STATISTICAL APPLICATIONS TO CARDIOVASCULAR DISEASE RESEARCH

BY KEZHEN LIU

A Dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Statistics and Biostatistics

Written under the direction of

Javier Cabrera

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

January, 2014

ABSTRACT OF THE DISSERTATION

Statistical Applications to Cardiovascular Disease Research

By KEZHEN LIU

Dissertation Director: Javier Cabrera

Cardiovascular disease (CVD) is the most frequent cause of deaths worldwide [1]. Scientists have done and are stilling doing a high volume of research on this area, hoping to help people who are already suffering from the disease and also to prevent those at high risk of getting CVD. Statistical applications play a very important role in most of these research activities and a better utilization of the right statistical methodology for a specific study would definitely make the research outcomes more reliable and eventually being beneficial to the human kind. This dissertation studies several scenarios in cardiovascular disease research where traditional statistical methods may not be applicable. And we proposed corresponding practical solutions or modifications to existing methods to better fit the problems case by case.

In the first part, we are focusing on using the gain in life expectancy to assess the treatment effect of an antihypertensive therapy for stroke. We first propose a framework for estimating this quantity by calculating the area between estimated survival curves given by two comparative treatments. And then, in order to better assess the variability of our

estimate especially with small sample size, we propose a new bootstrap method for obtaining confidence interval for this quantity. We also propose the corresponding bootstrap testing procedure to test the null hypothesis.

The second part of the dissertation is about meta-analysis in CVD research. We discover the non-normal behavior of the test statistics when sample size in each study of the meta-analysis is small. We use t distribution to approximate the underlying distribution and propose a simple formula to calculate the degree of freedom of the t distribution based on the sample size in each study as well as the number of studies.

Finally, we modify a new clinical design called Simultaneous Global Drug Development Program (SGDDP) which can be more efficient for evaluating the treatment effect on diseases such as CVD where ethnicity have a potential impact. We add an additional assumption to the original test to make it unbiased. We also show the performance of the program after the modification.

Acknowledgements

First of all, I would like to express the deepest appreciation to my advisor, Dr. Javier Cabrera, who led me into the world of statistical research and guided me all the way along. I could not have finished this dissertation without his help and support. He inspired me to find my own ways of conducting research with passion and persistence. His encouragement kept me going and helped me conquer the difficulties not only in the research but also in my life.

I would like to thank Dr. Jerry Cheng. It was a great pleasure to work with him on the bootstrap project which eventually became the first part of my dissertation. I would also like to thank the other two committee members: Dr. John Kolassa and Dr. Lee Dicker. Their feedbacks on my dissertation were very helpful.

My gratitude also goes to my summer intern supervisors and colleagues from Pfizer and Janssen Research: Dr. Michael Gaffney, Dr. Gang Chen and Dr. Zhilong Yuan. They spent a lot of time and energy working with me even after my internship ended. Without them, I could never be able to extend the intern projects into two parts of my dissertation. And I'm also very grateful for the opportunity to work in the industrial environment with them. Those experiences I gained and advices I got from them would definitely be beneficial for my future career.

Finally, I want to say "thank you" to all faculty and staff members, as well as my fellow Ph.D. students, in the department of statistics and biostatistics at Rutgers. Thank you all

for making my time at Rutgers much easier and joyful. I will always cherish the wonderful memories I had here.

Dedication

This work is dedicated to my parents, Shanyue Liu and Ping Jin, my beloved girlfriend, Xiaoqing Tang, for their unconditional love and support.

Table of Contents

Abstract ii

Acknowledgements iv

Dedication vi

List of Tables xi

List of Figures xii

1 Introduction 1

1.1 Motivation 1

1.2 Assessing Life Extension in Survival Analysis 2

1.3 Meta-Analysis with small sample sizes 4

1.4 Weighted-Z test in Clinical Design 7

2 Assessing Life Extension from Medical Interventions 10

2.1 Introduction 10

2.2 Methods 13

2.2.1 Notations 13

2.2.2 Estimating Survival Functions 14

2.2.3 Restricted Mean Survival Time 16

2.2.4 Estimation of the Area Between Two Survival Curves 17

2.2.5 Confidence Interval and p-value by Normal Approximation 18

2.2.6 Bootstrap Confidence Interval and p-value 19

2.3 Asymptotic Properties of the TSG estimators 22

2.4 Simulations 25

2.4.1 Bias of Estimation from KM Estimator and Cox Regression with Balanced Design 25

2.4.2 Bias of Estimation from Cox Regression with Imbalanced Design 27

2.4.3 Coverage of Bootstrap Confidence Intervals and Power of Bootstrap Testing 29

2.4.4 Comparing Results from Bootstrap Method and Normal Approximation.. 31

2.5 Application to a Real Dataset 37

2.5.1 Data Description 37

2.5.2 Analyses and Results 38

2.6 Discussions 40

3 Meta-Analysis with Small Sample Sizes 42

3.1 Introduction 42

3.2 Method 44

3.2.1 Data Structure 44

3.2.2 Inverse-variance Method 45

3.2.3 Using Pooled Standard Error from All Studies 46

3.2.4 Using Standard Error from Individual Study 47

3.2.5 Letter Values Approximation 48

3.2.6 Finding an Empirical Formula for the Degree of Freedom 49

3.2.7 Data Simulation 49

3.3 Simulation Results 50

3.3.1 Type I error 50

3.3.2 The t-distribution Approximation 54

3.3.3 Formula for Calculating Degree of Freedom 56

3.4 Simple Data Example 59

3.5 Discussion 60

4 Statistical Properties of the Design for Simultaneous Global Drug Development Program 62

4.1 Introduction 62

4.1.1 Background on Global Clinical Trials 62

4.1.2 SGDDP Design and Statistical Method 65

4.1.3 Issues about the SGDDP 67

4.2 Modification to the Hypothesis in SGDDP Design 68

4.2.1 Distribution of Weighted-Z test 68

4.2.2 Proportionality Assumption 69

4.2.3 Sample Size Evaluation 70

4.3 Power Consideration 74

4.3.1 The Test with Optimum Weight 75

4.3.2 Comparing Powers of Weighted Z test to Optimum Test..... 76

4.4 Discussion and Final Remarks 81

References 85

List of Tables

2.5.1 *TSG* from Kaplan-Meier Approach 38

2.5.2 *TSG* from Cox Partial Regression 39

3.2.1 Data from Study i 44

3.3.1 Sample Sizes in Different Simulation Sets 51

3.3.2 Comparing results from PROC GLM and Meta-analysis with MSE 51

3.3.3 Letter Value Spreads and Ratios of Test Statistic from D1 and D2 55

3.3.4 Approximate Degree of Freedom from Different Sample Size Settings 57

3.3.5 Testing Results of df from Different Formulas 58

3.4.1 Summary Statistics from Individual Study 59

3.4.2 Testing Results from Different Approximation Methods 60

4.2.1 Sample Size for LCT when MRCT has 20% TE Patients 73

4.2.2 Sample Size for LCT when MRCT has 10% TE Patients 73

4.3.1 Powers of the Weighted-Z Test with $n_{1p} = 150$ 78

4.3.2 Powers of the Weighted-Z Test with $n_{1p} = 200$ 78

4.3.3 Powers of the Weighted-Z Test with $w = 45\%$ and Different n_{1p} 79

4.3.4 Powers of the UMP tests with Minimum n_{1p} 80

List of Figures

2.4.1 *TSGs* from Two Methods--1 27

2.4.2 *TSGs* from Two Methods--2 28

2.4.3 Bootstrap Confidence Interval Coverage Rates 29

2.4.4 Bootstrap Testing Powers 30

2.4.5 Testing Powers from Both Methods--1.....32

2.4.6 Coverage of Confidence Intervals from Both Methods--1 33

2.4.7 Testing Powers from Both Methods--2.....34

2.4.8 Coverage of Confidence Intervals from Both Methods--2 34

2.4.9 Testing Powers from Both Methods--3.....36

2.4.10 Coverage of Confidence Interval from Both Methods--3.....36

3.3.1 Type I Errors of Different Methods on Datasets 1 53

3.3.2 Type I Errors of Different Methods on Datasets 2 53

3.3.3 Type I Errors of Different Methods on Datasets 3 54

3.3.4 Type I Errors of Meta-Analysis under t-distribution Approximation 55

3.3.5 The Relationship between *sn* and *df* with Fixed Values of *cn* 57

3.3.6 The Relationship between *cn* and *df* with Fixed Values of *sn* 58

4.1.1 Flow Chart of SGDDP (Huang et al. 2012).....65

4.4.1 Type I Error of the SGDDP when the Proportional Assumption doesn't Hold 83

4.4.2 Power of the SGDDP when the Proportional Assumption doesn't Hold 84

Chapter 1

Introduction

1.1 Motivation

Cardiovascular disease (CVD) is a group of diseases or disorders involving the heart and blood vessels [2]. It also refers to diseases that affect the cardiovascular systems in the human body [3]. Cardiovascular diseases draw a lot of attentions from scientists in different areas not only because they have serious symptoms such as a heart attack or stroke, but more importantly, they are the leading cause of deaths globally [1, 2] and are projected to remain the single leading cause of death at least in the following 10-15 years[4]. Scientists have done and are stilling doing various types of research on this area, trying their best to help people who are suffering from the CVD and also to prevent those at high risk of getting it.

For many years, statistical applications have played a very important role in most of these research activities. For example, statistical data analysis on the information from patients with CVD can help identify the potential risk factors that will increase the probability of CVD. Clinical trials with appropriate designs can evaluate the most efficient intervention or treatment for certain cardiovascular disease or for patients from specific sub-population. And a better utilization of the right statistical methodology for the specific objective in this kind of research would definitely make the outcomes or conclusions more reliable and eventually be beneficial to the human kind. There are already plenty of works have been done on how to apply different statistical methods to medical research

in general [5-7]. However, there are always times when those classical methods don't fit so well for specific practical problems. The motivation of this dissertation is to make contributions to statistical applications in medical research, or more specifically, for cardiovascular disease research by finding better solutions or methods for the practical problems that we found while collaborating with cardiologists at the cardiovascular institute of New Jersey.

This dissertation studies several scenarios in cardiovascular disease researches where traditional statistical methods may not be applicable. And we proposed corresponding practical solutions or modifications to existing methods to make them fit the problems better case by case.

1.2 Assessing Life Extension in Survival Analysis

The Systolic Hypertension in the Elderly Program (SHEP) trial was a randomized, placebo-controlled, clinical trial designed to assess the effect of antihypertensive drug treatment in reducing the risk of stroke in patients with isolated systolic hypertension [8, 9]. While, during the data analysis project whose objective is to show the benefits of treatment in terms of survival time [9], average life extension is believed to be the best statistic to be used here. Because this was a study with a very long follow-up time (22 years) [9] and the objective it to evaluate the treatment effect not only during the clinical trial but also the for the "legacy effect" during the follow-up time, which has already been showed in other studies [10-13]. And the gain in life expectancy or average life exten-

sion is the most intuitive and reasonable statistic to show that effect and it can be easily interpreted.

It turns out that this statistic is equivalent to the difference of mean survival times, or restricted mean survival times [14, 15] in some case scenarios, between two groups. And because the mean survival time, or restricted mean survival time is equal to the area under survival curve [16], we first propose a framework for estimating the average life extension by calculating the area between estimated survival curves given by two groups, treatment and placebo. Both Kaplan-Meier estimator and Cox proportional hazard model are utilized to estimate the survival curves. And then, in order to assess the variability of our estimate, we propose a new Bootstrap method for obtaining bootstrap confidence interval for this quantity, instead of using normal approximation. We also propose the corresponding bootstrap testing procedure to test the null hypothesis that two groups have the same expected survival.

In the estimating step, we estimate the survival curves first via non-parametric Kaplan-Meier estimator to reflect the observed survival probabilities in the study. We then use semi-parametric Cox proportional hazard model and obtain the direct adjusted average survival curves. By doing this, we can adjust for the potential imbalance of covariates between the two treatments and also we can predict survival probabilities.

With the survival curves estimated, the gain of treatment in terms of lifetime extension is estimated by the difference of the area under them. Furthermore, in order to assess the variability of our estimate, the easier way would be estimating the variance of the restricted mean survival time for each curve and then used normal approximation for hy-

pothesis testing or confidence interval calculation. But considering that we don't know the underlying distribution of the true survival time and it may not be normally distributed. The approximation may lead to a less accurate result especially when the sample size is not large. To avoid such potential problem, we propose a new procedure based on bootstrap method for obtaining confidence interval for this quantity. We also propose the corresponding bootstrap testing procedure to test the null hypothesis that two treatments have the same expected survival.

The bootstrap method is specially recommended for small and moderate sample sizes and datasets from clinical trials like SHEP are often large. But if we constrain ourselves to subjects who live longer than a fix period of K years, then the datasets will become smaller and the bootstrap can be very useful. In cardiovascular disease research, this is a very timely application because although it is known that cardiovascular procedures tend to prolong life by solving critical health issues, it is not proven that in the long run, the side effects of the cardiovascular procedures would not worsen the outcome. For example the treatment may onset other diseases such as cancer and the treatment benefits may be short-lived. As part of our ongoing collaboration with the cardiologists, we are currently looking at these issues with our bootstrap method and hopefully we will publish these results in the near future.

1.3 Meta-Analysis with Small Sample Sizes

Meta-analysis is a statistical technique to combine results of individual studies that with similar or related research hypotheses under a given single set of assumptions and condi-

tions. The advantage of this method is to increase the power of the analysis by making the best use of all the information we have gathered across all different individual studies. It is widely used in different research areas including cardiovascular disease research. And there are plenty of good examples in existing literatures. For example, Ernst et al (1993) [17] conducted a meta-analysis with six epidemiologic studies that were found from 1980 to 1992 on fibrinogen and cardiovascular disease to show the association between fibrinogen and subsequent myocardial infarction or stroke. Wald et al (2002) [18] showed the significant associations between homocysteine and three cardiovascular diseases with a meta-analysis using (*a*) 72 studies in which the prevalence of a mutation in the MTHFR gene (which increases homocysteine) was determined in cases (n=16849) and controls, and (*b*) 20 prospective studies (3820 participants) of serum homocysteine and disease risk.

There are several well developed statistical algorithms in meta-analysis to combine the information from different studies, given the types of data sets. One of them is the inverse-variance method, which weights the estimate of the effect of interest in each study by its own variance and adds them together to get the summary effect estimate. This is the most commonly used one for datasets with continuous response variable. And the test statistic is just the summary estimate divided by its standard deviation, which is assumed to follow a standard normal distribution under the null hypothesis. This can be justified be the central limit theorem because it can be considered as a weighted average. And when the number of studies and sample size in each study are large, the approximation should work well.

In most practical studies, researchers usually apply this method directly without considering whether it is appropriate for the particular case. While, unfortunately, people should be more precautious if they want the test to have a correct level of type I error. Because our study shows that when the sample size in each individual study is not large enough, and the number of studies is also not sufficiently large, the distribution of the test statistic is far away from the standard normal distribution. As a result, it could lead to an inaccurate level of type one error. This is not a rare scenario in practice, because the original motivation of a meta-analysis is to more powerfully estimate the true effect size as result from a single study may not be reliable, especially when the sample size is small. Basically, the application of meta-analysis is mostly essential when there is a set of single studies with small sample sizes, which happens sometime in cardiovascular disease researches. As sometimes, the treatment or risk factor we are interested in for CVD is rare, the sample sizes won't be very large. For example, for studying the relationship between hamstring anterior cruciate ligament (ACL) reconstructions in females and magnetic resonance imaging (MRI), we find 4 similar studies with sample sizes smaller than 10 for each single study [19-22].

In this dissertation, we first prove the severity of this problem by a set of simulations. Then in order to avoid that problem, we propose a new method to approximate the underlying true distribution of the test statistic by a t-distribution. We also propose a simple formula to calculate the degree of freedom of the t distribution based on the sample sizes in each study as well as the number of studies.

1.4 Weighted-Z test in Clinical Design

Ethnicity has always been believed to be a factor that has potential impact on the treatment effect. It was not a major concern for a clinical trial when it is performed in a region with the proportions of different races similar across all subareas and the treatment will only be applied in the same region. However, as nowadays, everything goes globe and the distribution of different races are also changing dramatically in some regions. Ethnic factor becomes more and more important in medical research and drug development. And as for the cardiovascular disease research we are focusing here, there are already plenty of studies showed that ethnicity has effect on different CVDs directly or associating with those well known risk factors [61-64]. And in order to control or study the relationship between ethnicity and treatment effect, some statistical methodology has already been proposed. Part of this dissertation will focus on how to combine efficacy information from different ethnic groups to help testing the treatment effect on one targeted ethnic group.

Huang et al (2012) [65] proposed a new clinical design for the Simultaneous Global Drug Development Program (SGDDP) to assess the impact of ethnic factors on the effect of a new treatment for a targeted ethnic (TE) population. The idea is to borrow efficacy information from clinical trials that have already been carried out in other countries or regions, so that the sample sized need for the local clinical trial with patients only from targeted ethnic population can be reduced. While at the same time, the power of the test remains at certain level. They used weighted-Z test to combine the information collected

from the TE and non-TE (NTE) subgroups in the SGDDP based on the fundamental assumption on their shared biological commonality.

Several designs with similar objective have already been proposed in literatures. For example, Hsiao et al. (2005) [66] proposed a two-stage design and provided sample size calculation for the LCT at the second stage. Lan et al. (2005) [67] applied weighted-Z tests to combine the information from the MRCT and the LCT. They all share the same idea as in Huang et al (2012) [65], which is to borrow information from other studies by down weighting it and use a weighted-Z test to combine all the information together. While, the main different here is the way how the information is combined or grouped for the final test.

In those previous designs, they believed that the stage or the location of the study is key factor that has potential impact on the treatment effect. So the information from previous stage should be down weighted. While in the SGDDP design, because the studies in different location and stage have the same key design features, the ethnic factor is the one believed to be the most important. As a result, after all the information is collected from all studies, patients will be grouped based on their ethnicities and those from non-targeted ethnic population will be down weighted.

Overall, to justify the way that the information is used for the final efficacy test, we must to emphasize the fundamental assumption. On one hand, all patients from different ethnic population share certain level of biological commonality so that we can borrow information from other ethnic group to help testing the treatment effect for the targeted group. While on the other hand, there are still potential differences between them and that’s why

we cannot simply treat every patient equally, and have to down weight patients from other ethnic groups. This assumption makes this SGDDP design reasonable but also makes the distribution of the final test statistics much more complicated. It turns out that if we treat it simply as having a standard normal distribution under the null hypothesis, the test would be biased, even if we assume that the endpoints for all ethnic groups having normal distribution.

In order to make the test unbiased, we mathematically formulated the fundamental assumption as proportional treatment effects between both subgroups. We used it to more rigorously describe the hypotheses, and showed the unbiasedness of the weighted-Z test under the new assumption. Moreover, to study the loss of efficiency from down weighting the NTE information in Huang (2012)'s design, we compare the power of their test with that of the uniformly most powerful (UMP) test, which we prove to be also a weighted-Z test. We discuss that the choice of weight should balance the maximization of power when the proportional assumption holds and the minimization of bias otherwise.

Chapter 2

Assessing Life Extension from Medical Interventions

2.1 Introduction

Comparing outcomes of several types of treatments or interventions is an important task in clinical trials as well as retrospective cohort studies in epidemiology. For simplicity and without loss of generality, in our study, we compare two treatments and refer them as the treatment and the control respectively. Among various measurements in assessing the relative efficacy for a time to event outcome, the median survival time and the hazard ratio are two commonly used statistics. While for studies with long time follow-up after the clinical trial, the overall gain in life expectancy is often more of interest, especially when the treatment effect is believed to persist even after the trial. This, also called "legacy effect", has already been discovered in several cardiovascular disease related area, such as hypertension, hyperlipidemia [9-13]. And besides, this quantity is also much straight forward for people to link it with the treatment effect, because it can directly represent the number of days that the participants in the active group lived longer compared with control.

