ASSESSMENT OF PREDICTIVE ACCURACY OF MIXTURE CURE MODELS

by

Anjali Kulkarni

A report submitted to the Department of Community Health and Epidemiology

In conformity with the requirements for

the degree of Master of Science

Queen's University

Kingston, Ontario, Canada

(August, 2011)

Supervisors: Dr. Paul Peng and Dr. Wenyu Jiang

Acknowledgements

I first thank my supervisors, Dr. Paul Peng and Dr. Wenyu Jiang for their continuous support for the duration of my master's program. The advice and feedback given by them was invaluable and enriched my learning experience. As well, a special thanks to my friends (and co-biostatistics students) Katelyn Balchin and Brianne Wood for the support and help during the summer. I also thank my friends Kendra Barrick and Holly Garnett for being there during all the stressful times, particularly with my mishaps with Kingston Transit. Lastly, a thanks to my parents, brother and friends for the unwavering support and encouragement provided throughout the years.

Table of Contents

Acknowledgements ii
Chapter 1 Introduction 1
Chapter 2 Background 3
Chapter 3 Simulation Study Based on Exponential Distribution 9
Chapter 4 Simulation Study Based on Weibull Distribution 19
Chapter 5 Application: Bone Marrow Transplant Study 29
Chapter 6 Conclusions and Future Directions 37
Bibliography 39

List of Figures

Figure 1: Kaplan-Meier curves of training data for exponential simulation study 11
Figure 2: Exponential mixture cure model predictive distributions for exponential simulation study 13
Figure 3: Exponential predictive distributions for exponential simulation study 14
Figure 4: Naïve and Graf estimates of Brier score for exponential simulation study 16
Figure 5: Kaplan-Meier curves of test data of exponential simulation study 18
Figure 6: Kaplan-Meier curves of training data for Weibull simulation study 20
Figure 7: MCM Weibull predictive distributions for Weibull simulation study 22
Figure 8: Weibull predictive distributions for Weibull simulation study 23
Figure 9: Naïve and Graf estimates of Brier score for Weibull simulation study 26
Figure 10: Kaplan-Meier curves of test data of Weibull simulation study compared to MCM Weibull and Weibull predictive models 27
Figure 11: Kaplan-Meier curves for transplant data 30
Figure 12: Naïve and Graf estimates of Brier score for transplant study 34
Figure 13: Kaplan-Meier curves of transplant data with MCM Weibull and Weibull predictive models 35

List of Tables

Table 1: Maximum likelihood estimates of parameters of exponential mixture cure model for simulation study 12
Table 2: Maximum likelihood estimates of exponential model for simulation study 13
Table 3: Average estimates of Brier scores for exponential simulation study 14
Table 4: Maximum likelihood estimates of Weibull mixture cure model for simulation study 21
Table 5: Maximum likelihood estimates of Weibull model for simulation study 22
Table 6: Average estimates of Brier scores for Weibull simulation study 24
Table 7: Maximum likelihood estimates of Weibull mixture cure model for transplant study 31
Table 8: Maximum likelihood estimates of Weibull model for example study 32
Table 9: Average estimates of Brier scores for example study 32

Chapter 1

Introduction

Regression models are often used to estimate the relationship between a set of covariates, denoted Z , and a response variable, denoted Y . When the response variable is time to some event, survival analysis (or time-to-event analysis) is used to analyze and predict data. In survival analysis, Y could be the survival time of a patient recently diagnosed with a particular disease. Z will contain information on the patient's age, sex, stage of disease at diagnosis and other covariates that could be related to the patient's survival time. Most regression analysis work involves estimating model parameters and associated confidence intervals for each covariate to determine their effects on the response variable (Lawless, 2003).

In survival analysis, the response variable Y (or time to event variable) is usually subject to censoring. More specifically, right censoring is very common in survival analysis data; one type of right censoring is when the time to event is greater than the length of observation. When right censoring occurs, the exact time to event is unknown due to reasons such as people dropping out of the study, loss to follow-up, or study termination (Lawless & Yuan, 2010). For a dataset with censoring distribution, a censoring indicator δ_i is used to indicate whether censoring has occurred or not. Let Y_i be the time to event for the i th observation and let C_i be the censoring time.

$$\delta_i = I(Y_i \leq C_i);$$

$\delta_i = 0$ if the subject is censored and $\delta_i = 1$ if the subject is uncensored. In this report, we will assume that the censoring distribution is independent of the times to event and the associated covariates (Graf, Schmoor, Sauerbrei, & Schumacher, 1999).

Survival models are often used to form the basis of predictions relating to patient survival. An example of the use of survival analysis models is when physicians wish to categorize newly diagnosed patients into risk groups based on various features of the patients. The survival analysis model would take into account several features (covariates) and will form a time to event (usually death in survival analysis) related prediction. It is important to assess the accuracy of

these predictions; this will help to determine if the “right” model is being used, or the “right” prediction method (Lawless et al, 2010).

Another set of models used in survival analysis are mixture cure models. These models are used when a sizable portion of the population of interest are subject to censoring even with long-term follow-up. This fraction of the population is considered to be cured and no longer at risk of experiencing the event of interest (Farewell, 1982). Mixture cure models incorporate both the cure fraction and the survival function for the uncured proportion of the population. They are used to form predictions relating to patient survival for diseases which have cures (Yu & Peng, 2008) (Peng, Taylor & Yu, 2007).

Because of the use of survival models and mixture cure models for forming predictions, prediction error or predictive accuracy of these models is also of interest to statisticians. This measure determines how accurate the prediction method is and indicates if changes need to be made (i.e. more or less covariates are needed or parameter estimates need to be modified or if a different model needs to be used).

One field in which predictive accuracy has been given much attention has been oncology. Predictive accuracy is important in this field for three main reasons: 1) it could influence treatment decisions for each individual; 2) it could contribute to the efficiency of health care programs; 3) the accuracy of predictions could assist patients and their families in making personal decisions with regards to their disease (Mackillop & Quirt, 1997).

This area has been researched greatly in regards to survival models, but there has been no work looking into the predictive accuracy of predictions made based on mixture cure models. For this project, we will be exploring predictive accuracy measurements for mixture cure models. We will be using two predictive accuracy measures that are used for survival models and will apply them to mixture cure models. Doing so will allow us to assess the usefulness of mixture cure models for forming predictions as well as the performance of the predictive accuracy measures for mixture cure models.

Chapter 2

Background

Survival Models

There are common models which are often used to model survival rates of a given population. Methods of statistical analysis can be parametric or semi- or non-parametric. Parametric methods are used when the form of the risk (hazard) function is assumed, whereas semi- or non-parametric methods are used when you do not assume complete knowledge about the risk.

A common non-parametric method of estimating $S(y) = P(Y \geq y)$, which is the probability of survival until time y , is the Kaplan-Meier estimate. Given a set of censored and uncensored observations, let t_i denote the i th distinct censored or uncensored observation. The Kaplan-Meier estimate of $S(y)$ is given by:

$$\hat{S}(y) = \prod_{t_i < y} \frac{n_i - d_i}{n_i}$$

where n_i is the number of patients at risk of experiencing the event before time t_i and d_i are the number that experienced the event at time t_i . Based on this estimate, the survival function decreases whenever a subject experiences the event of interest (Tableman & Kim, 2004). Plots of this estimate, known as Kaplan-Meier curves, are often used to demonstrate the survival rates of populations.

The simplest parametric regression models are based on the exponential and Weibull distributions. The survival function for the exponential distribution is:

$$S(y|z) = \exp(-\lambda(z)y),$$

where $\lambda(z)$ is either a constant or an expression based on a set of covariates, z . The survival function for the Weibull distribution is:

$$S(y|z) = \exp(-(\lambda(z)y)^p),$$

where p is a constant and again, $\lambda(z)$ is either a constant or an expression based on a set of covariates, z. Typically, for the exponential and Weibull distributions, $\lambda(z) = \exp(\beta z)$, where β is a set of coefficients for the covariates, z.

The most common semi-parametric regression model is the Cox proportional hazards (Cox PH) model. This model is semi-parametric because the baseline hazard function is not specified. The survival function takes on the formula

$$S(y|z) = S_0(y)^{\exp(\beta z)}.$$

Rather than absolute risk, Cox PH models estimate relative risk and also assume that the relative risk stays constant over time.

Mixture Cure Models

The presence of a cured group in a population is generally indicated by the Kaplan-Meier curves levelling off at a value greater than 0 – meaning that a sizable fraction of the population is right-censored near the end of the observation time (Peng, Dear & Carriere, 2001).

The general form of the survival function of the mixture cure model (the probability of surviving until time y) is:

$$P(Y \geq y) = S(y|x, z) = \pi(z)S_u(y|x) + 1 - \pi(z).$$

In this function, $\pi(z)$ is the proportion of uncured individuals in the population, which can be dependent on a set of covariates z and $S_u(y|x)$ is the survival function for the uncured individuals. The variables x and z are sets of covariates associated with the survival rates of the uncured population and with the cure rate, respectively. The survival rate of the uncured fraction tends to 0 as time approaches infinity (Farewell, 1982). On the other hand, $1 - \pi(z)$ corresponds to the fraction of cured individuals who are considered to not be at risk of experiencing the event of interest (Farewell, 1982).

Predictions Using Survival Models

In the standard prediction procedure, we wish to make a prediction about the response variable Y (for example, time to death) based on a set of given covariates $Z = z$ (for example, age and sex of the patient). We take a random sample $D = \{(y_i, z_i), i = 1, \ldots, n\}$ from the population; D is known as the training data, and predictions for the rest of the population are

made based on this data. There are two main types of predictions: point and distributional predictors (Lawless et al, 2010).

Distributional predictors uses the training data D to obtain $\hat{F}(y|z)$ which estimates $F(y|z) = \Pr(Y \leq y|Z = z)$, the cumulative distribution function for the survival distribution. This would, in turn, give the estimate of the survival function $\hat{S}(y|z)$. $\hat{F}(y|z)$ is of the form $F(y|z;\hat{\theta})$; $\hat{\theta}$ are the estimates of the parameters of the distribution, derived from the training data.

Point predictors give a predicted value $\hat{Y} = \hat{G}(z;\hat{\theta})$ for Y. $\hat{G}(z;\hat{\theta})$ is a value derived based on a set of covariates, Z, from the test data, and a set of parameters $\hat{\theta} = \hat{\theta}(D)$, which are derived from the training data. Some common point predictors are the mean $\hat{\mu}(z)$ and median $\hat{m}(z)$ of $\hat{F}(y|z)$.

