FINITE MIXTURE MODELS IN SURVIVAL DATA ANALYSIS

By

BENJAMIN YONGBIN LI

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Statistics

Written under the direction of

Javier Cabrera

And Approved by

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2019

Abstract of the Dissertation

Finite Mixture Models in Survival Data Analysis

by BENJAMIN YONGBIN LI

Dissertation Director:

Javier Cabrera

In the pharmaceutical industry, cost-effectiveness analysis is an important step in the development of new health interventions. It is a method for assessing the gains in health relative to the costs of different health interventions. This assessment helps the regulators, providers, and potential users to make informed decisions. Health gains can be measured in several ways. One of them is the estimated gained life expectancy due to the intervention. Although the randomized controlled trials (RCTs) are considered to be the most reliable sources of the evidence to be used in the cost-effectiveness analysis, data collected from these trials are often incomplete due to censoring and truncation. This requires the extrapolation of the survival probability beyond the time frame of the RCTs. For this purpose, parametric models are necessary to estimate the survival functions. Although there exist several single parametric models (such as the Weibull, Gamma, and lognormal) that can perform this task, they fail to provide accurate estimates when the survival data are heterogeneous. In these situations, the finite mixture models fit the data better and therefore their results are more consistent and reliable.

This dissertation studies the implementation of the finite mixture models in survival data analysis. It discusses in detail how to estimate the parameters of a finite mixture models through the expectation and maximization (EM) algorithm. These steps are flexible to account for the effects of covariates. In addition, we propose a new approach via censored quantile regression for finding the initial values of the EM algorithm. This method takes into consideration the special features of survival data and therefore will help improve the efficiency of the EM algorithm. We also demonstrate how to construct the desired confidence intervals of the estimates through bootstrapping.

In oncology as well as other therapeutic areas, some patients will not experience the relapse of the disease after being treated. These patients are considered to be cured. It is of interest to know both the cure rate and the survival function of the patients who are not cured by the intervention. We study the mixture cure model in the general framework of finite mixture models as a special case, and provide the modified EM algorithm to estimate both the cure rate and the survival function of the uncured patients.

Acknowledgement

Throughout the writing of this dissertation I have received a great deal of support and assistance. Firstly, I would like to express my sincere gratitude to my advisor Prof. Javier Cabrera for the continuous support of my PhD study and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this dissertation. I could not have imagined having a better advisor and mentor for my PhD study.

Besides my advisor, I would like to thank the rest of my dissertation committee: Prof. John Kolassa, Prof. William Kostis , and Dr. Birol Emir, for their insightful comments and encouragement, but also for the hard question which incented me to widen my research from various perspectives.

Last but not the least, I would like to thank my family: my parents, my sister, and my brother for supporting me spiritually throughout writing this dissertation and my life in general.

Dedication

This work is dedicated to my wife, Qing Wei, for her unyielding support through the writing process of this dissertation, during my study in the PhD program, and in all other aspects of life.

Table of Contents

List of Tables

List of Illustrations

Chapter 1: Introduction

1.1 Motivation

Survival analysis is a collection of statistical methods for data analysis. The random variable of interest in this type of analysis is time until an outcome event occurs. The term "survival analysis" came into being from initial studies where the event of interest was death. Later, the scope of the survival analysis has been broadened to include other fields, such as engineering, economics, and sociology. This topic is called reliability analysis in engineering, duration analysis in economics, and event history analysis in sociology. Depending on the research question in each specific study, the outcome event might be defined as death , the failure of a mechanical system, crash of the stock market, or duration of first marriage (Singh and Mukhopadhyay 2011).

One unique feature about survival analysis is that quite often than not the researcher works with incomplete data, in the sense that for some patients the exact time of the event of interest is not observed during the study period. These are called censored observations or censored times. There are three types of censoring: left censoring, right censoring, and interval censoring. Left censoring occurs when a study subject already experienced the event of interest before entering into the study, but the exact time is unknown. Interval censoring occurs when an event of interest is known to have occurred between two timepoints, t_1 and t_2, but again the exact time is unknown. The last type of censoring is right censoring, which occurs when a study subject does not experience the event of interest by the end of the study, or the subject is lost to follow-up during the study period. The rest

of the dissertation will focus exclusively on analysis of survival data that are right censored, as this is the type of censoring most frequently seen in clinical trials.

Within the pharmaceutical industry, the randomized controlled trials (RCTs) are a crucial part of the new drug development process. Survival analysis has been a major tool in RCTs to assess the benefits and risks of new interventions. It was originally motivated by the need to analyze time-to-death data in clinical trials. Since then, it has been extensively used in trials where the primary outcome measure is the time to a clinically important event, such as death, progression of a disease, serious adverse events, or response to the treatments. There has been enormous growth in the field of survival analysis over the past several decades. The most significant milestones in the development of new methodologies are the Kaplan-Meier (KM) estimator of the survival function, the log-rank test for comparing two survival distributions, and the Cox proportional hazard model for evaluating the effects of covariates on the survival time (Fleming and Lin 2000). The KM estimate and the log-rank test are nonparametric, while the Cox proportional hazard model is semi-parametric. In addition, there are parametric models where the survival time is assumed to follow certain distributions. The most common distributions in this category include exponential, Weibull, lognormal, log-logistic, gamma, Gompertz, generalized gamma, and generalized F. Between the nonparametric/semiparametric and parametric methods, the former are predominantly applied in RCTs where one of the endpoints is time to an event of interest, due to the fact that it is difficult to specify the parametric form given the complex nature of human diseases.

With the advancement of research in biology, chemistry, and physics, the competition in the pharmaceutical industry is getting increasingly intense. Pharmaceutical

companies need to price their new products properly and formulate marketing strategies in order to stay competitive (Beck 1990). At the same time, policy makers and healthcare professionals are faced with the question of how to allocate limited resources to the interventions that provide the most health benefits. Cost-effectiveness analysis offers a means to assist the decision making for both the sponsors and users of new interventions. It is a method for assessing the gains in health relative to the costs of different health interventions. The basic calculation involves dividing the cost of an intervention in monetary units by the expected heath gain, which can be measured in several ways. One of them is the estimated gained life expectancy due to the intervention. This requires the prediction of the survival probability for patients receiving the intervention at a future time point. The RCTs are considered to be the most reliable sources of the evidence to be used in the cost-effectiveness analysis. However, the nonparametric/semiparametric models are not suitable for making predictions of the survival probability, given the incomplete nature of the survival data. For example, since data are truncated at the termination of the trial, only the restricted mean survival time can be directly estimated. For this purpose, the survival time needs to be assumed to follow some parametric form. Once this parametric form is determined and its parameters are estimated based on data collected from the RCTs, extrapolation of the survival probability can be carried out.

There used to be no common practice in survival data extrapolation. Sponsors of new interventions would follow their own standards. And sometimes the selected models are not well justified. In 2011, the Decision Support Unit (DSU) of the National Institute for Health and Clinical Excellence (NICE) of UK published its Technical Support Document 14. In this document, the DSU gives its guideline on extrapolating patient level

survival data collected in RCTs. The recommended procedure consists of a sequence of steps, from visually checking the log-cumulative hazard plots or suitable residual plots, to running model diagnostics such as Akaide's Information Criterion (AIC) and Bayesian Information Criterion (BIC). However, the focus of this guidance is on fitting the data to a single parametric model or a piecewise parametric model, where a single model is estimated within a given period. At the same time, the DSU acknowledges that more flexible models can be very useful.

There are apparently advantages associated with a single parametric model. First, the estimation procedure is relatively easy. All classical parametric survival models can be estimated in standard statistical software, such as R and SAS. Second, the density, survival, and hazard function are relatively simple. Third, the interpretation of a single model is more intuitive and straightforward than that of a more complex model. However, there are situations where survival time is heterogeneous and therefore a single model is not sufficient. One example is discussed in Blackstone et al. (1986). They discovered that the risk of death or time-related events following a major surgical intervention or acute illness often can be characterized in 3 phases. Risk is high after the intervention or onset of illness, then falls to a lower level, and later rises again. The distinction among these phases is not clear so each phase cannot be individually modeled by a separate parametric model. Another example of heterogeneous survival time is competing risk of failure, where the event of interest could be attributed to multiple causes. For instance, if the primary outcome in a study is all-cause mortality, then a death attributable to cardiovascular causes and a death attributable to non-cardiovascular causes are competing events to each other, as the occurrence of one precludes the occurrence of the other. Finally, in some studies, the

patients will never experience the event of interest after the intervention. These patients are assumed to be “cured”, and the statistical model needs to account for this fact as the survival function approaches to a non-zero value when time goes to infinity (Othus et al. 2012). In these cases, a single parametric model is not capable of estimating accurately the underlying survival function and therefore tends to either over- or under-estimate the mean survival time. As a result, cost-effectiveness analysis relying on a single parametric model is more likely to draw wrong conclusions, which can lead to decisions that are detrimental to both the providers and potential users of new medical interventions.

There have been a variety of statistical methods developed to cope with the heterogeneous data in survival analysis. Among them, the finite mixture model has been proposed by many researchers. Erişoğlu et al. (2011) demonstrated that a mixture of two different distributions (Exponential-Gamma, Exponential-Weibull, and Gamm-Weibull) is appropriate for the heterogeneous survival times. Farewell (1982, 1986) and Yu and Tiwari (2007) utilized a mixture cure model to estimate the proportion of patients who are cured from the disease of interest. Marín et al. (2005) showed how to use Bayesian methods to fit a mixture of Weibull model with an unknown number of component. McLachlan and Peel (2000) discussed in detail how to apply the finite mixture model in survival analysis with competing risks. These researchers have clearly demonstrated that in certain situations, the mixture models possess advantages over the single models. They have employed a variety of methods to estimate the parameters of the mixture models, such as directly solving the likelihood function to get the maximum likelihood estimators (MLEs), utilizing the expectation and maximization (EM) algorithm to find the MLEs, and the Bayesian estimators.

This dissertation builds upon the previous work in analyzing survival data with finite mixture models, and makes contribution to this line of research in three areas. First, to improve the likelihood and speed of convergence of the EM algorithm, we recommend a new approach to find the initial values needed to start the EM iteration. Most of the current methods of finding the initial values are borrowed from implementations of the EM algorithm in areas where data are not censored and often assumed to follow a normal distribution. These assumptions are not valid any more in survival data analysis. To account for the special features of the survival data, we suggest looking for initial values of the EM algorithm through censored quantile regressions, which take data censoring into consideration and can fit parametric models suitable for the time-to-event random variables. Second, most of the previous work estimating the mean survival function without examining the effects of covariates on either the mixing weights or the distribution parameters for each individual component. In contrast, we build into the EM algorithm steps that can incorporate any number of covariates, both continuous and categorical. Third, previous research work on the mixture cure model focused almost exclusively on estimating the cure rate. A missing piece of important information is the survival function of the patients that are not cured. In a typical mixture model, each individual component follows a distribution. In the mixture cure model, the cure rate is a constant. In this sense, the mixture cure model is a special case in the more general mixture model framework. We demonstrate that, not just the cure rate, but also the survival function of the uncured patients, which in turn can follow another mixture of distributions, can all be estimated through the EM algorithm.

1.2 Organization of the Dissertation

Chapter 2 discusses the need to extrapolate data from RCTs in the development of new health interventions. The RCTs are considered to be the gold standard in clinical trials. Compared with other trial designs, they possess a number of advantages that render the results reliable and interpretable. However, most RCTs typically have short follow-up periods and therefore cannot establish long-term benefits of the intervention. This question can be answered by the cost-effectiveness analysis, which quite often needs to make extrapolation of the data collected from RCTs. In cost-effectiveness analysis, benefits of an intervention typically are measured by gained life expectancy due to the intervention. Data extrapolations are often conducted by assuming that the survival time follows a certain distribution. In the situation of heterogeneous survival data, a model of mixture distributions will provide better fit and estimates than a single distribution model.

Chapter 3 includes a brief introduction of the finite mixture model. It also discusses the EM algorithm and its theoretical properties. This algorithm is used primarily in this dissertation to find the MLEs of the parameters for the finite mixture model. Detailed steps of the estimation processes are given in cases without and with covariates. Also, a new method of finding the initial values of the EM algorithm through the censored quantile regressions is introduced. The chapter also covers how to construct confidence intervals for the estimates through bootstrapping.

Chapter 4 compares the finite mixture models with the other single parametric models most frequently used in survival data analysis, in the setting of simulated data as well as empirical data. The EM algorithm is implemented for the estimation, both with and without covariates. The results show that when survival data are heterogeneous, the finite

mixture models provide better fit, and more accurate extrapolations of the survival function and mean survival time.

In Chapter 5, the mixture cure model is discussed in detail. Both simulated and empirical data are used to demonstrate that when there exists a cure portion, the mixture cure model is superior to the single parametric models. In addition, as a special case of the general mixture model, the mixture cure model can also estimate the survival function for the uncured patients, no matter whether this function follows a single distribution or another mixture distribution.

Chapter 2: Extrapolation of Heterogeneous Survival Data

2.1 The Need for Extrapolation of Survival Data from Randomized Controlled Trials

In the development of new health intervention, the RCTs are considered to be the most reliable method to demonstrate efficacy and safety of the new intervention over the alternative. Sponsors of new interventions rely heavily on RCTs to answer patient-related questions and form the basis for regulatory authorities' decisions on approval (Kabisch et al. 2011). RCTs minimize bias and confounding factors. They allow for direct comparison of different interventions to establish superiority. Adequately powered RCTs avoid both type I and type II errors, and statistical test of significance is readily interpretable (Bulpitt 1996). However, RCTs also have their limitations. Besides being expensive and taking long time to finish, the patients in RCTs are not followed for a long period of time. The maximum follow-up in RCTs is typically only 1 to 5 years, while the choice of an intervention will often affect outcomes over a longer period (Jackson et al. 2017). This shortcoming of the RCTs becomes more prominent when there is need to establish the long-term benefits of the new intervention and its associated costs. This is important to the policy makers, the providers, and potential users of the new intervention. The policy makers will consider the costs and benefits related to the new interventions to decide how to allocate resources. The providers need this piece of information in price negotiations. The potential users weigh the benefits against the costs to decide whether to switch from their current treatment. The results from cost-effectiveness analysis will help provide answers to this question. Cost-effectiveness analysis compares the health effects of an intervention with the resources that must be invested to adopt the intervention (Beck 1990). It is one of the fastest-growing fields in health research. The analysis takes multiple

elements into consideration and boils down to a cost-effectiveness ratio, which is generally expressed as

$$\frac{Cost\ of\ intervention\ 2 - Cost\ of\ intervention\ 1}{Quality\ adjusted\ life\ expectancy\ 2 - Quality\ adjusted\ life\ expectancy\ 1}$$

Quality adjusted life expectancy is used in the denominator because it is the standard unit of effectiveness. It is a combined measure of quality of life and quantity of life (Muennig 2007). While RCTs are the major sources of information for both quality of life and quantity of life to be used in cost-effectiveness analysis, the focus of this dissertation is on the latter. Quantity of life, or life expectancy is typically measured by the expected overall survival time or expected survival time from certain causes over a long period of time, quite often a lifetime horizon (Jackson et al. 2017). However, the survival data collected from the RCTs, due to the limitation of the follow-up, only cover a short period of time. Extrapolating the RCT data from the trial period to the lifetime of the patients is needed to bridge the gap. The traditional non-parametric or semi-parametric survival data analysis techniques, such as the KM estimator and the Cox proportional hazard model, are not suitable for this purpose. In order to perform the extrapolation, the survival data need to be assumed to follow certain distributions. Based on this assumption, statistical procedures can be taken to estimate the parameters of the distribution, which in turn will enable the estimation of the expected survival time, or survival probability at a given time. It is obvious that the assumption about the underlying distributions is vital to the success of data extrapolation.

There is a wide range of parametric models in survival data analysis. These models include the exponential, gamma, Weibull, lognormal, log-logistic, Gompertz, generalized

gamma, and generalized F. They all have been implemented in practice to various extent. In 2011, the Decision Support Unit (DSU) of the National Institute for Health and Clinical Excellence (NICE) of UK published its Technical Support Document 14. In this document, the DSU reviews the extrapolation methods used in 45 NICE Technology Appraisals (TAs). Among them, 23 (51%) used the Weibull, 20 (44%) used the exponential, 9 (20%) used log-logistic, 6 (13%) used the Gompertz or lognormal, 2 (4%) used gamma, and 1 (2%) used piecewise modelling. In addition, the DSU provided its model selection algorithm as guidance on how to properly select the parametric model. This algorithm is broken down into multiple steps, including visual inspection of the log-cumulative hazard plots and checking the model fitting with AIC/BIC. The DSU recommends that all parametric models should be included in a systematic manner in the model selection process, and other novel survival modelling methods should be considered if these existing models fail to provide a good fit to the data.

2.2 Extrapolation of Heterogeneous Survival Data

One situation where novel survival modelling is needed is when the survival data from the RCTs are likely to be heterogeneous. In clinical research, heterogeneity may exist in a variety of cases. Some of them are discussed in section 1.1. Hougaard (1991) also gave some examples of the existence of heterogenous data in studies of myocardial infarction, diabetic nephropathy, and occupational mortality. One common feature of these cases is that the patients are exposed to multiple risks, and that these risks have different impacts on the patients' survival probability. Under this situation, the single parametric models discussed in the previous section are unlikely to provide a good fit. A much more suitable

approach is to adopt the finite mixture models, which allow for multiple distributions to have a composite effect on the marginal distribution of the survival time.

In the following chapters, we will develop such finite mixture models and compare their performances with those of the single parametric models. The comparison will be done first with simulated data. In addition, we will also apply the finite mixture models to analyze two sets of empirical data that are collected from RCTs.

The first empirical data set comes from the Systolic Hypertension in the Elderly Program (SHEP). SHEP is a double-blind, randomized, placebo-controlled trial of the treatment for isolated systolic hypertension (ISH) in persons 60 years of age and older. The primary endpoint of this trial is to determine whether anti-hypertensive drug treatment reduces the risk of total stroke in patients with ISH within this age group. Participants were randomized to either the active treatment arm treated with either chlorthalidone or atenolol, or the matching placebo arm. There were 2365 patients in the active treatment arm, and 2371 patients in the placebo arm. The randomized phase started in 1985 and concluded in 1990. Results from this trial indicate that anti-hypertensive treatment helps reduce the incidence of stroke. Although all-cause mortality and cardiovascular related mortality also favor the treatment arm, the results are not statistically significant (SHEP Cooperative Research Group 1991). After that, all patients were advised to receive active therapy and have been followed-up and their mortality and cause of death information is collected through the National Death Index (NDI). Using this information up to December 31st, 2006, Kostis et al. (2011) found that anti-hypertensive treatment for the ISH patients increases their long-term survival from cardiovascular related deaths, and that there is no significant difference in overall survival between the active therapy and placebo patients.

It remains to be an interesting question whether there is a method that can accurately predict the long-term survival probability from both cardiovascular related mortality and all-cause mortality, so the benefits of the anti-hypertensive treatment to the ISH patients can be clearly estabelished. At the time of writing this dissertation, we have information for SHEP patients about their mortality and cause of death up to December 31st, 2014. The total follow-up time from the beginning of the trail to this date is approximately 29.7 years. The KM estimates of the survival function for cardiovascular related survival and overall survival are plotted in Figure 1 and Figure 2, respectively. As can be seen in these figures, there is some separation in the respective survival curves between the two groups up to 20 years from the start of the trail, after that the curves are hardly separable. Since the placebo patients started active therapy at the conclusion of the randomized phase, they have been treated with anti-hypertensive drugs for more than 20 years. Therefore, there should be minimal treatment effects remaining between the active treatment and placebo patients, and we therefore could combine the two groups and examine their long-term life expectance together.

SHEP patients are recruited from 16 clinical centers. They consist of both males and females and different races. There are large variations in their medical history and physical conditions (SHEP Cooperative Research Group 1991). Therefore, there are strong reasons to believe there exists heterogeneity among the patients. Amaratunga and Cabrera (2015) noticed this feature of the SHEP trial and recommended using the mixture models to make extrapolation from the data. In addition, some SHEP patients are lost to follow-up during the years and their mortality information becomes unknow. Including these patients in the analysis will distort the estimation and extrapolation. The finite mixture model is

capable of estimating this proportion by treating these patients as “cured”. Thus the extrapolation of survival probability will only be for the patients with explicit mortality information.

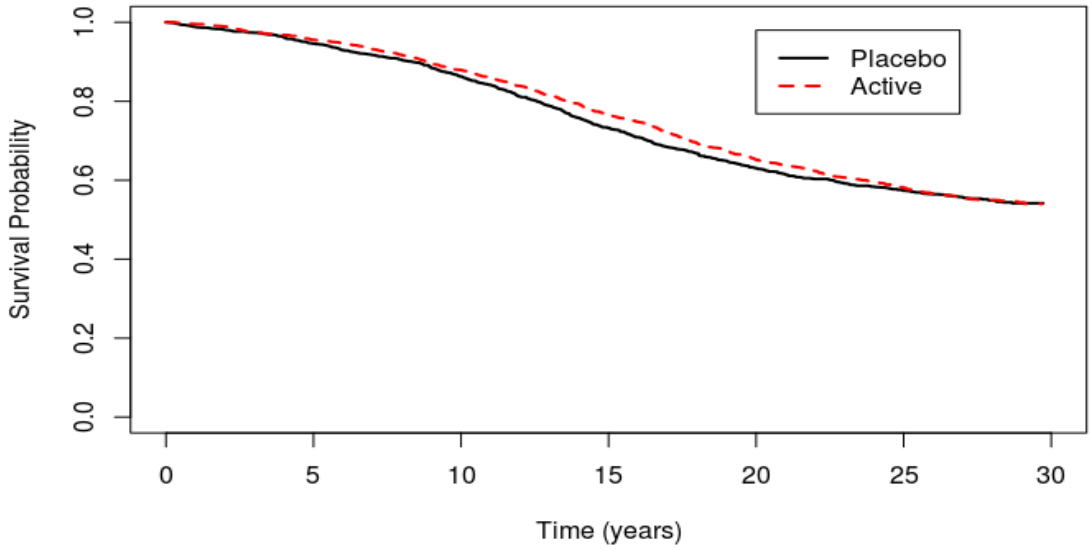

Figure 1 Kaplan-Meier Curves of CV-related Survival – SHEP

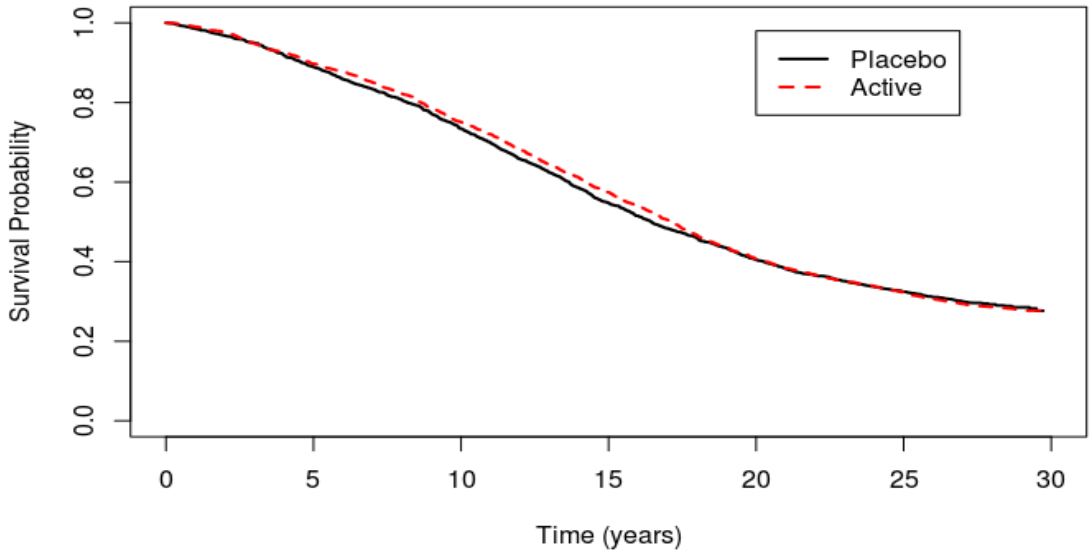

Figure 2 Kaplan-Meier curves of Overall Survival – SHEP

The second empirical data set is based on Hodi et al. (2010), which published the results of a double-blind Phase III clinical trial that investigated the efficacy and safety of ipilimumab in patients with previously treated metastatic melanoma. Ipilimumab works by blocking cytotoxic T-lymphocyte-associated antigen 4 to potentiate an antitumor T-cell response. A total of 676 eligible patients in the trial were randomized in a 3:1:1 ratio to receive ipilimumab plus a glycoprotein 100 peptide vaccine (IPI+GP100, 403 patients), ipilimumab alone (IPI, 137), or glycoprotein 100 alone (GP100, 136). Patients were followed up for up to 55 months. The primary endpoint is the overall survival, and one of the secondary endpoints is progression free survival. The study results show that ipilimumab plus GP100 or ipilimumab alone improves the overall survival in this patient population.

Figure 3 is the original figure in Hodi et al. (2010) that includes the KM curves for overall survival and progression free survival. We re-create the patient level trial data by digitizing these curves. The KM curves based on the digitized data are displayed in Figure 4. These re-generated KM curves based on the digitized patient level data are very close to the original curves in Figure 3. We can see that the progression-free survival curves for all three treatment arms follow an inverse "S" shape. This is an indication that a finite mixture model might fit the data better and generate more accurate predictions (Kleinbaum 2005). The overall survival curves for all three groups have a flat tail that does not approach 0 at the end of the follow-up period. This is indicative of a potential "cured" portion in the patients. Like with the SHEP data, it is an interesting and important question to estimate this proportion, and the finite mixture model is the right tool for this task.

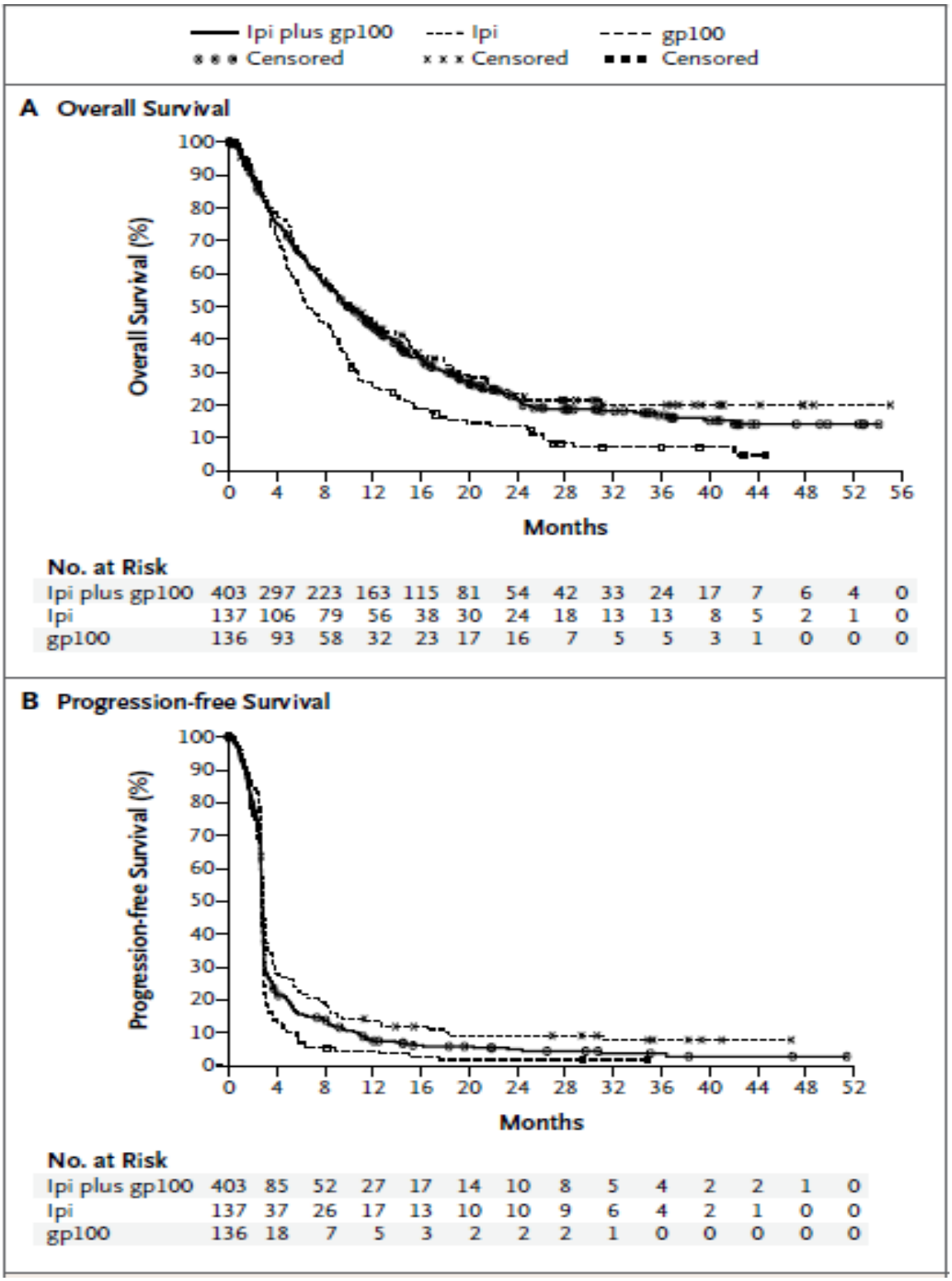

Figure 3 Kaplan-Meier Curves for Overall Survival and Progression-free Survival

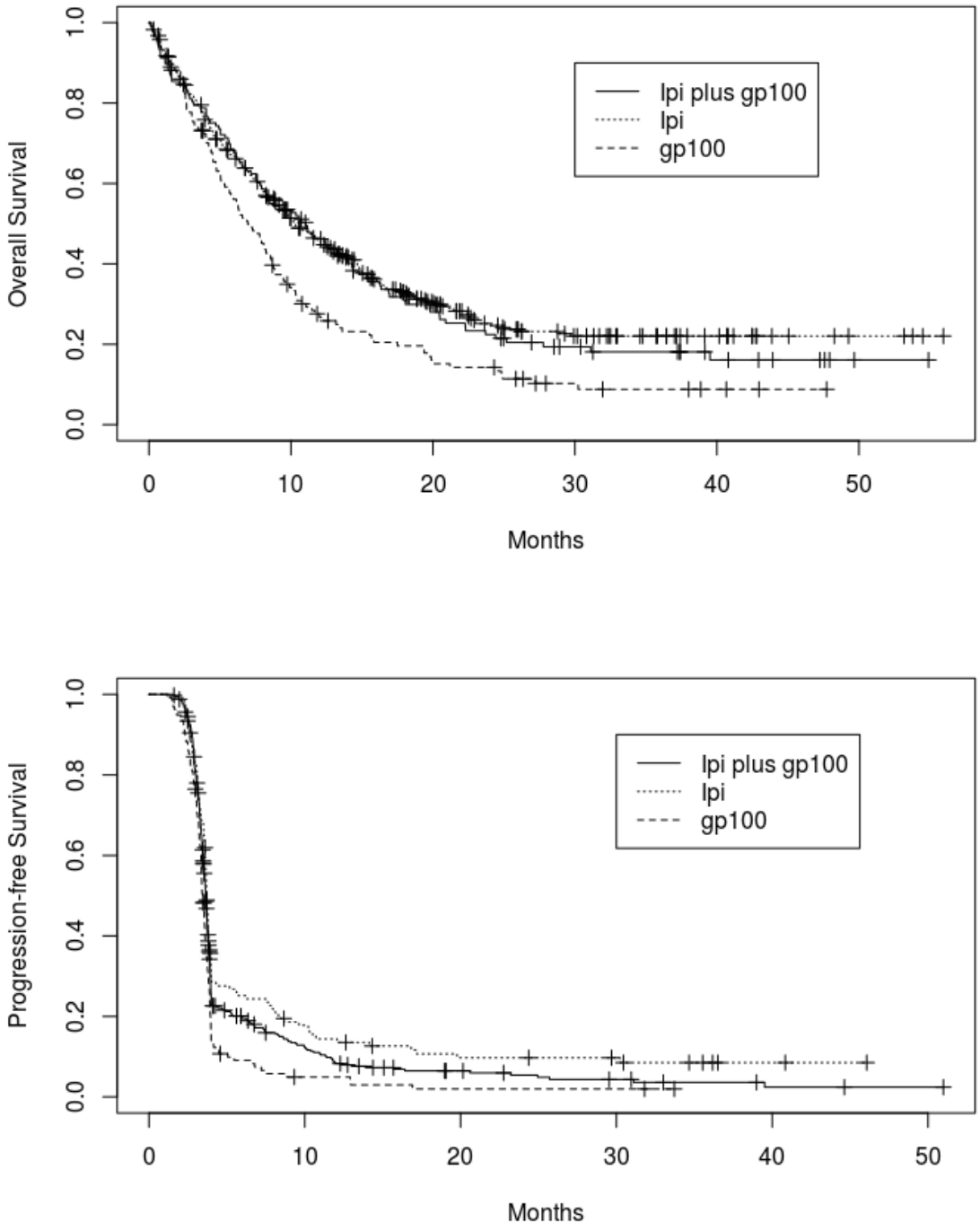

Figure 4 Digitized Kaplan-Meier Curves for Overall Survival and Progression-free Survival

Chapter 3: Finite Mixture Models and the Expectation and Maximization Algorithm

In this chapter we briefly introduce the finite mixture models and the different methods to estimate their parameters. The expectation and maximization (EM) algorithm and its application in finding the maximum likelihood estimators for the finite mixture models is discussed in detail. We also introduce our approach of choosing the initial values for the EM algorithm. Compared with other methods, this approach builds upon the censored quantile regressions and therefore is capable of handling the censoring feature in survival data. We finish this chapter by introducing the most popular parametric models in survival data analysis, whose performances will be compared with those of the finite mixture models in the following chapters.

