METHODS FOR SURVIVAL ANALYSIS IN SMALL SAMPLES

Rengyi Xu

A DISSERTATION

in

Epidemiology and Biostatistics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2017

Supervisor of Dissertation

Pamela A. Shaw

Associate Professor of Biostatistics

Co-Supervisor of Dissertation

Devan V. Mehrotra

Adjunct Associate Professor of Biostatistics

Graduate Group Chairperson

Nandita Mitra, Professor of Biostatistics

Dissertation Committee

Sharon X. Xie, Professor of Biostatistics

Warren B. Bilker, Professor of Biostatistics

Shannon L. Maude, Assistant Professor of Pediatrics

METHODS FOR SURVIVAL ANALYSIS IN SMALL SAMPLES

ACKNOWLEDGEMENT

I would like to express my deepest appreciation to my dissertation advisors, Dr. Pamela Shaw and Dr. Devan Mehrotra, for their guidance, dedication and support throughout the process of completing this thesis. I thank them for devoting an incredible amount of energy and time into my dissertation research, and also for teaching me become an independent biostatistician. I would also like to thank my committee members, Dr. Sharon Xie, Dr. Warren Bilker and Dr. Shannon Maude, for their insightful suggestions and feedbacks that greatly improved my dissertation.

My sincere appreciation goes to the faculty, staff and students in the Divison of Biostatistics, including my research advisors, Dr. Mary Putt, Dr. Kathleen Propert and Dr. Sarah Ratcliffe, and my master's thesis advisor, Dr. Rui Feng, for the collaborative research opportunity and the invaluable advice and guidance. I would also like to thank Dr. Wei-Ting Hwang and the Superfund Research Program for providing me with the opportunity to collaborate on interesting projects in environmental health. I would also like to acknowledge the friendship and support from my fellow students; my graduate school journey would not be complete without the lovely people I have met at the University of Pennsylvania.

Finally, I would like to give my most appreciative and loving gratitude to my parents for their unconditional love and support throughout my life. I thank them for always being there for me every step of the way. This dissertation is dedicated to them.

ABSTRACT

METHODS FOR SURVIVAL ANALYSIS IN SMALL SAMPLES

Rengyi Xu

Pamela A. Shaw

Devan V. Mehrotra

Studies with time-to-event endpoints and small sample sizes are commonly seen; however, most statistical methods are based on large sample considerations. We develop novel methods for analyzing crossover and parallel study designs with *small* sample sizes and *time-to-event outcomes*. For two-period, two-treatment (2×2) crossover designs, we propose a method in which censored values are treated as missing data and multiply imputed using pre-specified parametric failure time models. The failure times in each imputed dataset are then log-transformed and analyzed using ANCOVA. Results obtained from the imputed datasets are synthesized for point and confidence interval estimation of the treatment-ratio of geometric mean failure times using model-averaging in conjunction with Rubin's combination rule. We use simulations to illustrate the favorable operating characteristics of our method relative to two other existing methods. We apply the proposed method to study the effect of an experimental drug relative to placebo in delaying a symptomatic cardiac-related event during a 10-minute treadmill walking test. For parallel designs for comparing survival times between two groups in the setting of proportional hazards, we propose a refined generalized log-rank (RGLR) statistic by eliminating an unnecessary approximation in the development of Mehrotra and Roth's GLR approach (2001). We show across a variety of simulated scenarios that the RGLR approach provides a smaller bias than the commonly used Cox model, parametric models and the GLR approach in small samples (up to 40 subjects per group), and has notably better efficiency relative to Cox and parametric models in terms of mean squared error. The RGLR approach also consistently delivers adequate confidence interval coverage and type I error control. We further show that while the performance of the parametric model can be significantly influenced by misspecification of the true underlying survival distribution, the RGLR approach provides a consistently low bias and high relative efficiency. We apply all competing methods to data from two clinical trials studying lung cancer and bladder cancer, respectively. Finally, we further extend the

RGLR method to allow for stratification, where stratum-specific estimates are first obtained using RGLR and then combined across strata for overall estimation and inference using two different weighting schemes. We show through simulations the stratified RGLR approach delivers smaller bias and higher efficiency than the commonly used stratified Cox model analysis in small samples, notably so when the assumption of a constant hazard ratio across strata is violated. A dataset is used to illustrate the utility of the proposed new method.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1

INTRODUCTION

Most statistical methods are developed based on large sample considerations. In real data applications, however, clinical trials and epidemiological studies with time-to-event outcomes do not always meet the large sample size requirement due to various reasons. In fact, trials with small samples are often encountered, due to reasons such as rarity of the disease and nature of the trial design (e.g., an early phase or pilot clinical trial). Commonly used methods in a typical survival analysis include Cox proportional hazards model (Cox, 1972) and parametric regression. Inference is typically based on large sample theory, and may not be appropriate under small samples. Crossover trials also often have small samples, as every subject serves as his or her own control to help achieve higher efficiency than parallel designs. However, there is a lack of existing methods for analyzing crossover studies with time-to-event outcomes. We focus on three specific types of designs with time-to-event response variables and small sample sizes: two-period two-treatment crossover trials, two-group parallel designs in the setting of proportional hazards without stratification, and two-group parallel designs with stratification and proportional hazards.

1.1. Background

1.1.1. Crossover Designs

Crossover designs compare treatment effects on the same subject over different treatment periods. For trials with limited recruitment, crossover designs are ideal to use for higher efficiency than parallel designs. Under the commonly used two-period, two-treatment (2×2) crossover designs, patients are randomized to one of two sequences, AB or BA, where A and B are the treatment labels. A 'washout' period is included between the two periods to ensure no carry-over effects. The use of a period-specific baseline measurement, which is taken before the subject is given the treatment in each period, has been shown to increase statistical power (Chen, Meng, and Zhang, 2012; Kenward and Roger, 2010; Senn, 2002). There are many existing methods for handling baseline information in the analysis of crossover trials with *continuous* endpoints, including ignoring baseline measurements, analyzing the change from baseline, using a function of the baselines as a covariate, and joint-modeling of baseline and post-treatment responses (Chen, Meng, and Zhang,

2012; Hills and Armitage, 1979; Kenward and Roger, 2010; Metcalfe, 2010; Yan, 2013). Some recommended methods for providing higher efficiency than others include analysis of covariance (ANCOVA) and joint-modeling of the within-subject difference in treatment responses and difference in baseline responses (Mehrotra, 2014). The commonly used method, analysis of the change from baseline, suffered from poor efficiency, as also discussed by Kenward and Roger (2010) and Metcalfe (2010).

All methods mentioned above are for continuous endpoints, but crossover trials with time-to-event endpoints are also commonly encountered in research. Existing literature for examining treatment differences in crossover trials with censored time-to-event endpoints includes both regression-based and test-based approaches. France, Lewis, and Kay (1991) used stratified Cox regression, where each subject was treated as a stratum. Feingold and Gillespie (1996) proposed an approach based on the generalized Wilcoxon test. More recently, Brittain and Follmann (2011) proposed a hierarchical rank (H-R) test, where each patient is assigned a rank based on whether and when he or she has an event. The first ordering of the rank is determined by whether the individual has an event during any of the two periods, and the second ordering of the rank is based on the times of the events. A two-group Wilcoxon test is performed using the assigned ranks to test for the treatment effect. Brittain and Follmann (2011) showed that the H-R test has similar or greater power than both the Feingold and Gillespie's method and stratified Cox method under certain censoring patterns. However, none of these approaches utilizes baseline information, and whether and how to utilize baseline information in crossover trials with time-to-event outcomes has not been studied.

1.1.2. Parallel Designs under Proportional Hazards without Stratification

In parallel designs with time-to-event outcomes comparing two treatment groups, the parameter of interest is usually the hazard ratio, also referred to as relative risk. Under the proportional hazards assumption, i.e., the hazard ratio is constant throughout time, it is conventional to use the Cox proportional hazards model for estimation of relative risk and the log-rank test for hypothesis testing (Cox, 1972). However, Cox regression is a large sample method, and may not provide an appropriate result in small samples. Johnson et al. (1982) investigated the Cox model with one binary indicator as the covariate, and found that in small samples, there are non-trivial differences between the actual and asymptotic formula-based variances for the estimated log(hazard ratio). Another commonly used method is parametric regression, but parametric models are subject to

bias when the true underlying survival distribution is misspecified. Therefore, it is important to study analysis methods for failure time data in small samples, which are quite common in real data applications. Early phase clinical trials usually have less than 100 subjects per treatment group (Pocock, 1983), and cancer trials might have limited recruitment as well if the disease is rare.

Mehrotra and Roth (2001) proposed a method based on a generalized log-rank (GLR) statistic for the 2-group comparison to improve estimation and inference of hazard ratio in small sample studies. They showed that even though asymptotically the GLR method has similar performance to the Cox approach, when the sample size is small, GLR is notably more efficient than the Cox approach, in terms of mean squared error (MSE) for the log relative risk when there are no tied event times. A deficiency in the GLR method is that it uses an unnecessary approximation involving nuisance parameters which contributes to bias in the estimated hazard ratio. Furthermore, estimation of the nuisance parameters follows a non-intuitive path. These collectively offer opportunities to improve the GLR method in a tangible way with practical benefits.

1.1.3. Parallel Designs with Stratification and Proportional Hazards

In parallel designs comparing two treatments, when the risk of having an event is known to be affected by a prognostic factor, such as gender or race, stratification is employed in the design stage. Subjects are first divided into each stratum based on his or her prognostic factor characteristics, and then within each stratum, randomized to receive one of the treatments. The goal is estimation and inference involving the true 'overall' hazard ratio, defined as the exponent of the weighted mean of the stratum-specific true log hazard ratios using population relative frequency weights. Under the assumption of proportional hazards in each stratum, stratified Cox model is used to analyze such data. The stratified Cox model assumes that the hazard ratio is constant across strata, which is not always true. If there exists a stratum-treatment interaction, the conventional stratified Cox model tends to provide biased and less efficient results. Mehrotra, Su, and Li (2012) proposed a two-step stratified Cox approach to allow for different hazard ratios across strata, and combine the stratum-specific log hazard ratio by two weighting options, sample size weights and minimum risk weights (Mehrotra and Railkar, 2000). The two-step analysis provides comparable power to the one-step Cox analysis when there is no stratum-treatment interaction, but notably higher power when there is an interaction. It also delivers a point estimator for the overall treatment effect with very small bias. However, the Mehrotra, Su, and Li (2012) approach is based on large sample theory and

hence not ideal for small studies.

1.2. Novel Developments

In this dissertation, we focus on developing methods for analyzing studies with time-to-event outcomes in small samples, and specifically in three settings: 2×2 crossover designs with baseline measurements, parallel designs for two group comparison under proportional hazards without stratification, and stratified parallel designs under proportional hazards.

In Chapter 2, we propose a regression-based method using multiple imputation (MI) of censored event times in conjunction with analysis of covariance (ANCOVA) to incorporate baseline measurements into the analysis of crossover studies with time-to-event outcomes. In the imputation step, we propose to fit multiple candidate survival models, and use frequentist model averaging to pool the final results. Unlike Bayesian model averaging (Bates and Granger, 1969; Raftery, Madigan, and Hoeting, 1997), which requires setting a prior probability to each candidate model, frequentist model averaging does not need any priors (Buckland, Burnham, and Augustin, 1997; Burnham and Anderson, 2003; Hjort and Claeskens, 2003). The final point estimator is obtained by averaging across the imputations and a variance estimator is created that accounts for the uncertainty from both model averaging and imputation. We show that there can be a great efficiency gain in using baseline information for time-to-event endpoints in crossover trials, compared to H-R test and stratified Cox model. Furthermore, our proposed method delivers a point and confidence interval estimate with small-to-no-bias of the treatment-ratio of geometric mean event times.. We demonstrate the impressive performance of our proposed method through simulation studies and apply it to data from a crossover trial studying a new drug's effect on delaying a symptomatic cardiac-related event during a 10-minute treadmill walking test.

In Chapter 3, we focus on the parallel design for the two group comparison without stratification and propose a refined GLR (RGLR) method by replacing the 'approximate' nuisance parameters with 'exact' counterparts in the original GLR statistic. We develop the RGLR statistic for settings with and without tied event times, and show through extensive simulations that our proposed RGLR method provides notably smaller bias than GLR, Cox and parametric models, and provides a high relative efficiency and adequate 95% confidence interval coverage rate. We also provide further insights by developing an alternate and intuitive approach to estimate the nuisance parameter in

the GLR statistic. Finally, we illustrate the method in two clinical trials studying lung cancer and bladder cancer.

We further extend the RGLR method to allow for stratification factors in the analysis of a parallel two-group design in Chapter 4. Instead of assuming a constant hazard ratio across all strata as done by the commonly used stratified Cox model analysis, we allow the hazard ratio to vary across stratum, and use two weighting schemes to combine the stratum-specific estimates of the log(hazard ratio). We show through simulations that RGLR-based estimators provide smaller bias and higher relative efficiency than both the conventional one-step and the two-step stratified Cox model estimator. We apply the proposed method to a simulated data example for illustration.

We provide concluding remarks in Chapter 5 and discuss potential directions for future research.

CHAPTER 2

INCORPORATING BASELINE MEASUREMENTS IN CROSSOVER TRIALS WITH TIME-TO-EVENT ENDPOINTS

2.1. Introduction

Crossover designs are commonly seen in clinical trials to compare the treatment effects on the same subject over different treatment periods. For trials with limited recruitment, crossover designs are ideal to use for higher efficiency than parallel designs. The ability of each person to serve as his or her own control also mitigates the influence of potential confounding factors. In commonly used two-period, two-treatment (2×2) crossover designs, subjects are randomized to one of two sequences, AB or BA, where A and B are the treatment labels. A 'washout' period is included between the two periods to ensure no carry-over effects. The use of a period-specific baseline measurement, which is taken before the subject is given the treatment in each period, is often considered. However, whether and how to use a baseline measurement is often challenging, given the extra cost and the need to determine which statistical methods can be used to fully utilize the information from the baselines. For a 2×2 crossover trial, each subject has four responses: baseline (i.e., pre-treatment) in period 1, post-treatment in period 1, baseline in period 2 and post-treatment in period 2. There are many existing methods for handling baseline information in the analysis of crossover trials with *continuous* endpoints, including ignoring baseline measurements, analyzing the change from baseline, using a function of the baselines as a covariate, and joint-modeling of baseline and post-treatment responses (Chen, Meng, and Zhang, 2012; Hills and Armitage, 1979; Kenward and Roger, 2010; Metcalfe, 2010; Senn, 2002; Yan, 2013).

Mehrotra (2014) evaluated and compared 13 different methods for analyzing 2×2 crossover trials to incorporate baseline measurements with continuous endpoints. Among all the competing methods, two methods were shown to have the highest efficiency: analysis of covariance (ANCOVA) with the within-subject difference in baseline responses used as a covariate, and joint-modeling of the within-subject difference in treatment responses and difference in baseline responses. The commonly used method, analysis of the change from baseline, was shown to have poor efficiency, as also discussed by Kenward and Roger (2010) and Metcalfe (2010).

All methods mentioned above are for continuous endpoints, but crossover trials with time-to-event endpoints are also commonly encountered in research. For example, blood thinners like Warafin are important in preventing outcomes such as blood clots and stroke, but can also induce undesirable increases in bleeding time from simple cuts or other injuries. In this setting, researchers are sometimes interested in studying the effect of an experimental anticoagulant drug on bleeding time using a crossover design with a baseline measurement at the beginning of each period. Kimchi et al. (1983) and Markman et al. (2015) both studied a drug's effect in a crossover trial with a time-to-event outcome and collected baseline meausurements. However, neither incorporated the baseline information into their analysis. Our motivating data example is a crossover trial studying a drug's effect in preventing cardiac-related symptoms in a treadmill walking test. The outcome of interest for each subject is time to a specific cardiopulmonary event, with the outcome recorded as '$>$10 minutes' (i.e., right censored) if the event has not yet occurred after 10 minutes of observation. Existing literature for examining treatment differences in crossover trials with censored time-to-event endpoints includes both regression-based and test-based approaches. France, Lewis, and Kay (1991) used a stratified Cox regression, where each subject was treated as a stratum. Feingold and Gillespie (1996) proposed an approach based on the generalized Wilcoxon test. More recently, Brittain and Follmann (2011) proposed a hierarchical rank (H-R) test, which they showed to have similar or greater power than both the Feingold and Gillespie's method and stratified Cox method under certain censoring patterns. The main idea behind the H-R test is that avoiding an event is more clinically meaningful than delaying an event. Therefore, each patient is assigned a rank that orders how much better an individual does on the novel treatment. The first order of ranking is based on whether patients have an event, and second order of ranking is based on the times of the events. Patients who do not have an event on either treatment receive the same rank. With assigned ranks for everyone, a two-group Wilcoxon test is then performed to test for a treatment effect. However, none of these Wilcoxon-type approaches utilizes baseline information.

In this research, we propose a regression-based method using multiple imputation (MI) of censored values and analysis of covariance (ANCOVA) to incorporate baseline measurements into the analysis of 2$\times$2 crossover studies with censored time-to-event response outcomes. There is often uncertainty about the true underlying survival distribution in real data applications, and misspecification of the distribution can lead to a biased point estimator and/or inefficient analysis. To mitigate this issue, we propose to fit multiple survival models in the imputation step, and use fre-

quentist model averaging to pool the final results from the ANCOVA step. Unlike Bayesian model averaging (Bates and Granger, 1969; Raftery, Madigan, and Hoeting, 1997), which requires setting a prior probability for each candidate model, frequentist model averaging does not require any priors (Buckland, Burnham, and Augustin, 1997; Burnham and Anderson, 2003; Hjort and Claeskens, 2003). To implement model averaging in the presence of multiple imputation, we need to account for both the uncertainty from model averaging and imputation.

We show that there is a great efficiency gain in using baseline information for time-to-event endpoints in crossover trials compared to the H-R test and stratified Cox model. Furthermore, our proposed method is also able to provide a point and confidence interval estimate of a meaningful parameter of interest (treatment-ratio of geometric mean event times). Section 2.2 presents details of the proposed method. In Section 2.3, we contrast the numerical performance of our proposed method with that of the H-R test and stratified Cox model through simulation studies. Section 2.4 includes results from applying the different methods to our motivating real data example. Section 2.5 includes conclusions.

2.2. Methods

We consider a 2×2 crossover trial with two treatments, denoted by A and B. Subjects are randomized to either the AB or BA sequence, with a wash-out period between period 1 and 2. Let X_{ijk} and Y_{ijk} denote baseline and post-treatment event times, respectively, for subject j from sequence k in period i, where $i = 1, 2$; $j = 1, 2, \ldots, n$; and $k = 1, 2$. It is sufficient to assume that, after a log transformation, $(X_{1j1}, Y_{1j1}, X_{2j1}, Y_{2j1})^T$ and $(X_{1j2}, Y_{1j2}, X_{2j2}, Y_{2j2})^T$ follow a multivariate distribution with different means and same variance-covariance structure Σ. We assume there is no censoring at baseline, and in each period, subjects without a post-treatment event are censored at the end of period, denoted by time τ.

We propose a three-step procedure using multiple imputation and ANCOVA to estimate the ratio of geometric means of the event times for treatment A relative to B, denoted as θ, and test the null hypothesis $H_0 : \theta = 1$. For distributions that are symmetric on the log scale, the geometric mean is equivalent to the median. Thus, our parameter of interest can be used to approximate the ratio of median survival of the two treatments, which is commonly of interest in survival analysis. To implement our proposed method, we perform the following steps for each imputation iteration,

details of which are given in the sub-sections below.

Step 1: Fit two candidate parametric event models, log-normal and Weibull, to impute the post-treatment censored values sequentially, conditioning on the baseline event time in period 1 for period 1 imputation, and both baseline event times and post-treatment event time in period 2 for period 2 imputation.

Step 2: With the completed dataset from each candidate model, perform ANCOVA on the log-transformed event times to estimate $\log\theta$ and obtain its standard error.

Step 3: Average across the $\log\theta$ estimates based on weights associated with Akaike information criterion (AIC) from each parametric model fit to get a model averaged estimate and standard error, and synthesize for overall point and confidence interval estimation across the multiply imputed datasets using Rubin's rule.

It is important to note that although we consider only two distributions in Step 1, our method can be easily generalized to include more pre-specified candidate models in the imputation step. We chose log-normal and Weibull because they are very flexible and in our experience provide reasonable fitting models for capturing commonly seen event time data. Through numerical studies in Section 2.3, we show that even averaging over a small number of models can deliver a good performance.

2.2.1. Imputation

We generate M imputed data sets for each candidate model. Let $Z_{ijk} = 0, 1$ denote treatment A and B, respectively for subject j in period i and sequence k. We impute the censored values in period 1 first, and then impute the censored values in period 2. In the m-th imputed data set, we use the baseline value in period 1 and treatment indicator, Z_{1jk}, as covariates and fit two candidate parametric survival models, log-normal and Weibull, respectively, to Y_{1jk}.

$$\log Y_{1jk} = \beta_{s,0} + \beta_{s,1} Z_{1jk} + \beta_{s,2} U_{s,1jk} + \sigma_{s,1} W_{s,1jk}, \tag{2.1}$$

where $s = 1, 2$ denotes the log-normal and Weibull model, respectively, $W_{s,1jk}$ is the error distribution and $U_{s,1jk}$ is the baseline covariate in the s-th model. $W_{1,1jk}$ has the standard normal distribution for the log-normal distribution and $W_{2,1jk}$ has the standard extreme value distribution for the Weibull distribution. $U_{1,1jk} = \log X_{1jk}$ for the log-normal distribution, and $U_{2,1jk} = X_{1jk}$ for

the Weibull distribution; sample R code for implementation is provided in Appendix A. Equation (2.1) is a representation of the log-normal and accelerated failure time (AFT) model framework for the Weibull model that highlights the common linear regression model on the log-scale. For fitting the parametric model, we analyze the log event times for the log-normal model and fit the traditional Weibull model for the event times on the original scale. We use robust sandwich standard errors in both candidate models to correct for potential model misspecification.

Let $\hat{\boldsymbol{\beta}}_s = (\hat{\beta}_{s,0}, \hat{\beta}_{s,1}, \hat{\beta}_{s,2}, \hat{\sigma}_{s,1})^T$ and $\hat{\Sigma}_s$ denote the estimated coefficients and variance-covariance matrix in the s-th candidate model, respectively. We draw $\hat{\boldsymbol{\beta}}_s^*$ from a multivariate normal distribution $N(\hat{\boldsymbol{\beta}}_s, \hat{\Sigma}_s)$. For subject with a censored post-treatment value, we then impute a right-censored value with an uncensored value by using $\hat{\beta}_s^*$, treatment indicator Z_{1jk} and subject-specific period 1 baseline values in equation (2.1). The corresponding uncensored post-treatment values in period 1 are denoted by $Y_{s,1jk}^{(m)}$.

Now, with complete data in period 1, we can then use the observed/imputed post-treatment values in period 1, baseline values in both period 1 and period 2 as covariates, to impute post-treatment censored values in period 2 by fitting the s-th model,

$$\log Y_{2jk} = \alpha_{s,0} + \alpha_{s,1} Z_{2jk} + \alpha_{s,2} U_{s,1jk} + \alpha_{s,3} V_{s,2jk} + \alpha_{s,4} R_{s,1jk}^{(m)} + \sigma_{s,2} W_{s,2jk}, \tag{2.2}$$

where $U_{1,1jk} = \log X_{1jk}, V_{1,2jk} = \log X_{2jk}, R_{s,1jk}^{(m)} = \log Y_{s,1jk}^{(m)}$ for log-normal distribution, and $U_{1,1jk} = X_{1jk}, V_{2,2jk} = X_{2jk}, R_{s,1jk}^{(m)} = Y_{s,1jk}^{(m)}$ for Weibull distribution, and Z_{2jk} is the treatment indicator in period 2.

The imputation procedure described above for period 1 is now implemented using random draws from the assumed multivariate normal distribution of the vector of estimated regression coefficients in equation (2.2) for each of the two parametric models. The corresponding uncensored post-treatment values in period 2 are denoted by $Y_{s,2jk}^{(m)}$.

2.2.2. ANCOVA

After each imputation, we have two sets of complete data on every subject from the two candidate models, log-normal and Weibull. Each imputed dataset is analyzed using ANCOVA on the log-transformed event times. Specifically, we regress the difference between post-treatment event

times, $\Delta_{s,jk}^{(m)} = \log Y_{s,1jk}^{(m)} - \log Y_{s,2jk}^{(m)}$ on the difference between baseline measurements, $D_{jk} = \log X_{1jk} - \log X_{2jk}$ and the sequence indicator Q_j.

$$\Delta_{s,jk}^{(m)} = \gamma_{s,0} + \gamma_{s,1} D_{jk} + \gamma_{s,2} Q_j + \epsilon_{s,jk}, \tag{2.3}$$

where $\epsilon_{s,jk} \sim N(0, \eta^2)$.

The point estimator from the s-th model in the m-th imputed data set is $\log \hat{\theta}_s^{(m)} = \hat{\gamma}_{s,2}^{(m)}/2$, which is the logarithm of the ratio of geometric means for treatment A relative to B. The corresponding variance estimate for $\log \hat{\theta}_s^{(m)}$ from the s-th model in the m-th imputed data set is $\hat{v}_s^{(m)}$.

2.2.3. Model Averaging and Rubin's Combination Rule

For overall estimation and inference, we first combine the two estimators from the candidate models in each imputed data set, then pool the model-averaged estimators from all the imputed data sets and obtain the pooled variance estimate that accounts for both the uncertainty from model averaging and imputation (Schomaker and Heumann, 2014).