An example of the use of distributional and point predictors is in the case of a newly diagnosed cancer patient. The distributional predictor would be the probability of the patient surviving 1 year (or however many years are of interest). The point predictor, on the other hand, would be the estimated survival time of the patient (Lawless et al, 2010).

Prediction Error/Predictive Accuracy

The usefulness of survival models in forming predictions has been highly debated. Predicting the duration of survival for individual patients has been shown to be less accurate than using patient-specific survival probabilities as predictions (Graf et al, 1999).

Loss functions $L(Y, \hat{Y})$ are used to assess the accuracy of a point predictor $\hat{Y}$. Squared error loss $(Y - \hat{Y})^2$ and absolute error loss $|Y - \hat{Y}|$ are two common loss functions. The expected value of these loss functions is equal to the predictive accuracy of point predictors (Lawless et al, 2010).

There is no method for measuring predictive accuracy of predictive distributions that is commonly agreed upon by statisticians (Schumacher, Graf & Gerds, 2003). For distributional predictors, the most common prediction error used is the expected Brier score (EBS). For survival status at time y^*,

$$EBS(y^*) = E[(I(Y \geq y^*) - \hat{S}(y^*|z))^2]$$

where $I(Y \geq y^*)$ is an indicator function for the survival status at time y^* and $\hat{S}(y^*|z)$ is the distributional predictor for the probability of survival to time y^* (Lawless et al, 2010). The expectation is taken with respect to Y. If $EBS(y) = 0$, then the prediction model is considered perfect. If there is no knowledge about the disease of interest, a trivial prediction of $\hat{S}(y|z) = 0.5$ is used for every event – this corresponds to a Brier score of 0.25 (or poor predictive accuracy) (Graf et al, 1999). This formula compares individual vital status at a certain time point with predicted survival probabilities of that time. This score has been shown to be useful and has a meaningful interpretation even when the survival model is mis-specified. The Brier score can be computed at any time point at which there is at least one patient that is event-free and predicted survival probabilities are available for that time. Use of this score allows for the comparison of various regression models used to form predictions as well as assessing the effect of a specific covariate on patient survival time (Schumacher et al, 2003).

One of the main problems with calculating Brier score for survival data is the presence of censored observations. Few works have dealt with the effects of censoring on assessing prediction error (Lawless et al, 2010). Because these patients either drop out of the study, are lost to follow-up, or never experience the event of interest during the course of the study, it is difficult to incorporate them into the Brier score calculation, not knowing what their survival status is at a given time point (Schumacher et al, 2003). One of the options for dealing with censored data would be to simply eliminate those observations from the dataset when calculating the Brier score (Graf et al, 1999). We call this option to be the naïve estimate of the expected Brier score and it will be calculated according to this formula:

$$NBS(y^*) = \frac{1}{m}\sum_{i=1}^{m}\{[I(Y_i \geq y^*) - \hat{S}(y^*|z_i)]^2\}$$

(**Error! Reference source not found.**)

The survival status of observations censored before y^* is unknown, so these observations are eliminated from the dataset. m is then the number of observations which failed before time y^* and all observations which had an event occur or were censored after time y^*.

Another way to deal with censored observations is to incorporate a weighting scheme into the expected Brier score calculation, to account for the loss of information due to censoring. Graf et al developed a weighted Brier score which is based on an inverse probability of censoring

weighting scheme. Graf's weighted Brier score divides the population of interest into three categories. Let Y_i represent the observed failure time for patient i , y^* be a certain fixed time point for which we want to assess the predictive accuracy and δ_i be the censoring indicator, as explained previously. The three categories of patients are as follows (Graf et al, 1999):

Category 1: $Y_i \leq y^*$ and $\delta_i = 1$;

Category 2: $Y_i > y^*$ and $\delta_i = 1$ or $\delta_i = 0$;

Category 3: $Y_i \leq y^*$ and $\delta_i = 0$.

The patients in the first category are the uncensored patients who experience the event of interest before the fixed time y^*. The second category consists of all the censored and uncensored patients who experience the event of interest or are censored after the fixed time y^*. The third category is of all the patients that were censored before y^*.

The inverse probability of censoring weighting scheme is used for Graf`s Brier score. $\hat{G}(y)$ denotes the Kaplan-Meier estimate of the censoring distribution, based on $(Y_i, 1 - \delta_i)$. The weight of $\frac{1}{\hat{G}(Y_i)}$ is applied to patients in category 1. Observations from category 2 get a weight of $\frac{1}{\hat{G}(y^*)}$.

The weighted Brier score proposed by Graf et al is then:

$$BS(y^*) = \frac{1}{n}\sum_{i=1}^{n}\left[\left(0 - \hat{S}(y^*|z_i)\right)^2 I(Y_i \leq y^*, \delta_i = 1)\left(\frac{1}{\hat{G}(Y_i)}\right) + \left(1 - \hat{S}(y^*|z_i)\right)^2 I(Y_i > y^*)\left(\frac{1}{\hat{G}(y^*)}\right)\right] \quad (2)$$

With this estimation of the Brier score, all observations in category 1 contribute the weights $\frac{1}{n\hat{G}(Y_i)}$ and all observations in category 2 contribute the weights $\frac{1}{n\hat{G}(y^*)}$. The indicator functions allow only observations in categories 1 and 2 to contribute their estimated survival probabilities, $\hat{S}(y^*|z_i)$. The event status of patients in category 3 at time y^* is unknown. Therefore, this category makes no contribution to the weighted Brier score. The weights are included in order to compensate for the loss of information due to censoring. If there was no

censoring, the Graf estimate would reduce to the naïve estimate of the Brier score (Graf et al, 1999).

Prediction Error for Mixture Cure Models

The objectives of this work are to:

1) assess the predictive accuracy of mixture cure models compared to regular survival models and
2) compare the two different methods of estimating the expected Brier score, as outlined above.

We will use the two methods for estimating Brier score, (1) and (2), **Error! Reference source not found.**and will apply them to mixture cure models. We will be calculating a naïve estimate of the expected Brier score and will compare this to the Graf Brier score. For the naïve estimate, all observations which were censored before the time of interest will be eliminated and no weighting will be applied to the remaining data. For diseases with cures, a high level of censoring tends to occur near the end of the observation time. Therefore, for higher values of y^*, the naïve estimate of the Brier score will be calculated over fewer observations, without compensating for the loss of information. This could lead to inaccurate assessment of the predictive accuracy of the model, particularly when censoring rates are high (Graf et al, 1999).

Our hypotheses are as follows:

1) we expect that the Graf Brier score will provide a more accurate measure of prediction accuracy as opposed to the naïve estimate, particularly during the times when the censoring rate is high;
2) we expect that using mixture cure models as the predictive distributions as opposed to regular survival models will result in higher predictive accuracy of the models, when cured patients are present.

We will test these hypotheses by conducting two simulation studies in Chapter 3 and Chapter 4. In Chapter 5 we will apply the Brier score calculations to a dataset of 91 leukemia patients who received autologous or allogeneic bone marrow transplants.

Chapter 3

Simulation Study Based on Exponential Distribution

Methods

We will use a simulation study to assess the predictive accuracy of the mixture cure rate model through the Brier scores. We created two sets of data: the training data and test data. Both consist of 500 datasets, each containing 200 patients. One covariate was included, indicating if the patient was in the control or treatment group. In each dataset, 100 patients were in the control group and 100 in the treatment group. A parametric mixture cure model was used to simulate the datasets. The baseline distribution for the uncured group was set as an exponential distribution, corresponding to the following survival function:

$$S_u(y|z) = \exp(-\exp(\beta_0 + \beta_1 z)y)$$

(Error! Reference source not found.)

We set $\beta_0 = 0$ and $\beta_1 = \log\left(\frac{1}{2}\right)$. The covariate $z = 0$ or $z = 1$ indicates if the patient is in the control or treatment group, respectively. The proportion of uncured patients, $\pi(z)$ is assumed to have a logistic form:

$$\log\left(\frac{\pi(z)}{1-\pi(z)}\right) = \gamma_0 + \gamma_1 z$$

(Error! Reference source not found.)

We set $\gamma_0 = 2$ and $\gamma_1 = -1$. This corresponds to a cure rate of 11.9% in the control group and 26.9% in the treatment group.

The overall survival function then has the form:

$$S(y|z) = \frac{\exp(\gamma_0 + \gamma_1 z)}{1 + \exp(\gamma_0 + \gamma_1 z)}\exp(-\exp(\beta_0 + \beta_1 z)y) + 1 - \frac{\exp(\gamma_0 + \gamma_1 z)}{1 + \exp(\gamma_0 + \gamma_1 z)}$$

(Error! Reference source not found.)

We used a uniform distribution from 0 to 25, $U[0,25]$, to generate the censoring times for the dataset.

We use the data from the training data in the mixture cure model log likelihood function:

$$L = \log \prod_{i=1}^{n} [\pi(z_i) f_u(y_i|z_i)]^{\delta_i} [\pi(z_i) S_u(y_i|z_i) + 1 - \pi(z_i)]^{1-\delta_i}$$

(Error! Reference source not found.)

By maximizing this function, we obtain the maximum likelihood estimates: $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\gamma}_0$ and $\hat{\gamma}_1$ for each dataset. We use these estimates to obtain 500 predictive distributions, $\hat{S}(y|z)$ in a mixture cure model exponential form, based on each training dataset.

For comparison, we obtain estimates, $\hat{S}(y|z)$, from the training data based on the assumption that $S(y|z)$ has an exponential distribution. This will also be used to predict survival probabilities of the test data. This predicted survival function takes on the form:

$$\hat{S}(y|z) = \exp(-\exp(\hat{\alpha}_0 + \hat{\alpha}_1 z)\, y)$$

(Error! Reference source not found.)

Naïve Estimate of Brier score

First, we calculated a naïve estimate of the expected Brier score at various time points, y^*, using (1)**Error! Reference source not found.**.

Graf Brier score

We will be using Graf's weighted Brier score (2) for assessing the predictive accuracy of both the exponential mixture cure model distribution and the exponential distribution. This was calculated at the same time points, y^*, that were used to calculate the naïve estimate of the expected Brier score values with no weighting, to allow for comparison between the two values. The censoring weights, $\hat{G}(y)$, were obtained from censoring distributions of the training datasets.

