12-15-2014

Mixture Cure Models: Simulation Comparisons of Methods in R and SAS

Myra Robinson
University of South Carolina - Columbia

MIXTURE CURE MODELS: SIMULATION COMPARISONS OF METHODS IN R AND SAS

by

Myra Robinson

Bachelor of Science
University of South Carolina, 2012

Submitted in Partial Fulfillment of the Requirements

For the Degree of Master of Science in Public Health in

Biostatistics

The Norman J. Arnold School of Public Health

University of South Carolina

2014

Accepted by:

Jiajia Zhang, Director of Thesis

Alexander McLain, Reader

James Symanowski, Reader

Lacy Ford, Vice Provost and Dean of Graduate Studies

ACKNOWLEDGEMENTS

My thesis committee, in addition to my friends and family, has provided significant support throughout the course of this thesis. I would like to sincerely thank Dr. Jiajia Zhang, my thesis advisor, for all of her guidance and advice. Without her support and suggestions, I would not have explored the idea of mixture cure models and would not have gained the deeper understanding of them and their applications. Additionally, I am grateful for the advice and comments from my committee members Alexander McLain and Jim Symanowski.

I would also like to acknowledge Levine Cancer Institute and Jim Symanowski specifically for providing me with the internship opportunity last summer during which I was introduced to the sarcoma data set that ultimately motivated my thesis topic.

Finally, I could not have made it through the long hours of initial research, data collection, writing, and preparation without the support of my close friends and family. They have been there for me throughout the entire endeavor and I am indebted to them for their unwavering support.

ABSTRACT

Typical survival methods have the assumption that every subject will eventually experience the event of interest, given enough follow-up time. However, there are some occasions in which a proportion of the population of interest will never experience the event of interest. Therefore, the incorporation of a "cure" fraction in a statistical model is necessary. In this thesis, I comprehensively evaluate mixture cure models in two different statistical software programs: the smcure package in R and the PSPMCM macro in SAS. Extensive simulation studies in R and SAS allow evaluation of the performance of these two models. An additional aspect of this thesis involves application of the mixture cure models in R and SAS to a new real data set of soft tissue sarcoma patients. The results from the models fitted to the sarcoma data set in R and in SAS will then be compared.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

In the field of biostatistics, the analysis of survival data is often the goal of studies. The methods currently available to do this analysis are numerous and varied. Some of most commonly used methods in survival analysis include the proportional hazards (PH) model and the accelerated failure time (AFT) model. Both of these methods assume that every subject will eventually experience the event of interest, given enough follow-up time. There are some instances, especially with the advancements in modern medicine, in which a proportion of the population of interest are "cured" and will therefore never experience the event of interest. This situation motivates the incorporation of a cure fraction in a statistical model in order to analyze the ability of a certain treatment to cure a disease of interest. Once that model is defined, the next step is to develop procedures to fit the model to study datasets by utilizing popular statistical software. This chapter will review the mixture cure model and explain its implementation in two common statistical software programs, R and SAS. Additionally, a motivating data set for this study will be introduced and evaluated using the more typical survival analysis tools.

1.1 Mixture Cure Models

As previously stated, the motivation behind mixture cure modeling is the desire to address the situations in which there are cured proportions of individuals and the resulting consequence that those individuals will never experience the event of interest.[1] This led to exploration into cure rate estimation and development of the first mixture cure models by Boag, Berkson and Gage, and Haybittle.[2,3,4] From those initially developed models, various studies have proposed and assessed parametric and semiparametric mixture cure models.[5] Several authors have studied the parametric approach to mixture cure models [1,6,7] however, semiparametric models, are often of greater interest than parametric models since the parametric assumption can be hard to meet. Therefore, many studies more recently have explored modeling and estimation with semiparametric mixture cure models.[8,9,10,11,12]

To start, we give the expression for the mixture cure model. Let T denote the failure time for the event of interest and let Y be the indicator of an individual's susceptibility to the event of interest (Y=1 for susceptible, while Y=0 for not susceptible). Also, define $1-\pi(\boldsymbol{z})$ as the probability of being cured given the vector of covariates $\boldsymbol{z}$. $S(t|Y=1,\boldsymbol{x})$ gives the survival probability for susceptible, uncured patients at time t, given a certain covariate vector $\boldsymbol{x}$.

Covariate vectors $\boldsymbol{x}$ and $\boldsymbol{z}$ may affect the survival and the cure function, respectively. The expression for the mixture cure model is as follows:

$$S_{pop}(t|\boldsymbol{x},\boldsymbol{z}) = \pi(\boldsymbol{z})S(t|Y=1,\boldsymbol{x}) + \big(1-\pi(\boldsymbol{z})\big) \tag{1}$$

where $S_{pop}(t|\boldsymbol{x},\boldsymbol{z})$ is the unconditional survival function of T for the entire population. Here, $S(t|Y=1,\boldsymbol{x})$ is defined as the latency and $\pi(\boldsymbol{z})$ is defined as the incidence.

The modeling strategy for the mixture cure model involves separately modeling the cure proportion and the survival distribution of the uncured patients. Starting with the incidence portion of the model, the effects of the covariate vector **z** on the cure proportion is typically modeled using a logit link function

$$\pi(\boldsymbol{z}) = \frac{e^{\boldsymbol{bz}}}{1+e^{\boldsymbol{bz}}}$$

where **b** is a vector of unknown parameters associated with the covariate vector **z**. However, other link functions can be used as well, including the probit link

$$\phi^{-1}\big(\pi(\boldsymbol{z})\big) = \boldsymbol{bz}$$

where ϕ denotes the standard normal cumulative distribution function, and the complementary log-log link

$$\log\big(-log\big(1-\pi(\boldsymbol{z})\big)\big) = \boldsymbol{bz}$$

The latency portion of the model can be defined to be the proportional hazards (PH) model or the accelerated failure time (AFT) model. Let $S_0(t)$ be the baseline survival function for uncured (susceptible) individuals. When $S(t|Y = 1, \boldsymbol{x}) = S_0(t)^{e^{\boldsymbol{\beta x}}}$, the proportional hazards mixture cure (PHMC) model is selected and when $S(t|Y = 1, \boldsymbol{x}) = S_0\left(te^{\boldsymbol{\beta x}}\right)$, the accelerated failure time mixture cure (AFTMC) model is selected.

Computational Methods

Likelihood

The methods of estimating parameters vary between the SAS macro and the R package, but the general procedure involves maximizing the likelihood. Therefore, we will start by expressing the full likelihood function for the observed data. Let observed data for the ith individual, i = 1, …, n, have the form $\mathbf{O} = (t_i, \delta_i, \boldsymbol{z_i}, \boldsymbol{x_i})$. The observed survival time for the ith individual is t_i, while the censoring indicator, δ_i, is 1 if the event occurred and 0 if the individual is censored. Let $\boldsymbol{\Theta} = (\mathbf{b}, \boldsymbol{\beta}, S_0(t))$, the unknown parameters to be estimated, and let Y be the indicator of susceptibility as previously described, where Y = 1 if the individual will eventually experience the event of interest and Y = 0 if the individual will not. Given a vector $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ and the observed data **O**, the likelihood function is expressed in Equation 2.

$$L(\boldsymbol{b},\boldsymbol{\beta}) = \prod_{i=1}^{n}[1-\pi_i(\boldsymbol{z_i})]^{1-y_i}\,\pi_i(\boldsymbol{z_i})^{y_i} h(t_i|Y=1,\boldsymbol{x_i})^{\delta_i y_i} S(t_i|Y=1,\boldsymbol{x_i})^{y_i} \tag{2}$$

where h(·) is the hazard function that corresponds to S(·). The expression for the logarithm of this complete likelihood function is shown in two parts, Equations 3 and 4.

Incidence log likelihood

$$l_{c1}(\boldsymbol{b};O,y) = \sum_{i=1}^{n} y_i \log[\pi(z_i)] + (1-y_i)\log[1-\pi(z_i)] \tag{3}$$

Latency log likelihood

$$l_{c2}(\boldsymbol{\beta};O,y) = \sum_{i=1}^{n} y_i\,\delta_i \log[h(t_i|Y=1,x_i)] + y_i \log[S(t_i|Y=1,x_i)] \tag{4}$$

Estimation Procedures

Having expressed the full likelihood and log-likelihood functions, we will now explore the different estimation procedures. The Expectation-Maximization (EM) algorithm is utilized because of the incorporation of the latent variable, **y**. This unobservable variable is replaced by the expectation in the EM algorithm. The differences between R and SAS arise in the M-step of the EM algorithm. Obviously, the PHMC model utilizes semiparametric estimation consistently between R and SAS. However, for the AFTMC model, the SAS macro utilizes parametric optimization methods to obtain maximum likelihood estimates[5] while

the R package utilizes the rank-based estimation method proposed by Zhang and Peng.[12]

a. Semiparametric PHMC Model

The conditional expectation of the complete log-likelihood, with respect to the y_i's and given the observed data, **O,** and current estimates of the parameters, $\boldsymbol{\Theta}^{(m)}$**,** is computed with the expectation step (E-step) of the EM algorithm. The conditional expectation of y_i is sufficient as the complete log-likelihoods in equations (3) and (4) are linear functions of y_i. Therefore, the expectation of y_i is shown in equation 5.

$$w_i^{(m)} = E\left(y_i \middle| \boldsymbol{O}, \boldsymbol{\Theta}^{(m)}\right) = \delta_i + (1-\delta_i)\frac{\pi(z_i)S(t_i|Y=1,x_i)}{1-\pi(z_i)+\pi(z_i)S(t_i|Y=1,x_i)}\Bigg|_{(\boldsymbol{O},\boldsymbol{\Theta}^{(m)})} \tag{5}$$

For those who have experienced the event, $\delta_i = 1$, the expectation is 1, and for those censored observations, $\delta_i = 0$, the expectation is the probability of patients being uncured. The second part of equation 5 is interpreted as the ith individual's conditional probability of remaining uncured at the m[th] iteration of the algorithm. This expression is used in the expectations of equations 3 and 4, shown below in equations 6 and 7, respectively.

$$E(l_{c1}) = \sum_{i=1}^{n} w_i^{(m)} \log[\pi(z_i)] + \left(1 - w_i^{(m)}\right)\log[1-\pi(z_i)] \tag{6}$$

$$E(l_{c2}) = \sum_{i=1}^{n} \delta_i \log\left[w_i^{(m)} h(t_i|Y=1,x_i)\right] + w_i^{(m)} \log[S(t_i|Y=1,x_i)] \tag{7}$$

Equations 6 and 7 are maximized separately with respect to the unknown parameters in the maximization step (M-step) of the EM algorithm. For the incidence part shown in equation 6, in R, the 'link' option in the 'glm' function is utilized to estimate the parameters whereas in SAS, PROC LOGISTIC can be used to maximize the equation. In order to estimate $\boldsymbol{\beta}$ from equation 7 without having to specify the baseline hazard, a previously proposed partial likelihood type method is utilized.[10,13] Equation 8 shows the estimating equation for this.

$$log \prod_{i=1}^{n} \left[h_0(t_i) \exp\left(\boldsymbol{\beta x_i} + \log\left(w_i^{(m)}\right)\right)\right]^{\delta_i} S_0(t_i)^{\exp(\boldsymbol{\beta x_i} + \log\left(w_i^{(m)}\right))} \tag{8}$$

Equation 8 is similar to the standard PH model log-likelihood function, with an additional offset variable, $\log\left(w_i^{(m)}\right)$. This similarity allows use of the 'coxph' function in R and PROC PHREG in SAS to estimate the parameters in equation 7. Details on this estimation are available in other studies. [10,12,15]

To proceed back to the E-step, the estimated survival function must be updated. Let $t_{(1)} < t_{(2)} < \cdots < t_{(k)}$ denote the distinct uncensored failure times. Also, let $d_{t_{(j)}}$ be the number of events and $R(t_{(j)})$ be the risk at time $t_{(j)}$. Equation 9 gives a Breslow-type estimator for $S_0(t|Y=1)$.

$$\hat{S}_0(t|Y=1) = \exp\left(- \sum_{j:t_{(j)} \le t} \frac{d_{t_{(j)}}}{\sum_{i \in R(t_{(j)})} w_i^{(m)} e^{\hat{\beta}_m x_i}} \right) \tag{9}$$

Because of the cure proportion, $\hat{S}_0(t|Y=1)$ may not approach 0 as $t \to \infty$. Therefore, to avoid identifiability problems we can set $\hat{S}_0(t|Y=1) = 0$ for $t > t_{(k)}$, where $t_{(k)}$ is the last observed failure time.[4] The estimated survival function then becomes $\hat{S}(t|Y=1) = \hat{S}_0(t|Y=1)^{\exp(\hat{\boldsymbol{\beta}}\boldsymbol{x})}$. Note that this zero-tail constraint makes the assumption that individuals with survival times greater than the last observed failure time are all non-susceptible.

b. Semiparametric AFTMC Model

For AFTMC parameter estimation, the incidence part in equation 6 is estimated the same way as described for the PHMC model. The latency part from equation 7 must be demonstrated separately. The rank-based estimation method proposed by Zhang and Peng is used to estimate $\boldsymbol{\beta}$ in the M-Step.[12] This method involves rewriting equation 7 as a log-likelihood function for a standard semiparametric AFT model that contains an additional constant term, $w_i^{(m)}$. This is shown in equation 10.

$$log \prod_{i=1}^{n} [w_i^{(m)} h(\log(t_i) - \boldsymbol{\beta}\boldsymbol{x_i})]^{\delta_i} [S(\log(t_i) - \boldsymbol{\beta}\boldsymbol{x_i})^{w_i^{(m)}}] \tag{10}$$

Estimation of $\boldsymbol{\beta}$ in the M-step is then accomplished by the previously described methods of semiparametric estimation with AFT models.[8] Zhang and Peng recommend maximizing the convex function $G(\boldsymbol{\beta})$ shown in equation 11.[12]

$$G(\boldsymbol{\beta}) = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \delta_i w_i^{(m)} |\varepsilon_i - \varepsilon_j| I(\varepsilon_i - \varepsilon_j) \tag{11}$$

Utilizing this method, the maximization of equation 7 can be accomplished by maximizing equation 11 through a linear programming method in R.