For model averaging, we need to assign a standardized weight. There are many different options for the choice of weights, including an information criterion (Buckland, Burnham, and Augustin, 1997), Mallows' criterion (Hansen, 2007; Mallows, 1973) and cross-validation criterion (Hansen and Racine, 2012). We propose to use the straightforward and commonly used Akaike Information Criterion (AIC) (Akaike, 1974) to assign weights. Let $\boldsymbol{I}_s$ denote the AIC for the ANCOVA regression, equation (3), from the s-th candidate model, then the weight is defined as (Buckland, Burnham, and Augustin, 1997)

$$w_s = \frac{\exp(-\boldsymbol{I}_s/2)}{\sum_{i=1}^{2} \exp(-\boldsymbol{I}_i/2)}.$$

The model averaged estimator in the m-th imputed data set is $\log \hat{\theta}^{(m)} = \sum_{s=1}^{2} w_s \log \hat{\theta}_s^{(m)}$, and the variance for the model averaging estimator is estimated by (Buckland, Burnham, and Augustin, 1997)

$$\hat{\text{Var}}(\log \hat{\theta}^{(m)}) = \left[\sum_{s=1}^{2} w_s \sqrt{\hat{\text{Var}}(\log \hat{\theta}_s^{(m)}) + (\log \hat{\theta}_s^{(m)} - \log \hat{\theta}^{(m)})^2} \right]^2. \tag{2.4}$$

Now, we can pool the model averaged estimators across the M imputed data sets, with the final

estimator calculated as (Schomaker and Heumann, 2014)

$$\log \bar{\hat{\theta}} = \frac{1}{M} \sum_{m=1}^{M} \log \hat{\theta}^{(m)}. \tag{2.5}$$

When there is no model averaging, we can use Rubin (1987) to combine the results from multiple imputation. As noted earlier, with the presence of model averaging, the uncertainty from both model averaging and imputation needs to considered. The between-imputation variance is

$$v_{btw} = \frac{1}{M-1} \sum_{m=1}^{M} (\log \hat{\theta}^{(m)} - \log \bar{\hat{\theta}})^2.$$

The within-imputation variance is the average of the estimated variance from equation (4) across M imputed data sets

$$v_{within} = \frac{1}{M} \sum_{m=1}^{M} \hat{\text{Var}}(\log \hat{\theta}^{(m)}).$$

Therefore, the total variance of the estimator after multiple imputation is (Schomaker and Heumann, 2014)

$$v_{total} = \frac{M+1}{M(M-1)} \sum_{m=1}^{M} (\log \hat{\theta}^{(m)} - \log \bar{\hat{\theta}})^2 + \frac{1}{M} \sum_{m=1}^{M} \left[\sum_{s=1}^{2} w_s \sqrt{\hat{\text{Var}}(\log \hat{\theta}_s^{(m)}) + (\log \hat{\theta}_s^{(m)} - \log \hat{\theta}^{(m)})^2} \right]^2. \tag{2.6}$$

To test the null hypothesis $H_0 : \theta = \theta_0$ (with $\theta_0 = 1$ in our application), we carry out a t-test with test statistic $(\log \bar{\hat{\theta}} - \log \theta_0)/\sqrt{v_{total}}$. To calculate the degrees of freedom d^* for the t-test, we follow Barnard and Rubin (1999) so that $d^* = (1/d + 1/\hat{d}_{obs})^{-1}$, where $d = (M-1)[1 + \frac{v_{within}}{(1+1/M)v_{btw}}]^2$ and $\hat{d}_{obs} = (1-(1+1/M)v_{btw}/v_{total})(\frac{d_{com}+1}{d_{com}+3})d_{com}$, and d_{com} is the degrees of freedom for $\bar{\hat{\theta}}$ when there are no missing values.

2.3. Simulation

2.3.1. Simulation Set-up

To compare the performance of our proposed approach to the H-R test and stratified Cox model, we carried out a simulation study to examine type I error and power among all three methods. Since our method utilized baseline information, we also included the period-specific baseline event

times, in addition to the treatment indicator, as covariates in the stratified Cox model to make a fair comparison. The H-R test, however, does not incorporate baseline information, and thus, we used the method as is. We also examined the bias and 95% confidence interval (C.I.) coverage probability from our proposed estimator; of note, the other two methods cannot deliver an estimate of our parameter of interest (θ).

We simulated three underlying distributions for event times, namely log-normal, exponential and gamma. Two of the distributions, log-normal and exponential (a special case of the Weibull), are included in the candidate models in our method, while the gamma distribution is not. The density curves for each of the three distributions are shown in Supplementary Figure A.1 in Appendix A. Under the log-normal distribution, for each of the N subjects in sequence AB and BA, we generated correlated log event times from a multivariate normal distribution with mean parameter $(0, \log\theta, 0, 0)^T$ for AB sequence and $(0, 0, 0, \log\theta)^T$ for BA sequence and common variance-covariance structure with common variance 1 and correlation coefficients $\rho_{12}, \rho_{13}, \rho_{14}, \rho_{23}, \rho_{24}, \rho_{34}$. We considered three correlation structures, compound symmetry (CS), first-order autoregressive (AR(1)), and equipredictability (EP), where $\rho_{12} = \rho_{13} = \rho_{14} = \rho_{23} = \rho_{24} = \rho_{34} = \rho$ for CS, $\rho_{12} = \rho_{23} = \rho_{34} = \rho, \rho_{13} = \rho_{24} = \rho^2, \rho_{14} = \rho^3$ for AR(1), and $\rho_{23} = \rho_{14}, \rho_{24} = \rho_{13}, \rho_{34} = \rho_{12}$ for EP. The correlation structures are as follows:

$$\Sigma_{CS} = \begin{pmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{pmatrix} \quad \Sigma_{AR} = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{pmatrix} \quad \Sigma_{EP} = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} \\ \rho_{12} & 1 & \rho_{14} & \rho_{13} \\ \rho_{13} & \rho_{14} & 1 & \rho_{12} \\ \rho_{14} & \rho_{13} & \rho_{12} & 1 \end{pmatrix}.$$

We assumed no censoring in baseline event times in each period, and the post-treatment event times were right-censored at time τ. As discussed in the previous section, the parameter of interest θ is the ratio of the geometric means of the event times for treatment A and treatment B, and under the log-normal distribution, it is equivalent to the ratio of median event times.

For the exponential distribution, we used copulas (Sklar, 1973) to generate correlated event times from a multivariate exponential with mean $(2, 2\theta, 2, 2)^T$ for AB sequence and $(2, 2, 2, 2\theta)^T$ for BA sequence and common variance-covariance structure and correlation coefficients as specified above. Note that the ratio of arithmetic means is equivalent to the ratio of geometric means under expo-

nential distribution. Since copulas only preserves the rank correlation coefficient but not the linear correlation coefficient (Genest and MacKay, 1986), the correlated exponential data follows approximately, but not exactly, the specified variance-covariance structure.

To further illustrate the performance of our proposed method, we also considered an underlying gamma distribution, which is not included in our two candidate models from the imputation step. Specifically, we used a gamma distribution with scale of 0.7 and shape of 2 for subjects in treatment B. Event times for subjects in treatment A was generated from a gamma distribution with scale of 0.7θ and shape of 2. We again used copulas to generate the correlated event times. For AB sequence, the simulated event times followed a multivariate gamma distribution with mean $(1.4, 1.4\theta, 1.4, 1.4)^T$, and for BA sequence, the event times follows a gamma distribution with mean $(1.4, 1.4, 1.4, 1.4\theta)^T$. Note that it can be shown that the ratio of arithmetic means is equivalent to the ratio of geometric means in the setting that event times in treatment A and B follow a gamma distribution with the same shape parameter and ratio of scale parameter of θ. Again, the event times in the two sequences followed a common variance-covariance structure and correlation coefficients as specified above.

We varied the sample size, percentage of censoring, θ, correlation structure, and compared the performance of the different methods. Sample size per sequence was varied as $N = 12, 24, 48$, and percentage of censoring was controlled by changing the time τ, to generate 10% and 50% censoring for the total sample.

The mean pairwise correlation coefficient $\bar{\rho}$ took values of 0.5 and 0.7. Under CS, $\rho = \bar{\rho}$. For AR(1), $\rho = 0.7$ for $\bar{\rho} = 0.5$ and $\rho = 0.83$ for $\bar{\rho} = 0.7$. For EP, we set $\rho_{12} = 0.6, \rho_{13} = 0.5, \rho_{12} = 0.4$ when $\bar{\rho} = 0.5$, and $\rho_{12} = 0.8, \rho_{13} = 0.7, \rho_{12} = 0.6$ when $\bar{\rho} = 0.7$. We generated $M = 50$ imputed datasets within each of the 5000 replications. Under the null hypothesis, $\theta = 1$. Under the alternative hypothesis, we chose a value of θ such that the power was about 80% for the H-R test, given the true underlying distribution, $\Sigma, \bar{\rho}$ and percentage censoring.

2.3.2. Simulation Results

Table 2.1 reports type I error for the three distributions for the H-R test, stratified Cox model with baseline adjustment and our proposed multiple imputation and model averaging and ANCOVA method. As shown in Table 2.1, the stratified Cox model analysis had non-converge (NC) issues

under several scenarios when the sample size was 12 and 24 subjects per sequence with 50% censoring, and had an inflated type I error when there were 24 subjects per sequence with 10% censoring, $\bar{\rho} = 0.5$ and CS structure under exponential distribution. When the true distribution was gamma, the stratified Cox model analysis was associated with inflated type I error under CS structure with 24 subjects per sequence and $\bar{\rho} = 0.7$, 10% censoring, and with 48 subjects per sequence and $\bar{\rho} = 0.7$, 50% censoring. The H-R test and our proposed model averaging method controlled type I error throughout all the scenarios considered. Table 2.1 also reports the bias in the estimate of $\log \theta$ using our proposed method under the null hypothesis. The bias was negligible under all simulated scenarios.

Figures 2.1, 2.2 and 2.3 show the power for the three different methods for $N = 24$ subjects per sequence and different combinations of percentage censoring and variance-covariance structure under the log-normal, exponential and gamma distributions, respectively; results for other sample sizes are provided in Appendix A.

As shown in Figure 2.1, when the true distribution was log-normal, our proposed method always provided a higher or similar power than the H-R test and stratified Cox model. For cases where the H-R test or stratified Cox failed to deliver 80% power, our method was able to achieve power close to or above 80%. The increase in power using our method was more significant under AR(1) and EP structures than under CS structure. The power gain compared to the H-R test likely comes from the fact that the H-R test fails to utilize baseline information. Likewise, our proposed method has a substantially higher power than the stratified Cox model that adjusts for baseline covariates in part because our method makes better use of the baseline information. In addition, the model averaging aspect provides the flexibility of assuming more than one distribution and further improves the efficiency of the analysis. Results from assuming only one distribution, either log-normal or Weibull, is more prone to model misspecification in the imputation step.

Figure 2.2 displays the results when the true distribution was exponential. In this case, the true variance-covariance structure and percentage censoring affected the relative performance of the considered methods. When the true structure was CS, H-R test delivered higher power than the other considered methods. Of note, CS structure usually does not capture the true correlation pattern in most real data examples, since it assumes equal correlation among all pairs of with-subject event times, which has low plausibility. When the true structure was AR(1) or EP, which

are a more realistic representation of the correlation structure in real data applications, our method again showed a substantial power gain compared to the H-R test and stratified Cox model under 50% censoring. When the percentage censoring was 10%, our method delivered similar power as the H-R test. For all the other scenarios, where the stratified Cox model did not have non-convergence issues, our proposed method was consistently more powerful than the stratified Cox model.

Finally, when the underlying distribution was gamma, our proposed method still provided higher power than the stratified Cox model throughout all scenarios, but slightly lower power than the H-R test under CS structures, as shown in Figure 2.3. Under AR and EP structures, using multiple imputation, model averaging and ANCOVA approach delivered a more efficient analysis than both the H-R test and stratified Cox model. Recall that the true distribution, gamma, is not included as one of the candidate models in the imputation step; however, we are still able to provide a comparably efficient result. Additionally, our proposed method is able to provide a point and CI estimate of the treatment effect, while the other two methods do not.

Table 2.2 reports percentage bias and 95% C.I. coverage probability for $\log\theta$ using our proposed method under the alternative hypothesis. Our method was able to control bias within 10% under log-normal and exponential distribution. When the true distribution was gamma, it controlled bias within 10% under 10% censoring, and under 50% censoring, bias was no larger than 11%. Importantly, the 95% C.I. coverage probability was maintained at or above the nominal level under all the scenarios considered.

2.4. Data Application

We apply the three methods considered to a 2×2 crossover clinical trial of an investigation drug. The trial recruited 40 subjects in total, and randomly assigned 20 to the placebo then drug sequence and 20 to the drug then placebo sequence. The outcome variable was time until a symptomatic cardiac-related event of interest during a 10-minute treadmill walking test. Each subject also had a measurement at baseline before taking the treatment. Figure 2.4 displays the Kaplan-Meier curves for post-treatment event times for placebo and drug in period 1 and period 2, separately.

The H-R test delivers a p-value of 0.052, indicating that there is not enough evidence at the two-

Table 2.1: Type I error (target=5%) for the hierarchical rank test (H-R), stratified Cox model with baseline adjustment (SCB) and proposed multiple imputation with model averaging and ANCOVA (MI^{MA}) for log-normal, exponential and gamma distributions under the null hypothesis $H_0 : \theta = 1$ and bias in the estimate of $\log\theta$ using the proposed method (5000 simulations).

			$\overline{\rho} = 0.5$						$\overline{\rho} = 0.7$					
			10% Censoring			50% Censoring			10% Censoring			50% Censoring		
Distribution	Σ	Method \ N/seq	12	24	48	12	24	48	12	24	48	12	24	48
Log-normal	CS	H-R	4.6	4.3	4.9	4.4	4.3	4.8	4.5	5.0	4.3	4.6	4.8	4.8
		SCB	4.5	4.4	4.9	NC	4.4	4.6	4.5	5.0	4.8	NC	4.5	4.9
		MI^{MA}	4.3	4.8	4.9	2.4	3.9	4.9	4.8	4.9	4.8	1.9	3.8	4.2
		Bias	-0.002	-0.002	0.000	0.001	0.002	-0.002	0.005	-0.002	-0.001	0.001	0.003	0.002
	AR(1)	H-R	5.0	4.0	4.8	5.0	4.7	4.8	4.7	4.8	5.0	4.8	4.6	4.8
		SCB	4.2	4.4	4.5	NC	4.6	4.9	3.6	4.2	5.1	NC	4.6	4.7
		MI^{MA}	4.6	4.4	4.7	2.8	3.8	4.8	4.7	4.7	5.0	2.6	4.0	4.5
		Bias	0.002	0.002	0.001	-0.007	0.000	-0.002	-0.001	-0.002	0.001	0.001	-0.002	-0.001
	EP	H-R	4.4	5.1	4.7	4.8	4.4	4.4	4.8	4.3	4.4	4.4	5.1	4.9
		SCB	NC	4.9	4.5	NC	4.7	4.6	NC	3.7	4.3	NC	NC	4.8
		MI^{MA}	4.5	5.0	5.0	2.7	4.4	4.7	4.7	4.4	4.7	1.4	3.2	3.7
		Bias	-0.001	0.001	-0.002	0.001	0.001	0.001	0.001	-0.001	0.001	0.002	-0.000	0.002
Exponential	CS	H-R	4.8	4.8	5.0	4.9	4.5	4.8	4.7	4.6	5.0	4.1	4.3	4.3
		SCB	5.1	(5.6)	4.6	NC	4.7	5.0	4.0	4.7	5.3	NC	4.4	5.0
		MI^{MA}	4.7	5.0	4.4	2.2	3.8	5.1	4.4	4.8	4.5	1.8	2.9	4.0
		Bias	-0.003	-0.002	0.001	-0.071	-0.006	0.002	0.001	-0.001	0.002	0.063	0.002	-0.002
	AR(1)	H-R	4.6	4.6	4.6	4.6	5.0	4.4	4.3	5.0	5.1	4.5	4.5	4.8
		SCB	NC	4.5	5.2	NC	4.7	4.6	NC	4.8	5.2	NC	NC	4.7
		MI^{MA}	4.9	4.7	4.2	2.0	3.5	4.2	4.4	4.5	4.5	1.9	2.9	3.0
		Bias	-0.001	0.004	0.002	0.002	0.001	0.002	-0.001	-0.002	0.002	0.004	-0.001	-0.003
	EP	H-R	4.2	4.3	4.4	4.5	4.2	4.4	4.8	4.7	4.9	4.4	4.5	4.9
		SCB	NC	4.4	4.8	NC	4.3	4.5	NC	NC	4.2	NC	NC	4.2
		MI^{MA}	4.4	4.3	4.5	1.8	3.3	4.3	4.4	4.6	4.6	1.4	2.1	2.4
		Bias	-0.004	0.001	0.001	0.003	0.003	-0.001	0.002	-0.001	0.001	-0.015	0.001	0.001
Gamma	CS	H-R	4.2	4.4	4.7	4.4	4.4	4.6	4.2	4.7	4.4	4.1	4.8	5.1
		SCB	4.2	4.9	4.6	NC	4.6	4.8	NC	(5.9)	4.7	NC	4.9	(5.6)
		MI^{MA}	4.5	4.7	4.9	4.2	4.7	5.0	4.7	5.0	4.8	3.2	4.4	4.8
		Bias	0.001	0.002	-0.002	0.001	-0.000	0.001	0.002	0.002	-0.000	-0.002	-0.003	0.002
	AR(1)	H-R	4.3	4.6	4.4	4.7	5.1	4.5	4.4	4.2	4.3	4.3	4.7	4.8
		SCB	3.7	5.1	4.8	NC	4.1	4.5	NC	4.6	4.1	NC	4.1	5.1
		MI^{MA}	4.7	5.1	4.6	3.8	4.6	4.7	4.9	4.7	4.1	3.3	4.6	4.6
		Bias	-0.000	-0.000	-0.001	-0.006	0.001	0.000	-0.001	0.001	0.001	0.001	0.000	0.001
	EP	H-R	4.9	4.6	5.2	4.7	4.5	4.6	4.7	4.2	4.7	4.6	5.1	4.4
		SCB	4.3	5.0	5.2	NC	4.4	4.5	NC	3.8	5.0	NC	3.7	4.8
		MI^{MA}	4.7	4.3	4.9	3.5	4.7	5.0	4.8	4.7	4.5	3.0	3.9	4.0
		Bias	0.000	0.001	0.001	0.003	-0.000	-0.003	-0.000	-0.001	-0.001	0.001	-0.000	-0.001

Type I error more than $Z_{0.975}$ standard errors above 5% level is in parentheses. NC: non convergence. CS: compound symmetry covariance structure. AR(1): first-order autoregressive covariance structure. EP: equipredicability covariance structure. $\bar{\rho}$: mean pairwise correlation.

tailed 5% level of significance to show a difference between the drug and placebo in delaying the event of interest. On the other hand, stratified Cox model adjusting for period-specific baseline and our proposed method deliver a p-value of 0.020, and 0.005, respectively. The ratio of geometric mean of time to the cardiac-related event for patients taking the drug to patients on placebo was estimated to be 1.67, with 95% C.I of (1.18, 2.35). The raw data from this trial are provided in Table 2.3, and R code used to generate the analysis results for all the three methods are provided in Appendix A.

Table 2.2: Percentage bias and 95% C.I. coverage probability under the alternative hypothesis $H_1 : \theta \neq 1$ for the estimate of $\log \theta$ using the proposed method under log-normal, exponential and gamma distributions (5000 simulations).

			$\overline{\rho} = 0.5$						$\overline{\rho} = 0.7$					
			10% Censoring			50% Censoring			10% Censoring			50% Censoring		
Distribution	Σ	N/seq / Method	12	24	48	12	24	48	12	24	48	12	24	48
Log-normal	CS	%Bias	-4.9	-4.7	-3.2	-8.6	-7.9	-7.2	-4.1	-3.3	-4.0	-8.0	-8.3	-9.5
		Coverage	95.3	94.0	94.3	95.6	94.4	94.4	95.3	94.8	95.1	96.6	95.0	94.9
	AR(1)	%Bias	-4.5	-4.0	-4.1	-9.4	-8.2	-8.5	-2.2	-2.3	-2.5	-5.5	-6.1	-5.5
		Coverage	95.2	94.9	94.6	95.4	94.7	94.4	95.2	95.3	94.5	96.8	95.8	95.4
	EP	% Bias	-4.6	-4.8	-3.5	-9.3	-8.1	-7.3	-2.0	-5.0	-2.2	-5.5	-5.1	-5.7
		Coverage	95.1	94.8	94.7	96.1	94.4	94.4	95.8	96.5	95.3	97.8	96.8	96.0
Exponential	CS	%Bias	2.1	1.1	-0.1	1.4	0.8	0.8	1.7	1.3	0.9	2.4	2.7	3.3
		Coverage	95.0	95.2	94.7	97.0	95.3	95.1	95.0	95.3	94.2	96.6	96.2	95.6
	AR(1)	%Bias	0.6	1.8	0.4	2.2	3.1	2.9	1.4	1.2	1.3	3.9	5.9	5.7
		Coverage	94.8	94.9	95.2	96.3	95.7	95.4	94.9	95.3	95.1	97.1	96.3	95.7
	EP	%Bias	1.9	0.9	-0.1	1.5	1.8	1.4	1.8	1.7	2.4	3.6	4.1	4.1
		Coverage	95.3	95.2	94.8	96.6	95.1	95.0	95.1	95.2	95.0	97.5	97.0	97.1
Gamma	CS	%Bias	-5.4	-3.8	-3.6	-8.5	-3.3	-4.2	-3.9	-3.8	-3.0	-11.5	-5.1	-3.8
		Coverage	95.0	94.9	95.0	94.9	94.7	94.6	95.0	95.6	95.3	94.8	95.5	94.8
	AR(1)	%Bias	-5.2	-3.9	-3.3	-7.0	-4.8	-4.6	-3.7	-2.6	-1.7	-10.4	-10.6	-10.0
		Coverage	95.0	94.6	95.0	95.5	94.5	94.7	95.2	94.9	95.4	95.1	94.6	94.4
	EP	%Bias	-5.5	-3.9	-2.8	-7.4	-4.0	-4.4	-2.9	-2.7	-1.9	-9.9	-9.7	-7.9
		Coverage	94.8	94.9	95.3	95.1	94.7	94.6	95.3	94.7	94.7	95.3	94.5	95.4

CS: compound symmetry covariance structure. AR(1): first-order autoregressive covariance structure. EP: equipredicability covariance structure. $\bar{\rho}$: mean pairwise correlation. True values of θ used for all the simulated scenarios are provided in Table A.1 in Appendix A.

2.5. Discussion

While there are many methods for analyzing crossover trials with continuous endpoints, there are few studying crossover trials with time-to-event outcomes, which are often seen in practice. In this paper, we have proposed a method using multiple imputation, assuming two candidate parametric event time models, to impute censored post-treatment values. For each imputed dataset, ANCOVA, with difference in period-specific baseline responses as a covariate, is applied to log-transformed event times to estimate the log treatment-ratio of geometric means. Frequentist model averaging with AIC weighting in conjunction with Rubin's combination rule for multiple imputation is used for overall estimation and inference. We showed that by utilizing baseline information, our method provided a more efficient or as efficient result than some other existing methods, including H-R test and stratified Cox model, across different combinations of variance-covariance structures, percentage censoring and sample sizes. By using model averaging, we are able to provide a more flexible method than assuming only one distribution in the imputation step, which can be subject to mis-

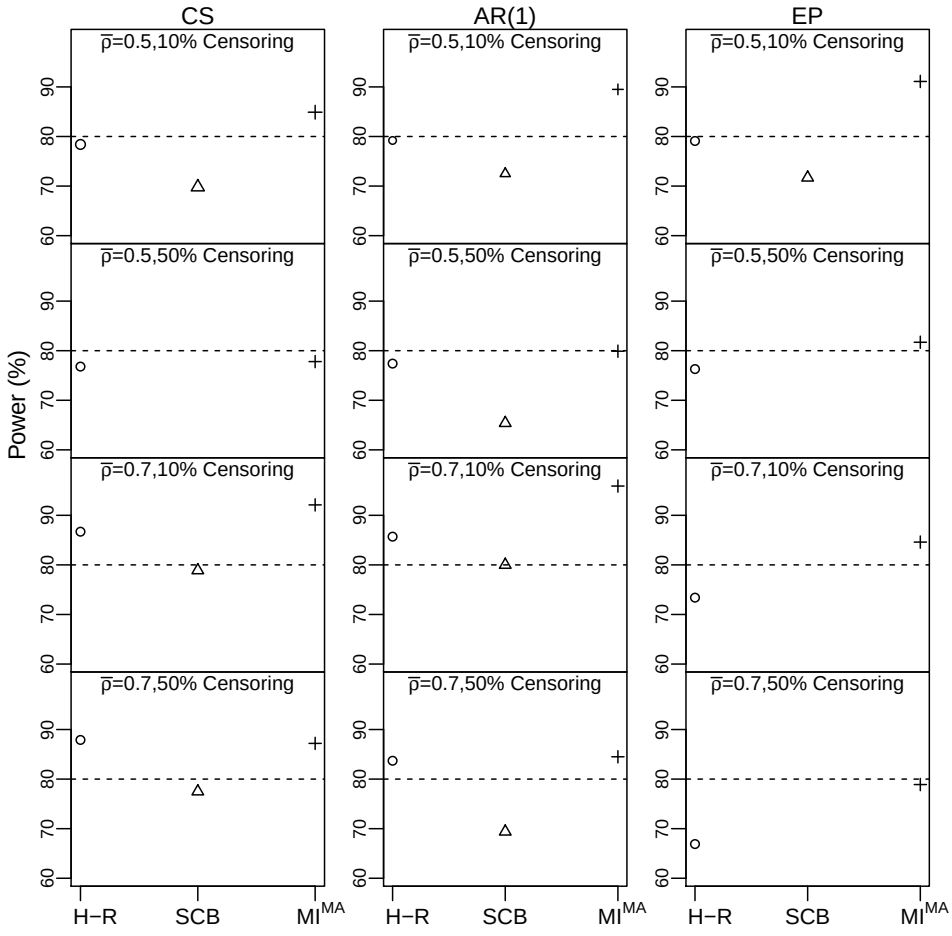


Figure 2.1: Power comparison for the Hierarchical Rank test (H-R), stratified Cox model (SCB) and proposed multiple imputation with model averaging and ANCOVA (MI^{MA}) under a log-normal distribution and varying assumptions for the true variance structure (compound symmetry (CS), first-order autoregressive (AR(1)), equipredictability (EP), mean pairwise correlation of baseline and post-treatment values across the two periods ($\bar{\rho} = 0.5, 0.7$) and percentage censoring (10%,50%), with 24 subjects per sequence. Stratified Cox model had non-convergence issues under CS structure with $\bar{\rho} = 0.5$ and 50% censoring, and under EP structure with $\bar{\rho} = 0.5$ and 50% censoring, $\bar{\rho} = 0.7$ and 10% censoring and $\bar{\rho} = 0.7$ and 50% censoring, and hence power is not reported.

specification of the true underlying distribution. Furthermore, the H-R approach does not provide a point estimator, while our regression-based method delivers an estimated ratio of geometric means of event times for one treatment relative to the other with small or no bias and adequate 95% C.I. coverage. The ratio of geometric means is a useful parameter in that it is equivalent to the ratio of median event times under a log-normal distribution and other distributions that are symmetric on the log-scale.

