CURRENT TOPICS ON TREATMENT COMPARISON FOR SURVIVAL ENDPOINTS WITH NON-PROPORTIONAL HAZARDS

By
JINGYU HUANG

A Dissertation submitted to the
School of Public Health
and the
School of Graduate Studies
Rutgers, The State University of New Jersey
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

Written under the direction of
Professors Yong Lin and Weichung Joe Shih
and Approved by

________________________
________________________
________________________
________________________
________________________

New Brunswick, New Jersey
October, 2021
ABSTRACT OF THE DISSERTATION

Current Topics on Treatment Comparison for Survival Endpoints with Non-Proportional Hazards

by Jingyu Huang

Dissertation Directors: Professors Yong Lin and Weichung Joe Shih

The proportional hazards (PH) assumption is the key assumption which may need to be examined in each survival analysis. It is the underlying assumption for the Cox proportional hazards and the log-rank test achieves its maximum efficiency when the PH assumption is satisfied. However, when the PH assumption is violated, we may not be able to provide a valid clinical interpretation for the hazard ratio of treatment effect, and the log-rank test may no longer achieve its desired level of power. Many alternative tests and measures have been developed for the non-proportional hazards situation. Four of them, Weighted log-rank test, Max-Combo Test (MX), Maximin Efficiency Robust Test (MERT) and The Restricted Mean Survival Time (RMST), are studied in this dissertation. Since previous papers only introduced MERT results by using uncensored case correlation matrix, MERT result with real censored correlation is discussed in this dissertation. The performance of the four non-proportional approaches, namely, Weighted log-rank test, MX, MERT, and RMST, is evaluated under the null and six typical proportional/non-proportional hazards conditions. The strength and weakness for each approach is explored in this dissertation. In addition, this proposal compares the test performance between Fleming-Harrington family $G_{0,0}, G_{1,0}, G_{0,1}, G_{1,1}$ and family $G_{0,0}, G_{2,0}, G_{0,2}, G_{2,2}$ for Weighted log-rank test, MX and MERT. The best non-proportional hazards test among those four tests and two families is
explored in this dissertation.

In clinical trials, two stage design is commonly used to dismiss ineffective treatment early at the end of the first stage, in order to avoid the high cost and long trial duration. Conditional power is the conditional probability of a significant result at the end of the trial given the data observed thus far. It is commonly used to evaluate the possibility of stopping the trial for futility. It also can be used to calculate the new sample size in the sample size re-estimation at the interim stage. The RMST is a robust measure of the survival time distribution. It provide a clinically interpretable summary and is widely used in the non-proportional hazards situation, since it does not rely on the PH assumption. This dissertation describes how to calculate the CP for the RMST endpoint at the interim stage for non-proportional hazards models. An approach for predicting survival curve at the final stage is developed in order to calculate the correlation in test statistics between the interim and final stages.
Acknowledgements

First and foremost, I would like to express my sincere gratitude to my advisors, Dr. Yong Lin and Dr. Weichung Joe Shih, whose continuous encouragement, guidance and support enabled me to finish this dissertation. Dr. Lin started to train me in his longitudinal class and Dr. Shih first brought me into the clinical trial design field in his class. They helped to make statistics easier for me with their patient teaching, sound advice, and great efforts to explain things succinctly. This thesis is the result of their enormous time and effort which were spent to help me in dealing with issues from detailed questions to huge obstacles. Without their guidance, patience, and support, it would not have been possible for me to accomplish the goal of this research.

It is difficult to overstate my deep gratitude to Dr. Shou-En Lu, who provided critical ideas that improved the validity of my work. Her insights encouraged me to think broadly and critically about my research. Furthermore I want to thank Dr. Lu’s guidance during statistics theory class which helped me lays a solid statistic theory foundation for my research.

I express my enormous gratitude to Dr. Sinae Kim, who provide valuable comments on my PhD research. I also would like to thank Dr. Kim’s guidance and support in the regression class and my PhD qualify exam.

Deep appreciation goes to Dr. Kaifeng Lu for agreeing to serve on my committee as the outside member and spent enormous time and efforts to help and guide me to accomplish my research.

Finally I would like to acknowledge and thank to all my school division for allowing me to conduct my research and provide all assistance requested.
Dedication

This thesis is dedicated to my parents whose words of encouragement and push for tenacity ring in my ears. Without their endless love and encouragement I would never have been able to complete my graduate studies. I dedicate this work to my wonderful son Vincent who came to this world during my doctoral program and has made me stronger, better, and more fulfilled than I could ever imagined. I also dedicate this dissertation to my uncle for leading me to the statistics field and spending the many hours of proofreading, and my aunt and their whole family who help and support me all the time.

I dedicate this dissertation to my friends, my teachers, and to all those who made this dissertation possible. Finally, to my caring, loving, and supportive wife, Yinzi. my deepest gratitude. Thank you for taking caring our son all the time and allowing me time away from it to research and write.
Table of Contents

Abstract ................................................................. ii
Acknowledgements ........................................................ iv
Dedication ................................................................. v
List of Tables .............................................................. x
List of Figures ............................................................... xi

1  Evaluation of Methods of Testing Treatment Effect for Non-proportional Hazards Models ................................................... 1

1. Introduction .............................................................. 2
   1.1. Challenge for Traditional Proportional Hazard Model ................. 2
   1.2. Available Solutions ................................................. 2
   1.3. Research Motivations ................................................. 3
   1.4. Research Objectives .................................................. 4
      1.4.1. Objective ....................................................... 4
      1.4.2. Specific aims .................................................. 4

2. Literature Review ....................................................... 6
   2.1. Weighted Log-rank Test (WLT) ..................................... 6
      2.1.1. Fleming Harrington (FH) Weighted Log-rank Test ............ 8
   2.2. Maximin Efficiency Robust Test (MERT) ............................ 11
      2.2.1. Application to Survival Analysis ............................... 13
   2.3. Max-Combo Test (MX) .............................................. 15
   2.4. Restricted Mean Survival Time (RMST) ............................ 16
2.4.1. Restricted standard deviation of survival time (RSDST) ....... 17

2.5. Estimation for Covariates with Non-proportional Hazards ........ 18

\[2.5.1. \text{The Estimation of Average Hazard Ratio (AHR) by Weighted Cox Regression (WCR)} \] ........ 19

\[2.5.2. \text{The Estimation of Covariates by RMST Regression} \] ........ 20

3. Maximin Efficiency Robust Test by Using Censored Case Correlation ... 22

\[3.1. \text{Notation} \] .................................................. 22

\[3.2. \text{Maximin Efficiency Robust Test by Using Real Correlation} \] ........ 23

\[3.2.1. \text{Difference Between Uncensored Correlation and Censored Correlation} \] 23

\[3.2.2. \text{Derive MERT Statistics by Using Real Censored Correlation} \] ....... 26

4. Evaluation of NPH Methods Performance when Survival Time Follows a Piecewise Exponential Distribution ................................. 28

\[4.1. \text{Method and Simulation Setup} \] .................................. 28

\[4.2. \text{Results} \] ......................................................... 31

\[4.2.1. \text{Performance of Tests} \] ...................................... 31

\[4.2.2. \text{Comparison of Performance} \] ................................ 31

5. Evaluation of NPH Methods Performance when Treatment Time Follows a Time-Dependent Exponential Distribution .................. 35

\[5.1. \text{Method and Simulation Setup} \] .................................. 35

\[5.2. \text{Results} \] ......................................................... 35

\[5.2.1. \text{Performance of Tests} \] ...................................... 36

\[5.2.2. \text{Comparison of Performance} \] ................................ 36

6. Evaluation of NPH Methods Performance by Using Real Samples ..... 39

\[6.1. \text{Example} \] ....................................................... 39

\[6.2. \text{Method} \] ......................................................... 39

\[6.2.1. \text{Performance of Tests} \] ...................................... 41

\[6.2.2. \text{Comparison of Performance} \] ................................ 42
7. Conclusion ................................................................. 43

II Two-stage in RMST with Conditional Power 44

8. Introduction ................................................................. 45
   8.1. Background ............................................................. 45
   8.2. Research Objectives ................................................... 46

9. Literature Review ........................................................ 47
   9.1. Two-Stage Design in RMST with Conditional Power (CP) ....... 47
      9.1.1. Notation ............................................................. 47
      9.1.2. Interim Stage with RMST ....................................... 48
      9.1.3. Final Stage with RMST .......................................... 49
      9.1.4. Conditional Power (CP) ......................................... 50
   9.2. Calculating Correlation in RMST Difference Between Interim and Final Analysis ............. 52
      9.2.1. Counting Process Approach ...................................... 52
      9.2.2. Martingale Integration Approach ............................... 53

10. Asymptotic Covariance of RMST Difference in Two-Stage Design ............... 56
    10.1. Notation ............................................................. 56
    10.2. Asymptotic Covariance by Martingale Integration Approach ................. 58
        10.2.1. Calculation of Asymptotic covariance by given $\Delta_F$ and RMST at final analysis for control .................. 66
        10.2.2. Calculation of Asymptotic Covariance by Given $\Delta_F$ and with Known Survival Distribution for Control Group .... 68
    10.3. Simulation Setup .................................................... 70
    10.4. Results ............................................................... 71

11. Conditional Power in Interim Analysis ........................................ 75
    11.1. Calculation of conditional power ..................................... 75
List of Tables

2.1. Correlation Matrix (Uncensored Case) . . . . . . . . . . . . . . . . . . . . . 14
3.1. Correlations . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 25
4.1. Survival Configuration for Each Model in Simulation . . . . . . . . . . . . . 28
4.2. Power Estimates for each Model . . . . . . . . . . . . . . . . . . . . . . . . 32
5.1. Survival Configuration for Each Model in Simulation . . . . . . . . . . . . . 35
5.2. Power Estimates for each Model . . . . . . . . . . . . . . . . . . . . . . . . 37
6.1. Test One-sided P-value for each Example . . . . . . . . . . . . . . . . . . . . 41
6.2. Regression Coefficients with Standard Errors for Two Examples . . . . . . 41
10.1. Survival Configuration for Each Model in Simulation . . . . . . . . . . . . . 70
10.2. Covariance by Models . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 73
11.1. Conditional Power for all Models at \( v_I = 1.3 \) . . . . . . . . . . . . . 77
11.2. Conditional Power for all Models by Interim Analysis cut Times . . . . . . 79
12.1. Asymptotic Variance for Model D by Different Turing Points . . . . . . . . 82
List of Figures

2.1. Difference between real time $t$ and relative time $v$, subjects B and D had the event of interest, subject E was administratively censored . . . . . . . . . . 7

2.2. Weight functions $W(t, v)$ in the $G^{o7}$ class of statistics . . . . . . . . . . . . 9

4.1. Simulation scenarios . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 29

5.1. Simulation scenarios . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 36

6.1. KM plots for Example 1 and Example 2 . . . . . . . . . . . . . . . . . . . . 40

10.1. One Example: Predicted Survival Curves at Interim and Final Stage by Models 72

10.2. Example: KM Curves at Interim and Final Stage . . . . . . . . . . . . . . . 74

12.1. Predicted Curve and True KM Curve at Final Stage in Model D . . . . . . . 83
Part I

Evaluation of Methods of Testing Treatment Effect for Non-proportional Hazards Models
Chapter 1

Introduction

1.1 Challenge for Traditional Proportional Hazard Model

Quantifying or measuring the difference between the survival curves of the randomized arms is usually set as the primary objective in a randomized clinical trial with a time-to-event endpoint. The Cox proportional hazards model is the most common test for survival analysis, which has been used for many years. The Proportional hazards (PH) assumption implies that the ratio of the two hazard functions is constant over time, and it is the underlying condition for the Cox proportional hazards model. The log-rank test is the most powerful non-parametric test for detecting alternatives and is the most commonly used testing procedure for PH model. When the PH assumption holds, we may easily estimate the survival curve difference between treatment groups by using the hazard ratio (HR) and test the PH alternative by using the log-rank test. However, when the PH assumption is violated, then it is difficult to interpret the clinical meaning for HR from the Cox PH model and the log-rank test may no longer provide the desired level of power.

1.2 Available Solutions

Many alternative measures or tests have been proposed to address this issue. To increase the power of the log-rank test for non-proportional hazards (NPH) model, the weighted log-rank test has been developed which uses a specific weight function to the log-rank test statistics for each target non-proportional hazards models. Different non-proportional hazards models require different weighted functions. Max-Combo Test (MX) and Maximin Efficiency Robust Test (MERT) are two general tests which are developed to fit all model conditions in one test. The restricted mean survival time (RMST), sometimes called the
restricted mean event time, is an alternative measure that can be reliably estimated and more efficient than the hazard ratio in a clinical trial, whereas the log-rank test may fail to detect the treatment difference and the hazard ratio becomes meaningless under the NPH model. Detail reviews of all these alternative measures and tests are provided in the literature review section.

1.3 Research Motivations

As one of the general test for the NPH model, MERT has barely been mentioned in the literature. It has been introduced to handle survival analysis in Gastwirth (1985). In the paper, $G^\tau$ as a weight function statistics, which was introduced by Harrington and Fleming, was used to generate the MERT statistics. The result indicates that the MERT had higher power than the log-rank test. The weight function which was used to generate the MERT statistics had been updated from $G^\tau$ to $G^{\rho,\tau}$ in Freidlin et al. (1999). The simulated power of MERT and MX have been evaluated under the family $G^{0.0}, G^{2.0}, G^{0.2}, G^{2.2}$ and uncensored correlation. The result showed MX had slightly better performance under early difference, middle difference, and late difference non-proportional models. Since the paper only discusses the result by using correlation matrix for uncensored case, questions about whether there is a difference in correlation matrix between censored and uncensored case and whether we can reach the same conclusion by using the real censored correlation remain.

Regarding to the weighted log-rank test, MERT, MX and RSMT, most literatures were focused on evaluating performance on comparing the performance between two of them, and no paper has included all of them in the comparison. Besides, for the NPH condition, most literatures only considered four typical situations: constant difference, early difference, late difference, and middle difference and ignored the crossing over situation. The question arises as among those four tests, which one will be the best general test that has the best power performance over all typical non-proportional conditions?

Weighted log-rank test, MERT and MX all used the test statistics $G^{\rho,\tau}$ to generate their test statistics. There are no other restrictions on the parameter values of $\tau$ and $\rho$ other than that they have to be nonnegative. We found that there are some papers used family
and some used family \( G^{0,0}, G^{1,0}, G^{0,1}, G^{1,1} \) during our literature review. Both families are able to present constant difference, early difference, late difference, and middle difference condition, respectively. The question arises as to what the differences are between those two families and which one has better power performance?

1.4 Research Objectives

1.4.1 Objective

In Part I of this document, the following objectives were set:

**Objective I:** To derive the MERT by using real censored correlation matrix

**Objective II:** To evaluate and compare the performance of NPH model methods including weighted log-rank test, MERT, MX, and RMST on testing treatment difference for all typical NPH conditions.

**Objective III:** To evaluate and compare the performance of NPH methods including weighted log-rank test, MERT, MX between family \( G^{0,0}, G^{1,0}, G^{0,1}, G^{1,1} \) and \( G^{0,0}, G^{2,0}, G^{0,2}, G^{2,2} \) for all typical NPH conditions.

1.4.2 Specific aims

The Specific aims for Part I of this document are described as follows:

1. To derive the MERT by using real censored correlation matrix.

2. To compare the difference in correlation matrix between uncensored and real censored one.

3. To find the MERT extreme pair under different maximum boundary of \( \tau \) and \( \rho \) in \( G^{0,\tau} \).

4. To compare the performance of weighted log-rank test, MERT, MX, and RMST on testing treatment difference for all typical NPH conditions.
5. To find the most powerful general test among weighted log-rank test, MERT, MX, and RMST for all typical NPH conditions.

6. To compare performance of weighted log-rank test, MERT, MX between family $G^{0,0}, G^{1,0}, G^{0,1}, G^{1,1}$ and $G^{0,0}, G^{2,0}, G^{0,2}, G^{2,2}$ for all typical NPH conditions.

7. To compare performance of RMST with different cut time strategies for each typical NPH conditions.

8. To compare performance of all tests under crossing curve condition.
Chapter 2

Literature Review

2.1 Weighted Log-rank Test (WLT)

Assume the two-sample general random censoring model with staggered entry. Observations are made on \( n_i \) individuals from sample \( i \) which \( i = 1 \) indicates the treatment group, otherwise \( i = 2 \) represents the control. Let \((W_{ij}, V_{ij}, C_{ij}), j = 1, \ldots, n_i, i = 1, 2,\) denote the real time of entry during the accrual period \([0, t_a]\), the real time from entry until failure, and the time from entry to censoring for subject \( j \) in treatment \( i \) respectively. It assume that \( W_{ij}, V_{ij}, C_{ij} \) are all nonnegative random variables.

At a given real study time \( t \), let

\[
X_{ij}(t) = \max(\min(V_{ij}, C_{ij}, t - W_{ij}), 0)
\]

and

\[
\delta_{ij}(t) = I\{V_{ij} \leq \min(C_{ij}, t - W_{ij})\}
\]

where \( I\{E\} \) is the indicator function of event \( E \). \( I\{E\} \) is 1 if the event \( E \) occurs and 0 otherwise.

Let \( v_{ij} \) be the relative times (time measured from the real time of entry) for subject \( j \) in treatment \( i \). We compared the difference between \( t \) and \( v \) in Figure 2.1.

By counting processes, \( N_{ij}(t, v) \) is defined as \( N_{ij}(t, v) = I\{X_{ij}(t) \leq v, \delta_{ij}(t) = 1\} \), with associated at-risk processes \( Y_{ij}(t, v) = I\{X_{ij}(t) \geq v\} \), where \( v \leq t \). We also let \( \bar{N}_i(t, v) = \sum_j N_{ij}(t, v), \bar{N}(t, v) = \sum_i \bar{N}_i(t, v) \), with similar definitions for \( \bar{Y}_i(t, v) \) and \( \bar{Y}(t, v) \). The hypothesis of interest is that the two survival distributions are equal. That is, \( H_0 : S_1(t, v) = S_2(t, v) \), where \( S_i(t, v) = Pr(X_{ij}(t) > v) \). The weighted log-rank
Figure 2.1: Difference between real time $t$ and relative time $v$, subjects B and D had the event of interest, subject E was administratively censored.

Statistics can be expressed in [Fleming and Harrington (1991)] as

$$W_{WLT}(t) = c \int_{v=0}^{v=1} W(t, v) \frac{Y_1(t, v) Y_2(t, v)}{\overline{Y}_1(t, v) + \overline{Y}_2(t, v)} \left\{ \frac{dN_1(t,v)}{\overline{Y}_1(t,v)} - \frac{dN_2(t,v)}{\overline{Y}_2(t,v)} \right\}, \quad (2.1)$$

where $c = \sqrt{\frac{1}{n_1(t)} + \frac{1}{n_2(t)}}$. $n_1(t)$ and $n_2(t)$ are the total number of subjects in treatment group $i = 1$ and $i = 2$ at a given study time $t$ respectively. $W(t, v)$ is the weighting function which is a nonnegative function of survival time. If the weighting function is equal to 1 at every time point, the weighted log-rank statistics will reduce to the log-rank statistics which can be represented as follows:

$$W_{LT}(t) = c \int_{v=0}^{v=1} \frac{Y_1(t, v) Y_2(t, v)}{\overline{Y}_1(t, v) + \overline{Y}_2(t, v)} \left\{ \frac{dN_1(t,v)}{\overline{Y}_1(t,v)} - \frac{dN_2(t,v)}{\overline{Y}_2(t,v)} \right\},$$

Under the statistical hypothesis that is being tested, the WLT statistic is approximately normally distributed with mean 0 and a variance that can be estimated as follows:

$$\hat{\sigma}^2_{WLT}(t) = c^2 \int_{v=0}^{v=1} W(t, v)^2 \frac{Y_1(t, v) Y_2(t, v)}{\overline{Y}_1(t, v) + \overline{Y}_2(t, v)} \times \left( 1 - \frac{dN_1(t,v)}{\overline{Y}_1(t,v)} + \frac{dN_2(t,v)}{\overline{Y}_2(t,v)} - 1 \right) \frac{dN_1(t,v)}{\overline{Y}_1(t,v)} + \frac{dN_2(t,v)}{\overline{Y}_2(t,v)}, \quad (2.2)$$

As we known, the log-rank test statistics is approximately normally distributed with mean 0 and a variance can be represented by using above equation 2.2 when the weighting
function equal to 1 at every time point as follows:

\[
\hat{\sigma}_{LT}^2(t) = c^2 \int_{v=0}^{v=t} \frac{Y_1(t,v)Y_2(t,v)}{\overline{Y}_1(t,v) + \overline{Y}_2(t,v)} \left( 1 - \frac{d\overline{N}_1(t,v) + d\overline{N}_2(t,v) - 1}{\overline{Y}_1(t,v) + \overline{Y}_2(t,v) - 1} \right) \frac{d\overline{N}_1(t,v) + d\overline{N}_2(t,v)}{\overline{Y}_1(t,v) + \overline{Y}_2(t,v)},
\]

The WLT at a given real study time \( t \) is defined as

\[
Z_{WLT}(t) = \frac{W_{WLT}(t)}{(\hat{\sigma}_{WLT}^2(t))^{0.5}},
\]

where \( Z_{WLT}(t) \approx N(0,1) \) under the null hypothesis \( H_0 : S_1(t,v) = S_2(t,v) \).

Similar test statistic to log-rank test as follow:

\[
Z_{LT}(t) = \frac{W_{LT}(t)}{(\hat{\sigma}_{LT}^2(t))^{0.5}},
\]

where \( Z_{LT}(t) \approx N(0,1) \) under the null hypothesis \( H_0 : S_1(t,v) = S_2(t,v) \).

### 2.1.1 Fleming Harrington (FH) Weighted Log-rank Test

Fleming and Harrington (1991) defined a class of the weighted log-rank statistics in which the weight function \( W(t,v) = \{\hat{S}(t,v-)^\rho(1-\hat{S}(t,v-))\}^\gamma \) for \( \rho \geq 0, \gamma \geq 0 \), where \( \hat{S}(t,v-) \) is the left-continuous version of the Kaplan and Meier (1958) estimator in the pooled sample, and called the \( G^{\rho,\gamma} \) statistics.

\[
G^{\rho,\gamma}(t) = c \int_{v=0}^{v=t} \{\hat{S}(t,v-)^\rho(1-\hat{S}(t,v-))\}^\gamma \frac{Y_1(t,v)Y_2(t,v)}{\overline{Y}_1(t,v) + \overline{Y}_2(t,v)} \left\{ \frac{d\overline{N}_1(t,v)}{\overline{Y}_1(t,v)} - \frac{d\overline{N}_2(t,v)}{\overline{Y}_2(t,v)} \right\}
\]

When \( \gamma = 0 \), this class reduces to the \( G^\rho \) class (Harrington and Fleming, 1982), and \( \rho = 0 \) and \( \rho = 1 \) correspond to the log-rank and Prentice-Wilcoxon statistics, respectively.

The common \( G^{\rho,\gamma} \) class of statistics are: \( G^{0,0}, G^{1,0}, G^{0,1}, G^{1,1}, G^{2,0}, G^{0,2}, \) and \( G^{2,2} \). They are divided into two families as \( (G^{0,0}, G^{1,0}, G^{0,1}, G^{1,1}) \) and \( (G^{0,0}, G^{2,0}, G^{0,2}, G^{2,2}) \). Either of them can cover a wide range of possible differences in the survival distributions. Specifically, those statistics are designed to detect constant difference, early difference, late
difference, and middle difference, respectively (below table).

<table>
<thead>
<tr>
<th>$(p, \gamma)$</th>
<th>$W(t, v)$</th>
<th>Distribution of weights</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(0, 0)$</td>
<td>$1$</td>
<td>evenly distributed (log rank)</td>
</tr>
<tr>
<td>$(1, 0)$</td>
<td>${\hat{S}(t, v^-)}$</td>
<td>Gehan-Wilcoxon</td>
</tr>
<tr>
<td>$(1, 0), (2, 0)$</td>
<td>${\hat{S}(t, v^-)}, {\hat{S}(t, v^-)}^2$</td>
<td>emphasize early difference</td>
</tr>
<tr>
<td>$(0, 1), (0, 2)$</td>
<td>${1 - \hat{S}(t, v^-)}, {1 - \hat{S}(t, v^-)}^2$</td>
<td>emphasize late difference</td>
</tr>
<tr>
<td>$(1, 1), (2, 2)$</td>
<td>${\hat{S}(t, v^-)}{1 - \hat{S}(t, v^-)}, {\hat{S}(t, v^-)}^2{1 - \hat{S}(t, v^-)}^2$</td>
<td>emphasize middle difference</td>
</tr>
</tbody>
</table>

Based on Figure 2.2, we can see that weight function value constantly equal to 1 for $(p, \gamma) = (0, 0)$ which indicate it provide evenly distributed weight. Both $(p, \gamma) = (1, 0)$ and $(p, \gamma) = (2, 0)$ have a monotone increase trend in weight function value indicate that they emphasize the early difference. On the contrary, a monotone decrease trend in weight function for both $(p, \gamma) = (0, 1)$ and $(p, \gamma) = (0, 2)$ indicate that they emphasize the late
difference. Both \((\rho, \gamma) = (1, 1)\) and \((\rho, \gamma) = (2, 2)\) have highest weight function in the middle which indicate they emphasize the middle difference. Regarding to the difference between \((1, 0)\) and \((2, 0)\); \((0, 1)\) and \((0, 2)\); \((1, 1)\) and \((2, 2)\), we will discuss them in Section 4.2.2.

The \(G^{\rho, \gamma}\) statistics in equation 2.4 can also be written as:

\[
G^{\rho, \gamma}(t) = c \int_{v=0}^{t} \left\{ \left[ \hat{S}(t,v-) \right]^{\rho} \left[ 1 - \hat{S}(t,v-) \right]^{\gamma} \left( \frac{Y_1(t,v) - \overline{Y_1}(t,v)}{\overline{Y_1}(t,v)} dN_1(t,v) + \frac{Y_2(t,v) - \overline{Y_2}(t,v)}{\overline{Y_2}(t,v)} dN_2(t,v) \right) \right\}. \tag{2.5}
\]

where \(c = \sqrt{\frac{1}{n_1(t)} + \frac{1}{n_2(t)}}\).

