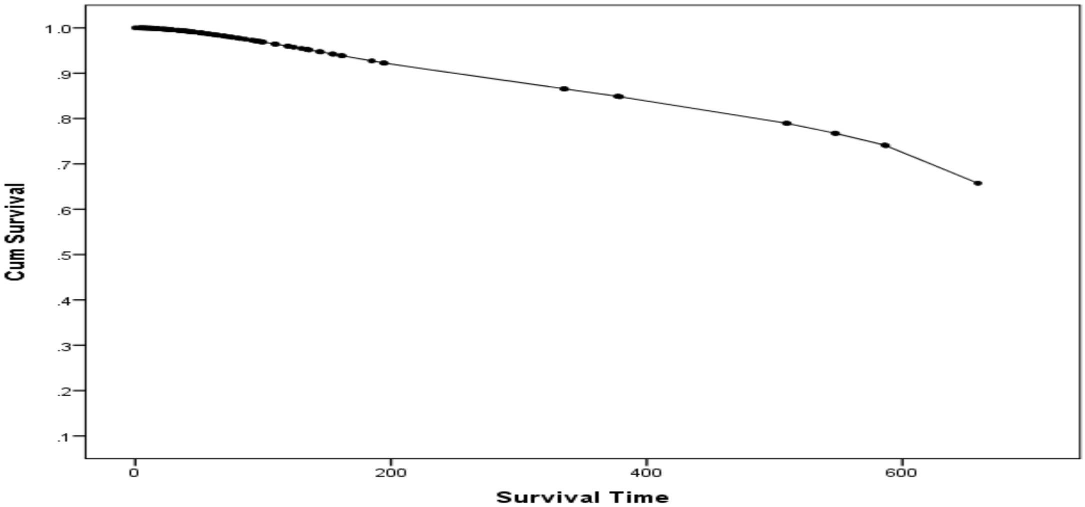


Figure 6-43 Cumulative survival function for medium tumour grade in the adjusted Hewa data

Survival for the large tumour grade between the German and adjusted Hewa data is shown in Figures 6.44 and 6.45 below. For the German data the cumulative survival curves for tumour grade large up to 50, 100 and 700 days are equal to 0.997, 0.995 and 0.811 respectively. Whilst for the adjusted Hewa data tumour grade large the cumulative survival probabilities are 0.988, 0.969 and 0.664, respectively.

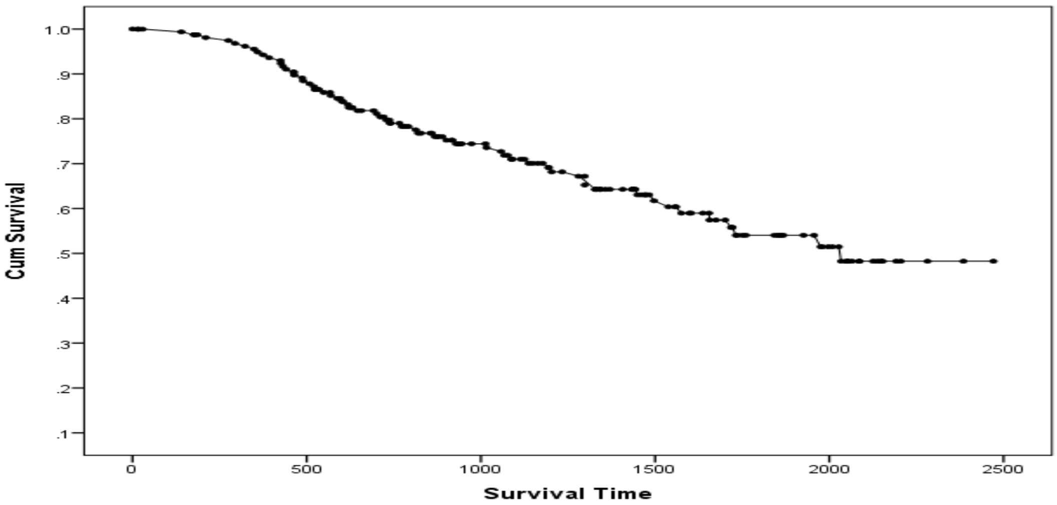


Figure 6-44 Cumulative survival function for the large tumour grade in the German data

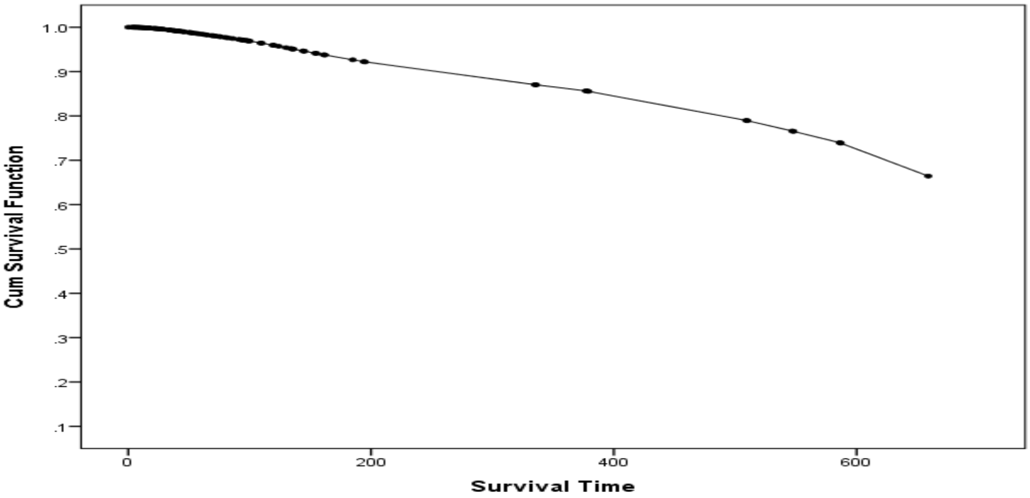


Figure 6-45 Cumulative survival function for the large tumour grade in the adjusted Hewa data

6.3.4 Connections between German, Nanakaly and Hewa data for survival analysis

Finally we consider a comparison between the three sets of data German, Nanakaly and Hewa. There were 686, 713 and 1163 patients in the German, Nanakaly and Hewa hospitals, respectively. The age variable is the only common variable amongst them. Figure 6.43 illustrates the cumulative survival function for three data sets, where the Nanakaly and Hewa data are unadjusted.

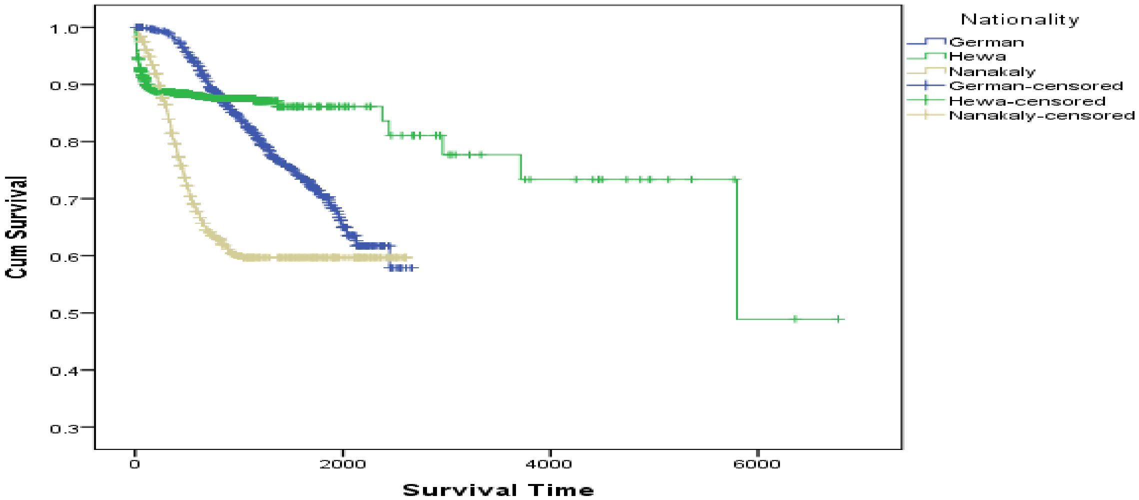


Figure 6-46 Cumulative survival function curves for the German, Nanakaly and Hewa data: the Nanakaly and Hewa data are unadjusted

Figures 6.47, 6.48 and 6.49 represent the cumulative survival function for the German data with the adjusted Nanakaly data and the adjusted Hewa data. The value of the cumulative survival function at 100 days for the German data is equal to 0.999 while for the adjusted Nanakaly data it is equal to 0.965 and for the adjusted Hewa data it is equal to 0.971. At day 500 the cumulative survival function value is 0.956 for the German data, 0.680 for the adjusted Nanakaly data and 0.866 for the adjusted Hewa data. Finally the cumulative survival probability on day 700 is 0.902 for the German, 0.524 for the adjusted Nanakaly data and 0.707 for the adjusted Hewa data. In general the survival curve for the German data is higher than for both the adjusted Nanakaly and Hewa data. On the other hand the survival function curve for Hewa data is higher than that for the Nanakaly data. This is likely to be because the data from Hewa is not reliable, due to the fact that we do not have the real time of death rather than representing a real large difference. Specifically the first adjusted survival curve model for the Nanakaly data are discussed in Section 5.1.1 and the model for the Hewa data is discussed in Section 5.2.1.

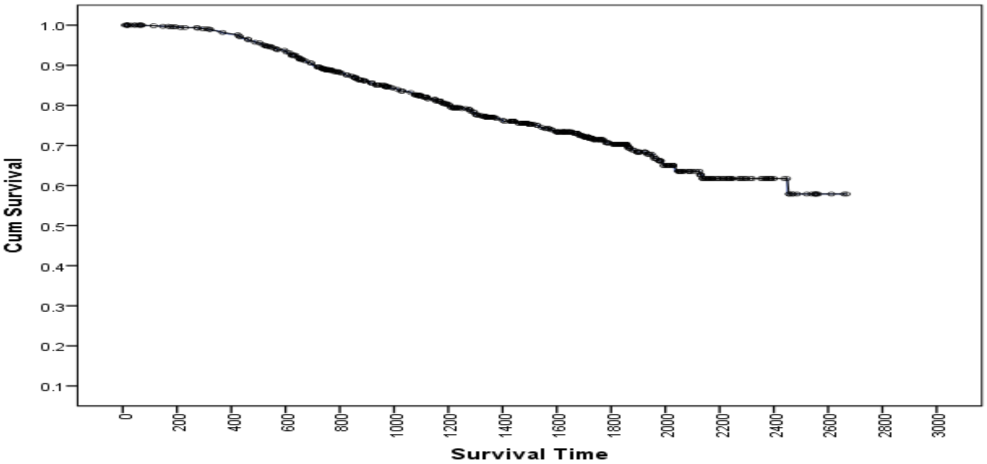


Figure 6-47 Cumulative survival function curve for the German data (this is a copy for Figure 6.9)

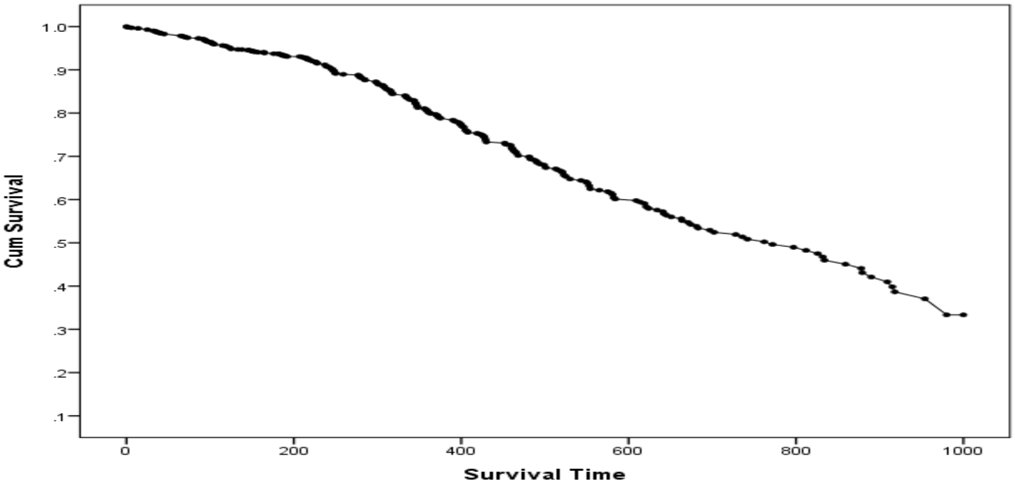


Figure 6-48 Cumulative survival function curve for the adjusted Nanakaly data from the first model (this is a copy of Figure 5.4)

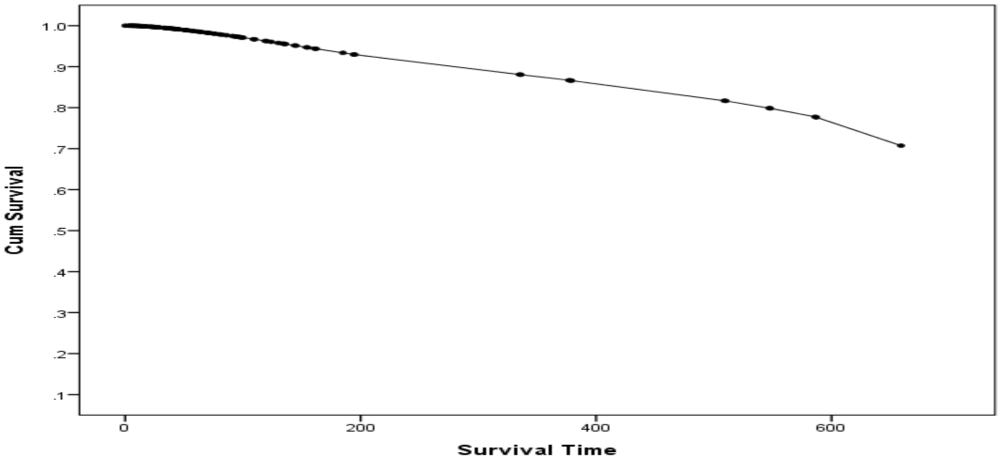


Figure 6-49 Cumulative survival function curve for the adjusted Hewa data from the first model (this is a copy of Figure 5.15)

The cumulative survival function for the unadjusted Hewa data for the two age classes is shown in Figure 6.50. The data is split into two parts at 48 based on the median age of all patients from the Nanakaly data into less than or equal to 48 years old and greater than 48 years old.

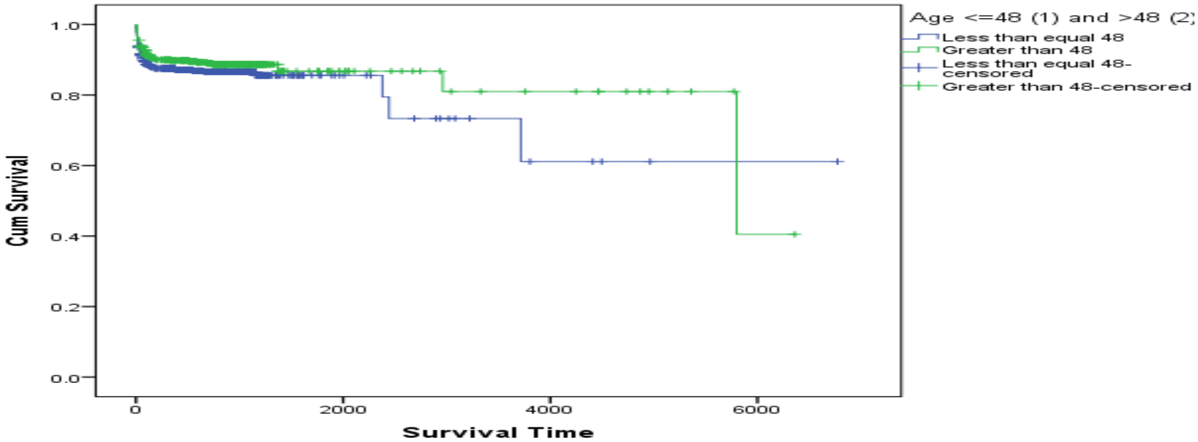


Figure 6-50 Cumulative survival function for age less than or equal to and greater than 48 years for the Hewa data for the unadjusted data

Figures 6.51, 6.52 and 6.53 display the cumulative survival functions for the age class less than or equal to 48 years old and greater than 48 years old for the patients in the German, adjusted Nanakaly and Hewa data respectively. For the age class of less than or equal to 48 years old, the cumulative survival probabilities for the German, and both the adjusted Nanakaly and Hewa data are given for three different periods and they are; at 250 days where there are equal to 0.996, 0.871 and 0.921 for German, Nanakaly and Hewa data respectively. At 500 days the corresponding values are equal to 0.961, 0.693 and 0.853. Finally at 700 days the cumulative survival probabilities are 0.908, 0.592 and 0.691 respectively. Comparing the age class greater than 48 years old, for the same time points the cumulative survival probability for 250 days shows 0.993, 0.865 and 0.938 respectively. The values for 500 days are 0.953, 0.626 and 0.878 respectively and the final values for 700 days are 0.896, 0.427 and 0.645 respectively.

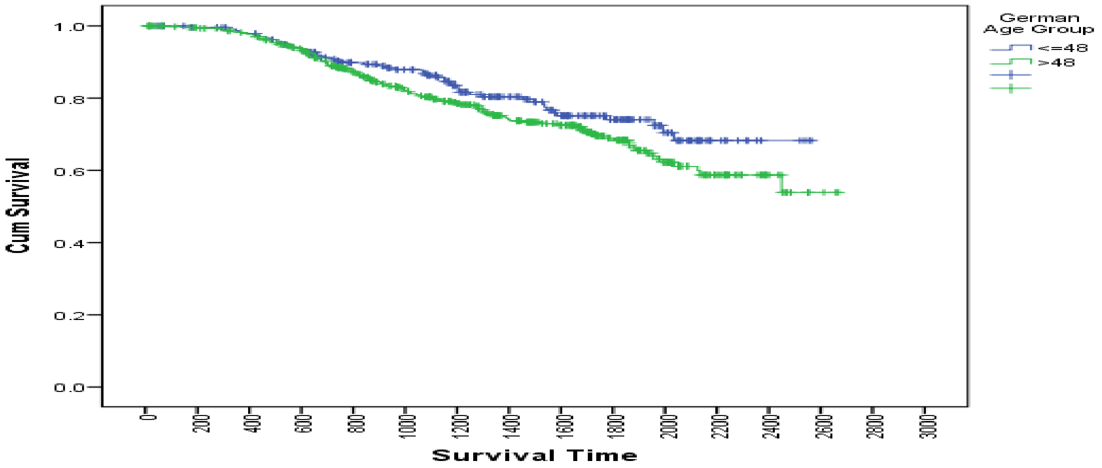


Figure 6-51Cumulative survival function for age less than or equal to 48 years and greater than 48 years for the German data

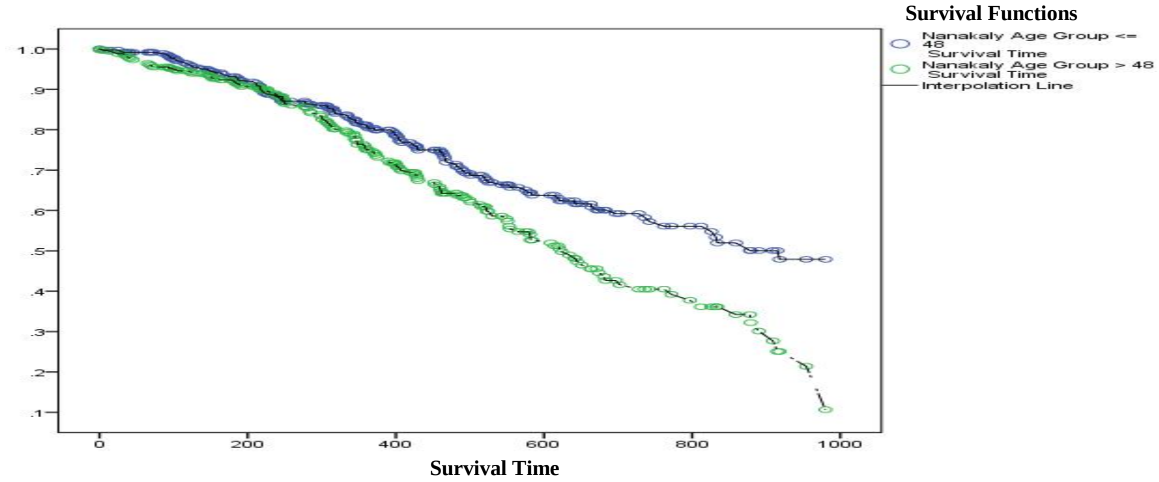


Figure 6-52 Cumulative survival function for age less than or equal to 48 years and greater than 48 years for the Nanakaly data: The curves are for the adjusted data

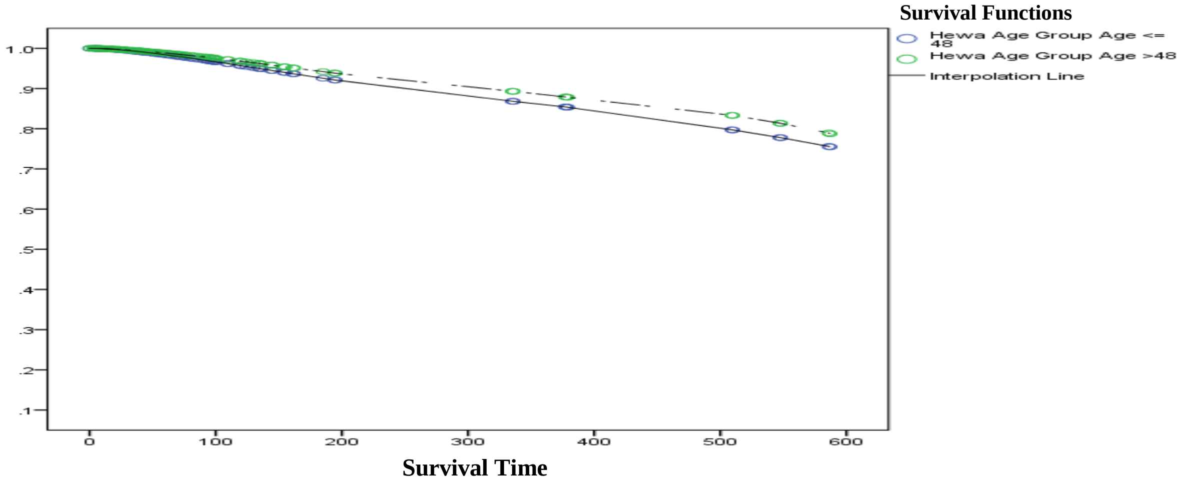


Figure 6-53 Cumulative survival function for age less than or equal to 48 years and greater than 48 years for the Hewa data: The curves are for the adjusted data

6.3.5 Comparison to previous studies on age in different countries

Age is considered one of the main factors when we are applying survival analysis, so a number of studies in various countries have been done. For example, in southern Iran, (Heydari et al., 2009) recorded that the survival probabilities of breast cancer were 0.970, 0.671, 0.453 and 0.253 for one, 5, 10 and 15 years. The five-year survival probability was similarly 0.751 in other areas of Iran as reported by Mousavi et al. (2006). While, Ghavam-Nasiri (2005) stated that in Mashhad, North-east of Iran, in general a five-year survival rate was 0.477. He also reported that the survival rate was not affected by age. Ueno et al. (2007) found the five-year survival probability of 0.803 and 0.670 between the periods 1982-1989 and 1990-2003 which was related to age at the time of diagnosis. For the purpose of comparison, Taylor et al. (2003) indicated that the five-year survival rate in New South Wales was 0.750, whereas in Western Sydney, Australia the probability was 0.791 (Clayforth et al. 2007). Also, Vahidian and Montazeri (2004) recorded that the survival rate of 0.62 of Iranian patients with breast cancer was lower than Western but higher than eastern European countries. The survival rate for Ugandans, Algerian and Gambian women are 0.462, 0.391 and 0.122 respectively (American Cancer Society, 2011) This indicates that the Iranian survival ratio is higher than them.

Using data from other cities and countries, Ziaei et al. (2013) investigated the survival rates in Tabriz (Northwest Iran) by a sample consisted of 271 breast cancer patients who visited a university clinic between 1997 and 2008. The survival rates for one three, five, seven and ten years were taken and they are significantly lower than cities in Europe and United States. This sample, however, is too small for a reliable survival analysis, in particular for the age group of less than 40 years old. From comparing our results with the previous studies, we see that the survival rate from breast cancer for South Iranian women is higher than for our patients from the Kurdistan Region of Iraq, whose survival rates appear closer to the African rates above.

6.3.6 Connection between unadjusted and adjusted data for Nanakaly and Hewa hospitals

In Section 6.1 we applied the SPSS program package for the datasets from Nanakaly and Hewa hospitals. We started with the unadjusted Nanakaly data and finding the survival curve (Kaplan Meier), and applying the Cox regression model, we found that age is a statistically significant variable in breast cancer risk.

The proportional hazard function is represented by the following equation (6.3) as described in Section 3.4.4,

$$\phi(x) = \phi_0 \exp(\sum_{i=1}^{p} \beta_i x_i). \qquad (6.3)$$

For the Nanakaly data there is only one variable, age. The relative risk is thus given by $\exp(\beta_1 Age)$. In particular we found that $\beta_1 = 0.032$, i.e. $\phi(x) = \exp(0.032Age)$, and that its inclusion in the model was statistically significant. In Section 5.2.1 we have seen that the original model as stated was not adequate because of the problems of "lost" individuals associated with the data. Can we adapt this analysis to use the above result?

Recall that we adjusted the data by using two models. The first model was based on estimating the hazard rate using $d_t / \tilde{n}_t$ instead of d_t / n_t, where $\tilde{n}_t$ is the adjusted measure of the number of individuals at risk, taking "lost" individuals into account. Here we can see that, considering subclasses of individuals, all risks are multiplied by a constant $n_t / \tilde{n}_t$ at each time point. Thus if age has a higher risk in the original model it also has higher risk in the adjusted model. Moreover this increased hazard will be unchanged, and so we argue that our parameter estimate of 0.032 is still valid, assuming that lost individuals are equally likely across all ages. Note that in model two we needed to use the Markov Chain process to estimate the value of the rate of recorded death and the rate of losing individuals (*p* and *l* respectively).

However, the results for the second model cannot be adjusted in a simple way to preserve the hazard ratios. Thus in principle the second model cannot be used in the same way. We saw that the two models produced very similar results, however, as we can see in the survival function curves in Figures 5.4 and 5.8. Thus we can conclude that the proportional hazard coefficient of 0.032 for age is still valid.

For the Hewa dataset there were two problems with the data, the problem of "lost" individuals as mentioned above, but also the problem of the absence of definitive times of death. We developed two models that tried to overcome these issues. From the original SPSS analysis, the Cox regression model showed that estrogen receptor, smoking and tumour grade were the significant risk factors for breast cancer as indicated in the following hazard function:

$$\begin{aligned}\phi(x) = \phi_0 \exp(-0.003\,Estogen\ \mathrm{Re}\,ceptor) * \exp(0.540\,Smoking\ Code) * \exp(-0.578\,Tumor\,Grade(1)) \\ * \exp(0.022\,Tumor\,Grade(2)) \cdot\end{aligned}$$

Our interpretation of the above risk factors is that the death rates for those individuals who are smokers are higher than for non-smokers, while the individuals with tumour grade 1 have a lower risk than for tumour grade 2 and tumour grade 3, whose risks are approximately the same. The estrogen receptor has a negative risk factor, which means that breast cancer appears to be hormonal-receptor-negative.

The first of our two models adjusted the true number of individuals at risk $\tilde{n}_t$ and the estimated number of deaths $\tilde{d}_t$, which depends upon the rate of transition from the Recorded Death class to the Death class (*z*). The estimated hazard rate was then adjusted from that using the original data d_t / n_t to $\tilde{d}_t / \tilde{n}_t$. As for the Nakakaly data there is a consistent scaling which preserves the order of the risk factors, though not in quite as straightforward a manner, as $\tilde{d}_t$ is a weighting of a number of recorded deaths from different time periods. Thus the conclusions of the SPSS analysis might be considered to be valid, except that the uncertainty about the interpretation of the Recorded Death category means that this is still questionable. For the second model we used the Markov Chain to estimate the rate of recorded death, the rate of censoring, the rate of losing individuals and the rate of death ($p, q, l\ and\ z$) in our model. As for the Nanakaly data, the hazard ratios are not preserved using this method. Further, we saw that the survival curves arising from the two models were considerably different (see Figures 5.14 and 5.20), partly at least as a consequence of the non-homogeneous censoring in the data. Thus the conclusions from our analysis of the risk factors in the Hewa data are not robust. We believe that it is not possible to obtain such robust conclusions from the current data.

7 CHAPTER 7: A proposed data collection methodology

7.1 Introduction

In this chapter we discuss the problems related to the data used in this study. To apply survival analysis methodology on breast cancer patients between women in the Kurdistan Region of Iraq, data was collected from the official database of the two main hospitals located in Nanakaly and Hewa.

The problems stem from an incomplete database due to various reasons based on what the Region has gone through in the past. As mentioned earlier in the problem statement of chapter one, the Region separated from the central Government in 1991 causing an internal conflict after that era. Every sector has been affected by this abnormal environment including the health sector. In comparison to various other diseases, the rate of breast cancer among women in the Region has risen dramatically. This is why we chose this disease to study in Kurdish society. The research involves the application of survival analysis in order to find new tools and ways to illustrate the importance of knowing the survival rate. When we first started collecting the data, we found that some data was missing which may be due to either those recording the data not realising the importance of the details or the patients not returning to the hospitals for a follow-up check. We have therefore made some adjustments to the data by proposing the use of a new model developed using mathematical methods.

7.2 Nanakaly hospital data

Nanakaly Hospital for Hematology and Oncology is a government hospital located in Azady-Hawler, Kurdistan Region of Iraq. It treats patients with blood disease, leukaemia & hemophilia. It was built by Hajji Ahmad Ismail Nanakaly and opened on 16 May 2004.

Nanakaly hospital announced in its annual cancer conference on Wednesday February 4, 2015, that the hospital contained three main departments which are blood diseases, cancer and child care. The first department deals with all blood problems, the second one deals with all types of cancer and the third one covers all diseases related to blood problems in children and all types of cancer in children.

Doctor Sami Ahmed, the director of the Nanakaly hospital talked about how the medical staff are distributed between the three department and how the medical services are provided to the patients. The hospital publicises that it is a charity and does not charge patients for chemotherapy or medication. All treatments, whether it be chemical or medicines, are available to patients for free.

The hospital also announced that the number of patients coming to the hospital that had benefited from the services provided to them in the year 2014 was more than 51,500 patients. During that year 1,839 new patients were recorded of which 855 were having blood problems and hemophilia. It was also stated that 40,790 patients visited the Hematologic, Chemistry, Bacteriology, Viruses, Serology and Parasite units. While 3,180 patients visited the ultrasound unit, 2,698 patients visited the X-ray unit and 537 patients visited the blood donation unit. Finally, 2,882 patients had been subjected to a bone marrow examination. The current member of staff serving as senior doctors and specialists comprised about 32 physicians, whilst there are only 48 medical assistants working three shifts. The shifts start and end at the following times: first shift from 8:15am to 1:15pm, second shift from 1:15pm to 7:15pm, and the third shift from 7:15pm to 8:15am the following morning. Work continues on Fridays and public holidays. The main problem at this point is that it is difficult for 48 nurses and their assistants to provide medical and therapeutic services to more than 5000 patients of the hospital which receives between 180 to 190 patients daily.

The data has been collected from the patients through direct contact between the specialist doctor and the patients. During this consultation the doctor will ask the patient some basic questions about age, weight, height, residency, marital status, and when she had realised that she had symptoms. After the consultation, the doctor will send her to the lab in order to take a specific blood test. After a couple of weeks the results of the test will come back from the lab. If the result is positive the doctor will suggest her to do a scan or mimeograph scan test to make sure the results are correct. The doctor will then advise the patients to come for follow-up tests. The patients are sent home if the blood tests are negative. During this process, the nurse will record all information which the doctor has requested from the patients, which includes extra information such as about breast feeding and number of births before transferring it to the computer database. During the process of recording the information and transferring the data to computer databases a lot of data is lost.

7.3 Hewa hospital data

The Hewa hospital in Sulaimaniyah, whose name (Kurdish for “hope”) recalls the agenda of the humanitarian services, specialises in oncology and haematology.

To identify the nature of work at Hewa hospital and how the medical services are provided to patients, the Union News Paper visited Hewa hospital in Suleimaniyah and met with its director Dr.Tawfiq Tarq and explained the following. The Hewa hospital of oncology and Haematology, established in 2007, was part of the public hospital in Suleimaniyah where the number of patients with cancerous diseases was around 500. The capacity was insufficient so the Ministry of health of the Kurdistan Regional Government in coordination with the Health Office in Suleimaniyah customised buildings belonging to the Health Office in Suleimaniyah to be a temporary building for Hewa hospital. However, they are still having problems with capacity for patients because this building was established to be a department of Health Office not a hospital (Al-Riad al Sharif, 2012).

Dr. Tawfiq talked about the capacity of the hospital and its sections saying that the hospital has 70 beds for patients who require special care, need necessary chemical treatments frequently and who have difficulty to come for a follow-up, especially those who live in areas far from the centre of Suleimaniyah or who are from other Governorates of Iraq. Patients who do not need to be in the hospital for control purposes undergo a 21 day treatment course. He added that, since 2007, there are over 10,000 patients who have benefited from the services provided by the hospital where 6,000 of them are from the Governorate of Kurdistan and 4,000 from other Governorates of Iraq. The hospital receives between 1,500 to 2,000 patients per year.

About the medical staffs who work at the hospital and its departments, he adds that there are 15 doctors and specialists. Ten of them specialize in oncology and haematology, and five physicians specialize in laboratory analysis of blood and bone marrow. There are also 60 health members working as doctor’s assistants, associate pharmacists or radiologist and others. In addition to that, there are 20 undergraduate nurses.

On the chemotherapy or medication, Dr.Tawfiq explained that the hospital does not charge patients. All treatments, whether chemical or medicinal, are available to patients free of charge. The cost for a course of treatment that takes 21 days is 3,000 US dollars and is paid for by the Ministry of Health with the support of the Kurdistan Regional Government. A large proportion of patients are from a low-income background. The support provided treats all patients equally and includes those from other parts of Iraq.

Dr. Tawfiq indicated that most doctors are specialists who have decided to continue to work at the hospital in the evening and not go to their private clinics. This allows them to provide medical and therapeutic services for patients who require continuous direct supervision.

In reply to a question on whether there are foreign medical experts in the hospital, he said that in general all doctors and other members in the hospital are Kurdish doctors, who are graduates of Iraqi Universities and Kurdistan Region Universities. Specifically, there are two Arab doctors, Dr. Basil and Dr. Sarmad, working in the hospital and there is no other foreign experience at the hospital.

There are currently about 25 physicians serving as senior doctors, specialists and senior evaluators. There are 100 medical assistants continuously working two shifts in the morning and the evening on a daily basis. This results in unforeseen pressure and stress for them as the hospital continually receives patients on a daily basis including Fridays and public holidays. It is therefore difficult for 100 nurses and their assistants to provide medical and therapeutic services for the cancer patients.

With regards to collecting the data from the patients, it is essentially the same process as Nanakaly Hospital which is described in the section above. Data is collected through direct contact between the doctors and their patients and is recorded in the same way but there are additional problems regarding the registration unit in this case. The regular employees in this unit are not aware of the importance of every detail about patients. This is partly due to the small size of the hospital and the large number of patients. Meanwhile their database includes many more details about breast cancer; time of diagnosis, time of death, smoking, drinking, exercise, family history, tumour size, tumour grade, lymph nodes, menopause, estrogens, progesterone, ethnicity, religion, hormone, income, occupation, BMI and blood type. However, when the time comes for a researcher to collect the data the basic problem still exists regarding the patients follow up as reflected from the survival function curve (see Figure 5.8).

7.4 General procedures to collect the data

The ability of information technology to deliver information in the health care system depends on the accuracy of recorded data and its accessibility for both patients and doctors. However, there is considerable evidence that the quality of current data is far from perfect, limiting the credibility of the data routinely collected for clinical use. The lack of information will affect the validity of the analysis adversely. Nevertheless, access to the clinical databases will be useful for patients' care at the national health services as it will keep doctors and nurses up to date on the current state of research and the development of new treatments.

One of the more challenging but nonetheless important tasks is going to be the setup of a centralized system for collecting, storing and sharing patients' records among the Kurdish hospitals. That can be achieved through development and use of electronic health records (EHR), which is constantly recording the data in all contexts. Applying this system for clinical structure will help to manage the information between patients and doctors. This system should be unified and strongly established to record clinical information in a way that can be shared and secured. Also it has to reflect the way that patients and doctors are working together to achieve the best health care.

It is necessary for electronic health records to protect the safety of patients and the good quality of care. In addition to keep all clinical records protected, such as management, planning, policy, commissioning, and research, for all uses, data must be appropriate for this purpose.

The best way to collect data that we can depend on is the use electronic health records (EHR) which is recognized by the Academy of Medical Royal College in 2008 (www.aomrc.org.uk/publications/ statements/doc_view/217-academy-statement-the-case-and-vision-for-patient-focusedrecords.html). The most important thing we must look at is the list of clinical record headers, each with a description of what should be logged under each header. In addition to the clinical categories, the full set of record headers should include admission, handover, discharge, outpatient, referrals, communications, and space for special remarks. Presently, the recorded data in Kurdish hospitals Nanakaly and Hewa do not meet these standards, as mentioned in section 7.2 and 7.3. In detail, the records in question must include the following:

1. Admission record: Standardised headers for the clinical situation to be recorded under when a patient is admitted. Not all headings are necessary in all care setting and situations. The

order they are listed in EHR applications, communications and letters can also be arranged by system providers and users.

2. Handover record: Standardised headers for the clinical data to be recorded under when a patient is being transferred from one professional team or the other, including in-hospital transfers at nights, weekends, or between consultants.

3. Discharge records: Standardised headers for the clinical data that should be held in the discharge record and included in the discharge summary from hospital to patient or their GP's.

4. Outpatient record: Standardised headers include the initial visit and follow up, this information should be included in the outpatient letter to GP and patient. This section should also include administrative information for the attributes of outpatient and ambulatory care sessions.

5. Referral record: Standardised headers are intended to log the clinical data in referrals between Gp_{s} and hospitals, with a copy to the patient. It should also be suitable after adaptation for specialist referral.

6. Core clinical record. They are priorities to be included in EHRs as they are in most countries.

The complete (EHR) record could include all the sections listed above for data recording, reviewing and communicating, their order adapted in context, all (EHR) logs should have the date time and the person's identity automatically registered. Finally, after preparing these standard records they will be reviewed and signed by a number of trusted organizations before making them available for use, for example Royal College of Anaesthetists, Royal College of General Practitioners, Royal College of Midwives, Royal College of Nursing and Royal College of Obstetricians and Gynaecologists (HSCIC, 2013).

7.5 Data required for survival analysis

Survival analysis describes the analysis of data that corresponds to the time from when an individual enters a study until the occurrence of some particular event or end-point. It is concerned with the comparison of survival curves for different combinations of risk factors. Analysis of survival data is complicated by the presence of censoring (patients leaving the study) (see e.g. Marmdan and Garibaldi, 2009).

Collecting data in western countries, for survival analysis and to follow up patients, there are different procedures that contain general guidelines to specific research questions, in order to make it more research focused.

In the study of cancer and other diseases, it is important to measure the time between response to treatment and recurrence or disease-free survival time, rather than just time to death (Clark et al., 2003). The recording of the type of event and when the period of observation starts and ends is necessary in survival analysis. All individuals with cancer cannot be observed for the same length of time, because some individuals are diagnosed at the beginning of the period under study, some near the end and others may be diagnosed at any time during the study. Basically, survival data contains uncensored and censored observations. Uncensored observations involve patients who are observed until they reach the end of the study. On the other hand, censored observations involve patients who survive beyond the end or who are lost to follow up at some point.

There are two major reasons for modelling survival data. First, we want to determine which combination of potential explanatory variables affects the form of the hazard function and, second, we want to estimate the hazard function for an individual in addition to their survival function (Collet, 1994). The methods used in survival analysis are semi-parametric, non-parametric and parametric methods, where each needs the same types of data. In western nations like Germany, the United Kingdom and the United States of America patients have rights and expectations, for example, as described in the NHS Constitution. These include "convenient and easy access to health services, free of charge and within maximum waiting times; a good quality of care and environment based on best practice; not to be discriminated against on the grounds of gender, race, religion and belief, sexual orientation, disability or age; to receive drugs and treatment as recommended by the National Institute for Health and Clinical Excellence (NICE) for use in the NHS if a specialist feels it is clinically appropriate for the patients; decisions made in a clear and transparent way so the patients can understand how services are planned and delivered; to be treated with dignity and respect in accordance with patients human rights; the right to privacy and confidentiality," (Collet 1994).

With respect to the above mentioned responsibilities of patients and hospitals, a lot of data is collected about patients that can be analysed. According to the International Agency for Research on Cancer (IARC) based in the United Kingdom there are some basic variables which are important in the statistical analysis of breast cancer such as date of diagnosis, recovery, death date, age, menopause, hormone treatment, survival time, tumour size, tumour grade, lymph node, estrogen, progesterone and time of censoring. Secondary variables of interest include tobacco use, alcohol use, infections, radiation exposures, occupational exposures, and medications (IARC, 2014). The World Cancer Research Fund/American Institute for Cancer Research (WCRF/AICR) identified additional factors such as diet, weight, and physical exercise (WCRF/AICR, 2007). IARC and WCRF/AICR evaluations set the standard in cancer epidemiology.

We collected the data to apply survival analysis from Kurdish hospitals in Nanakaly and Hewa. The variables in the Nanakaly hospital was time of diagnosis, time of death, residency and age of the patients only. The data contained many missing variables such as the variables related to the type of hormone treatment, and specific time of follow up. Also the data regarding social activity life for the patients, i.e., smoking, drinking, breast feeding, number of birth, abortion, life statues, habits of eating, doing exercises and etc was not available. However, in Hewa hospital there were a good number of variables but, crucially, the time of death was missing from their data. As a result it can be said that the best recorded data that can be relied on is the Nanakaly data from the Governorate of Hawler.

For each set of data we tried to compensate for the problems using a Markov chain model. The Nanakaly data structure of the random process uses, four states and its structure is shown in Figure 5.7, and the Hewa model structure uses five states (see Figure 5.19). Whilst these models and the associated methods described in sections 5.1 and 5.2 can be used to adjust for problems in the data, it would clearly be better to rectify those problems at source by collecting the data in a more efficient and complete way.

7.6 Solving problems in the Nanakaly and Hewa data

There are basic common issues for both hospitals which directly affect the quality of data causing problems in the analysis. Generally the hospitals are responsible for encouraging the patients to understand the necessity of follow-up checks and consider it as part of their treatment process. However, the problem concerns weak interaction between the patients and the doctors, between the doctors and the nurses, or both. Another problem is associated with the staff responsible for recording data, who must establish a strong, easy, understandable and multidimensional database of the patients. The basic problem in the data is that they were not recorded by the doctors themselves but by the nurses who have difficulty understanding the importance of each specific detail supplied by the patients. This is because they are not trained for such kinds of tasks. The reality is that when one comes to collect the data for research or any other academic purposes they are not helped to do so by the hospitals and need to depend on many other sources such as good personal relations with the nursing staff to obtain it. However, when the data is made available to the researchers, it contains a lot of missing information.

7.7 Description and justification of data requirements

Major reasons for the bad quality of the available data include a lack of awareness of the importance of keeping records on the doctors' side and a lack of trust in the health services on the patients' side. The latter may be linked to poor health education in some areas of the country or an insufficient degree of understanding between doctors and patients. In any case, not all the apparent deaths can be taken at face value since many of them likely arise from incomplete hospital records or simply a patient not showing up for their follow up treatment. It is important for the doctors and patients to make the hospitals be a place of trust between them. In addition they are responsible to recommend to the patients about the necessity for follow up and encouraging them to think rationally not emotionally and they have to make more effort to help the patients understand the importance of giving correct information and they have to record each question in detail and take the responsibility of providing a correct answer. On the other hand I suggest that the doctors have to put a specific schedule based on weekly or monthly support to control patient's follow up. The important thing is that the policy makers support and encourage the hospitals in the Kurdistan Region to build a strong, accurate database between all the Governorates of the Region. Since the main issue in Hewa Hospital is building (as published in Al-Riad al Sharif, 2012), the Government should build a sizable hospital and prepare a desirable place for them. Also we suggest it is better for the Kurdish Government to send their candidate nurses to receive international training abroad to learn how to establish a strong database or following some developed country's database, for example the German hospital database which is strongly established and includes all types of cancer.

7.8 Plan of a data collection methodology

In Western countries it is standard academic protocol to extensively plan data collection beforehand; in particular that involves clear specification of who will be responsible for collecting, and recording the data. Generally, the doctors in western countries who are working in the general practitioner (GP) surgeries are responsible for collecting the data from the patients. The accuracy of these data may be related to the awareness of these doctors who are highly trained to provide health information to the patients. The process of gathering data starts when the patients visit the general practitioner surgeries for the first time. Here the nurse asks the patients to fill an application form, which contains all information about the patients as well as examples and guides on how to complete it, before making an appointment with the doctor for them. After that this form will be checked and stored by the administration. These data should be stored in a clear way which will be easily accessible to the user. For instance it is important to clarify whether it is collected by questionnaires, emails, recorded interviews, copies of official documents and stating the name and location of this information (Royal College of General Practitioners, 2011). The reliability of these data should be checked through their consistency and finding ways and methods for dealing with any suspect or wrong data (Royal College of General Practitioners, 2011).