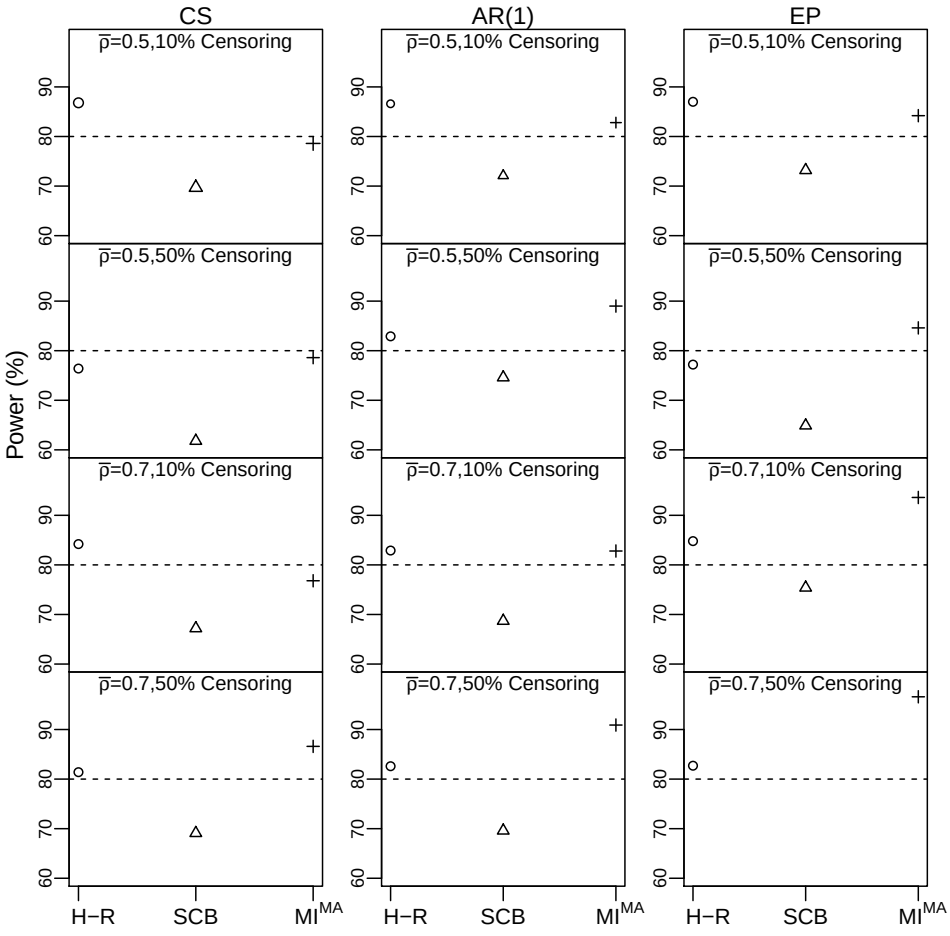


Figure 2.2: Power comparison for the Hierarchical Rank test (H-R), stratified Cox model (SCB) and proposed multiple imputation with model averaging and ANCOVA (MI^{MA}) under an exponential distribution and varying assumptions for the true variance structure (compound symmetry (CS), first order autoregressive (AR(1)), equipredictability (EP), mean pairwise correlation of baseline and post-treatment values across the two periods ($\bar{\rho} = 0.5, 0.7$) and percentage censoring (10%,50%), with 24 subjects per sequence. Stratified Cox model had non-convergence issues under EP structure with $\bar{\rho} = 0.7$ and 50% censoring, and hence power is not reported.

For our model-averaging approach, we only used two candidate models, log-normal and Weibull, to impute censored post-treatment values. More distributions can readily be used. The candidate distributions should include those that cover a spectrum of anticipated plausible shapes of the survival distribution for the outcome of interest. The relative success of our method, like other applications of multiple imputation, is not expected to perform well if the imputation model is grossly misspecified. We showed that using two candidate models provided efficient results with little bias for the

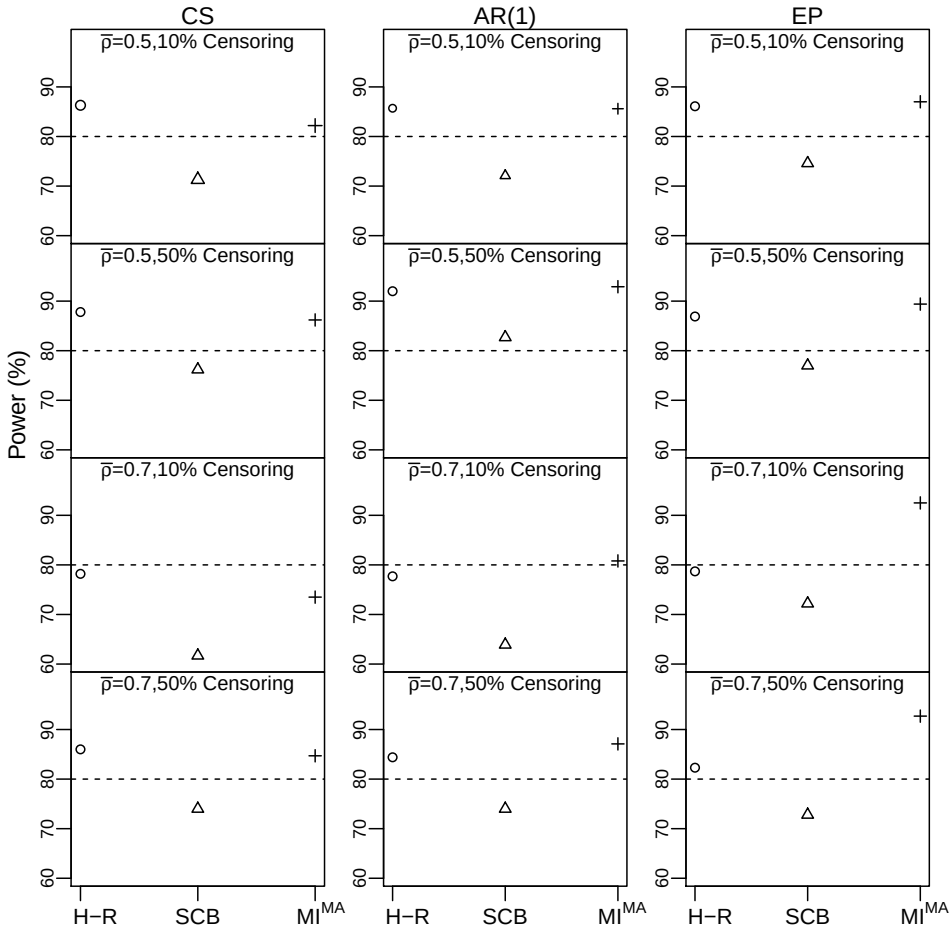


Figure 2.3: Power comparison for the Hierarchical Rank test (H-R), stratified Cox model (SCB) and proposed multiple imputation with model averaging and ANCOVA (MI^{MA}) under a gamma distribution and varying assumptions for the true variance structure (compound symmetry (CS), first order autoregressive (AR(1)), equipredictability (EP), mean pairwise correlation of baseline and post-treatment values across the two periods ($\bar{\rho} = 0.5, 0.7$) and percentage censoring (10%,50%), with 24 subjects per sequence.

settings considered, and thus, more candidate models could potentially improve these results. Although there is no upper limit on the number of models that can be fit, having an unnecessarily large amount of models is also not recommended, as it may increase the overall computation time without improving the power. It is also important to note that in order to properly use the model average approach to combine the parameter estimates, all of the candidate models need to estimate the treatment effect with the same parameter.

Table 2.3: Event times (minutes) for a 10-minute treadmill test in a 2×2 crossover clinical trial.

	Placebo-drug sequence					Drug-placebo sequence			
	Period 1 (placebo)		Period 2 (drug)			Period 1 (drug)		Period 2 (placebo)	
Subject	X_1	Y_1	X_2	Y_2	Subject	X_1	Y_1	X_2	Y_2
1	1.5	1	1	1.5	2	1	1	1	2.5
3	6	4	3.5	>10	4	6	>10	2.5	2.5
5	1	1	1.5	4.5	6	3	2	1	0.5
7	3.5	1.5	0.5	3	8	2.5	2.5	1.5	2
9	0.5	1	3.5	8	10	2	2.5	2.5	3
11	6	10	6	>10	12	1.5	4.5	2.5	1
13	0.5	0.5	1	>10	14	3.5	5.5	4.5	9.5
15	1	1	1	2.5	16	1	2	2	>10
17	1.5	1	0.5	0.5	18	6	>10	5	3.5
19	1	1.5	2	4	20	2	3	1.5	1.5
21	5	5.5	3	1.5	22	1.5	2.5	1.5	0.5
23	2.5	5	6	4.5	24	1.5	3.5	2.5	3
25	5	5.5	4.5	6	26	3.5	9	6	6
27	1	2	2.5	8.5	28	2	5.5	3.5	8
29	5	5.5	3.5	2	30	2.5	2.5	1	0.5
31	0.5	1	2	7.5	32	2.5	3.5	2.5	4
33	5	4	2	2	34	5.5	3	1	0.5
35	0.5	0.5	1	1.5	36	3	5.5	5	0.5
37	1.5	2	3	3	38	0.5	1	1	5.5
39	6	4	1.5	0.5	40	2.5	5	2.5	0.5
Median	1.5	1.75	2	3.5	Median	2.5	3.25	2.5	2.5

X_1: baseline response in period 1. Y_1: post-treatment response in period 1. X_2: baseline response in period 2. Y_2: post-treatment response in period 2.

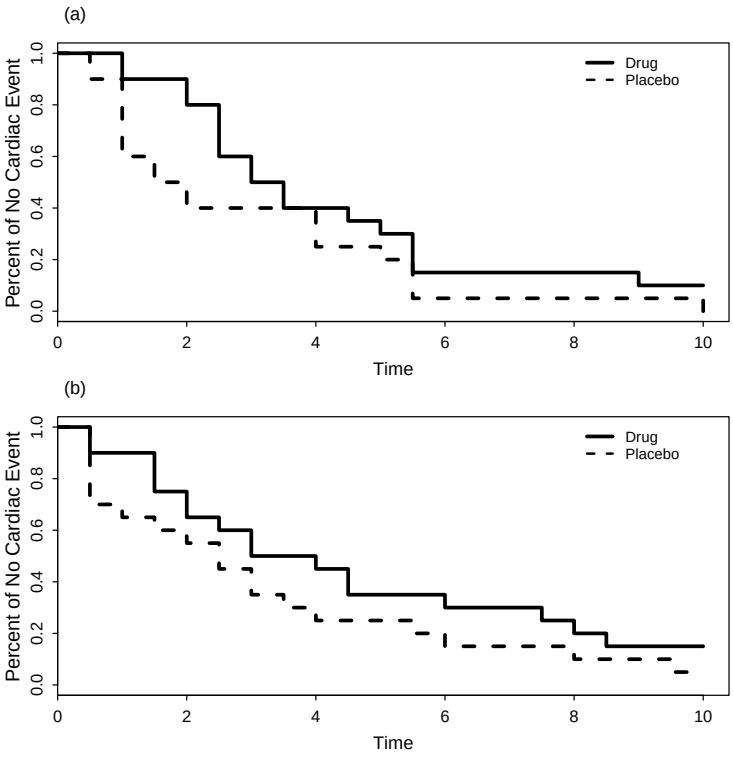


Figure 2.4: Kaplan-Meier curves for the time to a symptomatic cardiac-related event by treatment group from a 2×2 crossover trial; (a) is for period 1 and (b) is for period 2.

CHAPTER 3

HAZARD RATIO ESTIMATION IN SMALL SAMPLES

3.1. Introduction

In a typical survival analysis comparison of two groups, the hazard ratio, often called the relative risk, is generally the focus of inference. If the hazard ratio can be assumed constant throughout time, i.e., if the two groups have proportional hazard functions, it is conventional to use the Cox proportional hazards model for estimation of relative risk and the log-rank test for hypothesis testing; the latter can be derived as a score test via the Cox partial likelihood function (Cox, 1972). However, Cox regression is a large sample method and small sample sizes (10-100 subjects per group) are quite common in real data applications such as early phase clinical trials (Pocock, 1983). Besides randomized clinical trials, observational studies involving a rare disease also often have limited sample sizes. Therefore, it is important to study analysis methods for failure time data in small samples. Johnson et al. (1982) performed a simulation study to investigate the Cox model with one binary indicator as the covariate under small samples. They found that when total sample size exceeds 40, there is no censoring, and there are equal number of subjects in the two groups, the bias of the estimated log hazard ratio is reasonably low and the sample variance is similar to the asymptotic variance. However, in smaller samples, there are non-trivial differences between the actual and asymptotic formula-based variances.

To improve the estimation and inference of relative risk in studies with small sample sizes, Mehrotra and Roth (2001) proposed a method based on a generalized log-rank (GLR) statistic for the 2-group comparison. They showed that even though asymptotically the GLR method has similar performance to the Cox approach, when the sample size is small, GLR is notably more efficient than the Cox approach, in terms of mean squared error (MSE) for the log relative risk when there are no ties.

In this chapter, we refine the GLR method by replacing previously formulated 'approximate' nuisance parameters with 'exact' counterparts, for settings with and without tied event times. We show through numerical studies that the refined GLR (RGLR) statistic provides a notably smaller bias than the GLR statistic and more commonly used methods such as the Cox and parametric

models, while providing a high relative efficiency and maintaining coverage for 95% confidence intervals. We provide further insights into the GLR statistic by developing an alternate estimation approach for the nuisance parameters. We also compare the performance of the RGLR statistic to parametric models, the Cox model, and the GLR approach. Furthermore, we examine RGLR's performance with respect to type I error and confidence interval coverage, and we compare RGLR with correctly and incorrectly specified parametric models. Section 3.2 includes the derivation of RGLR statistic for testing and estimation, where we also provide a different approach for estimating the nuisance parameters. In Section 3.3, we study the numerical performance of the competing methods through a simulation study. In Section 3.4, we apply the different methods to data from two real data examples. Section 3.5 includes discussion and conclusions.

3.2. Methods

In this section we first develop the RGLR statistic under the assumption of no tied event times. We then extend the method to handle tied event times.

3.2.1. Refined GLR Statistic for Hypothesis Testing with No Tied Event Times

Suppose there are two treatment groups A and B, and we randomize N_A and N_B subjects to each of the groups, respectively. We assume for now that there are no tied observations. Let $t_1 < t_2 < \cdots < t_k$ denote the ordered observed event times for the combined data. Let T denote the random variable for the event time, and $S_B(t)$ and $h_B(t)$ denote the survival and hazard function for T in group B. By definition, we can write $S_B(t) = P(T > t) = \exp(-\int_0^t h_B(x)\mathrm{d}x)$, so that

$$P(t_{i-1} < T \leq t_i | T > t_{i-1}) = 1 - P(T > t_i | T > t_{i-1}) = 1 - \exp(-p_i), \tag{3.1}$$

where $p_i = \int_{t_{i-1}}^{t_i} h_B(x)\mathrm{d}x$. In the development of the original GLR statistic, $1 - \exp(-p_i)$ was simplified to p_i by invoking a first order Taylor series approximation (Mehrotra and Roth, 2001). In this paper, motivated by a desire to reduce bias, we use the exact value of $1 - \exp(-p_i)$ in a refined GLR statistic (RGLR).

Let the random variables D_{iA}, D_{iB} denote the number of events in group A and B at t_i, respectively, and let $D_i = D_{iA} + D_{iB}$. Let the random variables R_{iA}, R_{iB} denote the number of subjects still at risk at time t_i in group A and B, respectively. We then let r_{iA} and r_{iB} denote the observed number

of subjects at risk at time t_i in group A and group B, respectively, and the observed total number of events and observed total number of subjects at risk at time t_i as d_i and r_i, respectively. Under the no ties assumption, $d_i = 1\ \forall i$. At t_i, we can think of D_{iB} as following a binomial distribution with probability $\pi_{iB} = 1 - \exp(-p_i)$ and r_{iB} trials. Then, under the proportional hazards assumption, it follows that the number of events in group A, D_{iA}, will follow a binomial distribution with probability $\pi_{iA} = 1 - \exp(-\theta p_i)$ and r_{iA} number of trials, where θ is the hazard ratio for group A versus B. Let $G_i = \{j : \max(0, d_i - r_{iB}) \leq j \leq \min(d_i, r_{iA})\}$. Given $d_i, r_{iA}, r_{iB}, p_i, \theta$, the conditional distribution of D_{iA} follows a non-central hypergeometric distribution, and we can write the probability function as

$$\lambda_{iA} \equiv P(D_{iA} = d_{iA} | R_{iA} = r_{iA}, R_{iB} = r_{iB}, D_i = d_i, p_i, \theta) \tag{3.2}$$

$$= \frac{\binom{r_{iA}}{d_{iA}}\binom{r_{iB}}{d_{iB}}(1 - e^{-p_i})^{d_{iB}} e^{-p_i(r_{iB} - d_{iB})}(1 - e^{-\theta p_i})^{d_{iA}} e^{-\theta p_i(r_{iA} - d_{iA})}}{\sum_{j \in G_i}\binom{r_{iA}}{j}\binom{r_{iB}}{d_i - j}(1 - e^{-p_i})^{d_i - j} e^{-p_i(r_{iB} - d_i + j)}(1 - e^{-\theta p_i})^{j} e^{-\theta p_i(r_{iA} - j)}}. \tag{3.3}$$

Under the assumption of $d_i = 1\ \forall i$, the conditional mean and variance of D_{iA}, denoted by $E_{iA}(r_{iA}, r_{iB}, \theta, p_i)$ and $V_{iA}(r_{iA}, r_{iB}, \theta, p_i)$, can be derived as the following expressions:

$$E_{iA}(r_{iA}, r_{iB}, \theta, p_i) = \sum_{G_i} d_{iA}\lambda_{iA} = \frac{r_{iA}(e^{\theta p_i} - 1)}{r_{iA}(e^{\theta p_i} - 1) + r_{iB}(e^{p_i} - 1)} \tag{3.4}$$

$$V_{iA}(r_{iA}, r_{iB}, \theta, p_i) = \sum_{G_i} d_{iA}^2\lambda_{iA} - \left(\sum_{G_i} d_{iA}\lambda_{iA}\right)^2 = \frac{r_{iA}(e^{\theta p_i} - 1) r_{iB}(e^{p_i} - 1)}{[r_{iA}(e^{\theta p_i} - 1) + r_{iB}(e^{p_i} - 1)]^2}. \tag{3.5}$$

Note that the vector of nuisance parameters $\boldsymbol{p} = (p_1, p_2, \ldots, p_k)$ is unknown and needs to be estimated. We use an unconditional approach, as suggested by Mehrotra and Roth (2001). The estimate of nuisance parameter p_i is found by maximizing the product of two unconditional binomial likelihoods, Bin($r_{iA}, 1 - \exp(-\theta p_i)$) and Bin($r_{iB}, 1 - \exp(-p_i)$):

$$L(p_i|\theta) = \pi_{iA}^{d_{iA}}(1 - \pi_{iA})^{r_{iA} - d_{iA}} \pi_{iB}^{d_{iB}}(1 - \pi_{iB})^{r_{iB} - d_{iB}} \tag{3.6}$$

$$= (1 - e^{-\theta p_i})^{d_{iA}}(e^{-\theta p_i})^{r_{iA} - d_{iA}}(1 - e^{-p_i})^{d_{iB}}(e^{-p_i})^{r_{iB} - d_{iB}}. \tag{3.7}$$

Because we are assuming no ties, the solution can be simplified to

$$\tilde{p}_{i,\theta} = \begin{cases} \log\left(\frac{\theta r_{iA}+r_{iB}}{\theta r_{iA}+r_{iB}-1}\right), & \text{when } d_{iA}=0, d_{iB}=1 \\ \frac{1}{\theta}\log\left(\frac{\theta r_{iA}+r_{iB}}{\theta r_{iA}+r_{iB}-\theta}\right), & \text{when } d_{iA}=1, d_{iB}=0. \end{cases} \tag{3.8}$$

Let $\tilde{\boldsymbol{p}}(\theta)$ denote the estimated nuisance parameter vector $\boldsymbol{p}$, where $\tilde{\boldsymbol{p}}(\theta) = (\tilde{p}_{1,\theta}, \tilde{p}_{2,\theta}, \ldots, \tilde{p}_{k,\theta})$. Then, the RGLR test statistic for the general null hypothesis $H_0 : \theta = \theta_0$ is

$$RGLR[\theta_0, \tilde{\boldsymbol{p}}(\theta_0)] = \frac{\sum_{i=1}^{k}[d_{iA} - E_{iA}(r_{iA}, r_{iB}, \theta_0, \tilde{p}_{i,\theta_0})]^2}{\sum_{i=1}^{k} V_{iA}(r_{iA}, r_{iB}, \theta_0, \tilde{p}_{i,\theta_0})}. \tag{3.9}$$

The reference distribution for the RGLR statistic is approximated with an F-distribution with degrees of freedom 1 and k^*, where $k^* = \sum_i \min(d_i, r_i - d_i, r_{iA}, r_{iB})$. This is the same distribution as that used for the original GLR statistic (Mehrotra and Roth, 2001). We conjecture that our RGLR statistic has the same reference distribution as the GLR statistic because we only changed the approximate nuisance parameters in the original GLR formulation with 'exact' counterparts, which presumably should not affect the distribution. This is analogous to using different estimators of variance components (nuisance parameters) but the same reference null distributions in common linear mixed effects analyses. This conjecture is strongly supported via simulations in Section 3.3. Note that under the most commonly used null hypothesis $\theta_0 = 1$, estimation of the nuisance parameters is no longer required, and the RGLR statistic reduces to the usual log-rank test statistic (Mantel, 1966), which has an asymptotic distribution of χ^2_1.

Of note, it is easy to see that the equivalence between the RGLR statistic and Cox score statistic as sample size goes to infinity. As sample size increases, the estimate of p_i is approximately zero for most i's, because the time interval becomes smaller between two consecutive events and the probability of having an event in the interval approaches zero. It can be shown through L'Hopital's rule that when $p_i \to 0$, the RGLR statistic reduces to the score statistic from the Cox model. This demonstrates that the RGLR statistic is asymptotically similar to the Cox score statistic; this theoretical expectation is supported using simulations in Section 3.3.

3.2.2. Estimation of Nuisance Parameters

The development above is similar to the logic provided by Mehrotra and Roth (2001). However, to provide additional insight, we show that in the set up of Mehrotra and Roth's GLR statistic, the estimated nuisance parameter $\tilde{p}_{i,\theta}$ can also be estimated using the inverse-variance weighted average of the corresponding estimates of the failure probability in each group. Recall that we think of the number of events at time t_i in group A and B as following two binomial distributions with probability π_{iA} and π_{iB}, respectively. In the setting of GLR, $\pi_{iB} = p_i$ and $\pi_{iA} = \theta p_i$ using the Taylor approximation. Therefore, there are two natural estimates of the failure probability, $\hat{\pi}_{iB} = d_{iB}/r_{iB}$ from group B and $\hat{\pi}_{iA} = d_{iA}/r_{iA}$ from group A. Thus, we have two estimates of the nuisance parameter, namely $\hat{p}_{iB,\theta} = d_{iB}/r_{iB}$ and $\hat{p}_{iA,\theta} = d_{iA}/(\theta r_{iA})$. Hence,

$$\text{Var}(p_{iA}|R_{iA} = r_{iA}) = \frac{\text{Var}(D_{iA})}{\theta^2 r_{iA}^2} = \frac{p_i(1-\theta p_i)}{\theta r_{iA}} \quad \text{and}$$
$$\text{Var}(p_{iB}|R_{iB} = r_{iB}) = \frac{\text{Var}(D_{iB})}{r_{iB}^2} = \frac{p_i(1-p_i)}{r_{iB}}.$$

Accordingly, if we equate p_i with the inverse-variance weighted average of $\hat{p}_{iA,\theta}$ and $\hat{p}_{iB,\theta}$, i.e., set

$$p_i = \frac{\frac{\hat{p}_{iA,\theta}}{Var(p_{iA}|R_{iA}=r_{iA})} + \frac{\hat{p}_{iB,\theta}}{Var(p_{iB}|R_{iB}=r_{iB})}}{\frac{1}{Var(p_{iA}|R_{iA}=r_{iA})} + \frac{1}{Var(p_{iB}|R_{iB}=r_{iB})}},$$

and solve for p_i, we get the same estimated $\tilde{p}_{i,\theta}$ as that obtained via maximization of the product of the aforementioned two Binomial distributions (direct MLE approach). Since the formula for $\tilde{p}_{i,\theta}$ is somewhat complex, using the inverse-variance weighted average approach provides an intuitive and simple path to estimate the nuisance parameters.

In the setting of RGLR, however, these two approaches do not give the same estimates, because the relationship between the nuisance parameter p_i and failure probability π_{iB} is no longer linear. There are no simple closed-form solutions for the nuisance parameters using the inverse-variance weighted average approach. Although numerical solutions can still be achieved, we prefer the direct MLE approach because it delivers an exact closed-form solution. Further details of the derivation using the two approaches for the RGLR statistic can be found in Appendix B.

3.2.3. Inference using the Refined GLR Estimator for Relative Risk

The RGLR statistic is in quadratic form, which guarantees a unique minimum. Because small values of the RGLR statistic support the null hypothesis, we can derive an estimator for relative risk, $\hat{\theta}_{RGLR}$, by finding the θ that minimizes the RGLR test statistic.

The confidence interval of the RGLR estimator can then be calculated using the F$(1, k^*)$ as the reference distribution. Therefore, the 100$(1-\alpha)$% confidence interval for $\hat{\theta}_{RGLR}$ is $(\theta^L_{RGLR}, \theta^U_{RGLR})$, where

$$\theta^L_{RGLR} = \inf_{\theta}\{\theta : RGLR(\theta, \tilde{\boldsymbol{p}}(\theta)) \leq F_\alpha(1, k^*)\} \tag{3.10}$$

$$\theta^U_{RGLR} = \sup_{\theta}\{\theta : RGLR(\theta, \tilde{\boldsymbol{p}}(\theta)) \leq F_\alpha(1, k^*)\}. \tag{3.11}$$

3.2.4. Extension of RGLR to Accommodate Tied Event Times

In this section, we extend the RGLR statistic to allow for tied event times so that the method is more applicable for real data sets. There are several approaches for handling ties in the Cox model, including Breslow (1974), Efron (1977) and Kalbfleisch and Prentice (1973). Mehrotra and Roth (2011) extended the GLR statistic to incorporate ties following analogs of Kalbfleisch and Prentice's and Efron's approaches. We propose to use Efron's approach to handle ties for RGLR statistic, given that Efron's method is easier to implement.

With ties, the previous assumption that $d_i = 1\ \forall i$ no longer holds, and the conditional expected value and variance functions need to be updated to average over all possible orderings of tied event times at each time point i. Suppose in the time interval $(t_{i-1}, t_i]$, there are $d_i (> 1)$ event times given by $t_{i,1} < t_{i,2} < \cdots < t_{i,d_i}$. Now, if we construct an *average* 2×2 life table at the unobserved true event time $t_{i,j}$, the average number of failure event times for group A and B is d_{iA}/d_i and d_{iB}/d_i, respectively, and the average number of subjects still at risk is $r_{iA} - jd_{iA}/d_i$ and $r_{iB} - jd_{iB}/d_i$ for group A and B, respectively, where $j = 1, 2, \ldots, d_i$. Then, summing across the d_i time points, we get

$$\bar{E}_{iA}(\theta, p_{i,j}) = \sum_{j=1}^{d_i} E_{iA}\left(r_{iA} - (j-1)\frac{d_{iA}}{d_i}, r_{iB} - (j-1)\frac{d_{iB}}{d_i}, \theta, p_{i,j}\right) \tag{3.12}$$

$$\bar{V}_{iA}(\theta, p_{i,j}) = \sum_{j=1}^{d_i} V_{iA}\left(r_{iA} - (j-1)\frac{d_{iA}}{d_i}, r_{iB} - (j-1)\frac{d_{iB}}{d_i}, \theta, p_{i,j}\right), \tag{3.13}$$

where E_{iA} and V_{iA} are shown in equation (3.4) and (3.5), using the margins of the *average* 2×2 life table at each of the unobserved true event times for the d_i events.