Let \(v_1 < \ldots < v_m\) be the \(m\) distinct uncensored times in pooled data at a real study time \(t\). Then the equation 2.5 is equivalent to:

\[
G^{\rho, \gamma}(t) = c \sum_{r=1}^{m} \left\{ \left[ \hat{S}(t,v_r-) \right]^{\rho} \left[ 1 - \hat{S}(t,v_r-) \right]^{\gamma} \left( d_{1,r}(t) - \frac{n_{1,r}(t)}{n_r(t)} d_r(t) \right) \right\} \tag{2.6}
\]

where \(c = \sqrt{\frac{1}{n_1(t)} + \frac{1}{n_2(t)}}\). At a given study time \(t\), \(d_r\) stands for the number of deaths at time \(v_r\) and \(n_r\) gives the number of individuals at risk at time \(v_r\), that is, the number of individuals alive just prior to time \(v_r\). Both \(d_{1,r}\) and \(n_{1,r}\) correspond to the same definitions but restricted to observations in the group \(i = 1\) which is the treatment group. \(n_1\) and \(n_2\) are total number count of subjects in treatment group \(i = 1\) and \(i = 2\) at a given real study time \(t\), respectively. The Kaplan-Meier estimator in the pooled sample as \(\hat{S}(t,v_m-) = \prod_{r=1}^{m-1} \left( 1 - \frac{d_r(t)}{n_r(t)} \right)\).

The asymptotic null distribution of \(G^{\rho, \gamma}\) can be derived through a martingale central limit theorem. Under the null hypothesis, \(G^{\rho, \gamma}\) is asymptotically normal with zero mean and variance \(\hat{\sigma}_G^{2, \rho, \gamma}\) can be written by equation 2.2 as follows:

\[
\hat{\sigma}_G^{2, \rho, \gamma}(t) = c^2 \int_{v=0}^{t} \left( \left[ \hat{S}(t,v-) \right]^{\rho} \left[ 1 - \hat{S}(t,v-) \right]^{\gamma} \right)^2 \left( \frac{Y_1(t,v) - \overline{Y_1}(t,v)}{\overline{Y_1}(t,v)} dN_1(t,v) + \frac{Y_2(t,v) - \overline{Y_2}(t,v)}{\overline{Y_2}(t,v)} dN_2(t,v) \right) \times \left( 1 - \frac{dN_1(t,v) + dN_2(t,v) - 1}{\overline{Y_1}(t,v) + \overline{Y_2}(t,v) - 1} \right) \left( \frac{Y_1(t,v) + dN_1(t,v)}{\overline{Y_1}(t,v) + \overline{Y_2}(t,v)} \right) dN_1(t,v) + \frac{Y_2(t,v) - \overline{Y_2}(t,v)}{\overline{Y_2}(t,v)} dN_2(t,v). \tag{2.7}
\]

Similar to \(G^{\rho, \gamma}\) statistics, the \(\hat{\sigma}_G^{2, \rho, \gamma}\) in equation 2.7 is equivalent to:
\[ \sigma_{G,\gamma}^2(t) = c^2 \sum_{r=1}^{m} \left\{ \left( \hat{S}(t, v_r - \cdot) \right)^{\rho}[1 - \hat{S}(t, v_r - \cdot)]^{\gamma} \right\}^2 \frac{n_{1,r}(t) (n_r(t) - n_{1,r}(t)) (n_r(t) - d_r(t)) d_r(t)}{n_r^2(t) n_r(t) - 1} \]

(2.8)

Thus, by the WLT at a given real study time \( t \) in equation 2.3, we can write the FH test as follows:

\[
Z_{G,\rho,\gamma}(t) = \frac{W_{G,\rho,\gamma}(t)}{\left( \hat{\sigma}_{G,\rho,\gamma}^2(t) \right)^{0.5}} = \frac{c \sum_{r=1}^{m} \left\{ \left[ \hat{S}(t, v_r - \cdot) \right]^{\rho}[1 - \hat{S}(t, v_r - \cdot)]^{\gamma} \right\} \left( d_{1,r}(t) - \frac{n_{1,r}(t)}{n_r(t)} d_r(t) \right)}{\left\{ c^2 \sum_{r=1}^{m} \left\{ \left[ \hat{S}(t, v_r - \cdot) \right]^{\rho}[1 - \hat{S}(t, v_r - \cdot)]^{\gamma} \right\}^2 \frac{n_{1,r}(t) (n_r(t) - n_{1,r}(t)) (n_r(t) - d_r(t)) d_r(t)}{n_r^2(t) n_r(t) - 1} \right\}^{0.5}}
\]

(2.9)

where \( Z_{G,\rho,\gamma}(t) \approx N(0, 1) \) under the null hypothesis \( H_0 : S_1(t, v) = S_2(t, v). \)

### 2.2 Maximin Efficiency Robust Test (MERT)

The precise distribution underlying the data is often unknown, however we can specify a family of plausible alternatives: \( \Psi = \{f_i; i = 1, \ldots, I\} \). When model \( f_i \) is true, \( e(S, i) \) denote the asymptotic relative efficiency (ARE) between any test statistic \( S \) and the optimal test \( T_i \). When a model in \( \Psi \) is true, the lowest ARE \( S \) exists and can be denoted as \( e(S, \Psi) = \inf_{1 \leq i \leq I} \{e(S, i)\} \) (Freidlin et al., 1999). The maximin efficiency robust test (MERT) has the highest minimum efficiency, \( e(S, \Psi) \), called its maximin ARE, across all models in \( \Psi \) and possesses greater efficiency robustness than any of the optimum statistics \( \{T_i, i \in I\} \) by Gastwirth (1985).

Let \( \{T_i\} \) be the most powerful test statistics for model \( f_i, i \in I \). Under the null hypothesis, we assume that each \( T_i \) is asymptotically normally distributed. the standardized versions \( \{Z_i\} \) is derived as \( Z_i = [T_i - E(T_i)]/\sqrt{\text{Var}(T_i)} \), where \( E(T_i) \) and \( \text{Var}(T_i) \) are the mean and variance of \( T_i \). \( \{Z_i\} \) converges in law to \( N(0, 1) \) (Gastwirth, 1985). MERT exists when following three conditions hold: (1) exist asymptotically most powerful tests, \( \{T_i\}, i = 1, \ldots, I \), for the respective members of the alternative models \( f_i \) in family \( \Psi \); (2)
under the null hypothesis, standardized test statistics \( \{ Z_i \} \), of the \( \{ T_i \} \) are asymptotically jointly multivariate normal with the correlation matrix \( \{ \rho_{ij} \} \), where all \( \rho_{ij} > 0 \); and (3) the Pitman asymptotic relative efficiency (ARE) between test \( \{ Z_i \} \) and the optimal test \( \{ Z_j \} \) is \( \rho_{ij}^2 = \langle Z_i, Z_j \rangle^2 \). These three conditions are satisfied in a wide variety of survival data applications (Van Eeden, 1964; Gross, 1981).

Then MERT satisfies

\[
e (MERT, \Psi) = \sup_{Z \in \Gamma} \left[ \inf_{1 \leq i \leq I} \{ e (Z, i) \} \right]
\]

where \( \Gamma \) is the set of all consistent asymptotically normal tests for the problem. Gastwirth (1966) showed that, when the minimum correlation of the optimal tests \( Z_i \), \( \rho^* = \min(\rho_{ij}) \) is \( > 0 \), the MERT exists, is unique, and is a linear combination of the \( \{ Z_i \} \). Although we can use quadratic programming methods to obtain the MERT (Rosen, 1960), many studies still prefer to get the solution by using a linear combination of the \( Z_i \) that are optimum for two or three of the most "different" alternative models by Gastwirth (1985).

Let the tests \( Z_1 \) and \( Z_2 \), be the statistics for the most disparate alternatives, and called them extreme pair, which satisfy

\[
\rho_{12} = \langle Z_1, Z_2 \rangle = \inf_{i,j} \rho_{ij} > 0.
\]

Define

\[
R_{12} = [2 (1 + \rho_{12})]^{-\frac{1}{2}} (Z_1 + Z_2),
\]

(2.10)

If \( R_{12} \) is the MERT, then it has maximin efficiency \( (1 + \rho_{12})/2 \). A necessary and sufficient condition for the statistic \( R_{12} \) to be the MERT for the entire family \( \{ T_i \} \) Gastwirth (1970) is

\[
\rho_{1i} + \rho_{2i} \geq 1 + \rho_{12} \quad \text{for all } i \in I.
\]

(2.11)

When we fail to hold the condition in (2.11) the MERT for the extreme pair is not the
MERT for the entire family \( \{T_i\} \). There always is a test statistic \( T_3 \) which has less maximin efficiency than \( R_{12} \) from minimum correlation. The statistic of the form

\[
\sum_{i=1}^{3} a_i Z_i, a_i > 0, \tag{2.12}
\]

which \( \sum_{i=1}^{3} a_i Z_i \) has equal correlation with each of the three \( Z_i \) and its variance equals to 1. \( T_3 \) is the next candidate for MERT for the entire family. The MERT statistic is a linear combination of the standardized statistics, \( Z_i \), not the original \( \{T_i\} \). \( T_3 \) is only equally correlated with the extreme members, but not with all \( \{Z_i\} \) (Gastwirth, 1985).

### 2.2.1 Application to Survival Analysis

Applying the MERT to survival analysis, most common method for NPH is WLT statistics which has been introduced in Section 2.1. Similar to Section 2.1, we assume the two-sample general random censoring model with intermittent entry. Observations are made on \( n_i \) individuals from sample \( i \) which \( i = 1 \) indicate the treatment group, otherwise \( i = 2 \) represents the control. Let \((W_{ij}, V_{ij}, C_{ij}), j = 1, \ldots, n_i, i = 1, 2\), denote the real time of entry during the accrual period \([0, t_a]\), the real time from entry until failure, and the time from entry to censoring for subject \( j \) in treatment \( i \) respectively. It assume that \( W_{ij}, V_{ij}, C_{ij} \) are all nonnegative random variables.

At a given real study time \( t \), let

\[
X_{ij}(t) = \max(\min(V_{ij}, C_{ij}, t - W_{ij}), 0)
\]

and

\[
\delta_{ij}(t) = I\{V_{ij} \leq \min(C_{ij}, t - W_{ij})\}
\]

where \( I\{E\} \) is the indicator function of event \( E \). \( I\{E\} \) is 1 if the event \( E \) occurs and 0 otherwise.

Let \( v_{ij} \) be the relative times (time measured from the real time of entry) for subject \( j \) in treatment \( i \). By counting processes, \( N_{ij}(t,v) \) are defined as \( N_{ij}(t,v) = I\{X_{ij}(t) \leq v, \delta_{ij}(t) = 1\} \), with associated at-risk processes \( Y_{ij}(t,v) = I\{X_{ij}(t) \geq v\} \), where \( v \leq t \).
We also let $\overline{N}_i(t, v) = \sum_j N_{ij}(t, v)$, $\overline{N}(t, v) = \sum_i \overline{N}_i(t, v)$, with similar definitions for $\overline{Y}_i(t, v)$ and $\overline{Y}(t, v)$.

Fleming and Harrington (1991) defined a class of the weighted log-rank statistics in which the weight function $W(t, v) = \{\hat{S}(t, v)\}^\rho \{1 - \hat{S}(t, v)\}^\gamma$ for $\rho \geq 0, \gamma \geq 0$, where $\hat{S}(t, v)$ is the left-continuous version of the Kaplan and Meier (1958) estimator in the pooled sample, and called these $G^{\rho, \gamma}$ statistics.

\[
G^{\rho, \gamma}(t) = c \int_{v=0}^{v=t} \{\hat{S}(t, v)\}^\rho \{1 - \hat{S}(t, v)\}^\gamma \frac{\overline{Y}_1(t, v)\overline{Y}_2(t, v)}{\overline{Y}_1(t, v) + \overline{Y}_2(t, v)} \left\{ \frac{dN_1(t, v)}{\overline{Y}_1(t, v)} - \frac{dN_2(t, v)}{\overline{Y}_2(t, v)} \right\}
\]

In Freidlin et al. (1999), family of four members of FH weight functions $G^{0, 0}$, $G^{2, 0}$, $G^{0, 2}$, and $G^{2, 2}$ were used to generate the MERT. The four members of FH weight function cover a wide range of possible differences in the survival distributions. They are designed to detect constant difference, early difference, late difference, and middle difference, respectively. The correlation matrix (uncensored case) for the family is showed in Table 2.1.

Table 2.1: Correlation Matrix (Uncensored Case)

<table>
<thead>
<tr>
<th></th>
<th>$G^{0, 0}$</th>
<th>$G^{2, 0}$</th>
<th>$G^{0, 2}$</th>
<th>$G^{2, 2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$G^{0, 0}$</td>
<td>1</td>
<td>0.745</td>
<td>0.745</td>
<td>0.837</td>
</tr>
<tr>
<td>$G^{2, 0}$</td>
<td>1</td>
<td>0.167</td>
<td>0.535</td>
<td></td>
</tr>
<tr>
<td>$G^{0, 2}$</td>
<td>1</td>
<td></td>
<td>0.535</td>
<td></td>
</tr>
<tr>
<td>$G^{2, 2}$</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

While correlation is known, the extreme pair can be identified by selecting the lowest correlation among each pair in four members family group $G^{0, 0}$, $G^{2, 0}$, $G^{0, 2}$, and $G^{2, 2}$. Based on the Table 2.1, the correlation for the extreme pair has been found as $\rho_{12} = \rho_{G^{0, 2}, G^{2, 0}} = 0.167$. If $\rho_{12}$ meet the condition in equation 2.11, $R_{12}$ will be the MERT for the entire family and it has maximin efficiency $(1 + \rho_{12})/2$. However, $\rho_{G^{2, 0}, G^{2, 2}} + \rho_{G^{0, 2}, G^{2, 2}} = 0.535 + 0.535 = 1.07 < 1 + \rho_{G^{0, 2}, G^{2, 0}} = 1 + 0.167 = 1.167$ which fail to meet the condition at equation 2.11. Therefore, the MERT for the extreme pair is not the MERT for the entire family, we need calculate $T_3$ which is the next candidate for MERT for the entire family. For family group $G^{0, 0}$, $G^{2, 0}$, $G^{0, 2}$, and $G^{2, 2}$, we have $\alpha_i = (0, 0.5949, 0.5949, 0.1248)$ based on then uncensored case correlation matrix. The MERT statistic is approximately normally distributed to $N(0, 1)$ under the null hypothesis.
2.3 Max-Combo Test (MX)

In survival analysis, the Max-Combo test (MX), also named as Maximum of the Standardized Optimum Statistics, is another robust test for NPH [Tarone 1981]. It’s test statistics is the maximum of the standardized optimum statistics \( MX = \max_{1 \leq i \leq l} (Z_i) \) [Fleming and Harrington 1991]. Under \( H_0 : S_1(t) = S_2(t) \), the joint distribution of \((W_{K_1}, W_{K_2}, W_{K_3}, W_{K_4}) = (G^{0,0}, G^{1,0}, G^{0,1}, G^{1,1}) \) or \((W_{K_1}, W_{K_2}, W_{K_3}, W_{K_4}) = (G^{0,0}, G^{2,0}, G^{0,2}, G^{2,2}) \) is asymptotically multivariate normally distributed with mean zero and the variance-covariance given in of Theorem 7.5.1 of Fleming and Harrington [1991].

The variance-covariance between each two test statistics \( G^{ρ_1, γ_1} \) and \( G^{ρ_2, γ_2} \) can be represented as follows:

\[
\sigma_{G^{ρ_1, γ_1}, G^{ρ_2, γ_2}}(t) = c^2 \int_{v=0}^{t} W_{ρ_1, γ_1}(t,v) W_{ρ_2, γ_2}(t,v) \left( \frac{\bar{Y}_1(t,v)\bar{Y}_2(t,v)}{\bar{Y}_1(t,v) + \bar{Y}_2(t,v)} \right) \times \left( 1 - \frac{d\bar{N}_1(t,v) + d\bar{N}_2(t,v) - 1}{\bar{Y}_1(t,v) + \bar{Y}_2(t,v) - 1} \right) \frac{d\bar{N}_1(t,v) + d\bar{N}_2(t,v)}{\bar{Y}_1(t,v) + \bar{Y}_2(t,v)^2},
\]

Here by counting processes, \( N_{ij}(t,v) \) are defined as \( N_{ij}(t,v) = I\{X_{ij}(t) \leq v, δ_{ij}(t) = 1\} \), with associated at-risk processes \( Y_{ij}(t,v) = I\{X_{ij}(t) \geq v\} \), where \( v \leq t \). We also let \( \bar{N}_i(t,v) = \sum_j N_{ij}(t,v), \bar{N}(t,v) = \sum_i \bar{N}_i(t,v) \), with similar definitions for \( \bar{Y}_i(t,v) \) and \( \bar{Y}(t,v) \). \( W_{ρ_i, γ_i}(t,v) \) is a class of the weighted log-rank statistics in which the weight function \( W_{ρ_i, γ_i}(t,v) = \{\hat{S}(t,v-)^{ρ_i} (1-\hat{S}(t,v-))^{γ_i}\}^{γ_i} \) for \( ρ_i \geq 0, γ_i \geq 0 \), where \( \hat{S}(t,v-) \) is the left-continuous version of the Kaplan and Meier [1958] estimator in the pooled sample in a given study time \( t \).

As already known, MX test statistics is the maximum of the standardized statistics \( \{Z_I^* = \frac{W_{K_i}}{\sqrt{Var(W_{K_i})}}, l = 1, \ldots, 4\} \). Under the null hypothesis, \( \max\{Z_1^*, Z_2^*, Z_3^*, Z_4^*\} \) is asymptotically distributed as \( \max\{Z_1, Z_2, Z_3, Z_4\} \) as \( MN(0, \{ρ_{ij}\}) \) is a 4-variate Gaussian vector with zero mean and the covariance matrix equals to correlation matrix [Lee 1996]. Correlation matrix calculation is discussed in section 3.

Once the correlation matrix is known, we can use a program to calculate the multivariate normal probabilities \( P\{\max\{Z_1, Z_2, Z_3, Z_4\} \leq c\} \) for any \( c \) (Schervish, 1984). Then, the critical value \( c(a) \) can be found by \( P\{\max\{Z_1, Z_2, Z_3, Z_4\} \geq c(a)\} = a \). Suppose we compute a
maximum of the standardized statistics as \( \{Z_l^*, l = 1, \ldots, 4\} \) for one-sided tests, the null hypothesis that two survival distributions are equal is rejected if \( \max\{Z_1^*, Z_2^*, Z_3^*, Z_4^*\} \geq c(a) \). 

\[
\max\{|Z_1^*|, |Z_2^*|, |Z_3^*|, |Z_4^*|\}
\]
will be used instead in a similar manner for the two-sided tests (Lee 1996).

### 2.4 Restricted Mean Survival Time (RMST)

Restricted mean survival time (RMST) is a alternative measure which is proposed for NPH. Keeping the same setting with the above sections, we assume the two sample general random censoring model with intermittent entry. Observations are made on \( n_i \) individuals from sample \( i \) which \( i = 1 \) indicate the treatment group, otherwise \( i = 2 \) represents the control. Let \( (W_{ij}, V_{ij}, C_{ij}) \), \( j = 1, \ldots, n_i, i = 1, 2 \), denote the real time of entry during the accrual period \([0, t_a]\), the real time from entry until failure, and the time from entry to censoring for subject \( j \) in treatment \( i \) respectively. It is assuming that \( W_{ij}, V_{ij}, C_{ij} \) are all nonnegative random variables.

At a given real study time \( t \), let

\[
X_{ij}(t) = \max(\min(V_{ij}, C_{ij}, t - W_{ij}), 0)
\]

and

\[
\delta_{ij}(t) = I\{V_{ij} \leq \min(C_{ij}, t - W_{ij})\}
\]

where \( I\{E\} \) is the indicator function of event \( E \). \( I\{E\} \) is 1 if the event \( E \) occurs and 0 otherwise.

The restricted mean survival time \( \mu(B(t, v^*)) \) is the mean of the survival time \( B(t, v^*) = \min(X(t), v^*) \) limited to some horizon \( v^* > 0 \). It can be evaluated as the area under the Kaplan-Meier curve \( S(t, v) \) for \( v \) from 0 up to \( v^* \) (Royston and Parmar 2013). When \( B \) is the time to death, we may think of \( \mu(B) \) as the “\( v^* \) year life expectancy”. (Royston and Parmar 2011).
We have [Irwin 1949]:

\[
\mu(B(t,v^*)) = E(B(t,v^*)) = E[\min(X(t),v^*)] \\
= E(X(t)|X(t) \leq v^*) P(X(t) \leq v^*) + v^* P(X(t) > v^*) \\
= \int_0^{v^*} vf(t,v) dv + v^* S(t,v^*) \\
= \int_0^{v^*} S(t,v) dv
\]

Based on above equation we can see that the restricted mean survival time is a function of \(v^*\). In addition, \(\mu(B(t,v^*))\) increases monotonically with \(v^*\), and for large \(v^*\), it tends to the unrestricted mean survival time, \(\mu\).

A natural estimator for \(\mu\) is

\[
\hat{\mu}(B(t,v^*)) = \int_{v=0}^{v=v^*} \hat{S}(t,v) dv
\]

(2.13)

where \(\hat{S}(t,v)\) is the KM estimator for the survival function of \(X\) at given study time \(t\). Let \(v_1 < \ldots < v_m\) be the \(m\) distinct uncensored times in pooled data at a given study time \(t\). And in a given study time \(t\), \(d_r\) stands for the number of deaths at time \(v_r\) and \(n_r\) gives the number of individuals at risk at time \(v_r\), that is, the number of individuals alive just prior to time \(V_r\). The KM estimator in the pooled sample as \(\hat{S}(t,v_m) = \prod_{r=1}^{v=m} \left(1 - \frac{d_r(t)}{n_r(t)}\right)\).

### 2.4.1 Restricted standard deviation of survival time (RSDST)

To compute the variance, \(var(B(t,v^*))\), of the restricted survival time \(B(t,v^*)\), we may need to know \(E(B(t,v^*)^2)\) first:

\[
E\left(B(t,v^*)^2\right) = E\left(X(t)^2|X(t) \leq v^*\right) P(X(t) \leq v^*) + v^{*2} P(X(t) > v^*) \\
= \int_0^{v^*} v^2 f(t,v) dv + v^{*2} S(t,v^*) \\
= 2 \int_0^{v^*} v S(t,v) dv
\]

(2.14)
Then we have:

$$\text{Var}(B(t,v^*)) = E\left(B(t,v^*)^2\right) - [E(B(t,v^*))]^2 = 2 \int_0^{v^*} vS(t,v)\,dv - \left[\int_0^{v^*} S(t,v)\,dv\right]^2$$

(2.15)

The restricted standard deviation (RSDST) is $\sqrt{\text{Var}(B(t,v^*))}$. The variance of RMST can be estimated by $n^{-1}\sqrt{\text{Var}(B(t,v^*))}$

Through a martingale central limit theorem, the variance term can be estimated as below (Klein and Moeschberger, 2005)

$$SE(\hat{\mu}(B(t,v^*)))^2 = \text{Var}(\hat{\mu}(B(t,v^*))) = \sum_{r=1}^{m} \left[\int_{v_r}^{v^*} \hat{S}(t,v)\,dv\right]^2 \frac{d_r(t)}{n_r(t) (n_r(t) - d_r(t))}$$

(2.16)

where $d_r$ and $n_r$ are the number of events and number of patients at risk at $v_r$ in a given study time $t$, respectively. Then the test statistics can be written as:

$$\text{Test statistic} Z = \frac{\hat{\mu}_1(B(t,v^*)) - \hat{\mu}_2(B(t,v^*))}{\sqrt{SE(\hat{\mu}_1(B(t,v^*)))^2 + SE(\hat{\mu}_2(B(t,v^*)))^2}}$$

### 2.5 Estimation for Covariates with Non-proportional Hazards

The Cox PH regression is the most popular method for survival analysis with censored data. We have already known that the ratio of the two hazard functions is constant over time under the PH assumption, and it is the basis underlying condition for the Cox PH regression. In real data, the PH assumption is often violated when the effect of at least one of the prognostic factors in the Cox regression model changes overtime. As a consequence, we always under- or overestimate the prognostic factor by using the average HR (Michael et al., 2009). There are several alternative regression models to more appropriately cope with NPH, two of most popular ones which are the weighted cox regression and RMST regression. Both methods are reviewed in this section.
2.5.1 The Estimation of Average Hazard Ratio (AHR) by Weighted Cox Regression (WCR)

Let the sample size equals to \( n \), we observe \( m \) distinct and uncensored survival times \( v_j \) \((1 \leq j \leq m)\) among the \( n \) survival times \( v_i \) \((1 \leq i \leq n)\). Each individual has a covariate vector \( x_i = (x_{i1}, \ldots, x_{ir}, \ldots, x_{ik}) \) and \( \delta_i \) is a censoring indicator as: 1 for censored, 0 for dead. The risk set is denoted as \( R_j \) which is a set of individuals alive and uncensored prior to \( v_j \). \( \beta \) is a vector of \( k \) regression parameters estimates. We have the log partial likelihood for Cox’s model \((\text{Cox}, 1972)\) is defined as

\[
\log L(\beta) = \sum_{j=1}^{m} \left[ x_j \beta - \log \left( \sum_{h \in R_j} \exp(x_h \beta) \right) \right] = \sum_{j=1}^{m} l_j
\]

where \( l_j \) is the contribution to the log likelihood at failure time \( t_j \). Then we have the first derivatives of \( \log L(\beta) \),

\[
\frac{\partial \log L(\beta)}{\partial \beta_r} = \sum_{j=1}^{m} \left[ x_{jr} - \frac{\sum_{h \in R_j} x_{hr} \exp(x_h \beta)}{\sum_{h \in R_j} \exp(x_h \beta)} \right] = \sum_{j=1}^{m} \frac{\partial l_j}{\partial \beta_r}, 1 \leq r \leq k
\]

when setting equation to zero we can obtain estimates of \( \beta \) by iteration usually with the Newton-Raphson technique.

We can obtain the elements of the corresponding information matrix \( I \) as:

\[
I_{rs} = -\frac{\partial^2 \log L(\beta)}{\partial \beta_r \partial \beta_s} = \sum_{j=1}^{m} -\frac{\partial^2 l_j}{\partial \beta_r \partial \beta_s}
\]

\[
= \sum_{j=1}^{m} \left\{ \frac{\sum_{h \in R_j} x_{hr} x_{hs} \exp(x_h \beta)}{\sum_{h \in R_j} \exp(x_h \beta)} - \frac{\left[ \sum_{h \in R_j} x_{hr} \exp(x_h \beta) \right] \left[ \sum_{h \in R_j} x_{hs} \exp(x_h \beta) \right]}{\left[ \sum_{h \in R_j} \exp(x_h \beta) \right]^2} \right\}, \quad (1 \leq r, s \leq k)
\]

by setting the weighted score equation to zero, we can obtain the weighted maximum likelihood estimates \( \hat{\beta}_r \) of regression parameters \( \beta_r \), \( 1 \leq r \leq k \):

\[
\sum_{j=1}^{m} w(t_j) \frac{\partial l_j}{\partial \beta_r} = 0
\]
For censored samples, \( w(t) = \hat{S}(t) \hat{G}(t)^{-1} \) is set as the weighting function for a WCR. \( \hat{S}(t) \) is the KM estimator in the pooled sample. \( \hat{G}(t) \) denotes the KM estimator of the censoring which is estimated like \( \hat{S}(t) \) but with reversed censoring indicator \( \delta \) (Michael et al., 2009).

