A MODIFIED GENERALISED-GAMMA MIXTURE CURE MODEL FOR SURVIVAL DATA ANALYSIS

BY

Serifat Adedamola FOLORUNSO

B.Sc. Statistics (Abeokuta), M.Sc. Statistics (Ibadan)

(Matric Number: 159321)

A Thesis in the Department of Statistics

Submitted to the Faculty of Sciences

in Partial fulfillment of the requirement for the Degree of

DOCTOR OF PHILOSOPHY

of the

UNIVERSITY OF IBADAN

MAY, 2019

ABSTRACT

Cure models are special survival models developed to estimate cure rate in cancer research. Several cure models such as Lognormal-Mixture Cure-Models (LNMCM), Loglogistics-Mixture-Cure-Models (LLMCM), Weibull-Mixture-Cure-Models (WMCM) and Generalised-Gamma-Mixture-Cure-Models (GGMCM) have been used to study cure rates in epidemiology. The GGMCM has been established to have out-performed other parametric models in terms of Akaike Information Criterion (AIC) but could not handle acute- asymmetry in survival data. Therefore, this study was designed to develop a Modified Generalised-Gamma Mixture Cure Model (MGGMCM) that can handle acute-asymmetry in survival data.

The GGMCM: $g(t)=\dfrac{\beta}{\theta\,\Gamma(\alpha)}\left(\dfrac{t}{\theta}\right)^{\alpha\beta-1}e^{-\left(\frac{t}{\theta}\right)^{\beta}}$ whereα, β are the shape parameters and θ is the scale parameter was modified using a gamma generator: $f(x)=\dfrac{1}{\Gamma(\omega)}\left[-\log[1-G(t)]\right]^{\omega-1}g(t)$

, where ωisthe shape parameter, $G(t)$ and $g(t)$ are cumulative density function (cdf) and probabilty density function (pdf), respectively. Life ovarian cancer data was obtained from Department of Obstetrics and Gynaecology, University College Hospital, Ibadan, Nigeria covering the period 2000-2015. The diagnosis time (in months) was until death. The simulation study utilised data based on the continuous uniform distribution with $b=100$and $a = 1$ using samples of sizes of 10, 20 and 50 each in 50, 100 and 500 replicates, respectively. The hazard-function of the MGGMCM was derived by $\dfrac{pdf}{1-cdf}$ of MGGMCM; the cure model was given as $S(t) = c + (1 - c)S_u(t)$ where $S(t)$ is the survival function of the entire population, $S_u(t)$ is the survival function of the uncured patients and c is cure-rate.Parameters c, $\tilde{\mu}$ (median time-to-cure) and σ of MGGMCM were determined using Maximum Likelihood Estimation. The MGGMCM was exhaustively investigated in terms of its parametric essence and against extant models of similar intent using relevant assessment criteria like scope, general estimability, AIC and relative efficiency of estimates of cure rate, median time-to-cure, variances, bias and mean square error where applicable.

The hazard-function of the developed model was obtained as:

$$h(t)=\frac{-\log\left|1-\frac{\gamma\left(\alpha,e^{\frac{\log t-\mu}{\sigma}}\right)}{\Gamma(\alpha)}\right|^{\beta-1}\frac{1}{\Gamma(\alpha)}e^{\alpha\left(\frac{\log t-\mu}{\sigma}\right)}-e^{\frac{\log t}{\sigma}}}{\Gamma(\beta)-\gamma\left\lfloor-\log\left|\frac{\alpha,e^{\frac{\log t-\mu}{\sigma}}}{\Gamma(\alpha)}\right|,\beta\right\rfloor}.$$

The AIC, median to cure, c and variance(c) of ovarian cancer data were: WMCM (216.89, 60.07, 0.29, 0.097); LNMCM (205.98, 57.98, 0.37, 0.004); LLMCM (203.27, 56.87, 5.95, 0.052); GGMCM (206.20, 20.90, 0.11, 0.005) and MGGMCM (199.24, 11.82, 0.82, 0.001), respectively. For simulated data, Mean Square Error (MSE) and $|bias|$were obtained as follows: $n = 10, r = 50$: WMCM (772.11, 26.10); LNMCM (707.23, 26.70); LLMCM (791.30, 27.89); GGMCM (691.03, 25.33); MGGMCM (701.10, 25.81), respectively. At $n = 20, r = 50$: WMCM (611.59, 23.59); LNMCM (655.71, 25.31); LLMCM (695.0, 26.00); GGMCM (601.33, 23.57); MGGMCM (609.31, 23.89), respectively. At $n = 50, r = 50$: WMCM (700.18, 25.77); LNMCM (699.52, 25.83); LLMCM (719.52, 25.50); GGMCM (689.15, 25.59); MGGMCM (601.59, 23.19), respectively. At$n = 50, r = 500$: WMCM (623.90, 23.95); LNMCM (619.61, 24.01); LLMCM (644.59, 24.89); GGMCM (602.10, 24.01); MGGMCM (501.37, 22.11), respectively. The better model corresponds to the smallest MSE and $|bias|$ values as sample size and replicate increases.

The Modified Generalised-Gamma Mixture Cure Model was the better on the AIC criterion; the MGGMCM adequately handled the problem of acute-asymmetry associated with survival data and its robustness.

Keywords: Acute-asymmetry, Cure-rate, Diagnosis-time, Hazard-function, Median-to-cure

Word count: 490

DEDICATION

The study is dedicated solely to Almighty Allah, the heaven and earth architect with everything in it. It is also dedicated to my family (OLAGESHIN-FOLORUNSO) and all my teachers and patients used for the study.

CERTIFICATION

We certify that this work was carried out by Mrs. Serifat Adedamola Folorunso in the Departments of Statistics and Obstetrics and Gynaecology, Faculties of Sciences and Clinical Sciences, University of Ibadan

....

Supervisor

Angela U. Chukwu

B.Sc. (Calabar), M.Sc. (Ibadan), PhD (Ibadan)

Department of Statistics,
Faculty of Sciences, University
of Ibadan

....

Co-Supervisor

A. A. Odukogbe

MBBS (Ibadan), FWACS, FMCOG, MCommH (Liverpool)

Department of Obstetrics and Gynecology,

Faculty of Clinical Sciences, College of Medicine, University of Ibadan

ACKNOWLEDGEMENTS

Time and mercy has touched me greatly in my search for knowledge. Therefore, all praise and worship is due to Allah for the success of the work. Many people assisted me to wrap up this thesis. Too many to mention individually.

My supervisor, Dr. Angela Unna Chukwu, is owed a debt of gratitude for her endless patience, guidance, contributions and tolerating me throughout the work period. I took full advantage of her knowledge and experience.

The challenges of ill-health led me into the research. Therefore, I appreciate the efforts of Dr. Akin-Tunde Ademola Odukogbe and his team for saving my life and later agreeing to be my co-supervisor in this research. He generously assisted me in conducting the research that led to the result listed in this thesis. Dr. Timothy A. O. Oluwasola from O & G (UCH), Ibadan has made significant contributions to this thesis ' success.

It is my pleasure to show appreciation to the academic staff of Statistics Department under the headship of Dr. Angela U. Chukwu, Prof. Iyaniwura, Prof. T.A. Bamiduro, Prof. A.A. Sodipo, Prof. G. N. Amahia, Late Dr. A. A. Adelakun, Prof.O. I. Shittu, Prof. O. E. Olubusoye, Dr. Adedayo A. Adepoju, Dr. J. F. Ojo, Dr. Deji Osowole, Dr. Oluwayemisi O. Alaba, Dr. K. O. Obisesan, Dr. C. G. Udomboso, Dr. Oluwaseun O. Otekunrin, Dr. O.S. Yaya, Dr. F. J. Ayoola, Dr.O. B. Akanbi , Dr. S. O. Oyamakin, Dr. K. A. Adepoju and Dr. K.A. Osuolale. Their support throughout the study cannot be quantified.

I appreciate the contributions of the non-teaching staff in the Department, Mrs. Arewa F.A, Miss Kehinde Akinseye, Mr. Oluwasegun Aderinto, Mrs. Taye Olunubi, Mrs. Yemisi Adara, Mrs. Akinola Bukola and Mr. Akimu Umaru.

I commend the Biometry Unit of my Department, under the auspices of my supervisor for a progressive academic exercise to make the research a reality. The contribution of the University of Ibadan Laboratory for Interdisciplinary Statistical Analysis [UI-LISA] under the coordination of Prof. O.E. Olubusoye cannot be overlooked.

The other honorary consultants and the resident doctors in the Gynaecology Oncology Unit from O & G Department, UCH, Ibadan also assisted with data collection. Mrs.

Oluwole and her colleagues in the hospital's Health Records Department assisted with retrieval of patients' information.

I am very grateful to Ajimah Atolani who assisted with data collection in University College Hospital, Ibadan, Nigeria. An important ingredient in the success of the thesis is getting along with good people. Therefore, my gratitude goes to Dr. (Mrs) Bidemi Yusuf, Mrs Sakirat Fasasi, Miss Maryam Olofinjana, Bayo Ademola and Partners, Mr. & Mrs. Sikiru Animasaun, Mrs Abolore Jimoh, Mr. Suraju Olageshin, Mrs Alaba Lamidi-Sarumoh, Mr Kaseem Olageshin, Mr Dayo Akindele, Dr Zainab Hamzat, Mr Basheer Olowookere, Alhaji Sulaimon Agaka, Mr Idris idowu, Mrs Latifat Aremu, Mr. Nasamu, Mrs. Abosede Olugbeja, Sister Funmi, Mrs Abibat Jimoh and Dr Kola Aliu.

I am also very grateful to A.G Leventis (EGBA) Scholarship Scheme (Ph.D programme scholarship for three years (2016 – 2018), Conference of Muslim Organisations of Nigeria (CAMON) scholarship, the ISCB for giving me CFDC grant for funding my travelling expenses to attend the 38th Annual Conference of the International Society for Clinical Biostatistics (ISCB2017) at Vigo, Spain, the IBS for giving me travel grant to attend a conference of Biometrics by Boarder, New South Wales, Australia and the Royal Statistical Society (RSS) for travel grant to attend 2018 RSS Conference at Cardiff, Wales, UK. I appreciate the financial support for local and international conferences from University of Ibadan Post Graduate College.

I am indebted to the efforts of the following people with methodological issues of planning, typing of Equations, development of codes, analysis, printing, editing, proof-reading and binding; Dr. Badmus Nafiu, Elder-Adeniran Tajudeen, Durojaiye Lekan, Olusoji Femi, Akinpeloye Olajide, Hamid Oyebanjo, Mr Amoo, Adio Muizi, Umaru Oyekoya, Akintande Olalekan, Afolabi Saheed, Olanrewaju Rasak, Ayodeji F. Ayoola, Ogbonaya Ezichi, Engineer Steven, Taiwo, Wale, Anu and Owolabi among others.

I recognise my friends ' commitment because friendship without self-interest is one of life's rare and precious things. In that case, I am magnanimously indebted to my friend and colleagues: Dr. Adewale F. Lukman, Dr. Kazeem Osuolale, Miss Abisola Agbonde, Mrs. Victoria Olaoye, Mr. Peter Koleosho, Mrs. Oluwaseun Oyegbile, Mr. Saheed AjoboMrs. Yemisi Akinlade, Mr. Isenah Godwin, Mrs. Kemi Olokuntoye, Mr. Yemi Okunlola, Mr. Wale Onantunji, among others.

This acknowledgement will not be complete if I fail to applaud the effort of Dr. Ogunjuyigbe A.S.O., Dr. K.M Samuel, Mr. Soji Adisa and Dr. Ilori for their moral support at a very critical stage of this thesis. My endless gratitude goes to the three Sub-dean (postgraduate) that worked on this thesis, Prof Oladosu did my conversion examination on February 19, 2016, Dr Ogunjobi did my Postgraduate Abstract Approval on July 19, 2019 and Dr. Olatunji did my PhD Oral examination on May 31, 2019. The Lord Almighty will reward you abundantly.

In the cause of the research, the support of my husband, my children, my siblings, my in-law and extended family is greatly appreciated. He laid the foundation for “freedom” as he encouraged my educational pursuit, despite the challenges of home keeping and children up-keep. My special gratitude is due to my mother Idiat Olageshin for her prayers, encouragement and her emotional support. I am grateful to my niece, Kareemat Adedeji for her endless support.

I cannot but also commend the effort of those whose names could not be reflected here but in one way or the other contributed to the success of my degree. May Allah be with you all, may the peace and blessing of Allah be upon our noble Prophet Muhammed (S.A.W), his companions, his households and the Muslims who adhere to his teachings (Amen).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

1.1 Background of Study

Survival analysis is a type of data analysis in which response variable becomes the time prior to the actual occurrence of an event of interest. This event might be death, disease incidence, marriage, divorce, among others. Survival analysis has become frequently beneficial in a variety of fields including biology, medicine, public health, as well as epidemiology.

A typical analysis of survival data involves the modeling of time-to-event data, such as the time until death. The time to the event of interest is called either survival time or failure time;possibly the time of diagnosis of a disease such as cancer.

Years, months, weeks, days etc. can be used as standard for the observation of time to event. If the observation at hand is ovarian cancer, then the survival times in years until the patients develop such cancer. The outcome variable may be observed from two perspectives. It may be time to event or the status of event, which reveals the condition of event of interest. Survival time or failure time is the time to the event of interest (Zhao, 2008). It frequently occurs that not all individuals under medical observation will be observed until time of failure; that is, some individuals may be lost to follow-up, or may drop out of the study. This is known as Censoring.

The common types of censoring are right censoring, left censoring, right truncation and left truncation. The details of these will be discussed in the next chapter.

For illustration, right censoring occurs when an individual decides to drop in a clinical trial before the event of interest occurs. It is the moment when the time-to-event of the individual has not taken place and leaves the trial. Let's say, if we consider time until death, and the subjects are admitted gynecological cancer patients aged 65 years or older, then every individual who dies before 65 years cannot be observed, this illustration gives what is meant by left-truncated data.

Right censored data have been considered throughout this thesis. The data used can be in the form of accurate data, censored data, or truncated data in survival analysis. Exact data set for survival can be identified similarly as uncensored data set for survival that takes place with specific time, pending interest time. With the set of censored survivaldata, the subjects ' time is unknown until the event of interest occurs but in a certain time (Zhao, 2008).

The ultimate objective of cancer research is to accomplish cure where feasible or manage a substantial proportion of patients such that the symptoms of the disease wholly disappear and the disease never resurfaces. Also, for cancer patients and the physician society alike, the odds of becoming cured and indeed the length of time of survival from diagnosis are also of great concern.

Researchers are therefore interested in benefiting from medical intervention in good number of patients. In survival analysis, it is obvious to have subjects with censored observations. These are subjects that will either be lost to follow-up by the time of the analysis and maybe have not experienced the major interest event yet. Special models in survival analysis established as cure models directly measure this proportion (Ibrahim, et.al. 2001 &Maller and Zhou 1996).

Cure models are unique forms of the approach of survival data analysis where it might be expected that certain fraction of subjects may still not experience the occurrence of interest. Thus, a plateau finally reaches the survival curve (Lambert et al, 2007).
Cure has turned out to be an important long-term, therapy-based survival management procedure. Many clinical researches have also lately concentrated on evaluating the fraction of cured patients, and many are not probably experiencing the event of interest again. Presently, there exist two (2) key categories of cure fraction models; the mixture cure fraction model and the non -mixture cure fraction model.

Boag (1949) first established a model that could be used to estimate the fraction of cure and in 1952 Berkson and Gage revised that model which was called mixture cure model. This one was identified similarly as the standard model of cure rate.

Non-mixture cure model widely known as Bounded Cumulative Hazard (BCH) model was another mixture cure model established by Yakovlev *et. al.*(1993).

Claeskens and Keilegom (2016) gave different names to the very same model; it is termed the Cure model in Biostatistics, the Split Population Model in Econometrics and Limited Failure Population Model in Reliability Engineering. Lambert *et.al.*(2006) established that cure is proven to happen any time and that the diseased category rate of mortality in individual's yields accurately same level as that expected in the over-all population.

Conventionally, proportion of cured patients from a disease refers to the cure fraction, along with being suitable techniques to measure the developments in survival of curable diseases.

An accurate and informative conclusion is an extension derived from cure models. These conclusions would otherwise be unobtainable from any analysis that fails to account for a cured fraction in the population. If a cured component is not present, the analysis reduces to standard approaches of survival analysis. Most cure models assume that the susceptible individuals are homogeneous in risk (Wienke *et al.* 2003).

Survival time is the time from the initial event until failure occurs say time from birth until death; time from diagnosis of ovarian cancer until the medical intervention; time from start of treatment until remission of disease. Survival time is the main interest in medical statistics and it is rarely normally distributed because it is a continuous measurement that cannot assume negative values.

Lucijanic and Petrovecki (2012) had already established that time is the response variable which mostly showcases the length of time before an event happens which could be empirically reviewed as one of the many research variables. The achievements and concern in clinical studies are indeed the survival of patients and thus no researcher is interested in death because this is an undesired event. Events could also be regarded as an end point as long as they could be characterized as dichotomous, e.g. "event takes place", is assigned one (1) and zero (0) otherwise. There are so many possible uses of such a sort of analysis in biomedical research findings, including evaluating time to recover following a long therapeutic procedure, time to achieve predetermined serum levels of a substance, time to recover from disease, time to discharge from hospital, and so many others.

A time trend is a significant techniqueused to assess advancements throughout cancer therapy for patients. The net survival change estimate is the focus when analysing time trends in survival analysis of patients with cancer. Net survival at time ‘t’ refers to the proportion of patients who would have survived up to that time ‘t’ if the only possible cause of death is the disease of interest that is there is no other cause of death. The analysis of Parametric Cure Model (PCM) was established 50 years ago.This technique is used for enhancement of appropriate requirements for analysis of clinical investigation. Also, it can be utilised for analysis from several cancers studies where achievements of cure are possible such as gynecology, pediatrics, etc (Maetani and Gamel, 2013)

1.2 Statement of Problem

In medical studies, researchers are concerned with analysing the progression of a disease and achievement of cure from management of the disease. Thus, we usually wish to calculate the probability of an individual person surviving for a given length of time. This is the ultimate objective of cancer research, which would be to achieve cure in a substantial proportion of patients undergoing cancer management trials. In other words, the disease's symptoms and signs completely disappear and the disease never recurs. The chance of healing and the number of years of diagnosis survival are of great interest to cancer patients as well as to the health community. In clinical studies, it is common to have some subjects that are censored in nature. These occurred when some subjects or observations are incomplete as a result of several random causes. In typical systems of analysis, censoring causes must be independent of the event of interest. Also, it must be noted that censored subjects resulted either from death from another cause differently from the cause of interest; subject survives when the study terminate or the subject is lost to the study, by dropping out from the study, moving to a different area, etc. Special models of survival analysis known as cure models easily calculate this proportion (Maller and Zhou, 1996& Ibrahim et al., 2001).

Going by common opinion, cancers have always been deemed as terminal diseases and not curable. Cancers however have differing cure rates. The survival or cure status can be measured scientifically giving an avenue for health professionals to evaluate the efficacy of treatment. Thus, mixture cure fraction model will be employed because of its ability to simultaneously estimate the cure fraction and identify the distribution of the uncured. Also, a new modified parametric mixture cure models is requiredto check the key features of asymmetry in survival data.

1.3 Motivation

In the past, very little attention was paid to models that could handle acute-asymmetry features of survival data by many authors. Zografos and Balakrishnan (2009), a skewed distribution family that is being applied to merge two distributions, have recently proposed the Gamma Generated link function. The proposed mixture cure model was generated using the gamma generator. Also, there have been high demands in cancer cure research across the globe.

This being a fact, in any epidemiology research the success of cure management is vital. Thus, this research draws motivation in seeking for alternative estimation method for the cure fraction within population based cancer investigation.

1.4 Justification of Study

Recently, the improvement of novel therapies has brought about reduction in the death of cancer patients. The inspiration for this research originated from the point that the success from management of cure remains vital from every clinical work / investigation, hence the term, curative Medicine used when individuals have developed the disease. This contrasts with preventive Medicine.

An operative health measure exists when a vital cohort of patients are being managed from a specific ailment, however with respect to several features including differing factors and different body chemistry, some patients will not benefit from the particular treatments. Advancing in health managements, it is pertinent for individual subject to detach between changes within cure possibility and rise of expected survival time for uncured patients. Thus; Mixture cure fraction model will be employed because of its ability to simultaneously estimate a cure fraction and identify the distribution of the uncured. Also, the major characteristics of survival data is acute-asymmetry, hence, this require a parametric mixture cure model that can handle the acute-asymmetry in the survival data.

1.5 Scope of Study

This study covers four prevailing parametric Mixture Cure Models (MCM) which becomes the competing models for the new proposed one. These competing parametric cure models are lognormal MCM (LNMCM), loglogistic MCM (LLMCM), Weibull MCM (WMCM) and generalised-gamma MCM (GGMCM). These models had been used previously on gastric cancer data and it was discovered that generalised-gamma MCM performed better than lognormal MCM, loglogistic MCM, Weibull MCM but it could not control acute-asymmetry features in survival data. The present study will use real life ovarian cancer data to determine the flexible model that can compete with the existing models. Also, simulation study will be done; the simulation sample size will mimic the real life ovarian cancer data. The parameters estimation for the study will make use of Maximum likelihood estimation.

1.6 Definitions of Some Concepts

Survival time: It is the time from the initial event until failure occurs such as time from birth until death; time from HIV infection until the development of AIDS; time from start of treatment until remission of disease. Survival time is rarely normally distributed because it is a continuous measurement that cannot assume negative values. In cancers, it is the time from diagnosis.

Censoring: It is the incomplete observation of time to failure. Since one does not have complete observation which can give full information hence there is partial information contained in censored observations.

Cure: This is the accomplishment of a health condition. The component or process that terminates the health situation could be a drug prescription, medical managements, lifestyle changes, as well as philosophical frame of mind, meant to finish a person's suffering. The situation might also be referred to the level of being healthy or treated.

Remission: This is a short-term completion to the health indications and warnings of a life-threatening disease.

Relapse: A disease is said to be lethal if there exists an odds of the patient relapsing, no matter how elongated the patient has been in remission. Relapse (National Cancer Institute, 2018): This can also be referred to reoccurrence of a disease or reappearance of the signs and symptoms of a disease after a period of improvement.

Expected survival: It is obtained from national population life tables stratified by age, sex, calendar year and possibly other covariates.

Statistical cure: It is when the mortality rate observed in the patients eventually returns to the same level as that in the general population.

Cure fraction: The proportion of diseased cohorts that are treated through a particular management is refers to as cure fraction or cure rate. This is arisen via linking the disease-free survival of healed populaces against a corresponding control unit that have not experience the disease of interest. A different approach for determining cure fraction stands by means of determining at what time is the hazard rate of individuals diseased group resulted to the measurement of general population hazard rate. [Lambert. (2007) and Smoll *et.al.* (2012)]

1.7 Advantages of Parametric Cure Models

The advantages of parametric over nonparametric/semi parametric cure models are:

1. Flexibility of its hazard functions.
2. It is possible to estimate its parameters while the cox model can only provide estimates of the hazard differences between two or more groups.
3. We usually have data sets in health-life research that require more sophisticated parametric models.
4. It has long term validity and is more useful when compared with nonparametric and semi parametric cure models.
5. It is used to distinguish between curative and life prolongation therapies where by giving clinicians a false estimate of the best regimen, traditional methods fail to do this.
6. It can be used to predict time for survival.

1.8 Aim and Objectives

1.8.1 Aim

The aim of the study is to develop a modified generalised-gamma mixture cure model that is explicitly competent to handle acute-asymmetry in survival data. This in turn will give comprehensive and additional report/evidence for the cured patients' proportion that have benefited after health managements.

1.8.2 Specific Objectives

1. To derive a new parametric mixture cure fraction model capable of handling acute asymmetry in survival data.
2. To investigate the properties of the new parametric cure fraction models.
3. To compare the modified model with the existing ones.

CHAPTER TWO

LITERATURE REVIEW

2.1 Survival Analysis: An Overview

Survival analysis can be described as a variety of techniques for data analysis whereby the response variable becomes period before an event of interest takes place. This type of analysis generally centers on time to event data. It is a discipline of statistics that deals with data from time to event. From a broad perspective, it consists of techniques for positive, valued random variables, such as: time to death, time to recovery from a disease, hospital admission, duration of strike by workers,duration of doctoral study, time of diagnosis until death and so on. In medical research, times until death or recurrence of the ailment are the major category of time-to-event data where the concerned random variable remains a continuous positive random variable “T” (Claeskens and Keilegom, 2016).

The hazard functions and survival functions stand as the fundamental theories in survival analysis aimed at showing event times distribution. The probability of surviving, that is, when the patients are not experiencing the event is specified by the survival function. The hazard function provides the possibilities for the event to occur per time unit, given that an individual has survived up to the specified time. Studies show that, it is regularly of direct concern. There are other concepts e.g., median survival, cumulative hazard function, etc., that might be measured from knowing either the hazard or survival function.

In survival analysis, it normally happens that somehow the occurrence of interest may not be experienced by several individuals under study; they are referred to being' cured.' Therefore, the population is a mixture of two population units, whereby, one is a cured entity and the other a' susceptible' entity (Patilea and Keilegom, 2017). The time from entry into a study until an individual has a particular outcome is referred to as ‘time-to-event’.

2.1.1 Survival Function (S(t))

The survival function explains the probability of an individual surviving beyond a specified time t. Here, T is represented as the random variable specifying survival time. Therefore, ‘t’ is the time until the event of interest. The probability of experiencing the event of interest beyond time t is presented by the survival function (Coolen, 2012). The statistical expression of the survival function is shown in equation (2.1) below;

$$S(t) = P(T > t) = 1 - F(t)(2.1)$$

Where $F(t)$ is the probability distribution function and it is given by; $F(t) = Pr(T \leq t)$

T is a continuous random variable. The survival function can be presented as it is in the equation below, where $S(t)$ is the integral of the probability density function (PDF), $f(t)$:

$$S(t) = P(T > t) = \int_0^\infty f(t)\, dt \qquad (2.2)$$

Characteristics of S(t)

Wang (2006) established the survival function with the following features:

1. $\mathrm{S(t)} = 1 \text{ if } \mathrm{t} < 0$; therefore, the survival function S(t) is a non-increasing function of that same value 1 at $t = 0$,that is. $S(0) = 1$.
2. For a proper random variable $T, S(\infty) = 0$, implying that ultimately everyone will experience the event.. However, we will also allow the possibility that $S(\infty) > 0$. This coincides to some kind of scenario in which there is a high likelihood that perhaps the event would not die and therefore not experience it. For example, if the event of interest is the time from both the response to the deterioration of the disease as well as the disease really does have a cure over a certain percentage of the population, then we will have $S(\infty) > 0$, where $S(\infty)$ corresponds to the percentage of the population of cured individuals or groups.

Generally speaking, the survival function S(t) provides important information like those of average survival rates, t-year death rate, etc.

2.1.2 The Hazard Function (h(t)):

Wang (2006} described the function of hazard as the rate of instant failure. In the real sense, a person is more likely to experience the event of interest while 't' has not occurred . Hazard function is represented by;

$$h(t) = P\left(t < T < \frac{t + \Delta}{T} > t\right) \tag{2.3}$$

$$= \frac{f(t)}{1 + F(t)}$$

$$h(t) = \frac{f(t)}{s(t)} \tag{2.4}$$

The instantaneous failure rate $(h(t))$ refers to outcome of the risk concept in an interval after time ‘t’ . This is condition on the subject having survived to that time ‘t’. Such outcome could be death, failure, hospitalisation and so on.

The hazard function corresponds to the phenomenon of outcome risk (e.g., death, failure, hospital admission) at a time t interval, conditional on the subject having survived to time t.

Also, it provides the possibility that there is occurrence of death between t and$(t + \Delta)$, divided by the probability that the individual survived beyond time ‘t’. The $(h(t))$ contributes additional essential to the general usage in survival analysis than the density function. It is also used to compute the instantaneous risk.

2.2 Forms of Survival analysis

Blayney (2012) stressed that several models in survival time are available for use in the relationship between the set of explanatory variables. He furthered categorized them as parametric, nonparametric and semi parametric approaches.

2.2.1 The Parametric Survival Analysis

In parametric techniques of survival analysis, the random variable T is assumed to follow known probability distributions. The common ones are the Log-Logistic, Weibull, and Gamma distributions among others.

The description of probability distributions for outcome variable is a function of the predictors of interest (Fieller, 2010). In these conditions, estimations of parameter requires an appropriate modification of maximum likelihood.

Normally, there is need to estimate unknown parameter θ from the data because the probability density function relies on it. There are lots of estimation techniques; however, Fieller (2010) asserted that maximum likelihood estimation (M.L.E) is the best with the justification of asymptotic properties i.e. large samples.

In summary, the likelihood of a parameter θ for data $x_1, x_2, \dots \dots, x_n$, is the probability of observing the data $x_1, x_2, \dots \dots, x_n$This probability is calculated in terms of unknown quantity θ and so will be a function of it, $L(\theta)$ say. We can now maximize $L(\theta)$ with respect to θ, and the value that produces the maximum say, θ, is the maximum likelihood estimate of θ. It can be thought of as the most probable value of θ, in the light of the data just obtained.

Examples of Parametric Survival

2.2.1.1 Log-logistics distribution

The distribution of log-logistic probability density function is specified as:

$$f(x) = \frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1}\left(\left(1+\frac{x}{\beta}\right)^{\alpha}\right)^{-2} \qquad x, \alpha, \beta > 0 \qquad (2.5)$$

The cumulative distribution function $(F(x))$ of log-logistic distribution is specified as:

$$F(x) = \int_0^x f(t)dt$$

$$\therefore F(x) = \int_0^x \frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1}\left(\left(1+\frac{x}{\beta}\right)^{\alpha}\right)^{-2} dx$$

$$F(x) = \frac{\alpha}{\beta\beta^{\alpha-1}}\int_0^x x^{\alpha-1}\left(\left(1+\frac{x}{\beta}\right)^{\alpha}\right)^{-2} dx$$

$$F(x) = \left(\left(1+\frac{\beta}{x}\right)^{\alpha}\right)^{-1} \qquad (2.6)$$

The survival function $S(x)$ of log-logistic distribution is specified as:

$$1 - F(x)$$

Where

$$S(x) = 1 - \left(1 + \left(\frac{\beta}{x}\right)^{\alpha}\right)^{-1}$$

$$S(x) = \frac{1}{\left(1 + \left(\frac{\beta}{x}\right)^{\alpha}\right)}$$

$$= \frac{\left(\frac{\beta}{x}\right)^{\alpha}}{1 + \left(\frac{\beta}{x}\right)^{\alpha}} \qquad (2.7)$$

The hazard function $h(x)$ of log-logistic distribution is specified as:

$$h(x) = \frac{f(x)}{1 - F(x)}$$

$$h(x) = \frac{\frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1}\left(\left(1 + \frac{\beta}{x}\right)^{\alpha}\right)^{-2}}{\frac{\left(\frac{\beta}{x}\right)^{\alpha}}{1+\left(\frac{\beta}{x}\right)^{\alpha}}}$$

$$h(x) = \frac{\frac{\alpha}{x}\left(\frac{x}{\beta}\right)^{\alpha}\left(\left(1 + \frac{\beta}{x}\right)^{\alpha}\right)^{-2}\left(1 + \left(\frac{\beta}{x}\right)^{\alpha}\right)}{\left(\frac{\beta}{x}\right)^{\alpha}}$$

$$h(x) = \frac{\frac{\alpha}{x}\left(\frac{x}{\beta}\right)^{2\alpha}\left(1 + \left(\frac{\beta}{x}\right)^{\alpha}\right)}{\left(\left(1 + \frac{\beta}{x}\right)^{\alpha}\right)^{2}} \qquad (2.8)$$

2.2.1.2 The Weibull Distribution

The distribution of Weibull probability density function is specified as:

$$f(x) = \frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1} \exp\left(-\left(\frac{x}{\beta}\right)^{\alpha}\right) \;\; t > 0, \alpha > 0, \beta > 0 \tag{2.9}$$

The cumulative distribution function $F(x)$ of Weibull distribution is specified as:

$$F(x) = \int_0^{\infty} f(x)dx$$

$$= \int_0^{\infty} \frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1} \exp\left(-\left(\frac{x}{\beta}\right)^{\alpha}\right) dx$$

$$= \frac{\alpha}{\beta\beta^{\alpha-1}} \int_0^{x} x^{\alpha-1} \exp\left(-\left(\frac{x}{\beta}\right)^{\alpha}\right) dx$$

$$F(x) \; = 1 - \exp\left(-\left(\frac{x}{\beta}\right)^{\alpha}\right) \tag{2.10}$$

$$S(x) = 1 - F(x)$$

$$= 1 - \left(1 - \exp\left(-\left(\frac{x}{\beta}\right)^{\alpha}\right) \;\right)$$

$$S(x) = \; e^{-\left(\frac{x}{\beta}\right)^{\alpha}} \tag{2.11}$$

The hazard function $h(x)$ of Weibull distribution is specified as:

$$h(x) = \frac{f(x)}{S(x)} = \frac{f(x)}{1 - F(x)}$$

$$= \frac{\frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1} \exp\left(-\left(\frac{x}{\beta}\right)^{\alpha}\right)}{\exp\left(-\left(\frac{x}{\beta}\right)^{\alpha}\right)}$$

$$h(x) = \frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1} \tag{2.12}$$

2.2.2 The Non Parametric Survival Analysis

Coolen (2012) described Kaplan Meier as a technique which is generally used as a function of time in non-parametric survival models to evaluate survival probabilities. This technique provides a graphical survival probabilities as well as obtaining descriptive results for the survival data. The method is very useful for comparison of two or more survival experience. It can also be used for comparing survival curve for two or more groups of subjects.