3.1 Finite Mixture Models

A random variable Y follows a finite mixture distribution if the density $f(y)$ of Y can be written in the form

$$f(y) = \sum_{i=1}^{g} \pi_i f_i(y), \tag{3.1}$$

where the $f_i(y)$ are densities and the π_i satisfy such conditions that

$$0 \leq \pi_i \leq 1 \ (i = 1, \dots, g) \tag{3.2}$$

and

$$\sum_{i=1}^{g} \pi_i = 1. \tag{3.3}$$

The $f_i(y)$ are called the component densities of the mixture, and the π_i the mixing proportions or weights. It is clear that $f(y)$ is a density, which is called a g-component

finite mixture density. Its corresponding distribution function $F(y)$ is called a g-component finite mixture distribution. In some applications the number of components g is considered fixed, in others the value of g can be unknown and needs to be inferred from data. In this and the following chapters we treat g as fixed.

One way to generate a random variable Y following a g-component mixture distribution is through a random vector $\boldsymbol{Z} = (Z_1, \dots, Z_g)^T$, which follows a multinomial distribution consisting of one draw on g categories with corresponding probability $\boldsymbol{\pi} = (\pi_1, \dots, \pi_g)^T$. That is, Z_i can only take the value of 0 or 1, and $P(Z_i = 1) = \pi_i, (i = 1, \dots, g)$. Therefore,

$$P(\boldsymbol{Z} = \boldsymbol{z}) = {\pi_1}^{z_1}{\pi_2}^{z_2} \dots {\pi_g}^{z_g}. \tag{3.4}$$

The random vector $\boldsymbol{Z}$ can be viewed as the component label of the g-component finite mixture distribution of Y. When $Z_i = 1 \; and \; Z_{j \neq i} = 0 \; (i, j = 1, \dots, g)$, Y follows density $f_i(y) \; (i = 1, \dots, g)$. In other words, the conditional distribution of Y given $\boldsymbol{Z}$ is

$$f(y|\boldsymbol{Z} = \boldsymbol{z}) = f_i(y) \; (i = 1, \dots, g). \tag{3.5}$$

And the marginal distribution of Y has the g-component mixture form (3.1) (McLachlan and Peel 2000).

Since its first introduction by Karl Pearson in the late 1900's, the finite mixture model has been continuously receiving popularity as a methodology to model heterogeneous data. Due to its extreme flexibility, it has been widely applied in astronomy, biology, genetics, medicine, psychiatry, economics, engineering, and marketing (McLachlan and Peel 2000, sec 1.1). Along with its development, the fitting methods of

the finite mixture model have also evolved from the initial method of moments by Pearson (1894), which is rather calculation intensive, to the direct application of the maximum likelihood method by Wolfe (1965) and Day (1969), and eventually to the popular EM algorithm developed by Dempster (1977) to find the maximum likelihood estimators (MLEs) efficiently.

3.2 The Maximum Likelihood Estimators

The method of maximum likelihood is the most popular technique for deriving estimators (Casella and Berger 2002, p.315). If $Y_1, \dots, Y_n$ are an independent and identically distributed (iid) sample from a population with density $f(y|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of unknown parameters, the likelihood function of the sample point $\boldsymbol{y} = (y_1, \dots, y_n)^T$ is defined as

$$L(\boldsymbol{\theta}|\boldsymbol{y}) = f(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{j=1}^{n} f(y_j|\boldsymbol{\theta}). \tag{3.6}$$

This is the sample density function considered as a function of $\boldsymbol{\theta}$ for fixed $\boldsymbol{y}$. The log likelihood function is defined as

$$l(\boldsymbol{\theta}|\boldsymbol{y}) = \log[f(\boldsymbol{y}|\boldsymbol{\theta})] = \sum_{j=1}^{n} \log\left[f\left(y_j\middle|\boldsymbol{\theta}\right)\right]. \tag{3.7}$$

An MLE $\hat{\theta}(\boldsymbol{y})$ maximizes $L(\boldsymbol{\theta}|\boldsymbol{y})$ (equivalently $l(\boldsymbol{\theta}|\boldsymbol{y})$). Possible candidates for $\hat{\theta}(\boldsymbol{y})$ can be obtained as a solution to the likelihood equation

$$\frac{\partial}{\partial \boldsymbol{\theta}} L(\boldsymbol{\theta}|\boldsymbol{y}) = \mathbf{0}, \tag{3.8}$$

or equivalently,

$$\frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta}|\boldsymbol{y}) = \mathbf{0}. \tag{3.9}$$

When regularity conditions hold, if the likelihood equation has a unique root, then this root is the MLE, which is consistent and asymptotically efficient. In the case the likelihood equation has multiple roots, there exist approaches such as the Newton-Raphson iterative process that leads to an estimator that is consistent, asymptotically normal, and efficient (Lehmann and Casella 1998).

Suppose that $\boldsymbol{y} = (y_1, \dots, y_n)^T$ is a random sample from a population with a g-component mixture density, that is, for $j = 1, \dots, n$,

$$f(y_j|\boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_i f_i(y_j|\boldsymbol{\theta}_i), \tag{3.10}$$

where $\boldsymbol{\Psi} = (\pi_1, \dots, \pi_{g-1}, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_g)^T$ is the vector containing all the unknown parameters. The π_i are the weights, and the $\boldsymbol{\theta}_i$ contain the respective parameters of each component density. The log likelihood function of the random sample $\boldsymbol{y}$ is given by

$$l(\boldsymbol{\Psi}|\,\boldsymbol{y}) = \sum_{j=1}^{n} \log\big[f(y_j|\boldsymbol{\Psi})\big] = \sum_{j=1}^{n} \log\big[\sum_{i=1}^{g} \pi_i f_i(y_j|\boldsymbol{\theta}_i)\big]. \tag{3.11}$$

The MLE of $\boldsymbol{\Psi}$, $\widehat{\boldsymbol{\Psi}}$, is the solution to the likelihood equation,

$$\frac{\partial}{\partial \boldsymbol{\Psi}}\, l(\boldsymbol{\Psi}|\,\boldsymbol{y}) = \mathbf{0}. \tag{3.12}$$

It can be manipulated so that $\widehat{\boldsymbol{\Psi}}$ satisfies

$$\hat{\pi}_i = \sum_{j=1}^{n} \tau_{ij}(y_j|\widehat{\boldsymbol{\Psi}})/n \quad (i = 1, \dots, g), \tag{3.13}$$

and

$$\sum_{j=1}^{n} \tau_{ij}(y_j|\,\widehat{\boldsymbol{\Psi}})\, \partial \log\big[f_i(y_j|\widehat{\boldsymbol{\theta}}_i)\big]/\partial \boldsymbol{\theta}_i = \mathbf{0} \quad (i = 1, \dots, g), \tag{3.14}$$

where

$$\tau_{ij}(y_j|\boldsymbol{\Psi}) = \frac{\pi_i f_i(y_j|\boldsymbol{\theta}_i)}{\sum_{h=1}^{g} \pi_h f_h(y_j|\boldsymbol{\theta}_h)} \qquad (i = 1, \ldots, g; j = 1, \ldots, n) \tag{3.15}$$

is the posterior probability that y_j belongs to the i[th] component of the mixture (McLachlan and Krishnan 1997, Sec. 1.4). See Appendix A for detailed derivation of (3.13), (3.14), and (3.15).

Equations (3.13) and (3.14) suggest an iterative computation of the solution. For an initial value $\boldsymbol{\Psi}^{(0)}$ of $\boldsymbol{\Psi}$, a new estimate $\boldsymbol{\Psi}^{(1)}$ can be computed for $\boldsymbol{\Psi}$, which in turn can be substituted into (3.13) and (3.14) to produce a new updated $\boldsymbol{\Psi}^{(2)}$. This process is repeated until convergence is achieved. It turns out that this iterative computation is a direct application of the EM algorithm for finding solutions to likelihood equations.

3.3 The EM Algorithm for Finding the MLEs

The EM algorithm of Dempster et al. (1977) is a procedure of iterative computation to calculate the MLEs in cases where the observed data are deemed incomplete. We let $\boldsymbol{X}$ be a random vector from sample space $\mathcal{X}$, and $\boldsymbol{Y}$ be another random vector from sample space $\mathcal{Y}$. Rather than observing the complete data vector $\boldsymbol{x}$ in $\mathcal{X}$, we only observe the incomplete data vector $\boldsymbol{y}$ in $\mathcal{Y}$. We denote the density function of $\boldsymbol{x}$ by $g_c(\boldsymbol{x}|\boldsymbol{\Psi})$**,** and the density function of $\boldsymbol{y}$ by $g(\boldsymbol{y}|\boldsymbol{\Psi})$, where $\boldsymbol{\Psi} = (\Psi_1, \ldots, \Psi_d)^T$ is a vector of unknown parameters in parameter space Ω. There is a many-to-one mapping from $\mathcal{X}$ to $\mathcal{Y}$ such that $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x})$**.** The complete-data density $g_c(\boldsymbol{x}|\boldsymbol{\Psi})$ is related to the incomplete-data density $g(\boldsymbol{y}|\boldsymbol{\Psi})$ by

$$g(\boldsymbol{y}|\boldsymbol{\Psi}) = \int_{\boldsymbol{X}_{(y)}} g_c(\boldsymbol{x}|\boldsymbol{\Psi})\, \boldsymbol{dx}, \tag{3.16}$$

where $\mathcal{X}_{(\boldsymbol{y})}$ is the subset of $\mathcal{X}$determined by $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x})$.

We denote the log likelihood function of $\boldsymbol{x}$ by $l_c(\boldsymbol{\Psi}|\boldsymbol{x})$, and the log likelihood function of $\boldsymbol{y}$ by $l(\boldsymbol{\Psi}|\boldsymbol{y})$. The EM algorithm takes iterative steps to find the MLE for $\boldsymbol{\Psi}$, not by directly solving $\frac{\partial l(\boldsymbol{\Psi}|\boldsymbol{y})}{\partial \boldsymbol{\Psi}} = 0$, but through the utilization of $l_c(\boldsymbol{\Psi}|\boldsymbol{x})$. Each iteration of the algorithm consists of two steps, the expectation step (E-step) and the maximization step (M-step). On the first iteration, the E-step takes in $\boldsymbol{\Psi}^{(0)}$, an initial value of $\boldsymbol{\Psi}$, for the calculation of

$$Q\left(\boldsymbol{\Psi};\ \boldsymbol{\Psi}^{(0)}\right) = \boldsymbol{E}_{\boldsymbol{\Psi}^{(0)}}[(l_c(\boldsymbol{\Psi}|\boldsymbol{y})]. \tag{3.17}$$

$Q\left(\boldsymbol{\Psi};\ \boldsymbol{\Psi}^{(0)}\right)$ is the expected value of the complete-data log likelihood, given $\boldsymbol{y}$ and $\boldsymbol{\Psi}^{(0)}$. The M-step then maximizes $Q\left(\boldsymbol{\Psi};\ \boldsymbol{\Psi}^{(0)}\right)$ with respect to $\boldsymbol{\Psi}$ over the parameter space Ω. We denote the resulting estimate of $\boldsymbol{\Psi}$ from this M-step $\boldsymbol{\Psi}^{(1)}$. The iteration is then repeated, but with $\boldsymbol{\Psi}^{(0)}$ replaced by $\boldsymbol{\Psi}^{(1)}$ in the new E-step. Continuing in this fashion, on the $(k+1)^{th}$ iteration, the following calculations are carried out:

E-step Take conditional expectation of $l_c(\boldsymbol{\Psi}|\boldsymbol{y})$, based on the observed data $\boldsymbol{y}$ and the fitted $\boldsymbol{\Psi}^{(k)}$ from the previous step.

$$Q\left(\boldsymbol{\Psi};\ \boldsymbol{\Psi}^{(k)}\right) = \boldsymbol{E}_{\boldsymbol{\Psi}^{(k)}}\{(l_c(\boldsymbol{\Psi}|\boldsymbol{y})\}. \tag{3.18}$$

M-step Choose $\boldsymbol{\Psi}^{(k+1)}$to maximize $Q\left(\boldsymbol{\Psi};\ \boldsymbol{\Psi}^{(k)}\right)$, that is

$$Q\left(\boldsymbol{\Psi}^{(k+1)};\ \boldsymbol{\Psi}^{(k)}\right) \geq\ Q\left(\boldsymbol{\Psi};\ \boldsymbol{\Psi}^{(k)}\right). \tag{3.19}$$

The iteration continues until $l\left(\boldsymbol{\Psi}^{(k+1)}\middle|\boldsymbol{y}\right) - l\left(\boldsymbol{\Psi}^{(k)}\middle|\boldsymbol{y}\right)$ is less than an arbitrarily small amount (McLachlan and Krishnan 2008). Dempster et al. (1977) proved that $l(\boldsymbol{\Psi}|\boldsymbol{y})$

is a non-decreasing function in the EM iteration, so convergence can be achieved with a sequence of likelihood values that are bounded above.

The EM algorithm replaces one difficult to solve likelihood maximization with a sequence of easier maximizations whose limit is the answer to the original question. Although it works particularly well in "missing data" problems, the definition of "missing data" can be stretched to accommodate many models (Casella and Berger 2002, sec. 7.2.4). For example, when $l(\boldsymbol{\Psi}|\boldsymbol{y})$ is difficult to work with, we could augment the observed data $\boldsymbol{y}$ into the complete data $\boldsymbol{x}$, and create a new log likelihood function $l_c(\boldsymbol{\Psi}|\boldsymbol{x})$ that has a simpler form (Lehmann and Casella 1998, p.457).

3.4 Applying the EM Algorithm to the Finite Mixture Model

3.4.1 Without Covariates

As discussed in Section 3.1, the finite mixture distribution can be generated through a random vector $\boldsymbol{Z} = (Z_1, \dots, Z_g)^T$, which follows a multinomial distribution with $n = 1$ and probability $\boldsymbol{\pi} = (\pi_1, \dots, \pi_g)^T$. That is, $P(Z_i = 1) = \pi_i, (i = 1, \dots g)$. $\boldsymbol{Z}$ can be viewed as the component label of the mixture distribution. If a random variable Y follows a g-component finite mixture distribution, when $Z_i = 1$, the density of Y comes from the ith component $f_i(Y)$. However, we don't observe $\boldsymbol{z}$, the realized value of the random vector $\boldsymbol{Z}$. Therefore $\boldsymbol{z}$ can be viewed as missing data. Following the above notation within the EM framework, we observe the incomplete data vector $\boldsymbol{y} = (y_1, \dots y_j)^T$, not the complete data vector $\boldsymbol{x} = \left(y_1, \dots y_j, \boldsymbol{z}_1, \dots, \boldsymbol{z}_j\right)^T$. Based on (3.4) and (3.5), the complete-data likelihood function is given by

$$L_c(\boldsymbol{\Psi}|\boldsymbol{x}) = f_c(\boldsymbol{x}|\boldsymbol{\Psi}) = \prod_{j=1}^{n} \prod_{i=1}^{g} \left[\pi_i f_i(y_j|\boldsymbol{\theta}_i)\right]^{z_{ij}}, \tag{3.20}$$

and the complete-data log likelihood function is

$$l_c(\boldsymbol{\Psi}|\boldsymbol{x}) = \sum_{j=1}^{n}\sum_{i=1}^{g} z_{ij}\{\log(\pi_i) + \log[f_i(y_j|\boldsymbol{\theta}_i)]\}, \tag{3.21}$$

where $\boldsymbol{\Psi} = (\pi_1, \dots, \pi_{g-1}, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_g)^T$ is the vector containing all the unknown parameters. This setup enables us to take advantage of the EM algorithm to find the MLEs.

The E-step

On the first iteration of the E-step, we take the conditional expectation of the complete-data log likelihood as defined in (3.17), given $\boldsymbol{y}$ and $\boldsymbol{\Psi}^{(0)}$, the initial value of $\boldsymbol{\Psi}$. Since $l_c(\boldsymbol{\Psi}|\boldsymbol{x})$ is linear in z_{ij}, the expectation of the complete-data log likelihood can be simplified to the conditional expectation of Z_{ij} given $\boldsymbol{y}$ and $\boldsymbol{\Psi}^{(0)}$. Since each Z_{ij} is a Bernoulli random variable, its expectation is equal to the posterior probability that y_j follows the distribution of the i[th] component of the mixture.

It follows that on the $(k+1)^{th}$ iteration, the E-step requires the computation of the conditional expectation of Z_{ij} given $\boldsymbol{y}$ and $\boldsymbol{\Psi}^{(k)}$, where $\boldsymbol{\Psi}^{(k)}$ is from the k^{th} iteration. Like in (3.15), we could write this expectation as

$$E_{\boldsymbol{\Psi}^{(k)}}(Z_{ij}|\,\boldsymbol{y}) = P_{\boldsymbol{\Psi}^{(k)}}(Z_{ij} = 1|\boldsymbol{y}) = \tau_{ij}\left(y_j\middle|\boldsymbol{\Psi}^{(k)}\right), \tag{3.22}$$

where

$$\tau_{ij}\left(y_j\middle|\boldsymbol{\Psi}^{(k)}\right) = \frac{\pi_i^{(k)} f_i\left(y_j\middle|\boldsymbol{\theta}_i^{(k)}\right)}{\sum_{h=1}^{g}\pi_h^{(k)} f_h\left(y_j\middle|\boldsymbol{\theta}_h^{(k)}\right)} \quad (i = 1, \dots, g; j = 1, \dots, n). \tag{3.23}$$

Substituting (3.23) into (3.21), we obtain

$$Q\left(\boldsymbol{\Psi};\, \boldsymbol{\Psi}^{(k)}\right) = \sum_{j=1}^{n}\sum_{i=1}^{g}\tau_{ij}\left(y_j\middle|\boldsymbol{\Psi}^{(k)}\right)\{\log(\pi_i) + \log[f_i(y_j|\boldsymbol{\theta}_i)]\}. \tag{3.24}$$

The M-step

On the $(k+1)^{th}$ iteration of the M-step, we maximize (3.24) with respect to $\boldsymbol{\Psi}$ in the parameter space Ω. By taking partial derivative of (3.24) with respect to π_i and equating it to 0, we get

$$\pi_i^{(k+1)} = \frac{\sum_{j=1}^{n} \tau_{ij}\left(y_j \middle| \boldsymbol{\Psi}^{(k)}\right)}{n} \ (i = 1, \dots, g). \tag{3.25}$$

Similarly, we establish that θ_i $(i = 1, \dots, g)$ satisfies

$$\sum_{j=1}^{n} \sum_{i=1}^{g} \tau_{ij}\left(y_j \middle| \boldsymbol{\Psi}^{(k)}\right) \left\{\frac{\partial}{\partial \boldsymbol{\theta}_i} \log\left[f_i(y_j|\boldsymbol{\theta}_i)\right]\right\} = \mathbf{0}. \tag{3.26}$$

Quite often the solution to (3.26) exists in closed form (McLachlan and Peel 2000). These values will be used in the E- and M-steps of the next iteration, and the process is repeated until $l(\boldsymbol{\Psi}^{(k+1)}) - l(\boldsymbol{\Psi}^{(k)})$ is less than an arbitrarily small amount.

3.4.2 With Covariates

In many applications, both the weights and the means of the component distributions can be modeled as functions of covariates. Let $\boldsymbol{w} = (w_1, \dots, w_m)^T$ and $\boldsymbol{v} = \left(v_1, \dots, v_p\right)^T$ be two vectors of covariates such that

$$\pi_{ij} = \pi_i\left(\boldsymbol{\alpha}_i \middle| \boldsymbol{w}_j\right) \ (i = 1, \dots, g; j = 1, \dots, n) \tag{3.27}$$

and

$$\boldsymbol{\theta}_{ij} = \boldsymbol{\theta}_i\left(\boldsymbol{\beta}_i \middle| \boldsymbol{v}_j\right) \ (i = 1, \dots, g; j = 1, \dots, n) \ . \tag{3.28}$$

The two vectors $\boldsymbol{w}$ and $\boldsymbol{v}$ may or may not have elements in common. The logistic model is quite often selected for the function $\pi_i\left(\boldsymbol{\alpha}_i \middle| \boldsymbol{w}_j\right)$. With the presence of covariates, the vector

of unknown parameters becomes $\boldsymbol{\Psi} = (\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_{g-1}, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_g)^T$, where $\boldsymbol{\alpha}_i = (\alpha_{i1}, \dots, \alpha_{im})^T$ $(i = 1, \dots, g-1)$ and $\boldsymbol{\beta}_i = (\beta_{i1}, \dots, \beta_{ip})^T$ $(i = 1, \dots, g)$. The complete-data log likelihood function (3.21) becomes

$$l_c(\boldsymbol{\Psi}|\boldsymbol{x}) = \sum_{j=1}^{n} \sum_{i=1}^{g} z_{ij}\{\log(\pi_{ij}) + \log[f_i(y_j|\boldsymbol{\theta}_{ij})]\}. \tag{3.29}$$

The E-step

The E-step is essentially the same as in Section 3.4.1. On the $(k+1)^{th}$ iteration, we take the conditional expectation of complete-data log likelihood function (3.29), given the data $\boldsymbol{y}$ and $\boldsymbol{\Psi}^{(k)}$**,** the estimated parameters from the previous iteration. After this E-step we get

$$Q(\boldsymbol{\Psi};\ \boldsymbol{\Psi}^{(k)}) = \sum_{j=1}^{n} \sum_{i=1}^{g} \tau_{ij}(y_j|\boldsymbol{\Psi}^{(k)})\{\log(\pi_{ij}) + \log[f_i(y_j|\boldsymbol{\theta}_{ij})]\}, \tag{3.30}$$

where

$$\tau_{ij}(y_j|\boldsymbol{\Psi}^{(k)}) = \frac{\pi_{ij}^{(k)} f_i(y_j|\boldsymbol{\theta}_{ij}^{(k)})}{\sum_{h=1}^{g} \pi_{hj}^{(k)} f_h(y_j|\boldsymbol{\theta}_{hj}^{(k)})} \quad (i = 1, \dots, g; j = 1, \dots, n). \tag{3.31}$$

The M-step

The M-step on the $(k+1)^{th}$ iteration involves finding the $\boldsymbol{\Psi}^{(k+1)}$ that maximizes (3.30). This requires solving the two systems of equations

$$\sum_{j=1}^{n} \sum_{i=1}^{g} \tau_{ij}(y_j|\boldsymbol{\Psi}^{(k)}) \frac{\partial}{\partial \boldsymbol{\alpha}}\{\log(\pi_{ij})\} = \mathbf{0} \tag{3.32}$$

and

$$\sum_{j=1}^{n} \sum_{i=1}^{g} \tau_{ij}(y_j|\boldsymbol{\Psi}^{(k)}) \frac{\partial}{\partial \boldsymbol{\beta}}\{\log[f_i(y_j|\boldsymbol{\theta}_{ij})]\} = \mathbf{0}. \tag{3.33}$$

Both equation (3.32) and (3.33) can be solved by procedures that fit the generalized linear models and are available in most statistical software such as SAS and R.

The difference between the cases without covariates and the cases with covariates is that in (3.23) and (3.24), π_i and $\boldsymbol{\theta}_i$ take the same value for all y_j, whereas in (3.30) and (3.31), they vary among y_j as a function of $\boldsymbol{w}_j \; and \; \boldsymbol{v}_j$.

3.4.3 Likelihood Function of Survival Data

Equation (3.6) describes the likelihood function in its general form. One unique feature of the survival analysis is that the data are often censored. Censoring takes place because patients in an experiment are not followed for the entire lifespan of the event of interest. Most of the time censoring is assumed to be random. Depending on the timing, censoring can be classified as right censoring, left censoring, and interval censoring (Lee and Wang 2003). Among them, right censored data are most often seen in clinical trials where patients are followed up to a certain timepoint, beyond which it becomes unknow whether or when the patients will experience the event of interest, if they have not done so during the follow-up period. Let T denote the time to the event of interest that is subject to random right censoring, and let C be the random censoring time. Instead of observing T directly, we observe $Y = \min(T, C)$, and $\Delta = I(T \leq C)$, where $I(.)$ is an indicator function. Assume T follows a distribution with density function $f(t)$ and survival function $S(t)$, we observe $f(t)$ and $S(t)$, when $\Delta = 1$ and $\Delta = 0$, respectively. Therefore for an iid sample of size n that consists of observed pairs of $(y_j, \delta_j), j = 1, \dots, n$, the likelihood function can be written as

$$\prod_{j=1}^{n} f(y_j)^{\delta_j} S(y_j)^{1-\delta_j} \tag{3.34}$$

and the likelihood function of an iid sample following a finite mixture distribution is simply

$$\prod_{j=1}^{n}[\sum_{i=1}^{g}\pi_i f_i(y_j)]^{\delta_j}\,[\sum_{i=1}^{g}\pi_i S_i(y_j)]^{1-\delta_j} \tag{3.35}$$

Substituting (3.35) for the density function in (3.23) and (3.31) gives the calculation of τ_{ij} when using right censored survival data.

3.5 Choosing Initial Values for the EM Algorithm via Censored Quantile Regression

As shown in (3.17), the EM algorithm requires $\boldsymbol{\Psi}^{(0)}$, the initial values of the unknown parameters, to start the iteration process. The choice of the initial values is crucial to the speedy convergence to the global maxima. Starting from a poor set of initial values will often lead to convergence to local maxima or no convergence at all (McLachlan and Peel 2000, Hipp and Bauer 2006). Over the years, researchers have developed several methods of choosing the initial values. One common technique is the random starting value approach, which divides the data into g clusters and randomly assign each observation into a cluster. Parameters estimated based on the initial random cluster assignment are used as the initial values for the EM algorithm. A related method that builds on the random starting value approach is the iteratively constrained EM technique. It involves running several iteratively constrained EM algorithms and selecting the parameters from the best-fitting solutions as the initial values (Lubke and Muthén 2007). The k-means clustering technique uses results from the k-means algorithm as the initial values for the EM algorithm, and has been implemented in the R package "mixture" (Browne et al. 2014). Another R package "mclust", adopts the agglomerative hierarchical clustering technique for selection of the initial values (Fraley and Raftery 2006). Other popular techniques include utilizing principal component analysis (McLachlan 1988), using results from moment estimation (Lindsay and Basak 1993, Furman and Lindsay 1994), and starting with well-separated

values (Böhning et al. 1994). These methods are evaluated and compared by Karlis and Xekalaki (2003) and Shireman et al. (2017). They find that each technique has its strengths and drawbacks. While some techniques outperform the others in simulations, it is difficult to characterize situations where a technique can be expected to outperform the others when working with empirical data.