Three approaches are considered to define a covariance matrix for WCR estimates: (i) is via a sandwich estimate used by Lin (1991) and Sasieni (1993), (ii) is based on robust covariance matrices as introduced by Lin and Wei (1989), and (iii) is based on the jackknife by Michael et al. (2009).

### 2.5.2 The Estimation of Covariates by RMST Regression

Let \( v^* \) be a prespecified time point of interest. Assume \( X_{ij}(t) \) is the time-to-event variable for the \( j \)th subject at \( i \)th treatment group at study time \( t \). The subject-specific RMST at \( v^* \) is defined by \( RMST_{ij}(v^*) = E[min(X_{ij}(t), v^*)] \) and can be modeled via a generalized linear model as

\[
g[RMST_{ij}(v^*)] = x'_{ij} \beta
\]

Under the natural logarithm link which \( g(\cdot) = \log(\cdot) \), the model is as

\[
\log[RMST_{ij}(v^*)] = x'_{ij} \beta
\]

We can interpret the parameter estimates from the fitted log-linear regression as log ratios of the RMST.

Under the identity or linear link, the model is as follows

\[
RMST_{ij}(v^*) = x'_{ij} \beta
\]

We can interpret the parameter estimates from the fitted linear regression as ratios of the RMST.

The PROC RMSTREG in SAS program can be directly used to produce the estimates and Wald confidence intervals for the parameters. An option of pseudo-value regression
method is provide in PROC RMSTREG in fitting generalized linear models to time-to-event data.
Chapter 3

Maximin Efficiency Robust Test by Using Censored Case Correlation

The uncensored case correlation in any two of four members of FH weight functions $G^{0,0}$, $G^{2,0}$, $G^{0,2}$, and $G^{2,2}$ has been used to obtain MERT in section 2.2. In this chapter, both uncensored and censored case correlation matrix will be derived for both $G^{0,0}$, $G^{1,0}$, $G^{0,1}$, and $G^{1,1}$ and $G^{0,0}$, $G^{2,0}$, $G^{0,2}$, and $G^{2,2}$. Their difference in correlation between uncensored case and censored case will be discussed by simulation for all typical non-proportional conditions.

3.1 Notation

Similar to Section 2.1, we assume the two-sample general random censoring model with intermittent entry. Observations are made on $n_i$ individuals from sample $i$ which $i = 1$ indicate the treatment group, otherwise $i = 2$ represents the control. Let $(W_{ij}, V_{ij}, C_{ij})$, $j = 1, \ldots, n_i, i = 1, 2$, denote the real time of entry during the accrual period $[0, t_a]$, the real time from entry until failure, and the time from entry to censor for subject $j$ in treatment $i$ respectively. It is assume that $W_{ij}, V_{ij}, C_{ij}$ are all nonnegative random variables.

At a given real study time $t$, let

$$X_{ij}(t) = \max(\min(V_{ij}, C_{ij}, t - W_{ij}), 0)$$

and

$$\delta_{ij}(t) = I\{V_{ij} \leq \min(C_{ij}, t - W_{ij})\}$$

where $I\{E\}$ is the indicator function of event $E$. $I\{E\}$ is 1 if the event $E$ occurs and 0 otherwise.
Let \( v_{ij} \) be the relative times (time measured from the real time of entry) for subject \( j \) in treatment \( i \). By counting processes, \( N_{ij}(t) \) are defined as \( N_{ij}(t, v) = I\{X_{ij}(t) \leq v, \delta_{ij}(t) = 1\} \), with associated at-risk processes \( Y_{ij}(t, v) = I\{X_{ij}(t) \geq v\} \), where \( v \leq t \). We also let \( \tilde{N}_i(t, v) = \sum_j N_{ij}(t, v) \), \( \tilde{N}(t, v) = \sum_i \tilde{N}_i(t, v) \), with similar definitions for \( \tilde{Y}_i(t, v) \) and \( \tilde{Y}(t, v) \).

Fleming and Harrington (1991) defined a class of the weighted log-rank statistics in which the weight function \( W(t, v) = \{\hat{S}(t, v-)\}^\rho \{1 - \hat{S}(t, v-)\}^\gamma \) for \( \rho \geq 0, \gamma \geq 0 \), where \( \hat{S}(t, v-) \) is the left-continuous version of the Kaplan and Meier (1958) estimator in the pooled sample, and called these \( G^{\rho, \gamma} \) statistics.

\[
G^{\rho, \gamma}(t) = c \int_{v=0}^{t} \{\hat{S}(t, v-)\}^\rho \{1 - \hat{S}(t, v-)\}^\gamma \frac{\tilde{Y}_1(t, v)\tilde{Y}_2(t, v)}{\tilde{Y}_1(t, v) + \tilde{Y}_2(t, v)} \left\{ \frac{dN_1(t, v)}{\tilde{Y}_1(t, v)} - \frac{dN_2(t, v)}{\tilde{Y}_2(t, v)} \right\}
\]

### 3.2 Maximin Efficiency Robust Test by Using Real Correlation

#### 3.2.1 Difference Between Uncensored Correlation and Censored Correlation

For MERT, there are two groups of four members of FH weight functions which are commonly used. One is \( G^{0, 0}, G^{1, 0}, G^{0, 1}, \) and \( G^{1, 1} \); the other one is \( G^{0, 2}, G^{2, 0}, G^{0, 2}, \) and \( G^{2, 2} \). Both groups’ four cases cover a wide range of possible differences in the survival distributions. The correlation between \( G^{\rho_1, \gamma_1}(t) \) and \( G^{\rho_2, \gamma_2}(t) \) can be represented (Fleming and Harrington, 1991) as

\[
\text{CORR} (G^{\rho_1, \gamma_1}(t), G^{\rho_2, \gamma_2}(t)) = \rho_{G^{\rho_1, \gamma_1}(t), G^{\rho_2, \gamma_2}(t)} = \frac{\int_{v=0}^{t} W_{\rho_1, \gamma_1}(t, v) W_{\rho_2, \gamma_2}(t, v) \frac{Y_1(t, v)Y_2(t, v)}{Y_1(t, v) + Y_2(t, v)} \left( 1 - \frac{dN_1(t, v)}{Y_1(t, v)} + \frac{dN_2(t, v)}{Y_2(t, v)} - 1 \right) \frac{dN_1(t, v)}{Y_1(t, v) + Y_2(t, v)} - \frac{dN_2(t, v)}{Y_1(t, v) + Y_2(t, v)} \right) \times \frac{1}{\sqrt{\int_{v=0}^{t} W_{\rho_1, \gamma_1}(t, v) Y_1(t, v) \frac{Y_1(t, v)Y_2(t, v)}{Y_1(t, v) + Y_2(t, v)} \left( 1 - \frac{dN_1(t, v)}{Y_1(t, v)} + \frac{dN_2(t, v)}{Y_2(t, v)} - 1 \right) \frac{dN_1(t, v)}{Y_1(t, v) + Y_2(t, v)} - \frac{dN_2(t, v)}{Y_1(t, v) + Y_2(t, v)}}}
\]

where

\[
W_{\rho_1, \gamma_1}(t, v) = \{\hat{S}(t, v-)\}^{\rho_1} \{1 - \hat{S}(t, v-)\}^{\gamma_1}
\]
where $B$ is the beta function with parameters $\alpha$ and $\beta$. This equation indicates that the uncensored case correlations only depend on the $\rho$ and $\tau$. The distribution of the survival curve will not affect value in the uncensored case correlation matrix.

Let $v_1 < \ldots < v_m$ be the $m$ distinct uncensored times in pooled data at a given study time $t$. Based on [Tarone (1981)] and the correlation between two statistics can be estimated as

$$
\hat{\rho}_{G^\rho_1 \cdot \gamma_1 (t), G^\rho_2 \cdot \gamma_2 (t)} = \frac{\sum_{r=1}^{m} \left\{ w_{\rho_1, \gamma_1, r} (t) w_{\rho_2, \gamma_2, r} (t) \frac{n_{1,r}(t) \left(n_r(t) - n_{1,r}(t)\right)}{n_r(t)} \frac{(n_r(t)-d_r(t))d_r(t)}{n_r(t)-1} \right\}}{\sqrt{\sum_{r=1}^{m} \left\{ w_{\rho_1, \gamma_1, r}^2 (t) \frac{n_{1,r}(t) \left(n_r(t) - n_{1,r}(t)\right)}{n_r(t)} \frac{(n_r(t)-d_r(t))d_r(t)}{n_r(t)-1} \right\}}} \times \frac{1}{\sqrt{\sum_{r=1}^{m} \left\{ w_{\rho_2, \gamma_2, r}^2 (t) \frac{n_{1,r}(t) \left(n_r(t) - n_{1,r}(t)\right)}{n_r(t)} \frac{(n_r(t)-d_r(t))d_r(t)}{n_r(t)-1} \right\}}} \tag{3.2}
$$

where

$$
w_{\rho_1, \gamma_1, r} = \left\{ \hat{S}(t, v_r -) \right\}^{\rho_1} \left( 1 - \hat{S}(t, v_r -) \right)^{\gamma_1} \text{ and } w_{\rho_2, \gamma_2, r} = \left\{ \hat{S}(t, v_r -) \right\}^{\rho_2} \left( 1 - \hat{S}(t, v_r -) \right)^{\gamma_2}
$$

and at a given study time $t$, $d_r$ stands for the number of deaths at time $v_r$ and $n_r$ gives the number of individuals at risk at time $v_r$, that is, the number of individuals alive just prior to time $v_r$. Both $d_{1,r}$ and $n_{1,r}$ correspond to the same definitions but restricted to observations in the $1^{st}$ group which is the treatment group. The KM estimator in the pooled sample as

$$
\hat{S}(t, v_m -) = \prod_{r=1}^{r=m-1} \left( 1 - \frac{d_r(t)}{n_r(t)} \right).
$$
Uncensored and censored correlation matrices are calculated based on simulations. Correlation between each two FH weighted log-rank statistics was presented under uncensored case and seven censored case models for both \( (G^{0,0}, G^{1,0}, G^{0,1}, G^{1,1}) \) and \( (G^{0,0}, G^{2,0}, G^{0,2}, G^{2,2}) \) families in Table 3.1. Correlation values in Table 3.1 were calculated as the mean of all simulations. Based on the result in Table 3.1 we can see a significant difference in correlation value between uncensored and any censored correlation matrices. However, when we focus on seven censored case correlation matrices, the difference is very small between any two correlation matrices among them. Censored case correlation matrices have higher value in all correlation pairs than the uncensored case one for ether

<table>
<thead>
<tr>
<th>Correlation for uncensored case</th>
<th>Correlation for Model A (constant diff.)</th>
<th>Correlation for Model B (early diff.)</th>
<th>Correlation for Model C (late diff.)</th>
<th>Correlation for Model D (middle diff.)</th>
<th>Correlation for Model E (cross favor treatment)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( G^{1,0} )</td>
<td>0.866</td>
<td>0.866</td>
<td>0.945</td>
<td>0.950</td>
<td>0.933</td>
</tr>
<tr>
<td>( G^{0,1} )</td>
<td>0.866</td>
<td>0.913</td>
<td>0.855</td>
<td>0.849</td>
<td>0.856</td>
</tr>
<tr>
<td>( G^{1,1} )</td>
<td>0.745</td>
<td>0.745</td>
<td>0.930</td>
<td>0.932</td>
<td>0.937</td>
</tr>
<tr>
<td>( G^{2,0} )</td>
<td>0.837</td>
<td>0.725</td>
<td>0.872</td>
<td>0.848</td>
<td>0.935</td>
</tr>
<tr>
<td>( G^{0,2} )</td>
<td>1</td>
<td>0.725</td>
<td>0.718</td>
<td>0.705</td>
<td>0.852</td>
</tr>
<tr>
<td>( G^{2,2} )</td>
<td>1</td>
<td>0.725</td>
<td>0.718</td>
<td>0.705</td>
<td>0.852</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Correlation for Model F (cross favor control)</th>
<th>Correlation for Null</th>
<th>Correlation for Model B (early diff.)</th>
<th>Correlation for Model C (late diff.)</th>
<th>Correlation for Model D (middle diff.)</th>
<th>Correlation for Model E (cross favor treatment)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( G^{1,0} )</td>
<td>0.933</td>
<td>0.933</td>
<td>0.933</td>
<td>0.933</td>
<td>0.933</td>
</tr>
<tr>
<td>( G^{0,1} )</td>
<td>0.856</td>
<td>0.856</td>
<td>0.856</td>
<td>0.856</td>
<td>0.856</td>
</tr>
<tr>
<td>( G^{1,1} )</td>
<td>0.833</td>
<td>0.833</td>
<td>0.833</td>
<td>0.833</td>
<td>0.833</td>
</tr>
<tr>
<td>( G^{2,0} )</td>
<td>0.721</td>
<td>0.721</td>
<td>0.721</td>
<td>0.721</td>
<td>0.721</td>
</tr>
<tr>
<td>( G^{0,2} )</td>
<td>0.876</td>
<td>0.876</td>
<td>0.876</td>
<td>0.876</td>
<td>0.876</td>
</tr>
<tr>
<td>( G^{2,2} )</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
(G^{0,0}, G^{1,0}, G^{0,1}, G^{1,1}) or (G^{0,0}, G^{2,0}, G^{0,2}, G^{2,2}) families. By comparing the correlation matrix between (G^{0,0}, G^{1,0}, G^{0,1}, G^{1,1}) and (G^{0,0}, G^{2,0}, G^{0,2}, G^{2,2}) families in each condition, (G^{0,0}, G^{1,0}, G^{0,1}, G^{1,1}) family has slightly higher correlation value than the (G^{0,0}, G^{2,0}, G^{0,2}, G^{2,2}) at the same location in the matrix among each models. Pair of G^{0,1} and G^{1,0} always has the smallest correlation in (G^{0,0}, G^{1,0}, G^{0,1}, G^{1,1}) family and Pair of G^{0,2} and G^{2,0} always has the smallest correlation in (G^{0,0}, G^{2,0}, G^{0,2}, G^{2,2}) family among each model. Among one uncensored case and seven censored case models, uncensored case correlation matrix has the smallest correlation for either pair of G^{0,1} and G^{1,0} or pair of G^{0,2} and G^{2,0}.

3.2.2 Derive MERT Statistics by Using Real Censored Correlation

While correlation is calculated in Section 3.2.1 we can identify the extreme pair by selecting the lowest correlation among each pair in four members family group G^{0,0}, G^{1,0}, G^{0,1}, and G^{1,1} and G^{0,0}, G^{2,0}, G^{0,2}, and G^{2,2}. Once the correlation for the extreme pair has been identified as ρ_{12}, if ρ_{12} met the condition in equation 2.11 then the MERT for the entire family as R_{12}

\[ R_{12} = [2 (1 + ρ_{12})]^{-\frac{1}{2}} \left( Z_1 + Z_2 \right), \]

It has maximin efficiency \((1 + ρ_{12})/2\).

Otherwise, if the equation 2.11 is not satisfy, then the MERT for the extreme pair is not the MERT for the entire family, there often is a test statistic T_3 which is the next candidate for MERT for the entire family.

Among correlation matrix for each model in table 3.1 we can find that the pair of G^{1,0} and G^{0,1} always has the smallest correlation value in the correlation matrix for (G^{0,0}, G^{1,0}, G^{0,1}, G^{1,1}) family and the pair of G^{2,0} and G^{0,2} always has the smallest correlation value in the correlation matrix for (G^{0,0}, G^{2,0}, G^{0,2}, G^{2,2}) family. It indicates that (G^{1,0} and G^{0,1}) and (G^{2,0} and G^{0,2}) is the extreme pair for (G^{0,0}, G^{1,0}, G^{0,1}, G^{1,1}) and (G^{0,0}, G^{2,0}, G^{0,2}, G^{2,2}), respectively.

Under uncensored condition, by using the simplified correlation equation 3.1 we have ρ_{1i} + ρ_{2i} = 1 + ρ_{12} when ρ = γ = 1.584. It indicates that the equation R_{12} =
\[2 (1 + \rho_{1,2})^{-\frac{1}{2}} (Z_1 + Z_2)\] will be used to calculate the MERT statistic when upper boundary of \( \rho = \gamma \leq 1.584 \). Otherwise, then the form \( \sum_{i=1}^{3} a_i Z_i, a_i > 0 \) will be used to calculate the candidate statistics of MERT. For censored models, most of MERT for extreme pair of \( G^{1,0} \) and \( G^{0,1} \) are the MERT for family \((G^{0,0}, G^{1,0}, G^{0,1}, G^{1,1})\) by checking the MERT condition in equation 2.11. Similar to uncensored case condition, most extreme pair of \( G^{2,0} \) and \( G^{0,2} \) is not the MERT for the entire family \((G^{0,0}, G^{2,0}, G^{0,2}, G^{2,2})\) in censored case models. Therefore, the MERT for family \((G^{0,0}, G^{2,0}, G^{0,2}, G^{2,2})\) is calculated based on the form \( \sum_{i=1}^{3} a_i Z_i, a_i > 0 \) for most censored model.
Chapter 4
Evaluation of NPH Methods Performance when Survival Time Follows a Piecewise Exponential Distribution

4.1 Method and Simulation Setup

Table 4.1: Survival Configuration for Each Model in Simulation

<table>
<thead>
<tr>
<th>Model</th>
<th>λ for Treatment</th>
<th>λ for Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model A (constant diff.)</td>
<td>0.3</td>
<td>0.4</td>
</tr>
<tr>
<td>Model B (early diff.)</td>
<td>0.2 + 0.4I(t ≥ 1.75) − 0.2I(t ≥ 3.5)</td>
<td>0.4</td>
</tr>
<tr>
<td>Model C (late diff.)</td>
<td>0.4 − 0.25I(t ≥ 2) − 0.1I(t ≥ 2.2)</td>
<td>0.4</td>
</tr>
<tr>
<td>Model D (middle diff.)</td>
<td>0.4 − 0.3I(t ≥ 1) + 0.9I(t ≥ 3) − 0.6I(t ≥ 4)</td>
<td>0.4</td>
</tr>
<tr>
<td>Model E (cross at middle and overall favor treatment)</td>
<td>0.2 + 0.65I(t ≥ 2.2)</td>
<td>0.4</td>
</tr>
<tr>
<td>Model F (cross at middle and overall favor control)</td>
<td>0.2 + I(t ≥ 1.7)</td>
<td>0.4</td>
</tr>
<tr>
<td>Null</td>
<td>0.4</td>
<td>0.4</td>
</tr>
</tbody>
</table>

Enrollment distribution ∼ uniform(0, 3); Censor distribution ∼ min(exp(1/35), 5); Maximum study time = 5; Total number per group = 300; Simulation numbers for alternative = 5,000; Simulation numbers for null = 10,000

Monte Carlo simulations were performed for a null model and six alternative models in order to compare the powers of the FH tests, MERT, MX, and RMST under a variety of possible differences between the survival distributions of the treatment and control groups. In the null model, the data for both treatment and control groups were generated from the same exponential distribution. In the alternative ones, survival time was assumed to follow
Figure 4.1: Simulation scenarios

a piecewise exponential distribution with piecewise constant hazard

$$\lambda_i(v|v_0, v_1, \ldots, v_J) = \sum_{j=1}^{j=J} I_{[v_{j-1}, v_j]}(v) \lambda_{ij}$$

where $i = 1$ or 2 as label for treatment arm and control arm, respectively. $v$ is the relative times (time measured from the real time of entry). $I_A(v)$ is an indication function with value 1 if $v \in A$ and 0 if $v \notin A$. Time $v_1, v_2, \ldots, v_J, v_m$ are change points for the piecewise exponential model. $v_0 = 0$ and $v_J = \infty$. They are the boundary values for formulation purpose [Huang and Kuan 2017]. The survival function of the probability of
survival at time $v$ can be derived as follows:

$$S_i(v) = \exp(-\lambda_i(v|v_0, v_1, \ldots, v_J)) = \exp\left(-\sum_{j=1}^{J} I_{[v_{j-1}, v_j]}(v) \lambda_{ij}\right)$$

Six different types of piecewise exponential distributions were selected to represent six typical survival models. They are one proportional hazards model, and five non-proportional hazards models which represent: early difference, late difference, middle difference, cross at middle overall favor treatment, and cross at middle overall favor control.

Furthermore, for both null and alternative cases, censoring and staggered entering condition were considered. Censoring was exponentially distributed with $\lambda = 1/35$ in all models. For staggered entering, the enrollment distribution was Uniform$(0, 3)$ and the study duration was $t = 5$ for all models. censoring time, the accrual and event time from different subjects were independent. All these setting resulted in a average of 30% censoring rate among all seven simulated models. The detail description for each model was presented in Table 4.1 and shown graphically in Figure 4.1.

The FH tests were performed for FH families $(G_{0,0}, G_{0,1}, G_{1,0}, G_{1,1})$ and $(G_{0,0}, G_{2,0}, G_{0,2}, G_{2,2})$ for all seven survival models. The correlation between any two FH statistics in each family was calculated based on both uncensored case and censored case correlation matrices. Based on those correlation matrices, MERT and MX test were performed for all four members family which marked as MERT4 and MX4, respectively. Three members families $G_{0,0}, G_{1,0}, G_{1,1}$ and $G_{0,0}, G_{2,0}, G_{2,2}$ were considered for without late difference condition models. MERT and MX tests for those three members families were marked as MERT3 and MX3, respectively. Furthermore, WLT with $w(t) = \hat{S}(t) \hat{G}(t)^{-1}$ was added as one of alternative test which marked as WCR. Its test statistics combined with four FH ones and added into MX test, marked as MX5, to evaluate their performance in all seven conditions.

As we known in section 2.4, each RMST has a restricted time $v_{cut}$ in order to have a closed form area under the KM curves. Many papers recommended the $v_{cut}$ should be clinical meaningful and closer to the end of the study follow-up in order to cover most survival outcomes. It should also be prespecified at the study design, so that the integrity of the study can be protected and the selection bias can be minimized.
In my simulation study, three common $v_{\text{cut}}$ were considered to generate the RMST. They were: (a) end of the study follow-up, (b) minimum of the maximum event time of each arm (minimax event time), (c) minimum of the maximum observed (event or censored) time (minimax observed time) of each arm.

Similar configurations for constant difference, early difference, late difference and middle difference were previously considered by Fleming et al. (1987). Two crossing model (cross at middle overall favor to treatment, and cross at middle overall favor to control) were added to evaluate the performance for each test under crossing survival models. For all seven models, three hundred subjects were set as the sample size for each arm, and 5000 simulations were performed for each alternatives and 10000 simulations were performed for the null. All tests were performed using a one-sided test with a 0.025 significance level.

4.2 Results

4.2.1 Performance of Tests

The performance of each test was evaluated by their simulated power estimates in each typical survival models. Detail result can be seen in below table 4.2.

4.2.2 Comparison of Performance

Compared all tests in each simulated condition, FHT $G^{1,0}$ whose weighting function emphasizes early difference, has the highest power in its corresponding simulated model B for early separation. FHT $G^{0,1}$ whose weighting function emphasizes late difference, has the highest power in its corresponding simulated model C for late separation. FHT $G^{1,1}$ whose weighting function emphasizes middle difference, has the highest power in its corresponding simulated model D for middle separation among all tests. Similar trend can be found in ($G^{0,0}, G^{2,0}, G^{0,2}, G^{2,2}$) family. Compared all tests in each simulated condition, $G^{2,0}, G^{0,2}, G^{2,2}$ has the highest power in its corresponding designed model $B, C$, and $D$, respectively. Power of those FHT test in ($G^{0,0}, G^{2,0}, G^{0,2}, G^{2,2}$) family is slightly higher than the power of test in ($G^{0,0}, G^{1,0}, G^{0,1}, G^{1,1}$) family.