The Royal College of General Practitioners (2011) stated that the data collected on general practitioner surgeries in the United Kingdom indicates very high standards of patient diagnosis, treatment, care, and support. The most important thing is that, if necessary, they may refer the patients to the specialists or to the community services. The doctors share patients situation information and recording all necessary information involved in patients treatment.

Hospitals in the Kurdistan Region can follow exactly the same methods as Community Health Centres (CHCs) in the United Kingdom in order to have accurate data about the patients. Because of the great diversity of patients seen at CHCs and it is important to have their boards of directors representing their communities; CHCs appear to be ahead of the curve in collecting different information on each patient. Perhaps the most practical approach for the hospitals in the Kurdistan Region is to follow the Health and Social Care Information Center (HSCIC) method to collect the data. Also it would be helpful for the Kurdish hospitals to adapt one of the following ways to get access to the data: by using a Secure File Transfer Mechanism or by accessing the hospital episode (period) statistics (HES) data using the hospital episode statistics (HES) data interrogation system.

7.9 General flow chart for breast cancer

When establishing a design to collect the data on breast cancer it is necessary to take all related factors, causing it which are specified by the physicians, into consideration. The following graphs represent two subjective flow charts for that reason. The first flow chart contains general information to use for collecting the data. The second flow chart includes information about the individual in more detail. In the general data collection design flow chart, the data from the patient will be collected and recorded in a computer database. If we use the first choice then we have to think of the historical background of the patients or making questionnaire forms to collect the data. In the second case, we only have to make a database request, taking into account the typical structure of the data associated to a patient. For instance, in the UK a breast cancer patient’s file is typically subdivided into five categories, namely breast cancer health history, physical examination, case management, risk assessment and referral and treatment. After collection by either method the data will be ready to use in a specific model and to apply it for breast cancer patients (see Figure 7.1).

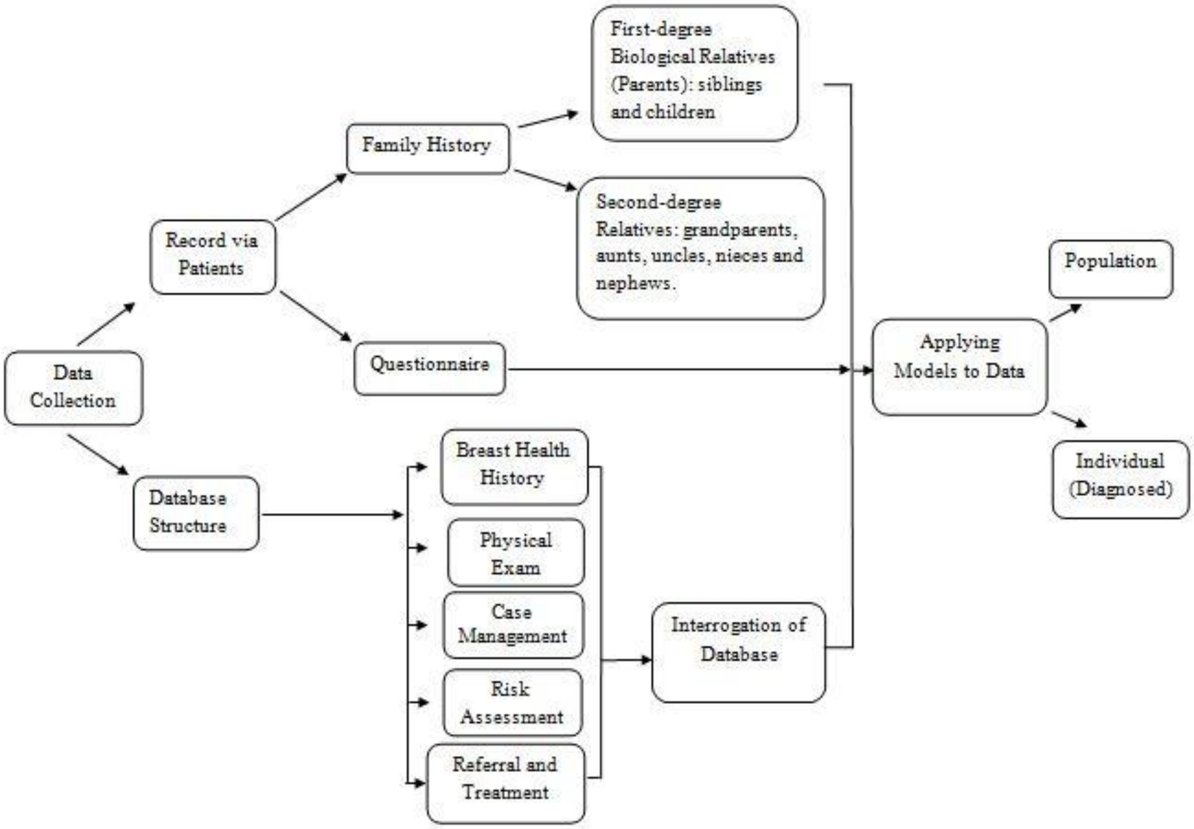


Figure 7-1 General flow chart for data collection design

7.9.1 Specific flow chart for breast cancer patients

When collecting the data using the form designed for patients individually, the information will be recorded in much more detail than the general flow chart. For first time patients when arriving at hospital answer to the most common questions about the date, name, age and address will be recorded. Then the person will he asked about the eleven main factors which are considered to be the basic cause of breast cancer, after which the individual will be asked about Gender, Smoking, Drinking alcohol, Age, Genetic risk, Marital status, Menopause, Number of children, Occupation and if they have undergone any surgery in addition to recording other health problems. Individuals undergo a number of processes, starting with a hormone test for Estrogen (+/-), Progesterone(+/-), HER2(+/-), FISH, LH, Thyroid Function and Prolacting. In the case of a positive diagnosis, the current stage of illness will be recorded. After the type, size and grade of the tumour have been specified the patients will go through to the treatment stage; there are four kinds of treatment; surgery, chemotherapy, radiotherapy and hormone therapy. It is crucial at this stage that the treatment is well-recorded in terms of dates and dosages. If the patient recovers then the patient will be sent home, the date and the time of recovery will recorded and they will be asked for a follow up every 6 months. But if the specific treatment is not successful, the dosage or type of treatment has to be changed and any alteration needs to be documented. If the patient recovers after this new action the date and time of recovery will be recorded and the patient will be asked for follow up every 6 months. If this process is not successful, the treatment will again be changed and the process continues. Finally this processes will continue until the patient either survives or dies (see Figure 7.2).

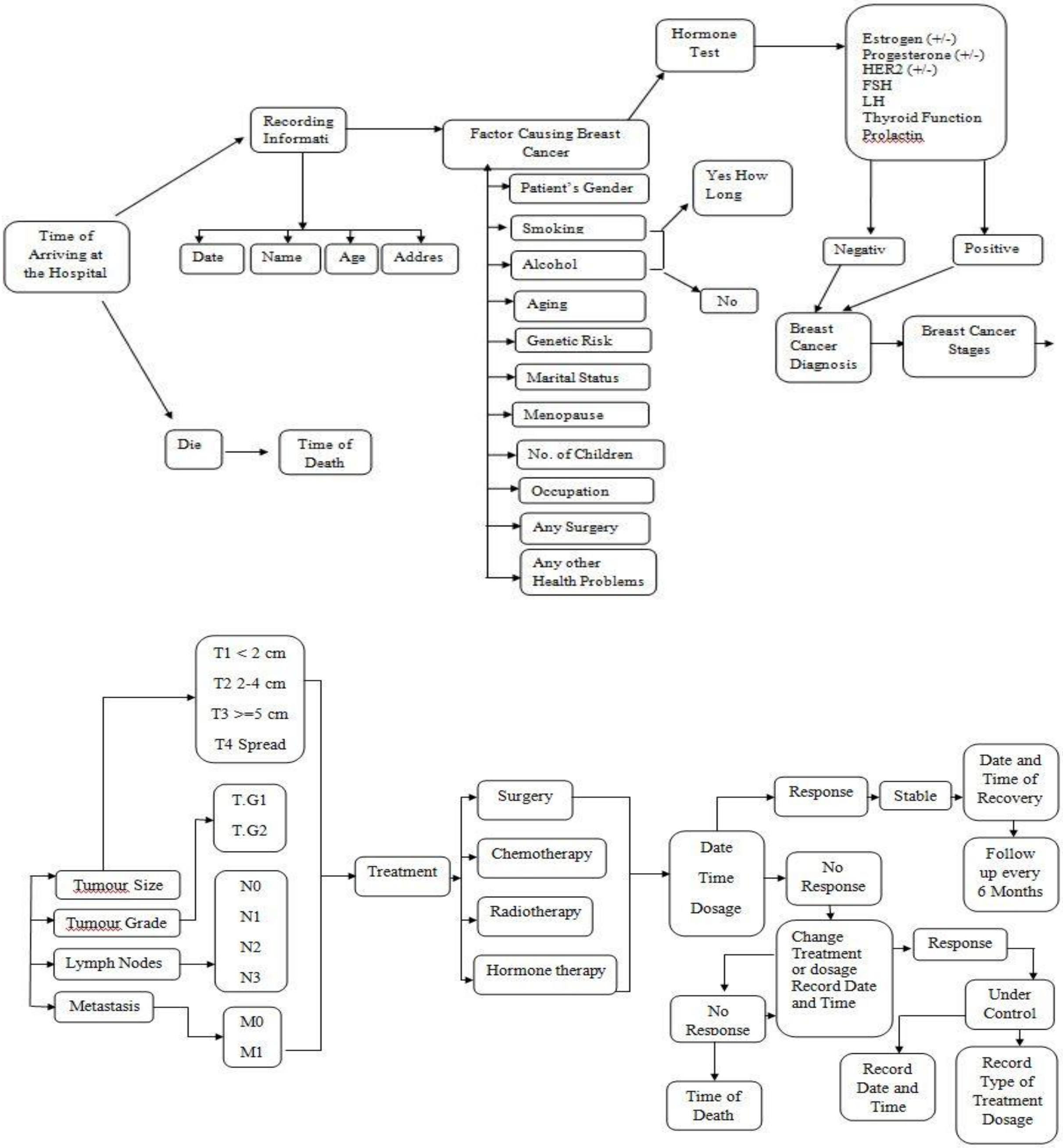


Figure 7-2 Specific flow chart for data collection design patient (individual)

7.9.2 Explanation of the flow charts

In UK hospitals, they follow electronic health records (EHRs), developing and using data that are recorded consistently and electronically across all contexts within a national standard. The implementation of a national standard can help facilitate shared care regardless of location and context, and give comparable data to support nationwide management and monitoring of health services. This in turn will benefit patients, clinicians, professions and the health service in general. In Kurdish hospitals, the traditional way (which is still widely adhered to) does not involve obtaining inclusive information about the patient's status. For instance, while in Nanakally hospital the patient's form features boxes for circumstantial information (such as age, gender, weight, height, allergy, diagnosis, stage of tumour, chemotherapy protocol, number of cycles) and treatment details (chemotherapeutic agent given, dose, way of delivery); frequently this information is not transferred into the database or is simply left empty altogether. We chose our flow chart of data collection design based on the UK hospitals method for collecting data, because it is very comprehensive. However, while in the UK the data is collected on five separate forms specialized to different areas, we would like to apply this design for Kurdish hospitals using only a single form while remaining thorough and applicable to the Kurdistan environment and culture, in order to make data collection easier for future researchers.

7.10 Feasibility (achievability) of the plan

It is practical for Hospitals in the Kurdistan Region to follow the Health and Social Care Information Centre (HSCIC) in the United Kingdom as they have launched a programme to look at other ways in which data can be securely accessed. The Data Access Request Service (DARS) provides customers with a single point of access for all new data applications. The information control panel has been created to provide greater visibility of the type and volume of requested data that are needed. The control panel shows the level of requested data that are necessary for its application and the displayed information is updated regularly.

Furthermore the policy of accessing the data in HSCIC is supported by the following principles:

1. Share information to support the provision of health and social care and the promotion of health.

2. Audit the data receiver to ensure loyalty to the terms of their contracts and agreements; ensure that any data are deleted at the end of the data sharing agreement under which it was released.
3. Provide a clear, simple, efficient, and transparent service for access to data delivered to publicly-stated service levels.
4. Use the latest technologies to create interconnected records as a means to providing safe and secure access to data.
5. Listen to feedback from customers, patients, and the public and involve them in the continuous improvement of the HSCIC system.

Following this plan is well-suited to create a strong health system and easy access to the accurate data if it is properly implemented in the Kurdistan Hospitals. Still, there are many more aspects to contemplate before putting this plan into action. The main problem might be the lack of awareness of the importance of this plan among authorities responsible for the funding. Another problem this plan could face is a lack of compliance by the doctors themselves; since many of them are using the most basic techniques to record patient's information it might take some time to effectively popularize the use of computers and databases when it comes to collecting and storing data. Other obstacles include the ignorance of patient's histories, furthermore it is sometimes regarded as a violation of trust to get correct information from the patients. The main point here is that a lack of trust or collaboration between doctor and patient may discourage objective bookkeeping. That in itself causes many of the problems we face when dealing with patients' records. Since a GP employed at a government hospital is generally paid relatively little, many of them work in private clinics in the afternoons, which only few people can afford to visit. On the other hand there is some competition between the doctors as a significant number of them came to the region only recently from the other parts of Iraq. The aforementioned obstacles are not the only ones. Even in the case of extensive government support, there is still a need for experts in statistics and software who are trained in the use of the equipment and the analysis of the resulting data.
In addition to that, the hospitals in the region generally do not meet the standards of modern healthcare. While many hospital buildings themselves may be neat and tidy, the patient's rooms are often damp and dark as opposed to light and airy, especially the waiting rooms.

The achievability of the suggested general flow chart for data collection relies heavily on the method of data collection; gathering information through personal interviews seems much more promising than through questionnaires, which might be answered incompletely or incorrectly. At the same time, choosing a database structure when collecting the data might be straightforward to do for some categories such as breast health history, case management, risk assessment while many patients may be unable to give the exact date of their last physical examinations or referrals, or the precise type of medication and dosage they have been given. Focus on the interviewing stage of the data collection process is indispensable in order to gain accurate data at the end.

In the following, we address some of the challenges which may arise during the initial phase of data collection, i.e. when the patient arrives at the hospital for the first time. Here, the first two steps (regarding recording basic information or the time of death, respectively) may be straightforward, while in step three, when it comes to assessing various factors causing breast cancer, the patient, for cultural reasons or lack of medical understanding, might take issue with the questions about smoking habits, alcohol consumption, age, genetic risk, menopause and so on, and choose to give inaccurate answers or entirely refuse to answer them. When we move on to more medical recording issues, starting with hormone tests, it may not be easy to get all the information because the doctors may not request all the tests to diagnose the breast cancer. In the case when breast cancer is diagnosed, there will be no problem in recording the information about the tumour size, tumour grade, lymph nodes and metastasis. Unfortunately, it may be difficult to actually treat the breast cancer because as soon as the patients are made aware of their illness, many of them go abroad looking for treatments, which means that the records will remain blank. Because of limited availability of expensive specific treatments to poor patients, they might serve as a comparison group. Finally, we also might face problems in recording the date and time of change of treatment type and dosages.

8 CHAPTER 8: Conclusion and future work

8.1 Conclusion

All over the world, especially in the developing regions, the rates of breast cancer, the most common malignancy in women, constituting just under one fifth of cancers in females, are on the rise. Kurdistan-Iraq is no exception, with an age-adjusted incidence rate of 68.9 per 100,000 year; in fact, breast cancer is the most prevalent cancer among the population (affecting about one third of female cancer patients), with particularly alarming rates among the younger demographic, according to the Kurdistan-Iraqi Cancer Registry. In order to tackle this problem, a precise understanding of the survival rate is essential. In order to do so, in this work we adopt the Cox regression and the Kaplan-Meier methods.

The main conclusion is that we have developed a new method for performing a survival analysis on a set of data where there are important unknown factors; namely hidden censoring of the data, so that the number of individuals apparently at risk is greater than those actually at risk. In particular we have shown how to adjust a Kaplan-Meier analysis to find a survival curve in such circumstances, and also shown how to estimate a true hazard (survivor) function from the biased one obtained directly from the data. For Nanakaly and Hewa data we generated a new model in two cases; with and without censoring. For without censoring in Nanakaly data we estimate a number of observations while for Hewa data we estimate the number of observation and the number of deaths. Examining the results for our data, we conclude that the survival rate of breast cancer in Erbil and Suleimania are lower at age 48 and above years. The findings of the present study suggest that age, smoking, estrogen receptors and tumour grade have an effect on breast cancer survival.

In order to ascertain the validity of the models we constructed, we considered different simulation techniques applied to the Nanakaly data. Because of the availability of a good survival function, we chose to work with a German data set. For each different simulation method, a distribution was chosen and the 'lost' patients were subsequently simulated from this distribution. Thus, death is only observed if the individual has not been 'lost', otherwise it is not. We see that our models perform well in many circumstances, and even when less accurate, are always an improvement on considering the apparent survival curves from the unadjusted data. For the Hewa data we need to estimate crucial parameter values. For some estimates we get realistic survival function curves. However, estimates are made with little information. Thus while survival curves are plausible we cannot rely on them.

As mentioned above, the data we work with, provided by the Nanakaly and Hewa hospitals, generally does not meet the standard of comprehensive data collection adopted e.g. in the UK. While in Britain, five different forms are in use in order to capture every accessible piece of information, we attempted to condense this into one form intended for use in Kurdistan-Iraq while keeping it as detailed as possible. The intention is to adopt the system which is in use in western countries, where the recording of data is the responsibility of the GP's, who generally are already well-versed in health education. The process of data collection is intended to start at a patient's first visit to a practitioner's surgery. Further details can be found in Figures 7.1 and 7.2 (chapter 7).

While the models from chapter 5 (with and without censoring) in question are easily implemented and don't constitute a substantial workload increase for the doctors, they still are well-suited to the task of keeping track of a patient's health records, including dealing with hidden or censored data. However, it is necessary to obtain government funding, a highly trained staff and the statistical expertise in order to fully implement the proposed models.

8.2 Future research

This study provided an understanding of the factors identified by primary care providers that negatively affect the primary care system in the Kurdistan region of Iraq. Primary care providers have a major role to play as shown by experience in distributing information in the community and in informing governing bodies about the main problems affecting the system. Obviously, recommendations for improvements to the health care system in general must be taken to a national level for a more comprehensive strategy for improving primary health care in Iraq. Therefore future validation of the suggested models may be conducted using cross validation or by using new data. Applying the recommended flow charts could provide more relevant information on breast cancer, thus providing a more comprehensive understanding of breast cancer. With the increased number of variables and need to identified attributable variables, taking into account the extensive use of the accelerated death model in research and literature, future versions of statistical analysis software such as SPSS or R may implement the needed models.

In this report a new method for performing a survival analysis on a dataset with important unknown factors is developed. Specifically, the system of data collection we are dealing with exhibits hidden censoring, making the number of individuals at risk appears greater than it actually is. To validate our methods, we applied them to data provided by German health services, for which an accurate survival function is known. For each simulation, we chose a distribution and artificially implemented censoring of individuals from the comparison dataset according to it. Subsequent research might include the systematic investigation of such models and how they react to censoring of the input data. Extensions might include the use of Cure models which model the survival time taking into account the nonzero probability of a patient's total recovery, which can be assumed to account for some of the missing records.

Appendix

A1: SPSS Hewa data analysis: these are intermediate tables for the analysis from Chapter 6, Section 6.2.

Table A1.1 Variables in the Cox regression model for the Hewa data

	B	SE	Wald	df	*p*. value	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
Progesterone Receptor	-.001	.001	3.571	1	.059	.999	.997	1.000
Estrogen Receptor	-.002	.001	3.181	1	.075	.998	.996	1.000
Menopause	-.217	.243	.802	1	.371	.805	.500	1.295
Hormone	.421	.232	3.301	1	.069	1.524	.967	2.401
Tumour size	-.004	.006	.589	1	.443	.996	.985	1.007
Lymph Nodes	-.012	.012	.952	1	.329	.988	.964	1.012
Religion Code			2.572	2	.276			
Religion Code(1)	.922	.819	1.265	1	.261	2.513	.504	12.521
Religion Code(2)	-.309	.419	.542	1	.462	.735	.323	1.670
Smoking Code	-.541	.293	3.414	1	.065	.582	.328	1.033
Drinking Code	-1.119	.795	1.980	1	.159	.327	.069	1.552
Weight	.002	.002	.973	1	.324	1.002	.998	1.006
Height	-.007	.012	.301	1	.583	.993	.969	1.018
BMI	-.001	.002	.119	1	.730	.999	.995	1.004
Family History Code	.270	.307	.776	1	.378	1.310	.718	2.391
Income Code			4.922	3	.178			
Income Code(1)	-.765	.447	2.927	1	.087	.465	.194	1.118
Income Code(2)	-.341	.281	1.479	1	.224	.711	.410	1.232
Income Code(3)	-.505	.255	3.921	1	.048	.604	.366	.995
Marital Status Code			4.964	3	.174			
Marital Status Code(1)	.077	.467	.027	1	.869	1.080	.432	2.700
Marital Status Code(2)	.524	.238	4.849	1	.028	1.689	1.059	2.693
Marital Status Code(3)	.097	.445	.048	1	.827	1.102	.461	2.637
Exercise Code	.075	.197	.146	1	.703	1.078	.732	1.588
Breast Feeding Code	.155	.258	.363	1	.547	1.168	.705	1.935
Tumour Grade Code			8.489	2	.014			
Tumour Grade Code(1)	-.641	.306	4.391	1	.036	.527	.289	.959
Tumour Grade Code(2)	-.053	.263	.041	1	.839	.948	.566	1.587

Where the column headers represent the following:

B: Parameter Coefficient, SE: Standard Error, Wald: Statistics Test, df: Degree of Freedom, *p*-value: Significant Value, Exp(B): Exponential Parameter, CI: Confidence Interval.

Table A1.2 Variables in the Cox regression model for the Hewa data

	B	SE	Wald	df	*p*. value	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
Progesterone Receptor	-.001	.001	3.570	1	.059	.999	.997	1.000
Estrogen Receptor	-.002	.001	3.199	1	.074	.998	.996	1.000
Menopause	-.217	.243	.797	1	.372	.805	.500	1.296
Hormone	.423	.232	3.318	1	.069	1.526	.968	2.405
Tumour Size	-.004	.006	.608	1	.435	.996	.985	1.007
Lymph Nodes	-.012	.012	.962	1	.327	.988	.964	1.012
Religion Code			2.545	2	.280			
Religion Code(1)	.913	.819	1.242	1	.265	2.492	.500	12.410
Religion Code(2)	-.309	.419	.544	1	.461	.734	.323	1.669
Smoking Code	-.540	.293	3.391	1	.066	.583	.328	1.035
Drinking Code	-1.113	.795	1.958	1	.162	.329	.069	1.562
Weight	.002	.002	.848	1	.357	1.002	.998	1.006
Height	-.004	.008	.245	1	.621	.996	.981	1.012
Family History Code	.265	.306	.751	1	.386	1.304	.716	2.376
Income Code			5.010	3	.171			
Income Code(1)	-.771	.447	2.977	1	.084	.462	.193	1.111
Income Code(2)	-.344	.280	1.506	1	.220	.709	.409	1.228
Income Code(3)	-.509	.255	3.990	1	.046	.601	.365	.990
Marital Status Code			4.984	3	.173			
Marital Status Code(1)	.066	.467	.020	1	.888	1.068	.427	2.670
Marital Status Code(2)	.524	.238	4.859	1	.027	1.689	1.060	2.693
Marital Status Code(3)	.100	.445	.051	1	.822	1.106	.462	2.644
Exercise Code	.071	.197	.130	1	.718	1.074	.730	1.580
Breast Feeding Code	.155	.258	.361	1	.548	1.167	.705	1.934
Tumour Grade Code			8.568	2	.014			
Tumour Grade Code(1)	-.642	.306	4.409	1	.036	.526	.289	.958
Tumour Grade Code(2)	-.051	.263	.038	1	.846	.950	.568	1.590

Table A1.3 Variables in the Cox regression model for the Hewa data

	B	SE	Wald	df	*p*. value	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
Progesterone Receptor	-.001	.001	3.555	1	.059	.999	.997	1.000
Estrogen Receptor	-.002	.001	3.136	1	.077	.998	.996	1.000
Menopause	-.214	.242	.783	1	.376	.807	.502	1.297
Hormone	.440	.226	3.791	1	.052	1.553	.997	2.420
Tumour Size	-.005	.006	.629	1	.428	.996	.984	1.007
Lymph Nodes	-.013	.012	1.017	1	.313	.988	.964	1.012
Religion Code			2.459	2	.292			
Religion Code(1)	.895	.817	1.199	1	.274	2.447	.493	12.143
Religion Code(2)	-.301	.418	.517	1	.472	.740	.326	1.681
Smoking Code	-.546	.293	3.458	1	.063	.580	.326	1.030
Drinking Code	-1.085	.792	1.879	1	.170	.338	.072	1.594
Weight	.002	.002	.801	1	.371	1.002	.998	1.006
Height	-.004	.008	.244	1	.622	.996	.981	1.011
Family History Code	.259	.306	.720	1	.396	1.296	.712	2.360
Income Code			5.470	3	.140			
Income Code(1)	-.792	.444	3.186	1	.074	.453	.190	1.081
Income Code(2)	-.362	.276	1.722	1	.189	.696	.405	1.196
Income Code(3)	-.526	.250	4.425	1	.035	.591	.362	.965
Marital Status Code			4.907	3	.179			
Marital Status Code(1)	.077	.466	.028	1	.868	1.081	.433	2.695
Marital Status Code(2)	.520	.238	4.793	1	.029	1.683	1.056	2.681
Marital Status Code(3)	.107	.444	.058	1	.809	1.113	.466	2.660
Breast Feeding Code	.145	.256	.320	1	.572	1.156	.700	1.907
Tumour Grade Code			8.556	2	.014			
Tumour Grade Code(1)	-.642	.306	4.407	1	.036	.526	.289	.958
Tumour Grade Code(2)	-.052	.263	.038	1	.845	.950	.568	1.589

Table A1.4 Variables in the Cox regression model for the Hewa data

	B	SE	Wald	df	*p*. value	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
Progesterone Receptor	-.001	.001	3.744	1	.053	.999	.997	1.000
Estrogen Receptor	-.002	.001	3.067	1	.080	.998	.996	1.000
Menopause	-.212	.242	.770	1	.380	.809	.504	1.299
Hormone	.434	.226	3.682	1	.055	1.544	.991	2.405
Tumour Size	-.004	.006	.598	1	.439	.996	.985	1.007
Lymph Nodes	-.012	.012	.982	1	.322	.988	.964	1.012
Religion Code			2.499	2	.287			
Religion Code(1)	.884	.818	1.169	1	.280	2.421	.487	12.024
Religion Code(2)	-.315	.418	.567	1	.452	.730	.321	1.657
Smoking Code	-.544	.295	3.415	1	.065	.580	.326	1.034
Drinking Code	-1.081	.792	1.860	1	.173	.339	.072	1.604
Weight	.002	.002	.787	1	.375	1.002	.998	1.005
Family History Code	.254	.305	.691	1	.406	1.289	.708	2.344
Income Code			5.472	3	.140			
Income Code(1)	-.791	.444	3.180	1	.075	.453	.190	1.082
Income Code(2)	-.360	.276	1.698	1	.193	.698	.406	1.199
Income Code(3)	-.526	.250	4.418	1	.036	.591	.362	.965
Marital Status Code			4.954	3	.175			
Marital Status Code(1)	.090	.466	.038	1	.846	1.095	.439	2.728
Marital Status Code(2)	.524	.238	4.855	1	.028	1.688	1.060	2.690
Marital Status Code(3)	.116	.444	.068	1	.795	1.122	.470	2.682
Breast Feeding Code	.149	.255	.339	1	.560	1.160	.704	1.912
Tumour Grade Code			8.528	2	.014			
Tumour Grade Code(1)	-.643	.305	4.443	1	.035	.526	.289	.956
Tumour Grade Code(2)	-.055	.262	.045	1	.833	.946	.566	1.582

Table A1.5 Variables in the Cox regression model for the Hewa data

	B	SE	Wald	df	*p*. value	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
Progesterone Receptor	-.001	.001	3.862	1	.049	.999	.997	1.000
Estrogen Receptor	-.002	.001	2.936	1	.087	.998	.996	1.000
Menopause	-.164	.227	.521	1	.470	.849	.544	1.325
Hormone	.423	.226	3.516	1	.061	1.527	.981	2.376
Tumour Size	-.004	.006	.612	1	.434	.996	.985	1.007
Lymph Nodes	-.011	.012	.827	1	.363	.989	.965	1.013
Religion Code			2.388	2	.303			
Religion Code(1)	.861	.816	1.112	1	.292	2.365	.477	11.715
Religion Code(2)	-.308	.418	.542	1	.462	.735	.324	1.668
Smoking Code	-.537	.294	3.343	1	.067	.585	.329	1.039
Drinking Code	-1.057	.791	1.789	1	.181	.347	.074	1.636
Weight	.002	.002	.767	1	.381	1.002	.998	1.005
Family History Code	.252	.305	.681	1	.409	1.286	.707	2.340
Income Code			5.363	3	.147			
Income Code(1)	-.788	.443	3.158	1	.076	.455	.191	1.085
Income Code(2)	-.353	.276	1.635	1	.201	.703	.409	1.207
Income Code(3)	-.517	.250	4.289	1	.038	.596	.366	.973
Marital Status Code			5.636	3	.131			
Marital Status Code(1)	.122	.463	.070	1	.792	1.130	.456	2.802
Marital Status Code(2)	.442	.195	5.140	1	.023	1.556	1.062	2.281
Marital Status Code(3)	.147	.441	.110	1	.740	1.158	.488	2.750
Tumour Grade Code			8.372	2	.015			
Tumour Grade Code(1)	-.639	.305	4.386	1	.036	.528	.290	.960
Tumour Grade Code(2)	-.058	.263	.048	1	.826	.944	.564	1.579

Table A1.6 Variables in the Cox regression model for the Hewa data

	B	SE	Wald	df	*p*. value	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
Progesterone Receptor	-.001	.001	3.620	1	.057	.999	.997	1.000
Estrogen Receptor	-.002	.001	3.326	1	.068	.998	.996	1.000
Hormone	.336	.192	3.068	1	.080	1.399	.961	2.036
Tumour Size	-.004	.006	.521	1	.470	.996	.985	1.007
Lymph Nodes	-.013	.012	1.121	1	.290	.987	.964	1.011
Religion Code			2.486	2	.288			
Religion Code(1)	.883	.819	1.162	1	.281	2.418	.486	12.042
Religion Code(2)	-.313	.418	.561	1	.454	.731	.322	1.659
Smoking Code	-.526	.292	3.248	1	.071	.591	.333	1.047
Drinking Code	-1.085	.792	1.875	1	.171	.338	.072	1.597
Weight	.002	.002	.759	1	.384	1.002	.998	1.005
Family History Code	.262	.305	.740	1	.390	1.300	.715	2.362
Income Code			5.260	3	.154			
Income Code(1)	-.779	.444	3.085	1	.079	.459	.192	1.094
Income Code(2)	-.354	.276	1.642	1	.200	.702	.409	1.206
Income Code(3)	-.514	.250	4.233	1	.040	.598	.367	.976
Marital Status Code			6.376	3	.095			
Marital Status Code(1)	.167	.459	.132	1	.716	1.182	.480	2.909
Marital Status Code(2)	.470	.192	6.011	1	.014	1.599	1.099	2.328
Marital Status Code(3)	.183	.440	.173	1	.677	1.201	.507	2.847
Tumour Grade Code			8.720	2	.013			
Tumour Grade Code(1)	-.655	.304	4.636	1	.031	.519	.286	.943
Tumour Grade Code(2)	-.064	.262	.060	1	.806	.938	.561	1.568

Table A1.7 Variables in the Cox regression model for the Hewa data

	B	SE	Wald	df	*p.* value	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
Progesterone Receptor	-.001	.001	3.797	1	.051	.999	.997	1.000
Estrogen Receptor	-.002	.001	3.279	1	.070	.998	.996	1.000
Hormone	.328	.191	2.937	1	.087	1.388	.954	2.019
Lymph Nodes	-.012	.012	.990	1	.320	.988	.965	1.012
Religion Code			2.635	2	.268			
Religion Code(1)	.899	.819	1.205	1	.272	2.457	.494	12.235
Religion Code(2)	-.327	.418	.613	1	.434	.721	.318	1.635
Smoking Code	-.515	.291	3.139	1	.076	.598	.338	1.056
Drinking Code	-1.117	.791	1.994	1	.158	.327	.069	1.542
Weight	.002	.002	.640	1	.424	1.002	.998	1.005
Family History Code	.261	.305	.736	1	.391	1.299	.715	2.360
Income Code			5.181	3	.159			
Income Code(1)	-.761	.443	2.957	1	.085	.467	.196	1.112
Income Code(2)	-.352	.276	1.629	1	.202	.703	.409	1.208
Income Code(3)	-.513	.250	4.229	1	.040	.598	.367	.976
Marital Status Code			6.162	3	.104			
Marital Status Code(1)	.171	.459	.139	1	.709	1.187	.483	2.918
Marital Status Code(2)	.460	.191	5.810	1	.016	1.585	1.090	2.304
Marital Status Code(3)	.171	.440	.151	1	.697	1.187	.501	2.808
Tumour Grade Code			9.066	2	.011			
Tumour Grade Code(1)	-.673	.303	4.917	1	.027	.510	.282	.925
Tumour Grade Code(2)	-.074	.262	.079	1	.779	.929	.556	1.553

Table A1.8 Variables in the Cox regression model for the Hewa data

	B	SE	Wald	df	*p.* value	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
Progesterone Receptor	-.001	.001	3.873	1	.049	.999	.997	1.000
Estrogen Receptor	-.002	.001	3.249	1	.071	.998	.996	1.000
Hormone	.325	.191	2.903	1	.088	1.385	.952	2.013
Lymph Nodes	-.012	.012	.975	1	.324	.988	.965	1.012
Religion Code			2.636	2	.268			
Religion Code(1)	.901	.819	1.210	1	.271	2.462	.495	12.257
Religion Code(2)	-.327	.418	.610	1	.435	.721	.318	1.637
Smoking Code	-.515	.291	3.140	1	.076	.598	.338	1.056
Drinking Code	-1.113	.791	1.981	1	.159	.329	.070	1.548
Family History Code	.263	.305	.748	1	.387	1.301	.716	2.364
Income Code			5.010	3	.171			
Income Code(1)	-.752	.442	2.886	1	.089	.472	.198	1.122
Income Code(2)	-.350	.276	1.606	1	.205	.705	.411	1.211
Income Code(3)	-.503	.249	4.082	1	.043	.605	.371	.985
Marital Status Code			6.089	3	.107			
Marital Status Code(1)	.160	.459	.122	1	.727	1.174	.478	2.883
Marital Status Code(2)	.457	.191	5.722	1	.017	1.579	1.086	2.295
Marital Status Code(3)	.166	.439	.143	1	.705	1.181	.499	2.794
Tumour Grade Code			9.200	2	.010			
Tumour Grade Code(1)	-.673	.303	4.928	1	.026	.510	.281	.924
Tumour Grade Code(2)	-.069	.262	.070	1	.791	.933	.558	1.559

Table A1.9 Variables in the Cox regression model for the Hewa data

	B	SE	Wald	df	*p*. value	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
Progesterone Receptor	-.001	.001	3.947	1	.047	.999	.997	1.000
Estrogen Receptor	-.002	.001	3.207	1	.073	.998	.996	1.000
Hormone	.321	.191	2.823	1	.093	1.379	.948	2.007
Lymph Nodes	-.012	.012	1.022	1	.312	.988	.965	1.011
Religion Code			2.686	2	.261			
Religion Code(1)	.912	.819	1.241	1	.265	2.489	.500	12.389
Religion Code(2)	-.328	.418	.615	1	.433	.721	.318	1.635
Smoking Code	-.513	.290	3.134	1	.077	.599	.339	1.056
Drinking Code	-1.131	.789	2.054	1	.152	.323	.069	1.515
Income Code			5.222	3	.156			
Income Code(1)	-.769	.442	3.030	1	.082	.463	.195	1.102
Income Code(2)	-.379	.274	1.909	1	.167	.685	.400	1.172
Income Code(3)	-.518	.249	4.331	1	.037	.596	.366	.970
Marital Status Code			6.126	3	.106			
Marital Status Code(1)	.151	.459	.109	1	.741	1.163	.474	2.858
Marital Status Code(2)	.457	.191	5.740	1	.017	1.580	1.087	2.296
Marital Status Code(3)	.166	.440	.143	1	.705	1.181	.499	2.796
Tumour Grade Code			9.499	2	.009			
Tumour Grade Code(1)	-.679	.304	5.003	1	.025	.507	.280	.919
Tumour Grade Code(2)	-.064	.262	.060	1	.806	.938	.561	1.567

Table A1.10 Variables in the Cox regression model for the Hewa data

	B	SE	Wald	df	*p*. value	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
Progesterone Receptor	-.001	.001	3.878	1	.049	.999	.997	1.000
Estrogen Receptor	-.002	.001	3.727	1	.054	.998	.995	1.000
Hormone	.294	.190	2.404	1	.121	1.342	.925	1.947
Religion Code			2.734	2	.255			
Religion Code(1)	.877	.820	1.144	1	.285	2.404	.482	11.995
Religion Code(2)	-.359	.417	.738	1	.390	.699	.308	1.583
Smoking Code	-.533	.290	3.380	1	.066	.587	.332	1.036
Drinking Code	-1.101	.791	1.936	1	.164	.333	.071	1.568
Income Code			5.280	3	.152			
Income Code(1)	-.781	.442	3.122	1	.077	.458	.193	1.089
Income Code(2)	-.388	.274	2.003	1	.157	.678	.396	1.161
Income Code(3)	-.520	.249	4.362	1	.037	.595	.365	.969
Marital Status Code			6.002	3	.112			
Marital Status Code(1)	.171	.459	.139	1	.709	1.187	.483	2.915
Marital Status Code(2)	.454	.191	5.664	1	.017	1.575	1.083	2.290
Marital Status Code(3)	.171	.441	.150	1	.698	1.187	.500	2.817
Tumour Grade Code			9.857	2	.007			
Tumour Grade Code(1)	-.709	.302	5.496	1	.019	.492	.272	.890
Tumour Grade Code(2)	-.089	.261	.115	1	.735	.915	.549	1.527

Table A1.11 Variables in the Cox regression model for the Hewa data

	B	SE	Wald	df	*p*. value	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
Progesterone Receptor	-.001	.001	3.758	1	.053	.999	.997	1.000
Estrogen Receptor	-.002	.001	3.770	1	.052	.998	.995	1.000
Hormone	.299	.190	2.475	1	.116	1.348	.929	1.955
Smoking Code	-.538	.289	3.460	1	.063	.584	.331	1.029
Drinking Code	-.021	.307	.005	1	.945	.979	.536	1.788
Income Code			4.964	3	.174			
Income Code(1)	-.780	.441	3.126	1	.077	.458	.193	1.088
Income Code(2)	-.394	.274	2.067	1	.151	.674	.394	1.154
Income Code(3)	-.496	.248	4.006	1	.045	.609	.374	.990
Marital Status Code			6.524	3	.089			
Marital Status Code(1)	.196	.458	.182	1	.669	1.216	.495	2.985
Marital Status Code(2)	.479	.190	6.353	1	.012	1.615	1.112	2.344
Marital Status Code(3)	.323	.420	.594	1	.441	1.382	.607	3.146
Tumour Grade Code			9.722	2	.008			
Tumour Grade Code(1)	-.698	.302	5.336	1	.021	.498	.275	.900
Tumour Grade Code(2)	-.080	.261	.093	1	.760	.923	.554	1.540

Table A1.12 Variables in the Cox regression model for the Hewa data

	B	SE	Wald	df	*p*. value	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
Progesterone Receptor	-.001	.001	3.756	1	.053	.999	.997	1.000
Estrogen Receptor	-.002	.001	3.790	1	.052	.998	.995	1.000
Hormone	.298	.190	2.470	1	.116	1.347	.929	1.954
Smoking Code	-.552	.199	7.661	1	.006	.576	.389	.851
Income Code			4.963	3	.175			
Income Code(1)	-.781	.441	3.128	1	.077	.458	.193	1.088
Income Code(2)	-.394	.274	2.066	1	.151	.674	.394	1.154
Income Code(3)	-.496	.248	4.004	1	.045	.609	.374	.990
Marital Status Code			6.525	3	.089			
Marital Status Code(1)	.197	.458	.184	1	.668	1.217	.496	2.987
Mairtal Status Code(2)	.479	.190	6.356	1	.012	1.615	1.113	2.345
Mairtal Status Code(3)	.323	.420	.593	1	.441	1.382	.607	3.146
Tumour Grade Code			9.721	2	.008			
Tumour Grade Code(1)	-.697	.302	5.331	1	.021	.498	.276	.900
Tumour Grade Code(2)	-.080	.261	.093	1	.760	.923	.553	1.540

Table A1.13 Variables in the Cox regression model for the Hewa data

	B	SE	Wald	df	*p*. value	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
Progesterone Receptor	-.001	.001	3.267	1	.071	.999	.997	1.000
Estrogen Receptor	-.002	.001	3.606	1	.058	.998	.995	1.000
Hormone	.290	.190	2.332	1	.127	1.336	.921	1.937
Smoking Code	-.543	.199	7.424	1	.006	.581	.393	.859
Marital Status Code			5.704	3	.127			
Marital Status Code(1)	.148	.457	.105	1	.746	1.160	.474	2.838
Marital Status Code(2)	.444	.189	5.502	1	.019	1.558	1.076	2.257
Marital Status Code(3)	.295	.419	.496	1	.481	1.343	.591	3.051
Tumour Grade Code			9.318	2	.009			
Tumour Grade Code(1)	-.664	.300	4.907	1	.027	.515	.286	.926
Tumour Grade Code(2)	-.058	.260	.050	1	.823	.944	.567	1.571

Table A1.14 Variables in the Cox regression model for the Hewa data

	B	SE	Wald	df	*p*. value	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
Progesterone Receptor	-.001	.001	2.763	1	.096	.999	.998	1.000
Estrogen Receptor	-.002	.001	4.159	1	.041	.998	.995	1.000
Hormone	.229	.179	1.631	1	.202	1.257	.885	1.785
Smoking Code	-.507	.199	6.523	1	.011	.602	.408	.889
Tumour Grade Code			10.821	2	.004			
Tumour Grade Code(1)	-.702	.300	5.496	1	.019	.495	.275	.891
Tumour Grade Code(2)	-.049	.260	.035	1	.852	.952	.572	1.587

Table A1.15 Variables in the Cox regression model for the Hewa data

	B	SE	Wald	df	*p*. value	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
Progesterone Receptor	-.001	.001	2.657	1	.103	.999	.998	1.000
Estrogen Receptor	-.002	.001	3.639	1	.056	.998	.995	1.000
Smoking Code	-.534	.198	7.317	1	.007	.586	.398	.863
Tumour Grade Code			10.130	2	.006			
Tumour Grade Code(1)	-.666	.299	4.965	1	.026	.514	.286	.923
Tumour Grade Code(2)	-.033	.260	.016	1	.899	.967	.581	1.612

Table A1.16 Variables in the Cox regression model for the Hewa data

	B	SE	Wald	df	*p*. value	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
Estrogen Receptor	-.003	.001	5.339	1	.021	.997	.995	1.000
Smoking Code	-.540	.197	7.490	1	.006	.583	.396	.858
Tumour Grade Code			9.000	2	.011			
Tumour Grade Code(1)	-.578	.294	3.869	1	.049	.561	.315	.998
Tumour Grade Code(2)	.022	.258	.007	1	.933	1.022	.616	1.696

A2: SPSS German data analysis: these are intermediate tables for the analysis from Chapter 6, section 6.3.

Table A2.1 Variables in the Cox regression model for the German data

	B	SE	Wald	df	*p*. value	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
Age	.009	.008	1.438	1	.230	1.009	.994	1.024
Hormone	.270	.168	2.579	1	.108	1.309	.942	1.819
Tumour Size	.013	.005	7.890	1	.005	1.014	1.004	1.023
Tumour Grade Code			8.403	2	.015			
Tumour Grade Code (1)	-1.130	.442	6.537	1	.011	.323	.136	.768
Tumour Grade Code(2)	-.344	.168	4.175	1	.041	.709	.510	.986
Lymph Nodes	.051	.009	30.106	1	.000	1.053	1.034	1.072
Progesterone Receptor	-.005	.001	21.894	1	.000	.995	.992	.997

Table A2.2 Variables in the Cox regression model for the German data

	B	SE	Wald	df	*p*. value	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
Hormone	.230	.165	1.951	1	.162	1.259	.911	1.739
Tumour Size	.013	.005	7.714	1	.005	1.013	1.004	1.023
Tumour Grade Code			8.090	2	.018			
Tumour Grade Code(1)	-1.115	.442	6.360	1	.012	.328	.138	.780
Tumour Grade Code(2)	-.334	.168	3.937	1	.047	.716	.515	.996
Lymph Nodes	.052	.009	30.151	1	.000	1.053	1.034	1.073
Progesterone Receptor	-.005	.001	21.660	1	.000	.995	.992	.997

Table A2.3 Variables in the Cox regression model for the German data

	B	SE	Wald	df	*p*. value	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
Tumour Size	.013	.005	8.130	1	.004	1.014	1.004	1.023
Tumour Grade Code			8.958	2	.011			
Tumour Grade Code(1)	-1.150	.441	6.798	1	.009	.317	.133	.752
Tumour Grade Code(2)	-.359	.167	4.605	1	.032	.698	.503	.969
Lymph Nodes	.051	.009	29.193	1	.000	1.053	1.033	1.072
Progesterone Receptor	-.005	.001	21.566	1	.000	.995	.992	.997

A3: The following figures represent simulations of various difference loss distribution there are ten simulations for an exponential distribution with means equal to 2000, 1000 and 500, respectively and ten for a gamma distribution with mean 1000 (parameters 3,3/1000).