The Kaplan Meier method is a typical example of nonparametric where the outcome variable does not follow any underlying distribution.Kaplan & Meier (1958) proposed the technique. It is distribution-free and many other statistics could be estimated using the Kaplan Meier method, such as median survival time, to describe the survival data. There is other nonparametric test for survival models such as log-rank test which is a type of chi-square test. This method is used for comparing estimated Kaplan-Meier curves in respective group of subjects (Coolen, 2012).

2.2.2.1 The Kaplan Meier Estimate

The likelihood of staying alive until the end of each observation is estimated using the techniques of Kaplan Meir. The time for each surviving patient at the beginning of that time, could be evaluated in days, months or years (Blayney , 2012). The approach is considered conditional likelihood meanwhile the survival estimate at the end of the current observation period would be based on the patient's last survival condition.

Censored patients make a significant contribution to the likelihood of surviving for every observation time of which they are followed throughout the study. Patients experiencing the "event" will reduce the likelihood of survival for the next observation time. The term cumulative probability of survival is the combination of all the conditional probabilities of surviving for each observation time and is represented by S(t) notation. Assuming time was classified into the following intervals $t_0 < t_1 < \cdots < t_m$ where starting time is 't_0' and the study ends at time 't_m'. Actually, the time periods considered corresponds to death or censoring events. This indicate that no one dies or censored between times t_j and t_{j+1}. The set of all subjects who survive to the time just before time t_j is the risk set $\mathrm{R_j}$. Thus it consists of all those who die at time t_j or who are censored or die after time t_j.

We further define n_j = the number of subjects at the time $t_j$and d_j = the number who die at time t_j. Thus, $S(t_j)$ could therefore be computed iteratively as follows:

$$s(t_0) = 1 \tag{2.13}$$

$$s(t_{j+1}) = s(t_j).\left(1 - \frac{d_j}{n_j}\right) \tag{2.14}$$

Based on the above definitions,

For any $t,\ t_k \leq t < t_{k+1}$ for $k = 1, \dots, m-1$

$$s(t) = \prod_{j=1}^{k}\left(1 - \frac{d_j}{n_j}\right) \tag{2.15}$$

$S(t) = 1\, for\, t < t_1$. If t_m is a censored time, then $S(t) = 0$ for $t \geq t_m$. Otherwise, $S(t)$ is undefined for any $t > t_m$.

Also note that if there are no censored data, then $n_{j+1} = n_j - d_j$. Thus since $n_1 = n$, it follows that

$$s(t) = \prod_{j=1}^{k}\left(1 - \frac{d_j}{n_j}\right) = \frac{n_2}{n_1} . \frac{n_3}{n_2} \dots \frac{n_{k+1}}{n_k} = \frac{n_{k+1}}{n_1} = \frac{n_{k+1}}{n} \tag{2.16}$$

2.2.3 Semi-Parametric Survival Analysis

Cox proportional hazards regression models remains a widely used regression technique for the analysis of survival data.

It facilitates analysis of differences in survival times for various groups of interest, while allowing adjustment for covariates of interest.

Bewick *et.al.*(2004) described the model of Cox regression as a semi-parametric model with far less assumptions than conventional parametric methods. It also has more assumptions than previously explained non-parametric methods. This does not make any assumptions about some of the description from the so-called baseline hazard function as opposed to parametric models

The Cox regression model is advantageous and straightforward to interpret by providing predictors with details about the relationship of the hazard function. A nonlinear relationship is assumed between the function of hazard and the predictors. The hazard ratio comparing any two observations is constant over time in the setting where the predictor variables remain constant. This assumption is called the proportional hazards assumption. Verifying this assumption is an important part of Cox regression analysis (Wolsztynski , 2015).

The Cox Proportional Hazard Regression Model

The model of Cox regression is a strategy frequently used for the analysis in survival. Its attraction is analogous to conventional methods of regression in which one or more independent variables explain a response variable.

Cox Regression Model

$$h(t; x) = h_0(t) ex\,p\{B_1 x_1 + \cdots + B_k x_k\} \tag{2.17}$$

where:

h_0(t) is the baseline hazard function, i.e. the hazard function when all covariates equal zero

exp is the exponential function (exp(x)=e^x)

x_i is the i^{th} covariate in the model, and

$B_i$is the regression coefficient for the i^{th} covariate , x_i

The peculiarities of the Cox Regression Model

The Cox Model varies greatly from normal regression in such a way that somehow the covariates are being used to estimate the hazard function, instead of just dependent variable Y.

The hazard function of the baseline can sometimes take any form, but it might not be negative. The covariates ' exponential function is often used to ascertain that the hazard is positive.

There is no intercept in the Cox Model. ((Any intercept might be consumed into the baseline hazard. The fundamental Cox Model assumes that somehow the hazard functions of two covariate levels are proportional to all t values.

The Cox Proportional Hazard Model

Hence, the following generalization is considered;

$$h(t,x) = h_0(t,\propto) ex\,p(B^T x) \qquad (2.18)$$

Where $\propto$ are some parameters influencing the baseline hazard function. The hazards are decomposed into a product of 2 items;

$h_0(t,\propto)$a term that depends on time but not the covariates and;

$exp\ (B^T x)$: : a term that depends on the covariates but not time.

The hazard function is the probability that an individual will experience an event (for example, death) within a small time interval: given that the individual has survived up to the beginning of the interval. It can therefore be interpreted as the risk of dying at time 1. The hazard function denoted by h(t) —can be estimated using the following equation:

$$h(t) = \frac{number\ of\ individual\ sex\ periencing\ an\ event\ in\ interval\ beginning\ at\ t}{(number\ of\ individuals\ surving\ at\ time\ t)\ X\ (interval\ width)}$$

To see the proportional hazards property analytically, take the ratio of $h(t;x)$ for two different covariate values:

$$\frac{h(t;x_i)}{h(t;x_j)} = \frac{h_0(t)exp\,\{B_1 x_{i1} + \cdots + B_k x_{ik}\}}{h_0(t)exp\,\{B_1 x_{j1} + \cdots + B_k x_{jk}\}} \tag{2.19}$$

$$= ex\,p\{B_i(x_{i1-}x_{j1}) + \cdots\ B_i(x_{ik-}x_{jk})\}$$

$h_0(x)$ cancels out => the ratio of those hazards is the same at all time points. For a single dichotomous cavariate, say with values 0 and 1, the hazard ratio is

$$\frac{h(t;x_{i=1})}{h(t;x_{i=0})} = \frac{h_0(t)e^{B*1}}{h_0(t)e^{B*0}} = \frac{e^B}{e^0} = e^B \tag{2.20}$$

2.3 Concepts of the Log Rank Test

The log rank test is used to examine null hypothesis of no difference in survival time between two or more predictors group.

The test compares the entire survival experience between groups. It can also be used to check if the survival curves are identical (overlapping) or not. Survival curves are estimated for each group separately, using the Kaplan Meier method and compared statistically using the log rank test. The log rank test is a non-parametric test which is distribution free. In essence, the log rank test compares the observed number of events in each group to what would be expected if the null hypothesis were true i.e. If the survival curves were identical (Bewick et. al, 2004).

H_0 : The two survival curves are identical

H_1 : The two survival curves are not identical

$(\propto= 0.05)$ Level of significant

Chi-square Test

Chi-square is a test commonly used to compare observed data that would be expected according to specific hypothesis. The chi-square test is always testing the null hypothesis which states that there is no significant difference between the expected and the observed result.

Chi-Square is denoted as χ^2 and the basic computational formula is given as

$$\chi^2 = \frac{(O_i - E_i)^2}{E_i}$$

Where O_i represented observed data or frequency, E_i represent expected data or frequency.

2.4 Concepts of Censoring

Cook (2008) perceived that subjects had always been followed within a length of time thus this is the main purpose of the interest in the analysis of survival. The use of a linear regression model is therefore not appropriate for survival time as a function of a set of explanatory variables since survival times are typically positive integers ; ordinary linear regression might not have been the best alternative unless these times are first transformed to remove this restriction, and ordinary linear regression cannot adequately justify censoring observations. Cook (2008) had said that the censored observation is termed incomplete information in an observation. There seem to be four popular types of censoring, viz right censoring, left censoring, right truncation and left truncation.

For quite number of reasons, the study focuses exclusively on right censoring. Moreover, the most easily understood of all four types of censuring is right censoring, and if a researcher can understand the concept comprehensively, ability to understand the other three types becomes much easier. Therefore, if an observation is right censored, it indicates that somehow the information is incomplete in some way because the subject did not experience an event throughout the whole period. The point of the analysis of survival is to follow subjects over time and observe when they experience the event of interest. Sometimes it tends to happen that, for all of the subjects throughout the study, the study sometimes doesn't take approximately sufficient time to observe the occurrence. This might have been due to a number of reasons. For reasons unrelated to the study, subjects may drop out from the study (i.e. patients moving to another area and leaving no address).

The unique feature behind all these instances is that when the subject might continue to stay in the study, then this would ultimately had been possible to study the time of the event (Cook, 2008).

2.4.1 Types of Censoring

Right-Censoring is an observation that is popular. If patients are followed in a study for a few months and do not experience the event of interest for the duration of the study, it is called right censored.

The patient's survival time is considered based on the study length of time. A good example of right censoring is that when a patient decides to leave the study before the end of the observation time and has not experienced the event. The survival time of the patient is said to be censored because there was no event of interest. (Cook, 2008).

Right censoring may take place because of the following:

- The study terminates whereas the subject survives
- The subject dies from a different cause.
- The subject is lost to the study by willingly dropping out.

This is called right-censoring because while the real un-observed occurrence is on the right side of our censoring time; i.e. at the end of the follow-up, not the entire event actually occurred.

Andersen and Keiding (1998) provided practical illustrations of right censoring. If one models the time for occurrences of failure, there is a justified reason for censoring, to be specific; one would not actually expect all participants to fail. Assuming the effect of children's vaccines is screened and a randomized controlled trial should be conducted implying that the last individuals in the study would die a hundred years or more from the start of the study. This naturally introduces censorship and the case is right-censoring signifying that "one does not know how long this person is going to live; only one knows that the individual is still alive."

Right-censoring could also happen if subjects are lost to follow-up in the randomized controlled trial, e.g. they might need to temporarily suspend or move away from the study.

Andersen and Keiding (1998) referred to all of these as examples of right-censorship, the main interest is in the longevity of the subjects, but due to practical circumstances one only has censored observations, meaning that for some subjects one will never know when they die, only that they were still alive at some point in time (the censoring period). Thus, one knows that the data point (time of death) for a censored individual is significantly larger than a certain value (time of censoring).

Left-Censoring: Cook (2008) sees that it may be too short to observe the failure time. Consider, for instance, a study in which interest centers on the time after surgical removal of the primary tumor recurring of a particular cancer. The patients are examined a few months after the operation to determine whether another cancer has recurred. Let T = time from surgery to cancer recurrence. Some of the patients could be found to have a recurrence at this time and hence the actual time from operation to examination is less than the time. It would seem that such situations are left censored.

In other words, Cook (2008) said that Left censoring actually takes place when subjects are only observed at a fixed appointment and it is only then discovered that death occurred sometime before, so survival times are less than the observation period.

Andersen and Keiding (1998) also gave as an example of left-censoring assuming some baboon troop always sleeps in the trees. one's interest is to estimate at what time in the morning they descend from the trees, and assuming that they do descend every day. Following them for couples of days, however, while sleeping, meaning that some days they descend before one even arrives at the scene. If one arrives at 9 a.m. on a particular day *x* and the baboons already descended, that gives left-censored observations. One needs to know when they descended, but all result is an upper limit (9 a.m.), because since the time of arrival they had already descended. Analogously, one now know that the data point (time of descent on day *x*) is smaller than a certain value (9. a.m.).

Interval Censoring

In Interval censoring, the exact time of occurrence of event is not known precisely, but an interval bounding this time is known. If interval is very short, e.g. one day, it is common to ignore this form of censoring and pick one end point of the interval consistently. Examples of interval censoring include: infection with a sexually-transmitted disease such as HIV/AIDS with regular testing (e.g. annually); and failure of a machine during the Japanese New Year (Cook, 2008).

Censoring versus Truncation

On the final note, Cook (2008) generally compared censoring to truncation and said that Censoring is when an observation is incomplete due to some random cause. The cause of the censoring must be independent of the event of interest if we are to use standard methods of analysis.

Truncation, on the other hand, is a variant of censoring which often occurs when the observer's incomplete nature is due to a systematic selection process inherent in the design of the study.

2.5 Reviews on Cure Models

Survival data models with surviving proportion can be referred to as models capturing the rate of cure as well as long-term survival models which can handle essential task in survival analysis. Boag (1949) developed the model and it was revised by Berkson and Gage (1952) who extended the long-term survival models as the mixture model to study events in which a percentage of the patients are cured.

In this specific model, the survival at time (t) is equivalent to the cured and uncured individuals. The moment all the uncured individuals have died or re-developed the disease, leaving only the completely cured members of the population then the Disease Free Survival (DFS) shape is going to be completely leveled. The original time that the shape proceeds to being leveled is the stage at which almost all outstanding disease-free survivors are admitted to be completely cured. In the case where the shape never goes leveled, the disease is definitely regarded incurable with the prevailing therapies. After a couple of years, researchers like Yakovlev *et al.* (1994), Yakovlev and Tsodikov (1996) and Chen *et. al.* (1999) in conjunction with Ibrahim et al. (2001) who developed an exceptional instance of the weighted Poisson distribution and expressed the long-term survival function.

Tsodikov, *et.al.* (1998) reviewed a prevailing cure fraction model and attention from proportion of survivor of long-term in phase moving hazard associated to the relapse of observed subjects of leukaemia in Hodgkin's disease cured patients. Their models were applied to Hodgkin's disease data from the International Data-base.

In 2000, Ibrahim, and Chen developed methods of Bayesian cure fraction models designed for multivariate failure time data using right censoring subjects. They offered a different model, termed as the multivariate cure rate model, and also provided a naturalmotivation and interpretation of it. Their proposed model was an introduction of a frailty term with the assumption of a positive stable distribution to create the association structure between the failure times in possession. The outcome from the correlation system is categorized with the frailty term and this outcome is quite innovative and generates a very suitable characterization of the association between the failure times.

Ibrahim et al. (2001) further confirmed that the Bayesian cross-validated predictive density based method of Conditional Predictive Ordinate (CPO) is problematic predominantly designed for models with cure proportion because the CPO for a subject can be either density-value (for uncensored non-cured) or probability (for either censored or cured).

The most acceptable techniques in statistic established designed for calculating cure otherwise prognostic properties originating aftermath from cancer emanated from the logrank test and Cox regression analyses was established by Sposto (2002). This depends on the proportional hazards (PH) assumption and implicitly which emphasize covariate effects on failure times rather than their effects on the proportion of cures. Classes of two parametric cure models (PCMs) were used to analyse Children's Cancer data which were compared to analyses of Cox regression model and concluded that PCM analyses results are similar or identical to Cox regression analysis when the PH assumption is appropriate and inappropriate, PCMs can offer a coherent way to investigate and report covariate effects on the proportion cured separately from their effect on time to failure. Despite their reliance on explicit parametric forms, PCMs often provide a good description of cancer outcome, and are insensitive to lack of fit provided that follow-up is sufficient.

In 2004, Binbing *et. al.* progressed from the semi parametric model to parametric cure model by recommending utmost frequently its usage among which are lognormal, loglogistic, weibull and generalised-gamma distributions and concluded that most estimates acquired from cure proportion especially from generalised-gamma distribution are established to be relatively characterized with statistics robustness. They

finally recommended that several cautions should be taken on the global use of such type of cure fraction models. Their study employed simulation.
Yingwei and Keith (2000) asserted that nonparametric techniques had attracted fewer attentions compared to the parametric techniques for analysis of cure rate. Therefore, they considered a non-parametric convectional mixture cure model where the proportional hazard assumption is used to model covariate effects on the failure time of patients who are not cured. To estimate parameters of interest in the model, the EM algorithm, the marginal likelihood approach, and multiple imputations were used. Yingwei and Keith (2000) also expanded the proportional hazard regression model of Cox by enabling a percentage of event-free patients and examining covariate impacts on that percentage.Their model has indeed been affirmed by data on breast cancer as well as its parameter estimation procedure and simulations have also been investigated. Their result incorporates comparisons with previous analyses with a parametric model, and other researchers support the conclusions of the parametric model but not those of the existing nonparametric model.

Odukogbe et.al. (2001) reported that in the developing world there have been poor and very few reports of previous ovarian cancer studies and suggested that motivating revitalized ovarian cancer studies in less developed countries such as Nigeria is absolutely essential for so many reasons.

Judy and Jeremy (2004) modeled the probability of incidence and parametric failure with the assumption of binary distribution to model latency and improved maximum likelihood procedures for joint estimation using a non-parametric probability form and an Expectation Maximization (EM) algorithm. They also affirmed that parametric survivorship analyses of clinical trials usually consist of the assumption of a hazard function constant with time. When the empirical curve obviously levels off, one can transform the hazard function model with the use of Gompertz or Weibull distribution with hazard decreasing over time.

Odukogbe *et.al.* (2004) reported that the Gynecology Oncology Units (GOU) of Department of Obstetrics and Gynaecology, University College of Hospital, Ibadan, Nigeria that is saddled with responsibility of services of cancer in the ovary had commenced their premier longitudinal study for this life threatening female cancer starting in December 1st, 1998 so as to inaugurate a regional cancer research study

Centre. They affirmed that the case fatality rate 6 month after surgery was 76% for ovarian cancer patients according to their study in Ibadan.

Betensky and Schoenfeld (2001) employed a risk model that is competing in nature and can be referred to a corresponding mixture model which was established for cure data. Furthermore, they obtained an estimator for the variance of the cumulative incidence function from the competing risks model, along with the cure rate, based on elementary computations.

Yin and Ibrahim (2005) worked on unified cure rate model and Cancho et.al (2011) modified the Conway–Maxwell–Poisson (COMPoisson). The new model is called The Conway–Maxwell–Poisson-generalised gamma regression model. Theproposed Poissons (COMPoisson) cure rate model is a flexible substitute for the unified cure rate model reviewed by Yin and Ibrahim (2005). Their model was able to account for over-dispersion and under-dispersion which is mostly encountered in discrete data.

Lambert *et. al.* (2007) upgraded non-mixture cure fraction model within the parametric framework in the direction of integrating background mortality, via stimulating estimations from cure fraction within population based cancer investigations. They studied estimations from relative survival and the cure fraction amidst two forms of model as well as investigating modeling significance of parameters ancillary for both types of model over selected parametric distribution.

Yi and Tiwari (2007) developed an innovative risk cure fraction models that is competing in nature and they modeled the dependence amongst the time of censoring and the survival time with the aids of a class models from Archimedean copula. Therefore, estimating parameter from sample with large result with the means of Matingale principle through the simulation study was considered.

Cooner, Banerjee, Carlin, and Sinha (2007) proposed a set of unifying cure proportion models which can expedite highest structure of hierarchical model building together with mutually prevailing cure model sets as exceptional conditions. This uniting group facilitates model robustness by means of detecting for improbability that resulted from some underlying management of cure. Concerns such as regressing on the cure fraction and propriety of the associated posterior distributions under different modeling assumptions finally offered a simulation study for illustration with two datasets (on melanoma and breast cancer) that revealed their framework's capability to differentiate between fundamental procedures resulted to deterioration and cure.

Andersson et.al. (2011) used cure fraction models to analyse time trends and gave valued information to unravel certain problems with the cure fraction models. They extended the flexible parametric survival model by including cure as a special case to evaluate the proportion of cure and the "uncured" survival. Robust parametric survival models use splines to model the underlying hazard function and therefore need not specify any parametric distribution.

Abu-Bakar, Salah, Ibrahim and Haron (2009) assessed the cure fraction models and revised them towards checking developments from survival of cancer patient over time, besides concentrating on a number of difficulties arose from the usage of the models.

Ortega, Cancho and Lachos (2008) reconsidered non informative prior's assumptions for Bayesian analysis designed using parameters from cure fraction model. The fresh methodology existed with the aids of Markov Chain Monte Carlo Procedures using an algorithms step of Metropolis-Hasting towards achieving summaries of interest from posterior. Aiming at cure fraction and covariates in survival data analysis, specific influence procedures like the local influence, individual total local influence, predictions local influence and generalised leverage were derived, analysed and discussed.

Shuangge (2009) affirmed that survival data for current status can be measured when the time from censoring random cause as well as the time from event censoring can be observed.

Thus, they considered cure model data for current status, with percentage of the cohorts that are not predisposed to the event of interest. They also assumed cure probability with generalised linear model and proposed new techniques of parameter estimation known as (penalized) maximum likelihood. Their results indicated that parametric regression coefficients estimates yields inconsistency, asymptotically normal and efficient.

Cancho, Ortega and Bolfarine (2009) developed the log exponentiated-Weibull regression model to allow the possibility that long term survivors were present in the data. The modification led to a log-exponentiated-Weibull regression model with cure rate, as special cases. The effects of covariates on the acceleration/deceleration of the timing of a given event and the surviving fraction were simultaneously estimated with their models; that is, the proportion of the population for which the event never occurs.

Ranganathan, Rajaraman and Perumal (2010) worked on mixture cure model by estimating the cure fraction and comparing it with other approaches like Kaplan-Meier method which was used for event free survival probability. Lognormal distribution for survival time was used to estimate both the cure fraction and the survival function for the uncured. They concluded that PH conditions are violated, analysis using a non PH model is advocated and mixture cure models are useful in estimating the cure fraction and constructing survival curves for non-cures.

Yu Gu and Banerjee (2010) developed a new universal forms of cure rate models using a proportional likelihoods arrangement where the models preserved all the benefits of stable proportional odds model designed for survival data analysis. This model correspondingly derived proportional odds model using cure rate initiating from the latent factors model, which incorporated several features of the earlier model by Cooner et al (2007). Yu Gu and Banerjee (2010) also established techniques for model selection criteria in the Bayesian framework where attention is at censoring as a result for comparison for models' performance with the criteria of Posterior predictive loss via Markov Chain Monte Carlo samples. Breast cancer data was used to validate the model and the results revealed that their proportional odds model with cure rate was adequately fitted compared to the other two competing models.

Wenbin (2010) studied the cure fraction model with accelerated failure time using the parameter estimation techniques established for kernel nonparametric maximum likelihood and the expectation maximization (EM) algorithm. This techniques was established to evaluate estimates of both the regression parameters as well as the unidentified error density in which kernel-smooth conditional profile probability was maximized in the M-step.He showed that the resulting estimates were consistent and asymptotically normal with a proper selection of the kernel bandwidth parameter. By inverting the empirical Fisher information matrix obtained from the probability of the profile using the EM algorithm, the asymptotic covariance matrix could be consistently estimated.

Aljawadi *et.al*. (2012) developed a Bounded Cumulative Hazard (BCH) model that is more appropriate than a mixture cure model whenever there are long-term survivors or cured in the interest population and proposed this cure rate model based on Weibull distribution with censored interval data. The technique for calculating the maximum likelihood estimation (MLE) was proposed to estimate the parameters throughout the algorithm of expectation-maximization (EM). The method of Newton Raphson has also been used. Their result showed that the cure fraction cannot be obtained analytically, but may be obtained from the numerical solution of the estimated equations. A simulation study was also provided for assessing the efficiency of the proposed estimation procedure.

In the presence of cure fraction, censored data and covariates, Achcar, Coelho-Barros and Mazucheli (2012) introduced Weibull distributions. They explored mixture and non-mixture models, and inferences were obtained using standard MCMC (Markov Chain Monte Carlo) methods for the proposed models under the Bayesian approach. A lifetime data set was used to illustrate the proposed methodology.

Datta (2013) introduced cure rate models to deal with survival models and compared the PH cure rate model performance with case weights to the standard unweighted PH cure rate model through simulation studies. Results of these studies suggest that when the sample size is relatively small, adding case weights in the PH cure rate model improves the estimation of the latency parameter.

Modern medical treatments have significantly improved cure rates for many chronic diseases and have generated increased interest in adequate statistical models for handling non-negligible cure fractions of survival data (Chao, 2013). The mixture cure models are therefore designed to model such data set, which assumes that the population being studied is a mixture of those cured and uncured. This led to Chao’s contribution (2013), who developed two programs in R called smcure and NPHMC. The first program aims to facilitate the estimation of two popular models of mixture cure: the model of proportional hazards (PH) mixture cure and the model of accelerated failure time (AFT). The second program focuses on designing the sample size required by the PH mixture cure model and standard PH model in survival trials with or without cure fractions. Extensive simulation settings and real data sets had been used to evaluate the two programs (Chao,*et.al.*2015). R CRAN (https:/cran.r-project.org/web/packages/NPHMC/NPHMC.pdf) is currently available for download. Ortega,*et.al.* (2014) recently introduced a new generalised distribution of binomial gamma. This is a flexible survival ratio model for cure. The underlying assumption with the model is that the number of competing causes of the event of interest following the negative binomial distribution also follows a generalised gamma distribution.

Adekanbi *et.al.* (2014) reported studies in Ibadan, Nigeria. the long-term survival of ovarian cancer patients in their study. The survival period was estimated to be a post-operative intervention of 250 weeks. This time was shorter than those of other studies and thus the shortest-term survivor discovered to be 120 weeks. The deficiencies of their study were the fact that the suboptimal surgery was performed by some patient. Also, since the operations were performed outside UCH, the magnitude of the operation could not be determined. Another unmatched co-founder was the varying skill levels among the surgeons; in addition, the death record dates were crude estimates.

Taweab *et.al.* (2015) investigated a cure fraction survival model as well as change-point effect established on the model for bounded cumulative hazard (BCH). They used maximum likelihood method to estimate the unknown parameters and the key difficulty in their work was that the likelihood function was not differentiable with respect to a change point parameter. Thus, a smoothed likelihood approach was proposed to address this problem. A simulation study was conducted to evaluate the efficacy of the estimators under various practical situations. Numerical results exhibited the satisfying performance of the proposed estimates and that the proposed model represents a useful addition to the literature of the BCH model.

Hsu *et.al.* (2016) confirmed that recently, in oncology studies, the appraisal of cure fractions underneath prominent cure rate model contributes a substantial recognition now within the past studies amidst literatures, however, ultimate prevailing testing techniques decided on restrictive assumptions, thus they extended previous studies by improving detection of cure fraction models with a score based techniques integrating covariate data including the prevailing process aiding as distinct case. However, the limitations of this extension lead to the fact that the hypotheses definition are not similar and the testing conditions may not hold as well as conditions of standard regularity might be complicated. By means of empirical findings, they constructed a sup-score test statistic for cure fractions and proved its limiting null distribution as an efficient mixture of chi-square methods.

Elangovan and Jayakumar (2016) claimed the advantages of models of cure rates over conventional survival analytical methods, including the well-known Cox regression model which assumes that no patient is cured but that all remain at risk of death or relapse in which they are concerned with survival only and do not accommodate the possibility of cure. Whereas in some types of cancer, like choriocarcinoma, breast and leukemia, a substantial number of patients may now be cured by treatment, i.e., cured proportion. The patients who are cured are called immunes or long-term survivors, while the remaining patients who develop a recurrence of the diseases are termed susceptible. In Cure rate models, the subjects of interest are therefore split into two forms i.e. cured subjects and uncured subjects. This model provides suitable methods in such scenario.

Gallardo *et.al.* (2017) launched the new WeibullYuleSimon distribution. Estimation of maximum probability for model parameters is performed. The estimation approach involved a small-scale simulation study indicating satisfactory parameter recovery. Results are applied to a real data set (melanoma) demonstrating the implication that in terms of model fitting, the proposed model can outperform conventional alternative models.

2.6 Reviews on Gamma and Generalised Gamma Family

The gamma function was firstly introduced by the Swiss mathematician Leonhard Euler (1707 − 1783). This was reported by Sebah and Gourdon (2002) in their review. It is notable that we need to generalise the factorial to non-integer values. Later, because of its extra ordinary significance, it was studied by other eminent mathematicians like Adrien−Marie Legendre (1752−1833), Carl Friedrich Gauss (1777 − 1855), Christoph Gudermann (1798 − 1852), Joseph Liouville (1809 − 1882), Karl Weierstrass (1815 − 1897), Charles Hermite (1822 − 1901), as well as many others. Roynette, et.al. (2009) described a family of generalised gamma convoluted (abbreviated as GGC) variables where they were able to prove that several random variables, related to the length of excursions away from 0 for a recurrent linear diffusion on *R*+, are GGC.

Zografos-Balakrishnan-G (2009) and Ristic-Balakrishnan−G (2011) studied some mathematical properties and presented special sub-models with an extra positive parameter which provides a comprehensive treatment of general mathematical properties of Zografos−Balakrishna distributions.

Barriga *et.al.* (2018) asserted that several attempts have been made to define new classes of distributions that provide more flexibility for modelingskewed data in practice. In their study, they defined a new extension of the generalised gamma distribution where their new lifetime model is very flexible distribution and was to fit real data from several fields, such asengineering, hydrology and survival analysis.
Balakrishnan and Pal (2015) measured the cure rate and assuming the time-to-event to follow the gamma distribution, they developed exact likelihood inference based on the EM algorithm. Their study employed an extensive Monte Carlo simulation study to discover the technique of inference developed. Model discrimination between different cure rate models is carried out by means of likelihood ratio test and Akaike and Bayesian information criteria.

Other parametric cure fraction models have been studied in literatures among which are Destructive weighted Poisson cure rate models which was done and discussed by Rodrigues et.al. (2009). Left truncated and right censored Weibull data and likelihood inference with an illustration have been discussed by Balakrishnan et.al. (2012). EM algorithm-based likelihood estimation for some cure rate models was discussed also discussed by Balakrishnan et.al. (2012).

Lognormal lifetimes and likelihood-based inference for flexible cure rate models based on COMPoisson family was discussed by Balakrishnan and Pal (2013). Expectation maximization-based likelihood inference for flexible cure rate models with Weibull lifetimes is discussed by Balakrishnan and Pal (2013).

