STATISTICAL METHODS TO STUDY HETEROGENEITY OF TREATMENT EFFECTS

Lin H. Taft

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the Department of Biostatistics,
Indiana University

August 2016

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Changyu Shen, PhD, Chair

Xiaochun Li, PhD

Doctoral Committee

Peng-Sheng Chen, MD

September 25, 2015

Jennifer Wessel, PhD

DEDICATION

To My Family

ACKNOWLEDGEMENTS

I would like to express my sincere appreciation to my advisor and mentor, Dr. Changyu Shen, for his guidance and advice throughout my Ph.D. program. I thank Dr. Xiaochun Li, Dr. Peng-Sheng Chen, and Dr. Jennifer Wessel for serving in my research committee. I also thank Boston Scientific CRM for permission to use their data.

I am very grateful to my parents for their love, support and inspiration. They taught me to always be positive and optimistic.

I thank the entire Department of Biostatistics for creating a welcome, collaborative, but challenging atmosphere, which I truly enjoyed for the past five years. My gratitude goes out to all my friends, for their support, encouragement, and all the laughter they brought me especially during the occasional hardship throughout the Ph.D. program.

Last but not least, I would like to give special thanks to my wonderful husband, who has been giving me incessant help and support during my Ph.D. journey.

Lin H. Taft

STATISTICAL METHODS TO STUDY HETEROGENEITY OF TREATMENT EFFECTS

Randomized studies are designed to estimate the average treatment effect (ATE) of an intervention. Individuals may derive quantitatively, or even qualitatively, different effects from the ATE, which is called the heterogeneity of treatment effect. It is important to detect the existence of heterogeneity in the treatment responses, and identify the different sub-populations. Two corresponding statistical methods will be discussed in this talk: a hypothesis testing procedure and a mixture-model based approach. The hypothesis testing procedure was constructed to test for the existence of a treatment effect in sub-populations. The test is nonparametric, and can be applied to all types of outcome measures. A key innovation of this test is to build stochastic search into the test statistic to detect signals that may not be linearly related to the multiple covariates. Simulations were performed to compare the proposed test with existing methods. Power calculation strategy was also developed for the proposed test at the design stage. The mixture-model based approach was developed to identify and study the sub-populations with different treatment effects from an intervention. A latent binary variable was used to indicate whether or not a subject was in a sub-population with average treatment benefit. The mixture-model combines a logistic formulation of the latent variable with proportional hazards models. The parameters in the mixture-model were estimated by the EM algorithm. The properties of the estimators were then studied by the simulations. Finally, all above methods were applied to a real randomized study in a low ejection fraction

population that compared the Implantable Cardioverter Defibrillator (ICD) with conventional medical therapy in reducing total mortality.

Changyu Shen, Ph.D., Chair

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1. INTRODUCTION

In clinical trials, it is common practice to report the average treatment effect (ATE) for the whole trial population. However, ATE may fail when different subgroups of patients have different treatment effects. It is widely accepted that individuals may derive quantitatively, or even qualitatively, different effects from the ATE. Detecting the heterogeneity of the treatment effects (HTE) is of increasing interests and crucial in evaluating and selecting treatments for individuals. For example, if a treatment is expensive, and may have adverse effects for some, we should obviously apply it only when it will improve the health outcome of interest. Our goal is to determine which patients will benefit from treatment and which not, so as to not do more harm than good.

The first step toward the goal is to develop a test, which can detect if there are beneficial effects and harmful effects from the treatment compared to the control. In addition, the test may help with power calculations and further study design. As a second step, we want to develop models to identify who benefits from the treatment and who does not, and the estimated average treatment effect in different sub-populations.

1.1 A Non-parametric Statistical Test of Null Treatment Effect in Sub-Populations

It is common that a medical intervention has a treatment benefit only for some patients in the intended patient population, whereas the rest do not derive a benefit and some are even harmed by the intervention.

Intuitively, to test the interaction effects between the treatment and subgroup, one can either partition patients into subgroups, to test the treatment effect in each subgroup, or add treatment-covariate interaction terms in the multiple regression. However, there are a lot of problems with the simple subgroups analysis methods. Not only does multiple comparison inflate the type I error, and creates selection bias, but also each subgroup may have a limited sample size making the test unreliable.

A lot of researchers have been working on improving the tests. There are generally two types of tests, one is to pre-specify a fixed number of subgroups before the statistical analysis, and another one is a post hoc subgroup search. Most publications in these two types heavily rely on model assumptions. However, it is hard to choose the right form of the interaction terms of the correct covariates, which substantially affects the study results. In this dissertation, we propose a non-parametric test with built-in stochastic search to avoid model assumptions and apply the test on a dataset with time-to-event outcome.