We derive the estimated nuisance parameter using the likelihood approach as before, where now

$$L(p_{i,j}|\theta) = \pi_{iA}^{d_{iA}/d_i}(1-\pi_{iA})^{r_{iA}-jd_{iA}/d_i}\pi_{iB}^{d_{iB}/d_i}(1-\pi_{iB})^{r_{iB}-jd_{iB}/d_i} \tag{3.14}$$

$$= (1-e^{-\theta p_{i,j}})^{d_{iA}/d_i}(e^{-\theta p_{i,j}})^{r_{iA}-jd_{iA}/d_i}(1-e^{-p_{i,j}})^{d_{iB}/d_i}(e^{-p_{i,j}})^{r_{iB}-jd_{iB}/d_i}. \tag{3.15}$$

To find the nuisance parameter that maximizes equation (3.14), we take the log and the first-order derivative respect to $p_{i,j}$ and set it to zero. The estimating equation is:

$$\frac{d_{iA}}{d_i} \cdot \frac{\theta e^{-\theta p_{i,j}}}{1-e^{-\theta p_{i,j}}} - \theta\left(r_{iA} - j\frac{d_{iA}}{d_i}\right) + \frac{d_{iB}}{d_i(1-e^{-p_{i,j}})} - \left(r_{iB} - j\frac{d_{iB}}{d_i}\right) = 0 \tag{3.16}$$

The estimating equation (3.16) is a nonlinear function of $p_{i,j}$, and there is no closed-form solution. Therefore, we use a numerical approach to solve for $p_{i,j}$ at $t_{i,j}$; let $\tilde{\boldsymbol{p}}(\theta)$ denote the estimated nuisance parameter matrix, where entry (i,j) is denoted as $\tilde{p}_{i,j,\theta}$. Therefore, using Efron's approach to extend RGLR for tied event times, the RGLR^E test statistic for the null hypothesis $H_0 : \theta = \theta_0$ is

$$RGLR^E[\theta_0, \tilde{\boldsymbol{p}}(\theta_0)] = \frac{\sum_{i=1}^{k}[d_{iA} - \bar{E}_{iA}(\theta_0, \tilde{p}_{i,j,\theta_0})]^2}{\sum_{i=1}^{k}\bar{V}_{iA}(\theta_0, \tilde{p}_{i,j,\theta_0})}. \tag{3.17}$$

The reference distribution used for RGLR^E is an F-distribution with degrees of freedom 1 and k^*, where $k^* = \sum_i \min(d_i, r_i - d_i, r_{iA}, r_{iB})$. This is the same distribution as that used for RGLR with no ties. Again, this approximation is based on a conjecture that is supported by simulations, as shown later. Of note, RGLR^E and RGLR are identical when there is no tied event times.

3.3. Simulation Study

We first compared the performance of the RGLR statistic to the Cox proportional hazards and parametric models, and to the GLR approach when there are no tied observations. We carried out a simulation study to examine issues of bias, efficiency, type I error and the nominal 95% confidence interval coverage. For estimation with the parametric model, we examined estimation under the true versus a misspecified distribution for the simulated survival times.

For each of the N_A and N_B subjects in group A and B, independent entry time e_{ij} was generated from a uniform distribution on (0, T), where i indicates subject and $j = 1, 0$ indicates group A or B, respectively. Independent of the entry time, survival time s_{iA} was generated from Weibull (rate= 0.5θ, shape=2), and s_{iB} was generated from Weibull (rate=0.5, shape=2), so that the hazard ratio was θ. Note that the probability density function of a Weibull distribution with shape parameter α and rate parameter λ is $f(t) = \alpha\lambda t^{\alpha-1}\exp(-\lambda t^{\alpha})$. The trial time for each subject was hence $t_{ij} = \min(s_{ij}, T - e_{ij})$.

We varied the sample size, percentage of censoring and the hazard ratio between the two groups to compare the performance of the different methods. Sample size per group was varied as $N_A = N_B = 10, 20, 40, 100$. We considered percentage of censoring for the total sample of 0% and 50%. The percentage of censoring was controlled by changing the final analysis time T. For example, for 20 subjects per group with true log hazard ratio of 0.6, with T=2 the mean censoring was 50.7% and the average number of events was 20.3. The log hazard ratio, denoted by $\ln(\theta)$, took values of 0, 0.6 and 1.2. Simulation results are based on 5000 replications.

Given the small sample sizes, a problem referred to as 'monotone likelihood' was encountered in some simulated datasets, where the highest event time in one group precedes the smallest event time in the other group (Bryson and Johnson, 1981). Under this scenario, the hazard ratio estimate from the Cox model is infinite and not reliable. Therefore, we deleted any simulated dataset in which this occurred, and if for a set of parameters of interest, there were more than 1% simulated datasets with a monotone likelihood, the results were not reported. For this reason results for 10 subjects per group are not considered for scenarios with 50% censoring.

For each simulation scenario, we compare the empirical bias, relative efficiency and the empiri-

cal coverage probability for the 95% confidence interval for all scenarios considered for the parametric (Weibull) regression model, Cox model, GLR and RGLR. The estimated log hazard ratio from fitting the Weibull regression is estimated by dividing the negative of the coefficient for the covariate Z, the group indicator, by the estimated scale parameter. The estimated log hazard ratio from the Cox model is the estimated coefficient for Z. Bias was reported for the case of $\ln(\theta) = 0$ and percentage bias, defined as 100 times the ratio of bias to the true value, was reported for $\ln(\theta) = 0.6$ and 1.2. The relative efficiency was calculated as the ratio of the MSE of the Cox model estimator and the estimator of the given competing method, i.e., $\%$RMSE $=$ $100 \times$ MSE of Cox$/$MSE of competing method. Accordingly, %RMSE$>$100% indicates that the target method is more efficient than the Cox model.

3.3.1. Results on Estimation without Tied Event Times

For the results shown in Table 3.1, the RGLR statistic always has the smallest bias among the four methods and provides higher efficiency relative to the Cox model, even with 100 subjects per group. Compared to the parametric model, RGLR still has a higher relative efficiency in small samples (fewer than 20 subjects per group under 0% censoring and fewer than 40 subjects per group under 50% censoring). While GLR has the highest relative efficiency under small samples, it has a bigger bias than RGLR, and fails to maintain the nominal 95% coverage rate in some scenarios, which will be further discussed in Section 3.3.2. It should also be noted that the results of the parametric method are based on the true distribution. For real data examples, it is quite difficult to make a correct assumption about the true distribution when sample size is small. When a wrong distribution is assumed, we would expect the parametric method to perform worse. Thus, the parametric method carries the risk of making the wrong assumption for the true distribution, whereas the RGLR method does not require any knowledge about the underlying distribution. We will examine the impact of misspecification of the survival distribution later in Section 3.3.3.

3.3.2. Results on Hypothesis Testing without Tied Event Times

Table 3.1 also reports the empirical coverage probability for the 95% confidence interval (C.I.). Note that under the null hypothesis of $H_0 : \ln(\theta) = 0$, i.e., the hazard ratio is 1, and a two-tailed 5% significance level, 100 minus the coverage probability is equal to the type I error rate. Therefore, a coverage probability below 95% under the null indicates an inflated type I error. In Table 3.1, a value

Table 3.1: Empirical bias, percent ratio of MSE relative to Cox model and coverage probability for 95% C.I. for $\ln(\theta) = 0, 0.6, 1.2$ based on 5000 simulations and an underlying Weibull distribution for the survival times.

			$\ln(\theta) = 0$			$\ln(\theta) = 0.6$			$\ln(\theta) = 1.2$		
Censoring	N	Method	Bias	%RMSE	Cov	%Bias	%RMSE	Cov	%Bias	%RMSE	Cov
0%	10	Cox (Wald)	-0.000	100	94.2	8.42	100	94.7	8.30	100	96.5
		Cox (Score)	-0.000	100	[93.3]	8.42	100	[93.5]	8.30	100	94.4
		Weibull	-0.001	102	[92.7]	11.3	104	[93.1]	11.1	114	[93.5]
		GLR	-0.000	135	95.1	-6.50	134	94.8	-7.04	143	[93.3]
		RGLR	-0.000	114	95.1	1.52	114	95.2	1.49	117	95.7
	20	Cox (Wald)	0.001	100	94.6	4.36	100	94.6	3.99	100	95.0
		Cox (Score)	0.001	100	[94.2]	4.36	100	[94.0]	3.99	100	94.4
		Weibull	0.001	101	[93.4]	5.63	104	[93.7]	5.54	111	[93.9]
		GLR	0.001	119	94.9	-3.69	115	[94.0]	-3.46	116	[93.1]
		RGLR	0.001	108	94.9	0.53	108	94.7	0.42	108	94.7
	40	Cox (Wald)	-0.005	100	94.8	1.01	100	94.6	1.38	100	95.0
		Cox (Score)	-0.005	100	94.6	1.01	100	94.8	1.38	100	94.7
		Weibull	-0.005	101	94.5	1.76	104	94.7	2.19	109	94.5
		GLR	-0.005	111	95.1	-3.42	107	[94.3]	-2.45	106	[94.0]
		RGLR	-0.005	105	95.1	-1.12	104	94.8	-0.49	104	94.7
	100	Cox (Wald)	-0.001	100	95.1	0.75	100	94.6	0.75	100	95.0
		Cox (Score)	-0.001	100	95.0	0.75	100	94.6	0.75	100	94.9
		Weibull	-0.001	101	94.9	0.96	102	94.6	1.03	108	94.8
		GLR	-0.001	105	95.2	-1.23	103	94.5	-0.86	102	94.7
		RGLR	-0.001	102	95.2	-0.21	101	94.6	-0.05	102	95.0
50%	20	Cox (Wald)	-0.001	100	95.4	4.91	100	95.3	5.51	100	96.2
		Cox (Score)	-0.001	100	94.4	4.91	100	[94.3]	5.51	100	94.9
		Weibull	-0.000	99	[94.0]	7.24	104	[94.1]	7.40	108	94.5
		GLR	-0.001	123	96.1	-5.67	125	95.5	-5.51	128	94.6
		RGLR	-0.001	110	96.1	0.13	110	95.7	0.66	112	95.9
	40	Cox (Wald)	-0.005	100	95.5	1.45	100	95.5	1.95	100	95.2
		Cox (Score)	-0.005	100	95.0	1.45	100	95.2	1.95	100	94.9
		Weibull	-0.004	100	95.1	2.94	101	95.1	3.23	105	94.7
		GLR	-0.004	112	95.6	-4.21	112	95.6	-3.78	112	94.7
		RGLR	-0.004	105	95.6	-1.18	105	95.8	-0.69	106	95.3
	100	Cox (Wald)	0.001	100	95.3	0.57	100	94.9	0.81	100	95.2
		Cox (Score)	0.001	100	95.2	0.57	100	94.8	0.81	100	95.1
		Weibull	0.001	100	95.0	1.10	101	94.7	1.35	104	94.8
		GLR	0.001	105	95.5	-1.94	105	95.0	-1.66	104	94.6
		RGLR	0.001	102	95.5	-0.62	102	95.1	-0.37	103	95.1

Bias is reported for $\ln(\theta) = 0$, and percentage bias is reported for $\ln(\theta) = 0.6$ and 1.2. %RMSE $= 100 \times$ MSE of Cox/MSE of competing method. Results for 10 per group with 50% censoring are not reported due to monotone likelihood problems in more than 1% of the simulated datasets. Coverage probability more than $Z_{0.975}$ standard errors below 95% is in square brackets. N: sample size per group. Cov: coverage probability for 95% C.I. Cox (Wald): Cox proportional hazards model with Wald test. Cox (Score): Cox proportional hazards model with Score test. Weibull: Weibull regression. GLR: Generalized log-rank approach. RGLR: Refined GLR approach.

in square brackets indicates that the coverage probability is more than $Z_{0.975}$ standard errors less than the nominal rate of 95%, which implies that the type I error rate is more than $Z_{0.975}$ standard errors above the nominal rate of 5%. We performed a Wald test for the estimated θ using parametric (Weibull) regression and both Wald and Score tests using Cox model. When sample size was 10 and 20 per group, the Wald test from Weibull regression and Cox Score and Cox Wald tests tended to provide an inflated type I error rate, while our RGLR statistic controlled the type I error rate under 5%. For $\ln(\theta) = 0.6$ and 1.2, RGLR consistently maintained at least 95% coverage rate across all simulated scenarios. On the other hand, GLR, Cox and parametric model failed to maintain the 95% coverage rate when sample size was small.

3.3.3. Misspecification of the Failure Time Distribution (No Tied Event Times)

As mentioned earlier, it is not always possible to assume the correct distribution when using a given parametric approach in a real data situation. When a wrong parametric model is fit to the data, we would expect the resulting estimator to be biased. On the other hand, the RGLR approach does not make any assumption about the underlying survival distribution. We carried out a simulation study on the effect of misspecification, where the data were generated from a Gompertz distribution. The survival time in group A was generated from a Gompertz(shape=0.5, rate=0.2θ), and the survival time in group B was generated from a Gompertz(shape=0.5, rate=0.2), so that proportional hazards still holds with hazard ratio θ. Each subject also had an independent entry time, and the trial was administratively censored by a fixed time T.

We again considered three different values for the log hazard ratio: $\ln(\theta) = 0, 0.6, 1.2$, percentage censoring of 0% and 50%, and varied the number of subjects per group as 10, 20, 40, 100. For each simulation, we fit the exponential, Weibull and Cox models, and applied the GLR and RGLR methods. Figure B.1 in the Appendix B shows the different hazard functions from Gompertz, Weibull and exponential distributions.

When Gompertz was the true distribution, fitting exponential and Weibull regression under 0% censoring resulted in large bias and low percent RMSE when $\ln(\theta) > 0$, as shown in the Table 3.2. The percentage bias from fitting exponential regression was as large as 40%, and its percent RMSE ranged from 14% to 210%. However, with a percentage bias around 30-40%, the high percent RMSE is largely meaningless. On the other hand, when the log hazard ratio was 0, exponential

regression had a very small absolute bias and a high percent RMSE. However, given its poor performance in the case of non-zero log hazard ratio, this behavior indicates a tendency towards attenuation bias. Li, Klein, and Moeschberger (1996) examined the behavior of exponential regression under misspecification in the context of hypothesis testing, and found that exponential regression notably underestimates the nominal 5% alpha level when the true distribution is Gompertz and substantially overestimates when the hazard rate is decreasing. This is consistent with our finding that the exponential model performed poorly for non-zero log hazard ratio scenarios. Weibull regression, although more stable than exponential regression, still resulted in a bias of 10% or more when the sample size was at least 20 subjects per group under 0% censoring. It also started to lose efficiency as sample size increased, for example, with %RMSE=63% when $\ln(\theta) = 1.2$ under 0% censoring.

Compared to the parametric approach, the Cox model, GLR and RGLR approaches are not subject to misspecification of the underlying distribution and thus provided much more stable results. The bias of RGLR approach was the smallest across all the simulated scenarios, and it also delivered a higher relative efficiency than the Cox model and Gompertz model when there were fewer than 100 subjects per group. The efficiency of the Gompertz model relative to the Cox model increased above 100% when the sample size reached 100 per group, which is expected.

When percentage censoring increased to 50%, all methods performed better than with 0% censoring. This could be due to the fact that some extreme values were censored under 50% censoring. However, exponential and Weibull regression were still the least ideal approaches. RGLR, on the other hand, consistently showed the lowest bias and high relative efficiency, relative to Cox and Gompertz model.

As shown in Table 3.2 and mentioned earlier, the exponential model underestimated Type I error and had poor coverage. The Cox model, especially using the score test, and GLR tended to provide a slightly lower coverage than desired. On the other hand, the RGLR approach is more stable and was able to maintain at least 95% coverage.

Figure 3.1 panels (a)-(c) show the empirical densities of the estimators from the different methods with an underlying Gompertz distribution and 0% censoring and 20 subjects per group for $\ln(\theta) = 0, 0.6,$ and 1.2, respectively. The vertical line is drawn at the true hazard ratio. As noted

in the simulation results, in all three cases, the exponential model was adversely impacted by mis-specification of the underlying true distribution. The RGLR estimates centered more closely around the true value than those from the Cox model.

Table 3.2: Empirical bias, percent ratio of MSE relative to Cox model and coverage probability for 95% C.I. for $\ln(\theta) = 0, 0.6, 1.2$ based on 5000 simulations and an underlying Gompertz distribution for the survival times.

			$\ln(\theta)=0$			$\ln(\theta)=0.6$			$\ln(\theta)=1.2$		
Censoring	N	Method	Bias	%RMSE	Cov	%Bias	%RMSE	Cov	%Bias	%RMSE	Cov
0%	10	Cox (Wald)	-0.000	100	[94.3]	8.98	100	95.1	7.96	100	96.8
		Cox (Score)	-0.000	100	[93.4]	8.98	100	[94.2]	7.96	100	94.5
		Gompertz	0.005	94	[92.3]	15.5	94	[93.0]	14.1	99	[93.7]
		Exp	-0.002	350	99.7	-39.8	210	98.5	-35.8	139	[93.4]
		Weibull	0.001	140	95.5	-5.10	138	95.2	-3.95	144	94.7
		GLR	0.000	135	95.2	-6.14	135	95.0	-7.36	143	[93.5]
		RGLR	0.000	114	95.2	2.04	115	95.7	1.17	117	96.0
	20	Cox (Wald)	0.003	100	94.5	4.06	100	95.0	3.68	100	95.1
		Cox (Score)	0.003	100	[94.0]	4.06	100	94.7	3.68	100	94.7
		Gompertz	0.005	97	[93.7]	7.36	98	[94.1]	6.90	102	94.6
		Exp	0.003	324	99.8	-39.6	132	96.9	-35.7	71	[81.2]
		Weibull	0.003	142	96.8	-10.3	133	95.7	-9.02	132	[94.0]
		GLR	0.003	119	94.9	-4.02	116	94.5	-3.55	116	[93.4]
		RGLR	0.003	108	94.9	0.21	107	95.1	0.37	108	94.9
	100	Cox (Wald)	0.002	100	94.9	0.94	100	94.9	0.81	100	95.1
		Cox (Score)	0.002	100	94.8	0.94	100	94.8	0.81	100	95.0
		Gompertz	0.002	100	94.8	1.51	100	94.7	1.41	102	94.7
		Exp	0.001	294	99.9	-40.3	34	[65.9]	-36.3	14	[5.4]
		Weibull	0.002	140	97.5	-14.2	95	[93.6]	-12.8	63	[83.6]
		GLR	0.002	105	94.9	-1.04	103	94.5	-0.80	103	94.7
		RGLR	0.002	102	94.9	-0.02	102	94.8	0.01	102	94.9
50%	20	Cox (Wald)	0.005	100	95.6	4.91	100	95.5	4.81	100	95.8
		Cox (Score)	0.005	100	94.8	4.91	100	94.7	4.81	100	94.9
		Gompertz	0.007	97	94.5	7.99	98	94.8	7.55	96	94.9
		Exp	0.006	142	97.9	-7.24	134	96.9	-7.71	138	95.8
		Weibull	0.006	112	95.7	2.93	112	95.7	1.95	119	95.5
		GLR	0.004	121	96.2	-4.35	122	95.6	-4.80	124	95.1
		RGLR	0.005	109	96.2	0.69	109	95.6	0.53	111	95.8
	40	Cox (Wald)	-0.010	100	95.0	0.45	100	95.2	1.14	100	95.5
		Cox (Score)	-0.010	100	94.6	0.45	100	94.8	1.14	100	95.0
		Gompertz	-0.010	99	94.5	2.22	99	94.8	2.64	97	94.6
		Exp	-0.008	138	97.3	-10.9	124	96.3	-10.4	116	[94.2]
		Weibull	-0.009	111	95.7	-2.28	109	95.5	-2.26	114	95.7
		GLR	-0.009	111	95.2	-4.41	111	95.4	-3.91	112	94.6
		RGLR	-0.010	105	95.2	-1.83	105	95.5	-1.18	105	95.3
	100	Cox (Wald)	0.000	100	95.1	0.62	100	95.0	0.69	100	94.8
		Cox (Score)	0.000	100	94.9	0.62	100	94.7	0.69	100	94.7
		Gompertz	0.000	100	94.7	1.26	99	94.7	1.40	99	94.6
		Exp	0.001	136	97.5	-11.1	112	95.4	-10.7	90	[91.4]
		Weibull	0.000	111	95.5	-3.11	108	95.1	-3.25	109	94.8
		GLR	0.000	105	95.1	-1.48	104	94.9	-1.47	104	94.4
		RGLR	0.000	102	95.1	-0.39	102	95.0	-0.34	102	94.9

Bias is reported for $\ln(\theta) = 0$, and percentage bias is reported for $\ln(\theta) = 0.6$ and 1.2. $\%\text{RMSE} = 100 \times$ MSE of Cox/MSE of competing method. Results for 10 per group with 50% censoring are not reported due to monotone likelihood problems in more than 1% of the simulated datasets. Coverage probability more than $Z_{0.975}$ standard errors below 95% is in square brackets. N: sample size per group. Cov: coverage probability for 95% C.I. Cox (Wald): Cox proportional hazards model with Wald test. Cox (Score): Cox proportional hazards model with Score test. Weibull: Weibull regression. GLR: Generalized log-rank approach. RGLR: Refined GLR approach.

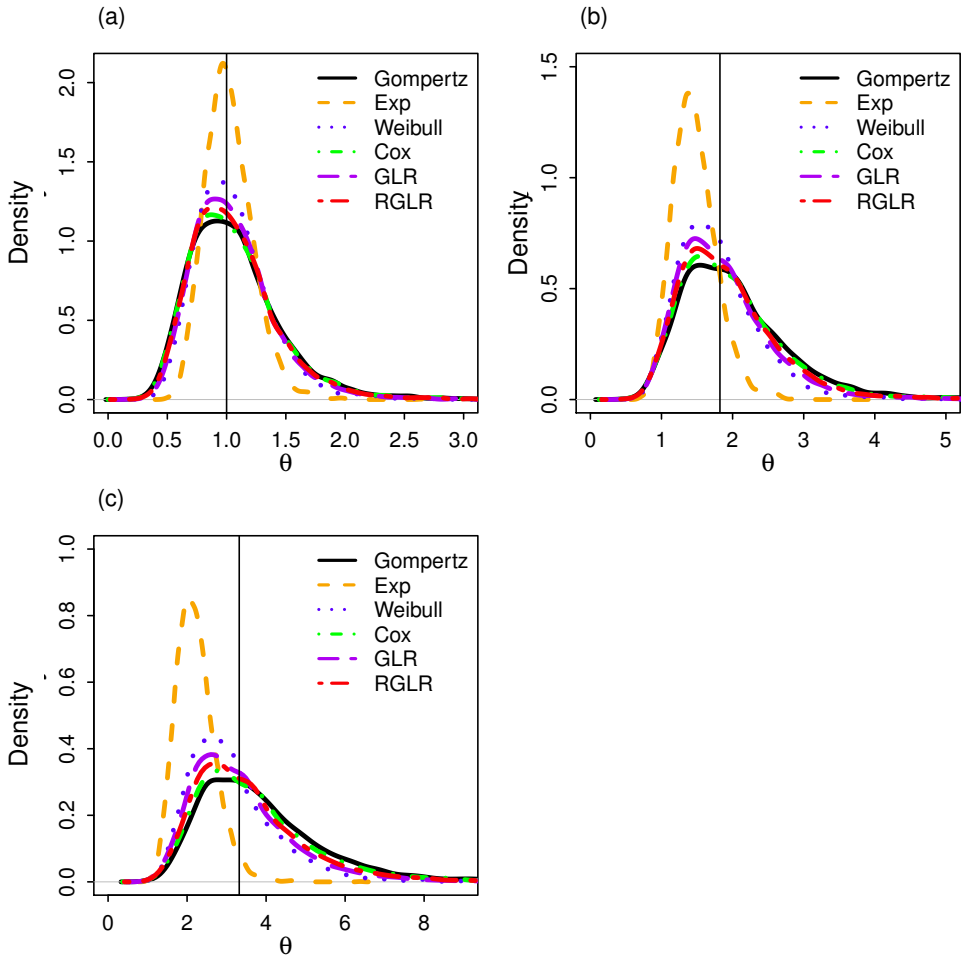


Figure 3.1: Empirical densities of estimators from the Gompertz, exponential, and Weibull parametric survival models, Cox model, generalized log-rank (GLR) and refined GLR (RGLR) (5000 simulations for 20 subjects per group with 0% censoring and an underlying Gompertz distribution) with a true hazard ratio of (a) 1 (b) 1.82 (c) 3.32. A vertical line is drawn at the true hazard ratio.

3.3.4. Simulation Results with Tied Event Times

To compare the performance of RGLR^E to competing methods when ties in the event times are present, we again generated the data from a Weibull distribution. The set up was the same as the scenario with no tied observations, where survival time in group A was from Weibull (rate=0.5θ, shape=2), and survival time in group B was from Weibull (rate=0.5, shape=2). Ties were created by rounding the event times to one digit after the decimal place, which is equivalent to rounding

to the nearest month if the trial time unit is in years. There were approximately 15-20% tied event times, calculated as the percentage of non-unique event times in group A and B, in the the simulation studies. We compared the proposed RGLR extension for ties, RGLR^E, Weibull regression, Cox model and GLR extension for ties, the latter two using Efron's approximation. The pattern of simulation results are very similar to that under no ties, and results are reported in Table 3.3.

When ties are present, with small sample sizes, RGLR^E still delivered the smallest bias among all the methods considered, and provided higher efficiency than both Cox model that adjusts for ties using Efron's approximation and Weibull regression. It also controlled type I error and maintained at least 95% coverage rate, while both Cox model and Weibull tended to have inflated type I error under small samples; of note, GLR^E failed to deliver adequate 95% confidence interval coverage in some cases.

3.4. Application to Two Real Datasets

We apply the RGLR and other competing methods to data from two clinical trials involving lung cancer (Kalbfleisch and Prentice, 1980) and bladder cancer (Pagano and Gauvreau, 2000).

3.4.1. Lung Cancer Clinical Trial

Kalbfleisch and Prentice (1980) reported results for a lung cancer trial with 137 male patients. There were 69 patients randomized patients to a standard chemotherapy and 68 patients to a test chemotherapy. Patients were categorized into four histological tumor types: squamous, small cell, adenoma and large cell. The outcome variable was time to death (in days). Kaplan-Meier curves comparing patients on standard and test chemotherapy with different cell types are presented in Figure 3.2.