With the estimated $\boldsymbol{\beta}$ from the rank method described above, an estimator for the survival function can be acquired based on the residuals, $\tau_i = \log t_i - \boldsymbol{\beta} x_i$, i = 1, …, n. Let $\tau_1 < \tau_2 < \cdots < \tau_k$ be the distinct uncensored failure residuals. Also let d_{τ_j} denote the number of failures and $R(\tau_j)$ be the risk set at τ_j. Equation 12 gives an estimator of $S_0(\varepsilon|Y = 1)$. This updated survival function is then used in the E-step.

$$\hat{S}_0(\varepsilon|Y = 1) = \exp\left(- \sum_{j:t_{(j)} < \varepsilon} \frac{d_{\tau_j}}{\sum_{i \in R(\tau_j)} w_i^{(m)}} \right) \tag{12}$$

Similarly to the semiparametric PHMC model, for $\varepsilon > \tau_k$, we can set $\hat{S}_0(\varepsilon|Y = 1) = 0$. Therefore, the estimated survival function is $\hat{S}(t|Y = 1) = \hat{S}_0(\varepsilon|Y = 1)$.

c. Parametric AFTMC Model

The parametric method of maximum likelihood estimation involves assuming a specific distribution for the failure time of uncured patients.[5] The options in the SAS macro for this distribution are lognormal, log-logistic, exponential, and Weibull. With this distribution selected, we can specify $S(\cdot)$, $h(\cdot)$ in equation 7 by a few unknown parameters associated with that specified distribution. Therefore, the M-step of the EM algorithm involves obtaining $\beta^{(m+1)}$ by maximizing equation 7, as well as other unknown parameters for the specified distribution. SAS uses PROC NLMIXED to obtain these maximum likelihood estimates.

Variance Estimation

The standard errors of the estimated parameters are not directly available from the estimating equations used in the EM algorithm. Therefore, bootstrap methods are employed. The R-package draws random bootstrap samples from the original dataset with replacement using the 'sample' function while the SAS-macro employs non-parametric bootstrap sampling methods and resamples from the original dataset using PROC MULTITEST with the BOOTSTRAP option.

1.2 Software Syntax

As of 2012, a package for modeling semiparametric mixture cure models is available in R, called smcure.[14] Additionally, a macro for modeling parametric

and semiparametric mixture cure models in SAS, called PSPMCM, was developed in 2007.[15]

smcure R-package Syntax

In order to call the smcure function within the smcure R-package, the following syntax is used:

```
smcure(formula, cureform, offset=NULL, data,
na.action=na.omit, model=c("ph", "aft"), link="logit",
Var=TRUE, emmax=50, eps=1e-7, nboot=100)
```

The required arguments in this statement are the **FORMULA**, **CUREFORM**, **DATA**, and **MODEL** portions. The **FORMULA** argument specifies the latency variable(s) and the associated survival response. The **CUREFORM** object allows specification of incidence variables. The **DATA** argument denotes the data frame that contains the incidence and latency variable(s). The **MODEL** statement allows specification of the model, with options of either "ph" for the proportional hazards model or "aft" for the accelerated failure time model.[14]

There are also several optional arguments within the smcure function. The **OFFSET** argument identifies variable(s) with a coefficient 1 in both the incidence and latency parts of the PHMC or AFTMC models. By default, **OFFSET=NULL**. The argument **NA.ACTION** specifies how to deal with missing data with the default option being omission (**NA.OMIT**). The **LINK** option allows specification of the link function used for the incidence component; options include logit (the

default), probit, or complimentary loglog (cloglog). If **VAR= TRUE**, which is the default setting, then bootstrap standard errors are returned for the estimated coefficients; for **VAR=FALSE**, only the coefficient estimates are returned. The argument **EMMAX** specifies the maximum number of iterations of the EM algorithm. If the convergence criterion, specified by **EPS** (default value of **EPS** = 1e-7), is not met after the **EMMAX** specified iterations, the EM iteration will be stopped and the last maximum likelihood estimates will be used in the output.[14] Section 4.1 contains an example of the syntax used for the smcure package with the motivating dataset.

PSPMCM SAS Macro Syntax

In order to invoke the PSPMCM SAS macro which fits parametric AFTMC and semiparametric PHMC models, the following statement is used:

```
%PSPMCM
      (DATA= , ID= , CENSCOD= , TIME= ,
       VAR= , INCPART= , SURVPART= ,
       TAIL= , SU0MET= ,
       MAXITER= , CONVCRIT= , ALPHA= ,
       FAST= , BOOTSTRAP= , NSAMPLE= , BOOTMET= ,
       GESTIMATE= , STRATA= , JACKDATA= , BASELINE= ,
       SPLOT= , PLOTFIT= );
run;
```

Similarly to the smcure package in R, the required arguments include **DATA**, **ID**, **TIME**, **CENSCOD**, **VAR**, **INCPART**, and **SURVPART**, as well as **ALPHA**, **BASELINE**, **SPLOT**, and **PLOTFIT**. These parameters are defined in the

following ways. The dataset name is specified with the **DATA** option; it is required that the dataset is structured so that there is one record per individual. **ID** is the individual's identification number. **CENSCOD** is the censoring indicator where 0 represents censoring, while 1 represents event occurrence. The variable that denotes the time to failure or time to censoring is defined in the **TIME** statement. The **VAR=** statement is used for covariate definition and indication of inclusion in the incidence (I) and/or in the latency (S) parts of the mixture cure model. Names of variables are separated by spaces. Also defined in the **VAR** statement is the value that the covariate will take on in later survival plots. **INCPART** is the selection of the model for the incidence part of the mixture cure model. Options include logit (LOGIT), probit (PROBIT), and complementary log log (CLOGLOG), with a default option is logistic regression (LOGIT). **SURVPART** selects the baseline survival function. The Cox proportional hazards model (COX) selection is the semiparametric option, while the parametric options include lognormal (LOGN), loglogistic (LLOG), exponential (EXP), and Weibull (WEIB).[15]

Additionally required, specification of the significance level, **ALPHA**, is used for the hazard ratio and odds ratio confidence limits. By default, **ALPHA=** 0.05 for 95% confidence intervals, but any value between 0 and 1 can be chosen. The value for the **BASELINE** argument is either Y or N, with a default of N.

When the value is set to Y, the baseline survival function estimate, $\hat{S}_0(t|U=1)$, along with parameter estimates for parametric models, are written to a dataset called BASELINE. If bootstrap resampling is selected, the dataset BASELINE_T will also contain estimates for bootstrap replicates. For a value of Y, the **SPLOT** argument plots the estimated conditional survival curve, $\hat{S}_0(t|U=1,\boldsymbol{x_i},\boldsymbol{z}_i)$, as well as the marginal survival curve, $\hat{S}_0(t|\boldsymbol{x_i},\boldsymbol{z}_i)$ for an individual with covariate vectors $\mathbf{x}_i$ and $\mathbf{z}_i$. The values for these covariate vectors are specified in the **VAR** statement for each covariate following incidence/latency specification and after the comma. However, the default value for **SPLOT** is N. The last required argument is the **PLOTFIT** statement. Although the default value is N, a value of Y results in computation of the observed marginal survival curve, $S^{(obs)}(t|\boldsymbol{x_i},\boldsymbol{z_i})$ for each stratum defined by the covariate vectors. A plot of this observed marginal survival as well as the estimated marginal survival, $\hat{S}(t|\boldsymbol{x_i},\boldsymbol{z_i})$, against time allows for examination the model's prediction abilities. A correlation coefficient between the observed and expected survival probabilities is also calculated for each stratum as a quantifier of goodness of fit.[15]

There are also several optional statements that can be utilized and specified in the SAS macro. The **SU0MET** option allows selection of the method of conditional baseline survival function estimation; either Breslow-type (CH) or product limit estimator (PL) can be selected, with a default of PL. To select a

constraint or tail completion method to be used to estimate $\hat{S}_0(t|U=1)$, the **TAIL** option is defined. The default value, ZERO, specifies that the zero tail constraint is used, while ETAIL and WTAIL select the exponential and Weibull tail completion methods, respectively. NONE is an option to indicate that no constraint is used, but this can cause identifiability and convergence issues. The maximum number of iterations to be performed by the EM algorithm is defined by **MAXITER**. The last maximum likelihood iteration is used in the output if the convergence criterion, which can be defined by **CONVCRIT** (default value is a relative change of less than 1e-5), is not met after the defined maximum number of iterations. By default, **MAXITER** = 200. The **FAST** option is, by default, set to Y in order to write parameter estimates and their standard errors to datasets called FAST_INCI and FAST_SURV.[15]

Additionally, for the PHMC model, there are several options available when bootstrap confidence intervals are requested. If **BOOTSTRAP**=Y, this option invokes the performance of non-parametric resampling with replacement from the original dataset. The default, however, is N. **NSAMPLE** specifies the number of bootstrap replicates produced. The **STRATA** option identifies a stratification variable to use in resampling. The bootstrap confidence interval type is specified by **BOOTMET**, with options including bias corrected (BC), normalized bias corrected (BOOTN), accelerated bias corrected (BCA), hybrid

method (HYB), and percentile (PTCL), as well as Jackknife after bootstrap (JACK). There is also the option to select all methods by specifying **BOOTMET** = ALL. Finally, the option **GESTIMATE**=Y outputs Q-Q plots and bar charts of the distribution of parameter estimates over the bootstrap replicates. This allows graphical checking of the validity of the bootstrap confidence intervals. With overdispersion relative to the normal distribution, percentile based confidence intervals become questionable in their validity.[15] For an example of this syntax utilized in a real setting, see Section 4.2 for the use of the PSPMCM macro with the motivating dataset.

1.3 Motivating Example

Limb-sparing resection is the preferred method of treatment for adult soft tissue sarcomas in extremities. Overall treatment goals are focused on complete resection of the tumor while simultaneously preserving of limb function and also maximizing survival. External beam radiation as an adjuvant therapy for soft tissue sarcomas was first explored in the mid-twentieth century[16,17] and had the goal of reducing local recurrence (LR) rates and assisting in local control in situations where margins were very close or positive for the resected tumor. In 1996, Yang et al. published a study prospectively evaluating the use of radiation therapy versus no radiation therapy in patients with resected soft tissue sarcomas.[18] The conclusion, based solely on log rank tests and Kaplan Meier

curves, was that there was a marked decrease in local recurrence in both high grade and low grade tumors associated with radiation therapy. As evidenced by the low event rate observed in the study population of Yang's paper—1 of 70 patients who received radiation therapy developed a LR while 17 of 71 patients who did not receive XRT developed a LR over the 12 year follow up period[18]—there appears to be evidence of a potential cure fraction in soft tissue sarcoma patients undergoing limb salvage surgery, and that may or may not be a result of radiation therapy.

The motivating dataset obtained from Levine Cancer Institute (LCI) is a result of an IRB approved retrospective chart review of all patients presenting to LCI between 1992 and 2010 who were diagnosed with extremity soft tissue sarcomas. In order to be included in the study, patients must have had the following characteristics: complete medical records, limb sparing surgery for an extremity soft tissue sarcoma, absence of metastatic disease, and be at least 18 years old at presentation time. Additionally, chemotherapy treatment excluded patients from the study. Patients received radiation therapy based on physician discretion and the National Comprehensive Cancer Network (NCCN) guidelines.

A total of 162 patients with soft tissue sarcomas met the inclusion and exclusion criteria mentioned previously. Our analysis will focus on those with

pathologically high grade soft tissue sarcomas, defined as grade 2 or grade 3; this results in a total of 120 patients used in the analysis. The main outcome of interest to be assessed is recurrence free survival, with an event occurring if the patient experienced a tumor recurrence, either local or systemic, or death. A surviving patient received censored status if they did not have a recurrence at last follow up. Table 1.1 describes the variables in the sarcoma dataset that were used or evaluated as potential covariates in a multivariable model.

Table 1.1. Soft Tissue Sarcoma Dataset Variables.

Variable	**Variable Name**	**Code**
Status indicator for any recurrence	cens_rfs	1 = recurrence event 0 = no recurrence (censored)
Time to recurrence	rfs	Continuous
Radiation therapy	rad_01	1 = received radiation therapy 0 = no radiation therapy
Age at surgery	age_01	1 = > 50 years old 0 = ≤ 50 years old
Gender	gender_01	1 = male 0 = female
Categorized tumor size	calc_ts_class	1 = tumor size > 5 cm 0 = tumor size ≤ 5 cm
Tumor site	d_site_01	1 = lower body 0 = upper body
Race	race_01	1 = White 0 = Other

The median follow up time for the recurrence free survival outcome in this study was 3.68 years (range: 0.13 to 20.32 years). There were 24 patients in the group that received no radiation therapy, while 96 patients received radiation therapy. The overall censoring rate was 57.5%, while the censoring rates in the

radiation therapy group versus the no radiation therapy group were 60.42% and 45.83%, respectively. Additionally, information regarding patient demographics and tumor characteristics was obtained and is shown in Table 1.2.

Table 1.2. Study population demographics for the high grade sarcoma patients, stratified by radiation treatment status.

	XRT, n (%)	No XRT, n (%)	**Total, n (%)**
Gender			
Female	45 (47)	19 (79)	**64 (53)**
Male	51 (53)	5 (21)	**56 (47)**
Race			
White	64 (67)	12 (50)	**76 (63)**
Other	28 (29)	10 (42)	**38 (32)**
*Missing***	*4 (4)***	*2 (8)***	***6 (5)****
Age			
≤ 50	34 (35)	10 (42)	**44 (37)**
> 50	62 (65)	14 (58)	**76 (63)**
Tumor Size			
≤ 5 cm	30 (31)	15 (62)	**45 (38)**
> 5 cm	66 (69)	9 (37)	**75 (62)**
Tumor Site			
Upper	23 (24)	6 (25)	**29 (24)**
Lower	73 (76)	18 (75)	**91 (76)**
TOTAL	**96**	**24**	**120**

From this demographics table, there are six patients who are missing information regarding their race. Because of this missing data, race will not be used in the multivariable models. Most studies involving the analysis of soft tissue sarcoma and race show differences in incidence of certain types of sarcomas, but less conclusive results for overall incidence of soft tissue

sarcomas.[19] However, all other covariates beyond the main exposure of interest, radiation therapy, will be evaluated in multivariable models later in this study.

As evidenced by the Kaplan Meier curve shown in Figure 1.1, there may be a difference in recurrence events between the subjects who received radiation therapy and those who did not receive radiation therapy. A log rank test found that this difference between treatment groups is not quite significant, with a p value = 0.0692. However, we can see that the radiation therapy group (XRT) has consistently higher survival probability than the no radiation therapy group (nXRT). Also notable in the Kaplan Meier survival curves are the plateaus for both treatment groups at values much greater than zero. This leveling out of survival curves occurred after about 3 years of follow up in the nXRT group and after about 13 years of follow up in the XRT group. This indicates that some patients will likely not experience a recurrence in either treatment group. This suggests that a proportion of cured patients may exist in those who received radiation therapy and those who did not.