Compared all general tests (log-rank Test, MERT4, MX4, and RMST), the log-rank
### Table 4.2: Power Estimates for each Model

<table>
<thead>
<tr>
<th>Model</th>
<th>A (constant diff.)</th>
<th>B (early diff.)</th>
<th>C (late diff.)</th>
<th>D (middle diff.)</th>
<th>E (Cross at middle overall favor treatment)</th>
<th>F (Cross at middle overall favor control)</th>
<th>Null</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Overall Censored rate</td>
<td>33.9%</td>
<td>30.8%</td>
<td>36.0%</td>
<td>31.9%</td>
<td>30.1%</td>
<td>23.5%</td>
</tr>
<tr>
<td>$G^{0.0}$, $G^{1.0}$, $G^{2.0}$, $G^{3.0}$ family</td>
<td>$G^{0.0}$ (log-rank test) FHT</td>
<td>0.3922</td>
<td>0.6540</td>
<td>0.7272</td>
<td>0.8140</td>
<td>0.5744</td>
<td>&lt; 0.0002</td>
</tr>
<tr>
<td></td>
<td>$G^{1.0}$ (early diff.) FHT</td>
<td>0.7622</td>
<td>0.9522</td>
<td>0.2654</td>
<td>0.7448</td>
<td>0.9762</td>
<td>0.2676</td>
</tr>
<tr>
<td></td>
<td>$G^{2.0}$ (late diff.) FHT</td>
<td>0.6870</td>
<td>0.0156</td>
<td>0.9854</td>
<td>0.7126</td>
<td>0.0014</td>
<td>&lt; 0.0002</td>
</tr>
<tr>
<td></td>
<td>$G^{3.1}$ (middle diff.) FHT</td>
<td>0.7516</td>
<td>0.8160</td>
<td>0.8486</td>
<td>0.9600</td>
<td>0.1960</td>
<td>&lt; 0.0002</td>
</tr>
<tr>
<td></td>
<td>WCR</td>
<td>0.7960</td>
<td>0.9036</td>
<td>0.6190</td>
<td>0.6238</td>
<td>0.8736</td>
<td>0.1134</td>
</tr>
<tr>
<td></td>
<td>MERT4 (uncensored corr)</td>
<td>0.8232</td>
<td>0.4932</td>
<td>0.8766</td>
<td>0.8348</td>
<td>0.3786</td>
<td>0.7960</td>
</tr>
<tr>
<td></td>
<td>MX4 (uncensored corr)</td>
<td>0.7580</td>
<td>0.9080</td>
<td>0.9690</td>
<td>0.9264</td>
<td>0.9522</td>
<td>0.1672</td>
</tr>
<tr>
<td></td>
<td>MERT4</td>
<td>0.7982</td>
<td>0.4588</td>
<td>0.8556</td>
<td>0.8128</td>
<td>0.3512</td>
<td>&lt; 0.0002</td>
</tr>
<tr>
<td></td>
<td>MX4</td>
<td>0.7804</td>
<td>0.9172</td>
<td>0.9734</td>
<td>0.9342</td>
<td>0.9566</td>
<td>0.1780</td>
</tr>
<tr>
<td></td>
<td>MX5</td>
<td>0.8000</td>
<td>0.9218</td>
<td>0.9788</td>
<td>0.9324</td>
<td>0.9542</td>
<td>0.1732</td>
</tr>
<tr>
<td></td>
<td>MERT3 ($G^{0.0}$, $G^{1.0}$, $G^{2.0}$, $G^{3.0}$)</td>
<td>0.7988</td>
<td>0.6966</td>
<td>0.9164</td>
<td>0.7448</td>
<td>0.0024</td>
<td>0.0256</td>
</tr>
<tr>
<td></td>
<td>MX3 ($G^{0.0}$, $G^{1.0}$, $G^{2.0}$, $G^{3.0}$)</td>
<td>0.7910</td>
<td>0.9280</td>
<td>0.9430</td>
<td>0.9624</td>
<td>0.2002</td>
<td>0.0254</td>
</tr>
<tr>
<td>$G^{0.0}$, $G^{1.0}$, $G^{2.0}$, $G^{3.0}$ family</td>
<td>$G^{1.0}$ (log-rank test) FHT</td>
<td>0.8022</td>
<td>0.6540</td>
<td>0.7272</td>
<td>0.8140</td>
<td>0.5744</td>
<td>&lt; 0.0002</td>
</tr>
<tr>
<td></td>
<td>$G^{2.0}$ (early diff.) FHT</td>
<td>0.6794</td>
<td>0.9892</td>
<td>0.0886</td>
<td>0.5140</td>
<td>0.9956</td>
<td>0.8436</td>
</tr>
<tr>
<td></td>
<td>$G^{3.2}$ (late diff.) FHT</td>
<td>0.5352</td>
<td>0.0012</td>
<td>0.9984</td>
<td>0.2776</td>
<td>&lt; 0.0002</td>
<td>0.0231</td>
</tr>
<tr>
<td></td>
<td>WCR</td>
<td>0.7960</td>
<td>0.9036</td>
<td>0.6190</td>
<td>0.6238</td>
<td>0.8736</td>
<td>0.1134</td>
</tr>
<tr>
<td></td>
<td>MERT4 (uncensored corr)</td>
<td>0.8278</td>
<td>0.4716</td>
<td>0.9522</td>
<td>0.6922</td>
<td>0.2566</td>
<td>&lt; 0.0002</td>
</tr>
<tr>
<td></td>
<td>MX4 (uncensored corr)</td>
<td>0.7332</td>
<td>0.0706</td>
<td>0.9944</td>
<td>0.9682</td>
<td>0.9862</td>
<td>0.7218</td>
</tr>
<tr>
<td></td>
<td>MERT4</td>
<td>0.7902</td>
<td>0.4648</td>
<td>0.9334</td>
<td>0.5414</td>
<td>0.2434</td>
<td>&lt; 0.0002</td>
</tr>
<tr>
<td></td>
<td>MX4</td>
<td>0.7512</td>
<td>0.9740</td>
<td>0.9954</td>
<td>0.9710</td>
<td>0.9874</td>
<td>0.7348</td>
</tr>
<tr>
<td></td>
<td>MX5</td>
<td>0.7728</td>
<td>0.9714</td>
<td>0.9968</td>
<td>0.9602</td>
<td>0.9892</td>
<td>0.7252</td>
</tr>
<tr>
<td></td>
<td>MERT3 ($G^{0.0}$, $G^{2.0}$, $G^{2.0}$, $G^{2.0}$)</td>
<td>0.7968</td>
<td>0.7400</td>
<td>0.9378</td>
<td>0.8250</td>
<td>0.0094</td>
<td>0.0258</td>
</tr>
<tr>
<td></td>
<td>MX3 ($G^{0.0}$, $G^{2.0}$, $G^{2.0}$, $G^{2.0}$)</td>
<td>0.7718</td>
<td>0.9804</td>
<td>0.9748</td>
<td>0.9910</td>
<td>0.7680</td>
<td>0.0246</td>
</tr>
<tr>
<td></td>
<td>RMST (study End)</td>
<td>0.8046</td>
<td>0.7298</td>
<td>0.8186</td>
<td>0.7230</td>
<td>0.6544</td>
<td>0.0014</td>
</tr>
<tr>
<td></td>
<td>RMST (minimax Event)</td>
<td>0.7932</td>
<td>0.8402</td>
<td>0.3264</td>
<td>0.8300</td>
<td>0.8528</td>
<td>0.0184</td>
</tr>
<tr>
<td></td>
<td>RMST (minimax Observed)</td>
<td>0.7940</td>
<td>0.8968</td>
<td>0.4442</td>
<td>0.8278</td>
<td>0.8494</td>
<td>0.0170</td>
</tr>
<tr>
<td></td>
<td>RMST-T (study end)</td>
<td>2.59</td>
<td>2.52</td>
<td>2.61</td>
<td>2.54</td>
<td>2.47</td>
<td>2.03</td>
</tr>
<tr>
<td></td>
<td>RMST-C (study end)</td>
<td>2.16</td>
<td>2.16</td>
<td>2.16</td>
<td>2.16</td>
<td>2.16</td>
<td>2.16</td>
</tr>
<tr>
<td></td>
<td>T-C (study end)</td>
<td>0.42</td>
<td>0.35</td>
<td>0.45</td>
<td>0.37</td>
<td>0.31</td>
<td>-0.14</td>
</tr>
</tbody>
</table>

MERT4 and MX4: MERT and MX are calculated based on 4 member of FH statistics family, respectively. MERT4 and MX4 (uncensored corr): MERT4 and MX4 are calculated based uncensored correlation matrix. RMST-T (study end) is average of RMST for treatment arm by $v_{\text{cut}} = \text{study end time}$. RMST-C (study end) is average of RMST for control arm by $v_{\text{cut}} = \text{study end time}$. T-C (study end) is average of RMST difference between treatment and control.

The test remains the most powerful test under the PH scenario in model A when we ignore the MERT4 by uncensored correlation matrix. Under all NPH scenarios (model B to model F), MX4 is the most powerful test in all four general tests regardless of the treatment effect. Since the model F is a model that two arms cross at middle of study and overall favor to control arm, the one-sided test power should be less than the type one error rate for all tests. However, power estimates are all greater than its one-sided type one error rate 0.025 in all MX tests for model F. Those results indicate that those MX tests may be not able to handle model F (cross at middle and overall favor control) and test statistic may lead reverse result to the real in model F. Compared power performance in all MX tests between
\((G^{0.0}, G^{2.0}, G^{0.2}, G^{2.2})\) family and \((G^{0.0}, G^{1.0}, G^{0.1}, G^{1.1})\), \((G^{0.0}, G^{2.0}, G^{0.2}, G^{2.2})\) family has slightly higher power than the \((G^{0.0}, G^{1.0}, G^{0.1}, G^{1.1})\) family one.

Compared log-rank test and MERT4, log-rank test is more powerful than MERT4 under all models except for model C (late diff.) which indicate MERT4 don’t have many power advantage to log-rank test in general. However, if we have already known that there is no late separation in the study and with removing \(G^{0.1}/G^{0.2}\) from the family for MERT, then MERT3 is more powerful than log-rank test under both model B (early diff.) and model D (middle diff.). It indicates, by removing late different test statistics in MERT, MERT3 have gained power advantage to log-rank test in early separation and middle separation models.

Power of most MERT4/MERT3 test in \((G^{0.0}, G^{2.0}, G^{0.2}, G^{2.2})\) family is higher than the power of test in \((G^{0.0}, G^{1.0}, G^{0.1}, G^{1.1})\) family except for model A and model D in MERT4, and model A in MERT3.

Compared log-rank test and WCR, WCR test gains about 30% power in model B (early diff.), but loses about 10% and 20% power in model C (late diff.) and model D (middle diff.) in both \((G^{0.0}, G^{1.0}, G^{0.1}, G^{1.1})\) and \((G^{0.0}, G^{2.0}, G^{0.2}, G^{2.2})\) families, respectively. Power performance for MX5 is close to the power performance for MX4 in all seven scenarios for both \((G^{0.0}, G^{1.0}, G^{0.1}, G^{1.1})\) and \((G^{0.0}, G^{2.0}, G^{0.2}, G^{2.2})\) families.

Compared power estimates in MERT4 test between using uncensored correlation matrix and using censored correlation matrix, uncensored correlation one has slightly higher power then the censored correlation MERT4 in all seven scenarios for either \((G^{0.0}, G^{1.0}, G^{0.1}, G^{1.1})\) or \((G^{0.0}, G^{2.0}, G^{0.2}, G^{2.2})\) families. However, the power estimates is 0.0305 for MERT4 with uncensored correlation under the null scenario. It shows that the uncensored correlation MERT4 can not preserve the type one error rate which indicate the uncensored correlation matrix may be not suitable for censored case data. Under the null, all rest tests can preserve the type I error.

Compared three RMST cut strategies, it is hard to say which strategy is the best. RMST (study end) is more powerful than log-rank test under all model except model D (middle diff). However, both RSMT (Minimax Event) and RSMT (Minimax Observed) have higher power than log-rank test in model B and model D. Compared RSMT (Minimax Event) and RSMT (Minimax Observed) with RSMT (study End), both improved 10% power compared
to RSMT (study End) in models (B, D, E), but lost about 30% power in model C.
Chapter 5

Evaluation of NPH Methods Performance when Treatment Time Follows a Time-Dependent Exponential Distribution

5.1 Method and Simulation Setup

<table>
<thead>
<tr>
<th>Table 5.1: Survival Configuration for Each Model in Simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Model A</strong> (Converging Hazards)</td>
</tr>
<tr>
<td>( \lambda ) for Treatment ( (1.1 - \frac{1.1}{(1+30)}) \times 0.4 )</td>
</tr>
<tr>
<td>( \lambda ) for Control ( 0.4 )</td>
</tr>
<tr>
<td><strong>Model B</strong> (Diverging Hazards)</td>
</tr>
<tr>
<td>( 0.3 )</td>
</tr>
<tr>
<td>( 0.3 + 0.06t )</td>
</tr>
<tr>
<td><strong>Model C</strong></td>
</tr>
<tr>
<td>( \frac{(t^2-5t+25/2)\times4}{125} )</td>
</tr>
<tr>
<td>( 0.4 )</td>
</tr>
</tbody>
</table>

Enrollment distribution \( \sim \) uniform\((0, 3)\); Censor distribution \( \sim \) min\((\exp(1/35), 5)\); Maximum study time = 5; Total number per group = 300; Simulation numbers for alternative = 5,000

Additional monte Carlo simulations were performed for three alternative models in order to compare the powers of the FH tests, MERT, MX, and RMST under a variety of possible differences between the survival distributions of the treatment and control groups. In those alternative models, survival time is assumed to follow a time-dependent exponential distribution. Those three NPH models are represented: converging hazards, diverging hazards, half diverging and half converging hazards. The same simulation procedures and evaluation methods in the section 4.1 were used. The detail description for each model was presented in Table 5.1 and shown graphically in Figure 5.1

5.2 Results
5.2.1 Performance of Tests

The performance of each test was evaluated by their simulated power estimates in each typical survival models. Detail result can be seen in below table 5.2.

5.2.2 Comparison of Performance

Compared all tests, $G^{1,0}, G^{0,1}, G^{0,1}$ has the highest power in model A (converging hazards), B (diverging hazards), and C (half diverging half converging hazards) for $(G^{0,0}, G^{1,0}, G^{0,1}, G^{1,1})$ family, respectively. $G^{2,0}, G^{0,2}, G^{2,2}$ has the highest power in model A (converging hazards), B (diverging hazards), and C (half diverging half converging hazards) for $(G^{0,0}, G^{2,0}, G^{0,2}, G^{2,2})$ family, respectively. Power of those corresponding FH test in $(G^{0,0}, G^{2,0}, G^{0,2}, G^{2,2})$ family is slightly higher than the power of test in $(G^{0,0}, G^{1,0}, G^{0,1}, G^{1,1})$ family except for model B.
Table 5.2: Power Estimates for each Model

<table>
<thead>
<tr>
<th>Model</th>
<th>A (Converging Hazards)</th>
<th>B (Diverging Hazards)</th>
<th>C (Half Diverging Half Converging Hazards)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Drop out rate</td>
<td>Overall Censored rate</td>
<td></td>
</tr>
<tr>
<td>$G^{0.0}$, $G^{1.0}$, $G^{1.1}$, $G^{1.2}$ family</td>
<td>$G^{0.0}$ (log-rank test) FHT</td>
<td>0.7550</td>
<td>0.6704</td>
</tr>
<tr>
<td></td>
<td>$G^{1.0}$ (early diff.) FHT</td>
<td>0.9204</td>
<td>0.4396</td>
</tr>
<tr>
<td></td>
<td>$G^{0.1}$ (late diff.) FHT</td>
<td>0.1438</td>
<td>0.8238</td>
</tr>
<tr>
<td></td>
<td>$G^{1.1}$ (middle diff.) FHT</td>
<td>0.2748</td>
<td>0.7608</td>
</tr>
<tr>
<td></td>
<td>MERT4 (uncensored corr)</td>
<td>0.6802</td>
<td>0.7538</td>
</tr>
<tr>
<td></td>
<td>MX4</td>
<td>0.8534</td>
<td>0.7510</td>
</tr>
<tr>
<td></td>
<td>MERT3 ($G^{0.0}$, $G^{1.0}$, $G^{1.1}$)</td>
<td>0.6984</td>
<td>0.6558</td>
</tr>
<tr>
<td></td>
<td>MX3 ($G^{0.0}$, $G^{1.0}$, $G^{1.1}$)</td>
<td>0.8830</td>
<td>0.7138</td>
</tr>
<tr>
<td>$G^{0.0}$, $G^{1.0}$, $G^{2.0}$, $G^{2.2}$ family</td>
<td>$G^{0.0}$ (log-rank test) FHT</td>
<td>0.7550</td>
<td>0.6704</td>
</tr>
<tr>
<td></td>
<td>$G^{2.0}$ (early diff.) FHT</td>
<td>0.9652</td>
<td>0.2634</td>
</tr>
<tr>
<td></td>
<td>$G^{0.2}$ (late diff.) FHT</td>
<td>0.0654</td>
<td>0.8096</td>
</tr>
<tr>
<td></td>
<td>$G^{2.2}$ (middle diff.) FHT</td>
<td>0.1680</td>
<td>0.7632</td>
</tr>
<tr>
<td></td>
<td>MERT4 (uncensored corr)</td>
<td>0.7286</td>
<td>0.7772</td>
</tr>
<tr>
<td></td>
<td>MX4 (uncensored corr)</td>
<td>0.9186</td>
<td>0.7566</td>
</tr>
<tr>
<td></td>
<td>MERT3 ($G^{0.0}$, $G^{2.0}$, $G^{2.2}$)</td>
<td>0.7670</td>
<td>0.6258</td>
</tr>
<tr>
<td></td>
<td>MX3 ($G^{0.0}$, $G^{2.0}$, $G^{2.2}$)</td>
<td>0.9390</td>
<td>0.7064</td>
</tr>
<tr>
<td></td>
<td>RMST (study End)</td>
<td>0.7794</td>
<td>0.6728</td>
</tr>
<tr>
<td></td>
<td>RMST (minimax Event)</td>
<td>0.8406</td>
<td>0.5852</td>
</tr>
<tr>
<td></td>
<td>RMST (minimax Observed)</td>
<td>0.8384</td>
<td>0.5894</td>
</tr>
<tr>
<td></td>
<td>RMST T</td>
<td>2.03</td>
<td>2.09</td>
</tr>
<tr>
<td></td>
<td>RMST C</td>
<td>2.16</td>
<td>2.24</td>
</tr>
<tr>
<td></td>
<td>T-C</td>
<td>0.39</td>
<td>0.35</td>
</tr>
</tbody>
</table>

MERT4 and MX4: MERT and MX are calculated based on 4 member of FH statistics family, respectively.
MERT4 and MX4 (uncensored corr): MERT4 and MX4 are calculated based uncensored correlation matrix.
RMST T (study end) is average of RMST for treatment arm by $v_{cut} = $ study end time.
RMST C (study end) is average of RMST for control arm by $v_{cut} = $ study end time.
T-C (study end) is average of RMST difference between treatment and control.

Comparing all general tests (log-rank Test, MERT4, MX4 and RMST), MX4 is the most powerful in all models. Power of those MX tests in ($G^{0.0}, G^{2.0}, G^{0.2}, G^{2.2}$) family is higher than the power of test in ($G^{0.0}, G^{1.0}, G^{0.1}, G^{1.1}$) family except for model C.

Compared log-rank test and MERT4, MERT4 is more powerful than log-rank test under model B and C. However, if we have already known that there is no late difference in the study and removed the $G^{0.1}/G^{0.2}$ from the family for MERT, then MERT3 is more powerful than log-rank test under model A (converging hazards) in ($G^{0.0}, G^{2.0}, G^{0.2}, G^{2.2}$) family.

MERT4 and MX4 are more powerful than MERT3 and MX3 in both model B and C for both ($G^{0.0}, G^{1.0}, G^{0.1}, G^{1.1}$) and ($G^{0.0}, G^{2.0}, G^{0.2}, G^{2.2}$) families.

RMST (study end) is more powerful than log-rank test under all model except model C.

Compared RSMT (Minimax Event) and RSMT (Minimax Observed) with RSMT (study
End), both improved 5% power compared to RSMT (study End) in converging hazards model A.

Compared power estimates in MERT4 test between using uncensored correlation matrix and using censored correlation matrix, uncensored correlation one has slightly higher power than the censored correlation MERT4 in all seven scenarios for either \((G^{0,0}, G^{1,0}, G^{0,1}, G^{1,1})\) or \((G^{0,0}, G^{2,0}, G^{0,2}, G^{2,2})\) families except model A in \((G^{0,0}, G^{2,0}, G^{0,2}, G^{2,2})\) family.
Chapter 6
Evaluation of NPH Methods Performance by Using Real Samples

6.1 Example

Two real data examples were used to evaluate the performance of NPH methods.

Example 1 (Biofeedback) is a subset of the data from Videoendoscopic Biofeedback study [Denk and Kaider 1997]. Total 33 subjects were included in this subset. The effect of biofeedback treatment on time until treatment success was evaluated in patients suffering from aspiration after head and neck surgery in this study. The primary event was the time from start of treatment until the patient achieved full swallowing rehabilitation. Patients were randomized into two groups: one group of patients received videoendoscopic biofeedback treatment; the other group received the conservative treatment including thermal stimulation with ice and exercises for the lips, tongue, laryngeal closure and elevation. Treatment was started as soon as the healing process after surgery was finished.

Example 2 (Gastric) is a subset of the data from a gastric cancer study comparing treated with chemotherapy plus radiation versus chemotherapy alone in patients with locally advanced, nonresectable gastric carcinoma [Group 1982]. A total 90 subjects were included in this subset. The outcome of interest is survival time of patients with locally advanced, nonresectable gastric carcinoma. Patients were randomized into two groups: one group of patients treated chemotherapy; the other group treated chemotherapy plus radiation.

6.2 Method

The FH tests were performed for FH families \((G^0,0, G^1,0, G^0,1, G^1,1)\) and \((G^0,0, G^2,0, G^0,2, G^2,2)\) for two examples. The correlation between any two FH statistics in each family was
calculated based on censored case equation in 3.2. Based on those correlation matrices, MERT and MX test were performed for all four members family which marked as MERT4 and MX4, respectively. Three members families $G^{0,0}, G^{1,0}, G^{1,1}$ and $G^{0,0}, G^{2,0}, G^{2,2}$ were considered without late difference condition models. MERT and MX tests for those 3 members family were marked as MERT3 and MX3, respectively.

As we known in section 2.4, each RMST has a restricted time $v_{cut}$ in order to have a closed form area under the KM curves. Many papers recommended the $v_{cut}$ should be clinical meaningful and closer to the end of the study follow-up in order to cover most survival outcomes. It should also be prespecified at the study design, so that the integrity of the study can be protected and the selection bias can be minimized (Huang and Kuan, 2017). In my simulation study, Three common $v_{cut}$ were considered to generate the RMST. They were: (a) end of the study follow-up, (b) minimum of the maximum event time of each arm (minimax event time), (c) minimum of the maximum observed (event or censored) time (minimax observed time) of each arm.

Let conservative group as the treatment and biofeedback as the control in example 1. Let chemotherapy plus radiation as the treatment and chemotherapy alone as the control in example 2. All tests were performed using a one-sided test with a 0.025 significance level under the null hypothesis $H_0 : S_T (t,v) \leq S_C (t,v)$, and alternatives $H_a : S_T (t,v) > S_C (t,v)$. Treatment effect for each example was estimated by using WCR and RMST regression under the linear link. Details of the covariates estimation methods have been

![Figure 6.1: KM plots for Example 1 and Example 2](image-url)
described in section 2.5

6.2.1 Performance of Tests

The performance of each test was evaluated by their tests’ one-sided p-value in each example. Detail result can be seen below table 6.1.

Table 6.1: Test One-sided P-value for each Example

<table>
<thead>
<tr>
<th>Example</th>
<th>Example 1 (Biofeedback)</th>
<th>Example 2 (Gastric)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$G_{0,0}, G_{0,1}, G_{1,0}, G_{1,1}$ family</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$G_{0,0}$ (log-rank test) FHT</td>
<td>0.2293</td>
<td>0.3150</td>
</tr>
<tr>
<td>$G_{1,0}$ (early diff.) FHT</td>
<td>0.0681</td>
<td>0.0228</td>
</tr>
<tr>
<td>$G_{0,1}$ (late diff.) FHT</td>
<td>0.7776</td>
<td>0.9237</td>
</tr>
<tr>
<td>$G_{1,1}$ (middle diff.) FHT</td>
<td>0.4666</td>
<td>0.5420</td>
</tr>
<tr>
<td>MERT4</td>
<td>0.3449</td>
<td>0.3736</td>
</tr>
<tr>
<td>MX4</td>
<td>0.1175</td>
<td>0.0474</td>
</tr>
<tr>
<td>MERT3 ($G_{0,0}, G_{1,0}, G_{1,1}$)</td>
<td>0.2021</td>
<td>0.1578</td>
</tr>
<tr>
<td>MX3 ($G_{0,0}, G_{1,0}, G_{1,1}$)</td>
<td>0.1042</td>
<td>0.0396</td>
</tr>
<tr>
<td>$G_{0,0}, G_{2,0}, G_{1,0}, G_{1,1}$ family</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$G_{0,0}$ (log-rank test) FHT</td>
<td>0.2293</td>
<td>0.3150</td>
</tr>
<tr>
<td>$G_{2,0}$ (early diff.) FHT</td>
<td>0.0337</td>
<td>0.0047</td>
</tr>
<tr>
<td>$G_{0,2}$ (late diff.) FHT</td>
<td>0.9310</td>
<td>0.9766</td>
</tr>
<tr>
<td>$G_{2,2}$ (middle diff.) FHT</td>
<td>0.5888</td>
<td>0.6396</td>
</tr>
<tr>
<td>MERT4</td>
<td>0.4156</td>
<td>0.3486</td>
</tr>
<tr>
<td>MX4</td>
<td>0.0767</td>
<td>0.0137</td>
</tr>
<tr>
<td>MERT3 ($G_{0,0}, G_{2,0}, G_{2,2}$)</td>
<td>0.1791</td>
<td>0.0995</td>
</tr>
<tr>
<td>MX3 ($G_{0,0}, G_{2,0}, G_{2,2}$)</td>
<td>0.0635</td>
<td>0.0106</td>
</tr>
<tr>
<td>RMST (study End)</td>
<td>0.2190</td>
<td>0.985</td>
</tr>
<tr>
<td>RMST (minimax Event)</td>
<td>0.0230</td>
<td>0.597</td>
</tr>
<tr>
<td>RMST (minimax Observed)</td>
<td>0.2190</td>
<td>0.312</td>
</tr>
</tbody>
</table>

Table 6.2: Regression Coefficients with Standard Errors for Two Examples

<table>
<thead>
<tr>
<th>Example</th>
<th>WCR</th>
<th>RMSTREG Linear (Study End)</th>
<th>RMSTREG logit (Study End)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coefficient (SE)</td>
<td>exp(coef) (95% CI)</td>
<td>Coefficient (SE)</td>
</tr>
<tr>
<td>Example 1</td>
<td>Treatment</td>
<td>-0.60 (0.40)</td>
<td>0.55 (0.24, 1.20)</td>
</tr>
<tr>
<td>Example 2</td>
<td>Treatment</td>
<td>-0.46 (0.24)</td>
<td>0.63 (0.39, 1.00)</td>
</tr>
</tbody>
</table>
6.2.2 Comparison of Performance

Example 1 (Biofeedback) represent a typical early separation NPH condition. $G^{1,0}$ FHT has lowest p-value among all tests in $(G^{0,0}, G^{1,0}, G^{0,1}, G^{1,1})$ family. $G^{2,0}$ FHT has the lowest p-value among all tests in $(G^{0,0}, G^{2,0}, G^{0,2}, G^{2,2})$ family. Overall, RMST which cut at minimax event has the best test performance. It has the lowest p-value among all tests in both $(G^{0,0}, G^{1,0}, G^{0,1}, G^{1,1})$ and $(G^{0,0}, G^{2,0}, G^{0,2}, G^{2,2})$ families. This is because in example 1, the minimax event time is around time 35, which is in the early period of the study and there is a significant difference between two arms before this cut time point. Compared MERT4 and MX4, MX4 has lower p-value than MERT4 in both $(G^{0,0}, G^{1,0}, G^{0,1}, G^{1,1})$ and $(G^{0,0}, G^{2,0}, G^{0,2}, G^{2,2})$ families. P-value decreased from MERT4 to MERT3 and from MX4 to MX3 in both $(G^{0,0}, G^{1,0}, G^{0,1}, G^{1,1})$ and $(G^{0,0}, G^{2,0}, G^{0,2}, G^{2,2})$ families as well. Compared log-rank test, MERT4, and MERT3, log-rank test has slightly lower p-value than MERT4, but has slightly higher p-value than MERT3.