An EM algorithm for the estimation of parameters of a flexible cure rate model with generalised gamma lifetime and model discrimination using likelihood-and information-based methods has been discussed by Balakrishnan and Pal (2015). Latent cure rate model under repair system and threshold effect have been discussed by Balakrishnan, *et. al.*(2015).

CHAPTER THREE
METHODOLOGY

3.1 Study Site

The site of this study is the University of Ibadan, Teaching Hospital at the Department of Obstetrics and Gynaecology. This is one of Nigeria's and West Africa's leading tertiary cancer therapy center. That Department in University College Hospital (UCH) accepts medical transfers from all states within the country and at large. For a substantial percentage of patients with cancer, they mark it as the right choice. The Gynaecology Oncology Unit (GOU) of UCH manages Patients with ovarian cancer included in this study.

3.2 Study Design

Life ovarian cancer data was obtained from Department of Obstetrics and Gynaecology, University College Hospital, Ibadan, Nigeria covering the period 2000-2015. The diagnosis time (in months) was until death. This study is retrospective in nature as well as descriptive non-intervention. It comprises the usage of accessible facts from ovarian cancer patient's medical registers including the procedure of follow-up of patient's status by means of contacting and communicating with the patients or their relatives in order to regulate the patients' health status should in case there might be missing observations and uncertain facts originating from accessible registers. The telephone method was also performed as a follow-up to the members of the family of the patient whenever the need came to light.

3.3 Overview of Cure Model

A patient who has survived for five years after a cancer diagnosis is not necessarily medically cured but is considered statistically cured because the five−year relative survival analysis is considered a good indication that the cancer is responding to treatment and that the treatment is successfully extending the life of the cancer patient. The survival figures so obtained are utilized in choosing treatment types and regimes, doses, in discriminating between the side effects profiles and cost effectiveness.

Cure models connotes the distinct survival models established to evaluate cure rate in cancer research. The estimation of the probability of survival as well as the cured proportion is done with the aid of this model.The model is split into:

Mixture cure model: This particular form of cure model is developed to evaluate the percentage of cured patients including percentage of the population in uncured patients ' survival function [Boag (1949) and Berkson and Gage (1952)].

Non-Mixture Cure Model: The Bounded Cumulative Hazard (BCH) can be portrayed as a model of non-mixture cure fraction. (Yakovlev et. al., 1993)

3.4 Importance of measuring Statistical Cure

In cancer studies, there was some necessary significance of statistical cure among which are listed below:

- The survival of patients is paramount and the most relevant queries.
- The model gives additional facts about the survival of patient.
- Cure models calculate the percentage of cured patients from cancer.
- The model predicts the time until cured is achieved.
- The model is used for estimation of uncured patient’s survival time.
- The patient’s “Net survival” estimate is not influenced by other cause mortality.

3.5 Ethical Consideration

The results from this study will not be generally accepted without the ethical approval. The Institute for Advanced Medical Research and Training (IAMRAT) granted this ethical approval, College of Medicine, University of Ibadan, Nigeria. UI/UCH EC gave full approval for an ethical consideration of benefits of subjects, their details on non-maleficence, fairness and privacy had been adhered strictly to when managing the study subjects and in administration of data. During data entry, observations were numbered serially; statistical analysis and study reports are to be defended even when publishing so as to safeguard the patients’ documentation as anonymous. This ethical approval is to have access to patient treatment and to collect real-life data on ovarian cancer along with other types of gynaecological cancer. (The copy of the ethical approval is attached and can be found in Appendix: UI/UCH EC approval by the IAMRAT.)

3.6 The Model: Mixture Cure Model

Boag (1949) was the first author that developed this model and it was revised by Berkson & Gage (1952) as:

$$S(t) = c + (1 - c)Su(t)(3.6.1)$$

Where

- $S(t)$ = Population Survival function
- $Su(t)$ = Uncured patients Survival function
- c =Proportion cured

In a mixture cure model, the population involves two components where those with probability $P(1 - c)$ are uncured component and $P(c)$ are cure component.

3.6.1 Estimations Procedure of Mixture Cure Model (MCM)

A technique used for parameter estimation in this study is Maximum Likelihood Estimation (MLE) method. Conventionally, estimations of mixture cure model parameters can either follow parametric or non-parametric techniques. The present study goes through parametric system of estimation. Considering the given equation (3.6.1) above, parameter c of the mixture cure model can be acquired as follow:

$$c = \frac{S(t) - S_u(t)}{1 - S_u(t)}(3.6.2)$$

3.6.2 Estimation of MCM's Likelihood Function.

In MCM, the likelihood function is given as follow:

$$L=\prod_{i=1}^{n}[f(t_i)]^{di}\left[S(t_i)\right]^{1-di} \qquad (3.6.3)$$

Where $d_i = \begin{cases} 0, & if\ patient\ is\ cured \\ 1, & otherwise \end{cases}$

d_i is 0 if the patient is cured and 1 if the patients is not cured.

$f(t) = pdf\ of\ parametric\ cure\ model$

$$S(t_i) = survival\ function$$

Recall that S(t) = 1 – F(t), hence equation (3.6.1) becomes;

$$1 - F(t) = c + (1-c)\big(1 - F_u(t)\big)(3.6.4)$$

Differentiating(3.6.4), $we\ have$;

$$f(t) = \big((1-c)f_u(t)\big)(3.6.5)$$

Put (3.6.5) in equation (3.6.3) to obtain

$$L=\prod_{i=1}^{n}\left[(1-c)f_u(t)\right]^{di}\left[c+(1-c)S_u(t)\right]^{1-di} \qquad (3.6.6)$$

The log-likelihood function of mixture cure model can be gotten with the aid of taking logarithms of equation (3.6.6) as follow:

$$LogL=\sum_{i=1}^{n}d_i(1-c)+d_i\log f_u(t)+\sum_{i=1}^{n}(1-d_i)\log\left[c+(1-c)S_u(t)\right] \qquad (3.6.7)$$

3.7 Review of Some Existing Parametric Mixture Cure Fraction Models

Several MCM amidst Lognormal MCM (LNMCM), Loglogistics MCM (LLMCM), Weibull MCM (WMCM) and Generalised-Gamma MCM (GGMCM) have been used to study cure rates in epidemiology. In MCM, estimation of corresponding c that is patients cure proportion, median time-to-cure and variances will be estimated from the examined distributions mentioned earlier. Some transformation were done in order to incorporate the distribution in cure fraction models, thus each parametric distribution were embedded in cure fraction models, which transformed a distribution to a model. The existing parametric cure models considered are reviewed below:

3.7.1 The LogLogistic MCM

The probability density function that is Loglogistic distributed is defined as:

$$f(t;\alpha,\beta)=\frac{\frac{\alpha}{\beta}\left(\frac{t}{\beta}\right)^{\alpha-1}}{\left[\left(1+\frac{t}{\beta}\right)^{\alpha}\right]^{2}};\ \alpha,\beta>0 \qquad (3.7.1)$$

with $\beta = \lambda^{-1}$ and $\lambda = \frac{1}{\beta}$; then we have

$$f(t)=\frac{\alpha\lambda(\lambda t)^{\alpha-1}}{\left(1+\lambda t\right)^{2\alpha}} \qquad (3.7.2)$$

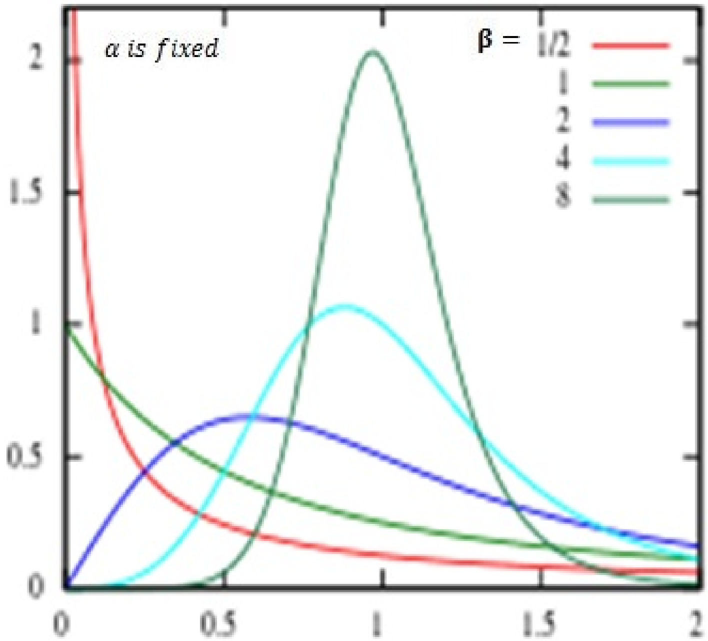


Fig 3.1: Density function of Loglogistic distribution

The above Fig 3.1 is the density function of loglogistic distribution. The figure shows the shapes of the distribution on how far it is from normal curve.

All the different values shown in the graph are the varying values of the shape parameters and it was discovered that this valueincreases; log-logistic distribution approaches normal distribution. This is clearly seen that when the value increases 8, the curve approaches a bell shape.

This is in agreement with the concept of "Central limit theorem" in the theory of classical inference which state that "as the sample size n becoming large, then the population of all possible sample statistic is approximately normal regardless of what probability distribution that describe the sample population

Using the transformation $\mu = -\log\lambda$ which implies $\lambda = e^{-\mu}, \sigma = \frac{1}{\alpha}$ and $y = \log t \Rightarrow \exp^{y} = \mathrm{t}$

$$f(t;\mu,\sigma) = \frac{\frac{1}{\sigma}\left(e^{\frac{\log t-\mu}{\sigma}}\right)^{1-\sigma}}{\left[\left(e^{\frac{\log t-\mu}{\sigma}}\right)\right]^{2}} \quad \mu,\sigma,t > 0 \qquad (3.7.3)$$

$$= \frac{\frac{1}{\sigma}(e^{y-\mu})^{\frac{1}{\sigma}-1}}{\left[1+\left(e^{\frac{y-\mu}{\sigma}}\right)\right]^{2}} (3.7.4)$$

$$f(t;\mu,\sigma) = \frac{\frac{1}{\sigma}\left(e^{\frac{\log t-\mu}{\sigma}}\right)^{1-\sigma}}{\left[1+\left(e^{\frac{\log t-}{\sigma}}\right)\right]^{2}} (3.7.5)$$

The Cumulative Density Function (cdf) can be achieved using:

$$P(T \le t) = F(t) = \int_{0}^{t} f(x)dx$$

$$P(T \le t) = F(t) = \int_0^t \frac{\alpha\lambda(\lambda x)^{\alpha-1}}{[1+(\lambda x)^{\alpha}]^2}dx \qquad (3.7.6)$$

$$= \alpha\lambda\int_0^t \frac{(\lambda x)^{\alpha-1}}{[1+(\lambda x)^{\alpha}]^2}dx \qquad (3.7.7)$$

Let

$$y = (\lambda x)^{\alpha}, \frac{dy}{dx} = \alpha\lambda^{\alpha}x^{\alpha-1} \Rightarrow dx = \frac{dy}{\alpha\lambda^{\alpha}x^{\alpha-1}}$$

And

$x = 0 \Leftrightarrow y = 0, x = t \Leftrightarrow (\lambda t)^{\alpha} = \mathrm{y}$

$$F(t) = \alpha\lambda \int_{0}^{(\lambda t)^{\alpha}} \frac{y\lambda^{-1}x^{-1}}{(1+y)^2} \frac{dy}{\alpha\lambda^{\alpha}x^{\alpha}x^{-1}}$$

$$= \alpha\lambda \int_{0}^{(\lambda t)^{\alpha}} \frac{y\lambda^{-1}x^{-1}.dy}{(1+y)^2\alpha y x^{-1}} = \int_{0}^{(\lambda t)^{\alpha}} \frac{1}{(1+y)^2} dy$$

$$= \int_{0}^{(\lambda t)^{\alpha}} (1+y)^{-2} dy$$

$$= -\frac{1}{1+(\lambda t)^{\alpha}} - -\frac{1}{1+(0)}$$

$$= 1 - \frac{1}{1+(\lambda t)^{\alpha}}$$

$$F(t) = 1 - [1+(\lambda t)^{\alpha}]^{-1} (3.7.8)$$

Therefore, the survival function is now obtained using

$$S(t) = 1 - F(t) (3.7.9)$$

$$= [1+(\lambda t)^{\alpha}]^{-1} (3.7.10)$$

Using the transformation $\mu = -\log\lambda$ which implies $\lambda = e^{-\mu}, \sigma = \frac{1}{\alpha}$ and $y = \log t \Rightarrow \mathrm{e}^{y} = t$

Then,

$$S(t;\mu,\sigma) = \left[1 + (e^{-\mu}e^{y})^{\frac{1}{\sigma}}\right]^{-1} (3.7.11)$$

$$= \left(1 + e^{\frac{y-\mu}{\sigma}}\right)^{-1}$$

$$= \left[1 + e^{\frac{\log t - \mu}{\sigma}}\right]^{-1} (3.7.12)$$

$$G(Z) = (1+e^{Z})^{-1} (3.7.13)$$

Where $Z = \frac{\log t -}{\sigma}$

Using equation (3.6.2), the $\hat{c}$ for logistic becomes

$$\text{clog}\log istic \ = 1 - \frac{1 - s(t_i)}{1 - \left[1 + e^{\frac{\log t - \mu}{\sigma}}\right]^{-1}} \text{(3.7.14)}$$

3.7.2 Lognormal MCM

The probability density function (pdf) that is lognormal distributed is defined as:

$$f(t; \mu, \sigma) = \frac{1}{t\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\log t -}{\sigma}\right)^2} t > 0, \mu > 0, \sigma > 0 \qquad (3.7.15)$$

The Cumulative Density Function (cdf) is obtained following the next procedures:

$$P(T \leq t) = F(t) = \int_0^t f(x)dx$$

$$= \int_0^t \frac{1}{x\sigma\sqrt{2\pi}} e^{\frac{-1}{2}\left(\frac{\log x -}{\sigma}\right)^2} dx \qquad (3.7.16)$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_0^t \frac{1}{x} e^{-\frac{1}{2}\left(\frac{\log x - \mu}{\sigma}\right)^2} dx$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_0^{\log t} e^{-\frac{1}{2}\left(\frac{\log x -}{\sigma}\right)^2} dx$$

$$Let \ z = \log x$$

$$dz = \frac{1}{x} dx$$

$$\therefore = \frac{1}{\sigma\sqrt{2\pi}} \int_0^{\log t} e^{-\frac{1}{2}\left(\frac{z - \mu}{\sigma}\right)^2} dz$$

$$\Phi\left(\frac{z - \mu}{\sigma}\right)$$

From the above, the survival function can then be gotten/obtained using

$$S(t) = 1 - F(t)$$

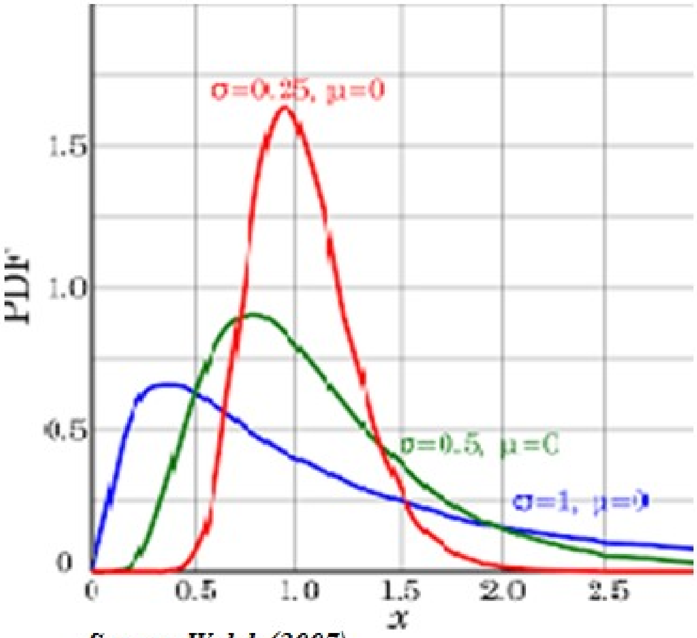
$$S(t; \lambda, \sigma) = 1 - \Phi\left(\frac{\log t - \mu}{\sigma}\right) \qquad (3.7.18)$$

where $Z = \left(\frac{\log t -}{\sigma}\right)$, we have

$$G^*(Z) = 1 - \Phi(Z)$$

Using equation (3.6.2), the $\hat{c}$ for log normal becomes

$$c \ \ logNormal == 1 - \frac{1 - s(t_i)}{\Phi\left(\frac{\log t -}{\sigma}\right)} \text{(3.7.19)}$$



Source: Walck (2007)

Figure 3.2: Density function of Lognormal distribution

In the figure 3.2 above shows the log-normal distribution for the basic form, with $\mu = 0$ and $\sigma = 1$, where the variable $x > 0$ and the parameters $\mu$and $\sigma > 0$ all are real numbers. It is sometimes denoted $\Lambda(\mu, \sigma^2)$ and can also be denoted as normally distributed variable by $N(\mu, \sigma^2)$.

3.7.3 The Weibull MCM

The probability density function (pdf) that is Weibull distributed is defined as:

$$f(t) = \frac{\alpha}{\beta^{\alpha}} t^{\alpha-1} e^{-\left(\frac{t}{\beta}\right)^{\alpha}} \qquad t > 0, \alpha > 0, \beta > 0 \tag{3.7.20}$$

The Cumulative Density Function is then obtained using

$$P(T \leq t) = F(t) = \int_0^t f(x)dx$$

$$= \int_0^t \frac{\alpha}{\beta^{\alpha}} x^{\alpha-1} e^{-\left(\frac{x}{\beta}\right)^{\alpha}} dx$$

$$= \frac{\alpha}{\beta^{\alpha}} \int_0^t x^{\alpha-1} e^{-\left(\frac{x}{\beta}\right)^{\alpha}} dx$$

Let $y = \left(\frac{x}{\beta}\right)^{\alpha}$

$$\frac{dy}{dx} = \frac{\alpha x^{\alpha-1}}{\beta^{\alpha}} \Longrightarrow \frac{\beta^{\alpha} dy}{\alpha x^{\alpha-1}}$$

Hence,

$$F(t) = \frac{\alpha}{\beta^{\alpha}} \int_0^{\left(\frac{t}{\beta}\right)^{\alpha}} x^{\alpha-1} \frac{e^{-y}\beta^{\alpha}}{\alpha(x)^{\alpha-1}} dy$$

$$= \int_0^{\left(\frac{t}{\beta}\right)^{\alpha}} e^{-y} dy$$

$$= -e^{-\left(\frac{t}{\beta}\right)^{\alpha}} - -e^{0}$$

$$F(t) = -e^{-\left(\frac{t}{\beta}\right)^{\alpha}} + 1$$

$$F(t) = 1 - e^{-\left(\frac{t}{\beta}\right)^{\alpha}} \tag{3.7.21}$$

Therefore, the survival function can be obtained using:

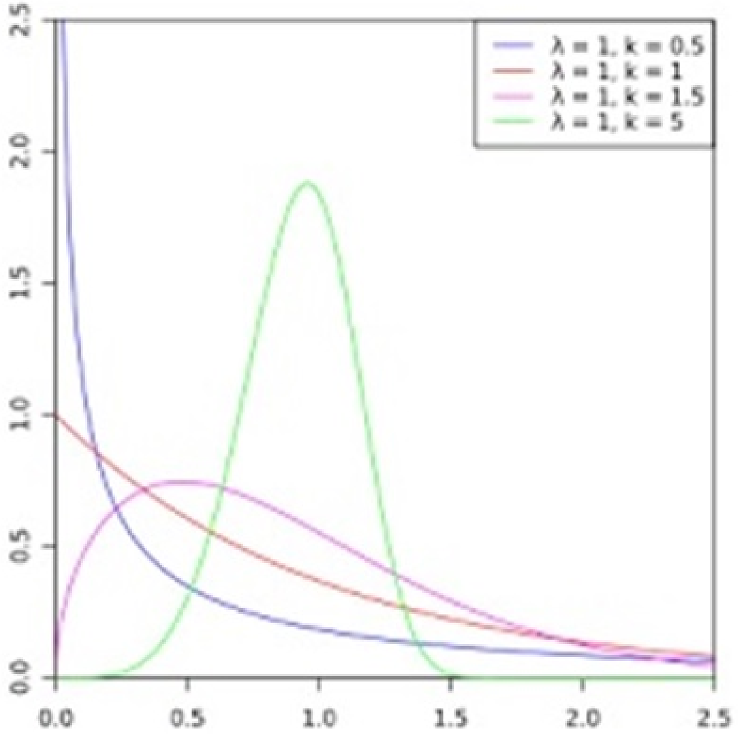
$$P(T > t) = S(t) = 1 - F(t)$$

$$S(t) = 1 - \left(1 - e^{-\left(\frac{t}{\beta}\right)^{\alpha}}\right)$$

$$S(t) = e^{-\left(\frac{t}{\beta}\right)^{\alpha}} \tag{3.7.22}$$

Using equation (3.6.2), the $\hat{c}$ for Weibull becomes

$$\text{cWeibull } = 1 - S(t) = 1 - e^{-\left(\frac{t}{\beta}\right)^{\alpha}} \tag{3.7.23}$$



Source: Walck (2007)

Figure 3.3: Weibul Distribution Density Function

The above fig 3.3 is the density function of Weibull distributionwhere$\lambda\ = 1$ that is, it is fixed, meaning that the scale parameter is fixed, and the shape parameter k is varying. All the different values shown in the graph are the varying values of k and it was discovered that as parameter k increases Weibull distribution approaches normal distribution.

3.7.4 The Generalised-Gamma MCM

The distribution of generalised-gamma probability density function is defined as:

$$f(t) = \frac{\beta}{\theta\Gamma(\alpha)}\left(\frac{t}{\theta}\right)^{\alpha\beta-1} e^{-\left(\frac{t}{\beta}\right)^{\alpha}} \qquad t > 0, \qquad \theta > 0, \alpha > 0, \qquad \beta > 0 \qquad (3.7.24)$$

Given $\theta > 0$ as scale parameter, then shape parameters are $\beta > 0\ and\ \alpha > 0$ and the gamma function of x is given as $\Gamma(x)$.

The cumulative density function can also be obtained using:

$$P(T \leq t) = F(t) = \int_0^t \frac{\beta}{\theta\Gamma(\alpha)}\left(\frac{y}{\theta}\right)^{\alpha\beta-1} e^{-\left(\frac{y}{\theta}\right)^{\beta}} dy$$

If we let $m = \left(\frac{y}{\theta}\right)^{\beta}\ y^{\beta} = m\theta^{\beta}$

$$y = m^{\frac{1}{\beta}\theta}$$

$$\frac{dm}{dy} = \frac{\beta y^{\beta-1}}{\theta^{\beta}}$$

$$dy = \frac{\theta^{\beta} dm}{\beta y^{\beta-1}} = \frac{\theta^{\beta} dm}{\beta m^{1-\frac{1}{\beta}}\theta^{\beta-1}}$$

$$F(t) = \int_0^{\left(\frac{t}{\theta}\right)^{\beta}} \frac{\beta}{\theta\Gamma(\alpha)}\left(m^{\frac{1}{\beta}}\right)^{\alpha\beta-1} e^{-m} \frac{\theta^{\beta} dm}{\beta m^{1-\frac{1}{\beta}}\theta^{\beta-1}}$$

$$F(t) = \frac{1}{\Gamma(\alpha)}\int_0^{\left(\frac{t}{\theta}\right)^{\beta}} \frac{m^{\alpha-\frac{1}{\beta}} e^{-m} dm}{m^{1-\frac{1}{\beta}}}$$

$$F(t) = \frac{1}{\Gamma(\alpha)}\int_0^{\left(\frac{t}{\theta}\right)^{\beta}} m^{\alpha-1} e^{-m} dm$$

$$F(t) = \frac{\gamma(\alpha, m)}{\Gamma(\alpha)}$$

$$F(t) = \frac{\gamma\left[\alpha, \left(\frac{t}{\theta}\right)^{\beta}\right]}{\Gamma(\alpha)}$$

The survival function can then be obtained using:

$$S(t) = 1 - F(t)$$

$$= 1 - \frac{\gamma\left[\alpha, \left(\frac{t}{\theta}\right)^{\beta}\right]}{\Gamma(\alpha)}$$

With $y = \log t \implies t = e^{y}$and $\mu = log\theta \implies \theta = e^{\mu}$

Also $\beta = \frac{1}{\sigma}$, we have

$$S(t^{1}, \mu, \sigma) = 1 - \frac{\gamma\left[\alpha, \left(\frac{e^{y}}{e^{\mu}}\right)^{\frac{1}{\sigma}}\right]}{\Gamma(\alpha)}$$

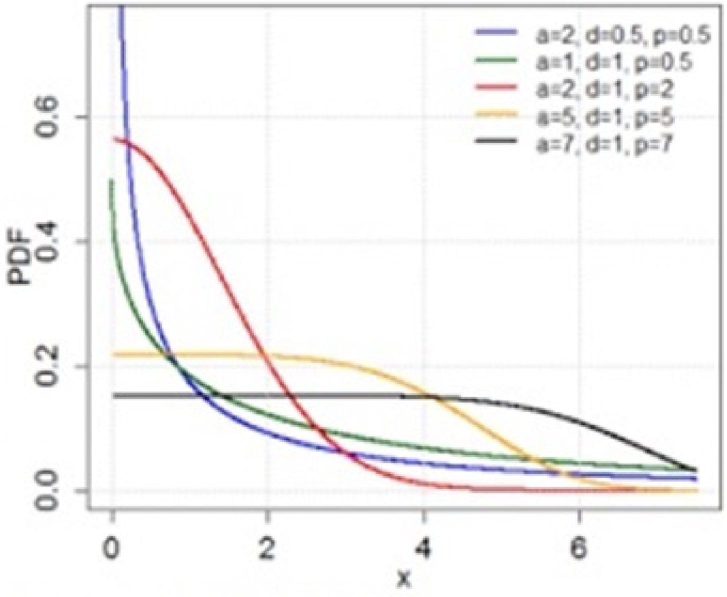
$$S(t^{1}, \mu, \sigma) = 1 - \frac{\gamma\left[\alpha, e^{\frac{y-\mu}{\sigma}}\right]}{\Gamma(\alpha)}$$

$$S(t^{1}, \mu, \sigma) = 1 - \frac{\gamma\left[\alpha, e^{\frac{log\, t-\mu}{\sigma}}\right]}{\Gamma(\alpha)}$$

where $Z = \frac{log\, t-\mu}{\sigma}$, we have

$$G(Z) = 1 - \frac{\gamma[\alpha, e^{Z}]}{\Gamma(\alpha)}$$

$$G(Z) = 1 - \frac{\gamma\left[\alpha, e^{\frac{log\, t-\mu}{\sigma}}\right]}{\Gamma(\alpha)}$$



Source: Walck (2007)

Figure 3.4: Density Function of Generalised-Gamma Distribution

The above fig 3.4 is the density function of generalised-gamma distribution. The figure demonstrates the shapes of the distribution on how far it is from normal curve.

3.8 Parameter Estimations and Inferences

3.8.1 Parameter Estimations

The key three parameters that were obtained from transformation earlier discussed, that is, $\mu, \sigma, and\ c$ will be estimated. Differentiating the baseline distribution in equation 3.6.7 that is the cure model log-likelihood functions with respect to the parameters.

Log Likelihood Functions of Lognormal MCM

Given the density function and survival function of lognormal distribution, and plugging them into the log likelihood functions of cure models in equation 3.6.7, it gives equation 3.8.1 below:

$$f_u(t) = \frac{1}{t\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\log t-\mu}{\sigma}\right)^2}$$

$$S_u(t) = 1-\Phi\left(\frac{\log t-\mu}{\sigma}\right)$$

$$Log\ L = \sum_{i=1}^{n} d_i \log(1-c) + \sum_{i=1}^{n} d_i\ log\ f_u(t_i) + \sum_{i=1}^{n}(1-d_i)\log\left[c+(1-c)S_u(t)\right]$$

$$= \sum_{i=1}^{n} d_i \log(1-c) + \sum_{i=1}^{n} d_i \log\frac{1}{t\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\log t-}{\sigma}\right)^2} + \sum_{i=1}^{n}(1$$

$$- d_i)\log\left[c+(1-c)\left(1-\Phi\left(\frac{\log t-\mu}{\sigma}\right)\right)\right]$$

$$= \sum_{i=1}^{n} d_i \log(1-c) + \sum_{i=1}^{n} d_i \left[\log\left(t_i\sigma\sqrt{2\pi}\right)^{-1} - \frac{1}{2}\left(\frac{\log t-\mu}{\sigma}\right)^2\right] + \sum_{i=1}^{n}(1-d_i)\log\left[c\right.$$

$$\left. + (1-c)\left(1-\Phi\left(\frac{\log t-\mu}{\sigma}\right)\right)\right]$$

$$Log\ L = \sum_{i=1}^{n} d_i \log(1-c) + \sum_{i=1}^{n} d_i \log\left(t_i\ \sigma\sqrt{2\pi}\right) - \frac{1}{2}\sum_{i=1}^{n} d_i \left(\frac{\log t-\mu}{\sigma}\right)^2$$

$$+ \sum_{i=1}^{n}(1$$

$$- d_i)\log\left[c+(1-c)\left(1-\Phi\left(\frac{\log t-\mu}{\sigma}\right)\right)\right] \qquad (3.8.1)$$

Differentiating with respect to parameters $\mu, \sigma\ and\ c,$ we then have equations 3.8.2, 3.8.3, and 3.8.4.

$$\frac{\partial \log L}{\partial \mu} = \frac{1}{2\sigma^2} 2 \sum_{i=1}^{n} d_i \left(\log t_i - \mu\right) = 0$$

$$= \sum_{i=1}^{n} d_i \left(\log t_i - \hat{\mu}\right) = 0 \qquad (3.8.2)$$

$$\frac{\partial \log L}{\partial \sigma} = -\sum_{i=1}^{n} \frac{d_i t_i \sqrt{2\pi}}{t_i \sigma \sqrt{2\pi}} + \frac{1}{2\sigma^3} \ \sum_{i=1}^{n} d_i \left(\log t_i - \mu\right)^2$$

$$+ \sum_{i=1}^{n} (1 - d_i) \frac{\frac{\partial}{\partial \sigma}\left[c + (1-c)\left(1 - \Phi\left(\frac{\log t - \mu}{\sigma}\right)\right)\right]}{\left[c + (1-c)\left(1 - \Phi\left(\frac{\log t - \mu}{\sigma}\right)\right)\right]}$$

$$\frac{\partial \log L}{\partial \sigma} = -\frac{1}{\sigma} \sum_{i=1}^{n} d_i + \frac{1}{2\sigma^3} \ \sum_{i=1}^{n} d_i \left(\log t_i - \mu\right)^2 \qquad (3.8.3)$$

$$\frac{\partial \log L}{\partial c} = -\sum_{i-1}^{n} \frac{d_i}{(1-c)} + \sum_{i=1}^{n} 1 - \Phi\left(\frac{\log t - \mu}{\sigma}\right)$$

$$\frac{\partial \log L}{\partial c} = -\frac{1}{1-c} \sum_{i=1}^{n} d_i + \sum_{i=1}^{n} \Phi\left(\frac{\log t - \mu}{\sigma}\right) \qquad (3.8.4)$$

Equations 3.8.2, 3.8.3 and 3.8.4 do not have a close form solution; that is, the parameters can only be solved with numerical approach.