Most of the current techniques for choosing the initial values and their comparisons are implemented in the setting of finite mixture of normal distributions, as this is the most commonly employed mixture model (Biernacki et al. 2003). Although these techniques can still be applied in the analysis of survival data, they are unlikely to provides the optimal set of initial values due the fact that none of them considers the censoring feature of the survival data. For this reason, we propose an alternative procedure to select the initial values, which is based on the censored quantile regression of survival data.

For a random variable Y, the τ-th quantile is the value y such that $P(Y \leq y) = F_Y(y) = \tau$. The quantile function is defined as

$$Q_Y(\tau) = F_Y^{-1}(\tau) = inf\{y: F(y) > \tau\} \tag{3.36}$$

for $\tau \in [0,1]$. Another presentation of quantiles treats the τ-th quantile as the solution to the minimization problem below

$$q_\tau = \underset{c}{\operatorname{argmin}}\, E[\rho_\tau(Y - c)], \tag{3.37}$$

where $\rho_\tau(.)$ is an asymmetric absolute loss function defined as

$$\rho_\tau(Y - c) = [\tau - I(Y \leq c)](Y - c) = [(1 - \tau)I(Y \leq c) + \tau I(Y > c)]|Y - c|.$$

This is a weighted sum of absolute deviations, where negative deviations are weighted by $(1 - \tau)$ and positive deviations are weighted by τ. The definition of the median is a special

case of (3.37) when $\tau = 0.5$ (Davino 2014). Under this loss function, the quantiles can be viewed as particular centers of the distribution (Hao 2007).

In survival data analysis with covariates, where Y is time to the event of interest and $\boldsymbol{x}$ is a vector consisting of the covariates, a quantile regression model linearly links $Q_Y(\tau)$ to $\boldsymbol{x}$. Quite often, Y is transformed into $\log Y$, so

$$Q_{logY}(\tau|\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta}(\tau) + Q_{\boldsymbol{\varepsilon}}(\tau), \tag{3.38}$$

where $Q_{\boldsymbol{\varepsilon}}(\tau)$ is the τ-th quantile of the error term ε on the log scale (Xue et al. 2016). It follows that equation (3.37) can be extended to

$$\widehat{\boldsymbol{\beta}}(\tau) = \underset{\boldsymbol{\beta}(\tau)}{\operatorname{argmin}} E[\rho_\tau(\log(\boldsymbol{Y}) - \boldsymbol{x}^T\boldsymbol{\beta}(\tau)]. \tag{3.39}$$

There are two popular procedures for the estimation of $\widehat{\boldsymbol{\beta}}(\tau)$. Portnoy (2003) used a recursively reweighted estimation procedure, while Peng and Huang (2008) adopted a martingale-based estimating equation. The two methods often result in similar estimates, and both are available in R and SAS (Xue et al. 2016).

Our proposed approach of selecting the initial values of the EM algorithm builds upon the results of the censored quantile regression. This approach takes advantage of the fact that quantiles can be viewed as special centers of the distribution. By grouping the data into selected quantiles that represent the number of the components in the mixture model, we could estimate both the weights and parameter values for each individual component and use these estimates to start the EM algorithm. Assuming we have an iid sample of $Y_1, Y_2, \ldots, Y_n$ observations where Y_j is the observed survival time (censored or not) to the event of interest, this procedure is implemented in the following steps. For illustration purposes, we assume the mixture has two components.

1. Fit censored quantile regression for the different quantiles in a set Q of size 30 to 100 and extract the residuals. The distribution model chosen in the quantile regression will correspond to that in the finite mixture model. The effects of covariates on the weights, if any, are incorporated by adding these covariates to the quantile regression. Define q_i as the i-th quantile, r_{ij} as the residual of y_j for q_i.
2. For every subset S of Q, where $S = \{q_i, q_k\}$ for $i \neq k$, calculate $C(S) = \sum_{j=1}^{n}\left[\min(|r_{ij}|, |r_{kj}|)\right]^2$, the sum of squares of the minimum absolute residual between q_i and q_k. Select the set $\hat{S} = \{\hat{q}_i, \hat{q}_k\}$ that minimizes $C(S)$. The quantiles $\hat{q}_i$ and $\hat{q}_k$ roughly represent the two components.
3. Define $w_{ij} = \frac{|r_{ij}|}{|r_{ij}|+|r_{kj}|}$. This is the weight, or the probability that Y_j belongs to $\hat{q}_i$. Define w_{kj} in a similar fashion.
4. Fit two parametric survival models with the weights from Step 3 respectively and get the parameter estimates for each individual model.
5. Use the weight and parameter estimates from Steps 3 and 4 as the initial value for the EM algorithm.

Compared with the other existing techniques of finding the initial values for the EM algorithm, this procedure possesses a major advantage. Censoring is a special feature of the survival data and none of the current methods has the mechanism to include its impact on finding the initial values. Our proposed approach considers the effects of censoring in the observed data through the censored quantile regression.

3.6 Construction of Confident Intervals

In the literature of finite mixture models there are two approaches to obtain the confidence intervals of $\widehat{\boldsymbol{\Psi}}$, or any function of it $\tau(\widehat{\boldsymbol{\Psi}})$. One is based on the observed information matrix (Dietz and Böhning 1995, Liu 1998), the other is based on bootstrap (McLachlan and Krishnan 2008). Basford et al. (1997) compared these two approaches and concluded that the results from the information-based approach is unstable unless the sample size was every large. The bootstrap method was first introduced by Efron (1979). Efron (1981) also studied how to conduct bootstrap with censored data and proposed a technique he called the Percentile Method to construct the confidence intervals for small samples. We follow Efron's method as outlined below to obtain the confidence intervals for the $\tau(\widehat{\boldsymbol{\Psi}})$ of interest.

1. Given a sample of size n consisting of pairs of observed time to event (censored or not) and censoring indicators, estimate $\tau(\widehat{\boldsymbol{\Psi}})$ following the methods discussed in sections 3.4 and 3.5.
2. From the given sample, draw independently pairs of the time to event and the censoring indicator n times with replacement. This creates a bootstrap sample of the observed data.
3. Estimate $\tau(\boldsymbol{\Psi}^{\wedge})$ based on this bootstrap sample and call it $\tau(\widehat{\boldsymbol{\Psi}}^{*})$.
4. Repeat steps 2 and 3 a large number of times, for example $N = 1000$, yielding bootstrap values $\tau(\widehat{\boldsymbol{\Psi}}^{*}{}_{1}), \tau(\widehat{\boldsymbol{\Psi}}^{*}{}_{2}), \dots, \tau(\widehat{\boldsymbol{\Psi}}^{*}{}_{N})$ from which we can get the bootstrap distribution of $\tau(\widehat{\boldsymbol{\Psi}})$ and the associated percentiles. The confidence intervals can then be constructed from the percentiles. For example, the 95% confidence interval is given by $\{\tau(\widehat{\boldsymbol{\Psi}}^{*})_{0.025}, \tau(\widehat{\boldsymbol{\Psi}}^{*})_{0.975}\}$, where $\tau(\widehat{\boldsymbol{\Psi}}^{*})_{0.025}$ and

$\tau(\widehat{\boldsymbol{\Psi}}^*)_{0.975}$ are the 2.5% and 97.5% percentiles of the bootstrap distribution, respectively.

3.7 Common Parametric Survival Models

As discussed briefly in the previous section, several non-parametric and parametric models have been developed over the course of survival data analyses. These models generally fall into two families: the proportional hazard model or the accelerated failure time (AFT) model. Let y be the survival time, $\boldsymbol{x}$ a vector of the covariates. Under the proportional hazard model, the hazard function is written as

$$h(y) = h_{\mathbf{0}}(\boldsymbol{y})g(\boldsymbol{x}), \tag{3.40}$$

where $h_0(y)$ is the baseline hazard function or the hazard function of a reference group, and $g(\boldsymbol{x})$ as a function of only the covariates reflects the effects of these covariates on the underlying risk $h_0(y)$. If subject j has survival time y_j and covariates $\boldsymbol{x}_j$, and subject k has survival time y_k and covariates $\boldsymbol{x}_k$, then the hazard ratio of the two patients is $h(y_j)/h(y_k) = g(x_j)/g(x_k)$. This ratio is constant over time and only a function of the covariates, hence the name of this family of models.

Depending on the form of $h_0(y)$, there could be both nonparametric/semi-parametric and parametric models in the proportional hazard family. The well-known Cox proportional hazard model is an example of the former. Under this model, $h_0(y)$ is unspecified and $g(\boldsymbol{x})$ is equal to $e^{\boldsymbol{x}^T\boldsymbol{\beta}}$, where $\boldsymbol{\beta}$ is a vector of the unknown parameters. While the Cox proportional hazard model has been widely implemented to assess the impacts of covariates on survival risk, it cannot make predictions about the survival time as $h_0(y)$ is unspecified.

When both $h_0(y)$ and $g(\boldsymbol{x})$ assume to follow certain parametric forms, some parametric models also possess the feature of proportional hazard. One such an example is the Weibull model. There are two parameters in the Weibull distribution, the shape parameter γ, and the scale parameter λ. The Weibull baseline hazard function can be written as $h_0(y) = \lambda \gamma y^{\gamma-1}$. When $g(\boldsymbol{x}) = e^{\boldsymbol{x}^T\boldsymbol{\beta}}$, as in the Cox proportional hazard model, the hazard function becomes $h(y) = \lambda e^{\boldsymbol{x}^T\boldsymbol{\beta}} \gamma y^{\gamma-1}$. It is clear that Y follows another Weibull distribution with shape parameter γ and scale parameter $\lambda e^{\boldsymbol{x}^T\boldsymbol{\beta}}$. Therefore, when the shape parameter γ is assumed to be constant and the scale parameter λ is a function of the covariates, the Weibull model follows a parametric distribution which also presents the trait of proportional hazard.

In contrast, the AFT family contains primarily parametric models. The general form of models in this family links the logarithm of the survival time to the covariates by the equation

$$\log(y) = \boldsymbol{x}^T\boldsymbol{\beta} + \sigma\varepsilon, \tag{3.41}$$

where, σ is an unknown scale parameter, and ε is a random variable with a known density function $f(\epsilon)$, which determines the distribution that y follows, such as the Weibull, Gamma, Lognormal, and Log-logistic distribution. The AFT family gets its name because we can re-write (3.41) as $y = e^{\boldsymbol{x}^T\boldsymbol{\beta}+\sigma\varepsilon}$, so y is increased or decreased when $\boldsymbol{x}^T\boldsymbol{\beta} > 0$ or $\boldsymbol{x}^T\boldsymbol{\beta} < 0$, respectively (Lee and Wang 2003).

Table A. 16 in Appendix B summarizes these parametric proportional hazard and AFT models. Among them, the Weibull and the exponential distribution can be parameterized both as a proportional hazard and AFT model. The exponential distribution is a special case of the Weibull distribution when $\gamma = 1$. The Gompertz distribution is

closely related to the Weibull distribution and sometimes is called the log-Weibull distribution. The gamma distribution simplifies to the exponential distribution when $\alpha = 1$. The Generalized gamma distribution reduces to the gamma, lognormal, and Weibull when $Q = 0$, $Q = 1$, and $Q = \sigma$, respectively. The Generalized F distribution is equivalent to the Generalized gamma distribution with $P = 0$. It is also equivalent to the log-logistic distribution when $P = 1$.

The individual component of a finite mixture model most often comes from one these models. In the following chapters, we will fit finite mixture models as well as these single parametric models to various survival data and compare their performances under different conditions.

Chapter 4: Analyzing Survival Data with the Finite Mixture Models

Chapter 3 introduces the finite mixture models and the EM algorithm as a popular technique to estimate their parameters. A new approach based on the censored quantile regression is also proposed to choose the initial values for the EM algorithm. This approach takes into consideration the feature of censoring in survival data, and therefore will improve the likelihood of convergence of the EM algorithm to the global maximum.

In this chapter, we will apply the methods developed in Chapter 3 and fit finite mixture models to several sets of survival data. These data sets consist of both simulated and empirical data. Together they represent some typical situations where the finite mixture models can provide more accurate estimates and extrapolations. The common parametric survival models in section 3.6 are also fit to these data, and the results are compared to those of the finite mixture models.

The component distribution in the finite mixture model is chosen to be the Weibull, due to its flexibility and intuitive interpretation (Marín et al. 2005).

4.1 Simulated Data

An iid sample of 300 observations is generated to simulate a time-to-event random variable T which follows a mixture of two Weibull distributions. For the first Weibull distribution, the weight π_1 is set to equal 0.75, shape parameter $\gamma_1 = 1.5$ and scale parameter $\lambda_1 = 2$. For the second Weibull distribution, $\pi_2 = 0.25$, $\gamma_2 = 1$, and $\lambda_2 = 10$. Censoring is through another random variable T^{censor}, which also follows a mixture of two Weibull distributions. The two component Weibull distributions in the mixture for T^{censor} have scale 12 and 20, respectively. They share the same weights and shape

parameters as their counterparts in the mixture for T. In addition, the data are truncated at $T = 5$. Therefore, an observation with either T> T^{censor} or T> 5 is censored. This results in approximately a 23% censoring rate. What we can directly observe from the simulation is not T, instead, we observe pairs of $(y_j, \delta_j), j = 1, \dots, 300$, where $y_j = \min(t_j, t_j^{censor}, 5)$ and $\delta_j = 1\ if\ t_j < \min(t_j^{censor}, 5)$. No covariates are included in this simulation.

Under this setting, the parameter vector is $\boldsymbol{\Psi} = (\pi_1, \gamma_1, \lambda_1, \gamma_2, \lambda_2)^T$, which includes one of the weights, the shape and scale parameter for each of the two component Weibull distributions. To start the EM algorithm, we follow the steps outlined in section 3.5 to find the initial value $\boldsymbol{\Psi}^{(0)}$. The stopping rule for the EM algorithm is set to be $l(\boldsymbol{\Psi}^{(k+1)}) - l(\boldsymbol{\Psi}^{(k)}) < 0.0001$. That is, when the difference in the log likelihood between the $(k+1)^{th}$ and the k^{th} iteration is less than 0.0001, we consider the EM algorithm has reached convergence and set $\widehat{\boldsymbol{\Psi}} = \boldsymbol{\Psi}^{(k+1)}$. Based on $\widehat{\boldsymbol{\Psi}}$, we can estimate any function $\tau(\boldsymbol{\Psi})$ by $\tau(\widehat{\boldsymbol{\Psi}})$, such as the density, survival function, hazard function, restricted mean survival time, and the mean survival time. If $k > 5000$ and the stopping rule is still not satisfied, the EM algorithm is considered failing to converge.

The parametric models covered in section 3.7 are fit to the simulated data and their performance are compared with that of the mixture model. As the exponential distribution is a special case of the Weibull distribution, it is not included in the comparison. The KM estimate is a non-parametric maximum likelihood estimator of the survival function. It is widely used as a surrogate of the true survival function (Lee and Wang 2003). For this reason, we include the KM estimator as a benchmark in addition to the true underlying survival distribution.

Survival function plots from all models are generated and overlaid to that of the true mixture distribution as well as the KM estimate to give a visual presentation of the model fitting. We also choose log likelihood and AIC as model fitting diagnostics (Solka et al. 1998). In addition, we compare restricted mean survival time and extrapolated mean survival time calculated from these models, as these are the most important measures of benefits of health interventions and hence the primary reason for utilizing the parametric models.

The results are included in Figure 5-A to C as well as Table 1. For the mixture of two Weibull model, the 95% confidence intervals of the survival function, restricted mean survival time at T=5, and mean survival time are constructed following the bootstrap approach discussed in section 3.6.

The survival functions for the true mixture distribution and each of the models are plotted in Figures 5-A to C. For each model, parameters are estimated with observed data up to $T = 5$, and extrapolations about the survival function are made up to $T = 35$. Among all the models, the mixture of two Weibull model yields a survival function that most closely traces the true function, as shown in Figure 5-A. In comparison, some of the single parametric models generate good estimate of the survival function as well, such as the log-logistic and Generalized F, while others fail to do so, such as the lognormal and Gompertz. Model fitting diagnostics are displayed in Table 1. The mixture of two Weibull model has the best log likelihood (-452.64), followed by the log-logistic (-456.24). In terms of AIC, the mixture of two Weibull model also claims the best value (915.29). The true restricted mean survival time at $T = 5$ is 2.33. This is the time point up to which data are observed. Again, the mixture of two Weibull model provides very close estimates (2.32). When it

comes to the mean survival time – one of the most important in the survival analysis, the mixture of two Weibull has an extrapolated value of 3.89, which is the closest to the true mean (3.85). The Generalized F and Gompertz model, on the other hand, are unable to estimate this value.

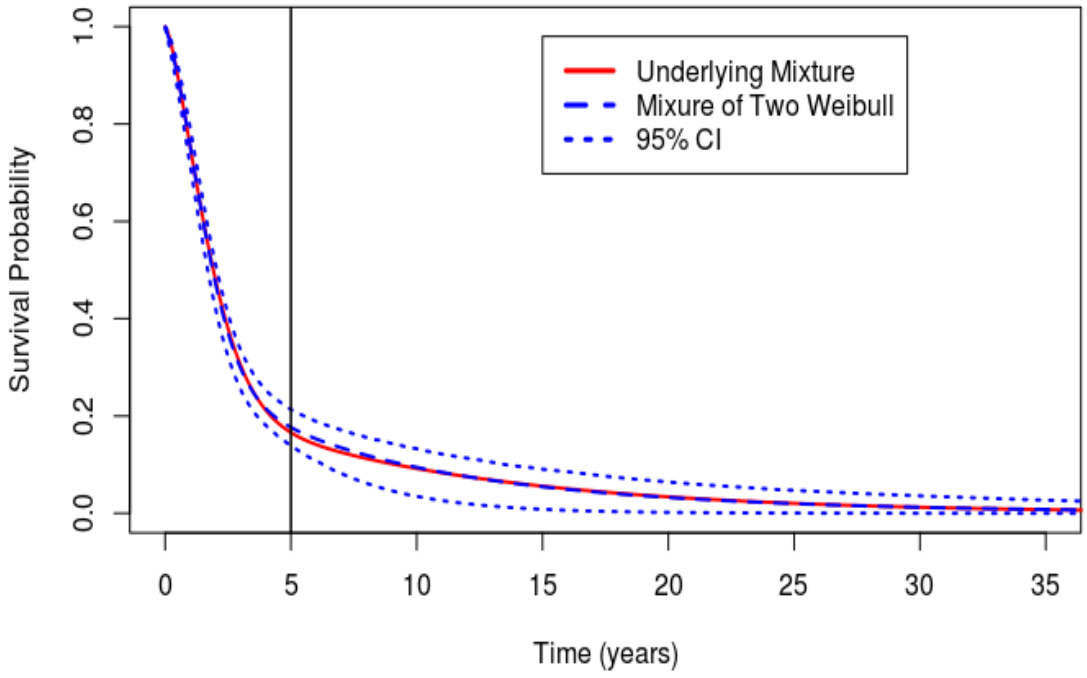

Figure 5-A Extrapolation of Survival Probability – Simulated Data Without Covariates

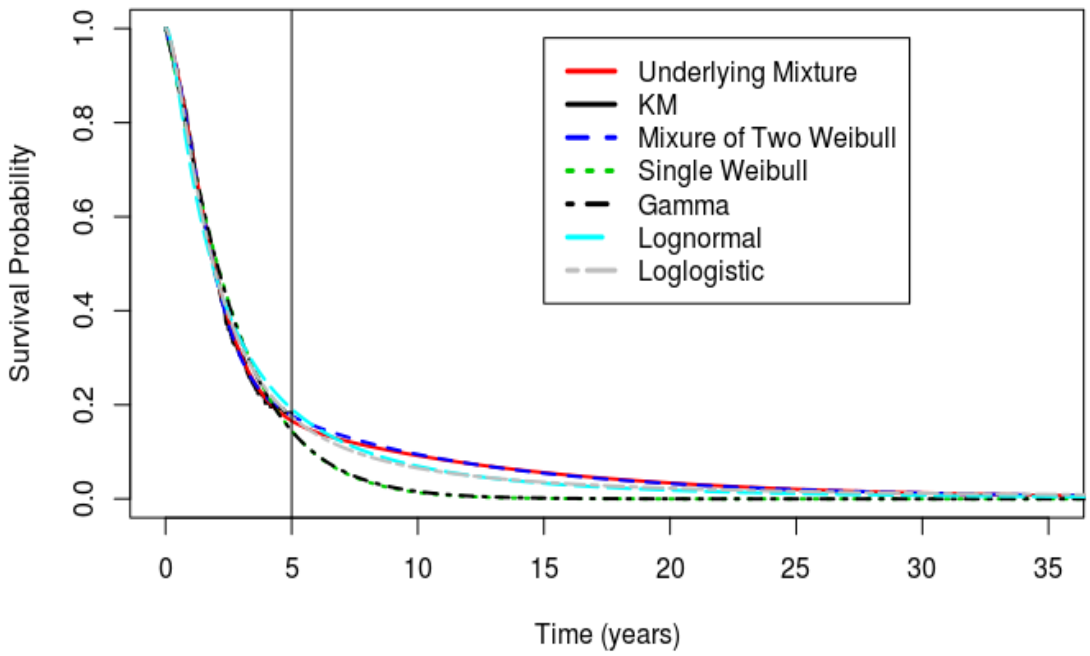

Figure 5-B Extrapolation of Survival Probability – Simulated Data Without Covariates

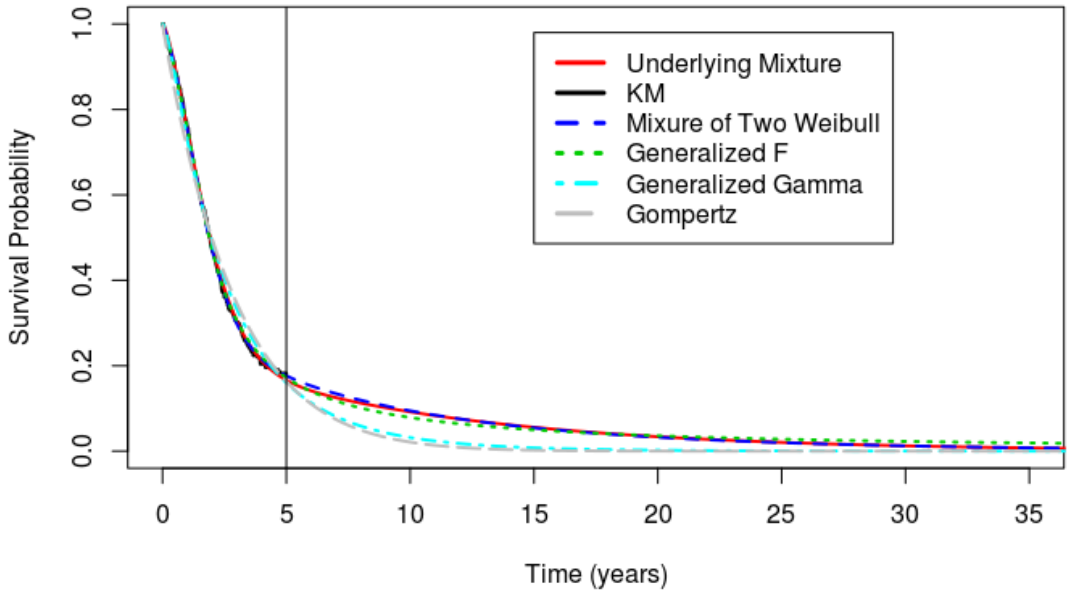

Figure 5-C Extrapolation of Survival Probability – Simulated Data Without Covariates

Table 1 Comparison of Mixture of Two Weibull and Common Parametric Models Simulated Data Without Covariates

Models	Log Likelihood	AIC	Restricted Mean (95% CI) T=5	Extrapolated Mean (95% CI) Lifetime
Underlying model			2.33	3.85
KM			2.31	
Mixture of two Weibull	-452.64	915.29	2.32 (2.16-2.47)	3.89 (2.75-5.29)
Single Weibull	-461.13	926.25	2.38 (2.19-2.56)	2.69 (2.39-3.01)
Gamma	-459.94	923.88	2.37 (2.20-2.56)	2.70 (2.40-3.04)
Lognormal	-462.75	929.51	2.35 (2.16-2.56)	3.54 (2.97-4.46)
Log-logistic	-456.24	916.47	2.35 (2.15-2.54)	4.04 (3.24-5.35)
Generalized F	-456.92	921.84	2.30 (2.15-2.56)	
Generalized Gamma	-458.23	922.46	2.36 (2.18-2.56)	2.86 (2.47-3.46)
Gompertz	-464.13	932.26	2.35 (2.14-2.55)	

Next we report the results from a simulation where both the weights and the shapes of the mixture model are functions of covariates. Like in the previous case, the random variable T follows a mixture of two Weibull distribution. The iid sample of observed pairs (y_j, δ_j) contains 300 observations. The covariate vector is $\boldsymbol{X} = (X_1, X_2)^T$, where X_1 is a continuous variable following a normal distribution $N(5,1)$, and X_2 follows a Bernoulli distribution with $P(X_2 = 1) = 0.5$. The weight π_1 is linked to the covariates by $\pi_1 = \frac{\exp(\alpha_1+\alpha_2*X_1+\alpha_3*X_2)}{1+\exp(\alpha_1+\alpha_2*X_1+\alpha_3*X_2)}$, and $\pi_2 = \frac{1}{1+\exp(\alpha_1+\alpha_2*X_1+\alpha_3*X_2)}$, where $\alpha_1 = 0.2$, $\alpha_2 = 0.1$, and $\alpha_3 = 0.3$, respectively. The first Weibull distribution has $\gamma_1 = 1.5$ and the second Weibull distribution has $\gamma_2 = 2$. The scale parameter λ_i is linked to the covariates by $\lambda_i =$

$e^{\beta_{i1}+\beta_{i2}*X_1+\beta_{i3}*X_2}$, where $\beta_{11}=\beta_{21}=\log(2), \beta_{12}=0.1, \beta_{13}=0.2, \beta_{22}=0.3$, and $\beta_{23}=0.5$, respectively. The censoring variable T^{censor} also follows a mixture of two Weibull distribution. Each of the two components of T^{censor} has the same shape parameter as their counterparts in T. The two scale parameters are linked to the covariates by $\lambda_1^{censor}=e^{\log 2+0.3*X_1+0.8*X_2}$ and $\lambda_2^{censor}=e^{\log 2+0.5*X_1+X_2}$. The data are truncated at $T=8$. If T$>T^{censor}$ or T>8, T is censored. This results in approximately a 23% censoring rate. The parameter vector of the mixture of two Weibull model becomes $\boldsymbol{\Psi}=(\alpha_1,\alpha_2,\alpha_3,\beta_{11},\beta_{12},\beta_{13},\beta_{21},\beta_{22},\beta_{23},\gamma_1,\gamma_2)^T$.

The mixture model as well as the other parametric models are fit to the simulated data. The estimation is based on observed values up to $T=8$. The extrapolation of survival function is made up to $T=35$. To plot the survival function at given time $T=t$, we take the average of the estimated survival probability at t and $\boldsymbol{X}_j=\boldsymbol{x}_j, j=1,\ldots,300$. That is, the survival probability at a time t is the average of the probabilities at this time point across all observed values of the covariates. The results are displayed in Figures 6-A to C and Table 2.

In Figures 6-A to C, the red solid line represents the true underlying survival function. During the period when data are available, all models slightly over-estimate the survival probability. During the extrapolation period, the mixture of two Weibull model rather accurately predicts the survival probability at all time points. However, the other parametric models either over-estimate (such as the lognormal or Generalized F) or under-estimate (such as the single Weibull or Gompertz) it. In Table 2, the mixture model reports the best log likelihood (-554.05) and AIC (1130.09), despite the fact that it has the largest number of parameters to estimate.