The predictions were made based on the characteristics of patients in the 500 test datasets. The survival probabilities of patients in each test dataset was calculated by $\hat{S}(y|z)$, which is estimated from the training datasets; the value of the covariate, z, comes from the test dataset.

For the naïve Brier score and the Graf Brier score, the predicted probabilities were obtained from the estimated MCM exponential distribution and the estimated exponential distribution for comparison.

Results

The training data had a cure rate of 12.044% in the control group and 26.512% in the treatment group. In addition, the simulation produced an overall censoring rate of 24.076% with 15.794% of the control group and 32.358% of the treatment group being censored in the training data. The Kaplan-Meier curves of the training data for the 500 simulation runs are shown below:

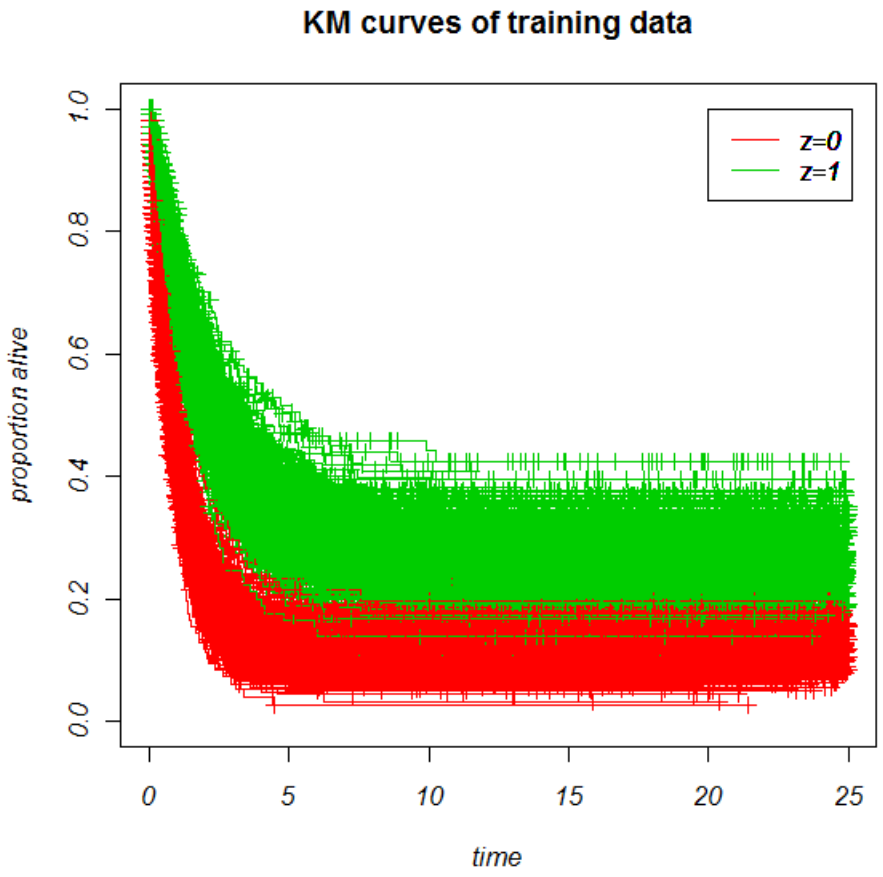

Figure 1: Kaplan-Meier curves of training data for exponential simulation study

The maximum likelihood estimates for the exponential mixture cure model (5**Error! Reference source not found.**), along with their biases and variances, are shown in the table below.

Table 1: Maximum likelihood estimates of parameters of exponential mixture cure model for simulation study

	Average	Bias	Variance
$\hat{\beta}_0$	0.007717957	-0.007717957	0.01286742
$\hat{\beta}_1$	-0.698216	0.005068773	0.03082995
$\hat{\gamma}_0$	2.027913	-0.0279135	0.1251805
$\hat{\gamma}_1$	-0.9878687	-0.01213133	0.1827085

The estimated MCM exponential distributions (based on the maximum likelihood estimates) are shown in the graph below.

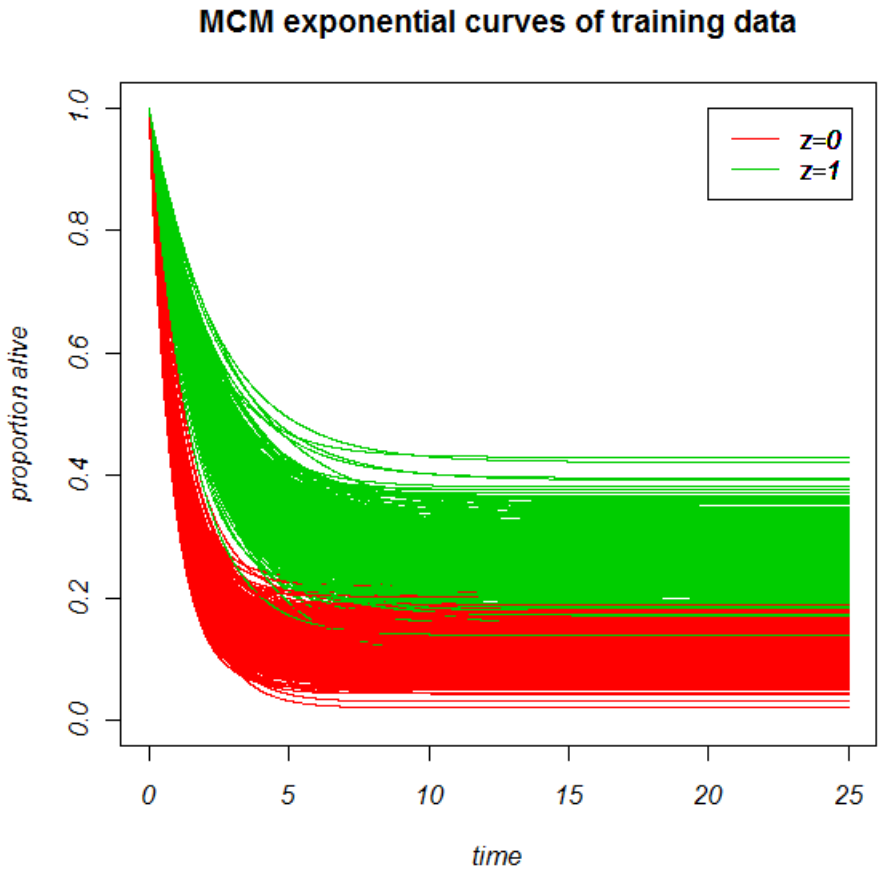

Figure 2: Exponential mixture cure model predictive distributions for exponential simulation study

The estimates for the exponential distribution (7**Error! Reference source not found.**) and their variances are in the following table.

Table 2: Maximum likelihood estimates of exponential model for simulation study

	Average	Variance
$\hat{\alpha}_0$	0.3743178	0.008019858
$\hat{\alpha}_1$	-0.9151651	0.08345706

The Kaplan-Meier curves of the training data and the estimated exponential distribution are shown in the graph below.

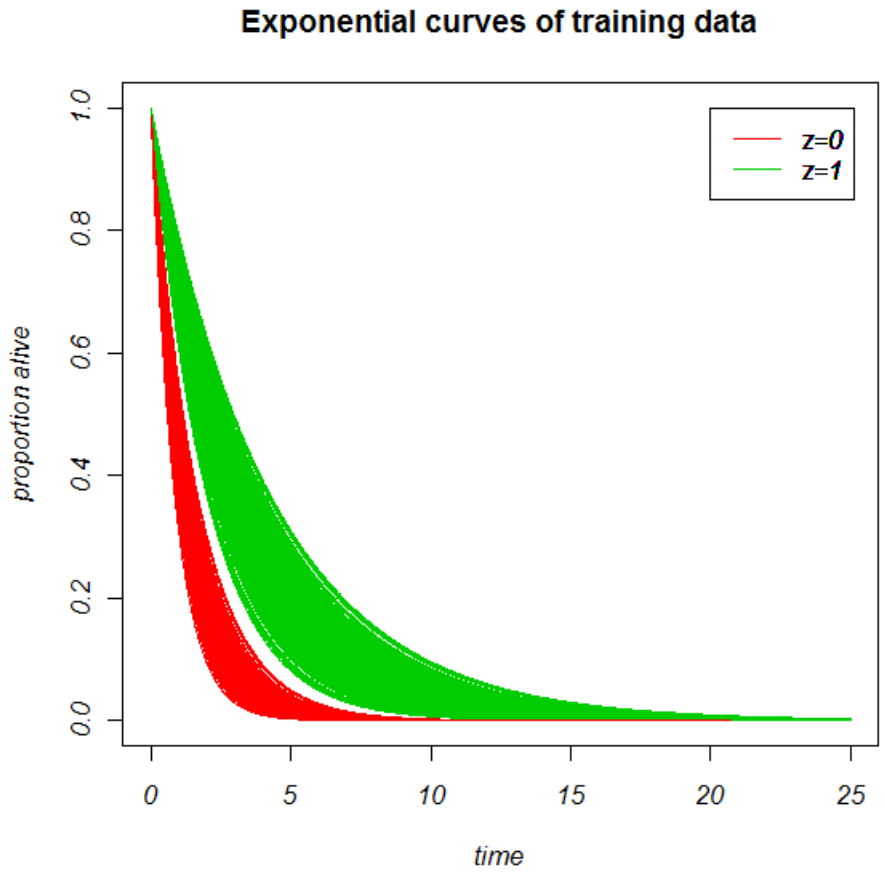

Figure 3: Exponential predictive distributions for exponential simulation study

The naïve expected Brier score and Graf Brier score were calculated for various time points, using the estimated exponential mixture cure model and estimated exponential distribution as predictive probability distributions. The results of the Brier score calculations and their variances are shown in the table below. Each score displayed is the average of the 500 scores that were calculated (one for each training/test dataset pairing).