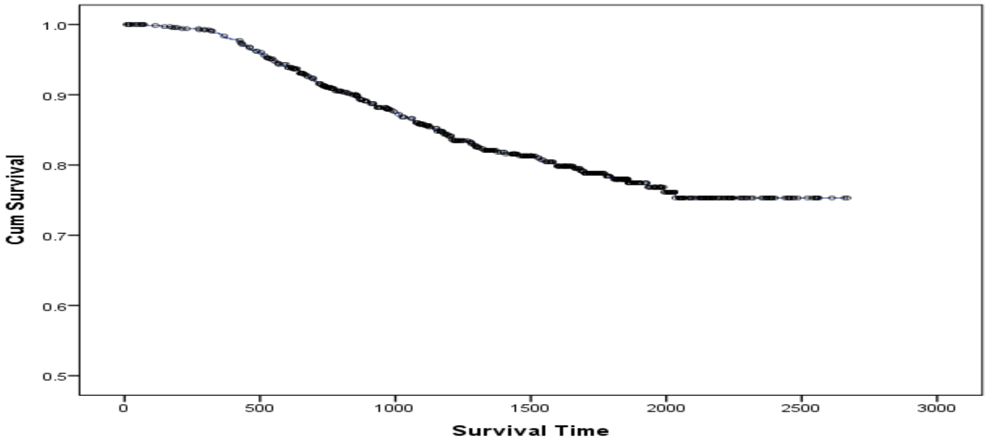


Figure A3.1 Uncorrected cumulative survival for exponential loss distribution using mean 2000 simulation a1

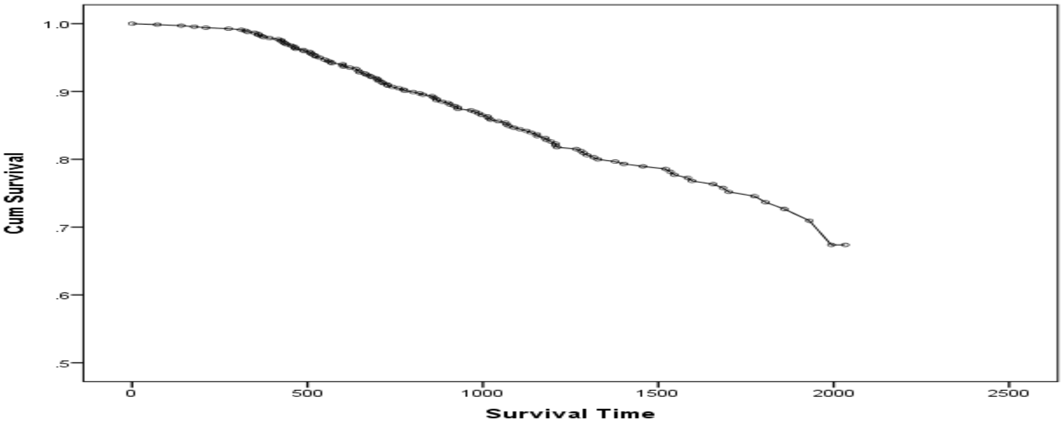


Figure A3.2 Cumulated survival for exponential loss distribution using mean 2000: corrected using the first model simulation a1

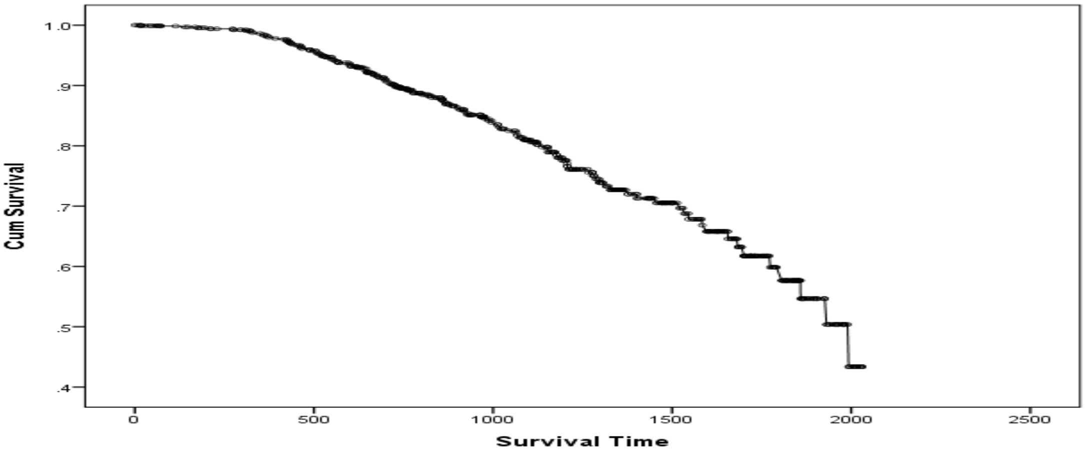


Figure A3.3 Cumulated survival for exponential loss distribution using mean 2000: corrected using the second model simulation a1

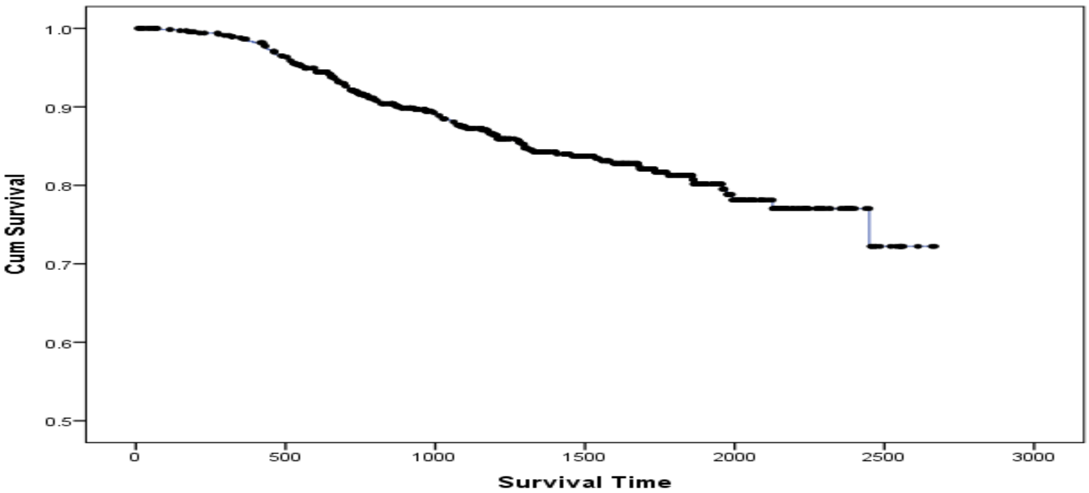


Figure A3.4 Uncorrected cumulative survival for exponential loss distribution using mean 2000 simulation a2

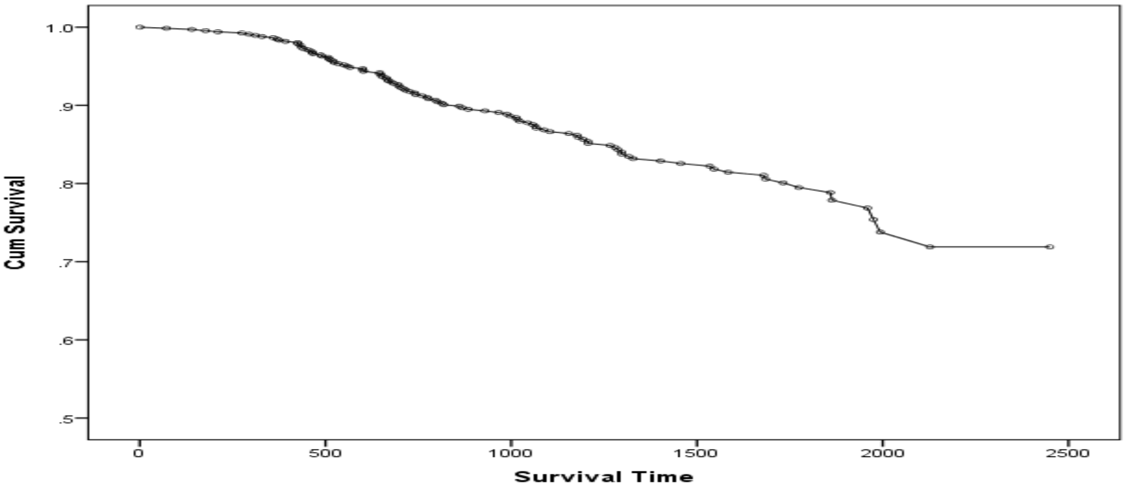


Figure A3.5 Cumulated survival for exponential loss distribution using mean 2000: corrected using the first model simulation a2

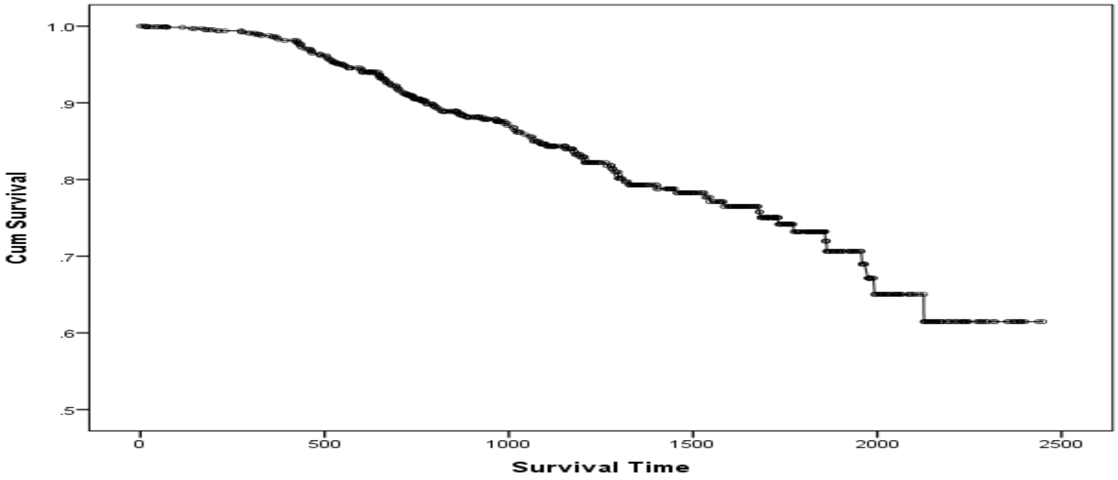


Figure A3.6 Cumulated survival for exponential loss distribution using mean 2000: corrected the second model simulation a2

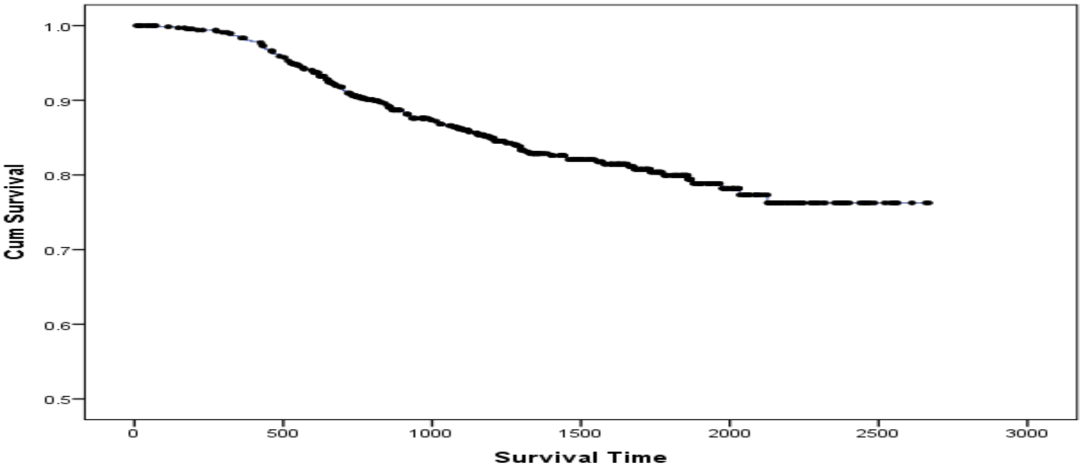


Figure A3.7 Uncorrected cumulative survival for exponential loss distribution using mean 2000 simulation a3

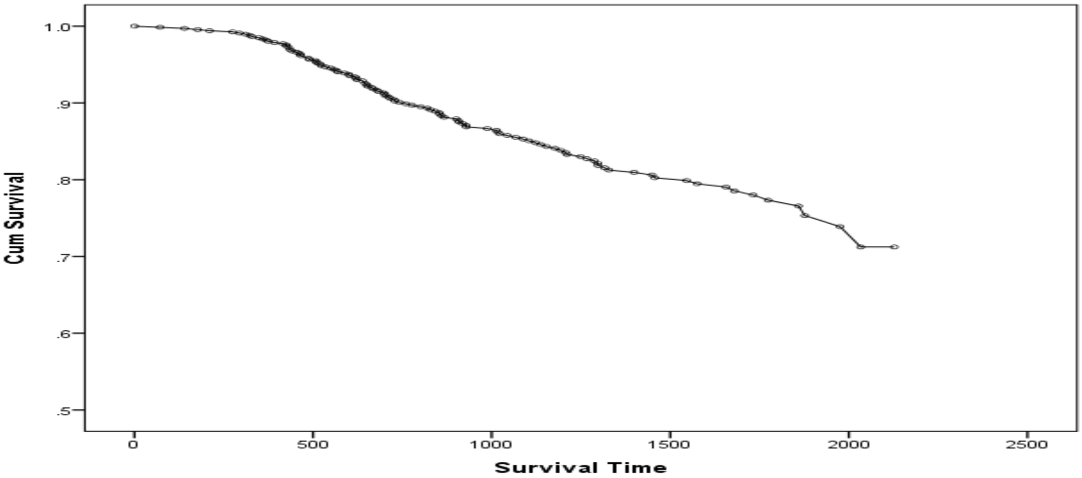


Figure A3.8 Cumulated survival for exponential loss distribution using mean 2000: corrected the first model simulation a3

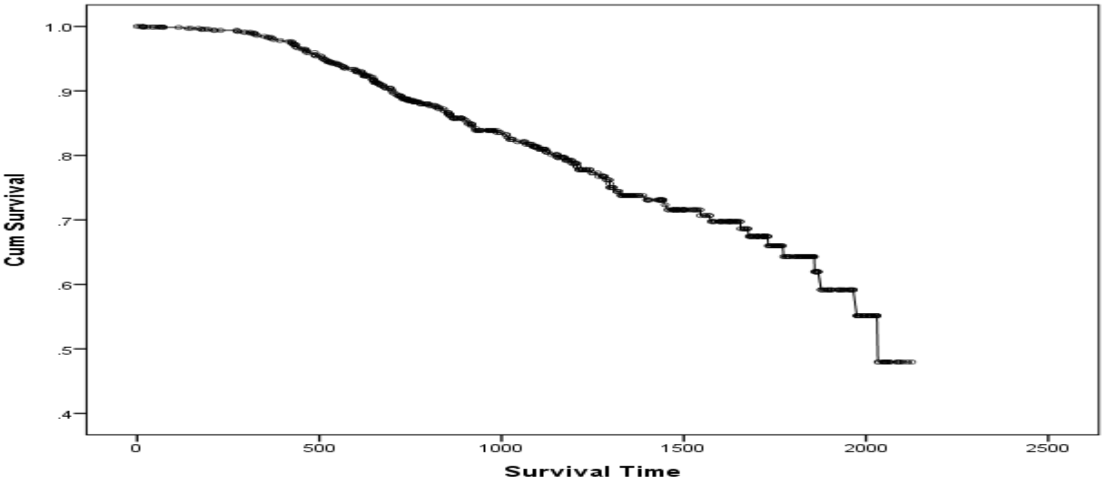


Figure A3.9 Cumulated survival for exponential loss distribution using mean 2000: corrected the second model simulation a3

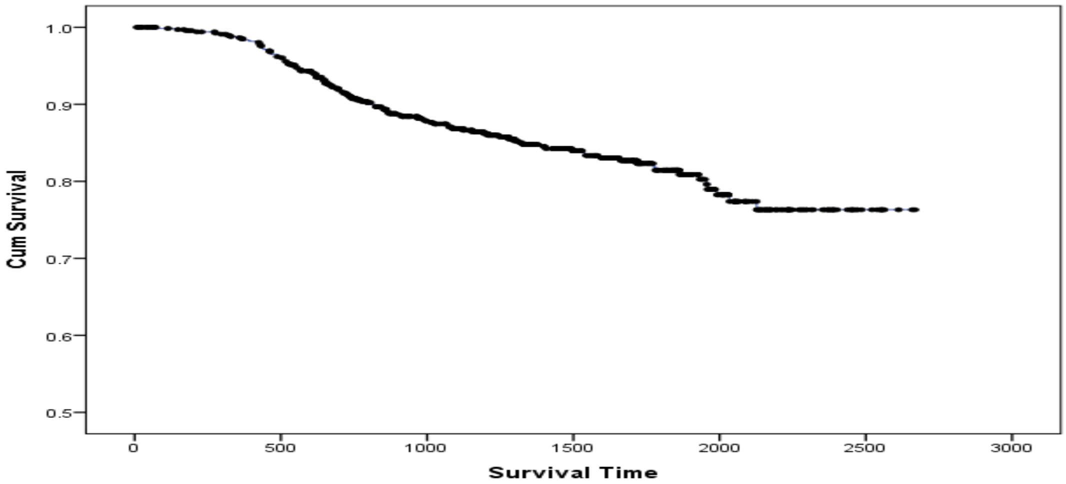


Figure A3.10 Uncorrected cumulative survival for exponential loss distribution using mean 2000 simulation a4

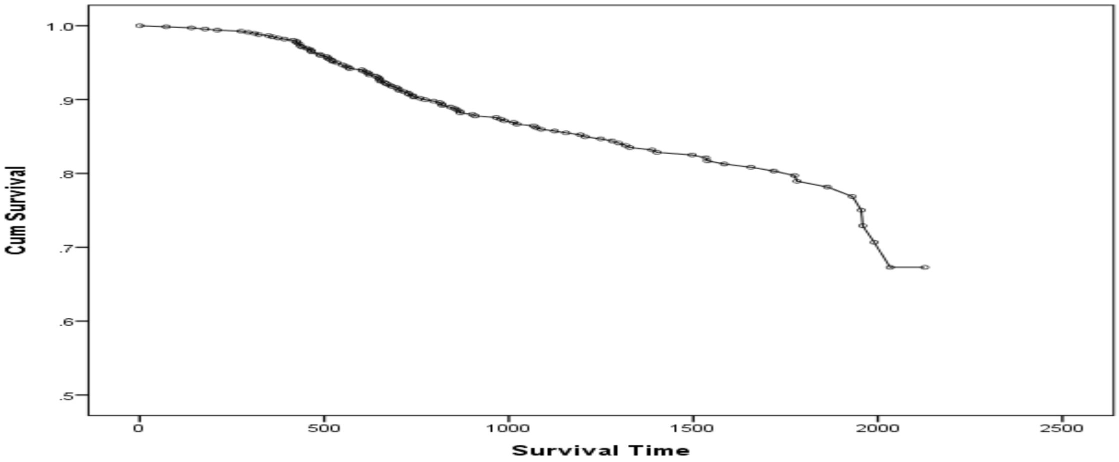


Figure A3.11 Cumulated survival for exponential loss distribution using mean 2000: corrected the first model simulation a4

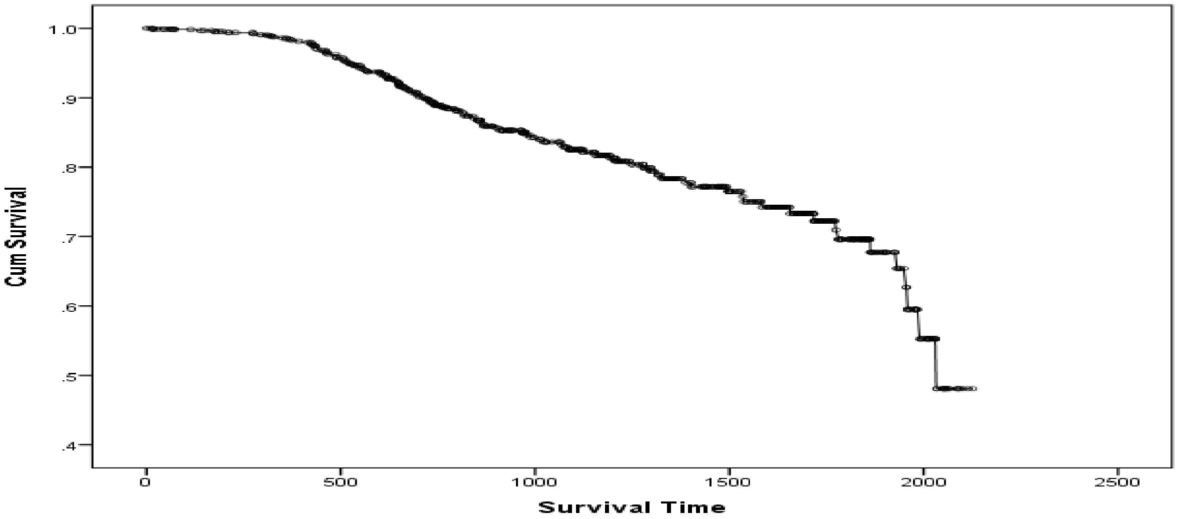


Figure A3.12 Cumulated survival for exponential loss distribution using mean 2000: corrected the second model simulation a4

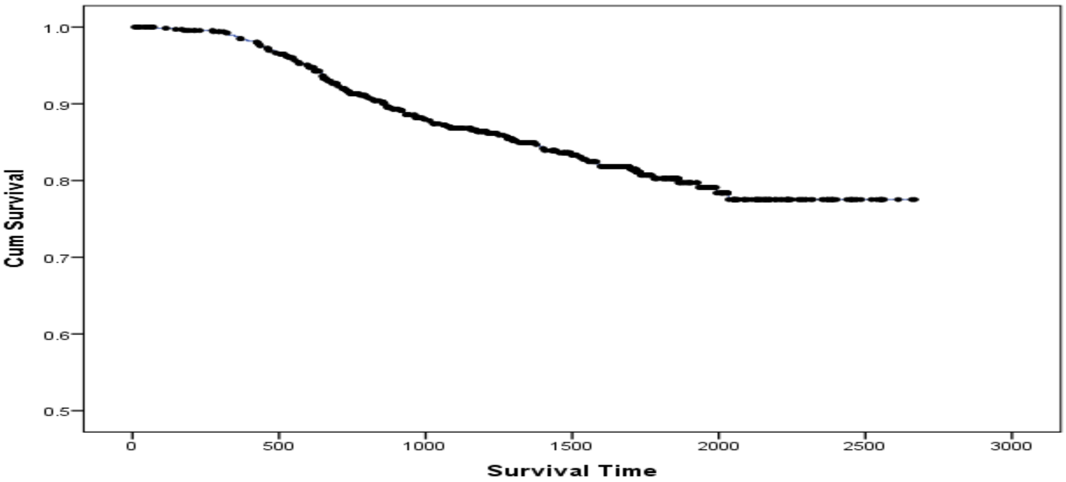


Figure A3.13 Uncorrected cumulative survival for exponential loss distribution using mean 2000 simulation a5

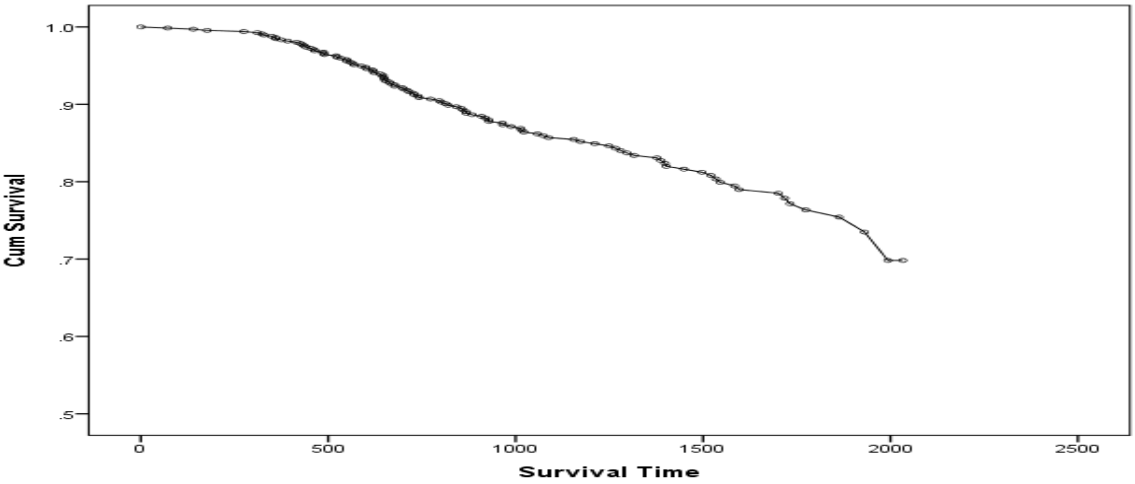


Figure A3.14 Cumulated survival for exponential loss distribution using mean 2000: corrected the first model simulation a5

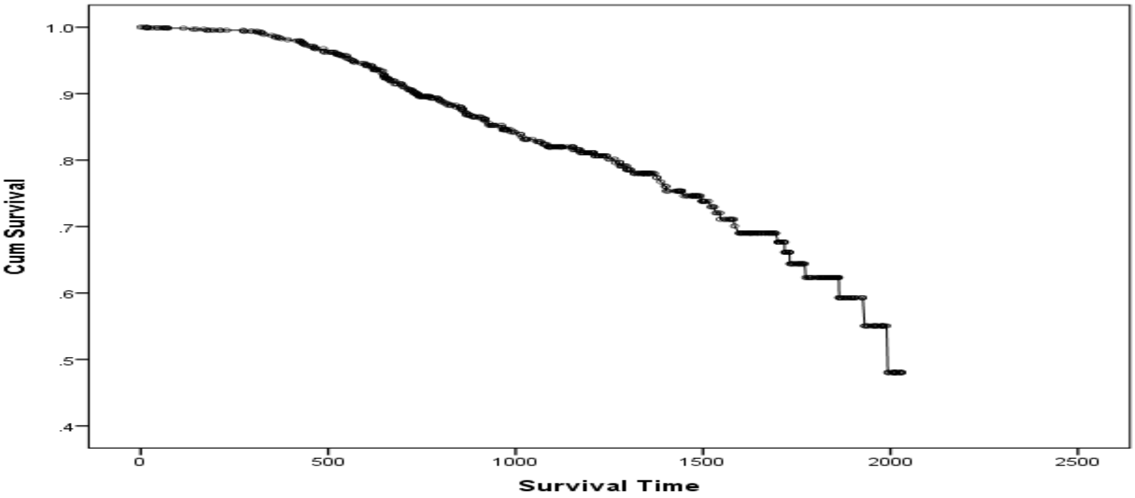


Figure A3.15 Cumulated survival for exponential loss distribution using mean 2000: corrected the second model simulation a5

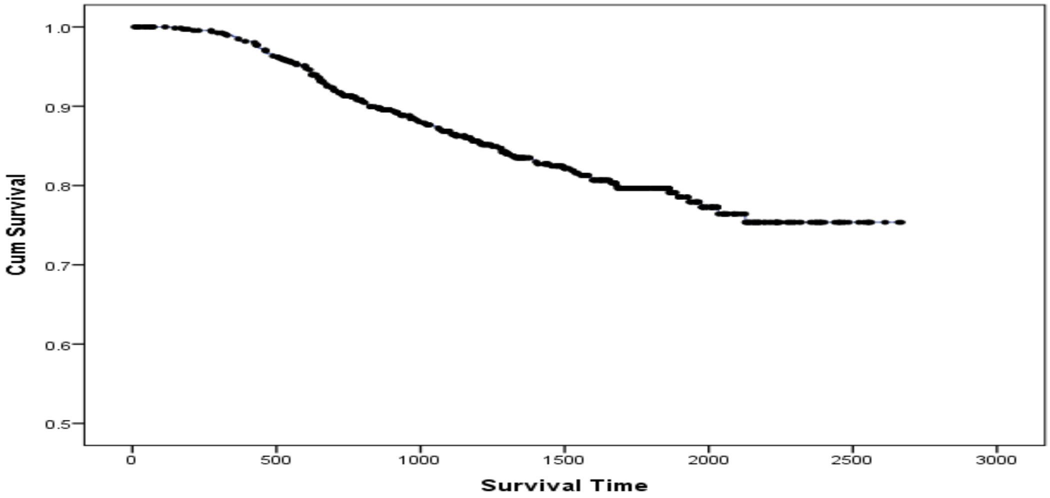


Figure A3.16 Uncorrected cumulative survival for exponential loss distribution using mean 2000 simulation a6

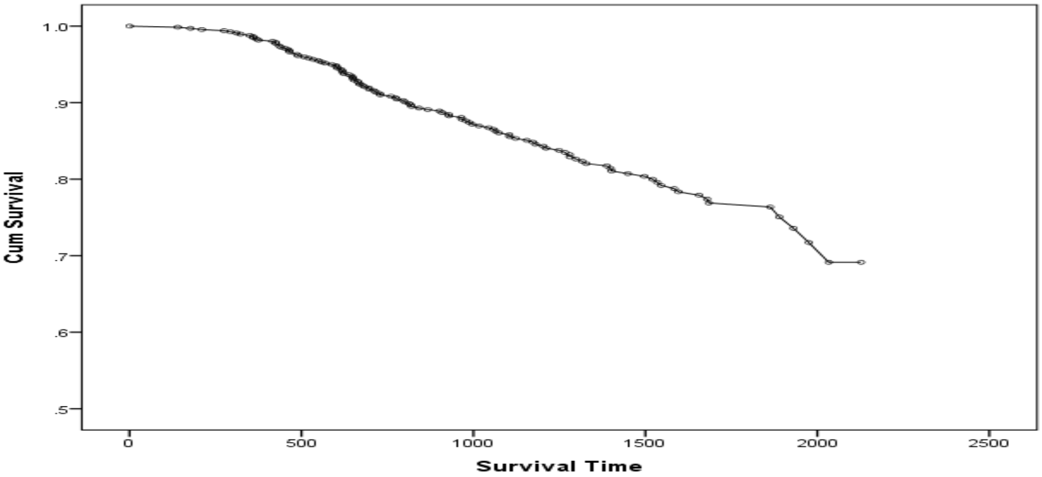


Figure A3.17 Cumulated survival for exponential loss distribution using mean 2000: corrected the first model simulation a6

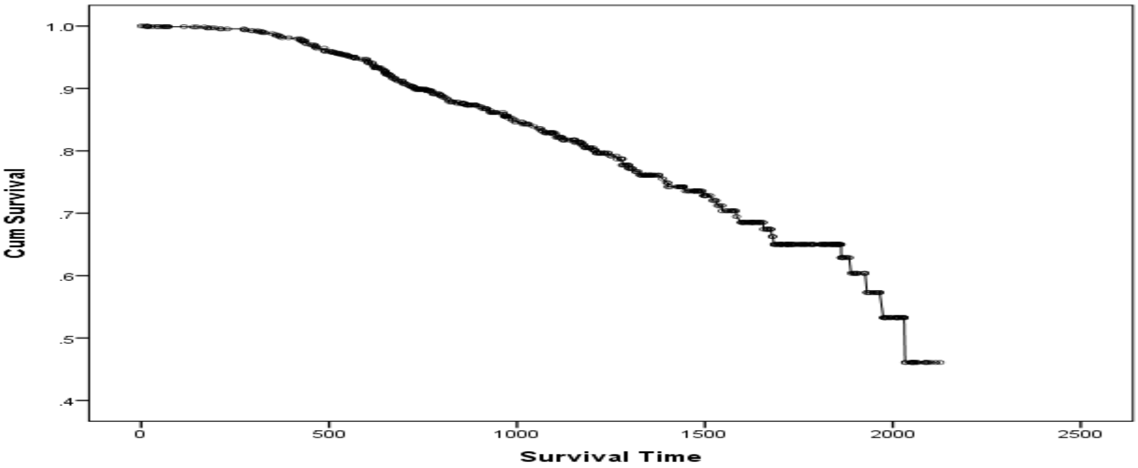


Figure A3.18 Cumulated survival for exponential loss distribution using mean 2000: corrected the second model simulation a6

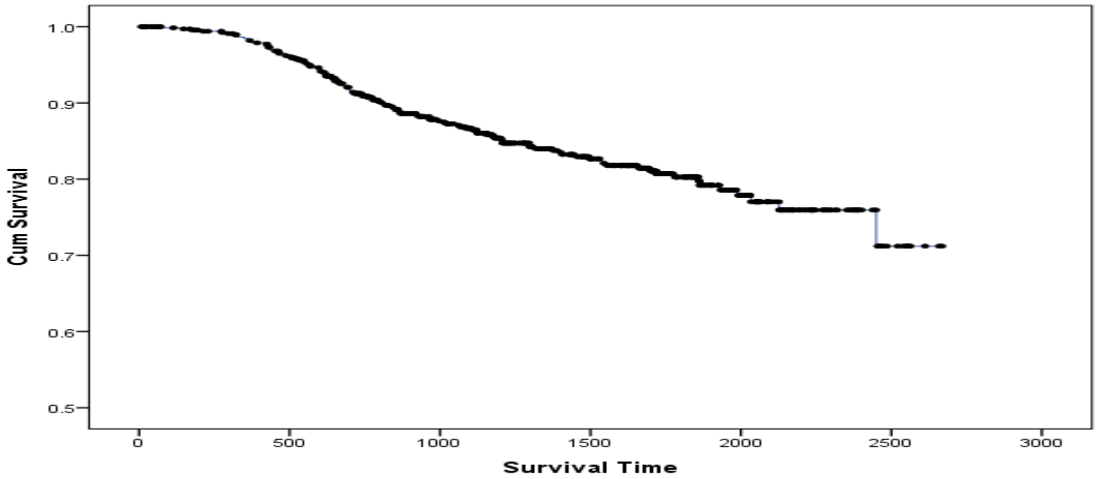


Figure A3.19 Uncorrected cumulative survival for exponential loss distribution using mean 2000 simulation a7

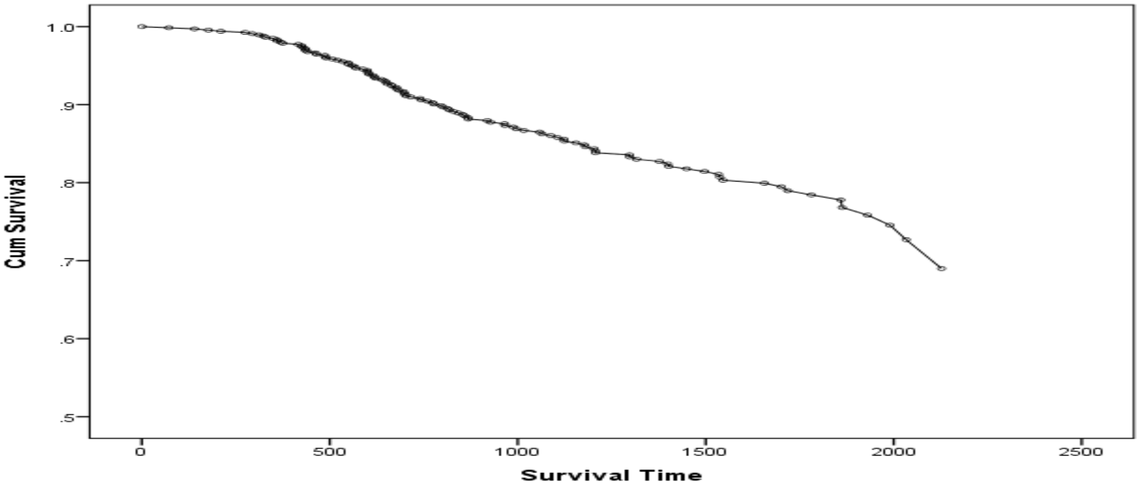


Figure A3.20 Cumulated survival for exponential loss distribution using mean 2000: corrected the first model simulation a7

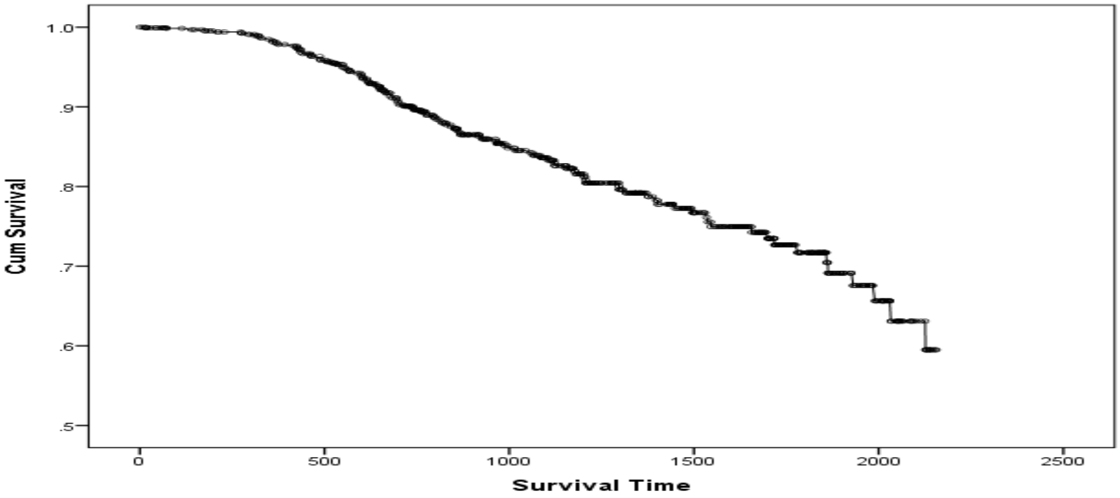


Figure A3.21 Cumulated survival for exponential loss distribution using mean 2000: corrected the second model simulation a7

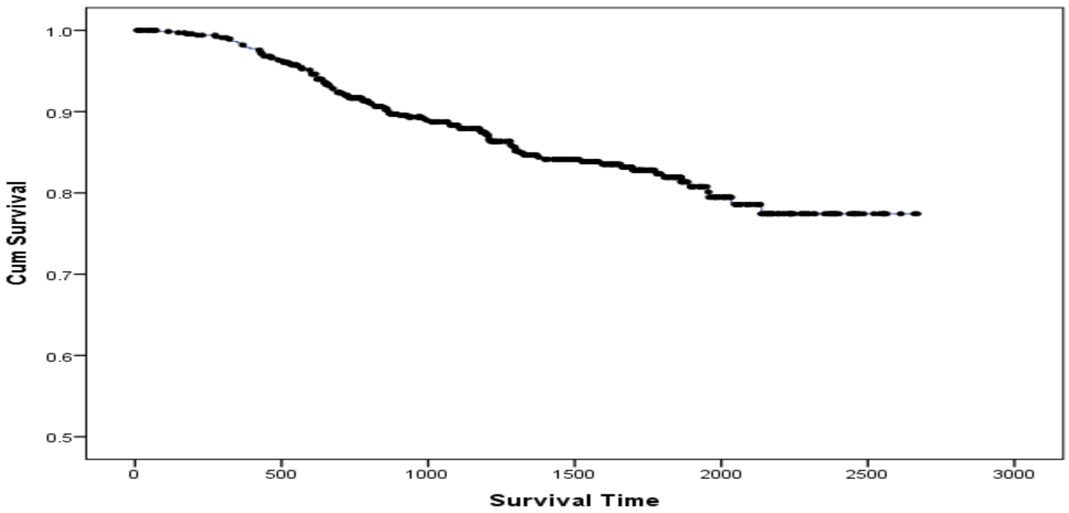


Figure A3.22 Uncorrected cumulative survival for exponential loss distribution using mean 2000 simulation a8

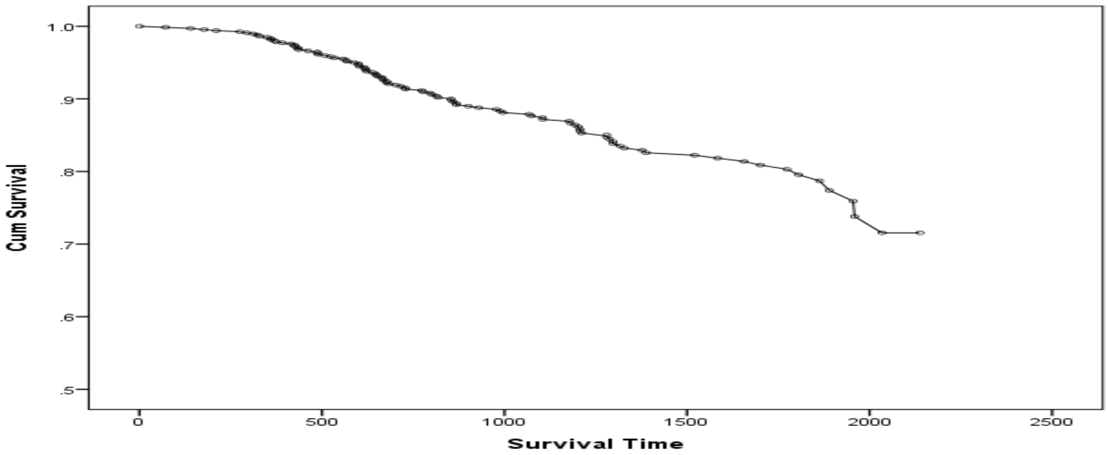


Figure A3.23 Cumulated survival for exponential loss distribution using mean 2000: corrected the first model simulation a8

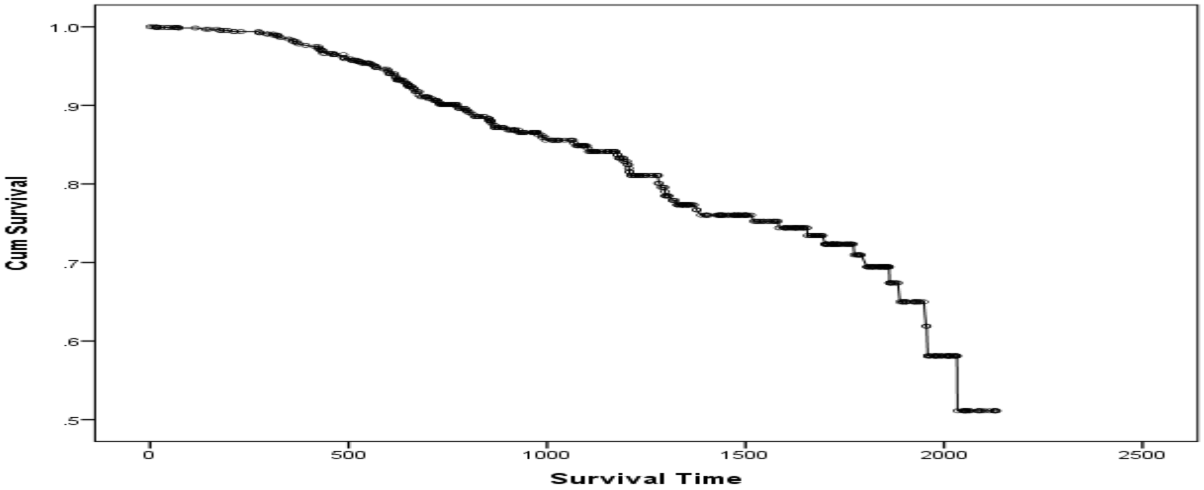


Figure A3.24 Cumulated survival for exponential loss distribution using mean 2000: corrected the second model simulation a8

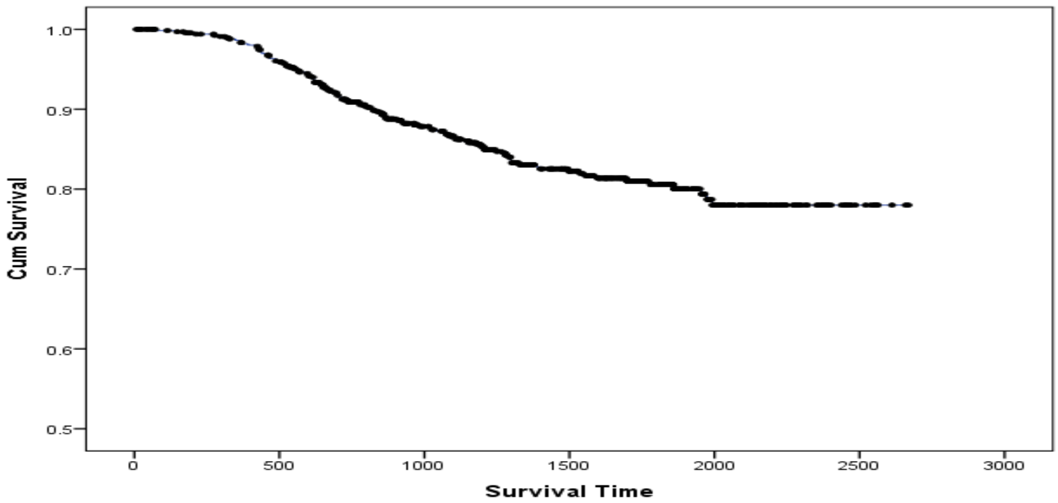


Figure A3.25 Uncorrected cumulative survival for exponential loss distribution using mean 2000 simulation a9.