This dataset was also fit with the Cox Proportional Hazards model with only radiation therapy and a hazard ratio for treatment was found to be 0.560 (95% CI: 0.297, 1.055, p = 0.0729). These results, prior to modeling with a mixture cure model, suggest that for those patients receiving radiation therapy, the risk

reduction of a recurrence is not quite significant, but still notably trending towards nearly a 50% reduction.

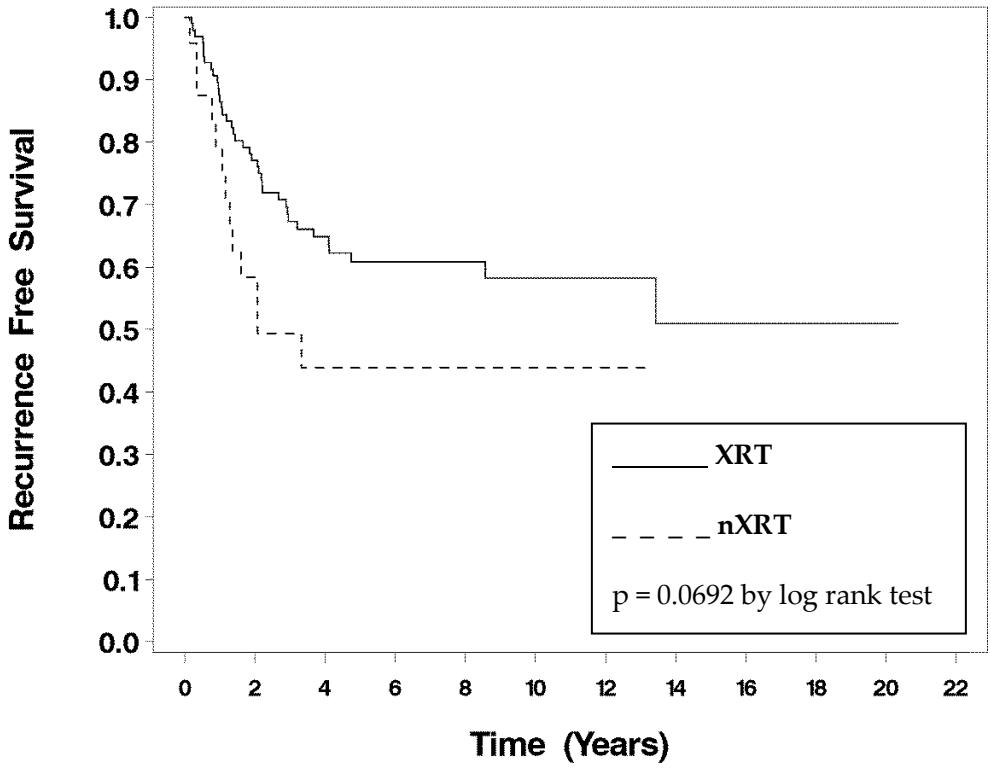

Figure 1.1. Kaplan-Meier recurrence free survival for the radiation therapy group (XRT) and the no radiation therapy group (nXRT).

Model selection procedures with the previously mentioned covariates of interest resulted in a model including radiation therapy, gender, categorized age, and categorized tumor size. The reduction in risk of a recurrence in those who received radiation therapy versus those who did not receive radiation therapy is over 65% (HR = 0.339, p = 0.0018, 95% CI: 0.172, 0.669), after adjusting for age, gender, and tumor size. This significant risk reduction with the radiation treatment is definitely noteworthy. Since this motivating dataset displays the

possibility of a cure proportion, a mixture cure model should be evaluated and implemented in order to accurately estimate that cure proportion as well as the survival probability of uncured patients.

1.4 Outline of Thesis

The main aims of this study are to compare and contrast the R package and the SAS macro, to determine optimal study settings for each of these programs, to suggest updates to the smcure package in R if weaknesses are found, and, finally, to apply the mixture cure model to a new data set with cure tendencies. These aims are accomplished in the following chapters. Chapter 2 outlines the simulation settings used for the comparison of the R package and SAS macro. Chapter 3 contains the results from each of the planned settings. The application of the mixture cure models in R and SAS to the sarcoma data set introduced in Section 1.3 is addressed in Chapter 4. Finally, Chapter 5 summarizes and concludes the study.

CHAPTER 2: SIMULATION PLAN AND METHODOLOGY

In this chapter, the settings used for the simulation studies are outlined. The goal of this study is to compare and contrast the current methods of mixture cure modeling in R and SAS by assessing parameter estimation and standard error estimation. Simulations allowed us to assess the limitations of these two packages, as well as to find their optimal utilization settings. Section 2.1 explains the simulation plan and methods for the PHMC model while Section 2.2 describes the simulation strategy for the AFTMC model. Finally, the expected results of this study are defined in Section 2.3

2.1 PHMC Model Simulation Study Settings

The following aims were investigated for both the SAS macro and the R package with the PHMC model:

- *Baseline survival function distributions*

Survival times of the uncured patients were generated from the Weibull distribution, Wei(1,2), and the standard lognormal distribution, logN(0, 1).

- *Covariates*

Simulation results for the PHMC model were evaluated in two covariate vector settings. The first covariate setting was a single binary covariate from a binomial

distribution with probability 0.5, resulting in the following vectors: $\mathbf{z} = (z_0, z_1)$ and $\mathbf{x} = x_1$. The second setting had two covariates, one categorical covariate, again from a binomial distribution with probability 0.5, and one continuous covariate generated from a standard normal distribution, N(0,1). Therefore, the covariate vectors were $\mathbf{z} = (z_0, z_1, z_2)$ and $\mathbf{x} = (x_1, x_2)$.

- *Cure rate*

The cure rate was varied through definition of the **b** vector, assessing slight and moderate cure rates. For the single binary covariate setting, the slight cure rate for the treatment group ($z_1 = 1$) was 27% and for the control group ($z_1 = 0$), 12%, resulting from defining $\mathbf{b} = (2, -1)$. The moderate cure rate for the treatment group was 40% and 20% for the control group, when $\mathbf{b} = (1.3863, -1)$. The cure rate when there is a continuous variable is found using a numeric integral of the cure rate function. The formula code used to determine this can be found in Appendix C. The resulting slight and moderate cure rates when the continuous variable's coefficient was defined to be 0.3 were the same as with the single covariate setting: 12%/27% and 20%/40%.

- *Censoring rate*

Slight and moderate censoring rates, relative to the defined cure rates, and their impacts were evaluated.

- *Sample size*

Samples of two different sizes were generated, including n = 200 and n = 500, in order to observe the effect of sample size on parameter estimation.

- *Link Function*

The cure proportion of all the data sets was generated from a logistic model. A sensitivity analysis with respect to the different link functions available for the incidence portion was then performed by comparing cure rate estimates resulting from misspecified link functions to the correctly specified logit link estimates.

- *Computation time*

The computation time required for different sample sizes and settings was compared between the R package and the SAS macro.

2.2 AFTMC Model Simulation Study Settings

The following aims were investigated for both the SAS macro and the R package with the AFTMC model:

- *Failure time error distribution*

The error distribution for the survival times of the uncured patients followed the Extreme Value/log-Weibull distribution, defining $\lambda = 1$ and $k = 1$.

- *Covariates*

Simulation results for the AFTMC model was evaluated in only one covariate vector setting: a single binary covariate from a binomial distribution with probability 0.5.

- *Cure rate*

The cure rate was again varied through definition of the **b** vector, assessing slight (12%/27%) and moderate (20%/40%) cure rates, as previously defined.

- *Censoring rate*

Slight and moderate censoring rates, relative to the defined cure rates, and their impacts were evaluated.

- *Sample size*

Similarly to the PHMC model settings, data sets of sample size n = 200 and n = 500 were generated to observe the effect of sample size.

- *Link Function*

The cure proportion of all the data sets was generated from a logistic model and only the correctly specified logit link function was selected from the model options.

- *Computation time*

As the methods of AFTMC model parameter estimation are completely different between R and SAS (semiparametric versus parametric, respectively), the

computation times are not comparable, so no computation time analysis was performed.

2.3 *Expected Results*

The results obtained from these simulation studies included estimate biases, mean square error, confidence interval coverage probabilities, and computation times. Bias is simply defined as the difference between the estimated parameter and the true defined parameter. Mean square error, MSE, is commonly defined as the square of the bias of the estimate plus the variance of the estimate. The coverage probabilities were calculated by finding a 95% confidence interval for each of the parameter estimates and determining the frequency in which the true parameter value was captured. For all computation time data, simulations in R and in SAS were completed using computers with an Intel Core i7-4770S at 3.10 GHz processor, 8.00 GB RAM. R version 3.0.2 was used for the smcure package and SAS version 9.4 was used for the PSPMCM macro. These results allowed some conclusions to be made in the comparison of the statistical software programs.

CHAPTER 3: SIMULATION RESULTS AND DISCUSSION

From the simulation plan described in Chapter 2, the results for each of the combinations of settings gave an extensive assessment of the impact of certain factors on the models' performance. This chapter presents the obtained results and briefly interprets their meaning for the scope of this study. Section 3.1 focuses on the results of the simulations associated with the proportional hazards mixture cure model, while Section 3.2 focuses on the results of the accelerated failure time mixture cure model simulations.

3.1 PHMC Simulation Study

For the proportional hazard mixture cure model simulation study, the following settings were utilized. As previously stated in Chapter 2, the logistic model was used to generate the probability of cure. Settings included both 1-covariate data sets and 2-covariate data sets. Tables 3.1-3.4 give the results from a single z covariate for the cure portion which was a binary variable generated from a binomial distribution with probability 0.5. That same generated variable was used as the **x** covariate in the survival portion. Tables 3.5-3.8 give the results from data sets with 2 covariates generated. The first covariate, z_1, was the same as previously described, a binary variable, while the second covariate, z_2, was a

continuous covariate generated from a standard normal distribution. Again, both of these generated variables were used as the x_1 and x_2 covariates, respectively. A uniform distribution, $U[c_1, c_2]$, was used to generate censoring times, with c_1 and c_2 defined in order to give chosen censoring rates. Survival times of the uncured patients were generated from the Weibull distribution with $\lambda = 2$ and $k = 1$ as well as from the standard Lognormal distribution, logN(0, 1). Generated sample sizes included n = 200 and n = 500. One hundred bootstrap samples were used for all simulation settings as a previous study had found the difference between 100, 200 and 500 samples in the R package to be trivial.[14]

The estimated biases, mean square error (MSE), and confidence interval capture rate for the defined parameters from the PHMC model, are presented in the following tables. In Tables 3.1 and 3.2, the covariate vectors have true values of $b_0 = 2$, $b_1 = -1$, and $\beta = 2$ for a cure rate of 12%, $1 - \pi(\boldsymbol{z} = 0) = 0.12$, in the control group (z = 0) and 27%, $1 - \pi(\boldsymbol{z} = 1) = 0.27$, in the treatment group. In Tables 3.3 and 3.4, the parameters were defined to have true values of $b_0 = 1.3863$, $b_1 = -1$, and $\beta = 2$, therefore, the **b** vector of parameters gives a cure rate of 20% in the control group and a cure rate of 40% in the treatment group. These results include data sets of sample size n = 200 and n = 500, with 500 replications each from both the R package and the SAS macro.

Table 3.1. Estimates from the smcure PHMC model where cure rate is 12% in control and 27% in treatment group. (2,-1,2)

Survival Distr	Censoring Rate	Parameter	True Value	R					
				n = 200			n = 500		
				Bias	MSE	CI Cap	Bias	MSE	CI Cap
Weibull	Slight	$\widehat{b_0}$	2	0.0372	0.4597	93.4	0.0167	0.0792	90.2
	U[0,20]	$\widehat{b_1}$	-1	0.0253	0.6017	96.8	-0.0108	0.1383	94.6
	22.2%	$\widehat{\beta_1}$	2	0.0339	0.1079	95.4	0.0118	0.0413	94.8
	Mod	$\widehat{b_0}$	2	0.0651	2.2389	97.4	0.0315	0.1402	94.0
	U[0,4]	$\widehat{b_1}$	-1	-0.0495	2.4398	98.8	-0.0208	0.2004	95.6
	32.4%	$\widehat{\beta_1}$	2	0.0315	0.1269	92.8	0.0088	0.0437	96.0
LNorm	Slight	$\widehat{b_0}$	2	0.1130	1.6654	95.8	0.0430	0.1519	89.0
	U[0,25]	$\widehat{b_1}$	-1	-0.1037	1.8373	97.8	-0.0358	0.2101	93.0
	23.1%	$\widehat{\beta_1}$	2	0.0137	0.1090	94.8	0.0141	0.0405	93.6
	Mod	$\widehat{b_0}$	2	-0.2311	1.7622	67.6	-0.3475	0.3203	57.2
	U[0.5,5]	$\widehat{b_1}$	-1	0.2472	1.9410	83.8	0.3580	0.3796	68.2
	29.5%	$\widehat{\beta_1}$	2	-0.0435	0.1194	94.0	-0.0754	0.0519	89.2

Table 3.2. Estimates from the PSPMCM PHMC model where cure rate is 12% in control and 27% in treatment group. (2,-1,2)

Survival Distr	Censoring Rate	Parameter	True Value	SAS					
				n = 200			n = 500		
				Bias	MSE	CI Cap	Bias	MSE	CI Cap
Weibull	Slight	$\widehat{b_0}$	2	0.0455	0.2415	95.6	0.0158	0.0893	95.4
	U[0,20]	$\widehat{b_1}$	-1	-0.0601	0.3512	96.4	-0.0016	0.1314	94.8
	22.2%	$\widehat{\beta_1}$	2	0.0177	0.1022	96.2	0.0002	0.0396	93.6
	Mod	$\widehat{b_0}$	2	0.0867	0.6708	94.6	0.0497	0.1521	96.6
	U[0,4]	$\widehat{b_1}$	-1	-0.0634	0.7741	95.2	-0.0416	0.2088	95.4
	32.4%	$\widehat{\beta_1}$	2	0.0246	0.1201	95.6	0.0180	0.0458	95.2
LNorm	Slight	$\widehat{b_0}$	2	0.1349	0.7428	93.1	0.0264	0.1262	94.4
	U[0,25]	$\widehat{b_1}$	-1	-0.1375	0.8535	95.1	-0.0185	0.1716	94.1
	23.1%	$\widehat{\beta_1}$	2	0.0180	0.1049	94.1	0.0085	0.0401	93.6
	Mod	$\widehat{b_0}$	2	-0.1837	1.3239	78.0	-0.2701	0.6469	66.4
	U[0.5,5]	$\widehat{b_1}$	-1	0.1846	1.4418	83.0	0.2885	0.6914	73.0
	29.5%	$\widehat{\beta_1}$	2	-0.0371	0.1128	94.4	-0.0546	0.0481	94.4