Example 2 (Gastric) represent a typical crossover at middle and overall favor control NPH condition. P-value for MX3, MX4, $G^{1,0}$ FHT, and $G^{2,0}$ FHT are all less than or close to 0.025, which lead reverse result to the real in all those tests. All rest tests fail to reject the null and display a true result to the real.

Regression coefficient estimates with standard errors are display in table 6.2 for Example 1 and Example 2. For the WCR, an average of hazard ratio at all death time between treatment and control is 0.55 for example 1 and is 0.63 for example 2. In the RMSTREG with linear link, the ratio of the RMST between treatment and control is 15.78 for example 1 and is -6.69 for example 2. In the RSMTREG with logit link, the log ratio of the RMST between treatment and control is 0.32 for example 1 and is -0.01 for example 2.
Chapter 7
Conclusion

Overall, based on above results, we can conclude that, if we have already known which model the study data will go, the corresponding FH test will be the first choice except for the model F (cross at middle and overall favor control). If model F will happen then log-rank test, MERT, or RMST will be used. If we are not sure what kind of model will be, but we definitely sure the model F (cross at middle and overall favor control) situation will not happen in the study, then the MX4 will always be our best general test. Furthermore, if we are not sure whether model F situation will happen or not, by considering MX may lead reverse one sided test result to the real in model F, the RMST or MERT can be one of the supportive general tests. Regarding to RMST cut point selection, we may use study end time when we are not sure what kind of model will be. Both time of last event and time of last observed are better choice than study end when we are sure early difference or middle difference will happen in the study. If we are sure model C (late difference) will not happen in the study, considering MX3 may lead reverse one sided test result to the real in model F, we recommend to use MERT3, RMST(last Event), or RMST(last observed). Above conclusion can be made under either \((G^{0,0}, G^{1,0}, G^{0,1}, G^{1,1})\) family or \((G^{0,0}, G^{2,0}, G^{0,2}, G^{2,2})\) family. \((G^{0,0}, G^{2,0}, G^{0,2}, G^{2,2})\) family is recommended to use, since it is more powerful in all above suggested MERT, MX tests than those tests in \((G^{0,0}, G^{1,0}, G^{0,1}, G^{1,1})\) family.
Part II

Two-stage in RMST with Conditional Power
Chapter 8

Introduction

8.1 Background

Two-stage designs are commonly used in clinical trials, especially in Phase II oncology studies. In two-stage design, patients are divided into two stages. At the completion of the first stage, an interim analysis is conducted with either stop for futility or stop for efficacy. The interim analysis result will be used to determine whether the second stage should be conducted or not. The trial may be stopped at the first stage if the result is not as good as expected. Conditional power as one of the common tool in two-stage design, is wildly used in monitoring futility or sample size re-estimation at the interim stage. Conditional power is also commonly used in methods for flexible sample design in clinical trials: likelihood, weighted, dual test, and promising zone approaches (Shih et al., 2016). Furthermore, the conditional power is also used in evaluation different assumptions about the trend of future data. The trend can be assumed under the current data, under the null hypothesis, or under the alternative hypothesis.

The RMST is a robust summary measurement for the survival time distribution. It does not rely on the PH assumption and can be interpreted perfectly in clinical. The RMST can be estimated even under heavy censoring (Huang and Kuan, 2017). Compared to log-tank test and other commonly used alternative methods for NPH, a test of the difference in RMST between the treatment arm and control arm may be more appropriate in recent complex data environment. The results in chapter 7 showed RMST had excellent performance in all typical NPH models. Method of designing a two-stage randomized clinical trial based on the difference in RMST is still a very new and hot topic. It is still not clear how to calculate conditional power based on observed interim data. In this section, we will discuss calculating covariance in RMST difference between interim and final analyses based on
observed data and using CP in a two-stage design with RMST endpoint.

8.2 Research Objectives

In Part II of this document, the following objectives were set:

Objective I: To derive asymptotic covariance of RMST test statistics between the interim and final analyses.

Objective II: To evaluate the accuracies of the approximations of covariance through simulations.

Objective III: To derive conditional power based on asymptotic covariance and evaluate its accuracies through simulations.
Chapter 9

Literature Review

9.1 Two-Stage Design in RMST with Conditional Power (CP)

In two-stage design with oncology trials, the commonly used method is based on the log-rank test which is event driven and the variance-covariance matrix of the test statistics under the null can be derived based on the number of events. The interim analysis is conducted when the planned number of events is reached. However, two stage design for RMST is a total different story. The interim cut is changed to a specific time point. The structure of the variance-covariance matrix of the test statistics become more complex. In addition, there is no general independent increment structure among test statistics in interim and final analyses.

9.1.1 Notation

Similar to Section 2.4, we assume the two-sample general random censoring model with staggered entry. Observations are made on $n_i$ individuals from sample $i$ which $i = 1$ indicate the treatment group, otherwise $i = 2$ represents the control. Let $(W_{ij}, V_{ij}, C_{ij}), j = 1, \ldots, n_i, i = 1, 2$, denote the real time of entry during the accrual period $[0, t_a]$, the real time from entry until failure, and the time from entry to censor for subject $j$ in treatment $i$ respectively. It is assumed that $W_{ij}, V_{ij}, C_{ij}$ are all nonnegative random variables.

At a given real study time $t$, let

$$X_{ij}(t) = \max(\min(V_{ij}, C_{ij}, t - W_{ij}), 0)$$

and

$$\delta_{ij}(t) = I\{V_{ij} \leq \min(C_{ij}, t - W_{ij})\}$$
where \( I \{ E \} \) is the indicator function of event \( E \). \( I \{ E \} \) is 1 if the event \( E \) occurs and 0 otherwise. Let \( v_{ij} \) is the relative times (time measured from the real time of entry) for subject \( j \) in treatment \( i \).

Suppose we are monitoring futility by using RMST at both interim and final analysis in survival data. Let the interim stage is cut at study time \( t_I \) and final study time is \( t_F \), where \( t_F \geq t_I \). Then we have \( \hat{S}_i (t_I, v) \) and \( \hat{S}_i (t_F, v) \) be the KM estimator of the survival function in treatment \( i \) at relative time \( v \) for interim and final analyses respectively. Where \( i = 1 \) indicate the treatment arm, otherwise \( i = 2 \) represents the control.

Let \( v_1 < \ldots < v_{m_I} \) be the \( m_I \) distinct uncensored times at interim stage cut time \( t_I \), \( d_{i,r} (t_I) \) stands for the number of deaths at time \( v_r \) for treatment \( i \) and \( n_{i,r} (t_I) \) gives the number of individuals at risk at time \( v_r \) for treatment \( i \), that is, the number of individuals alive just prior to time \( v_r \). The KM estimator \( \hat{S}_i (t_I, v_{m_I}) = \prod_{r=1}^{r=m_I} \left( 1 - \frac{d_{i,r} (t_I)}{n_{i,r} (t_I)} \right) \). Similar to interim stage, \( d_{i,r} (t_F) \) and \( n_{i,r} (t_F) \) correspond to the same definitions but restricted to the final stage at study end time \( t_F \). The KM estimator at final stage for treatment \( i \) as \( \hat{S}_i (t_F, v_{m_F}) = \prod_{r=1}^{r=m_F} \left( 1 - \frac{d_{i,r} (t_F)}{n_{i,r} (t_F)} \right) \), where \( v_1 < \ldots < v_{m_F} \) be the \( m_F \) distinct uncensored times at final stage study time \( t_F \).

The restricted mean survival time \( \mu(B(t, v^*)) \) is the mean of the survival time \( B(t, v^*) = \min(X(t), v^*) \) limited to some horizon \( v^* > 0 \). Let the restricted time limit \( v^* = v_I \) for interim stage and \( v^* = v_F \) for final stage.

**9.1.2 Interim Stage with RMST**

Let \( B_{I,i} \) be the restricted survival time for treatment \( i \) at interim stage. Based on the equation [2.13], the natural estimator of RMST at interim for treatment \( i \) can be obtained as

\[
\hat{\mu} (B_{I,i}) = \hat{\mu} (B_i (t_I, v_I)) = \int_{v=0}^{v=v_I} \hat{S}_i (t_I, v) \, dv
\]

Based on the equation [2.16] the variance estimator at interim for treatment \( i \) is

\[
Var (\hat{\mu} (B_{I,i})) = \sum_{r=1}^{m_I} \left[ \int_{v=v_{r}}^{v=v_I} \hat{S}_i (t_I, v) \, dv \right]^2 \frac{d_{i,r} (t_I)}{n_{i,r} (t_I) (n_{i,r} (t_I) - d_{i,r} (t_I))}
\]
where \( v_1 < \ldots < v_{m_v} \) be the \( m_v \) distinct event times at interim stage restricted time limit \( v_I \), \( d_{i,r}(t_I) \) stands for the number of deaths at time \( v_r \) for treatment \( i \) and \( n_{i,r}(t_I) \) gives the number of individuals at risk at time \( v_r \) for treatment \( i \), that is, the number of individuals alive just prior to time \( v_r \). Then we have the estimator of RMST difference between two arms at interim stage \( \hat{\Delta}_I \) as

\[
\hat{\Delta}_I = \hat{\mu}(B_{I,1}) - \hat{\mu}(B_{I,2}) = \int_{v=0}^{v=v_I} \hat{S}_1(t_I, v) dv - \int_{v=0}^{v=v_I} \hat{S}_2(t_I, v) dv = \int_{v=0}^{v=v_I} \{ \hat{S}_1(t_I, v) - \hat{S}_2(t_I, v) \} dv \tag{9.1}
\]

Since two arms are independent, the standard deviation of RMST difference at interim stage \( \hat{\sigma}_I \) can be obtained as

\[
\hat{\sigma}_I = \sqrt{\sum_{i=1}^{2} \left( \sum_{r=1}^{m_{v_I}} \int_{v=v_r}^{v=v_I} \hat{S}_i(t_I, v) dv \right)^2 \frac{d_{i,r}(t_I)}{n_{i,r}(t_I)(n_{i,r}(t_I) - d_{i,r}(t_I))}} \tag{9.2}
\]

Finally, we get the test statistics at interim stage \( Z_I \) as

\[
Z_I = \frac{\hat{\Delta}_I}{\hat{\sigma}_I} = \frac{\int_{v=0}^{v=v_I} \{ \hat{S}_1(t_I, v) - \hat{S}_2(t_I, v) \} dv}{\sqrt{\sum_{i=1}^{2} \left( \sum_{r=1}^{m_{v_I}} \int_{v=v_r}^{v=v_I} \hat{S}_i(t_I, v) dv \right)^2 \frac{d_{i,r}(t_I)}{n_{i,r}(t_I)(n_{i,r}(t_I) - d_{i,r}(t_I))}}}
\]

### 9.1.3 Final Stage with RMST

Similar to interim stage, the natural estimator of RMST for treatment \( i \) at final can be obtained as

\[
\hat{\mu}(B_{F,i}) = \hat{\mu}(B_i(t_F, v_F)) = \int_{v=0}^{v=v_F} \hat{S}_i(t_F, v) dv
\]

Based on the equation 2.16 the variance estimator at final for treatment \( i \) is

\[
\text{Var} (\hat{\mu}(B_{F,i})) = \sum_{r=1}^{m_{v_F}} \left( \int_{v=v_r}^{v=v_F} \hat{S}_i(t_F, v) dv \right)^2 \frac{d_{i,r}(t_F)}{n_{i,r}(t_F)(n_{i,r}(t_F) - d_{i,r}(t_F))}
\]

where \( v_1 < \ldots < v_{m_{v_F}} \) be the \( m_{v_F} \) distinct event times at interim stage restricted time limit \( v_F \), \( d_{i,r}(t_F) \) stands for the number of deaths at time \( v_r \) for treatment \( i \) and \( n_{i,r}(t_F) \)
gives the number of individuals at risk at time $v_r$ for treatment $i$, that is, the number of individuals alive just prior to time $v_r$. Then we have the estimator of RMST difference between two arms at final stage $\hat{\Delta}_F$ is

$$\hat{\Delta}_F = \hat{\mu}(B_{F,1}) - \hat{\mu}(B_{F,2}) = \int_{v=0}^{v=v_F} \hat{S}_1(t_F, v) \, dv - \int_{v=0}^{v=v_F} \hat{S}_2(t_F, v) \, dv = \int_{v=0}^{v=v_F} \left\{ \hat{S}_1(t_F, v) - \hat{S}_2(t_F, v) \right\} \, dv \quad (9.3)$$

The standard deviation of RMST difference at final stage $\hat{\sigma}_F$ is

$$\hat{\sigma}_F = \sqrt{\sum_{i=1}^{2} \{ \sum_{r=1}^{m_{v_F}} \left[ \int_{v=v_r}^{v=v_F} \hat{S}_i(t_F, v) \, dv \right]^2 \frac{d_{i,r}(t_F)}{n_{i,r}(t_F)(n_{i,r}(t_F)-d_{i,r}(t_F))} \} } \quad (9.4)$$

Finally, we have the test statistics at final stage $Z_F$

$$Z_F = \frac{\hat{\Delta}_F}{\hat{\sigma}_F} = \frac{\int_{v=0}^{v=v_F} \left\{ \hat{S}_1(t_F, v) - \hat{S}_2(t_F, v) \right\} \, dv}{\sqrt{\sum_{i=1}^{2} \{ \sum_{r=1}^{m_{v_F}} \left[ \int_{v=v_r}^{v=v_F} \hat{S}_i(t_F, v) \, dv \right]^2 \frac{d_{i,r}(t_F)}{n_{i,r}(t_F)(n_{i,r}(t_F)-d_{i,r}(t_F))} \} }}$$

### 9.1.4 Conditional Power (CP)

CP is defined as the probability of rejecting $H_0$ at the end of the trial when total information is obtained, conditional on the partial information accumulated up to an interim point.

As we know, the CP can be written as

$$CP = P(Z_F > c|Z_I).$$

where $Z_I$ is given and $c$ is the critical value which usually set to 1.96 for one-sided test. The joint distribution of $Z_I$ and $Z_F$ approximately follows a bivariate normal distribution as

$$\begin{pmatrix} Z_I \\ Z_F \end{pmatrix} \sim AMN \left( \begin{pmatrix} \frac{\Delta_I}{\sigma_I} \\ \frac{\Delta_F}{\sigma_F} \end{pmatrix}, \begin{pmatrix} 1 & \rho_{IF} \\ \rho_{IF} & 1 \end{pmatrix} \right).$$
Then we have

\[ Z_F | Z_I \sim N \left( \frac{\Delta F}{\sigma_F} + \rho_{IF} \left( Z_I - \frac{\Delta I}{\sigma_I} \right), 1 - \rho_{IF}^2 \right) \]

where \( \rho_{IF} \) is correlation between \( Z_I \) and \( Z_F \). The conditional power can be derived by plugging the above mean and standard error part into the formula below

\[
P(Z_F > c | Z_I) = P \left( \frac{Z_F - \Delta F - \rho_{IF} \left( Z_I - \frac{\Delta I}{\sigma_I} \right)}{\sqrt{1 - \rho_{IF}^2}} > \frac{c - \Delta F - \rho_{IF} \left( Z_I - \frac{\Delta I}{\sigma_I} \right)}{\sqrt{1 - \rho_{IF}^2}} | Z_I \right) \\
= 1 - \Phi \left( \frac{c - \Delta F - \rho_{IF} \left( Z_I - \frac{\Delta I}{\sigma_I} \right)}{\sqrt{1 - \rho_{IF}^2}} \right) \\
= \Phi \left( \frac{\Delta F}{\sigma_F} + \rho_{IF} Z_I - \rho_{IF} \frac{\Delta I}{\sigma_I} - c}{\sqrt{1 - \rho_{IF}^2}} \right) \\
= \Phi \left( \frac{\Delta F}{\sigma_F} - \rho_{IF} \frac{\Delta I}{\sigma_I} - c}{\sqrt{1 - \rho_{IF}^2}} + \rho_{IF} Z_I \right).
\]

Finally, we obtain the general CP formula for RMST:

\[
P(Z_F > c | Z_I) = P \left( \frac{\Delta F - \rho_{IF} \frac{\Delta I}{\sigma_I} - c}{\sqrt{1 - \rho_{IF}^2}} + \frac{\rho_{IF}}{\sqrt{1 - \rho_{IF}^2}} Z_I \right). \tag{9.5}
\]

From above equation, we can see that the CP not only depends on the interim observed \( Z_I \), but also depends on the true treatment effects at both interim and final analyses as: \( \Delta F \) and \( \Delta I \), and the true correlation between interim and final \( \rho_{IF} \). Survival distributions for two arms usually will be assumed in the study design. We can easily generate the treatment effects for interim and final analysis based on known survival distributions. The correlation between interim and final also can be calculated by counting process approach which is mentioned in following section 9.2.1.

When we set the true treatment effect at interim stage \( \frac{\Delta I}{\sigma_I} = Z_I \), then we can simplify the CP equation as

\[
P(Z_F > c | Z_I) = \Phi \left( \frac{\Delta F - \rho_{IF} \frac{\Delta I}{\sigma_I} - c}{\sqrt{1 - \rho_{IF}^2}} + \frac{\rho_{IF}}{\sqrt{1 - \rho_{IF}^2}} Z_I \right) = \Phi \left( \frac{\Delta F - c}{\sqrt{1 - \rho_{IF}^2}} \right). \tag{9.6}
\]

Based on above equation, we can see that the simplified CP depends on two parts: one is
the true treatment effects at final stage $\frac{\Delta F}{\sigma_F}$, the other one is the true correlation between interim and final $\rho_{IF}$.

9.2 Calculating Correlation in RMST Difference Between Interim and Final Analysis

For calculating $\hat{\rho}_{IF}$, since the test statistics $Z_I$ and $Z_F$ under the null hypothesis follow a bivariate normal distribution as

$$\begin{pmatrix} Z_I \\ Z_F \end{pmatrix} | H_0 = MN \begin{pmatrix} 0 & 1 \\ 1 & \rho_{IF} \end{pmatrix}$$

implies

$$\begin{pmatrix} \hat{\Delta}_I - \Delta_I \\ \hat{\Delta}_F - \Delta_F \end{pmatrix} | H_0 = MN \begin{pmatrix} \sigma_I^2 & \sigma_{IF} \\ \sigma_{IF} & \sigma_F^2 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

where $\sigma_{IF}$ is covariance in RMST difference between interim and final analyses. The correlation can be obtained as

$$\rho_{IF} = \frac{\sigma_{IF}}{\sigma_I \sigma_F}, \quad (9.7)$$

When both interim and final data are observed, counting process approach can be used to estimate the variance-covariance matrix in RMST difference between interim and final analyses. Another alternative method Martingale integration approach was introduced to estimate the variance of RMST difference.

9.2.1 Counting Process Approach

Based on Lu and Tian’s paper [Lu and Tian 2020], we can estimate the variance-covariance matrix by using counting process. we note the expansion

$$\begin{pmatrix} \hat{\Delta}_I - \Delta_I \\ \hat{\Delta}_F - \Delta_F \end{pmatrix} = -2 \sum_{i=1}^{2} \frac{1}{n_i} \sum_{j=1}^{n_i} \left( \int_{0}^{v_I} \int_{P(X_{ij}^{(I)} \geq v; W_{ij} < t_i)}^{v_I} \int_{P(X_{ij}^{(F)} \geq v)} \int_{P(X_{ij}^{(F)} \geq v)}^{v_F} S_j(s) ds dM_{ij}^{(I)}(v) \right) + O_p(n^{-1/2}) \quad (9.8)$$
where $M_{ij}^{(I)}(v)$ and $M_{ij}^{(F)}(v)$ are the basic martingale for treatment group $i$ and subject $j$ at interim or final stage, respectively. We can have them as

$$M_{ij}^{(I)}(v) = I_{[X_{ij}^{(I)} \leq v, \delta_{ij}=1]} - \int_0^v h(s) I_{[X_{ij}^{(I)} \geq s]} ds$$

and

$$M_{ij}^{(F)}(v) = I_{[X_{ij}^{(F)} \leq v, \delta_{ij}=1]} - \int_0^v h(s) I_{[X_{ij}^{(F)} \geq s]} ds$$

where $h(s)$ is the hazard at time $s$.

The variance-covariance matrix can be estimated as

$$\sum_{i=1}^2 \sum_{j=1}^{n_i} \left( \int_0^{v_i} \int_0^{v_j} \frac{\hat{S}_i(s) ds}{\hat{P}(X_{ij}^{(I)} \geq v, W_{ij} \leq \tau_i)} \frac{d\hat{M}_{ij}^{(I)}(v)}{d\hat{M}_{ij}^{(I)}(v)} \right) \otimes 2$$

(9.9)

It can be estimated by integration directly with all required information. However, the other more consentient method is using Monte-Carlo simulation.

### 9.2.2 Martingale Integration Approach

The Martingale integration approach was introduced to estimate the variance of RMST in [Fleming and Harrington (1991)]. We can simplify the natural estimator for RMST in equation 2.13 as

$$\hat{\mu}(B(t,v^*)) = \int_{v=0}^{v=v^*} \hat{S}(t,v) dv \Rightarrow \hat{\mu}(B(v^*)) = \int_{v=0}^{v=v^*} \hat{S}(v) dv$$

For the right censored data, $\hat{F}_{n,i}(v)$ is the KM estimator of $F_i(v)$ for treatment $i$. We have

$$\hat{\mu}(B_i) = \hat{\mu}(B_i(v^*)) = \int_{v=0}^{v=v^*} [1 - \hat{F}_{n,i}(v)] dv$$

After standardization, we have

$$\sqrt{n_i} [\hat{\mu}(B_i) - \mu(B_i)] = \int_{v=0}^{v=v^*} \sqrt{n_i} \left[ F_i(v) - \hat{F}_{n,i}(v) \right] dv$$
\[- \int_{v=0}^{v=v^*} \sqrt{n_i} \frac{\hat{F}_{n,i}(v) - F_i(v)}{1 - F_i(v)} [1 - F_i(v)] dv \]

where \( n_i \) is total number of subjects for treatment \( i \). By the martingale CLT which was mentioned in Theorem 6.3.1 of Fleming and Harrington (1991), we know that

\[ \sqrt{n_i} \hat{F}_{n,i}(v) - F_i(v) \approx BM(C_i(v)) \]

where Brownian Motion (BM) is a stochastic process of normally distributed random variable. It is a martingale and has stationary with independent increments, and

\[ C_i(v) = \int_0^v \frac{dH_i(a)}{P_i(X \geq a)} = n_i \sum_{r=2}^{m_v} \hat{F}_{i}(v_r) - \hat{F}_{i}(v_{r-1}) \left( 1 - \hat{F}_{i}(v_{r-1}) \right) n_{i,r} \]

thus, the distribution of \( \sqrt{n_i} [\hat{\mu}(B_i) - \mu(B_i)] \) converges to the distribution of

\[- \int_{v=0}^{v=v^*} BM(C_i(v)) [1 - F_i(v)] dv \]

Based on BM’s properties, it is obvious that above integral have mean zero and are normally distributed. We now compute the covariance

\[ n_i \text{VAR}(\hat{\mu}(B_i) - \mu(B_i)) \]

\[ = n_i \text{COV}(\hat{\mu}(B_i) - \mu(B_i), \hat{\mu}(B_i) - \mu(B_i)) \]

\[ = E \int_0^{v^*} BM(C_i(v)) [1 - F_i(v)] dv \int_0^{v^*} BM(C_i(s)) [1 - F_i(s)] ds \]

\[ = E \int_0^{v^*} \int_0^{v^*} BM(C_i(v))BM(C_i(s)) [1 - F_i(v)] [1 - F_i(s)] dvds \]

\[ = \int_0^{v^*} \int_0^{v^*} E \{ BM(C_i(v))BM(C_i(s)) \} [1 - F_i(v)] [1 - F_i(s)] dvds \]

Since \( E \{ BM(A)BM(B) \} = \min(A, B) \) for Brownian Motion on the standard clock, we have

\[ E \{ BM(C_i(v))BM(C_i(s)) \} = \min(C_i(v), C_i(s)) \]
Thus

\[ n_i \text{VAR} (\hat{\mu}(B_i) - \mu(B_i)) = \int_0^{v^*} \int_0^{v^*} \text{min}(C_i(v), C_i(s)) [1 - F_i(v)] [1 - F_i(s)] dv ds \]

\[ = \int_0^{v^*} \int_0^{v^*} \text{min} \left( \int_0^v \frac{dH_i(a)}{P_i(X \geq a)}, \int_0^s \frac{dH_i(a)}{P_i(X \geq a)} \right) S_i(v) S_i(s) dv ds \]

since \( S_i(s) \) and \( S_i(v) \) are the same survival function \( S_i(t) \). After some calculation we obtain

\[ \text{Var}(\hat{\mu}(B_1) - \hat{\mu}(B_2)) = \sum_{i=1}^{2} \left\{ \frac{1}{n_i} \int_0^{v^*} \left[ \int_a^{v^*} S_i(v) dv \right]^2 \frac{dH_i(a)}{P_i(X \geq a)} \right\} \]  \hspace{1cm} (9.10)\]

Finally, the variance of RMST difference between two arms can be estimated as

\[ \hat{\sigma}^2 = \sum_{i=1}^{2} \left\{ \frac{1}{n_i} \int_0^{v^*} \left[ \int_a^{v^*} \hat{S}_i(v) dv \right]^2 \frac{d\hat{H}_i(a)}{\hat{P}_i(X \geq a)} \right\} \]  \hspace{1cm} (9.11)
Chapter 10
Asymptotic Covariance of RMST Difference in Two-Stage Design

Based on section 9, we know that covariance of RMST difference between interim and final analyses can be estimated by using counting process approach when both interim and final data are observed. In the study design, the variance-covariance matrix can be obtained either by direct integration from equation 9.9 with one simulation data or more conveniently by Monte-Carlo simulation based on assumed survival distribution for each arm. Either of them are complex and time-consuming. Furthermore, neither counting process approach or simulation is able to calculate the covariance with observed data at interim stage alone. In this chapter, we will introduce how to estimate the asymptotic covariance in RMST difference between interim and final analysis at interim stage by using the Martingale integration approach with predicted curves. The accuracy of the asymptotic covariance is evaluated by simulations.