Statistically, this gain in life expectancy can be measured by estimating the average number of days that the survival of treatment patients exceeded that of control patients, or equivalently by the area between the two survival curves from the two treatment arms [23, 24]. While in practice, there is usually a maximum follow-up time, like the end of a clini-

cal trial. The survival curves estimated from those kinds of data are limited up to certain time point $T_{\max}$. As a result, it is hard to estimate the actual mean survival time without extrapolate the survival curve beyond time point $T_{\max}$. Irwin (1949) [14] proposed an estimator for the expectation of life restricted to this time $T_{\max}$, called restricted mean survival time. It is equal to the area under the survival curve form 0 to $T_{\max}$. Then more precisely, the gain in life expectancy obtained from this method should be consider as a restricted gain up to $T_{\max}$.

There are several steps in the estimating and testing of the survival gains: 1) estimating the survival probabilities for the subjects from the two treatment groups; 2) calculating the area under the two curves and its standard error; 3) computing the p-value of the null hypothesis that there is no survival gain between the two treatments.

In the first step, the Kaplan-Meier estimator is the most straightforward method. It represents the observed survival estimates and is quite accurate when the distributions of covariates in the two treatment arms are balanced. When we have imbalance of the distributions, we need to adjust for the covariates with regression models such as semi-parametric model (Cox regression) and parametric models (Weibull, exponential, etc.) [25]. Individual survival curves are predicted from the regression results and expected survival curves can be obtained in several ways such as mean covariate method [26] or direct adjustment [28]. The mean covariate method applies the parameter estimates from a regression to produce one survival curve in each treatment arm for a “typical” participant who assumes average values for all the covariates. Though it is easy to calculate, it lacks good interpretation and can be misleading in some circumstances [26, 27]. The di-

rect adjustment method computes a weighted average of the individual survival curves, with weights proportional to the number of individuals at each level of the covariates. Since it is a clear improvement over the average covariate method, we adopt the approach in our study. As for the choices of regression models, parametric models are only occasionally used in the analysis of survival data although they may offer advantages over Cox model. However, they involve stronger assumptions [29]. Therefore, we will only consider Cox regression model here.

With the estimated survival curves from both treatment arms, the gain of survival is estimated by the difference of the area under them. Furthermore, we would like to estimate the precision of this estimator and conduct statistical hypothesis testing for no treatment effect. Because of the correlation, the variance estimation of the survival curves is quite complicated [30] and it is even more so when we take consideration of entire time region and the difference of the area under two curves. To overcome this problem, we propose to adopt the Bootstrap sampling method [31] to obtain a bootstrap confidence interval of the survival gain and a bootstrap p-value to test a hypothesis that the two treatment arms have the same expected survival.

The rest of the chapter is arranged as follows. After reviewing the basic background in survival analysis, section 2.2 presents the framework to estimate survival curves, the area differences between curves, a bootstrap confidence intervals and a p-value for hypothesis testing. Section 2.3 provides a theoretical proof of consistency of bootstrap estimator. Section 2.4 conducts simulation studies to evaluate the effectiveness of this framework. Section 2.5 analyzes a real data set from a clinical trial. Section 2.6 concludes the chapter with discussion and future research directions.

2.2 Methods

2.2.1 Notations

Assume that there are *n* subjects receiving a same treatment in a study, which studies an event of interest, for example, death due to some cause. Let T_i denote the survival time for the *i*th subject. Assume that T_i, ..., T_n are continuous random variables identically distributed and with a cumulative distribution function $F(.)$ and a density function $f(.)$. Define the survival function

$$S(t) = P(T > t) = 1 - F(t) = \int_t^{\infty} f(u)du \qquad (2.2.1)$$

Since time to event data is sometimes censored due to end of the follow-up period of the study or dropout of subjects from the study, we generally observe a sample of pairs (T_i, δ_i), $i = 1, ..., n$ where $\delta_i = 1$ if the subject has an event and $\delta_i = 0$ if the subject is censored. Note that there are several types of censorship [33]. In this dissertation we focus on the right censoring type of time to event data. In addition we observe a list of covariates denoted by X_i that identify the demographic and medical characteristics of the *i*th patient.

2.2.2 Estimating survival functions

Assume the observed times for the n subjects are $t_1 \le t_2 \le \ldots \le t_n$. For each t_i, we denote n_i as the number of subjects who are at risk just prior to time t_i, and d_i as the number of events at time t_i. The Kaplan-Meier estimator [32] is the nonparametric maximum likelihood estimate of $S(t)$ with a product of the form

$$\hat{S}(t) = \prod_{t<t_i} \frac{n_i - d_i}{n_i} \tag{2.2.2}$$

Note that when there is no censoring, n_i is just the number of survivors just prior to time t_i. With censoring, n_i is the number of survivors less the number of losses (censored cases). It is only those surviving cases that are still being observed (have not yet been censored) that are "at risk" of an (observed) event.

The Kaplan-Meier is quite accurate when the distributions of covariates in the two treatment arms are balanced. When we have imbalance of the distributions, we need to adjust for the covariates with regression models. We define the hazard function $h(.)$ as

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}\log S(t), \tag{2.2.3}$$

which is the ratio of the density function to the survival function. Hence, the hazard function is related to the survival function: $S(t) = \exp[-\int_0^t h(u)du]$. The semi-parametric Cox proportional hazard model [34] incorporates covariates X in the form:

$$h(t|X) = h_0(t)\exp(\beta^T X), \tag{2.2.4}$$

where β is a vector of regression coefficients and $h_0(t)$ is a baseline survival function. The survival function can be written in terms of a base survival function $S_0(t) = \exp[-\int_0^t h_0(u)du]$:

$$S(t|X) = \exp(-\int_0^t h_0(u)\exp(\beta^T X)du) = S_0(t)^{\exp(\beta^T X)} \tag{2.2.5}$$

This can be estimated by $\hat{S}(t|X) = \hat{S}_0(t)^{\exp(\hat{\beta}^T X)}$ where $\hat{S}_0(t)$ is the estimated baseline survival function by the Aalen-Nelson estimator [35] and $\hat{\beta}$ is the estimated coefficient from Cox regression based on a partial likelihood approach.

Based on the direct adjustment method, we can obtain the average survival curves for the subjects in one treatment arm by averaging the individual curves as

$$\hat{S}(t|X) = \frac{1}{n}\sum_{i=1}^{n}\hat{S}_0(t)^{\exp(\hat{\beta}^T X_i)} \tag{2.2.6}$$

When some of the predictors do not satisfy proportional hazards assumption, we may stratify them to get around the problem. In the case of a stratified Cox regression model [36], the above becomes

$$\hat{S}(t|X) = \frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\hat{S}_0^j(t)^{\exp(\hat{\beta}^T X_{ij})} \tag{2.2.7}$$

where J is the number of strata, n_j is the number of subjects in the jth stratum, n is the total number of subjects, $\hat{S}_0^j(t)$ is the estimated baseline survival function for the jth stratum.

2.2.3 Restricted mean survival time

One important issue in analyzing survival data is to compare the survival function $S_{trt}(t)$ of a treatment group with that of a control group $S_{ctr}(t)$. The quantity to compare here is the expectation of time to event variable T. Since

$$E(T) = \int_0^\infty uf(u)du = \int_0^\infty (1 - F(u))du = \int_0^\infty S(u)du, \tag{2.2.8}$$

which is the area under the survival curve, therefore treatment survival gain (TSG) is defined as

$$TSG = E(T_{trt} - T_{ctr}) = \int_0^\infty (S_{trt(u)} - S_{ctr}(u))du \tag{2.2.9}$$

which is the area between two survival curves from the two treatment groups.

In order to estimate the $E(T)$ or TSG, we have to first estimate the survival curves from 0 to ∞. However in practice, this cannot be achieved directly most of the times. Because in most studies, there is a maximum follow-up time and the survival curves estimated from those studies are only up to certain time point T. There are different methods proposed to deal with this problem [35, 37, 38]. In this chapter, we are going to use restricted

mean survival time proposed by Irwin (1949) [14]. Let $T_{\max}$ be the maximum follow-up time, then the restricted mean of survival time up to $T_{\max}$, $\mu(T_{\max})$ is defined as

$$\mu(T_{\max}) = E(\min(u, T_{\max})) = \int_0^{T_{\max}} S(u)du. \tag{2.2.10}$$

Then, the TSG based on restricted means from both arms will simply follow by its definition:

$$TSG = \mu_{trt}(T_{\max}) - \mu_{ctr}(T_{\max}) \tag{2.2.11}$$

And $S(u)$ here can be estimated by either KME or Cox regression model. While, do notice that when covariates are involved, (2.2.6) or (2.2.7) shows that the estimated survival function depends on the values of X. When the distribution of X is not balanced for the two treatment arms, the KM estimator approach produces misleading results which compares the survival gains between two difference groups of subjects.

2.2.4 Estimation of the area between two survival curves

We estimate the TSG in the following steps:

1. Obtain the two estimated curves $\hat{S}_{trt}(t)$ and $\hat{S}_{ctr}(t)$ either by their corresponding Kaplan-Meier estimators or by direct adjusted survival curves from Cox regression described in Section 2.2.2. In general, the two estimated curves are expressed as step functions. We assume that $\hat{S}_{trt}(t)$ takes value y_i in the interval $[t_{i-1}, t_i)$ for

$i = 1, ..., n$ and $t_0 = 0$; $\hat{S}_{trt}(t)$ takes value z_j in the interval $[s_{j-1}, s_j)$ for $j = 1, ..., m$ and $s_0 = 0$. And assume there is a maximum follow-up time, or end time for whole study ω, where $\omega \geq \max(s_m, t_n)$.

2. Suppose that $t_n < \omega$ then we set the interval for the integral to be $[0, t_{n^*}]$, with $t_{n^*} = \omega$, and $y_{n^*} = y_n$. Essentially we extend the estimated survival curve of the treatment group to the maximum follow-up time ω based on the last point of that curve. Note that $n^* = n + 1$. Same procedures will be taken if $s_m < \omega$.
3. Estimate TSG by the trapezoidal method over the interval $[0, \omega]$:

$$\widehat{TSG} = \sum_{t=1}^{n^*} y_i (t_i - t_{i-1}) - \sum_{j=1}^{m^*} z_j (s_j - s_{j-1}) \qquad (2.2.12)$$

This algorithm uses the fact that the estimated survival curves in real data analysis regardless which estimating method is utilized are step functions and hence the area under the curve can be calculated without error as a sum of rectangular areas.

2.2.5 Confidence interval and p-value by normal approximation

By the procedure above, the estimate for the treatment gain in terms of life extension can be easily calculated. The next step would naturally be making some statistical inference based on that to see how precise the estimate is or whether the TSG is statistically significant. In order to do that, we need to know the underlying distribution of the estimate or at least an approximate one. However, by (2.2.12) we can see that the distribution of $\widehat{TSG}$

won't be in a simple form, especially when the underlying distribution of time to event is unknown. But since this statistic is still a mean of survival times in a general way, normal approximation would definitely to be considered first.

Now let's just consider the simplest case here. Suppose y_i and z_j here in (2.2.12) are all from Kaplan-Meier estimator, then the variance of the restricted mean can be estimated by

$$\hat{V}(\hat{u}) = \sum_{i=1}^{D} \left[\int_{t_i}^{\omega} \hat{S}(t)dt \right]^2 \frac{d_i}{n_i(n_i - d_i)} \tag{2.2.13}$$

in general [39]. Here the $\hat{S}(t)$ is the Kaplan-Meier estimate of the survival function. $t_i s$ are the event times and ordered as $0 < t_1 < t_2 < \cdots < t_D$ and d_i are the number of individuals with an event at corresponding event time t_i. While n_i is the number of individuals at risk just before event time t_i. And since $\widehat{TSG}$ is just a linear function of two independent restricted means, the variance of $\widehat{TSG}$ the summation of the variances from both treatment and control groups calculated by (2.2.13). And once we have $\hat{V}(\widehat{TSG})$, the procedure for hypothesis test and confidence interval calculation would easily follow.

2.2.6 Bootstrap confidence interval and p-value

Since $\widehat{TSG}$ is just a simple function of Kaplan-Meier estimates from two groups, the large sample behavior would be similar to KME, which has already been shown to have an asymptotic normal distribution [41, 42]. So the hypothesis test and confidence interval

based on normal approximation discussed in previous section should work well when the sample size is large. However, we can't make any comments on that when the sample size is small, especially, when we don't have any information regarding the true distribution of the survival time. In those case scenarios, Bootstrap method would be a good alternative because it is commonly used for estimating the variance or confidence interval for an estimate when the underlying distribution is unknown.

Efron [31] proposed to bootstrap the survival function by sampling the pairs of censoring indicators and observed times to event with replacement. He also showed that this is equivalent to sample from the distribution of survival times (denote x_i^* as the samples), and sample from the censoring time (denote u_i^* as the samples), and then assign $t_i^* = \min(x_i^*, u_i^*)$, $\delta_i^* = 1$ if $t_i^* = x_i^*$ and 0 otherwise. This algorithm has been applied by Utzek and Sanchez [43] to estimate a bootstrap confidence envelop of the survival curve.

Denote the upper bound in follow-up times for both arms as ω. Here we apply Efron's algorithm to estimate a confidence interval of the area between two survival curves as follows:

1. Using the algorithm in Section 2.2.4 to calculate the estimate $\widehat{TSG}_{obs}$ with the upper bound in time for both arms as ω.
2. Use Efron's method to select two bootstrap samples $\{(t_i^*, \delta_i^*),\ i = 1, ..., n\}$ and $\{(s_j^*, v_j^*),\ j = 1, ..., m\}$ from the treatment group and the control group respectively. Order the pairs by the t_i^* s and the s_j^* s and estimate the survival curves for the both bootstrap samples. If $t_n^* < \omega$, we add the point (y_n^*, ω) to the estimated sur-

vival curve for the treatment sample and do the same operation to the estimated survival curve for the sample from the control group.

3. Calculate the survival gain from the bootstrap samples - $\widehat{TSG}^{(b)}$ over the interval $[0, \omega]$ using algorithm described in Section 2.2.4.
4. Repeat steps 2 and 3 B (at least 1000) times and order the estimates increasingly as $\{\widehat{TSG}_i^{(b)} : i = 1, ..., B\}$ from which we estimate the sample quantiles as $\widehat{TSG}_{0.025B}^{(b)}$ and $\widehat{TSG}_{0.975B}^{(b)}$. Then $(\widehat{TSG}_{0.025B}^{(b)}, \widehat{TSG}_{0.975B}^{(b)})$ is a 95% bootstrap confidence interval for TSG.

We also propose a similar algorithm for testing the null hypothesis that $TSG = 0$ or there is no survival gain from treatment comparing to the control, *vs* the one-sided alternative hypothesis that $TSG > 0$. This algorithm is an adaptation of the general bootstrap testing algorithm that can be found in [44]. We modify the above algorithm as follows

- 2*. Using Efron's method to select two bootstrap samples $\{(t_i^*, \delta_i^*),\ i = 1, ..., n\}$ and $\{(s_j^*, v_j^*),\ j = 1, ..., m\}$ both from the control. If $t_n^* < \omega$, we add the point (y_n^*, ω) to the estimated survival curve for the treatment sample and do the same operation to the estimated survival curve for the sample from the control group.
- 4*. Repeat steps 2* and 3 B times, where B is a large number at least 1000. Observe the sample $\{\widehat{TSG}_1^{(b)}, ..., \widehat{TSG}_B^{(b)}\}$ from which we estimate the bootstrap one-sided p-value as the $\#(\widehat{TSG}_j^{(b)} > \widehat{TSG}_{obs}) / B$.

2.3 Asymptotic Properties of the TSG Estimators

Although, the asymptotic properties of Kaplan-Meier estimates, the general bootstrap procedure and results from that procedure applied on survival analysis have already been studied in different literatures [45-48], we still need to show the asymptotic properties or convergence of the estimate based on our procedure since we made several modifications comparing to the general method. And it will help to justify the confidence interval and hypothesis testing based on our procedure.

Denote $\widehat{TSG}*$ as the bootstrap version of $\widehat{TSG}$ in (2.2.12). We will show that 1) $\widehat{TSG}$ is an asymptotically consistent estimator; 2) $\widehat{TSG}*$ converges to $\widehat{TSG}$ in the same way as $\widehat{TSG}$ converges to TSG asymptotically. Note in 1) it suffices to show that in each arm, the area under the estimated survival curve is a consistent estimator for the expected survival time. Denote MSD_{trt} as $\int_0^T S_{trt}(u)du$ for the treatment group where T is the end of follow-up and $\widehat{MSD}_{trt}$ as $\sum_{i=1}^{n^*} y_i(t_i - t_{i-1})$. We need to show that $\widehat{MSD}_{trt}$ is a consistent estimator of MSD_{trt}. With 2), the bootstrap confidence interval based on $\widehat{TSG}*$ will provide accurate coverage for the true value and the bootstrap p-value is available to conduct hypo these testing. We first use Kaplan-Meier estimator for the survival function. Then we extend the conclusion to the direct adjusted survival curves based on Cox regression.

First assume that S_{trt} is absolutely continuous from 0 to T and $(y_n, t_{n^*} = T)$ is added if $t_n < T$. There are two different situations about we should consider.

1. $S_{trt}(t)$ is strictly decreasing around time point T

   Note that $\sum_{i=1}^{n} y_i(t_i - t_{i-1})$ is just a simple function of the *KME*, which is a uniformly consistent estimator of S_{trt} in $[0, t_n]$[49]. As a result, the sum converges to $\int_0^{t_n} S_{trt}(u)du$ in probability.

   When $t_n = T$, by definitions, $\widehat{MSD}_{trt}$ is a consistent estimator for MSD_{trt}.

   When $t_n < T$, decompose MSD_{trt} and $\widehat{MSD}_{trt}$ as:

$$MSD_{trt} = \int_0^{t_n} S_{trt}(u)du + \int_{t_n}^{T} S_{trt}(u)du \tag{2.3.1}$$

$$\widehat{MSD}_{trt} = \sum_{i=1}^{n} y_i(t_i - t_{i-1}) + y_n(T - t_n) \tag{2.3.2}$$

   Since $dS_{trt}(T^-)/dt < 0$, we have density $f_{trt}(T^-) > 0$. It follows that $P(\lim_{n\to\infty} t_n = T) = 1$. Therefore the second terms in (2.3.1) and (2.3.2) will converge to zero with probability one with the finite values of $S_{trt}(t)$ and y_n. Then the proof is completed.

2. $S_{trt}(t)$ is constant around time point T

   Now that the density function $f_{trt}(T^-) = 0$, there are no observations prior to T in a close neighborhood. Define $T^* = \arg\max\{ f_{trt}(t) : f_{trt}(t) > 0\}$, then we can decompose MSD_{trt} and $\widehat{MSD}_{trt}$ as:

$$MSD_{trt} = \int_0^{t_n} S_{trt}(u)du + \int_{t_n}^{T^*} S_{trt}(u)du + S_{trt}(T^*)(T - T^*) \tag{2.3.3}$$

$$\widehat{MSD}_{trt} = \sum_{i=1}^{n} y_i(t_i - t_{i-1}) + y_n(T^* - t_n) + y_n(T - T^*) \tag{2.3.4}$$

Using the same arguments as in 1, the first two terms in (2.3.4) converges to those in (2.3.3) consistently. In addition, since $S_{trt}(t)$ is continuous, $P(\lim_{n\to\infty} S_{trt}(t_n) = S_{trt}(T^*)) = 1$. And since y_n converges to $S_{trt}(t_n)$, it converges to $S_{trt}(T^*)$ as well. Therefore, the third term in (2.3.4), $y_n(T - T^*)$ converges to that in (2.3.3) consistently. With that, the proof is completed.

The asymptotic properties of the bootstrap version of *KME* have been studied in [50, 51]. Denote $\widehat{S_{km}(t)}$ as the Kaplan-Meier estimator of a survival function $S(t)$ and $\widehat{S^*_{km}(t)}$ as its bootstrap version. Then $\sqrt{n}[\widehat{S^*_{km}(t)} - \widehat{S_{km}(t)}]$ converges to $\sqrt{n}[\widehat{S_{km}(t)} - S(t)]$. Since $\widehat{MSD}^*_{trt}$, $\widehat{MSD}_{trt}$ and MSD_{trt} are simple functions of $\widehat{S^*_{km}(t)}$, $\widehat{S_{km}(t)}$ and $S(t)$, it follows that $\sqrt{n}[\widehat{MSD}^*_{trt} - \widehat{MSD}_{trt}]$ converges to $\sqrt{n}[\widehat{MSD}_{trt} - MSD_{trt}]$. The conclusion holds for the control group as well. Therefore $\sqrt{n}[\widehat{TSG}^* - \widehat{TSG}]$ converges to $\sqrt{n}[\widehat{TSG} - TSG]$ asymptotically. This validates the bootstrap confidence interval and p-value that generated by our algorithms.