Table 3.3. Estimates from the smcure PHMC model where cure rate is 20% in control and 40% in treatment group. (1.3863,-1,2)

Survival Distr	Censoring Rate	Parameter	True Value	R					
				n = 200			n = 500		
				Bias	MSE	CI Cap	Bias	MSE	CI Cap
Weibull	Slight	$\widehat{b_0}$	1.3863	0.0389	0.1287	91.0	0.0171	0.0466	89.6
	U[0,20]	$\widehat{b_1}$	-1	-0.0453	0.2398	94.8	-0.0181	0.0911	94.6
	32.6%	$\widehat{\beta_1}$	2	0.0394	0.1272	95.0	0.0051	0.0459	95.8
	Mod	$\widehat{b_0}$	1.3863	0.0348	0.1626	93.2	0.0157	0.0547	92.2
	U[0,7]	$\widehat{b_1}$	-1	-0.0198	0.2816	96.6	-0.0229	0.1034	95.0
	37.3%	$\widehat{\beta_1}$	2	0.0352	0.1405	94.2	0.0033	0.0506	93.6
LNorm	Slight	$\widehat{b_0}$	1.3863	0.0287	0.1641	90.8	0.0242	0.0588	88.6
	U[0,25]	$\widehat{b_1}$	-1	-0.0276	0.2883	95.6	-0.0073	0.1028	96.0
	33.1%	$\widehat{\beta_1}$	2	0.0112	0.1223	95.8	0.0047	0.0442	96.0
	Mod	$\widehat{b_0}$	1.3863	-0.0001	0.4483	85.2	-0.0286	0.1185	85.8
	U[0,8]	$\widehat{b_1}$	-1	0.0100	0.5873	92.4	0.0321	0.1720	91.0
	39.6%	$\widehat{\beta_1}$	2	0.0138	0.1490	94.2	0.0074	0.0538	95.0

Table 3.4. Estimates from the PSPMCM PHMC model where cure rate is 20% in control and 40% in treatment group. (1.3863,-1,2)

Survival Distr	Censoring Rate	Parameter	True Value	SAS					
				n = 200			n = 500		
				Bias	MSE	CI Cap	Bias	MSE	CI Cap
Weibull	Slight	$\widehat{b_0}$	1.3863	0.0448	0.1547	95.6	0.0148	0.0530	96.2
	U[0,20]	$\widehat{b_1}$	-1	-0.0482	0.2414	95.2	-0.0006	0.0848	97.2
	32.6%	$\widehat{\beta_1}$	2	0.0408	0.1237	94.8	0.0135	0.0461	94.4
	Mod	$\widehat{b_0}$	1.3863	0.0433	0.1827	96.6	0.0057	0.0635	95.4
	U[0,7]	$\widehat{b_1}$	-1	-0.0375	0.2724	96.4	-0.0063	0.0991	95.6
	37.3%	$\widehat{\beta_1}$	2	0.0187	0.1332	95.6	0.0075	0.0457	96.2
LNorm	Slight	$\widehat{b_0}$	1.3863	0.0335	0.1733	96.8	0.0145	0.0698	92.6
	U[0,25]	$\widehat{b_1}$	-1	-0.0247	0.2541	97.6	-0.0101	0.1051	94.8
	33.1%	$\widehat{\beta_1}$	2	0.0153	0.1258	94.4	0.0081	0.0457	95.4
	Mod	$\widehat{b_0}$	1.3863	0.0451	0.8141	91.0	-0.0059	0.4047	89.2
	U[0,8]	$\widehat{b_1}$	-1	-0.0480	0.8692	93.2	0.0104	0.4540	90.8
	39.6%	$\widehat{\beta_1}$	2	0.0017	0.1364	94.4	-0.0051	0.0557	93.6

From the results presented in Tables 3.1-3.4, we can see some trends. In both R and SAS, the estimate biases are very small, consistently less than 0.2, with most less than 0.05. With only a few exceptions, the bias decreases with increasing sample size, as expected. For all settings in both programs, the mean square error for each parameter estimate decreases with increasing sample size and increases with increasing censoring rate from slight to moderate. Confidence interval capture rates are relatively similar between the R and SAS settings. However, the SAS macro seems to have closer to accurate (95%) capture rates than the R package in some of the settings with higher censoring rates or higher cure rates.

The following tables give the results from the two-covariate setting. The **z** and **x** vectors were generated as previously defined in section 2.1. In Tables 3.5 and 3.6, the covariate vectors have true defined values of $\mathbf{b} = (2, -1, 0.3)$ and $\beta = (2, 0.5)$, which still gives a cure rate of 12% in the control group ($z = 0$) and 27% in the treatment group when the value for the continuous variable was chosen to be the mean, which is zero. In Tables 3.7 and 3.8, the parameters were defined to have true values of $\mathbf{b} = (1.3863, -1, 0.3)$ and $\beta = (2, 0.5)$; therefore, the **b** vector of parameters has a cure rate of 20% in the control group and a cure rate of 40% in the treatment group. Again, 500 replications each were performed for sample sizes $n = 200$ and $n = 500$.

Table 3.5. Estimates from the smcure PHMC model where cure rate is 12% in control and 27% in treatment group. (2,-1,0.3,2,0.5)

Survival Distr	Censoring Rate	Parameter	True Value	R					
				n = 200			n = 500		
				Bias	MSE	CI Cap	Bias	MSE	CI Cap
Weibull	Slight	$\widehat{b_0}$	2	0.0604	0.4220	93.6	0.0159	0.0680	91.4
	U[0,20]	$\widehat{b_1}$	-1	-0.0514	0.5818	96.0	-0.0068	0.1229	96.6
	22.3%	$\widehat{b_2}$	0.3	0.0086	0.0834	95.0	0.0054	0.0287	94.8
		$\widehat{\beta_1}$	2	0.0429	0.1065	94.4	0.0110	0.0381	94.2
		$\widehat{\beta_2}$	0.5	-0.0003	0.0176	94.0	0.0007	0.0062	95.4
	Mod	$\widehat{b_0}$	2	0.0990	2.0239	95.4	0.0441	0.1661	93.4
	U[0,4]	$\widehat{b_1}$	-1	-0.0756	2.1783	96.8	-0.0370	0.2292	96.8
	32.1%	$\widehat{b_2}$	0.3	0.0184	0.1067	95.2	0.0066	0.0359	96.6
		$\widehat{\beta_1}$	2	0.0298	0.1226	94.6	0.0149	0.0427	96.0
		$\widehat{\beta_2}$	0.5	0.0036	0.0205	94.8	0.0018	0.0079	94.8
LNorm	Slight	$\widehat{b_0}$	2	0.0953	2.4130	93.0	-0.0161	0.1065	89.2
	U[0,25]	$\widehat{b_1}$	-1	-0.0656	2.5766	94.0	0.0156	0.1620	95.2
	24.1%	$\widehat{b_2}$	0.3	0.0484	0.0982	95.8	0.0201	0.0332	94.4
		$\widehat{\beta_1}$	2	0.0372	0.1140	93.0	0.0087	0.0402	94.2
		$\widehat{\beta_2}$	0.5	-0.0057	0.0189	94.8	-0.0050	0.0070	94.0
	Mod	$\widehat{b_0}$	2	0.3435	1.5454	63.0	-0.3795	0.3396	49.4
	U[0.5,5]	$\widehat{b_1}$	-1	0.3787	1.7530	74.8	0.3910	0.4043	65.2
	31.3%	$\widehat{b_2}$	0.3	0.1213	0.1169	92.6	0.0770	0.0438	90.8
		$\widehat{\beta_1}$	2	-0.0436	0.1225	93.6	-0.0731	0.0477	91.0
		$\widehat{\beta_2}$	0.5	-0.0276	0.0242	92.2	-0.0277	0.0085	94.8

Table 3.6. Estimates from the PSPMCM PHMC model where cure rate is 12% in control and 27% in treatment group. (2,-1,0.3,2,0.5)

Survival Distr	Censoring Rate	Parameter	True Value	SAS					
				n = 200			n = 500		
				Bias	MSE	CI Cap	Bias	MSE	CI Cap
Weibull	Slight	$\widehat{b_0}$	2	0.0283	0.2458	95.8	0.0062	0.0846	95.2
	U[0,20]	$\widehat{b_1}$	-1	-0.0150	0.3691	96.2	0.0056	0.1252	96.8
	22.3%	$\widehat{b_2}$	0.3	0.0223	0.0814	96.6	0.0030	0.0289	96.6
		$\widehat{\beta_1}$	2	0.0167	0.0990	95.0	0.0129	0.0398	94.0
		$\widehat{\beta_2}$	0.5	0.0061	0.0176	93.6	0.0054	0.0066	95.8
	Mod	$\widehat{b_0}$	2	0.1585	0.8130	95.0	0.0320	0.1630	96.0
	U[0,4]	$\widehat{b_1}$	-1	-0.1379	0.9496	94.4	-0.0288	0.2092	96.8
	32.1%	$\widehat{b_2}$	0.3	0.0285	0.1080	97.0	0.0156	0.0390	95.8
		$\widehat{\beta_1}$	2	0.0452	0.1186	95.8	0.0138	0.0466	94.0
		$\widehat{\beta_2}$	0.5	-0.0037	0.0218	94.4	0.0003	0.0082	94.8
LNorm	Slight	$\widehat{b_0}$	2	0.0508	0.5075	94.2	-0.0274	0.1188	91.8
	U[0,25]	$\widehat{b_1}$	-1	-0.0452	0.6181	94.8	0.0384	0.1666	93.8
	24.1%	$\widehat{b_2}$	0.3	0.0405	0.0886	95.4	0.0285	0.0340	93.6
		$\widehat{\beta_1}$	2	0.0354	0.1059	93.8	-0.0046	0.0388	95.6
		$\widehat{\beta_2}$	0.5	-0.0090	0.0182	94.4	-0.0103	0.0071	93.4
	Mod	$\widehat{b_0}$	2	-0.3118	1.3787	67.8	-0.3417	0.6848	55.6
	U[0.5,5]	$\widehat{b_1}$	-1	0.3428	1.5409	72.0	0.3670	0.7421	64.6
	31.3%	$\widehat{b_2}$	0.3	0.1255	0.1117	92.4	0.0826	0.0435	90.4
		$\widehat{\beta_1}$	2	-0.0604	0.1223	92.6	-0.0579	0.0467	95.0
		$\widehat{\beta_2}$	0.5	-0.0305	0.0247	93.2	-0.0231	0.0097	92.6

Table 3.7 Estimates from the PHMC model where cure rate is 20% in control and 40% in treatment group. (1.3863,-1,0.3,2,0.5)

Survival Distr	Censoring Rate	Parameter	True Value	R					
				n = 200			n = 500		
				Bias	MSE	CI Cap	Bias	MSE	CI Cap
Weibull	Slight	$\widehat{b_0}$	1.3863	0.0408	0.1288	90.4	0.0091	0.0486	87.2
	U[0,20]	$\widehat{b_1}$	-1	-0.0327	0.2444	97.6	-0.0191	0.0900	95.0
	34.0%	$\widehat{b_2}$	0.3	0.0069	0.0582	96.4	0.0044	0.0230	95.0
		$\widehat{\beta_1}$	2	0.0348	0.1272	95.2	0.0141	0.0449	95.0
		$\widehat{\beta_2}$	0.5	0.0051	0.0209	95.2	0.0063	0.0079	93.6
	Mod	$\widehat{b_0}$	1.3863	0.0451	0.1732	92.0	0.0256	0.0585	89.4
	U[0,7]	$\widehat{b_1}$	-1	-0.0439	0.2942	95.4	-0.0179	0.1041	95.8
	37.0%	$\widehat{b_2}$	0.3	0.0099	0.0662	96.4	0.0007	0.0257	95.0
		$\widehat{\beta_1}$	2	0.0280	0.1362	95.2	0.0189	0.0486	94.6
		$\widehat{\beta_2}$	0.5	0.0035	0.0239	93.8	0.0023	0.0083	95.2
LNorm	Slight	$\widehat{b_0}$	1.3863	0.0186	0.1869	90.2	-0.0044	0.0543	91.8
	U[0,25]	$\widehat{b_1}$	-1	-0.0100	0.3113	95.8	0.0099	0.0983	96.2
	33.9%	$\widehat{b_2}$	0.3	0.0280	0.0665	95.8	0.0243	0.0246	97.2
		$\widehat{\beta_1}$	2	0.0277	0.1162	96.6	0.0064	0.0463	93.4
		$\widehat{\beta_2}$	0.5	0.0083	0.0240	92.4	0.0082	0.0082	96.8
	Mod	$\widehat{b_0}$	1.3863	-0.1085	0.3203	80.0	-0.1321	0.1039	79.8
	U[0,8]	$\widehat{b_1}$	-1	0.1086	0.4412	91.8	0.1243	0.1501	89.8
	39.7%	$\widehat{b_2}$	0.3	0.0761	0.0797	94.6	0.0506	0.0334	90.6
		$\widehat{\beta_1}$	2	-0.0238	0.1384	94.8	-0.0249	0.0471	95.6
		$\widehat{\beta_2}$	0.5	-0.0282	0.0273	93.8	-0.0229	0.0104	92.8

Table 3.8 Estimates from the PSPMCM PHMC model where cure rate is 20% in control, 40% in treatment group. (1.3863,-1,0.3,2,0.5)