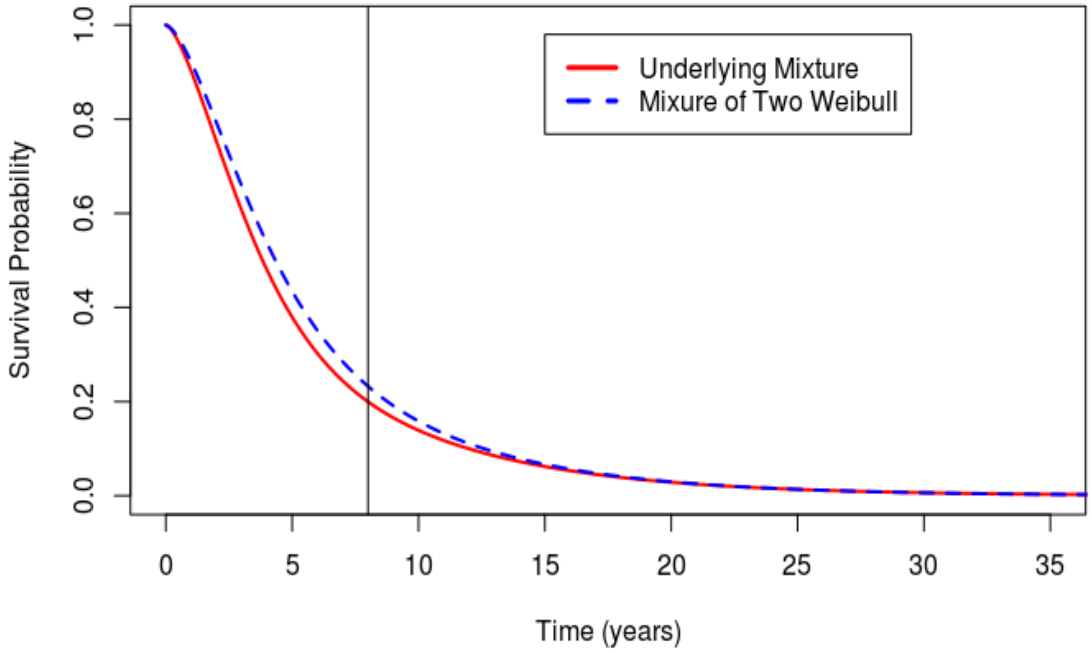

Figure 6-A Extrapolation of Survival Probability – Simulated Data with Covariates

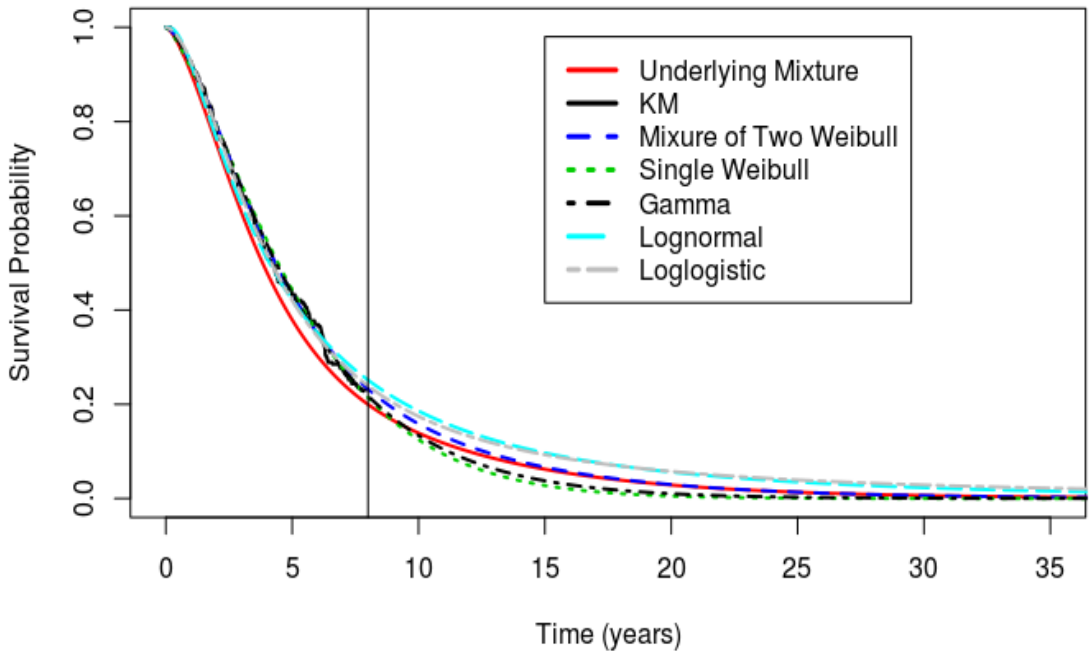

Figure 6-B Extrapolation of Survival Probability – Simulated Data with Covariates

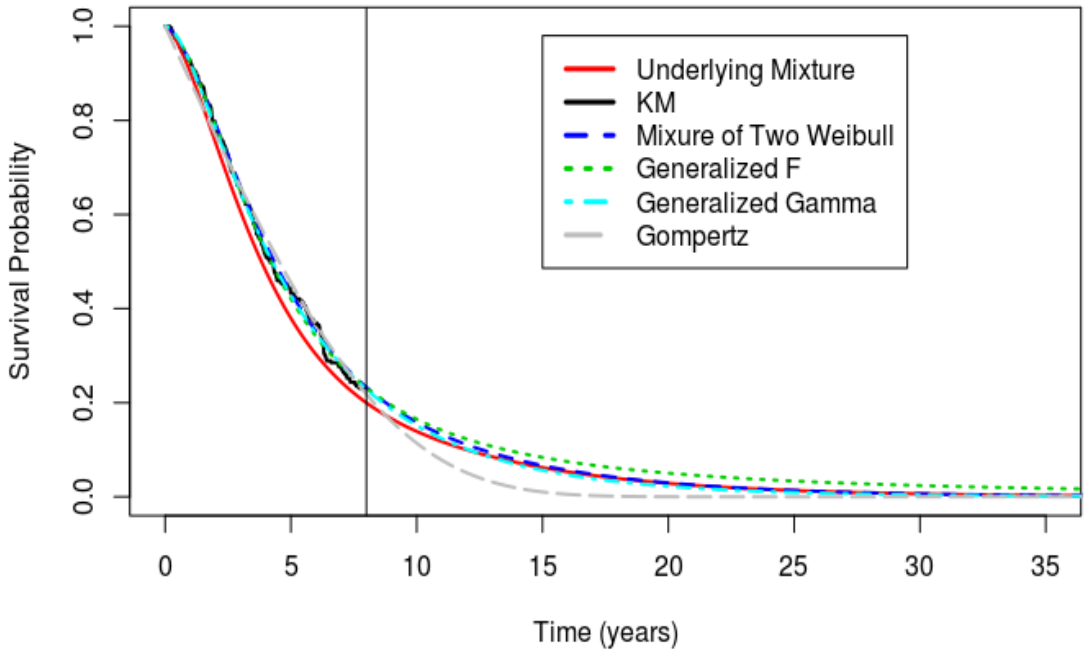

Figure 6-C Extrapolation of Survival Probability – Simulated Data with Covariates

Table 2 Comparison of Mixture of Two Weibull and Common Parametric Models Simulated Data with Covariates

Models	Log Likelihood	AIC
Mixture of two Weibull	-554.05	1130.09
Single Weibull	-568.53	1145.07
Gamma	-567.14	1142.28
Lognormal	-569.05	1146.11
Log-logistic	-566.49	1140.97
Generalized F	-566.23	1144.46
Generalized Gamma	-566.43	1142.86
Gompertz	-574.39	1156.78

4.2 Empirical Data

In the previous section we test the mixture of two Weibull model on simulated data and compared its performance with that of the common single parametric models. In this

section we repeat the exercise by fitting these models to data collected from SHEP, the RCT introduced in section 2.2.

At the time of writing this dissertation, we have information about mortality and cause of death for SHEP patients until December 31st, 2014. To test whether the mixture models are at an advantage over single parametric models in making extrapolations, we fit all these models to SHEP data up to December 31st, 2010 and predict survival function from 2011 to the end of 2014.

We first examine the survival probability from cardiovascular related mortality, and fit all the candidate models to the data without considering the effects of covariates. Since the true underlying survival distribution is unknown, we use the KM estimate as the benchmark. The results are displayed in Figures 7-A to C and Table 3. As shown in Figure 7-A, the mixture of two Weibull model accurately estimate the survival probability during the observation period. It also provides very close prediction from 2011 to 2014. The 95% confidence interval is quite narrow due to the large sample size. Among the single parametric models, the generalized F model fails to converge, the others either over or under estimate the survival function during the observation period, and they all make under prediction between 2011 and 2014. Both the log likelihood and AIC in Table 3 indicate that the mixture model provides the best fit (-7396.99 and 14803.99, respectively), followed by the Log-logistic (-7429.83 and 14863.66, respectively). The extrapolated restricted mean survival time up to the end of the observation period ($T = 26$) by these models are close to that of the KM model. This is because the over and under estimation during the observation period cancels each other out. The extrapolated mean survival time

from cardiovascular related mortality is 50 years based on the mixture model. The single models estimate this value as low as 28 and as high as 70.

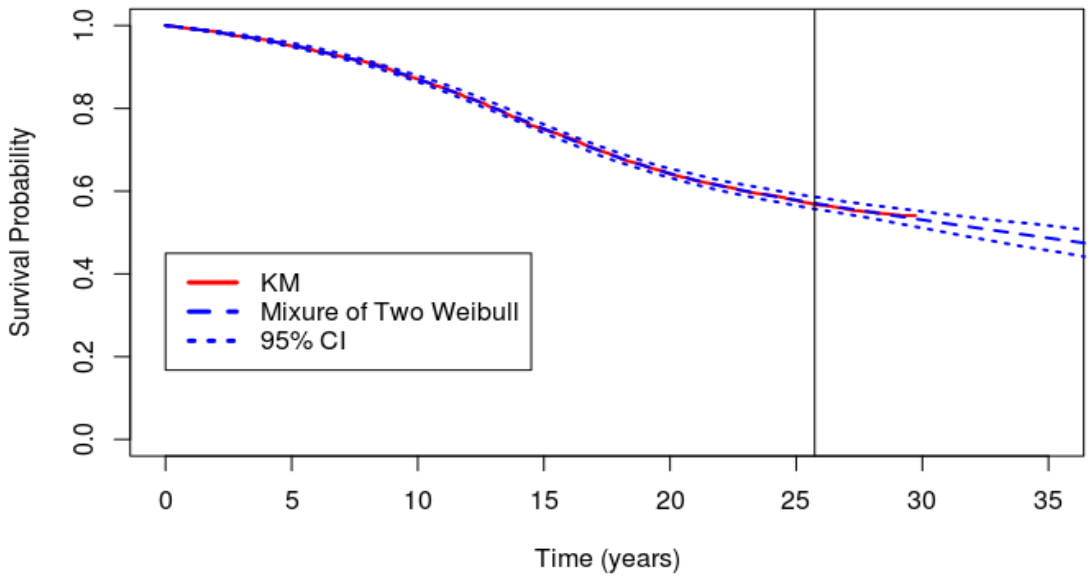

Figure 7-A Extrapolation of Survival Probability 2011-2014 in SHEP – CV-related Mortality Without Covariates – All Patients

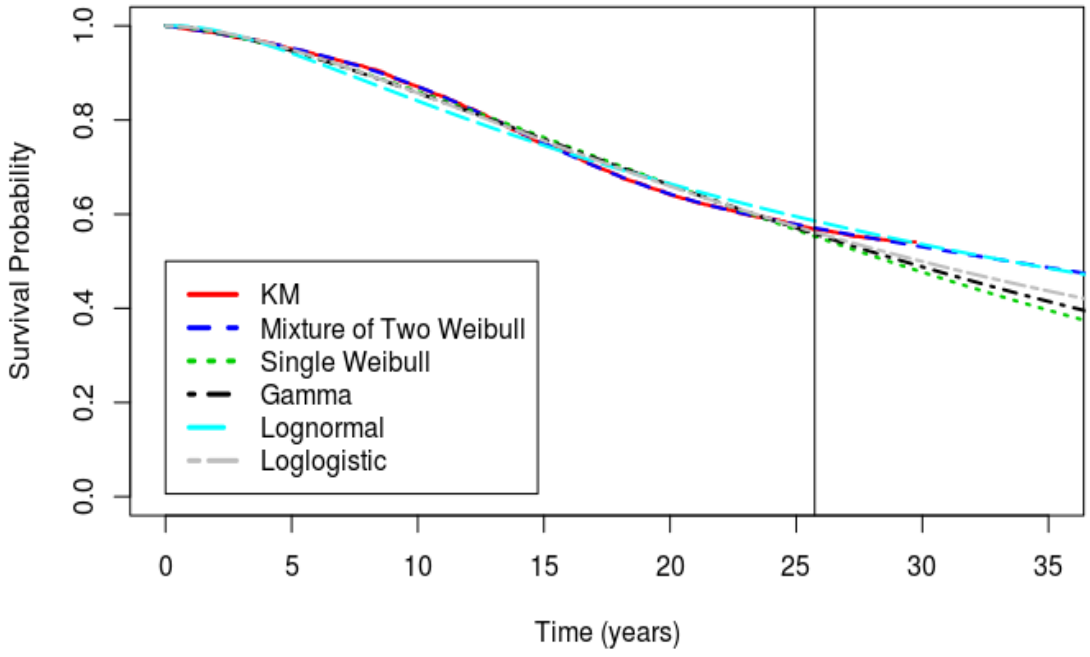

Figure 7-B Extrapolation of Survival Probability 2011-2014 in SHEP – CV-related Mortality Without Covariates – All Patients

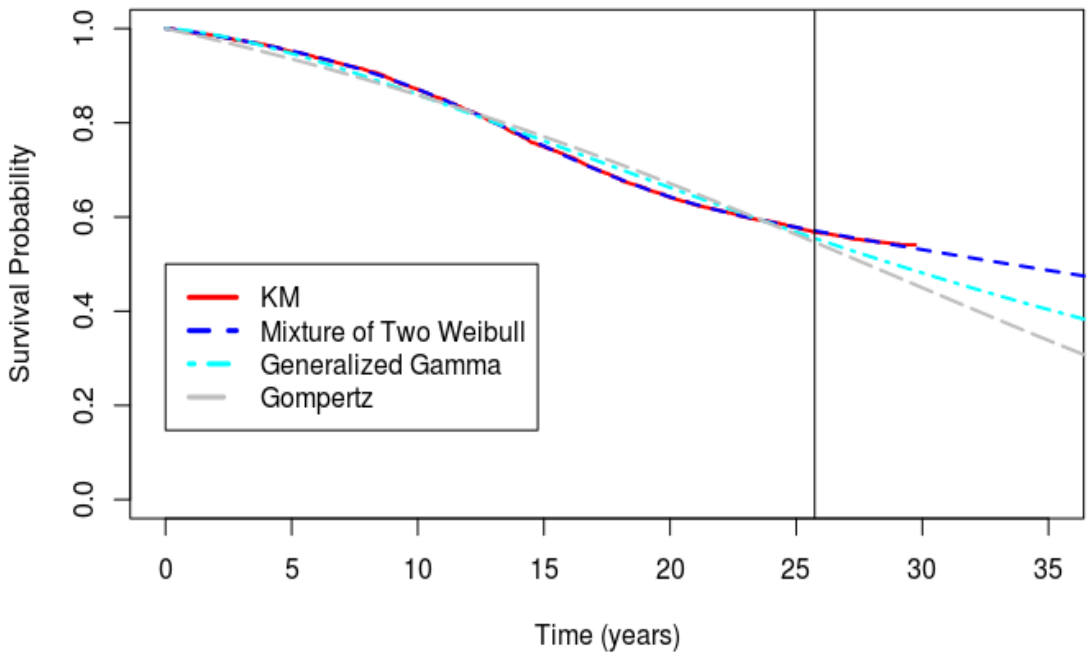

Figure 7-C Extrapolation of Survival Probability 2011-2014 in SHEP – CV-related Mortality Without Covariates – All Patients

Table 3 Comparison of Mixture of Two Weibull and Common Parametric Models CV-related Mortality - SHEP Data Without Covariates

Models	Log Likelihood	AIC	Restricted Mean (95% CI) T=26	Extrapolated Mean (95% CI) Lifetime
KM			20.7	
Mixture of two Weibull	-7396.99	14803.99	20.46 (20.29-20.68)	50.01 (42.25-70.48)
Single Weibull	-7432.72	14869.44	20.53 (20.30-20.75)	33.43 (31.85-35.16)
Gamma	-7433.90	14871.81	20.5 (20.30 – 20.70)	36.41 (34.73 – 38.22)
Lognormal	-7492.32	14988.63	20.41 (20.18-20.65)	69.74 (63.07-77.90)
Log-logistic	-7429.83	14863.66	20.49 (20.26-20.70)	61.30 (55.70-68.31)
Generalized F	Cannot be estimated			
Generalized Gamma	-7432.30	14870.61	20.52 (20.30-20.75)	34.55 (31.77-38.07)
Gompertz	-7465.07	14934.15	20.52 (20.29-20.73)	28.02 (26.91-29.32)

Next, we examine the effects of covariates on the survival function. Based on recommendations from experts in cardiovascular disease, we select treatment arm, body mass index (BMI), and age as the covariates. This increases the number of parameters to 14 in the finite mixture model. The survival function at a given time t is calculated in the same fashion as in the simulated data case.

The results are displayed in Figures 8-A to C and Table 4. During both the observation and extrapolation period, the estimate from the mixture model is slightly below the KM estimate. The single parametric models, while being able to trace the KM estimate closely during the observation period, all fall below the KM curve by a large amount during the extrapolation period. This indicates that with available data up to the end of 2010, these single parametric models will under estimate the survival probability from 2011 to 2014.

In terms of model fitting, the mixture model is better than the other models by a large margin. As shown in Table 4, for the mixture model, the log likelihood is -7151.51 and AIC is 14331.01. The next best values are from the log-logistic, with log likelihood=-7221.28 and AIC=14452.57. This is indicative that the SHEP patients are heterogeneous, and that by considering the effects of covariates that are related to this heterogeneity, the finite mixture model provides a better fit and more reliable extrapolation.

Similar results are obtained when we repeat the exercise on all-cause mortality, as shown in Figures 9-A to 10-C, and Table 5 and Table 6.

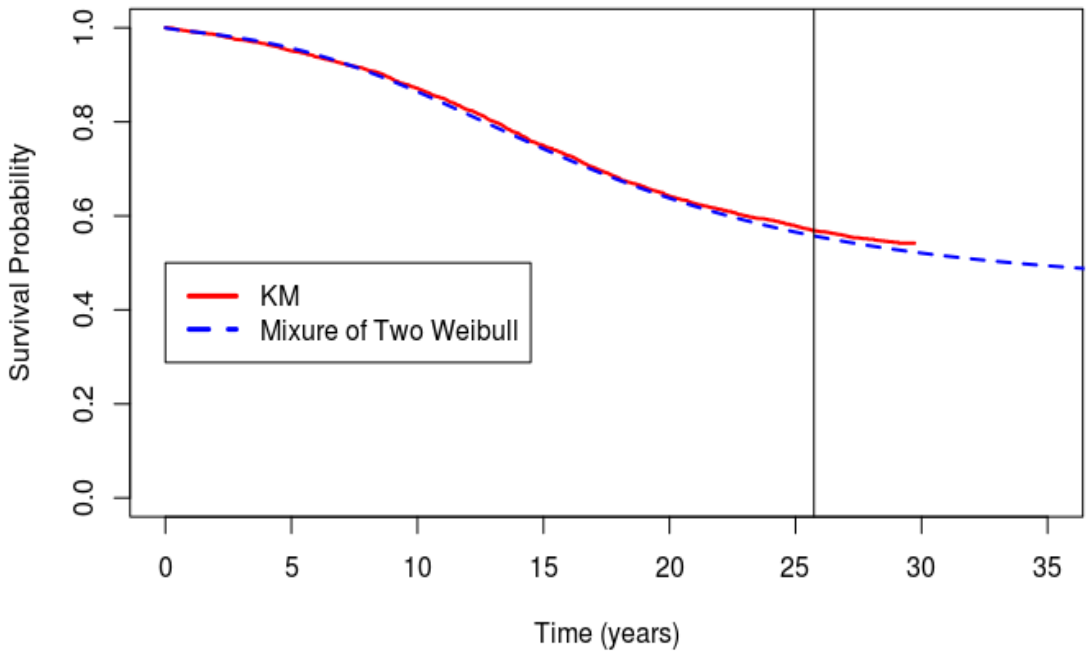

Figure 8-A Extrapolation of Survival Probability 2011-2014 in SHEP – CV-related Mortality with Covariates – All Patients

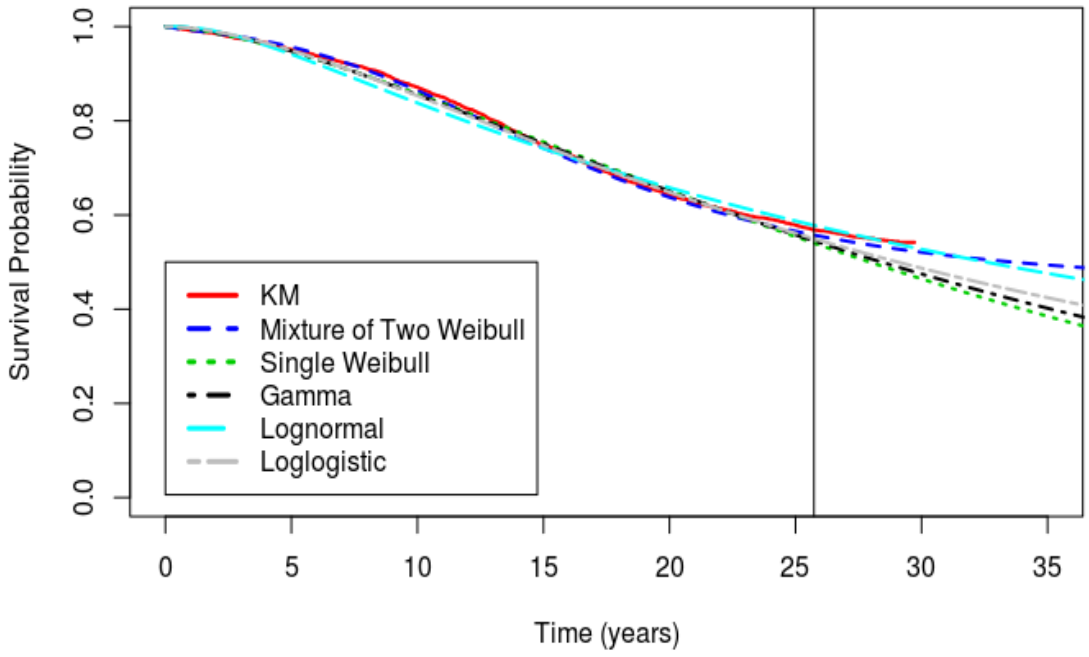

Figure 8-B Extrapolation of Survival Probability 2011-2014 in SHEP – CV-related Mortality with Covariates – All Patients

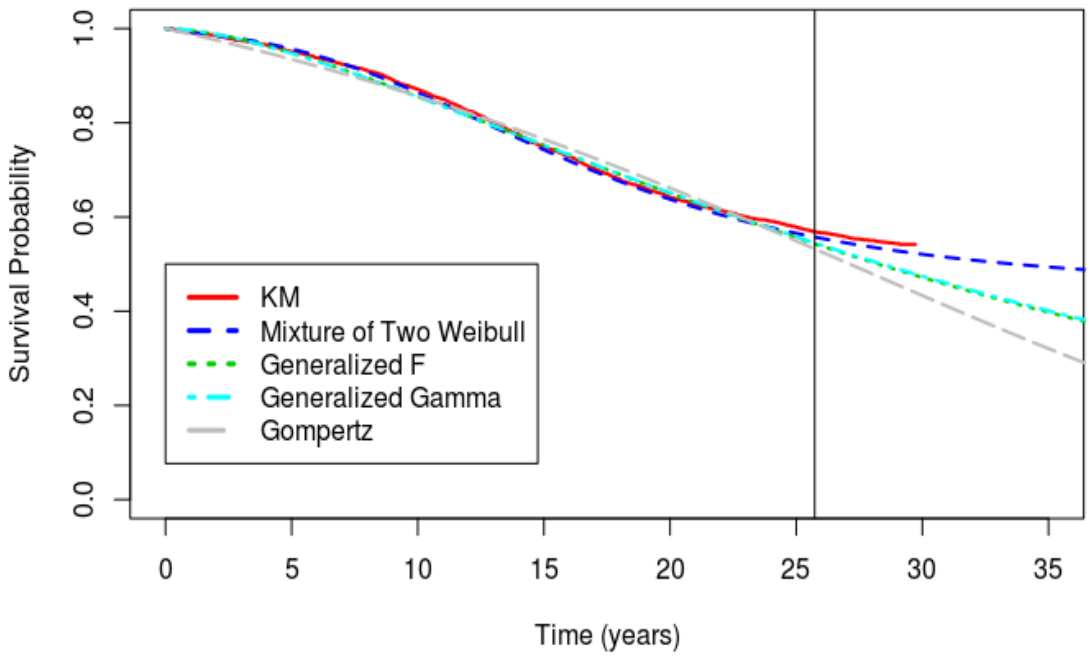

Figure 8-C Extrapolation of Survival Probability 2011-2014 in SHEP – CV-related Mortality with Covariates – All Patients

Table 4 Comparison of Mixture of Two Weibull and Common Parametric Models CV-related Mortality - SHEP Data with Covariates

Models	Log Likelihood	AIC
Mixture of two Weibull	-7151.51	14331.01
Single Weibull	-7236.66	14483.33
Gamma	-7233.91	14477.83
Lognormal	-7294.28	14598.56
Log-logistic	-7221.28	14452.57
Generalized F	-7232.98	14479.96
Generalized Gamma	-7233.69	14479.38
Gompertz	-7271.38	14552.77

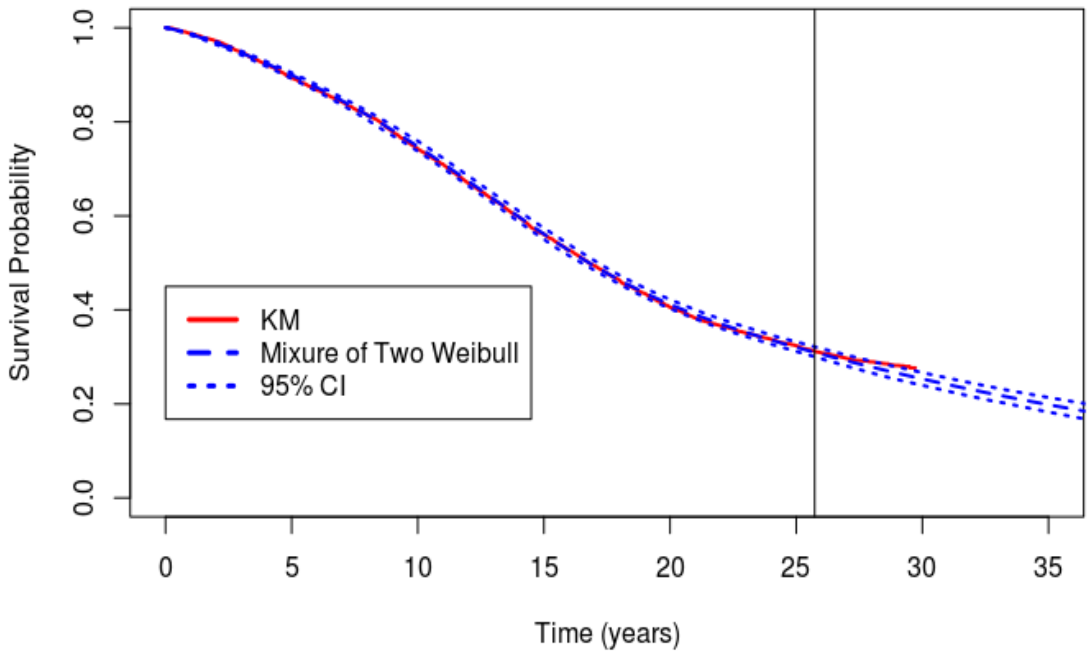

Figure 9-A Extrapolation of Survival Probability 2011-2014 in SHEP – All-cause Mortality Without Covariates – All Patients

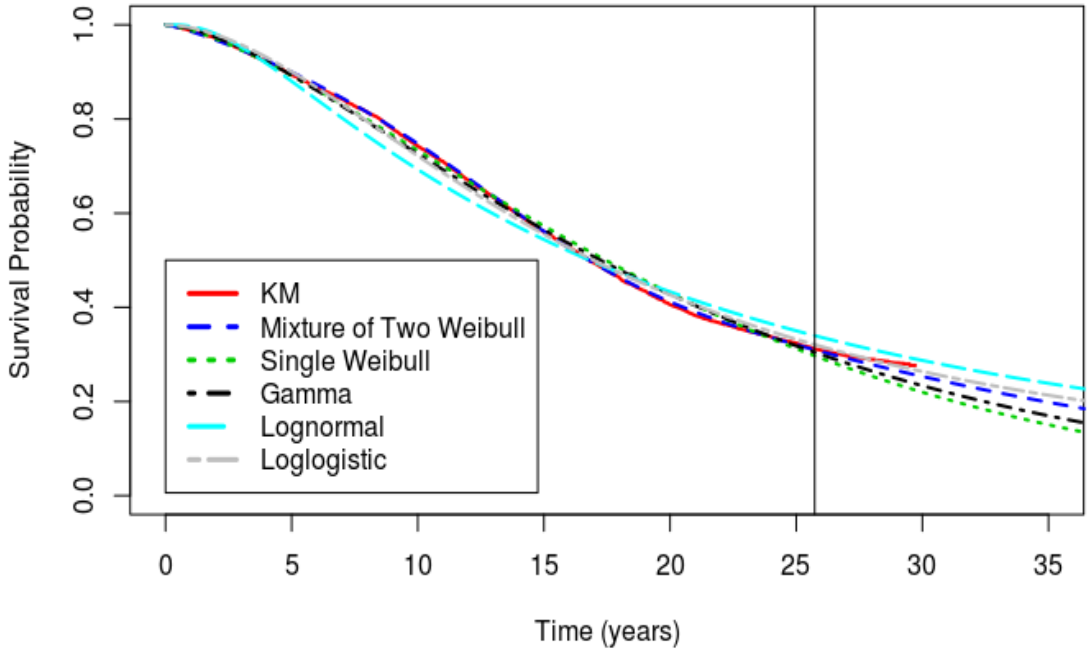

Figure 9-B Extrapolation of Survival Probability 2011-2014 in SHEP – All-cause Mortality Without Covariates – All Patients