All the arguments above are based on *KME* which is nonparametric estimator. While, there already have been some applications on using parametric or semi-parametric models for estimating restricted mean survival time to adjust for the covariates [52, 53]. Let's take Cox model for example. Under general conditions, the direct adjusted survival curve

from the Cox model converges to KME in probability [54]. The estimated survival function and its bootstrap counterpart based on Cox regression will converges in probability to their KME versions respectively. Therefore the asymptotic properties for the Cox regression based survival estimators remain the same as that for the KME based ones.

2.4 Simulations

In this section, we present simulation studies to demonstrate that proposed procedures effectively estimate survival gains and its confidence interval as well as p-value for hypothesis testing. Intuitively, factors such as number of subjects in a trial and censoring rates of lifetime will directly affect the estimation accuracy and power. Also, since our estimator is restricted by the follow-up time, it is very likely that the $T_{\max}$ may affect the power and the performance of our procedure. Besides, how the survival curve is estimated under different circumstances would also be important. We study whether and how they all relate to the performance of our procedure so that future users can decide when and where to apply it.

2.4.1 Bias of estimation from KM Estimator and Cox regression with balanced design

We simulate samples of a size n for both treatment and control arm with life time and censoring time from some known distributions. The true TSG can be calculated. We compute the estimated TSG based on KME and Cox regression respectively. Repeat this operation for N times, we get the average TSG and its standard error. We vary n and re-

peat the whole steps. We also use censoring time with different distributions. The purpose of this simulation is to show the different results from the two methods in estimating survival functions and study the biases.

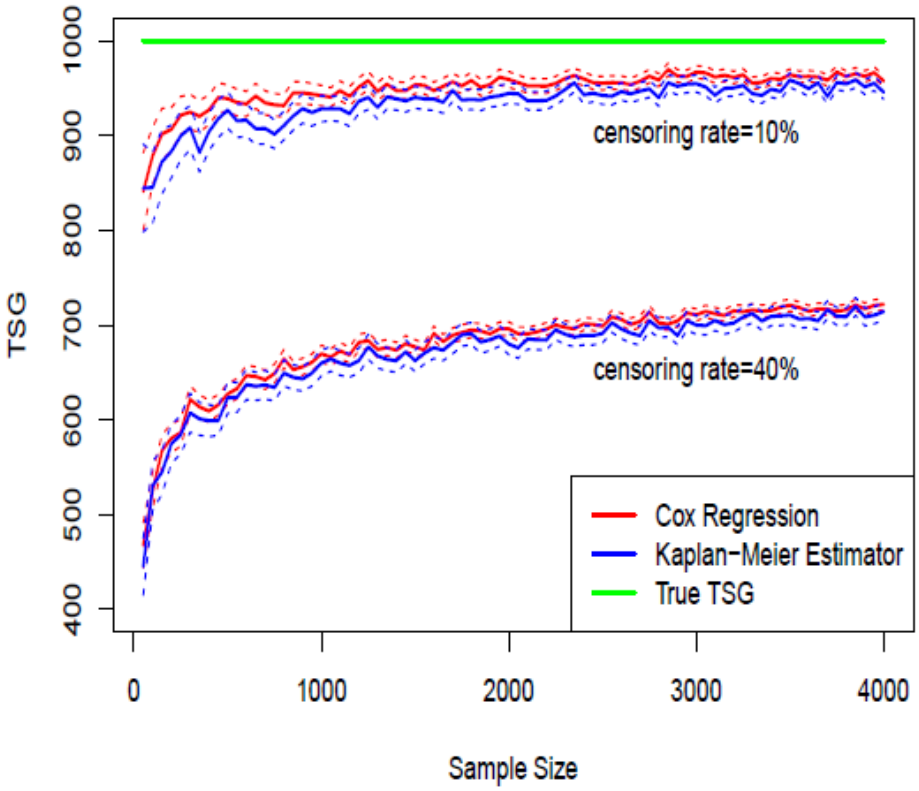
Assume life time $y_{trt}(t) \sim \exp(\gamma_{trt})$ for a treatment group, where $\gamma_{trt} = \gamma_0 \exp(\alpha + \beta x_{trt})$; life time $y_{ctr}(t) \sim \exp(\gamma_{ctr})$ for a control group, where $\gamma_{ctr} = \gamma_0 \exp(\beta x_{ctr})$; covariate $x_{trt} \sim normal(\mu_{trt,}\sigma_{trt}^2)$ for the treatment group; covariate $x_{ctr} \sim normal(\mu_{ctr,}\sigma_{ctr}^2)$ for the control group; and censoring time $y_{cen}(t) \sim \exp(\gamma_{cen})$. Then the theoretical survival gain is

$$E(y_{trt}) - E(y_{ctr}) = 1/\gamma_0 \exp(-\alpha - \mu_{trt}\beta - 0.5\beta^2\sigma_{trt}^2) - 1/\gamma_0 \exp(-\mu_{ctr}\beta - 0.5\beta^2\sigma_{ctr}^2). \quad (2.4.1)$$

We vary n from 50 to 4000. For each n, we generate the y_{trt} or y_{ctr} (denoted by y) with the equal probability using $\alpha = \log(0.5)$, $\beta = 1$, $\mu_{trt} = \mu_{ctr} = 2$, $\sigma_{trt} = \sigma_{ctr} = 1$, $\gamma_0 = 2.23 \times 10^{-4}$, and $\gamma_{cen} = 10^{-4}$. Then the observed time to event $t = \min(y, y_{cen})$, censoring indicator $\delta = I\{y \geq y_{cen}\}$ where $I(.)$ is an indication function which takes the value of 1 when the argument is true, and 0 otherwise. For each value of n, calculate $\widehat{TSG}_{km}$ and $\widehat{TSG}_{cox}$. Repeat this for 2000 times to get the mean of estimates and their standard errors. The true TSG is calculated as 1000 with censoring rate of 0.11. Thus we plot in Figure 2.4.1 the true TSG, $\widehat{TSG}_{km}$, $\widehat{TSG}_{cox}$ and their 95% confidence bands. We observe that 1) when sample size increases, the estimates from both methods get close to the true value; 2) Cox regression tends to achieve less bias and produce narrower confidence bands though the two methods do not differ significantly.

To examine the effects of censoring, we repeat the simulations using the same settings as above except for $\gamma_{cen} = 7.14 \times 10^{-4}$. The censoring rate increases to 40%. We plot the results in the same figure and observe that the bias is bigger when the censoring is severer.

Figure 2.4.1 *TSGs* from Two Methods--1



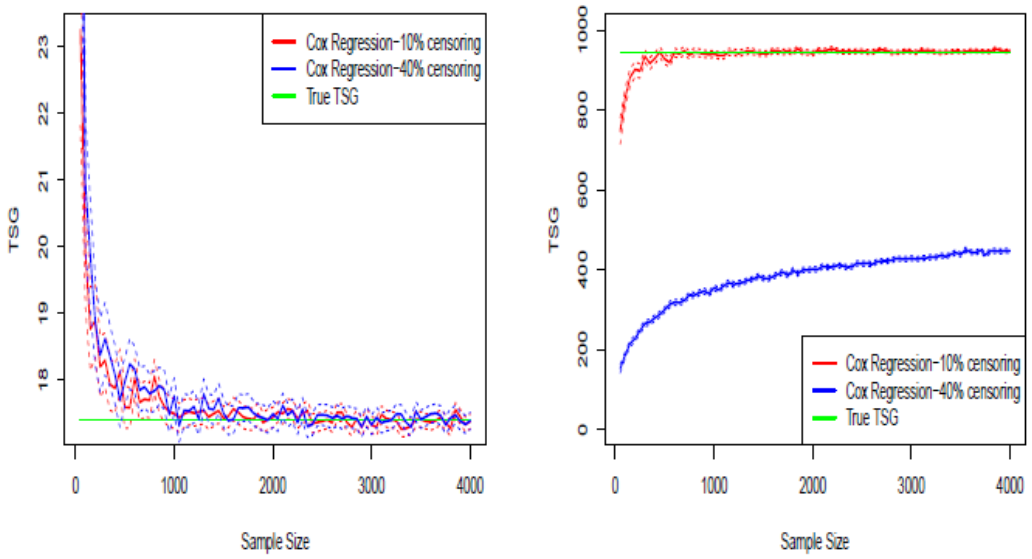
2.4.2 Bias of estimation from Cox regression with imbalanced design

When the design is not balanced in the two treatment arms, the result from *KME* is not easy to interpret. Cox regression method, on the other hand, can be applied to estimate the *TSG* on a particular subgroup of subjects from both arms.

Assume that X in the active treatment arm takes values of 0 and -2 with equal probability, 0 and 2 in the control arm. We might be interested to know the TSG when X takes on -2 or 2. To do so, we can fit a Cox regression model to available data, then using the concept of "counter-factual" by assigning a different treatment type to the same subgroup of subjects.

Similar setup as the previous setting except for the construction of covariate X and $\gamma_0 = 7.8\times10^{-3}$, the true overall $TSG = 1000$, the true TSG for the subgroup is 945.0 for $X = -2$, 17.3 for $X = 2$. With $\gamma_{cen} = 1/4000$ the censoring is set to 10%. We repeat the simulations with $\gamma_{cen} = 1/350$ to achieve 40% censoring. The estimated $TSGs$ for both subgroups are shown in Figure 2.4.2.

Figure 2.4.2 $TSGs$ from Two Methods--2



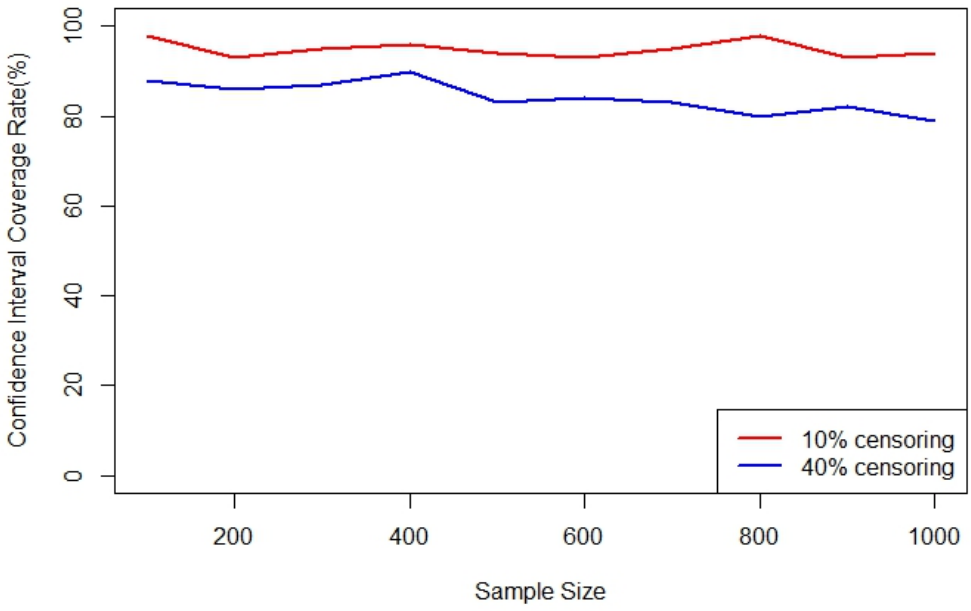
(a) X=2 (b) X=-2

It can be observed that the censoring rate makes difference in estimating TSG for subset $X = -2$ while no difference for the other subset. The reason is because of the different lifetime distributions.

2.4.3 Coverage of bootstrap confidence intervals and power of bootstrap testing

We want to study the accuracy of the Bootstrap method. We concentrate on the case when the covariates are balanced in both treatment arms and use the KME approach. Using the similar setup, we use Bootstrap steps to obtain a 95% confidence interval (CI) and check if the CI covers the true TSG. Repeat these steps 100 times, we can calculate the percentage of a correct coverage of the CI. We plot the results in Figure 2.4.3

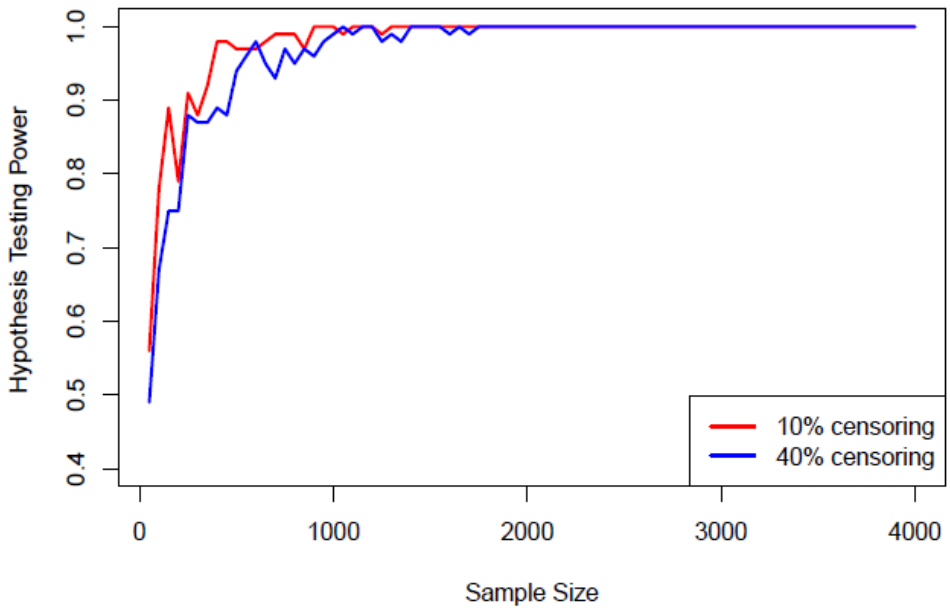
Figure 2.4.3 Bootstrap Confidence Interval Coverage Rates



We observe that censoring rate of the life time data plays an important role in the correct coverage of the Bootstrap CI. For example, with a low censoring rate of 10%, the coverage percentage achieves around 90 to 95% even with moderate sample size. However, the higher rate of 40% makes the coverage percentage a little bit lower.

Next, we evaluate the Bootstrap p-value calculation for hypothesis testing. We use the same simulation setting as before and obtain the percentage of Bootstrap p-value less than 0.05 as the power of hypothesis test in Figure 2.4.4. We observe that the testing power increases rapidly to 0.9 and above with the sample size and low censoring rate data achieves high powers than the high censoring rate data. We also try other settings with different true TSG with similar power curves.

Figure 2.4.4 Bootstrap Testing Powers



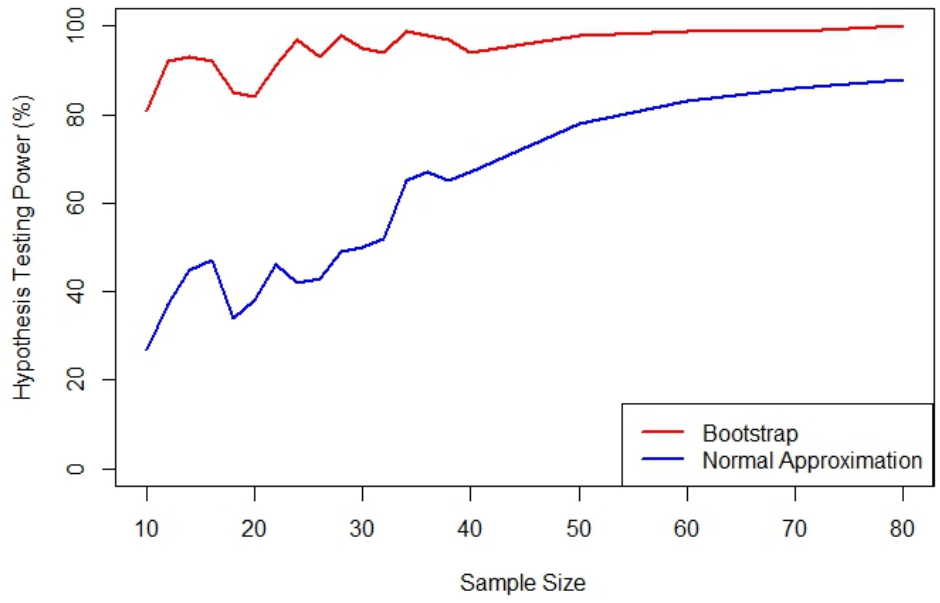
2.4.4 Comparing results from bootstrap method and normal approximation

One of the motivations that we propose this bootstrap procedure is that we believe the statistical inference based on bootstrap is more reliable than simply using the normal approximation, especially when the sample size is small. In this section, we are going to compare the coverage of confidence intervals and the powers of hypothesis testing from both methods by simulation with relatively small sample sizes. And we also want to study whether the underlying distribution of the time to event and the censoring distribution will affect their performances. In this part of the study, we only focus on using KME to estimate the survival curve.

We start with simple setting where there is no censoring distribution, only end of the study. The survival times are randomly generated from χ_2^2 and χ_3^2 , and then multiplied by 1000, for the control and treatment group respectively. The maximum follow-up time is $T_{\max} = 4000$. By (2.2.10) and (2.2.11), the true *TSG* can be easily calculated using numerical integrating method such as the "integrate( )" function in R, which equals to 164. And we set the sample sizes for both groups to be equal and changing from 10 to 80 gradually. Still, for each sample size setting, we repeated the simulation 100 times to get the percentage of the confidence intervals which cover the true *TSG* and the tests reject the null hypothesis, which is equivalent to power. The results are showed below in Figure 2.2.5 and Figure 2.2.6 correspondingly.

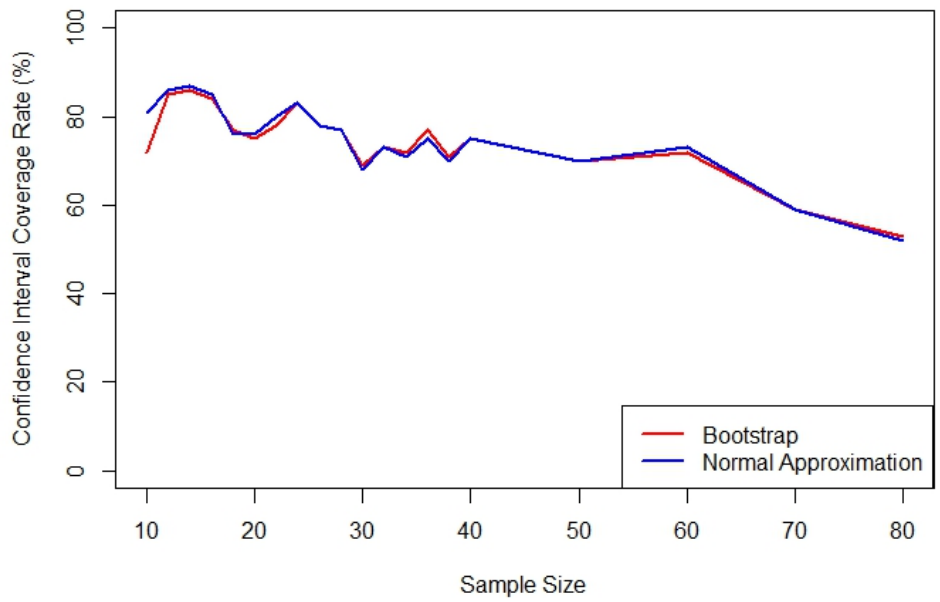
From Figure 2.2.5, we can see that the bootstrap test performs very well. Even when the sample size is very small, only above 10, its power is still higher than 80%. While the power from the normal approximation test is always lower than that from the bootstrap test. And the different is significant when the sample size is small, almost 40% to 80% when the sample size is between 10 and 20. Although as the sample size increasing to above 50, the difference is getting much smaller. But overall, the bootstrap test does out-perform the normal test in this set of simulation and the advantage is obvious.