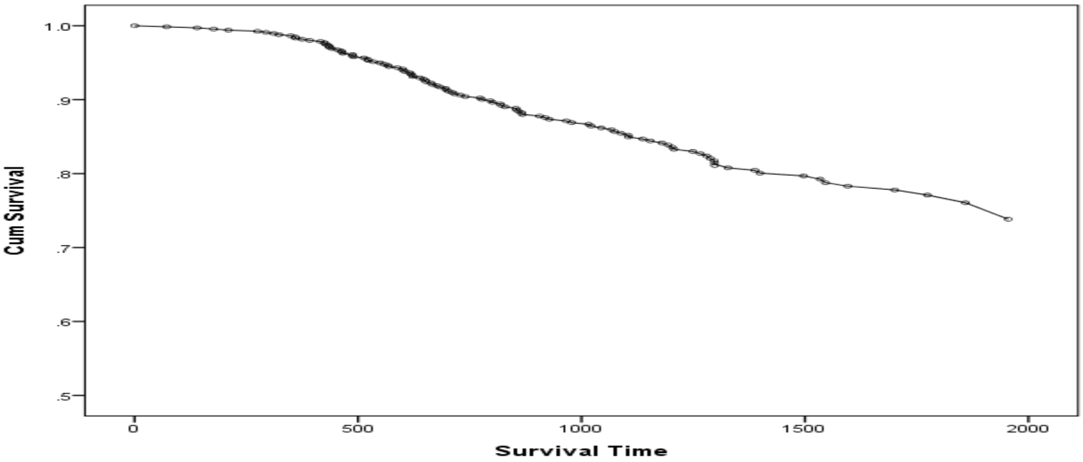


Figure A3.26 Cumulated survival for exponential loss distribution using mean 2000: corrected the first model simulation a9

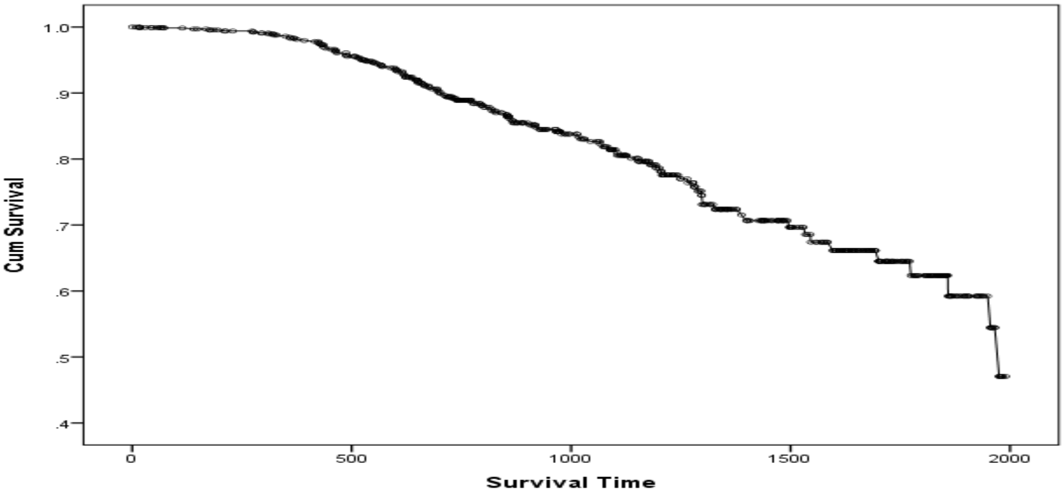


Figure A3.27 Cumulated survival for exponential loss distribution using mean 2000: corrected the second model simulation a9

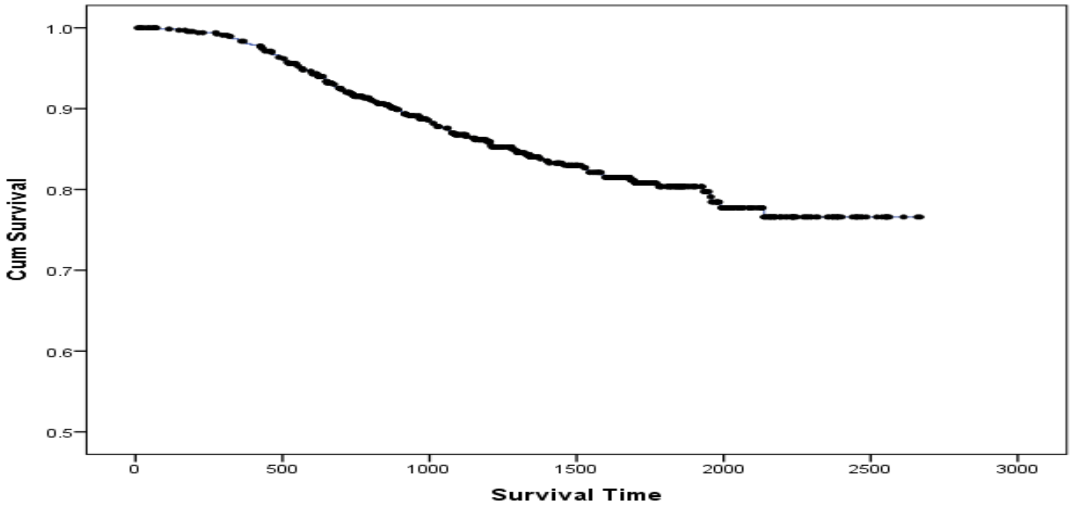


Figure A3.28 Uncorrected cumulative survival for exponential loss distribution using mean 2000 simulation a10

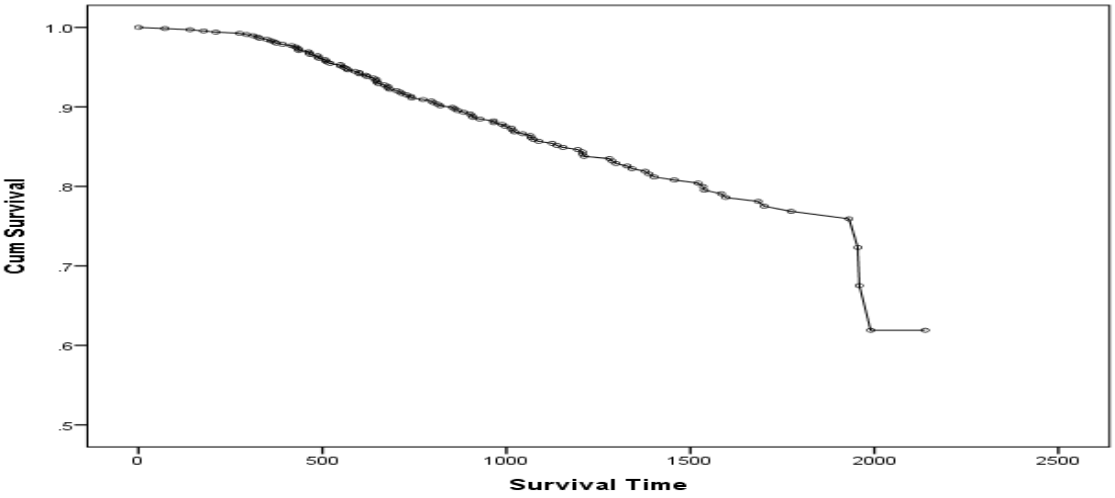


Figure A3.29 Cumulated survival for exponential loss distribution using mean 2000: corrected the first model simulation a10

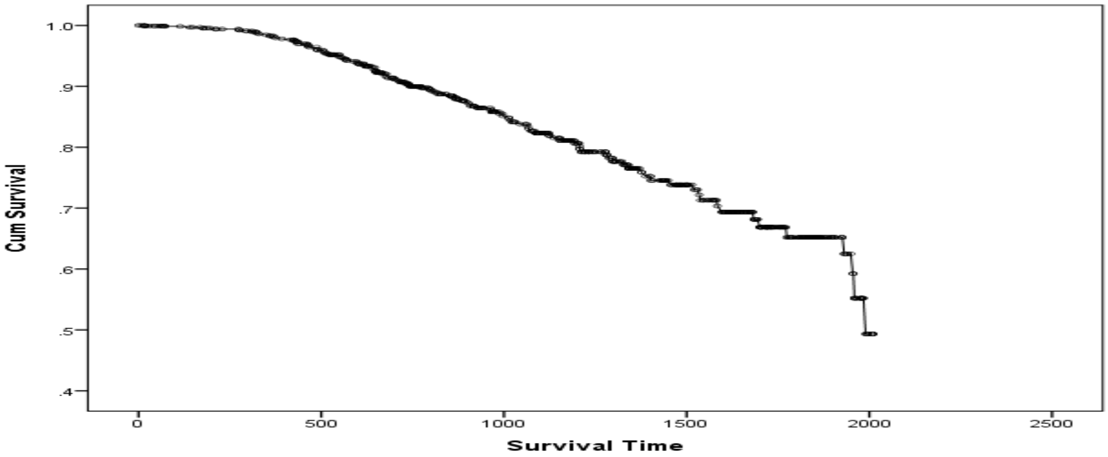


Figure A3.30 Cumulated survival for exponential loss distribution using mean 2000: corrected the second model simulation a10

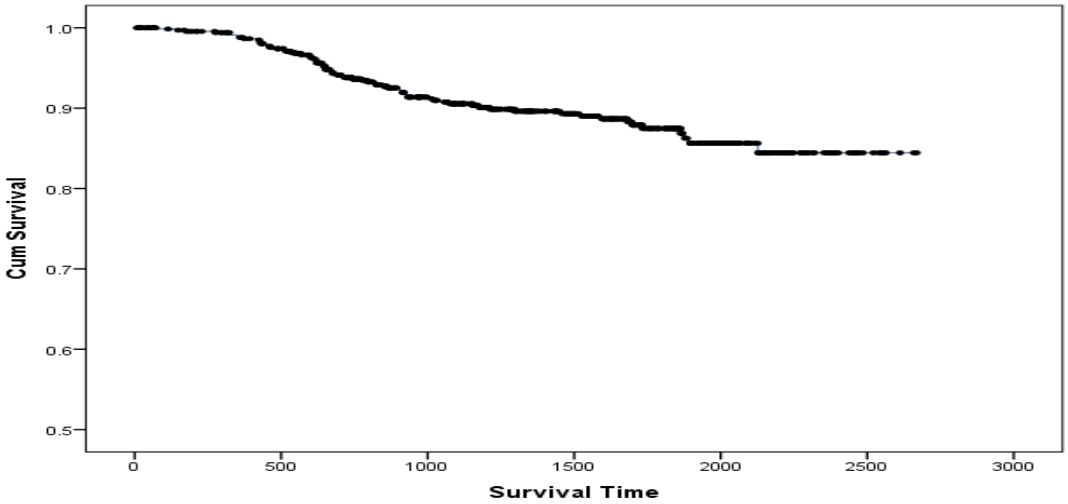


Figure A3.31 Uncorrected cumulative survival for exponential loss distribution using mean 1000 simulation b1

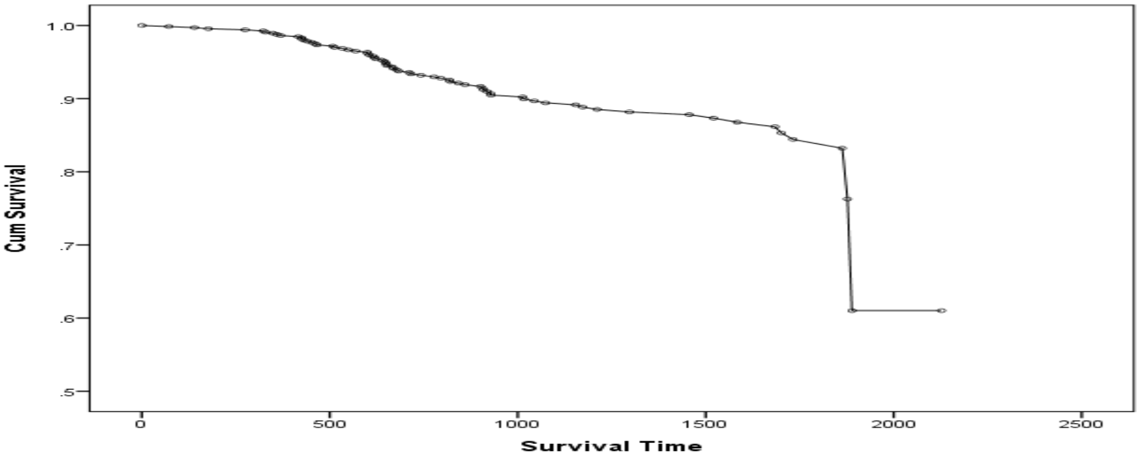


Figure A3.32 Cumulated survival for exponential loss distribution using mean 1000: corrected the first model simulation b1

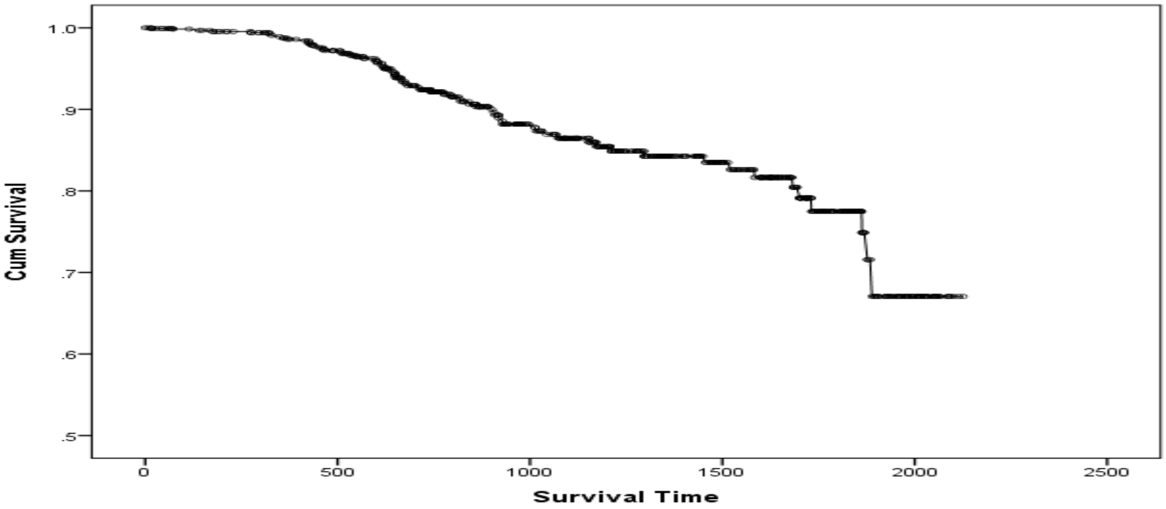


Figure A3.33 Cumulated survival for exponential loss distribution using mean 1000: corrected the second model simulation b1

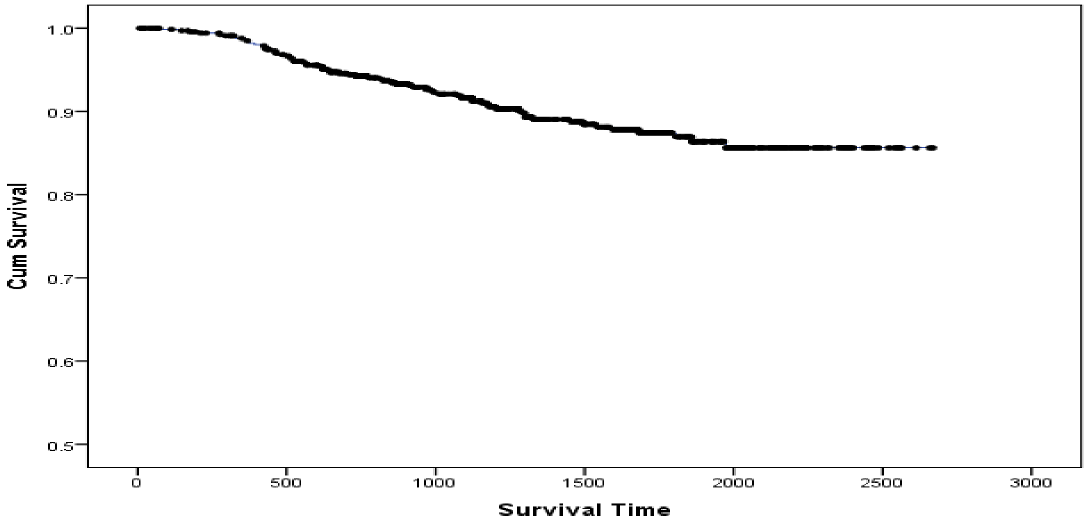


Figure A3.34 Uncorrected cumulative survival for exponential loss distribution using mean 1000 simulation b2

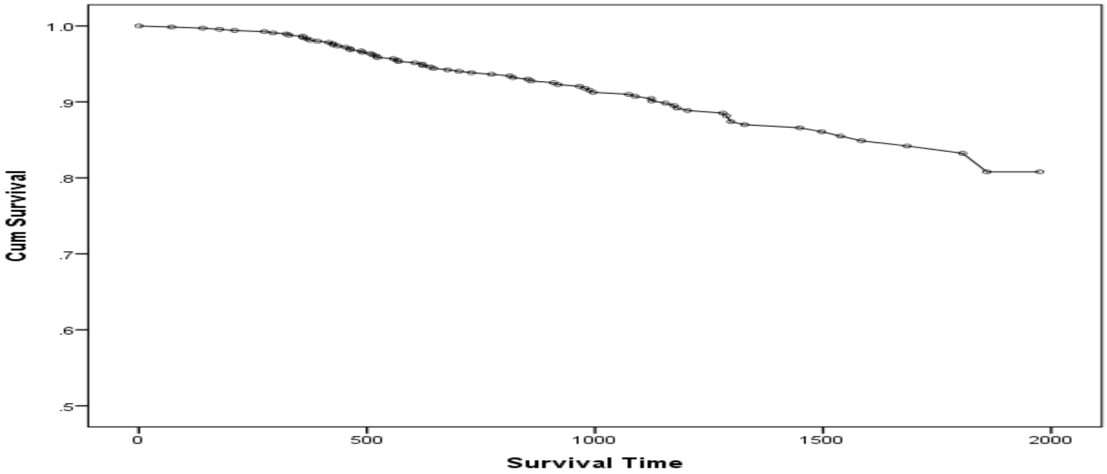


Figure A3.35 Cumulated survival for exponential loss distribution using mean 1000: corrected the first model simulation b2

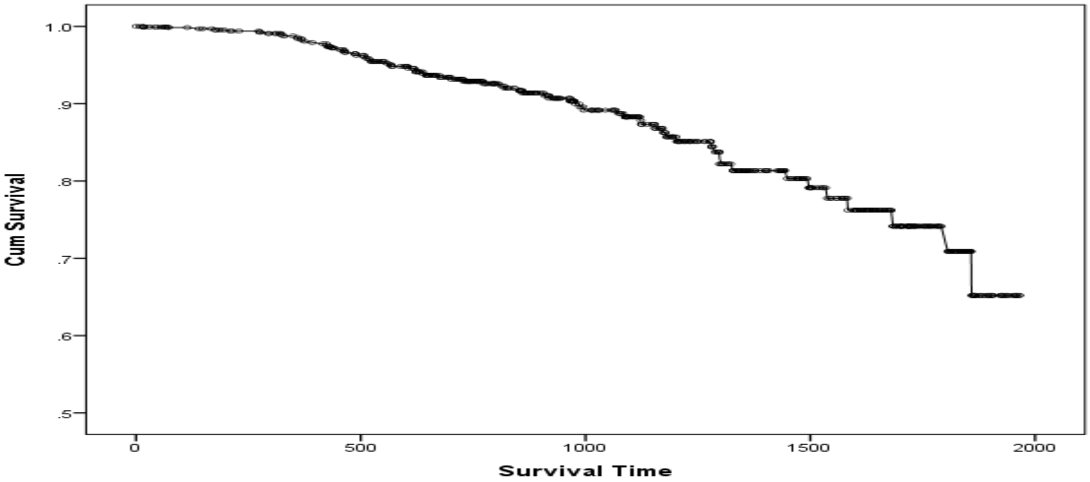


Figure A3.36 Cumulated survival for exponential loss distribution using mean 1000: corrected the second model simulation b2

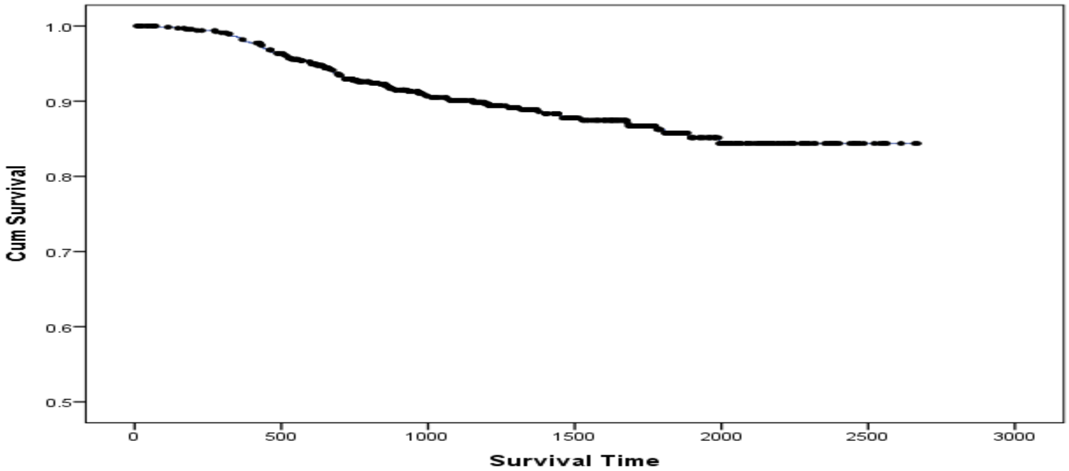


Figure A3.37 Uncorrected cumulative survival for exponential loss distribution using mean 1000 simulation b3

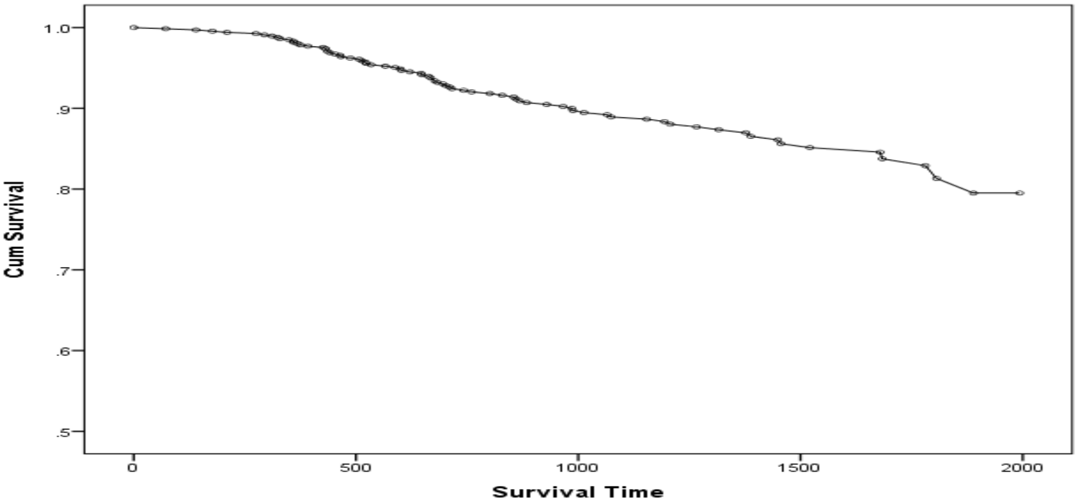


Figure A3.38 Cumulated survival for exponential loss distribution using mean 1000: corrected the first model simulation b3

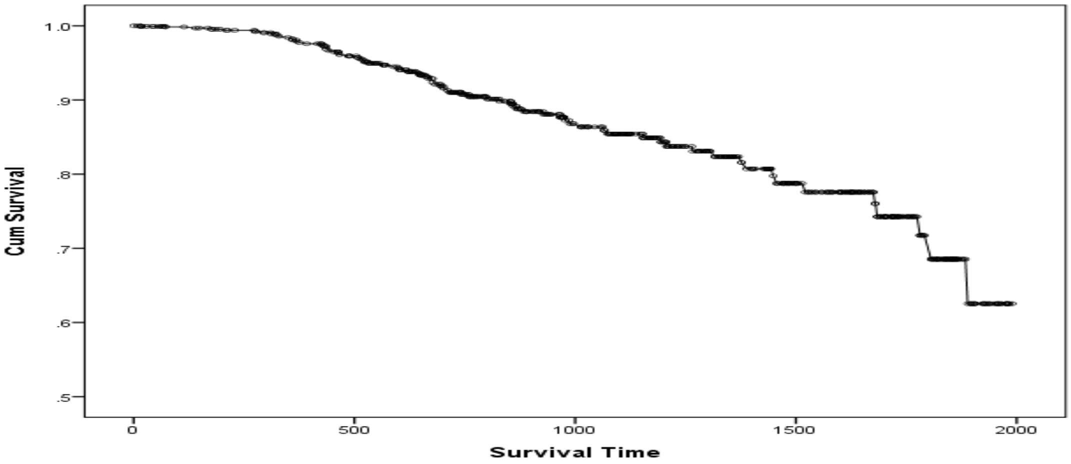


Figure A3.39 Cumulated survival for exponential loss distribution using mean 1000: corrected the second model simulation b3

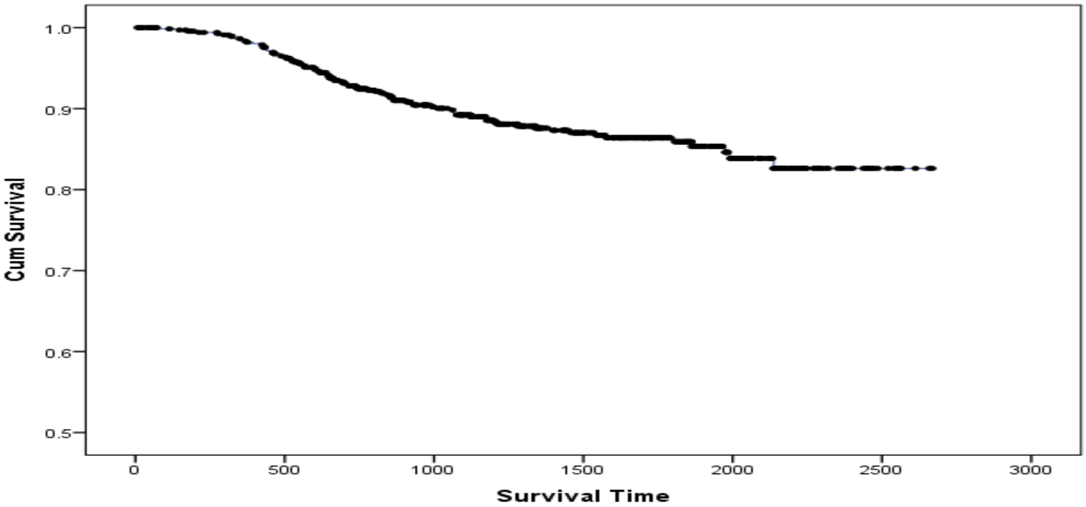


Figure A3.40 Uncorrected cumulative survival for exponential loss distribution using mean 1000 simulation b4

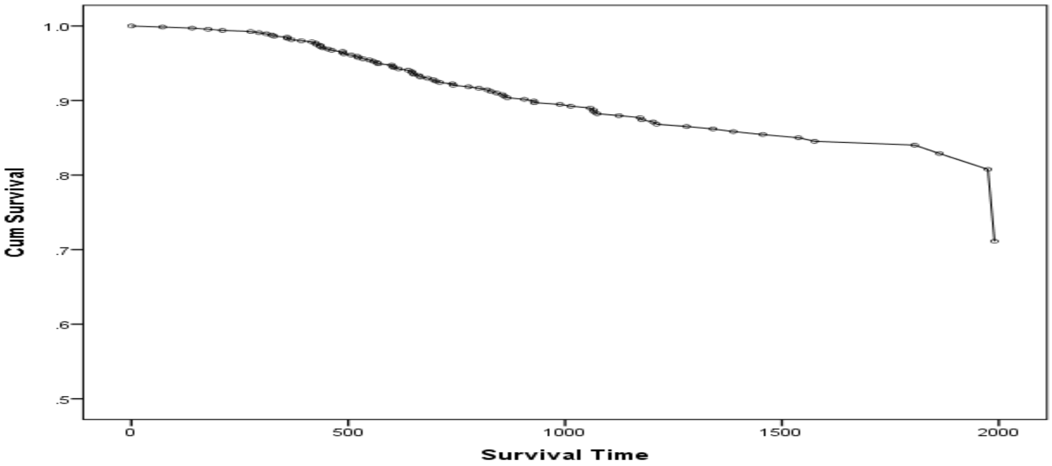


Figure A3.41 Cumulated survival for exponential loss distribution using mean 1000: corrected the first model simulation b4

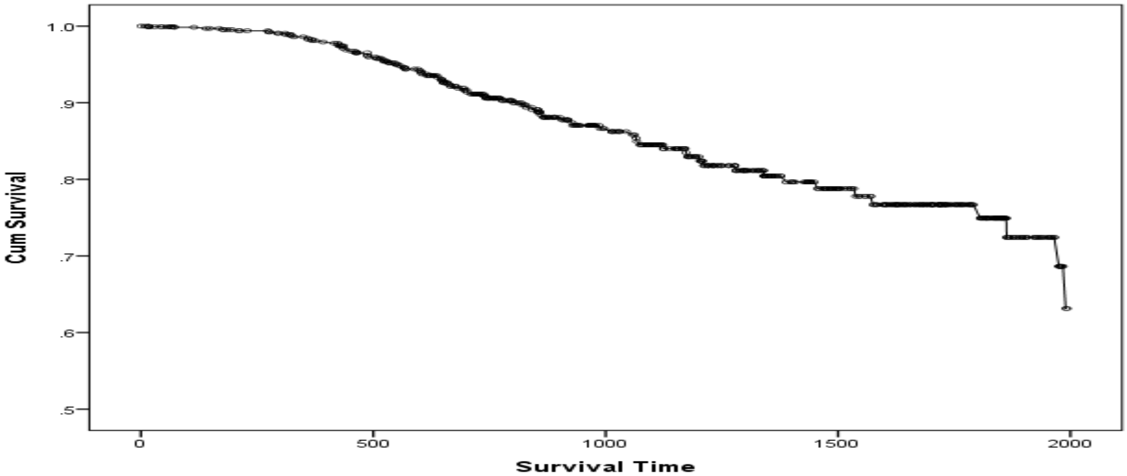


Figure A3.42 Cumulated survival for exponential loss distribution using mean 1000: corrected the second model simulation b4

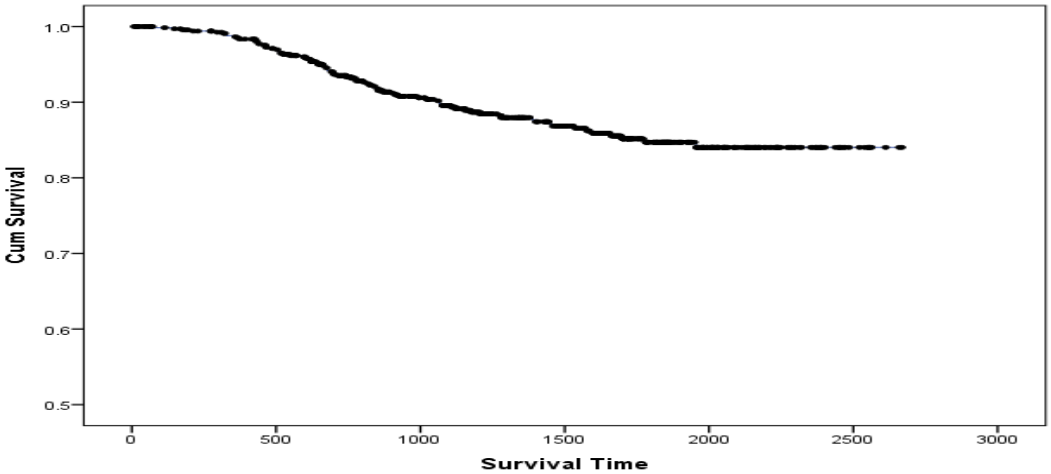


Figure A3.43 Uncorrected cumulative survival for exponential loss distribution using mean 1000 simulation b5

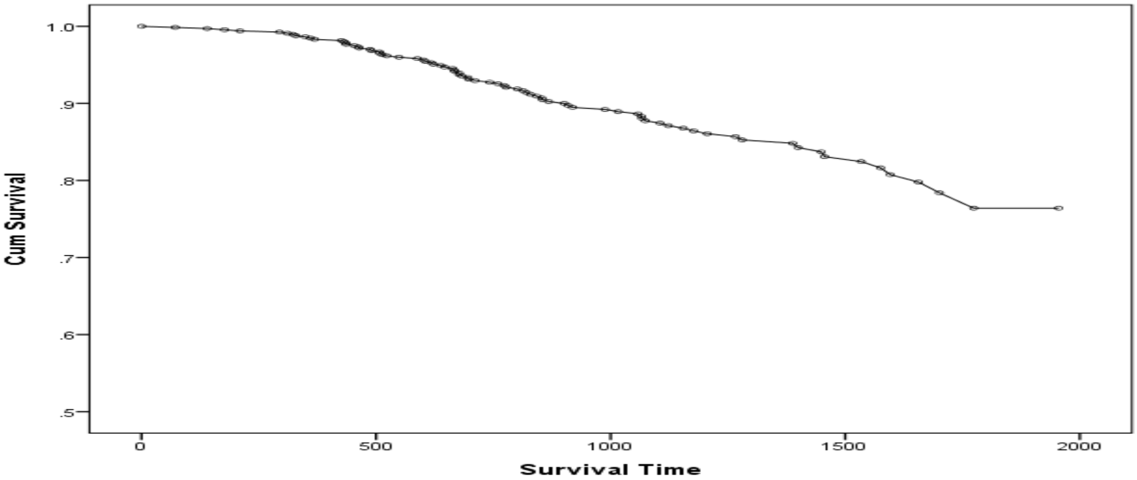


Figure A3.44 Cumulated survival for exponential loss distribution using mean 1000: corrected the first model simulation b5

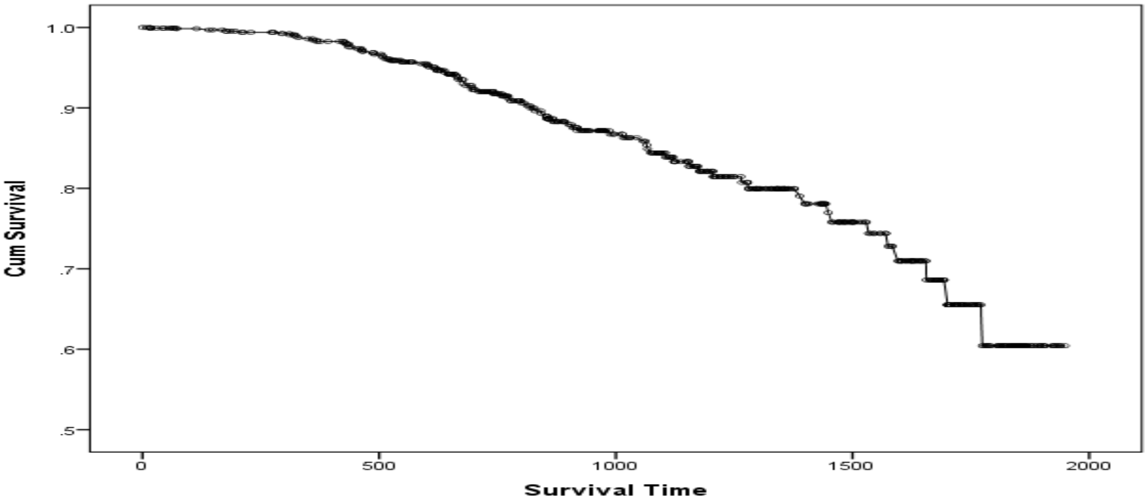


Figure A3.45 Cumulated survival for exponential loss distribution using mean 1000: corrected the second model simulation b5

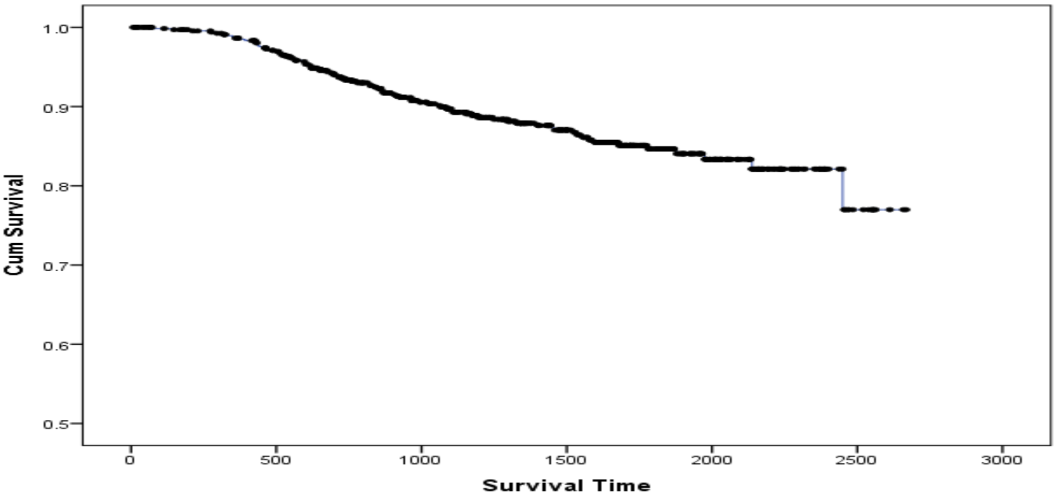


Figure A3.46 Uncorrected cumulative survival for exponential loss distribution using mean 1000 simulation b6

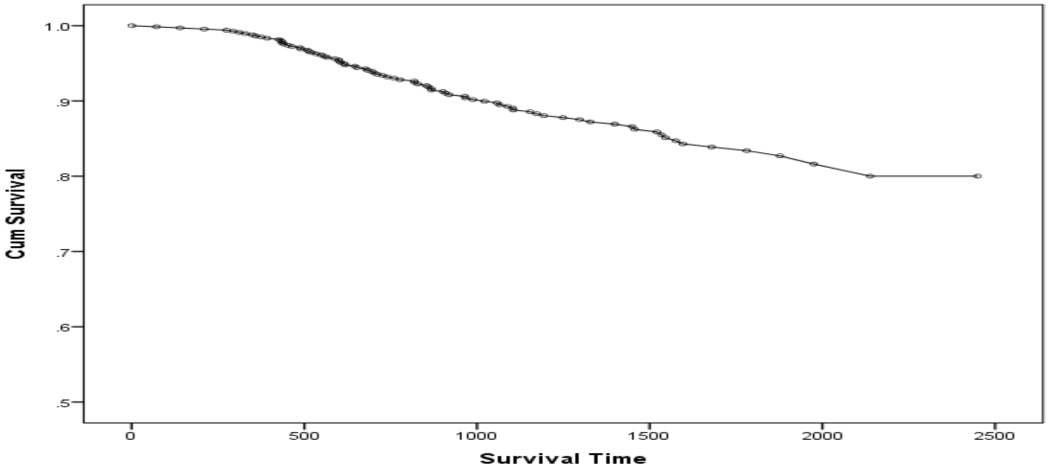


Figure A3.47 Cumulated survival for exponential loss distribution using mean 1000: corrected the first model simulation b6

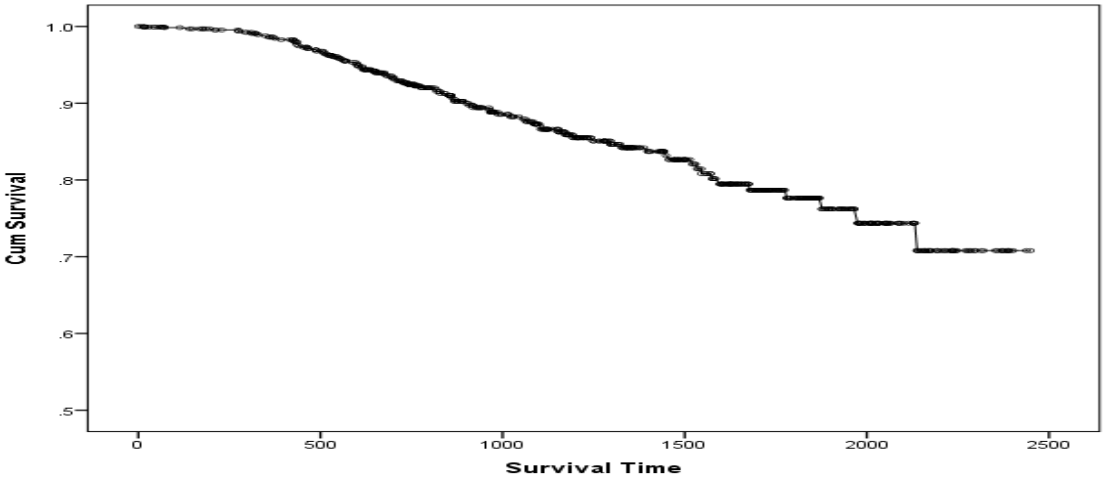


Figure A3.48 Cumulated survival for exponential loss distribution using mean 1000: corrected the second model simulation b6

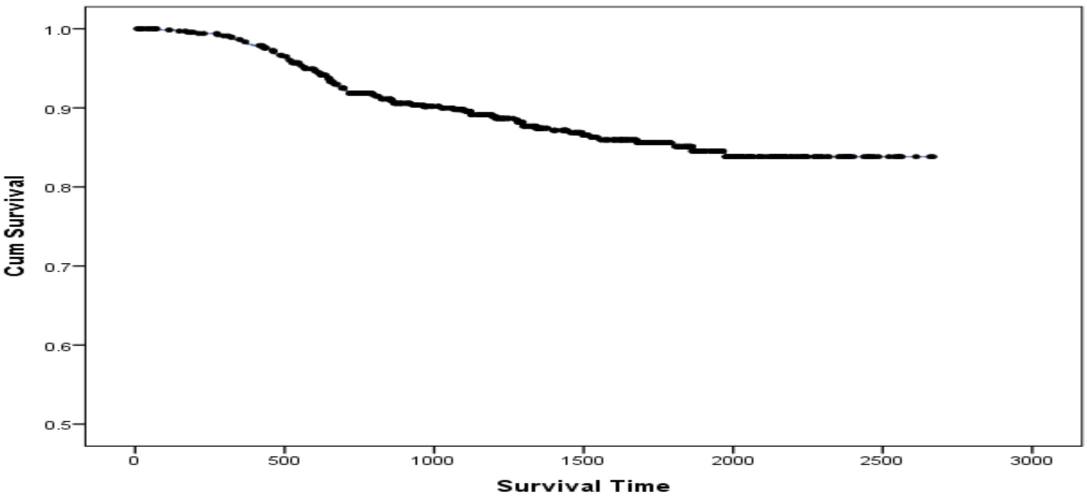


Figure A3.49 Uncorrected cumulative survival for exponential loss distribution using mean 1000 simulation b7

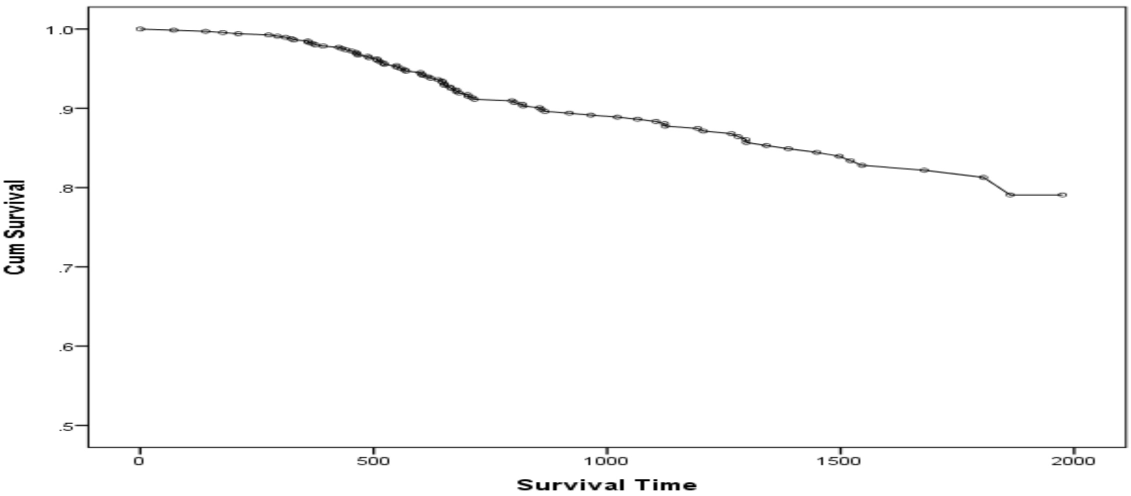


Figure A3.50 Cumulated survival for exponential loss distribution using mean 1000: corrected the first model simulation b7

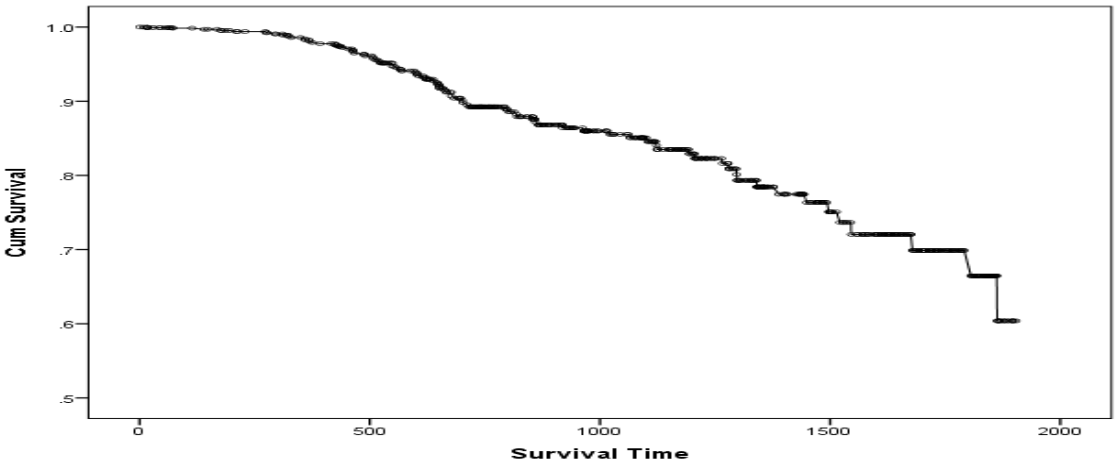


Figure A3.51 Cumulated survival for exponential loss distribution using mean 1000: corrected the second model simulation b7