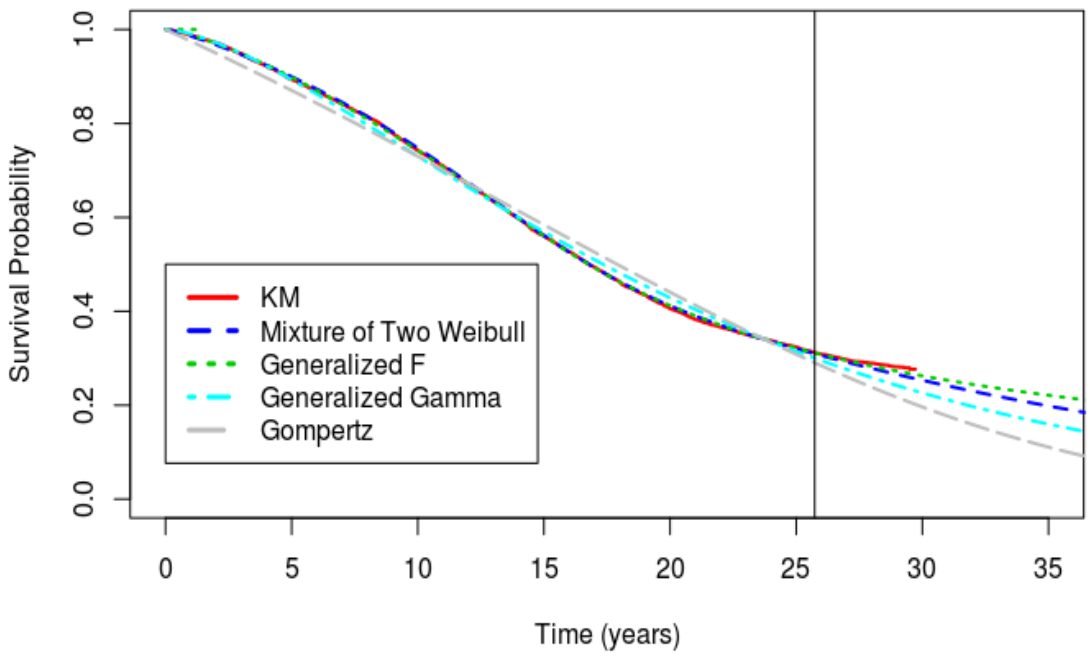

Figure 9-C Extrapolation of Survival Probability 2011-2014 in SHEP – All-cause Mortality Without Covariates – All Patients

Table 5 Comparison of Mixture of Two Weibull and Common Parametric Models All-cause Mortality - SHEP Data Without Covariates

Models	Log Likelihood	AIC	Restricted Mean (95% CI) T=26	Extrapolated Mean (95% CI) Lifetime
KM			16.60	
Mixture of two Weibull	-12959.48	25928.96	16.63 (16.45-16.82)	23.76 (21.71-23.72)
Single Weibull	-13002.90	26009.80	16.67 (16.45-16.90)	20.41 (19.86-20.95)
Gamma	-13002.93	26009.85	16.61 (16.37-16.83)	21.25 (20.67-21.84)
Lognormal	-13135.68	26275.35	16.42 (16.18-16.65)	28.63 (27.32-30.02)
Log-logistic	-13012.29	26028.57	16.62 (16.38-16.84)	30.02 (28.63-31.71)
Generalized F	-12967.31	25942.62	16.62 (16.33-16.86)	
Generalized Gamma	-13001.11	26008.22	16.61 (16.37-16.83)	20.79 (20.19-21.52)
Gompertz	-13070.52	26145.03	16.61 (16.37-16.85)	19.06 (18.70-19.49)

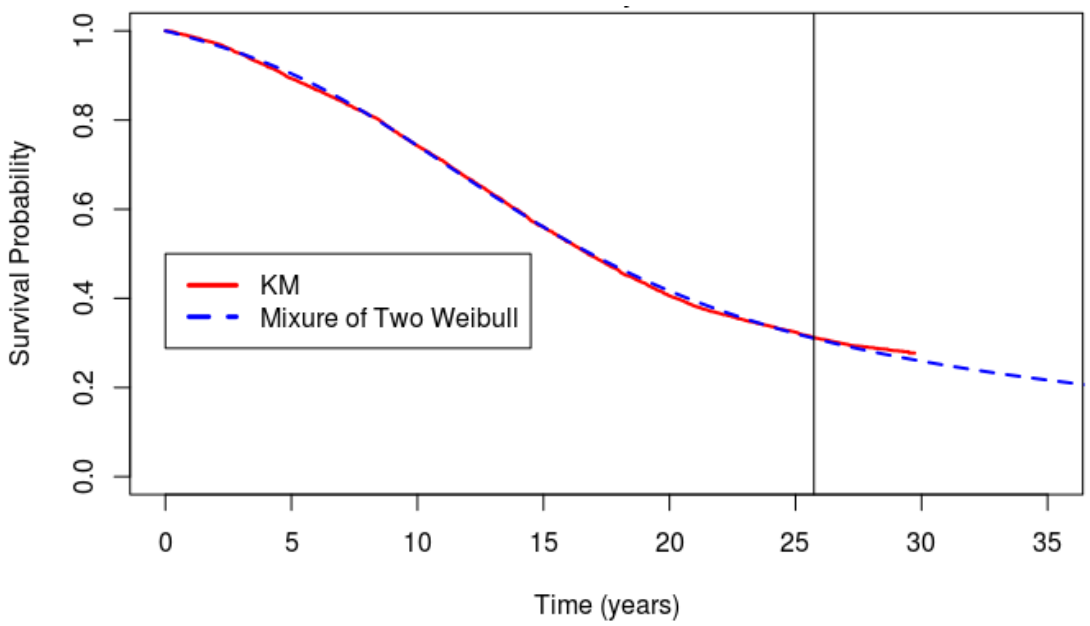

Figure 10-A Extrapolation of Survival Probability 2011-2014 in SHEP – All-cause Mortality with Covariates – All Patients

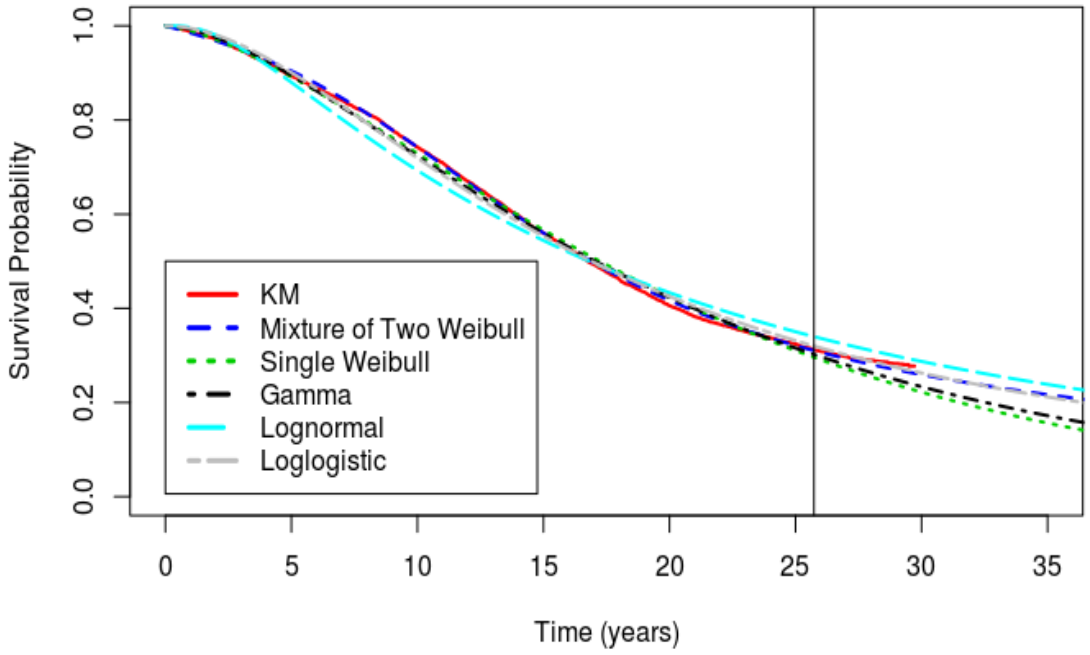

Figure 10-B Extrapolation of Survival Probability 2011-2014 in SHEP – All-cause Mortality with Covariates – All Patients

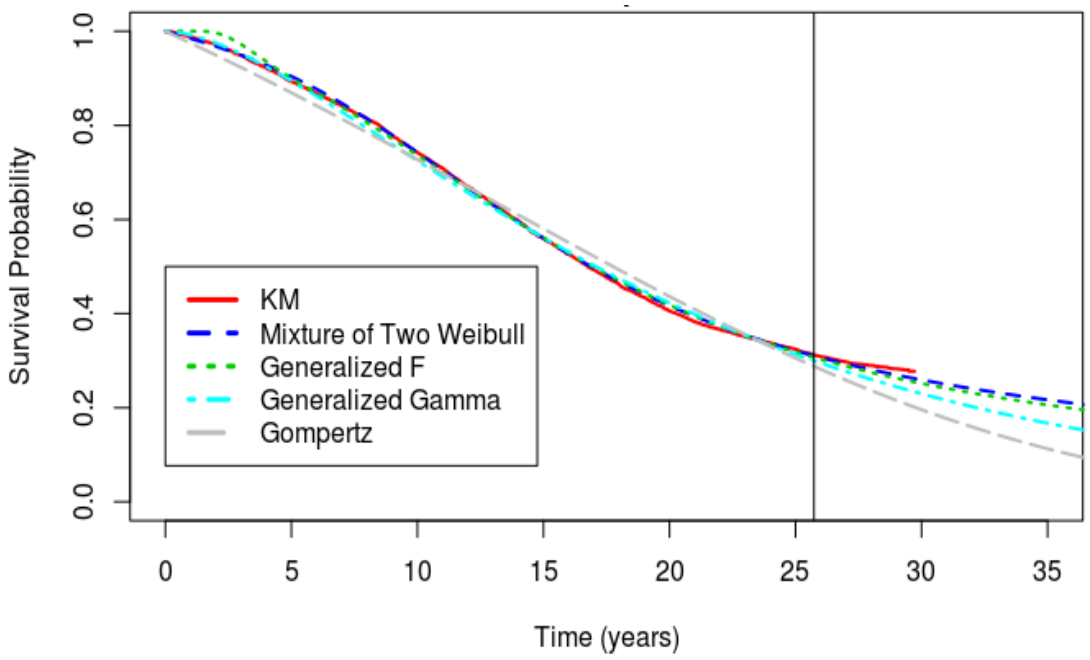

Figure 10-C Extrapolation of Survival Probability 2011-2014 in SHEP – All-cause Mortality with Covariates – All Patients

Table 6 Comparison of Mixture of Two Weibull and Common Parametric Models All-cause Mortality - SHEP Data with Covariates

Models	Log Likelihood	AIC
Mixture of two Weibull	-12548.31	25124.63
Single Weibull	-12667.54	25345.09
Gamma	-12659.77	25329.69
Lognormal	-12796.00	25602.23
Log-logistic	-12642.72	25295.43
Generalized F	-12557.43	25128.87
Generalized Gamma	-12659.77	25331.53
Gompertz	-12739.72	25489.44

4.3 Finite Mixture Models with More Than Two Components

Determining the number of components g in a mixture is an important but challenging task which has yet been completely resolved (McLachlan and Peel 2000). When the mixture model is employed as an alternative method to estimate unknown distributions, the commonly used criteria, such as AIC and BIC, would be adequate for

choosing the number of components g (Ćwik and Koronacki 1997, Biernacki et al. 1998, Solka et al. 1998). Making extrapolations about the survival probability with the mixture model falls into this category. When applying the finite mixture models to the survival data, we could fit a set of models with a varying number of components and select the optimal one based on model fitting diagnostics. In this section we will present a couple situations where we demonstrate when to pick a mixture of 3 Weibull model versus a mixture of 2 Weibull model.

For this purpose, we use the digitized data based on Hodi et al. (2010), also introduced in section 2.2. To determine whether a mixture model with 3 components possesses advantages over a mixture model with 2 components as well as the other single parametric models, we first fit a mixture model of 3 Weibull, a mixture model of 2 Weibull, and all the single models to the digitized progression-free data.

For the IPI arm, Figures 11-A to 11-C clearly show that the estimated survival functions from both mixture models closely mimic the KM estimation. The single models, on the other hand, are far off from the KM curve. Table 7 also shows that both mixture models provide a substantially better fit to the data than the single models. The log likelihood values for the 3-mixture and 2-mixture models are -231.97 and -239.12, respectively. The next best single-model log likelihood value is -270.51 from the generalized F. In terms of AIC, the 3-mixture model has the best value of 479.94, followed by the 2-mixture model (488.24) and generalized F (549.03). Between the two mixture models, the 3-mixture model outperforms the 2-mixture model in both log likelihood and AIC, even though it has 3 more parameters (1 for the weights and 2 for the third Weibull) to estimate. Although both have a restricted mean of 8.52 at $T = 46$, which is very close

to the KM estimate of 8.55, for most of the time the 2-mixture model either under-estimates or over-estimates the survival function. In comparison, the 3-mixture model's estimates are consistently close to the KM curve over time. Therefore, when there is need to extrapolate survival probability beyond the observation period or the mean survival time, results from the 3-mixture model are more reliable. For example, the estimated mean survival time based on the 3-mixture model is 13.6 months, while it is only 9.37 months based on the 2-mixutre model. This is a substantial difference for this type of patients, and the choice of model will have a huge impact on the cost effectiveness profile of the intervention.

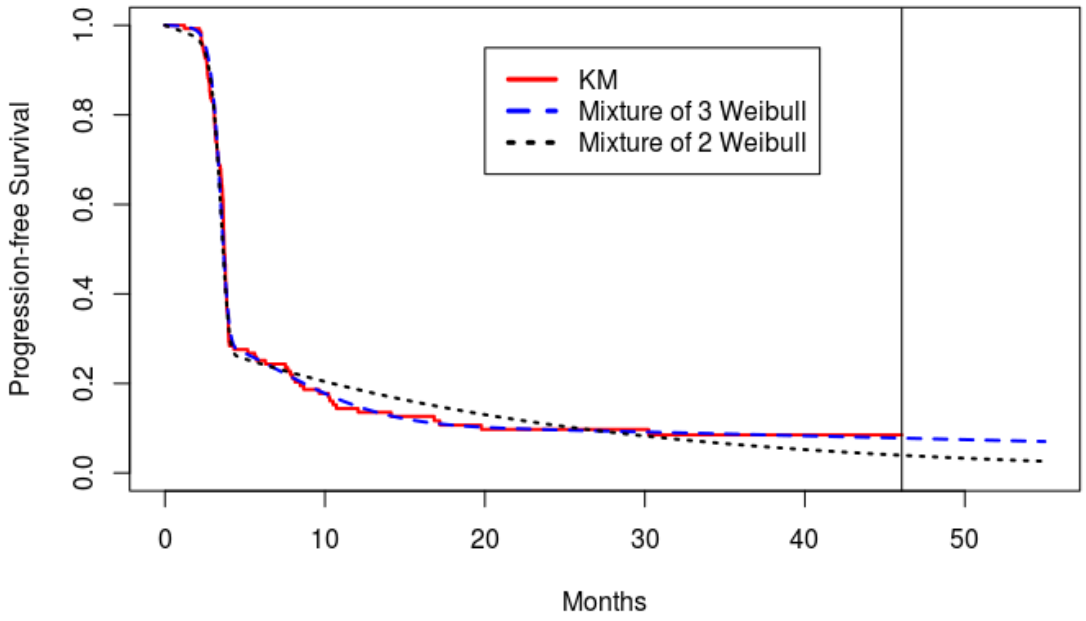

Figure 11-A Progression-free Survival – IPI

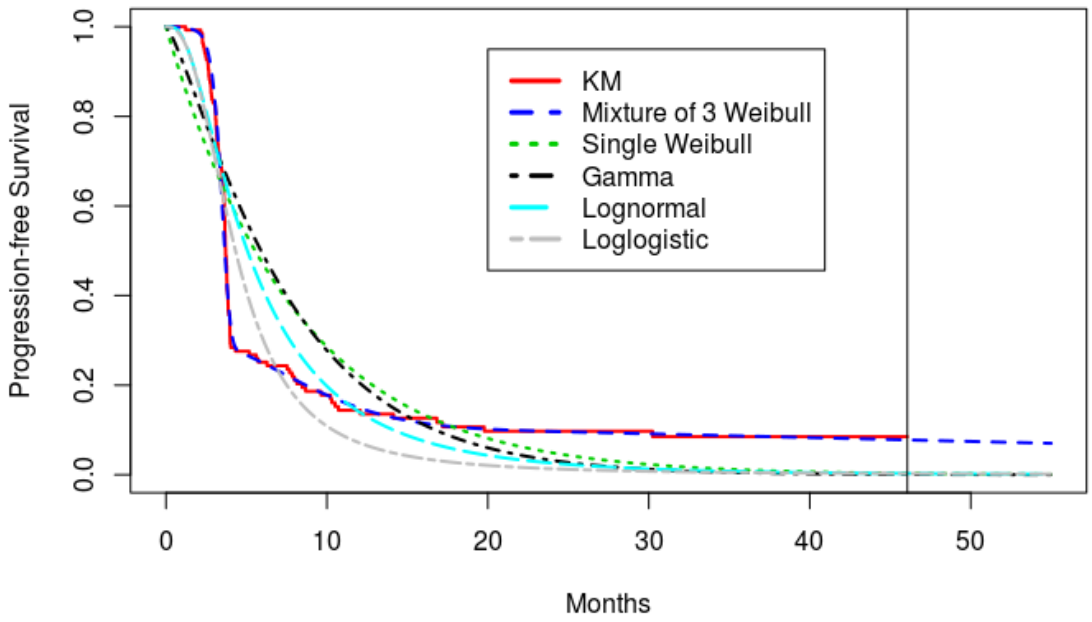

Figure 11-B Progression-free Survival – IPI

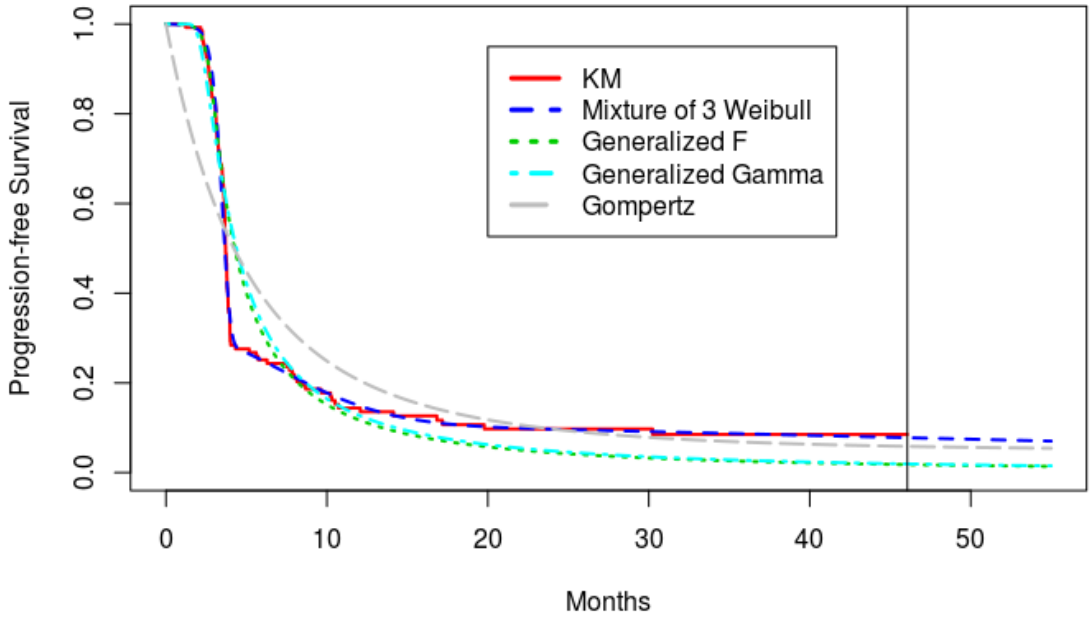

Figure 11-C Progression-free Survival – IPI

Table 7 Progression Free Survival – IPI

Models	Log Likelihood	AIC	Restricted Mean (95% CI) T=46	Extrapolated Mean (95% CI) Lifetime
KM			8.55	
Mixture of 3 Weibull	-231.97	479.94	8.52 (6.5 – 9.97)	13.6 (7.75 – $2.88x10^5$)
Mixture of 2 Weibull	-239.12	488.24	8.52 (6.68 – 9.84)	9.37 (6.93 – 12.49)
Single Weibull	-359.99	723.97	7.95 (6.59 – 9.51)	7.98 (6.69 – 8.33)
Gamma	-357.49	718.97	7.76 (6.68 – 9.02)	7.77 (6.7 – 8.98)
Lognormal	-324.09	652.18	6.94 (5.92 – 8.01)	6.98 (5.95 – 8.33)
Log-logistic	-313.80	631.59	5.62 (4.81 – 6.56)	5.7 (4.88 – 6.73)
Generalized F	-270.51	549.03	6.99 (4.59 – 8.26)	
Generalized Gamma	-284.04	574.08	7.19 (5.93 – 8.82)	9.4 (6.42 – 23)
Gompertz	-347.96	699.93	8.74 (6.68 – 11.1)	

Similar results are observed for the IPI+GP100 arm. As shown in Figures 12-A to 12-C, the two mixture models fit the data decently well, while all single models fail to do so. In Table 8, both log likelihood and the AIC for the two mixture models are much better than those for the single models. Between the two mixture models, the 3-mixture model has better log likelihood and AIC. It also has higher restricted mean and mean survival time than the 2-mixture model.

When these models are compared for the GP100 arm, although the mixture models maintain their advantages over the single models, as shown in Figures 13-A to 13-C and Table 9, the difference between themselves become minimal. The log likelihood, AIC, restricted mean, and mean survival time are all similar between them. Therefore, for the GP100 arm, the 2-mixture model should be adopted since it can perform as well as the 3-mixture model but has less parameters to estimate.

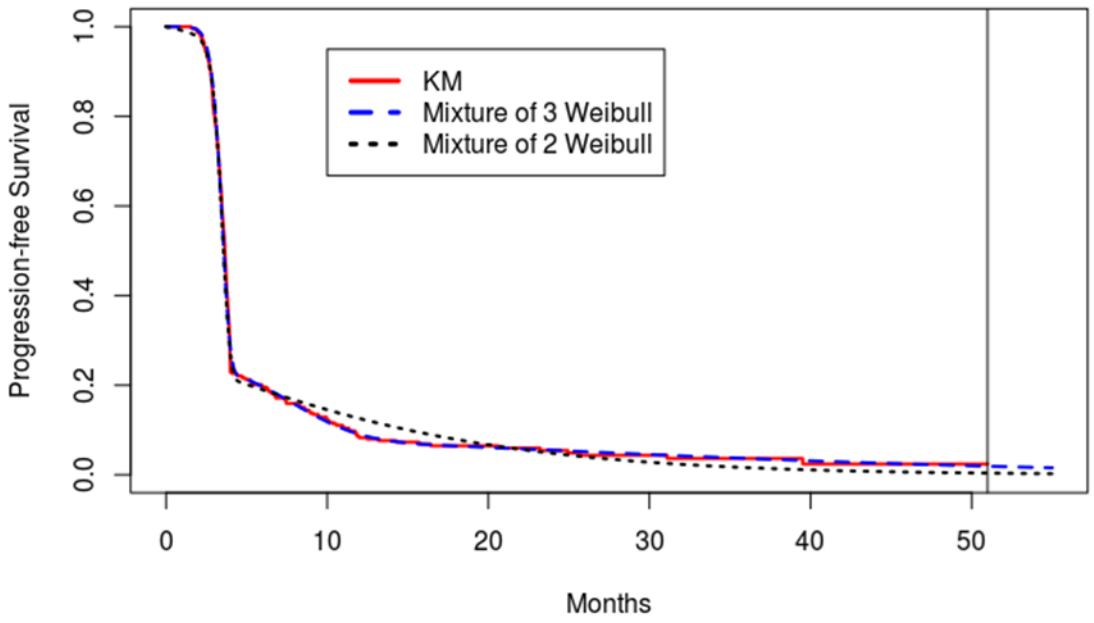

Figure 12-A Progression-free Survival – IPI+GP100

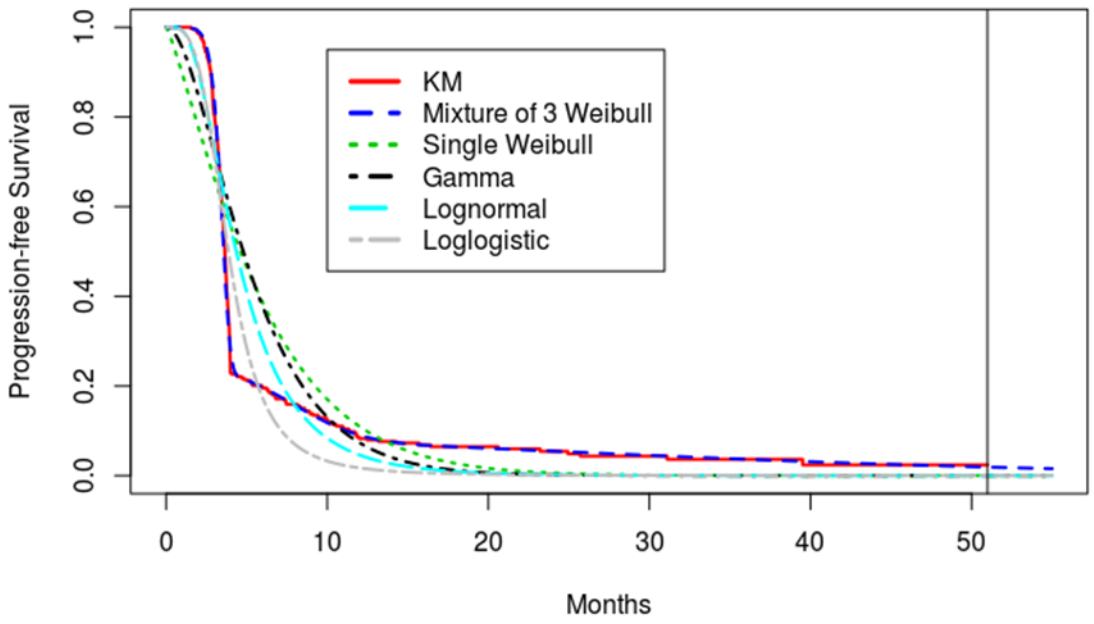

Figure 12-B Progression-free Survival – IPI+GP100

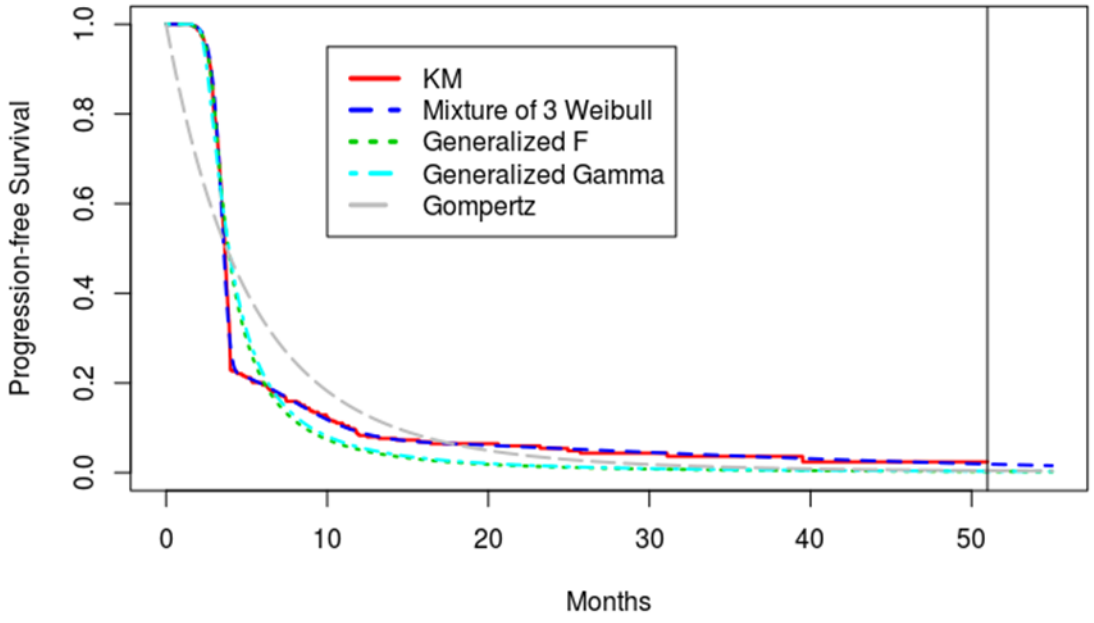

Figure 12-C Progression-free Survival – IPI+GP100

Table 8 Progression Free Survival – IPI+GP100

Models	Log Likelihood	AIC	Restricted Mean (95% CI) T=51	Extrapolated Mean (95% CI) Lifetime
KM			6.47	
Mixture of 3 Weibull	-615.95	1247.90	6.51 (5.55 – 7.61)	6.84 (5.7 – 8.16)
Mixture of 2 Weibull	-631.04	1272.07	6.26 (5.4 – 6.76)	6.29 (5.4 – 6.89)
Single Weibull	-988.70	1981.40	5.86 (5.36 – 6.37)	5.86 (5.37 – 6.37)
Gamma	-955.92	1915.84	5.67 (5.25 – 6.09)	5.67 (5.30 – 6.10)
Lognormal	-861.76	1727.51	5.22 (4.89 – 5.57)	5.22 (4.88 – 5.60)
Log-logistic	-816.97	1637.93	4.41 (4.18 – 4.69)	4.42 (4.18 – 4.66)
Generalized F	-707.59	1423.18	5.25 (4.64 – 5.72)	5.39 (4.68 – 6.01)
Generalized Gamma	-736.31	1478.63	5.35 (4.91 – 5.91)	5.52 (4.97 – 6.28)
Gompertz	-996.38	1996.75	6.12 (5.70 – 7.00)	