Survival Distr	Censoring Rate	Parameter	True Value	SAS					
				n = 200			n = 500		
				Bias	MSE	CI Cap	Bias	MSE	CI Cap
Weibull	Slight	$\widehat{b_0}$	1.3863	0.0453	0.2240	96.0	0.0136	0.0777	95.0
	U[0,20]	$\widehat{b_1}$	-1	-0.0386	0.3177	95.8	-0.0162	0.1174	93.0
	34.0%	$\widehat{b_2}$	0.3	0.0334	0.0754	94.2	-0.0034	0.0288	93.4
		$\widehat{\beta_1}$	2	0.0481	0.1279	96.0	0.0118	0.0475	93.8
		$\widehat{\beta_2}$	0.5	-0.0005	0.0222	94.6	-0.0021	0.0087	95.2
	Mod	$\widehat{b_0}$	1.3863	0.0528	0.1918	97.4	0.0202	0.0680	95.2
	U[0,7]	$\widehat{b_1}$	-1	-0.0479	0.2772	97.0	-0.0103	0.1052	95.8
	37.0%	$\widehat{b_2}$	0.3	0.0133	0.0695	95.0	0.0020	0.0248	96.4
		$\widehat{\beta_1}$	2	0.0373	0.1318	94.4	0.0081	0.0503	94.6
		$\widehat{\beta_2}$	0.5	0.0151	0.0237	93.0	0.0016	0.0086	94.0
LNorm	Slight	$\widehat{b_0}$	1.3863	0.0417	0.2131	96.2	0.0006	0.0703	94.2
	U[0,25]	$\widehat{b_1}$	-1	-0.0740	0.3157	95.4	0.0102	0.1094	95.6
	33.9%	$\widehat{b_2}$	0.3	0.0404	0.0655	96.4	0.0132	0.0245	95.8
		$\widehat{\beta_1}$	2	0.0255	0.1166	96.2	-0.0020	0.0448	95.6
		$\widehat{\beta_2}$	0.5	-0.0053	0.0230	93.2	-0.0014	0.0083	94.6
	Mod	$\widehat{b_0}$	1.3863	-0.0849	0.5257	88.8	-0.1251	0.1339	85.6
	U[0,8]	$\widehat{b_1}$	-1	0.1014	0.6344	90.8	0.1369	0.1742	88.8
	39.7%	$\widehat{b_2}$	0.3	0.0605	0.0764	96.4	0.0409	0.0289	96.2
		$\widehat{\beta_1}$	2	-0.0193	0.1312	96.4	-0.0194	0.0533	94.2
		$\widehat{\beta_2}$	0.5	-0.0256	0.0271	94.4	-0.0148	0.0098	94.4

Some trends are also evident in Tables 3.5-3.8 with the two-covariate setting results. The estimate biases were smaller than 0.4 in all settings; however most biases were smaller than even 0.15. Again, mean square error decreased with larger sample sizes and increased with increasing censoring rate. Confidence interval coverage probabilities were, for the most part, relatively good for both the R package and the SAS macro, although the lognormal settings with higher censoring rates typically saw much lower coverage probabilities than expected (95%). Additionally, the PSPMCM macro was consistently closer in all settings to the accurate capture rate.

Sensitivity Analysis

In order to assess the effect of link function specification on model estimation, we used the same data generation described for Tables 3.1-3.4, but we focused only on the sample size of 500, a single binary covariate, and the Weibull survival distribution. As previously described, the probability of cure was generated from a logistic model, but as the smcure package and the PSPMCM macro both have three options for the link function (logit, probit, and cloglog), we re-estimated the unknown parameters using each of those link functions in order to simulate misspecification of the link function chosen for the probability of cure. The cure rates for the control group and the treatment group were then calculated using the appropriate link functions. The resulting estimates of the **b**

parameters within each of the previously specified settings are summarized in Tables 3.9-3.12.

Table 3.9. Estimated cure rates for different link functions as compared to true cure rate (12%/27%); slight censoring with Weibull survival distribution (n=500).

Censoring Distribution	Parameter	True Value	Estimated Values					
			R			SAS		
			Logit	Probit	Cloglog	Logit	Probit	Cloglog
U[0,20]								
22.2%	$\widehat{b_o}$	2	2.0167	1.1627	0.7248	2.0158	1.1857	0.7585
	$\widehat{b_1}$	-1	-1.0108	-0.5476	-0.4581	-1.0016	-0.5618	-0.4798
	Cure Rate							
	Control	0.1192	0.1175	0.1225	0.1269	0.1176	0.1179	0.1182
	Treatment	0.2689	0.2678	0.2692	0.2710	0.2662	0.2664	0.2668

Table 3.10. Estimated cure rates for different link functions as compared to true cure rate (12%/27%); moderate censoring with Weibull survival distribution (n=500).

Censoring Distribution	Parameter	True Value	Estimated Values					
			R			SAS		
			Logit	Probit	Cloglog	Logit	Probit	Cloglog
U[0,4]								
32.3%	$\widehat{b_o}$	2	2.0315	0.9726	0.5054	2.0497	1.2024	0.7708
	$\widehat{b_1}$	-1	-1.0208	-0.3851	-0.2810	-1.0416	-0.5823	-0.4957
	Cure Rate							
	Control	0.1192	0.1159	0.1654	0.1906	0.1141	0.1146	0.1152
	Treatment	0.2689	0.2668	0.2784	0.2861	0.2674	0.2676	0.2680

Table 3.11. Estimated cure rates for different link functions as compared to true cure rate (20%/40%); slight censoring with Weibull survival distribution (n=500).

Censoring Distribution	Parameter	True Value	Estimated Values					
			R			SAS		
			Logit	Probit	Cloglog	Logit	Probit	Cloglog
U[0,20]								
32.4%	$\widehat{b_o}$	1.3863	1.4034	0.8273	0.4480	1.4011	0.8492	0.4813
	$\widehat{b_1}$	-1	-1.0181	-0.5884	-0.5557	-1.0006	-0.5992	-0.5734
	Cure Rate							
	Control	0.2000	0.1973	0.2040	0.2091	0.1976	0.1979	0.1983
	Treatment	0.4046	0.4048	0.4056	0.4074	0.4012	0.4013	0.4017

Table 3.12. Estimated cure rates for different link functions as compared to true cure rate (20%/40%); moderate censoring with Weibull survival distribution (n=500).

Censoring Distribution	Parameter	True Value	Estimated Values					
			R			SAS		
			Logit	Probit	Cloglog	Logit	Probit	Cloglog
U[0,7]								
37.3%	$\widehat{b_0}$	1.3863	1.4020	0.7669	0.3616	1.3920	0.8438	0.4764
	$\widehat{b_1}$	-1	-1.0229	-0.5372	-0.4920	-1.0063	-0.6029	-0.5782
	Cure Rate							
	Control	0.2000	0.1975	0.2216	0.2380	0.1991	0.1994	0.1998
	Treatment	0.4046	0.4063	0.4092	0.4157	0.4048	0.4048	0.4053

From these tables, despite the large bias in parameter estimation, the estimated cure rates associated with each of the different link functions overall are quite similar to the true cure rates. With the moderate censoring distributions, there was slightly more bias in the cure rate estimates from the R package than the SAS macro, as seen in Tables 3.10 and 3.12. In these cases where the bias in the cure rate estimates is larger in the R package, the complimentary log log estimates were more biased than the probit estimates, while the logit was predictably the most accurate. However, the most inaccurate cure rate estimate, observed for the control group with the complimentary log log link in Table 3.10, was still relatively close to the true value, (0.1906 versus 0.1159).

Computation Time Analysis

In order to assess and compare the time-restrictions of the PHMC models in R and SAS, the average time elapsed to get parameter estimates and bootstrapped standard error was recorded for the smcure package and the

PSPMCM macro. For each of the following settings, 500 data sets were generated as previously described. The average computation times for the parameter estimates and their bootstrapped standard errors with each of the models (R-smcure, SAS-PSPMCM) are recorded in Table 3.13.

Table 3.13. Average computation time for model parameter estimation, comparing R and SAS PHMC models.

Survival Distribution	Covariate Setting, True Vector	Computation Time (seconds)			
		R		SAS	
		n = 200	n = 500	n = 200	n = 500
Weibull	Cens U[0,20]				
	1 Covariate (2, -1, 2)	3.99	9.42	2.99	3.37
	2 Covariates (2, -1, 0.3, 2, 0.5)	4.19	9.24	3.47	4.16
	Cens U[0,4]				
	1 Covariate (2, -1, 2)	8.53	16.50	6.32	7.58
	2 Covariates (2, -1, 0.3, 2, 0.5)	9.67	19.36	9.56	10.73
Lognormal	Cens U[0,25]				
	1 Covariate (2, -1, 2)	6.55	15.37	5.31	6.56
	2 Covariates (2, -1, 0.3, 2, 0.5)	7.78	16.07	7.85	8.38
	Cens U[0.5,5]				
	1 Covariate (2, -1, 2)	15.23	40.49	12.03	17.92
	2 Covariates (2, -1, 0.3, 2, 0.5)	13.47	32.61	13.83	18.62

As seen in Table 3.13, the average computation times for the PSPMCM macro in SAS are shorter than the computation times for the smcure package in R. The data generated with a logistic-Lognormal data typically took longer than the logistic-Weibull data, especially when censoring rate was higher and sample size

was larger. Despite some noticeable differences in computation times between R and SAS, for the big picture, even the setting with the longest time took, on average, much less than a minute (40.5 seconds) to obtain parameter estimates and standard errors.

3.2 AFTMC Simulation Study

For the accelerated failure time mixture cure model simulation study, the following settings were utilized. As with the PHMC model simulations, the logistic model was used to generate the probability of cure. Only a single binary covariate **z** was generated from a binomial distribution with probability 0.5. That same generated variable was used as the **x** covariate in the survival portion. A uniform distribution, $U[c_1, c_2]$, was used to generate censoring times, with constants c_1 and c_2 defined in order to give chosen censoring rates. The error distribution for the failure times of the uncured patients followed the extreme value distribution. In Tables 3.14-3.15, we have regression parameters defined as $b_0 = 2$, $b_1 = -1$, and $\beta_1 = 2$ for a cure rate of 12% in the control group ($z = 0$) and 27% in the treatment group. Similarly in Tables 3.16-3.17, regression parameters are defined as $b_0 = 1.3863$, $b_1 = -1$, and $\beta_1 = 2$, for cure rates of 20% and 40% in the control and treatment groups, respectively. The logistic-Extreme data was generated in samples sizes of $n = 200$ and $n = 500$, with 500 replications. Additionally, 100 bootstrap samples were chosen.

Table 3.14. Estimates from the smcure AFTMC model where cure rate is 12% in control and 27% in treatment group. (2,-1,2)

Censoring Rate	Parameter	True Value	R					
			n = 200			n = 500		
			Bias	MSE	CI Cap	Bias	MSE	CI Cap
Slight	$\widehat{b_0}$	2	0.0400	0.5275	93.4	0.0300	0.0783	92.6
U[0.5,30]	$\widehat{b_1}$	-1	-0.1181	0.9785	94.2	0.0092	0.2051	96.4
28.7%	$\widehat{\beta_1}$	2	-0.0318	0.1298	92.4	-0.0065	0.0833	92.0
Moderate	$\widehat{b_0}$	2	0.1527	2.4471	94.8	0.0301	0.1118	93.0
U[0.5,9]	$\widehat{b_1}$	-1	-0.4970	3.4686	85.0	-0.5031	0.6138	63.0
43.2%	$\widehat{\beta_1}$	2	-0.2469	0.3514	78.0	-0.2697	0.1743	65.8

Table 3.15. Estimates from the PSPMCM AFTMC model where cure rate is 12% in control and 27% in treatment group. (2,-1,2)

Censoring Rate	Parameter	True Value	SAS					
			n = 200			n = 500		
			Bias	MSE	CI Cap	Bias	MSE	CI Cap
Slight	$\widehat{b_0}$	2	0.0446	0.2163	95.6	0.0199	0.0853	94.6
U[0.5,30]	$\widehat{b_1}$	-1	0.0456	0.9093	95.8	0.0256	0.2385	96.2
28.7%	$\widehat{\beta_1}$	2	0.0112	0.1247	94.0	0.0130	0.0467	95.8
Moderate	$\widehat{b_0}$	2	0.0552	0.3129	97.2	0.0191	0.1188	94.2
U[0.5,9]	$\widehat{b_1}$	-1	-0.5700	4.8323	82.4	-0.2304	3.2700	84.6
43.2%	$\widehat{\beta_1}$	2	-0.3504	0.5829	85.6	-0.1617	0.2761	88.8

Table 3.16. Estimates from the smcure AFTMC model where cure rate is 12% in control and 27% in treatment group. (1.3863,-1,2)

Censoring Rate	Parameter	True Value	R					
			n = 200			n = 500		
			Bias	MSE	CI Cap	Bias	MSE	CI Cap
Slight	$\widehat{b_0}$	1.3863	0.0318	0.1193	90.2	0.0322	0.0527	86.0
U[0.5,30]	$\widehat{b_1}$	-1	0.0274	0.3859	94.4	0.0112	0.1397	95.2
37.9%	$\widehat{\beta_1}$	2	0.0180	0.1657	89.8	-0.0186	0.0638	93.0
Moderate	$\widehat{b_0}$	1.3863	0.0183	0.2086	91.8	0.0156	0.0579	90.0
U[0.5,9]	$\widehat{b_1}$	-1	-0.1627	0.8810	80.0	-0.3019	0.3822	73.0
49.8%	$\widehat{\beta_1}$	2	-0.1865	0.4448	75.2	-0.2421	0.1987	74.6

Table 3.17. Estimates from the PSPMCM AFTMC model where cure rate is 12% in control and 27% in treatment group. (1.3863,-1,2)

Censoring Rate	Parameter	True Value	SAS					
			n = 200			n = 500		
			Bias	MSE	CI Cap	Bias	MSE	CI Cap
Slight	$\widehat{b_0}$	1.3863	-0.0012	0.1287	96.6	0.0004	0.0573	94.2
U[0.5,30]	$\widehat{b_1}$	-1	0.0396	0.6169	96.8	-0.0033	0.1339	94.2
37.9%	$\widehat{\beta_1}$	2	-0.0015	0.1530	95.6	0.0035	0.0539	96.6
Moderate	$\widehat{b_0}$	1.3863	0.0453	0.1741	96.4	0.0213	0.0693	95.0
U[0.5,9]	$\widehat{b_1}$	-1	0.3111	4.3420	86.0	-0.0710	2.7025	89.2
49.8%	$\widehat{\beta_1}$	2	-0.2895	0.7862	84.8	-0.1113	0.3780	89.4

The results from the AFTMC model simulations show some similar trends as the PHMC model results. Tables 3.14-3.17 show relatively small biases overall. Again, mean square error decreased with increasing sample size and increased with higher censoring rates. Confidence interval coverage probabilities were acceptable for the semiparametric methods in R, but they were somewhat concerning in the parametric methods in SAS. It is expected that those capture rates would be much closer to accurate because of the fully parametric methods used in the SAS macro.