Table 3: Average estimates of Brier scores for exponential simulation study

Predictive distributions:	**MCM Exponential**		**Exponential**	
y^*	**Average naïve**	**Average Graf**	**Average naïve**	**Average Graf**

	estimate (variance)	**estimate (variance)**	**estimate (variance)**	**estimate (variance)**
1	0.2276933 (0.0001111937)	0.2275078 (0.0001246475)	0.27310758 (0.0006350960)	0.2707867 (0.0006475710)
2	0.2143839 (0.0001367306)	0.2176023 (0.0001835196)	0.27277453 (0.0006508592)	0.2711268 (0.0006444379)
3	0.1863555 (0.0001910286)	0.1932529 (0.0002575364)	0.23278384 (0.0005128121)	0.2338234 (0.0004997375)
4	0.1662064 (0.0002123714)	0.1768016 (0.0003149035)	0.19837043 (0.0003934002)	0.2030017 (0.0003900555)
5	0.1520928 (0.0002182118)	0.1663882 (0.0003635569)	0.17294731 (0.0003280802)	0.1820086 (0.0003645048)
6	0.1426036 (0.000232923)	0.1605292 (0.0004259282)	0.15514110 (0.0002998388)	0.1691836 (0.0004005527)
7	0.1352621 (0.000237035)	0.1565761 (0.0004776012)	0.14229196 (0.0002956682)	0.1616663 (0.0004809716)
8	0.1289458 (0.0002410116)	0.1538611 (0.0005609607)	0.13226417 (0.0003030718)	0.1575897 (0.0006265665)
9	0.1234623 (0.0002459546)	0.1526355 (0.000637775)	0.12416854 (0.0003230699)	0.1565387 (0.0007987697)
10	0.1187095 (0.0002359055)	0.1520455 (0.0007420051)	0.11758397 (0.0003289320)	0.1572713 (0.0010481101)
11	0.1136474 (0.0002365726)	0.1510704 (0.0008238036)	0.11093924 (0.0003462608)	0.1581433 (0.0012845981)
12	0.1089713 (0.0002352198)	0.1504019 (0.0009291851)	0.10482789 (0.0003623071)	0.1596978 (0.0015741527)
13	0.1041333 (0.0002272226)	0.1499839 (0.001096411)	0.09847850 (0.0003573427)	0.1616103 (0.0019708574)
14	0.09957615 (0.0002223198)	0.1499041 (0.001192826)	0.09239299 (0.0003505671)	0.1640538 (0.0022819010)
15	0.09472293 (0.0002178313)	0.1503255 (0.001445600)	0.08565029 (0.0003391631)	0.1669271 (0.0028804293)
16	0.08961762 (0.0002093514)	0.1507063 (0.001770856)	0.07832163 (0.0003259573)	0.1697561 (0.0037159676)
17	0.084981 (0.0002035064)	0.1505321 (0.001866152)	0.07151644 (0.0003079752)	0.1714855 (0.0040472564)
18	0.07994344 (0.0002086972)	0.1522061 (0.002590140)	0.06392788 (0.0002894862)	0.1758052 (0.0056311177)
19	0.07475355 (0.0002057451)	0.1557184 (0.004050653)	0.05593788 (0.0002661373)	0.1824511 (0.0085640356)
20	0.06924136 (0.0001996191)	0.1559889 (0.006345177)	0.04733007 (0.0002306520)	0.1847419 (0.0132493112)
21	0.06386552 (0.0002024436)	0.1590379 (0.008067667)	0.03871215 (0.0002026792)	0.1907229 (0.01738908)
22	0.05847720 (0.0001962808)	0.1632127 (0.012394528)	0.02994575 (0.0001674280)	0.1980461 (0.02867353)

Discussion

The maximum likelihood estimates obtained from the training data for the exponential mixture cure model, as seen in Table 1, had small variances and little bias from the exact values

used to generate the data. As well, the estimated parameters for the exponential distribution, seen in Table 2, that was fit to the data had little variation. These distributions were then used to form predictions on the test data.

We had hypothesized that the estimated Brier scores using the exponential mixture cure model as the predictive model would be lower than the estimated Brier scores based on the exponential model – thus indicating that the mixture cure model has greater predictive accuracy. The estimated Brier scores calculated for times $y^* = 1, 2, \ldots, 22$ (Table 3) are displayed in the graph (Figure 4) below.

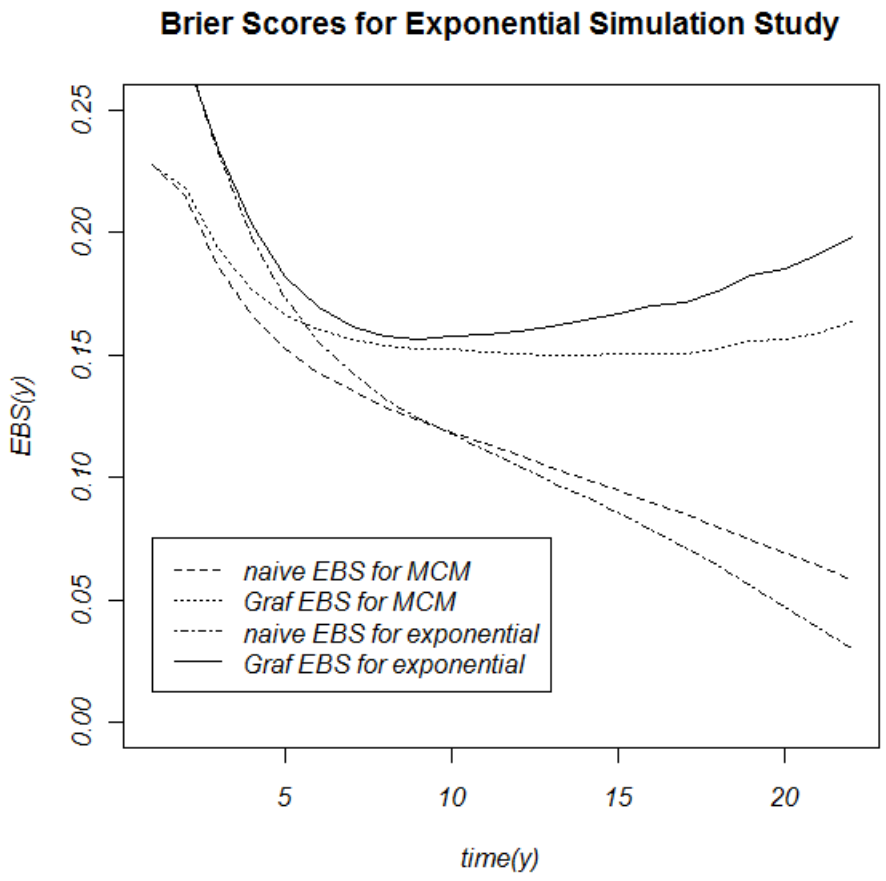

Figure 4: Naïve and Graf estimates of Brier score for exponential simulation study

The naïve estimates suggest that the predictive accuracy of both models is steadily increasing over time. There is no significant difference in the naïve estimates for the two different

models for most of the observation period. During the end of the observation time, however, the naïve estimates suggest that the predictive accuracy of the exponential model is better than the predictive accuracy of the exponential mixture cure model (the “correct” model).

Unlike the naïve estimate, the difference seen between the Graf Brier score estimates during later times indicates that the exponential mixture cure model has greater predictive accuracy. This difference was expected, given that we know the exponential mixture cure model to be “correct,” based on how the data was simulated.

These results suggest that the naïve estimate of the Brier score is less effective at differentiating between different predictive models. Conversely, the Graf Brier score is better at detecting model mis-specification. This was particularly apparent in the tail-end of the distribution, which is where the exponential distribution diverges from the true Kaplan-Meier curves the most. Also, the tail-end of the distribution is where majority of the censoring occurs – this suggests that incorporating a weighting scheme for censoring provides a more appropriate predictive accuracy measure compared to simply eliminating the censored observations.

The following graph displays the Kaplan-Meier curves for the test data of the exponential simulation study.

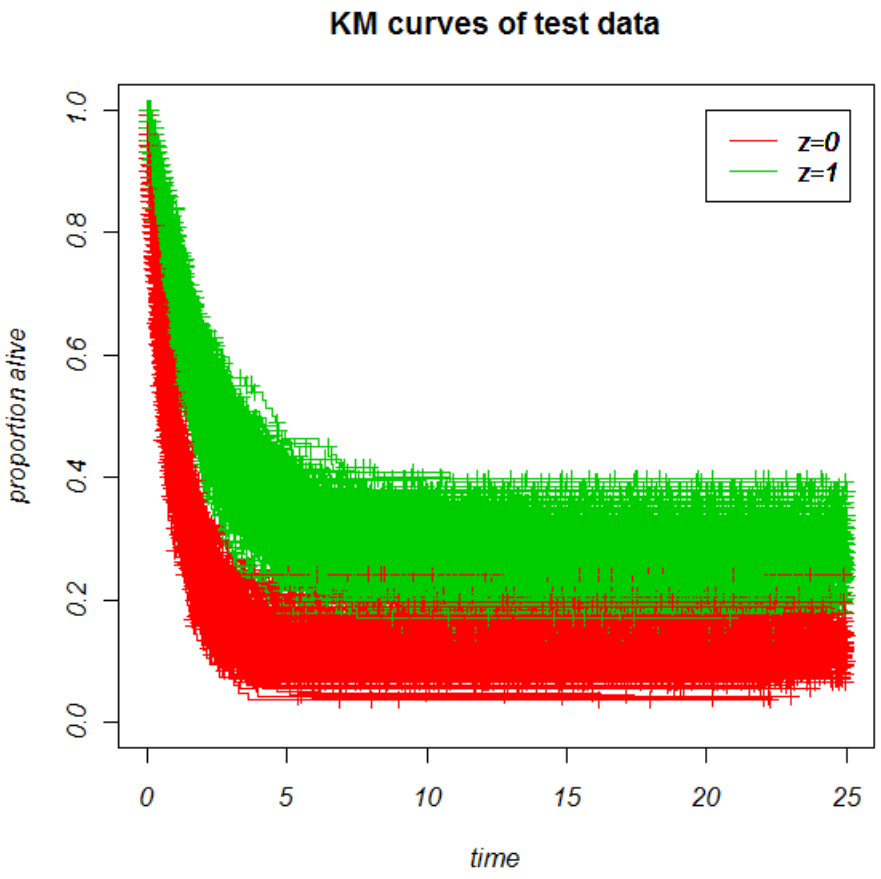

Figure 5: Kaplan-Meier curves of test data of exponential simulation study

These curves demonstrate that the predictive models based on the exponential mixture cure model (Figure 2) are a good fit to the test data, compared to the curves of the exponential model (Figure 3), implying that the mixture cure model would make better predictions – this is illustrated by the Graf estimate of the Brier score. The exponential model curves appear to be a good fit near the beginning of the observation time, when censoring rates are low in the test data compared to event rates. When the event rate decreased and the survival status levelled off in the test data, the exponential curves continue to decrease to zero, whereas the exponential mixture cure model plateaus at values greater than 0. This resulted in the Graf estimate having higher values for the exponential model compared to the values for the mixture cure model at these times.