Figure 2.4.5 Testing Powers from Both Methods--1



On the other side, the coverage of the confidence intervals from both methods seems to be almost the same as showed in Figure 2.2.6. This means that the confidence interval and hypothesis test are not equivalent [56]. Besides, as the sample size increasing, the coverage rate drops a little bit.

Figure 2.4.6 Coverage of Confidence Intervals from Both Methods --1



Now, we change the data a little more complicated by adding an underlying censoring distribution. This time, we use the setup similar to that we used in section 2.4.3. Let $y_{trt}(t) \sim \exp(\gamma_{trt})$ with $\gamma_{trt} = 1/3311$, $y_{ctr}(t) \sim \exp(\gamma_{ctr})$ with $\gamma_{ctr} = 1/1656$ and the censoring time $y_{cen}(t) \sim \exp(\gamma_{cen})$ with $\gamma_{cen} = 1/1000$. The maximum follow-up time is set to be $T_{\max} = 5000$. Then the theoretical survival gain is $TSG = 144.5$. Similar simulations are performed as above and the results are plotted in Figure 2.4.7 and Figure 2.4.8.

Similar as Figure 2.4.5, Figure 2.4.7 also shows that the bootstrap test performs much better comparing to the normal test even when the sample size is moderate large, like 80. And moreover, the power difference between two methods does not become smaller as sample size increasing, as what we observed in Figure 2.4.5.

Figure 2.4.7 Testing Powers from Both Methods--2

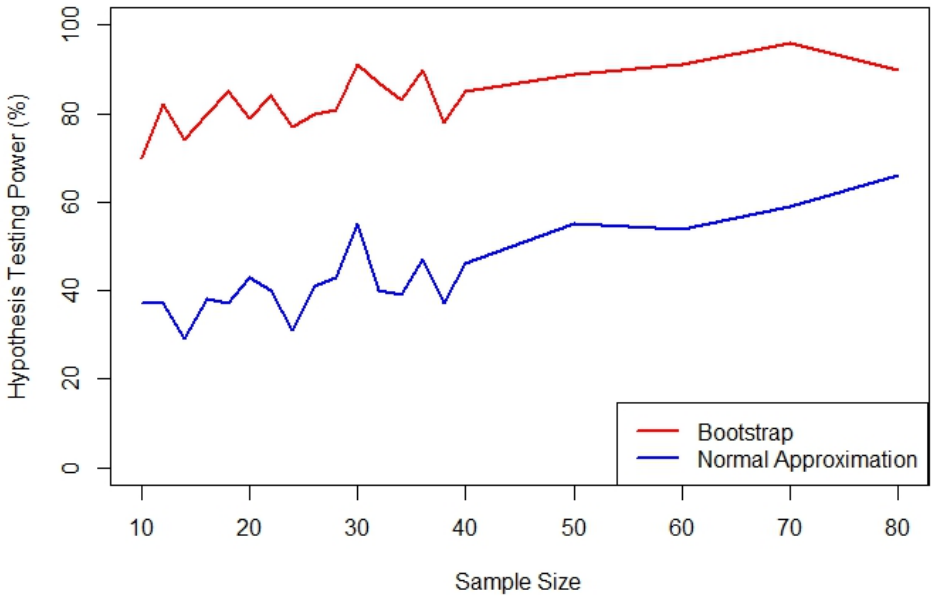
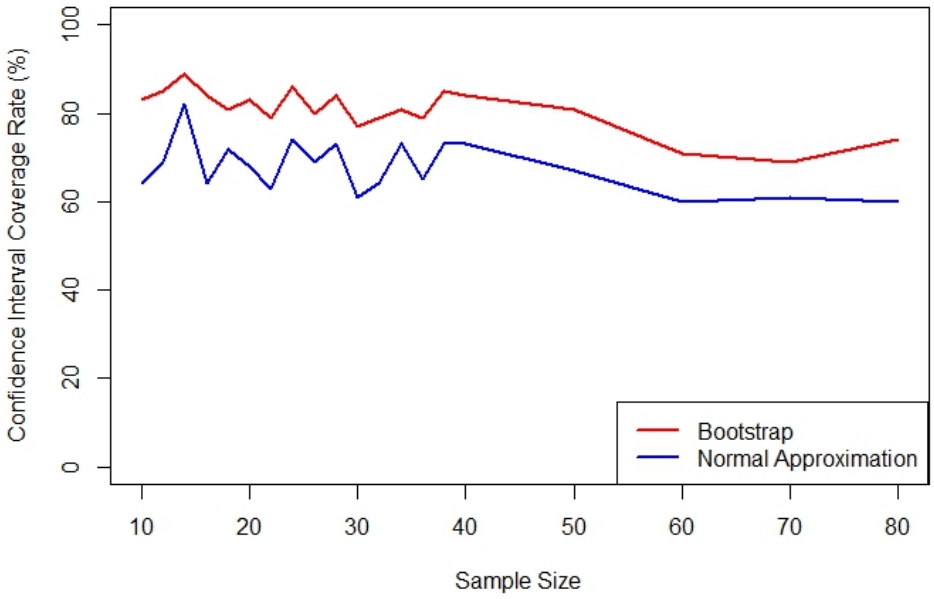


Figure 2.4.8 Coverage of Confidence Intervals from Both Methods--2



While, Figure 2.4.8 shows a different pattern from that in Figure 2.4.6. In this second set of simulation, the confidence intervals from bootstrap method do have a higher coverage rate than that from normal approximation. The coverage rates from both methods and the difference between them almost stay the same when the sample size increases.

The previous two sets of simulations show that the test from bootstrap always works better than that from normal approximation. While in terms of confidence interval, on the data generated from chi-squared distribution without censoring, those two methods work almost the same. But on the data from exponential distribution and with censoring, the confidence from bootstrap has higher coverage. The might indicate that the underlying distribution of survival time and the censoring mechanism do have some influence on how these two methods perform.

So, in the third set of simulations, we keep the most settings same as that in the second set of simulations. The only thing we change is the censoring distribution, which become as $y_{cen}(t) \sim \exp(\gamma_{cen})$ with $\gamma_{cen} = 1/1500$. This means that comparing to the second set of simulations, the average censoring rate here should be lower. But the true *TSG* stays the same. The results are plotted in Figure 2.4.9 and 2.4.10.

Comparing Figure 2.4.9 to Figure 2.4.7, we can see that the patterns of the power change for both methods stay the same, as well as the difference between them. The only difference is that, as the censoring rate lowers, the powers from both methods increase a little bit but not significant. As for the coverage rate in Figure 2.4.10, it decreases from both methods as sample size increases. But the result from bootstrap method still has a higher rate all the time.

Figure 2.4.9 Testing Powers from Both Methods--3

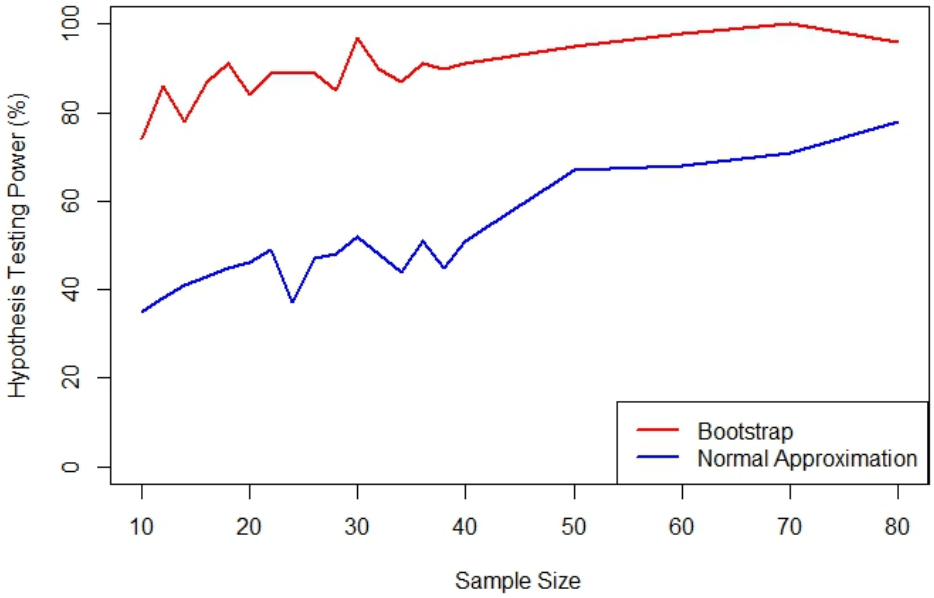
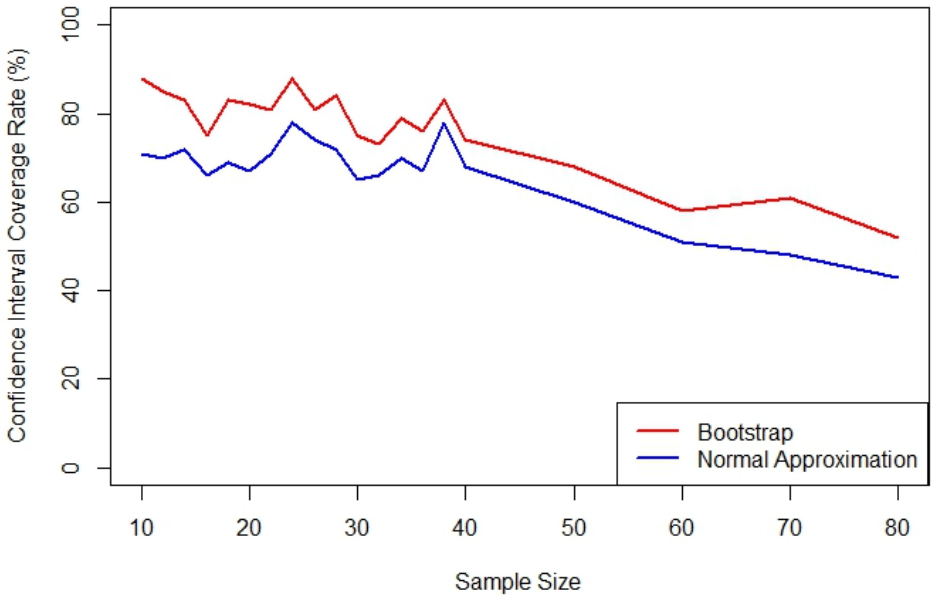


Figure 2.4.10 Coverage of Confidence Interval from Both Methods --3



2.5 Applications to a Real Dataset

2.5.1 Data Description

The Systolic Hypertension in the Elderly Program (SHEP) was a randomized, double blinded placebo controlled trial in older patients with isolated systolic hypertension with the primary endpoint of fatal or non-fatal stroke. The investigators randomized 4736 participants (56.8% women) with systolic blood pressure (SBP) 160 mm Hg or higher and diastolic blood pressure 90 mm Hg or lower to stepped care antihypertensive therapy based on chlorthalidone or matching placebo.

Initially, all participants received chlorthalidone (or placebo) to decrease the SBP by at least 20 mm Hg and to below 160 mm Hg. If this was not achieved, a second drug, atenolol, (reserpine if atenolol was contraindicated) or placebo was added. During an average follow-up of 4.5 years there was a significant decrease in stroke with relative risk (RR) of 0.63 (95% confidence interval of 0.49 to 0.82) but the non-significant effects on all cause (RR=0.87, 0.73-1.05), cardiovascular (0.80, 0.60-1.05) or non-cardiovascular (1.05, 0.80-1.38) mortality.

After the end of the study all participants were advised to take active treatment at the discretion of their physician. Recruitment begun on March 1, 1984 and vital status, date of death and cause of death were ascertained using the NDI through the end of 2006. The total duration of follow-up was 21 years and 10 months. Death was classified as cardiovascular if it was due to *International Classification of Diseases, Ninth Revision* codes

290 to 459 or *International Statistical Classification of Disease, 10th Revision* codes I00 to I99.

2.5.2 Analyses and Results

In order to access the efficacy of the treatment, we are interested in estimating the net gain in life expectancy free from cardiovascular death in the active therapy group by calculating the area between survival curves of the two interventions expressed as the mean number of days that the survival of a patient in the active treatment group exceeded that of a patient in the placebo group.

First, we fit Kaplan-Meier survival curves for cardiovascular (CV) death only on the treatment group and control group. We calculate the areas under the two curves and take the difference and follow the Bootstrap steps to obtain the confidence interval and p-value for testing whether there is no difference in the two groups in term of survival gains. We repeat the same set of analyses for the end point of all-cause mortality. The results are shown in Table 2.5.1:

Table 2.5.1 *TSG* from Kaplan-Meier Approach

		Bootstrap		
End Point	*TSG*	Mean	95% CI	p-value
All-Cause Death	104.7	105.6	(-39.2,241.8)	0.073
CV Death	158.9	157.6	(36.4,286.6)	0.009

Next we use Cox partial regression approach to correct any imbalance of the covariates between the two treatment groups. After using all the significant variables and checking the proportional hazard assumption, we stratify age using two categories of older than 71 and the rest, and race with 3 categories of white, black and others. The two covariates are sex and indicator of whether the patient previously has myocardial infarction (histmi). For the end point of cardiovascular death, we stratify age and use sex and race as covariates. The results are in Table 2.5.2.

Table 2.5.2 *TSG* from Cox Partial Regression

End Point	Covariates (besides treatment)	*TSG*	Bootstrap Mean	Bootstrap 95% CI	Bootstrap p-value
All-Cause Death	sex, histmi (race and age-stratified)	64.9	67.1	(-62.6,190.7)	0.158
CV Death	sex, race (age-stratified)	146.0	145.5	(14.9,276.1)	0.016

From the results in Table 2.5.1 and 2.5.2, chlorthalidone reduces CV death significantly and does not reduce all-cause mortality. The Cox partial regression approach achieves a slightly tighter confidence interval of the *TSG* for CV death. The estimated *TSG* for all-cause death differs quite a lot from the two methods.

2.6 Discussions

This chapter proposes a bootstrap-based method to estimate the survival gain of a treatment *vs*. its control and assess the precision of this estimator. In the estimating step, the Kalplan-Meier approach is straightforward and less computational intensive. Under the assumption of balanced study covariates, we can use this approach to estimate the survival gains for a similar group of participants. However, when this assumption does not hold, the Kaplan-Meier method produces misleading results. To solve this problem, we propose an alternative Cox partial regression approach. With that, we can deal with the possible imbalance between the two groups.

Furthermore, the ultimate goal of proposing this method is that we can make inference of the survival gain of a hypothetical participant or group of participants with similarity to the counterfactual causal inference framework. And because of the properties of bootstrap method, the underlying distribution of the time to event and the size of the sample would not have large influence on the results based on our procedures, which is often the case when we simply use a normal approximation to get the p-value or confidence interval. And based on our simulation results, our bootstrap procedure does have advantage over the normal approximation. Not only our method has a significant higher power of testing but also has higher coverage from the confidence interval, even when the sample size is moderate large.

In our study, we propose to use the restricted mean of survival time to avoid making any assumptions regarding the survival curve past the maximum follow-up time. But it also

brings limitation. Because it limits our estimate to only assess the life extension or treatment gain up to this up time bond. However, the ideal goal or ultimate objective for a study like SHEP is to assess the overall life extension brought by the treatment. In order to do that, we must find a good way to estimate and extrapolate the rest part of survival function which beyond follow-up. By doing so, we can more accurately access the survival gain of a treatment. And this should be further explored.

Chapter 3

Meta-Analysis with small sample sizes

3.1 Introduction

Meta-analysis is a statistical technique to combine results of individual studies that with similar or related research hypotheses. The advantage of method is increasing the power of the analysis by making the best use of all the information we have gathered across all those individual studies. There are several well developed statistical algorithms for different type of datasets [58, 59]. One of them is the inverse-variance method which is also the most widely used one. It is used to combine the effect estimates, such as log odds ratios, risk differences, mean differences, etc., together by weighted averaging with weights equal to the variance of the effect estimates. And the weighted average is called the summary effect estimate from the set of studies. Then to test the summary effect, the test statistic is calculated by dividing the summary effect estimate by its standard deviation. This statistic is usually considered as following a standard normal distribution, at least approximately or asymptotically.

While using meta-analysis to a real data application, we discovered that when the number of observations in each study and the number of studies are not sufficiently large, the distribution of the test statistic could be far away from the standard normal distribution. And we believe that researchers should pay attention to this when they are doing meta-analysis. Although, in the literature, most previous meta-analysis have been published do

have large sample sizes, we also have examples with small sample sizes. Like the one we were trying to perform. We want to study the relationship between hamstring anterior cruciate ligament (ACL) reconstructions in females and magnetic resonance imaging (MRI). And we only find 4 similar studies with sample sizes smaller than 10 for each single study [19-22]. Similar situation could also be found in cardiovascular disease researches, especially when the treatment or the risk factor of interest is rare. These all make it important to take a close look of the true distribution of the summary statistic in meta-analysis with small sample sizes. After all, the application of meta-analysis is essential when there is a set of single studies with small sample sizes, and the evidence gathered from each study individually doesn't have enough statistical power to prove anything.

In this chapter, we are going to show the non-normal behavior of the summary test statistics in meta-analysis, when the number of observations in each study and the number of studies are not sufficiently large. For the purpose of simplicity, we chose the difference of means between two groups as the effect size that we are interested in, and assume that the dataset is the combination of the individual study data with the same or similar two treatment groups.

We show the results in two steps. First, we prove that if we use the pooled standard error from different studies to get the variance of the mean difference for each study, the test statistic actually followed a t-distribution with certain degree of freedom. Second, if the variance of the mean difference is estimated from individual study data, then the true distribution is not easy to specify but surely not close to standard normal distribution when the sample size is not large. It will then be proved by the result from simulations.

Then, in order to control the type I error of the test, we propose to use t-distributions to better approximate the underlying true distribution of the test statistic with certain degree of freedom. And a simple formula is also proposed to calculate the degree of freedom.

3.2 Method

3.2.1 Data structure

Suppose the meta-analysis was carried on a set of I studies with continues outcome and two treatment groups. Then the number of observations, the means and the standard deviations of the response in each study can be showed in the table below.

Table 3.2.1 Data from Study i

Study i	Group size	Mean	Standard Deviation
Treatment	K_{1i}	m_{1i}	sd_{1i}
Control	K_{2i}	m_{2i}	sd_{2i}

Then the treatment effect we are interested in is the difference of the means between treatment and control groups: $\hat{\theta}_i = m_{1i} - m_{2i}$. And the standard error of that can be estimated by

$$SE\{\hat{\theta}_i\} = \sqrt{MSE * (\frac{1}{K_{1i}} + \frac{1}{K_{2i}})} \tag{3.2.1}$$

or

$$SE\{\hat{\theta}_i\} = \sqrt{\frac{sd_{1i}^2}{K_{1i}} + \frac{sd_{2i}^2}{K_{2i}}}. \tag{3.2.2}$$

Here, the MSE is the mean square error calculated based on all the datasets in the meta-analysis.