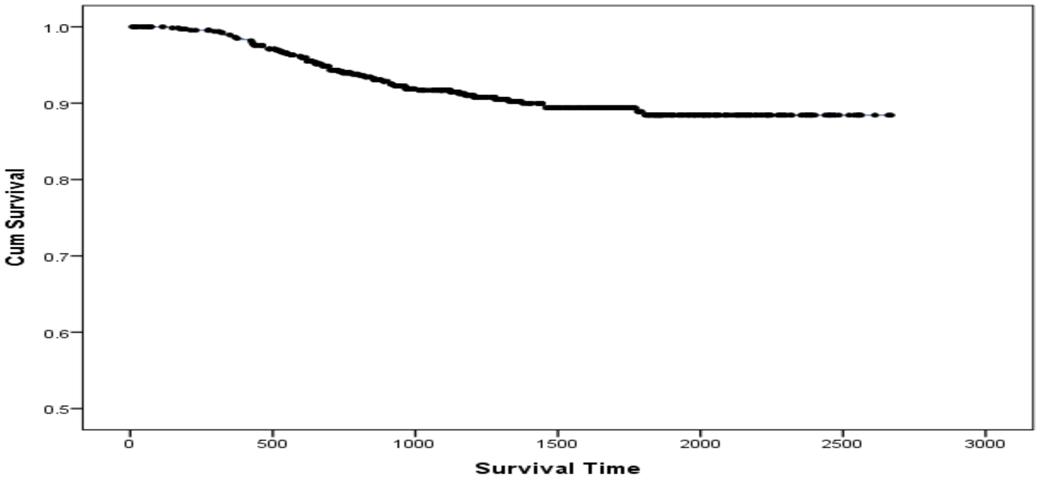


Figure A3.52 Uncorrected cumulative survival for exponential loss distribution using mean 1000 simulation b8

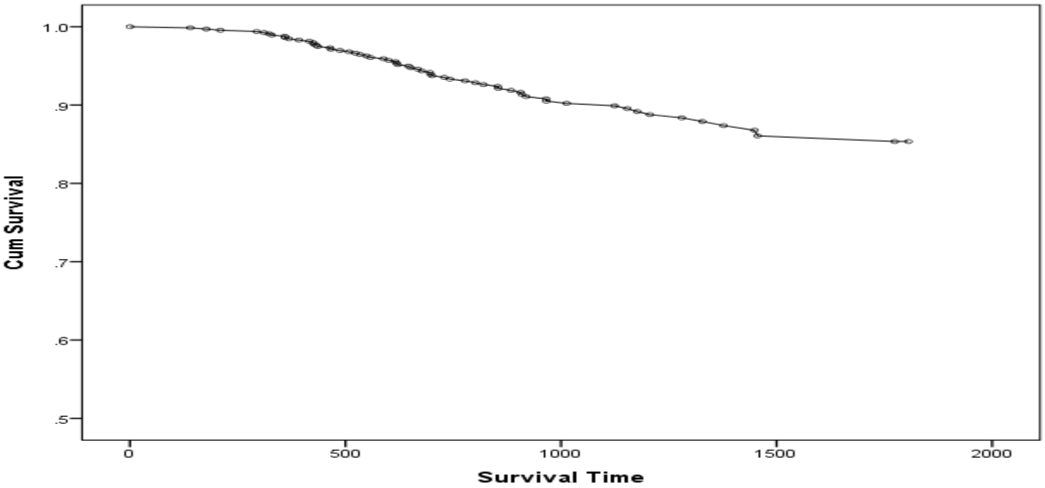


Figure A3.53 Cumulated survival for exponential loss distribution using mean 1000: corrected the first model simulation b8

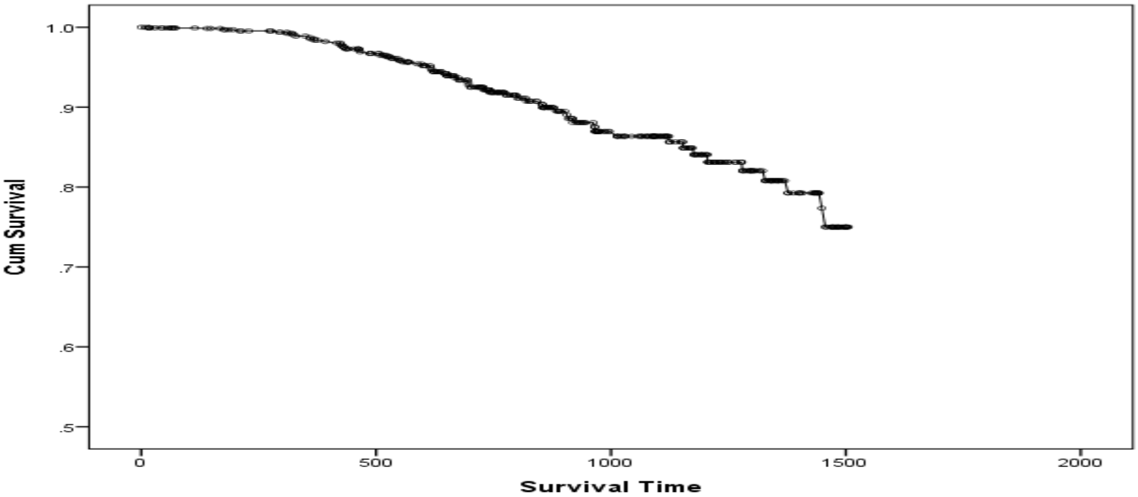


Figure A3.54 Cumulated survival for exponential loss distribution using mean 1000: corrected the second model simulation b8

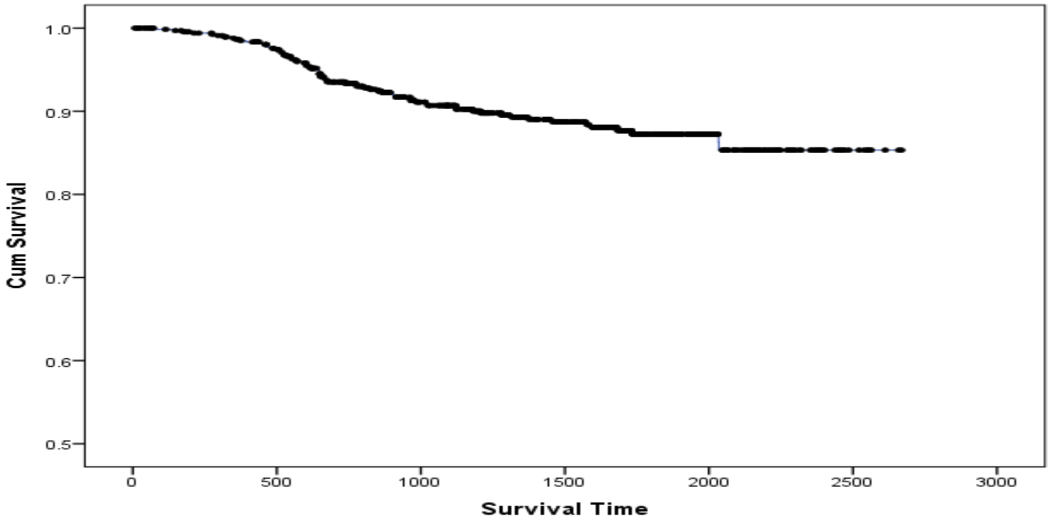


Figure A3.55 Uncorrected cumulative survival for exponential loss distribution using mean 1000 simulation b9

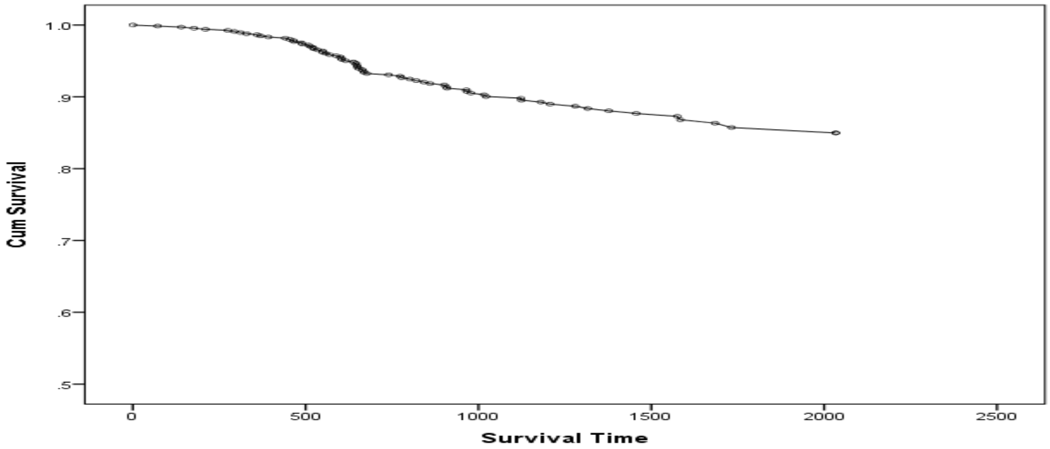


Figure A3.56 Cumulated survival for exponential loss distribution using mean 1000: corrected the first model simulation b9

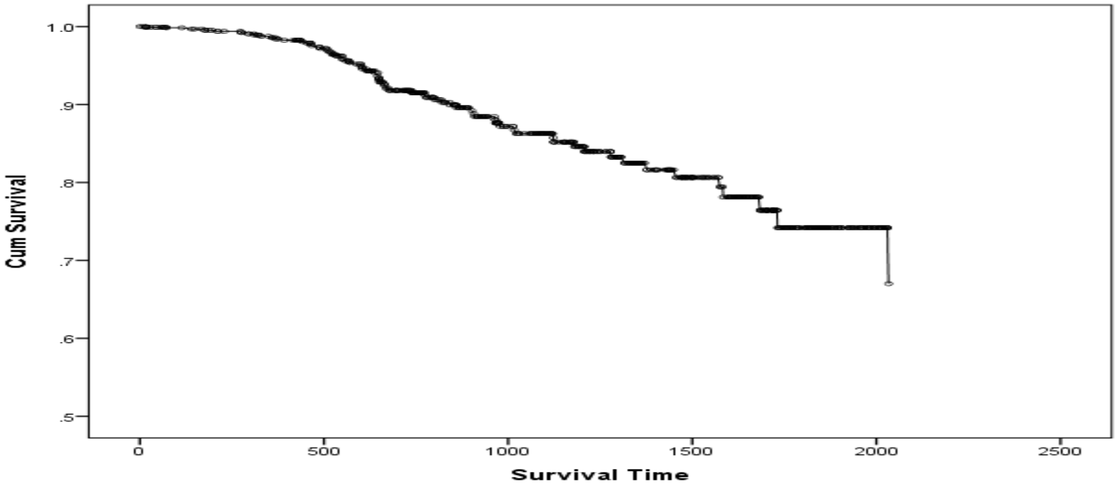


Figure A3.57 Cumulated survival for exponential loss distribution using mean 1000: corrected the second model simulation b9

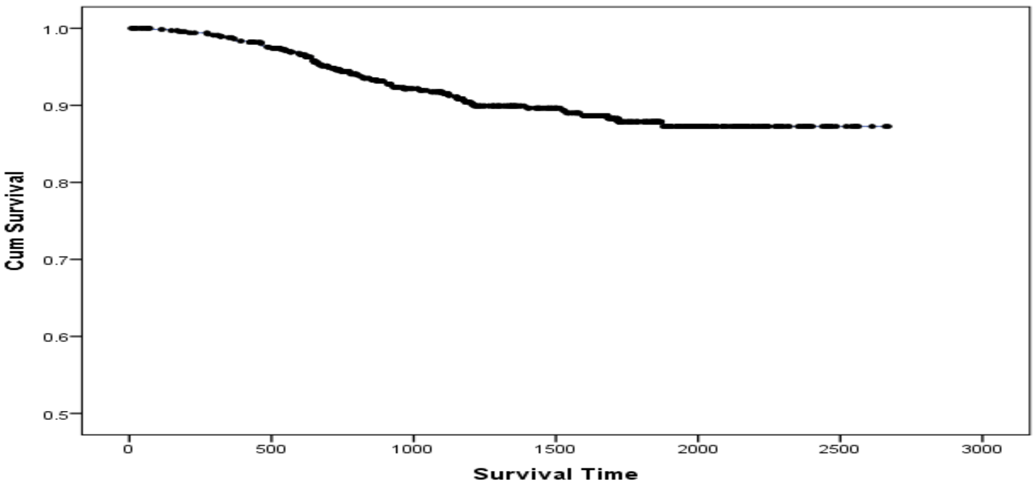


Figure A3.58 Uncorrected cumulative survival for exponential loss distribution using mean 1000 simulation b10

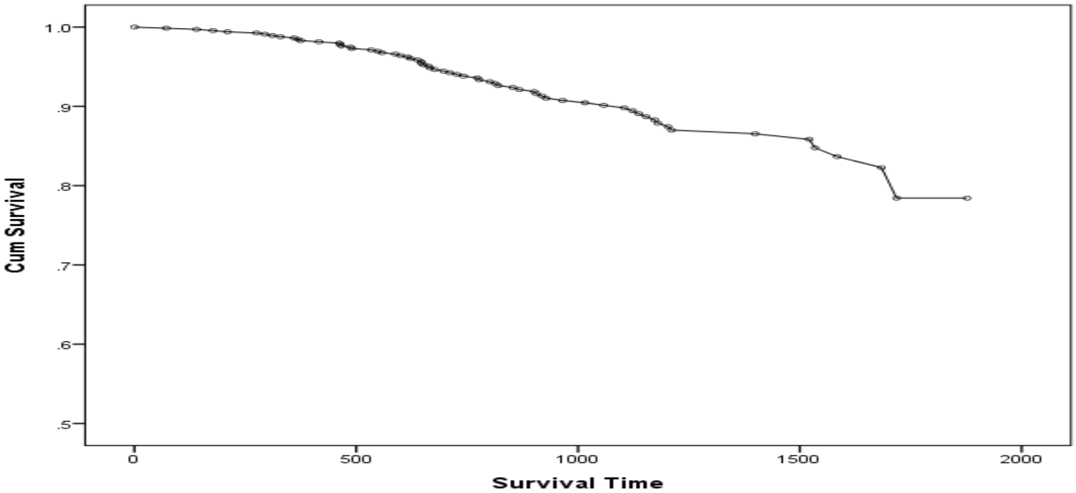


Figure A3.59 Cumulated survival for exponential loss distribution using mean 1000: corrected the first model simulation b10

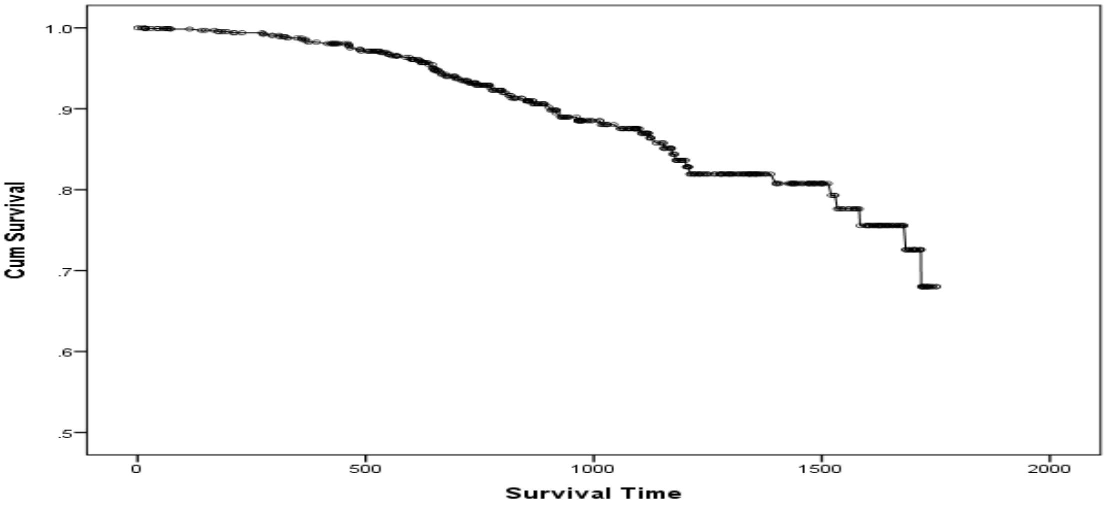


Figure A3.60 Cumulated survival for exponential loss distribution using mean 1000: corrected the second model simulation b10

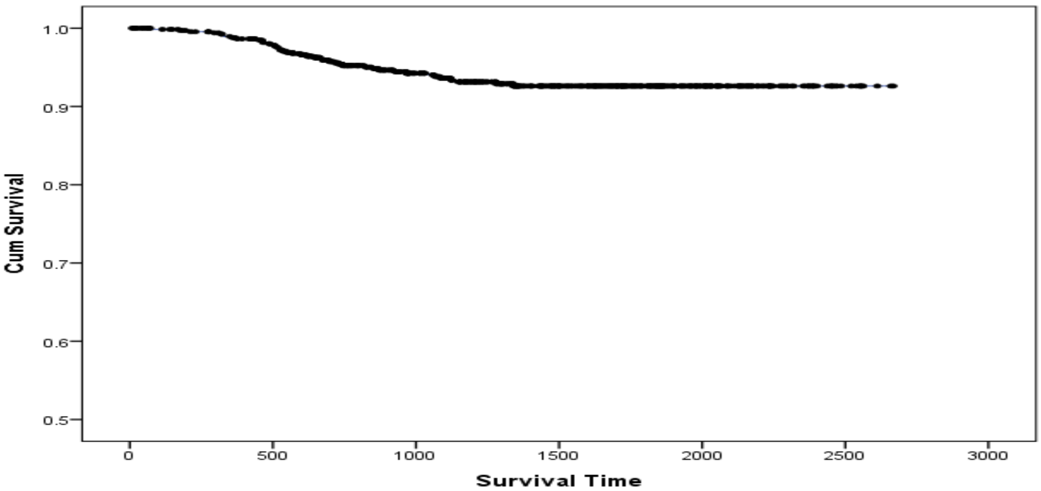


Figure A3.61 Uncorrected cumulative survival for exponential loss distribution using mean 500 simulation c1

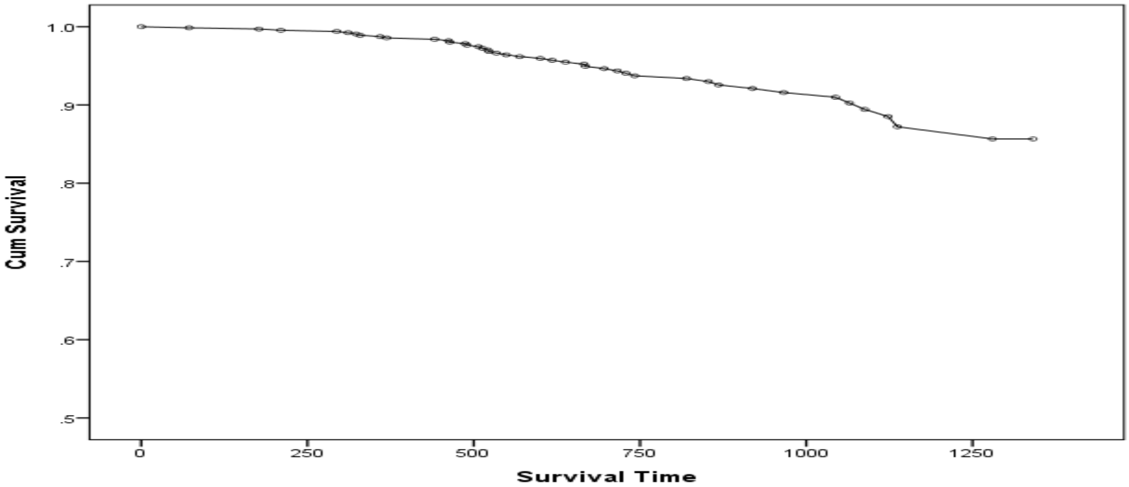


Figure A3.62 Cumulated survival for exponential loss distribution using mean 500: corrected the first model simulation c1

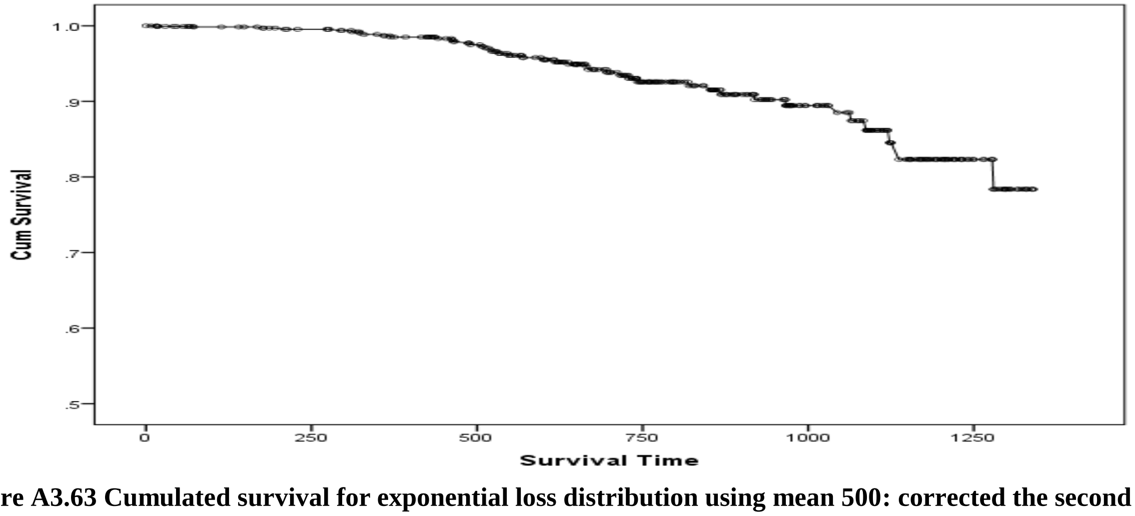


Figure A3.63 Cumulated survival for exponential loss distribution using mean 500: corrected the second model simulation c1

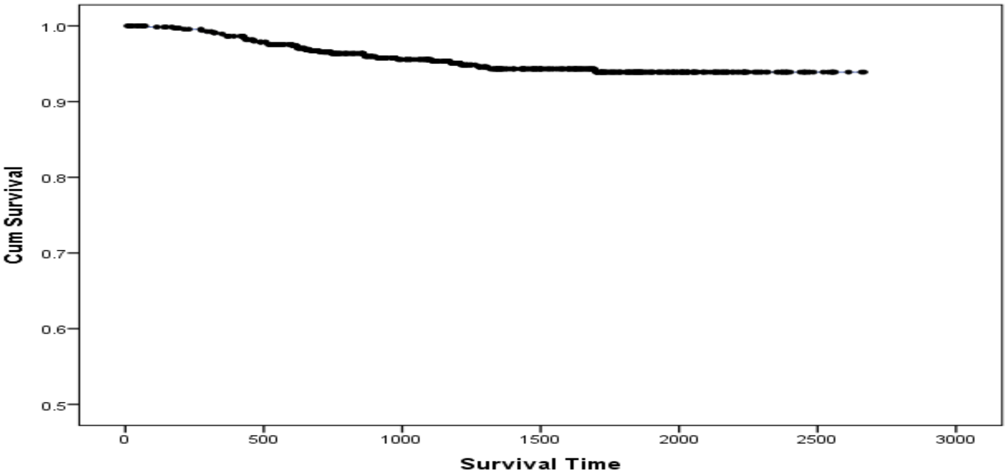


Figure A3.64 Uncorrected cumulative survival for exponential loss distribution using mean 500 simulation c2

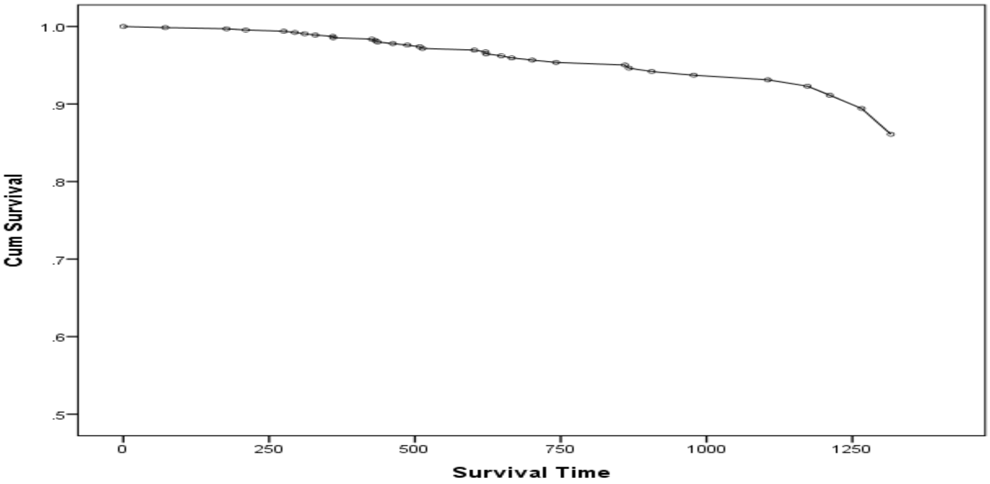


Figure A3.65 Cumulated survival for exponential loss distribution using mean 500: corrected the first model simulation c2

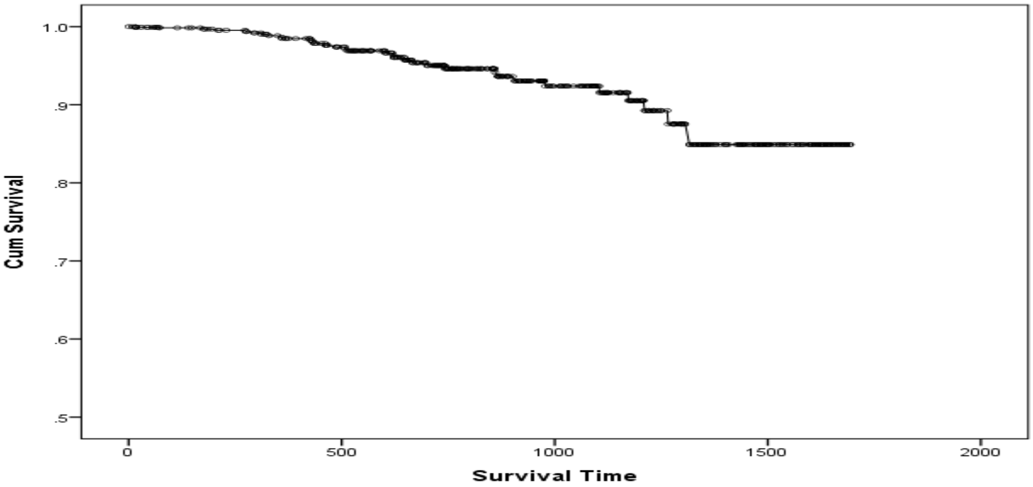


Figure A3.66 Cumulated survival for exponential loss distribution using mean 500: corrected the second model simulation c2

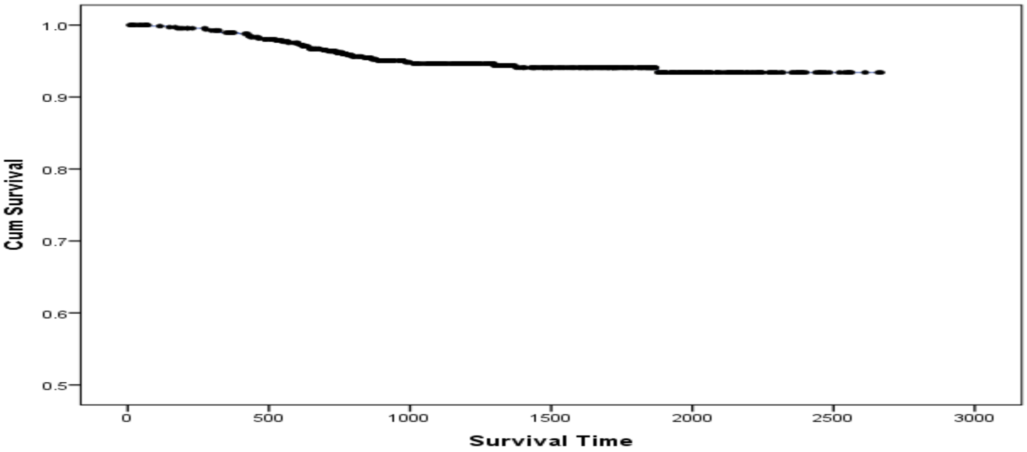


Figure A3.67 Uncorrected cumulative survival for exponential loss distribution using mean 500 simulation c3

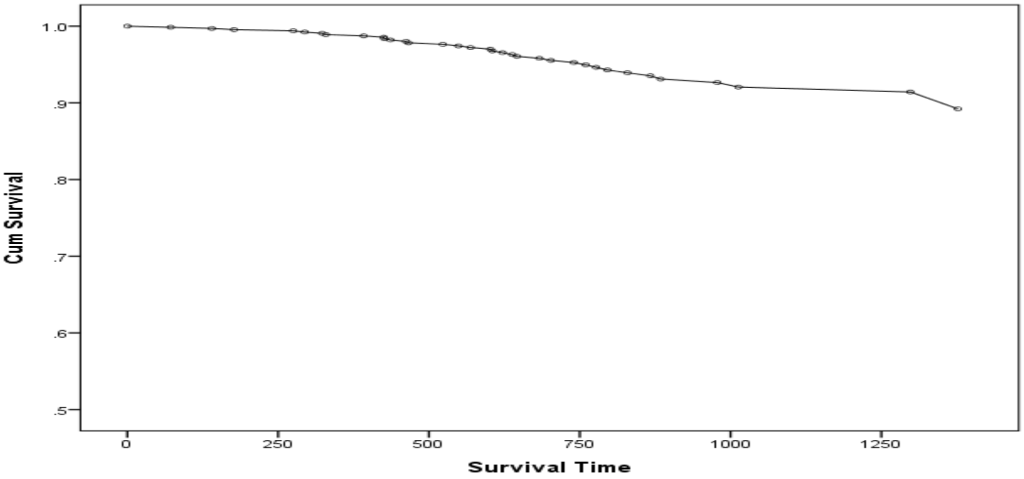


Figure A3.68 Cumulated survival for exponential loss distribution using mean 500: corrected the first model simulation c3

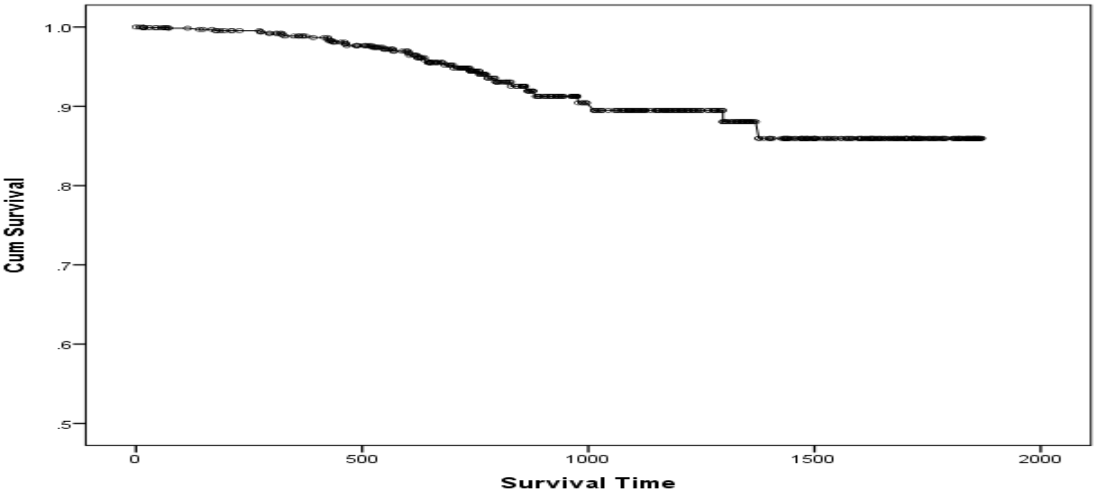


Figure A3.69 Cumulated survival for exponential loss distribution using mean 500: corrected the second model simulation c3

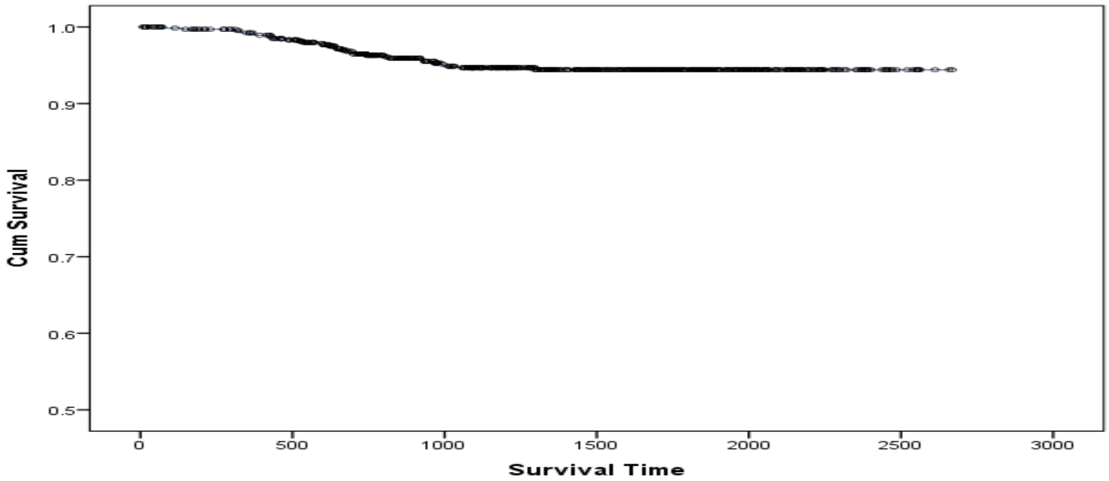


Figure A3.70 Uncorrected cumulative survival for exponential loss distribution using mean 500 simulation c4

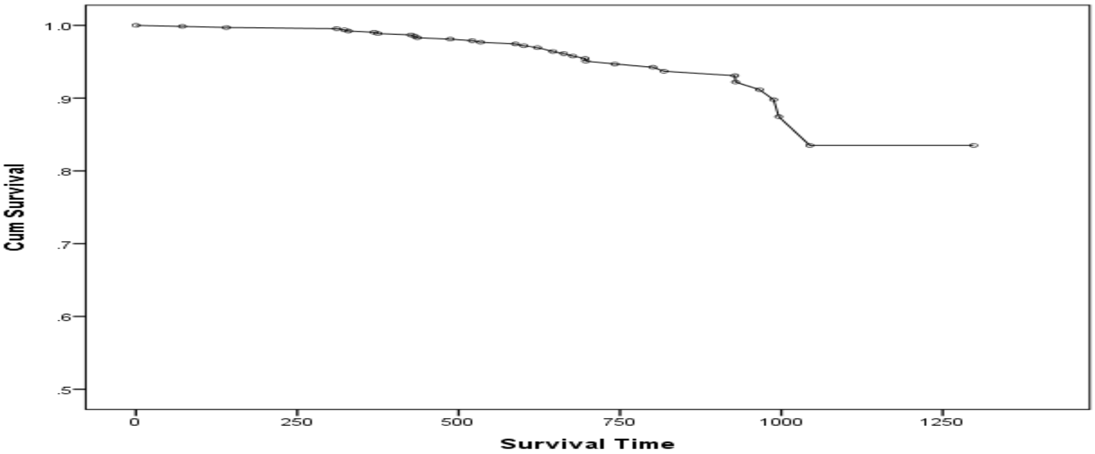


Figure A3.71 Cumulated survival for exponential loss distribution using mean 500: corrected the first model simulation c4

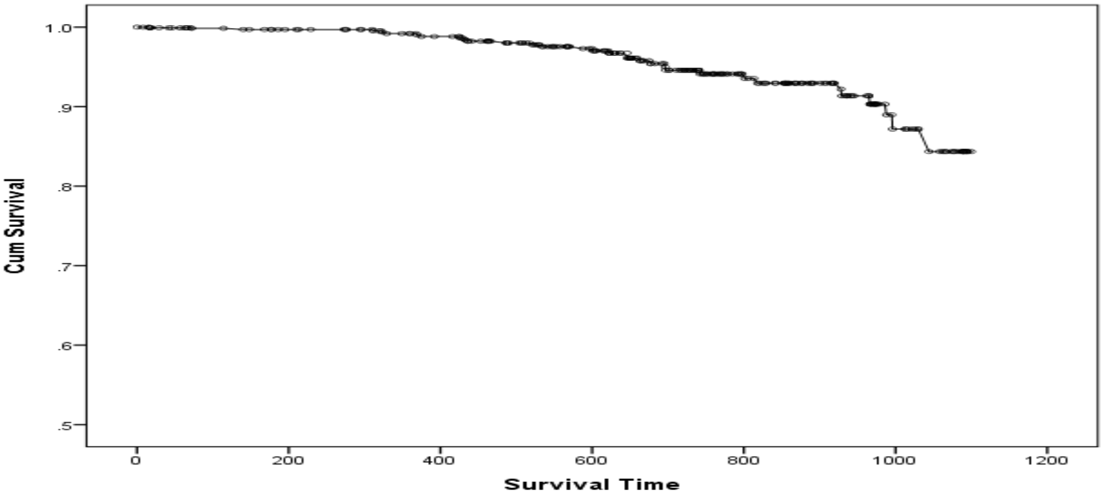


Figure A3.72 Cumulated survival for exponential loss distribution using mean 500: corrected the second model simulation c4

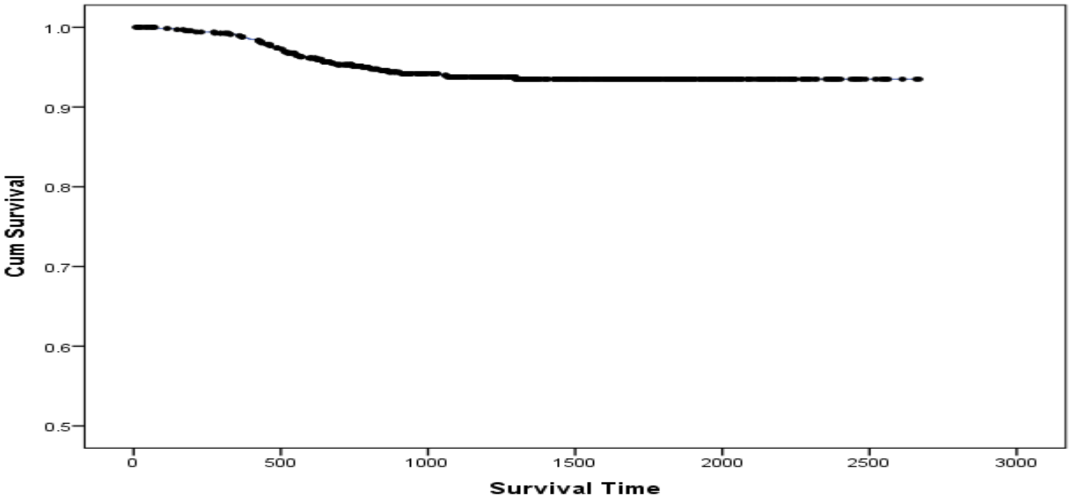


Figure A3.73 Uncorrected cumulative survival for exponential loss distribution using mean 500 simulation c5

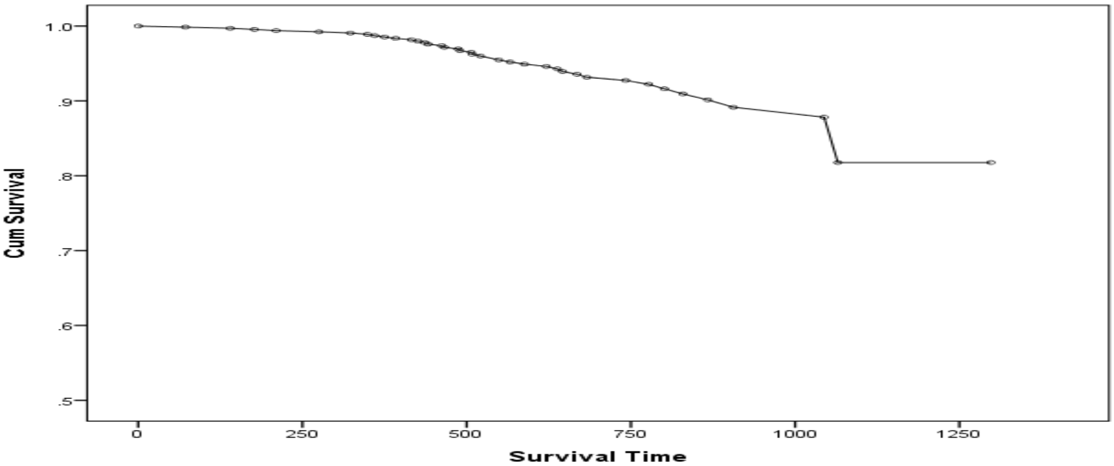


Figure A3.74 Cumulated survival for exponential loss distribution using mean 500: corrected the first model simulation c5

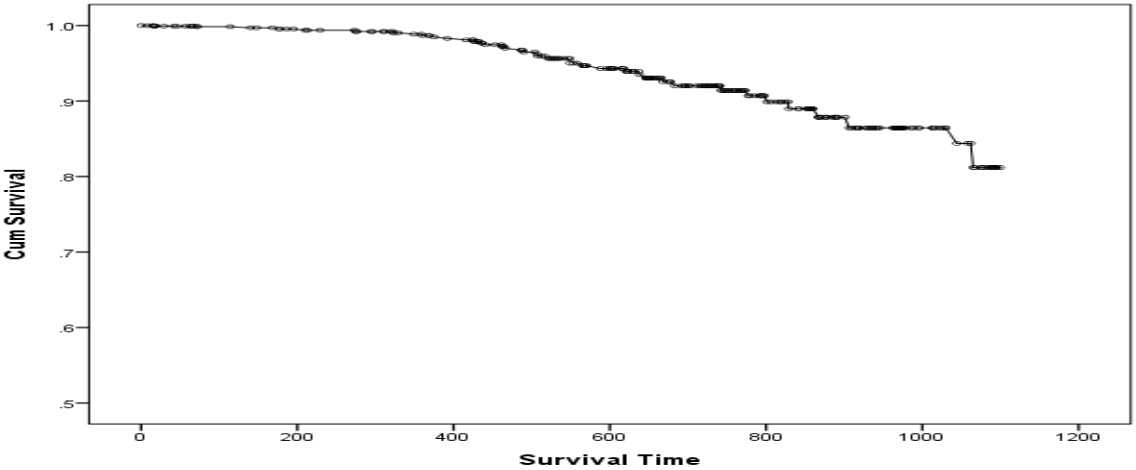


Figure A3.75 Cumulated survival for exponential loss distribution using mean 500: corrected the second model simulation c5

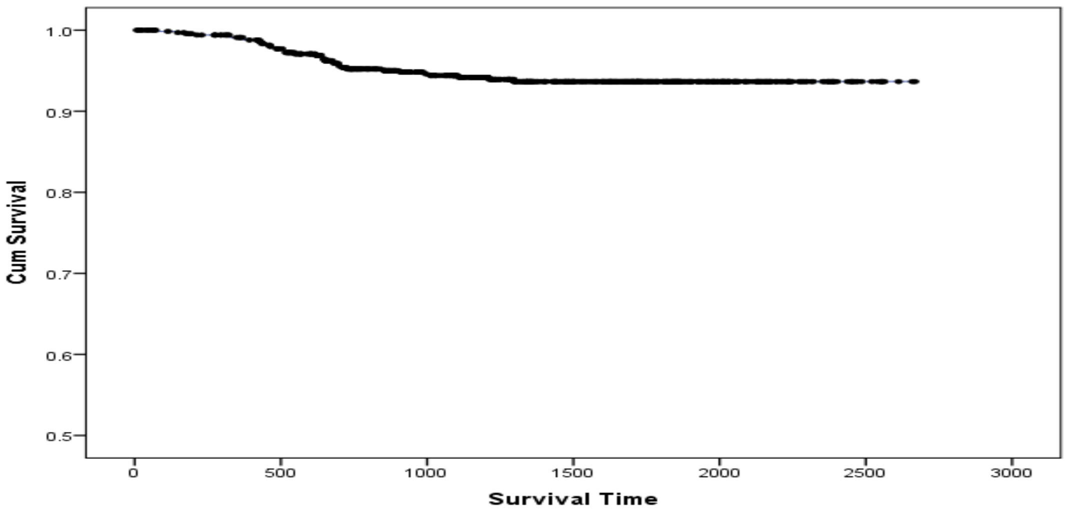


Figure A3.76 Uncorrected cumulative survival for exponential loss distribution using mean 500 simulation c6

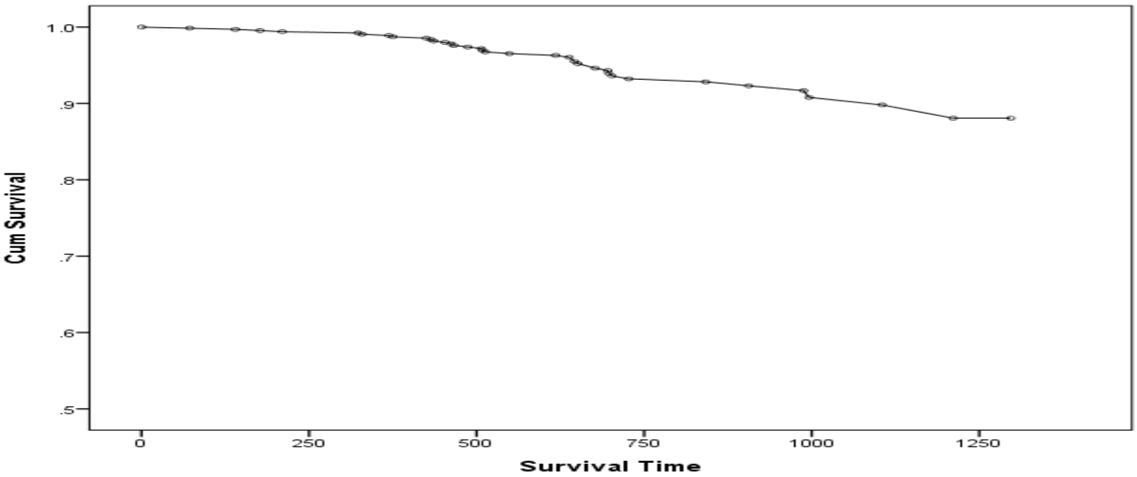


Figure A3.77 Cumulated survival for exponential loss distribution using mean 500: corrected the first model simulation c6