There were no tied event times in the large cell group, so we applied Weibull regression, Cox model, GLR and RGLR. The remaining groups all had some tied event times; therefore, we applied Weibull regression, Cox model with Efron's approximation for ties, GLR^E, RGLR^E. For patients with large cell group, GLR and RGLR provided a smaller estimated hazard ratio (test/standard) and narrower 95% C.I. than Weibull and Cox model, as shown in Figure 3.3 (b). The estimated hazard ratio (95% C.I.) was 1.64 (0.76, 3.55) using Weibull regression, 1.54 (0.69, 3.41) using Cox regression, 1.44 (0.71, 2.96) using GLR and 1.49 (0.69, 3.22) using RGLR. For patients with squamous, adenoma

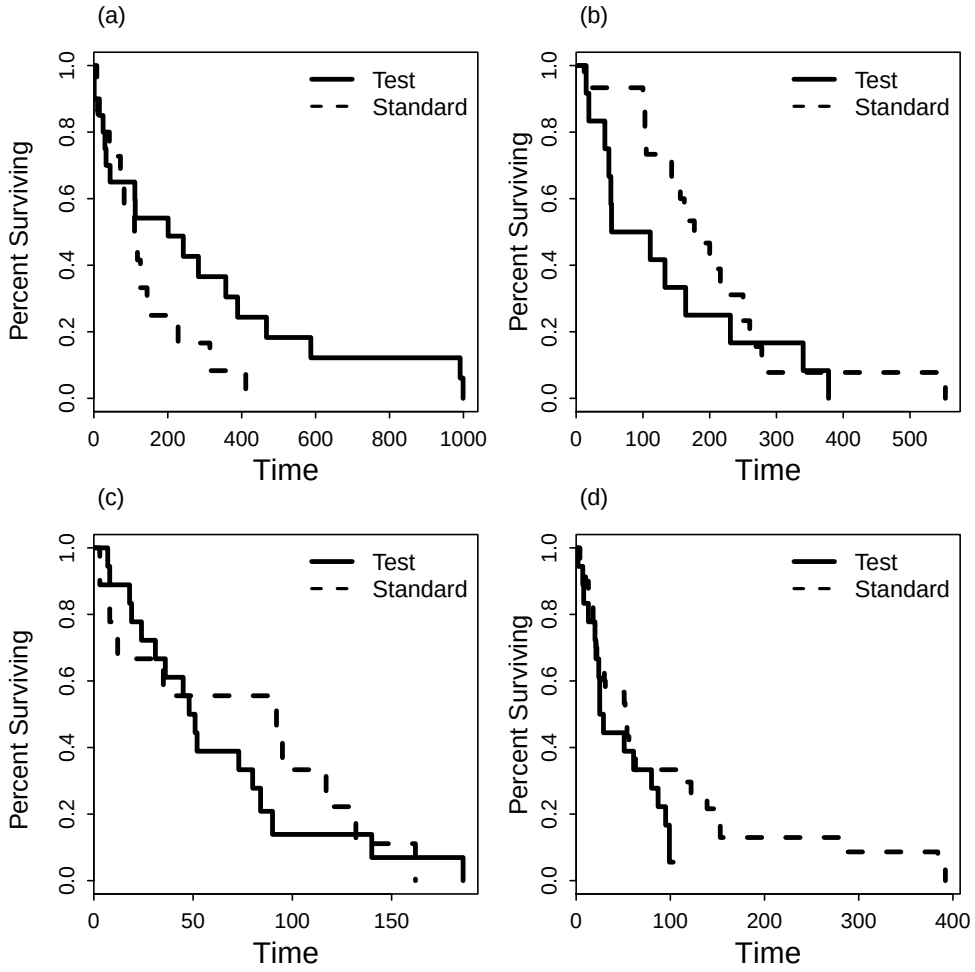
Table 3.3: Empirical bias, percent ratio of MSE relative to Cox model and coverage probability for 95% C.I. for $\ln(\theta) = 0, 0.6, 1.2$ based on 5000 simulations and an underlying Weibull distribution for the survival times with tied observations.

			$\ln(\theta) = 0$			$\ln(\theta) = 0.6$			$\ln(\theta) = 1.2$		
Censoring	N	Method	Bias	%RMSE	Cov	%Bias	%RMSE	Cov	%Bias	%RMSE	Cov
0%	10	Cox^E (Wald)	-0.001	100	94.7	7.30	100	94.6	6.36	100	96.6
		Cox^E (Score)	-0.001	100	[93.7]	7.30	100	[93.8]	6.36	100	94.7
		Weibull	-0.001	99	[92.7]	11.5	101	[93.0]	11.3	107	[93.3]
		GLR^E	-0.000	138	95.3	-8.46	136	94.6	-9.50	140	[92.6]
		RGLR^E	-0.000	116	95.3	-0.09	116	95.0	-0.70	117	95.7
	20	Cox^E (Wald)	0.001	100	94.6	3.52	100	94.6	2.91	100	94.8
		Cox^E (Score)	0.001	100	[94.1]	3.52	100	[94.3]	2.91	100	94.4
		Weibull	0.002	99	[93.4]	5.69	102	[93.7]	5.62	106	[93.8]
		GLR^E	0.001	120	94.9	-4.95	115	[94.0]	-4.74	114	[92.7]
		RGLR^E	0.001	109	95.0	-0.55	108	94.7	-0.75	108	94.7
	100	Cox^E (Wald)	-0.001	100	95.0	0.52	100	94.7	0.65	100	95.0
		Cox^E (Score)	-0.001	100	95.0	0.52	100	94.6	0.65	100	95.0
		Weibull	-0.001	101	94.9	0.96	102	94.6	1.05	108	94.8
		GLR^E	-0.001	106	95.2	-1.21	102	94.5	-1.78	102	94.7
		RGLR^E	-0.001	103	95.1	-0.23	101	94.7	-0.35	102	95.0
50%	20	Cox^E (Wald)	-0.001	100	95.4	3.91	100	95.4	4.19	100	96.3
		Cox^E (Score)	-0.001	100	94.5	3.91	100	94.6	4.19	100	95.1
		Weibull	-0.000	96	[94.0]	7.45	100	[93.8]	7.67	101	94.5
		GLR^E	-0.001	123	96.0	-6.61	124	95.5	-6.72	126	94.6
		RGLR^E	-0.001	110	96.0	-0.89	110	95.8	-0.67	113	96.0
	40	Cox^E (Wald)	-0.004	100	95.6	0.90	100	95.7	1.06	100	95.2
		Cox^E (Score)	-0.004	100	95.2	0.90	100	95.2	1.06	100	94.8
		Weibull	-0.004	99	94.9	3.22	99	94.7	3.41	101	94.5
		GLR^E	-0.004	112	96.0	-4.70	111	95.6	-4.55	111	94.5
		RGLR^E	-0.004	106	96.0	-1.73	105	95.8	-1.52	106	95.2
	100	Cox^E (Wald)	0.001	100	95.3	0.79	100	95.5	0.63	100	95.2
		Cox^E (Score)	0.001	100	95.1	0.79	100	95.2	0.63	100	95.1
		Weibull	0.001	100	95.0	1.93	101	95.2	1.93	101	95.2
		GLR^E	0.001	105	95.4	-1.64	105	95.5	-1.76	104	95.4
		RGLR^E	0.001	102	95.4	-0.38	102	95.6	-0.52	102	95.5

Bias is reported for $\ln(\theta) = 0$, and percentage bias is reported for $\ln(\theta) = 0.6$ and 1.2. %RMSE $= 100 \times$ MSE of Cox/MSE of competing method. Results for 10 per group with 50% censoring are not reported due to monotone likelihood problems in more than 1% of the simulated datasets. Coverage probability more than $Z_{0.975}$ standard errors below 95% is in square brackets. N: sample size per group. Cov: coverage probability for 95% C.I. Cox^E (Wald): Cox proportional hazards model using Efron's method for ties with Wald test. Cox^E (Score): Cox proportional hazards model using Efron's method for ties with Score test. Weibull: Weibull regression. GLR^E: Generalized log-rank approach using Efron's method for ties. RGLR^E: Refined GLR approach using Efron's method for ties.

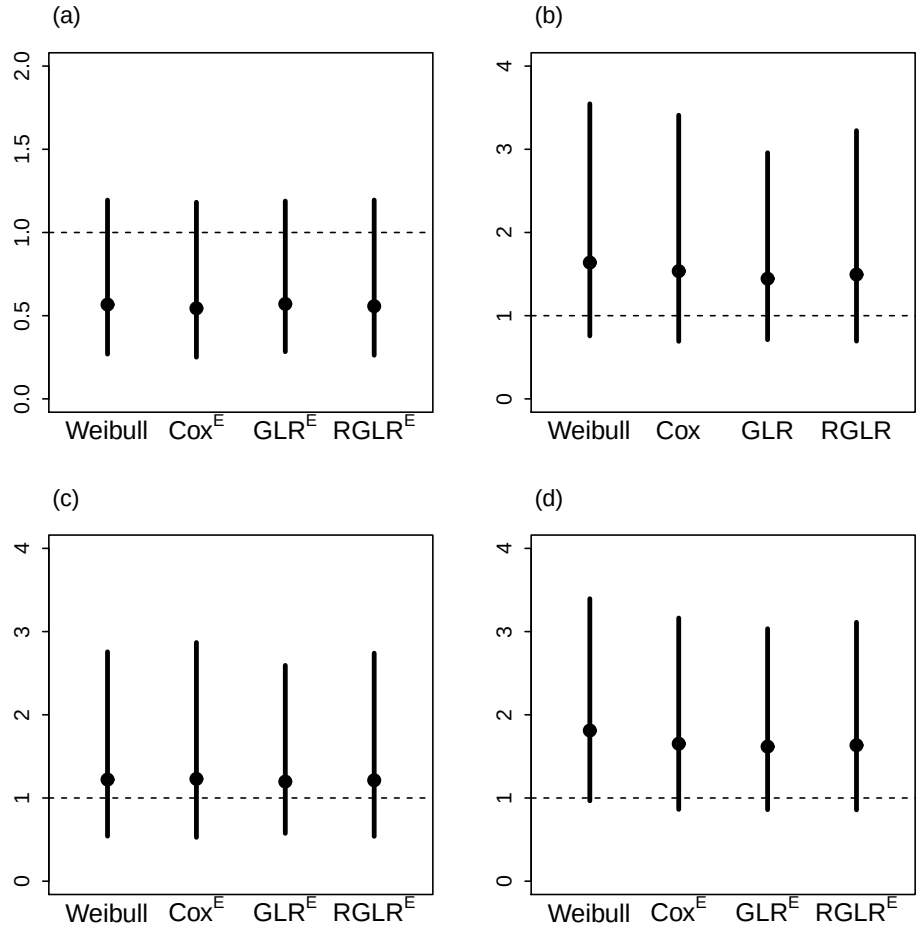
and small cell types, four methods, Weibull, Cox model with Efron's approximation, GLR^E and RGLR^E provided similar results, as shown in Figure 3.3 panels (a), (c) and (d). The true hazard ratio is unknown in a real data example, but based on our simulation results, the RGLR approach has the smallest bias and maintains coverage for 95% C.I. in small samples, and thus, is expected to be closer to the truth.

Figure 3.2: Lung cancer data example: Kaplan-Meier curves for time to death comparing test to standard chemotherapy by cell types.



(a) squamous cell (b) large cell (c) adenoma cell and (d) small cell. Data from Kalbfleisch and Prentice (1980).

Figure 3.3: Lung cancer data example: Estimated hazard ratio and 95% confidence interval comparing test to standard chemotherapy by cell types.



(a) squamous cell (b) large cell (c) adenoma cell and (d) small cell. Cox: Cox regression. Weibull: Weibull regression. GLR: Generalized log-rank approach. RGLR: Refined GLR approach. Cox^E: Cox regression using Efron's method to adjust for tied events. Weibull: Weibull regression. GLR^E: Generalized log-rank approach using Efron's method to adjust for tied events. RGLR^E: Refined GLR approach using Efron's method to adjust for tied events. Data from Kalbfleisch and Prentice (1980).

3.4.2. Bladder Cancer Clinical Trial

Pagano and Gauvreau (2000) reported results on a bladder cancer clinical trial. The study included 86 patients in total, who were assigned to either placebo or chemotherapy (Thiotepa) after surgery. The outcome of interest was time to recurrence (in months). We further divided the subjects into two groups according to the number of tumors removed at surgery, one or multiple, and assessed

the treatment effect. Among patients with one tumor removed, 26 patients were on placebo and 23 were on chemotherapy. Among those with multiple tumors removed, 22 patients were on placebo and 15 were on chemotherapy. Figure 3.4 panels (a) and (b) present the Kaplan-Meier curves comparing patients on placebo and chemotherapy with one or multiple tumors removed.

Because of the tied event times in the data set, we again applied Weibull regression, Cox model with Efron's approximation for ties, GLR^E, RGLR^E. The four methods provided similar results among patients with one tumor removed, but quite different results for those with multiple tumors removed. For patients with only one tumor removed, the estimated hazard ratio (95% C.I.) of recurrence (placebo/chemotherapy) was 1.28 (0.62, 2.66) using Weibull regression, 1.28 (0.61, 2.69) using the Cox model with Efron's approximation, 1.27 (0.61, 2.63) using GLR^E and 1.27 (0.61, 2.68) using RGLR^E. For those with multiple tumors removed, the corresponding results were 2.37 (0.84, 6.70) using Weibull regression, 1.96 (0.70, 5.51) using Cox model with Efron's approximation, 3.50 (1.27, 9.85) using GLR^E and 3.60 (1.27, 10.25) using RGLR^E. As shown in Figure 3.4 (d), both GLR^E and RGLR^E provided statistical evidence of a treatment difference based on the C.I. excluding one, while Weibull regression and Cox model did not.

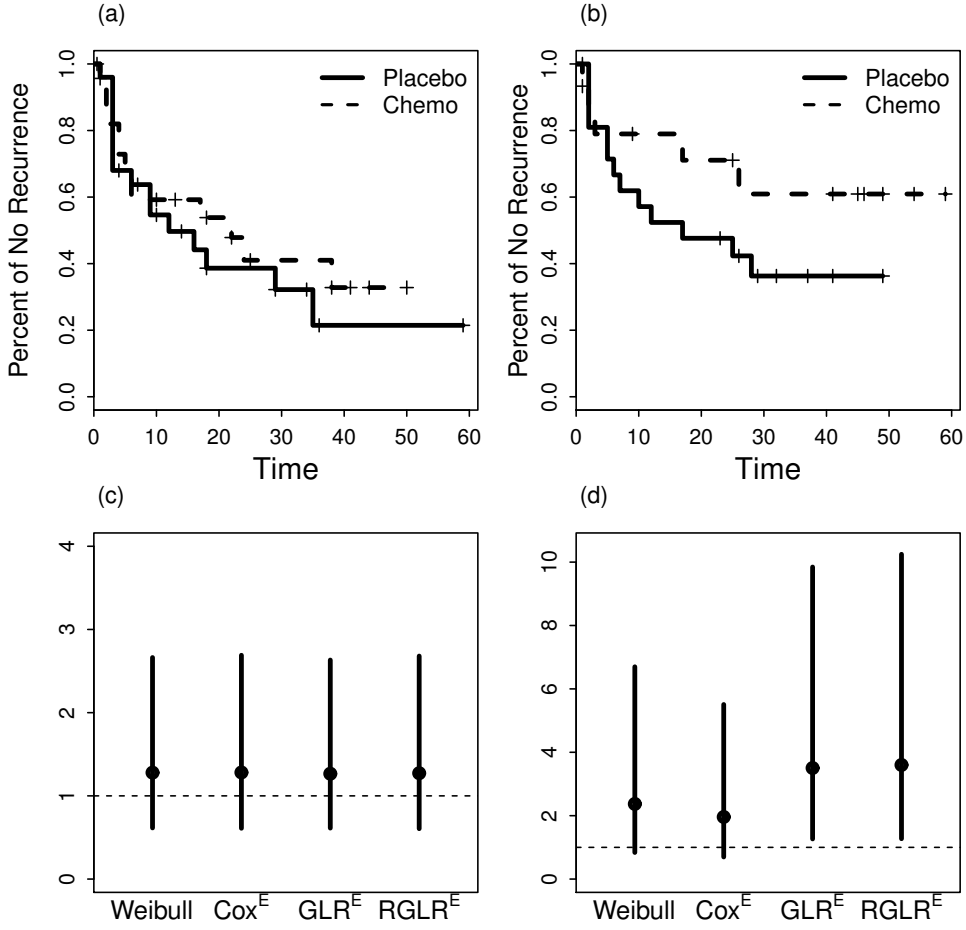
While Weibull and Cox regressions generated a narrower confidence interval, both of the methods tend to have inflated type I error and lower coverage probability for 95% C.I. in small samples, as shown in our simulation studies (Section 3.3). Therefore, our numerical results suggest RGLR^E is expected to be closer to the truth in this example.

Of note, in both real data examples, the estimated HR for GLR and GLR^E was always closer to one than that for RGLR and RGLR^E. This is consistent with the simulation results in Section 3.3 which showed that GLR and GLR^E tend to underestimate true hazard ratios that are greater than one (and, by analogy, overestimate true hazard ratios that are less than one).

3.5. Discussion

Small sample studies of time-to-event outcomes are quite common in early phase clinical trials and observational studies of rare diseases. Thus, it is important to have methods that provide efficient hazard ratio estimation, control type I error and maintain confidence interval coverage in small sample settings. In this chapter, we developed the RGLR statistic, and extended the method

Figure 3.4: Bladder cancer data example: Kaplan-Meier survival curves and estimated hazard ratio and 95% confidence interval comparing placebo and chemotherapy by number of tumors removed at surgery.



(a) and (c): one tumor removed at surgery. (b) and (d): multiple tumors removed at surgery. Cox^E: Cox regression using Efron's method to adjust for tied events. Weibull: Weibull regression. GLR^E: Generalized log-rank approach using Efron's method to adjust for tied events. RGLR^E: Refined GLR approach using Efron's method to adjust for tied events. Data from Pagano and Gauvreau (2000).

to allow for ties. RGLR reduces bias while maintaining high relative efficiency versus the Cox model by eliminating an unnecessary approximation in the GLR statistic. We also provided a more intuitive development using inverse-variance weighting to estimate the nuisance parameters for GLR. In addition, we have also demonstrated control of type I error rate and 95% C.I. coverage in small samples for RGLR and explored the effect of misspecification of the underlying distribution on parametric models. Through simulation studies, we have shown that the RGLR approach provides

smaller bias relative to the Cox and true parametric models, and GLR, when the sample size per group is around 40 or less and comparable performance for larger samples. RGLR was able to consistently keep the type I error at or below the 5% nominal level in extensive simulations, while the parametric and Cox models were observed to have an inflated type I error rate in small samples. Furthermore, in real data applications, it is often challenging to know the true underlying distribution. We have illustrated through simulations that when an incorrect distribution is used by a parametric regression, it can result in large bias for the estimated hazard ratio. On the other hand, the RGLR approach does not require any assumption about the true distribution, and consistently delivers a very low bias with better efficiency relative to the Cox model. We recommend the use of RGLR in the setting of two-group comparisons with survival outcomes in small samples over the commonly used parametric and Cox models.

CHAPTER 4

HAZARD RATIO ESTIMATION IN STRATIFIED PARALLEL DESIGNS UNDER PROPORTIONAL HAZARDS

4.1. Introduction

In randomized clinical trials with a time-to-event endpoint, it is particularly important to incorporate stratification when the risk of having the event of interest is affected by a certain prognostic factor, such as race, gender, baseline disease severity, and so on. Several studies have shown that omitting important covariates can lead to potentially spurious results (Bretagnolle and Huber-Carol, 1988; Ford and Norrie, 2002; Pocock et al., 2002; Schumacher, Olschewski, and Schmoor, 1987; Struthers and Kalbfleisch, 1986). For example, Schumacher, Olschewski, and Schmoor (1987) showed that the estimated hazard ratio is attenuated if a prognostic factor is omitted, and this result is also confirmed by Bretagnolle and Huber-Carol (1988). A commonly used approach for analyzing stratified trials with time-to-event outcomes is the stratified Cox proportional hazard model (Cox, 1972), which makes the assumption of proportional hazards within each stratum. It also imposes an additional assumption that the hazard ratio is exactly the same across all strata, which seems implausible in many practical settings. When there is a treatment by stratum interaction, i.e., the hazard ratio differs by stratum, using the conventional stratified Cox model analysis can lead to a biased and/or less efficient result.

To ensure unbiased and efficient results even when there exists a treatment by stratum interaction, Mehrotra, Su, and Li (2012) proposed a two-step approach to allow for different hazard ratios across strata. Their procedure entails fitting a Cox model separately for each stratum and then combining the stratum-specific log hazard ratio estimates to obtain an estimate of the overall log hazard ratio; the latter is defined later in this chapter and is presumed to be the parameter of interest. They considered two weighting schemes: sample size (SS) weights and minimum risk (MR) weights (Mehrotra and Railkar, 2000); both of these are described in the next section. The Mehrotra, Su, and Li (2012) method was developed for large sample applications; however, many randomized clinical trials involve relatively small samples (50-200 patients per treatment group) (Pocock, 1983). It is also common at the discovery stage of a drug, for there to be known prognostic factors, for

example, in settings such as cardiovascular disease or cancer, that may require a stratified design. In Chapter 3, we developed a method for improving hazard ratio estimation using a refined generalized log-rank (RGLR) statistic in small randomized clinical trials without stratification, and showed that it provides higher efficiency and smaller bias than the Cox proportional hazards model analysis in small samples. In this chapter, we extend the RGLR statistic to handle stratification and explore its performance in small samples. An additional contribution is the theoretical development of a (remarkably accurate) approximation for the variance of the RGLR-based estimate of a log hazard ratio. Section 4.2 includes details of the two-step RGLR approach for both the SS and MR weighting schemes. In Section 4.3, we explore the relative performance of the competing methods, namely the conventional stratified Cox model and two-step Cox model and corresponding two-step RGLR analyses, through simulations. We then apply the methods to a real data example from a colon cancer clinical trial in Section 4.4. Section 4.5 includes summary remarks and discussion.

4.2. Methods

Suppose there are $i = 1, 2, \ldots, S$ strata, and within stratum i, we randomize n_{iA} and n_{iB} subjects to treatment A and B, respectively; by design, the ratio n_{iA}/n_{iB} is constant across strata. Denote the sample size in stratum i as $n_i = n_{iA} + n_{iB}$ and total sample size as $n = \sum_{i=1}^{S} n_i$. Within stratum i, let $t_{i1} < t_{i2} < \cdots < t_{ik_i}$ denote the ordered observed event times for the combined group across treatments. Let β_i denote the log hazard ratio in stratum i with $\theta_i = \exp(\beta_i)$ representing the hazard ratio. If there is no treatment by stratum interaction, i.e., if $\beta_i = \beta$ for all i, there is no ambiguity about the definition of the overall log hazard ratio. However, in the presence of an interaction, i.e, if $\beta_i \neq \beta$ for at least one i and i^*, it is natural to define the target parameter as a population weighted average of the β_i's, i.e., $\bar{\beta} = \sum_{i=1}^{S} f_i \beta_i$, where f_i is the fraction of subjects in the target population that are from stratum i ($\sum_{i=1}^{S} f_i = 1$). The overall hazard ratio is defined as $\theta_{overall} = \exp(\bar{\beta})$.

The conventional stratified Cox model analysis assumes no treatment by stratum interaction, and this can (and often does) result in a biased estimate of $\bar{\beta}$. To allow for a potential treatment by stratum interaction, we propose to use RGLR to estimate the log hazard ratio in each stratum, and combine the stratum-specific point estimates using a weighted average to estimate the overall log hazard ratio:

$$\hat{\bar{\beta}} = \sum_{i=1}^{S} w_i \hat{\beta}_i. \tag{4.1}$$

Following Mehrotra, Su, and Li (2012), we consider two weighting schemes: sample size (SS) and minimum risk (MR). Sample size weighting uses the sample size in each stratum relative to the whole sample as the weight, i.e., $\hat{w}_i^{SS} = f_i$. While sample size weighting provides an unbiased estimator of $\bar{\beta}$, assuming simple random sampling, it can suffer from a needlessly large variance. The MR weights proposed by Mehrotra and Railkar (2000) are intended to minimize mean-squared error; for our stratified time-to-event setting, the weights are calculated as:

$$\hat{w}_i^{MR} = \frac{a_i}{\sum_{i=1}^{S} \hat{V}_i^{-1}} - \frac{b_i \hat{V}_i^{-1}}{\sum_{i=1}^{S} \hat{V}_i^{-1} + \sum_{i=1}^{S} b_i \hat{\beta}_i \hat{V}_i^{-1}} \cdot \frac{\sum_{i=1}^{S} \hat{\beta}_i a_i}{\sum_{i=1}^{S} \hat{V}_i^{-1}}, \tag{4.2}$$

where $b_i = \hat{\beta}_i \sum_{i=1}^{S} \hat{V}_i^{-1} - \sum_{i=1}^{S} \hat{\beta}_i \hat{V}_i^{-1}$, $a_i = \hat{V}_i^{-1}(1 + b_i \sum_{i=1}^{S} \hat{\beta}_i n_i / n)$, and $\hat{V}_i$ is the estimated variance for $\hat{\beta}_i$.