Log likelihood functions of Weibull MCM

Given the density function and survival function of Weibull distribution and plugging them into the log likelihood functions of cure models in equation 3.6.7, it gives equation 3.8.5 below:

$$f_u(t) = \frac{\alpha}{\beta^{\alpha}} t^{\alpha-1} e^{-\left(\frac{t}{\beta}\right)^{\alpha}}$$

Let $\alpha = \dfrac{1}{\sigma}$, $\beta = \dfrac{1}{\lambda}$, $t = e^{y}$ and $y = \log t$

$$f_u(\mathrm{y}) = \frac{1/\sigma}{\left(1\!\!\diagup\!\lambda\right)^{1/\sigma}} \left(e^{y}\right)^{1/\sigma-1} e^{-\left(\frac{e^{y}}{1\!\diagup\!\lambda}\right)^{1\!\diagup\!\sigma}}$$

$$S(t) = e^{-\left(\frac{t}{\beta}\right)^{\alpha}}$$

$$S(y) = e^{-\left(\frac{e^{y}}{1\!\diagup\!\lambda}\right)^{1\!\diagup\!\sigma}}$$

$$S(y) = e^{-\left(\lambda e^{y}\right)^{1\!\diagup\!\sigma}}$$

Let $\lambda = e^{-\mu} \Rightarrow \mu = -\log\lambda$

$$\Rightarrow S(y) = e^{-\left(e^{-\mu} e^{y}\right)^{1\!\diagup\!\sigma}}$$

$$S(y) = e^{-\left(e^{\frac{y-\mu}{\sigma}}\right)}$$

$$F(t) = 1 - e^{-\left(\frac{t}{\beta}\right)^{\alpha}}$$

$$\Rightarrow F(y) = 1 - e^{-\left(e^{-\mu} e^{y}\right)^{1\!\diagup\!\sigma}}$$

$$\Rightarrow F(y) = 1 - e^{-\left(e^{\frac{y-\mu}{\sigma}}\right)}$$

Let $m = e^{\frac{y-\mu}{\sigma}}$

$$F(\mathrm{y}) = 1 - e^{-m}$$

$$\frac{dF(y)}{dm} = e^{-m} \frac{dm}{dy} = \frac{1}{\sigma} e^{\frac{y-\mu}{\sigma}}$$

$$\frac{dF(y)}{dm} = \frac{1}{\sigma} e^{\frac{y-\mu}{\sigma}} e^{-e^{\frac{y-\mu}{\sigma}}}$$

Recall that y = logt

Therefore;

$$
\begin{aligned}
Log\ L = &\sum_{i=1}^{n} d_i \log(1-c) \\
&+ \sum_{i=1}^{n} d_i \log\left[\frac{1}{\sigma} e^{\frac{logt-\mu}{\sigma}-e^{\frac{log}{\sigma}}}\right] \\
&+ \sum_{i=1}^{n} (1-d_i)\ log\ (c+(1-c))\left[e^{-\left(e^{\frac{\log t-\mu}{\sigma}}\right)}\right] \qquad (3.8.5)
\end{aligned}
$$

Differentiating 3.8.5 with respect to parameters $\mu, \sigma\ and\ c,$ we then have (3.8.6), (3.8.7), and (3.8.8)

$$
\frac{\partial \log L}{\partial C} = \frac{-\sum_{i=1}^{n} d_i}{1-C} + \sum_{i=1}^{n}(1-d_i)\left[\left(\frac{1-\left(1+e^{\frac{\log t}{\sigma}}\right)^{-1}}{C+(1-C)\left(1+e^{\frac{\log t-\mu}{\sigma}}\right)^{-1}}\right)\right] \qquad (3.8.6)
$$

$$
\begin{aligned}
\frac{\partial \log L}{\partial \mu} = \sigma \sum_{i=1}^{n} d_i \frac{\frac{\partial}{\partial \mu}\left(e^{\frac{t_i-\mu}{\sigma}} - e^{\frac{t_i-\mu}{\sigma}}\right)}{e^{\frac{t_i-\mu}{\sigma}} - e^{\frac{t_i-\mu}{\sigma}}} + \sum_{i=1}^{n}(1-d_i)\frac{\partial}{\partial \mu}\Big(C + (1 \\
- C)\left(1+e^{\frac{\log t-\mu}{\sigma}}\right)^{-1}\Big) \qquad (3.8.7)
\end{aligned}
$$

$$
\frac{\partial \log L}{\partial \sigma} = \sum_{i=1}^{n} d_i \frac{\frac{\partial}{\partial \mu}\left(e^{\frac{t_i-\mu}{\sigma}} - e^{\frac{t_i-\mu}{\sigma}}\right)}{e^{\frac{t_i-\mu}{\sigma}} - e^{\frac{t_i-\mu}{\sigma}}} + \sum_{i=1}^{n}(1-d_i)\frac{\partial}{\partial \mu}\log\left[C+(1-C)\left(1+e^{\frac{\log t-\mu}{\sigma}}\right)^{-1}\right] \qquad (3.8.8)
$$

From the results above, it is shown that it does not have a close form solution; that is, the parameters can only be solved with numerical approach.

Log likelihood functions of Generalised Gamma MCM

Given the density function and survival function of generalised gamma distribution and plugging them into the log likelihood functions of cure models in (3.6.7), we have (3.8.8) below:

$$
f_u(t) = \frac{\beta}{\theta\Gamma(\alpha)}) \left(\frac{t}{\theta}\right)^{\alpha\beta-1} e^{-\left(\frac{t}{\theta}\right)^{\beta}} t > 0,\ \theta > 0, \beta > 0
$$

$$
F_U(t) = \frac{\gamma\left[\alpha,\left(\frac{t}{\theta}\right)^{\beta}\right]}{\Gamma(\alpha)} S(t,\mu,\sigma) = 1 - \frac{\gamma\left[\alpha, e^{\frac{\log t-}{\sigma}}\right]}{\Gamma(\alpha)}
$$

Therefore;

$$Log\ L = \sum d_i \log(1-C) + \sum d_i\ log\left[e^{(\alpha Z - log\ t_i - e^{Z})}.\frac{1}{\sigma\Gamma(\alpha)}\right]$$

$$+\sum(1-\ d_i)\log\left[C+(1-C)\left[1-\frac{\upgamma(\alpha,\ e^{z})}{\Gamma(\alpha)}\right]\right]$$

$$=\sum d_i\left[\log(1-c)+\left(\frac{1}{\sigma\Gamma(\alpha)}\right)+\alpha\left(\frac{\log t_i-\mu}{\sigma}\right)-logt_i-\exp\left(\left(\frac{\log t_i-\mu}{\sigma}\right)\right)\right]$$

$$+\sum(1-\ d_i)\log\left[c+\ (1-c)\left(1-\frac{\upgamma}{\Gamma(\alpha)}\right)\right] \qquad (3.8.9)$$

Differentiating (3.8.9) with respect to parameters μ, σ and C. we then have (3.8.10), (3.8.11) and (3.8.12)

$$\frac{\partial\ Log\ L}{\partial\ C} = -\ \frac{1}{1-C}\sum d_i + \frac{1}{\Gamma(\alpha)} + \sum(1-\ d_i)\frac{\gamma(\alpha,\ e^{Z})}{C+(1-C)\left[1-\frac{\gamma(\alpha,\ e^{Z})}{\Gamma(\alpha)}\right]} \qquad (3.8.10)$$

$$=-\ (1-c)^{-1}\Sigma d_i + \frac{1}{\Gamma(\alpha)}\Sigma(1-d_i)\gamma(\alpha,e^{Z})\left[c+(1-c)\left[1-\frac{\gamma(\alpha,\ e^{Z})}{\Gamma(\alpha)}\right]\right]^{-1}$$

$$\frac{\partial\ Log\ L}{\partial\ \mu} = \Sigma d_i\left[-\frac{\alpha}{\sigma}-\ e^{\left(\frac{\log t_i-\mu}{\sigma}\right)}.\left(-\frac{1}{\sigma}\right)\right]+\Sigma(1-\ d_i)\left[c+\ (1-c)\left(1-\frac{\gamma}{\Gamma(\alpha)}\right)\right]^{-1}\frac{(-1)(1-C)\gamma\mu}{\Gamma(\alpha)}(3.8.11)$$

$$\frac{\partial\ Log\ L}{\partial\ \sigma}=\frac{-\ \Sigma d_i}{\sigma}\left[\alpha-\ e^{\left(\frac{lo\quad i-\mu}{\sigma}\right)}\right]\frac{-1}{\Gamma(\alpha)}\Sigma(1-\ d_i)(1-c)\,\gamma\mu\left[c+(1-c)\left(1-\frac{\gamma}{\Gamma(\alpha)}\right)\right]^{-1} \qquad (3.8.12)$$

3.9 Hessian Matrix

The derivatives from the second order of equations (3.8.10), (3.8.11) and (3.8.12) produced the Fisher Information Matrix in equation (3.91); the result is the diagonal elements of the Fishers information matrix obtained as follows. The matrix is also known as the Hessian matrix. This is a square matrix of a function's partial second-order derivatives. It describes a function of several parameters in the local curve.

Hessian matrix (or Variance-Covariance) for each parameter:

$$\begin{bmatrix} \frac{\partial^2 \log L}{\partial c^2} & \frac{\partial^2 \log L}{\partial c\partial\mu} & \frac{\partial^2 \log L}{\partial c\partial\sigma} \\ \frac{\partial^2 \log L}{\partial\mu\partial c} & \frac{\partial^2 \log L}{\partial\mu^2} & \frac{\partial^2 \text{lo}}{\partial\mu\partial\sigma} \\ \frac{\partial^2 \text{lo}}{\partial\sigma\partial c} & \frac{\partial^2 \log L}{\partial\sigma\partial\mu} & \frac{\partial^2 \log L}{\partial\sigma^2} \end{bmatrix} \qquad (3.9.1)$$

$$-\frac{\partial^2 Log\ L}{\partial c^2} = \frac{1}{(1-C\ )^2}\Sigma d_i - \frac{1}{\Gamma(\alpha)^2}\Sigma(1-\ d_i)[\gamma(\alpha,\ e^{z})]^2\left[c+(1+c)\left[1-\frac{\gamma(\alpha,\ e^{z})}{\Gamma(\alpha)}\right]\right]^2(3.9.2)$$

$$-\frac{\partial^2 Log\ L}{\partial\mu^2} = -\Sigma\frac{d_i}{\sigma^2}e^{\left(\frac{\log t_i-\mu}{\sigma}\right)-(1-c)\Gamma(\alpha)\Sigma(1-d_i)\frac{\gamma\mu\mu\Gamma(\alpha)+(1+c)[\gamma^2\mu\,\gamma\mu\mu\gamma]}{[\Gamma(\alpha)-\gamma(1-c)]^2}}(3.9.3)$$

and

$$-\frac{\partial^2 Log\ L}{\partial\sigma^2} = \frac{1}{\sigma^4}\Sigma d_i\left[\sigma^2 + (\log t_i\ -\mu)\left(\ 2\alpha\sigma-(2\alpha+\log t_i-\mu)e^{\left(\frac{\log t_i-\mu}{\sigma}\right)}\right)\right]$$

$$-(1-\ c)(1-d_i)\frac{\gamma\sigma\sigma\Gamma(\alpha)+(1-c)[\gamma^2\mu\gamma\mu\mu\gamma]}{[\Gamma(\alpha)-\gamma(1-c)]^2}(3.9.4)$$

3.10 Gap to Fill

In practice, highly skewed data requires skewed distribution to handle it, thus, having confirmed that survival data like cancer data is highly skewed, thus, the data should then be handled with distributions that can accommodate high degree of asymmetry. Therefore, there is an utmost purpose for us to modify $f(x)$ that is the convectional distribution (generalised-gamma), with the assumptions that we are going to combine existing $G(x)\ and\ g(x).$ In this case, the gamma generator known as the Gamma-Generated Gamma is the link function that will be used. The process is to achieve the study second objective that is to derive an innovative parametric mixture cure fraction model that can control acute-asymmetry in survival data. The gamma generator is defined in the equation 3.10.1 below'

$$f(x) = \frac{1}{\Gamma(\omega)}[-\log[1-G(x)]]^{\omega-1}g(x)(3.10.1)$$

where:

ω = Shape parameter

$G(x)$ = reference distribution cdf

$g(x)$ = reference distribution pdf

If the shape parameter ωbecomes 1, at that point,

$$f(x) = \frac{1}{\Gamma(1)}\times\ [-\log[1-G(x)]]^{1-1}\ g(x) =\ [-\log[1-G(x)]]^0 g(x)$$

$$\therefore f(x) = g(x)$$

The essence of shape parameters is to control the asymmetry in the survival data. In a situation where the degree of asymmetry is so high and conventional distribution cannot handle it, one need to seek for distribution that is explicitly competent to regulate and control acute-asymmetry in survival data. This in turn will give comprehensive and additional report/evidence for the cured patients' proportion that have benefited from medical interventions.

3.11 Development of Modified GGMCM

This study developed an innovative cure model termed modified GGMCM. The following pdf and cdf were used jointly to achieve the proposed model. Following the assumption that random variable t follows a generalised-gamma distribution, and then its density function is defined follows as:

$$f(t) = \frac{\beta}{\theta\Gamma\alpha}\left(\frac{t}{\theta}\right)e^{-\left(\frac{t}{\beta}\right)^{\beta}} \qquad (3.11.1)$$

and its cdf as;

$$F(t) = \frac{\gamma\left[\alpha,\left(\frac{t}{\theta}\right)^{\beta}\right]}{\gamma(\alpha)} \qquad (3.11.2)$$

But in scale location form, equations (3.11.1 & 3.11.2) becomes

$$f(t) = \frac{1}{\sigma\Gamma\alpha}e^{-\alpha\left(\frac{\log t-}{\sigma}\right)} - e^{\frac{\log t-\mu}{\sigma}} \qquad (3.11.3)$$

and

$$F(t) = \frac{\gamma\left[\alpha,\ e^{\frac{\log t-}{\sigma}}\right]}{\Gamma(\alpha)} \qquad (3.11.4)$$

If we put the pdf and cdf in equation (3.11.3 & 3.11.4) into (3.10.1), we have

$$g(t) = \frac{1}{\Gamma\beta}\cdot\left[-\log\left[1-\frac{\gamma\left[\alpha,\ e^{\frac{\log t-\mu}{\sigma}}\right]}{\Gamma\alpha}\right]\right]^{\omega-1}\frac{1}{\sigma\Gamma\alpha}e^{-\alpha\left(\frac{\log t-\mu}{\sigma}\right)} - e^{\frac{\log t-}{\sigma}} \qquad (3.11.5)$$

Where in the above equation (3.11.2), the parameter β is equal to the parameter α in the same equation ,Thus, cdf of gamma generalised link function is

$$G(t) = \frac{\gamma(-\log(F(t)),\ \beta)}{\Gamma\beta} \qquad (3.11.6)$$

Putting equation (3.11.4) in scale location form through equation (3.11.6), we obtain the cdf of modified generalised Gamma as

$$G(t) = \frac{\gamma\left[-\log\left[\frac{\gamma\left[\alpha,\ e^{\frac{\log t-}{\sigma}}\right]}{\Gamma\alpha}\right],\ \ \beta\right]}{\Gamma\beta} \qquad (3.11.7)$$

Equations (3.11.5) and (3.11.7) above provide a new model of parametric mixture cure fraction. This is good for the management of acute asymmetry in survival data. They are the density function (pdf) and cumulative distribution function (cdf) of the modified GGMCM

To derive a new parametric mixture cure fraction model that would control acute asymmetry in survival data.

Using the Known function of density and survival from Modified-Generalised Gamma Mixture Cure Models and plugging them into the functions of log likelihood from cure models in equation 3.6.7, it gives equation (3.11.8) below:

Using the log-likelihood parameter function, the following procedures were obtained:

$$\log L = \Sigma d_i \log(1-C) + \Sigma d_i \log g(t_i) + \Sigma(1-\ d_i)\log[C+(1-C)S(t_i)] + \Sigma(1-\ d_i)\log[C+(1-C)S(t_i)]$$

$$\log L = \Sigma d_i \log(1-c\ ) + \Sigma d_i\ log\left[\gamma\frac{1}{\Gamma\beta}\Gamma\alpha e^{\alpha\left(\frac{\log t-\mu}{\sigma}\right)} - \left(\frac{\log t-\mu}{\sigma}\right)\left[-\log\left[1-\frac{\gamma\left(\alpha,\ e^{\left(\frac{\log t-\mu}{\sigma}\right)}\right)}{\Gamma\alpha}\right]\right]\right]^{\beta-1} +$$

$$\Sigma(1-\ d_i)\log[C+(1-C)]\left[1-\gamma\left[-\log\left[\frac{\gamma\left(\alpha,\ e^{\left(\frac{\log t-}{\sigma}\right)}\right)}{\Gamma\beta}\right]\ ,\ \beta\right]\right] \qquad (3.11.8)$$

The differentiation of equation (3.11.8) for c and solve the equation to zero provides the likelihood equation as

$$\frac{\partial \log L}{\partial C} = -\ \frac{\Sigma d_i}{1-C} + \frac{\Sigma(1-d_i)\left[1-\left(1-\gamma\left[-\log\left[\frac{\gamma\left(\alpha,\ e^{\left(\frac{\log t-}{\sigma}\right)}\right)}{\Gamma\beta}\right],\ \beta\right]\right)\right]}{C+(1-C)\left(1-\gamma\left[-\log\left[\frac{\gamma\left(\alpha,\ e^{\left(\frac{\log t-}{\sigma}\right)}\right)}{\Gamma\beta}\right],\ \beta\right]\right)} \qquad (3.11.9)$$

Furthermore, the second derivatives equation (3.12.8) above for c is

$$\frac{\partial^2 \log L}{\partial C^2} = -\frac{\Sigma d_i}{(1-C)^2} + \frac{\partial}{\partial C}\frac{\Sigma(1-d_i)\left[1-\gamma\left[-\log\left[\frac{\gamma\left(\alpha,\ e^{\left(\frac{\log t-}{\sigma}\right)}\right)}{\Gamma\alpha}\right],\ \beta\right]\right]}{C+(1-C)\,\gamma\left[-\log\left[\frac{\gamma\left(\alpha,\ e^{\left(\frac{\log t-}{\sigma}\right)}\right)}{\Gamma\alpha}\right],\ \beta\right]} \qquad (3.11.10)$$

It is not possible to achieve equation (3.11.10) in a close form after having equated it to zero when resolving for c, it is therefore resolved through numerical repetitive process.

Reasons for Modified-Generalised Gamma Distribution

1. Due to its robustness.
2. It has wider scope of applicability.
3. It is used for a good description of life time scenario such as survival data.
4. It is a distribution used to handle skewness in data.

3.12 Statistical Properties

In this section, the fourth objectives to investigate the properties of the new parametric cure fraction models will be examined.

3.12.1 Proper Density function for Modified GGMCM.

If the proposed MGGMCM is a proper PDF, this procedure should be followed:

From equation (3.11.5) above, the probability density function of the proposed model (MGGMCM) is given by:

$$g(t) = \frac{1}{\Gamma\beta}\left[-\log\left(1 - \frac{\gamma\left(\alpha,\ e^{\frac{e^{\log t-}}{\sigma}}\right)}{\Gamma\alpha}\right)\right]^{\beta-1} \times \frac{1}{\sigma\Gamma\alpha} e^{\frac{\log t-\mu}{\sigma}} - e^{\frac{\log t-\mu}{\sigma}} \quad \alpha, \beta > 0$$

To prove that the proposed (MGGMCM) is a proper density function, it must satisfied equation (3.12.1) condition.

$$\int_{-\infty}^{\infty} g(t)dt = 1 \tag{3.12.1}$$

$$= \int_0^{\infty} \frac{1}{\Gamma\beta}\left[-\log\left(1 - \frac{\gamma\left(\alpha,\ e^{\frac{e^{\log t-\mu}}{\sigma}}\right)}{\Gamma\alpha}\right)\right]^{\beta-1} \times \frac{1}{\Gamma\alpha} e^{\alpha\left(\frac{\log t-}{\sigma}\right)} - e^{\frac{\log t-}{\sigma}} dt$$

$$Let \qquad F = \frac{\gamma\left(\alpha,\ e^{\frac{e^{\log t-}}{\sigma}}\right)}{\Gamma\alpha}$$

$$f = \frac{dF}{dt} = \frac{1}{\sigma\Gamma\alpha} e^{\alpha\left(\frac{\log t-}{\sigma}\right) - e^{\frac{logt-}{\sigma}}}$$

Where;

$$dF = \frac{1}{\sigma\Gamma\alpha} e^{\alpha\left(\frac{\log t-}{\sigma}\right) - e^{\frac{\log t-}{\sigma}}} dt$$

Hence;

$$dt = \frac{dF}{\frac{1}{\sigma\Gamma\alpha} e^{\alpha\left(\frac{\log t-\mu}{\sigma}\right) - e^{\frac{logt-}{\sigma}}}}$$

$$dt = \frac{dF}{f}$$

By substitution

$$\int_0^{\infty} \frac{1}{\Gamma_{\beta}} [-\log(1-F)]^{\beta-1} \times f \times \frac{dF}{f}$$

$$\int_0^{\infty} \frac{1}{\Gamma_{\beta}} [-\log(1-F)]^{\beta-1} dF$$

$$let\ y = -\log(1-F)$$

$$-y = \log(1-F)$$

Taking the exponential of both sides

$$e^{-y} = e^{\log(1-F)}$$

$$e^{-y} = 1 - F \qquad (3.12.2)$$

From

$$y = -\log(1-F)$$

$$\frac{dy}{dF} = \frac{1}{(1-F)} \times -1$$

$$\frac{dy}{dF} = \frac{1}{1-F}dy$$

From (3.12.2)

$$1 - F = e^{-y}$$

$$\therefore dF = e^{-y}dy$$

By substitution

$$\int_0^\infty \frac{1}{\Gamma\beta}[y]^{\beta-1}e^{-y}dy$$

$$= \frac{1}{\Gamma\beta}\int_0^\infty y^{\beta-1}e^{-y}dy$$

By applying gamma function

$$\Gamma\alpha = \int_0^\infty e^{-x}x^{\alpha-1}dx$$

$$\therefore = \frac{1}{\Gamma\beta} \times \Gamma\beta$$

$$= 1. \qquad Proved$$

3.12.2 Asymptotic Properties

Taking limit of equation (3.13.2) as $t \to 0$ *and at* $t \to \infty$

$$g(t) = \frac{1}{\Gamma R}\left\{-\log\left[\frac{\gamma\left(\alpha,\ e^{\frac{logt-\mu}{\sigma}}\right)}{\Gamma\alpha}\right]\right\}^{\beta-1} \frac{1}{\sigma\Gamma\alpha}e^{\alpha\left(\frac{logt-}{\sigma}\right)-e^{\frac{logt}{\sigma}}} \qquad (3.12.2)$$

$$\lim_{t\to 0} g(t) = \lim_{t=0}\frac{1}{\Gamma R}\left\{-\log\left[\frac{\gamma\left(\alpha,\ e^{\frac{log0-\mu}{\Gamma\alpha}}\right)}{\Gamma\alpha}\right]\right\}^{\beta-1}\frac{1}{\sigma\Gamma\alpha}e^{\alpha\left(\frac{log0-\mu}{\sigma}\right)-e^{\frac{log0-\backslash mu}{\sigma}}}$$

$$= \frac{1}{\Gamma\beta}(-log\infty)^{\beta-1}\frac{1}{\sigma\Gamma\alpha}e^{\infty-e^{\infty}}$$

$$= \frac{\infty}{\Gamma\beta\sigma\Gamma\alpha}$$

$$\lim_{t\to 0} g(t) = \infty$$

As $t \to \infty$

$$\lim_{t=\infty} = \lim_{t=\infty}\frac{1}{\Gamma R}\left\{-\log\left[\frac{\gamma\left(\alpha,\ \frac{log\infty-\mu}{\sigma}\right)}{\Gamma\alpha}\right]\right\}^{\beta-1}\frac{1}{\sigma\Gamma\alpha}e^{\alpha\left(\frac{log\infty-\mu}{\sigma}\right)-e^{\frac{log\infty-}{\sigma}}}$$

$$= \frac{1}{\Gamma\beta}(0).\frac{1}{\sigma\Gamma\alpha}(0)$$

$$= 0$$

3.12.3 Hazard Function

Hazard $=\frac{Pdf}{Survival}$ and $Survival = 1 - CDF$

$$G(t) \equiv pdf = \frac{1}{\Gamma\beta}\left[\left[-\log\left(1 - \frac{\gamma\left(\alpha,\ \frac{e^{logt-}}{\sigma}\right)}{\Gamma\alpha}\right]\right]^{\beta-1}\frac{1}{\sigma\Gamma\alpha}e^{\alpha\left(\frac{e^{logt-\mu}}{\sigma}\right)-e^{\frac{logt-\mu}{\sigma}}dt}$$

$$G(t) = \frac{\left(\gamma\left[-\log\left(\frac{\gamma\left(\alpha,e^{\frac{logt-\mu}{\Gamma\alpha}}\right)}{\Gamma\alpha}\right)\right],\ \beta\right)}{\Gamma\beta}$$

$$s(t) = 1 - G(t)$$

$$= 1 - \frac{\gamma\left(\left[-log\left(\frac{\alpha,\ e^{\frac{logt-\mu}{\sigma}}}{\Gamma\alpha}\right)\right],\ \beta\right)}{\Gamma\beta}$$

Hazard or $h(t) = \frac{g(t)}{s(t)}$

$$h(t) = \frac{\frac{1}{\Gamma\beta}\left[-\log\left(1 - \frac{\gamma\left(\alpha,\ e^{\frac{logt-}{\Gamma\alpha}}\right)}{\gamma\alpha}\right)\right]^{\beta-1} \frac{1}{\sigma\Gamma\alpha} e^{\alpha\left(\frac{log}{\sigma}\right)-e^{\frac{logt-\mu}{\sigma}}}}{1 - \frac{\gamma\left(\left[-\log\left(\frac{\gamma\left(\alpha,e^{\frac{logt-\mu}{\sigma}}\right)}{\Gamma\alpha}\right)\right],\beta\right)}{\Gamma\beta}}$$

$$= \frac{\frac{1}{\Gamma\beta}\left[-\log(1-F)^{\beta-1}f\right]}{\frac{\Gamma\beta-\gamma([-\log F,\ \beta)}{\Gamma\beta}}$$

$$= \frac{-\log{(1-F)^{\beta-1}f}}{\Gamma - \gamma([E - \log F],\ \beta)} \times \frac{\Gamma\beta}{\Gamma\beta}$$

$$h(t) = \frac{-\log{(1-F)^{\beta-1}f}}{\Gamma\beta - \gamma([E - \log F],\ \beta)}$$

$$h(t) = \frac{-log\left(1 - \frac{\gamma\left(\alpha,e^{\frac{logt-}{\sigma}}\right)}{\Gamma\alpha}\right)^{\beta-1} \frac{1}{\Gamma\alpha} e^{\alpha\left(\frac{logt-}{\sigma}\right)-e^{\frac{logt-\mu}{\sigma}}}}{\Gamma\beta - \gamma\left(\left[-\log\left(\frac{\gamma\left(\alpha,e^{\frac{logt-\mu}{\sigma}}\right)}{\Gamma\alpha}\right),\ \beta\right]\right)}$$

3.12.4 Entropy

The Renyi Entropy is examined in this section; which is a measure of uncertainty variation. It is also used to measure the degree of disorderliness in a model. The equation below gives the formulae used to estimate it .

$$Ent = \frac{1}{1 - S(t)} \log[f(x)]$$

$$Since \qquad F(t) = \frac{\gamma\left[\alpha,\ e^{\frac{logt-}{\sigma}}\right]}{\Gamma\alpha}$$

The survival functions at time t;

$$S(t) = 1 - F(t)$$

$$S(t) = 1 - \frac{\gamma\left[\alpha,\ e^{\frac{logt-}{\sigma}}\right]}{\Gamma\alpha}$$

$$Ent = \frac{1}{1 - S(t)}\log[f(x)] \equiv \frac{1}{1 - S(t)}\log g(t)$$

$$Ent = \frac{1 \times \frac{1}{\sigma\Gamma\alpha} e^{\alpha\left(\frac{logt-\mu}{\sigma}\right) - e^{\frac{logt-\mu}{\sigma}}}}{1 - \frac{\gamma\left[\alpha,\ e^{\frac{logt-\mu}{\sigma}}\right]}{\Gamma\alpha}}$$

$$Ent = \frac{e^{\alpha}\left(\frac{logt-\mu}{\sigma}\right) - e^{\frac{logt-}{\sigma}}}{\sigma\gamma\left[log, e^{\frac{logt-\mu}{\sigma}}\right]}$$

3.13 Order Statistics of the proposed Model

$$g(y_r) = \frac{n!\,[1 - F(Y_r)]^{n-r}[F(y_r)]^{r-1} f(y_r)}{(n-r)!\,(r-1)!}$$

$$f(t) = \frac{\beta}{\Gamma k\theta}\left(\frac{t}{\theta}\right)^{k\beta-1} e^{-\left(\frac{t}{\theta}\right)^{\beta}}$$

(3.14.1)

$$\log t = y \rightarrow t = e^{y}$$

$$\log\theta = \mu \Rightarrow \theta = e^{\mu}$$

$$\beta = \frac{1}{Y}$$

$$f(y) = \frac{1}{\gamma\Gamma k} e^{-\mu}\left(\frac{e^{y}}{e^{\mu}}\right)^{\frac{k}{Y}-1} e^{-\left(\frac{e^{y}}{e^{\mu}}\right)^{\frac{1}{Y}}} \tag{3.14.2}$$

$$f(y) = \frac{1}{\gamma\Gamma k} e^{-\mu} e^{(y-\mu)^{\frac{k}{Y}-1} e^{-e^{\frac{y-\mu}{\gamma}} e^{y}}}$$

$$= \frac{1}{\gamma\Gamma k} e^{-\mu} e^{k\left(\frac{y-\mu}{\gamma}\right)^{\frac{k}{Y}-1} e^{-y} e^{\mu} e^{-e^{\frac{y-\mu}{\gamma}} e^{y}}}$$

$$f(y) = \frac{1}{\gamma\Gamma k} e^{k\left(\frac{y-\mu}{\gamma}\right) - e^{\frac{y-\mu}{\gamma}}}$$

The above is the Modified Generalised-Gamma in Scale Location Form.

3.14 Setting up Monte Carlo Experiment

In this simulation study, the uniform distribution with value of b=100 and a=1 are considered for data generation. Each data set contained 10, 20, and 50 observations with 50, 100, and 500 replications, each of whose different censoring rates depends on the value of a and b. For the purpose of regulating the generation process, assumptions were made by subjecting that our true survival time t will have to follow uniform distribution. The following were the algorithms used for data generation:

1. Generate from U (10, 1,100), U (20, 1,100) and U (50, 1,100) with (50,100,500) replications each.
2. Generate from each sample size the censoring time of 3+20[U (10, 1,100)].
3. We conditioned the censoring time for if else statement,
   If else (censoring < survival time, 1, 0)

CHAPTER FOUR

ANALYSIS, RESULTS AND DISCUSSION

4.1 Introduction

This chapter focuses on the results of data analysis described in the last chapter. The analysis tracks results from life data as well as simulation study outcomes. These results were presented in tables, plots, charts, and figures.

We used simulated data from the experiment discussed earlier to show capability of the proposed Modified Generalised-Gamma MCM (MGGMCM). In order to compare the models, we considered MSE, RMSE and Absolute BIAS as some criteria to check for the best across the replications considered. The efficiency of the model is determined from model with least criterion value.

We also used thirty seven (37) life ovarian cancer data which was gotten from Department of Obstetrics and Gynaecology, University College Hospital, Ibadan, Nigeria covering the period 2000-2015. The survival period for cohort of patients is from the date of diagnosis until death in months and it was done using the prevailing four models of parametric mixture cure compared to the model proposed. The simulation was carried out with R codes, some of the outputs were copied to excel spread sheet in other to calculate some necessary statistics such as the average parameters. Data cleaning of the thirty seven life ovarian cancer data was done with the use of Microsoft excel spread sheet and the analysis was carried out using R software. The R codes were used to estimate all the parameters such as μ,σ,c and median time-to-cure from the considered models as they were showcased in the previous sections for both simulated data and ovarian cancer data. . The following statistics were estimated for each of the parametric distribution; Akaike Information Criterion (AIC), log-likehood Density Curves, Survival Curves, Hazard Curves, median, Median time to recuperate, and variances of c in all the considered models were also obtained.