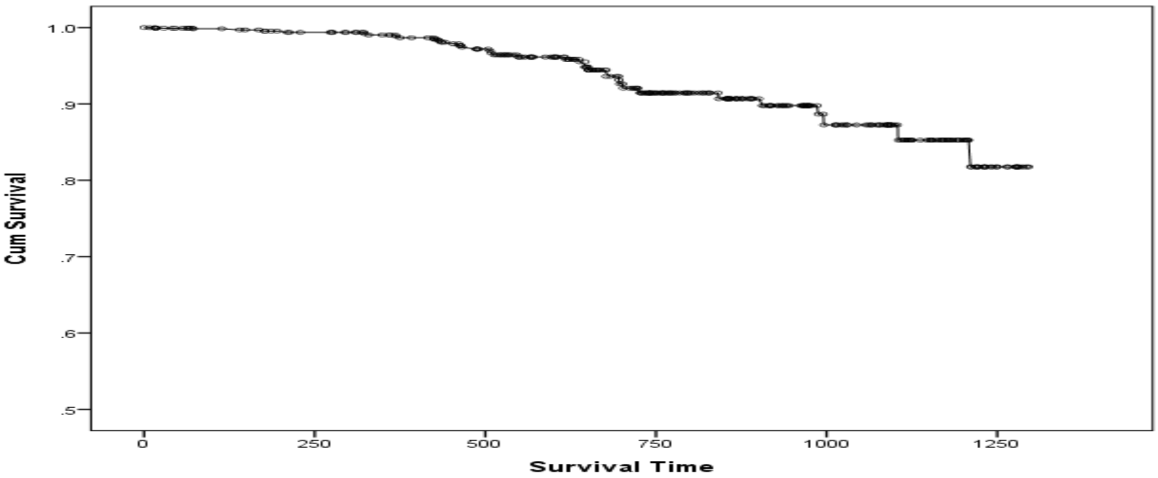


Figure A3.78 Cumulated survival for exponential loss distribution using mean 500: corrected the second model simulation c6

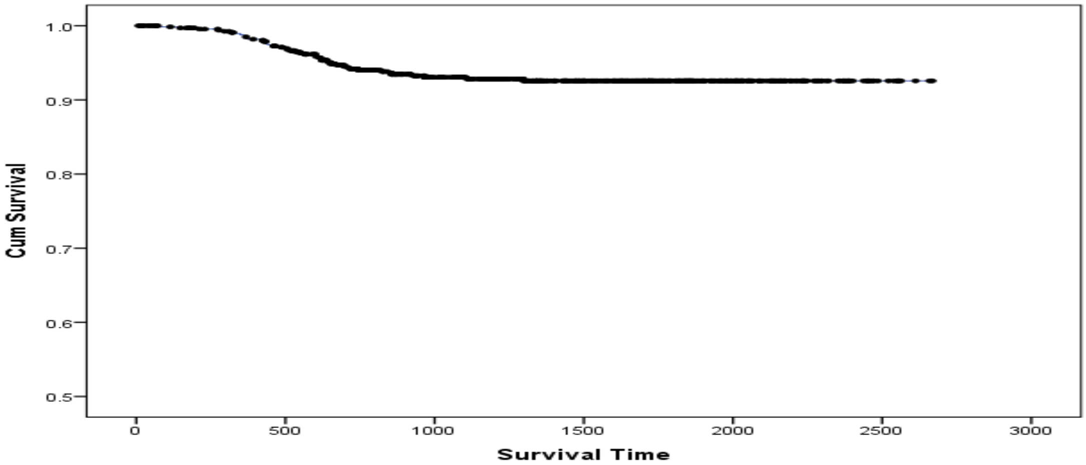


Figure A3.79 Uncorrected cumulative survival for exponential loss distribution using mean 500 simulation c7

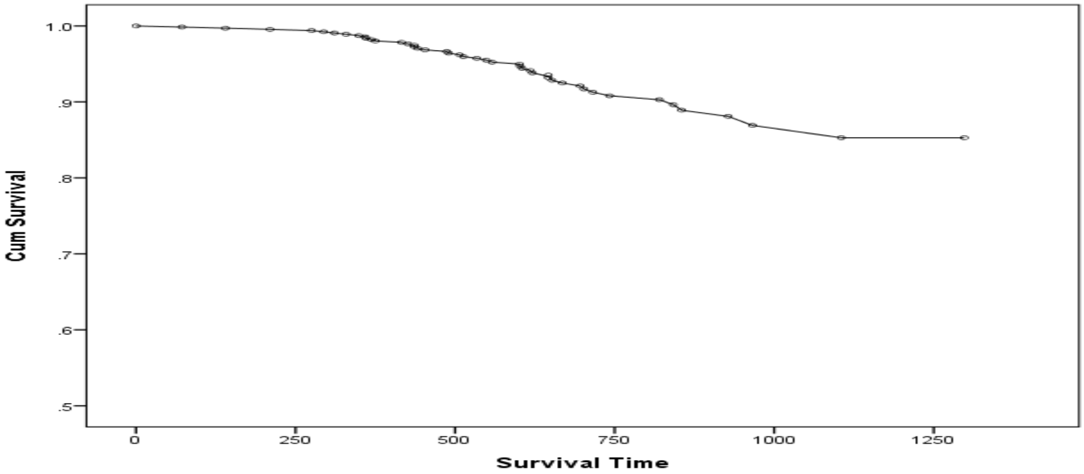


Figure A3.80 Cumulated survival for exponential loss distribution using mean 500: corrected the first model simulation c7

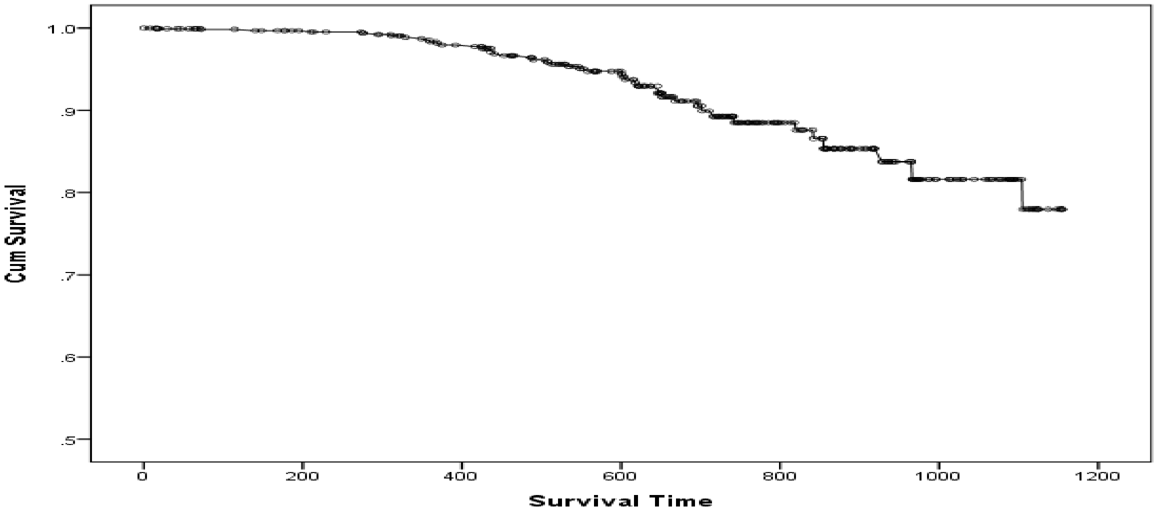


Figure A3.81 Cumulated survival for exponential loss distribution using mean 500: corrected the second model simulation c7

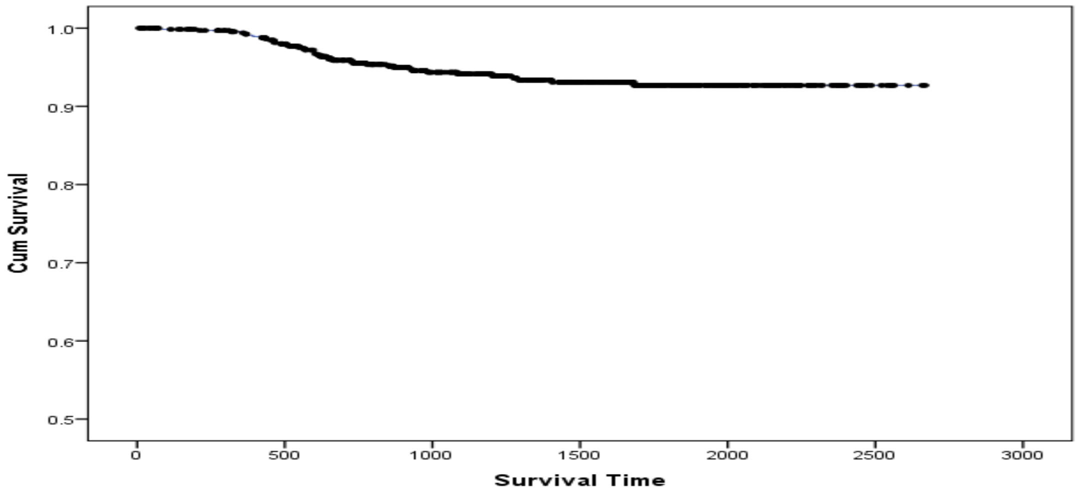


Figure A3.82 Uncorrected cumulative survival for exponential loss distribution using mean 500 simulation c8

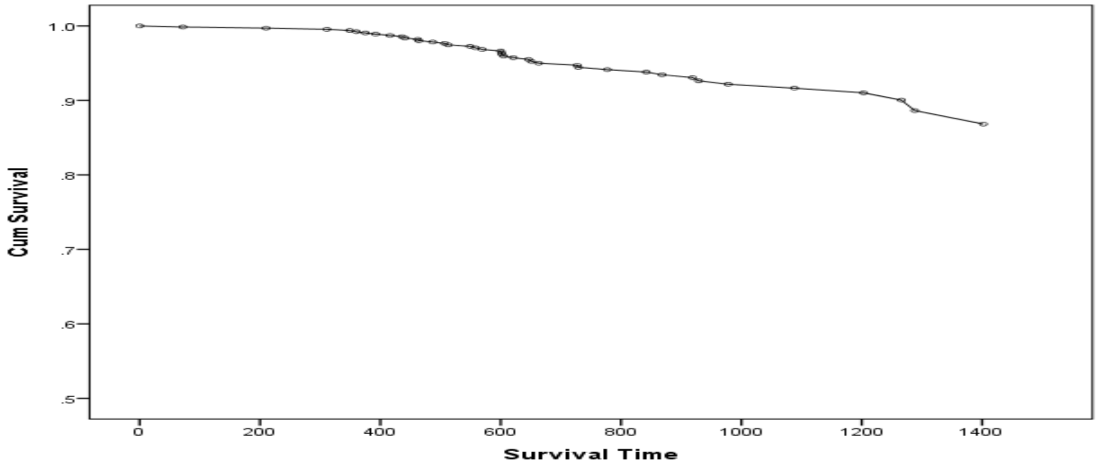


Figure A3.83 Cumulated survival for exponential loss distribution using mean 500: corrected the first model simulation c8

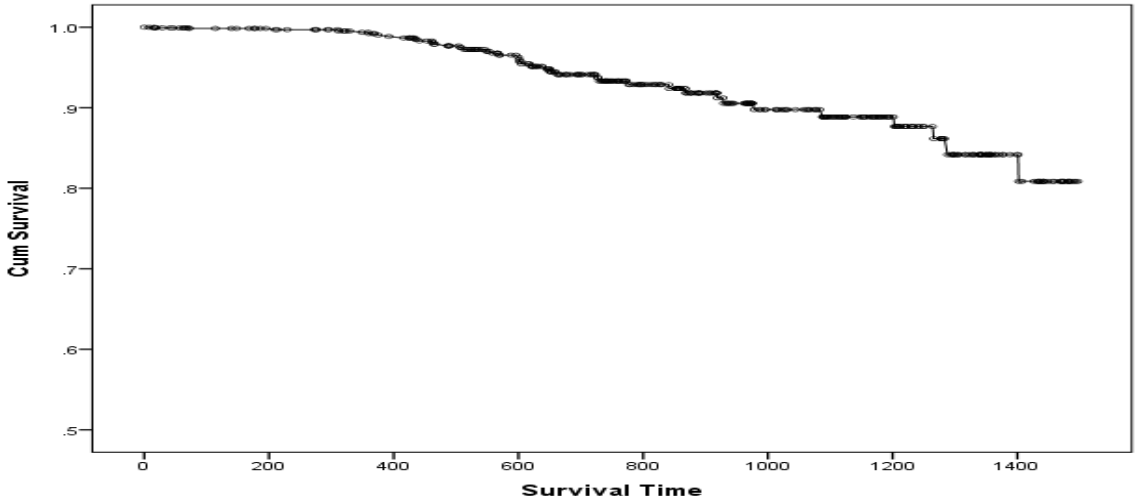


Figure A3.84 Cumulated survival for exponential loss distribution using mean 500: corrected the second model simulation c8

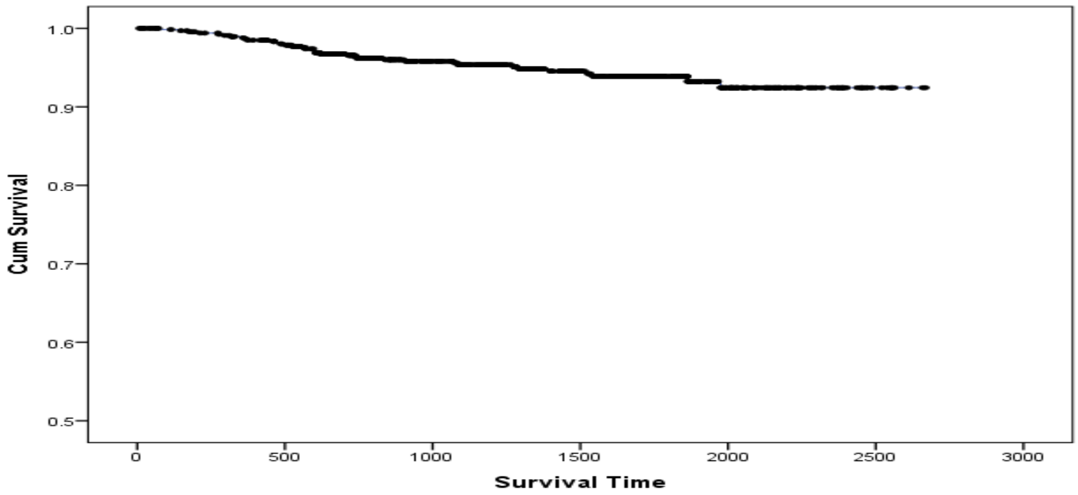


Figure A3.85 Uncorrected cumulative survival for exponential loss distribution using mean 500 simulation c9

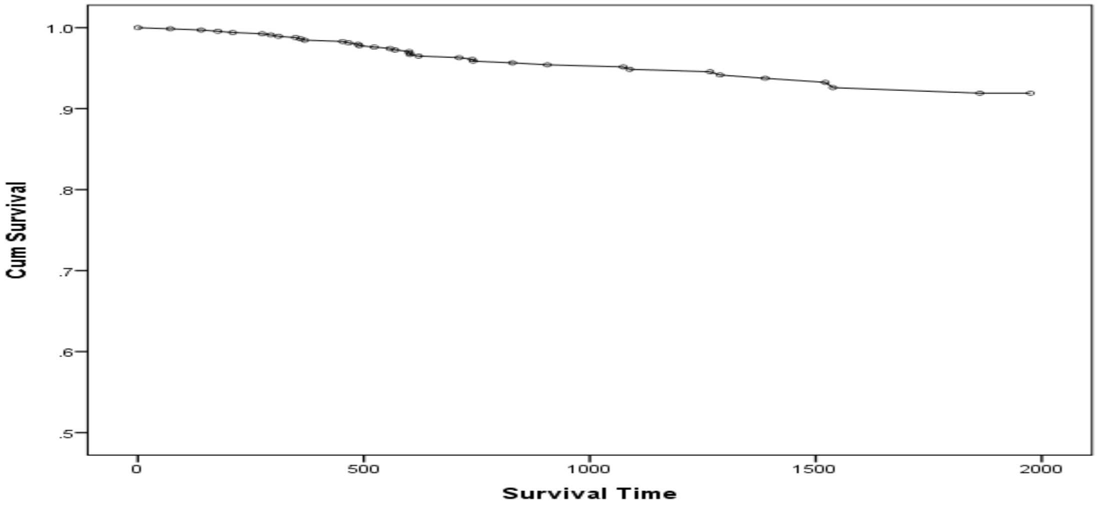


Figure A3.86 Cumulated survival for exponential loss distribution using mean 500: corrected the first model simulation c9

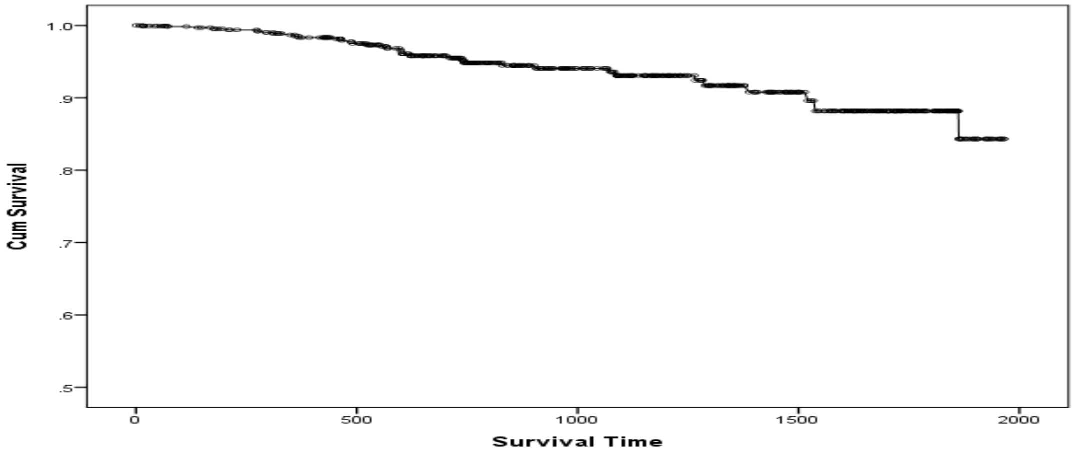


Figure A3.87 Cumulated survival for exponential loss distribution using mean 500: corrected the second model simulation c9

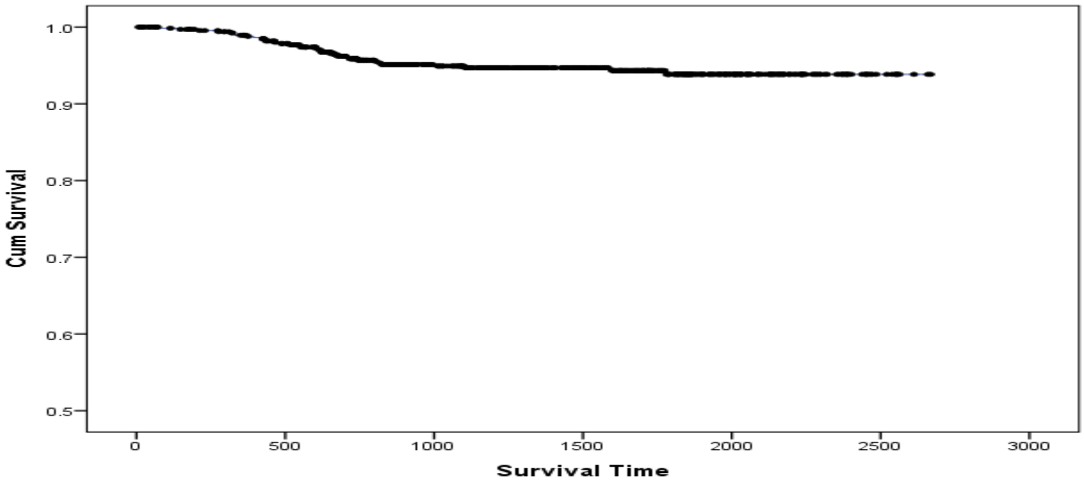


Figure A3.88 Uncorrected cumulative survival for exponential loss distribution using mean 500 simulation c10

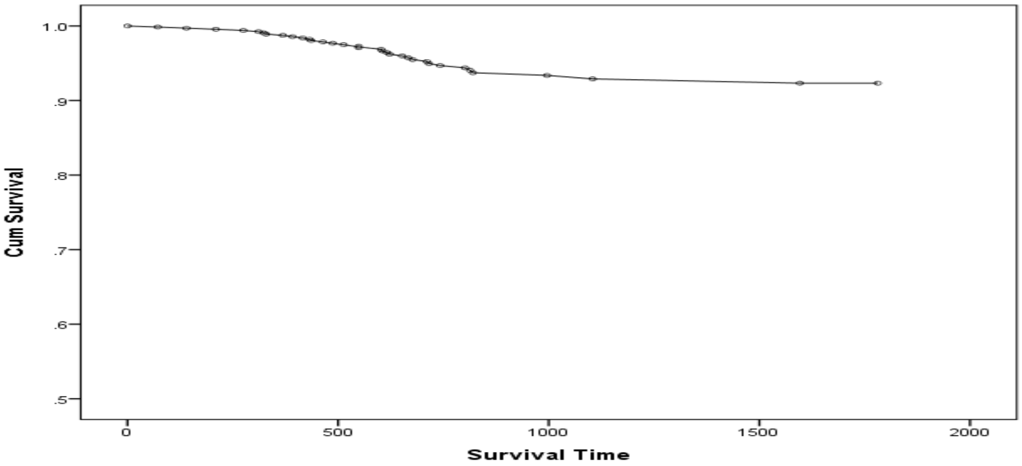


Figure A3.89 Cumulated survival for exponential loss distribution using mean 500: corrected the second model simulation c10

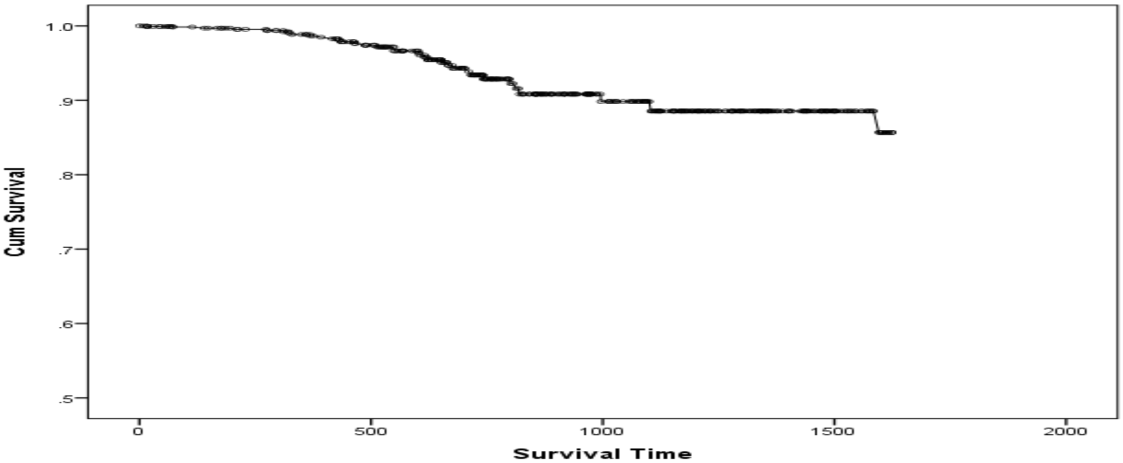


Figure A3.90 Cumulated survival for exponential loss distribution using mean 500: corrected the second model simulation c10

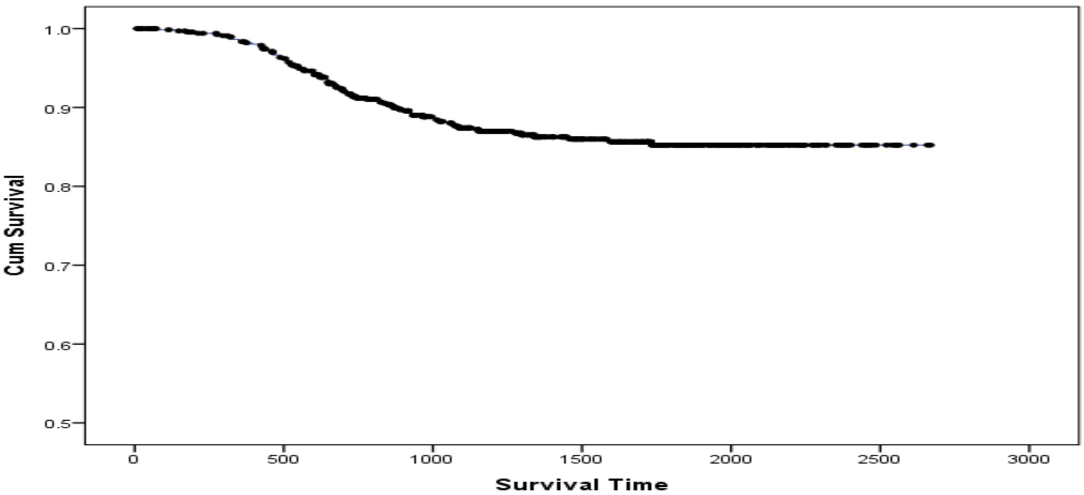


Figure A3.91 Uncorrected cumulative survival for gamma (3, 3/1000) loss distribution i.e. with mean 1000 simulation d1

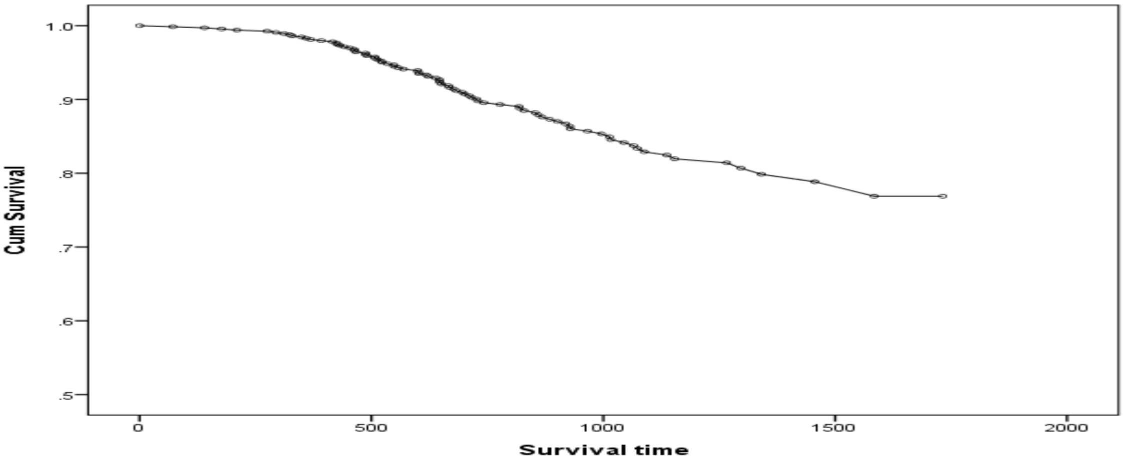


Figure A3.92 Cumulated survival for gamma (3, 3/1000) loss distribution i.e. with mean 1000: corrected the first model simulation d1

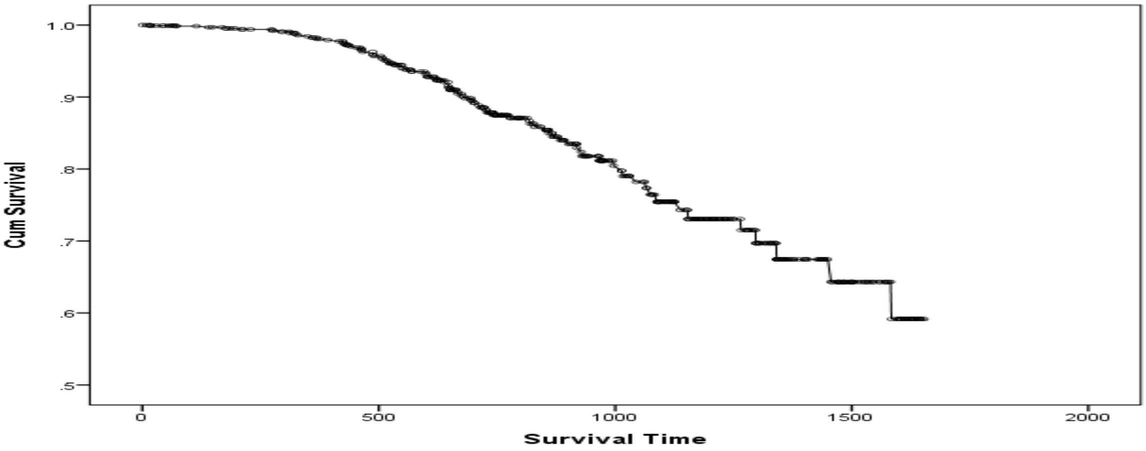


Figure A3.93 Uncorrected cumulative survival for gamma (3, 3/1000) loss distribution i.e. with mean 1000 The second model simulation d1

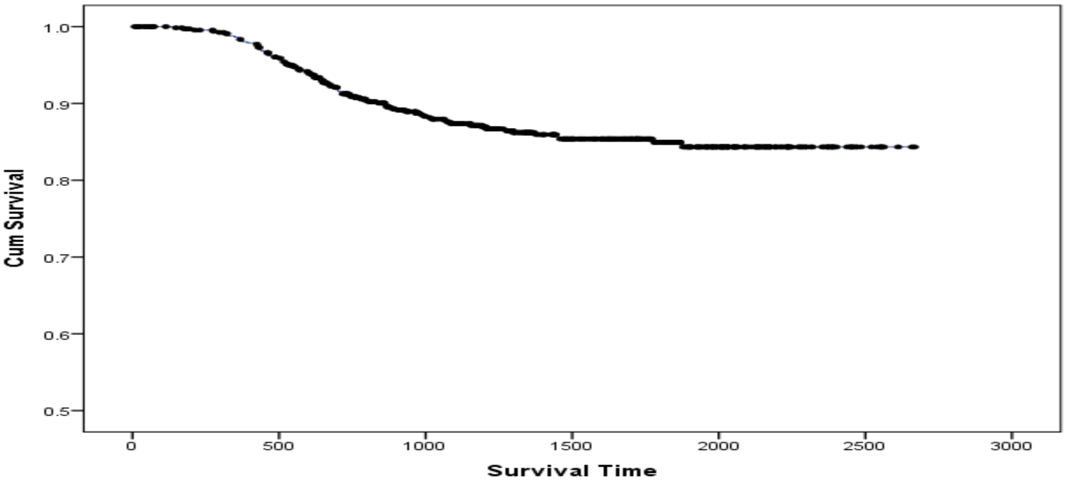


Figure A3.94 Uncorrected cumulative survival for gamma (3, 3/1000) loss distribution i.e. with mean 1000 simulation d2

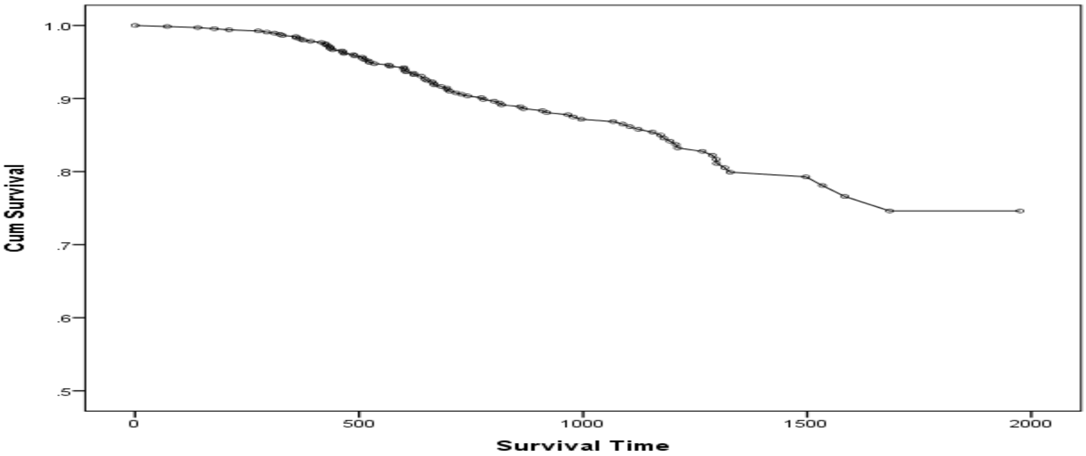


Figure A3.95 Cumulated survival for gamma (3, 3/1000) loss distribution i.e. with mean 1000: corrected the first model simulation d2

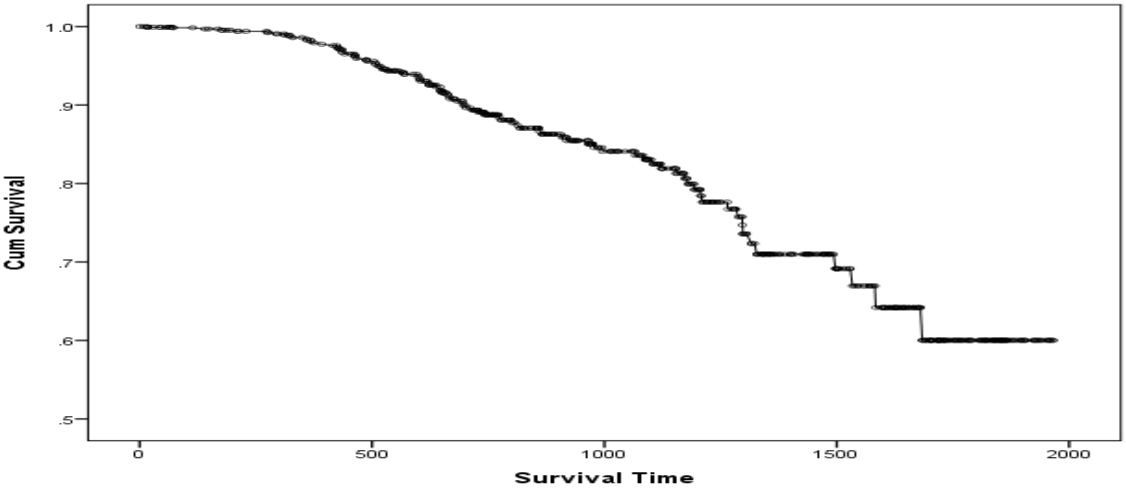


Figure A3.96 Cumulated survival for gamma (3, 3/1000) loss distribution i.e. with mean 1000: corrected the second model simulation d2

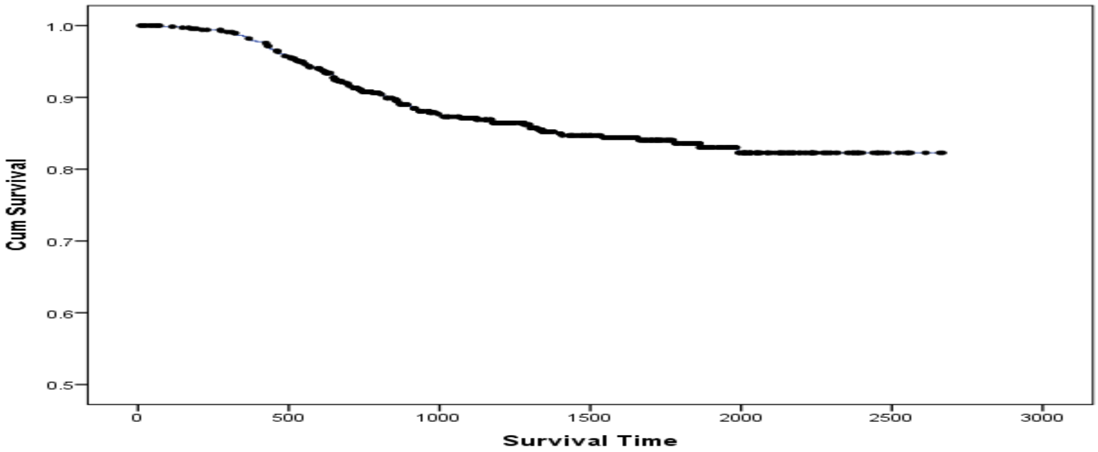


Figure A3.97 Uncorrected cumulative survival for gamma (3, 3/1000) loss distribution i.e. with mean 1000 simulation d3

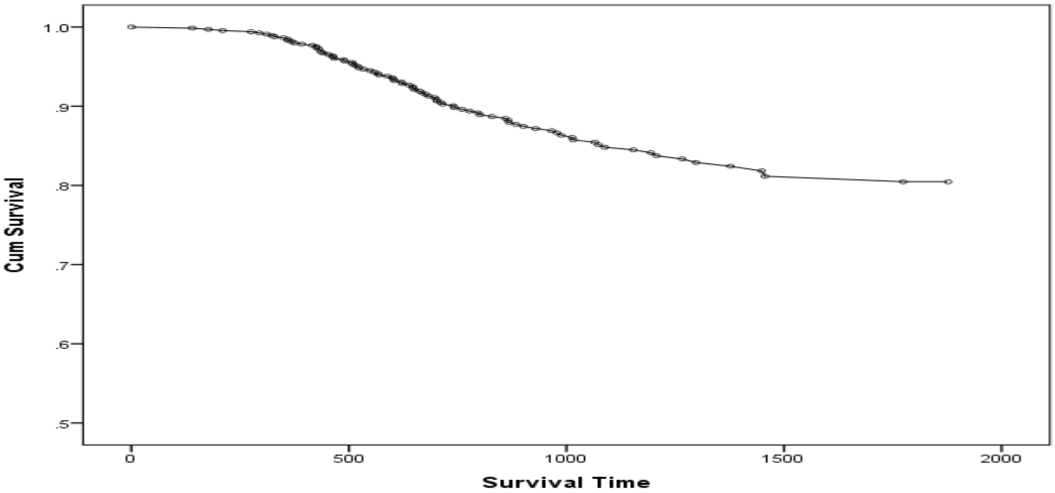


Figure A3.98 Cumulated survival for gamma (3, 3/1000) loss distribution i.e. with mean 1000: corrected the first model simulation d3

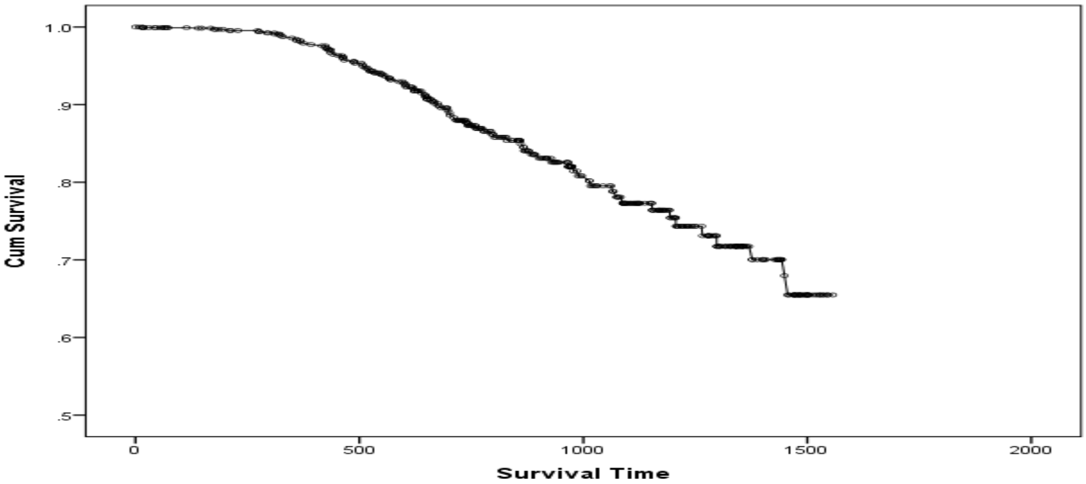


Figure A3.99 Cumulated survival for gamma (3, 3/1000) loss distribution i.e. with mean 1000: corrected the second model simulation d3

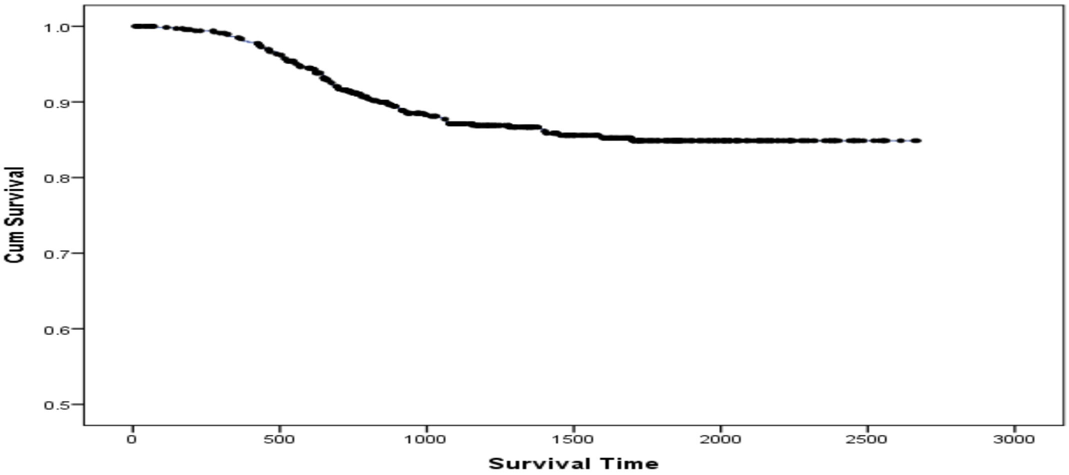


Figure A3.100 Uncorrected cumulative survival for gamma (3, 3/1000) loss distribution i.e. with mean 1000 simulation d4

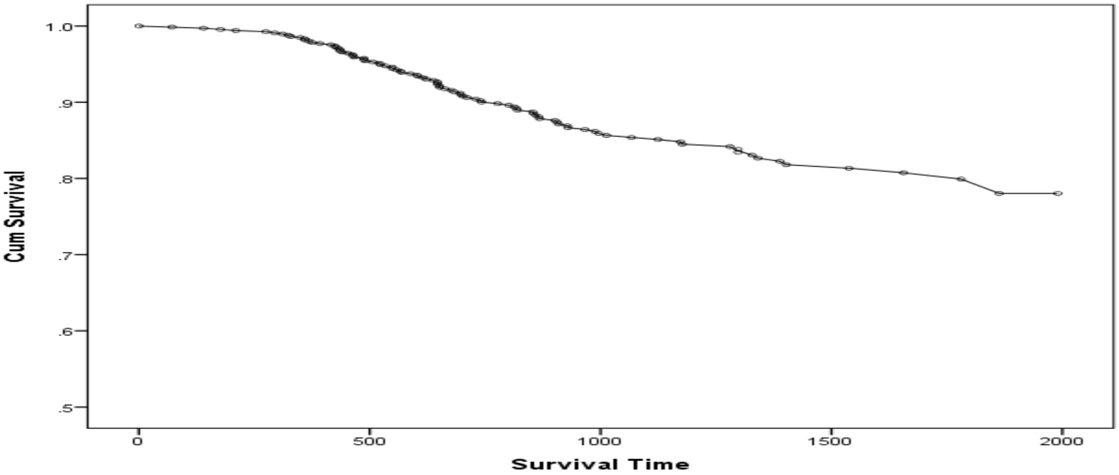


Figure A3.101 Cumulated survival for gamma (3, 3/1000) loss distribution i.e. with mean 1000: corrected the first model simulation d4

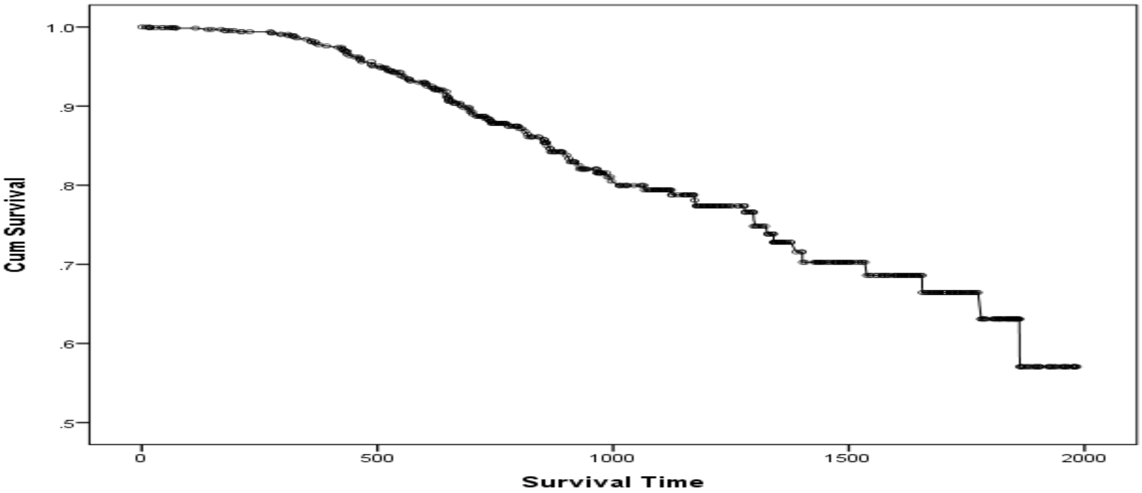


Figure A3.102 Cumulated survival for gamma (3, 3/1000) loss distribution i.e. with mean 1000: corrected the second model simulation d4