3.2.2 Inverse-variance method

In meta-analysis, there are different ways to define the weight for the effect estimate in each study. For inverse-variance method, it uses the estimated variance of the main response. Taking a data with structure as in Table 3.2.1 as an example, the individual effect estimate $\hat{\theta}_i$ is weighted by

$$w_i = \frac{1}{\left(SE\left\{ \hat{\theta}_i \right\} \right)^2} \tag{3.2.3}$$

Then apply the weight defined in (3.2.3) to every $\hat{\theta}_i$ to get the summary estimate as a weighted mean, which is given by

$$\hat{\theta} = \frac{\sum w_i \hat{\theta}_i}{\sum w_i} \tag{3.2.4}$$

with

$$SE\left\{\hat{\theta}\right\} = \frac{1}{\sqrt{\sum w_i}} \tag{3.2.5}$$

As a result, the final test statistic is for testing the treatment effect across all studies in the meta-analysis will be:

$$z = \frac{\hat{\theta}}{SE\left\{\hat{\theta}\right\}} = \frac{\sum w_i \hat{\theta}_i}{\sqrt{\sum w_i}} \tag{3.2.6}$$

3.2.3 Using pooled standard error from all studies

Obviously, in this inverse-variance method showed above, how the variance or standard error of $\hat{\theta}_i$ is estimated will directly affect the calculation of weight for study i. And as we already showed in previous section, there are two ways to do that which is by either (3.2.1) or (3.2.2). While, in order to apply (3.2.1), we have to calculate the Mean Square Error (*MSE*) first, which can be obtained as follow:

$$MSE = \frac{\sum_{i=1}^{I} sd_{1i}^2 \times (K_{1i} - 1) + sd_{2i}^2 \times (K_{2i} - 1)}{\sum_{i=1}^{I} \left(K_{1i} + K_{1i}\right) - (I - 1) - 1} \tag{3.2.7}$$

Then, after the *MSE* is available, we can get weight for each study by

$$w_i = \frac{1}{\left(SE\left\{ \hat{\theta}_i \right\} \right)^2} = \frac{1}{MSE \times (\frac{1}{K_{1i}} + \frac{1}{K_{2i}})}. \tag{3.2.8}$$

And the final test statistic would become

$$z = \frac{\sum_{i=1}^{I} w_i \hat{\theta}_i}{\sqrt{\sum_{i=1}^{I} w_i}} = \frac{\sum_{i=1}^{I} (m_{i1} - m_{i2})}{\sqrt{MSE \sum_{i=1}^{I} (\frac{1}{K_{i1}} + \frac{1}{K_{i2}})^{-1}}} \tag{3.2.9}$$

which follows t-distribution with degree of freedom $\sum \left(K_{1i} + K_{1i} \right) - I$. This should be equivalent to the test for the treatment effect in the generalized linear model, when study factor is treated as a simple $I - 1$ dimensional categorical variable.

3.2.4 Using standard error from individual study

If in some situation, we don't think using the pooled standard error is such a good approach, maybe because we believe that all studies are not that similar. Then we can use (3.2.2) to calculate the standard error for $\hat{\theta}_i$. And the weight for study i becomes

$$w_i = \frac{1}{\left(SE\left\{ \hat{\theta}_i \right\} \right)^2} = \frac{1}{\frac{sd_{1i}^2}{K_{1i}} + \frac{sd_{2i}^2}{K_{2i}}} \tag{3.2.10}$$

And the overall test statistic would become

$$z = \frac{\sum_{i=1}^{I} w_i \hat{\theta}_i}{\sqrt{\sum_{i=1}^{I} w_i}} = \frac{\sum_{i=1}^{I} (\frac{sd_{1i}^2}{K_{1i}} + \frac{sd_{2i}^2}{K_{2i}})^{-1} (m_{i1} - m_{i2})}{\sqrt{\sum_{i=1}^{I} (\frac{sd_{1i}^2}{K_{1i}} + \frac{sd_{2i}^2}{K_{2i}})^{-1}}} \quad (3.2.11)$$

with a distribution that is not easy to be described. And when $K_{ij}s$ are not so large, it will distribute far away from the standard normal distribution.

3.2.5 Letter values approximation

Since the true distribution is complicated and certainly non-normal, to get a better control of the type I error of the final test in practice, we must find other standard distribution which is closer to the true distribution and get a better approximation then using standard normal distribution. While, t-distribution is good candidate since in 3.2.3, the test statistics actually do belong to that family. And we find there is a way to calculate the degree of freedom of the close t-distribution to any empirical distribution, which is called letter values approximation. The idea is to get the spread of the distribution of interest at each letter value, which is the $1/2^m$ quantile ($m = 2,... 10.$) and calculate the ratios between these and that from normal distribution. Then based on those ratios, there is an empirical formula to get the degree of freedom of close T-distribution.

Here are the main steps:

- Get the Letter Values and Spread from the distribution of interest.
- Calculate the ratios of the Spreads for targeted distribution over standard normal distribution.

- Fit a linear model between $\log_2(ratio)\log 2$ and $-\log_2(tail\ probability)$ and get the slope the fitted line.
- Degree of freedom of a close t-distribution: $1/2+1/(2.25\times slope)$.

3.2.6 Finding an empirical formula for the degree of freedom

Based on this letter values approximation, we can find a close t-distribution for certain test statistic in (3.2.11), with specific number of sample sizes. But it still cannot be used directly in practice. So, with this objective and the idea from letter values approximation method, we propose the following procedure to find an empirical formula for the degree of freedom of a close t-distribution. First of all, we simulate the distributions of the test statistic with different number of studies and different number of observations in each study. Then, using the letter values approximation, we get the degree of freedom of the t-distribution which is close to the true distribution under each sample size setting. And based on these data, we try different formulas and find the best one to describe the relationship between number of studies, number of observations and degree of freedom.

3.2.7 Data simulation

All data sets used in this chapter are simulated by SAS with the model

$$Y_{ijk} = \gamma_i + \mu_j + \varepsilon_{ijk} \tag{3.2.12}$$

where Y_{ijk} is the main response of observation in study i and treatment j with γ_i, μ_j and ε_{ijk} representing study effect, treatment effect and random error respectively. We use different numbers of observations to see how they affect the type I error. Also, we simulate datasets with different combinations of number of studies and number of observation in each study to find the relationship between the distribution and those numbers.

3.3 Simulation Results

In the section, we will first show some simulation results to prove that no matter which method you use to calculate the weight ((3.2.8) or (3.2.10)), the type one error would not be the same as we desired, which support our statement that using a standard normal approximation is not a good approach, especially when the sample sizes are relatively small. Then, after our motivation is fully justified, we are going to show the results by using the method we proposed to get a better approximation with t-distribution. And finally, we will present the empirical formula we find for calculation the degree of freedom.

3.3.1 Type I error

First of all, Table 3.3.1 shows the different settings of sample sizes we use in our simulations. They changes from very small to acceptable large in terms of rule of thumb for normal approximation. Then, Table 3.3.2 shows the result of testing the treatment effect

on the first five replicates of the simulation with sample sizes from "Data sets1", using both PROC GLM and Meta-Analysis with the weights calculated based on MSE. We set $\mu_1 = \mu_2 = 11$ and γ_i for $i = 1$ to 6 has he value of (1.5, 1.0, 0.5, -0.5, -1.0, -1.5) correspondingly. And $\varepsilon_{ijk}s$ are random numbers from $N(0, 25)$. It is clear to see that results from both approaches are exactly the same, which confirms our conclusion in section 3.2.3.

Table 3.3.1 Sample Sizes in Different Simulation Sets

	Data sets 1		Data sets 2		Data set3	
Study#	Group1	Group2	Group1	Group2	Group1	Group2
1	3	4	7	8	30	35
2	4	2	5	9	35	20
3	5	3	7	7	40	40
4	6	3	10	8	25	30
5	3	4	8	4	30	20
6	5	4	6	9	25	30

Table 3.3.2 Comparing results from PROC GLM and Meta-analysis with MSE

	PROC GLM			Meta-Analysis using MSE		
Replicate	Estimate	S.D.	t_value	Estimate	S.D.	z_value
1	-0.40	1.54	-0.26	-0.40	1.54	-0.26
2	-2.13	1.40	-1.52	-2.13	1.40	-1.52
3	1.29	1.77	0.73	1.29	1.77	0.73
4	-3.62	1.45	-2.50	-3.62	1.45	-2.50
5	2.22	1.54	1.44	2.22	1.54	1.44

Then, in Figure 3.3.1, we show the type I errors from PROC GLM and meta-analysis with both ways of calculating weights. Here, in order to show the severity of the type I error inflation, the sample sizes used in this first set of simulations are very small and unbalanced (Table 3.3.1 under Data sets 1) and we replicated 50 times with each replication consisting 1000 individual simulation to calculate the type I error. Notice that "Meta1" here represents the type I error of using the test statistics calculate by (3.2.9) and treating it as standard normal, while "Meta2" is based on the calculation from (3.2.11). Clearly, the type I error from "Meta2" is around 0.2 and sometimes even close to 0.25 which is way above the desired level (0.05). This may be somehow extreme scenarios. But it does support our concerns that simply treating the weighted average as standard normal will cause serious problem sometime.

Another interesting point worth noticing is that, although the values of test statistics in "GLM" and "Meta1" are exactly the same as proved in Table 3.3.2, the type I errors from "Meta1" are higher. It is simply because, instead of compared to the true t-distribution as it is in "GLM", the test statistic in "Meta1" is incorrectly compared to standard normal distribution. However, the difference is not very significant because the degree of freedom of the true t-distribution it follows (t-distribution with $df = 40$) is fairly large in spite of the small sample sizes in individual studies.

And if we increase the sample sizes in each individual study, the type I errors, especially that of "Meta2", will become smaller and closer to 0.05 level as in Figure 3.3.2. And if the sample sizes are large enough, there won't be much difference between these three methods (Figure 3.3.3). Details of sample sizes are also in Table 3.3.1. And same simulation parameters are used as for Figure 3.3.1.

Figure 3.3.1 Type I Errors of Different Methods on Datasets 1

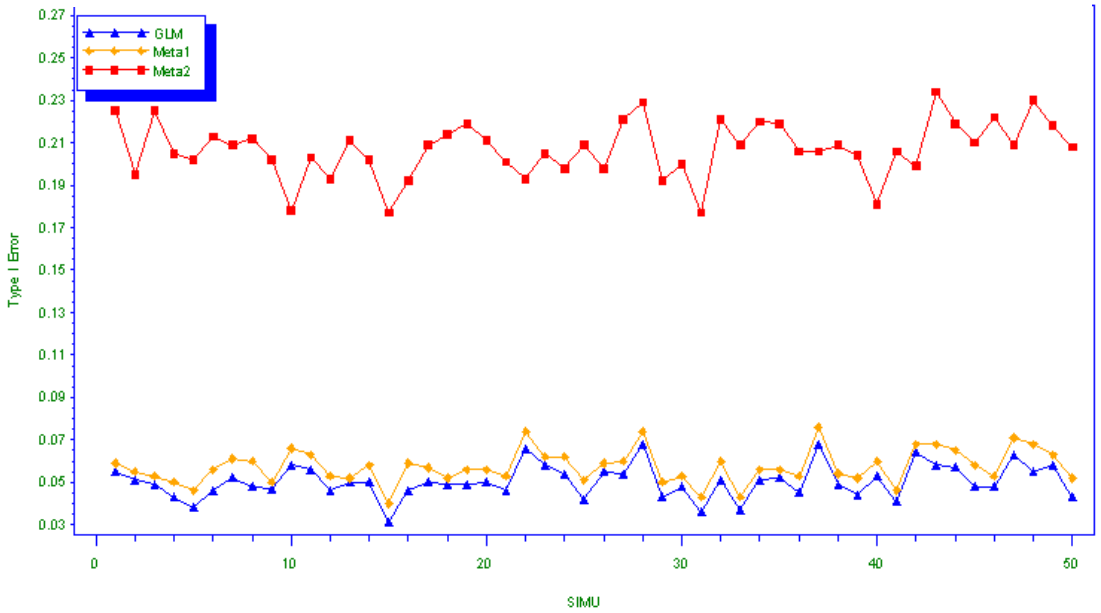


Figure 3.3.2 Type I Errors of Different Methods on Datasets 2

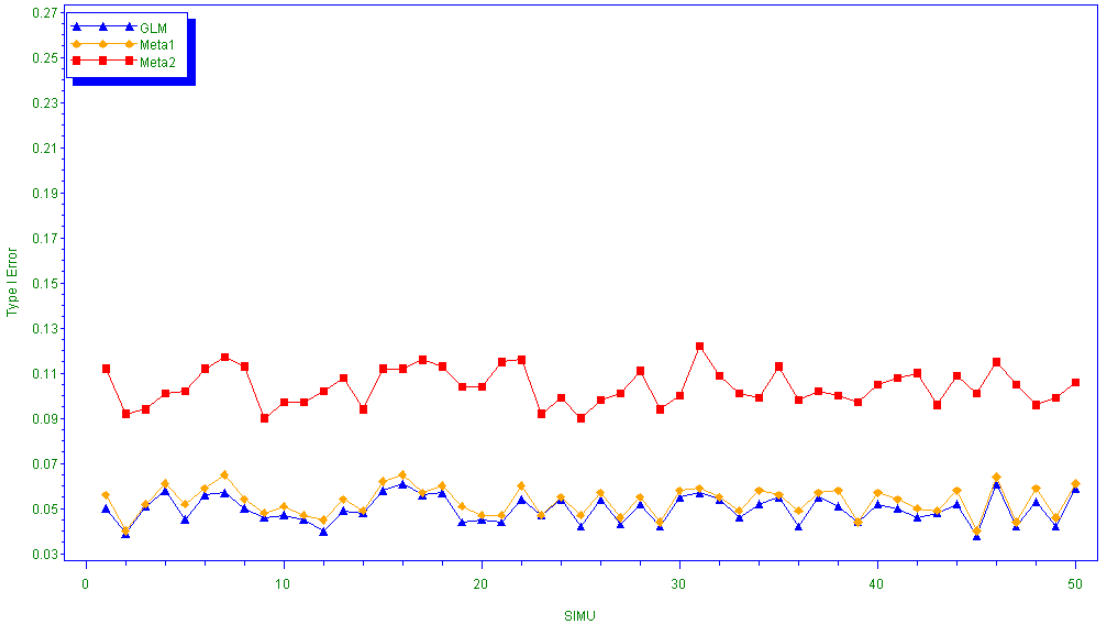
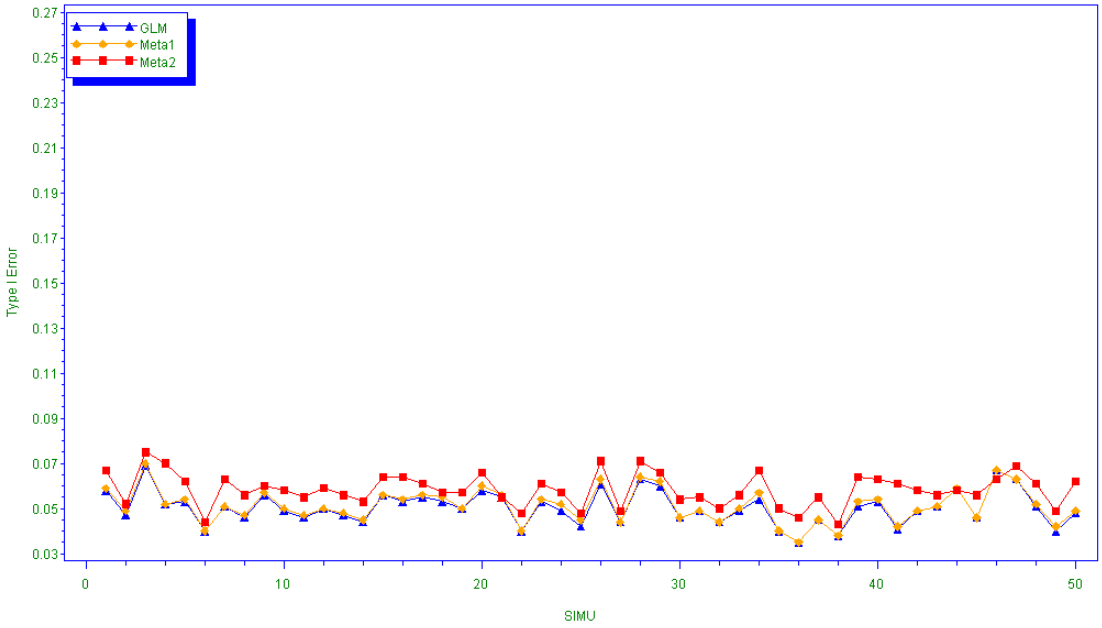


Figure 3.3.3 Type I Errors of Different Methods on Datasets 3



3.3.2 The t-distribution approximation

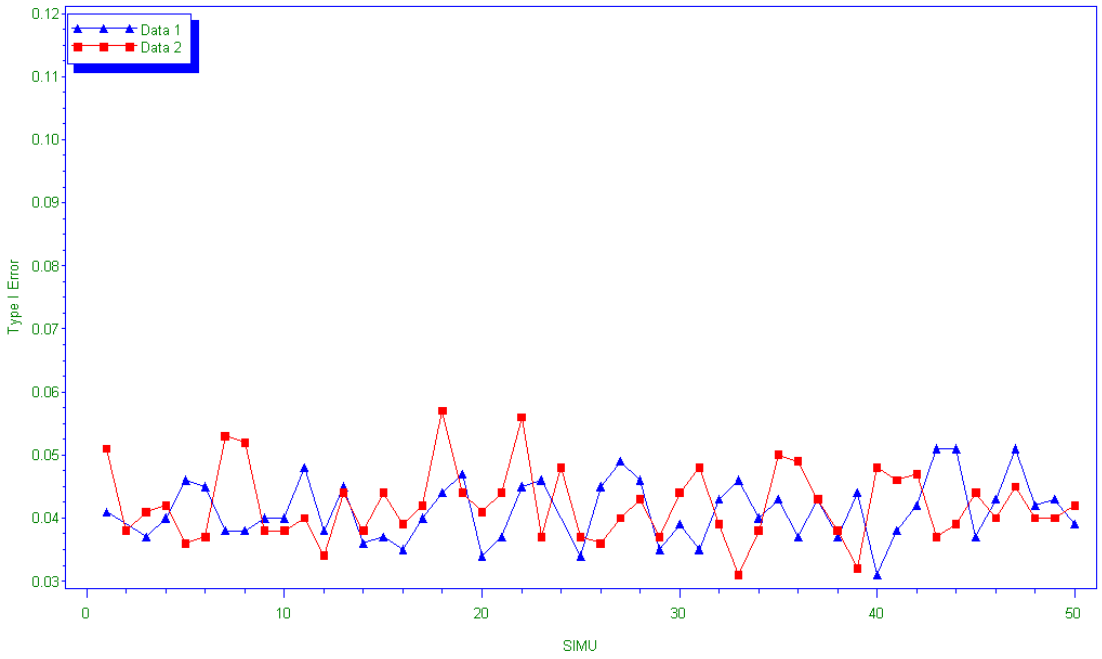
For both sample size settings “datasets 1” and “datasets 2”, we actually simulated 500000 (50×1000) times each and calculated same numbers of the final test statistic values which leads to the type I error plots we see above. They also gave us very good empirical distribution of the test statistic from method “Meta2” under two different simple size settings. We can get the letter values and theirs spreads of these two distributions easily. And then after applying the letter value approximation method, we got the approximated degree of freedom for them as 4 for that from “datasets 1” and 22 for “datasets 2”, along with scales of 1.3 and 1.2 respectively. And from the type I error curves in Figure 3.3.4, we can say that the approximation works very well.

Table 3.3.3 Letter Value Spreads and Ratios of Test Statistic from D1 and D2

	Distribution			Ratio to Normal	
Letter Value	Normal	D 1	D 2	D 1	D 2
Fspread	1.35	2.00	1.58	1.48	1.17
Espread	2.30	3.53	2.71	1.53	1.18
Dspread	3.07	4.92	3.68	1.60	1.20
Cspread	3.73	6.31	4.54	1.69	1.22
Bspread	4.31	7.76	5.31	1.80	1.23
Aspread	4.84	9.55	6.08	1.97	1.26
Zspread	5.32	11.68	6.72	2.20	1.26
Yspread	5.77	14.28	7.39	2.48	1.28
Xspread	6.19	17.27	8.15	2.79	1.32

*Here, D 1 and D 2 stands for Datasets 1 and Datasets 2

Figure 3.3.4 Type I Errors of Meta-Analysis under t-distribution Approximation



3.3.3 Formula for calculating degree of freedom

The above results show that if we use the t-distribution with the degree of freedom from the Letter Values Approximation method as the standard distribution to compare with, the type I error from "Meta2" will be well controlled to the desired level. This gives us enough reason to keep using t-distribution to approximate the true distribution of the test statistics. But in order to make it more practical, we need to find an easy formula to calculate the degree of freedom (df) base on the observed data, mostly the samples sizes, which including the number of studies (sn) and number of observations in each study and each treatment group (cn), since we don't have the empirical distribution to start with in real data analysis.

To make it simple, here we only consider the situation that each study and each group has the same number of observations. We simulate sets of data with different combinations of sn and cn, and applied our approximation method on each datasets to get the corresponding degree of freedoms. Results are listed in Table 3.3.4.