4.2 Presentation of Results (Simulated Data)

Table 4.1: Model Evaluation of Simulated Data

Sample Size	Rep	MSE					RMSE					|Bias|				
		Llogis	Weib	Lnorm	GG	MGG	Llogis	Weib	Lnorm	GG	MGG	Llogis	Weib	Lnorm	GG	MGG
n=10	50	791.300	772.110	707.230	691.030	701.100	28.130	27.790	26.480	26.290	26.480	27.890	26.100	26.700	25.330	25.810
	100	801.900	777.030	700.520	700.100	710.400	28.320	27.890	26.470	26.450	27.650	27.100	26.230	25.900	25.890	25.970
	500	856.200	819.450	810.010	711.320	750.610	29.260	28.630	28.460	26.670	26.400	27.150	26.550	27.200	25.910	26.380
n=20}	50	695.200	611.590	655.710	601.330	609.310	26.370	24.730	25.610	24.520	24.640	26.000	23.590	25.310	23.570	23.890
	100	720.500	774.630	671.910	623.500	620.900	26.840	27.830	25.920	24.970	24.920	26.230	26.190	25.500	23.270	23.130
	500	751.300	71.980	704.130	671.450	663.530	27.410	26.700	26.530	25.910	25.760	26.710	25.550	25.890	24.890	23.800
n =50	50	719.640	700.180	699.520	689.150	601.590	26.830	26.460	26.450	26.250	24.530	25.500	25.770	25.830	25.590	23.150
	100	703.900	707.480	700.850	610.790	598.400	26.530	26.600	26.470	24.710	24.460	25.900	25.980	25.800	24.080	23.030
	500	644.590	623.900	619.610	602.100	501.370	25.390	24.980	24.890	24.540	22.390	24.890	23.950	24.010	24.000	22.110

llogis:	LLMCM
lognorm:	LNMCM
Weibull:	WMCM
GG:	GGMCM
MGG:	MGGMCM

The result in the table 4.1 above gives the simulation result where values are consequently generated through a uniform distribution of which different censoring rates depends on the value of a and b. Simulation were done as it was explained in previous chapter under setting up the Monte Carlo Experiment. The data set were generated with sample sizes 10, 20 and 50 with the replication's levels of 50, 100 and 500. In this simulation settings, censoring times were uniformly distributed. We mimicked the situation of the real life ovarian cancer data resulting to having a scenario of when the data set is relatively large with sample size of 37. That informed the choice of using sample size not more than 50 observation.

From the table, using MSE criterion, when the sample size n=10, with replications' level of 50, 100 and 500. The results show that GGMCM has the least values with 690.030, 700.100 and 711.320 respectively. For RMSE criterion, with the same sample size n=10, and replications' level of 50, 100 and 500, the results show that GGMCM still has the least values with 26.290, 26.450 and 26.670, respectively. Also, for absolute bias as criterion, when the sample size n=10, with replications' level of 50, 100 and 500. The results show that GGMCM has the least values with 25.330, 25.890 and 25.910, respectively.

When the sample size increases to n=20, using MSE, RMSE and absolute bias criterion, with replications' level of 50, GG still gives the least values with 601.330, 24.520, 23.570, respectively. However, as the sample size increases to n=20 for replications level of 100 and 500. The results show that the proposed MGGMCM has the least values with 620.900 and 663.530, respectively. Also, for RMSE criterion, the results show that MGGMCM has the least values with 24.920 and 25.760, respectively. And finally for the absolute bias criterion, with the same sample size n=20, and replications level of 100 and 500, the results show that MGGMCM has the least values with23.130 and 23.800, respectively.

As soon as the sample size further increases to 50, results from the MSE criterion with replications level of 50, 100 and 500 show that the proposed MGGMCM has the least values with 601.590, 598.400 and 501.370 respectively. The same result was gotten for RMSE criterion, with the same sample n=50 and replications level of 50, 100 and 500. The results are 24.530, 24.460 and 22.390 respectively for MGGMCM which has the least values. Finally, using absolute bias criterion, when the sample size n=50 and replications level is 50, 100 and 500, the results show that MGGMCM has the least values with23.150, 23.030 and 22.110, respectively.

On the final note, it was observed that the mean square error (MSE) criterion for each sample sizes increases as the number of replication increases. Consistently, the proposed estimators outperforms other convectional models unless there is a very small sample size.

Table 4.2: Average Parameters

Sample	Rep			$\hat{\mu}$					$\hat{\sigma}$					c		
Size		Llogis	Weib	Lnorm	GG	MGG	Llogis	Weib	Lnorm	GG	MGG	Llogis	Weib	Lnorm	GG	MGG
n =10	50	4.981	3.451	4.267	5.885	5.203	25.883	24.910	25.557	27.490	27.089	0.016	$27 \times 10^{(-5)}$	0.051	0.154	0.127
	100	5.020	4.116	5.013	5.910	5.414	26.117	25.100	26.350	27.890	27.158	0.009	$9.2 \times 10^{(-5)}$	0.089	0.553	0.501
	500	3.510	2.584	3.391	4.831	4.549	25.339	24.024	25.150	26.001	25.911	0.001	0.000	0.094	0.468	0.531
n=20	50	5.319	5.100	5.417	5.810	5.900	24.389	24.219	25.380	26.800	27.580	0.001	0.000	0.004	0.007	0.493
	100	5.511	4.777	5.031	5.883	6.316	25.019	25.100	26.017	26.933	27.877	0.005	0.000	0.009	0.101	0.524
	500	5.826	4.966	5.225	6.010	6.515	25.111	24.361	26.232	27.100	28.203	0.008	0.000	0.019	0.358	0.612
n =50	50	5.818	5.300	5.712	6.223	6.400	25.419	25.012	26.817	27.111	27.821	0.012	0.003	0.051	0.444	0.571
	100	5.908	5.450	5.410	6.298	6.792	26.101	25.714	26.114	27.495	28.400	0.053	0.006	0.055	0.481	0.716
	500	6.044	5.555	5.700	6.546	6.889	26.349	25.422	27.036	27.811	28.759	0.077	0.002	0.093	0.500	0.890

llogis:	LLMCM
lognorm:	LNMCM
Weibull:	WMCM
GG:	GGMCM
MGG:	MGGMCM

In the previous chapters, some transformations were done to have the same parameters across the models. This led to the process of getting the three salient parameters of interest for this study. The parameters are $\mu, \sigma\ and\ c.$ The result depicts the estimate of the aforementioned parameters: the location parameter (μ), the scale parameter (σ) and most importantly the cure fraction parameter (c) for all the model explored with the proposed model MGGMCM. The scaling factor σ in the subject matter is used for the purpose of standardization across the models.

When considering the location parameter (μ), with sample size n=10, and replications' level of 50, 100 and 500. The results show that GGMCM has the highest average values with 5.885, 5.910 and 4.831, respectively. Also, the scale parameter (σ) when the sample size n=10, with same replications' level are 27.490, 27.890 and 26.001. The results also show that GGMCM has the highest values with 24.530, 24.460 and 22.390, respectively. For Cure fraction parameter (c), when the sample size n=10, with replications' level of 50, 100 and 500. The highest value of c still belongs to GGMCM with 0.154, 0.553 and 0.468, respectively.

For the location parameter (μ), when the sample size n=20, with replications' level of 50, 100 and 500. The results show that the proposed MGGMCM has the highest average values with 5.900, 6.316 and 6.515, respectively. Similarly, the scale parameter (σ) when the sample size n=20, and the level of replications 50, 100 and 500 gives the proposed MGGMCM to have the highest values with 27.580, 27.877 and 28.230, respectively. Furthermore, for cure fraction parameter (c), when the sample size n=20, with replications level of 50, 100 and 500. The results also show that MGGMCM has the highest values with 0.493, 0.524 and 0.612, respectively.

Steadily, for the location parameter (μ), when the sample size n=50, with replications' level of 50, 100 and 500. The results show that the proposed MGGMCM has the highest numerical values with 6.400, 6.792 and 6.889, respectively. In the same vein, the scale parameter (σ) when the sample size n=50, with same level of replications show that MGGMCM has the highest values with 27.821, 28.400 and 28.759 respectively. On a final note, consistently, cure fraction parameter (c), when the sample size n=50, with replications level of 50, 100 and 500. The results show that MGGMCM has the highest values with 0.571, 0.716 and 0.890, respectively.

The results in table 4.2 above indicates that when the sample sizes increases to 20 and 50 at all levels of replications, the proposed MGGMCM give estimates that is higher in numerical value than the existing models with larger dispersion. A case when n=50 at 500 level of replication, the results presents cure fraction parameter (c) under the MGGMCM to be 0.8902, a result very close to 1The cure fraction that corresponds to the proportion of patients cured of theirdisease is 0.89 while those that are uncured of their disease are 0.11. this result corroborate with the definition of mixture cure model by Boag (1949) & Berkson and Gage (1952 which gives it as the evaluation of the proportion of cured patients along with proportions of uncured patients' survival function. This result is equivalently calculating the percentage of cured patients from cancer that is their cure rate. As it was established in previous chapters that the cure (c) lies between 0 and 1, and the closer to the value of 1, the better the proportion of cure; in this case, the proposed model MGGMCM gives better rate.

It can be deduce from the result in table 4.2 that in a small sample size study when the sample size is relatively small, say when n=10, distribution like GGMCM can over-estimate parameters of interest or better put that the proposed MGGMCM will under estimate parameters of the models considered with less level of variety. This is evidence from table 4.2 when n=10 at every levels of replication.

However, as sample size n increases to n= 20 and n=50, irrespective of the replication's level, the proposed model MGGMCM gives the best with the competing models, across all the parameters of interest considered for this study.

On the final note, it was observed that cure fraction (c) associated with the convectional models for each sample sizes increases as the number of replication increases. Consistently, cure fraction (c) associated with the proposed gives the highest value of c. Since c represent the proportion of patients who can benefit from medical intervention, having c = 0.890 implies that the convectional models underestimate the cure fraction. This indicated that about 89% patients were able to be managed or cured after medical intervention using the proposed MGGMCM.

Table 4.3: Median Time to Cure (Recovery Time of Patients)

Distribution	$n = 10$			$n = 20$			$n = 50$		
	50	100	500	50	100	500	50	100	500
	$e\hat{\mu}$	$e\hat{\mu}$	$e\hat{\mu}$	$e\hat{\mu}$	$e\hat{\mu}$	$e\hat{\mu}$	$e\hat{\mu}$	$e\hat{\mu}$	$e\hat{\mu}$
Weibull	61.07060	29.18130	24.72350	22.14910	20.32320	20.00190	20.39150	15.19490	10.18180
IIogis	67.61500	25.23670	19.30010	19.10100	17.19910	15.51900	11.00100	10.33330	8.32110
Lognorm	53.91510	19.75680	16.39150	15.15651	11.51170	10.11110	10.55190	9.91000	8.01290
GG	24.58100	12.66860	12.28070	10.39880	8.22920	8.31070	7.77810	7.00210	6.11900
MGG	10.72780	12.00980	9.67230	7.19990	6.71580	6.00130	6.11710	5.82510	3.10590

llogis:	LLMCM
lognorm:	LNMCM
Weibull:	WMCM
GG:	GGMCM
MGG:	MGGMCM

This table 4.3 above gives the result of median time-to-cure that is the recovery time of patients at which cure happens across the models. This is the amount of time after which some proportions of the patients have died and some proportion has recovered after therapy (medical intervention). The time that it takes for cure to occurs. From the result, when considering sample size n=10 with replications' level of 50, 100 and 500. The results show that the proposed MGGMCM has the minimum recovery time that is the values of median time-to-cure are 10.72780, 12.00980 and 9.67230 respectively. Also, for sample size n=20 with the same replications' level, the results are 7.19990, 6.71580 and 6.00130. As n increases, with sample size n=50, under all the levels of replications considered, the median time-to-cure gives better result which has the least median time to cure. The results are 6.11710, 5.82510 and 3.10590

It can be deduced that table 4.3 presents results on the longevity of time required to cure a particular ovarian cancer patients that is the recovery time. The smaller the value under a particular model or distribution, the better it (distribution) is. Hence, it is evidenced within the scope of statistical investigation and analysis of this research that proposed MGGMCM proved beyond any reasonable doubt the reason for its adoption. This is crystal clear as its median time to cure is the least among the considered models across all the sample sizes and level of replications. According to this study, other competing models were experiencing longer recovery time than that of the new proposed MGGMCM.

This result implies that MGGMCM has recovery time of approximately 10 months, 12 months and 9 months when sample size n= 10, with 50, 100 and 500 replications respectively. The recovery time when sample size n= 20, with 50, 100 and 500 replications are 7 months, 6.5 months and 6 months respectively. Finally, MGGMCM has recovery time of approximately 6 months, 5 months and 3 months when sample size n= 50, with 50, 100 and 500 replications respectively. The result on the above table 4.3

confirms the Yigzaw et.al (2019) studywhich established that the minimum recovery time to cure a particular disease is substantial.

In this study, we are trying to compare on the average those patients who can benefit from medical intervention that is the proportion of patients that can be managed from cancer or cured. We don't want a situation whereby the patients will stay a longer time because of the pain they undergone, we need a model that can guarantee quick recovery time that is a model that can improve their healing. With the above result, we realize that many patients were able to cure faster on the average using the proposed model comparing to other convectional models.

On the final note, it was observed that the median time to cure ($e\hat{\mu}$) associated with the convectional models for each sample sizes decreases as the number of replication increases. Consistently, median time to cure ($e\hat{\mu}$) associated with the proposed model gives the minimum value of$e\hat{\mu}$. Since $e\hat{\mu}$ represent the median time to cure of patients who can benefit from medical intervention, having small value of median time to cure implies that a lot of patients were able to cure on time and this shows that proposed model is optimal in estimating the proportion of those who can benefit from medical intervention.

4.3 Analysis of exploratory data (Ovarian Cancer)

Exploratory data analysis (EDA) is an approach for analysing data sets by summarising their key features, often with pictorial procedures. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task. It is techniques that help one to understand the data and assist to understand the events that generated the data. In this section, descriptive statistics of the ovarian cancer were done. The figures in this section clearly show the spread of the distribution for the observed data.

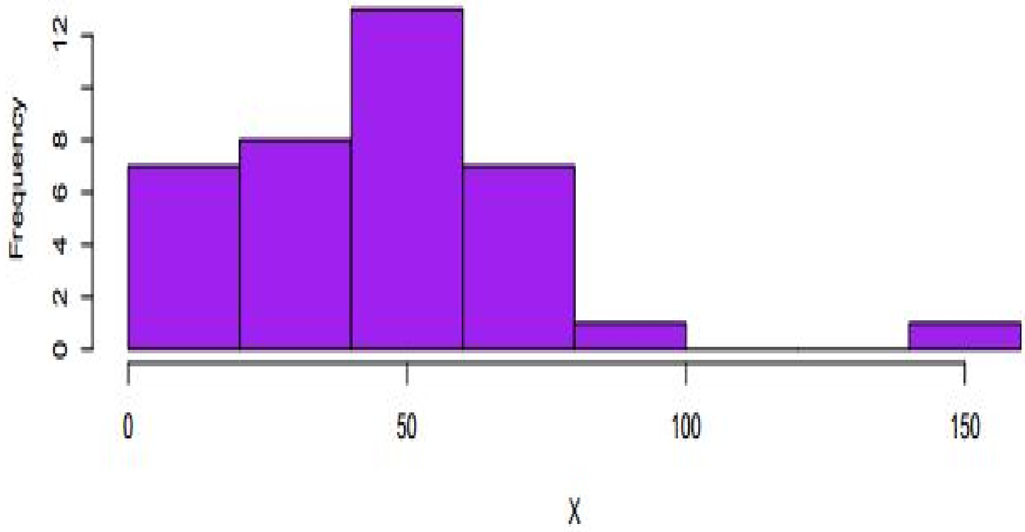


Figure 4.1. Histogram plot for the survival time

The figure 4.1 above represents the histogram plot of the survival time for the real life ovarian cancer observation. In the figure above, the ovarian cancer data is clearly shown and the spread of the distribution for the observed data were shown graphically.

It is evident from the plot that the data is skewed to the right. Similarly, this result claimed that survival data were rightly skewed. Also, there is a wide gap between the observation with 100 months and 159 months. The gap indicates that there is an outlier case.

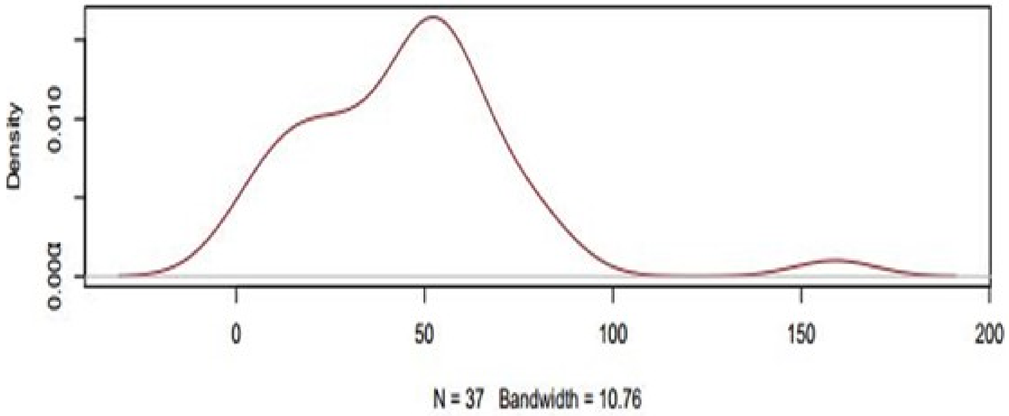


Figure 4.2 Survival Period Density Plot

The Figure 4.2 clearly demonstrates the survival density plot. This plot proves that the real life ovarian cancer data is far from normal density curve. The curve also shows the right long tail that was accrued from the outlier observed from the data, that is, the observation with 159 months. The non-normality in the data is more strikingly shown with this curve.

From the figure 4.2 above, the density curve in survival analysis always describes the transition times, and a sharp peak, for instance around a particular value means that most survival times are found around. It could be observed that the ovarian cancer cure rate rises rapidly to the peak around 50 followed by another rapid then decline which seems to level off. The curve assists us to see the true pictures of the characteristics of the ovarian cancer data. It also helps us to know the center and the spread about this central value. Since our interest is to investigate outliers or study the distribution or pattern of the ovarian cancer data values, several plots are available to allow the study of the distribution. One such plot is the density plot.

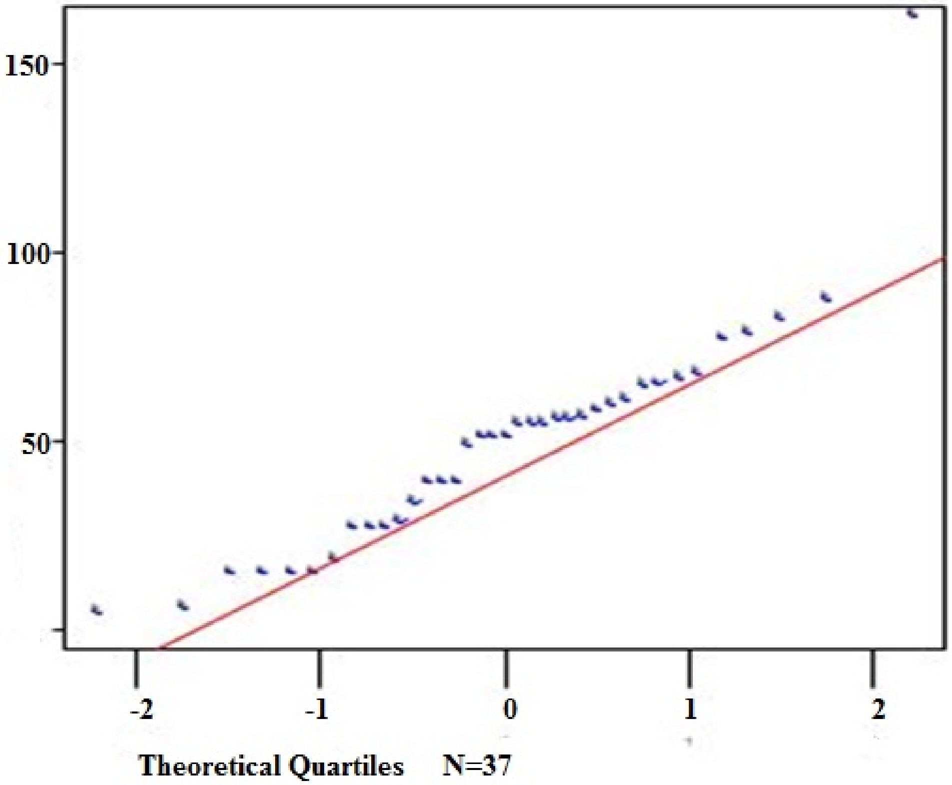


Figure 4.3 Survival Period Normal Q−Q Plot

The above figure 4.3 displays the Q-Q plot which checks for normality status of the data. In most cases, survival data is characterized with skewed data and prior expectation is that ovarian cancer data should have same features. The result from the figure depicted that the real life ovarian cancer shows that it is skewed to the right having much longer tail due to the outlier case observed with (159 months). Thus, the prior expectation is true that the data exhibits non-normality. The point at the extreme end to the right is the outlier month mentionedearlier.

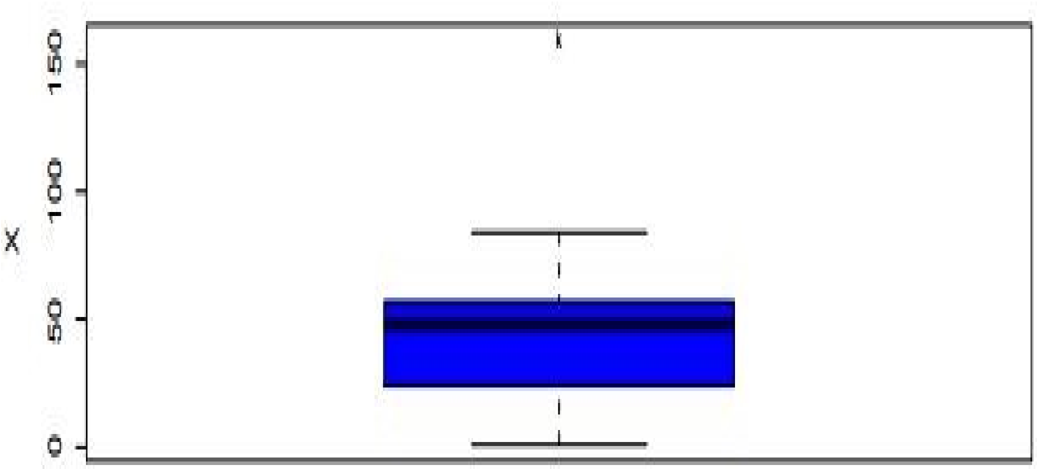


Figure 4.4 Survival Period Box-Plot

Figure 4.4 demonstrates the survival period boxplot. This plot is often used to detect outliers in a dataset, but the outliers observed from the data on ovarian cancer can be clearly located outside the plot displaying outliers as dots outside the whiskers.

Figure 4.4 above also provides pictorial information about ovarian cancer in real life scenerio from the result. There are two lines beyond the box and they are called the whiskers. These lines show the minimum and maximum observation The plot represents the five number summaries of statistics in a dataset namely minimum, Q1, Q2, Q3 and maximum respectively.. From the result of this figure 4.4, it can be deduced that there is distinct case of ovarian cancer that has survived up till 159 months after being diagnosed of the diseases, in this study, that observation was seen as an outlier.

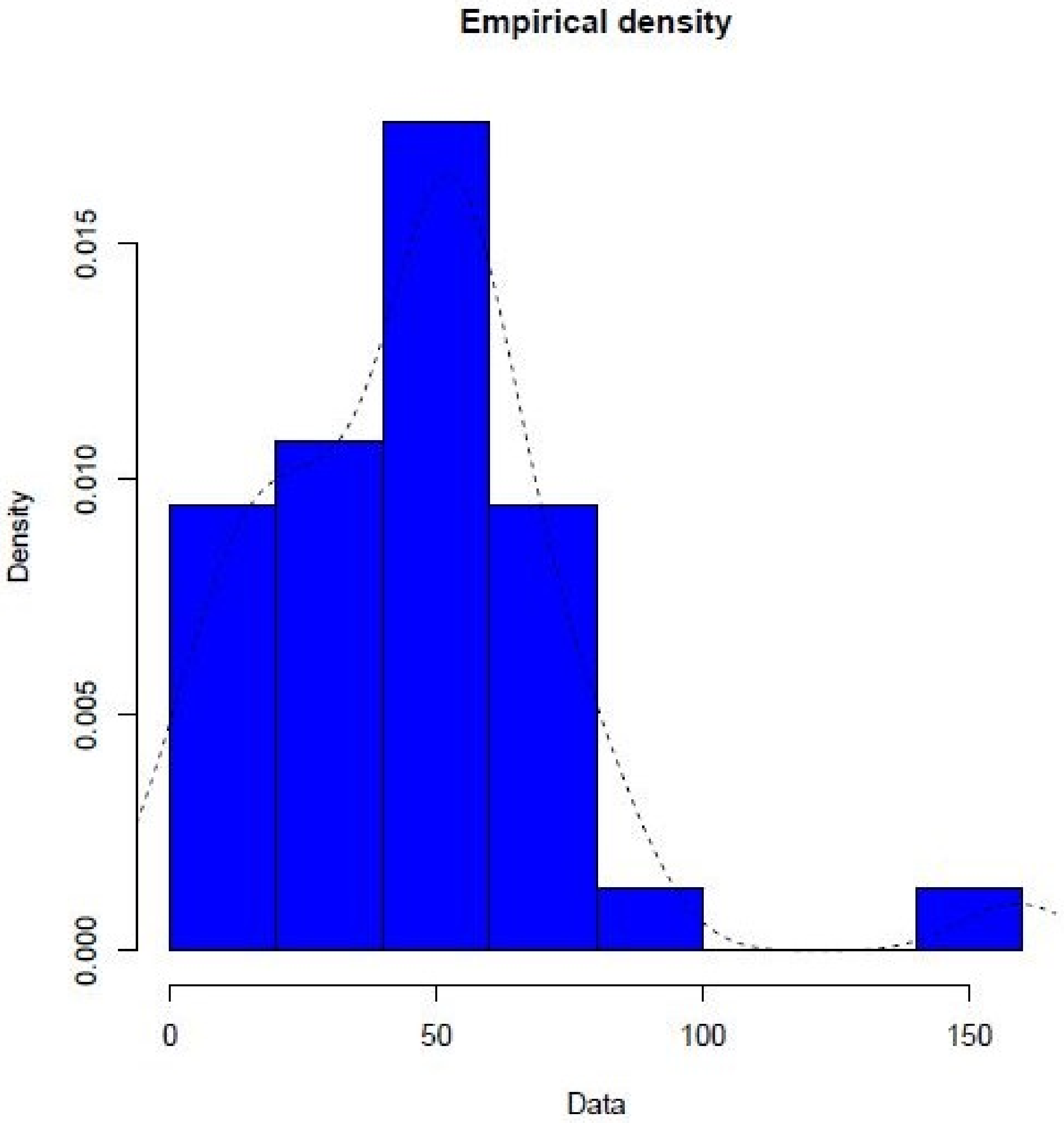


Figure 4.5 Empirical Density

Figure 4.5 above illustrates the life data based on observation density plot that shows the histogram and the curve of density embedded on it. This plot clearly shows how the distribution is far from normality which illustrates the shapes of the distribution and how it spread. It also shows the length of the tails and which side the tails move either to the right or left. This result conformed to the previous results that show that it is positively right skewed. The gap in the plot shows that the data has an outlier case.

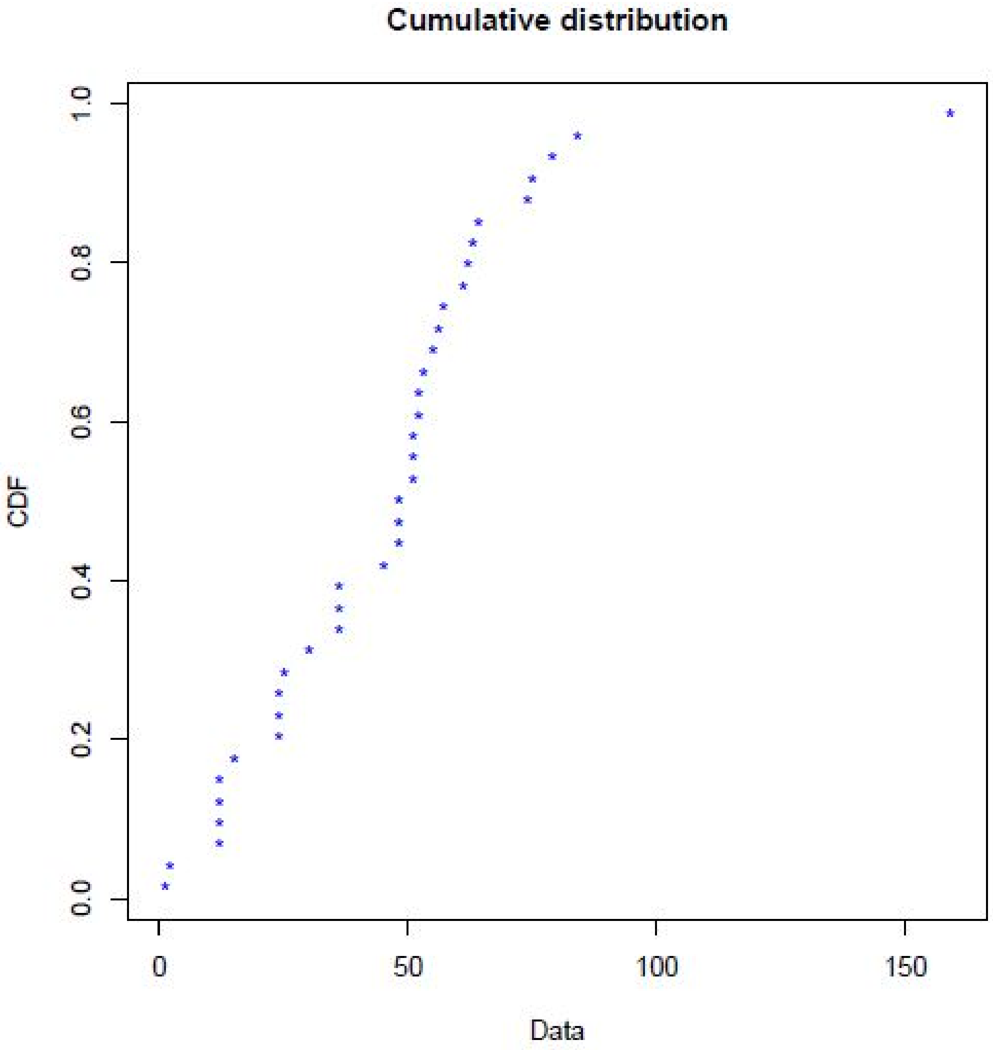


Figure 4.6 Cumulative Distribution plot

Figure 4.6 above shows the cumulative density plot of the data on ovarian cancer, the plot displays that it converges to 1, and so this indicates a perfect distribution estimate on the data for all patients observed. In cumulative distribution function, the curve is a concave up parabola which lies between $-1 < x \leq 0$ and a concave down parabola which also lies between $0 < x < 1$. Hence, the figure 4.6 gives the curve that is a concave down parabola which lies between $0 < x < 1$. The x-axis signifies time in months, and the y-axis displays the cumulative distribution function of surviving or the percentage of patients benefiting from medical intervention.