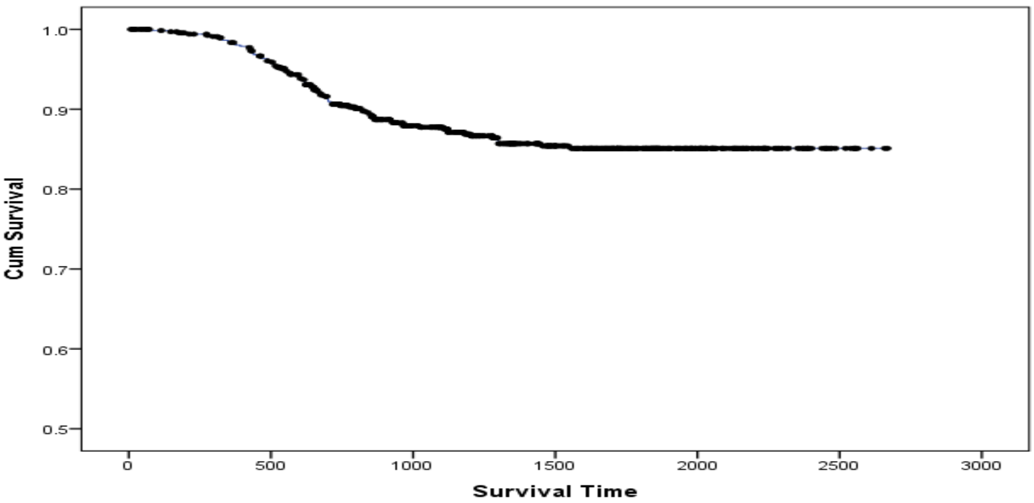


Figure A3.103 Uncorrected cumulative survival for gamma (3, 3/1000) loss distribution i.e. with mean 1000 simulation d5

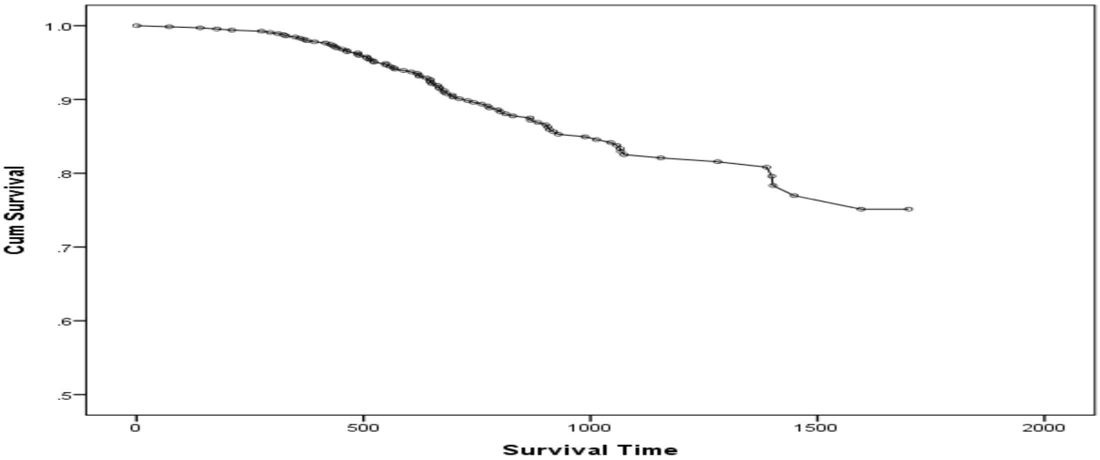


Figure A3.104 cumulative survival for gamma (3, 3/1000) loss distribution i.e. with mean 1000: corrected the first model simulation d5

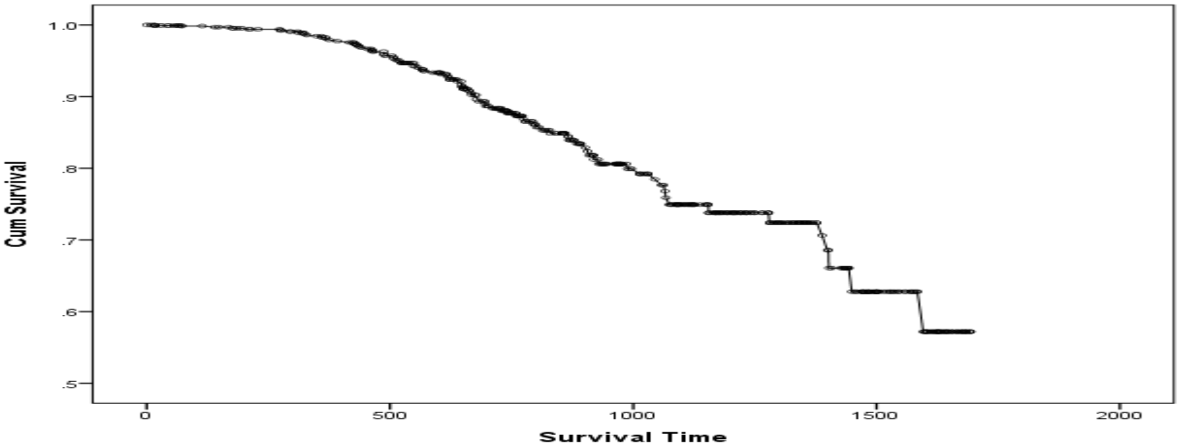


Figure A3.105 Cumulated survival for gamma (3, 3/1000) loss distribution i.e. with mean 1000: corrected the second model simulation d5

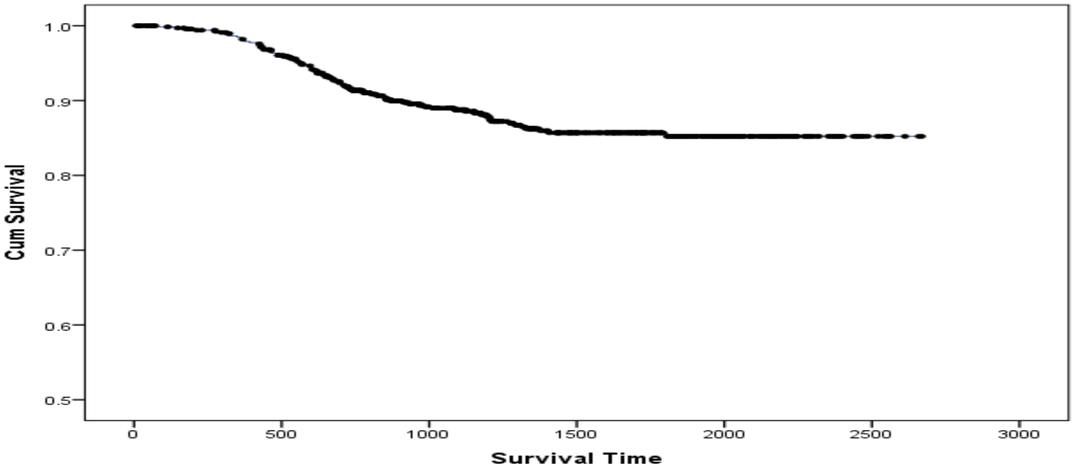


Figure A3.106 Uncorrected cumulative survival for gamma (3, 3/1000) loss distribution i.e. with mean 1000 simulation d6

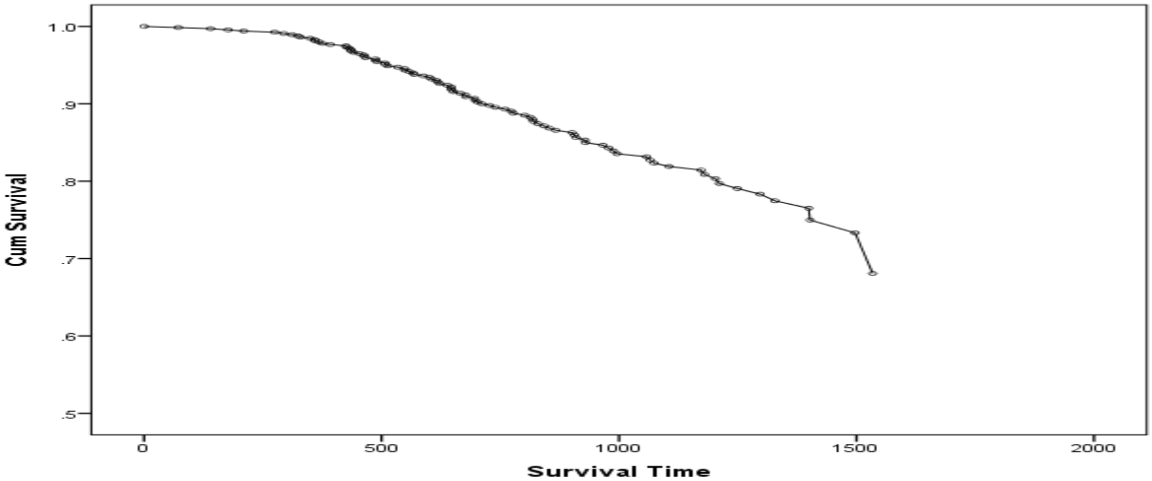


Figure A3.107 Cumulated survival for gamma (3, 3/1000) loss distribution i.e. with mean 1000: corrected the first model simulation d6

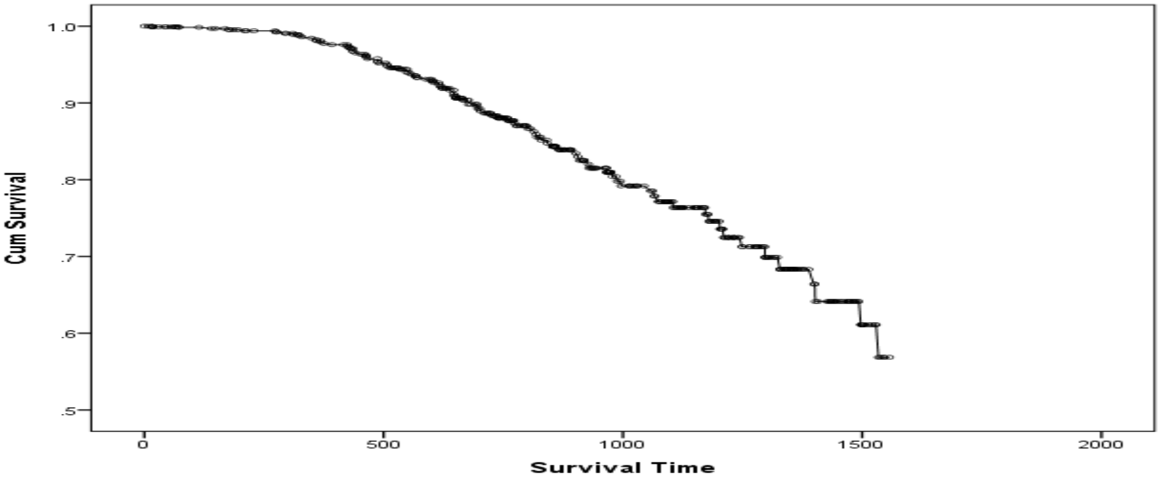


Figure A3.108 Cumulated survival for gamma (3, 3/1000) loss distribution i.e. with mean 1000: corrected the second model simulation d6

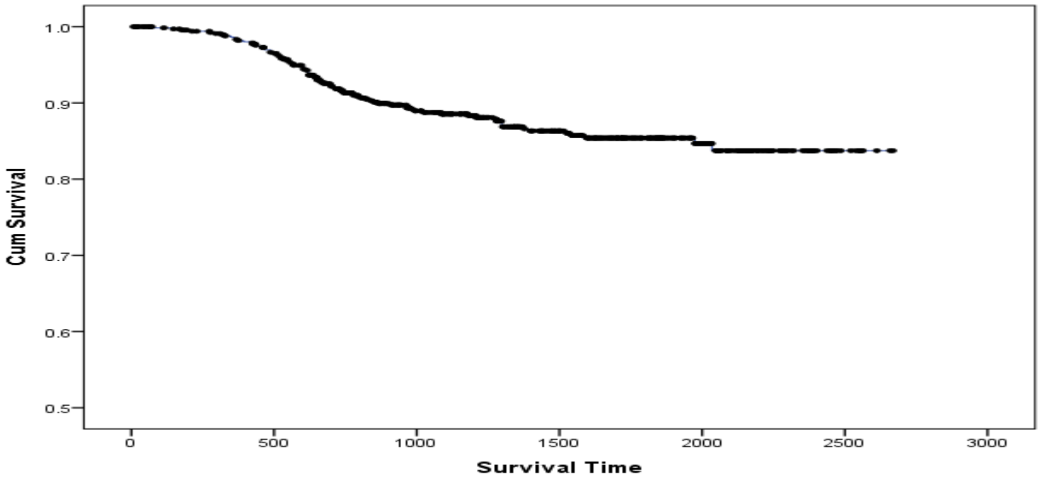


Figure A3.109 Uncorrected cumulative survival for gamma (3, 3/1000) loss distribution i.e. with mean 1000 simulation d7

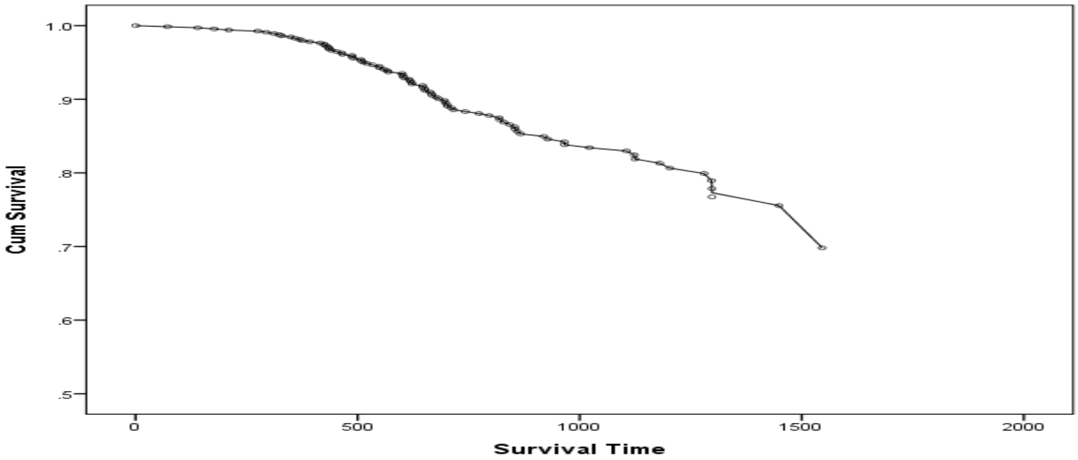


Figure A3.110 Cumulated survival for gamma (3, 3/1000) loss distribution i.e. with mean 1000: corrected the first model simulation d7

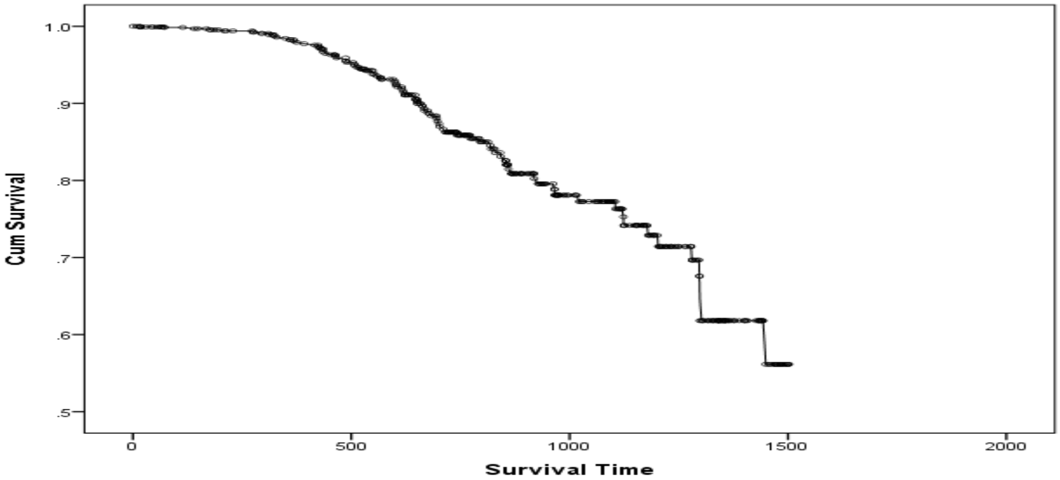


Figure A3.111 Cumulated survival for gamma (3, 3/1000) loss distribution i.e. with mean 1000: corrected the second model simulation d7

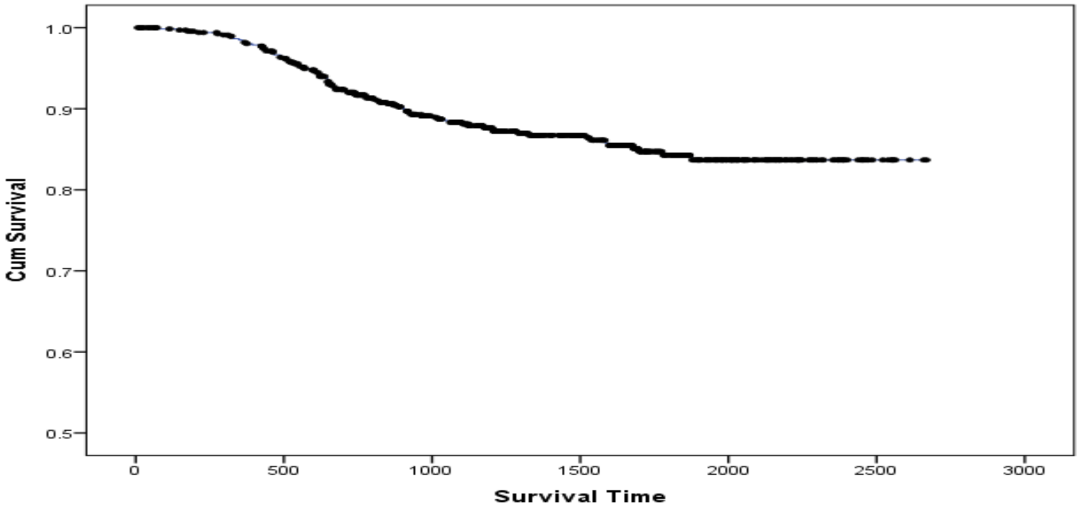


Figure A3.112 Uncorrected cumulative survival for gamma (3, 3/1000) loss distribution i.e. with mean 1000 simulation d8

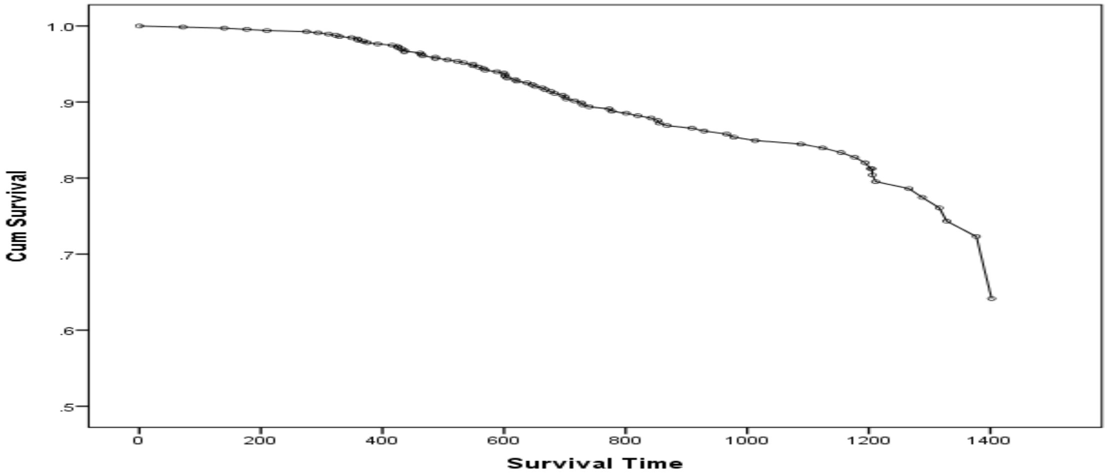


Figure A3.113 Cumulated survival for gamma (3, 3/1000) loss distribution i.e. with mean 1000: corrected the first model simulation d8

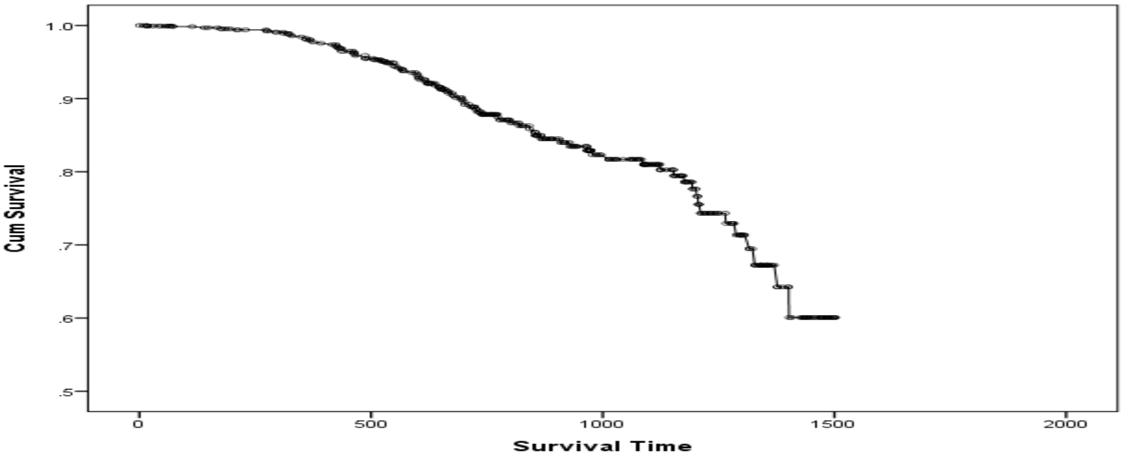


Figure A3.114 Cumulated survival for gamma (3, 3/1000) loss distribution i.e. with mean 1000: corrected the second model simulation d8

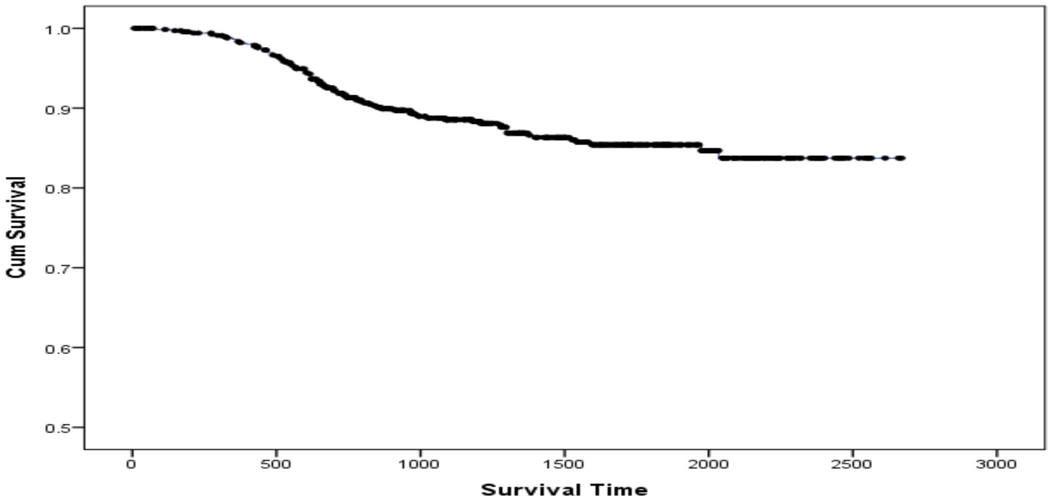


Figure A3.115 Uncorrected cumulative survival for gamma (3, 3/1000) loss distribution i.e. with mean 1000 simulation d9

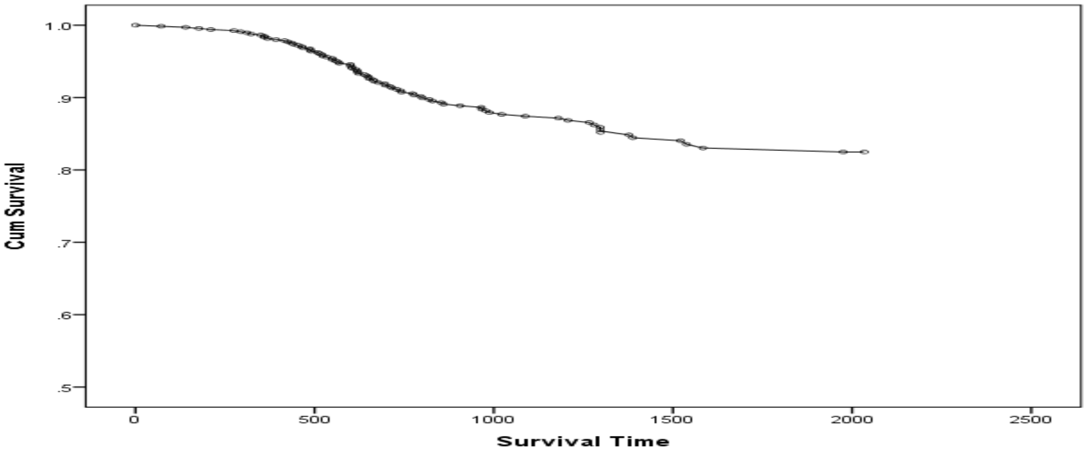


Figure A3.116 Cumulated survival for gamma (3, 3/1000) loss distribution i.e. with mean 1000: corrected the first model simulation d9

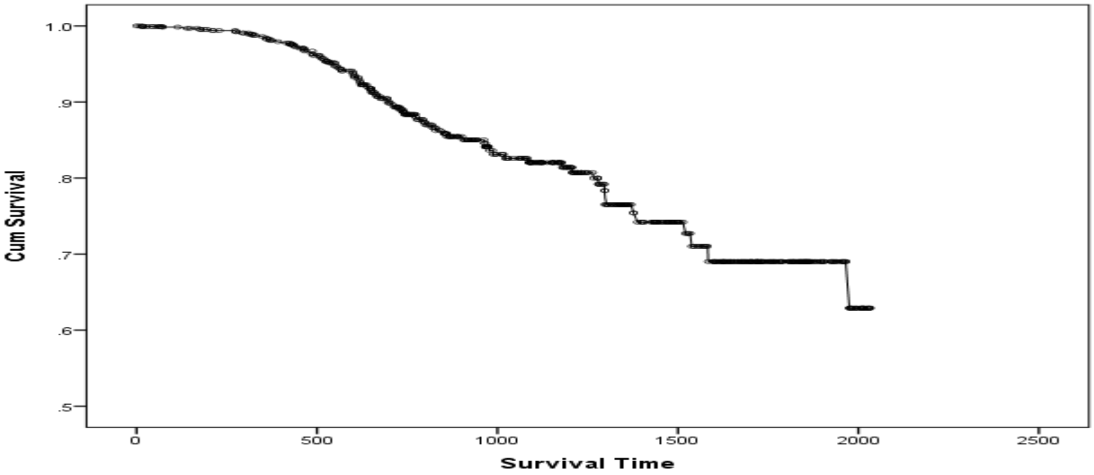


Figure A3.117 Cumulated survival for gamma (3, 3/1000) loss distribution i.e. with mean 1000: corrected the second model simulation d9

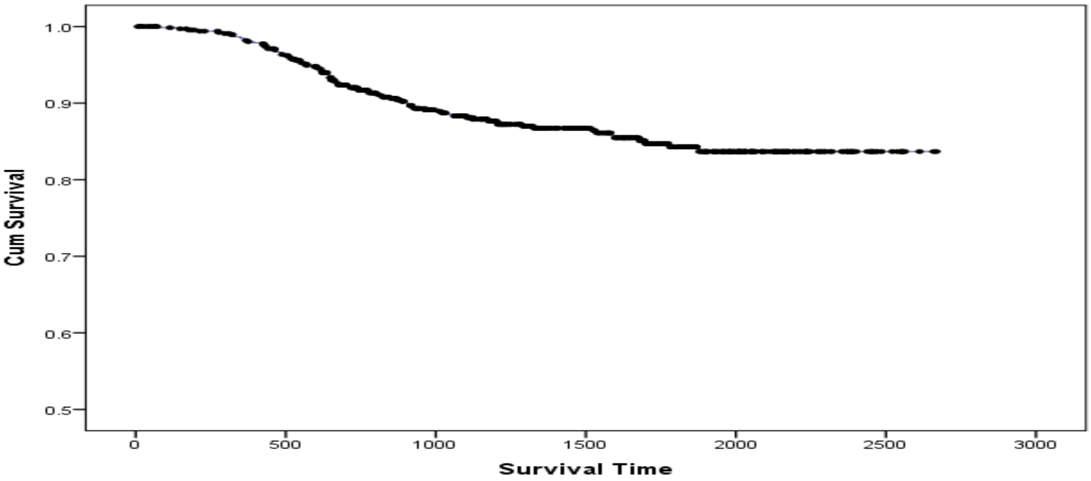


Figure A3.118 Uncorrected cumulative survival for gamma (3, 3/1000) loss distribution i.e. with mean 1000 simulation d10

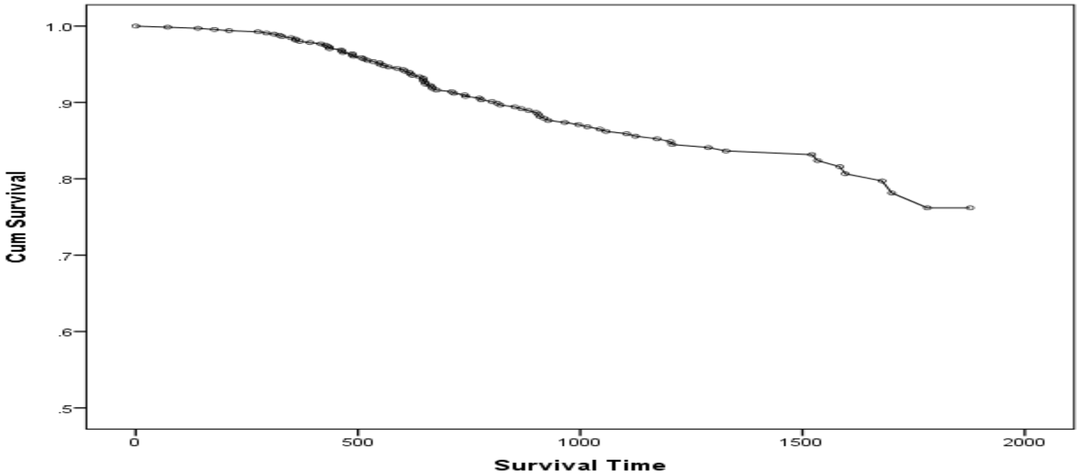


Figure A3.119 Cumulated survival for gamma (3, 3/1000) loss distribution i.e. with mean 1000: corrected the first model simulation d10

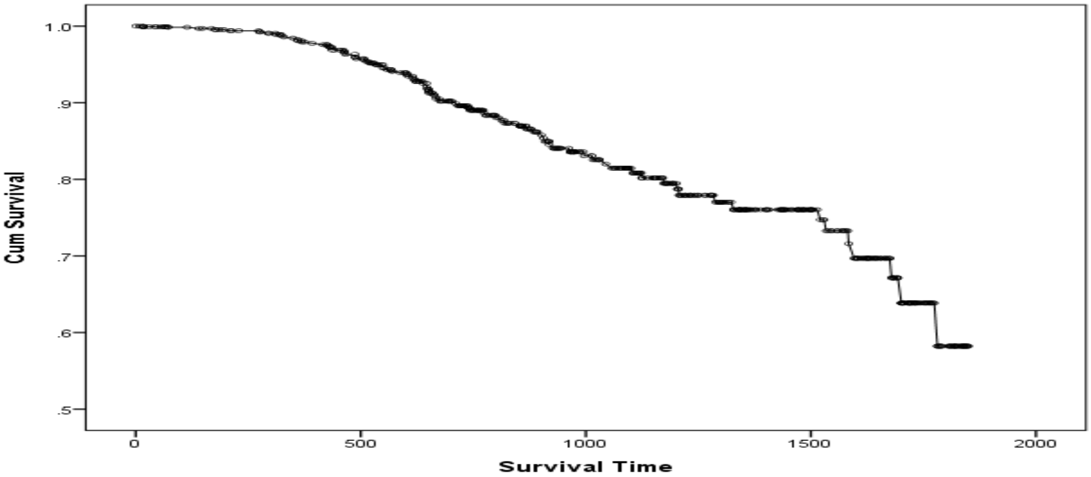


Figure A3.120 Cumulated survival for gamma (3, 3/1000) loss distribution i.e. with mean 1000: corrected the second model simulation d10

References

Abadi, A., Farzaneh, A., Chris, B. and Parvin, Y. (2012). Breast Cancer Survival Analysis: Applying the Generalized Gamma Distribution under Different Conditions of the Proportional Hazards and Accelerated Failure Time Assumptions. International Journal of Preventive Medicine, Vol. *3,* No. 9, pp. 644–651

Abdelkrim, S., Amel, T., Nabiha, M., Nadia, B., Ahlam, B., Affissath, A., Wafa, J. and Moncef, M. (2010). Distribution of Molecular breast cancer subtypes among Tunisian women and correlation with histopathological parameters: A study of 194 patients. *Pathology-Research and Practice, Vol. 206*, pp. 772-775

About Nanakaly Hospital for Hematology& Oncology https://en.wikipedia.org/w/index.php?title=Nanakaly_Hospital_for_Hematology_%26_Oncology&gettingStartedReturn=true

Abuelghar, W. M., Elsaeed, M. M., Tamara, T. F., Elaithy, M. I. and Ali, M. S. (2013). Measurement of serum estradiol / progesterone ratio on the day of embryo transfer to predict clinical pregnancies in injection (ICSI) cycles. Is this of real clinical value? *Middle East Fertility Society Journal*. *Vol.* 18, No. 1, pp. 31-37.

Abu-Taleb, A.A., Smadi, M.M. and Alawneh, A.J. (2007). Bayes estimation of the lifetime parameters for the exponential distribution. *J. Math. Statist.*, *Vol.3,* No.3, pp. 106-108.

Aggarwal, A., Karen, F., Alicia, S., Lucille L. A., Ana, M. L., Lawrence, S. L., Judith, O., Robert B. W., Carla D. W. and Denise, E. B. (2008). Are Depressive Symptoms Associated with Cancer Screening and Cancer Stage at Diagnosis among Postmenopausal Women? The Women’s Health Initiative Observational Cohort. *Journal of Women's Health, Vol. 17*, No. 8, pp. 1353-1361.

Ahmed, E.S., Volodin, A.I. and Hussein, A. (2005). Robust weighted likelihood estimation of exponential parameters, *IEEE Trans. Reliab, Vol. 54,* No.3, pp. 389-395.

Al-Eid, H. S. (2012). *Cancer Incidence and Survival Report Sau- di Arabia 2007*. Accessed from http://www.scr.org.sa/reports/SCR2007.pdf

Al Tamimi, D., Mohamed, A., Ayesha, A., Ammar, K. and Amal, A. (2010). Portion expression profile and prevalence pattern of the molecular classes of breast cancer - a Saudi population based study, *BioMed Central Cancer, Vol. 10,* No. 223. pp. 1-13.

Al-Humadi, A. H. (2009). Epidemiology of Colon & Rectal Cancer in Iraq. *World Journal of Colorectal Surgery*, *Vol. 15,* No.1.

Ali, M. M., Woo, J. and Nadarajah, S. (2005). Bayes estimators of the exponential distribution, *J. Statist. Mgmt. Sys., Vol. 8,* No. 1. pp. 53-58.

Al-Riad al Sharif, (2012). Investigations and reportages: Hospital (hiwa) Oncology and Hematology in Sulaimaniyah. The Union, daily political newspaper. No. 3707 issued on Sep. 16, 2012. Sited in Jan. 18, 2015 from http://www.alitthad.com/paper.php?name=News&file=article&sid=125630

Althius, M. D., Dozier, J. M., Anderson, W. F., Devessa, S. S. and Brinton, L. A. (2005). Global trends in breast cancer incidence and mortality 1973-1997. *International Journal of Epidemiology*, *Vol.* 34, pp. 405-412.

Alwan, N. A., Al-Kubaisy, W., Al-Rawaq, K. (2000). Assessment of response to tamoxifen among Iraqi patients with advanced breast cancer. *East Mediterr Health Journal*, *Vol. 6*, pp. 475–482.

Alwan, N. A. (2010). Breast Cancer Incidence Among Iraqi Women Profiled. Accessed from http://www.sciencedaily.com/releases/2010/03/100311074127.htm on April, 25, 2013.

American Cancer Society (ACS) (2006). Breast Cancer: Treatment Guidelines for

American Cancer Society (ACS), Inc, (2011). Cancer in Africa. pp. 1-16.

American Cancer Society, *Breast Cancer Facts & Figures 2007-2008.*

America's Health Insurance Plans (2004). Reviewed 10 March 2015. Available at http://www.ahip.org/content/fileviewr.aspx.

Andersen, P. K., Borgan, O., Gill, R. D. and Keiding, N. (1992). *Statistical Models based on Counting Processes*. New York: Springer.

Anderson, B. O., Yip, C. H., Smith, R. A., Shyyan, R., Sener, S. F., Eniu, A., Carlson, R.W., Azavedo, E. and Harford, J. (2008). Guideline implementation for breast healthcare in low-income and middle-income countries: overview of the Breast Health Global Initiative Global Summit 2007. *Cancer, Vol, 113*, pp. 2221–2243.

Annual Cancer Conference in Erbil February 4, 2015.

Arjas, E. and Gasbarra, D. (1994). Nonparametric Bayesian inference from right censored survival data using the Gibbs sampler. *Statist. Sin., Vol, 4*, pp. 505-524.

Arjas, E. and Heikkinen, J. (1997). An algorithm for nonparametric Bayesian estimation of a Poisson intensity. *Comput. Statist*., *Vol. 12*, pp. 385-402.

Baglietto, L., Dallas, R. E., Dorota, M. G., John, L. H., Graham, G. G. (2005). Does dietary folate intake modify effect of alcohol consumption on breast cancer risk? Prospective cohort study. *British Medical Juornal. Vol. 331*, No. 7520, pp. 807-810.

Bala, M. (2005). Data Collection Procedures. Indira Gandhi National Open University. Sited on March 25, 2015 from http://www.celt.mmu.ac.uk/researchmethods/Modules/Data_collection/index.php

Bellizzi, K.M., Rowland, J.H., Jeffery, D.D., McNeel, T. (2005). Health behaviors of cancer survivors: examining opportunities for cancer control intervention. *Journal of Clinical Oncology, Vol. 23*, pp. 8884-8893

Biggar**,** R**.,** Elisabeth, W. A.**,** Niels, K., Jan, W. and Mads, M. (2013). Breast cancer in women using digoxin: tumor characteristics and relapse risk. *BioMed Central Cancer, Vol*. 15, No. R13, pp. 1-9.

Blanchard, C.M., Denniston, M.M., Baker, F., Ainsworth, S.R., Courneya, K.S., Hann, D.M., Gesme, D.H., Reding, D., Flynn, T. and kennedy, J.S. (2003). Do adults change their lifestyle behaviors after a cancer diagnosis? *American Journal of Health Behavior, Vol. 27*, No. 3 pp. 246-256.

Bocchino, C. (2004). Racial and Ethnic Data Collection by Health Plans. In: VerPlog M, Perrin E, editors. Eliminating Disparities: Measurement and Data Needs. Washington, DC: National Academies Press; National Research Council Report. pp. 272–87. http://www.jstor.org/stable/25460784

Boffetta, P. Franco, M., Regina, W., Corrado, M., Alberto, P. and Benedetto, T. (1993). Survival of breast cancer patients from Piedmont, Italy, *Cancer Causes & Control, Vol. 4,* No. 3, pp 209-215. http://link.springer.com/article/10.1007%2FBF00051315

Bower, J. E., Ganz, P. A., Desmond, K. A., Bernaards, C., Rowland, J.H., Meyerowitz, B. E. and Belin, T.R. 2006. Fatigue in long-term breast carcinoma survivors: a longitudinal investigation. *Cancer*, *Vol. 106,* No. 4, pp. 751-758.

Breslow, N. E. (1972). Contribution to the discussion of paper by D. R. Cox, Regression Models and Life Tables. *Journal of the Royal Statistical Society, Series B 34*, pp. 216-217.

Breslow, N. E. (1974). Covariance analysis of censored survival data. *Biometrics, Vol. 30,* pp. 89-100.

Breslow, N. E. (1975). Analysis of survival data under a proportional hazards model. *International Statistical Review, Vol. 43*, pp. 45-57.

Brown, M. L., Goldie, S. J., Draisma, G., Harford, J. and Lipscomb, J. (2006). *Chapter 29. Health service interventions for cancer control in developing countries. Disease Control Priorities in Developing Countries.* 2nd ed. New York: Oxford University Press/World Bank.

Caan, B., Sternfeld, B., Gunderson, E., Coates, A., Quesenberry, C., Slattery, M.L.,(2005). Life after cancer epidemiology (LACE) study: a cohort of early stage breast cancer survivors (United States). *Cancer Causes and Control, Vol. 16*, pp. 545-556.

Calle, E.E., Rodriguez, C., Walker-Thurmond, K. and Thun, M.J. (2003). Overweight, obesity and mortality from cancer in a prospectively studied cohort of U.S. adults. *N Engl J Med. Vol. 348,* No. 17, pp. 1625-38.

Carey, L. A., Perou, C.M., Livasy, C.A., Dressler, L.G., Cowan, D., Conway, K., Karaca, G., Troester, M.A., Tse, C.K., Edmiston, S., Deming, S.L., Geradts, J., Cheang, M.C., Nielsen, T.O., Moorman, P.G., Earp, H.S., Millikan, R.C.,(2006). Race, Breast Cancer Subtypes, and Survival in the Carolina Breast Cancer Study. *The Journal of the American Medical Association, Vol. 295*, No. 21, pp. 2492-2502.

Carol, T., Karen, K. and M.D. (2005). The Encyclopedia of Breast Cancer, Facts on File, Inc, USA.

Chlebowski, R.T., Aiello, E., McTiernan, A. (2002). Weight loss in breast cancer patient management. *Journal of Clin Oncol*, *Vol. 20*, pp. 1128-43.

Chlebowski, R.T., Aiello, E., McTiernan, A. (2002).Weight loss in breast cancer patientmanagement. *Journal of Clinical Oncology, Vol. 20,* pp. 1128-1143.

Clark, T. G., Bradburn, M. J., Love, S. B., and Altman, D. G. (2003). Survival analysis parti: Basic concepts and _rst analyses. *British Journal of Cancer, Vol. 89,* No. 2, pp. 232-238.

Clark, T.G., Bradburn, M.J., Love, S.B., and Altman, D.G. (2003). Survival Analysis Part I: Basic concepts and first analyses. *British Journal of Cancer, Vol. 89,* No. 3, pp. 431–436.

Clayforth, C., Fritschi, L., McEvoy, S.P., Byrne, M.J., Ingram, D., Sterrett, G., Harvey, J.M., Joseph, D. and Jamrozik, K.(2007). Five-year survival from breast cancer in Western Australia over a decade. *The Breast. Vol. 16,* No.4, pp 375-381.

Clayton, D. G. (1991). A Monte Carlo method for Bayesian inference in frailty models. *Biometrics, Vol. 47,* pp. 467-485.

Cleves, M, A., Gould, W.W and Guitierrez, R. G. (2002). *An Introduction to Survival Analysis using Stata*, Stata Press, Texas : College Station.

Coleman, A.L., Cummings, S.R., Yu, F., Kodjebacheva, G., Ensrud, K.E., Gutierrez, P., Stone, K.L., Cauley, J.A., Pedula, K.L., Hochberg, M.C., Mangione, C.M. (2007). Study Group of Osteoporotic Fractures. Binocular visual-field loss increases the risk of future falls in older white women. *Journal of the American Geriatrics Society, Vol. 55,* No. 3, pp. 357–364.

Coleman, M.P., Quaresma, M., Berrino, F., Lutz, J.M., De Angelis, R., Capocaccia, R., Baili, P., Rachet. B., Gatta, G., Hakulinen, T., Micheli, A., Sant, M., Weir, H.K., Elwood, J.M., Tsukuma H, Koifman, S., e Silva,G.A., Francisci, S., Santaquilani, M., Verdecchia, A., Storm,H.H., Young, J.L. and the CONCORD Working Group. (2008).Cancer survival in five continents: a worldwide population-based study (CONCORD).*The Lancet.Vol.9,* pp.730-756. 13 March 2015 from www.thelancet.com/oncology

Collet, D. (1994). *Modelling Survival Data In Medical Research*. Chapman & Hall, London.

Compton, C.C., Byrd, D.R, Garcia-Aguilar, J., Kurtzman, S., Olawaiye, H.A. and Washington, M.K. (2012). *Cancer Survival Analysis*. 21 September 2013 from http://www.springer.com/978-1-4614-2079-8

Couch, F. J., DeShano, M. L., Blackwood, M. A., Calzone, K., Stopfer, J., Campeau, L., Ganguly, A., Rebbeck, T., Weber, B. L., Jablon, L., Cobleigh, M. A., Hoskins, K., and Garber, J. E. (1997), BRCA1 Mu- tations in Women Attending Clinics That Evaluate the Risk of Breast Cancer, *New England Journal of Medicine, Vol. 336*, pp. 1409-15.

Coups, E.J., Ostroff, J.S. (2005). A population-based estimate of the prevalence of behavioral risk factors among adult cancer survivors and noncancer controls. *Preventive Medicine, Vol. 40,* pp. 702-711.