Chapter 4

Simulation Study Based on Weibull Distribution

Methods

Our second simulation study was run using a similar method as the first one – the difference is that this simulation study used the Weibull mixture cure model, as opposed to the exponential mixture cure model of the first simulation. The training data and test data both consist of 100 datasets, each with 200 patients. Similar to the first study, 1 covariate was included, indicating if the patient was in the control or treatment group. The survival function for the uncured group followed a Weibull regression model:

$$S_u(y|z) = \exp(-[(\exp](\beta_0 + \beta_1 z)\, y)^p)$$

(Error! Reference source not found.)

We set $p = 0.5$, $\beta_0 = \log(2)$ and $\beta_1 = 0$. The covariate $z = 0$ or $z = 1$ indicates if the patient is in the control or treatment group, respectively.

The proportion of uncured patients, $\pi(z)$, and the censoring distribution were derived using the same method as the first simulation, the logistic form (4), as well as the same parameters ($\gamma_0 = 2$ and $\gamma_1 = -1$).

The overall survival function then has the form:

$$S(y|z) = \frac{\exp(\gamma_0 + \gamma_1 z)}{1 + \exp(\gamma_0 + \gamma_1 z)} \exp(-[(\exp](\beta_0 + \beta_1 z)\, y)^p) + 1 - \frac{\exp(\gamma_0 + \gamma_1 z)}{1 + \exp(\gamma_0 + \gamma_1 z)}$$

(Error! Reference source not found.)

For comparison, we obtained estimates, $\hat{S}(y|z)$, from the training data based on the assumption that $S(y|z)$ follows a Weibull distribution:

$$S(y|z) = \exp\left(-\left(\exp\left(-(\alpha_0 + \alpha_1 z)\right) y\right)^{\frac{1}{\sigma}}\right)$$

(Error! Reference source not found.)

We then used the naïve estimate and Graf estimate of the Brier score to assess and compare the predictive accuracy of these two predictive distributions.

Results

The training data had a cure rate of 12.06% in the control group and 27% in the treatment group. In addition, the simulation produced an overall censoring rate of 22.93% with 15.75% of the control group and 30.11% of the treatment group being censored in the training data. The Kaplan-Meier curves of the training data are shown in the following graph.

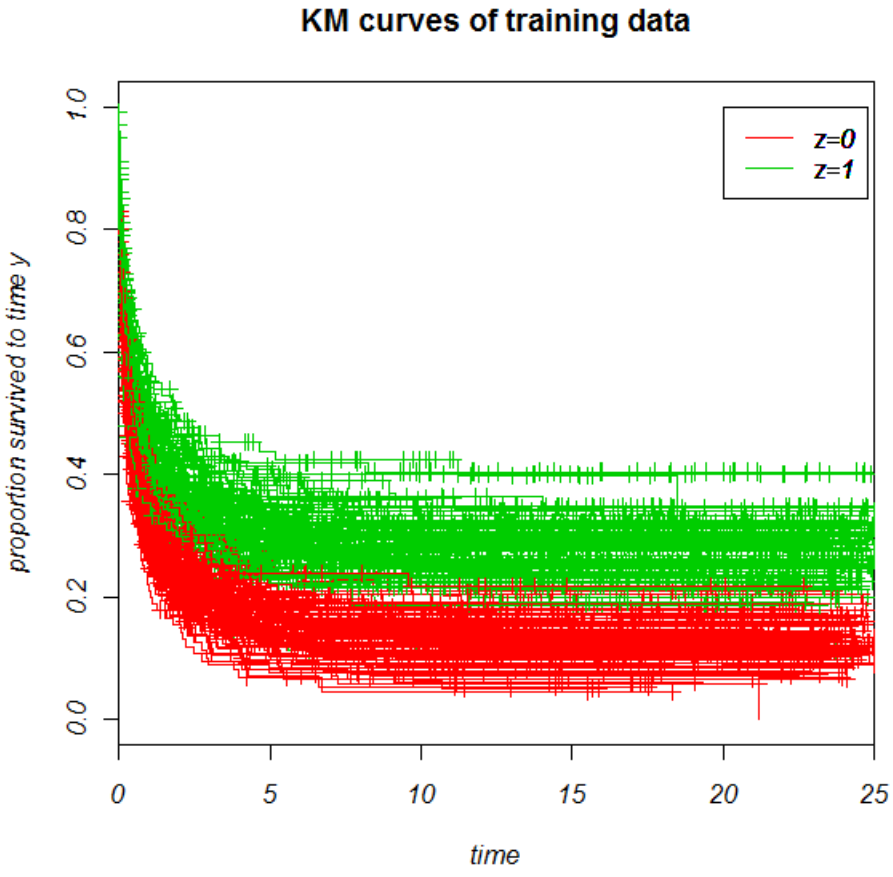

Figure 6: Kaplan-Meier curves of training data for Weibull simulation study

The average of the 100 maximum likelihood estimates of the Weibull mixture cure model (9**Error! Reference source not found.**), based on the training data, are shown in the table below, along with their biases and variances.

Table 4: Maximum likelihood estimates of Weibull mixture cure model for simulation study

	Average	Bias	Variance
$\hat{p}$	0.5035522	-0.003552237	0.00201947
$\hat{\beta}_0$	0.7055685	-0.01242136	0.04698409
$\hat{\beta}_1$	0.3113423	-0.3113423	0.09452987
$\hat{\gamma}_0$	2.016294	-0.01629411	0.1263324
$\hat{\gamma}_1$	-1.060478	0.0604781	0.1281705

The Weibull mixture cure model predictive distributions are shown in the graph below.

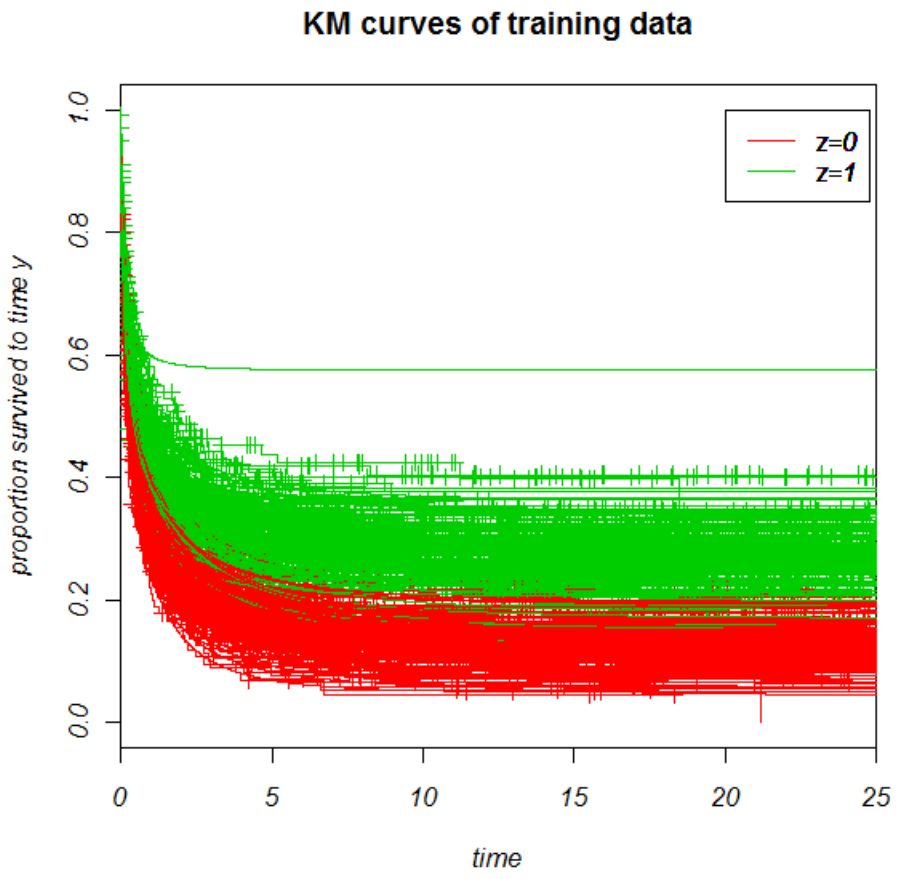

Figure 7: MCM Weibull predictive distributions for Weibull simulation study

The averages of the maximum likelihood estimates of the Weibull regression model (9**Error! Reference source not found.**) fit to the training data are shown in the table below, with their variances.

Table 5: Maximum likelihood estimates of Weibull model for simulation study

	Average	Variance
$\hat{\alpha}_0$	0.1539787	0.1120935
$\hat{\alpha}_1$	1.199192	0.2578648
$\hat{\sigma}$	2.792974	0.02625505

The Weibull regression model predictive distributions are shown in the graphs below.

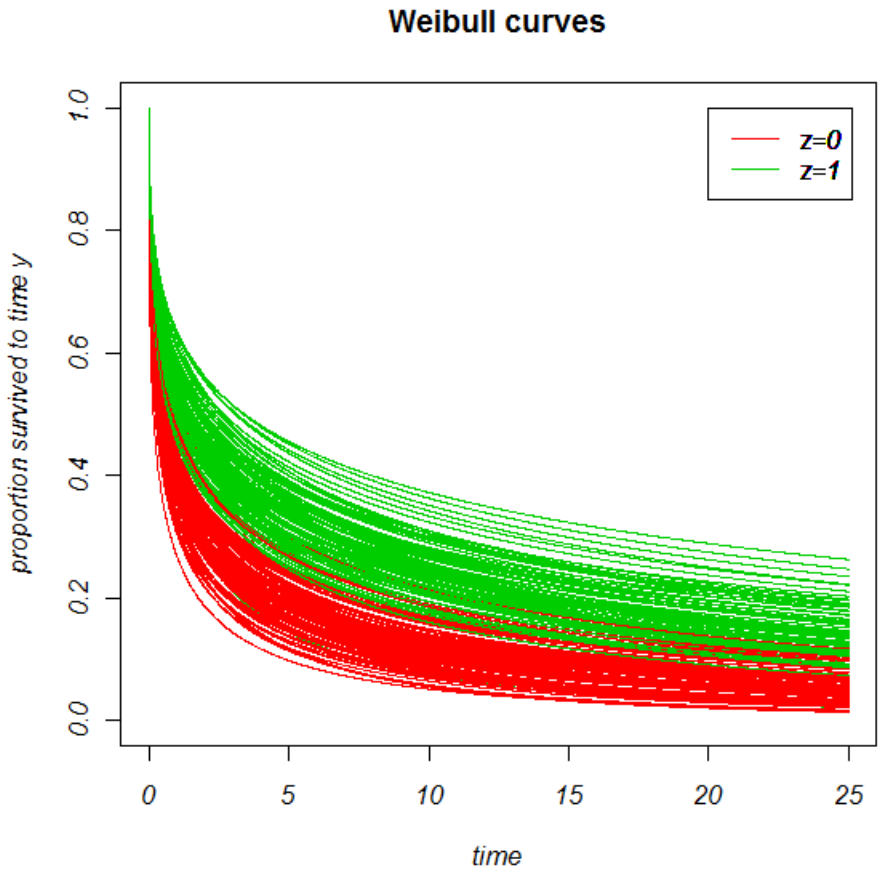

Figure 8: Weibull predictive distributions for Weibull simulation study

The naïve expected Brier score and Graf Brier score were calculated for various time points, using the estimated Weibull mixture cure model and estimated Weibull distribution as predictive probability distributions. The results of the Brier score calculations and their variances are shown in the table on the next page. Each score displayed is the average of the 100 scores that were calculated (one for each training/test dataset pairing).