To implement equation (4.2) we need to derive the variance of the stratum-specific RGLR estimate of the log hazard ratio. In previous work (Chapter 3), we have shown that the (single-stratum) RGLR statistic requires estimation of a nuisance parameter at time t_j, $p_j = \int_{t_{j-1}}^{t_j} h_B(x)\mathrm{d}x$, where $h_B(t)$ is the hazard function for group B. Let random variables D_{jA} and D_{jB} denote the number of events at time t_j in group A and B, respectively, and let $D_j = D_{jA} + D_{jB}$. Let random variables R_{jA} and R_{jB} denote the number of subjects at risk at time t_j in group A and B, respectively, and r_{jA}, r_{jB} denote the observed number of subjects at risk at time t_j in group A and B. Under the assumption of no tied event times, given $d_j, r_{jA}, r_{jB}, p_j, \beta$, D_{jA} follows a non-central hypergeometric distribution, and the conditional mean and variance of D_{jA} is

$$E_{jA} = \frac{r_{jA}(1 - e^{-p_j e^{\beta}})e^{-p_j}}{r_{jA}(1 - e^{-p_j e^{\beta}})e^{-p_j} + r_{jB}(1 - e^{-p_j})e^{-p_j e^{\beta}}} \tag{4.3}$$

$$V_{jA} = \frac{r_{jA}(1 - e^{-p_j e^{\beta}})e^{-p_j} r_{jB}(1 - e^{-p_j})e^{-p_j e^{\beta}}}{[r_{jA}(1 - e^{-p_j e^{\beta}})e^{-p_j} + r_{jB}(1 - e^{-p_j})e^{-p_j e^{\beta}}]^2}. \tag{4.4}$$

We showed in Chapter 3 that the nuisance parameter can be estimated using an unconditional approach. Let $\tilde{\mathbf{p}}$ denote the estimated nuisance parameter vector, and define $S_k(\beta, \tilde{\mathbf{p}}) = \sum_{j=1}^{k}(d_{jA} - E_{jA})$ and $I_k(\beta, \tilde{\mathbf{p}}) = \sum_{j=1}^{k} V_{jA}$. Note that $E[S_k(\beta, \tilde{\mathbf{p}})] = 0$, and therefore, we can estimate the log hazard ratio β by solving the moment equation $S_k(\beta, \tilde{\mathbf{p}}) = 0$. Denote the moment log hazard ratio

estimator by $\hat{\beta}^{RGLR}$. Then by the first order Taylor expansion, we have

$$\hat{\beta}^{RGLR} - \beta \cong -\frac{S_k(\beta, \tilde{\mathbf{p}})}{\frac{\partial S_k(\beta, \tilde{\mathbf{p}})}{\partial \beta}}, \tag{4.5}$$

where

$$-\frac{\partial S_k(\beta, \tilde{\mathbf{p}})}{\partial \beta} = \sum_{j=1}^{k} \frac{p_j e^{\beta} e^{-p_j e^{-\beta}}}{1 - e^{-p_j e^{\beta}}} V_{jA}. \tag{4.6}$$

As sample size gets large, $p_j \to 0$, and by L'Hopital's rule, we have that $\lim_{p_j \to 0} \frac{p_j e^{\beta} e^{-p_j e^{-\beta}}}{1 - e^{-p_j e^{\beta}}} = 1$. Thus,

$$-\frac{\partial S_k(\beta, \tilde{\mathbf{p}})}{\partial \beta} \cong \sum_{j=1}^{k} V_{jA} = I_k(\beta, \tilde{\mathbf{p}}). \tag{4.7}$$

Combining equations (4.5) and (4.7) gives us

$$\sqrt{I_k(\beta, \tilde{\mathbf{p}})}(\hat{\beta}^{RGLR} - \beta) \cong \frac{S_k(\beta, \tilde{\mathbf{p}})}{\sqrt{I_k(\beta, \tilde{\mathbf{p}})}}. \tag{4.8}$$

Since $S_k(\beta, \tilde{\mathbf{p}})$ converges to the Cox score as $n \to \infty$, and $\tilde{\mathbf{p}}$ goes to 0, an argument similar to that in Andersen and Gill (1982) can be applied to show asymptotic normality of the RGLR estimator for β. Specifically, by the Martingale Central Limit Theorem,

$$\frac{S_k(\beta, \tilde{\mathbf{p}})}{\sqrt{I_k(\beta, \tilde{\mathbf{p}})}} \xrightarrow{D} N(0, 1). \tag{4.9}$$

Therefore, denoting $\hat{V}(\hat{\beta}^{RGLR}, \tilde{\mathbf{p}}) = I_k^{-1}(\hat{\beta}^{RGLR}, \tilde{\mathbf{p}})$ and combining equations (4.8) and (4.9), we have

$$\frac{\hat{\beta}^{RGLR} - \beta}{\sqrt{\hat{V}(\hat{\beta}^{RGLR}, \tilde{\mathbf{p}})}} \xrightarrow{D} N(0, 1). \tag{4.10}$$

And we conjecture the variance for the RGLR estimator to be

$$\hat{V}(\hat{\beta}^{RGLR}, \tilde{\mathbf{p}}) \approx \left(\sum_{j=1}^{k} \frac{r_{jA}(1 - e^{-\tilde{p}_j e^{\hat{\beta}^{RGLR}}}) e^{-\tilde{p}_j} r_{jB}(1 - e^{-\tilde{p}_j}) e^{-\tilde{p}_j e^{\hat{\beta}^{RGLR}}}}{[r_{jA}(1 - e^{-\tilde{p}_j e^{\hat{\beta}^{RGLR}}}) e^{-\tilde{p}_j} + r_{jB}(1 - e^{-\tilde{p}_j}) e^{-\tilde{p}_j e^{\hat{\beta}^{RGLR}}}]^2} \right)^{-1}. \tag{4.11}$$

With the variance formula now established for the RGLR estimator in each stratum, we can now

calculate the MR weights using equation (4.2). For both weighting schemes, we do hypothesis testing ($H_0 : \bar{\beta} = 0$ vs. $H_1 : \bar{\beta} \neq 0$), where $\hat{\bar{\beta}}^{RGLR}$ is the weighted sum of the S independent stratum-specific estimates $\hat{\beta}_i^{RGLR}$, as shown in equation (4.1). The variance of $\hat{\bar{\beta}}^{RGLR}$ is calculated as

$$\hat{V}(\hat{\bar{\beta}}^{RGLR}) = \sum_{i=1}^{S} \hat{w}_i^2 \hat{V}(\hat{\beta}_i^{RGLR}, \tilde{\mathbf{p}}). \tag{4.12}$$

Then, confidence interval calculations can be done using Wald tests implied by equation (4.10). A numerical study of the empirical accuracy of the variance formula (4.11) is provided in Appendix C.

4.3. Simulations

4.3.1. Simulation Set-up

We performed a simulation study to examine the bias, relative efficiency and nominal 95% confidence interval (C.I.) coverage probability of the two-step RGLR using SS weights and MR weights, and compared the performance of our proposed methods to the conventional stratified Cox proportional hazards analysis and the two-step method of Mehrotra, Su, and Li (2012) in which stratum-specific Cox model estimates are combined using SS weights or MR weights.

We considered the case of 2 strata and 4 strata in the simulation study. Usually, in the presence of stratification, only the total number of subjects per group and randomization ratio ($= 1$ here) is fixed by design. Therefore, we used a similar simulation set-up as Mehrotra and Railkar (2000) and treated the number of subjects in each stratum as a random variable. Specifically, n pairs of subjects were first assigned to stratum i with probability f_i ($\sum f_i = 1$), where $i = 1, 2$ for 2 strata and $i = 1, 2, 3, 4$ for 4 strata, and then, within each pair, one subject was randomly assigned to treatment A and the other to treatment B with equal probability. Thereafter, for subject j in stratum i and randomized to treatment q (q =A or B), we generated an entry time e_{ijq} from a uniform distribution (0, T). For 2 strata, survival times s_{ijq} for subject j under treatment A and treatment B in stratum i were generated from Weibull (scale= $\lambda_i/\sqrt{\theta_i}$, shape=2) and Weibull (scale=λ_i, shape=2) respectively, where $\lambda_1 = 0.6, \lambda_2 = 1.2$. Note that the hazard function for Weibull (scale=λ, shape=γ) is $\gamma x^{\gamma-1}/\lambda^{\gamma}$, so the hazard ratio of treatment A relative to B in stratum i is θ_i. The follow-up time for a subject j randomized to treatment q in stratum i was $t_{ijq} = \min(s_{ijq}, T - e_{ijq})$.

For 4 strata, we used the same procedure for generating number of subjects per stratum, entry time

and survival time as described above for the two strata simulations. Survival time s_{ijq} for subject j in stratum i under treatment A and B was generated from Weibull (scale=$\lambda_i/\sqrt{\theta_i}$, shape=2) and Weibull (scale=λ_i, shape=2), respectively, where now with $\lambda_1 = 0.6, \lambda_2 = 0.8, \lambda_3 = 1$, and $\lambda_4 = 1.2$.

We varied the stratum-specific relative frequency and true log hazard ratio, along with total sample size and overall percentage censoring. Both equal (Scenario 1) and unequal stratum sizes (Scenario 2) were considered. For 2 strata, we set $f_1 = f_2 = 0.5$ and $f_1 = 0.7, f_2 = 0.3$ for Scenario 1 and 2, respectively. For 4 strata, we set $f_1 = f_2 = f_3 = f_4 = 0.25$ and $f_1 = 0.15, f_2 = 0.35, f_3 = 0.35, f_4 = 0.15$ for Scenario 1 and 2, respectively. Under the null hypothesis, stratum-specific and overall log hazard ratio was 0 in all cases. Under the alternative hypothesis, we considered two settings: the same log hazard ratio across strata (Alt 1) and different log hazard ratios across strata (Alt 2). The stratum-specific log hazard ratios in each scenario are summarized in Table 4.1; of note, the overall log hazard ratio ($\bar{\beta}$) was fixed at -0.7 in every case, which corresponds to an overall hazard ratio of $\exp(-0.7) = 0.5$. Subjects per treatment group was varied as 50, 100 for 2 strata, and 100, 200 for 4 strata. Two percentage censoring values were considered: 25% and 50%. 5000 replications were generated. Hypothesis testing was done at the $\alpha = 0.05$ level. Results for bias (under the null hypothesis), percent bias (under the alternative hypothesis), type I error rate, power, relative efficiency and coverage probability for the 95% confidence interval (C.I.) for 2 and 4 strata were obtained. Here, relative efficiency refers to 100 times the ratio of the mean squared error (MSE) for the estimator of $\bar{\beta}$ using the stratified Cox model relative to that using the given alternative method of estimation. Thus, relative efficiency estimators greater than 100% represent an improvement over the stratified Cox model.

4.3.2. Simulation Results

Table 4.2 shows the results for the 2 strata case under the null hypothesis and the two alternative hypotheses for both equal (Scenario 1) and unequal (Scenario 2) relative frequency in each stratum. In Scenario 1, under the null hypothesis, all methods were associated with very small bias. Our proposed two-step RGLR provided similar efficiency relative to the stratified Cox model, and higher efficiency than the two-step Cox model method under both weighting schemes. Our proposed method also controlled the type I error rate under 5% across all simulated scenarios, while both the stratified Cox model and the two-step Cox model method had inflated type I error for 50 subjects per treatment group and 25% censoring. Under the alternative hypothesis with no stratum by treatment

Table 4.1: True log hazard ratio in each stratum and overall under the null and alternative hypotheses.

2 strata

Scenario 1: Equal stratum sizes

Stratum	Relative frequency	Null (no interaction)	Alt 1 (no interaction)	Alt 2 (interaction)
1	0.5	0	-0.7	-0.2
2	0.5	0	-0.7	-1.2
Overall		0	-0.7	-0.7

Scenario 2: Unequal stratum sizes

Stratum	Relative frequency	Null (no interaction)	Alt 1 (no interaction)	Alt 2 (interaction)
1	0.7	0	-0.7	-0.4
2	0.3	0	-0.7	-1.4
Overall		0	-0.7	-0.7

4 strata

Scenario 1: Equal stratum sizes

Stratum	Relative frequency	Null (no interaction)	Alternative 1 (no interaction)	Alternative 2 (interaction)
1	0.25	0	-0.7	-0.3
2	0.25	0	-0.7	-0.4
3	0.25	0	-0.7	-0.8
4	0.25	0	-0.7	-1.3
Overall		0	-0.7	-0.7

Scenario 2: Unequal stratum sizes

Stratum	Relative frequency	Null (no interaction)	Alternative 1 (no interaction)	Alternative 2 (interaction)
1	0.15	0	-0.7	-0.3
2	0.35	0	-0.7	-0.4
3	0.35	0	-0.7	-0.8
4	0.15	0	-0.7	-1.65
Overall		0	-0.7	-0.7

Note: under all the alternative hypotheses for both 2 strata and 4 strata, the overall log hazard ratio $\bar{\beta}$ is fixed at -0.7.

interaction (Alt 1), the stratified Cox is expected to have the best performance, and the two-step RGLR provided very similar efficiency relative to the stratified Cox model. The two-step RGLR also delivered a percentage bias less than 2% and maintained adequate coverage probability for the 95% C.I., while the stratified Cox model failed to do so under equal stratum sample size with 50 subjects per treatment and 25% censoring. When there was interaction between treatment and stratum (Alt 2), the proposed two-step RGLR provided notably better efficiency and smaller bias than all the other competing methods. Both the stratified and two-step Cox model methods had issues with maintaining adequate 95% C.I. coverage probability in several simulated scenarios, but the two-step RGLR with SS weights maintained adequate coverage probability throughout all simulated settings. The two-step RGLR with MR weights also performed well but it failed to maintain adequate coverage probability in the scenario with 100 subjects per treatment and 50% censoring. With 100 subjects per treatment and 50% censoring, the two-step RGLR with SS weights delivered 42% higher efficiency than the stratified Cox model, with a percentage bias of 0.8%, comparing to

-28.3% bias from the stratified Cox model. The performance of the methods for unequal relative frequency in each stratum was similar to that for equal relative frequency described above.

Table 4.3 shows the results for the 4 strata case. Under both equal and unequal relative stratum frequency, our two-step RGLR provided the smallest bias and higher relative efficiency compared to the stratified Cox model. When there was a treatment by stratum interaction, the stratified Cox model had a bias as large as -16.9%, while the two-step RGLR controlled the bias under 8%. In terms of type I error, the stratified and two-step Cox model methods had inflated type I error issues with smaller sample sizes (100 subjects per treatment group with 25% and 50% censoring under Scenario 1), while our two-step RGLR did not. In terms of coverage probability, the two-step RGLR maintained adequate coverage probability for 95% C.I. throughout all scenarios, while the stratified Cox model failed to do so under several scenarios.

We also examined power among the methods. Table 4.4 shows the results for 100 subjects per treatment with 50% censoring for 2 strata and 4 strata cases; results under other simulated scenarios (not shown) did not provide additional insights and are hence not shown. When there was no interaction between treatment and stratum, our two-step RGLR provided similar power as the stratified Cox model. When there is interaction, using two-step RGLR delivered a power increase of at least 5 percentage points relative to the stratified Cox model. While the two-step Cox model method seemed to have slightly better power than the two-step RGLR, the former also had inflated type I error rate while our two-step RGLR did not.

4.4. Application

We apply the stratified Cox model, the Mehrotra, Su, and Li (2012)'s two-step Cox model method and our proposed two-step RGLR method, with both two-step methods using sample size (SS) and minimum risk (MR) weights, to a clinical trial involving resected colon cancer (Lin et al., 2016). The data set included 154 patients with stage C colon cancer who were randomized to receive placebo or levamisole combined with fluorouracil therapy, with 77 patients in each group. The outcome of interest was overall survival. Patients were stratified by the number of lymph nodes involved (≤ 4 vs >4). Table 4.5 summarizes the results from applying all the methods. The stratified Cox model provided an estimated overall hazard ratio (therapy:placebo) of $\exp(-0.64) = 0.53$ (95% C.I.: 0.31, 0.90), with a p-value of 0.021. On the other hand, the two-step Cox and two-step RGLR,

Table 4.2: Bias (% bias), percent ratio of MSE relative to one-step stratified Cox model and coverage probability for 95% C.I. for overall log hazard ratio $\bar{\beta}$ for 2 strata based on 5000 simulations.

Scenario 1: Equal stratum sizes ($f_1 = f_2 = 0.5$)

			Null			Alt 1 (no interaction)			Alt 2 (interaction)		
Censoring	N/trt	Method	Bias	%RE	Cov	%Bias	%RE	Cov	%Bias	%RE	Cov
25%	50	Stratified Cox	-0.001	100	[94.2]	1.8	100	94.6	-12.7	100	[92.8]
		2-step Cox (SS wts)	-0.001	95	[93.9]	3.7	94	[94.0]	4.2	93	[94.1]
		2-step RGLR (SS wts)	-0.001	102	94.7	0.1	101	95.0	0.5	100	94.8
		2-step Cox (MR wts)	-0.001	97	[93.9]	2.8	97	[94.3]	0.7	97	[94.3]
		2-step RGLR (MR wts)	-0.001	105	94.9	-0.8	104	95.1	-3.0	103	94.7
	100	Stratified Cox	-0.001	100	95.0	0.9	100	94.8	-13.5	100	[90.1]
		2-step Cox (SS wts)	-0.002	97	94.8	1.8	97	94.5	2.5	110	[94.0]
		2-step RGLR (SS wts)	-0.002	102	95.3	-0.2	100	95.0	0.5	115	94.4
		2-step Cox (MR wts)	-0.001	99	94.8	1.54	98	94.5	0.5	113	[94.1]
		2-step RGLR (MR wts)	-0.001	103	95.3	-0.6	102	95.0	-1.5	117	94.4
50%	50	Stratified Cox	0.000	100	94.4	2.0	100	94.7	-27.1	100	[89.8]
		2-step Cox (SS wts)	0.001	86	94.7	4.3	85	95.1	4.8	97	95.7
		2-step RGLR (SS wts)	0.001	93	95.6	0.4	93	96.0	1.0	106	96.3
		2-step Cox (MR wts)	0.000	94	94.4	2.8	93	94.8	-4.3	109	94.5
		2-step RGLR (MR wts)	0.000	102	95.5	-1.0	102	95.3	-8.3	116	94.8
	100	Stratified Cox	-0.001	100	95.0	1.1	100	94.9	-28.3	100	[82.7]
		2-step Cox (SS wts)	-0.003	89	94.8	2.4	89	94.6	2.9	135	94.9
		2-step RGLR (SS wts)	-0.003	93	95.3	0.3	93	95.1	0.8	142	95.2
		2-step Cox (MR wts)	-0.002	95	94.7	1.7	95	94.6	-3.0	141	[93.5]
		2-step RGLR (MR wts)	-0.002	99	95.1	-0.4	99	94.9	-5.2	145	[93.6]

Scenario 2: Unequal stratum sizes ($f_1 = 0.7, f_2 = 0.3$)

			Null			Alt 1 (no interaction)			Alt 2 (interaction)		
Censoring	N/trt	Method	Bias	%RE	Cov	%Bias	%RE	Cov	%Bias	%RE	Cov
25%	50	Stratified Cox	0.002	100	[94.2]	1.6	100	94.6	-12.5	100	[93.4]
		2-step Cox (SS wts)	0.003	93	[94.2]	3.5	92	94.3	4.6	91	[94.2]
		2-step RGLR (SS wts)	0.002	101	94.9	0.0	99	95.0	0.0	100	94.9
		2-step Cox (MR wts)	0.002	96	[94.3]	2.6	96	94.5	0.2	98	[94.2]
		2-step RGLR (MR wts)	0.002	104	94.7	-0.9	103	95.0	-4.4	105	94.4
	100	Stratified Cox	0.003	100	95.3	0.6	100	95.5	-12.1	100	[91.3]
		2-step Cox (SS wts)	0.002	96	95.0	1.5	96	95.3	2.5	106	94.4
		2-step RGLR (SS wts)	0.002	101	95.6	-0.5	100	95.7	0.0	112	94.8
		2-step Cox (MR wts)	0.002	98	95.0	1.1	98	95.4	0.1	111	94.4
		2-step RGLR (MR wts)	0.002	103	95.5	-0.9	102	95.5	-2.4	115	94.7
50%	50	Stratified Cox	-0.001	100	95.0	1.5	100	95.0	-21.7	100	[90.0]
		2-step Cox (SS wts)	-0.003	87	95.2	3.2	89	95.0	-0.3	104	95.9
		2-step RGLR (SS wts)	-0.003	95	96.1	-0.5	97	95.8	-4.7	113	96.2
		2-step Cox (MR wts)	-0.002	95	95.0	2.0	96	94.7	-8.4	112	[94.3]
		2-step RGLR (MR wts)	-0.002	103	95.7	-1.6	104	95.5	-12.8	116	94.4
	100	Stratified Cox	0.004	100	95.2	0.4	100	95.1	-21.0	100	[87.9]
		2-step Cox (SS wts)	0.004	88	95.0	1.3	89	94.9	4.1	106	95.5
		2-step RGLR (SS wts)	0.004	92	95.7	-0.7	93	95.3	1.4	113	95.9
		2-step Cox (MR wts)	0.004	94	94.9	0.8	95	94.8	-2.3	116	[94.3]
		2-step RGLR (MR wts)	0.003	99	95.3	-1.2	99	95.3	-5.1	121	94.4

Trt=treatment group; bias is reported under the null hypothesis, and percentage bias is reported under the alternative hypothesis. %RE is 100 times MSE of stratified Cox/MSE of competing method. Coverage probability more than $Z_{0.975}$ standard errors below 95% is in square brackets. Each 2-step method uses a weighted average of stratum-specific log hazard ratio estimates; SS wts=sample size weights, MR wts=minimum risk weights.

for both SS and MR weights, provided a non-significant p-value (>0.05). The estimated hazard ratio in stratum 1 from using Cox and RGLR were $\exp(-0.27) = 0.76$ and $\exp(-0.26) = 0.77$, respectively, with corresponding estimates of the hazard ratio in stratum 2 being $\exp(-1.16) = 0.31$ and $\exp(-1.14) = 0.32$, respectively. The Kaplan-Meier curves by stratum in Figure 4.1 appear to support the differential treatment effect across the two strata, i.e, they suggest evidence of a

Table 4.3: Bias (% bias), percent ratio of MSE relative to one-step stratified Cox model and coverage probability for 95% C.I. for overall log hazard ratio $\bar{\beta}$ for 4 strata based on 5000 simulations.

Scenario 1: Equal stratum sizes ($f_1 = f_2 = f_3 = f_4 = 0.25$)

			Null			Alt 1 (no interaction)			Alt 2 (interaction)		
Censoring	N/trt	Method	Bias	%RE	Cov	%Bias	%RE	Cov	%Bias	%RE	Cov
25%	100	Stratified Cox	0.000	100	94.8	0.7	100	95.2	-8.5	100	[93.1]
		2-step Cox (SS wts)	0.001	93	[94.0]	3.3	91	94.6	3.6	92	[94.0]
		2-step RGLR (SS wts)	0.001	101	94.8	-0.3	99	95.3	-0.1	100	94.8
		2-step Cox (MR wts)	0.001	95	[94.1]	2.6	94	94.9	1.8	95	[94.1]
		2-step RGLR (MR wts)	0.001	103	94.9	-1.0	101	95.4	-1.9	102	94.6
	200	Stratified Cox	0.001	100	94.8	0.4	100	95.2	-9.2	100	[91.6]
		2-step Cox (SS wts)	0.001	97	94.7	1.6	96	94.9	1.6	114	[94.2]
		2-step RGLR (SS wts)	0.001	102	95.2	-0.4	100	95.4	-0.4	119	94.7
		2-step Cox (MR wts)	0.001	98	94.7	1.3	97	95.0	0.6	116	[94.1]
		2-step RGLR (MR wts)	0.001	103	95.1	-0.7	101	95.3	-1.5	119	94.4
50%	100	Stratified Cox	-0.001	100	94.9	1.1	100	94.7	-16.0	100	[90.7]
		2-step Cox (SS wts)	-0.001	87	94.5	4.1	85	94.5	4.3	94	94.9
		2-step RGLR (SS wts)	-0.001	94	95.4	0.2	93	95.3	0.4	104	95.7
		2-step Cox (MR wts)	-0.001	92	[94.3]	2.9	90	[94.2]	0.2	102	[93.9]
		2-step RGLR (MR wts)	-0.001	99	95.3	-1.0	99	94.9	-3.8	110	94.6
	200	Stratified Cox	0.000	100	94.9	0.4	100	95.2	-16.9	100	[86.4]
		2-step Cox (SS wts)	-0.001	94	94.9	1.9	91	94.8	2.2	129	95.0
		2-step RGLR (SS wts)	-0.001	98	95.4	-0.3	96	95.4	0.1	136	95.3
		2-step Cox (MR wts)	-0.001	96	94.7	1.3	95	94.6	-0.5	134	[94.3]
		2-step RGLR (MR wts)	-0.001	100	95.1	-0.8	99	95.1	-2.6	138	94.5

Scenario 2: Unequal stratum sizes ($f_1 = 0.15, f_2 = 0.35, f_3 = 0.35, f_4 = 0.15$)

			Null			Alt 1 (no interaction)			Alt 2 (interaction)		
Censoring	N/trt	Method	Bias	%RE	Cov	%Bias	%RE	Cov	%Bias	%RE	Cov
25%	100	Stratified Cox	0.000	100	95.1	0.8	100	95.6	-9.8	100	[92.4]
		2-step Cox (SS wts)	0.000	93	94.2	3.5	91	94.9	3.4	95	[94.3]
		2-step RGLR (SS wts)	0.000	101	95.0	-0.1	99	95.6	-0.6	104	94.9
		2-step Cox (MR wts)	0.000	95	94.4	2.6	94	95.2	1.2	100	[94.3]
		2-step RGLR (MR wts)	0.000	103	95.1	-0.9	101	95.5	-2.8	107	94.6
	200	Stratified Cox	-0.000	100	95.5	0.5	100	95.2	-9.8	100	[91.1]
		2-step Cox (SS wts)	-0.001	97	95.2	1.7	96	94.9	2.0	113	94.9
		2-step RGLR (SS wts)	-0.000	101	95.6	-0.2	100	95.2	-0.2	119	95.3
		2-step Cox (MR wts)	-0.000	98	95.1	1.4	97	94.8	0.9	116	94.8
		2-step RGLR (MR wts)	-0.000	102	95.8	-0.6	101	95.4	-1.4	120	95.1
50%	100	Stratified Cox	0.002	100	95.0	0.5	100	95.3	-16.8	100	[90.9]
		2-step Cox (SS wts)	0.002	91	94.8	3.4	89	95.2	-0.3	112	95.4
		2-step RGLR (SS wts)	0.002	98	95.7	-0.4	97	95.9	-4.2	119	95.9
		2-step Cox (MR wts)	0.002	95	94.6	2.2	94	94.9	-4.1	116	94.7
		2-step RGLR (MR wts)	0.002	102	95.4	-1.6	102	95.6	-8.0	119	94.7
	200	Stratified Cox	0.001	100	95.0	0.2	100	95.4	-16.5	100	[86.8]
		2-step Cox (SS wts)	0.002	93	94.9	1.7	93	95.2	1.9	128	95.4
		2-step RGLR (SS wts)	0.002	97	95.3	-0.4	97	95.4	-0.4	136	95.5
		2-step Cox (MR wts)	0.002	96	94.6	1.2	96	95.0	-0.9	134	94.4
		2-step RGLR (MR wts)	0.002	100	95.1	-0.9	100	95.2	-3.2	138	94.7

Trt=treatment group; bias is reported under the null hypothesis, and percentage bias is reported under the alternative hypothesis. % RE is 100 times MSE of stratified Cox/MSE of competing method. Coverage probability more than $Z_{0.975}$ standard errors below 95% is in square brackets. Each 2-step method uses a weighted average of stratum-specific log hazard ratio estimates; SS wts=sample size weights, MR wts=minimum risk weights.

treatment by stratum interaction. The overall hazard ratio from the two-step RGLR with SS and MR weights was estimated to be $\exp(-0.50) = 0.61$ (95% C.I.: 0.34, 1.06) and $\exp(-0.53) = 0.59$ (95% C.I.: 0.83, 1.02), respectively.

Table 4.4: Power comparisons among the competing methods based on 100 subjects per treatment group and 50% censoring with 5000 simulations for 2 strata (top panel) and 4 strata (bottom panel).

	2 strata			
	Scenario 1: $f_1 = f_2 = 0.5$		Scenario 2: $f_1 = 0.7, f_2 = 0.3$	
Method	Alt 1 (no interaction)	Alt 2 (interaction)	Alt 1 (no interaction)	Alt 2 (interaction)
Stratified Cox	92.5	[66.8]	96.2	[80.5]
2-step Cox (SS wts)	90.4	86.2	94.9	90.7
2-step RGLR (SS wts)	89.8	85.0	94.4	89.5
2-step Cox (MR wts)	91.8	[84.2]	95.6	[89.9]
2-step RGLR (MR wts)	91.3	[83.1]	95.1	88.9

	4 strata			
	Scenario 1: $f_1 = f_2 = f_3 = f_4 = 0.25$		Scenario 2: $f_1 = 0.15, f_2 = 0.35, f_3 = 0.35, f_4 = 0.15$	
Method	Alt 1 (no interaction)	Alt 2 (interaction)	Alt 1 (no interaction)	Alt 2 (interaction)
Stratified Cox	91.2	78.8	90.8	[80.6]
2-step Cox (SS wts)	89.8	86.9	89.9	87.3
2-step RGLR (SS wts)	88.5	85.4	88.4	85.6
2-step Cox (MR wts)	[90.9]	[87.2]	90.8	87.4
2-step RGLR (MR wts)	89.7	85.5	89.3	85.6

Square brackets indicate the case where the coverage probability is more than $Z_{0.975}$ standard errors below 95%.