CHAPTER 4: REAL DATA ANALYSIS- LEVINE CANCER INSTITUTE DATA

As described in Section 1.3, the soft tissue sarcoma dataset from Levine Cancer Institute appeared to have cure tendencies as seen by the plateauing in the Kaplan Meier survival curves. Therefore, it was used as an illustration for the application of the semiparametric PH mixture cure model, utilizing both the smcure package in R and the PSPMCM macro in SAS. For comparison, the options available were defined similarly between R and SAS. To get bootstrap standard error estimates, 100 bootstrap samples were selected. The maximum number of iterations was set at 200 and the convergence criterion was defined as 1e-7. Since the smcure method of conditional baseline survival function estimation is the Breslow type, the CH option was selected in the PSPMCM macro.

4.1 smcure Package Application

First, using the smcure R package, the PHMC model was fit to the motivating sarcoma dataset. The variables used in this mixture cure model statement are described above in Table 1.1; however, the first model that was run included only the main treatment factor: radiation therapy. The following shows an example of the syntax used in R:

```
smcurehisarc<- smcure(Surv(rfs,cens_rfs)~rad_01,
cureform=~rad_01, data=hisarc, model="ph", emmax=200, eps=1e-7,
nboot=100)
```

The resulting parameter estimates and bootstrapped standard errors obtained from invoking this model are shown in Table 4.1.

Table 4.1. smcure PHMC model parameter estimates and bootstrapped standard errors for the simple model, along with Z-values and associated p-values.

Model	Parameter	Estimate	Standard Error	Z Value	Pr(|Z|)
Cure Portion	Intercept	0.34996	0.49094	0.71283	0.47595
	rad_01	-0.32083	0.60437	-0.53084	0.59553
Survival	rad_01	-0.76114	0.40139	-1.89629	0.05792

In order to obtain the standard errors of the estimated parameters, 100 bootstrap samples were acquired. From the parameter estimates in the cure probability model, the cure rate can be calculated for each of the treatment groups. The cure rate for the radiation group is calculated to be $1 - \hat{\pi}(\boldsymbol{z}) = 1 - e^{0.34996-0.32083}/(1 + e^{0.34996-0.32083}) = 0.493$, suggesting that 49.3% of patients receiving radiation therapy are cured. The cure rate for the group that did not receive radiation therapy was calculated similarly, but was found to only be 41.3%. However, the p-value for the radiation therapy variable is not significant at $\alpha = 0.05$ level of significance (p = 0.5955). Therefore we cannot truly conclude that there is a significant difference in cure rate between the two treatment groups when only assessing the radiation treatment.

We can also obtain the predicted survival probabilities for the radiation treatment groups using the `predictsmcure` command as shown below. The resulting predicted survival curves are shown with the Kaplan Meier curves in Figure 4.1.

```
predhisarc=predictsmcure(smcurehisarc, newX=c(1,0),
newZ=c(1,0),model="ph")
```

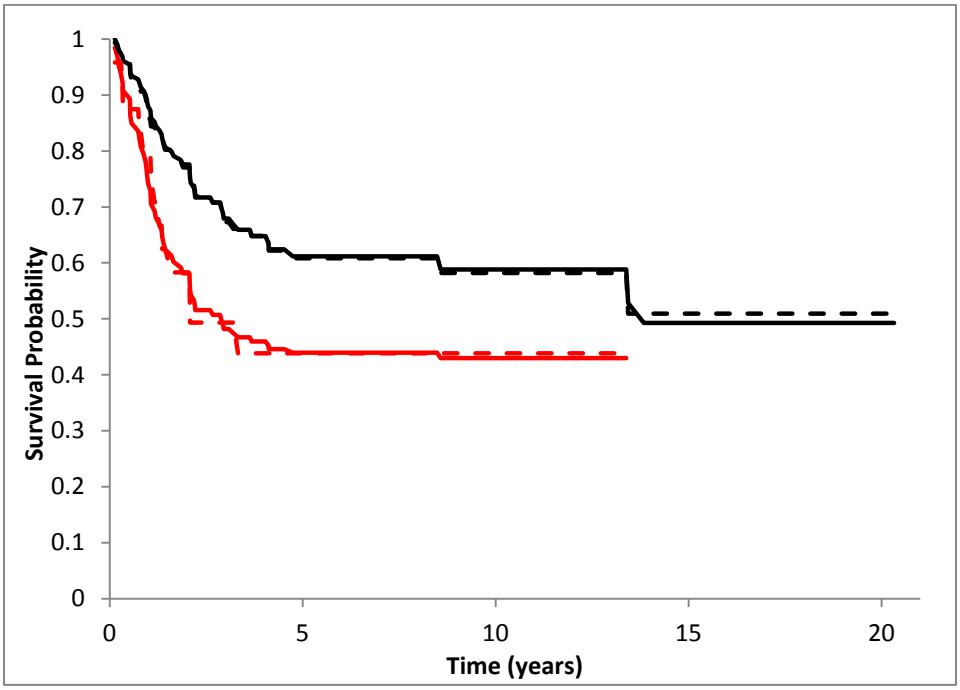

Figure 4.1 Predicted survival curves and Kaplan Meier curves for the sarcoma study, stratified by treatment group. Upper black lines are the XRT group while lower red lines are the nXRT group. The dashed curves are the Kaplan Meier estimates while the solid lines are the Cox mixture cure model estimates.

From the predicted survival curves in Figure 4.1, it would appear that those who receive radiation therapy have a better predicted survival probability, meaning they are less likely to experience a recurrence, than those who do not receive radiation therapy. Additionally, the logistic-Cox mixture cure model appears to be a good fit for the data since the KM curves and the MC model

estimates are very similar. The p-value for the radiation parameter estimate in the failure time distribution model is not quite significant (p = 0.0579). However, although a solid conclusion of significance is not possible, the near significant p-value suggests that radiation therapy could have a positive effect on recurrence free survival in the uncured population. The hazard ratio for receiving radiation therapy versus not receiving radiation therapy is $e^{-0.76114} = 0.47$ (95% CI: 0.213, 1.026), suggesting that the use of radiation therapy in patients who are not "cured" appears to slow down the recurrence of a soft tissue sarcoma. A multivariable mixture cure model that assesses the inclusion of variables described in Table 1.1 is also estimated in Section 4.3.

4.2 PSPMCM Macro Application

Utilizing the SAS PSPMCM macro, the PHMC model was again fit to the motivating sarcoma dataset. As previously stated, the variables used in this mixture cure model statement are described in Table 1.1, starting initially with only radiation therapy in the model. The following shows the statement used to invoke the PSPMCM macro in SAS.

```
%PSPMCM
      (DATA= hisarc, ID= id, CENSCOD= cens_rfs,
       TIME= rfs, VAR= rad_01(IS,1),INCPART= logit,
       SURVPART= Cox, TAIL= zero, SU0MET= ch,
       MAXITER= 200, CONVCRIT= 1e-7, ALPHA= 0.05,
       FAST= Y, BOOTSTRAP= Y, NSAMPLE= 100, BOOTMET= ALL,
       GESTIMATE= , STRATA= , JACKDATA= , BASELINE= ,
       SPLOT= , PLOTFIT= Y);
run;
```

Here, note that VAR=rad_01(IS,1), which means that the radiation treatment variable, rad_01, is included in both the incidence and the latency parts of the model. Also, the specified value after the comma is the value that the rad_01 variable will take on in any obtained survival function plots. The following Tables 4.2 and 4.3 show the parameter estimates output resulting from invocation of the SAS macro.

Table 4.2. PSPMCM macro incidence parameter estimates output from the LOGISTIC procedure for the sarcoma data.

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	0.3501	0.4145	0.7135	0.3983
rad_01	1	-0.3204	0.4621	0.4808	0.4880

Table 4.3. PSPMCM macro latency parameter estimates output from the PHREG procedure for the sarcoma data.

Analysis of Maximum Likelihood Estimates								
Parameter	**DF**	**Parameter Estimate**	**Standard Error**	**Chi-Sq**	**Pr > ChiSq**	**Hazard Ratio**	**95% HR Confidence Limits**	
rad_01	1	-0.65078	0.33982	3.6676	0.0555	0.522	0.268	1.015

The standard errors of these fast estimates in the output may be underestimated as they are based on the inverted Hessian matrix which was computed on the last maximum likelihood iteration. Therefore, the "fast" standard errors and confidence intervals should not be used for drawing conclusions about the effects of the variables of interest. Instead, the bootstrap standard errors and confidence limits should be used. These can be obtained from the "BOOTDIST"

data set that is written by the PSPMCM macro. The parameter estimates, boot strap standard errors, and p-values for the mixture cure model are given in Table 4.2, where the intercept and "L_rad_01" variables come from the incidence portion and the "S_rad_01" variable is from the latency portion of the model.

Table 4.4. PSPMCM model parameter estimates and bootstrapped standard errors for the simple model, along with Z-values and associated p-values.

Parameter	Estimate	Standard Error	Z Value	Pr(|Z|)
Intercept	0.3501	0.58566	0.59779	0.5500
L_rad_01	-0.3204	0.65882	-0.48632	0.6267
S_rad_01	-0.76135	0.45673	-1.66696	0.0955

These parameter estimates are consistent with the results from the smcure package. The cure rate for each of the treatment groups can be easily calculated from the parameter estimates for the incidence portion. For the radiation group, we calculated $1 - \hat{\pi}(\boldsymbol{z}) = 1 - e^{0.3501-0.3204}/(1 + e^{0.3501-0.3204}) = 0.493$, which indicates that 49.3% of patients receiving radiation therapy are cured. Comparatively, the patients who did not receive radiation therapy only had a cure rate of 41.3%. However, the radiation therapy parameter estimate is not significant in the incidence portion, so there does not appear to be a significant difference in cure rate between the two groups before adjusting for other covariates.

In order to obtain the bootstrap standard errors, 100 bootstrap samples were acquired from the macro. These standard errors are also comparable with

the smcure package standard errors and the same conclusions, in terms of significance, are drawn. There appears to be some potential positive effect of radiation therapy on the recurrence free survival in the uncured subset of patients, although it is not significant at $\alpha = 0.05$ level of significance (p = 0.0955).

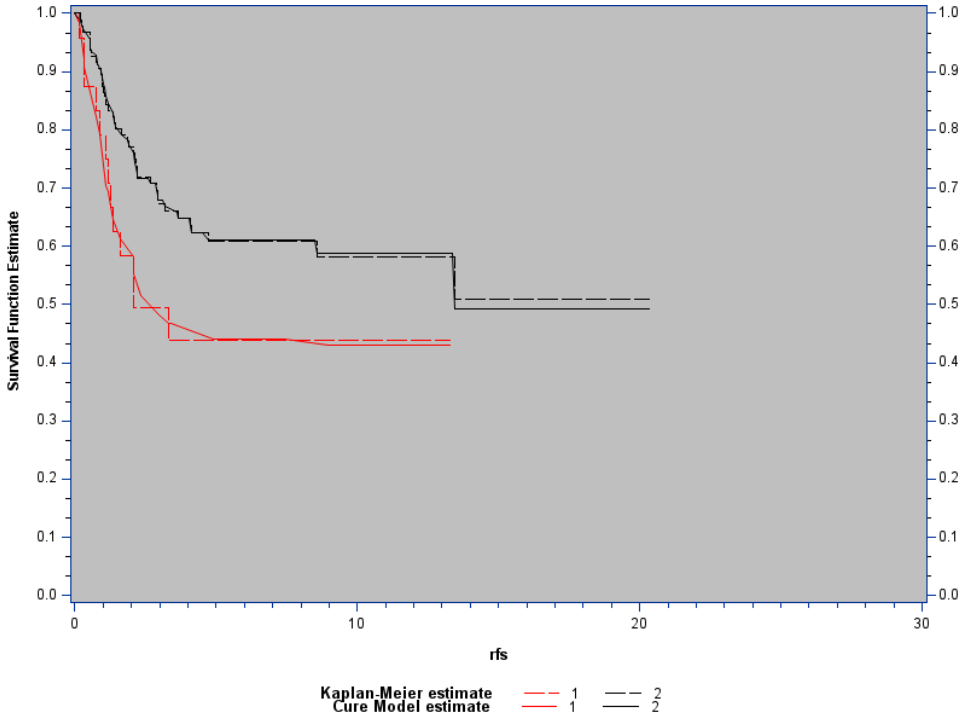

Figure 4.2. Marginal survival function curves for soft tissue sarcoma dataset. Upper black lines are the XRT group while lower red lines are the nXRT group. The dashed curves are the Kaplan Meier estimates while the solid lines are the Cox mixture cure model estimates.

The PSPMCM macro also gives the estimated and observed marginal survival functions, shown in Figure 4.2. The correlation between the estimated and the observed survival probabilities is also obtained and gives an idea of goodness of fit. The correlation between the observed and expected marginal survival curves for the nXRT group is 0.99365 while the correlation for the XRT

group is 0.99880. These high correlations are indicative of a good fit by the logistic-Cox mixture cure model.

4.3 Multiple Proportional Hazards Mixture Cure Models

In order to more completely analyze this soft tissue sarcoma dataset with the mixture cure model, multiple covariates will be assessed to select the most appropriate model. The covariates available in the dataset include several demographic variables such as age and gender and also tumor-related information such as tumor size and tumor site. The model selection procedure in started in R with all five variables (radiation therapy, categorized age, tumor size class, gender, and tumor site) in the "full" model. Table 4.3 gives the parameter estimates and standard errors for the full model.

Table 4.5. smcure PHMC model parameter estimates and bootstrapped standard errors for the full model, along with Z-values and associated p-values.