Cox, D.R. (1972). Regression models with life-tables (with discussion). *Journal of Research Statistics Society Series B Stat Methodology, Vol. 34,* pp. 269–76.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B 34*, 187-220.

Cox, D. R. (1975). Partial likelihood. *Biometrika, Vol. 62*, pp. 269-276.

Crowder, M. (2012). *Multivariate Survival Analysis and Computing Risks.* New York: CRC Press.

Curado, M. P.,Edwards, B., Shin, H.R., Storm, H., Ferlay, J., Heanue, M. and Boyle, P. (2007). Cancer incidence in five continents*, Scientific Publications, Vol. IX*, No. 160. IARC, Lyon, France.

Curado, M., Ebrahimi, M., Vahdaninia, M., Montazeri, A. (2002). Risk factors for breast cancer in Iran: a case control study. *Breast Cancer Res, Vol. 4,* p. R10.

Danaei, G., Stephen, V. H., Alan, D. L., Christopher, J. L., Majid, E., and the Comparative Risk Assessment collaborating group (Cancers) (2005). Causes of cancer in the world: comparative risk assessment of nine behavioural and environmental risk factors. *Lancet, Vol. 366*, pp. 1784–1793.

Darendeliler, E. and Ağaoğlu, F.Y. (2003). Meme Kanserinin Epidemiyolojisi ve Etyolojisi. Eds: Topuz E, Aydıner A, Dincer M. *Meme Kanseri.İstanbul: Nobel Tıp Kitabevleri*, pp. 13-33.

Demark-Wahnefried, W., Rimer, B.K., Winer, E.P. (1997). Weight gain in women diagnosed with breast cancer. *Journal of Am Diet Assoc*, Vol. 97, pp. 519-26.

Denmark-Wahnefried, W., Aziz, N.M., Rowland, J.H., Pinto, B.M., (2005). Riding the crest of the teachable moment: promoting long-term health after the diagnosis of cancer. *Journal of Clinical Oncology , Vol. 23,* pp. 5814-5830.

Denmark-Wahnefried, W., Jones, L.W., (2008). Promoting a healthy lifestyle among cancer survivors. *Hematolpgy/Oncology Clinics of North America, Vol. 22,* pp. 319-342

Desreux, J., Gaspard, U., Bleret, V., Van Cauwenberge, J.R., Thille, A., Herman, P., Lifrange, E.(2011). Breast cancer in Belgium: why are we the first in Europe? *NCBI, Vol. 66,* No. 5-6, pp. 231-237.

Dey, S., Soliman, A.S., Hablas, A., Seifeldin, I.A., Ismail, K., Ramadan, M., El-Hamzawy, H., Wilson, M.L., Banerjee, M., Boffetta, P., Harford, J. and Merajver, S.D. (2010). Breast Cancer Res Treat: Urban–rural differences in breast cancer incidence by hormone receptor status across 6 years in Egypt. *Breast Cancer Res Treat, Vol. 120,* pp. 149–160.

Dey, D. K., Miiller, P. and Sinha, D. (eds) (1998). *Practical Nonparametric and Semiparametric Bayesian Statistics*. New York: Springer.

Dunnwald, L., Rossing, M. and Li, C. (2007).Hormone receptor status, tumor characteristics, and prognosis: a prospective cohort of breast cancer patients. *Breast Cancer Research, Vol, 9,* No.R6, pp. 1-10

Ebrahimi, M., Vahdaninia, M and Montazeri, A. (2002). Risk Factors for Breast Cancer in Iran: a case-control study. *Article of Brest Cancer Research , Vol. 4,* No. 5, pp. 1-4.

Edwin, S., Iversen, Jr., Giovanni, P., Donald, A. Berry and Joellen M.Schildkraut (2000). Genetic Susceptibility and Survival: Application to Breast Cancer. Journal of the *American*

Statistical Association, Vol. 95, No. 449, pp. 28-42, http://www.jstor.org/stable/2669520 .Accessed: 14/05/2013

Eide, G. E., Omenaas, E. and Gulsvik, A. (1996). The semiparametric proportional hazards model revisited: practical reparameterisations. *Statist. Med., Vol. 15*, pp. 1771-1777.

El Fatemi, H., Ihsane, S., Soufia, El Jayi, Kaoutar, M., My Abdelilah M., Nadia, S., and Afaf, A. (2013). Diagnosis Error: Carcinoma or Primary Breast Lymphoma? A Case Report and Literature Review. *Advances in Breast Cancer Research, ,Vol. 2***,** No. 1, pp 11-14 . Accessed from http://www.scirp.org/journal/abcr

Farooq, S. and Coleman, M.P. (2005). Breast cancer survival in south Asian women in England and Wales. *Journal of Epidemiol Community Health, Vol, 59*, No.5, pp. 402-406.

Freedman, R.J., Aziz, N., Albanes, D., Hartman, T., Danforth, D., Hill, S., Serbing, N., Reynolds, J.C. and Yanovski, J.A. (2004). Weight and bosy composition changes during and after adjuvant chemotherapy in women with breast cancer. *Journal of Clinical Endocrinol Metab, Vol. 89,* No.5, pp. 2248-2253.

Fritz, A., Percy, C., Jack, A., Shanmugaratnam, K., Sobin, L., Parkin, D. M., Whelan, S. (2000). *International Classification of Diseases for Oncology (ICD-O),* Geneva (Switzerland), World Health Organization.

Gail, M. H., Brinton, L. A., Byar, D. P., Corle, D. K., Green, S. B., Schairer, C., and Mulvhill, J. J. (1989). Projecting Individualized Probabilities of Developing Breast Cancer for White Females Who are Being Ex- amined Annually, *Journal of the National Cancer Institute, Vol. 81,* pp.1879- 1886.

Gamerman, D. (1991). Dynamic Bayesian models for survival data. *Appl. Statist., Vol. 40,* pp. 63-79; correction, *Vol. 41*, No. 1992, p. 285.

Gary, R. J. (1992). Flexible Methods for Analyzing Survival Data Using Splines, With Applications to BreastCancer Prognosis. *Journal of the American Statistical Association, Vol. 87,* No. 420, pp. 942-951. Accessed: 29/05/2013, http://www.jstor.org/stable/2290630.

GDH ; General Director of Health-Hawler (2015). Sited from http://dohhawler.org/Babat-2714 on 05/02/ 2015.

Gelfand, A. E. and Mallick, B. K. (1995). Bayesian analysis of proportional hazards models built from monotone functions. *Biometrics, Vol. 51*, PP. 843-852.

Gennari, A., Maria, P., Paolo, P., Matteo, P., Mariantonietta, C., Ulrich, P., Paolo, B.(2008). HER2 Status and Efficacy of Adjuvant Anthracyclines in Early Breast Cancer: A Pooled Analysis of Randomized Trials. *Journal of the National Cancer Institute, Vol. 100,* No. 1, pp. 14-20.

Ghavam-Nasiri, M.R., Anvari, K., Nowferesti, G.H., Silanian-Toosi, M. (2005). Locally advanced breast cancer: An experience in Mashhad, North-East of Iran, 1995-1999. *Arch Iranian Med. Vol. 8,* No.3. pp. 206-210.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (eds) (1996). *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.

Gilligan, M.A., Joan,N., Xu, Z., Rodney, S., Purushottam, W. and Ann, B. (2007). Relationship between Number of Breast Cancer Operations Performed and 5-Year Survival after Treatment for Early-Stage Breast Cancer. *American Journal of Public Health, Vol. 97,* No. 3, pp. 539-544.

Godwin, J. D. and Brown, C. C. (1975). Some prognostic factors in survival of patients with cancer of the colon and rectum. *Journal of Chronic Diseases , Vol. 28*, pp. 441-454.

Goodwin, J.S,, Zhang, D. D. and Ostir, G. V., Effect of Depression on Diagnosis, Treatment, and Survival of Older Women with Breast Cancer (2004). *Journal of American Geriatrics Society. Vol. 52*, No. 1, pp. 106–111.

Gore, S. M., Pocock, S. J. and Kerr, G. R. (1982). Long term survival analysis: the curability of breast cancer. *Statist. Med., Vol. 1,* pp. 93-104.

Grambsch, P. M. and Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika,Vol. 81*, pp.515-526.

Gray, R. J. (1992). Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *J. Am. Statist. Ass.,Vol. 87*, pp. 942-951.

Habib, O.S., Jawad, K. and Mohammed, K. (2007). Cancer Registration in Basra 2005: Preliminary Results. *Asian Pacific Jounal of Cancer Prevention, Vol. 8*, pp.187-190.

Hadi, N., Sadeghi-Hassanabadi, A., Talei, A.R., Arasteh, M.M. and Kazerooni, T. (2002). Assessment of a breast cancer screening programme in Shiraz, Islamic Republic of Iran. *East Mediterr Health Journal, Vol. 8*, No.2 and 3, pp. 386-392.

Hamdan, H. and Garibaldi, J.M. (2009). Modelling Survival Prediction in Medical Data. *Intelligent Modelling and Analysis (IMA) Research Group.* University of Nottingham:UK.

Hankey, B. F. and Myers, M. H. (1971). Evaluating differences in survival between two groups of patients. *Journal of Chronic Diseases, Vol. 24*, pp. 523-531.

Harirchi, I., Kolahdoozan, S., Karbakhsh, M., Chegini, N., Mohseni, S.M., Montazeri, A., Momtahen, A. J., Kashefi, A. and Ebrahimi, M. (2011). Twenty years of breast cancer in Iran: downstaging without a formal screening program, *Ann Oncol, Vol. 22*, pp. 93-97.

Hasnain-Wynia, R. and Baker, D. (2006). Obtaining Data on Patient Race, Ethnicity, and Primary Language in Health Care Organizations: Current Challenges and Proposed Solutions. *Health Services Research. Vol. 41(4 Pt 1),* pp.1501-1518.

Hastie, T. J. and Tibshirani, R. (1986). Generalized additive models (with discussion). *Statist. Sci.,Vol. 1*, pp. 297-318.

Health & Social Care Information Center (hscic). Reviewed 13 March (2015). Available at http://www.hscic.gov.uk/dars.

Heideman, W.H., Russell, N.S., Gundy, C., Rookus, M.A. and Voskuil, D.W. (2009). The frequency, magnitude and timing of post-diagnosis body weight gain in Dutch breast cancer survivors. *EUROPEAN JOURNAL OF CANCER, Vol. 45,* pp. 119-126.

Helgeson, V.S., Tomich, P.L. (2005). Surviving cancer: a comparison of 5-year diseasefree breast cancer survivors with healthy women. *Psychooncology, Vol. 14,* pp. 307-317.

Hemming, K. and Shaw, J.(2002). A Parametric Dynamic Survival Model Applied to Breast Cancer Survival Times, *Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 51,* No. 4, pp. 421-435. http://www.jstor.org/stable/3592619

Hider, P., Nicholas, B. (1999). The early detection and diagnosis of breast cancer: a literature review - an update. *NZHTA Report , Vol. 2,* No. 2. pp. 5-26.

Hoang, J.K., Vauka, J., Ludwig, B. and Glastonbury, Ch. M. (2013). Evaluation of Cervical Lymph Nodes in Head and Neck Cancer With CT and MRI: Tips, Traps, and a Systematic Approach, American Roentgen Ray Society. From www.ajronline.org on 28/10/2015.

Holleczek, B. and Brenner, H. (2012). Trends of population-based breast cancer survival in Germany and the US: Decreasing discrepancies, but persistent survival gap of elderly patients in Germany, *BioMed Central Cancer, Vol. 12,* No. 317, pp. 1-11.

Holmes, M.D., Chen, W.Y., Feskanich, D., Kroenke, C.H., Colditz, G.A.,(2005). Physical activity and survival after breast cancer diagnosis. *JAMA, Vol. 293,* pp. 2479-2486.

Horov´a, I., Posp´ıˇsil, Z. and Zelinka, J.,(2007). Semiparametric Estimation of Hazard Function for Cancer Patients. *The Indian Journal of Statistics, Vol. 69,* Part 3, pp. 494-513.

Hosmer, D. W., Lemeshow, S. and May, S. (2008). *Applied Survival Analysis: Regression Modeling of Time-to-Event Data,* 2nd Edition. John Wiley & Sons, Inc.

Hosmer, D.W. and Lemeshow, S. (1999). *Applied Survival Analysis; Regression Modeling of Time to Event Data*. New York: John Wiley & Sons, Inc.

Hsairi, M., Fakhfakh, R., Ben Abdallah, M., Jlidi, R., Sellami, A., Zheni, S., Hmissa, S., Achour, N. , Nacef, T., (2002).Estimation a l'echelle national de l'incidence des cancers en Tunisie 1993-1997, *Journal of Tunisie Medical Vol. 80, No. 2, pp. 57-64.*

HSCIC; Health & Social Care Information Centre (2013). Standards for the clinical structure and content of patient records. *Royal College of Physicians.* Reviewed 16 April 2015. Available at https://www.rcplondon.ac.uk/sites/default/files/standards-for-the-clinical-structure-and-content-of-patient-records.pdf

HSCIC; Health & Social Care Information Center (2015). The national provider of information, data and IT systems for health and social care, Omnibus data collections. Sited on March 25, 2015 from http://www.hscic.gov.uk/dars

Huo, D., Francis, I., Andrey, K., Jean-Marie, D., Rita, N., James, D., Bifeng, Z., Tatyana, G., Chunling, Z., Olayiwola, O.,David, M., Sani, M., Abayomi, O., Adewumi, O., Festus, I., Adeyinka, F., Charles, M., and Olufunmilayo, I. (2009). Population Differences in Breast Cancer: Survey in Indigenous African Women Reveals Over-Representation of Triple-

Negative Breast Cancer. *Journal of Clinical Oncology by American Society of Clinical Oncology, Vol. 27,* No. 27, PP 4515–4521. From http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2754904/pdf/zlj4515.pdf

Hussaion, A. H. and Aziz, P. M. (2009). The Incidence Rate of Breast Cancer in Suleimani Governorate in 2006: Preliminary Study. *Journal of Zankoy Suleimani, Vol. 12,* No. 1, Part A, pp. 59-65.

IARC (2008). *World cancer report 2008*. Lyon, International Agency for Research on Cancer.

Ibrahim, J. G., Chen, M. H. and Sinha, D. (2001). Bayesian Survival Analysis. New York: Springer.

İğci A, Asoğlu O (2003). Meme Kanserinin Erken Tanısında Tarama Yontemleri. Eds: Topuz E, Aydıner A, Dincer M. Meme Kanseri. *İstanbul: Nobel Tıp Kitabevleri*, pp. 113-23.

Ihemelandu, C., Leffall, L., Dewitty, R., Naab, T., Mezghebe, H., Makambi, K., Adams-Campbell, L., Frederick, W. (2007). Molecular breast cancer subtypes in premenopausal and postmenopausal African-American women: age-specific prevalence and survival. *Journal of Surgical Research, Vol. 143,* No. 1, pp. 109-118.

Ingram, C. and Brown, J.K. (2004). Patterns of weight and body composition change in premenopausal women with early stage breast cancer. *Cancer Nursing, Vol. 27*, No. 6, pp. 483-490.

International Union Against Cancer, 2010. Available at: http://www.uicc.org/index. php?option¼com_content&task¼view&id¼13&Itemid¼113 (accessed 02. 06. 13).

Irwin, M.L., McTiernan, A., Baumgartner, R.N., Baumgartner, K.B., Bernstein, L., Gilliland, F.D. and Ballard-Barbash, *R.* (2005). Changes in body fat and weight after a breast cancer diagnosis: influence of demographic, prognostic, and lifestyle factors. *Journal of Clinical Oncology, Vol. 23,* No. 4, pp. 774-782.

Jemal, A., Freddie, B., Melissa, M. Jacques, F., Elizabeth, W., and David, F. (2011). Global Cancer Statistics. *American Cancer Society, Vol. 61,* No. 2, pp. 69-90.

Jeong, J. (2006). A New Parametric Family for Modelling Cumulative Incidence Functions: Application to Breast Cancer Data. *Journal of the Royal Statistical Society. Series A (Statistics in Society), Vol. 169,* No. 2, pp. 289-303. http://www.jstor.org/stable/3559674

Johannsson, O., Ranstam, J., Borg, A., and Olsson, H. (1998). Survival of BRCA1 Breast and Ovarian Cancer Patients: A Population-Based Study from Southern Sweden, *Lancet, Vol. 351,* pp.304-305.

John P. K. and Melvin L. M. (1997). *Survival Analysis*, Springer.

Kalbfleisch, J. D. (1978). Non-parametric Bayesian analysis of survival time data. *J R. Statist.Soc. B, Vol. 40,* pp. 214-221.

Kalbfleisch, J. D. and Prentice, R. L.(1973). Marginal likelihoods based on Cox's regression and life model. *Biometrika , Vol. 60,* pp. 267-278.

Kamangar, F., Cheng, C., Abnet, C.C., Rabkin, C.S.(2006). Interleukin-1B polymorphisms and gastric cancer risk--a meta-analysis. *Cancer Epidemiology Biomarkers Prevention, Vol. 15,* No. 10, pp. 1920-1928.

Kaplan, E.L., Meier, P. (1958). Non-parametric estimation from incomplete observations. *Journal of American Statistics Association. Vol. 53*, pp. 457–81.

Kleinbaum, D.G and Klein, M. (2005). *Survival Analysis, a self-Learning Text.* USA: Springer Science + Business Media, Inc.

Koifman, S. E., Silva, G.A., Francisci, S., Santaquilani, M., Verdecchia, A., Storm, H.H. and Young, J.L. CONCORD Working Group. (2008). Cancer survival in five continents: a worldwide population-based study (CONCORD). *Lancet Oncol, Vol. 9,* No. 8, pp. 730–756.

Konecny, G., Giovanni, P., Mark, P., Michael, U., Sugandha, D., Zuleima, A., Cindy, W., Hong-Mei, R., Ingo, B., Margret, F., He-Jing, W., Malgorzata, B., Ram, S., Herrmann, H.,Dennis, J. (2003). Quantitative Association Between HER-2/neu and Steroid Hormone Receptors in Hormone Receptor-Positive Primary Breast Cancer. *Journal of the National Cancer Institute, Vol. 95*, No. 2, pp. 142-153.

Kozusko, F. and Bajzer, Z. (2003). Combining Gompertzian growth and cell population dynamics. *Mathematical Biosciences, Vol. 185,* pp. 153–167.

Lacey, J., Aimee, R.**,** Saundra, S.**,** Pamela, M.**,** Shih-Chen, C.**,** Michael, F.**,** Robert, N.**,** Philip, C.**,** Christine, D. (2009). Breast cancer epidemiology according to recognized breast cancer risk factors in the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial Cohort. *BMC Cancer, Vol. 9,* p. 84.

Lambert, P.C. and Royston, P. (2009). Further development of flexible parametric models for survival analysis. *The Stata Journal, Vol. 9,* No. 2, pp. 265–290.

Lan, N., Laohasiriwong, W. and Stewart, J. (2013). Survival probability and prognostic factors for breast cancer patients in Vietnam. *Glob Health Action, Vol. 6,* p. 18860.

Lavigne, E., Eric, J., Sai, Y., Paul,J., Kenneth, C., Dean, A., Howard, M., and Jacques, B.(2013). Breast cancer detection and survival among women with cosmetic breast implants: systematic review and meta-analysis of observational studies. *British Medical Journal , pp. 1-12.*Available at http://www.bmj.com/content/346/bmj.f2399.

Lawless, J.F. (2003). *Statistical Models and Methods for Lifetime Data,* 2nd Ed. New Jersey: John Wiley & Sons, INC.

Lee, J., Wacholder, S., Struewing, J., McAdams, M., Pee, D., Brody, L., Tucker, M. and Hartge, P. (1999), Survival After Breast Cancer in Ashkenazi Jewish BRCA1 and BRCA2 Mutation Carriers, *Journal of the National Cancer Institute, Vol. 19,* pp. 259-263.

Leonard, T. (1978). Density estimation stochastic processes and prior information. *J. R. Statist. Soc. B, Vol. 40,* pp. 113- 146.

Leung, K.M., Elashoff, R.M., Afifi, A. A. (1997). Censoring issues in survival analysis. *Annual Review of Public Health. Vol. 18,* pp. 83–104.

Little, M.P. and Boice J.D. Jr. (1999). Comparison of breast cancer incidence in the Massachusetts tuberculosis fluoroscopy cohort and in the Japanese atomic bomb survivors. *Radiation Research, Vol.151*, No. 2, pp. 218-224.

Machin, D., Cheung, Y. B., Parmar, M. K. 2nd ed. (2006). *Survival analysis a practical approach.* West Sussex: John Wiley and Sons Ltd.

Majid, R., Mohammed, H., Hassan, H., Abdulmahdi, W., Rashid, R. and Hughson, M. (2012). A population-based study of Kurdish breast cancer in northern Iraq: Hormone receptor and HER2 status. A comparison with Arabic women and United States SEER data. *BMC Women's Health, Vol. 12,* No. 16, pp. 1-10. Accessed from http://www.biomedcentral.com/1472-6874/12/16

Majid, R., Mohammed, H., Saeed, H., Safar, B., Rashid, R., Hughson, M. (2009). Breast cancer in kurdish women of northern Iraq: incidence, clinical stage, and case control analysis of parity and family risk. *BMC Women's Health, Vol. 9,* No. 33.pp. 1-6.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports , Vol. 50,* pp. 163-170.

Masayoshi, T., Charles, E., Tsutomu, Y., Masahide, A., Shoji, T., Haruo, E. and Issei, N., (1987). Incidence of Female Breast Cancer among Atomic Bomb Survivors, Hiroshima and Nagasaki, 1950-1980. *Radiation Research, Vol. 112,* No.2, pp. 243-272.

Mathers, C., Lopez, A., Murray, C. (2006). *The Burden of Disease and Mortality by Condition: Data, Methods, and Results for 2001.* New York: Oxford University Press.

McCredie, M., Coates, M., Grulich, A. (1994). Cancer incidence in migrants to New South Wales (Australia) from the Middle-East. 1972- 91. *Cancer causes and control, Vol. 5,* pp. 414-421.

Meyerhardt, J.A., Giovannucci, E.L., Holmes, M.D., Chan, A.T., Chan, J.A., Colditz, G.A. and Fuchs, C.S. (2006). Physical activity and survival after colorectal cancer diagnosis. *Journal of Clinical Oncology, Vol. 24,* No.22, pp. 3527-3534.

Miecznikowski**,** J., Dan, W**.,** Song, L**.,** Lara, S., and David, G. (2010). Comparative survival analysis of breast cancer microarray studies identifies important prognostic genetic pathways. *BioMed Central Cancer, Vol. 10,* No. 573 , pp. 1-7.

Miller, L.D., Johanna, S., Joshy, G., Vinsensius, B., Liza, V., Alexander, P., Yudi, P., Per Hall, S., Edison, T. and Jonas, B. (2005). An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proceedings of the National Academy of Sciences of the United State of America, Vol. 102,*No. 38, pp. 13550-13555.

Miller, R. (1980). *Survival analysis.* California: Stanford University Press.

Montazeri, A., Ebrahimi, M., Mehrdad, N., Ansari, M. and Sajadian, A. (2003). Delayed presentation in breast cancer: a study in Iranian women. *BMC Women's Health, Vol. 3,* p. 4.

Montazeri, A., Mandana, E., Neda, M., Mariam, A. and Akram S. (2003). Delayed presentation in breast cancer: a study in Iranian women. *BioMed Central Cancer. Vol, 3,* No. 4, pp. 1-6. from http://www.biomedcentral.com/1472-6874/3/4

Mostert, P.J., Bekker, A. and Roux, J.J. (1998). Bayesian analysis of survival data using the Rayleigh model and Linex Loss, *South Africa Statist. J., Vol. 32,* No.1, pp. 19-42.

Mousavi, S., Mohagheghi, M., Mousavi-Jerrahi, A., Nahvijou, A., Seddighi, Z. (2006). Burden of breast cancer in Iran: a study of the Tehran population based cancer registry. *Asian Pac J Cancer Prev. Vol. 7* No. 4, pp. 571-574.

National Cancer Institute. Genetics of Breast and Ovarian Cancer (2012). Accessed at http://www.cancer.gov/cancertopics/pdq/genetics/breast-and-ovarian/HealthProfessional on 2/14/2012.

National Cancer Institute Fact Sheets (2012). Antiperspirants/Deodorants and Breast Cancer. Accessed at http://www.cancer.gov/cancertopics/factsheet/Risk/AP-Deo on February 16, 2012.

National Comprehensive Cancer Network (NCCN). Practice Guidelines in Oncology: Breast Cancer. Version 1.2012. Accessed at www.nccn.org on April 17, 2012.
National Comprehensive Cancer Network (NCCN). Practice Guidelines in Oncology: Genetic/Familial High-Risk Assessment: Breast and Ovarian. Version 1.2012. Accessed at www.nccn.org on 8/23/12

National Cancer Institute, (2012). *Surveillance, Epidemiology and End Results (SEER) Program*. Accessed from http://seer.cancer.gov/statfacts/html/breast.html#incidence-mortality.

National Cancer Center, (2009). Cancer Facts & Figures 2009 in the Republic of Korea. The Ministry for Health, Welfare and Family Affairs, Republic of Korea.

National Cancer Information Center, 2009. Cancer Prevalence (2007). The Ministry of Health & Welfare, Republic of Korea. Available at http://www.cancer.go.kr/cms/ statics/can_in_lif/index.html.

Nerenz, D. R, Gunter M, Garcia M, Green-Weir R, Wisdom K, Joseph C.(2002). *Developing a Health Plan Report Card on Quality of Care for Minority Populations.* The Commonwealth Fund.

Newby, M. (2010). *Bayesian Statistics. United Kingdom*. pp. 44-50.

NHS Choices (2011). Unhealthy lifestyles linked to UK cancer rates. *Accessed from* http://www.nhs.uk/news/2011/01January/Pages/unhealthy-lifestyles-linked-to-UK-cancer-rates.aspx .

NHS, National Health Service (2012) , Accessed at http://www.nhs.uk/Conditions/Cancer-of-the-breast-female/Pages/Introduction.aspx on December 2, 2013.

Nichols, D. (2010). Breast Cancer in Iraq Leads to Gulf War Veteran News Alert and Rep Boswell Legislation *American Association for Cancer Research*. Accessed from http://www.veteranstoday.com/2010/03/18/breast-cancer-in-iraq-leads-to-gulf-war-veteran

Non-Invasive or Invasive Breast Cancer (2012). BreastCancer.ORG. Accessed from http://www.breastcancer.org/symptoms/diagnosis/invasive

Ogle, K.S., Swanson, G.M., Woods, N., Azzouz, F.,(2000). Cancer and comorbidity: redefining chronic diseases. *Cancer, Vol. 88,* pp. 653-663.

Olsen, M.H., Bidstrup, P.E., Frederiksen, K., Rod, N.H., Grønbaek, M., Dalton, S.O. and Johansen, C.(2012). Loss of partner and breast cancer prognosis - a population-based study, Denmark, 1994-2010.*British Journal of Cancer, Vol. 106,* No. 9, pp. 1560-1563.

Omar, Z., Zainudin, M. and Nor, I. (2006). Malaysian Cancer Statistics-Data and Figure Peninsular Malaysia 2006. *National Cancer Registry*. Accessed at http://www.makna.org.my/PDF/MalaysiaCancerStatistics.pdf

Ora, P., Yehiel, F., Efrat, T., Micha, B., Xiaonan, X. and Susan, H. (2004). Cancer after pre-eclampsia: follow up of the Jerusalem perinatal. *BMJ,* pp. 328-919. Accessed from http://dx.doi.org/10.1136/bmj.38032.820451.7C

Othman, R.T., Abdulljabar, R., Saeed, A., Kittani, S.S., Sulaiman, H.M., Mohammed, S.A., Rashid, R.M. and Hussein, N.R.(2011). Cancer incidence rates in the Kurdistan region/Iraq from 2007-2009. *Asian pacific Journal of Cancer Prevention, Vol. 12,* No.5, pp. 1261-1264.

Özkan, A., Arzu T.M. Aysel, G. and Ayse, San. T. (2010). Do Turkish Nursing and Midwifery Students Teach Breast Self-Examination to Their Relatives? *Asian Pacific Journal of Cancer Prevention, Vol.* 11, pp. 1569-1573.

Özmen, V. (2006). Screening and registering programs for breast cancer in Turkey and in the world. *Journal of Breast Health, Vol, 2,* No. 2, pp. 55-58.

Özmen, V. (2008). Breast cancer in the world and Turkey. *Meme Sağlığı Dergisi, Vol. 4,* pp. 7-12.

Parkin, M., Freddie, B., Ferlay, J. and Paola, P. (2005). Global cancer statistics, 2002. *A Cancer Journal for Clinicians, Vol. 55,* No. 2, pp. 74-108.

Parmar, M. K., Torri, V. and Stewart, L. (1998). Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. *Statistics In Medicine . Vol. 17*, pp. 2815-2834.

Patients. Version VIII. From www.cancer.org on 28/10/2015.

Patnaik, J.L., Byers, T., DiGuiseppi, C., Denberg, T.D. and Dabelea, D. (2011). The influence of comorbidities on overall survival among older women diagnosed with breast cancer. *Journal of National Cancer Institute, Vol. 103,* pp. 1-11.

Patterson, R.E., Neuhouser, M., Hedderson, M.M., Schwartz, S.M., Standish, L.J. and Bowen, D.J. (2003). Changes in diet, physical activity, and supplement use among adults diagnosed with cancer. *Journal of the American Dietetic Association, Vol. 103,* pp. 323-328.

Perot, R T, and Youdelman, M. R. (2001). *Ethnic, and Primary Language Data Collection in the Health Care System: An Assessment of Federal Policies and Practices.* The Commonwealth Fund.

Peto, J. (2001). Cancer epidemiology in the last century and the next decade. *Nature, Vol. 411*, pp. 390–395.

Pictures of Breast Anatomy (2012). BreastCancer.ORG. Accessed from http://www.breastcancer.org/pictures/breast_anatomy

Pierce, J.P., Stefanick, M.L., Flatt, S.W., Natarajan, L., Sternfeld, B., Madlensky, L., Al-Delaimy,W.K.,Thomson, C.A., Kealey,S., Hajek, R.,Parker,B.A., Newman, V.A., Caan, B. and Rock, C.L. (2007). Greater survival after breast cancer in physically active women with high vegetable-fruit intake regardless of obesity. *Journal of Clinical Oncology, Vol. 25*, No. 17, pp. 2345-2351.

Pinto, B.M., Eakin, E., Maruyama, N.C., (2000). Health behavior changes after a cancer diagnosis: what do we know and where do we go from here. *Annals of Behavioral Medicine, Vol. 22*, pp. 38-52.

Porta, M., Manuel, G., Nuria, M. and Planas, J. (1991). Influence of "diagnostic delay" upon cancer survival: an analysis of five tumour sites. Journal of Epidemiol Community Health Vol. *45*, No. 3, pp. 225–230.

Porter, P. (2008). Westernizing women's risk? Breast cancer in lower-income countries. *New England Journal of Medicine, Vol. 358,* No. 3, pp. 213–16.

Prentice, R. L. and Gloeckler, L. A. (1978). Regression Analysis of Grouped Survival Data with Application to Breast Cancer Data Biometrics, Vol. 34, No. 1 (Mar., 1978), pp. 57-67, http://www.jstor.org/stable/2529588 .Accessed: 29/05/2013 10:13

Prinja, Sh., Gupta,N. and Verma, R. (2010). Censoring in Clinical Trials: Review of Survival Analysis Techniques. Indian Journal Community Medicine. Vol, 35, No.2, pp. 217–221.

Pritchard, K.I., Shepherd, L.E., O'Malley, F.P., Andrulis, I.L., Tu, D., Bramwell, V.H., Levine, M.N., (2006). HER2 and responsiveness of breast cancer to adjuvant chemotherapy. *The New England Journal of Medicine, Vol. 354,* No. 20, pp. 2103-2111.

Quantin, C. Thierry, M., Bernard, A., Jean, M. and Joseph, L.(1996). A Regression Survival Model for Testing the Proportional Hazards Hypothesis, *Biometrics, Vol. 52*, No. 3, pp. 874-885. Accessed on 14/05/2013 from http://www.jstor.org/stable/2533049.

Raqab, M.M., Ahsanullah, M. (2001). Estimation of location and scale parameters of generalized exponential distribution based on order statistic. *J. Statist. Comput. Simul. Vol.* 22, pp. 112-127.

Raza, M.S. and Broom, M. (2016). Survival Analysis Modeling With Hidden Censoring. *Journal of Statistical Theory and Practice, Vol. 10,* No. 2, pp. 375-388.

Rebecca, L., Siegel, Kimberly, D. Milllller and Ahmedin, J. (2016), Cancer Statistics,2016, CA: *A Cancer Journal for Clinicans. Vol*. 66, No.1, pp. 7-30.

Reed, T., Wagener, D. K., Donahue, R. P., and Kuller, L. H. (1986), Fam- ily History of Cancer Related to Cholesterol Level in Young Adults, *Genetic Epidemiology,Vol. 3,* pp. 63-71.

Renard, F., Vankrunkelsven, P., Van Eycken, L., Henau, K., Boniol, M. and Autier, P. (2010). Decline in breast cancer incidence in the Flemish region of Belgium after a decline in hormonal replacement therapy. *NCBI, Vol. 21,* No. 12, pp. 2356-2360.

Rennert, G. (2006). *Breast cancer*. In Cancer Incidence in the Four Member Countries (Cyprus, Egypt, Israel, and Jordan) of the Middle-East Cancer Consortium (MECC) compared with US SEER, Chapter 8. Edited by Friedman LS, Edwards BK, Reiss LAG, Young JL. Bethesda, MD: National Cancer Institute. NIH Pub No. 06–5873, pp.73–81.

Rezaianzadeh, A., Peacock, J., Reidpath, D., Talei, A., Hosseini, S.V. and Mehrabani, D. (2009). Survival analysis of 1148 women diagnosed with breast cancer in Southern Iran. *BMC Cancer, Vol. 9*, p. 168.

Ridder, G. (1990). The Non-Parametric Identification of Generalized Accelarated Failure-Time Models. *The Review of Economic Studies, Vol. 57,* No. 2, pp. 167-181.

Ries, L., Melbert, D., Krapcho, M., Mariotto, A., Miller, B., Feuer, E., Clegg, L., Horner, M., Howlader, N., Eisner, M., Reichman, M. and Edwards, B. (eds) (2007). SEER Cancer Statistics Review, 1975-2004, National Cancer Institute. Bethesda, MD, http://seer.cancer.gov/csr/1975_2004/, based on November 2006 SEER data submission, posted to the SEER web site, 2007.

Ries, L., Melbert, D., Krapcho, M., Stinchcomb, D., Howlader, N., Horner, M., Mariotto, A., Miller, B., Feuer, E., Altekruse, S., Lewis, D., Clegg, L., Eisner, M., Reichman, M. and Edwards, B., (eds) (2007). National Cancer Institute. Bethesda, MD: *SEER Cancer Statistics Review, 1975-2005* 2008 http://seer.cancer.gov/csr/1975_2005/.

Robb, C., Haley, W., Balducci, L., Extermann, M., Perkins, E., Small, B. and Mortimer, J. (2007). Impact of breast cancer survivorship on quality of life in older women. *Crit Rev Oncol Hematol , Vol. 62,* No. 1, pp. 84-91.

Robson, M., Chappuis, P., Satagopan, J., Wong, N., Boyd, J., Goffin, J.,Hudis, C., Roberge, D., Norton, L., Begin, L., offit,. K. and Foulkes, W. (2004). A combined analysis of outcome following breast cancer: differences in survival based on BRCA1/BRCA2 mutation status and administration of adjuvant treatment. *Breast Cancer Res. ,Vol. 6,* No. 1, pp. R8-R17 , Accessed from http://www.ncbi.nlm.nih.gov/pubmed/14680495

Rock, C. (2003). Diet and breast cancer: Can dietary factors influence survival? *J Mammary Gland Biol Neoplasia, Vol. 8,* pp. 119-32.

Rock, C., Flatt, S., Newman, V. , Caan, B.J., Haan, M.N., Stefanick, M.L., Faerber, S. and Pierce, J.P. (1999). Factors associated with weight gain in women after diagnosis of breast cancer. *J Am Diet Assoc, Vol. 99,* No. 10, pp. 1212-1221.

Rosmawati, N.H., (2010). Knowledge, attitudes and practice of breast self-examination among women in a suburban area in Terengganu, Malaysia. *Asian Pacific Journal Cancer Prev. Vol. 11,* No. 6, pp. 1503-1508.

Royal College of General Practitioners (2011). Magazine of It's Your Practice. Viewed on March 2015 from www.rcgp.org.uk

Rubin, S. C., Benjamin, I., Behbakht, K., Takahashi, H., Morgan, M. A., LiVolsi, V. A., Berchuck, A., Muto, M. G., Garger, J. E., Weber, B. L., Lynch, H. T., and Boyd, J. (1996). Clinical and Pathological Features of Ovarian Cancer in Women with Germline Mutations of BRCA1, *New England Journal of Medicine, Vol. 335,* pp. 1413-1416.

Rudat,V., Nuha, B., Saleh, T. and Mousa. A. (2013). Body Mass Index and Breast Cancer Risk: A Retrospective Multi-Institutional Analysis in Saudi Arabia. *Advances in Breast Cancer Research, Vol. 2, pp. 7-10. Accessed from* http://www.scirp.org/journal/abc

Saleem, M. and M. Aslam (2008b). Bayesian Analysis of the two component mixture of the Rayleigh Distribution with uniform and Jeffreys priors, Journal of *Applied Stat. Sci., Vol. 16,* No.4. pp.105-113.

Saleem, M. and M. Aslam, (2008a). On prior selection for the mixture of Rayleigh distribution using predictive Intervals, *Pakistan Journal of Stat. Vol. 24,* No. 1. pp. 21-35.

Saleem, M. and Raza, A. (2011). On Bayesian Analysis of the Exponential Survival Time Assuming the Exponential Censor Time. *Pakistan Journal of Science, Vol. 63,* No. 1. pp. 44-48.

Salih, K. (1995). Anfal: The Kurdish Genocide in Iraq. *Digest of Middle East Studies, Vol. 4,* pp. 24-39.

Sargent,D . J. (1997). A flexible approach to time-varying coefficients in the Cox regression setting. *Lifetime Data Anal., Vol. 3,* pp. 13-25.

Satia, J.A., Campbell, M.K., Galanko, J.A., James, A., Carr, C. and Sandler, R.W. (2004). Longitudinal changes in lifestyle behaviors and health status in colon cancer survivors. Cancer Epidemiology, *Biomarkers & Prevention, Vol. 13,* pp. 1022-1031.

Shabila, N., Namir G. Al-Tawil, Tariq, S. Al-Hadithi, Egbert, S. and Kelsey, V.(2012). Iraqi primary care system in Kurdistan region: providers' perspectives on problems and opportunities for improvement. *BioMed Central Women's Health, Vol. 12 , No.* 21, pp. 1-9. Accessed from http://www.biomedcentral.com/1472-698X/12/21

Sheila, M., Stuart, J., and Gillian, R. (1984). Regression Models and Non-Proportional Hazards in the Analysis of Breast Cancer Survival. *Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 33*, No. 2, pp.176-195 http://www.jstor.org/stable/2347444 Accessed: 07/05/2013 09:19.

Siddiqui, T., Sabih, M. and Khan, S. and Salam, A. (2001). A Survival Analysis of Metastatic Breast Cancer in Pakistani Patients, *Journal of Pakistan Medical Association, Vol. 51,* pp. 120.

Simsek, F. (2000). Five Year Survival Analysis of Patients with Clinical Stages I and IIA Breast Cancer who Received Initial Treatment at North Carolina Hospitals. North Carolina Public Health , *CHIS study*. No. 123, pp. 1-9.

Sinha, D., Chen, M.-H. and Ghosh, S. K. (1999) Bayesian analysis and model selection for interval censored survivald ata. *Biometrics,Vol. 5* , No.5, pp. 585-590.

Sinha, D., Ibrahim, J. G. and Chen , M. (2002). Models for Survival Data from Cancer Prevention Studies. *Journal of the Royal Statistical Society. Series B (Statistical*

Methodology), Vol. 64, No.3 , pp. 467-477 Published by: Wiley for the Royal Statistical Society Stable URL: http://www.jstor.org/stable/3088783

Slattery, M. L., and Kerber, R. A., (1993). A Comprehensive Evaluation of Family History and Breast Cancer Risk, *Journal of the American Medical Association, Vol. 270,* No. 13, pp. 1563-8.

Spitale, A., Mazzola, P., Soldini, D., Mazzucchelli, L. and Bordoni, A. (2009). Breast cancer classification according to immunohistochemical markers: clinicopathologic features and short-term survival analysis in a population-based study from the South of Switzerland. *US National Library of Medicine National Institutes of Health Search database, Vol. 20,* No. 4, pp. 628-635 from http://www.ncbi.nlm.nih.gov/pubmed/19074747

Stewart, B. and Kleihues, P.(2003). *World Cancer Report.* IARCPress. Lyon.

Stewart, B. and Kleihues, P. (2003). *The Global Burden of Cancer. World Cancer Report*. 1st ed. Lyon: IARC Press. pp. 12-7.

Strati, A., Sabine, K., Athina, M., Cleo, P**.** and Evi, L.(2013). Comparison of three molecular assays for the detection and molecular characterization of circulating tumor cells in breast cancer. *Breast Cancer Research , Vol.* **15**, No. R20, pp. 1-7.

Sughayer, M. A., Maha M. A., Suleiman M. and Mahmoud, A. (2006). Prevalence of Hormone Receptors and HER2/neu in Breast Cancer Cases in Jordan. *Pathology Oncology Research, Vol. 12,* No 2. , pp. 83-86. Article is available online at http://www.webio.hu/por/2006/12/2/0083

Tabatabai, M.A, Wayne M. Eby, Nadim Nimeh and Karan P. Singh (2012).Role of Metastasis in Hypertabastic Survival Analysis of Breast Cancer: Interaction with Clinical and Gene Expression Variables. *Libertas Academica Ltd, Vol,* 5, pp, 1-17.

Tableman, M. and Kim, J. s. (2004). *Survival Analysis Using S*. New York: Chapman & Hall/CRC.

Taylor, R., Davis, P., Boyages, J. (2003). Long-term survival of women with breast cancer in New South Wales. *Europ J Can. Vol. 39,* No. 2, pp 215–222.

Ueno, M., Kiba, T., Nishimura, T., Kitano, T., Yanagihara, K., Yoshikawa, K., Ishiguro, H., Teramukai, S., Fukushima, M., Kato, H. and Inamoto, T. (2007). Changes in survival during the past two decades for breast cancer at the Kyoto University Hospital. *EJSO. Vol. 33,* No. 6. pp. 696-699.

UNDP (2005). Iraq Living Conditions Survey 2004. Baghdad, Iraq, Central Organization for Statistics and Information Technology, Ministry of Planning and Development Cooperation.

Vahdaninia, M. and Montazeri, A. (2004). Breast cancer in Iran: a survival analysis. *Asian Pac J Cancer Prev.Vol. 5,* No. 2. pp. 223- 225.

Verweij, P. J. M. and van Houwelingen, H . C. (1995). Time-dependent effects of fixed covariates in Cox regression. *Biometrics, Vol. 51*, pp. 1550-1556.

Wang, J. and Li, Y. (2005). Estimators for survival function when censoring times are known, *Communications in Statistics-Theory and Methods*, *Vol. 34,* pp. 449-459.

Watson, M., Haviland, J.S., Greer, S., Davidson, J. and Bliss, J.M.(1999). Influence of psychological response on survival in breast cancer: a population-based cohort study. *Journal of NCBI. Vol. 354,* No. 9187, pp. 1331-1336.

Watson, P., Marcus, J., and Lynch, H. (1998). Prognosis of BRCA1 Hered- itary Breast Cancer, *Lancet*, *Vol. 351*, pp. 304-305.

WHO (2004*). The world health report 2004 - changing history*.

WHO (2007). *Cancer control: knowledge into action: WHO guide for effective programmes: early detection*.

WHO (2008). *The global burden of disease: 2004 update*. From www.who.int\evidence\bod on 08/05/2015.

Wikipedia (Breast Cancer) http://ar.wikipedia.org in 10-09-2011.

Yange, Q., Khoury, M., Rodriquez, C., Calle, E., Tathan, L., and Flanders, W. (1998). Family history score as a predictor of breast cancer mortality: prospective data from the cancer prevention study II, United States, 1982-1991. *Am J Epidemiol, Vol. 47,* pp. 652-659.

Yaw, Y. H., Mirnalini, K., Zalilah M. S., Chan, Y. M., Zailina, H., Rokiah, M.Y., Zabedah, O., Nurfaizah, S. and Yong, H. W. (2010). Pattern of Weight Changes in Women with Breast Cancer. *Asian Pacific J Cancer Prev, Vol.* 11, pp. 1535-1540.

Yi, M. and Kim, J. (2013). Factors influencing health-promoting behaviors in Korean breast cancer survivors. *European Journal of Oncology Nursing, Vol. 17,* No. 2, pp. 138-145.

Yip, C.H., Robert, A., Benjamin, O., Anthony, B., David, B., Eng-Suan, A., Rosemary, S., Marilys, C., Gary, L., and Anne, M. (2008). Guideline implementation for breast healthcare in low- and middle-income countries: early detection resource allocation. *Cancer*, *Vol. 113*, pp. 2244–2256.

Zhao, J., Liu, H., Wang, M., Gu, L., Guo, X., Gu, F., Fu, L.(2009). Characteristics and prognosis for molecular breast cancer subtypes in Chinese women. *Journal of surgical Oncology, Vol, 100,* No. 2, pp. 89-94.

Ziaei, J., Zohreh, S., Iraj, A., Saeed, D., Ali, P., and Jalil, V. (2013). Survival Analysis of Breast Cancer Patients in NorthwestIran. *Asian Pacific Journal of Cancer Preentionv,Vol.* 14, No. 1, pp. 39-42.