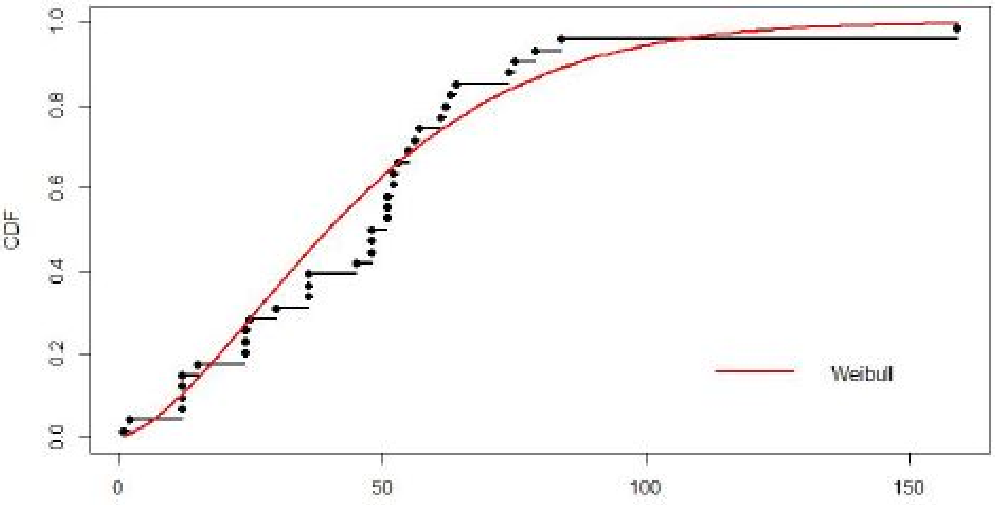


Figure 4.7 Cumulative Density Plot of Weibull Mixture Cure Model.

The figure 4.7 above gives the cumulative density plot of the ovarian cancer data associated with weibull MCM , the plot show the movement from 0 to 1, In cumulative distribution function, the curve is a concave up parabola which lies between $-1 < x \leq 0$ and a concave down parabola which also lies between $0 < x < 1$. Hence, the figure 4.6 gives the curve that is a concave down parabola which lies between $0 < x < 1$. The x-axis signifies time in months, and the y-axis displays the cumulative distribution function of surviving or the percentage of patients likely to enjoy the benefits of medical intervention. The lines represent survival curves of the ovarian cancer associated with weibull. A stepwise dotted black line in the plot indicates the observed event. It is observed that weibull MCM is closely converged to observed data. This implies that the model estimate the ovarian cancer but not as fast as the clinicians expect.

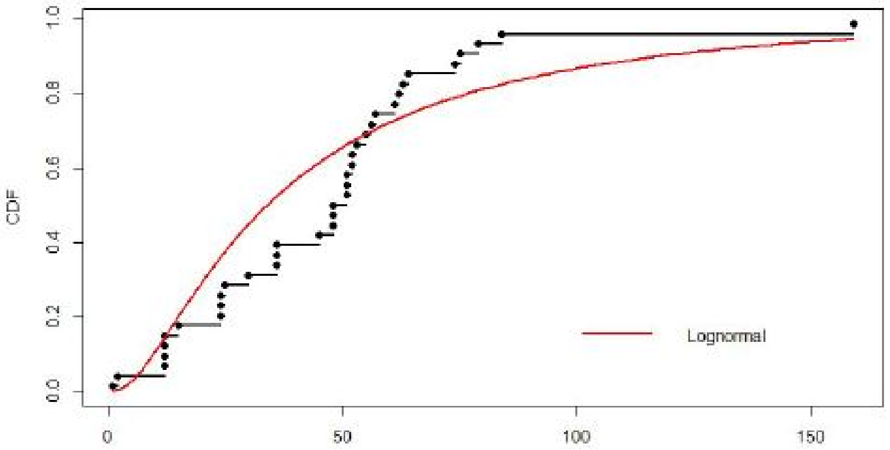


Figure 4.8 Cumulative Density Plot of Lognormal Mixture Cure Model.

The figure 4.8 above gives the cumulative density plot of the ovarian cancer data associated with lognormal MCM , the plot show the movement from 0 to 1, In cumulative distribution function, the curve is a concave up parabola which lies between $-1 < x \leq 0$ and a concave down parabola which also lies between $0 < x < 1$. Hence, the figure 4.6 gives the curve that is a concave down parabola which lies between $0 < x < 1$. The x-axis signifies time in months, and the y-axis displays the cumulative distribution function of surviving or the proportion of patients that have benefited from medical intervention.

The lines represent survival curves of the ovarian cancer associated with lognormal. A stepwise dotted black line in the plot indicates the observed event. It is observed that lognormal MCM does not closely converge to observed data. This implies that the model under estimate the ovarian cancer and this could undermine the clinician’s effort.

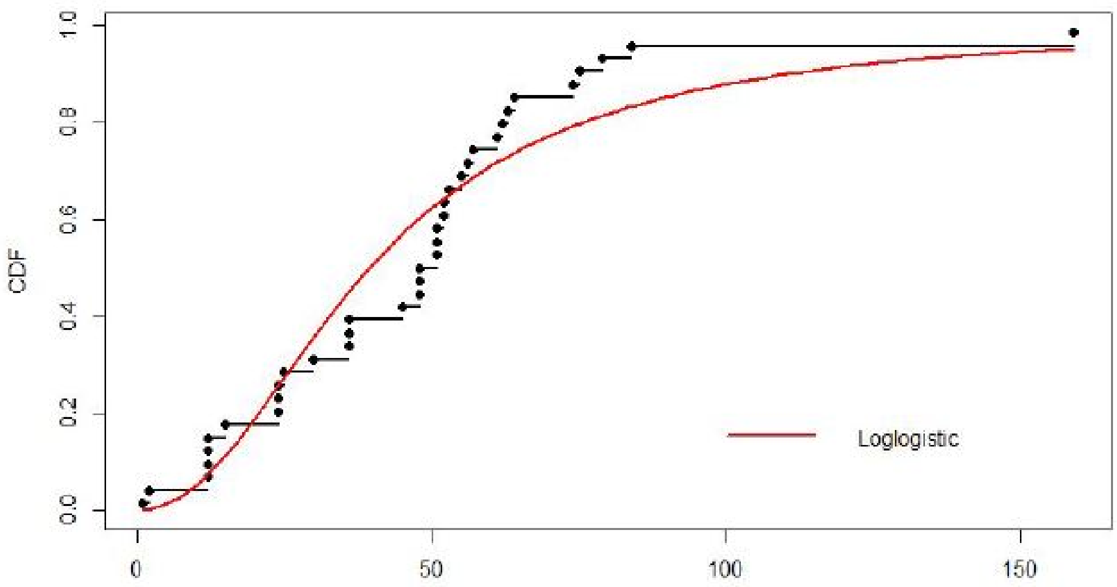


Figure 4.9 Cumulative Density Plot of Loglogistic Mixture Cure Model.

The figure 4.8 above gives the cumulative density plot of the ovarian cancer data associated with loglogistic MCM , the plot show the movement from 0 to 1, In cumulative distribution function, the curve is a concave up parabola which lies between $-1 < x \leq 0$ and a concave down parabola which also lies between $0 < x < 1$. Hence, the figure 4.6 gives the curve that is a concave down parabola which lies between $0 < x < 1$. The x-axis signifies time in months, and the y-axis displays the cumulative distribution function of surviving or the percentage of patients likely to enjoy the benefits of medical intervention.

The lines represent survival curves of the ovarian cancer associated with loglogistic. A stepwise dotted black line in the plot indicates the observed event. It is observed that loglogistic MCM does not closely converge to observed data. This implies that the model under estimate the ovarian cancer and this could undermine the clinician’s effort.

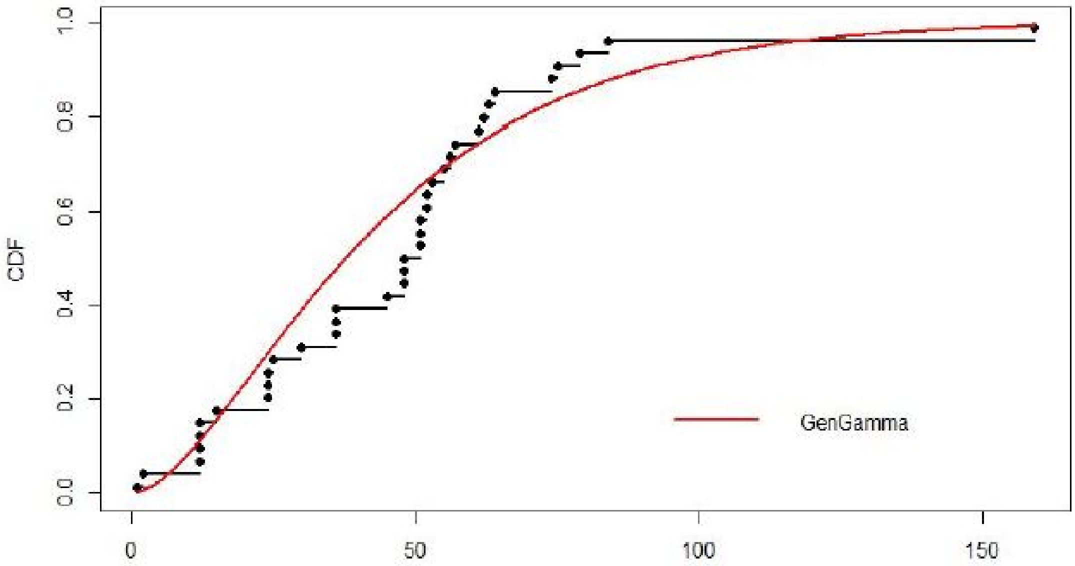


Figure 4.10 Cumulative Density Plot of Generalised - Gamma Mixture Cure Model

The figure 4.10 above gives the cumulative density plot of the ovarian cancer data associated with *Generalised - Gamma* MCM , the plot show the movement from 0 to 1, In cumulative distribution function, the curve is a concave up parabola which lies between $-1 < x \leq 0$ and a concave down parabola which also lies between $0 < x < 1$. Hence, the figure 4.6 gives the curve that is a concave down parabola which lies between $0 < x < 1$. The x-axis signifies time in months, and the y-axis displays the cumulative distribution function of surviving or the percentage of patients likely to enjoy the benefits of medical intervention. The lines represent survival curves of the ovarian cancer associated with *Generalised - Gamma*. A stepwise dotted black line in the plot indicates the observed event. It is observed that *Generalised - Gamma* MCM closely converge to observed data. This implies that the model estimate the ovarian cancer but not as fast as the clinicians expect.

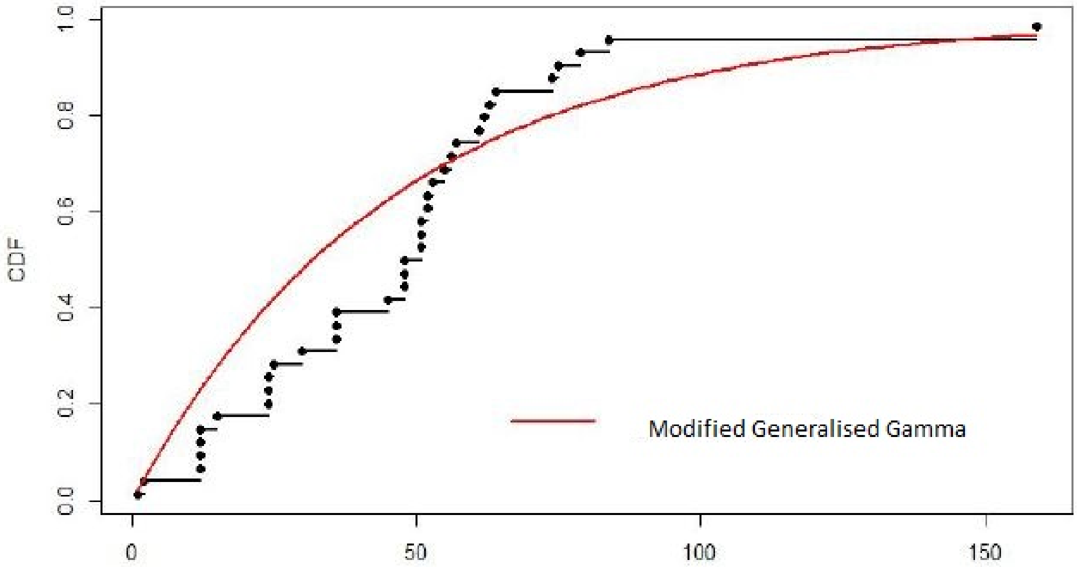


Figure 4.11: Cumulative Density Plot of Modified Generalised-Gamma Mixture Cure Model

The figure 4.11 above gives the cumulative density plot of the ovarian cancer data associated with the proposed model modified Generalised - Gamma MCM, the plot show the movement from 0 to 1, In cumulative distribution function, the curve is a concave up parabola which lies between $-1 < x \leq 0$ and a concave down parabola which also lies between $0 < x < 1$. Hence, the figure 4.6 gives the curve that is a concave down parabola which lies between $0 < x < 1$. The x-axis signifies time in months, and the y-axis displays the cumulative distribution function of surviving or the percentage of patients likely to enjoy the benefits of medical intervention. The lines represent survival curves of the ovarian cancer associated with modified Generalised - Gamma. A stepwise dotted black line in the plot indicates the observed event. It is observed that as fast as the clinician would anticipate, modified Generalised-Gamma MCM converges closely to observed data. This means the model estimates the data on ovarian cancer faster than the models from existing models aside from Generalised – Gamma that is robust when sample size is very small.

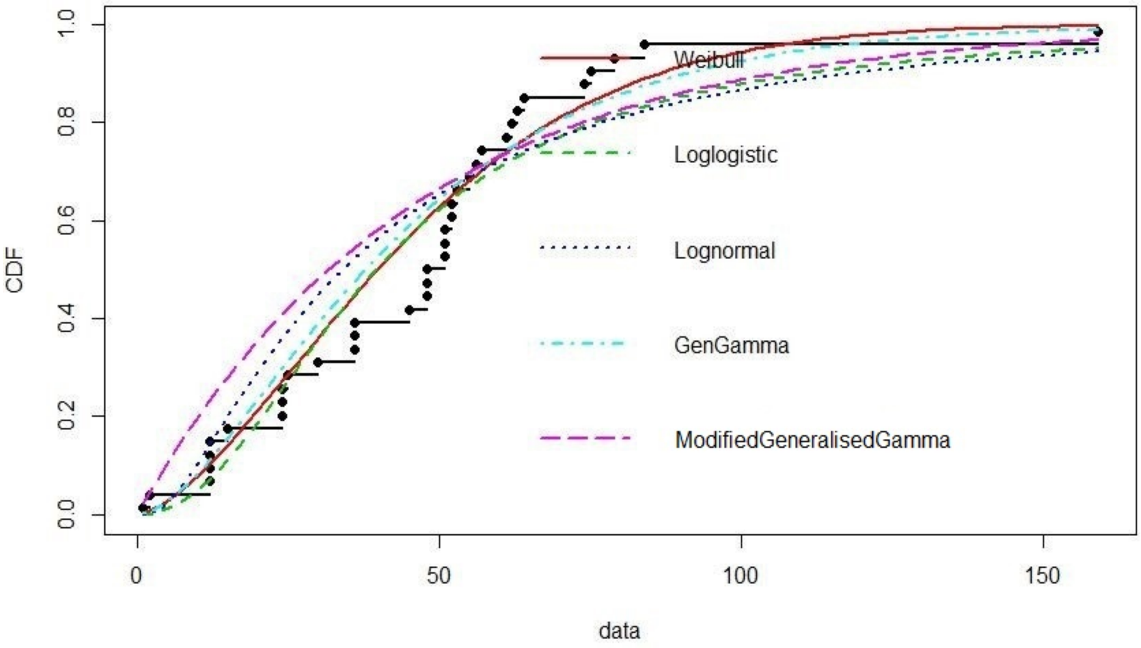


Figure 4.12: Cumulative plots of density for all models

Figure 4.12 above gives the competitive models and the proposed model. From the plots, the proposed MGGMCM is the dash line in lilac colour. The figure shows that MGGMCM converges to one more quickly than convectional models, which means that MGGMCM has outperformed. This implies that the distribution of the proposed model estimate perfectly on the data for all the observed patients. In cumulative distribution function, the curve is a concave up parabola which lies between $-1 < x \leq 0$ and a concave down parabola which also lies between $0 < x < 1$. Hence, the figure 4.6 gives the curve that is a concave down parabola which lies between $0 < x < 1$. The x-axis signifies time in months, and the y-axis displays the cumulative distribution function of surviving or the percentage of patients likely to enjoy the benefits of medical intervention.

The lines represent survival curves of the ovarian cancer associated with all the models. A stepwise dotted black line in the plot indicates the observed event. It is observed that the proposed modified Generalised - Gamma MCM closely converge to observed data as fast as the clinician would expect. This implies that the model estimate the ovarian cancer data faster than the convectional models aside from Generalised – Gamma that is robust when sample size is very small.

4.3.1 Summary of Statistics for Life Data

Table 4.4: Survival Time of Ovarian Cancer: Descriptive Statistics

Min	Q_1	Mean	Median	Q_3	Max	Skewness	Kurtosis
1month	24months	48months	45.65months	57months	159months	1.395	7.339

Table 4.4 above shows the summary of the ovarian cancer data. Based on results gotten from the ovarian cancer data used to validate the proposed MGGMCM, the minimum time of diagnosis of the diseases is 1 month. The average month of diagnosis is 48 months. The results from this descriptive statistics evidenced from histogram plot shows higher degree of departure from normality in the used data. Therefore, mean as a measure of location has to be accompanied by another important and better central of tendency that is useful for survival analysisin this case, is the median.The ovarian cancers survival times were sorted from shortest to longest that is it was ordered in magnitude and value for the median was computed to be 45.65 (in month of diagnosis). When the ovarian cancer was partitioned into four equal parts, the first end point of data is 24 months which gives with the upper third points is 57 points. The estimate of skewness, degree of asymmetry, is 1.395 which indicates that the ovarian cancer data is positively skewed and the kurtosis of 7.339 > 3 indicates that it is leptokurtic. The maximum time of diagnosis of the diseases is 159 months.

In conclusion, the result gives the exploratory data analysis of the ovarian cancer data. This results corroborates with figure 4.1 (histogram plot), figure 4.3 (density plot) and figure 4.4 (boxplot). All these depicts that the distribution of the ovarian cancer is positively skewed. Commensurately, the same table shows that median = 45.65 < mean = 48. This agreed with the fact that when mean > median > mode = positive skewness.

Table 4.5: Model Evaluation of the Ovarian Cancer Data

Model	AIC	-2loglike	μ.˜	σ.	$e\mu$˜.	c	Var(c.)
Weib	216.88450	210.88460	2.44000	68.2000	60.07060	0.28900	0.0968100
Lognor	205.97820	199.97820	4.04000	0.32800	57.97480	0.37120	0.0042610
Llog	203.27440	197.27440	5.97010	56.2070	56.86490	0.18100	0.2878700
GG	206.20140	198.20140	3.96350	0.30170	20.90111	0.10840	0.0052300
MGG	199.23530	194.47120	7.52400	15.2460	11.82010	0.82070	0.0001253

llogis:	LLMCM
Weibull:	WMCM
GG:	GGMCM
MGG:	MGGMCM
$\tilde{\mu}$:	Median
$e\tilde{\mu}$:	Median time-to-cure

The findings in Table 4.5 present the synopsis for evaluation of model on real-life ovarian cancer. In order for assessment of the model performances, we make reference to Akaike Information Criterion (AIC) and the log-likelihood of each of the model examined in this study. As a standard, the model with the lowest AIC is assumed to outperform others. With reference to the same Table 4.5, the value of AIC = 199.23530 and log-likelihood = 194.47120 for the proposed MGGMCM, this result outperformed other models considered in the study.

The interest of cancer study is to achieve a greater proportion of patients that have benefitted from medical intervention. That is proportion of patients that their cancer mass has been reduce to the nearest minimum. c is the proportion of those who are cured that is, it represents the proportion of patients who can benefit from medical intervention.

From the table 4.5 above c associated with all the models are ‘c’ associated with weibull is about 29%, ‘c’ associated with lognormal is almost 37%, c associated with loglogistic is 18%, ‘c’ associated with genaralised-gamma is 10% and ‘c’ associated with Modified generalised-gamma is 82%. This implies that 82.07% of patients can benefit and likely to gain more medical intervention using the proposed model. The existing models underestimate the cure fraction. The results show that the proposed model is optimal in estimating the proportion of those who can benefit from medical intervention.

In survival analysis, $e\mu^{\sim}$ is the median time to cure which gives the Average time we need for those patients who can benefit from medical intervention. The median time which can as well be regarded as recovery time of ovarian patients was discovered to take less than 12 months on the average for patients to be managed or cured using the proposed MGGMCM; this is the least recovery time among the considered models. The proposed model also has the smallest level of dispersion evidenced from variance measured for cure fraction parameter as var(c) = 0.0001253. The result in the table gives the median time to cure which is used for making an inference about the median time that a particular patient will be cured or will be managed, this is the average number of months that the patients can be managed before the occurrence of death.

On a final note, having obtained the least AIC value of the proposed MGGMCM, the least among all its competing frontiers similar models, supports its selection as the best distribution to ovarian cancer data/survival data. The minimum value of var (c) also makes it a better estimate than others.

4.4. Hypothesis Testing (Significance of the cure fraction 'c')

$$H_o: c\ is\ not\ significant$$

$$H_1: c\ is\ significant$$

Or

$$H_o: c = 0$$

$$H_1: c \neq 0$$

$t_{cal} = \frac{c}{s(c)}$ versus $t_{tab} = t_{\propto/_2}$, n – p

$= t_{0.975,37-1}$

$= t_{0.975,36}$

= 1.69

Table 4.6: Testing the Significance of Cure Fraction (c)

Model	**c**	$SE(c)$	$t_{cal} = \frac{c}{s(c)}$	**Remark**
Weibull	0.28900	0.3111431	0.9288	Not Significant (Accept H_0)
Lognormal	0.3712	0.0652763	5.6866	Significant **(Reject H_0)**
Llogis	0.18100	0.5365352	0.3373	Not Significant (Accept H_0)
GG	0.10840	0.0723187	1.4989	Not Significant (Accept H_0)
MGG	0.82070	0.0111938	73.3174	Significant **(Reject H_0)**

In statistics, making inference is the important aspect of the discipline. Therefore, statistical inference can be divided into estimation of parameters, or other characteristics of the density function that has been selected as a model for a random variable, and of testing hypothesis about the model. The parameter of interest in all the considered models is cure fraction, thatis, proportion of diagnosed ovarian cancer patients that were cured, Hence, it is essential to carry out the significance test for this parameter. The table (4,6) above, compute cure fraction "c" foreach model/distribution, its standard error, the value of t-test statistic and the criticalvalue. Under similar situation with the fundamental rule ofstatistical inference, the test is significance if the critical value is less than the test-statisticvalue. In this case, the critical value at 5% significance level is 1:96.

Weibull distributioncomputes the t-test statistic to be 0:9288 a value that is not greater than the critical value of 1:96and consequently falls in feasible region. Hence, the parameter c in question is not significant.That is, its inclusion in the Weibull distribution does not improve its modelling efficiency for theovarian cancer data. Lognormal on the other hand produced test-statistic of 5.6866 which isgreater than 1.96 critical value, then, it is significant. This means, the parameter c is pertinentin Lognormal distribution for modelling ovarian cancer data. Log-logistic distribution andGeneralised-gamma are both not significant under similar argument, as their respective test-statistic values and less than the critical point.

Unlike Log-logistic and Generalised Gamma, theproposed Modified Generalised-Gamma (MGG) distribution is highly significant among the class of compared models as the magnitude of the estimate of its test-statistic = 73:3174 > 1:96 and consequently fall in the critical region, indicating the significance of the test. That is, significance of the test parameter which gives information about its inclusion in the MGG distribution for modelling ovarian cancer data. In summary, the last column of table (4.6) gives a succinct remark about the significance of the parameter c under each distribution. Having estimate the point estimator of c = 0.82070 which is the proportion of patients who can benefit from medical intervention. It is pertinent to check for the efficiency of the estimator. This has been checked using the minimum variance criterion. And from the previous table 4.5 the minimum variance across the model is 0.0001253 which is the variance associated with the proposed model. Conducting a significance test is paramount and the result in the table 4.6 reported that the proposed new model is highly significant followed by ‘c’ associated with lognormal.

Table 4.7 Confidence Interval of Cure Fraction Model (c)

Model	c	Var(c)	SE(c)	$SE(c) \times Z_{\propto/2}$	$c \pm Z_{\propto/2} SE(c)$	CI
Weibull	0.28900	0.0986100	0.3111431	0.098405	0.28900 ± 0.6098405	[-0.3208405, 0.8988405]
Lognormal	0.37120	0.0042610	0.0652763	0.1279415	0.3712 ± 0.1279415	[0.2432585, 0.4991415]
Llogis	0.18100	0.2878700	0.5365352	1.05160900	0.18100 ± 1.05160900	[-0.870609, 1.232609]
GG	0.10840	0.0052300	0.0723187	0.1417447	0.10840 ± 0.1417447	[-0.0333447, 0.2501447]
MGG	0.82070	0.0001253	0.0111938	0.0219398	0.82070 ± 0.0219398	[0.7987602, 0.8426398]

In statistics analysis, the most important benefit of using confidence interval is that you provide a range of values with a known probability of capturing the population parameter that is you can claim to have 95% confidence that it will include the true population parameter. It also helps one not to be so confident that the population value is exactly equal to the single point estimate. That is, it makes us more careful in how we interpret our data and helps keep us in proper perspective. Therefore the table 4.7 above gives the 95% confidence interval of the point estimate ‘c’ for all the considered models. It is clearly observed that there is close margin of c in the proposed model. The margin of error associated with the proposed model MGGMCM is very small compare to the conventional models. The confidence interval of cure fraction with a very low margin of error gives a better one.

4.5 Discussion of Results

It is common in survival analysis that many subjects under study will not experience the event of interest. These subjects are termed “cured”. The cohort is divided into two units namely the cured unit and the uncured unit. It depends on the components: the probability of being cured and the conditional survival function of the susceptible subjects.

In this study, new model was proposed to estimate a mixture cure model when the data are subject to positive high asymmetry. The study used some parametric models for the cure proportion as shown in previous chapters. Modification was done on the generalised gamma mixture cure model and this model is referred to Modified Generalised – Gamma Mixture Cure Model (MGGMCM). Discussing the result of the study in line with the aim of the work, the study has been able to develop a modified generalised-gamma mixture cure model survival that is explicitly competent to regulate and control acute-asymmetry in survival data. The proposed which has been checked under various properties of new distribution has satisfied the condition of a proper density function in previous chapters. Other properties were checked as well.

The study employed a simulation study to validate the performance of the proposed model in terms of the estimators used for the study. The generation of data were done with sample size n = 10, 20 and 50 with 50, 100 and 500 replications using uniform distribution to constraint the censoring rates in the cure model. The choice of the sample size not more than 50 was as a fact that the real life ovarian cancer gotten from UCH was 37 observations and the simulation study must mimick the real life scenario.

From the simulation study, MSE, RMSE and absolute bias were used as the model selection criterion. The criterion may guide the choice of which model among the considered mixture cure models gives the best that can accommodate non-normality in the survival data since skewness is featured in the survival data and our interest is to focus on this. Therefore, all the results from simulation study showcase the efficiency of proposed model and how it outperformed other competing models having used the entire criterion.

The criteria used for the real life ovarian cancer data analysis are loglikelihood, cure rate (c), Akaike Information Criterion (AIC), , variances, mean time-to-cure, var (c) and median survival time. These criteria are used to determine efficiency of the model. The result from the real life ovarian data indicated that the data conform to prior expectation that the distribution of the data is far from normal, and that it is positively skewed. It is also deduced from the result that there is an outlier observation in the ovarian cancer data meaning that there was a particular patient that has survived for 159 month during the period of the study and the same patient is under cancer management at Gyneacology Oncology Unit (GOU) of UCH.

It was evidenced from the real life ovarian cancer data analysis that the model with the least AIC gives the best model. This agreed with Lore*et. al.* (2014) in their work "An Akaike information criterion for multipleevent mixture cure models" where they reported that after calculating the AIC values for each of the considered models, the models weresorted according to their resulting AIC values; the models were examined andcompared, the best models according to the AIC has the least value. The proposed MGGMCM has the least AIC. Also, the loglikelihood of MGGMCM gives the least values compared to the competing models. The var(c) gives the minimum variances among the parametric cure models considered.

From the exploratory data analysis of the real life ovarian cancer, the results comprises of the summary statistics that depicts the description information of the ovarian cancer survival time, this recapitulates the thirty seven life ovarian cancer data used in this research. It always captures the minimum observed data, maximum observation; central tendencies such as mean, median plus measures of spread like quartiles were also reported. The summarization includes other statistic which can illustrate the ovarian cancer information description such as the measure of symmetry such as skewness and kurtosis. Furthermore, all the plots showcase that the data is characterized with acute-asymmetry.

In the result, proposed model MGGMCM performs better when we use cumulative distribution function plot to estimate the real life ovarian cancer data. The plot from the cdf converges to one more quickly than convectional modelsThese results were in agreement with Seppa *et.al.* (2009).In their study, a random effect of mixture cure fraction model was examined. The study was modeled to cause-specific survival data.

The female breast cancer data used to validate the model is from the Finnish Cancer Registry. The study is a population based and it makes use of sets of random effect. This is used to capture the variation in the cure fraction and in the survival of the non-cured patients.

Furthermore**,** other results fromreal life ovarian cancer data showed that model selection criteria for this study established that convectional models were significantly less effective than the modified Gamma-Generalised Mixture Cure Model (MGGMCM) with the underlying assumptions. This indicated that the MGGMCM enriched progress and provides a better fit for the real life ovarian cancer data used in comparing the competing models. Therefore, the result implies that MGGMCM offers least criterion value. It is then established that the MGGMCM is efficient, more consistent and effectively solve acute-asymmetry problem associated with survival data.

The study revealed that the recovery time of ovarian patients was less than 12 months for the proposed MGGMCM. The other existing parametric mixture cure models considered has their patients having longer recovery time. According to this study, patients were experiencing longer recovery time under the existing parametric cure models. The implication of this is that, Cure time refers to the length of time needed for something to fully cure. In practice, the idea of a cure is the permanent end to the specific instance of the disease but in cure model, the cure refers to when a patient recovers from it, the person is said to be cured at that moment. The symptoms might reoccur especially in cancer cases like ovarian cancer. So the beauty of this is that, since researchers interest is on management of the disease which is cure, the time to cure should be another interest of researchers, what duration it takes for cure to occur. This is what i refer to mean time to cure or recovery time. The recovery time should be small in order to have an efficient cure model.

All statistics models cannot be 100% sufficient because they rely on so many assumptions, and any model that has so many assumptions will definitely have some limitations because of violations of assumptions. In that case, in a situation where the level of violations is much, that when there is high level of violations of assumption of models, one needs a robust model that is the main reason for adding the shape parameters in our link functions for the benefit of flexibility. In this study, we have been able to develop a robust point estimator for the proportion of patients who can benefit from medical intervention. A cohort of patients who need medical intervention from the proposed model was about 82%. It is believed that all the patients will not gain medical intervention that is 100% of them cannot, it is expected that the proportion should lie between 0 and 1. It can never be 0 and it can never be 1, when it is 0, it means that no one is benefited from medical intervention and when it is 1, it also means that all the cohort of patients will benefit but in reality, that is not possible.