Table 4.5: Log hazard ratio estimates for the Colon cancer data example in Lin et al. (2016).

	$N(\%)$	Stratified Cox	2-step Cox (SS wts)	2-step RGLR (SS wts)	2-step Cox (MR wts)	2-step RGLR (MR wts)
Stratum 1 $\hat{\beta}_1$	112 (73%)	-0.64*	-0.27	-0.26	-0.27	-0.26
Stratum 2 $\hat{\beta}_2$	42 (27%)	-0.64*	-1.16	-1.14	-1.16	-1.14
$\hat{\bar{\beta}}$		-0.64*	-0.51	-0.50	-0.54	-0.53
95% C.I.		(-1.18, -0.10)	(-1.08, 0.05)	(-1.07, 0.06)	(-1.10, 0.02)	(-1.09, 0.02)
P-value		0.021	0.075	0.080	0.057	0.060

*The stratified Cox model assumes $\beta_1 = \beta_2$; these are the implied stratum-specific estimates based on the overall estimate.

4.5. Discussion

The stratified Cox model is often used to analyze stratified randomized clinical trials with time-to-event data. However, the assumption of equal hazard ratios across strata may not be true in real applications. Therefore, it is important to develop methods to handle a treatment by stratum interaction, especially in relatively small stratified trials with low power to detect a treatment by stratum interaction. In this work, we proposed a two-step RGLR approach in which we estimate stratum-specific log hazard ratios using the RGLR approach and combine them across strata using SS or MR weights. Through simulation studies, we have shown that the two-step RGLR provides notably smaller bias and smaller mean squared error than the conventional stratified Cox model when there is a treatment-by-stratum interaction, with similar performance when there is no interaction. The stratified Cox model is subject to have inflated type I error in small samples, while the two-step RGLR does not. The stratified Cox model also has trouble with CI under-coverage in small samples, while the two-step RGLR with SS weights does not and with MR weights generally does not. The two-step RGLR method also delivers much higher power than the stratified Cox model when the hazard ratio differs across strata while suffering no material power loss in other cases. Finally,

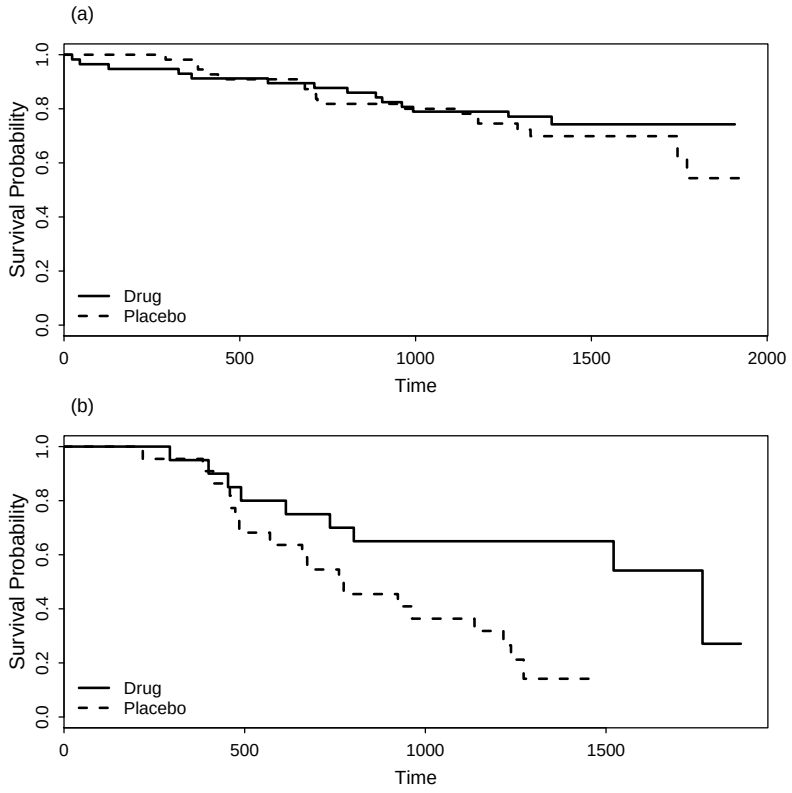


Figure 4.1: Kaplan-Meier survival curves by treatment group; (a) is for stratum 1 (b) is stratum 2.

the proposed method has similar or better performance than the two-step method of Mehrotra, Su, and Li (2012) in terms of bias and mean squared error; this to be be expected because within each stratum, the RGLR estimator outperforms the Cox model estimator in small to moderate sample sizes, notably so in small samples.

The two-step RGLR removes the restrictive assumption of equal hazard ratios across strata in the stratified Cox model analysis, and outperforms the stratified Cox model when there is an interaction between treatment and stratum. More importantly, the two-step RGLR also provides an estimated stratum-specific hazard ratio, while the stratified Cox model only provides an estimated overall hazard ratio. As shown in the colon cancer example, when the hazard ratio is different across stratum, using the two-step RGLR can provide additional insight into the difference across strata, while the stratified Cox model does not.

CHAPTER 5

DISCUSSION

5.1. Summary

In this dissertation, we developed methods for analyzing time-to-event data in small samples under both crossover and parallel designs. In Chapter 2, we examined situations in 2×2 crossover trials with time-to-event endpoints and proposed a regression-based method to incorporate baseline information to improve the efficiency. We proposed to use multiple imputation with multiple candidate models, to impute censored outcomes in post-treatment. In each imputed data set, we applied ANCOVA on the log-transformed event times, with the difference in period-specific baseline measurement as a covariate, to estimate the log treatment-ratio of geometric means. Finally, we used frequentist model averaging with AIC weighting and Rubin's combination rule for multiple imputation to combine the results from the candidate models. We compared our method to existing methods, the H-R test and stratified Cox model, and showed through extensive numerial studies that our method was able to deliver more or as efficient results. Additionally, we were also able to provide a point estimate on the ratio of geometric means between the two treatments, while H-R test fails to do so. For symmetric distributions, the ratio of geometric means is approximately equal to the ratio of medians, which is a commonly used measure for time-to-event outcomes. Therefore, we were able to provide a meaningful estimate of the treatment effect. For ease of illustration, we used the log-normal and Weibull as two candidate models to impute the censored values, because they are flexible to capture a variety of distribution shapes for in survival data. Even by using only two models, we delivered higher power than the stratified Cox model, and more or similar power as H-R test, even when the true underlying distribution is not included in the candidate model. In practice, the number and choice of candidate models can be changed to fit the anticipated potential distributions for a given setting, and we believe that adding more candidate models will only improve the efficiency of our proposed method more.

In Chapter 3, we focused on improving hazard ratio estimation in small parallel clinial trials in the setting of proportional hazards. We proposed a refined generalized log-rank (RGLR) statistic that replaced the estimation of nuisance parameters with the exact counterpart in the original general-

ized log-rank (GLR) statistic by Mehrotra and Roth (2001). We also provided a more intuitive development for the nuisance parameter estimation using inverse-variance weighting in GLR statistic. We showed that RGLR reduced bias significantly, compared to GLR, Cox and parametric models, and maintained high relative efficiency versus the Cox model in small samples. Our proposed RGLR also controlled the type I error rate and maintained the nominal coverage probability in small samples, while Cox and parametric models were subject to type I error inflation when there were fewer than 40 subjects per group. Additionally, the true underlying distribution is often unknown in real data applications, and thus, parametric models are subject to misspecification. We showed that our proposed method was not subject to misspecification, and consistently delivered low bias and higher efficiency relative to the Cox model.

In Chapter 4, we further extended the RGLR statistic to allow for stratification factors and proposed the two-step RGLR method, in which stratum-specific log hazard ratio was first obtained using RGLR and the overall log hazard ratio was combined using two different weighting schemes, sample size and minimum risk weights. In addition, we also developed a variance estimator for the RGLR estimate of the log hazard ratio and demonstrated its accuracy through simulation studies. We showed that the two-step RGLR method provided notably smaller bias and mean squared error than the conventional stratified Cox model in the presence of a treatment by stratum interaction, and delivered similar performance when there was no interaction. Compared to the two-step Cox model method by Mehrotra, Su, and Li (2012), our two-step RGLR also had a similar or better performance in terms of bias and mean squared error in small samples; the former was developed for larger sample sizes while the RGLR approach provided notably better performance in small samples as shown in Chapter 3. The stratified and two-step Cox model methods also suffered from inflated type I error, while our two-step RGLR did not. When there was an interaction between treatment and stratum, the two-step RGLR was able to deliver higher power than the stratified Cox model.

5.2. Future Directions

5.2.1. Non-parametric ANCOVA

There are several interesting directions to consider for future study of crossover studies with time-to-event outcomes. The method we proposed in Chapter 2 used parametric ANCOVA to estimate

the treatment effect between the two treatments. We can also use non-parametric ANCOVA to potentially further improve the efficiency of our proposed method. Parametric ANCOVA still poses some underlying normality assumptions on the event times, while non-parametric approach removes the assumption completely. It is expected that when the underlying normality assumption is truly not met, by non-parametric ANCOVA can provide a more robust and efficient result. However, the point estimator directly from non-parametric ANCOVA does not provide a meaningful interpretation. Thus, we also need to incorporate a method that can invert the test from non-parametric approach to have a meaningful point estimator on treatment effect.

5.2.2. Baseline Censoring

In Chapter 2, we assumed no censoring in the baseline measurement, and only imputed the censored values in post-treatment. However, it is possible that in some real data applications, censored values can also be observed in the baseline measurement. Therefore, we are interested in loosing this assumption, and extending our method to allow for censoring in the baseline measurement. One possible direction is to use other characteristics of the subjects, such as gender, age and sex, to first impute the censored baseline measurements, to have complete data in baseline, and then proceed with the method as proposed.

5.2.3. Comparison to Lin et al. (2016) Methods for Stratified Trials

Lin et al. (2016) recently proposed a solution to estimating a confidence interval based on the score test statistic from the stratified Cox model, and proposed to handle tied event times using the Breslow (1974) method. They assumed a constant hazard ratio across the strata and used a sub-optimal approach to handle tied event times. On contrary, we proposed a two-step RGLR method for stratified clinical trials with time-to-event outcomes in Chapter 4, and proposed to use Efron's method for handling ties in the RGLR method in Chapter 3. We allowed for a treatment by stratum interaction and demonstrated better performance for the two-step RGLR compared the conventional stratified Cox model. Thus, we are interested in comparing our proposed two-step RGLR method to the Lin et al. (2016) method.

APPENDIX A

SUPPLEMENTARY MATERIALS FOR CHAPTER 2

A.1. Supplementary Figure

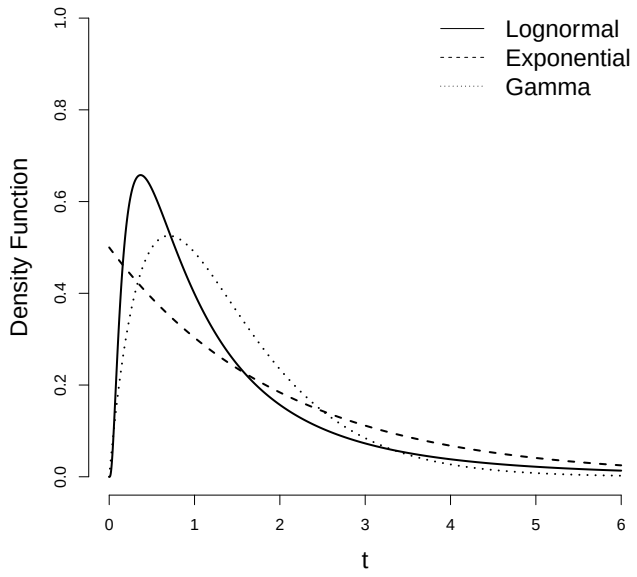


Figure A.1: Density curves for survival time under lognormal($\mu = 0, \sigma = 1$), where μ and σ denotes the mean and standard deviation on the log scale, exponential(rate=0.5) and gamma(shape=2, scale=0.7), respectively.

A.2. Supplementary Tables

Table A.1: True θ values used in the simulation study under the alternative hypothesis for each combination of distribution, covariance structure, $\bar{\rho}$, censoring and sample size per sequence ($\theta = 1$ under the null hypothesis.)

		$\bar{\rho} = 0.5$						$\bar{\rho} = 0.7$					
		10% Censoring			50% Censoring			10% Censoring			50% Censoring		
Distribution	N/seq / Σ	12	24	48	12	24	48	12	24	48	12	24	48
Log-normal	CS	1.85	1.6	1.4	1.95	1.7	1.5	1.65	1.5	1.3	1.9	1.6	1.35
	AR(1)	1.85	1.6	1.4	1.95	1.7	1.5	1.65	1.5	1.3	1.9	1.6	1.35
	EP	1.85	1.6	1.4	1.95	1.7	1.4	1.6	1.5	1.25	1.75	1.45	1.35
Exponential	CS	2	1.7	1.5	2.1	1.8	1.6	1.8	1.5	1.35	2	1.7	1.5
	AR(1)	2	1.7	1.5	2.1	1.8	1.6	1.8	1.5	1.3	2	1.7	1.4
	EP	2	1.7	1.4	2.1	1.8	1.6	1.8	1.5	1.3	2	1.7	1.4
Gamma	CS	2.7	1.8	1.5	3.7	2.2	1.6	2.2	1.5	1.35	3	1.8	1.5
	AR(1)	2.7	1.8	1.5	3.7	2.2	1.6	2.2	1.5	1.35	3	1.8	1.5
	EP	2.7	1.8	1.4	3.7	2.2	1.6	2	1.5	1.3	3	1.8	1.4

CS: compound symmetry covariance structure. AR(1): first-order autoregressive covariance structure. EP: equipredicability covariance structure. $\bar{\rho}$: mean pairwise correlation.

Table A.2: Power (%) for the hierarchical rank test (H-R), stratified Cox model with baseline adjustment (SCB) and proposed multiple imputation with model averaging and ANCOVA (MI^{MA}) under log-normal distribution based on 5000 simulations.

		$\bar{\rho} = 0.5$						$\bar{\rho} = 0.7$					
		10% Censoring			50% Censoring			10% Censoring			50% Censoring		
Σ	N/seq / Method	12	24	48	12	24	48	12	24	48	12	24	48
CS	H-R	69.3	78.4	83.1	71.6	76.8	79.6	72.7	86.7	81.5	84.4	87.9	80.3
	SCB	NC	69.8	74.3	NC	NC	71.3	NC	78.9	73.0	NC	77.5	73.0
	MI^{MA}	76.3	84.9	89.4	65.8	77.8	86.3	80.2	92.1	88.7	75.7	87.2	80.9
AR(1)	H-R	68.6	79.2	80.8	69.1	77.4	76.9	71.9	85.7	80.0	82.6	83.7	77.9
	SCB	NC	72.5	74.6	NC	65.4	69.8	NC	80.0	73.7	NC	69.4	71.9
	MI^{MA}	80.4	89.5	92.0	67.7	79.9	85.8	86.0	95.9	92.3	77.7	84.5	84.6
EP	H-R	70.2	79.1	75.1	70.9	76.3	79.8	68.1	73.4	67.9	82.8	66.9	78.6
	SCB	NC	71.7	67.2	NC	NC	72.4	NC	NC	68.9	NC	NC	75.9
	MI^{MA}	83.1	91.1	87.3	69.8	81.7	89.8	94.2	84.6	94.8	85.1	78.9	93.5

NC: Non-convergence issues. CS: compound symmetry covariance structure. AR(1): first-order autoregressive covariance structure. EP: equipredicability covariance structure. $\bar{\rho}$: mean pairwise correlation. True values of θ used under the alternative hypothesis are provided in Table A.1.

Table A.3: Power (%) for the hierarchical rank test (H-R), stratified Cox model with baseline adjustment (SCB) and proposed multiple imputation with model averaging and ANCOVA (MI^{MA}) under exponential distribution based on 5000 simulations.

		$\overline{\rho} = 0.5$						$\overline{\rho} = 0.7$					
		10% Censoring			50% Censoring			10% Censoring			50% Censoring		
Σ	N/seq / Method	12	24	48	12	24	48	12	24	48	12	24	48
CS	H-R	80.0	86.8	92.0	67.4	76.4	85.7	84.8	84.2	88.0	76.8	81.4	86.7
	SCB	NC	69.7	78.5	NC	61.8	75.1	NC	67.2	72.9	NC	69.1	76.4
	MI^{MA}	70.1	78.6	85.2	63.3	78.6	89.7	76.6	76.8	81.5	74.0	86.6	93.2
AR(1)	H-R	78.5	86.6	91.7	66.4	82.9	83.6	82.3	82.9	87.0	74.0	82.6	85.4
	SCB	NC	72.1	78.6	NC	74.6	73.7	NC	68.7	73.3	NC	69.6	76.7
	MI^{MA}	75.2	82.8	88.3	67.0	89.0	92.3	80.8	82.8	86.4	78.9	90.9	95.2
EP	H-R	78.8	87.0	92.4	84.7	77.2	85.2	84.2	84.8	77.3	75.0	82.7	71.0
	SCB	NC	73.2	80.0	NC	64.9	75.5	NC	75.4	71.3	NC	NC	64.1
	MI^{MA}	75.3	84.2	89.7	92.2	84.6	93.5	92.9	93.6	91.5	89.4	96.6	93.4

NC: Non-convergence issues. CS: compound symmetry covariance structure. AR(1): first-order autoregressive covariance structure. EP: equipredicability covariance structure. $\bar{\rho}$: mean pairwise correlation. True values of θ used under the alternative hypothesis are provided in Table A.1.

Table A.4: Power (%) for the hierarchical rank test (H-R), stratified Cox model with baseline adjustment (SCB) and proposed multiple imputation with model averaging and ANCOVA (MI^{MA}) under gamma distribution based on 5000 simulations.

		$\overline{\rho} = 0.5$						$\overline{\rho} = 0.7$					
		10% Censoring			50% Censoring			10% Censoring			50% Censoring		
Σ	N/seq / Method	12	24	48	12	24	48	12	24	48	12	24	48
CS	H-R	87.5	86.3	87.7	86.9	87.8	89.2	91.3	78.2	86.9	92.5	86.0	88.9
	SCB	NC	71.3	72.4	NC	76.2	78.6	NC	61.7	72.3	NC	74.0	79.4
	MI^{MA}	82.7	82.2	83.2	78.7	86.2	89.2	87.8	73.5	82.5	85.9	84.7	88.7
AR(1)	H-R	86.9	85.7	88.0	89.0	92.0	87.4	90.8	77.7	86.3	91.7	84.4	84.3
	SCB	NC	72.1	74.9	NC	82.7	78.3	NC	63.9	74.1	NC	74.0	76.0
	MI^{MA}	86.2	85.6	87.2	84.8	92.9	90.6	91.8	80.8	88.4	89.3	87.1	88.9
EP	H-R	87.6	86.1	79.8	88.2	86.9	88.1	85.3	78.7	75.1	91.4	82.3	78.7
	SCB	NC	74.6	66.1	NC	77.0	80.2	NC	72.2	69.6	NC	72.8	74.7
	MI^{MA}	87.0	87.0	80.8	86.5	89.4	92.7	95.9	92.5	90.8	93.0	92.7	92.4

NC: Non-convergence issues. CS: compound symmetry covariance structure. AR(1): first-order autoregressive covariance structure. EP: equipredicability covariance structure. $\bar{\rho}$: mean pairwise correlation. True values of θ used under the alternative hypothesis are provided in Table A.1.

A.3. R code for Data Application Example

```
###load packages
library(survival)
library(perm)
library(truncdist)
library(mvtnorm)

######read in data
```

```
data<-read.csv('treadmill dataset.csv',header=T)

################################
##########H-R Test#############
################################
n=40
t1=data$p1 ###post-trt time in period 1
t2=data$p2 ###post-trt time in period 2
delta1=data$deltap1  ###post-trt censoring indicator in period 1
delta2=data$deltap2  ###post-trt censoring indicator in period 2

hr_rank=rep(NA,n)
#patients with only one event get extreme ranks
ind_p2=which(delta1==0&delta2==1)
n01=length(ind_p2)
t2_sorted=sort(t2[ind_p2],index.return=T)
ind_p2rank=ind_p2[t2_sorted$ix]
hr_rank[ind_p2rank]=1:n01

ind_p1=which(delta1==1&delta2==0)
n10=length(ind_p1)
t1_sorted=sort(t1[ind_p1],index.return=T,decreasing=T)
ind_p1rank=ind_p1[t1_sorted$ix]
hr_rank[ind_p1rank]=(n-n10+1):n

#patients with events in both periods have less extreme ranks
ind_both1=which(delta1==1&delta2==1&t1>t2)
n11_1=length(ind_both1)
diff1=sort(t1[ind_both1]-t2[ind_both1],index.return=T,decreasing=T)
both1rank=ind_both1[diff1$ix]
hr_rank[both1rank]=(n01+1):(n01+n11_1)

ind_both2=which(delta1==1&delta2==1&t1<=t2)
n11_2=length(ind_both2)
diff2=sort(t2[ind_both2]-t1[ind_both2],index.return=T)
both2rank=ind_both2[diff2$ix]
hr_rank[both2rank]=(n-n10-n11_2+1):(n-n10)

#patients with no events in either period have the average rank
ind_neither=which(delta1==0&delta2==0)
hr_rank[ind_neither]=(n01+n11_1+1+n-(n10+n11_2))/2

##two group wilcoxon signed rank test
pval=permTS(hr_rank[1:(n/2)],hr_rank[(n/2+1):n],exact=T,
            control=permControl(setSEED=FALSE))$p.value

################################
####Stratified Cox#############
################################
####transform the data to long format
data_long=reshape(data,
```

```
                  varying=c("deltab1","b1","deltap1","p1",
                  "deltab2","b2","deltap2","p2"),
                  v.names=c("delta","event_time"),
                  timevar=c("b1","p1","b2","p2"),
                  times=c(1,2,3,4),
                  new.row.names=1:(4*n),
                  direction="long")
data_long <- data_long[order(data_long$id),]
data_long$trt_ind=c(rep(c(NA,0,NA,1),n/2),rep(c(NA,1,NA,0),n/2))
data_long$period_ind=rep(c(0,0,1,1),n)

model_scb=coxph(Surv(event_time[time==2|time==4],delta[time==2|time==4])
                ~trt_ind[time==2|time==4]+period_ind[time==2|time==4]
                +event_time[time==1|time==3]+strata(id[time==2|time==4]),
                data=data_long)

################################
#####Proposed Method###########
################################
tau=10 ##end of time
###period 1
num_missing_p1=length(which(data$deltap1==0))
##period 2
num_missing_p2=length(which(data$deltap2==0))
#####BA sequence, period 1
id_BA_p1=which(data$deltap1==0&data$seq=="ba")
num_missing_BA_p1=length(id_BA_p1)
#####AB sequence, period 2
id_AB_p2=which(data$deltap2==0&data$seq=="ab")
num_missing_AB_p2=length(id_AB_p2)
#####BA sequence, period 1
id_BA_p2=which(data$deltap2==0&data$seq=="ba")
num_missing_BA_p2=length(id_BA_p2)

#####function to resample coefficients and calculate mean
resample_p1<-function(model,b1){
  cov=summary(model)$var
  coef_all=summary(model)$table[,1]

  ##resample coef and scale(sigma) together from N
  coef_all_new=rmvnorm(1,mean=coef_all,cov)
  coef_hat_p1_new=coef_all_new[-length(coef_all_new)]
  sd_hat_p1_new=exp(coef_all_new[length(coef_all_new)])
  #get the mean for each person with censored data
  mu_hat_BA_p1=coef_hat_p1_new[1]+coef_hat_p1_new[2]+b1*coef_hat_p1_new[3]
  return(list(mu_hat_BA_p1,sd_hat_p1_new))
}

resample_p2<-function(model,b1_AB,p1_AB,b2_AB,b1_BA,p1_BA,b2_BA){
  cov_p2=summary(model)$var
  coef_all_p2=summary(model)$table[,1]
```

```
  ##resample coef and scale(sigma) together from N
  coef_all_p2_new=rmvnorm(1,mean=coef_all_p2,cov_p2)
  coef_hat_p2_new=coef_all_p2_new[-length(coef_all_p2_new)]
  sd_hat_p2_new=exp(coef_all_p2_new[length(coef_all_p2_new)])

  #get the mean for each person with censored data
  mu_hat_AB_p2=coef_hat_p2_new[1]+coef_hat_p2_new[2]+b1_AB*coef_hat_p2_new[3]+
    p1_AB*coef_hat_p2_new[4]+b2_AB*coef_hat_p2_new[5]

  mu_hat_BA_p2=coef_hat_p2_new[1]+b1_BA*coef_hat_p2_new[3]+
    p1_BA*coef_hat_p2_new[4]+b2_BA*coef_hat_p2_new[5]
  return(list(mu_hat_AB_p2,mu_hat_BA_p2,sd_hat_p2_new))
}

###impute from log-normal
impute_fn<-function(mu_hat,sd_hat){
  lowerbd=pnorm((log(tau)-mu_hat)/sd_hat)
  z_unif=sapply(lowerbd,function(x) runif(1,min=x))
  imputed=mu_hat+qnorm(z_unif)*sd_hat
  return(imputed)
}

########################################
######Impute assuming log-normal#########
########################################
#First, impute censored values in period 1
data$trt_ind=c(rep(0,20),rep(1,20))
model_p1=survreg(Surv(log(p1),deltap1,type='right')~trt_ind+log(b1),
                 dist='gaussian',robust=T,data=data)
out_ln_p1=resample_p1(model_p1,log(data$b1[data$id %in% id_BA_p1]))
mu_hat_BA_p1=out_ln_p1[[1]]
sd_hat_p1=out_ln_p1[[2]]

data$p1_imputed_ln=data$p1
data$deltap1_imputed=data$deltap1
#####################################################
########Impute from Weibull#########################
#####################################################
#First, impute censored values in period 1
######censored in post-trt period 1
model_p1_weib=survreg(Surv(p1,deltap1,type='right')~trt_ind+b1,
                          dist='weibull',robust=T,data=data)
out_weib_p1=resample_p1(model_p1_weib,data$b1[data$id%in%id_BA_p1])
mu_hat_BA_p1_weib=out_weib_p1[[1]]
sd_hat_p1_weib=out_weib_p1[[2]]
data$p1_imputed_weib=data$p1

####impute m times
m=50
aic_ln_p=rep(NA,m)
```

```
aic_weib_p=rep(NA,m)
beta_hat=rep(NA,m)
se_beta_hat=rep(NA,m)
beta_hat_log_weib=rep(NA,m)
se_beta_hat_log_weib=rep(NA,m)

for (j in 1:m){
  ##############################################
  ##########log-normal##########################
  ##############################################
  ####First, impute censored values in period 1
  data$p1_imputed_ln[id_BA_p1]=exp(impute_fn(mu_hat_BA_p1,sd_hat_p1))
  data$deltap1_imputed[id_BA_p1]=1

  ####Second,impute censored values in period 2
  data$trt_ind_p2=c(rep(1,20),rep(0,20))
  model_p2_censored=survreg(Surv(log(p2),deltap2,type='right')~trt_ind_p2
                            +log(b1)+log(p1_imputed_ln)
                            +log(b2),dist='gaussian',robust=T,data=data)
  out_ln_p2=resample_p2(model=model_p2_censored,
              b1_AB=log(data$b1[data$id%in%id_AB_p2]),
              p1_AB=log(data$p1_imputed_ln[data$id%in%id_AB_p2]),
              b2_AB=log(data$b2[data$id%in%id_AB_p2]),
              b1_BA=log(data$b1[data$id%in%id_BA_p2]),
              p1_BA=log(data$p1_imputed_ln[data$id %in% id_BA_p2]),
              b2_BA=log(data$b2[data$id%in%id_BA_p2]))
  mu_hat_AB_p2=out_ln_p2[[1]]
  mu_hat_BA_p2=out_ln_p2[[2]]
  sd_hat_p2=out_ln_p2[[3]]

  data$p2_imputed_ln=data$p2
  data$p2_imputed_ln[id_AB_p2]=exp(impute_fn(mu_hat_AB_p2,sd_hat_p2))
  data$p2_imputed_ln[id_BA_p2]=exp(impute_fn(mu_hat_BA_p2,sd_hat_p2))

  ######ANCOVA#############################
  logy=log(data$p1_imputed_ln)-log(data$p2_imputed_ln)
  logx=log(data$b1)-log(data$b2)
  model=lm(logy~logx+as.numeric(data$seq=="ab"))
  aic_ln_p[j]=AIC(model)
  beta_hat[j]=-coef(model)[3]/2
  se_beta_hat[j]=coef(summary(model))[3,2]/2

  ##############################################
  ##########Weibull#############################
  ##############################################
  ############First, impute censored values in period 1
  data$p1_imputed_weib[id_BA_p1]=sapply(exp(mu_hat_BA_p1_weib),function(x)
    rtrunc(1,spec="weibull",a=tau,shape=1/sd_hat_p1_weib,scale=x))

  ####Second, impute censored values in period 2
  model_p2_censored_weib=survreg(Surv(p2,deltap2,type='right')~trt_ind_p2+b1
```

```
                                +p1_imputed_weib+b2,
                                dist='weibull',robust=T,data=data)
  out_weib_p2=resample_p2(model=model_p2_censored_weib,
                        b1_AB=data$b1[data$id%in%id_AB_p2],
                        p1_AB=data$p1_imputed_weib[data$id%in%id_AB_p2],
                        b2_AB=data$b2[data$id%in%id_AB_p2],
                        b1_BA=data$b1[data$id%in%id_BA_p2],
                        p1_BA=data$p1_imputed_weib[data$id %in% id_BA_p2],
                        b2_BA=data$b2[data$id%in%id_BA_p2])
  mu_hat_AB_p2_weib=out_weib_p2[[1]]
  mu_hat_BA_p2_weib=out_weib_p2[[2]]
  sd_hat_p2_weib=out_weib_p2[[3]]

  data$p2_imputed_weib=data$p2
  data$p2_imputed_weib[id_AB_p2]=sapply(exp(mu_hat_AB_p2_weib),function(x)
    rtrunc(1,spec="weibull",a=tau,shape=1/sd_hat_p2_weib,scale=x))
  data$p2_imputed_weib[id_BA_p2]=sapply(exp(mu_hat_BA_p2_weib),function(x)
    rtrunc(1,spec="weibull",a=tau,shape=1/sd_hat_p2_weib,scale=x))

  ######ANCOVA#############################
  logy_weib=log(data$p1_imputed_weib)-log(data$p2_imputed_weib)
  logx_weib=log(data$b1)-log(data$b2)
  model_log_weib=lm(logy_weib~logx_weib+as.numeric(data$seq=="ab"))
  aic_weib_p[j]=AIC(model_log_weib)
  beta_hat_log_weib[j]=-coef(model_log_weib)[3]/2
  se_beta_hat_log_weib[j]=coef(summary(model_log_weib))[3,2]/2
}

#########################################################
#######Combine the beta's from model averaging ######
#########################################################
w_lognormal_p=exp(-aic_ln_p/2)/(exp(-aic_ln_p/2)+exp(-aic_weib_p/2))
w_weib_p=exp(-aic_weib_p/2)/(exp(-aic_ln_p/2)+exp(-aic_weib_p/2))
beta_hat_log_combined=beta_hat*w_lognormal_p+beta_hat_log_weib*w_weib_p
####Within imputation variance from model averaging
var_ln=rep(NA,m)
var_weib=rep(NA,m)
for(j in 1:m){
  var_ln[j]=w_lognormal_p[j]*sqrt(se_beta_hat[j]^2+(beta_hat[j]
               -beta_hat_log_combined[j])^2)
  var_weib[j]=w_weib_p[j]*sqrt(se_beta_hat_log_weib[j]^2
                    +(beta_hat_log_weib[j]-beta_hat_log_combined[j])^2)
}
####combine beta across multiple imputation
beta_hat_mean=mean(beta_hat_log_combined)

###calculate p-value
within_imp_var=mean((var_ln+var_weib)^2)
btw_imp_var=sum((beta_hat-beta_hat_mean)^2)/(m-1)
total_var=within_imp_var+(1+1/m)*btw_imp_var
gamma=(1+1/m)*btw_imp_var/total_var
```

```
dof_com=mean(n-num_missing_p1,n-num_missing_p2)-3
##degrees of freedom
dm=(m-1)*(1+within_imp_var/((1+1/m)*btw_imp_var))^2
##degrees of freedom for small samples
dof=(1-gamma)*((dof_com+1)/(dof_com+3))*dof_com
ts=abs(beta_hat_mean/sqrt(total_var))
dof_obs=(1-gamma)*((dof_com+1)/(dof_com+3))*dof_com
dof=1/(1/dof_obs+1/dm)
###Point estimate
exp(beta_hat_mean)
##P-value
pval_base=2*(1-pt(ts,df=dof))
##CI
exp(c(beta_hat_mean-qt(0.975,df=dof)*sqrt(total_var),
      beta_hat_mean+qt(0.975,df=dof)*sqrt(total_var)))
```