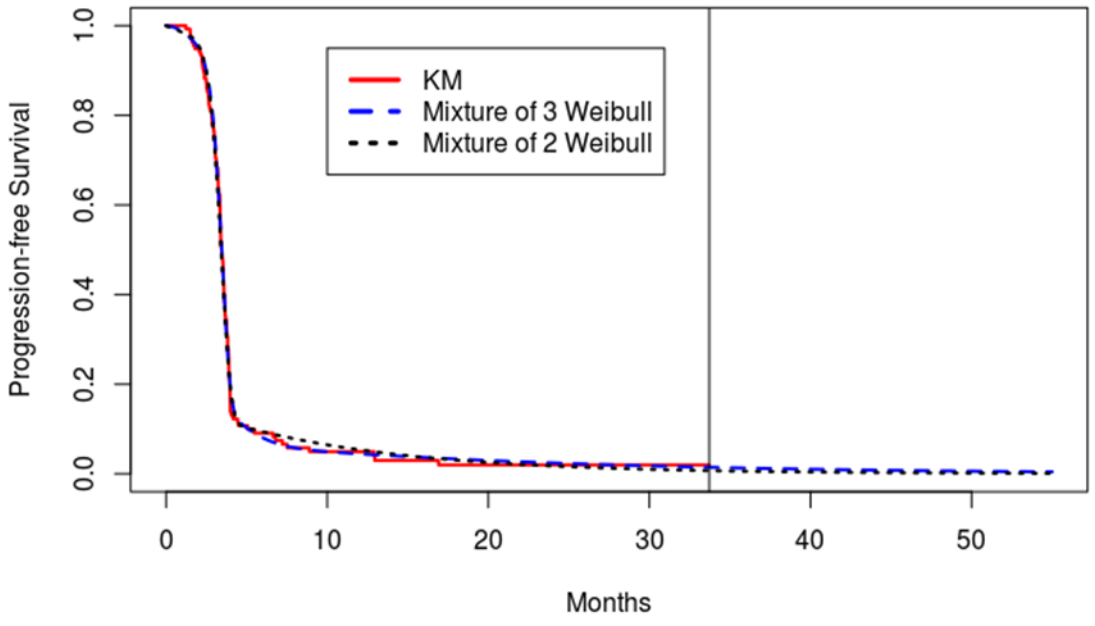

Figure 13-A Progression-free Survival – GP100

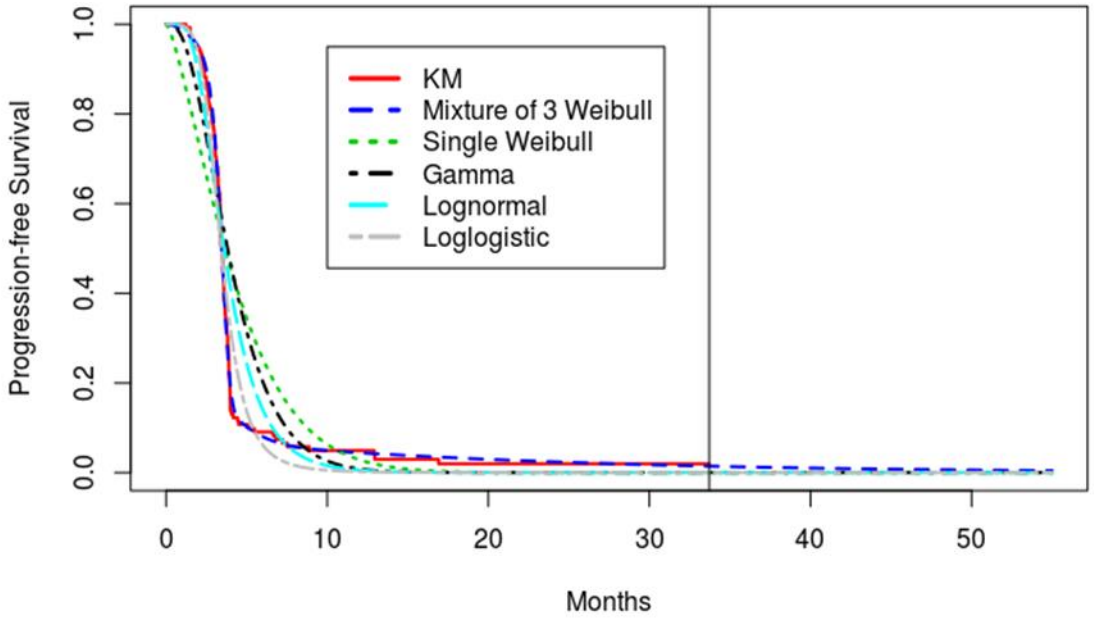

Figure 13-B Progression-free Survival – GP100

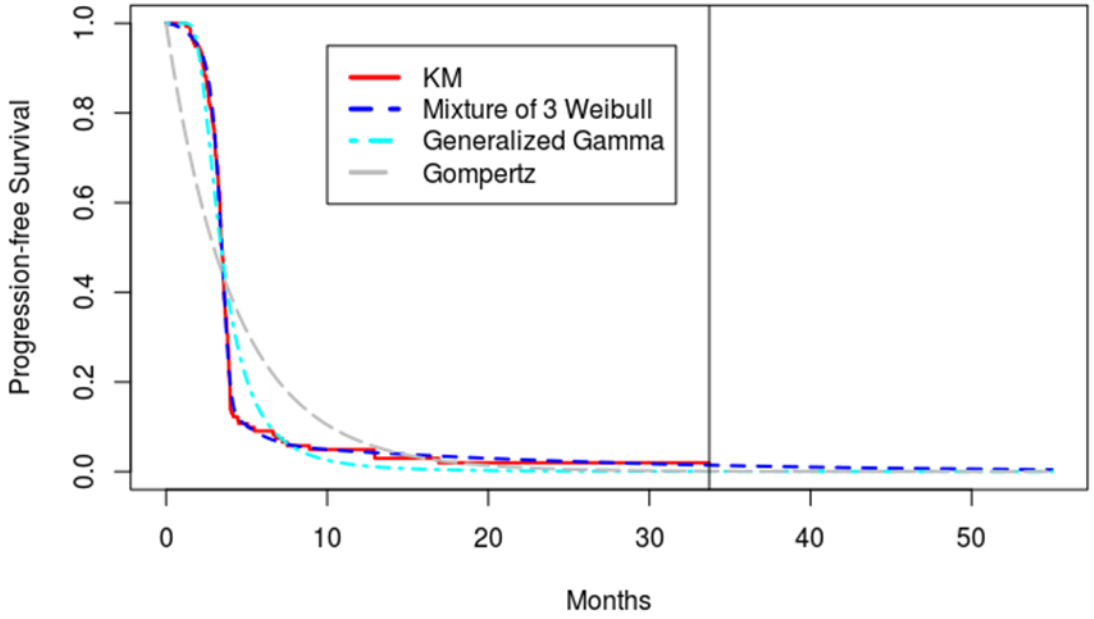

Figure 13-C Progression-free Survival – GP100

Table 9 Progression Free Survival – GP100

Models	Log Likelihood	AIC	Restricted Mean (95% CI) T=34	Extrapolated Mean (95% CI) Lifetime
KM			4.38	
Mixture of 3 Weibull	-190.51	397.02	4.45 (3.75 – 5.03)	4.71 (3.82 – $3.21x10^5$)
Mixture of 2 Weibull	-193.60	397.19	4.45 (3.79 – 5.3)	4.53 (3.84 – 6.23)
Single Weibull	-306.07	616.15	4.36 (3.83 – 4.96)	4.36 (3.84 – 4.96)
Gamma	-284.72	573.45	4.25 (3.86 – 4.69)	4.25 (3.85 – 4.66)
Lognormal	-253.07	510.13	4.02 (3.68 – 4.41)	4.02 (3.70 – 4.38)
Log-logistic	-229.47	462.94	3.68 (3.46 – 3.9)	3.68 (3.46 – 3.92)
Generalized F	Cannot be estimated.			
Generalized Gamma	-233.51	473.02	4.04 (3.66 – 4.52)	4.04 (3.68 – 4.52)
Gompertz	-319.23	642.47	4.4 (3.68 – 5.22)	

When it comes to overall survival, the 3-mixture model starts to lose its edge over the other models. For all three treatment arms, while the 3-mixture model still has the highest log likelihood, the best AIC values are taken by other models. Although by checking the plots of the estimated survival functions we observe that the two mixture of Weibull models still fit the data better than the single models, the log likelihood, AIC, and estimated restricted mean survival time among them are indistinguishable. Similar patterns are also present between the two mixture models, the 3-mixture model is not at a clear advantage any more. Based on these observations, the 2-mixture model should be selected for the estimation and extrapolations with regard to the overall survival for all three treatment arms.

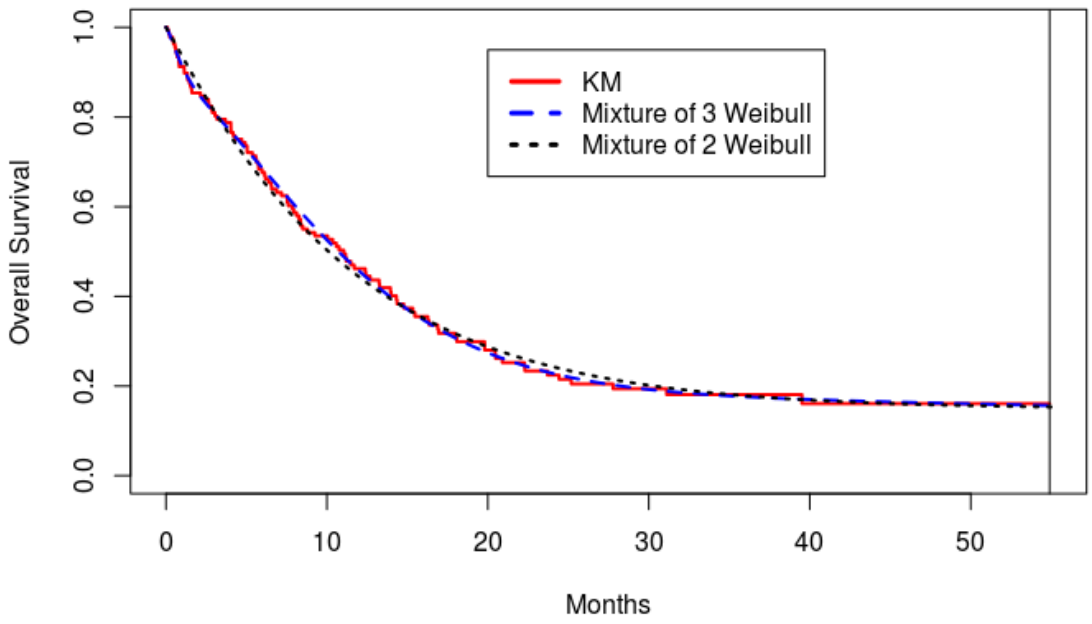

Figure 14-A Overall Survival – IPI

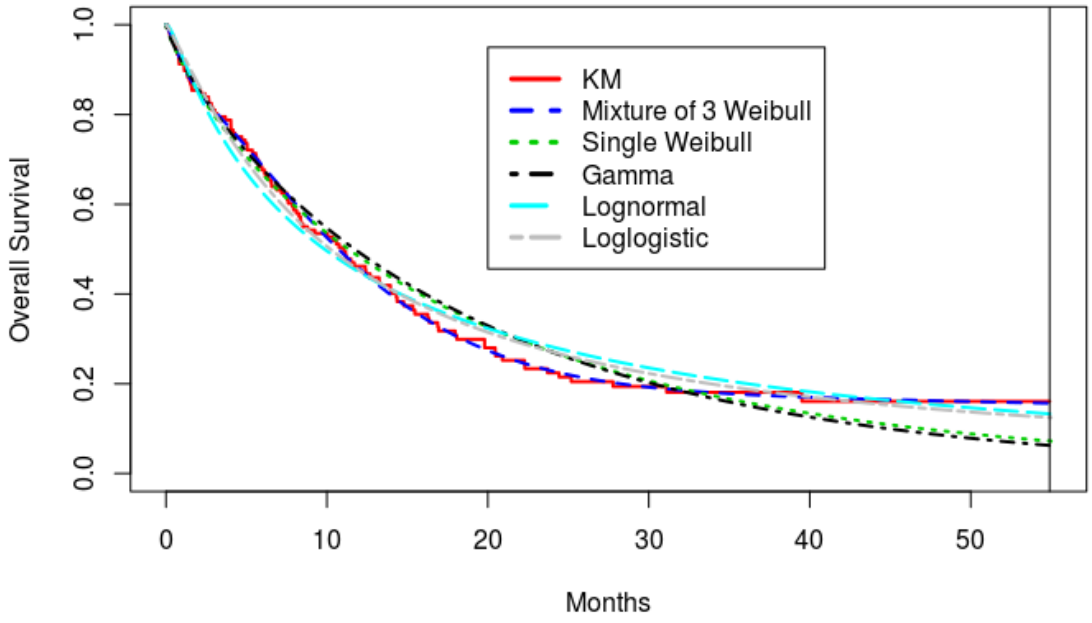

Figure 14-B Overall Survival – IPI

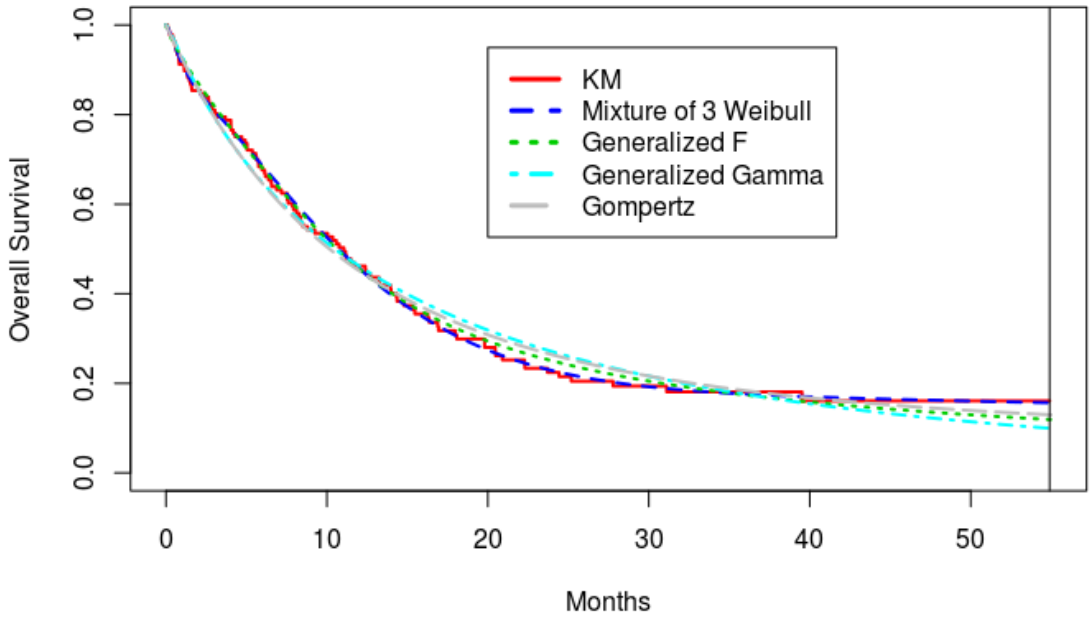

Figure 14-C Overall Survival – IPI

Table 10 Overall Survival – IPI

Models	Log Likelihood	AIC	Restricted Mean (95% CI) T=55	Extrapolated Mean (95% CI) Lifetime
KM			17.6	
Mixture of 3 Weibull	-390.20	796.38	17.6 (14.3 – 20.4)	49.9 (16.4- $6.06x10^7$)
Mixture of 2 Weibull	-392.5	794.44	17.6 (14.1 – 19.8)	42.2 (20.8 – $8.84x10^7$)
Single Weibull	-396.67	797.35	17.2 (14.2 – 20.4)	19.1 (15.1 – 24.3)
Gamma	-397.41	798.82	17.2 (14.2 – 20.1)	18.5 (14.6 – 23.1)
Lognormal	-396.38	796.77	18.1 (14.8 – 21.5)	32.5 (20.6 – 51.2)
Log-logistic	-394.76	793.52	17.9 (14.9 – 21.2)	66.6 (31.2 – 556)
Generalized F	-393.22	794.44	17.5 (14.7 – 23.8)	
Generalized Gamma	-395.12	796.25	17.5 (14.4 – 21.0)	22.3 (16.4 – 23.1)
Gompertz	-393.85	791.70	17.7 (14.7 – 20.9)	

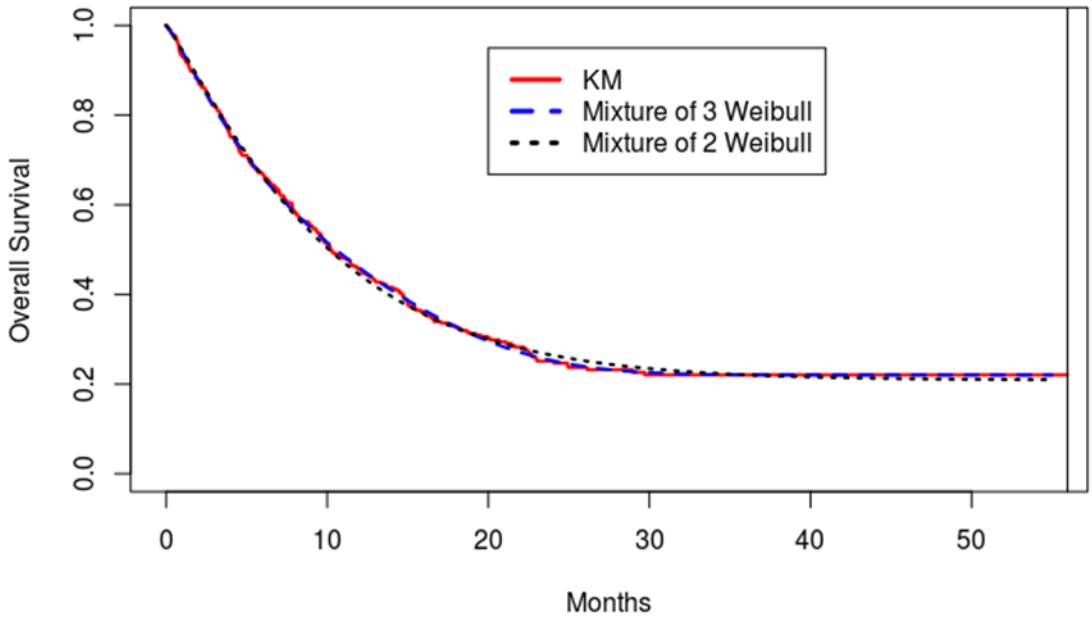

Figure 15-A Overall Survival – IPI+GP100

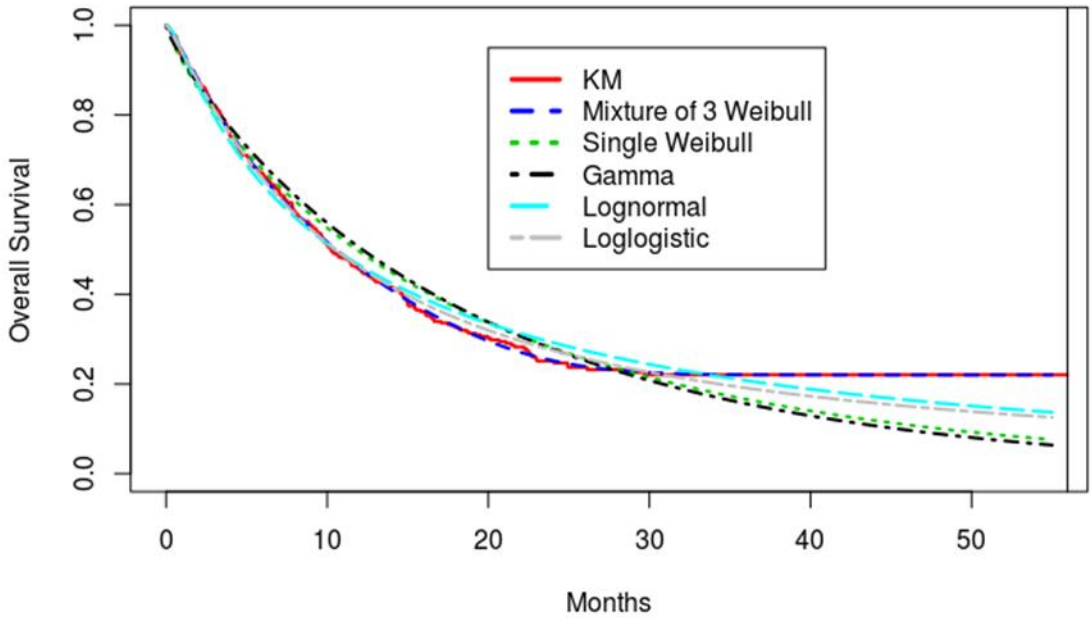

Figure 15-B Overall Survival – IPI+GP100

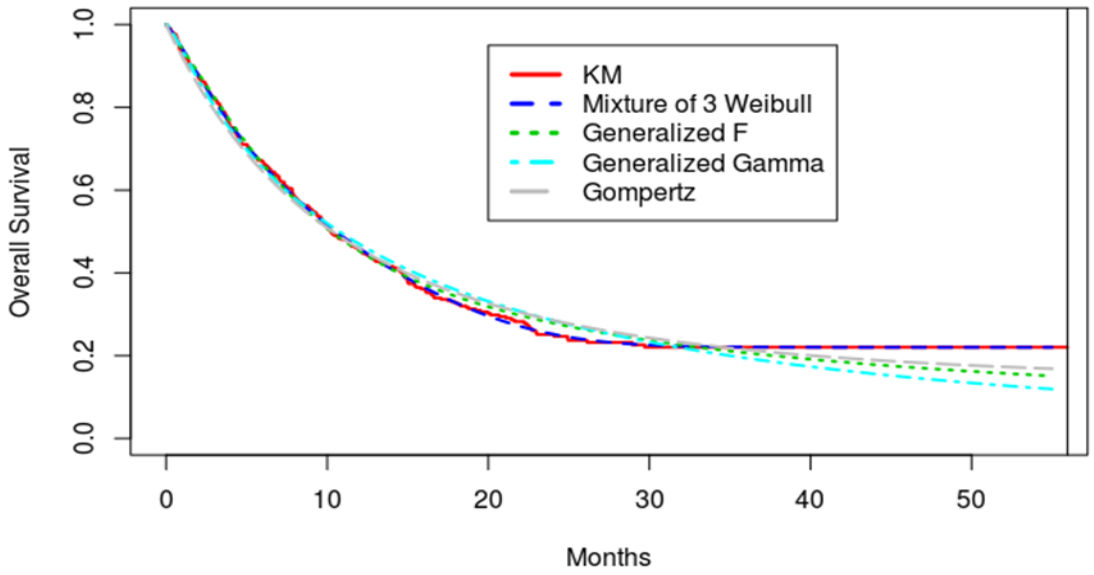

Figure 15-C Overall Survival – IPI+GP100

Table 11 Overall Survival – IPI+GP100

Models	Log Likelihood	AIC	Restricted Mean (95% CI) T=56	Extrapolated Mean (95% CI) Lifetime
KM			19.40	
Mixture of 3 Weibull	-1059.71	2135.42	19.44 (17.4 – 21.1)	16107666 ($4.17x10^5$ – $5.81x10^8$)
Mixture of 2 Weibull	-1062.17	2134.35	17.61 (17.4 – 20.9)	11799751 ($2.52x10^5$ – $7.14x10^8$
Single Weibull	-1080.59	2165.18	17.7 (15.9 – 19.6)	19.7 (16.8 – 23.1)
Gamma	-1082.96	2169.91	17.6 (15.9 – 19.6)	18.9 (16.6 – 21.7)
Lognormal	-1072.30	2148.60	18.8 (16.7 – 20.9)	32.9 (24.9 – 42.3)
Log-logistic	-1068.99	2141.99	18.3 (16.5 – 20.3)	63.5 (37.1 – 184)
Generalized F	-1067.58	2143.15	18.8 (16.8 – 21.0)	
Generalized Gamma	-1071.43	2148.87	18.4 (16.5 – 20.5)	26.5 (20.1 – 44.2)
Gompertz	-1068.35	2140.70	19.0 (17.0 – 21.2)	

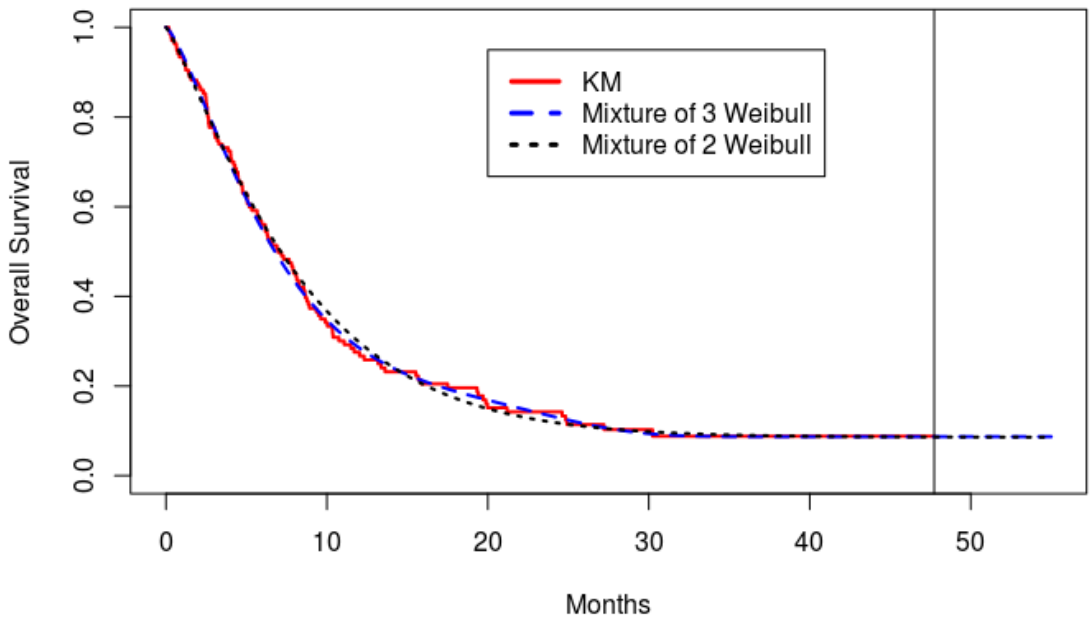

Figure 16-A Overall Survival – GP100

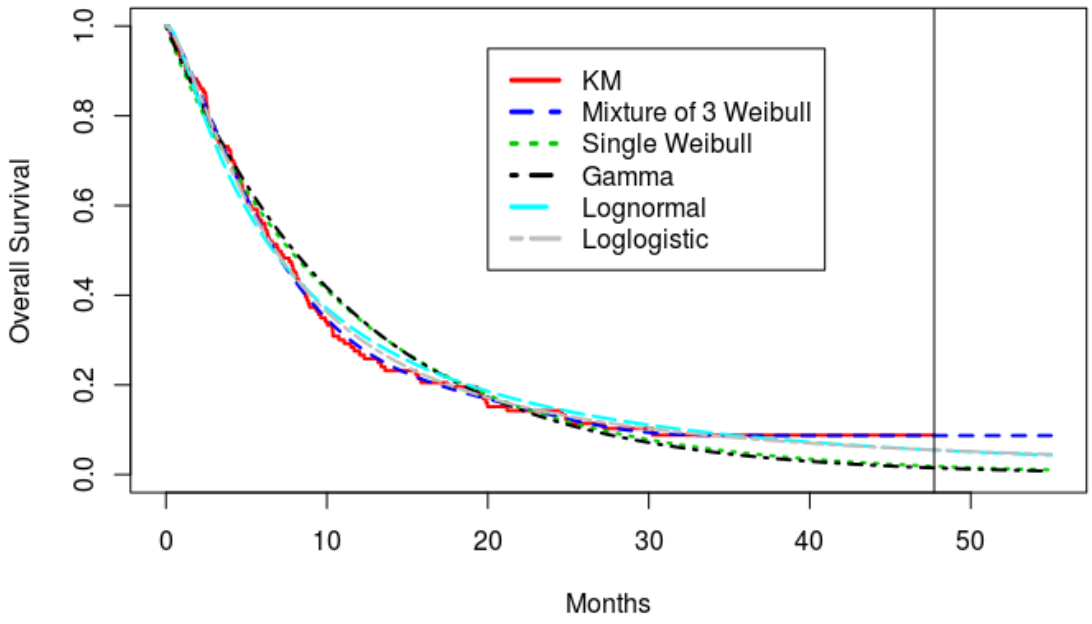

Figure 16-B Overall Survival – GP100

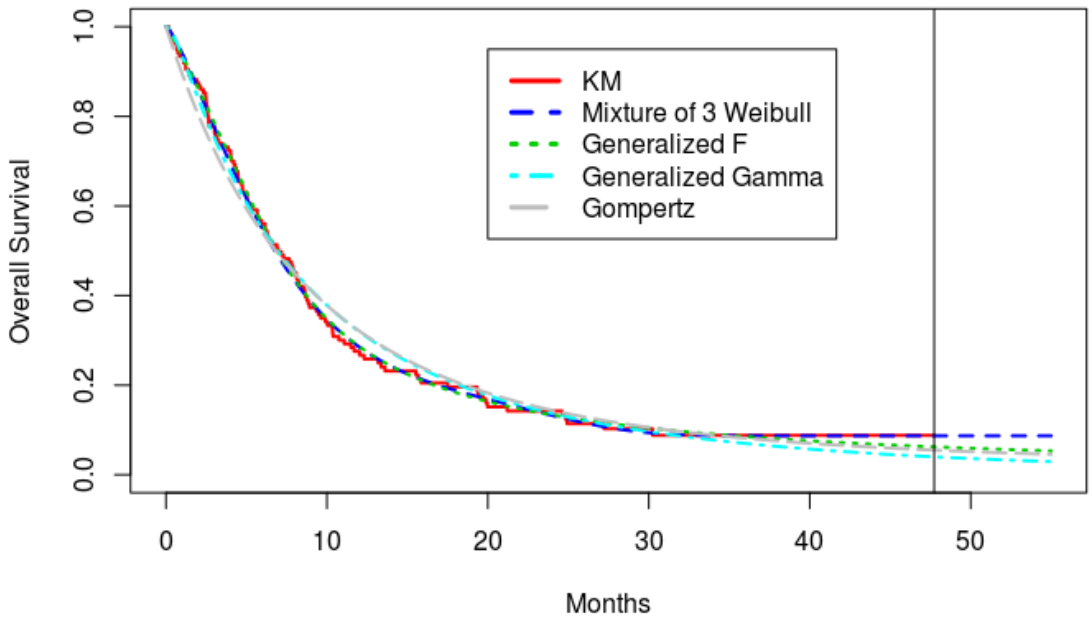

Figure 16-C Overall Survival – GP100

Table 12 Overall Survival – GP100

Models	Log Likelihood	AIC	Restricted Mean (95% CI) T=48	Extrapolated Mean (95% CI) Lifetime
KM			11.6	
Mixture of 3 Weibull	-383.17	782.35	11.59 (9.32 – 13.23)	370920 (13.2 – 4.51x10^6)
Mixture of 2 Weibull	-384.78	779.56	11.61 (9.88 – 13.54)	9112 (10.8 – 2.05x10^6)
Single Weibull	-391.44	786.88	11.3 (9.35 – 13.3)	11.5 (9.49 – 13.9)
Gamma	-391.60	787.20	11.2 (9.39 – 13.4)	11.4 (9.58 – 13.6)
Lognormal	-387.85	779.70	11.8 (9.54 – 14.0)	14.2 (10.8 – 19.9)
Log-logistic	-385.68	775.36	11.6 (9.53 – 13.7)	17.5 (12.1 – 29.2)
Generalized F	-384.69	777.39	11.5 (9.81 – 15.8)	
Generalized Gamma	-387.18	780.36	11.5 (9.56 – 13.7)	12.6 (9.85 – 19.0)
Gompertz	-388.42	780.84	11.6 (9.42 – 14.0)	

One phenomenon we notice in Tables 11 and 12 is that for the IPI+GP100 and GP100 arms, the extrapolated mean survival time is unrealistically large from both the 3-mixtue and the 2-mixture models. For the IPI+GP100 arm, the expected survival time is 16,107,666 and 11,799,751 months based on the 3-mixture and 2-mixture model, respectively. For the GPI arm, the 3-mixutre model estimates an expected survival time of 370,920 months, and the 2-mixture model comes up with an estimate of 9,112 months. These values are clearly impossible, despite these two models have the best model fitting values. Upon closer examination of the survival function plots, we realize that the KM curves do not approach 0 during the observation period. Instead, they reach a value above 0 at a certain time point and stop decreasing onward. For example, for the IPI+GP100 arm, as can be seen in Figure 15-A, the KM curve reaches and stays above 0.2 at approximately 30 months. This indicates that a portion of the patients will never experience the event of interest, or death in our case. In other words, these patients are considered "cured". While all patients will die eventually, this "cured" feature might be caused by lost to follow-up to these patients and as a result their death time is never known to the investigator. Similar concerns about this "cured" portion of patients due to lost to follow-up also exist for SHEP. If these patients are included in the extrapolation of the mean survival time, they will appear always alive and therefore will artificially inflate the estimation.