10.1 Notation

Similar to Section 2.4, we assume the two-sample general random censoring model with staggered entry. Observations are made on \( n_i \) individuals from sample \( i \) which \( i = 1 \) indicate the treatment group, otherwise \( i = 2 \) represents the control. Let \((W_{ij}, V_{ij}, C_{ij}), j = 1, \ldots, n_i, i = 1, 2\), denote the real time of entry during the accrual period \([0, t_a]\), the real time from entry until failure, and the time from entry to censor for subject \( j \) in treatment \( i \) respectively. It assumes that \( W_{ij}, V_{ij}, C_{ij} \) are all nonnegative random variables and are independent to each other.
At a given real study time $t$, let

$$X_{ij}(t) = \max (\min(V_{ij}, C_{ij}, t - W_{ij}), 0)$$

and

$$\delta_{ij}(t) = I\{V_{ij} \leq \min(C_{ij}, t - W_{ij})\}$$

where $I\{E\}$ is the indicator function of event $E$. $I\{E\}$ is 1 if the event $E$ occurs and 0 otherwise. Let $v_{ij}$ is the relative times (time measured from the real time of entry) for subject $j$ in treatment $i$.

Suppose we are monitoring futility by using RMST at both interim and final analysis in survival data. Let the interim stage is cut at study time $t_I$ and final study time is $t_F$, where $t_F \geq t_I$. Then we have $\hat{S}_i(t_I, v)$ and $\hat{S}_i(t_F, v)$ be the KM estimator of the survival function of treatment $i$ at relative time $v$ for interim and final analyses, respectively. Where $i = 1$ indicate treatment group, otherwise $i = 2$ represents control.

In interim analysis which cut at $t_I$ and let $v_1 < \ldots < v_{m_{t_I}}$ be the $m_{t_I}$ distinct uncensored times at interim stage study time $t_I$, $d_{i,r}(t_I)$ stands for the number of deaths at time $v_r$ for treatment $i$ and $n_{i,r}(t_I)$ gives the number of individuals at risk at time $v_r$ for treatment $i$, that is, the number of individuals alive just prior to time $v_r$. The Kaplan-Meier estimator $\hat{S}_i(t_I, v_{m_{t_I}}) = \prod_{r=1}^{r=m_{t_I}} \left(1 - \frac{d_{i,r}(t_I)}{n_{i,r}(t_I)}\right)$. Similar to interim analysis, $d_{i,r}(t_F)$ and $n_{i,r}(t_F)$ correspond to the same definitions but restricted to the final stage at study end time $t_F$. The Kaplan-Meier estimator at final analysis for treatment $i$ as $\hat{S}_i(t_F, v_{m_{t_F}}) = \prod_{r=1}^{r=m_{t_F}} \left(1 - \frac{d_{i,r}(t_F)}{n_{i,r}(t_F)}\right)$, where $v_1 < \ldots < v_{m_{t_F}}$ are the $m_{t_F}$ distinct uncensored times at final stage study time $t_F$.

The restricted mean survival time $\mu(B(t,v^*))$ is the mean of the survival time $B(t,v^*) = \min(X(t), v^*)$ limited to some horizon $v^* > 0$. Let real study time $t = t_I$ at interim stage and $t = t_F$ at final stage. In two-stage analyses, we also set the restricted time limit $v^* = v_I$ for interim stage and $v^* = v_F$ for final stage.
10.2 Asymptotic Covariance by Martingale Integration Approach

As we known that, the covariance of RMST difference between interim and final stages \( \sigma_{IF} \) can be written as:

\[
\sigma_{IF} = \text{COV} \left( \Delta_I - \Delta_I \right)
\]

\[
= \text{COV} \left( (\hat{\mu}(B_{I,1}) - \mu(B_{I,1})) - (\mu(B_{I,1}) - \mu(B_{I,2})),
(\hat{\mu}(B_{F,1}) - \mu(B_{F,2})) - (\mu(B_{F,1}) - \mu(B_{F,2})) \right)
\]

\[
= \sum_{i=1}^{2} \{ \text{COV} (\hat{\mu}(B_{I,i}) - \mu(B_{I,i}), \hat{\mu}(B_{F,i}) - \mu(B_{F,i})) \}
\]

Since two arms are independent, then

\[
\sigma_{IF} = \text{COV} (\hat{\mu}(B_{I,1}) - \mu(B_{I,1}), \hat{\mu}(B_{F,1}) - \mu(B_{F,1}))
+ \text{COV} (\hat{\mu}(B_{I,2}) - \mu(B_{I,2}), \hat{\mu}(B_{F,2}) - \mu(B_{F,2}))
\]

\[
= \sum_{i=1}^{2} \{ \text{COV} (\hat{\mu}(B_{I,i}) - \mu(B_{I,i}), \hat{\mu}(B_{F,i}) - \mu(B_{F,i})) \}
\]

Similar to estimating variance of RMST by using martingale integration approach in section 9.2.2, we can simplify the natural estimator for RMST in equation 2.13 as

\[
\hat{\mu}(B(t,v^*)) = \int_{v=0}^{v=v^*} \hat{S}(t,v) \, dv \Rightarrow \hat{\mu}(B(v^*)) = \int_{v=0}^{v=v^*} \hat{S}(v) \, dv
\]

Let the RMST cut time \( v^* = v_I \) in interim analysis and \( v^* = v_F \) in final analysis. \( \hat{F}_{n,I}(v) \) is the KM estimator for a unknown lifetime distribution \( F_I(v) \) in interim analysis and \( \hat{F}_{n,F}(v) \) is the KM estimator for the \( F_F(v) \) in final analysis. Then we have

\[
\hat{\mu}(B_{I,i}) = \hat{\mu}(B_{I,i}(v_I)) = \int_{v=0}^{v=v_I} [1 - \hat{F}_{n,I,i}(v)] \, dv
\]

Same with \( \hat{\mu}(B_{F,i}) \) in final analysis

\[
\hat{\mu}(B_{F,i}) = \int_{s=0}^{s=v_F} [1 - \hat{F}_{n,F,i}(s)] \, ds
\]
Since \((W_{ij}, V_{ij}, C_{ij}), j = 1, \ldots, n_i, i = 1, 2\), denote the real time of entry during the accrual period \([0, t_a]\), the real time from entry until failure, and the time from entry to censor for subject \(j\) in treatment \(i\) respectively. It assumes that \(W_{ij}\), \(V_{ij}\), \(C_{ij}\) are all nonnegative random variables and are independent to each other, then we have the cumulative hazard function in interim analysis

\[
\Lambda_{I,i}(v) = \int_0^v \frac{dH_{I,i}^1(a)}{P_{I,i}(X \geq a, W < t_I)} \\
= \int_0^v \frac{dP_I(\min(V, C, t_I - W) \leq a, \delta = 1, W < t_I)}{P_I(\min(V, C, t_I - W) \geq a, W < t_I)} \\
= \int_0^v \frac{d \int_0^a P_I(C \geq x) P_I(t_I - W \geq x, W < t_I) d(P_I(V \leq x))}{P_I(V \geq a) P_I(C \geq a) P_I(t_I - W \geq a, W < t_I)} \\
= \int_0^v \frac{P_I(C \geq a) P_I(t_I - W \geq a, W < t_I) d(P_I(V \leq a))}{P_I(V \geq a)}
\]

Similarly, for final analysis, we have

\[
\Lambda_{F,i}(v) = \int_0^v \frac{dH_{F,i}^1(a)}{P_{F,i}(X \geq a)} \\
= \int_0^v \frac{dP_I(\min(V, C, t_F - W) \leq a, \delta = 1)}{P_I(\min(V, C, t_F - W) \geq a)} \\
= \int_0^v \frac{d \int_0^a P_I(C \geq x) P_I(t_F - W \geq x) d(P_I(V \leq x))}{P_I(V \geq a) P_I(C \geq a) P_I(t_F - W \geq a)} \\
= \int_0^v \frac{P_I(C \geq a) P_I(t_F - W \geq a) d(P_I(V \leq a))}{P_I(V \geq a)}
\]

Thus,

\[
\Lambda_{I,i}(v) = \Lambda_{F,i}(v) \quad \text{for } 0 \leq v \leq t_I
\]

It implies \(F_{I,i}(v) = F_{F,i}(v)\) and \(S_{I,i}(v) = S_{F,i}(v)\) for \(0 \leq v \leq t_I\). Similar to section 9.2.2 after standardization the RMST difference equation for interim analysis, we have

\[
\sqrt{\bar{n}_i} [\mu(B_{I,i}) - \mu(B_{F,i})] = \int_{v=0}^{v=t_I} \sqrt{\bar{n}_i} \left[ F_{I,i}(v) - \hat{F}_{n,I,i}(v) \right] dv
\]
where \( n_i \) is total number of subjects for treatment \( i \). By the martingale CLT, we know that

\[
\sqrt{n_i} \left( \hat{F}_{n,I,i}(v) - F_{I,i}(v) \right) \approx BM \left( C_{I,i}(v) \right)
\]

where Brownian Motion (BM) is a stochastic process and \( C_{I,i}(v) \) is a function of \( v \). We have

\[
C_{I,i}(v) = \int_0^v \frac{dH_{I,j}(a)}{P_{I,j}(X \geq a, W < t_I)} = \int_0^v \frac{dH_i(a)}{P_i(V \geq a) P_i(C \geq a) P_i(t_I - W \geq a, W < t_I)}
\]

where \( H_i(a) \) is the cumulative hazard at time \( a \) for treatment \( i \) and we have \( H_i(a) = H_{I,j}(a) = H_{F,i}(a) \). For fixed \( C_{I,i}(v) \), \( BM \left( C_{I,i}(v) \right) \) just a normal random variable with mean 0 and variance equals to \( C_{I,i}(v) \).

Similarly, for final analysis,

\[
\sqrt{n_i} \left( \hat{F}_{n,F,i}(s) - F_{F,i}(s) \right) \approx BM \left( C_{F,i}(s) \right)
\]

where \( C_{F,i}(s) \) is

\[
C_{F,i}(s) = \int_0^s \frac{dH_{F,i}(a)}{P_{F,i}(X \geq a)} = \int_0^s \frac{dH_i(a)}{P_i(V \geq a) P_i(C \geq a) P_i(t_F - W \geq a)}
\]

The distribution of \( \sqrt{n_i} [\hat{\mu}(B_{I,i}) - \mu(B_{I,i})] \) converge to the distribution of

\[
- \int_{v=0}^{v=v_I} BM(C_{I,i}(v)) \left[ 1 - F_{I,i}(v) \right] dv
\]

The same for \( \sqrt{n_i} [\hat{\mu}(B_{F,i}) - \mu(B_{F,i})] \), it converges to the distribution of

\[
- \int_{s=0}^{s=v_F} BM(C_{F,i}(s)) \left[ 1 - F_{F,i}(s) \right] ds.
\]
we can compute the covariance

\[
\text{COV} (\hat{\mu}(B_{I,i}) - \mu(B_{I,i}), \hat{\mu}(B_{F,i}) - \mu(B_{F,i}))
\]

\[
= \frac{1}{n_i} \int_0^{v_f} \int_0^{v_f} \text{BM}(C_{I,i}(v)) [1 - F_{I,i}(v)] \text{BM}(C_{F,i}(s)) [1 - F_{F,i}(s)] ds dv
\]

For data without staggered entry, since \(P_i(t_I - W \geq a, W < t_I) = P_i(t_F - W \geq a) = 1\) for all \(0 \leq a \leq t_I\), we have \(C_{I,i}(v) = C_{F,i}(v)\) for all \(0 \leq v \leq t_I\). We know that \(E\{\text{BM}(A)\text{BM}(B)\} = \min(A,B)\) for Brownian Motion (Fleming and Harrington, 1991)

Then let \(s \leq v\), we have

\[
E\{\text{BM}(C_{I,i}(v))\text{BM}(C_{F,i}(s))\}
\]

\[
= \text{COV}(\text{BM}(C_{F,i}(v)), \text{BM}(C_{F,i}(s)))
\]

\[
= \text{COV}(\text{BM}(C_{F,i}(s)) + \text{BM}(C_{F,i}(v)) - \text{BM}(C_{F,i}(s)), \text{BM}(C_{F,i}(s)))
\]

\[
= \text{COV}(\text{BM}(C_{F,i}(s)), \text{BM}(C_{F,i}(s))) + \text{COV}(\text{BM}(C_{F,i}(v)) - \text{BM}(C_{F,i}(s)), \text{BM}(C_{F,i}(s)))
\]

\[
= \text{VAR}(\text{BM}(C_{F,i}(s)))
\]

\[
= C_{F,i}(s)
\]

Thus,

\[
E\{\text{BM}(C_{I,i}(v))\text{BM}(C_{F,i}(s))\} = \min(C_{I,i}(v), C_{F,i}(s)) = \min(C_{F,i}(v), C_{F,i}(s))
\]

For data with staggered entry, since \(t_I \leq t_F\), we have

\[
C_{F,i}(v) = \int_0^v \frac{dH_{F,i}(a)}{P_{F,i}(X \geq a)}
\]

\[
= \int_0^v \frac{dH_i(a)}{P_i(V \geq a) P_i(C \geq a) P_i(t_F - W \geq a)}
\]
\[
= \int_0^v \left( \frac{P_i (t_I - W \geq a, W < t_I)}{P_i (t_F - W \geq a)} \right) \frac{dH_i (a)}{P_i (V \geq a)} \frac{dH_i (a)}{P_i (C \geq a)} \frac{dH_i (a)}{P_i (t_I - W \geq a, W < t_I)}
\]

let

\[
C_{ItoF,i} (v) = C_{I,i} (v) - C_{F,i} (v) = \int_0^v (1 - \frac{P_i (t_I - W \geq a, W < t_I)}{P_i (t_F - W \geq a)}) \frac{dH_i (a)}{P_i (X \geq a, W < t_I)}
\]

We know, \(BM(C_{I,i} (v)) = BM(C_{F,i} (v)) + BM(C_{ItoF,i} (v))\), then we have

\[
Var(BM(C_{I,i} (v)))
\]

\[
= C_{I,i} (v) = Var(BM(C_{F,i} (v)) + BM(C_{ItoF,i} (v)))
\]

\[
= Var(BM(C_{F,i} (v))) + Var(BM(C_{ItoF,i} (v))) + 2COV(BM(C_{F,i} (v)), BM(C_{ItoF,i} (v)))
\]

\[
= C_{F,i} (v) + C_{ItoF,i} (v) + 2COV(BM(C_{F,i} (v)), BM(C_{ItoF,i} (v)))
\]

\[
= C_{I,i} (v) + 2COV(BM(C_{F,i} (v)), BM(C_{ItoF,i} (v)))
\]

it implies \(COV(BM(C_{F,i} (v)), BM(C_{ItoF,i} (v))) = 0\), thus \(BM(C_{F,i} (v))\) and \(BM(C_{ItoF,i} (v))\) are independent. By Brownian Motion properties and let \(s \leq v\), then we have

\[
E \{BM(C_{I,i} (v)) BM(C_{F,i} (s))\}
\]

\[
= COV(BM(C_{I,i} (s)) + BM(C_{I,i} (v)) - BM(C_{I,i} (s)), BM(C_{F,i} (s))
\]

\[
= COV(BM(C_{I,i} (s)), BM(C_{F,i} (s))) + COV(BM(C_{I,i} (v)) - BM(C_{I,i} (s)), BM(C_{F,i} (s)))
\]

\[
= COV(BM(C_{I,i} (s)), BM(C_{F,i} (s)))
\]

\[
= COV(BM(C_{F,i} (s)) + BM(C_{ItoF,i} (s)), C_{F,i} (s))
\]

\[
= COV(BM(C_{F,i} (s)), BM(C_{ItoF,i} (s)))
\]

\[
= VAR(BM(C_{F,i} (s)))
\]

\[
= C_{F,i} (s)
\]

it also implies \(COV(BM(C_{F,i} (s)), BM(C_{ItoF,i} (v))) = 0\). Similarly, we can get
COV(BM(C_{F,i}(v)), BM(C_{ItoF,i}(s))) = 0. Then for E\{BM(C_{I,i}(s))BM(C_{F,i}(v))\}, We have

\[ E\{BM(C_{I,i}(s))BM(C_{F,i}(v))\} \]
\[ = COV(BM(C_{F,i}(s)), BM(C_{ItoF,i}(s)), C_{F,i}(v)) \]
\[ = COV(BM(C_{F,i}(s)), BM(C_{F,i}(v)) + COV(BM(C_{ItoF,i}(s)), BM(C_{F,i}(v))) \]
\[ = E(BM(C_{F,i}(s)), BM(C_{F,i}(v))) \]
\[ = C_{F,i}(s) \]

it implies, for any v and s, we have

\[ E\{BM(C_{I,i}(v))BM(C_{F,i}(s))\} = \min(C_{F,i}(v), C_{F,i}(s)) \]

Thus,

\[
COV(\hat{\mu}(B_{I,i}) - \mu(B_{I,i}), \hat{\mu}(B_{F,i}) - \mu(B_{F,i})) \\
= \frac{1}{n_i} \int_0^{v_F} \int_0^{v_I} \min(C_{F,i}(v), C_{F,i}(s)) [1 - F_{I,i}(v)][1 - F_{F,i}(s)]dvds \\
= \frac{1}{n_i} \int_0^{v_F} \int_0^{v_I} \min(\int_0^{v} \frac{dH_i(a)}{P_{F,i}(X \geq a)}, \int_0^{s} \frac{dH_i(a)}{P_{F,i}(X \geq a)})S_{I,i}(v)S_{F,i}(s)dvds \\

we have

\[ \sigma_{IF} = \sum_{i=1}^{2} \left\{ \frac{1}{n_i} \int_0^{v_F} \int_0^{v_I} \min(\int_0^{v} \frac{dH_i(a)}{P_{F,i}(X \geq a)}, \int_0^{s} \frac{dH_i(a)}{P_{F,i}(X \geq a)})S_{I,i}(v)S_{F,i}(s)dvds \right\} \]

\[ (10.1) \]

we also can write it as follow:

\[ \sigma_{IF} = \sum_{i=1}^{2} \left\{ \frac{1}{n_i} \int_0^{v_F} \int_0^{v_I} S_{I,i}(v)S_{F,i}(s) \int_{\min(v,s)}^{\min(v,s)} \frac{dH_i(a)}{P_{F,i}(X \geq a)}dvds \right\} \]

The formula is consistent with the expression which was mentioned in Murray and Tsiatis
\[ \sigma_{IF} = \sum_{i=1}^{2} \left\{ \frac{1}{n_i} \int_0^{v_2} \int_0^{v_1} \min \left( \int_0^{v} \frac{dH_{I,i}(a)}{P_{F,i}(X \geq a)} , \int_0^{s} \frac{dH_{F,i}(a)}{P_{F,i}(X \geq a)} \right) S_{I,i}(v) S_{F,i}(s) dvds \right\} \]

where \( n_i(t_2) \) is the sample size in treatment group \( i \) at study time \( t_2 \), \( S_i \) is the KM survival function for treatment group \( i \). \( H_i(x) \) as the probability of remaining uncensored at time \( x \) for treatment group \( i \). \( \lambda_i(t) \) is the hazard for treatment group \( i \) at time \( t \). \( \tau_1 \) and \( \tau_2 \) are restricted time limit for interim and final analyses, respectively. Finally, the covariance can be consistently estimated as

\[ \hat{\sigma}_{IF} = \sum_{i=1}^{2} \left\{ \frac{1}{n_i} \int_0^{v_F} \int_0^{v_I} \min \left( \int_0^{v} \frac{dH_{I,i}(a)}{P_{F,i}(X \geq a)} , \int_0^{s} \frac{dH_{F,i}(a)}{P_{F,i}(X \geq a)} \right) \hat{S}_{I,i}(v) \hat{S}_{F,i}(s) dvds \right\} \]

(10.2)

Since \( S_{I,i}(v) = S_{F,i}(v) \) for \( 0 \leq v \leq v_I \), we can use the \( S_{I,i}(v) \) to replace the \( S_{F,i}(v) \) for \( 0 \leq v \leq v_I \). Then the equation [10.1] can be simplified as:

\[ \sigma_{IF} = \sum_{i=1}^{2} \left\{ \frac{1}{n_i} \int_0^{v_F} \int_0^{v_I} \min \left( \int_0^{v} \frac{dH_{I,i}(a)}{P_{F,i}(X \geq a)} , \int_0^{s} \frac{dH_{I,i}(a)}{P_{F,i}(X \geq a)} \right) S_{I,i}(v) S_{F,i}(s) dvds \right\} \]

\[ = \sum_{i=1}^{2} \frac{1}{n_i} \left\{ \int_0^{v_I} \int_0^{v_I} \min \left( \int_0^{v} \frac{dH_{I,i}(a)}{P_{F,i}(X \geq a)} , \int_0^{s} \frac{dH_{I,i}(a)}{P_{F,i}(X \geq a)} \right) S_{I,i}(v) S_{F,i}(s) dvds \right\} \]

\[ + \sum_{i=1}^{2} \frac{1}{n_i} \left\{ \int_0^{v_F} \int_0^{v_I} \min \left( \int_0^{v} \frac{dH_{I,i}(a)}{P_{F,i}(X \geq a)} , \int_0^{s} \frac{dH_{I,i}(a)}{P_{F,i}(X \geq a)} \right) S_{I,i}(v) S_{F,i}(s) dvds \right\} \]

\[ = \sum_{i=1}^{2} \frac{1}{n_i} \left\{ \int_0^{v_I} \int_0^{v_I} \min \left( \int_0^{v} \frac{dH_{I,i}(a)}{P_{F,i}(X \geq a)} , \int_0^{s} \frac{dH_{I,i}(a)}{P_{F,i}(X \geq a)} \right) S_{I,i}(v) S_{F,i}(s) dvds \right\} \]

\[ + \sum_{i=1}^{2} \frac{1}{n_i} \left\{ \int_0^{v_F} \int_0^{v_I} \min \left( \int_0^{v} \frac{dH_{I,i}(a)}{P_{F,i}(X \geq a)} , \int_0^{s} \frac{dH_{I,i}(a)}{P_{F,i}(X \geq a)} \right) S_{I,i}(v) S_{F,i}(s) dvds \right\} \]

\[ = \sum_{i=1}^{2} \frac{1}{n_i} \left\{ n_i \sigma_{F,i}^2 + \int_0^{v_I} \int_0^{v_I} \min \left( \int_0^{v} \frac{dH_{I,i}(a)}{P_{F,i}(X \geq a)} , \int_0^{s} \frac{dH_{I,i}(a)}{P_{F,i}(X \geq a)} \right) S_{I,i}(v) S_{F,i}(s) dvds \right\} \]

\[ = \sum_{i=1}^{2} \frac{1}{n_i} \left\{ n_i \sigma_{F,i}^2 + \int_0^{v_I} \int_0^{v_I} \left[ \int_0^{v} \frac{dH_{I,i}(a)}{P_{F,i}(X \geq a)} \right] S_{I,i}(v) S_{F,i}(s) dvds \right\} \]

\[ = \sum_{i=1}^{2} \frac{1}{n_i} \left\{ n_i \sigma_{F,i}^2 + \int_0^{v_I} \int_0^{v_I} \left[ \int_0^{v} \frac{dH_{I,i}(a)}{P_{F,i}(X \geq a)} \right] S_{I,i}(v) S_{F,i}(s) dvds \right\} \]

\[ = \sum_{i=1}^{2} \frac{1}{n_i} \left\{ n_i \sigma_{F,i}^2 + \int_0^{v_I} \left[ \int_0^{v} \frac{dH_{I,i}(a)}{P_{F,i}(X \geq a)} \right] S_{I,i}(v) S_{F,i}(s) dv \right\} \]
\[= \sum_{i=1}^{2} \frac{1}{n_i} \left\{ n_i \sigma_{F,i}^2 + \int_{0}^{\nu_I} \left[ \int_{0}^{\nu} \frac{dH_i(a)}{P_{F,i}(X \geq a)} \right] S_{I,i}(v)dv \left( \mu(B_{F,i}) - \mu(B_{I,i}) \right) \right\} \]

Thus, we have

\[\sigma_{IF} = \sum_{i=1}^{2} \frac{1}{n_i} \left\{ n_i \sigma_{F,i}^2 + \int_{0}^{\nu_I} \left[ \int_{0}^{\nu} \frac{dH_i(a)}{P_{F,i}(X \geq a)} \right] S_{I,i}(v)dv \left( \mu(B_{F,i}) - \mu(B_{I,i}) \right) \right\} \tag{10.3} \]

where \(B_{I,i}\) and \(B_{F,i}\) is restricted survival time in interim analysis and final analysis for treatment \(i\), respectively. The covariance can also be estimated consistently as

\[\hat{\sigma}_{IF} = \sum_{i=1}^{2} \frac{1}{n_i} \left\{ n_i \sigma_{F,i}^2 + \int_{0}^{\nu_I} \left[ \int_{0}^{\nu} \frac{\hat{d}H_i(a)}{P_{F,i}(X \geq a)} \right] \hat{S}_{I,i}(v)dv \left( \hat{\mu}(B_{F,i}) - \hat{\mu}(B_{I,i}) \right) \right\} \tag{10.4} \]

Based on the simplified equation \(10.3\), we found that the covariance is related to three parts:

1. Part I: \(\sigma_{F,i}^2\) is related to the variance of RSMT from time 0 to \(\nu_I\) in final survival curve. It can be calculated by data at interim stage for treatment \(i\) as

\[\sigma_{F,i}^2 = \frac{1}{n_i} \int_{0}^{\nu_I} \int_{0}^{v} \min \left( \int_{0}^{v} \frac{dH_i(a)}{P_{F,i}(X \geq a)} \right) S_{F,i}(v)S_{F,i}(s)dvds\]

\[= \frac{1}{n_i} \int_{0}^{\nu_I} \left[ \int_{a}^{v} S_{F,i}(v)dv \right]^2 \frac{dH_i(a)}{P_{F,i}(X \geq a)}\]

\[= \frac{1}{n_i} \int_{0}^{\nu_I} \left[ \int_{a}^{v} S_{F,i}(v)dv \right]^2 \left( \frac{P_{I}(t_{I} - W \geq a, W < t_I)}{P_{I}(t_{I} - W \geq a)} \right) \frac{dH_i(a)}{P_{I,i}(X \geq a)}\]

2. Part II: \(\int_{0}^{\nu_I} \left[ \int_{0}^{v} \frac{dH_i(a)}{P_{F,i}(X \geq a)} \right] S_{I,i}(v)dv\) relates to the survival curve from time 0 to \(\nu_I\) for treatment \(i\). It can be calculated by data at interim stage as well as

\[\int_{0}^{\nu_I} \left[ \int_{0}^{v} \frac{dH_i(a)}{P_{F,i}(X \geq a)} \right] S_{I,i}(v)dv\]

\[= \int_{0}^{\nu_I} \left[ \int_{0}^{v} \left( \frac{P_{I}(t_{I} - W \geq a, W < t_I)}{P_{I}(t_{I} - W \geq a)} \right) \frac{dH_i(a)}{P_{I,i}(X \geq a)} \right] S_{I,i}(v)dv\]

3. Part III: \(\mu(B_{F,i}) - \mu(B_{I,i})\) is the difference in RMST between interim RMST cut time \(v_I\) and final RMST cut time \(v_F\) for treatment \(i\).