Then, we try different simple models with df as response variable and sn, cn as covariates and picked the best one, which is

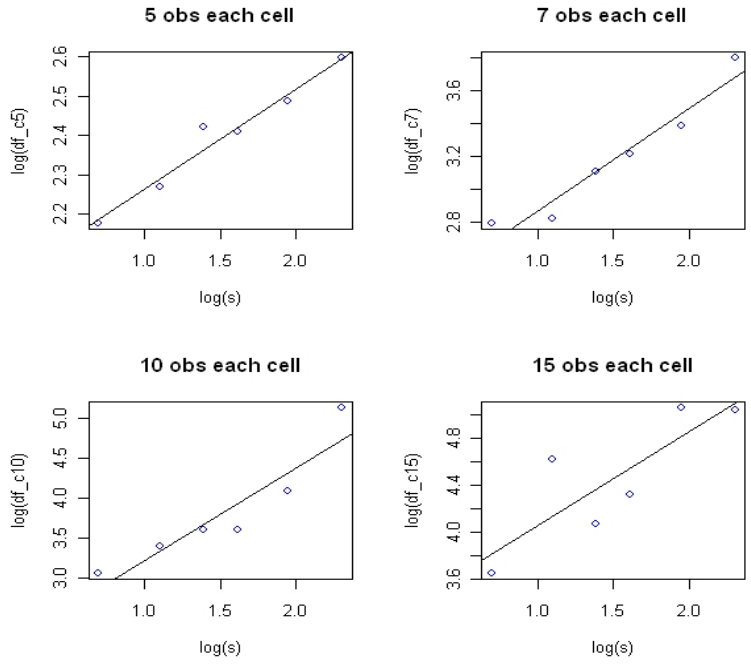
$$\log(df) \sim \log(sn) + \log(cn) + \log(sn) \times \log(cn) \qquad (3.3.1)$$

But still, there are different combinations of coefficients can be chose for the final formula. We then test the formula on another set of simulated data to see which one works better.

Table 3.3.4 Approximate Degree of Freedom from Different Sample Size Settings

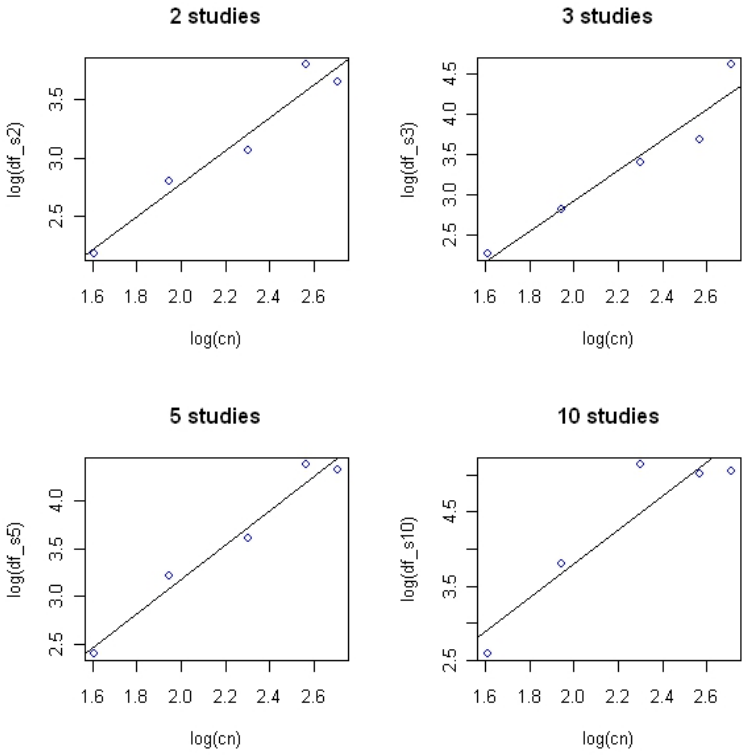
	# of Observations in each Cell (*cn*)				
#of Studies(*sn*)	5	7	10	13	15
2	8.826	16.354	21.547	44.742	38.692
3	9.682	16.824	29.817	40.016	101.209
4	11.253	22.311	36.855	70.412	58.504
5	11.131	24.798	37.047	79.148	75.046
7	12.010	29.439	59.461	69.206	156.837
10	13.409	44.503	168.717	148.854	154.026

Figure 3.3.5 The Relationship between *sn* and *df* with Fixed Values of *cn*



*Here, X is $\log(sn)$ and Y is $\log(df)$

Figure 3.3.6 The Relationship between cn and df with Fixed Values of sn



*Here, X is $\log(cn)$ and Y is $\log(df)$

Table 3.3.5 Testing Results of df from Different Formulas

		df from Different Methods			
sn	cn	DF 1	DF2	DF3	DF4
6	6	18.2	19.6	20.3	19.9
6	8	28.9	34.8	33.6	33.4
6	12	131.9	78.4	68.3	69.2
6	14	61.0	106.7	89.4	91.4
9	6	23.5	29.4	25.9	25.9
9	8	40.7	52.3	42.8	43.4
9	12	73.4	117.6	87.1	90.1
9	14	1919.8	160.0	114.1	118.9

*DF1: Using Letter Value Approximation; DF2: $df \leftarrow sn \times cn^2 \times \exp(-2.4)$

DF3: $df \leftarrow sn^{0.6} \times cn^{1.75} \times \exp(-1.2)$; DF4: $df \leftarrow sn^{0.65} \times cn^{1.8} \times \exp(-1.4)$

Table 3.3.5 shows that the when *sn* and *cn* are small all three formulas work very well. But as they becoming larger, sometimes, the approximation from those formulas works not so well. However, when sample sizes or number of studies are large, it is not necessary to apply our method anyway.

3.4 Simple data example

Now let's apply our method to the example we mentioned earlier. So we have four similar datasets all about studying the correlation between magnetic resonance imaging (MRI) cross-sectional area measurements and the intraoperative graft size in hamstring anterior cruciate ligament (ACL) reconstructions [19-22]. The original summary statistics used in all for studies were the person correlation. But for the illustrating purpose, we use the Fisher's transformation and get the summary statistics as below in Table 3.4.1.

Table 3.4.1 Summary Statistics from Individual Study

Study (Author, Year)	Summary Statistics			
	Delta	S.E.	*cn*	p-value
Bickel et al., (2008)	0.76	1.141	14	0.0135
Wenecke et al., (2011)	0.59	1.543	9	0.1422
Beyzadeoglu et al., (2012)	0.45	2.910	7	0.3494
Erquicia et al., (2013)	0.56	1.765	8	0.2008

And table 3.4.2 shows the results after we apply different methods using *cn* from the average of all four studies. Comparing the p-value, we can see that they are similar from all three formulas. And they are all a little bit higher than that from normal distribution.

Table 3.4.2 Testing Results from Different Approximation Methods

	Different Methods			
	DF2	DF3	DF4	Normal
df	29	32	31	NA
p-value	0.1027	0.1025	0.1026	0.1010

3.5 Discussion

In this chapter, we discuss the behavior of the summary test statistic of meta-analysis when using inverse-variance method. We show that if overall MSE is used in variance estimation, then the final test statistic would follow an exact t-distribution. However, if the variance is estimated within each study, the true underlying distribution of the weighted average would become very complicated. And the simulation results shows that it is not so close to standard normal distribution especially when sample sizes are small as well as the number of studies. We propose an approach using t-distribution to better approximate the true distribution with an empirical formula to calculate the degree of freedom. And it comes down to three similar formulas. All of them should lead to very close results. We suggest using the formula for DF2, which is a little bit simpler that the other two. The method works well with small sample sizes. However, since our formula comes from simple case scenario where each group in each study has the same

number of observations, more simulation may be needed to come up a formula for more complicated situations.

Chapter 4

Statistical Properties of the Design for Simultaneous Global Drug Development Program

4.1 Introduction

4.1.1 Background on Global Clinical Trials

Ethnicity has always been believed to be a factor that has potential impact on the treatment effect. And it is becoming more and more important as medical research and drug developments all go globally, because the proportions of different ethnic groups in the local population change dramatically from region to region. As for area like cardiovascular disease research we are focusing here, researchers have already started to put extra efforts on the potential effect brought by ethnic factor, since plenty of studies already showed that ethnicity has effect on different CVDs directly or associating with those well known risk factors [61-64]. One simple way to deal with that is to control the proportion of different races in the sample. For example, as in the Study CHART [68], one of their sites in San Diego, CA, was specifically designed to enroll more subjects from Hispanic or Latino origin than other site to represent the underlying population. And as for other studies, they may set an additional site in other country to study the effect on different ethnic group, such as the Syst-China trial which can be considered as the extension of the SHEP trail in China [69], which we studied in chapter 2. Another example is the study

that was carried out in a Han Chinese population living in rural areas of northern China to exam the association between renin-angiotensin-aldosterone system (RAAS) genes and salt-sensitivity of blood pressure (BP) [70].

Drug development also faces the same challenge brought by this ethnic factor. The design of clinical trials for new drug or treatment may be a well-established topic in statistics. But when it goes to global and with this ethnic factor to consider, it becomes more complicated, along with the troubles caused by the different regulations in different countries. Mainly for now, a multi-regional clinical trial (MRCT) with many participating countries or regions could be sufficient to obtain the approval of a new drug in the United States (US) and Europe Union (EU). However, when it comes to globe, especially for countries like China or Japan, additional local clinical trial (LCT) may be required to assess the potential impact of ethnic factors on treatment effect [71, 72]. The main reason that this is not a major issue in MRCT is because among those counties in MRCT, the proportions of different ethnic groups are similar. While, it is not the case for countries like China or Japan. Then it is reasonable to assume there would be potential difference on the safety and efficacy results from the MRCT and LCT. The well-known E5 guideline (1998) [73] published by the International Conference on Harmonization (ICH) provided a general principle and framework for evaluating such impact. The general idea behind that is to extrapolate the information from studies in one region to another, so that all the information can be utilized more efficiently. However, there are no commonly accepted statistical criteria on how to extrapolate or combine information from different studies, such as MRCT and LCT.

Several methods and designs based on this guideline have been proposed in literatures in recent years. For example, Shih (2001) [74] introduced the idea of a “consistency trial”. They suggested designing the new local trial, also known as bridging study, based on the consistency of the previous similar trials determined by a Bayesian prediction method. Chow et al. (2002) [75] proposed the use of a sensitivity index as a possible criterion to determine whether a bridging study is necessary, and the sample size for such study. A statistical method for the assessment of similarity of clinical results between regions was also proposed using the concept of population bioequivalence.

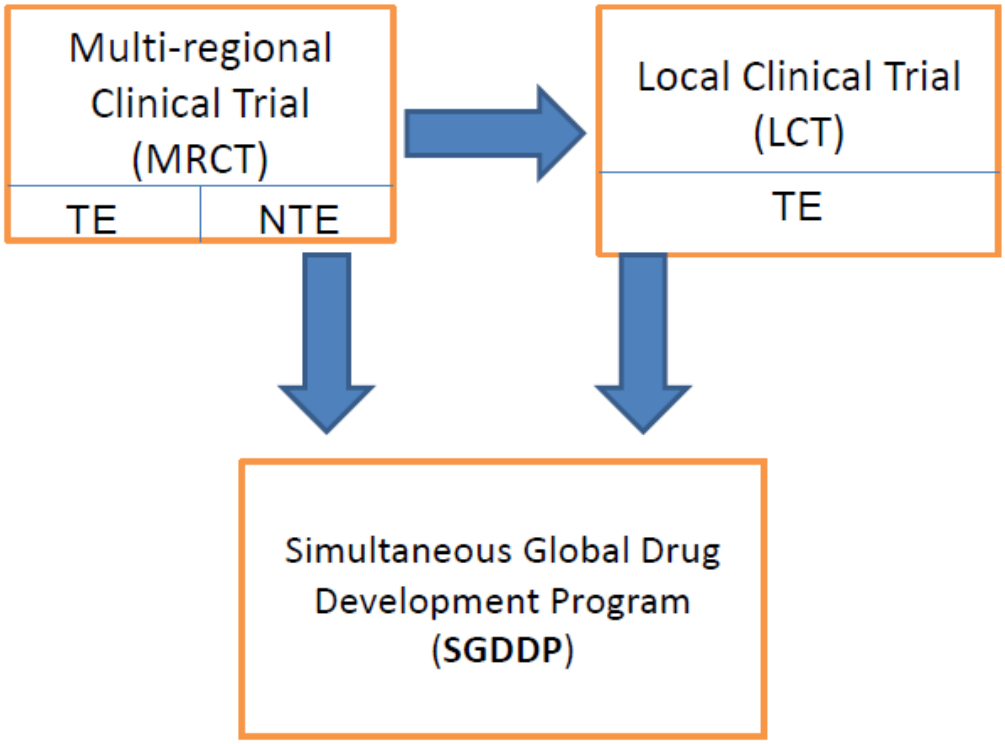
Hsiao et al. (2003, 2005) [66, 76] proposed a two-stage design and provided sample size calculation for the LCT at the second stage. Lan et al. (2005) [67] applied weighted-Z tests to combine the information from the MRCT and the LCT. The main theme of both papers was to borrow information from the MRCT to test the efficacy in the target ethnic (TE) population. However, both of them designed the LCT after the completion of the MRCT, which could lead to years of delay for local registration.

We can see that a better methodology for global clinical trial design and corresponding statistical analysis plan would really be beneficial to both the drug developer and the patients. For the developer side, a better way to use all the information from different location and ethnic group will evenly reduce the sample size needed in certain country which will lead to huge cut in the overall investment. And it will be even better if the overall time in the drug development can be shorted. While for the patients, especially those are suffering under disease such as CVD which is the number one cause of death worldwide, the early they get the new drug, the better chance they have to improve their life, not to mention that the efficacy results may be more reliable from those new designs.

4.1.2 SGDDP Design and Statistical Method

With the motivation of finding a good solution to the problem mentioned above, Huang et al. (2012) [65] proposed to design the MRCT and the LCT concurrently in a simultaneous global drug development program (SGDDP). It is outlined in Fig 4.1.1. Generally, the MRCT is a standard large phase III clinical trial with many participating regions while the LCT is a smaller trial carried out in the targeted region to collect more data in TE patients. The first major difference between this SGDDP design and the two-stage design or bridging study is that, SGDDP is designed prospectively and both the MRCT and the LCT share the same key design features. The results from the MRCT will be ana-

Figure 4.1.1 Flow Chart of SGDDP (Huang et al. 2012)



lyzed first, and if positive, used for regulatory registration in the US and the EU. The results of the SGDDP will be considered for further local regulatory registration only when the results of MRCT are at least promising (e.g., $p < 0.1$) [65]. Although still, the LCT will probably start a little bit later then MRCT, and whether LCT should be carried out until the end depending on the results from MRCT, the overall time length of the program will be significantly shortened.

While, regarding the statistical method for the efficacy evaluation, similar to traditional bridging study designs, the SGDDP is also trying to borrow efficacy information from the MRCT to test the treatment effect for TE patients, based on the fundamental assumption that patients in TE group and in NTE group share some level of biological commonality of humankind. However, instead of borrowing the MRCT as a whole, the SGDDP grouped patients from both the MRCT and the LCT by ethnicity into TE and non-TE (NTE) subgroups, and test statistics for both of them were then combined to get the weighted-Z statistic. Basically, the main difference here, in statistic point of view, is that in the bridging design, the location or the stage or the study is the main factor which we believe has a major impact on the efficacy result. While, that factor is no longer the major concern in SGDDP, because the MRCT and LCT are designed prospectively and both share the same key design features. Instead, because of the ultimate objective, or the initial motivation, is regarding the impact ethnic factor, it naturally becomes the key factor on how to combine the information.

Let Z_1 and Z_2 be the test statistics for the TE group and NTE group, respectively. Then, the weighted-Z statistic was formulated as follows:

$$Z = \sqrt{1-w}Z_1 + \sqrt{w}Z_2 \quad (4.1.1)$$

where w was a pre-specified weight, $0 \le w < 1$. This testing strategy relied on the fundamental assumption that patients from the TE and NTE groups should share biological similarity in the response to the treatment. To adjust for ethnic difference, the information from NTE group was down-weighted.

The above design has several advantages. First, it provides a possible solution to design bridging trials with statistical rigor, providing adequate statistical power with type I error rate controlled at a given level. Second, the simultaneously designed MRCT and LCT with similar design features ensured the validity of combining information together. It would also shorten the timeline for drug registration in the new region. Moreover, since the weight is pre-selected, it would avoid potential selection bias in Lan, et al. (2005).

4.1.3 Issues about the SGDDP

Despite these attractive features, issues remain in the proposed design for the SGDDP [77, 78]. One of them is that the fundamental assumption of the SGDDP design is not necessarily equivalent to identical treatment effect between the TE and NTE groups, due to the potential impact of ethnicity. This leads to complexity in hypothesis specification. Let δ_1 be the mean effect for TE patients, and then the hypothesis for the SGDDP can be formulated as follows:

$$H_0 : \delta_1 = 0 \ vs. \ H_a : \delta_1 > 0 \quad (4.1.2)$$

Since the H_0 is only for the TE group, the weighted-Z test statistic defined in (4.1.1) may not be distributed as standard normal since the distribution of Z_2 is not necessarily standard normal under (4.1.2) because there is no statement regarding the mean treatment effect for NTE patients neither in the underlying assumption nor in the null hypothesis. This may lead to biased tests.

Therefore, in this chapter, we propose a modification to (4.1.2) by translating the statement "both TE and NTE patients share some biological commonality" into a quantitative proportionality assumption under which the weighted-Z test would become unbiased. Additionally, we will derive the uniformly most powerful (UMP) test and use it to evaluate the performance of the weighted-Z test to illustrate the impact of different weights on power. Finally, we will discuss the situation when this proportionality assumption does not hold.

4.2 Modification to the Hypothesis in SGDDP Design

4.2.1 Distribution of Weighted-Z test

To closely study the distribution of this Weighted-Z test in the SGDDP, we should first fully understand the fundamental assumption which is patients from the TE and NTE groups should share biological similarity in the response to the treatment. Based on that, we can claim that the TE and the NTE population are independent, but not necessarily identically distributed in terms of treatment effect. Here, without loss of generality, let's focus on continuous endpoints and using the notation used in (4.1.1). Then the above

statements are equivalent to stating that the Z_1 and Z_2 are independent but not necessarily identically distributed due to the consideration of ethnic effect. A natural hypothesis for testing the treatment effect of the TE population is (4.1.2). And under the H_0, it's safe to claim that Z_1 would follow the standard normal distribution. However, due to the impact of ethnic factor, the mean of Z_2, E_2 is not necessarily equal to 0 under the H_0. We can only claim that Z_2 follows $N(E_2,1)$, with $E_2 = \frac{1}{2}\sqrt{N_2}\delta_2 / \sigma_2$ if the mean effect for NTE patients are distributed as normal with mean δ_2 and variance σ_2^2, where the number of patients in the NTE group is N_2 (Huang et al., 2012). As a result, the distribution of Z under H_0 would be $N(\sqrt{w}E_2,1)$, which is not $N(0,1)$ for $\delta_2 \neq 0$, leading to a biased test for the treatment effect for the TE population.

4.2.2 Proportionality Assumption

In order to make this test unbiased, the first possible solution would be adding $\delta_2 = 0$ into the null hypothesis, since this the reason causes this problem. Unfortunately, if we change H_0 like that, the alterative hypothesis would also be changed. It would become either one of the δ is not equal to 0, which is not the same as what we want to claim.

So, instead of changing the null hypothesis, the modification we are proposing is to quantify the assumption that the NTE and TE population share certain biological commonality. We assume that the treatment effects between two populations are proportional:

$$\delta_1 = \gamma * \delta_2 \tag{4.2.1}$$

where γ is an unknown positive constant. With (4.2.1), the following two null hypotheses are equivalent:

$$H_0 : \ \delta_1 = 0 \tag{4.2.2}$$

$$H_0 : \ \delta_1 = 0 \text{ and } \delta_2 = 0 \tag{4.2.3}$$

Therefore, the distribution of the weighted test statistic Z is standard normal under H_0 in (4.1.2) which makes the test unbiased.