Table 6: Average estimates of Brier scores for Weibull simulation study

Predictive distributions:	MCM Weibull		Weibull	
y^*	Average naïve estimate (variance)	Average Graf estimate (variance)	Average naïve estimate (variance)	Average Graf estimate (variance)
1	0.234562 (0.0001192478)	0.236602 (0.0001436231)	0.241485 (0.00006136293)	0.242351 (0.00008010483)
2	0.203338 (0.0002776093)	0.208442 (0.0003195601)	0.210697 (0.0001402258)	0.213777 (0.0001710910)
3	0.181563 (0.0003036084)	0.191621 (0.0004271848)	0.187793 (0.0001767553)	0.195301 (0.0002639764)
4	0.166531 (0.0003133733)	0.180286 (0.0004681693)	0.171029 (0.0002019391)	0.182297 (0.0003228286)
5	0.156259 (0.0003110577)	0.173348 (0.0004735165)	0.158918 (0.0002166731)	0.17411 (0.0003578231)
6	0.147641 (0.0003454360)	0.167489 (0.0004955986)	0.148776 (0.0002496958)	0.167598 (0.0003995823)
7	0.14018 (0.0003364451)	0.164054 (0.0005444583)	0.139856 (0.0002710756)	0.163823 (0.0005010500)
8	0.133834 (0.0003461761)	0.161017 (0.0006280170)	0.132128 (0.0002821128)	0.16063 (0.0006098755)
9	0.128455 (0.0003497034)	0.159371 (0.0007355028)	0.125391 (0.0003133122)	0.159123 (0.0007912322)
10	0.122914 (0.0003431191)	0.157855 (0.0008228798)	0.118441 (0.0003142068)	0.157766 (0.0009253452)
11	0.117785 (0.0003526544)	0.157597 (0.0010118910)	0.111874 (0.0003359614)	0.15799 (0.001200903)
12	0.112806 (0.0003397649)	0.155654 (0.0011183573)	0.105553 (0.0003159911)	0.156476 (0.001406673)
13	0.108144 (0.0003450561)	0.154282 (0.0010504669)	0.099399 (0.0003132640)	0.155326 (0.001385401)
14	0.103321 (0.0003376378)	0.155191 (0.0011756487)	0.093027 (0.0002955728)	0.157089 (0.001614147)
15	0.09798 (0.0003360664)	0.153435 (0.0013246478)	0.085842 (0.0002673394)	0.155561 (0.001838589)
16	0.093115 (0.0003492903)	0.156995 (0.0022875886)	0.079364 (0.0002716610)	0.161016 (0.003403603)
17	0.088628 (0.0003593661)	0.157846 (0.0028316907)	0.073207 (0.0002558021)	0.163223 (0.004503762)
18	0.083376 (0.0003891428)	0.157231 (0.0035024135)	0.065934 (0.0002568001)	0.16323 (0.005679514)
19	0.078119 (0.0003992559)	0.156938 (0.0046232527)	0.058569 (0.0002358261)	0.163636 (0.007592186)
20	0.073635 (0.0003979632)	0.161541 (0.0055898550)	0.051939 (0.0002013913)	0.169758 (0.008678515)
21	0.068419	0.157615	0.044228	0.16563

	(0.0003896729)	(0.0057444273)	(0.0001586443)	(0.009133175)

Discussion

The maximum likelihood estimates obtained in the Weibull simulation study for the Weibull mixture cure model, as shown in

Table 4, have small variances and biases, with the exception of the estimate for β_1. β_1 is the coefficient for the covariate, z. When simulating the time to event data, this coefficient was set equal to 0. However, the cure rate was assumed to have a logistic form which was dependent on z; this altered the regression models for $z = 0$ and $z = 1$, thus biasing the z coefficient away from 0. The maximum likelihood estimates for β_1 still had small variance and a small range.

Similar to the first simulation study, we had hypothesized that the estimated Brier scores using the Weibull mixture cure model as the predictive distribution would be lower than the estimated Brier scores based on the Weibull model. The estimated Brier scores calculated for times $y^* = 1,2, \dots, 21$ (Table 6) are displayed in the graph below.

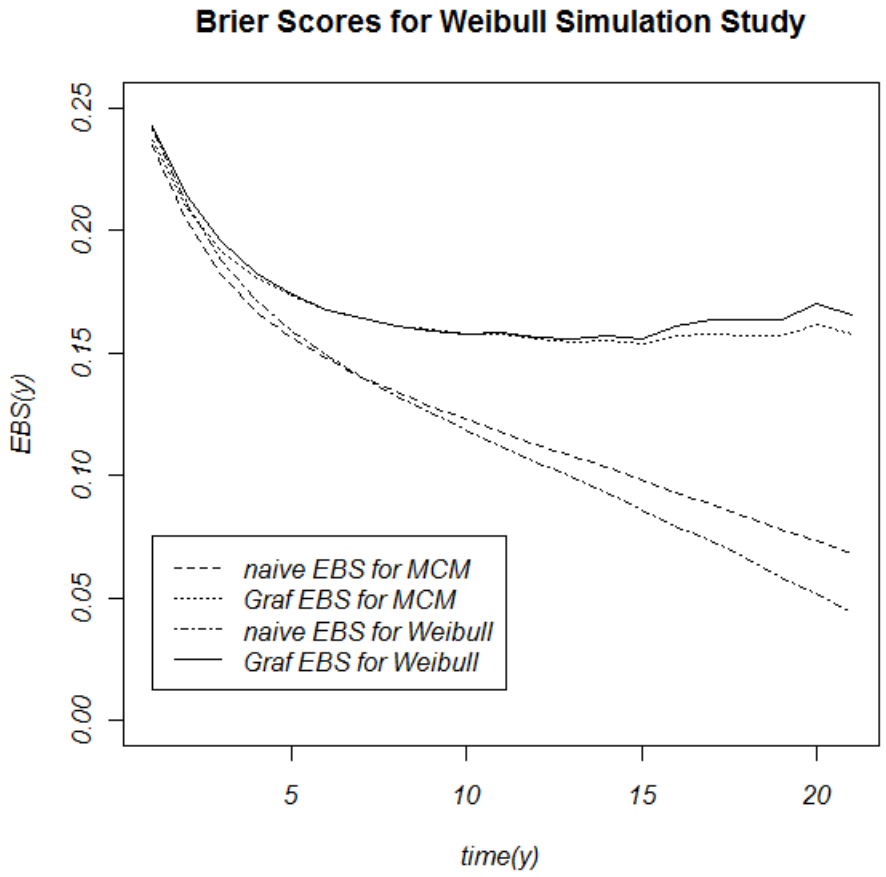

Figure 9: Naïve and Graf estimates of Brier score for Weibull simulation study

Similar to the exponential simulation study, these results suggest that the naïve estimate is less affected by model mis-specification. Also, the Graf Brier score estimates suggest that the Weibull mixture cure model has lower prediction error than the Weibull regression model in times late in the observation period.

For both the naïve and Graf estimates, there is very little difference between the predictive accuracy of the two models at earlier times. During later times, the difference seen between the two models in the naïve estimates indicate that the Weibull model has greater predictive accuracy than the mixture cure model – demonstrating the inaccuracy of the naïve estimate when censoring is high.

There begins to be a difference, although a small one, in the predictive accuracies of the two models based on Graf estimate at times very late in the observation period, although not as large as the difference seen in the exponential simulation study. This suggests that the Weibull regression model could be more flexible than the exponential model when applied to cure data.

The following graph displays the Kaplan-Meier curves for the test data of the Weibull simulation study.

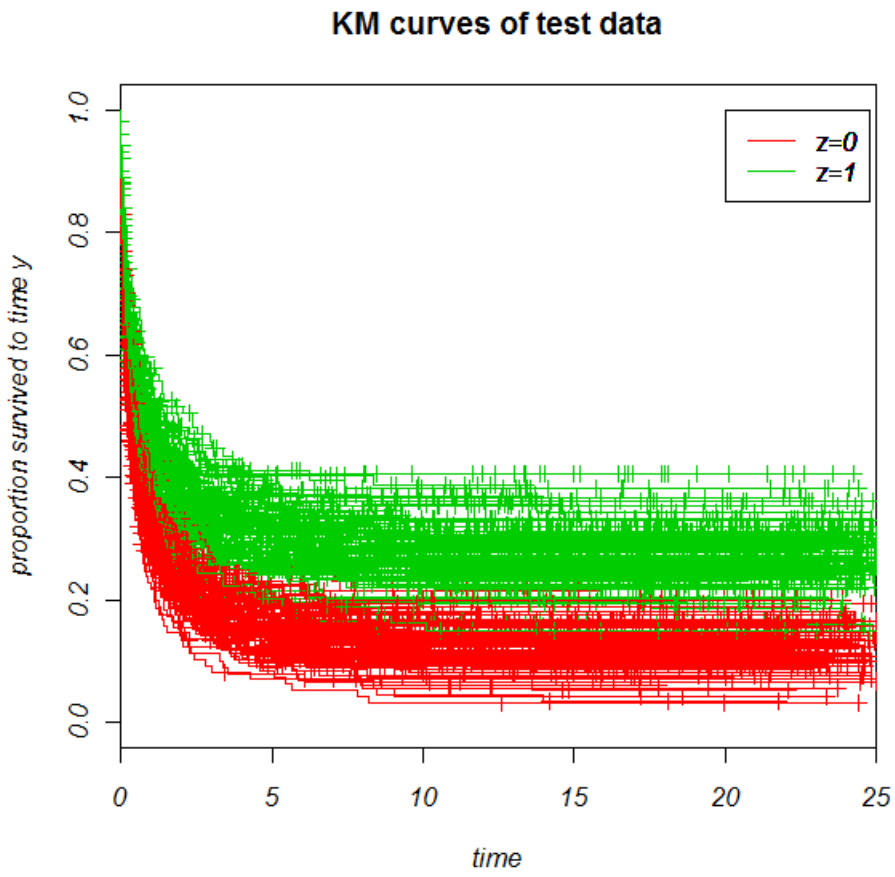

Figure 10: Kaplan-Meier curves of test data of Weibull simulation study compared to MCM Weibull and Weibull predictive models

These curves demonstrate that the predictive models based on the Weibull mixture cure model (Figure 7) are a good fit to the test data, compared to the curves of the Weibull model (Figure 8).