4.6 Implications from the Study

From the result of this study, the last objectives that is to make suggestions for policy makers, clinicians and academicians. This study has succeeded in developing a robust model that can be used to estimate the proportions of patients that can actually benefit from medical intervention Judging from the value of c, the proposed model has the highest value of c compared to the existing models that underestimate the proportion of c, which in turn will undermine the efforts of the clinicians. With the efforts put forward by the clinicians, we want to know the percentage of patients that they had been able to manage or cure after medical intervention. Within the time frame under which the patients are being managed, those values of ‘c’ that we arrived at are cure fraction. The clinicians have interest in those that are likely to gain from medical intervention and those that are not likely to gain. They need a model that can give the true picture of this proportion. Since cancer data are featured with high degree of asymmetry, we need a modification of the existing parametric cure models that will be flexibly robust to accommodate this feature. The study also reported that the time is going to take a particular patient to be cured is going to vary from patients to patients because patient have different body chemistry, thus, the clinicians also have interest in the recovery time. The recovery time associated with the proposed model on the average is very small compared to the existing models. this is a good one for the clinicians because, after medical intervention, they want a situation where their patients recover on time, they don’t want a situation where large proportion of patients are staying longer than necessary because the pains associated with ovarian cancer is severe. On a final note, the property of the robust estimator cure proportion shows that the variance of c is small and it is efficient. According to Cramer Rao, a lower value of the variance indicates that the estimator is efficient. This is a justification that the proposed is estimating the proportion of patients accurately and there will not be issue for undermining the clinician’s efforts.

CHAPTER FIVE
SUMMARY AND CONCLUSION

5.1 Summary

Based on the findings of this study as well as the facts from the discussion of study outcomes, we can affirm that after all the criterions used to determine the efficiency of the proposed model, it is appropriate to give the necessary information about the values of the criteria used. From the study, the output confirmed that the criteria for determining the flexible best model between the competing models and the proposed models show that the proposed model significantly perform better than the convectional models; having considered the underlying assumptions associated with criterion in question.

As a result of this outcome, the proposed Modified Generalised-Gamma Mixture Cure Model enriched improvement as well as gives appropriate estimates for the real life ovarian cancer data compared to the previous models. As a result of this, proposed model MGGMCM provides minimum numerical value. Also, with simulated data, it has the minimum MSE, RSME and absolute bias. Also, the significance of the cure fraction parameter 'c' associated with theproposed Modified Generalised-Gamma Mixture Cure Model (MGGMCM) is highly significant among competing models.

Furthermore, from confidence interval of cure fraction parameter 'c', the margin of error associated with the proposed model MGGMCM is very small compare to the conventional models. The confidence interval of cure fraction with a very low margin of error gives a better one. Therefore, this research has added to knowledge in terms of remarkable contribution of generalization and improvement on existing GGMCM. Furthermore, according to the results, the context affirmed that modified GGMCM accomplished extra consistent outcomes which can sufficiently control restrictions of non-normality related with survival data and to its robustness.

5.2 Conclusion

As mentioned earlier, this study is an improvement to the solutions associated with survival data analysis that is characterized with the problem of acute-asymmetry. Using the outcomes acquired from both the simulated data and real life data set, we can conclude that MGGMCM provides the model that is robust. This has been confirmed from the results of the analysis, using the real life data for ovary cancer as well as simulation study. The proposed converged to 1 quickly than the existing indicating that the model is very efficient. The MGGMCM have the ability to efficiently estimate sizeable real life ovarian cancer data. The study outcomes displayed that proposed MGGMCM stood as better model aimed at modeling survival data that is characterized with non-normality. Additionally, this new model is applicable for modeling of survival data analysis and appropriate statistical mechanisms to solve various data that exhibits acute-asymmetry. It also provides inference robustness, criterion robustness and model robustness.

In Survival analysis where some patients have not experienced the event of interest, we prefer to use cure models (Jahanjou et al 2014& Smoll et. al. 2012); since many of patients among those diagnosed with Ovary malignancies will not require hysterectomy, these cases assumed to be cured. Moreover, survival data requires more flexible and robust models that would adequately accommodate and isolate intractable distribution characteristics such that would engender complex asymetry, this is the reason for Modiified Generalised Gamma Mixture Cure Model that can embrace or accommodate the high degree of asymmetry that survival data exhibits. We studied the performance of the developed model MGGMCM and compared it with selected conventional models such as LNMCM, LLMCM, WMCM and GGMCM. With the aid of simulation from MCMC and analysis from real life ovarian cancer data gotten from Department of Obstetrics and Gynaecology, University College Hospital, Ibadan, Nigeria covering the period 2000-2015 with the help of R software code. The results showed that our new MGGMCM remained an upgraded statistical model for statistical modeling and statistical inference.

5.3 Limitation of the Study

From the result of the simulated data, we were able to establish that GGMCM gives the same or least in few cases for MSE, RSME, and absolute BIAS when the sample size is 10 and when the replication is 50 times. Contrarily, the proposed model gives the least MSE, RSME, and Absolute BIAS when the sample size is 20 and 50 at each level of the replication. These results indicate that the proposed model is better when the sample size is large. The model will underestimate when the sample size is too small.

5.4 Contribution to Knowledge

According to Chukwu & Folorunso (2015), Generalised-Gamma Mixture Cure Fraction Model has the highest cure proportion among the considered models, it also gives a better fit to the gastric cancer but it cannot capture the skewness that survival data exhibits.

- This work has provided an extension of GGMCM.
- The modified GGMCM was better on the Akaike Information Criterion and other criterion used in this study.
- The proposed model gives the minimum variance for the proportions of patients that can benefit from medical intervention.
- The cure fraction parameter ‘c’ associated with the proposed model is significant.
- The margin of error in the confidence interval for the cure proportion is also small
- It adequately handled limitations of non-normality connected with survival data and to its robustness.

.

5.5 Suggestion for Further Study

For any other researcher that wants to engage and pursue the related study, the researcher should take into account other types of cure model such as non-mixture models. Also, the covariates were not involved in the analysis either in the real life data set that is the ovarian cancer data set or the simulated data set under the right censoring case. This increased the necessity of such research. Also, the clinician should endeavor to collect the entire necessary attribute that might boost the study in terms of patients other details that will serve as the independent variable (covariates). The clinicians should endeavor to continue to do follow up study on the patients and recruit more patients to the study. The four stages of cancer should look into in order to test per stage. The added shape parameter in the proposed model should also be varied.

REFERENCES

Abu-Bakar, M.R., Salah, K.A., Ibrahim, N. A. & Haron, K. 2009. Bayesian Approach for Joint Longitudinal and Time-to-Event Data with Survival Fraction. Bull. *Malays.Math. Sci. Soc.* 32, 75-100.

Achcar, A.J., E.A. Coelho-Barros and J. Mazuchel, 2012. Cure fraction models using mixture and non-mixture models. Tatra Mt. Math. Publ., 51: 1-9.

Adekanbi A. A., Olayemi O., Okolo C. A. Fawole , A. O. , Odukogbe A. A. *et.al*. 2014. Survival of Ovarian Cancer patients in Ibadan: Clinical andpathological factors. *Journal of Obstetrics and Gynaecology*, 34: 57–59.

Akaike H. 1973. Information theory and extension of the maximum likelihood principle. In Petrov B. N. and Saki F. C *Second International Symposium on Information Theory* (pp. 267-281).

Aljawadi B.A.I., Abu-Bakar M.R., Ibrahim N.A., *et.al*. 2012. Parametric Maximum Likelihood Estimation of Cure Fraction Using Interval-Censored Data. *Science Alert.*

Andersen, P.K. and Keiding, N. 1998. *Survival analysis* Encyclopedia of Biostatistics 6. Wiley, pp. 4452-4461

Andersson Therese ML **,** Dickman Paul W**,** Eloranta SandraandLambert Paul C. 2011. Estimating and modelling cure in population-based cancer studies within the framework of flexible parametric survival models*BMC Medical Research Methodology* 2011, 11:96 doi:10.1186/1471-2288-11-96

Balakrishnan N. *et.al.* 2012. EM Algorithm-Based Likelihood Estimation for Some Cure Rate Models. *Journal of Statistical Theory and Practice.*

Balakrishnan N. and Pal Suvra. 2013. Lognormal lifetimes and likelihood-based inference for flexible cure rate models based on COM-Poisson family. *Computational Statistics & Data Analysis* 67, pages 41-67.

Balakrishnan N. and Pal S. 2015. An EM algorithm for the estimation of parameters of a flexible cure rate model with generalised gamma lifetime and model discrimination using likelihood- and information-based methods. *Computational Statistics* 30:1, pages 151-189.

Balakrishnan N. and Pal S. 2015. Likelihood Inference for Flexible Cure Rate Models with Gamma Lifetimes.*Communications in Statistics - Theory and Methods* . Volume 44, 2015 - Issue 19https://doi.org/10.1080/03610926.2014.964807

Barriga Gladys D.C., Cordeiro Gauss M., Dey Dipak K.,. Cancho Vicente G, Louzada Francisco, Suzuki Adriano K. 2018. The Marshall-Olkin generalised gamma distribution. *Communications for Statistical Applications and Methods*. 2018, Vol. 25, No. 3, 245–261https://doi.org/10.29220/CSAM.2018.25.3.245Print ISSN 2287-7843/Online ISSN 2383-4757

Berkson J. and Gage R.P. 1952. Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, 47:501515.

Betensky, R. A. and Schoenfeld, D.A. 2001. Nonparametric Estimation in a cure model with Random cure times. *Biometrics* 57, 282-286.

Bewick V., Cheek L. and Ball J. 2004.Statistics review 12: Survival analysis.*Critical Care* 2004, 8:389-394 (DOI 10.1186/cc2955

Binbing Y, Tiwari R, Cronin K*, et.al.* 2004. Cure fraction estimation from the mixture cure models for grouped survival data. *Stat Med;* 23: 173347.

Blayney J. 2012. *Survival Analysis: An Introduction.*

Boag J.W .1949. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society. Series B (Methodological)*, 11(1):1553, 1949.

Burnham K. and Anderson D. 2004. Multi-model inference: Understanding AIC and BIC in model selection. *Sociological Methods Research*, 33:261 304.

Cancho Vicente G., Ortega Edwin M.M., Barriga Gladys D.C., Hashimoto Elizabeth M.. (2011) The Conway–Maxwell–Poisson-generalised gamma regression model with long-term survivors. *Journal of Statistical Computation and Simulation* 81:11, pages 1461-1481

Cancho, V.G, Ortega E.M.M. and Bolfarine H. 2009. The Log-exponentiated-Weibull Regression Models with Cure Rate. R CRAN *(*https://cran.r-project.org/web/packages/NPHMC/NPHMC.pdf*).*

Chao C. 2013. Advanced Methodology Developments in Mixture Cure Models. University of South Carolina , Ph.D. Thesis Scholar Commons.

Chao C., Songfeng, W., Wenbin, L. *et.al.* 2015. Package 'NPHMC'. Title Sample Size Calculation for the Proportional Hazards Mixture. Cure Model. Version 2.2. Date 2013-09-23. *Description an R-package for calculating sample.*

Chen, M. H., Ibrahim, J. G., and Sinha, D. 1999. A new Bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association*, 94(447), 909919.

Chukwu, A. U. and Folorunso, S.A. 2015. Determinant of flexible Parametric Estimation of Mixture CureFraction Model: An Application of Gastric cancer Data.*West African Journal of Industrial & Academic Research* Vol.15 No.1.

Claeskens, G. I. and Keilegom, V. 2016. The Focused Information Criterion for a Mixture Cure Model. *Institute of Statistics, Biostatistics and Actuarial Sciences* Université catholique de Louvainmaller

Cook A. 2008.Censoring and Truncation; Introduction to Survival Analysis.

Coolen A. 2012. *Principles of Survival Analysis*.

Cooner, F., Banerjee, S., Carlin, B. P. and Sinha, D. 2007. Flexible Cure Rate Modelling under Latent Activation Schemes. *J. Amer. Statist*. Assoc. 102 560572.

Datta A. 2013. *A Study of the Cure Rate Model with Case Weights and Time-Dependent Weights*. Guelph, Ontario, Canada.

Elangovan, R. and Jayakumar, B. 2016. Cure Rate Models. *Asia Pacific Journal of Research Vol: Issue XXXVIII.*

Fieller N. 2010. *Survival Analysis* Course Booklet. Department of Probability and Statistics. University of Sheffield.

Gallardo D.I., Hctor W. Gmez and Heleno Bolfarine. 2017. A new cure rate model based on the YuleSimon distribution with application to a melanoma data set. *Journal of Applied Statistics 44:7, pages 1153-1164.*

Hsu, W., Todem,, D. and Kim K. 2016. A Sup-Score Test for the Cure Fraction in Mixture Models forLong-Term Survivors. *Biometrics* doi: 10.1111/biom.12514

Ibrahim J.G, and Chen M.H 2000. Power Prior Distributions for Regression Models. *Statistical Science*. 15:46–60.

Ibrahim J.G, Chen M.H. and Sinha D. Bayesian Semi-Parametric Models for Survival Data with a Cure Fraction. *Biometrics*. 2001b; 57: 383–388.

Ibrahim J.G, Chen M.H. and Sinha D.2001. Bayesian Survival Analysis. *Springer Series in Statistics. Springer*-Verlag, New York.

Judy, P. S. and Jeremy, M. G. T. 2004. Estimation in a Cox Proportional Hazards Cure Model. *Biometrics*https://doi.org/10.1111/j.0006-341X.2000.00227

Kaplan, E.L. & Meier, P. 1958. *Non-parametric estimation from incomplete observations*, J American Stats Assn. 53, pp. 457–481, 562–563.

Lambert, P, Thompson, J.R. and Weston, C.L. 2006. Estimating and modeling the cure fraction in population based cancer survival analysis*, Biostatistics, 8, 576-594.*

Lambert, P 2007. Modeling of the cure fraction in survival studies. *Stata J;7:125*

Lambert, P, Dickman, P and Osterlund, P, 2007. Temporal trends in the proportion cured for cancer of the colon and rectum: a population based study using data from the finish cancer registry. *Int J Cancer;121:20529.*

Lore D., Gerda C., Bart B. 2014. *An Akaike information criterion for multipleevent mixture cure models.*

LucijanicMarko and PetroveckiMladen 2012. Analysis of censored data. Biochemia Medica 22(2):151-5, DOI: 10.11613/BM.2012.018

Maetani, S. and Gamel, J. 2013. Parametric Cure Model versus Proportional Hazards Model in Survival Analysis of Breast Cancer and Other Malignancies. *Advances in Breast Cancer Research*, **2**, 119-125. doi: 10.4236/abcr.2013.24020.

Maller, R.A. and X. Zhou, 1996. Survival Analysis with Long-Term Survivors. Wiley, New York, ISBN: 9780471962014, Pages: 278.

National Cancer Institute, 2018. *NCI Dictionary of Cancer Terms.*

Odukogbe, A.A, Adewole, I.F., Ojengbede, O.A., Olayemi, O.,Oladokun, A., *et.al.* 2001. Grand-multi-parity trends and complications: a study in two hospital settings. *Journal of Obstetrics and Gynaecology*, *21, 361-367.*

Odukogbe, A.A., Adebamowo ,C.A., Ola, B., Olayemi, O.,Oladokun, A., Adewole, I.F., *et.al.* 2004. Ovarian cancer in Ibadan: characteristics and management. *Journal of Obstetrics and Gynaecology*, vol.24 N.3, 294-297.

Ortega Edwin M. M., Cancho Vicente G. and Lachos Victor Hugo 2008. Assessing influence in survival data with a cure fraction and covariates. SORT 32 (2) July-December 2008, 115-140

Ortega E. M .M., Barriga G. D. C., Hashimoto E. M, *et.al.* 2014. A New Class of Survival Regression Models with Cure Fraction. A perspective. *Eur J Cancer; 45:106779.*

Patilea V. and Keilegom I.V. 2017. *A General Approach for Cure Models in Survival Analysis.*

Ranganathan R., Rajaraman S. and Perumal V. 2010. Cure Models for Estimating Hospital-Based Breast Cancer Survival. *Asian Pacific journal of cancer prevention. VL – 11*

Ristic, M.M. and Balakrishnan, N. 2011. The gamma-exponentiated exponential distribution. *J. Statist. Comput. Simulation.*

doi:10.1080/00949655.2011.574633.

Rodrigues Josemar, Cancho Vicente G, Castro Mario de, Balakrishnan N. (2012) A Bayesian destructive weighted Poisson cure rate model and an application to a cutaneous melanoma data. *Statistical Methods in Medical Research* 21:6, pages 585-597.

Rodrigues, J., Cancho, V.G., De Castro, M. and Louzada-Neto, F., 2009. On the unification of the long-term survival models. Statistics and Probability. *Letters 79,753759.*

Roynette Bernard, Vallois Pierre, Yor Marc. 2009. A family of generalised gamma convoluted variables.Probability and Mathematical Statistics, 29 (2), pp.181-204. hal-00292334

Sebah Pascal and Gourdon Xavier. 2002. Introduction to the Gamma Function numbers computation free fr / Constants /.html

Shuangge, M. A. 2009. Cure Model with Current Status Data. *Statistica Sinica 19 , 233-249*

Smoll, N.R., Schaller, K. and Gautschi, O.P. 2012. The cure Fraction of Glioblastoma Multiforme *Neuroepidemiology* 2012;39:63–69 https://doi.org/10.1159/000339319

Sposto, R. 2002. Cure model analysis in cancer: an application to data from the Children Cancer Group. *Stat Med;21:293312.*

Taweab, F., Ibrahim, N. A. and Arasan J. 2015. A Bounded Cumulative Hazard Model with A change- Point According to a Threshold in a Covariate for Right-Censored Data

Tsodikov A, Loeffler M and Yakovlev A 1998. A Cure Model with Time-Changing Risk Factor: An Application to the Analysis of Secondary Leukemia. A Report from the International Database on Hodgkin's Disease. *Statistics in Medicine*.;17:27–40.PubMed

Walck C, 2007, *Hand-book on Statistical Distribution for Experimentalists*, Particle Physics Group, Fysikum University of Stockholm.

Wang M. 2006. Summary Notes for Survival Analysis. Department of BiostatisticsJohns Hopkins University.

Wenbin, L. 2010.Efficient Estimation for an Accelerated Failure Time Model with a Cure Fraction. *Statistica Sinica 20 661-674*

Wienke, A., Lichtenstein, P. and Yashin, A.I. 2003. A Bivariate Frailty Model With A Cure Fraction Modeling Familial Correlations In Diseases, *Biometrics* 59,1178-1189

Wolsztynski E. 2015. *ST3054/ST6004 - Survival Analysis*. Department of Statistics School of Mathematical Sciences University College Cork, Ireland.

Yakovlev, A.Y and Tsodikov, A.D. 1996. Stochastic Models of Tumor Latency and Their Biostatistical Applications. *World Scientific, Singapore.*

Yakovlev, A.Y. 1994. Parametric versus nonparametric methods for estimating cure rates based on censored survival-data. *Statistics in Medicine 13 (9), 983985*

Yakovlev, A.Y., Tsodikov, A.D. and Bass, L., 1993. A stochastic-model of hormesis. *Mathematical Biosciences 116 (2), 197219.*

Yi Li and Tiwari. 2007. *Mixture cure survival models with dependent censoring*.

Yigzaw A. L. and Demeke L. W. 2019. Survival analysis of time to cure on multi-drug resistance tuberculosis patients in Amhara region, Ethiopia *BMC Public Health* 19:165 https://doi.org/10.1186/s12889-019-6500-3

Yin, G. and Ibrahim, J.G., 2005. Cure rate models: a unified approach. *The CanadianJournal of Statistics 33 (4), 559570*

Yingwei P. and Keith B. G. 2000. A Nonparametric Mixture Model for Cure Rate Estimation *Biometrics* Vol. 56, No. 1 (Mar., 2000), pp. 237-243 https://www.jstor.org/stable/2677127

Yu-Gu, D. S. and Banerjee S.2010. *Analysis of Cure Rate Survival Data under Proportional Odds Model.*

Zhao, G.M.A. 2008. *Nonparametric and Parametric Survival Analysis of Censored Data with Possible Violation of Method Assumptions.*

Zografos, K. and Balakrishnan, N. 2009. On families of beta- and generalised gamma generated distributions and associated inference. *Stat. Method., 6, 344-362.*

APPENDICES

a. R Code used for Exploratory Data Analysis.

b. R Code for used for the Analysis

c. R code for Simulation Study

Appendix A

R Code for Exploratory Data Analysis

```
a=200 b=10 c=12 d=1.3 x=seq(0,5,0.01) bww1.pdf=function(x,a,b,c,d){ k1=(1-exp(-
x^d)) k2=(1-exp(-(1+c)*x^d))/(c+1) k3=((c+1)/c)*k1 k6=(1-exp(-c*x^d))
k7=((c+1)/c)*d*x^(d-1)*exp(-x^d) k8=(k3-k2)^(a-1) k9=(1-(k3-k2))^(b-1) k4=beta(a,b)
k5=1/k4 bww1.pdf=k5*k8*k9*k7*k6 }
plot(x,bww1.pdf(x,200,10,12,1.3),col="red",ylim=c(0,3.5),type="p
",xlab="x",ylab="pdf BWWD") lines(x,bww1.pdf(x,150,8,12,1.3),col="blue",lty=2)
lines(x,bww1.pdf(x,100,6,12,1.3),col="black",lty=4)
lines(x,bww1.pdf(x,10,5,12,1.3),col="green",lty=6)
legend("topright",inset=0.02,col=c("red","blue","black","green")
,legend=c("a=200,b=10,c=12,d=1.3","a=150,b=8,c=12,d=1.3","a=100,
b=6,c=12,d=1.3","a=10,b=5,c=12,d=1.3"),lty=1:2:4)
```

When a = 1

```
a=1 b=10 c=12 d=1.3 x=seq(0,5,0.01) bww1.pdf=function(x,a,b,c,d){ k1=(1-exp(-x^d))
k2=(1-exp(-(1+c)*x^d))/(c+1) k3=((c+1)/c)*k1 k6=(1-exp(-c*x^d))
k7=((c+1)/c)*d*x^(d-1)*exp(-x^d) k8=(k3-k2)^(a-1) k9=(1-(k3-k2))^(b-1) k4=beta(a,b)
k5=1/k4 bww1.pdf=k5*k8*k9*k7*k6 }
plot(x,bww1.pdf(x,1,10,12,1.3),col="red",ylim=c(0,3.5),type="p", xlab="x",ylab="pdf
LWWD (when a = 1) " ) lines(x,bww1.pdf(x,1,8,12,1.3),col="blue",lty=2)
lines(x,bww1.pdf(x,1,6,12,1.3),col="black",lty=4)
lines(x,bww1.pdf(x,1,5,12,1.3),col="green",lty=6)
legend("topright",inset=0.02,col=c("red","blue","black","green")
,legend=c("a=1,b=10,c=12,d=1.3","a=1,b=8,c=12,d=1.3","a=1,b=6,c=
12 ,d=1.3","a=1,b=5,c=12,d=1.3"),lty =1:2:4)
```

When b = 1

```
a=200 b=1 c=12 d=1.3 x=seq(0,5,0.01) bww1.pdf=function(x,a,b,c,d){ k1=(1-exp(-
x^d)) k2=(1-exp(-(1+c)*x^d))/(c+1) k3=((c+1)/c)*k1 k6=(1-exp(-c*x^d))
k7=((c+1)/c)*d*x^(d-1)*exp(-x^d) k8=(k3-k2)^(a-1) k9=(1-(k3-k2))^(b-1) k4=beta(a,b)
k5=1/k4 bww1.pdf=k5*k8*k9*k7*k6 }
plot(x,bww1.pdf(x,200,1,12,1.3),col="red",ylim=c(0,3.5),type="p"
,xlab="x",ylab="pdf EWWD (when b = 1) " )
lines(x,bww1.pdf(x,150,1,12,1.3),col="blue",lty=2)
lines(x,bww1.pdf(x,100,1,12,1.3),col="black",lty=4)
lines(x,bww1.pdf(x,10,1,12,1.3),col="green",lty=6)
```

```
legend("topleft",inset=0.02,col=c("red","blue","black","green"),
legend=c("a=200,b=1,c=12,d=1.3","a=150,b=1,c=12,d=1.3","a=100,b= 1
,c=12,d=1.3","a=10,b=1,c=12,d=1.3"),lty =1:2:4)
```

When a = b = 1

```
a=1 b=1 c=12 d=1.3 x=seq(0,5,0.01) bww1.pdf=function(x,a,b,c,d){ k1=(1-exp(-x^d))
k2=(1-exp(-(1+c)*x^d))/(c+1) k3=((c+1)/c)*k1 k6=(1-exp(-c*x^d))
k7=((c+1)/c)*d*x^(d-1)*exp(-x^d) k8=(k3-k2)^(a-1) k9=(1-(k3-k2))^(b-1) k4=beta(a,b)
k5=1/k4 bww1.pdf=k5*k8*k9*k7*k6 }
plot(x,bww1.pdf(x,1,1,12,1.3),col="red",ylim=c(0,3.5),type="p",x lab="x",ylab="pdf
WW (when a = b = 1) " )
lines(x,bww1.pdf(x,1,1,12,1.3),col="blue",lty=2)
lines(x,bww1.pdf(x,1,1,12,1.3),col="black",lty=4)
lines(x,bww1.pdf(x,1,1,12,1.3),col="green",lty=6)
legend("topleft",inset=0.02,col=c("red","blue","black","green"),
legend=c("a=1,b=1,c=12,d=1.3","a=1,b=1,c=12,d=1.3","
a=1,b=1,c=12,d=1.3","a=1,b=1,c=12,d=1.3"),lty=1:2:4)
```

Appendix B

R Code for the Analysis

```
library(tidyverse) library(readxl)
library(maxLik) library(survival)
library(gsl) library(flexsurvcure)
library(smcure) library(car)
library(gtools) library(xtable)
require("knitr")
require("markdown")
require("rattle")
require("xtable")
require("stringr")
require("fBasics")
require("MASS")
require("survival")
require("STAR")
require("gamlss.dist")
require("VGAM") getwd()
setwd("C:/Users/FOLORUNSO/
Desktop/AlhajaSF") cerv <-
read.csv('OvarianC.csv',
header=T) attach(cerv)
  General Log-Likelihood Function cureloglike <- function(stime,d,dist){ distri <- dist
function(p){ params <- p alpha <- params[1] if(distri=="lnorm"){
mu <- params[2] ; sigma <- params[3] fu <- dlnorm(stime,mu,sigma) su <- 1 -
plnorm(stime,mu,sigma) }else if(distri=="weibull"){ sh <- params[2] ; sc <- params[3]
fu <- dweibull(stime,sh,sc) su <- 1 - pweibull(stime,sh,sc) }else if(distri=="llogis"){ sh
<- params[2] ; sc <- params[3] fu <- dlogis(log(stime),log(sh),1/sc) su <- 1 -
plogis(log(stime),log(sh),1/sc)
}else if(distri=="gengamma"){ beta <- params[2] ; a <- params[3]; teta <- params[4]
 a1 <- beta/(teta*gamma(a))
 a2 <- (stime/teta)^(a*teta - 1) a3 <- exp(-(stime/teta)^beta) fu <-
dgengamma.orig(stime,shape=beta,scale=a,k=teta) su <- 1 -
pgamma(log(stime/a),shape=teta,scale=beta) su <- 1 -
pgengamma.orig(stime,shape=beta,scale=a,k=teta)
} else if(distri=="gamgengengam"){ work in progress}
else stop("Error:Undefined Distribution given")
cure <- exp(alpha)/(1 + exp(alpha)) ft <- (1 - cure)*fu st <- cure + (1 - cure)*su loglike
<- sum(log(ft^d * st^(1-d) )) return(-loglike) }
}
  Assuming the Censoring Indicator Equals Cure Indicator (Not a Good Option)
  Lognormal Cure clm <-
cureloglike(stime=cerv$Ovarian,d=cerv$Censoring,dist="lnorm") clm_results <-
nlm(clm, p=c(alpha=-2,mu=2.1,sigma=0.5), hessian=T,iterlim=1000) aiclognorm <- (-
2*-clm_results$minimum) + (2 *length(clm_results$estimate ))
 serror for meanlog and sdlog serror_lnorm <-
sqrt(diag(solve(clm_results$hessian)))[2:3] names(clm_results$estimate) <-
c("alpha","mu","sigma") alpha<- -291.0669819 cure_lnorm <-
exp(alpha)/(1+exp(alpha));cure_lnorm cure_lnorm <-
deltaMethod(clm_results$estimate, "exp(alpha)/(1+exp(alpha))")
  Survival Plot of lognormal Cure Model ( Ovarian ) fit<-survfit(Surv(Ovarian,
Censoring) ~ Group, data = cerv) plot(fit,lty = 2:3,
xlab="Time",ylab="S(t)",main="Survival Plot (Ovarian Cancer)") stsl <-
clm_results$estimate[1] + (1- clm_results$estimate[1])*(1-plnorm(cerv[,1],
clm_results$estimate[2],clm_results$estimate[3])) lines(Ovarian,stsl,col=3,lwd=2)
legend("topright",c("Kaplan-Meier","Lognormal Cure"),col=c(1,3),lwd=2)

  Weibull Results cwe <-
cureloglike(stime=cerv$Ovarian,d=cerv$Censoring,dist="weibull") cwe_results <-
nlm(cwe,p=c(alpha=0.2,sh=0.5,sc=5), hessian=T,iterlim=1000) aicws <- (-2*-
cwe_results$minimum) + (2 *length(cwe_results$estimate ))
 obtaining cure fraction and its serror serror_weibull <-
sqrt(diag(solve(cwe_results$hessian)))[2:3] names(cwe_results$estimate) <-
c("alpha","shape","scale") alpha<- -132.543320 cure_wei <-
exp(alpha)/(1+exp(alpha));cure_wei
  Survival Plot of weibull Cure Model ( Ovarian ) fit<-survfit(Surv(Ovarian,
Censoring) ~ Group, data = cerv) plot(fit,lty = 2:3, xlab="Time",ylab="S(t)") stwei <-
```

```
cwe_results$estimate[1] + (1- cwe_results$estimate[1])*(1-pweibull(cerv[,1],
cwe_results$estimate[2],cwe_results$estimate[3])) lines(Ovarian,stwei,col=2,lwd=2)
legend("topright",c("Kaplan-Meier","weibull Cure"),col=c(1,2),lwd=2)
new_data0 <- cerv[(cerv$Censoring==0),1] new_data1 <- cerv[(cerv$Censoring==1),1]
para1 <- fitdistr(new_data0,"weibull")$estimate cureloglike <-
function(new_data0,new_data1,fixed=c(F,F,F)){ params <- fixed function(p){
params[!fixed] <- p c <- params[1] alpha <- params[2] beta <- params[3] a <-
alpha/(beta^alpha) b <- new_data1^(alpha-1)
ft <- a * b*exp(-(new_data1/beta)^alpha) st <- 1 - pweibull(new_data0,alpha,beta) d <-
sum(log((1-c)*ft)) e <- sum(log(c + ((1-c) *st ))) loglike <- -( d+e ) loglike } } nLL <-
cureloglike(new_data0,new_data1) results1 <- nlm(nLL,c(c = 0.5,alpha = para1[1] ,beta
= para1[2])
,hessian=T,iterlim=1000)
fit<-survfit(Surv(Ovarian, Censoring) ~ Group, data = cerv) plot(fit,lty = 2:3,
xlab="Time",ylab="S(t)") stwc <- results1$estimate[1] + (1- results1$estimate[1])*(1-
pweibull(cerv[,1], results1$estimate[2],results1$estimate[3]))
lines(Ovarian,stwc,col=2,lwd=2) legend("topright",c("Kaplan-Meier","Weibull
Cure"),col=c(1,2),lwd=2)
  Log-Logistic Results cllogis <-
cureloglike(stime=cerv$Ovarian,d=cerv$Censoring,dist="llogis") cllogis_results <-
nlm(cllogis,p=c(alpha=-2,sh=2,sc=20), hessian=T,iterlim=1000) aicllog <- (-2*-
cllogis_results$minimum) + (2 *length(cllogis_results$estimate ))
 obtaining cure fraction and its serror serror_llogis <-
sqrt(diag(solve(cllogis_results$hessian)))[2:3] names(cllogis_results$estimate) <-
c("alpha","scale","shape") alpha<- -21.321942 cure_llog <-
exp(alpha)/(1+exp(alpha));cure_llog
fit<-survfit(Surv(Ovarian, Censoring) ~ Group, data = cerv) plot(fit,lty = 2:3,
xlab="Time",ylab="S(t)") stll <- cllogis_results$estimate[1] + (1-
cllogis_results$estimate[1])*(1-pllogis
( cerv[,1],cllogis_results$estimate[2],cllogis_results$estimate [3]))
lines(Ovarian,stll,col=4,lwd=2) legend("topright",c("Kaplan-Meier","LogLogistic
Cure"),col=c(1,4),lwd=2)
 new_data0 <- cerv[(cerv$Censoring==0),1] new_data1 <-
cerv[(cerv$Censoring==1),1] para1 <- fitdistr(new_data0,"llogis")$estimate para1 <-
```

```
llogisMLE(new_data0)$estimate cureloglike <-
function(new_data0,newdata1,fixed=c(F,F,F)){ params <- fixed
function(p){ params[!fixed] <- p c <- params[1] alpha <- params[2] beta <- params[3] a
<- ( alpha/beta)*((new_data1/beta)^(alpha -1)) b <- (1+(( new_data1/beta)^alpha))^ 2 ft
<- a/b st <- 1 - pllogis(new_data0,alpha,beta) d <- sum(log((1-c)*ft)) e <- sum(log(c +
((1-c) *st ))) loglike <- -( d+e ) loglike } } nLL <- cureloglike(new_data0,new_data1)
results1 <- nlm(nLL,c(c = 0.5,alpha = 0.075 ,beta = 0.2) ,hessian=T,iterlim =1000) fit<-
survfit(Surv(Ovarian, Censoring) ~ Group, data = cerv) plot(fit,lty = 2:3,
xlab="Time",ylab="S(t)") stll <- results1$estimate[1] + (1- results1$estimate[1])*(1-
pllogis(cerv[,1], results1$estimate[2],results1$estimate[3]))
lines(Ovarian,stll,col=4,lwd=2) legend("topright",c("Kaplan-Meier","LogLogistic
Cure"),col=c(1,4),lwd=2)
  Generalised Gamma Results clgengam <-
cureloglike(stime=cerv$Ovarian,d=cerv$Censoring,dist="gengamma")
clgengam_results <- nlm(clgengam,p=c(alpha=-3,a=0.2,beta=0.02,teta=6),
hessian=T,iterlim=1500,steptol=1e-5, gradtol=1e-6) aicgengam <- (-2*-
clgengam_results$minimum) + (2 *length(clgengam_results$estimate ))
 obtaining cure fraction and its serror
serror_gengamma <- sqrt(diag(solve(clgengam_results$hessian)))[2:4]
names(clgengam_results$estimate) <- c("alpha","shape","scale","teta") alpha<- -
5.839908e+02 cure_GG <- exp(alpha)/(1+exp(alpha));cure_GG
fit<-survfit(Surv(Ovarian, Censoring) ~ Group, data = cerv) plot(fit,lty = 2:3,
xlab="Time",ylab="S(t)") stgg <- clgengam_results$estimate[1] + (1-
clgengam_results$estimate [1])*
(1- pgengamma.stacy(cerv[,1],clgengam_results$estimate[2],clgengam_results$estimate
[3])) lines(Ovarian,stgg,col=5,lwd=2) legend("topright",c("Kaplan-Meier","GG
Cure"),col=c(1,5),lwd=2)
  new_data0 <- cerv[(cerv$Censoring==0),1] new_data1 <-
cerv[(cerv$Censoring==1),1] k <- exp(-1); Scale <- exp(1) gdata <-
data.frame(y=new_data0) para1 <-
coef(vglm(y~1,gengamma.stacy,gdata,shape=k,scale=Scale)) cureloglike <-
function(new_data0,newdata1,fixed=c(F,F,F,F)){ params <- fixed function(p){
params[!fixed] <- p c <- params[1] aa <- params[2] dd <- params[3] pp <- params[4] a
<- ( pp/(aa^dd))/gamma(dd/pp ) b <- ( new_data1^(dd-1))*(exp(-new_data1/aa)^pp ) ft
<- a*b st <- 1 - pgengamma.stacy(new_data0,aa,dd,pp) d <- sum(log((1-c)*ft)) e <-
sum(log(c + ((1-c) *st ))) loglike <- -( d+e ) loglike } } nLL <-
cureloglike(new_data0,new_data1) results1 <- optim(c(c = 0.5,0.5,0.5,0.5),
nLL,method="SANN",hessian=T, control=list(maxit=10000)) fit<-
survfit(Surv(Ovarian, Censoring) ~ Group, data = cerv) plot(fit,lty = 2:3,
xlab="Time",ylab="S(t)") stggs <- results1$estimate[1] + (1- results1$estimate[1])*(1-
pgengamma.stacy(cerv[,1], results1$estimate[2],results1$estimate[3]))
lines(Ovarian,stggs,col=5,lwd=2) legend("topright",c("Kaplan-Meier","GG
Cure"),col=c(1,5),lwd=2)
fit<-survfit(Surv(Ovarian, Censoring) ~ Group, data = cerv) plot(fit,lty = 2:3,
xlab="Time",ylab="S(t)") lines(cerv[,1],stsl,col=3,lwd=2)
lines(cerv[,1],stwei,col=2,lwd=2) lines(cerv[,1],stll,col=4,lwd=2)
lines(cerv[,1],stggs,col=5,lwd=2) legend("topright",c("Kaplan-
Meier","lnorm","weibull","llogis","GG"),col=c(1,3,2,4,5), lwd=c(1,2,2,2,2))
library(actuar)
fitf1 <- fitdist(Ovarian, "weibull") summary(fitf1) fitf2 <- fitdist(Ovarian,
"llogis",start=list(shape=1,scale=500)) summary(fitf2) fitf3 <- fitdist(Ovarian, "lnorm")
summary(fitf3) fitf4 <- fitdist(Ovarian, "gamma") summary(fitf4) fitf5 <-
fitdist(Ovarian, "pareto", start=list(shape=1,scale=500)) summary(fitf5) cdfcomp(fitf2,
lwd=2, legendtext="Loglogistic") cdfcomp(fitf3, lwd=2, legendtext="Lognormal")
cdfcomp(fitf4, lwd=2, legendtext="GenGamma") cdfcomp(fitf5, lwd=2,
legendtext="GammaGenGamma")
```