4.2.3 Sample size evaluation

In this section, we are going to show that under the proportionality assumption, the sample size calculation in Huang et al. (2012) [65] will be depending on not only the weight but the proportionality parameter γ as well. Taking continuous endpoints as an example, let p be the proportion of the TE patients in the MRCT with an overall effect size δ_0, and let the effect size of the TE group and NTE group be δ_1 and δ_2 respectively, then

$$\delta_0 = p * \delta_1 + (1 - p) * \delta_2 = \delta_1 * (p * (\gamma - 1) + 1) / \gamma \tag{4.2.4}$$

From (4.2.4) we can see that if the values of δ_0 and p are fixed, then δ_1 and γ are one-to-one corresponding to each other. As a result, any sample size calculation based on the parameter combinations of $w \setminus \gamma$ and $w \setminus \delta_1$ are equivalent.

Now let's take a look at the method Huang et al. (2012) [65] used to calculate the minimum sample size need in LCT to make the test having a power at least equal to β. First of all, because the power of the Weighted-Z test for treatment effect in TE population can be calculated by the formula below:

$$\beta = \Phi\left\{\sqrt{1-w}E_1 + \sqrt{w}E_2 - Z_{1-\alpha/2}\right\}, \tag{4.2.5}$$

then we have

$$E_1 = (Z_\beta + Z_{1-\alpha/2} - \sqrt{w}E_2)/\sqrt{1-w}. \tag{4.2.6}$$

Here, E_1 and E_2 are the corresponding means of Z_1 and Z_2. Z_β and $Z_{1-\alpha/2}$ are the $\beta\times100$ and $(1-\alpha/2)\times100$ percentiles of the standard normal distribution. And because we have

$$E_2 = \frac{1}{2}\sqrt{N_2}\delta_2/\sigma_2, \tag{4.2.7}$$

E_1 can be calculated by (4.2.6) if we know the values of N_2, δ_2 and σ_2. While, since we also have

$$E_1 = \frac{1}{2}\sqrt{n_{1p}}\delta_1/\sigma_1, \tag{4.2.8}$$

where, n_{1p} is the total sample size for TE subgroup, and σ_1 is the standard deviation of the end point in TE population. Then n_{1p} can be easily calculated by

$$n_{1p} = 4(E_1\sigma_1/\delta_1)^2 \tag{4.2.9}$$

Note that we assume the treatment group and control group have same number of subjects in both TE and NTE subgroup. Also note that n_{1p} is the sample size of the whole

TE subgroup. So the sample size of LCT would be n_{1p} subtracted by the number of TE subjects in MRCT.

To sum up, if we have the values of δ_0, p, N_2, σ_1, σ_2, α and β pre-fixed, then for the selected $w \backslash \gamma$, we can first calculated the values of δ_1 and δ_2 by (4.2.4) and then get E_2 by (4.2.7). After that, apply (4.2.6) to get the value of E_1 and finally get the sample size n_{1p} by (4.2.9), and the sample size for LCT will follow.

Table 4.2.1 and Table 4.2.2 below show the sample sizes for LCT if we use the same parameter settings as for table 1a and 1b in Huang et al. (2012) [65]. Here, we assume that $\sigma_1 = \sigma_2 = 1$. The type I error is set to $\alpha = 0.05$ and the desired power lever it set to $\beta = 80\%$. The values of γ are selected as $\gamma = 0.55,\ 0.76,\ 1.00$ and 1.26 to match the values of $\delta_1 = 0.15,\ 0.20,\ 0.25$ and 0.30 in Huang et al. (2012). We also set the overall sample size and effect size δ_0 of MRCT to be 500 and 0.25 so that the test for MRCT alone will also have a power equal to 80%.

As shown in Table 4.2.1, borrowing information from the MRCT can significantly reduce the sample size needed in LCT while still have the same level of power. For example, if the underlying $\gamma = 0.76$, which means $\delta_1 = 0.20$ and the treatment effect for TE group is smaller than that for NTE group. Under this case scenario, if we chose the weight $w = 40\%$, the LCT in this SGDDP would require 118 patients which is much fewer than 685 if no information from NTE patients are to be used.

Table 4.2.1 Sample Size for LCT when MRCT has 20% TE Patients

	Proportional Parameter (γ)			
Weight(w)	0.55	0.76	1.00	1.26
0%	1296	685	403	249
10%	638	332	188	108
20%	449	232	127	69
30%	327	166	88	43
40%	235	118	59	26
50%	162	79	37	12
60%	101	48	20	3
70%	49	23	8	0
80%	4	3	3	0

a) The MRCT has 500 patients with 20% of them belong to the targeted ethnic subgroup; b) the endpoints are normally distributed with the overall effect size of 0.25 and $\sigma_1 = \sigma_2 = 1$; c) has different underlying values of γ

Table 4.2.2 Sample Size for LCT when MRCT has 10% TE Patients

	Proportional Parameter (γ)			
Weight(w)	0.55	0.76	1.00	1.26
0%	1346	735	453	299
10%	683	371	225	144
20%	493	266	159	100
30%	370	198	117	72
40%	277	148	85	50
50%	203	107	60	34
60%	142	74	40	20
70%	90	45	23	10
80%	44	22	10	2

a) The MRCT has 500 patients with 10% of them belong to the targeted ethnic subgroup; b) the endpoints are normally distributed with the overall effect size of 0.25 and $\sigma_1 = \sigma_2 = 1$; c) has different underlying values of γ ;

Similar results can be found in Table 4.2.2 when there are fewer (10%) TE patients are enrolled in the MRCT. And by comparing Table 4.2.1 and Table 4.2.2, we can see that under the same value of $w \backslash \gamma$, more TE patients are need in Table 4.2.2, 148 patients comparing to 118 patients mentioned in previous example to maintain the same power. This is straightforward to understood since the n_{1p} is the summation of number of TE patients both from MRCT and LCT. When the number from MRCT drops, the one from LCT has to increase.

By comparing the numbers in Table 4.2.1 and Table 4.2.2 to Table 1a and Table 1b in Huang et al. (2012) [65] cell by cell, we can conclude that they are totally identical, which verified our conclusion earlier that sample size calculation based on the parameter combinations of $w \backslash \gamma$ and $w \backslash \delta_1$ are equivalent. It also means that our modification can justify the calculation in that paper.

4.3 Power Consideration

With the proportional assumption, the Weighted-Z test defined in (4.1.1) for hypothesis (4.1.2) becomes unbiased test and the type I error rate can be well controlled. Then in the following sections, we are going to study the performance of the test in terms of power, especially with the choice of weight changing. In order to do that, we are first going to find certain standard test to be comparing with. The UMP test becomes our first candidate. We also notice that Lan et al. (2005) [67] suggested an optimum choice w such that is $E\left(Z_w\right)$ maximized. And those two tests actually turn out to be equivalents which then

are defined to be our optimum test. Their powers are then compared to the Weighted-Z test in different ways to show that although we can don't achieve the maximum power in practice, the loss of power is acceptable even without complicated procedure to choose the weight.

4.3.1 The test with optimum weight

In this section, we will first derive the UMP test. Without loss of generality, let's assume $E_i = E(Z_i)$ be the expectation of Z_i in (4.1.1) with $i = 1, 2$. Then, based on the assumption of (4.2.1), the joint distribution of Z_1 and Z_2 can be expressed as:

$$f_{E_1,E_2}(z_1, z_2) = C(E_1, E_2)\exp\left(E_1 z_1 + \frac{E_1}{\gamma}\sqrt{\frac{N_2}{n_{1p}}}z_2\right)h(z_1, z_2) \tag{4.3.1}$$

where N_2 and n_{1p} are the numbers of patients in the NTE and TE groups, respectively. $C(E_1, E_2)$ and $h(z_1, z_2)$ are two real value functions. Since γ is a positive constant as we assumed, the UMP test can be constructed as following (Casella & Berger, 2008) [79]:

$$\varphi(z_1, z_2) = \begin{cases} 1? & if\ z_1 + \sqrt{\frac{N_2}{n_{1p}}}\, z_2 \Big/ \gamma > c_0 \\ 0? & if z_1 + \sqrt{\frac{N_2}{n_{1p}}}\, z_2 \Big/ \gamma < c_0 \end{cases} \tag{4.3.2}$$

where c_0 is a constant controlled by the size of the test. It is straightforward to see that the UMP test is equivalent to a weighed Z test with

$$\frac{\sqrt{w}}{\sqrt{1-w}} = \frac{1}{\gamma}\sqrt{\frac{N_2}{n_{1p}}} \tag{4.3.3}$$

Besides of the UMP test, we also found out that for given n_{1p} and N_2, Lan et al. (2005) had suggested an optimal choice of w such that $E(Z_w)$ is maximized, which is,

$$w^* = \frac{N_2\delta_2^{\,2}}{N_2\delta_2^{\,2} + n_{1p}\delta_1^{\,2}} \tag{4.3.4}$$

Under the proportional assumption, we then have

$$w^* = \frac{N_2\delta_2^{\,2}}{N_2\delta_2^{\,2} + n_{1p}(\gamma\delta_2)^2} = \frac{N_2}{N_2 + n_{1p}\gamma^2} \tag{4.3.5}$$

It's straightforward to verify that w^* is the solution to (4.3.3). Thus the "statistically" optimal Weighted-Z test is also the UMP test, which is reasonable because with larger expected value, the probability of the test statistic being greater than c_0 is also larger, so as the power. Next we will use the UMP test as a standard to evaluate the performance of the weighted-Z test with different choices of w.

4.3.2 Comparing powers of weighted Z test to optimum test

Since we assume the true treatment effect for the TE and NTE population is proportional, we can borrow the down-weighted information from the NTE group in the Weighted-Z test. In practice, the weight of the information from the NTE group should be capped by the actual proportion of NTE patients in the SGDDP. In other words, the cap for the choices of w is as follows:

$$w_0 = \frac{N_2}{N_2 + n_{1p}} \tag{4.3.6}$$

Because in practice, we do not know the true value of γ, as well as w^*, we cannot use the information from NTE group entirely to achieve the maximum power.

Again, taking continuous endpoints as an example, the power of the weighted-Z test can be calculated with (4.2.5) with $0 \le w \le w_0$. Here, we still have $E_1 = \frac{1}{2}\sqrt{n_{1p}}\delta_1 / \sigma_1$, $E_2 = \frac{1}{2}\sqrt{N_2}\delta_2 / \sigma_2$, and $\delta_i s$, are connected by (4.2.4). Without loss of generality, we still assume $\sigma_1 = \sigma_2 = 1$. The results below show the loss of power of the Weighted-Z tests with different weights comparing to the UMP test.

Table 4.3.1 lists the powers of the Weighed-Z test as w changes from 0 to 70%, which is less than the cap $w_0 = 72.7\%$, given that $\delta_0 = 0.25$, $N_2 = 400$ and $n_{1p} = 150$. From the table, we can see that if we do not borrow any information from the NTE group ($w = 0\%$), then the power would be very low. However, with appropriate selection of the weight, the loss of power is moderate. For example, assume that $\delta_1 = 0.76\delta_2$, then for $w = 50\%$, the power of the weighted Z test would be 0.777, only 5% less than that of the UMP test. We can also find that the closer the means of both TE and NTE groups are, the less the power loss is.

Similar results can be found in table 4.3.2, where n_{1p} is increased to 200 while the other parameters stay the same. Moreover, as the n_{1p} increased, the cap of the weight w_0 decreased. And with the same choice of $w \setminus \gamma$, the test in Table 4.3.2 is always having

Table 4.3.1 Powers of the Weighted-Z Test with $n_{1p} = 150$

	Proportional Parameter (γ)			
Weight(w)	0.55	0.76	1.00	1.26
0%	0.149	0.231	0.334	0.451
10%	0.413	0.513	0.611	0.703
20%	0.536	0.621	0.701	0.772
30%	0.624	0.692	0.755	0.810
40%	0.688	0.742	0.790	0.833
50%	0.737	0.777	0.813	0.846
60%	0.774	0.802	0.828	0.851
70%	0.801	0.818	0.834	0.849
Optimum test	0.826	0.826	0.834	0.851
$w_0 = 72.7\%$	0.806	0.821	0.834	0.847

a) The MRCT has 500 patients with 20% of them belong to the targeted ethnic subgroup and LCT has additional 50 patients ; b) the endpoints are normally distributed with the overall effect size of 0.25 and $\sigma_1 = \sigma_2 = 1$; c) has different underlying values of γ ;

Table 4.3.2 Powers of the Weighted-Z Test with $n_{1p} = 200$

	Proportional Parameter (γ)			
Weight(w)	0.55	0.76	1.00	1.26
0%	0.184	0.293	0.424	0.564
10%	0.466	0.584	0.694	0.789
20%	0.587	0.684	0.770	0.841
30%	0.668	0.746	0.813	0.868
40%	0.726	0.787	0.839	0.882
50%	0.769	0.815	0.855	0.889
60%	0.800	0.833	0.863	0.889
Optimum test	0.838	0.847	0.865	0.890
$w_0 = 66.7\%$	0.815	0.841	0.865	0.886

a) The MRCT has 500 patients with 20% of them belong to the targeted ethnic subgroup and LCT has additional 100 patients ; b) the endpoints are normally distributed with the overall effect size of 0.25 and $\sigma_1 = \sigma_2 = 1$; c) has different underlying values of γ ;

larger power than that in Table 4.3.1. This is also well expected since the sample size of TE patients is larger and that of NTE patients remains the same. And with same value of treatment effect for both subgroups, of course larger sample size brings higher power of the test.

Table 4.3.3 and Table 4.3.4 evaluate the performance of the Weighted-Z test in other perspectives, In Table 4.3.3, we compare the power of weighted-Z test with $w = 45\%$ to the UMP test as n_{1p} increasing from 100 to 400. Here Wz represents for Weighted-Z test and Opt for optimum test. The trend shown in the table is that as the n_{1p} size increases, the loss of power becomes smaller. This is because $w = 45\%$ is becoming closer to the optimal weight.

Table 4.3.3 Powers of the Weighted-Z Test with $w = 45\%$ and Different n_{1p}

n_{1p}	Proportional Parameter (γ)							
	0.55		0.76		1.00		1.26	
	Wz	Opt	Wz	Opt	Wz	Opt	Wz	Opt
100	0.670	0.813	0.706	0.802	0.740	0.798	0.772	0.802
150	0.714	0.826	0.761	0.826	0.803	0.834	0.840	0.851
200	0.749	0.838	0.802	0.847	0.848	0.865	0.886	0.890
250	0.778	0.850	0.835	0.865	0.882	0.890	0.918	0.919
300	0.802	0.860	0.861	0.882	0.907	0.911	0.941	0.941
350	0.823	0.870	0.883	0.897	0.927	0.928	0.957	0.957
400	0.841	0.880	0.900	0.910	0.942	0.942	0.968	0.969

a) The MRCT has 500 patients with 20% of them belong to the targeted ethnic subgroup and LCT has additional $(n_{1p} - 100)$ patients ; b) the endpoints are normally distributed with the overall effect size of 0.25 and $\sigma_1 = \sigma_2 = 1$; c) has different number of patients belong to the targeted ethnic subgroup and different underlying values of γ ;

For a pre-specified $w \backslash \gamma$, we already calculated the additional number of TE patients needed to make the corresponding Weighted-Z test having 80% power (Table 4.2.1). In Table 4.3.4, we based on this sample size and calculate the power of the corresponding UMP test to show the loss of power. For example, for the MRCT, assume that there are 500 patients, 100 of which (20%) are from TE population. And if we assume the overall effect size $\delta_0 = 0.25$, and the underlying $\gamma = 0.76$, then if we pre-specify that $w = 40\%$, we need to enroll at least 118 TE patients in the LCT to obtain $\beta = 80\%$ for the SGDDP. While, the power of the corresponding UMP test under this sample size setting, is 0.853, which is just slightly higher than 0.8. In a word, this table shows that with reasonable choice of weight, the loss of power for the SGDDP design is generally small.

Table 4.3.4 Powers of the UMP tests with Minimum n_{1p}

	Proportional Parameter (γ)			
Weight(w)	0.55	0.76	1.00	1.26
0%	0.975	0.97	0.964	0.957
10%	0.928	0.917	0.906	0.895
20%	0.904	0.891	0.879	0.866
30%	0.884	0.871	0.858	0.845
40%	0.867	0.853	0.840	0.828
50%	0.852	0.838	0.825	0.815
60%	0.838	0.824	0.813	0.805
70%	0.826	0.813	0.804	0.800

a) MRCT has 500 patients and 20% of which belong to the targeted ethnic subgroup; b) use a normally distributed endpoint with the overall effect size for MRCT equal to 0.25 and $\sigma_1 = \sigma_2 = 1$; c) has different underlying values of γ; d) numbers of patients in the LCT equal to numbers from corresponding cells in table 3.2.1.

Over all, these four tables show that, if the weight is pre-specified regardless the true value of γ, then there will be some loss of power comparing to the theoretically optimum test. This is unavoidable unless we know the true γ. However, if we pick an appropriate w, the loss of power is generally small.

4.4 Discussion and Final Remarks

The SGDDP design provides a possible solution on the assessment of the impact of potential ethnic factors. The program consists of two phases, a MRCT phase and a LCT phase and the results of the SGDDP can provide adequate evidence for local regulatory registration of a new treatment. The reason this design is valid is that we assume both the TE and NTE population share some level of biological commonality, which can be translated quantitatively that the treatment effect between the two populations are proportional. With this proportional assumption, we were able to make sure that the Weighted-Z test was unbiased. We also validated the sample size calculations in Huang et al. (2012) under this additional assumption. Besides, we showed that with a proper choice of weight, the loss of power of the weighted-Z test would be acceptable.

However, if the proportional assumption does not hold, then the problem can be more complicated and more investigation is needed. There may be two extreme scenarios. The first one is that the new treatment is effective in the NTE group but not in TE group, i.e., $\delta_1 = 0$ but $\delta_2 \neq 0$. In this case, the weighted test for the treatment effect for the TE population would be biased as discussed earlier. The distribution of the weighted-Z sta-

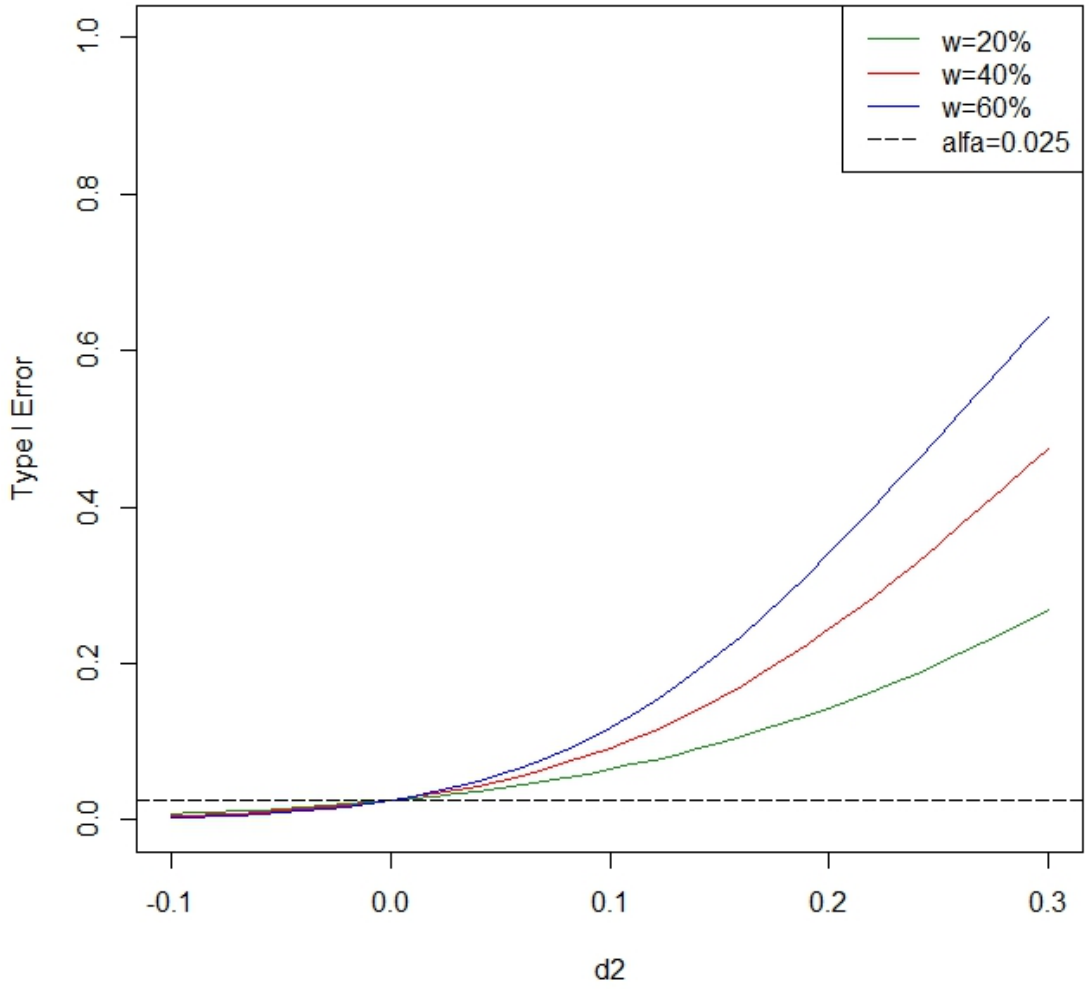
tistic will depend on the true value of δ_2. And if we treat it as standard normal regardless of δ_2, then the test is very like to be biased. From figure 4.4.1, we can see that when the value of δ_2 is large, the type I error can be unacceptable. For example, if $\delta_2 = 0.2$, the type one error would be close to 0.4.