Model	Parameter	Estimate	Standard Error	Z Value	Pr(|Z|)
Cure Portion	Intercept	-1.04005	3.93646	-0.26421	0.79162
	rad_01	-1.21938	5.86065	-0.20806	0.83518
	age_01	0.85081	0.91845	0.92636	0.35426
	gender_01	0.49043	2.03085	0.24149	0.80917
	ts_class	2.12661	3.11636	0.68240	0.49498
	d_site_01	0.02141	1.04421	0.02051	0.98364
Survival	rad_01	-0.89582	0.57766	-1.55075	0.12096
	age_01	0.33178	0.61650	0.53816	0.59047
	gender_01	-0.14530	0.56793	-0.25583	0.79808
	ts_class	-0.22466	0.72984	-0.30782	0.75822
	d_site_01	0.03264	0.71526	0.04563	0.96361

The least significant covariate that was first removed from full model was tumor site in the cure portion, p = 0.9836. Tumor site in the survival portion was the next variable to be removed (p = 0.9419). Following the complete removal of the tumor site variable, the categorical age variable was then removed from the cure portion (p = 0.9015) and then tumor size class was removed from the survival portion (p = 0.9728). Gender was also removed from both portions with insignificant p-values (p = 0.6891, 0.8518 for the survival and cure portions, respectively). The resulting model, shown in Table 4.4, does not contain all significant variables; however, it was chosen as the final reduced model since, with the adjustment for age, the radiation therapy variable is significant in the survival portion (p = 0.0055) suggesting that radiation therapy has a significant impact on survival probability in those who are uncured, after adjusted for age. Additionally, the inclusion of gender and tumor size class in the cure portion results in a noteworthy, although insignificant p-value (p = 0.1112) for the radiation therapy variable. This suggests that after adjusting for tumor size, there might be a difference in cure probability for those who receive radiation therapy and those who do not. For those with tumors less than 5 cm in diameter, the probability of cure is 80.2% for those who receive radiation therapy, and 58.4% for those who do not receive radiation therapy. Comparatively, for those who have tumors larger than 5 cm, the probability of cure in those who receive

radiation therapy is 34.8% and only 15.6% in those who do not receive radiation therapy. Appendix C, part (a) contains the details of these calculations.

Table 4.6. smcure PHMC model parameter estimates and bootstrapped standard errors for the reduced model, along with Z-values and associated p-values.

Model	Parameter	Estimate	Standard Error	Z Value	Pr(|Z|)
Cure Portion	Intercept	-0.3396	0.4745	-0.7157	0.4742
	rad_01	-1.0574	0.6639	-1.5927	0.1112
	ts_class	2.0255	0.5685	3.5627	0.0004
Survival	rad_01	-0.8726	0.3141	-2.7784	0.0055
	age_01	0.6692	0.4936	1.3557	0.1752

Utilizing this same model in the SAS macro finds the following full (Table 4.5) and reduced (Table 4.6) models with parameter estimates. We do not find quite the same significance or near significance from the PSPMCM reduced multivariable model that we had with the smcure package.

Table 4.7. PSPMCM PHMC model parameter estimates and bootstrapped standard errors for the full model, along with Z-values and associated p-values.

Model	Parameter	Estimate	Standard Error	Z value	Pr(|Z|)
Cure Portion	Intercept	-1.0392	0.99398	-1.04549	0.295795
	rad_01	-1.219	0.76143	-1.60094	0.109391
	age_01	0.8502	0.83643	1.016463	0.309409
	gender_01	0.4907	0.78237	0.627197	0.53053
	calc_ts_c	2.1261	0.82965	2.562647	0.010388
	d_site_01	0.0208	0.84674	0.024565	0.980402
Survival	rad_01	-0.8957	0.5601	-1.59918	0.109781
	age_01	0.3323	0.60639	0.547997	0.583694
	gender_01	-0.1456	0.52043	-0.27977	0.779655
	calc_ts_c	-0.2242	0.66474	-0.33727	0.73591
	d_site_01	0.0333	0.65947	0.050495	0.959728

The parameter estimates are very similar but the standard errors are larger which results in larger p-values. The p-value for the radiation therapy in the cure portion is more significant (p = 0.1038) and in the survival portion, where radiation therapy was significant in the smcure model, radiation therapy is also significant at α = 0.05 (p = 0.0432). Cure probability calculations are nearly identical between the two programs, which is to be expected with very similar estimates. Appendix C, part (b) contains the details of these calculations.

Table 4.8. PSPMCM PHMC model parameter estimates and bootstrapped standard errors for the reduced model, along with Z-values and associated p-values.

Model	Parameter	Estimate	Standard Error	Z value	Pr(|Z|)
Cure Portion	Intercept	-0.3397	0.5371	-0.6325	0.5271
	rad_01	-1.0572	0.6498	-1.6269	0.1038
	ts_class	2.0255	0.6237	3.2473	0.0012
Survival	rad_01	-0.8728	0.4317	-2.0220	0.0432
	age_01	0.6691	0.5231	1.2792	0.2008

CHAPTER 5: CONCLUSIONS AND FUTURE STUDIES

Mixture cure models have growing applications in the field of biostatistics as the advancement in medical treatments and technology has led to more diseases being cured rather than just mitigated. With the typical survival methods that assume all patients will eventually get the disease of interest, there is no way to account for a proportion of cured individuals who will never experience that disease occurrence. Mixture cure models allow both the cured proportion and the remaining uncured individuals to be modeled simultaneously with incidence and latency portions, respectively. The goal of this thesis was to compare methods of this mixture cure modeling in two popular statistical software programs, R and SAS. Utilizing simulations and real data analysis, that comparison was possible. This final chapter, divided into six sections, summarizes and explains the findings from the comparative methods used for the smcure package in R and the PSPMCM macro in SAS.

5.1 Overall Model Comparison

The PHMC models in R and SAS are consistent in terms of computational methods for parameter estimates. Only slight differences in parameter estimates

arise when the PSPMCM macro option for estimation of the conditional baseline survival function, **SU0MET,** is chosen to be "PL" for the product limit estimator. This is because the default and only option for the smcure package in R is the Breslow-type method, which is selected in the PSPMCM macro with "CH."

On the contrary, as discussed in Section 1.1, the smcure package has the advantage of implementing a semiparametric approach to the AFTMC model, which makes fewer assumptions than the parametric approach for the AFTMC model from the PSPMCM macro. Semiparametric approaches (or nonparametric when possible) are often preferred because of their flexibility since assumptions about the survival distribution are not necessary.

5.2 Syntax and Model Output Comparisons

Comparing the R package to the SAS macro in terms of the user interface, the PSPMCM macro was found to have many more options available for specification while the smcure package contained more "default only" settings. The additional complex plots available in SAS, especially the goodness of fit Q-Q plot, seem advantageous for certain study goals. Additionally, there is more freedom in the SAS macro for specification of certain computational methods. This includes multiple survival distributions as well as the multiple methods for estimation of the conditional baseline survival function. However, with more options to specify in the macro, it is potentially easier for the user to run into

issues with misuse of the macro or incorrect option selection. The simplicity and straightforward options in the smcure package might make it more ideal for someone looking only to obtain parameter estimates and standard errors.

Another aspect for comparison is the output from the two programs. In the PSPMCM macro, the output that the user sees initially will include the "Fast Estimates" for the incidence and latency parts. However, that output does not contain the bootstrapped standard errors, and instead, gives the standard errors based on the inverted Hessian matrix which was computed on the last maximum likelihood iteration. This could easily be misinterpreted by users who assume that, since they selected the bootstrap option, they would get bootstrap standard errors in the primary output. Instead bootstrap standard errors are only found through the "BOOTDIST" data set that is created with the selection of bootstrap resampling. The smcure output, however, directly contains the bootstrap standard errors and associated p-values for quicker and easier interpretation.

5.3 Estimate Bias and Mean Square Error Comparisons

Both the smcure package and the PSPMCM macro showed good performance in terms of bias and mean square error for a range of settings in both the PHMC model and the AFTMC, as seen previously in Sections 3.1 and 3.2. One notable difference though is the inflated mean square error of estimates for a few specific settings in R for the Weibull PHMC model. With moderate

censoring in the small sample size and smaller cure rate setting, the average mean square error for the cure portion estimates was much larger for the R model estimates than for those from SAS, in both covariate settings. There were also somewhat increased mean square errors in the lognormal data settings; however, these increases were more consistent between the two programs, R and SAS.

Additionally, for the AFTMC settings, the semiparametric model in R cannot estimate the intercept, β_0, as it is not identifiable. The intercept is identifiable in the SAS macro, so the intercept can be estimated; however, without a value from R for comparison, this estimate was not reported. Besides that discrepancy, the two AFTMC models are relatively similar in their performance with the estimates and standard errors. Although, some of the confidence interval capture rates in the fully parametric model from SAS are notably low for a parametric method. With fully parametric methods, it is expected to see more consistent results.

5.4 Computation Time Comparisons

From Section 3.1, we saw that for the PHMC model, the R package takes slightly longer to estimate parameters and standard errors than the SAS macro. However, for all settings and in both programs, it only takes a matter of seconds, on average, to get parameter estimates with 100 bootstrap samples. Therefore,

there does not appear to be a distinct advantage in terms of computation time for the SAS macro as compared to the R package. Also, note that the computation times were not compared in the AFTMC models in R and SAS as they are not comparable, having two difference estimation methods.

5.5 Real Data Analysis Comparisons

Without adjusting for other covariates, the parameter estimates were identical between the smcure package and PSPMCM macro for the simple PHMC model, which looked only at the impact of radiation therapy on the recurrence free survival in soft tissue sarcoma patients. Additionally, the conclusions drawn from the p-values between the two simple models were similar. It is noteworthy though that the standard errors from the smcure package were consistently smaller than those from the PSPMCM macro, which can lead to different conclusions in certain situations. This is noticed in Section 4.3 where the multivariable PHMC model was explored. Both R and SAS found the radiation therapy to be significant in the survival portion ($p = 0.0055$ and $p = 0.0432$, respectively). Additionally, both models suggest a possible trend for the radiation therapy variable in the cure portion ($p = 0.1112$ and $p = 0.1038$).

Because there were some inconsistencies in bootstrap standard errors in the smcure adjusted and the PSPMCM adjusted models, larger bootstrap samples were chosen, to see if the differences were a result of unstable estimates

due to small bootstrap sample size. In the process of choosing larger bootstrap sample sizes, it was found that the standard error estimates are not stable in the smcure package. Running the identical model with 100 bootstrap samples multiple times resulted in several different p-values for radiation therapy (as well as the other cure portion variables), ranging from 0.11 that resulted from the initial model fit to 0.72, as a result of different standard error estimates. When 500 bootstrap samples were used, the bootstrap standard error estimates and resulting p-values remained more consistent in the cure portion, but were actually very large, indicating little or no significance in any of the cure portion variables. This issue is inconsistent with the results from changing the bootstrap sample size to 500 in the SAS macro. The standard error estimates remained consistent with 100 bootstrap samples and only increased a little with 500 bootstrap samples. These issues motivate some of the future directions mentioned in Section 5.6.

5.6 Future Directions

Beyond the scope of this study, there are several directions in which the two models could be further evaluated and potentially improved upon. The impact of different bootstrap sample sizes should be evaluated both on the mean square error of estimates as well as the added computation time. Although a previous study had found the difference in standard errors between 100, 200 and

500 samples sizes with the R package to be trivial, there might be some impact and potential benefit in the SAS macro [3]. By comparing the expected positive effect that an increased bootstrap sample size has on the estimate's mean square error to the subsequent expected additional computation time, a "cost-benefit" analysis of sorts could be performed to identify ideal bootstrap sample sizes in the R package and the SAS macro.

One of the biggest factors that should be investigated further is the variance estimation procedures in the smcure package and PSPMCM macro. The bootstrapping methods utilized in R and those used in SAS differ enough that, despite similar parameter estimates, the estimate bootstrapped standard errors are dissimilar. A complete analysis of the sampling methods should be performed for better understanding of the differences between the R package and the SAS macro for mixture cure models.

Also, simulation studies should be performed to examine the bootstrap variance in comparison to the empirical variance. A brief investigation of the empirical variances of the estimates compared to the bootstrap variances resulting from this thesis is presented in Appendix D. No significant conclusions can be drawn from the limited results, although it is noteworthy that for small sample sizes the bootstrap variances are not comparable to the empirical variances. When there are inconsistencies between the bootstrap variance and the

empirical variance, the bootstrap variance may not be stable. This relationship should be further assessed in future studies to determine if there is some connection with model performance, such as confidence interval capture rates.

An additional evaluation that could be beneficial for a more complete assessment of the R package and the SAS macro would be a study of the survival distribution effect for the AFTMC models. By purposefully misspecifying the survival distribution of a generated data set and evaluating the estimate's biases and mean square errors, the robustness of smcure's semiparametric AFTMC model could be compared to PSPMCM's parametric AFTMC model. This would be carried out somewhat similarly to the sensitivity analysis performed in Section 3.1, only in this case, the focus would be on the failure time generation. Failure time errors would be generated from a lognormal distribution while the survival distribution would still be defined as the log Weibull/Extreme Value distribution.

Additionally, as the rank based estimation method for the semiparametric AFTMC model in R performs better with continuous covariates, it would be interesting to further study the performances of the semiparametric and parametric AFTMC models with a different covariate setting. Perhaps the smcure AFTMC model would have closer results, or even better results, to the parametric PSPMCM macro's results.

Finally, as evidenced by long computation times for the AFTMC model in the smcure package (ranging from approximately ten minutes to nearly two hours), there appears to be room for improvement in the source code to potentially speed up the estimation and bootstrapping times. The source code for the package was initially written in R rather than C or C++. Rewriting the package in C or C++, where the algorithms are typically faster, could result in an improvement in computation time. Additionally, there are aspects in the current bootstrapping source code that could be streamlined for efficiency.

REFERENCES

1. Farewell 1982, "The use of mixture models for the analysis of survival data with long-term survivors." Biometrics, 1982: 1041-1046.
2. Berkson J., and R. Gage. "Survival curve for cancer patients following treatment." Journal of the American Statistical Association, 1952: 501-515.
3. Boag, J. "Maximum likelihood estimates of the proportion of patients cured by cancer therapy." Journal of the Royal Statistical Society: Series B (Methodological), 1949:15-53.
4. Haybittle, J.L. "A two-parameter model for the survival curve of treated cancer patients." American Statistical Association, 1965: 16-26
5. Peng Y., and K.C. Carriere. "An empirical comparison of parametric and semiparametric cure models." Biometrical Journal, 2002: 1002-1014.
6. Peng, Y. Dear K.B.G., and J.W. Denham. "A generalized F mixture model for cure rate estimation." Statistics in Medicine, 1998: 813-830.
7. Yamaguchi K. "Accelerated failure-time regression models with a regression model of surviving fraction: An application to the analysis of 'Permanent Employment' in Japan." American Statistical Association, 1992: 284-292

8. Li, C., and J. Taylor. "A semi-parametric accelerated failure time cure model." Statistics in Medicine, 2002: 3235-3247.
9. Peng, Y. Fitting semiparametric cure models." Computational Statistics & Data Analysis, 2003: 481-490.
10. Sy J.P., and J.M.G. Taylor. "Estimation in a Cox Proportional Hazards Cure Model." Biometrics, 2000: 227-236.
11. Taylor, J. "Semi-parametric estimation in failure time mixture models." Biometrics: 1995: 899-907.
12. Zhang, J., and Y. Peng. "A new estimation method for the semiparametric accelerated failure time mixture cure model." Statistics in Medicine, 2007:3157-3171
13. Peng, Y, and K.B.G. Dear. "A nonparametric mixture model for cure rate estimation." Biometrics, 2000: 237-243.
14. Cai, C., Zou, Y., Peng, Y., and J. Zhang. "smcure: An R-package for estimating semiparametric mixture cure models." Computer Methods and Programs in Biomedicine, 2012: 1255-1260.
15. Corbière, F. and P. Joly. "A SAS macro for parametric and semiparametric mixture cure models." Computer Methods and Programs in Biomedicine, 2007: 173-180.