Modified Generalised Gamma

```
setwd("C:/Users/FOLORUNSO/Desk
top/AlhajaSF") cerv <-
read.csv('OvarianC.csv', header=T)
attach(cerv)
new_data0 <- cerv[(cerv$Censoring==0),1] new_data1 <- cerv[(cerv$Censoring==1),1]
k <- exp(-1); Scale <- exp(1) gdata <- data.frame(y=new_data0) para1 <-
coef(vglm(y~1,gengamma.stacy,gdata,shape=k,scale=Scale)) cureModelloglike <-
function(new_data0,new_data1,fixed=c(F,F,F,F,F)){ params <- fixed function(p){
params[!fixed] <- p a <- params[1] b <- params[2] mu <- params[3] sigma <- params[4]
m <- (1/gamma(b))*(-log(1 - dgamma(new_data1,a,exp((log(new_data1)-
mu)/sigma))))^(b-1) v <- (1 /sigma*gamma(a))*(exp(a*((log(new_data1)-mu)/sigma)-
exp((log(new_data1)-mu)/sigma )) fu <- m*v su <- 1 - ((dgamma((-log(1 -
dgamma(new_data1,a, exp((log(new_data1)-mu)/sigma)))), b))
*(gamma(-log(dgamma(new_data1,a, exp((log(new_data1)-mu)/sigma)))))/gamma(b))
cure <- exp(a)/(1 + exp(a)) ft <- (1 - cure)*(fu)
st <- cure + (1 - cure)*su loglike <-
sum(log(ft^(new_data1) * st^(1-(new_data1))))
return(-loglike)
} } nLL <- cureModelloglike(new_data0,new_data1) results11 <-
optim(c(c = 0.5,0.02,0.1,0.075), nLL,method="SANN",hessian=T,
control=list(maxit=10000))
 results1 <- nlm(nLL,p=c(a=100, b=121,mu =520,sigma = 102),
hessian=T,iterlim=1000) se <- sqrt(diag(solve(results1$hessian))) aicggs <- (-
2*results1$value) + (2 *length(results1$par ))
fit<-survfit(Surv(Ovarian, Censoring) ~ Group, data = cerv) plot(fit,lty
= 2:3, xlab="Time",ylab="S(t)") plot.survival(fit,lty = 2:3,
xlab="Time",ylab="S(t)") stggs <- results1$estimate[1] + (1-
results1$estimate[1])*(1-pgengamma.stacy
( cerv[,1],results1$estimate[2],results1$estimate [3]))
lines(Ovarian,stggs,col=5,lwd=2) legend("topright",c("Kaplan-
Meier","GG Cure"),col=c(1,5),lwd=2)
```

Appendix C

R Code for Simulation

```
 Simulation Study
 set.seed(123)
Ovarian10_50 <- matrix(replicate(50,
runif(10,1,100)), 10) censtimes <- 3 + 20 *runif
(10) ztimes <- pmin(Ovarian10_50, censtimes)
status <- ifelse(censtimes < Ovarian10_50, 1, 0)
;status simdata10<- data.frame(Ovarian10_50,
censoring);simdata10 write.csv(Ovarian10_50,
file="Ovarian10_50.csv")
exp.fit <-
survreg(Surv(cerv[,1],cerv[,2])~1,dist="weib",scale=1)
coeff <- exp.fit$coeff   muhat var <- exp.fit$var thetahat <-
exp(coeff)   exp(muhat) thetahat muhat <- exp.fit$coeff
model Evaluation bias.est=thetahat-cerv[,1]  subtract y
values bias=mean(bias.est)  average over all x values
bias2=bias^2  square
MSE<-var+bias2
RMSE<-sqrt(MSE)
cerv.u <-
cerv[,1][cerv$Censoring=
=1] nu <- length(cerv.u)
scalehat <-
rep(exp(muhat),nu)
Shat <- 1 - pweibull(cerv.u,1,scalehat)
LCL <- exp(log(Shat)*exp(1.96/sqrt(nu)))
UCL <- exp(log(Shat)*exp(-
1.96/sqrt(nu))) C.I.Shat <-
data.frame(cerv.u,Shat,LCL,UCL)
round(C.I.Shat,5)
qq.weibull(Surv(cerv[,1],cerv[,2]),scale
=1)
```

```
 Weibull fit weib.fit <-
survreg(Surv(cerv[,1],cerv[,2])~1,dist="weib")
summary(weib.fit)
 Estimated median along with a 95% C.I. (in months).
medhat <-
predict(weib.fit,type="uquantile",p=0.5,se.fit=T)
medhat1 <- medhat$fit[1] medhat1.se <-
medhat$se.fit[1] exp(medhat1)
C.I.median1 <- c(exp(medhat1),exp(medhat1-
1.96*medhat1.se), exp(medhat1+1.96*medhat1.se))
names(C.I.median1) <- c("median1","LCL","UCL")
C.I.median1
 qq.weibull(Surv(cerv[,1],cerv[,2]))
 weib.fit0 <- survReg(Surv(weeks,status) ~ 1 ,dist="weib" )
 summary(weib.fit0) weib.fit1 <-
survreg(Surv(cerv[,1],cerv[,2]) ~ Group,dist="weib")
summary(weib.fit1) weib.fit1$linear.predictors
weib.fit20 <-
survReg(Surv(weeks,status) ~ 1 ,
data=aml[aml$group==0,],dist="wei
b") weib.fit21 <-
survReg(Surv(weeks,status) ~ 1 ,
data=aml[aml$group==1,],dist="wei
b")
 lognorm fit.lognorm <-
survreg(Surv(cerv[,1],cerv[,2])~1,dist="lognormal")
qq.reg.resid.r(aml1,aml1$weeks,aml1$status,fit.lognorm,"q
norm", "standard normal quantile")
 Estimated median along with a 95% C.I. (in months).
medhat <- predict(fit.lognorm,type="uquantile",p=0.5,se.fit=T)
medhat1 <- medhat$fit[1]
medhat1.se <- medhat$se.fit[1]
exp(medhat1)
```

```
  Log-logistic fit loglogis.fit<-survreg(Surv(cerv[,1],cerv[,2])~1,dist="loglogistic")
summary(loglogis.fit)
 Estimated median along with a 95% C.I. (in weeks).
medhat <-
predict(loglogis.fit,type="uquantile",p=0.5,se.fit=T)
medhat1 <- medhat$fit[1] medhat1.se <-
medhat$se.fit[1] exp(medhat1)
C.I.median1 <- c(exp(medhat1),exp(medhat1-1.96*medhat1.se),
exp(medhat1+1.96*medhat1.se))
names(C.I.median1) <-
c("median1","LCL","UCL")
C.I.median1
qq.loglogistic(Surv(weeks,status))
  Generalised Gamma fit
GG.fit<-survreg(Surv(cerv[,1],cerv[,2])~1,dist="extreme")
summary(loglogis.fit)
 Estimated median along with a 95% C.I. (in weeks).
medhat <-
predict(GG.fit,type="uquantile",p=0.5,se.fit=T)
medhat1 <- medhat$fit[1] medhat1.se <-
medhat$se.fit[1] exp(medhat1)
C.I.median1 <- c(exp(medhat1),exp(medhat1-1.96*medhat1.se),
exp(medhat1+1.96*medhat1.se))
names(C.I.median1) <- c("median1","LCL","UCL")
C.I.median1 qq.loglogistic(Surv(weeks,status))
 Simulation estimate setwd("C:/Users/ FOLORUNSO /Desktop/AlhajaSF") simdata500
<- read.csv('simdata500.csv', header=T) attach(simdata500) survgengam<-
flexsurvcure(Surv(lifetimes, status) ~ Group, data=simdata500, dist="gengamma",
link="loglog", mixture = TRUE) survllogis<-flexsurvcure(Surv(lifetimes, status) ~
Group, data=simdata500, dist="llogis", link="loglog", mixture = TRUE) survwei<-
flexsurvcure(Surv(lifetimes, status) ~ Group, data=simdata500, dist="weibull",
link="loglog", mixture = TRUE) survlnorm<-flexsurvcure(Surv(lifetimes, status) ~
Group, data=simdata500, dist="lnorm", link="loglog", mixture = TRUE)
summary(survgengam)
```

```
  Alhaja’s Curves
library(fitdistrplus)
library(MASS) library(actuar)
library(moments)
library(extraDistr)  library(gbs)
library(stats) library(lattice)
getwd() setwd("C:/Users/
FOLORUNSO
/Desktop/AlhajaSF") cerv <-
read.csv(’OvarianC.csv’,
header=T) attach(cerv)
Dead <- cerv[(cerv$Censoring==0),1] Alive <-
cerv[(cerv$Censoring==1),1] plotdist(Dead,
histo = TRUE, demp = TRUE, col="yellow")
plotdist(Alive, histo = TRUE, demp = TRUE,
col="violet") densityplot(Ovarian, lwd=2, col =
"blue") library(reliaR)
  fitting distribution on Ovarian data library(actuar) fitf1 <-
fitdist(Ovarian, "weibull") summary(fitf1) fitf2 <- fitdist(Ovarian,
"llogis",start=list(shape=1,scale=500)) summary(fitf2) fitf3 <-
fitdist(Ovarian, "lnorm") summary(fitf3) fitf4 <- fitdist(Ovarian,
"gamma") summary(fitf4) fitf5 <- fitdist(Ovarian,
"gengammagen", start=list(shape=1,scale=500)) summary(fitf5)
x<-Ovarian par(mfrow=c(1,2),mar=c(5.1,5.1,4.1,2.1))
Make room for the hat.
  S(t), the survival curve curve((1-pgengamma.orig(x, scale=30.05032172,
shape=1.84062842,k=10.03142, lower.tail=FALS from=0, to=200, col=’red’, lwd=2,
ylab=expression(hat(S)(t)), xlab=’t’,bty=’n’,ylim=c(0,1)
  h(t), the hazard curve curve(dgengamma.orig(x, scale=30.05032172,
shape=1.84062842, k=10.03142)/(1-pgengamma.orig (x, scale=30.05032172,
shape=1.84062842,k=10.03142, lower.tail=FALSE)), from=0, to=200, col=’blue’,
lwd=2, ylab=expression(hat(h)(t)), xlab=’t’,bty=’n’)
```

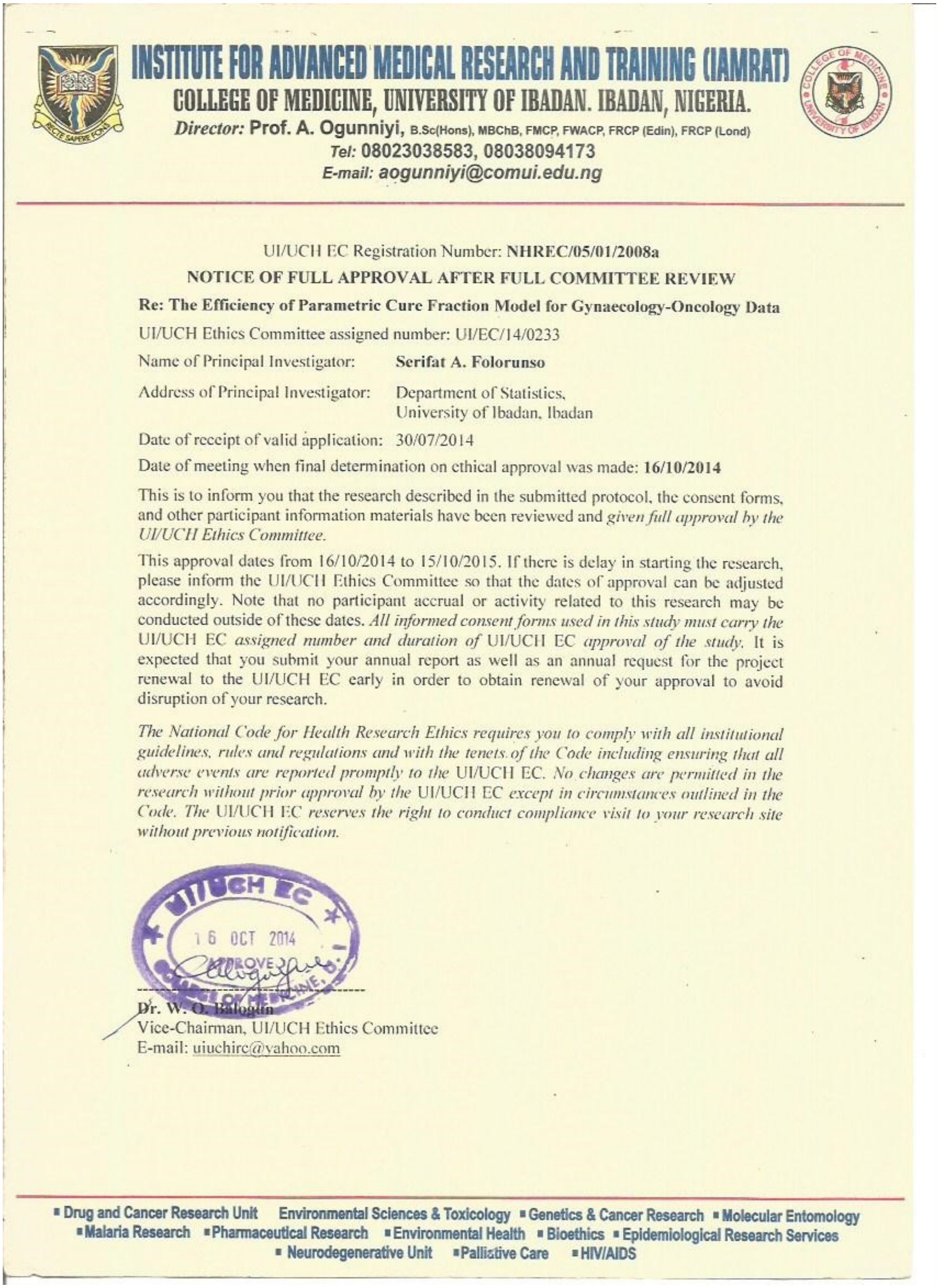
PDF plot of the models

```
m<-curve(dweibull(x, scale=50.339355, shape=1.553091),
col=’red’, ylab=’f(x)’, bty=’n’) p<-curve(dllogis(x, scale=39.513559, shape=2.142304),
col=’blue’, ylab=’f(x)’, bty=’n’) q<-curve(dlnorm(x, meanlog=3.5254222,
sdlog=0.9699595), col=’yellow’, ylab=’f(x)’, bty=’n’) r<-curve(dgamma(x,
rate=0.04032172, shape=1.84062842), col=’green’, ylab=’f(x)’, bty=’n’) s<-
curve(dgengamma.orig(x, scale=30.05032172, shape=1.84062842, k=10.03142),
col=’violet’, ylab=’f(x)’, bty=’n’)
plot(c(x,m,p,q,r,s), lwd=2, type="s", bty="n",
xlab="Ovarian Cancer", ylab="f(x)") lines(m, col =
"blue", lwd=2) lines(p, col = "violet", lwd=2) lines(q, col
= "yellow", lwd=2) lines(r, col = "green", lwd=2) lines(s,
col = "orange", lwd=2)
legend("topleft",c("Ovarian","Weibull","loglogistics","lo
gnormal
","GenGamma", "ModifiedGenGamma"),
col=c(1,"blue","violet","yellow","green","orange"),lwd=c(1,2,2,2,2,2), cex=0.45)
a<-dweibull(x, scale=50.339355, shape=1.553091) b<-
dllogis(x, scale=39.513559, shape=2.142304) c<-dlnorm(x,
meanlog=3.5254222, sdlog=0.9699595) d<-dgamma(x,
rate=0.04032172, shape=1.84062842) e<-dgengamma.orig(x,
scale=30.05032172, shape=1.84062842, k=10.03142)
plot(c(x,a,b,c,d,e),lwd=2, type="s",
bty="n") lines(x, col = "red",
lwd=2) lines(a, col = "blue",
lwd=2) lines(b, col = "violet",
lwd=2) lines(c, col = "yellow",
lwd=2) lines(d, col = "green",
lwd=2) lines(e, col = "orange",
lwd=2)
```

Simulation Study

```
                          set.seed(123)
Ovarian10_50 <- matrix(replicate(50,
runif(10,1,100)), 10) censtimes <- 3 + 20 *runif
```

```
(10) ztimes <- pmin(Ovarian10_50, censtimes)
status <- ifelse(censtimes < Ovarian10_50, 1, 0)
;status simdata10<- data.frame(Ovarian10_50,
censoring);simdata10 write.csv(Ovarian10_50,
file="Ovarian10_50.csv")
weib.fit <-
survreg(Surv(cerv[,1],cerv[,2])~1,dist="weib")
summary(weib.fit) coeff <- weib.fit$coeff   muhat
var <- weib.fit$var thetahat <- exp(coeff)
exp(muhat) thetahat muhat <- weib.fit$coeff
model Evaluation bias.est=thetahat-cerv[,1]
subtract y values bias=mean(bias.est)  average
over all x values bias2=bias^2  square MSE<-
var+bias2 RMSE<-sqrt(MSE) cerv.u <-
cerv[,1][cerv$Censoring==1] nu <- length(cerv.u)
scalehat <- rep(exp(muhat),nu) Shat <- 1 -
pweibull(cerv.u,1,scalehat)
LCL <- exp(log(Shat)*exp(1.96/sqrt(nu)))
UCL <- exp(log(Shat)*exp(-
1.96/sqrt(nu))) C.I.Shat <-
data.frame(cerv.u,Shat,LCL,UCL)
round(C.I.Shat,5)
 Estimated median along with a 95% C.I. (in weeks).
medhat <- predict(weib.fit,type="uquantile",p=0.5,se.fit=T) medhat1 <- medhat$fit[1]
medhat1.se <- medhat$se.fit[1] exp(medhat1)  median time to cure
C.I.median1 <- c(exp(medhat1),exp(medhat1-
1.96*medhat1.se), exp(medhat1+1.96*medhat1.se))
names(C.I.median1) <- c("median1","LCL","UCL")
C.I.median1
```



INSTITUTE FOR ADVANCED MEDICAL RESEARCH AND TRAINING (IAMRAT)

COLLEGE OF MEDICINE, UNIVERSITY OF IBADAN. IBADAN, NIGERIA.

Director: **Prof. A. Ogunniyi,** B.Sc(Hons), MBChB, FMCP, FWACP, FRCP (Edin), FRCP (Lond)
Tel: **08023038583, 08038094173**
E-mail: ***aogunniyi@comui.edu.ng***

UI/UCH EC Registration Number: **NHREC/05/01/2008a**

NOTICE OF FULL APPROVAL AFTER FULL COMMITTEE REVIEW

Re: The Efficiency of Parametric Cure Fraction Model for Gynaecology-Oncology Data

UI/UCH Ethics Committee assigned number: UI/EC/14/0233

Name of Principal Investigator: **Serifat A. Folorunso**

Address of Principal Investigator: Department of Statistics,
University of Ibadan, Ibadan

Date of receipt of valid application: 30/07/2014

Date of meeting when final determination on ethical approval was made: **16/10/2014**

This is to inform you that the research described in the submitted protocol, the consent forms, and other participant information materials have been reviewed and *given full approval by the UI/UCH Ethics Committee.*

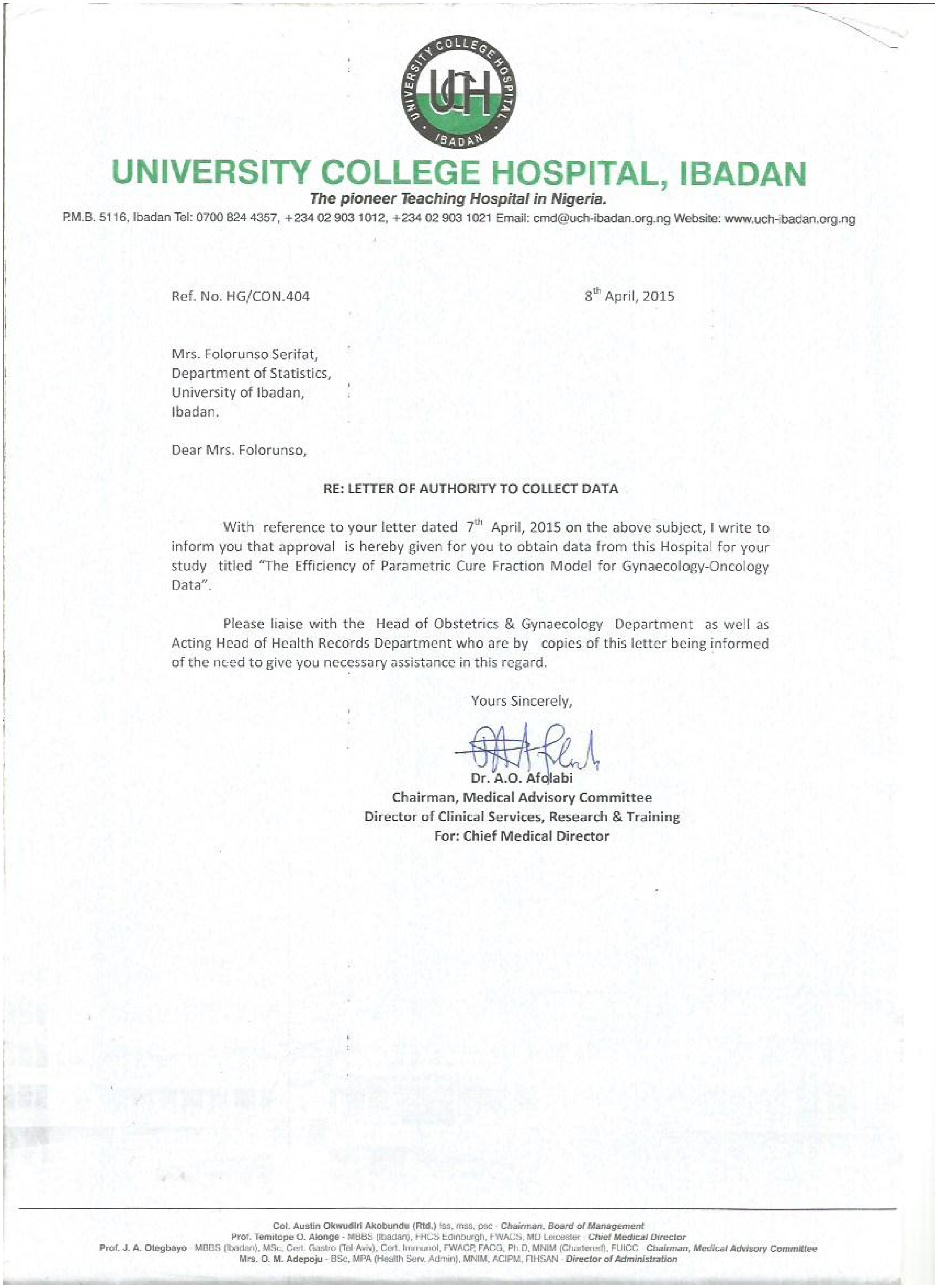
This approval dates from 16/10/2014 to 15/10/2015. If there is delay in starting the research, please inform the UI/UCH Ethics Committee so that the dates of approval can be adjusted accordingly. Note that no participant accrual or activity related to this research may be conducted outside of these dates. *All informed consent forms used in this study must carry the* UI/UCH EC *assigned number and duration of* UI/UCH EC *approval of the study.* It is expected that you submit your annual report as well as an annual request for the project renewal to the UI/UCH EC early in order to obtain renewal of your approval to avoid disruption of your research.

The National Code for Health Research Ethics requires you to comply with all institutional guidelines, rules and regulations and with the tenets of the Code including ensuring that all adverse events are reported promptly to the UI/UCH EC. *No changes are permitted in the research without prior approval by the* UI/UCH EC *except in circumstances outlined in the Code. The* UI/UCH EC *reserves the right to conduct compliance visit to your research site without previous notification.*

UI/UCH EC
16 OCT 2014
APPROVED
COLLEGE OF MEDICINE

Dr. W. O. Balogun
Vice-Chairman, UI/UCH Ethics Committee
E-mail: uiuchirc@yahoo.com

▪ Drug and Cancer Research Unit Environmental Sciences & Toxicology ▪ Genetics & Cancer Research ▪ Molecular Entomology
▪ Malaria Research ▪ Pharmaceutical Research ▪ Environmental Health ▪ Bioethics ▪ Epidemiological Research Services
▪ Neurodegenerative Unit ▪ Palliative Care ▪ HIV/AIDS



UNIVERSITY COLLEGE HOSPITAL, IBADAN

The pioneer Teaching Hospital in Nigeria.

P.M.B. 5116, Ibadan Tel: 0700 824 4357, +234 02 903 1012, +234 02 903 1021 Email: cmd@uch-ibadan.org.ng Website: www.uch-ibadan.org.ng

Ref. No. HG/CON.404

8th April, 2015

Mrs. Folorunso Serifat,
Department of Statistics,
University of Ibadan,
Ibadan.

Dear Mrs. Folorunso,

RE: LETTER OF AUTHORITY TO COLLECT DATA

With reference to your letter dated 7th April, 2015 on the above subject, I write to inform you that approval is hereby given for you to obtain data from this Hospital for your study titled "The Efficiency of Parametric Cure Fraction Model for Gynaecology-Oncology Data".

Please liaise with the Head of Obstetrics & Gynaecology Department as well as Acting Head of Health Records Department who are by copies of this letter being informed of the need to give you necessary assistance in this regard.

Yours Sincerely,

Dr. A.O. Afolabi
Chairman, Medical Advisory Committee
Director of Clinical Services, Research & Training
For: Chief Medical Director