APPENDIX B

SUPPLEMENTARY MATERIALS FOR CHAPTER 3

B.1. Two Approaches for Nuisance Parameters Estimation for the RGLR Statistic

We show here the details of the calculation for nuisance parameter in RGLR statistic using two different approaches: MLE from the likelihood of two Binomials, and inverse-variance weighted average.

B.1.1. MLE from Likelihood

We have two Binomial distributions for the number of events in group A and group B, namely $d_{iB} \sim$ Binomial (R_{iB}, π_{iB}), where $\pi_{iB} = 1 - e^{-p_i}$, and $d_{iA} \sim$ Binomial (R_{iA}, π_{iA}), where $\pi_{iA} = 1 - e^{-\theta p_i}$. Then, the likelihood from the two Binomial is

$$\begin{aligned} L &= \pi_{iA}^{d_{iA}}(1-\pi_{iA})^{R_{iA}-d_{iA}}\pi_{iB}^{d_{iB}}(1-\pi_{iB})^{R_{iB}-d_{iB}} \\ &= (1-e^{-\theta p_i})^{d_{iA}}(e^{-\theta p_i})^{R_{iA}-d_{iA}}(1-e^{-p_i})^{d_{iB}}(e^{-p_i})^{R_{iB}-d_{iB}} \end{aligned}$$

Take the log of the likelihood, we have

$$l = d_{iA}\log(1-e^{-\theta p_i}) - \theta p_i(R_{iA}-d_{iA}) + d_{iB}\log(1-e^{-p_i}) - p_i(R_{iB}-d_{iB})$$

Take derivative with respect to p_i,

$$\partial l/\partial p_i = \frac{d_{iA}\theta e^{-\theta p_i}}{1-e^{-\theta p_i}} - \theta(R_{iA}-d_{iA}) + \frac{d_{iB}e^{-p_i}}{1-e^{-p_i}} - (R_{iB}-d_{iB}) = 0 \tag{B.1}$$

Because there is only one person having an event at any time t_i, i.e., $d_{iA} = 0, d_{iB} = 1$ or $d_{iB} = 0, d_{iA} = 1$, we can use this to simplify the equation above to solve for p_i.

Case1: When $d_{iA} = 0, d_{iB} = 1$, equation (B.1) becomes

$$\begin{aligned} -\theta R_{iA} + \frac{e^{-p_i}}{1 - e^{-p_i}} - (R_{iB} - 1) &= 0 \\ (1 - e^{-p_i})(-\theta R_{iA} - R_{iB} + 1) + e^{-p_i} &= 0 \\ e^{-p_i}(\theta R_{iA} + R_{iB}) &= \theta R_{iA} + R_{iB} - 1 \\ p_i &= \log\left(\frac{\theta R_{iA} + R_{iB}}{\theta R_{iA} + R_{iB} - 1}\right) \end{aligned}$$

Case2: When $d_{iB} = 0, d_{iA} = 1$, equation (B.1) becomes

$$\begin{aligned} \frac{\theta e^{-\theta p_i}}{1 - e^{-\theta p_i}} - \theta(R_{iA} - 1) - R_{iB} &= 0 \\ \frac{\theta}{e^{\theta p_i} - 1} - \theta(R_{iA} - 1) - R_{iB} &= 0 \\ p_i &= \log\left(\frac{\theta R_{iA} + R_{iB}}{\theta R_{iA} + R_{iB} - \theta}\right) \end{aligned}$$

B.1.2. Inverse-Variance Weighting

Again, we have two Binomial distributions, $d_{iB} \sim$ Binomial (R_{iB}, π_{iB}), where $\pi_{iB} = 1 - e^{-p_i}$, and $d_{iA} \sim$ Binomial (R_{iA}, π_{iA}), where $\pi_{iA} = 1 - e^{-\theta p_i}$. Then, naturally, we have 2 point estimates for the nuisance parameter p_i from the two Binomial distributions:
$\hat{p}_{iB} = -\log\left(1 - \frac{d_{iB}}{R_{iB}}\right)$, and $\hat{p}_{iA} = -\frac{1}{\theta}\log\left(1 - \frac{d_{iA}}{R_{iA}}\right)$

We also know that Var(d_{iB}) $= R_{iB}\pi_{iB}(1 - \pi_{iB})$ and Var(d_{iA}) $= R_{iA}\pi_{iA}(1 - \pi_{iA})$, so we can take the average of the 2 point estimates weighted by inverse-variance.

To compute the variance for $\hat{p}_{iB}$, by definition, we have

$$\begin{aligned} \text{Var}(\hat{p}_{iB}) &= \text{Var}\left[-\log\left(1 - \frac{d_{iB}}{R_{iB}}\right)\right] \\ &= E\left[\left\{-\log\left(1 - \frac{d_{iB}}{R_{iB}}\right)\right\}^2\right] - \left\{E\left[-\log\left(1 - \frac{d_{iB}}{R_{iB}}\right)\right]\right\}^2 \end{aligned}$$

The formula for the two expectations seem very complex, but we can simplify and approximate

them using the assumption of no tied observations. By definition,

$$E\left[-\log\left(1-\frac{d_{iB}}{R_{iB}}\right)\right] \tag{B.2}$$

$$=-\log\left(1-\frac{0}{R_{iB}}\right)\cdot P(d_{iB}=0)-\log\left(1-\frac{1}{R_{iB}}\right)\cdot P(d_{iB}=1) \tag{B.3}$$

$$-\log\left(1-\frac{2}{R_{iB}}\right)\cdot P(d_{iB}=2)-\cdots-\log\left(1-\frac{R_{iB}}{R_{iB}}\right)\cdot P(d_{iB}=R_{iB}) \tag{B.4}$$

Because d_{iB} can only be 0 or 1, we can think of the probability of $d_{iB}>1$ to be very close to 0 and thus get rid of the cases where $d_{iB}>1$. Thus, equation (B.2) can be approximated by

$$E\left[-\log\left(1-\frac{d_{iB}}{R_{iB}}\right)\right]\approx 0-\log\left(1-\frac{1}{R_{iB}}\right)\cdot P(d_{iB}=1) \tag{B.5}$$

$$=-\log\left(1-\frac{1}{R_{iB}}\right)R_{iB}\pi_{iB}(1-\pi_{iB})^{R_{iB}-1} \tag{B.6}$$

$$\approx-\log\left(1-\frac{1}{R_{iB}}\right)R_{iB}\pi_{iB} \tag{B.7}$$

Again, given that the probability mass is mainly on 0 and 1, π_{iB} should be close to 0, and thus we can approximate $(1-\pi_{iB})$ to be 1 and simplify equation B.6 to B.7.

Similarly, we have

$$E\left[\left\{-\log\left(1-\frac{d_{iB}}{R_{iB}}\right)\right\}^2\right]\approx\left[\log\left(1-\frac{1}{R_{iB}}\right)\right]^2 R_{iB}\pi_{iB}$$

Therefore,

$$\begin{aligned}\text{Var}(\hat{p}_{iB}) &= \left[\log\left(1-\frac{1}{R_{iB}}\right)\right]^2 R_{iB}\pi_{iB}(1-R_{iB}\pi_{iB})\\ &= \left[\log\left(1-\frac{1}{R_{iB}}\right)\right]^2 R_{iB}(1-e^{-p_i})[1-R_{iB}(1-e^{-p_i})]\end{aligned}$$

Follow a similar logic, we can derive the variance of $\hat{p}_{iA}$

$$\begin{aligned}\text{Var}(\hat{p}_{iA}) &= \left[\log\left(1-\frac{1}{R_{iA}}\right)\right]^2 R_{iA}\pi_{iA}(1-R_{iA}\pi_{iA})\\ &= \left[\log\left(1-\frac{1}{R_{iA}}\right)\right]^2 R_{iA}(1-e^{-\theta p_i})[1-R_{iA}(1-e^{-\theta p_i})]\end{aligned}$$

Now, use the inverse-variance weighting, and equate the true p_i to the average,

$$p_i = \frac{\frac{\hat{p}_{iB}}{\mathsf{Var}(\hat{p}_{iB})} + \frac{\hat{p}_{iA}}{\mathsf{Var}(\hat{p}_{iA})}}{\frac{1}{\mathsf{Var}(\hat{p}_{iB})} + \frac{1}{\mathsf{Var}(\hat{p}_{iA})}}$$
$$p_i = \frac{\hat{p}_{iB}\mathsf{Var}(\hat{p}_{iA}) + \hat{p}_{iA}\mathsf{Var}(\hat{p}_{iB})}{\mathsf{Var}(\hat{p}_{iB}) + \mathsf{Var}(\hat{p}_{iA})}$$

Use the fact that there is only one event at each time point to simplify the equation,

Case1: When $d_{iB} = 0, d_{iA} = 1$, $\hat{p}_{iB} = -\log(1-0) = 0$,

$$p_i = \frac{\hat{p}_{iA}\mathsf{Var}(\hat{p}_{iB})}{\mathsf{Var}(\hat{p}_{iB}) + \mathsf{Var}(\hat{p}_{iA})} \tag{B.8}$$
$$p_i[\mathsf{Var}(\hat{p}_{iB}) + \mathsf{Var}(\hat{p}_{iA})] \tag{B.9}$$
$$= -\frac{1}{\theta}\log\left(1 - \frac{1}{R_{iA}}\right)\left[\log\left(1 - \frac{1}{R_{iB}}\right)\right]^2 R_{iB}(1 - e^{-p_i})[1 - R_{iB}(1 - e^{-p_i})] \tag{B.10}$$

The left-hand side of equation B.9 is a product of p_i and function of e^{-p_i} and $e^{-\theta p_i}$, and the right-hand side is a function of e^{-p_i}. On the other hand, the equation (1) from the likelihood is only a function of e^{-p_i} and $e^{-\theta p_i}$.

Case2: When $d_{iB} = 1, d_{iA} = 0$, $\hat{p}_{iA} = 0$,

$$p_i = \frac{\hat{p}_{iB}\mathsf{Var}(\hat{p}_{iA})}{\mathsf{Var}(\hat{p}_{iB}) + \mathsf{Var}(\hat{p}_{iA})} \tag{B.11}$$
$$p_i[\mathsf{Var}(\hat{p}_{iB}) + \mathsf{Var}(\hat{p}_{iA})] \tag{B.12}$$
$$= -\log\left(1 - \frac{1}{R_{iB}}\right)\left[\log\left(1 - \frac{1}{R_{iA}}\right)\right]^2 R_{iA}(1 - e^{-\theta p_i})[1 - R_{iA}(1 - e^{-\theta p_i})] \tag{B.13}$$

Again, the left-hand side of equation B.12 involves both p_i and function of e^{-p_i} and $e^{-\theta p_i}$, which does not agree with equation (B.1).

For RGLR statistic, the inverse-variance weighting average and likelihood MLE approach generate two different estimates for the nuisance parameters. The MLE approach is able to provide us a closed-form under no ties assumption. However, the inverse-variance weighting average approach

results in a somewhat complex form. Although we can still solve it through numerical solutions, it will introduce approximation in the nuisance parameters. Therefore, we recommend using the MLE approach to for nuisance parameter estimation for the RGLR statistic.

B.2. Supplementary Figure

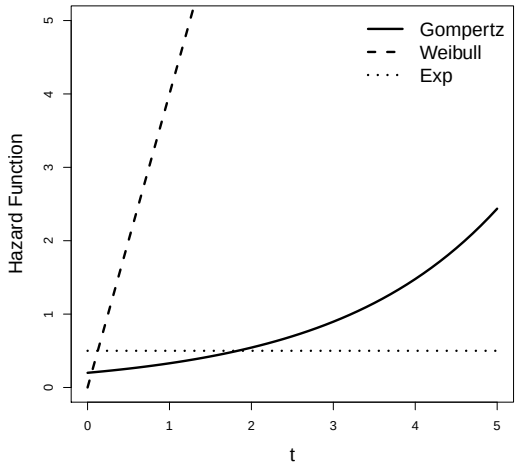


Figure B.1: Hazard function of Gompertz(shape=0.5, rate=0.2), Weibull(shape=2, rate=0.5) and Exponential(rate=0.5).

APPENDIX C

SUPPLEMENTARY MATERIALS FOR CHAPTER 4

C.1. Verification of the Variance Formula for the RGLR Estimator

To examine the accuracy of equation (4.11), we performed a simulation of 5000 replications with varying β, sample size and percentage censoring. We generated group B survival data from a Weibull distribution with shape parameter of 2 and scale parameter of 0.6; for group A, the Weibull parameters were chosen to ensure a constant hazard ratio (A:B) over time, with β =0, -0.3, -0.8 and -1.3. We studied sample sizes per group of 25, 50, 100, and 25% and 50% censoring. As shown in Table C.1, the mean of the estimated variance $\hat{V}$ from using equation (4.11) was very close to the empirical variance of $\hat{\beta}^{RGLR}$ (i.e., the variance of the 5000 $\hat{\beta}^{RGLR}$ values) for all simulated scenarios, even when the sample size was as small as 25 subjects per group with 25% censoring. This indicates that equation (4.11) is able to accurately estimate the true variance of $\hat{\beta}^{RGLR}$ within each stratum.

Table C.1: Comparison of the mean of the proposed variance estimator for the log hazard ratio to the empirical variance based on data from Weibull distribution and 5000 simulations.

N/trt	β	25% Censoring		50% Censoring	
		Var($\hat{\beta}^{RGLR}$)	Mean of $\hat{V}$	Var($\hat{\beta}^{RGLR}$)	Mean of $\hat{V}$
25	0	0.119	0.119	0.221	0.230
	-0.3	0.122	0.122	0.174	0.177
	-0.8	0.139	0.133	0.192	0.196
	-1.3	0.163	0.159	0.239	0.227
50	0	0.056	0.057	0.106	0.107
	-0.3	0.056	0.058	0.083	0.084
	-0.8	0.066	0.063	0.092	0.091
	-1.3	0.078	0.075	0.101	0.102
100	0	0.029	0.028	0.051	0.052
	-0.3	0.028	0.029	0.039	0.041
	-0.8	0.030	0.031	0.043	0.044
	-1.3	0.037	0.036	0.050	0.049

Trt: treatment group. β: log hazard ratio. Var($\hat{\beta}^{RGLR}$): empirical variance of estimated log hazard ratio. $\hat{V}$: proposed variance estimator for the refined generalized log-rank statistic (RGLR) estimator for log hazard ratio in equation (4.11).

BIBLIOGRAPHY

Akaike, H (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19.6, 716–723.

Andersen, PK and Gill, RD (1982). Cox's regression model for counting processes: a large sample study. *The Annals of Statistics*, 1100–1120.

Barnard, J and Rubin, DB (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika* 86.4, 948–955.

Bates, JM and Granger, CW (1969). The combination of forecasts. *Journal of the Operational Research Society* 20.4, 451–468.

Breslow, N (1974). Covariance analysis of censored survival data. *Biometrics*, 89–99.

Bretagnolle, J and Huber-Carol, C (1988). Effects of omitting covariates in Cox's model for survival data. *Scandinavian Journal of Statistics* 2, 125–138.

Brittain, E and Follmann, D (2011). A hierarchical rank test for crossover trials with censored data. *Statistics in Medicine* 30.30, 3507–3519.

Bryson, MC and Johnson, ME (1981). The incidence of monotone likelihood in the Cox model. *Technometrics* 23.4, 381–383.

Buckland, ST, Burnham, KP, and Augustin, NH (1997). Model selection: an integral part of inference. *Biometrics* 53.6, 603–618.

Burnham, KP and Anderson, DR (2003). *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media.

Chen, X, Meng, Z, and Zhang, J (2012). Handling of baseline measurements in the analysis of crossover trials. *Statistics in Medicine* 31.17, 1791–1803.

Cox, DR (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)* 34.2, 187–220.

Efron, B (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association* 72.359, 557–565.

Feingold, M and Gillespie, BW (1996). Cross-over trials with censored data. *Statistics in Medicine* 15.10, 953–967.

Ford, I and Norrie, J (2002). The role of covariates in estimating treatment effects and risk in long-term clinical trials. *Statistics in Medicine* 21.19, 2899–2908.

France, LA, Lewis, JA, and Kay, R (1991). The analysis of failure time data in crossover studies. *Statistics in Medicine* 10.7, 1099–1113.

Genest, C and MacKay, J (1986). The joy of copulas: bivariate distributions with uniform marginals. *The American Statistician* 40.4, 280–283.

Hansen, BE (2007). Least squares model averaging. *Econometrica* 75.4, 1175–1189.

Hansen, BE and Racine, JS (2012). Jackknife model averaging. *Journal of Econometrics* 167.1, 38–46.

Hills, M and Armitage, P (1979). The two-period cross-over clinical trial. *British Journal of Clinical Pharmacology* 8, 7–20.

Hjort, NL and Claeskens, G (2003). Frequentist model average estimators. *Journal of the American Statistical Association* 98.464, 879–899.

Johnson, ME, Tolley, HD, Bryson, MC, and Goldman, AS (1982). Covariate analysis of survival data: a small-sample study of Cox's model. *Biometrics* 38.3, 685–698.

Kalbfleisch, JD and Prentice, RL (1973). Marginal likelihoods based on Cox's regression and life model. *Biometrika*, 267–278.

Kalbfleisch, JD and Prentice, RL (1980). *The Statistical Analysis of Failure Time Data*. Hoboken, NJ: John Wiley & Sons.

Kenward, MG and Roger, JH (2010). The use of baseline covariates in crossover studies. *Biostatistics* 11.1, 1–17.

Kimchi, A, Lee, G, Amsterdam, E, Fujii, K, Krieg, P, and Mason, DT (1983). Increased exercise tolerance after nitroglycerin oral spray: a new and effective therapeutic modality in angina pectoris. *Circulation* 67.1, 124–127.

Li, YH, Klein, JP, and Moeschberger, M (1996). Effects of model misspecification in estimating covariate effects in survival analysis for small sample sizes. *Computational Statistics & Data Analysis* 22.2, 177–192.

Lin, DY, Dai, L, Cheng, G, and Sailer, MO (2016). On confidence intervals for the hazard ratio in randomized clinical trials. *Biometrics* 72.4, 1098–1102.

Mallows, CL (1973). Some comments on C p. *Technometrics* 15.4, 661–675.

Mantel, N (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports. Part 1* 50.3, 163–170.

Markman, JD, Frazer, ME, Rast, SA, McDermott, MP, Gewandter, JS, Chowdhry, AK, Czerniecka, K, Pilcher, WH, Simon, LS, and Dworkin, RH (2015). Double-blind, randomized, controlled, crossover trial of pregabalin for neurogenic claudication. *Neurology* 84.3, 265–272.

Mehrotra, DV (2014). A recommended analysis for 2$\times$ 2 crossover trials with baseline measurements. *Pharmaceutical Statistics* 13.6, 376–387.

Mehrotra, DV and Railkar, R (2000). Minimum risk weights for comparing treatments in stratified binomial trials. *Statistics in Medicine* 19.6, 811–825.

Mehrotra, DV and Roth, AJ (2001). Relative risk estimation and inference using a generalized log-rank statistic. *Statistics in Medicine* 20.14, 2099–2113.

Mehrotra, DV and Roth, AJ (2011). Improved Hazard Ratio Estimation with Tied Event Times in Small Trials. *Statistics in Biopharmaceutical Research* 3.3, 456–462.

Mehrotra, DV, Su, SC, and Li, X (2012). An efficient alternative to the stratified Cox model analysis. *Statistics in Medicine* 31.17, 1849–1856.

Metcalfe, C (2010). The analysis of cross-over trials with baseline measurements. *Statistics in Medicine* 29.30, 3211–3218.

Pagano, M and Gauvreau, K (2000). *Principles of Biostatistics*. Pacific Grove, CA: Duxbury.

Pocock, SJ (1983). *Clinical Trials: A Practical Approach*. Chichester, West Sussex, England: John Wiley & Sons.

Pocock, SJ, Assmann, SE, Enos, LE, and Kasten, LE (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in Medicine* 21.19, 2917–2930.

Raftery, AE, Madigan, D, and Hoeting, JA (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* 92.437, 179–191.

Rubin, DB (1987). *Multiple Imputation for Nonresponse in Surveys*. New York, NY, USA: John Wiley & Sons.

Schomaker, M and Heumann, C (2014). Model selection and model averaging after multiple imputation. *Computational Statistics & Data Analysis* 71, 758–770.

Schumacher, M, Olschewski, M, and Schmoor, C (1987). The impact of heterogeneity on the comparison of survival times. *Statistics in Medicine* 6.7, 773–784.

Senn, SJ (2002). *Cross-over Trials in Clinical Research*. Chichester, West Sussex, England: John Wiley.

Sklar, A (1973). Random variables, joint distribution functions, and copulas. *Kybernetika* 9.6, 449–460.

Struthers, CA and Kalbfleisch, JD (1986). Misspecified proportional hazard models. *Biometrika*, 363–369.

Yan, Z (2013). The impact of baseline covariates on the efficiency of statistical analyses of crossover designs. *Statistics in Medicine* 32.6, 956–963.