16. McNeer, G.P., Cantin, J., Chu, F. and J.J. Nickson. "Effectiveness of radiation therapy in the management of sarcoma of the soft somatic tissues." Cancer, 1968: 391-397.

17. Perry, H., and F.C.H. Chu. "Radiation therapy in the palliative management of soft tissue sarcomas." Cancer, 1962: 179-183.

18. Yang, J.C., Chang, A.E., Baker, A.R., et al. "Randomized prospective study of the benefit of adjuvant radiation therapy in the treatment of soft tissue sarcomas of the extremity." Journal of Clinical Oncology, 1996: 1679-1689.

19. Hsieh M.C., Wu X.C, Andrews P.A., and Chen V.W. "Racial and Ethnic Disparities in the Incidence and Trends of Soft Tissue Sarcoma Among Adolescents and Young Adults in the United States, 1995-2008." J Adolesc Young Adult Oncol. 2013: 89-94

APPENDIX A: SELECTED R CODE

Weibull PHMC model, 2 covariate simulation code

```
#######################################
###define data generation function#####
#######################################
generate<-function(n,intercept,bZ,beta,k,lambda,C,simu,nb){

est<-matrix(rep(0,simu*3),nrow=simu)
estbias<-matrix(rep(0,simu*3),nrow=simu)
var<-matrix(rep(0,simu*3),nrow=simu)
mse<-matrix(rep(0,simu*3),nrow=simu)
lcl<-matrix(rep(0,simu*3),nrow=simu)
ucl<-matrix(rep(0,simu*3),nrow=simu)
cap<-matrix(rep(0,simu*3),nrow=simu)
time_mod<-matrix(rep(0),nrow=simu)
censrate<-matrix(rep(0),nrow=simu)
  for(i in 1:simu){
      #cure rate indicator
      z<- rbinom(n,1,0.5)
      linpred<-cbind(1,z) %*% c(intercept,bZ)
      prob<-exp(linpred)/(1+exp(linpred))
      y<-rbinom(n=n, size=1, prob=prob)

      #survival probabilities Weilbull and time calculation
      u<-runif(n,0,1)
      x<-z
time1<-qweibull(1-exp(-exp(log(-log(u))-beta*x)),k,lambda)
      delta<-as.numeric(time1<=C)

#delta restriction based on cure indicator
      status<-ifelse(y==0,0,delta)
      T<-status*time1+(1-status)*C

      #final data to be run in smcure
      curedata<- data.frame(status,T,z,x)

      #censoring rate
      censrate[i]<-1-(sum(curedata$status)/n)
      Tdistr<-hist(T)

      ptm<-proc.time()
smcuremod<- smcure(Surv(T,status)~x,cureform=~z, data=curedata,
model="ph",link="logit",nboot=nb)
      time_mod[i]<-(proc.time()-ptm)[3]

      est[i,]<-c(smcuremod$b, smcuremod$beta)
```

```
      estbias[i,]<-est[i,]-c(intercept,bZ,beta)

      var[i,]<-c(smcuremod$b_var, smcuremod$beta_var)
      mse[i,]<-var[i,]+estbias[i,]*estbias[i,]

      lcl[i,]<-est[i,]-1.96*sqrt(var[i,])
      ucl[i,]<-est[i,]+1.96*sqrt(var[i,])

cap[i,]<-ifelse(lcl[i,]<=c(intercept,bZ,beta) &
ucl[i,]>=c(intercept,bZ,beta),1,0)
      }

##est bias##
biasvec<-apply(estbias,2,mean)
##est MSE##
MSEvec<-apply(mse,2,mean)
##est means##
meanvec<-apply(est,2,mean)
##mean censoring rate##
meanCR<-apply(censrate,2,mean)
##capture rate##
caprate<-apply(cap,2,function(x) sum(x)/simu)
##mean time##
meantime<-apply(time_mod,2,mean)

write.csv(est,"est.csv")
write.csv(estbias,"estbias.csv")
write.csv(mse,"mse.csv")
write.csv(time_mod,"time_mod.csv")


list(meanvec,biasvec,MSEvec,meanCR,caprate,meantime)
}
##############end function##############

#########PARAMETER DEFINITION##########
#define number of subjects
n<-200
####logistic model####
#define coefficients for logistic model
intercept<-2
bZ<--1
####failure time model####
#generate rv from weibull distribution
lambda<-2
k<-1
C<-runif(n,0,4)
#define coefficient for failure time distr
beta<-2
#number of simulations
simu<-500
#number bootstrap resamples
nb<-100

########################################
generate(n,intercept,bZ,beta,k,lambda,C,simu,nb)
```

APPENDIX B: SELECTED SAS CODE

Weibull PHMC model, 2 covariate simulation code

```
*generate from weibull distribution;
%Macro weibsimulate(n,numsim,f,g,intercept,bZ,beta,k,lambda);
options nonotes; ods graphics off; ods exclude all; ods noresults;
proc datasets nolist; delete outestw; run;
proc datasets nolist; delete outbootsew; run;
proc datasets nolist; delete estcrw; run;
proc datasets nolist; delete timew; run;

      %do id=1 %to &numsim;
            data temp;
                  do id=1 to &n;
                        uni=rand('UNIFORM'); /*U[0,1]*/
                        c= &f + (&g-&f)*uni; /*U[f,g]*/
                        zvar=rand('BERNOULLI',0.5);
                        linpred=&intercept+&bZ*zvar;
                        prob=exp(linpred)/(1+exp(linpred));
                        y=rand('BERNOULLI',prob);
                        u=rand('UNIFORM');
                        xvar=zvar;
                        time1=quantile('WEIBULL',1-exp(-exp(log(-
                        log(u))-&beta*xvar)),&k,&lambda);
                        if time1<=c then delta=1; else delta=0;
                        if y=0 then status=0; else status=delta;
                        T=status*time1+(1-status)*c;
                        output;
                  end;
            drop linpred uni u prob time1 delta;
            run;
            proc summary data=temp;
            var status; output out=sumoftemp sum=statsum; run;
            data censrate; set sumoftemp;
            cr = 1-(statsum)/&n; run;
            data a;
            starttime=%sysfunc(time()); run;

            %PSPMCM
                        (DATA=temp, ID=id, CENSCOD=status, TIME=T,
                         VAR= zvar(I) xvar(S),
                         INCPART=logit, SURVPART=Cox,
                         TAIL=zero, SU0MET=ch,
                         MAXITER=200, CONVCRIT=1e-5, ALPHA=0.05,
                         FAST=Y, BOOTSTRAP=Y, NSAMPLE=100, BOOTMET=,
                         GESTIMATE=, STRATA=,
                         JACKDATA=, BASELINE=, SPLOT=, PLOTFIT= );
            run;
```

```
            data b; set a;
            finishtime=%sysfunc(time());
            elapsedtime=finishtime-starttime; run;

            data fast_surv; set fast_surv;
            if Parameter="lpi" then delete;
            keep parameter estimate; run;

            data fast_inci; set fast_inci;
            Parameter=variable;
            keep parameter estimate; run;

            data est; set fast_inci fast_surv;
            sim=&id; run;

            proc means data=bootdist;
            var L_int ; output out=se1 std=sd; run;
            data se1; set se1;
            parameter='Intercept';
            drop _type_ _freq_; run;

            proc means data=bootdist;
            var L_zvar ; output out=se2 std=sd; run;
            data se2; set se2;
            parameter='zvar';
            drop _type_ _freq_; run;

            proc means data=bootdist;
            var S_xvar ; output out=se3 std=sd; run;
            data se3; set se3;
            parameter='xvar';
            drop _type_ _freq_; run;

            data se; set se1 se2 se3;
            sim=&id; run;

      proc append base=outestw data=est; run;
      proc append base=outbootsew data=se; run;
      proc append base=estcrw data=censrate; run;
      proc append base=timew data=b; run;
      %end;
options notes; ods graphics on; ods exclude none; ods results;
%mend weibsimulate;
%weibsimulate(200,500,0,20,2,-1,2,2,1)

title1 "Computation time: weibsimulate(200,500,0,20,2,-1,2,2,1)";
*simulation results;
title2 "Average Computation Time";
proc means data=timew;
var elapsedtime;
run;

data outestw2;
set outestw;
if parameter="Intercept" then truevalue=2;
      else if parameter="zvar" then truevalue=-1;
      else if parameter="xvar" then truevalue=2;
```

```
bias=estimate-truevalue;
biassq=bias*bias;
run;

proc sort data=outestw2;
by sim parameter;
run;

proc sort data=outbootsew;
by sim parameter;
run;

data finalw;
merge outestw2 outbootsew;
by sim parameter;
run;

data allw;
set finalw;
var = sd*sd;
mse = var+biassq;
lcl = estimate-1.96*sd;
ucl = estimate+1.96*sd;
if lcl <=truevalue and ucl>=truevalue then capture = 'yes';
      else capture = 'no';
run;

proc sort data=allw; by parameter; run;

title2 'Confidence Interval Capture';
proc freq data=allw; by parameter; tables capture; run;

title2 'Average Bias of Parameter Estimates';
proc means data=allw; by parameter; var bias; run;

title2 'Average MSE of Parameter Estimates';
proc means data=allw; by parameter; var mse; run;

title2 'Average Censoring Rate';
proc means data=estcrw; var cr; run;
```

APPENDIX C: CALCULATIONS AND REAL DATA ANALYSIS OUTPUT

Cure Rate Calculation (in R)

```
#case of (2,-1,0.3) (12%/27%)
#control
nn<-function(z){(1/(1+exp(2+0.3*z)))*1/sqrt(2*pi)*exp(-z^2/2)}
integrate(nn, -Inf,Inf)

#treatment
nn<-function(z){(1/(1+exp(1+0.3*z)))*1/sqrt(2*pi)*exp(-z^2/2)}
integrate(nn, -Inf,Inf)

#case of (1.3863,-1,0.3) (20%/40%)
#control
nn<-function(z){(1/(1+exp(1.3863+0.3*z)))*1/sqrt(2*pi)*exp(-
z^2/2)}
integrate(nn, -Inf,Inf)

#treatment
nn<-function(z){(1/(1+exp(0.3863+0.3*z)))*1/sqrt(2*pi)*exp(-
z^2/2)}
integrate(nn, -Inf,Inf)
```

Real Data Cure Rate Calculations

A. R- real data results

a. Simple model: cure rate calculations

Radiation:
$1 - \hat{\pi}(\boldsymbol{z}) = 1 - e^{0.3500-0.3208}/(1 + e^{0.3500-0.3208}) = 0.493$
No radiation:
$1 - \hat{\pi}(\boldsymbol{z}) = 1 - e^{0.3500}/(1 + e^{0.3500}) = 0.413$

b. Multiple covariate model: cure rate calculations

Tumors less than 5 cm
Radiation:
$1 - \hat{\pi}(\boldsymbol{z}) = 1 - e^{-0.3396-1.0574}/(1 + e^{-0.3396-1.0574}) = 0.802$

No radiation:

$$1 - \hat{\pi}(\boldsymbol{z}) = 1 - e^{-0.3396}/(1 + e^{-0.3396}) = 0.584$$

Tumors greater than 5 cm

Radiation:

$$1 - \hat{\pi}(\boldsymbol{z}) = 1 - e^{-0.3396-1.0574+2.0255}/(1 + e^{-0.3396-1.0574+2.0255}) = 0.384$$

No radiation:

$$1 - \hat{\pi}(\boldsymbol{z}) = 1 - e^{-0.3396+2.0255}/(1 + e^{-0.3396+2.0255}) = 0.156$$

B. SAS- real data results

a. Simple model: cure rate calculations

Radiation:

$$1 - \hat{\pi}(\boldsymbol{z}) = 1 - e^{0.3501-0.3204}/(1 + e^{0.3501-0.3204}) = 0.493$$

No radiation:

$$1 - \hat{\pi}(\boldsymbol{z}) = 1 - e^{0.3501}/(1 + e^{0.3501}) = 0.413$$

b. Multiple covariate model: cure rate calculations

Tumors less than 5 cm

Radiation:

$$1 - \hat{\pi}(\boldsymbol{z}) = 1 - e^{-0.3397-1.0572}/(1 + e^{-0.3397-1.0572}) = 0.802$$

No radiation:

$$1 - \hat{\pi}(\boldsymbol{z}) = 1 - e^{-0.3397}/(1 + e^{-0.3397}) = 0.584$$

Tumors greater than 5 cm

Radiation:

$$1 - \hat{\pi}(\boldsymbol{z}) = 1 - e^{-0.3397-1.0572+2.0255}/(1 + e^{-0.3397-1.0572+2.0255}) = 0.384$$

No radiation:

$$1 - \hat{\pi}(\boldsymbol{z}) = 1 - e^{-0.3397+2.0255}/(1 + e^{-0.3397+2.0255}) = 0.156$$

APPENDIX D: VARIANCE COMPARISON

Table D.1. Bootstrap versus empirical variance comparison for select simulation settings

Lognormal			**b0**	**b1**	**beta**
n=200	12%/27%	Bootstrap Variance	1.708806	1.941011	0.119375
	U[0.5,5]	Empirical Variance	0.913531	0.980333	0.059238
n=500	12%/27%	Bootstrap Variance	0.199578	0.251463	0.046256
	U[0.5,5]	Empirical Variance	0.130475	0.144098	0.024305
Weibull					
n=200	12%/27%	Bootstrap Variance	3.311067	3.551104	0.109674
	U[0,4]	Empirical Variance	0.676867	0.752525	0.048935
n=500	12%/27%	Bootstrap Variance	0.139222	0.199957	0.043578
	U[0,4]	Empirical Variance	0.069359	0.093058	0.020368