The above section points out that the covariance of RMST difference between interim and
final stages $\sigma_{IF}$ not only depends on the variance of RMST (Part I) and survival curve distribution (Part II) from time 0 to $v_I$, but also depends on the difference in RMST between interim cut time $v_I$ and final cut time $v_F$ (Part III) for two arms. Regarding to the survival distribution between $v_I$ and $v_F$, The covariance seems only depends on the area under the curve and nothing else. Since the covariance is related to three parts, $\sigma^2_{F,i}$ and $\int_0^{v_I} \left[ \int_0^v \frac{dH_i(a)}{PF_i(X \geq a)} \right] S_{I,i}(v) dv$ can be estimated by using the observed data at interim stage, when we assumed survival curve at interim stage is the same with the survival curve at final stage which is $\hat{S}_{I,i}(v) = \hat{S}_{F,i}(v)$ for $0 \leq v \leq v_I$. Regarding to the RMST difference between interim and final analysis for two arms, we may get them under two scenarios. One scenario is directly using the RMST at final analysis for control arm with given RMST difference at final analysis $\Delta_F$. Those parameters usually are given as assumptions in the study design. The other one is using the $\Delta_F$ with a known survival distribution for control group. Either of two scenarios provides enough information for part III. Once we had all three parts, the covariance can be calculated by using either equation [10.2] or equation [10.4] with the observed data at interim stage.

10.2.1 Calculation of Asymptotic covariance by given $\Delta_F$ and RMST at final analysis for control

Commonly, final analysis RMST will be given as an assumption in study design for either treatment or control groups. When both of them $\mu(B_{F,1})$ and $\mu(B_{F,2})$ are given, then the estimator of RMST difference between $v_I$ and $v_F$ in treatment group $\hat{\mu}(B_{F,1}) - \hat{\mu}(B_{I,1})$ can be calculated as follow

$$\hat{\mu}(B_{F,1}) - \hat{\mu}(B_{I,1}) = \mu(B_{F,1}) - \hat{\mu}(B_{I,1})$$

Similarly in control group,

$$\hat{\mu}(B_{F,2}) - \hat{\mu}(B_{I,2}) = \mu(B_{F,2}) - \hat{\mu}(B_{I,2})$$

If known $\mu(B_{F,2})$ and the RMST difference between treatment and control in final
analysis $\Delta F$, then \( \hat{\mu}(B_{F,1}) - \hat{\mu}(B_{I,1}) \) can be calculated as

\[
\hat{\mu}(B_{F,1}) - \hat{\mu}(B_{I,1}) = \Delta_F + \mu(B_{F,2}) - \hat{\mu}(B_{I,1})
\]

For control group,

\[
\hat{\mu}(B_{F,2}) - \hat{\mu}(B_{I,2}) = \mu(B_{F,2}) - \hat{\mu}(B_{I,2})
\]

Once we have Part III \( \hat{\mu}(B_{F,i}) - \hat{\mu}(B_{I,i}) \) for either \( i = 1 \) for treatment group or \( i = 2 \) for control, then the covariance can be computed by the formula in equation 10.4.

Instead of using formula in equation 10.4, the formula in equation 10.2 is another optional formula we can use to calculate the asymptotic covariance. Predicted survival curves with fixed final analysis RMST were developed in order to compute the covariance by using the formula 10.2. Since we have already known that \( S_{I,i}(v) = S_{F,i}(v) \) for \( 0 \leq v \leq t_I \), the distribution of predicted final survival curve from 0 to \( v_I \) is assumed to be identical with the observed curve at interim stage. Furthermore, because of covariance seems doesn’t depend on the survival curve distribution after interim RMST cut time \( v_I \) but the area under the curve only, the simple way to create the predicted curve is to set \( \hat{S}_{I,ItoF}(v) = \hat{S}_{I,i}(v_I)S_i(v - v_I) \) which follows an exponential distribution as \( S_i(v) = \exp(-\lambda_i v) \) from \( v_I \) to \( v_F \) in predicted curve for each arm. Then, we have the estimator of RMST at final analysis based on the predicted survival curve as

\[
\hat{\mu}(B_{F,i}) = \hat{\mu}(B_i(t_F,v_F)) = \int_{v=0}^{v=v_I} \hat{S}_i(t_F,v) \, dv \\
= \int_{v=0}^{v=v_I} \hat{S}_{I,i}(v) \, dv + \int_{v=v_I}^{v=v_F} \hat{S}_{I,ItoF}(v) \, dv \\
= \int_{v=0}^{v=v_I} \hat{S}_{I,i}(v) \, dv + \int_{v=v_I}^{v=v_F} [\hat{S}_{I,i}(v_I)S_i(v - v_I)] \, dv \\
= \int_{v=0}^{v=v_I} \hat{S}_{I,i}(v) \, dv + \hat{S}_{I,i}(v_I) \int_{v=v_I}^{v=v_F} S_i(v - v_I) \, dv \\
= \int_{v=0}^{v=v_I} \hat{S}_{I,i}(v) \, dv + \hat{S}_{I,i}(v_I) \int_{v=v_I}^{v=v_F} \exp(-\lambda_i(v - v_I)) \, dv \quad (10.5)
\]
Where \(i = 1\) for the treatment and \(i = 2\) for the control. Based on the predicted survival curve, the variance of RMST at final analysis can be estimated by following equation 2.14 as:

\[
\text{Var} \left( \hat{\mu} (B_{F,i}) \right) = \frac{1}{n_{F,i}} \text{Var} \left( B_i (t_F, v_F) \right)
\]

\[
= \frac{1}{n_{F,i}} \left\{ E \left( B_i (t_F, v_F)^2 \right) - \left[ E (B_i (t_F, v_F)) \right]^2 \right\}
\]

\[
= \frac{1}{n_{F,i}} \left\{ 2 \int_{v=0}^{v_I} v \hat{S}_{I,i} (v) \, dv + 2 \int_{v=0}^{v_F} v \hat{S}_{I,i} (v_I) \hat{S}_i (v - v_I) \, dv - \left[ \int_{v=0}^{v_I} \hat{S}_{I,i} (v) \, dv \right]^2 \right\}
\]

\[
= \frac{1}{n_{F,i}} \left\{ 2 \int_{v=0}^{v_I} v \hat{S}_{I,i} (v) \, dv + 2 \hat{S}_{I,i} (v_I) \int_{v=0}^{v_F} v \exp (-\lambda_i (v - v_I)) \, dv
\]

\[
- \left[ \int_{v=0}^{v_I} \hat{S}_{I,i} (v) \, dv + \hat{S}_{I,i} (v_I) \int_{v=0}^{v_F} \exp (-\lambda_i (v - v_I)) \, dv \right]^2 \right\} \quad (10.6)
\]

Once we have predicted survival curves for two treatment groups, we can use equation 10.2 to calculate the covariance.

### 10.2.2 Calculation of Asymptotic Covariance by Given \(\Delta_F\) and with Known Survival Distribution for Control Group

In some situations, we may know \(\Delta_F\) alone without \(\mu (B_{F,1})\) and \(\mu (B_{F,2})\) from study design. In this situation, the survival distribution from previous reference literatures will be used as the survival distribution in the current study for the control group. Once we had the survival distribution for the control, the covariance can be estimated by using either equation 10.2 or equation 10.4 with a given \(\Delta_F\).

Let the distribution for the control group is known as \(S_2 (v)\) during whole study time, then we have the survival function from interim analysis cut \(v_I\) to \(v_F\) as \(\hat{S}_{2,ItoF} (v)\) which \(\hat{S}_{2,ItoF} (v) = \hat{S}_{I,2} (v_I) S_2 (v - v_I)\). then the estimator of RMST difference between \(v_I\) and \(v_F\) in control group \(\hat{\mu} (B_{F,2}) - \hat{\mu} (B_{I,2})\) can be calculated as

\[
\hat{\mu} (B_{F,2}) - \hat{\mu} (B_{I,2}) = \hat{S}_{I,2} (v_I) \int_{v=v_I}^{v=v_F} S_2 (v - v_I) \, dv
\]
For treatment group, when $\Delta_F$ is given, we have

$$\hat{\mu}(B_{F,1}) - \hat{\mu}(B_{I,1}) = \Delta_F + \hat{\mu}(B_{I,2}) + \hat{S}_{I,2}(v_I) \int_{v=v_I}^{v=v_F} S_2(v - v_I) \, dv - \hat{\mu}(B_{I,1})$$

Similar to the section 10.2.1, once we had RMST in time interval between $v_I$ and $v_F$ for two arms, we can use equation 10.4 to compute the covariance. The formula in equation 10.2 is another optional formula to estimate the asymptotic covariance by developing predicted curve for treatment group. We still set the distribution of interim analysis $v_I$ to the final analysis $v_F$ as $\hat{S}_{1,ItoF}(v) = \hat{S}_{I,1}(v_I) S_1(v - v_I)$ which simply follows an exponential distribution as $S_1(v) = \exp(-\lambda_1 v)$ from $v_I$ to $v_F$ in predicted survival curve. Then the estimator of RMST at final analysis for treatment group can be calculated by equation 10.5 as:

$$\hat{\mu}(B_{F,1}) = \hat{\mu}(B_1(t_F, v_F)) = \int_{v=0}^{v=v_F} \hat{S}_1(t_F, v) \, dv$$

And the variance of RMST at final analysis for treatment group can be estimated by equation 10.6 as:

$$Var(\hat{\mu}(B_{F,1})) = \frac{1}{n_{F,1}} Var(B_1(t_F, v_F))$$

$$= \frac{1}{n_{F,1}} \left\{ \frac{2}{2} \int_{v=0}^{v=v_I} v \hat{S}_{I,1}(v) \, dv + 2 \hat{S}_{I,1}(v_I) \int_{v=v_I}^{v=v_F} v \exp(-\lambda_1(v - v_I)) \, dv \right.$$  

$$- \left[ \int_{v=0}^{v=v_I} \hat{S}_{I,1}(v) \, dv + \hat{S}_{I,1}(v_I) \int_{v=v_I}^{v=v_F} \exp(-\lambda_1(v - v_I)) \, dv \right]^2 \right\}$$

The RMST difference between treatment and control at final stage $\Delta_F = \hat{\Delta}_F = \hat{\mu}(B_{F,1}) - \hat{\mu}(B_{F,2})$ is given. By solving below equation we can get the estimator of $\lambda_1$.

$$\hat{\Delta}_F = \int_{v=0}^{v=v_I} \hat{S}_{I,1}(v) \, dv + \hat{S}_{I,1}(v_I) \int_{v=v_I}^{v=v_F} \exp(-\lambda_1(v - v_I)) \, dv - \int_{v=0}^{v=v_F} \hat{S}_2(t_F, v) \, dv$$

Once we have the predicted curve for treatment group, we may use the equation 10.2 to estimate the covariance directly.
10.3 Simulation Setup

Table 10.1: Survival Configuration for Each Model in Simulation

<table>
<thead>
<tr>
<th>Model</th>
<th>$\lambda$ for Treatment</th>
<th>$\lambda$ for Control</th>
<th>$\Delta F$</th>
<th>$\mu(B_{F2})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model A (constant diff.)</td>
<td>0.3</td>
<td>0.4</td>
<td>0.4287</td>
<td>2.16</td>
</tr>
<tr>
<td>Model B (early diff.)</td>
<td>0.2 + 0.4$I(t \geq 1.75) - 0.2I(t \geq 3.5)$</td>
<td>0.4</td>
<td>0.3549</td>
<td>2.16</td>
</tr>
<tr>
<td>Model C (late diff.)</td>
<td>0.4 - 0.08$I(t \geq 1) - 0.08I(t \geq 2)$</td>
<td>0.4</td>
<td>0.3233</td>
<td>2.16</td>
</tr>
<tr>
<td>Model D (middle diff.)</td>
<td>0.4 - 0.3$I(t \geq 1) + 0.9I(t \geq 3)$</td>
<td>0.4</td>
<td>0.3740</td>
<td>2.16</td>
</tr>
<tr>
<td>Model E (cross at middle and overall favor treatment)</td>
<td>0.2 + 0.65$I(t \geq 2.2)$</td>
<td>0.4</td>
<td>0.3056</td>
<td>2.16</td>
</tr>
<tr>
<td>Model F (cross at middle and overall favor control)</td>
<td>0.2 + $I(t \geq 1.7)$</td>
<td>0.4</td>
<td>-0.1363</td>
<td>2.16</td>
</tr>
</tbody>
</table>

Enrollment distribution $\sim$ uniform(0,3); Censor distribution $\sim$ min(exp(1/35), 5);
Maximum study time $t_F = 5$; Total number per group = 300;
Interim cut time $t_I = 2$; $\hat{\Delta}_F$ is RMST difference at final stage
Simulation numbers for alternative = 5,000
$\mu(B_{F2})$ is RMST at final stage for Control

Similar to section [4.1], Monte Carlo simulations were performed for six alternative models in order to compare the asymptotic covariance by Martingale integration approach with the true covariance by simulations. In the alternative models, survival time was assumed to follow a piecewise exponential distribution with piecewise constant hazard

$$\lambda_i(v|v_0, v_1, \ldots, v_J) = \sum_{j=1}^{J-I} I_{[v_{j-1}, v_j]}(v) \lambda_{ij}$$

where $i = 1$ or 2 as label for treatment arm and control arm, respectively. $v$ is the relative times (time measured from the real time of entry). $I_A(v)$ is an indication function with value 1 if $v \in A$ and 0 if $v \notin A$. Time $v_1, < \ldots, < v_J, < v_m$ are change points for the piecewise exponential model with $v_0 = 0$ and $v_J = \infty$. They are the boundary values for formulation purpose [Huang and Kuan, 2017]. The survival function as the probability of survival at time $v$ can be derived as follows:

$$S_t(v) = exp \left(-\lambda_i(v|v_0, v_1, \ldots, v_J)\right) = exp \left(-\sum_{j=1}^{j=J-I} I_{[v_{j-1}, v_j]}(v) \lambda_{ij}\right)$$
Six different types of piecewise exponential distributions were selected to represent six common survival models. They are one PH model, and five NPH models which represent: early difference, late difference, middle difference, cross at middle overall favor treatment, and cross at middle overall favor control.

Furthermore, censoring and staggered entering condition were considered. Censoring was exponentially distributed with \( \lambda = 1/35 \) in all models. For the staggered entering, the enrollment distribution was Uniform(0, 3) and the study duration was \( t_F = 5 \) for all models. The interim stage cut was at study time \( t_I = 2 \). Subjects whose event happen after 2 were considered as a censoring in interim analyses. KM curves in interim analysis were estimated by using the observed data at interim stage for each arm. Asymptotic covariance in RMST difference between interim and final analysis was estimated by using Martingale integration approach with different scenarios. Evaluations of accuracy were performed among covariance by simulation and three asymptotic covariances by two different scenarios. They are: 1) Using equation 10.2 based on predicted survival curves with given \( \Delta_F \) and RMST in final analysis for control arm; 2) Using equation 10.2 based on predicted survival curves with given \( \Delta_F \) and known survival distribution for control arm. Four fixed RMST cut points \( v_I = 1.3, v_I = 1.5, v_I = 1.7, \) and \( v_I = 2 \) were selected for covariance evaluation. RMST cut time was set to \( v_I = 5 \) for all final analyses.

All these setting resulted in an average of 30% censoring rate among all six simulated models. The detail description for each model and value of \( \Delta_F \) was presented in Table 10.1.

10.4 Results

Six KM plots are created for six common survival models examples in Figure 10.1. In each plot, solid curves are predicted survival curves for either treatment or control arm. Two dash curves are KM curves which are created from simulation data. Red color represents the treatment and blue is for the control. In treatment predicted curve (red solid line), the KM curve from 0 to \( v_I \) is the observed KM curve in interim analysis and rest part of the curve is the predicted part which followed a simple exponential distribution. Similarly, the control curve from time 0 to \( v_I \) is the observed KM curve in interim analysis and the
rest part followed a known distribution which is from study design or reference literature. When we focus on two red curves, we can see two treatment arm curves start to separate to each other at time $v = 1$ and a significant separation can be found after time $v = 1.5$ in all six models. Although survival distribution in interim analysis is theoretically identical to the distribution in final analysis during the same interim time period, difference still can be found between two KM curves in observed sample data and the difference is getting larger when the time is increasing. Similar trend can be found in two control arm curves as well. Furthermore, when we compared the predicted curve with its real one, a significant difference between two curves also can be found after the interim analysis in all six models. because the covariance in RMST difference between interim and final analysis seems only depends on the area under the curve and nothing else for the time period between $v_I$ and $v_F$, therefore, those significant differences will not affect the accuracy of asymptotic covariance once we fixed the RMST in final analysis.

Figure 10.1: One Example: Predicted Survival Curves at Interim and Final Stage by Models

Based on simulation results in Table 10.2, both two scenarios’ asymptotic covariances are close to its simulation ones at RMST cut time $v_I = 1.3$ for all six models. In the model A with constant difference, simulation covariance is close to those two asymptotic covariance which are calculated by using formula in 10.2 with predicted curves in all four RMST cutting times. Two predicted curve asymptotic covariance are identical for all RMST
Table 10.2: Covariance by Models

<table>
<thead>
<tr>
<th>Model</th>
<th>( \mu (B_{F,2}) )</th>
<th>( \Delta_F )</th>
<th>( \lambda_2 ) for control</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2.16</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>2.16</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>2.16</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>2.16</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>2.16</td>
<td>0.4</td>
<td>-0.1363</td>
</tr>
</tbody>
</table>

\( \Delta_F \) is the RMST difference between treatment and control at final stage

\( \mu (B_{F,2}) \) is the RMST for control at final stage

<table>
<thead>
<tr>
<th>Pred. Curve Mean</th>
<th>( \hat{\sigma}_{IF} )</th>
<th>( \hat{\sigma}_{IF} )</th>
<th>( \hat{\sigma}_{IF} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pred. Curve Mean</td>
<td>( \hat{\sigma}<em>{IF} ) Given ( \Delta_F, \mu (B</em>{F,2}) )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pred. Curve Mean</td>
<td>( \hat{\sigma}_{IF} ) Given ( \Delta_F, \lambda_2 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( v_I = 1.3 )</td>
<td>0.0034</td>
<td>0.0028</td>
<td>0.0037</td>
</tr>
<tr>
<td>( \sigma_{IF} ) by simulations</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pred. Curve Mean</td>
<td>( \hat{\sigma}_{IF} )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pred. Curve Mean</td>
<td>( \hat{\sigma}_{IF} )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( v_I = 1.5 )</td>
<td>0.0043</td>
<td>0.0035</td>
<td>0.0047</td>
</tr>
<tr>
<td>( \sigma_{IF} ) by simulations</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pred. Curve Mean</td>
<td>( \hat{\sigma}_{IF} )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pred. Curve Mean</td>
<td>( \hat{\sigma}_{IF} )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( v_I = 1.7 )</td>
<td>0.0053</td>
<td>0.0043</td>
<td>0.0057</td>
</tr>
<tr>
<td>( \sigma_{IF} ) by simulations</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pred. Curve Mean</td>
<td>( \hat{\sigma}_{IF} )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pred. Curve Mean</td>
<td>( \hat{\sigma}_{IF} )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( v_I = 2.0 )</td>
<td>0.0067</td>
<td>0.0056</td>
<td>0.0073</td>
</tr>
<tr>
<td>( \sigma_{IF} ) by simulations</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pred. Curve Mean</td>
<td>( \hat{\sigma}_{IF} )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pred. Curve Mean</td>
<td>( \hat{\sigma}_{IF} )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\( \Delta_F \) is the RMST difference between treatment and control at final stage

\( \mu (B_{F,2}) \) is the RMST for control at final stage

Pred. Curve \( \hat{\sigma}_{IF} \) Given \( \Delta_F, \mu (B_{F,2}) \) is calculated by 10.2

Pred. Curve \( \hat{\sigma}_{IF} \) Given \( \Delta_F, \lambda_2 \) is calculated by 10.2

cutting times \( v_I = 1.3, 1.5, 1.7, \) and 2.0. The difference between simulation covariance and other asymptotic are getting larger when the cutting time growing. Similar trend can be found in the rest five models. We found the asymptotic covariances start to lose their accuracy when the RMST cut time is close to the interim stage cut time. It can be anticipated. one of the possible reasons is because data didn’t follow a perfect homogenous entry. We used a uniform distribution to simulate subject enrollment. It is not a perfect method to simulate the homogenous entry. If more subjects are enrolled close to the interim cut, then we definitely marked them all as censoring and that would affect in our covariance calculation. Another reason is we may have many biases in discrete data. Although we have proved \( S_{I,i}(v) = S_{F,i}(v) \) for \( 0 \leq v \leq v_I \), it is still impossible to archive it in real observed data. We also can see it in figure [10.2]. In treatment group (red lines), the survival curve at interim stage is close and partially overlap with the survival curve at final stage during \( v \) from 0 to 1.5. When the \( v \) close to the interim cut time \( v = 2 \), those two lines start to separate. the same thing can be found in control group as well. Since two asymptotic
covariances are identical and closer to the simulation one. Both of them will be used in CP calculation at following section.
Chapter 11

Conditional Power in Interim Analysis

In the previous chapter, approach of calculating the covariance in RMST difference between interim and final analyses by using the predicted curve was introduced and evaluated its performance by simulation. In this chapter, the correlation in RMST difference between interim and final analyses and conditional power at interim stage will be calculated based on the predicted curve covariance. The performance of futility test based on CP will also be evaluated by simulation.

11.1 Calculation of conditional power

According to the section 9.1.4, CP can be calculated by using formula 9.5, as

\[
P(Z_F > c | Z_I) = \Phi \left( \frac{\frac{\Delta_F}{\sigma_F} - \rho_{IF} \frac{\Delta_I}{\sigma_I} - c}{\sqrt{1 - \rho_{IF}^2}} + \frac{\rho_{IF}}{\sqrt{1 - \rho_{IF}^2}} Z_I \right)
\]

When we set the true treatment effect at interim stage \(\frac{\Delta_I}{\sigma_I} = Z_I\), then the CP formula can be simplified to 9.6 as

\[
P(Z_F > c | Z_I) = \Phi \left( \frac{\frac{\Delta_F}{\sigma_F} - c}{\sqrt{1 - \rho_{IF}^2}} \right)
\]

According to section 9.2, the correlation in RMST difference between two stages can be derived by

\[
\rho_{IF} = \frac{\sigma_{IF}}{\sigma_I \sigma_F},
\]

As we known, the covariance in RMST difference can be estimated at interim analysis by using equation 10.2 with predicted curve in section 10.2. \(\hat{\sigma}_I\) is easily calculated by using
observed data at interim stage. Once $\sigma_F$ is given, we can estimate the correlation in RMST difference between two stages as:

$$\hat{\rho}_{IF} = \frac{\hat{\sigma}_{IF}}{\hat{\sigma}_I \hat{\sigma}_F},$$

(11.1)

where $\hat{\sigma}_F = \sigma_F$. $\sigma_F$ usually is given in study design.

Once we have the estimator of correlation, we can derive the CP either by plugging in the correlation estimate into the formula 9.5 with given treatment effect at interim, treatment effect at final, and observed $Z_I$, or by plugging in it into the formula 9.6 with assumed treatment effect at final.

### 11.2 Simulation plan

Similar to previous chapter, Monte Carlo simulations were performed for six alternative models. In these alternative models, survival time was assumed to follow a piecewise exponential distribution with piecewise constant hazard. Six different types of piecewise exponential distributions were selected to represent six common survival conditions. They are one PH model, and five NPH models which represent: early difference, late difference, middle difference, cross at middle overall favor treatment, and cross at middle overall favor control.

Furthermore, censoring was exponentially distributed with $\lambda = 1/35$ in all models. The enrollment distribution was Uniform(0, 3) and the study duration was $t_F = 5$ for all models. The interim stage cut was at study time $t_I = 2$. Subjects whose event happen after 2 was considered as a censoring at interim stage survival analyses. For all six models, the sample size was three hundred subjects in each group, and 5000 simulations were performed for each alternative. All tests were performed using a one-sided test with a 0.025 significance level. All given parameters were generated by simulation in its corresponding survival model or from simulation model set up directly. Observed parameters were from each randomly simulating sample. Covariance in RMST difference between two stages were calculated by using predicted curve either with given $\Delta_F$ and $\lambda_2$ or with given $\Delta_F$ and RMST in final analysis for the control. CP was calculated in formula 9.5 and formula 9.6 with either
correlation by simulation or correlations by predicted curve. CP results were evaluated in three RMST cut time point which are \( v = 1.3, t = 1.5, \) and \( t = 1.7. \) CP less than 10% and CP less than 30% were selected as the monitoring rule of stopping study for futility. The negative predictive value (NPV) was generated for stopping futility at 10% CP and 30% CP. The stopping rate and its NPV were evaluated and compared between correlation by simulation and correlation by predicted curve in all survival models.