The most noticeable difference between these curves and the curves from the exponential simulation study (Figure 2 and Figure 3) are the predictive models that do not incorporate cure rates. The Weibull regression models shown here decrease to zero at a much slower rate than the exponential models used in the first simulation study. Because of their slow rate at decreasing to zero, they don't diverge from the cure data as much as the exponential model; this results in a smaller difference in predictive accuracy between the Weibull mixture cure model and the Weibull model.

Chapter 5

Application: Bone Marrow Transplant Study

Methods

We obtained a dataset of 91 patients with refractory acute lymphoblastic leukemia. Treating acute lymphoblastic leukemia with chemotherapy has been shown to have high long-term disease-free survival rates. Chemoradiotherapy followed by a bone marrow transplant is used for the patients for whom primary chemotherapy was unsuccessful and for those considered to be at high risk of relapse. Allogeneic bone marrow transplants come from matched sibling donors of the patients. For the patients for whom such matches are unavailable, autologous transplants were used. This study compared the long-term survival rates of patients with refractory acute lymphoblastic leukemia who were treated with bone marrow transplants (Kersey et al, 1987).

In the sample of 91 patients, 46 of these patients received allogeneic bone marrow transplants, while the other 45 received autologous transplants. 22 of the patients were censored during the observation period, resulting in an overall censoring rate of 24.2%. 20% of those receiving autologous transplants were censored, compared to 28.3% of the patients receiving allogeneic transplants. The Kaplan-Meier curves for this dataset are shown below:

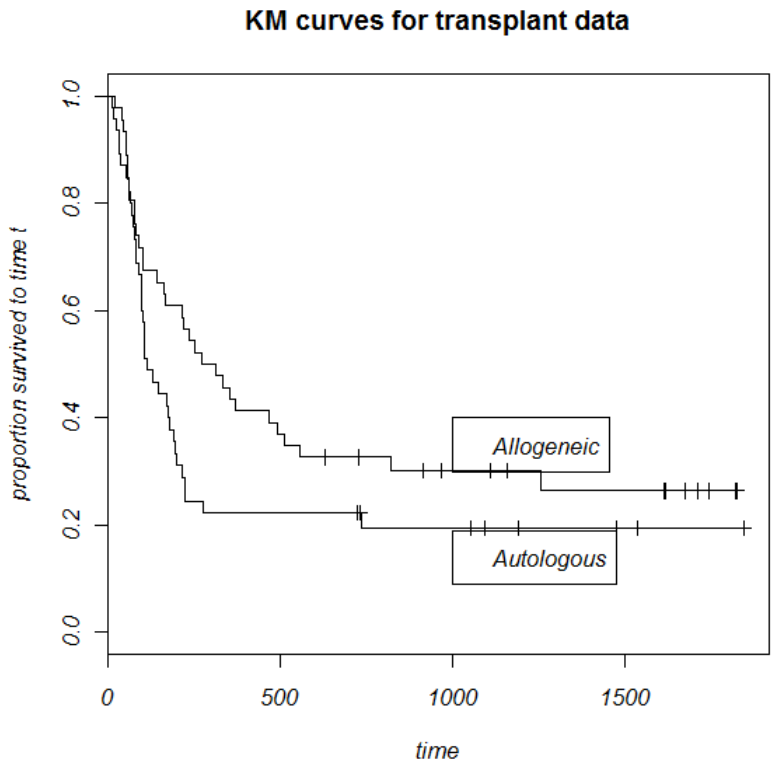

Figure 11: Kaplan-Meier curves for transplant data

The long-term censored survival times seen in these curves are indicative of the presence of cured patients. The levelling-off of the curves for the two treatment groups suggests that the autologous treatment has a cure rate of 19.4% and the cure rate of the allogeneic group is 26.3%. Previous analysis of this dataset suggested that the Weibull mixture cure model survival distribution provides a better fit to this data than the exponential mixture cure model, which was used for the simulated datasets (Peng et al, 2001).

In order to apply the different Brier scores to this dataset, we will use the leave-one-out cross-validation method. We partition the full dataset into 91 subsets, each with one observation from the full dataset excluded. A Weibull mixture cure model is fit to each subset of 90 patients. This estimated model will then be used to predict the survival status of the one observation excluded from the dataset (Simon et al, 2011).

The Weibull mixture cure models were obtained using the method of maximum likelihood estimation that was used for the simulation study. The survival distribution for the

uncured proportion was assumed to have a Weibull form (**Error! Reference source not found.**8).

Similar to the simulation studies, the proportion of uncured patients, $\pi(z)$, has a logistic form. The maximum likelihood estimates for each predictive distribution, $\hat{p}$, $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\gamma}_0$ and $\hat{\gamma}_1$, were obtained from maximizing the log-likelihood function over each leave-one-out dataset.

For comparison, we obtained estimates of the predictive distribution, $\hat{S}(y)$, with the assumption that the data followed a Weibull model (10**Error! Reference source not found.**). We will obtain the maximum likelihood estimates: $\hat{\alpha}_0$, $\hat{\alpha}_1$ and $\hat{\sigma}$ by maximizing the log likelihood function (6).

Results

We fit a Weibull mixture cure model to the 91 transplant datasets (of 90 patients each) (9). The average of the 91 maximum likelihood estimates are shown in the table below.

Table 7: Maximum likelihood estimates of Weibull mixture cure model for transplant study

	Average
$\hat{p}$	1.144919
$\hat{\beta}_0$	-4.928907
$\hat{\beta}_1$	-0.6558068
$\hat{\gamma}_0$	1.388453
$\hat{\gamma}_1$	-0.4173844

These estimates correspond to an approximate cure rate of 20% in the group receiving autologous transplants and 27.5% in the group receiving allogeneic transplants; these values are close to the hypothesized cure rates of the population, as estimated by the proportion of long-term survivors seen in the Kaplan-Meier curves (Figure 11). These estimates were then used to form the 91 predictive distributions which were used in the leave-one-out cross-validation estimation of the Brier score.

The maximum likelihood estimates and their variances for the parameters of the Weibull model (10) that was fit to the dataset are shown in the table below.

Table 8: Maximum likelihood estimates of Weibull model for example study

	Average
$\hat{\alpha}_0$	5.955005
$\hat{\alpha}_1$	0.6588245
$\hat{\sigma}$	1.546439

The following table has the average of the Brier score estimates for both predictive models. The Brier scores were calculated for various time points and their curves are shown below, as well.

Table 9: Average estimates of Brier scores for example study

Distribution used for predicted probabilities:	**MCM Weibull**		**Weibull**	
y^*	**Average naïve expected Brier Score**	**Average Graf Brier Score**	**Average naïve expected Brier Score**	**Average Graf Brier Score**
100	0.2455087	0.2455087	0.2430432	0.2430432
200	0.2370015	0.2370015	0.2555144	0.2555144
300	0.2221809	0.2221809	0.2434555	0.2434555
400	0.2171609	0.2171609	0.2311122	0.2311122
500	0.2120279	0.2120279	0.2202387	0.2202387
600	0.2051937	0.2051937	0.2111825	0.2111825
700	0.2019644	0.2059645	0.2054844	0.2086471
800	0.1808314	0.1967902	0.1836620	0.1975743
900	0.1759551	0.1915863	0.1776155	0.1926291
1000	0.1673226	0.1930550	0.1685035	0.1953927
1100	0.1550274	0.1905181	0.1529464	0.1920552
1200	0.1387005	0.1904314	0.1333691	0.1937806
1300	0.1332415	0.1815851	0.1254303	0.1845829
1400	0.1331783	0.1815979	0.1242324	0.1858785
1500	0.1263409	0.1798587	0.1141484	0.1844388
1600	0.1121874	0.1750778	0.09373566	0.1783260
1700	0.09441106	0.1771561	0.06850157	0.1837663
1800	0.08172313	0.1807896	0.04921626	0.1913718

Discussion

We estimated $S(y|z)$ for the transplant patient dataset under the assumption that $S(y|z)$ follows a Weibull mixture cure model and then a simple Weibull distribution, similar to the Weibull simulation study.

Based on the results of our simulation studies, we expected that naïve estimates of the Brier score would overestimate the predictive accuracy of the models, compared to the Graf estimate. As well, based on the Weibull simulation study, we did not expect to see a significant difference between the Brier scores of the two different predictive distributions. The estimated Brier scores calculated for times $y^* = 100, 200, \ldots, 1800$ (Table 9) are displayed in the graph below.