Survival models with a cure portion can be viewed as a special case of the finite mixture model. In the next chapter we will discuss in detail their estimations and applications.

Chapter 5: The Mixture Cure Models

5.1 The Finite Mixture Models and the Mixture Cure Models

In classical survival analysis, one assumption is that all patients will eventually experience the event of interest. However, there are situations where the subject may never experience it. For example, in a clinical trial where the medical intervention is efficacious against certain disease, some treated patients will not suffer a replace of the disease. These patients are then considered cured. If the endpoint of the trial is the time until recurrence of the disease, these patients will have infinite survival time. In economic studies where the interest is the time from unemployment to the next employment, some patients may never find a job and therefore will have infinite unemployment time. One common feature in the above cases is that if T is the time to the event of interest and $S(t) = P(T > t)$ is the survival function, then $\lim_{t\to\infty} S(t) > 0$. This positive, non-zero limiting value corresponds to the portion of patients who will never experience the event. This portion is often called the cure rate. And statistical models that take this feature into account are commonly referred to as the cure models (Amico 2018), although they bear different names in other fields, such as the split population models in economics (Schmidt and Witte 1989).

When there is a mixture of two groups in a population, one cured and one not cured, the survival function of the population $S_p(t)$ consists of two major parts, the probability of being uncured (often called the "incidence model") and the conditional survival function of the uncured (often called the "latency model") (Klein et al. 2016). Basically, $S_p(t)$ can be written as

$$S_p(t) = 1 - \pi + \pi S_u(t) \tag{5.1}$$

where π denotes the probability of not cured (the incidence model), and $S_u(t)$ is the survival probability of the uncured group (the latency model). Both π and $S_u(t)$ can also be modeled as functions of covariates. If $\boldsymbol{X}$ and $\boldsymbol{Z}$ are the set of covariates that π and $S_u(t)$ depend on, respectively, then (5.1) becomes

$$S_p(t|\boldsymbol{x},\boldsymbol{z}) = 1 - \pi(\mathbf{x}) + \pi(\boldsymbol{x})S_u(t|\boldsymbol{z}). \tag{5.2}$$

The logistic model is often assumed for $\pi(\mathbf{x})$, the effect of $\boldsymbol{X}$ on π. For the survival function of the uncured group $S_u(t)$, different assumptions have led to parametric, semi-parametric, and non-parametric mixture models (Amico 2018).

As discussed in Chapter 3, the classical survival analysis models are broadly classified as parametric and non-parametric. They can also be categorized as the proportional hazard models or the accelerated failure time (AFT) models. The Cox proportional hazard model is the most popular among the non-parametric models. While most parametric models fall into the AFT category, some of them, such as the Weibull, can be both proportional hazard and AFT. When these classical survival analysis models are extended to accommodate the cure portion, they form the proportional hazard mixture cure models (Anthony and Chen-Hsin 1992, Peng and Dear 2000, Sy and Taylor 2000, Corbière et al. 2009), the semi-parametric AFT mixture cure models (Li and Taylor 2002, Zhang and Peng 2007, Lu 2010) , and the full parametric mixture cure models (Boag 1949, Berkson and Gage 1952, Farewell 1977, 1982, Yamaguchi 1992, Ghitany et al. 1994, Peng et al. 1998). Among these, the full parametric mixture cure models are capable of making extrapolations of survival probability for the uncured at a future time. They can be treated as a special case of the more general mixture models discussed in Chapter 3.

The full parametric mixture models are comprised of two components: $1-\pi$, the probability of being cured, and $f_u(t)$, the density function of the survival time for the uncured. Since the survival function of the cured group is $S_c(t) = P(T > t) \equiv 1$, it follows a degenerate distribution. Therefore, when $g = 2$ in (3.1), we can get (5.1) by taking 1 minus the integration of each component. In the current literature, $S_u(t)$ is assumed to follow a single parametric model, such as the exponential, Weibull, generalized gamma, or generalized F. We can further extend this part of the mixture by assuming that $S_u(t)$ also follows a mixture of distributions. Thus the resulting parametric mixture models contain $g+1$components, one of them has a degenerate distribution corresponding to the cured portion, the remaining g components each represent an individual distribution.

The likelihood function for the finite mixture models with right censored data is given in (3.35) under section 3.4.3. It is derived based on the fact that due to right censoring, we don't observe T, the time to the event of interest. Instead, we observe $Y = \min(T, C)$ and $\Delta = I(T \leq C)$, where C is a random censoring variable and $I(.)$ is an indicator function. When $\Delta = 1$, we observe $f(t)$, otherwise, we observe $S(t)$. In a mixture model with a cured portion, the observed value of Δ gives different information about π_u. In a sample of observed pairs $(y_j, \delta_j), j = 1, \dots, n$, when $\delta_j = 1$, the j[th] subject is clearly not cured. When $\delta_j = 0$, the j[th] subject may or may not be cured. As a result, when data are right censored, the likelihood function of a mixture cure model can be written as

$$\prod_{j=1}^{n}[\sum_{i=1}^{g}\pi_i f_i(y_j)]^{\delta_j}\,[(1-\pi_u)+\sum_{i=1}^{g}\pi_i S_i(y_j)]^{1-\delta_j}. \qquad (5.3)$$

Following the procedures discussed in section 3.4, the EM algorithm can be applied to find the MLEs of the parameters $\boldsymbol{\Psi}$. The complete-data likelihood function of a mixture cure model with $g+1$ components can be written as

$$L_c(\boldsymbol{\Psi}|\boldsymbol{x}) = f_c(\boldsymbol{x}|\boldsymbol{\Psi}) = \prod_{j=1}^{n}(1-\pi_u)^{z_{cj}} \prod_{i=1}^{g}\left[\pi_i f_i(y_j)\right]^{z_{ij}}, \tag{5.4}$$

where π_u is the probability of being uncured, and Z_{cj} is an indicator variable which takes the value of 1 if the j[th] patient is cured, and 0 otherwise. And the loglikelihood function is

$$l_c(\boldsymbol{\Psi}|\boldsymbol{x}) = \sum_{j=1}^{n}\{z_{cj}\log(1-\pi_u) + \sum_{i=1}^{g} z_{ij}\{\log(\pi_i) + \log[f_i(y_j|\boldsymbol{\theta}_i)]\}\}, \tag{5.5}$$

where $\boldsymbol{\Psi} = (\pi_u, \pi_1, \dots, \pi_{g-1}, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_g)^T$ is the vector containing all the unknown parameters. In order to accommodate the existence of a cure portion, two modifications, one at each of the two steps within an iteration, need to be made to the EM algorithm.

At the E-step, the expected value of Z_{cj} and Z_{ij} need to be calculated differently. This is because only censored observations have a chance of being cured. By combining (3.23) and (5.3), we get

$$E_{\boldsymbol{\Psi}^{(k)}}(Z_{cj}|\boldsymbol{y}) = (1-\delta_j)\frac{1-\pi_u^{(k)}}{(1-\pi_u^{(k)})+\sum_{h=1}^{g}\pi_h^{(k)}S_h(y_j|\boldsymbol{\Psi}^{(k)})}, \tag{5.6}$$

and

$$E_{\boldsymbol{\Psi}^{(k)}}(Z_{ij}|\boldsymbol{y}) = \delta_i\frac{\pi_i^{(k)}f_i(y_j|\boldsymbol{\Psi}^{(k)})}{\sum_{h=1}^{g}\pi_h^{(k)}f_h(y_h|\boldsymbol{\Psi}^{(k)})} + (1-\delta_j)\frac{\pi_i^{(k)}S_i(y_j|\boldsymbol{\Psi}^{(k)})}{(1-\pi_u^{(k)})+\sum_{h=1}^{g}\pi_h^{(k)}S_h(y_j|\boldsymbol{\Psi}^{(k)})}. \tag{5.7}$$

At the M-step, when the mixture model does not contain a cure portion, no distinction needs to be made among the $\pi_i's$, because each π_i corresponds to an individual distribution and the mixture is not affected by the order of the individual component. Also,

there is no prior knowledge about the values of the $\pi_i's$. This is not the case when there is a cure portion in the mixture. From the tail of the KM estimate of the survival function, we know the upper bound of the cure rate. Therefore, among the updated $\pi_u^{(k+1)}, \pi_1^{(k+1)}, \ldots, \pi_g^{(k+1)}$, the one closest to the tail value of the KM estimate is assigned as the estimate of the cure portion at the $k+1$ iteration.

In order to demonstrate the advantage of the mixture cure model over single models in applicable situations, we test their performances in two settings, one with simulated data and one with empirical data.

5.2 Simulated Data

An iid sample of 300 observations is generated to simulate a time-to-event random variable T which follows a mixture of two Weibull distributions. In addition, 30% of the population is simulated to be cured. Thus π_u, the rate of being uncured, is equal to 0.7. For the first Weibull distribution, the weight π_1 is set to equal 0.2, shape parameter $\gamma_1 = 0.5$ and scale parameter $\lambda_1 = 12$. For the second Weibull distribution, $\pi_2 = 0.5$, $\gamma_2 = 5$, and $\lambda_2 = 10$. Censoring is through another random variable T^{censor} which also follows a mixture of two Weibull distributions with a 30% cure portion. The two component Weibull distributions in the mixture for T^{censor} have scale 60 and 40, respectively. They share the same weights and shape parameters as their counterparts in the mixture for T. In addition, the data are truncated at $T = 50$. Therefore, an observation with either $T > T^{censor}$ or T> 50 is censored. This results in approximately a 35% censoring rate. No covariates are included in this simulation. Under this setting, the parameter vector is $\boldsymbol{\Psi} = (\pi_u, \pi_1, \gamma_1, \lambda_1, \gamma_2, \lambda_2)^T$, which includes the rate of being uncured, the weight for one of the

component Weibull distribution, the shape and scale parameters for each of the two component Weibull distributions.

We fit a mixture cure model of 2 Weibull, a mixture cure model of 1 Weibull, and the single parametric models to the simulated data. Like in the previous chapters, we compare the survival function plots, the restricted mean survival time at $T = 50$, and the mean survival time of these models to those of the underlying true distribution. We also use log likelihood and AIC to evaluate model fitting. Finally, we compare the estimated cure rate from the mixture cure model of 2 Weibull and mixture cure model of 1 Weibull. The results are included in Figure 17-A to 17-C as well as Table 13. For the mixture of 2 Weibull model and the mixture of 1 Weibull model, the 95% confidence intervals of the cure rate, survival function, restricted mean survival time, and mean survival time are constructed following the bootstrap approach discussed in section 3.6.

Given the simulation is generated with a 30% cure rate, the restricted mean survival time and the mean survival time are conditional on that the patients are not cured. The true restricted mean survival time at $T = 50$ is 7.5, and the true mean survival time is 9.39. Among all the models, the mixture cure model of 2 Weibull generates the estimates that are closest to the true values (6.77 and 8.49, respectively). The next closest values are from the mixture cure model of 1 Weibull, which is 5.6 for both the mean and restricted mean time. The single parametric models, due to the fact they do not include the cure portion, result in over-estimating the mean and restricted mean survival time by a large amount. Some of them cannot even calculate these values. In terms of model fitting, the survival function plots in Figure 17-A, 17-B, and 17-C show that the two mixture cure models trace the true survival curve much more closely than all the other single parametric models. In

Table 13, both the log likelihood and AIC values suggest that the best two are the mixture models. Log likelihood is -688.03 for the mixture of 2 Weibull model, and -779.41 for the mixture of 1 Weibull model. AIC for the two models are 1388.06 and 1564.82, respectively. This indicates that between the two mixture models with a cure portion, the mixture of 2 Weibull provides a better fit. In addition, the mixture of 2 Weibull provides an estimated cure rate of 0.32, which is closer to the true rate of 0.3, compared with the estimate of 0.34 provided by the mixture of 1 Weibull model. It is clear from this simulated situation that when there is presence of a cure portion, the mixture cure models that accounts for this feature will provide more accurate estimates of the survival function in general, and the mean survival time in particular, as well as an estimate of the possibility that the patients will be cured.

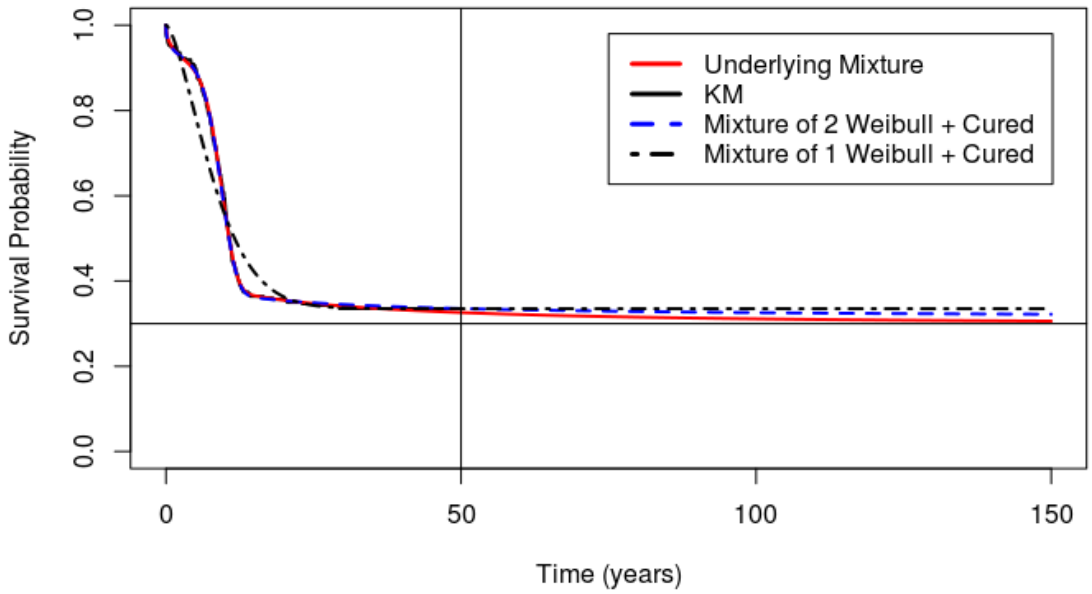

Figure 17-A Extrapolation of Survival Probability – Simulated Data with a Cure Portion

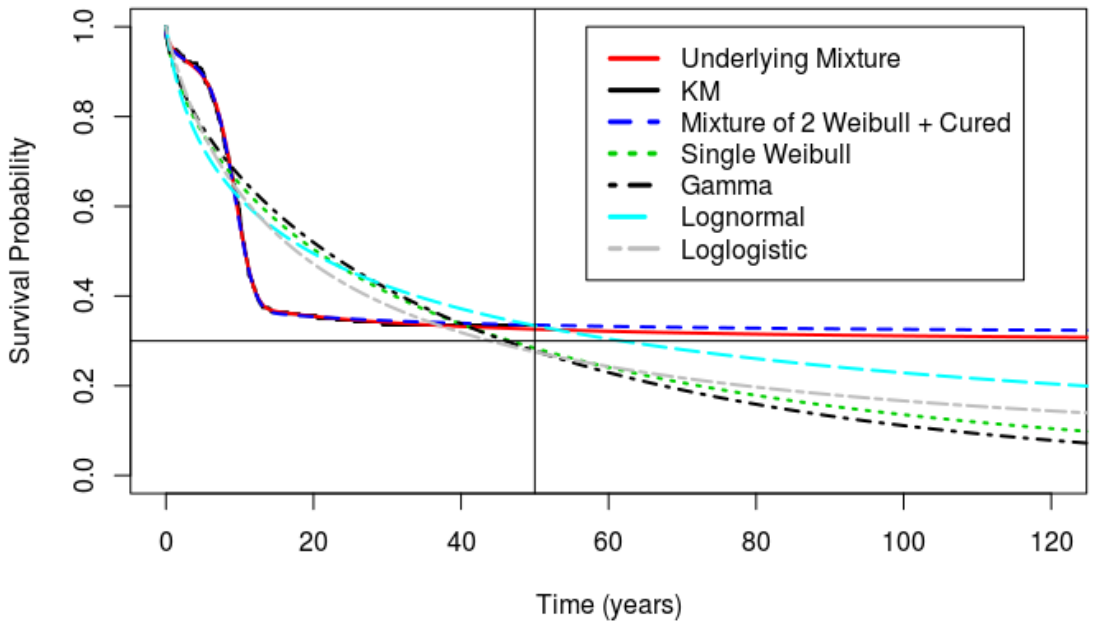

Figure 17-B Extrapolation of Survival Probability – Simulated Data with a Cure Portion

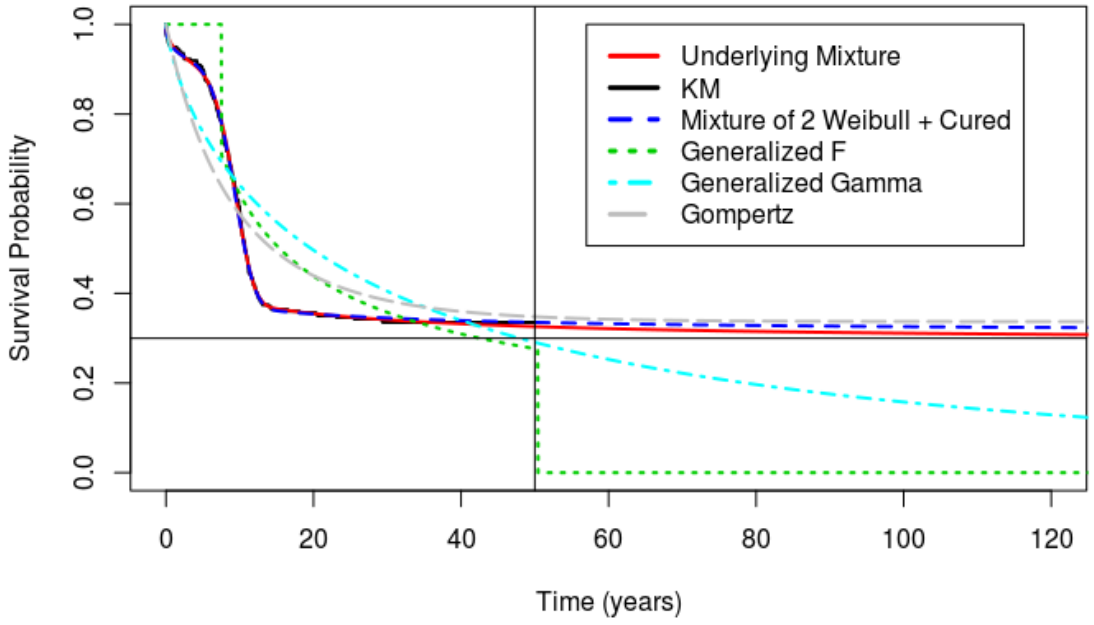

Figure 17-C Extrapolation of Survival Probability – Simulated Data with a Cure Portion

Table 13 Extrapolation of Survival Probability – Simulated Data with a Cure Portion

Models	Log Likelihood	AIC	Cure Rate	Restricted Mean (95% CI) T=50	Extrapolated Mean (95% CI) Lifetime
Underlying Model			0.3	7.50	9.39
KM				22.5	
Mixture of 2 Weibull + Cured	-688.03	1388.06	0.32 (0.17 – 0.36)	6.77 (5.66 – 15.13)	8.49 (95.91 – $6.71*10^8$)
Mixture of 1 Weibull + Cured	-779.41	1564.82	0.34 (0.28 – 0.37)	5.6 (5.14 – 6.19)	5.6 (5.14 – 6.19)
Single Weibull	-846.93	1697.85		24.8 (22.6 – 27)	47 (36.9 – 62.3)
Gamma	-852.15	1708.30		25.2 (23.1 – 27.6)	40.9 (33 – 50.4)
Lognormal	-853.86	1711.73		25.1 (22.7 – 27.4)	221 (122 – 462)
Log-logistic	-834.32	1672.65		23.9 (21.6 – 26)	
Generalized F	-794.39	1596.79			
Generalized Gamma	-844.02	1694.04		24.7 (22.2 – 26.9)	58.4 (41.1 – 99)
Gompertz	-810.33	1624.66		23.7 (21.3 – 26)	

5.3 Empirical Data

In Section 4.3, we observe that in the digitized data based on Hodi et al. (2010), both the IPI+GP100 and GP100 groups have an extremely large mean survival time for overall survival (Tables 11 and 12). Also, the KM estimates of the survival function have a long flat tail above 0. These are all signs that there might exist a cure portion in both groups. To test the existence of the potential cure rate, we fit a mixture cure model of 2 Weibull and a mixture cure model of 1 Weibull to the overall survival data for the IPI+GP100 group, along with the single parametric models. The log likelihood for the 2

Weibull mixture is -1062.15, and -1062.17 for the 1 Weibull mixture, AIC is 2136.29 for the 2 Weibull mixture and 2130.34 for the 1 Weibull mixture. Between the two mixture cure models, the mixture of 1 Weibull provides a decent fit to the data. The comparisons between the 1 Weibull mixture and the other single parametric models are displayed in Figures 18-A, 18-B and table 14. The 1 Weibull mixture cure model has the best values in both log likelihood and AIC. It comes up with an estimated cure rate of 0.21. For the uncured patients, the mean survival time becomes 7.65, and a restricted mean of 7.64 at $T = 56$, the ending time of the study. Based on what is observed in the data, the mixture cure model proves to be a better fit than the single models. It provides an estimate of the possibility of being cured, and more reasonable estimates of the mean survival time for patients who are not cured.

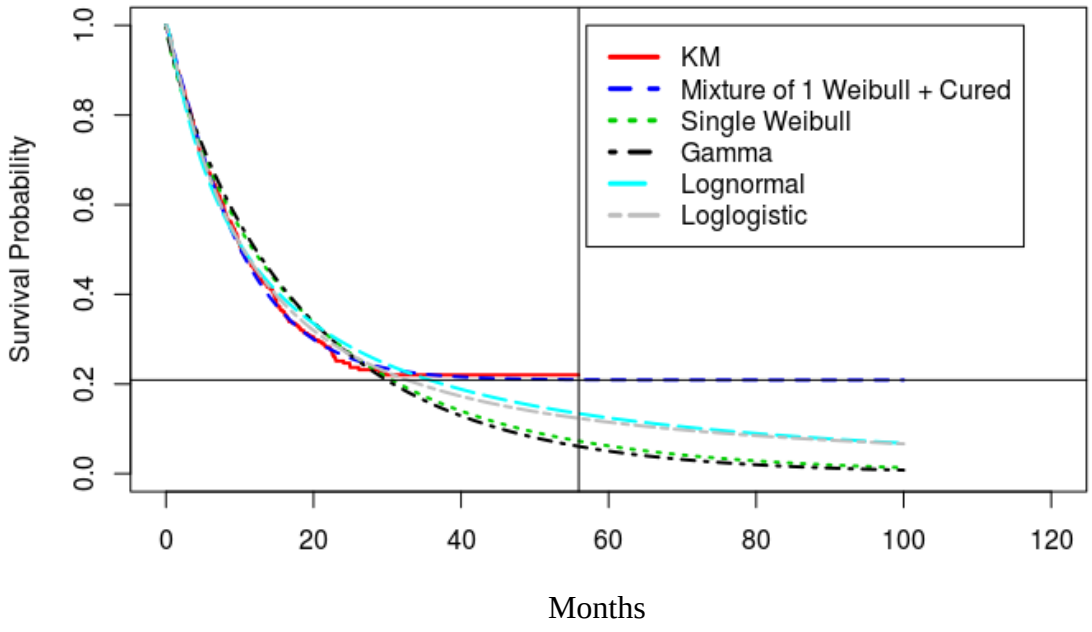

Figure 18-A Overall Survival with a Cure Portion – IPI+GP100

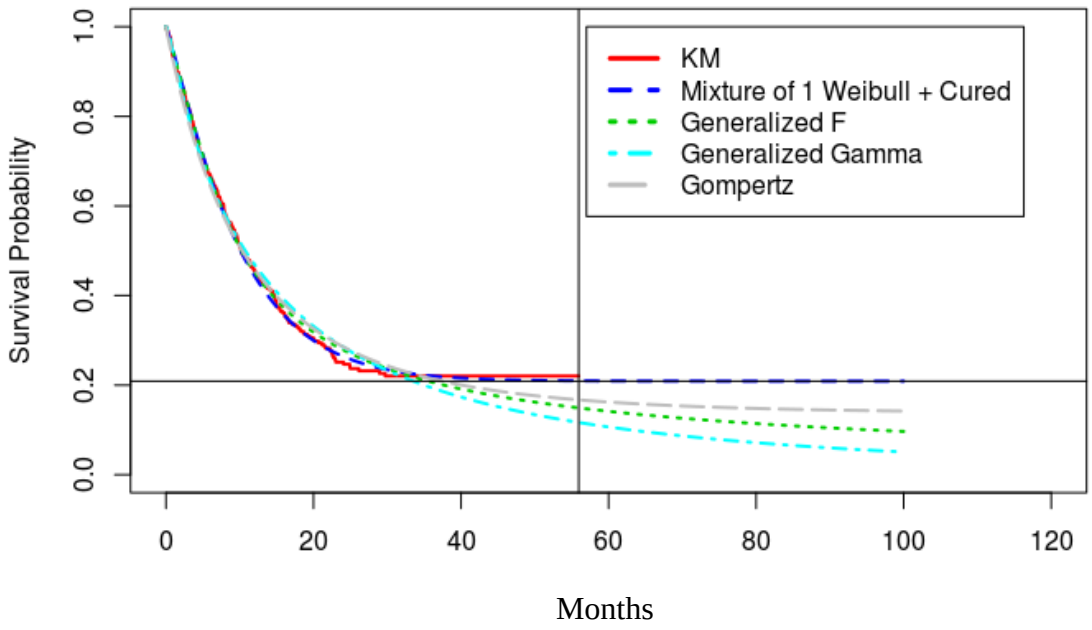

Figure 18-B Overall Survival with a Cure Portion – IPI+GP100

Table 14 Overall Survival with a Cure Portion – IPI+GP100

Models	Log Likelihood	AIC	Cure Rate	Restricted Mean (95% CI) T=56	Extrapolated Mean (95% CI) Lifetime
KM				19.4	
Mixture of 1 Weibull + Cured	-1062.17	2130.34	0.21 (0.16 – 0.26)	7.64 (6.5 – 8.77)	7.65 (6.5 – 8.78)
Single Weibull	-1080.59	2165.18		17.7 (15.9 – 19.6)	19.7 (16.8 – 23.1)
Gamma	-1082.96	2169.91		17.6 (15.9 – 19.6)	18.9 (16.6 – 21.7)
Lognormal	-1072.30	2148.60		18.8 (16.7 – 20.9)	32.9 (24.9 – 42.3)
Log-logistic	-1068.99	2141.99		18.3 (16.5 – 20.3)	63.5 (37.1 – 184)
Generalized F	-1067.58	2143.15		18.8 (16.8 – 21.0)	
Generalized Gamma	-1071.43	2148.87		18.4 (16.5 – 20.5)	26.5 (20.1 – 44.2)
Gompertz	-1068.35	2140.70		19.0 (17.0 – 21.2)	

The SHEP patients are followed for over 20 years after the conclusion of the randomized phase. Over this long period of time, some patients are lost to follow-up. As a result, their mortality information is unknown and they will appear to be always alive. In this sense, they should be treated as "cured", as including them in the extrapolation will result in over estimation. To evaluate the impact of the potential cure portion on the extrapolated survival function, we fit a 1 Weibull mixture cure model to the all-cause mortality data up to December 31st, 2014 in SHEP. The results are presented in Figure 19-A, 19-B, and Table 15. The 1 Weibull mixture cure model has the best log likelihood (-13917.73) and AIC value (27841.46). It estimates a cure rate of 0.216. The restricted mean survival time at $T = 29.7$ years and the extrapolated mean survival time are 11.3 and 11.7 respectively. These are survival times for SHEP patients who have mortality and cause of death information available. These estimates will be confirmed as data beyond 2014 for SHEP patients become available in the future.