11.3 Results

Table 11.1: Conditional Power for all Models at \( v_I = 1.3 \)

<table>
<thead>
<tr>
<th>Model</th>
<th>A (constant diff.)</th>
<th>B (early diff.)</th>
<th>C (late diff.)</th>
<th>D (middle diff.)</th>
<th>E (Cross at middle overall favor treatment)</th>
<th>F (Cross at middle overall favor control)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( v_I = 1.3 )</td>
<td>( \Delta_I )</td>
<td>( \Delta_F )</td>
<td>( \lambda ) for control</td>
<td>( \sigma_I )</td>
<td>( \sigma_F )</td>
<td>( \rho_{IF} )</td>
</tr>
<tr>
<td>Given parameters</td>
<td>0.0620</td>
<td>0.0620</td>
<td>0.0620</td>
<td>0.0620</td>
<td>0.0620</td>
<td>0.0620</td>
</tr>
<tr>
<td>Observed at Interim</td>
<td>0.0620</td>
<td>0.0620</td>
<td>0.0620</td>
<td>0.0620</td>
<td>0.0620</td>
<td>0.0620</td>
</tr>
<tr>
<td>By ( \mu_{IF} ) by simulations</td>
<td>0.8533</td>
<td>0.8533</td>
<td>0.8533</td>
<td>0.8533</td>
<td>0.8533</td>
<td>0.8533</td>
</tr>
<tr>
<td>Mean of CP</td>
<td>0.8533</td>
<td>0.8533</td>
<td>0.8533</td>
<td>0.8533</td>
<td>0.8533</td>
<td>0.8533</td>
</tr>
<tr>
<td>% Stop at CP10 (NPV%)</td>
<td>0.36 (61.1%)</td>
<td>0.36 (61.1%)</td>
<td>0.36 (61.1%)</td>
<td>0.36 (61.1%)</td>
<td>0.36 (61.1%)</td>
<td>0.36 (61.1%)</td>
</tr>
<tr>
<td>% Stop at CP30 (NPV%)</td>
<td>0.74 (76.3%)</td>
<td>0.74 (76.3%)</td>
<td>0.74 (76.3%)</td>
<td>0.74 (76.3%)</td>
<td>0.74 (76.3%)</td>
<td>0.74 (76.3%)</td>
</tr>
<tr>
<td>Simplify CP</td>
<td>0.8518</td>
<td>0.8518</td>
<td>0.8518</td>
<td>0.8518</td>
<td>0.8518</td>
<td>0.8518</td>
</tr>
<tr>
<td>By ( \tilde{\mu}_{IF} ) given ( \lambda_2, \Delta_F )</td>
<td>0.8459</td>
<td>0.8459</td>
<td>0.8459</td>
<td>0.8459</td>
<td>0.8459</td>
<td>0.8459</td>
</tr>
<tr>
<td>Mean of CP</td>
<td>0.8459</td>
<td>0.8459</td>
<td>0.8459</td>
<td>0.8459</td>
<td>0.8459</td>
<td>0.8459</td>
</tr>
<tr>
<td>% Stop at CP10 (NPV%)</td>
<td>0.74 (76.3%)</td>
<td>0.74 (76.3%)</td>
<td>0.74 (76.3%)</td>
<td>0.74 (76.3%)</td>
<td>0.74 (76.3%)</td>
<td>0.74 (76.3%)</td>
</tr>
<tr>
<td>% Stop at CP30 (NPV%)</td>
<td>0.45 (47.5%)</td>
<td>0.45 (47.5%)</td>
<td>0.45 (47.5%)</td>
<td>0.45 (47.5%)</td>
<td>0.45 (47.5%)</td>
<td>0.45 (47.5%)</td>
</tr>
<tr>
<td>Mean of simplify CP</td>
<td>0.8518</td>
<td>0.8518</td>
<td>0.8518</td>
<td>0.8518</td>
<td>0.8518</td>
<td>0.8518</td>
</tr>
</tbody>
</table>

CP was calculated based on covariance which was estimated by predicted curve with Martingale integration approach formula \[10.2\]. Covariance calculations were done with following two scenarios: 1) Given \( \Delta_F, \sigma_F \) and assumed survival distribution for control group is known. 2) Given \( \Delta_F, \sigma_F, \) and RMST for control group at final stage. CPs were calculated
by using the formula $9.5$ and the simplified formula $9.6$ with either simulation correlation or estimated correlations by predicted curve. Both 10% and 30% CP were selected as the futility stopping rule for monitoring futility at interim stage. Their performances were evaluated by their futility stopping rate and negative predictive value (NPV), and compared among CP by simulation and two CPs by predicted curves in different scenarios. Based on the simulation results in table 11.1, two estimated CPs are close to their corresponding simulation CP in all six models at restricted cut time $v_I = 1.3$. In the first setting with proportional hazards model, CPs calculated by the formula $9.5$ are close to the simplified one in all three different correlations scenarios: one is simulation correlation, the other two are estimated ones by predicted curves. CP which is estimated by predicted curve with known survival distribution for control group has the best performance in futility stopping. Compared to other two CPs, three of them all stopped nothing in 10% rule. However, the CP with known $\lambda_2$ has higher stopping rate and higher NPV rate than rest of two in 30% rule. By multiplying the stopping rate and NPV, number of true failed studies stopped for CP with known $\lambda_2$ is twice more than the number for rest of two CPs in 30% CP rule. The futility stopping performance for rest of two CPs are almost the same. The result can easily be explained. Different with CP by simulation and CP by given $\mu(BF,2)$, the control arm’s RMST value in final analysis is not fixed for CP by given $\lambda_2$. By letting the control arm’s predicted curve followed the known distribution, the predicted survival curve is closer to the real and increases the chance of stopping true futility. Rest of two CPs have very similar performance, since they have close RMST value for control arm. The similar trends were observed for rest five models except the model F (cross at middle overall favor to control). because the model F has extremely low test power in final analysis, all CPs have 100% stopping rate and 99.8% in NPV. Furthermore, compared to 30% CP rule, much less simulation cases are stopped in 10% CP rule but all of stopped ones have a higher NPV.

CP performances were also evaluated by three different interim analysis time points. Similar simulations were performed for $v_I = 1.5$ and $v_I = 1.7$. In simulation results table 11.2 all CPs are close to each other in the same cutting time. CPs are increasing slowly when the cutting time are growing. Similar trend with $v_i = 1.3$ was found in $v_I = 1.5$ and $v_I = 1.7$. CP by $\lambda_2$ always has the best performance in futility stopping rate among all
<table>
<thead>
<tr>
<th>$v_f = 1.3$</th>
<th>$v_f = 1.5$</th>
<th>$v_f = 1.7$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Model</strong></td>
<td><strong>A</strong> (constant diff.)</td>
<td><strong>B</strong> (early diff.)</td>
</tr>
<tr>
<td>Mean of CP</td>
<td>0.8532</td>
<td>0.7590</td>
</tr>
<tr>
<td>% Stop at $T_{10}$ (NPV%)</td>
<td>0</td>
<td>0.8 (95%)</td>
</tr>
<tr>
<td>% Stop at $T_{30}$ (NPV%)</td>
<td>0.36 (61.1%)</td>
<td>1.26 (76.2%)</td>
</tr>
<tr>
<td>Simplify CP</td>
<td>0.8518</td>
<td>0.7548</td>
</tr>
<tr>
<td>By $μ_{F}$ given $λ_2$, $Δ_F$</td>
<td>Mean of CP</td>
<td>0.8459</td>
</tr>
<tr>
<td>% Stop at $T_{10}$ (NPV%)</td>
<td>0</td>
<td>0.06 (100%)</td>
</tr>
<tr>
<td>% Stop at $T_{30}$ (NPV%)</td>
<td>0.74 (70.3%)</td>
<td>1.94 (76.3%)</td>
</tr>
<tr>
<td>Mean of simplify CP</td>
<td>0.8517</td>
<td>0.7556</td>
</tr>
<tr>
<td>By $μ_{F}$ given $μ(B_{F.2}), Δ_F$</td>
<td>Mean of CP</td>
<td>0.8527</td>
</tr>
<tr>
<td>% Stop at $T_{10}$ (NPV%)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>% Stop at $T_{30}$ (NPV%)</td>
<td>0.4 (65%)</td>
<td>1.3 (78.5%)</td>
</tr>
<tr>
<td>Mean of simplify CP</td>
<td>0.8518</td>
<td>0.7557</td>
</tr>
</tbody>
</table>

### Table 11.2: Conditional Power for all Models by Interim Analysis cut Times

<table>
<thead>
<tr>
<th>$v_f$</th>
<th>$v_f$</th>
<th>$v_f$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.3</td>
<td>1.5</td>
<td>1.7</td>
</tr>
</tbody>
</table>

### Notes:
- $μ_{F}$ given $λ_2$, $Δ_F$ is calculated in predicted curve with given $λ_2$ and $Δ_F$ by formula 10.2.
- $μ_{F}$ given $μ(B_{F.2}), Δ_F$ is calculated in predicted curve with given $μ(B_{F.2}), Δ_F$ by formula 10.2.
- $μ(B_{F.2})$ is RMST in final analysis for the control. $Δ_F$ is RMST difference in final analysis.
- % Stop at $T_{10}$ is percentage of simulations were stopped by using rule of stopping study when CP < 10%.
- % Stop at $T_{30}$ is percentage of simulations were stopped by using rule of stopping study when CP < 30%.
- NPV is negative predictive value which is (subjects true failed study)/(total subjects stopped study based on CP)*100.
- CP is calculated by formula 9.5. Simplify CP is calculated by formula 9.6.

Three cutting times. However, its NPV rate has been caught up and exceeded by rest two CPs in $v_f = 1.5$ and $v_f = 1.7$ for most models. All CPs have their best futility stopping rate and their top or close to the top NPV in $v_f = 1.5$ in model A (PH model) and C (NPH with late difference). For model B (NPH with early difference) and E (NPH with cross at middle and overall favor treatment), $v_f = 1.7$ has the best monitoring futility performance for all CPs. Stopping futility performances are slightly down from $v_f = 1.5$ to $v_f = 1.7$ in
all three CPs for most of models. It is anticipated. $v_I = 1.7$ is too close to the interim stage cut time $t_I = 2$. The number of subjects drop rapidly when study time gets close to the interim stage cut. $\hat{S}_I(v)$ will be totally different to $\hat{S}_F(v)$ when $v$ gets close to $t_I$. 
Chapter 12

Asymptotic variance of RMST Difference at final stage

As commonly known, $\sigma_F$ is usually given in the study design. We can barely calculate it at interim stage of the study. In the previous chapter, a predicted curve approach was proposed to calculate the covariance of RMST difference between two stages, in order to compute the CP at interim stage. In this chapter, the asymptotic variance of RMST difference in final analysis will be derived by using the predicted curve approach as well.

12.1 Calculation of Asymptotic variance of RMST test statistics at final stage

Based on the variance formula \[9.11\], $\sigma_F$ is depending on the survival distribution after interim stage for two arms. The estimator $\hat{\sigma}_F$ is close to the true one when the predicted curve is close to the real survival curve. By considering most real survival curve may have a turning point in the middle, we add a turning point in our predicted curve to improve the accuracy of estimation for $\hat{\sigma}_F$.

For treatment group, we set both survival distribution from interim RMST cut $v_I$ to the turning point $v_{turning}$ and survival distribution from $v_{turning}$ to final RMST cut $v_F$ to follow two different simple exponential distributions with $\lambda$ equal to $\lambda_{1a}$ and $\lambda_{1b}$, respectively. $\lambda_{1a}$ is equal to the instant hazard at $v_I$ for treatment group. When we fix the $\hat{\Delta}_F$, the estimator of $\lambda_{1b}$ can been gotten by solving below equation.

$$
\hat{\Delta}_F = \int_{v=0}^{v=v_I} \tilde{S}_{I,1} (v) \, dv + \tilde{S}_{I,1} (v_I) \int_{v=v_I}^{v=v_{turning}} \exp (-\lambda_{1a} (v - v_I)) \, dv 
+ \tilde{S}_{I,1} (v_I) \exp (-\lambda_{1a} (v_{turning} - v_I)) \int_{v=v_{turning}}^{v=v_F} \exp (-\lambda_{1b} (v - v_{turning})) \, dv 
- \int_{v=0}^{v=v_F} \tilde{S}_2 (t_F, v) \, dv
$$
Then, the estimator of variance at final stage for treatment can be derived as:

\[
\text{Var}(\hat{\mu}(B_{F,1})) = \frac{1}{n_{F,i}} \text{Var}(B_i(t_F, v_F))
\]

\[
= \frac{1}{n_{F,i}} \left( 2 \int_{v=0}^{v=v_f} v \hat{S}_{I,1}(v) \, dv + 2 \hat{S}_{I,1}(v_f) \int_{v=v_f}^{v_{\text{turning}}} v \exp (-\lambda_a (v - v_f)) \, dv \\
+ 2 \hat{S}_{I,1}(v_f) \exp (-\lambda_a (v_{\text{turning}} - v_f)) \int_{v=v_f}^{v_{\text{turning}}} v \exp (-\lambda_b (v - v_{\text{turning}})) \, dv \\
- \left[ \int_{v=0}^{v=v_f} \hat{S}_{I,1}(v) \, dv + \hat{S}_{I,1}(v_f) \int_{v=v_f}^{v_{\text{turning}}} \exp (-\lambda_a (v - v_f)) \, dv \\
+ \hat{S}_{I,1}(v_f) \exp (-\lambda_a (v_{\text{turning}} - v_f)) \int_{v=v_f}^{v_{\text{turning}}} \exp (-\lambda_b (v - v_{\text{turning}})) \, dv \right]^2 \right)
\]

(12.1)

For the control group, then we can use the general formula 10.6 to calculate the variance based on predicted curve for control

\[
\text{Var}(\hat{\mu}(B_{F,i})) = \frac{1}{n_{F,i}} \text{Var}(B_i(t_F, v_F))
\]

\[
= \frac{1}{n_{F,i}} \left( 2 \int_{v=0}^{v=v_f} v \hat{S}_{I,i}(v) \, dv + 2 \hat{S}_{I,i}(v_f) \int_{v=v_f}^{v_{\text{turning}}} v \exp (-\lambda_i (v - v_f)) \, dv \\
- \left[ \int_{v=0}^{v=v_f} \hat{S}_{I,i}(v) \, dv + \hat{S}_{I,i}(v_f) \int_{v=v_f}^{v_{\text{turning}}} \exp (-\lambda_i (v - v_f)) \, dv \right]^2 \right)
\]

12.2 Example

In figure 12.1 by fixing the \( \Delta_F = 0.3740 \) and knowing the survival distribution for control \( \lambda = 0.4 \), Six predicted curves were created for model D in 6 different turning point scenarios: without turning point, turning point at 2, turning point at 2.5, turning point at 3, turning point at 3.5, and turning point at 4.

<table>
<thead>
<tr>
<th>Table 12.1: Asymptotic Variance for Model D by Different Turing Points</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Model D With Different Turing Point</strong></td>
</tr>
<tr>
<td><strong>Predicted Without Turning</strong></td>
</tr>
<tr>
<td><strong>True Curve</strong> (Simulations)</td>
</tr>
<tr>
<td>( \hat{\sigma}^2_{F} )</td>
</tr>
<tr>
<td>( \hat{\sigma}^2_{IF} )</td>
</tr>
</tbody>
</table>

In table 12.1 when the \( \Delta_F \) and \( S_2(v) \) are fixed, covariance estimates are exactly the same among all six turning point scenarios and also are identical to the covariance by
Figure 12.1: Predicted Curve and True KM Curve at Final Stage in Model D

This result consolidated our findings in section 10.2. The survival distribution after interim analysis has no impact to the value of covariance, when the area under the curve is fixed. Range of estimated variance in RMST difference at final stage is [0.0200, 0.0208]. The variance by simulation is 0.0216. Although the true variance is outside the predicted range, but the difference between the true and the maximum of the range is small. Those asymptotic variances are generated by predicted curves with different turning points. Since the calculation is based on assumptions of homogeneous entry and $\hat{S}_I(v) = \hat{S}_F(v)$, the true final variance sometimes falls outside the range of estimated variance once these assumptions were violated. However, the difference between predicted variance and true one is still in a reasonable range. Therefore, those asymptotic variances are a good reference for us to select a reasonable $\sigma_F$ in study design or monitoring futility at interim stage.
Chapter 13

Conclusion

In Chapter 10, with observed interim data and given RMST difference at final analysis, covariances in the RMST difference between interim and final analyses were derived by predicted curve based on either known distribution of control or the RMST for control group at final analysis. The properties of the covariance were identified. The major property is that the covariance does not depend on the survival distribution after the RMST interim cut time once the area under the curve was fixed. These properties allow us to calculate the asymptotic covariance without assuming survival distribution for the part after RMST interim cut time. The accuracy of asymptotic covariance was evaluated by simulation with six different PH and NPH models. The simulation results showed that the asymptotic covariance is close to its corresponding simulation one in different RMST cut times for all six models.

In Chapter 11, conditional powers were calculated based on the asymptotic covariances which were derived by predicted survival curves. By using the simplified formula, the conditional power can be simply calculated based on observed interim data with assumed RMST difference and variance at final analysis. Simulation also was used to evaluate their accuracy. Based on the simulation results, all estimated CPs are close to their corresponding simulation CP in all six models. CP estimated by predicted curve with known control group’s survival distribution has the best performance in futility stopping. It has higher stopping rate and higher NPV rate than rest of two at RMST cutting point $v_I = 1.3$ and $v_I = 1.5$. The futility stopping performance for CP by simulation and CP by giving RMST at final for control arm are almost the same. Compared among three RMST cutting points, All CPs have their best futility stopping performance at $v_I = 1.5$. Stopping futility performances are slightly down from $v_I = 1.5$ to $v_I = 1.7$ in all three CPs for most of
models.

In Chapter 12, asymptotic variances of RMST difference at final stage were derived based on predicted curves. Different turning points were added to the predicted curve in order to fit more complex model. A range of asymptotic variance was created by using estimated variance from different turning points. Although the true variance may not fit in the predict range, it can still provide us a reference for final stage variance and will benefit us in developing variance assumption in any interim analyses.
Chapter 14
Discussion and Future Works

14.1 Discussion

In part I of this dissertation, the censored case correlation matrices for two of four members families $G^{0,0}, G^{1,0}, G^{0,1}, G^{1,1}$ and $G^{0,0}, G^{2,0}, G^{0,2}, G^{2,2}$ were derived. MERT test was compared with Max-combo test and log-rank test in four typical non-proportional hazards models in Freidlin et al. (1999) and was barely mentioned in last 20 years. In the paper, all MERT tests were performed based on uncensored case correlation matrices under the family $G^{0,0}, G^{2,0}, G^{0,2}, G^{2,2}$. In this dissertation, MERT test was performed and compared between plug in the uncensored case correlation matrices and censored correlation matrices for either $G^{0,0}, G^{1,0}, G^{0,1}, G^{1,1}$ or $G^{0,0}, G^{2,0}, G^{0,2}, G^{2,2}$. Based on 10000 simulations, the power of MERT test with uncensored case correlation was 0.0305 in the $G^{0,0}, G^{1,0}, G^{0,1}, G^{1,1}$ family and was 0.0323 in the $G^{0,0}, G^{2,0}, G^{0,2}, G^{2,2}$ family under the null hypothesis model. Both powers were significantly over our one-sided significant level which was set to 0.025. However, the same MERT tests were well preserved the type one error by using the censored correlation matrices. This result indicated that uncensored correlation is not appropriate to be used in the censored data.

Overall, the Max-Combo test has better power performance than Log-rank test, MERT, and RMST across different PH or NPH models scenarios investigated here. However, it still has some limitations especially as a test to detect crossing survival functions. There is one argument that Max-Combo test ignore the unfavored part, but is merely focused on testing favored treatment part in detecting crossing survival function models under one-sided test. This has been identified in this dissertation under the model F (cross middle overall favor to control). It has been found that all Max-Combo tests may lead reverse results to the real in model F. Similar case happened in using Max-combo test in real Gastric data. P-value
for all Max-Combo tests are less than or close to 0.025, which also lead reverse result to the real. In these cases, the short term and long-term survival benefits may be totally reversed. Considered the Max-Combo test has limitation in detecting crossing survival function, the RMST test can be used as an optional general test for all PH and NPH models. When we have already known delayed treatment benefit will not happen in the study, MERT test which only included either $G_{0,0}^0, G_{1,0}^1, G_{1,1}^1$ or $G_{0,0}^2, G_{2,0}^2, G_{2,2}^2$ will be our first choice of general test to detect one-sided effect for all rest PH and NPH models.

In this dissertation, performance for MERT test didn’t meet the research expectation. It didn’t beat other proposed general tests or measurement in test power as what it did in [Freidlin et al. (1999)]. One of the key reasons for not able to repeat Freidlin paper’s finding is that instead of using the uncensored case correlation, censored case correlation was used to generate MERT test in this dissertation. Based on simulation results, we can see that MERT has higher test power with the uncensored correlation than the test with the censored case correlation. However, it has been identified that it is not appropriate to use uncensored correlation, since it cannot preserve the type one error rate. The other reason is that we used piecewise exponential distribution to simulate each NPH models.

Survival time was set to follow a piecewise exponential distribution with piecewise constant hazard in order to simulate different NPH models. However, the simulated hazard changed roughly in piecewise distribution. Although it successfully simulated each typical NPH models, there is a huge difference in values of test statistics between model corresponding FH statistic and non-corresponding FH statistics. In most cases, corresponding statistic has a very high value and non-corresponding statistic values are extremely low. MERT is not able to achieve a good test performance in this extremely unbalanced situation, since it is a convex combination of standard test statistics from two or three of the most different alternative models. By changing the survival time from a piecewise exponential distribution to a time-dependent exponential distribution, we can see an improvement in MERT test performances.

Furthermore, the performance of RMST with different cut time strategies was also evaluated in typical proportional hazards and non-proportional hazards models. Three popular $v_{cut}$ strategies were evaluated in this dissertation. They are (a) $v_{cut}$ is the overall study
time, (b) $v_{cut}$ is the minimum of the maximum observed event time of two arms (minimax event time), (c) $v_{cut}$ is the maximum observed (event or censored) time (minimax observed time) of two arms. Based on simulation results, there was a significant test power difference between $v_{cut} = \text{final}$ and rest two strategies in model with delay treatment benefit. Compared to the RMST cut at final study time, RMST by minimax event time and RMST by minimax observed time lost 30% to 40% power. the reason is that last observed or event happened much earlier than the study end time in late difference model. We lost more information in these two RMST cutting strategies.

Based on simulation results in this dissertation, it is still not clear which $\rho$ and $\gamma$ combination is the best one for the FH test. Simulation results have already shown that Max-combo test or FH test in corresponding model have better test performance in $G_{0,0}^0, G_{2,0}^0, G_{0,2}^0, G_{2,2}^2$ family than test performance in $G_{0,0}^0, G_{2,0}^0, G_{0,2}^0, G_{2,2}^2$ family. We can easily predict that the test power will increase if we use $G_{0,0}^0, G_{3,0}^3, G_{0,3}^3, G_{3,3}^3$ family. Since the limitation for $\rho$ and $\gamma$ is that both values should greater or equal to zero, we can set the $\rho$ and $\gamma$ to a very large number in order to achieve high test power for either Max-combo test or FH test. However, both Max-combo test and FH test ignore the unfavored apart and are merely focused on testing favored treatment part in detecting crossing survival function models under one-sided test. This issue is getting worse when $\rho$ and $\gamma$ values are increasing. Therefore, more analyses are needed to find the best value for $\rho$ and $\gamma$.

RMST-based analysis become more popular nowadays. Compared to the traditional log-rank test and proportional hazard model, it is more clinical interpretable and has better test performance in nonproportional hazards models. However, there are still many challenges to using RMST as the primary analysis in a randomized clinical trial, especially for group sequential study based on RMST. In part II of this dissertation, I focused on how to conduct a two-stage design with the RMST endpoint. A new method of calculation asymptotic covariance in RMST difference between interim and final analyses was developed. CP was calculated based on asymptotic covariance at interim stage and has been evaluated by simulations.

One of major challenges for developing a two-stage study in RMST is how to select the RMST cut time point $v_I$ in the study, especially for the interim stage analysis. In this
dissertation, the asymptotic covariance was estimated in staggered entered data under the homogenous entry assumption. \( S_I(v) = S_F(v) \) for \( 0 \leq v \leq v \) has been proved in section 10.2. Based on this property, \( \hat{S}_I(v) = \hat{S}_F(v) \) for \( v \) has been implied and \( \hat{S}_I(v) \) was used as an estimator for \( S_F(v) \) for \( 0 \leq v \leq v \). Since most interim analyses are conducted during subjects enrollment period, most study subjects will become censoring when the relative time \( v \) is close to the interim stage cut time \( t_I \) at interim analyses. \( \hat{S}_I(v) = \hat{S}_F(v) \) will be violated heavily and the estimated covariance will become unreliable when the RMST cut time is close to interim stage cut time in interim analyses. However, in general, letting \( v_I \) be close to \( t_I \) as much as possible can avoid unnecessary loss of information. The minimum of maximum observed event time (minimax event time) of each arm was suggest to be used at \( v_I \) in \cite{Tian et al. 2020}. Since the minimax event time is not fixed during simulation, I am not able to evaluate its performance by simulation. Therefore, fixed cutting time points were selected in this dissertation.

Martingale integration approach was used in deriving formula for variance of RMST. I expanded it to the covariance calculation. By exploring properties of the covariance formula, I discovered that the survival distribution after the interim analysis had no impact to the covariance estimation when the area under the curve was fixed. With the given \( \Delta_F \), observed data at interim, and either survival distribution or the RMST for the control in final analysis, we are able to calculate the covariance in RMST difference between interim and final analyses. This new approach requires to create predicted curves for both arms. The predicted curve from time 0 to \( v_I \) is the observed KM plot in interim analysis, and the rest part of predicted curve follows a simple exponential distribution with a fixed under curve area.

Many benefits can be found by using the Martingale integration approach to generate the covariance in RMST difference between interim and final analyses. Compared with counting process approach, the Martingale integration approach’s formula is more straightforward and clear. Besides, it requires less information. By using the martingale integration approach with predicted curves, we can calculate the covariance in interim analyses without observing final data or assuming final survival distribution for treatment arm. We may be able to handle some complex analyses in interim analyses as well, like calculating covariance
by assuming survival curve follow current trend, or generating CP under the null hypothesis in interim analysis. The estimated covariance is also adjustable by changing the predicted curves with different RMST difference in final analysis. However, this method still has some limitations. \( \hat{S}_I(v) \) is set as the estimator for \( S_I(v) \). It was assumed \( \hat{S}_I(v) \approx \hat{S}_F(v) \) and replaced \( \hat{S}_F(v) \) by \( \hat{S}_I(v) \) for \( 0 \leq v \leq t_I \) in this method. Although these assumptions have been proved in theory, they can still be heavily violated when interim analysis is close to interim stage cut time or the study didn’t follow a homogenous entry. Since the RMST is the area under the survival curve. Different to other endpoints, RMST is less sensitive to the event counts, but more focus on the survival curve distribution. From this point of view, we may consider a new way in analyses with RMST, where rather than subject level event counting, may put more attentions on the under curve area on survival curves.

### 14.2 Future Work

Instead of assuming value of area under the curve for survival curves after interim analyses, we can use more complex distribution models to fit the interim data and explore their survival curves beyond the interim analyses cut \( v_I \). Either Weibull or log-normal is a good potential candidate for fitting distribution. In covariance calculation, Martingale integration approach will be used and the predicted curve will exactly follow the survival distribution which we found based on the interim data. Compared to assuming survival curve follows simple exponential distribution after \( v_I \), predicted curves will be more consistent and smooth with one fitted distribution throughout the entire study. This way may make the predicted curve more practical, however, it still need to be further explored and evaluated.

How to control the type I error and how to re-estimate sample size at interim stage are two hot topics for two-stage design. Regarding to the RSMT endpoint, the basic concept of these two topics have been introduced in paper [Lu and Tian (2020)](https://doi.org/10.1002/prev.202000015). However, the covariance of RMST difference between interim and final analyses was calculated based on the counting process approach in the paper. A detailed process in Martingale integration approach with predicted curves can be explored in future as well.
Bibliography


P. Royston and M. K. B. Parmar. The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. Statistics in Medicine, 30:2409–2421, 2011.