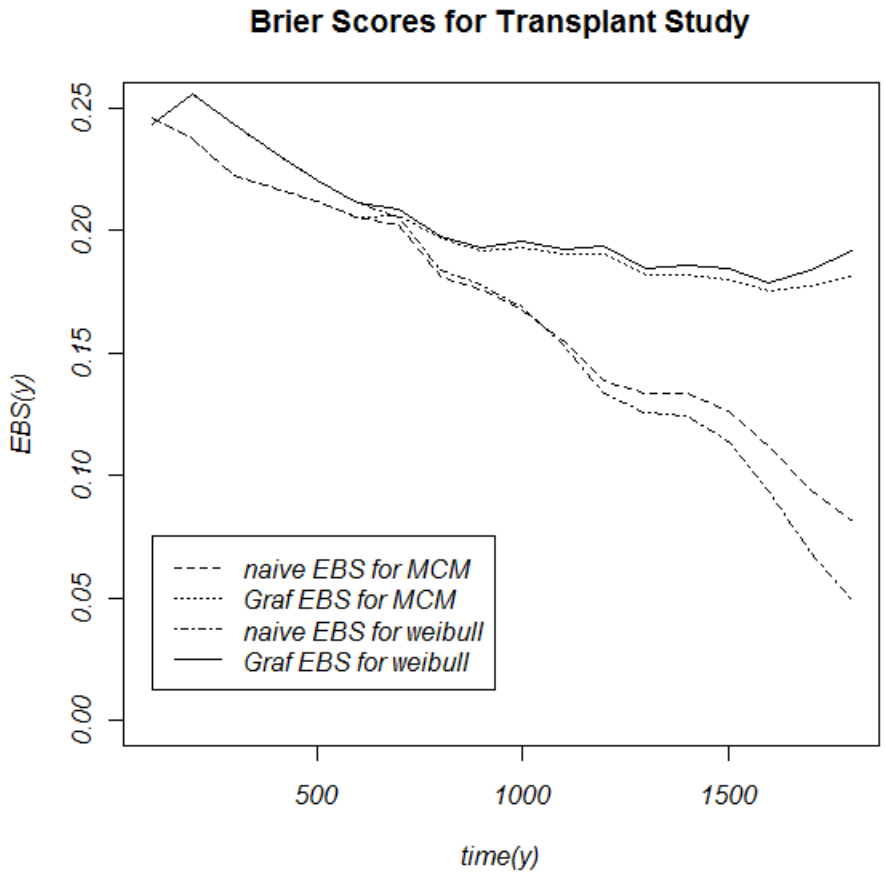

Figure 12: Naïve and Graf estimates of Brier score for transplant study

Similar to the simulation studies, the naïve estimates suggest that the predictive accuracy of both models increases over time. The Graf estimates also show a slight increase in predictive accuracy over time, though not as much as the naïve estimates. However, the scores do not differentiate well between the mixture cure model and the survival model which does not incorporate cure rates, which is similar to what was seen in the Weibull simulation study.

An interesting point to note for this dataset is that the naïve estimates and the Graf estimates are equal for each model from $y^* = 100, 200, \ldots, 600$. This happens because the first case of censoring occurs at $y = 628$, but several events occur before this time. So, when calculating the naïve estimate, no observations were eliminated from the dataset for $y^* < 628$. As well, the weights for the Graf Brier score are based on the Kaplan-Meier estimates of the

inverse censoring distribution of the dataset. Therefore, for $y^* < 628$, there were no observations in category 3 ($Y_i \leq y^*$ and $\delta_i = 0$) and the weights for all the observations was equal to 1, making this score equivalent to the naïve estimate.

The Kaplan-Meier curves of the transplant dataset are illustrated below, along with the estimated predictive models based on the Weibull mixture cure model (9**Error! Reference source not found.**) and the Weibull model (10).

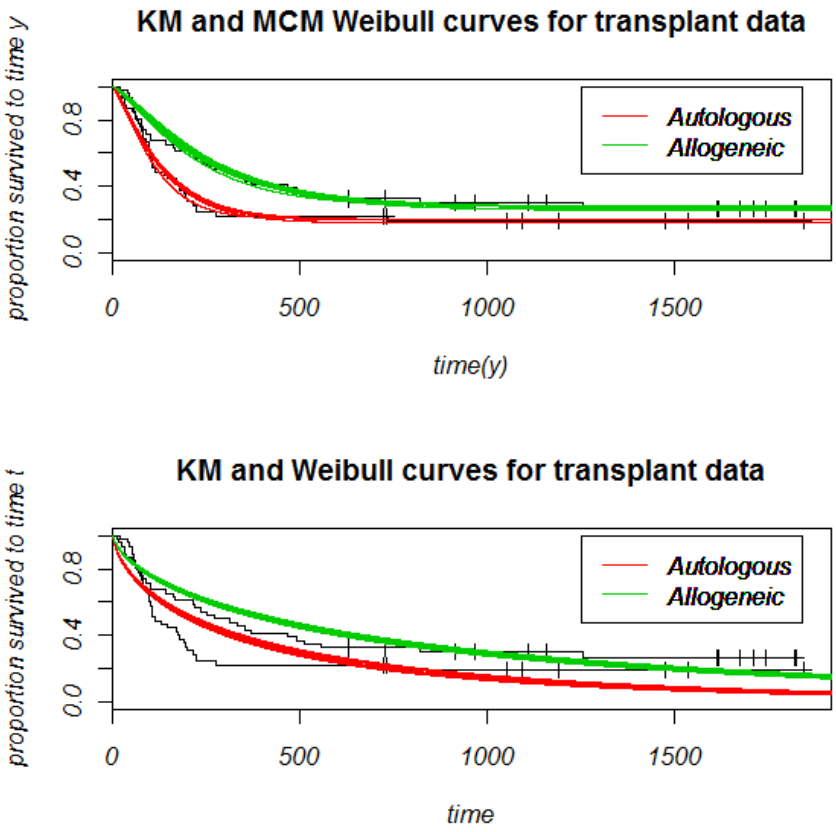

Figure 13: Kaplan-Meier curves of transplant data with MCM Weibull and Weibull predictive models

Upon visual inspection, the Weibull mixture cure model appears to fit the data well for both groups of patients. The Weibull model seems to have a slightly worse fit; however, this model also has a long tail-end, decreasing to 0 well beyond the observation time period. This model differs from the Weibull mixture cure model mostly at the beginning of the observation

period, as shown by Figure 7. This was reflected by the Brier scores calculated for $y^* < 600$ being lower for the mixture cure model.

Chapter 6

Conclusions and Future Directions

We had hypothesized that the Graf estimate of the Brier score would provide a more accurate assessment of the predictive accuracy of the two models. The results from both the simulation studies and the transplant dataset support this hypothesis. Particularly in the exponential simulation study, the Graf estimate is affected by using the incorrect model as the predictive distribution and better able at detecting which model is incorrect, compared to the naïve estimate. The naïve estimate simply ignores the effects of censoring on the predictive accuracy while the Graf estimate, on the other hand, compensates for the loss of information due to censoring by introducing a weighting scheme to the Brier score (Graf et al, 1999). The datasets we dealt with in all three studies had high censoring rates, resulting in the large difference between the naïve and Graf estimates.

We had also expected that the Brier score estimates calculated when using mixture cure models as the predictive distribution would be lower than the estimates calculated using regular survival models. This hypothesis was supported by our results of the exponential simulation study, mostly for later times. Our results for the analysis of the Weibull simulation study and the transplant dataset, however, indicate that using a mixture cure model as opposed to a survival model has little improvement in predictive accuracy.

The Weibull regression models used in the simulation study and the example study decrease to 0 less rapidly than the exponential regression model used in the first simulation, which means they could provide a more adequate fit to data involving cured patients than the exponential model; because it takes longer to decrease to 0, it might not diverge from cure data during the “plateau” as greatly as the exponential model would. Future research could involve extending the simulation study to include more complicated regression models, such as log-normal, gamma, etc.

As well, semi-parametric methods for modeling survival rates, such as the Cox proportional hazards model, could be used. Along with using different survival regression

models, there are also different cure models which can be used. The mixture cure model is one way of modelling survival rates of diseases with cures. The proportional hazards cure model is a semi-parametric method; future research could include comparisons of the predictive accuracy of the different types of cure models.

Another area that could be researched further would be datasets with varying cure rates, as well as the censoring rates; perhaps there would be greater differences between the fits of the mixture cure models and their survival model counterparts when the cure rate is lower or higher. As well, the simulation studies and example study we conducted included only one covariate in the analysis. Further research could extend the simulation studies to include more than one covariate.

Finally, another topic that was not addressed here was the decomposition of the expected Brier score. The expected Brier score formulas can be decomposed to give values for "resolution" and "calibration" – different aspects of predictive accuracy. Resolution is defined as the model's ability to discriminate between when the event will and will not occur. Calibration is the measure of how well the estimated predicted event-free probabilities, $\hat{S}(y|z)$,correspond to the true event-free probabilities, $S(y|z)$. Future research on the predictive accuracy of mixture cure models could look into changes in specific parts of the Brier score (Graf et al, 1999) (Yates, 1982).

Bibliography

Farewell, V. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics , 38*, 1041-1046.

Graf, E., Schmoor, C., Sauerbrei, W., & Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine , 18*, 2529-2545.

Kersey, J.H., Weisdore, D., Nesbit, M., LeBien, T.W., Woods, W.G., McGlave, P.B., Kim, T., Vallera, D.A., Goldman, A., Bostrom, B., Hurd, D. & Ramsay, N.K.C. (1987). Comparison of autologous and allogeneic bone marrow transplantation for treatment of high-risk refractory acute lymphoblastic leukemia. *The New England Journal of Medicine, 317(8), 461-467.*

Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data.* Hoboken, New Jersey: John Wiley & Sons, Inc.

Lawless, J. F., & Yuan, Y. (2010). Estimation of prediction error for survival models. *Statistics in Medicine , 29*, 262-274.

Mackillop, W. J., & Quirt, C. F. (1997). Measuring the accuracy of prognostic judgments in oncology. *Journal of Clinical Epidemiology , 50* (1), 21-29.

Peng, Y., Dear, K. B., & Carriere, K. (2001). Testing for the presence of cured patients: a simulation study. *Statistics in Medicine* (20), 1783-1796.

Peng, Y., Taylor, J. M., & Yu, B. (2007). A marginal regression model for multivariate failure time data with a surviving cure fraction. *Lifetime Data Analysis , 13*, 351-369.

Schumacher, M., Graf, E., & Gerds, T. (2003). How to assess prognostic models for survival data: a case study in oncology. *Methods Inf Med, 5, 564-571*.

Simon, R., Subramanian, J., Li, M.-C., & Menezes, S. (2011, January 7). Using cross-validation to evaluate predictive accuracy of survival risk classifiers based on high-dimensional data. *Briefings in Bioinformatics* , 1-12.

Tableman, M., & Kim, J. S. (2004). *Survival Analysis Using S: Analysis of Time-to-Event Data.* Washington DC: Chapman & Hall/CRC.

Yates, J. F. (1982). External Correspondence: Decompositions of the Mean Probability Score. *Organizational Behaviour and Human Performance* (30), 132-156.

Yu, B., & Peng, Y. (2008). Mixture cure models for multivariate survival data. *Computational Statistics and Data Analysis , 52*, 1524-1532.