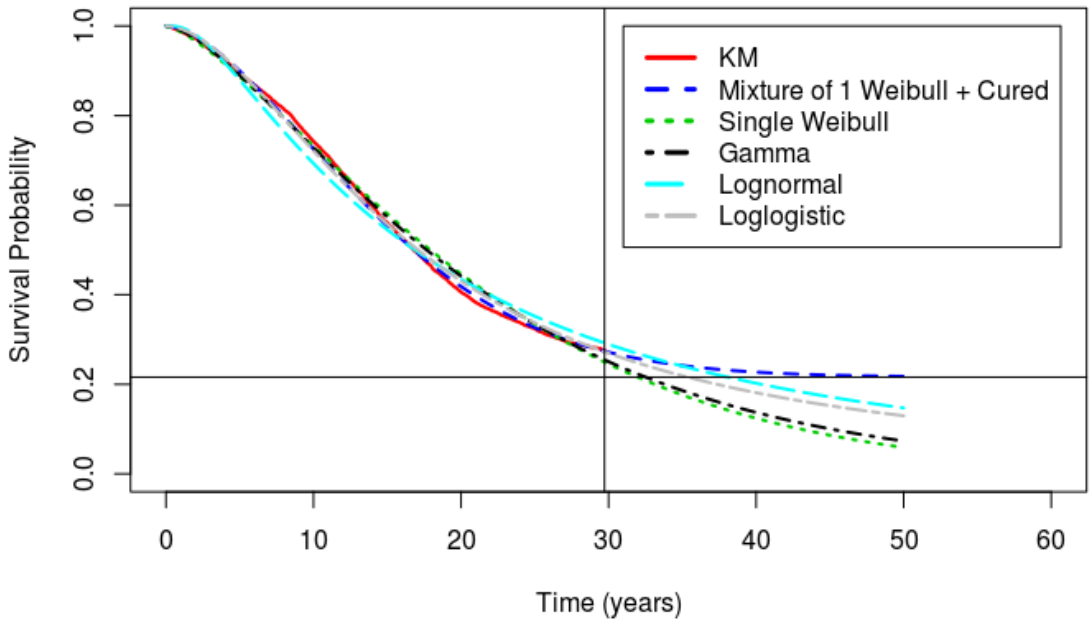

Figure 19-A All-cause Mortality with a Cure Portion – All Patients in SHEP

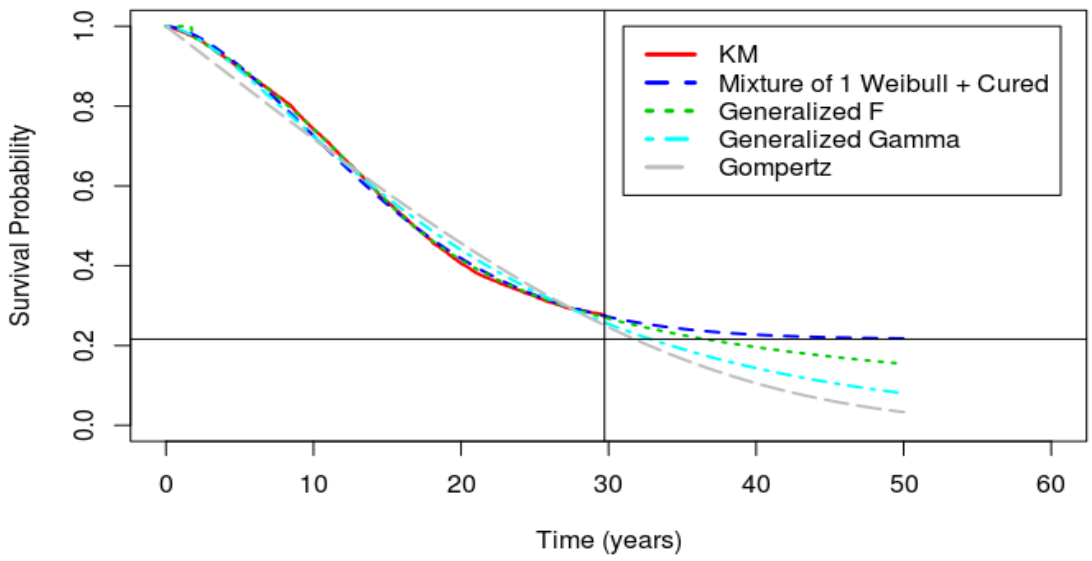

Figure 19-B All-cause Mortality with a Cure Portion – All Patients in SHEP

Table 15 Extrapolation of Mean Survival Time with a Cure Portion – All-cause Mortality – All Patients in SHEP

Models	Log Likelihood	AIC	Cure Rate (95% CI)	Restricted Mean (95% CI) T=29.7	Extrapolated Mean (95% CI) Lifetime
KM				17.8	
Mixture of 1 Weibull + Cured	-13917.73	27841.46	0.216 (0.198 – 0.23)	11.3 (10.9 – 11.9)	11.7 (11.2 – 12.4)
Single Weibull	-13973.41	27950.82		18 (17.7 – 18.2)	21.4 (20.9 – 21.9)
Gamma	-13960.55	27925.10		17.9 (17.6 – 18.2)	22 (21.5 – 22.6)
Lognormal	-14065.77	28135.55		17.7 (17.4 – 18)	28.9 (27.6 – 30.2)
Log-logistic	-13947.31	27898.63		17.8 (17.6 – 18.1)	30.2 (29.4 – 32.5)
Generalized F	-13897.10	27802.20		17.8 (17.5 -18.1)	
Generalized Gamma	-13959.37	27924.74		17.9 (17.6 – 18.2)	22.3 (21.6 – 23.2)
Gompertz	-14064.23	28132.46		17.8 (17.5 – 18.1)	20.4 (19.9 – 20.9)

Appendix A: Derivation of the Maximum Likelihood Estimators in the Finite Mixture Models

Suppose that $\boldsymbol{y} = (y_1, \dots, y_n)^T$ is an iid random sample from a population following a g-component mixture density, that is, for $j = 1, \dots, n$,

$$f(y_j|\boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_i f_i(y_j|\boldsymbol{\theta}_i), \tag{A.1}$$

where $\boldsymbol{\Psi} = (\pi_1, \dots, \pi_{g-1}, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_g)^T$ is the vector containing all the unknown parameters. The π_i are the weights, and the $\boldsymbol{\theta}_i$ contain the respective parameters of each component density. The log likelihood function of the random sample $\boldsymbol{y}$ is given by

$$l(\boldsymbol{\Psi}|\,\boldsymbol{y}) = \sum_{j=1}^{n} \log\left[f(y_j|\boldsymbol{\Psi})\right] = \sum_{j=1}^{n} \log\left[\sum_{i=1}^{g} \pi_i f_i(y_j|\boldsymbol{\theta}_i)\right]. \tag{A.2}$$

In order to get (3.13), we take partial derivatives of (A.2) with respect to π_i ($i = 1, \dots, g-1$), we get

$$\frac{\partial l(\boldsymbol{\Psi}|\,\boldsymbol{y})}{\partial \pi_i} = \sum_{j=1}^{n} \left[\frac{f_i(y_j|\boldsymbol{\theta}_i)}{\sum_{i=1}^{g} \pi_i f_i(y_j|\boldsymbol{\theta}_i)} - \frac{f_g(y_j|\boldsymbol{\theta}_g)}{\sum_{i=1}^{g} \pi_i f_i(y_j|\boldsymbol{\theta}_i)}\right]. \tag{A.3}$$

This is because $\pi_g = 1 - \sum_{i=1}^{g-1} \pi_i$. The MLE of $\pi_i, \hat{\pi}_i$, satisfies

$$\sum_{j=1}^{n} \left[\frac{f_i(y_j|\boldsymbol{\theta}_i)}{\sum_{i=1}^{g} \hat{\pi}_i f_i(y_j|\boldsymbol{\theta}_i)} - \frac{f_g(y_j|\boldsymbol{\theta}_g)}{\sum_{i=1}^{g} \hat{\pi}_i f_i(y_j|\boldsymbol{\theta}_i)}\right] = 0. \tag{A.4}$$

By multiplying $\hat{\pi}_i$ to both sides of (A.4) and $\hat{\pi}_g \big/ \hat{\pi}_g$ to the second part of the left side of (A.4), we obtain

$$\sum_{j=1}^{n} \left[\frac{\hat{\pi}_i f_i(y_j|\boldsymbol{\theta}_i)}{\sum_{i=1}^{g} \hat{\pi}_i f_i(y_j|\boldsymbol{\theta}_i)} - \frac{\hat{\pi}_i \hat{\pi}_g f_g(y_j|\boldsymbol{\theta}_g)}{\hat{\pi}_g \sum_{i=1}^{g} \hat{\pi}_i f_i(y_j|\boldsymbol{\theta}_i)}\right] =$$

$$\sum_{j=1}^{n}\left[\tau_i(y_j|\boldsymbol{\theta}_i)-\frac{\hat{\pi}_i}{\hat{\pi}_g}\tau_g(y_j|\boldsymbol{\theta}_i)\right]=0. \tag{A.5}$$

for $i=1,\ldots,g-1$, where

$$\tau_i(y_j|\boldsymbol{\theta}_i)=\frac{\hat{\pi}_i f_i(y_j|\boldsymbol{\theta}_i)}{\sum_{i=1}^{g}\hat{\pi}_i f_i(y_j|\boldsymbol{\theta}_i)}. \tag{A.6}$$

Since (A.5) also holds for $i=g$, we can sum over $i=1,\ldots,g$ in (A.5) to give

$$\sum_{j=1}^{n}\sum_{i=1}^{g}\left[\tau_i(y_j|\boldsymbol{\theta}_i)-\frac{\hat{\pi}_i}{\hat{\pi}_g}\tau_g(y_j|\boldsymbol{\theta}_g)\right]=\sum_{j=1}^{n}\left[1-\frac{1}{\hat{\pi}_g}\tau_g(y_j|\boldsymbol{\theta}_g)\right]=0. \tag{A.7}$$

from (A.7), we get

$$\hat{\pi}_g=\frac{\sum_{j=1}^{n}\tau_g(y_j|\boldsymbol{\theta}_g)}{n}. \tag{A.8}$$

Substitute (A.8) into (A.5) yields

$$\hat{\pi}_i=\frac{\sum_{j=1}^{n}\tau_i(y_j|\boldsymbol{\theta}_i)}{n}\ (i=1,\ldots,g-1). \tag{A.9}$$

Together, (A.8) and (A.9) give us (3.13).

To derive (3.14), we take partial derivatives of (A.2) with respect to $\boldsymbol{\theta}_i$ $(i=1,\ldots,g)$, we get

$$\frac{\partial l(\boldsymbol{\Psi}|\,y)}{\partial\boldsymbol{\theta}_i}=\sum_{j=1}^{n}\left[\frac{\frac{\partial}{\partial\boldsymbol{\theta}_i}\sum_{i=1}^{g}\pi_i f_i(y_j|\boldsymbol{\theta}_i)}{\sum_{i=1}^{g}\pi_i f_i(y_j|\boldsymbol{\theta}_i)}\right]=$$

$$\sum_{j=1}^{n}\left[\frac{\frac{\partial}{\partial\boldsymbol{\theta}_i}\pi_i f_i(y_j|\boldsymbol{\theta}_i)}{\sum_{i=1}^{g}\pi_i f_i(y_j|\boldsymbol{\theta}_i)}\right]\ (i=1,\ldots,g). \tag{A.10}$$

By multiplying $f_i(y_j|\boldsymbol{\theta}_i)$ to both the numerator and denominator of (A.10), we obtain

$$\frac{\partial l(\boldsymbol{\Psi}|\,\boldsymbol{y})}{\partial \boldsymbol{\theta}_i} = \sum_{j=1}^{n}\left[\frac{\pi_i f_i(y_j|\boldsymbol{\theta}_i)\frac{\partial}{\partial \boldsymbol{\theta}_i} f_i(y_j|\boldsymbol{\theta}_i)}{f_i(y_j|\boldsymbol{\theta}_i)\sum_{i=1}^{g}\pi_i f_i(y_j|\boldsymbol{\theta}_i)}\right] =$$

$$\sum_{j=1}^{n}\left\{\tau_i(y_j|\boldsymbol{\theta}_i)\left[\frac{\partial}{\partial \theta_i}\log f_i(y_j|\boldsymbol{\theta}_i)\right]\right\}. \qquad \text{(A.11)}$$

for $i = 1, \ldots, g$. And the MLE of $\boldsymbol{\theta}_i$, $\widehat{\boldsymbol{\theta}}_i$, satisfies

$$\sum_{j=1}^{n}\left\{\tau_i(y_j|\widehat{\boldsymbol{\theta}}_i)\left[\frac{\partial}{\partial \theta_i}\log f_i(y_j|\widehat{\boldsymbol{\theta}}_i)\right]\right\} = \mathbf{0}\ (i = 1, \ldots, g). \qquad \text{(A.12)}$$

Appendix B: Common Parametric Survival Models

A.16 Summary of Common Parametric Survival Models

Model	Parameters	Probability Density Function	Survival Function	Hazard Function	PH or AFT
Weibull	λ, γ	$f(t) = \lambda\gamma t^{\gamma-1} e^{-\lambda t^{\gamma}}$	$S(t) = e^{-\lambda t^{\gamma}}$	$h(t) = \lambda\gamma t^{\gamma-1}$	AFT/PH
Exponential	λ	$f(t) = \lambda e^{-\lambda t}$	$S(t) = e^{-\lambda t}$	$h(t) = \lambda$	AFT/PH
Lognormal	μ, σ	$f(t) = \frac{1}{\sqrt{2\pi}\sigma t} e^{-\frac{(logt-\mu)^2}{2\sigma^2}}$	$S(t) = 1 - \Phi(\frac{logt-\mu}{\sigma})$	$h(t) = \frac{\phi(\frac{logt-\mu}{\sigma})}{\sigma t[1-\Phi(\frac{logt-\mu}{\sigma})]}$	AFT
Gamma	α, β	$f(t) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} t^{\alpha-1} e^{-\frac{t}{\beta}}$, where $\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1} e^{-t} dt$	No closed form	No closed form	AFT
Log-logistic	α, β	$f(t) = \frac{\frac{\alpha}{\beta}(\frac{t}{\beta})^{\alpha-1}}{(1+(\frac{t}{\beta})^{\alpha})^2}$	$S(t) = \frac{1}{1+\left(t/\beta\right)^{\alpha}}$	$h(t) = \frac{\frac{\alpha}{\beta}(\frac{t}{\beta})^{\alpha-1}}{1+(\frac{t}{\beta})^{\alpha}}$	AFT
Gompertz	α, β	$f(t) = \beta e^{\alpha t - \frac{\beta}{\alpha(e^{\alpha t}-1)}}$	$S(t) = e^{-\frac{\beta}{\alpha(e^{\alpha t}-1)}}$	$h(t) = \beta e^{\alpha t}$	PH
Generalized Gamma	μ, σ, Q	If $\gamma \sim Gamma(Q^{-2}, 1), and\ w = \frac{\log(Q^2\gamma)}{Q}, then\ t = e^{\mu+\sigma w}$ follows the Generalized gamma distribution with probability density function $f(t\|\mu,\sigma,Q) = \frac{\|Q\|(Q^{-2})^{Q^{-2}}}{\sigma t\Gamma(Q^{-2})} e^{(Q^{-2}(Qw-e^{Qw}))}$	No closed form	No closed form	AFT
Generalized F	σ, μ, Q, P	If $y \sim F(2s_1, 2s_2), and\ w = \log(y), then\ t = e^{w\sigma+\mu}$ follows the Generalized F distribution. Let $s_1 = 2(Q^2 + 2P + Q\delta)^{-1}, s_2 = 2(Q^2 + 2P - Q\delta)^{-1}$. Equivalently, $Q = \left(\frac{1}{s_1} - \frac{1}{s_2}\right)\left(\frac{1}{s_1} + \frac{1}{s_2}\right)^{-\frac{1}{2}}, P = \frac{1}{s_1 + s_2}$ Define $\delta = (Q^2 + 2P)^{\frac{1}{2}}, and\ w = \frac{(\log(t)-\mu)\delta}{\sigma}$	No closed form	No closed form	AFT

		Then the probability density function of t is $f(t) = \frac{\delta\left(\frac{s_1}{s_2}\right)^{s_1} e^{s_1 w}}{\sigma t\left(1+\frac{s_1 e^{w}}{s_2}\right)^{(s_1+s_2)} B(s_1,s_2)}$ and $B(s_1, s_2) = \frac{\Gamma(s_1)\Gamma(s_2)}{\Gamma(s_1+s_2)}$ is the beta function.			

Note: PH = proportional hazard.

Bibliography

Amaratunga, D., and Cabrera, J. (2015), “Review of New Statistical Techniques for Analysis of Cardiovascular Trial and Registry Data,” *Current Hypertension Reports*, 17 (10), 1-6.

Amico, M., and Keilegom, I.V. (2018), “Cure Models in Survival Analysis,” *Annual Review of Statistics and Its Application*, 5:1, 311-342.

Anthony, Y. C. Kuk, and Chen, C. H. (1992), "A Mixture Model Combining Logistic Regression with Proportional Hazards Regression," *Biometrika*, 79 (3), 531-541.

Basford, K., Greenway, D. R., McLachlan, G., and Peel, D. (1997), “Standard Errors of Fitted Component Means of Normal Mixtures,” *Computational Statistics*, Vol. 12.

Beck, J. (1990), "How to evaluate drugs: Cost-effectiveness analysis," *JAMA*, 264 (1), 83-84.

Berkson, J., and Robert P. G. (1952), "Survival Curve for Cancer Patients Following Treatment," *Journal of the American Statistical Association*, 47 (259), 501-515.

Biernacki, C., Gilles, C., and Govaert, G. (1998), “Assessing a Mixture Model for Clustering With the Integrated Classification Likelihood,” *INRIA*, RR-3521.

—— (2003), "Choosing Starting Values for the EM Algorithm for Getting the Highest Likelihood in Multivariate Gaussian Mixture Models," *Computational Statistics and Data Analysis*, 41 (3), 561-575.

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), 15-44.

Böhning, D., Dietz, E., Schaub, R., Schlattmann, P., and Lindsay, B. (1994), "The Distribution of the Likelihood Ratio for Mixtures of Densities From the One-parameter Exponential Family," *Annals of the Institute of Statistical Mathematics*, 46 (2), 373-388.

Casella, G., and Lehmann, E. L. (2002), *Statistical inference* (2nd ed.), Duxbury advanced series, Australia ; Pacific Grove, CA: Thomson Learning.

Bulpitt, C. J. (1996), *Randomized Controlled Clinical Trials* (2nd ed.), New York, NY: Springer US.

Corbière, F., Commenges, D., Taylor, Jeremy M. G., and Joly, P. (2009), "A Penalized Likelihood Approach for Mixture Cure Models," *Statistics in Medicine*, 28 (3), 510-524.

Ćwik, J., and Koronacki, J., (1997), "A Combined Adaptive-mixtures/plug-in Estimator of Multivariate Probability Densities." *Computational Statistics and Data Analysis*, 26 (2), 199-218.

Davino, C. (2014), *Quantile Regression: Theory and Applications*, Wiley series in probability and statistics. Chichester, England: Wiley.

Day, N. E. (1969), "Estimating the Components of a Mixture of Normal Distributions," *Biometrika*, 56 (3), 463-474.

Dempster, A. P., Laird, N.M., and Rubin, D. B. (1977) "Maximum Likelihood From Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society Series B (Methodological)*, 39 (1), 1-38.

Dietz, E., and Dankmar, B. (1995), "Statistical Inference Based on a General Model of Unobserved Heterogeneity," in *Statistical Modelling Lecture Notes in Statistics*, vol 104, Springer, New York, NY

Efron, B. (1979), "Bootstrap Methods: Another Look at the Jackknife," *The Annals of Statistics*, 7 (1), 1-26.

—— (1981), "Censored Data and the Bootstrap," *Journal of the American Statistical Association*, 76 (374), 312-319.

Erişoğlu, E., Ülkü, M., and Erol, H. (2011), "A Mixture Model of Two Different Distributions Approach to the Analysis of Heterogeneous Survival Data," International Journal of Computational & Mathematical Sciences, Vol. 5, 75

Farewell, V. T. (1977), "A Model for a Binary Variable With Time-Censored Observations," *Biometrika*, 64 (1), 43-46.

—— (1982), "The Use of Mixture Models for the Analysis of Survival Data with Long-term Survivors," *Biometrics*, 38 (4), 1041-1046.

—— (1986), "Mixture Models in Survival Analysis: Are They Worth the Risk?" *Canadian Journal of Statistics*, 14 (3), 257-262.

Fleming, T. R., and Lin, D. Y. (2000), "Survival Analysis in Clinical Trials: Past Developments and Future Directions," *Biometrics*, 56 (4), 971-983.

Fraley, C., and Raftery, A. (2006), "MCLUST Version 3: An R Package for Normal Mixture Modeling and Model-based Clustering," *Univeristy of Washington Tech Report*, 504, 51.

Furman, W. D., and Lindsay B. G. (1994), "Testing for the Number of Components in a Mixture of Normal Distributions Using Moment Estimators," *Computational Statistics and Data Analysis*, 17 (5), 473-492.

Ghitany, M. E., Maller, R. A., and Zhou, S. (1994) "Exponential Mixture Models with Long-Term Survivors and Covariates," *Journal of Multivariate Analysis*, 49 (2),218-241.

Hao, L. X. (2007), “Quantile regression,” *Quantitative applications in the social sciences*, 149.

Hipp, J. R., and Bauer, D. J. (2006), "Local Solutions in the Estimation of Growth Mixture Models," *Psychological Methods*, 11 (1), 36-53.

Hodi, F. Stephen., O'Day, S.J., McDermott, D. F., Weber, R. W., Sosman, J. A., Haanen, J. B., Gonzalez, R., Robert, C., Schadendorf, D., Hassel, J. C., Akerley, W., van den Eertwegh, A. J. M., Lutzky, J., Lorigan, P., Vaubel, J. M., Linette, G. P., Hogg, D., Ottensmeier, C. H., Lebbé, C., Peschel, P., Quirt, I., Clark, J. I., Wolchok, J. D., Weber, J. S., Tian, J., Yellin, M. J., Nichol, G. M., Hoos, A., and Urba, W. J. (2010), "Improved Survival with Ipilimumab in Patients with Metastatic Melanoma," *New England Journal of Medicine*, 363 (8), 711-723.

Jackson, C., Stevens, J., Ren, S. J., Latimer, N., Bojke, L., Manca, A., and Sharples, L. (2017), "Extrapolating Survival from Randomized Trials Using External Data: A Review of Methods," *Medical Decision Making*, 37 (4),377-390.

Kabisch, M., Ruckes, B., Seibert-Grafe, M., and Blettner, M. (2011) "Randomized Controlled Trials: Part 17 of a Series on Evaluation of Scientific Publications." *Deutsches Arzteblatt international*, 108 (39), 663-668.

Dimitris, K., and Xekalaki, E., (2003), "Choosing Initial Values for the EM Algorithm for Finite Mixtures." *Computational Statistics & Data Analysis*, 41 (3), 577-590.

Klein, J. P., van Houwelingen, H. C., Ibrahim, J. G., and Scheike, T. H. (2016), *Handbook of Survival Analysis* (1st ed.), Chapman & Hall/CRC Handbooks of Modern Statistical Methods: CRC Press.

Kleinbaum, D. G., (2005), *Survival Analysis A Self-Learning Text* (2nd ed.), New York, NY: Springer New York.

Kostis, J. B., Cabrera, J., Cheng, J. Q., Cosgrove, N. M., Deng, Y. Z., Pressel, S. L., and Davis, B. R. (2011), "Association Between Chlorthalidone Treatment of Systolic Hypertension and Long-term Survival," *JAMA*, 306 (23), 2588-2593.

Lee, E. T., and Wang, J. W. (2003) *Statistical Methods for Survival Data Analysis* (3rd ed.), Wiley series in probability and statistics, New York: J. Wiley.

Li, C. S., and Taylor, J. M. G. (2002), "A semi-parametric Accelerated Failure Time Cure Model," *Statistics in Medicine*, 21 (21), 3235-3247.

Lindsay, B. G., and Basak, P. (1993), "Multivariate Normal Mixtures: A Fast Consistent Method of Moments," *Journal of the American Statistical Association*, 88 (422):468-476.

Liu, C. H. (1998), "Information Matrix Computation from Conditional Information via Normal Approximation," *Biometrika*, 85 (4), 973-979.

Lu, W. B. (2010) "Efficient Estimation for an Accelerated Failure Time Model With a Cure Fraction," *Statistica Sinica*, 20:661.

Lubke, G., and Muthén, B. O. (2007), "Performance of Factor Mixture Models as a Function of Model Size, Covariate Effects, and Class-Specific Parameters," *Structural Equation Modeling: A Multidisciplinary Journal*, 14 (1), 26-47.

Marín, J. M., Rodríguez-Bernal, M. T., and Wiper, M. P. (2005), "Using Weibull Mixture Distributions to Model Heterogeneous Survival Data," *Communications in Statistics - Simulation and Computation*, 34 (3), 673-684.

McLachlan, G. J. (1988), "On the Choice of Starting Values for the EM Algorithm in Fitting Mixture Models," *Journal of the Royal Statistical Society: Series D (The Statistician)*, 37 (4-5), 417-425.

McLachlan, G. J., and Krishnan, T., (2008), *The EM Algorithm and Extensions* (2nd ed,), Wiley series in probability and statistics. Hoboken, N.J.: Wiley-Interscience.

McLachlan, G. J., and Peel, D. (2000), *Finite Mixture Models*, New York: Wiley.

Muennig, P. (2007) *Cost-effectiveness Analyses in Health: A Practical Approach* (2nd ed.), Scitech Book News 32, no. 1

Othus, M., Barlogie, B., Leblanc, M. L., and Crowley, J. J. (2012) "Cure Models as a Useful Statistical Tool for Analyzing Survival," *Clinical Cancer Research*, 18 (14), 3731-6.

Pearson, K. (1894), "Contributions to the Mathematical Theory of Evolution," *Philosophical Transactions of the Royal Society of London*, 185:71-110.

Peng, L., and Huang, Y. (2008), "Survival Analysis With Quantile Regression Models," *Journal of the American Statistical Association*, 103 (482), 637-649.

Peng, Y., and Dear, Keith B. G. (2000), "A Nonparametric Mixture Model for Cure Rate Estimation," *Biometrics*, 56 (1), 237-243.

Peng, Y., Dear, Keith B. G., and Denham, J. W. (1998), "A Generalized F Mixture Model for Cure Rate Estimation," *Statistics in Medicine*, 17 (8), 813-830.

Portnoy, S. (2003), "Censored Regression Quantiles," *Journal of the American Statistical Association*, 98 (464), 1001-1012.

Schmidt, P., and Witte, A. D. (1989), "Predicting Criminal Recidivism Using 'Split Population' Survival Time Models," *Journal of Econometrics*, 40 (1), 141-159.

SHEP Cooperative Research Group (1991), "Prevention of Stroke by Antihypertensive Drug Treatment in Older Persons With Isolated Systolic Hypertension: Final Results of the Systolic Hypertension in the Elderly Program (SHEP)," *JAMA*, 265 (24),3255-3264.

Shireman, E., Steinley, D., and Brusco, M. (2017), "Examining the Effect of Initialization Strategies on the Performance of Gaussian Mixture Modeling," *Behavior Research Methods*, 49 (1), 282-293.

Singh, R., and Mukhopadhyay, K. (2011), "Survival Analysis in Clinical Trials: Basics and Must Know Areas," *Perspectives in Clinical Research*, 2 (4), 145-8.

Solka, J., Wegman, E., Priebe, C., Poston, W., and Rogers, G. (1998) "Mixture Structure Analysis Using the Akaike Information Criterion and the Bootstrap," *Statistics and Computing*, 8 (3), 177-188.

Sy, J. P., and Taylork Jeremy M. G. (2000), "Estimation in a Cox Proportional Hazards Cure Model," *Biometrics*, 56 (1), 227-236.

Wolfe, J. H., and Naval Personnel Research Activity San Diego (1965). “A Computer Program for the Maximum Likelihood Analysis of Types,” available at *http://handle.dtic.mil/100.2/AD620026.*

Xue, X., Xie, X., and Strickler, H. D. (2016), "A Censored Quantile Regression Approach for the Analysis of Time to Event Data," *Statistical Methods in Medical Research*, 27 (3), 955-965.

Yamaguchi, K. (1992), "Accelerated Failure-Time Regression Models With a Regression Model of Surviving Fraction: An Application to the Analysis of 'Permanent Employment' in Japan," *Journal of the American Statistical Association*, 87 (417), 284.

Zhang, J., and Peng, Y. (2007), "A New Estimation Method for the Semiparametric Accelerated Failure Time Mixture Cure Model," *Statistics in Medicine,* 26 (16), 3157-3171.