The other extreme scenario is that the new treatment is effective in the TE population but not in the NTE population, i.e., $\delta_1 \neq 0$ but $\delta_2 = 0$. In this case, the statistical test in MRCT phase could fail since the majority of the patients enrolled in the MRCT are NTE patients. Unless the treatment effect for the TE population is extremely large, the power of the weighted-Z test would be extremely low. From figure 4.4.2 we can see that with $w = 40\%$, even $\delta_1 = 0.25$, the power is only around 0.2.

Although, in application, we intend to believe that the treatment effects should be at least similar among patients from different ethnic groups, these two scenarios discussed above should still be kept in mind. One possible solution is to apply a test first, testing weather the ration between δ_1 and δ_2 is positive and finite. Then, overall bias might be controlled and the power would be increased.

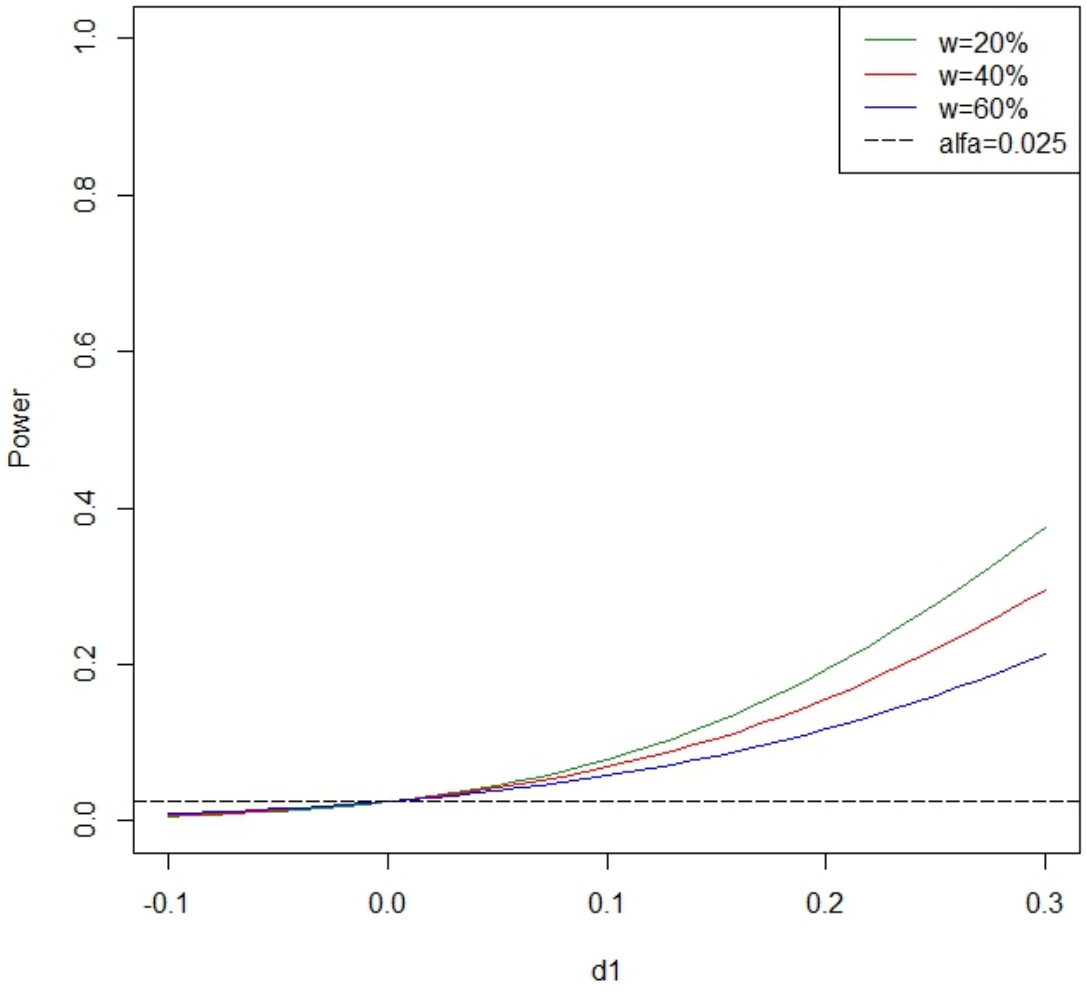
Another important aspect of this method should be noticed is that, to avoid selection bias, the weight w in the final Weighted-Z test should be pre-specified in the design stage. Since we don't know the true value of γ, how to select a w should be a further topic to work on. And estimating γ first by Bayesian Methods based on related studies might be helpful.

Figure 4.4.1 Type I Error of the SGDDP when the Proportional Assumption doesn't Hold



a) MRCT has 500 patients and 20% of which belong to the targeted ethnic subgroup; b) use a normally distributed endpoint with the overall effect size for MRCT equal to 0.25 and $\sigma_1 = \sigma_2 = 1$; c) $\delta_1 = 0$; d) numbers of TE patients in the LCT equal to 50.

Figure 4.4.2 Power of the SGDDP when the Proportional Assumption doesn't Hold



a) MRCT has 500 patients and 20% of which belong to the targeted ethnic subgroup; b) use a normally distributed endpoint with the overall effect size for MRCT equal to 0.25 and $\sigma_1 = \sigma_2 = 1$; c) $\delta_2 = 0$; d) numbers of TE patients in the LCT equal to 50.

References

[1] Global status report on noncommunicable disaeses 2010. Geneva, World Health Organization, 2011

[2] Anthea, Maton, et al. "Human biology and health." *Englewood: Prentice Hall* (1993).

[3] Fuster, Valentin, and Bridget B. Kelly, eds. "Promoting cardiovascular health in the developing world: a critical challenge to achieve global health." (2010).

[4] Mathers, Colin D., and Dejan Loncar. "Projections of global mortality and burden of disease from 2002 to 2030." *PLoS medicine* 3.11 (2006): e442.

[5] Altman, Duglas G. "Practical statistics for medical research." (1991).

[6] Rao, Calyampudi Radhakrishna, J. Philip Miller, and Dabeeru C. Rao. "Handbook of statistics: epidemiology and medical statistics." (2007).

[7] Bailar, John C., and David C. Hoaglin, eds. "Medical uses of statistics." (2012).

[8] Probstfield, J. L. "Prevention of stroke by antihypertensive drug-treatment in older persons with isolated systolic hypertension-final results of the Systolic Hypertension in the Elderly Program (SHEP)." *JAMA-JOURNAL OF THE AMERICAN MEDICAL ASSOCIATION* 265.24 (1991): 3255-3264.

[9] Kostis, John B., et al. "Association between chlorthalidone treatment of systolic hypertension and long-term survival." *JAMA: The Journal of the American Medical Association* 306.23 (2011): 2588-2593.

[10] Multiple Risk Factor Intervention Trial Research Group. Mortality after 10 1/2 years for hypertensive participants in the Multiple Risk Factor Intervention Trial. Circulation. 1990;82(5):1616-1628.

[11] Kostis, William J., et al. "Persistence of Mortality Reduction After the End of Randomized Therapy in Clinical Trials of Blood Pressure–Lowering Medications." *Hypertension* 56.6 (2010): 1060-1068.

[12] Chalmers, John, and Mark E. Cooper. "UKPDS and the legacy effect." *The New England journal of medicine* 359.15 (2008): 1618.

[13] Kostis, William J., et al. "Continuation of mortality reduction after the end of randomized therapy in clinical trials of lipid-lowering therapy." *Journal of clinical lipidology* 5.2 (2011): 97-104.

[14] Irwin, J. O. "The standard error of an estimate of expectation of life, with special reference to expectation of tumourless life in experiments with mice." *J Hyg* 47.2 (1949): 188-9.

[15] Zucker, David M. "Restricted mean life with covariates: modification and extension of a useful survival analysis method." *Journal of the American Statistical Association* 93.442 (1998): 702-709.

[16] Royston, Patrick, and Mahesh KB 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* 30.19 (2011): 2409-2421.

[17] Ernst, Edzard, and Karl Ludwig Resch. "Fibrinogen as a cardiovascular risk factor: a meta-analysis and review of the literature." *Annals of Internal Medicine* 118.12 (1993): 956-963.

[18] Wald, David S., Malcolm Law, and Joan K. Morris. "Homocysteine and cardiovascular disease: evidence on causality from a meta-analysis." *Bmj* 325.7374 (2002): 1202.

[19] Bickel, Brent A., et al. "Preoperative magnetic resonance imaging cross-sectional area for the measurement of hamstring autograft diameter for reconstruction of the adolescent anterior cruciate ligament." *Arthroscopy: The Journal of Arthroscopic & Related Surgery* 24.12 (2008): 1336-1341.

[20] Wernecke, Gregory, et al. "Using magnetic resonance imaging to predict adequate graft diameters for autologous hamstring double-bundle anterior cruciate ligament reconstruction." *Arthroscopy: The Journal of Arthroscopic & Related Surgery* 27.8 (2011): 1055-1059.

[21] Beyzadeoglu, Tahsin, et al. "Prediction of semitendinosus and gracilis autograft sizes for ACL reconstruction." *Knee Surgery, Sports Traumatology, Arthroscopy* 20.7 (2012): 1293-1297.

[22] Erquicia, Juan Ignacio, et al. "How to Improve the Prediction of Quadrupled Semitendinosus and Gracilis Autograft Sizes With Magnetic Resonance Imaging and Ultrasonography." *The American journal of sports medicine* (2013).

[23] Wright, Janice C., and Milton C. Weinstein. "Gains in life expectancy from medical interventions—standardizing data on outcomes." *New England Journal of Medicine* 339.6 (1998): 380-386.

[24] Naimark, David, Gary Naglie, and Allan S. Detsky. "The meaning of life expectancy." *Journal of general internal medicine* 9.12 (1994): 702-707.

[25] Meier, Paul, et al. "The price of Kaplan–Meier." Journal of the American Statistical Association 99.467 (2004).

[26] Nieto, F. Javier, and Josef Coresh. "Adjusting survival curves for confounders: a review and a new method." *American journal of epidemiology* 143.10 (1996): 1059-1068.

[27] Lee, James, et al. "Covariance adjustment of survival curves based on Cox′ s proportional hazards regression model." *Computer applications in the biosciences: CABIOS* 8.1 (1992): 23-27.

[28] Chang, I., Rebecca Gelman, and Marcello Pagano. "Corrected group prognostic curves and summary statistics." *Journal of chronic diseases* 35.8 (1982): 669-674.

[29] Nardi, Alessandra, and Michael Schemper. "Comparing Cox and parametric models in clinical studies." *Statistics in medicine* 22.23 (2003): 3597-3610.

[30] Gail, M. H., and D. P. Byar. "Variance calculations for direct adjusted survival curves, with applications to testing for no treatment effect." *Biometrical journal* 28.5 (1986): 587-599.

[31] Efron, Bradley. "Censored data and the bootstrap." *Journal of the American Statistical Association* 76.374 (1981): 312-319.

[32] Kaplan, Edward L., and Paul Meier. "Nonparametric estimation from incomplete observations." *Journal of the American statistical association* 53.282 (1958): 457-481.

[33] Collett, David. *Modelling Survical Data in Medical Research*. Vol. 57. CRC press, 2003.

[34] Cox, David R. "Regression models and life-tables." *Journal of the Royal Statistical Society. Series B (Methodological)* (1972): 187-220.

[35] Moeschberger, Melvin L., and John P. Klein. *Survival analysis: Techniques for censored and truncated data*. Springer, 2003.

[36] Therneau, Terry M. *Modeling survival data: extending the Cox model*. Springer, 2000.

[37] Miller Jr, Rupert G. *Survival analysis*. Vol. 66. John Wiley & Sons, 2011.

[38] Barker, Chris. "The mean, median, and confidence intervals of the Kaplan-Meier survival estimate—Computations and applications." The American Statistician 63.1 (2009).

[39] Andersen, Per Kragh, ed. *Statistical models based on counting processes*. Springer, 1993.

[41] Gill, Richard. "Large sample behaviour of the product-limit estimator on the whole line." The Annals of Statistics (1983): 49-58.

[42] Breslow, Norman, and John Crowley. "A large sample study of the life table and product limit estimates under random censorship." The Annals of Statistics 2.3 (1974): 437-453.

[43] Utzet, Frederic, and Álex Sánchez. "Some applications of the bootstrap to survival analysis." *Anuario de psicología* 55 (1992): 155-167.

[44] Davison, Anthony Christopher. *Bootstrap methods and their application*. Vol. 1. Cambridge university press, 1997.

[45] Efron, Bradley, and Robert Tibshirani. "Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy." Statistical science (1986): 54-75.

[46] Freitag, Gudrun, and Axel Munk. Consistency of bootstrap procedures for the nonparametric assessment of non inferiority with random censorship. Technical Report, Georg-Augaust-Universität Göttingen, Germany, 2005.

[47] Cheng, Guang, and Jianhua Z. Huang. "Bootstrap consistency for general semiparametric M-estimation." The Annals of Statistics 38.5 (2010): 2884-2915.

[48] Horváth, Lajos, and Brian S. Yandell. "Convergence rates for the bootstrapped product-limit process." The Annals of Statistics (1987): 1155-1173.

[49] Wang, Jia-Gang. "A note on the uniform consistency of the Kaplan-Meier estimator." *The Annals of Statistics* 15.3 (1987): 1313-1316.

[50] Lo, Shaw-Hwa, and Kesar Singh. "The product-limit estimator and the bootstrap: some asymptotic representations." *Probability Theory and Related Fields* 71.3 (1986): 455-465.

[51] Akritas, Michael G. "Bootstrapping the Kaplan—Meier Estimator." *Journal of the American Statistical Association* 81.396 (1986): 1032-1038.

[52] Karrison, Theodore. "Restricted mean life with adjustment for covariates." Journal of the American Statistical Association 82.400 (1987): 1169-1176.

[53] Andersen, Per Kragh, Mette Gerster Hansen, and John P. Klein. "Regression analysis of restricted mean survival time based on pseudo-observations." Lifetime Data Analysis 10.4 (2004): 335-350.

[54] Nielsen, Bent. "Expected survival in the Cox model." *Scandinavian journal of statistics* 24.2 (1997): 275-287.

[55] Grouven, U., et al. "Application of adjusted survival curves to renal transplant data." *Methods of information in medicine* 31.3 (1992): 210.

[56] Martin, Michael A. "Bootstrap hypothesis testing for some common statistical problems: A critical evaluation of size and power properties." Computational Statistics & Data Analysis 51.12 (2007): 6321-6342.

[57] Efron, Bradley, and Robert Tibshirani. *An introduction to the bootstrap*. Vol. 57. CRC press, 1993.

[58] Deeks, Jonathan J., and Julian PT Higgins. "Statistical algorithms in Review Manager 5." *2010-03-02].http://www, cochrane. org/sites/default/files/uploads/Statistical_*

Methods_in_RevManS. pdf (2010).

[59] Wolf, Fredric M. *Meta-analysis: Quantitative methods for research synthesis*. Vol. 59. Sage, 1986.

[60] Hedges, Larry V., et al. "Statistical methods for meta-analysis." (1985): 350-351.

[61]Kurian, Anita K., and Kathryn M. Cardarelli. "Racial and ethnic differences in cardiovascular disease risk factors: a systematic review." *Ethnicity and Disease* 17.1 (2007): 143.

[62]Sacco, Ralph L., et al. "Race-ethnic disparities in the impact of stroke risk factors the Northern Manhattan stroke study." *Stroke* 32.8 (2001): 1725-1731.

[63]Freedman, B. I., et al. "The impact of ethnicity and sex on subclinical cardiovascular disease: the Diabetes Heart Study." *Diabetologia* 48.12 (2005): 2511-2518.

[64] Burke, Gregory L., et al. "The impact of obesity on cardiovascular disease risk factors and subclinical vascular disease: the Multi-Ethnic Study of Atherosclerosis." *Archives of internal medicine* 168.9 (2008): 928.

[65] Huang, Qin, et al. "Design and Sample Size Considerations for Simultaneous Global Drug Development Program." *Journal of Biopharmaceutical Statistics* 22.5 (2012): 1060-1073.

[66] Hsiao, C.-F., Xu, J.-Z. and Liu, J.-P. (2005). A two –stage design for bridging studies. *Journal of Biopharmaceutical Statistics* 15, 75-83

[67] Lan, K.K.G., Soo, Y., Siu, C. and Wang, M. (2005). The use of weighted Z-tests in medical research. *Journal of Biopharmaceutical Statistics* 15: 625-639

[68] Riley, William T., et al. "Overview of the consortium of hospitals advancing research on tobacco (chart)." *Trials* 13.1 (2012): 122.

[69]Liu, Lisheng, et al. "Comparison of active treatment and placebo in older Chinese patients with isolated systolic hypertension." *Journal of hypertension* 16.12 (1998): 1823-1829.

[70] Gu, Dongfeng, et al. "Genetic variants in the renin-angiotensin-aldosterone system and salt-sensitivity of blood pressure." *Journal of hypertension* 28.6 (2010): 1210.

[71] Ministry of Health, Labor, and Welfare. (2007). *Basic Principles on Global Clinical Trials* Tokyo, Japan: MHLW.

[72] State Food and Drug Administration. (2007). *Drug Registration Regulation* Beijing, China: SFDA.

[73] ICH International Conference on Harmonization Tripartite Guidance E5. (1998) Ethnic factor in the acceptability of foreign data. *The US Federal Register*, 83: 31790–31796.

[74] Shih, Weichung Joe. "Clinical trials for drug registrations in Asian-Pacific countries: Proposal for a new paradigm from a statistical perspective." *Controlled Clinical Trials* 22.4 (2001): 357-366.

[75] Chow, Shein-Chung, Jun Shao, and Oliver Yoa-Pu Hu. "Assessing sensitivity and similarity in bridging studies." *Journal of Biopharmaceutical Statistics* 12.3 (2002): 385-400.

[76] Hsiao, Chin-Fu, Jia-Zhen Xu, and Jen-pei Liu. "A group sequential approach to evaluation of bridging studies." *Journal of biopharmaceutical statistics* 13.4 (2003): 793-801.

[77] Tsong, Yi. "Statistical Considerations on Design and Analysis of Bridging and Multiregional Clinical Trials." *Journal of Biopharmaceutical Statistics* 22.5 (2012): 1078-1080.

[78] Li, Ning, and William Wang. "Practical and Statistical Considerations on Simultaneous Global Drug Development." *Journal of Biopharmaceutical Statistics* 22.5 (2012): 1074-1077.

[79] Casella, G., Berger, R.L. (2008). *Statistical Inference.* Cengage Learning.

[80] FDA, US, and C. D. E. R. January. "Guidance for industry: statistical approaches to establishing bioequivalence." *Center for Drug Evaluation and Research, Food and Drug Administration, Rockville, Maryland* (2001).