1.2 Power Calculation for Study Design

As a post hoc test, the main purpose of the test is to help guide study design in the future. With some knowledge on the HTE, we can design a study to achieve the desired power to detect HTE. In this dissertation, we try to give recommendations on the sample sizes in study designs under the assumption that the treatment effects are normally distributed. We also use some computational methods to help deal with some very complex formulations.

1.3 Logistic-Cox Mixture Model

Mixture models have been applied to different areas in survival analysis, such as competing risks, and cure rate analysis.

In this dissertation, we develop the mixture model to model the heterogeneity of the treatment effect with the assumption that different groups have different average treatment effects. We focus on having only two sub-populations in this dissertation; one group of patients has beneficial effect from the treatment, whereas the other group has no effect or a harmful effect. We use constrains to separate the treatment effects into different ranges to create two groups, and calculate the posterior probability of each patient in each group to use as a weight to model patients in different groups with different survival models. The mixture model can be simply expanded to accommodate more sub-populations with different treatment effects, although more complex models may require more patient subjects to estimate. In addition, the constraints on different groups' treatment effects can be modified depending on different characteristics for different disease and different clinical trials.

My dissertation contains three related topics on the heterogeneity of treatment effects. In chapter 2, we introduce a statistical procedure to test the null treatment effect in sub-populations, and to determine the existence of sub-populations that have treatment effects in possibly different directions. In chapter 3, Study design and power calculations for the detection of HTE using the procedure described in chapter 2 are discussed. In chapter 4, a mixture model is developed to identify sub-populations with treatment benefit. Conclusions and discussions are summarized in chapter 5.

CHAPTER 2. A NON-PARAMETRIC STATISTICAL TEST OF NULL TREATMENT EFFECT IN SUB-POPULATIONS

2.1 Background

Randomized clinical trials (RCTs) are designed to estimate the average treatment effect (ATE) in a well-defined population. However, ATE may not reflect the impact of the treatment on every subject in the trials. It is well recognized that a medical intervention may have treatment benefit for some patients but not for others[1, 2]. For example, genetic variation can lead to different drug responses even in a relatively homogeneous population meeting the entry criteria of an RCT. In the extreme case, a medical intervention can even worsen the intended efficacy endpoints for some patients. In addition, medical interventions are applied to more heterogeneous populations in the real world, where the heterogeneity in treatment effect is likely enhanced. In principle, there are three types of patients, those who benefit from, are not affected by, and are harmed by the intervention. We will call them the "beneficial group", "neutral group" and "harmed group". The ATE is a net effect of the treatment effect in the three groups. In some clinical trials, treatment effect can be 'a mixture of substantial benefits for some, little benefit for many, and harm for a few'[3]. In this case, ATE is the net result of the competition of the three groups and the inferred ATE can be tremendously misleading for some patients.

Heterogeneity in treatment effect (HTE) is usually detected through the test of the interaction between treatment arm indicatory and a covariate whose value defines sub-populations. There are two types of interactions first introduced by Peto [4]: the *qualitative interaction* (*QLI*) and the *quantitative interaction* (*QNI*). Peto described QLI

as when the true treatment effects vary in direction among sub-populations, and he used QNI when variation is only in magnitude, but not in direction. Obviously, the existence of QLI is critical and the most detrimental, where a patient can be given a treatment considered to be effective based on ATE that will actually worsen the outcome.

The caveats of existing approaches subgroup analysis have been discussed in multiple publications[5-7]. Specifically for the detection of HTE, conventional statistical tests of interaction terms usually rely on correct specification of a parametric or semi-parametric model. This is a major limitation as the test is not very meaningful if the model is wrong.

Gail and Simon (G&S) [8] developed a likelihood ratio test to detect QLI in the setting of *I* fixed sub-populations. They defined two test statistics that summarize positive and negative standardized treatment differences over subgroups. The null hypothesis of consistent direction in treatment effect across the sub-populations is rejected if both statistics exceed critical values. Later Piantadosi and Gail (P&G)[9] proposed a standardized range test where the maximum and the minimum of the standardized treatment difference of each subgroup are test statistics. Li and Chan [10] proposed an extension to the range test to utilize all the observations rather than only use the max and min, to reach better power. Recently, Bayman et al. [11] proposed a method using Bayes factor to test for QLI when multiple subgroups were determined only by one variable.

All above methods are designed for pre-specified sub-populations. Alternative approaches have been developed to search for sub-populations (e.g. identify QLI with variable selection). Bonetti and Gelber[12] discussed the subpopulation treatment effect pattern plots (STEPP) approach. In this approach, they defined overlapping

subpopulations that contain patients with increasingly larger (or smaller) value of a specific covariate, to explore the interaction between this covariate and the treatment effect. Chen et al. [13] proposed a Bayesian approach to search for qualitative interactions in a multiple regression setting with adaptive decision rules. Tian et al. [14]developed a simple modified covariate method to estimate the covariates and treatment interaction without the need of main effect. Several tree-based methods were proposed to avoid the problem of incorrect model assumption, such as Virtual Twins [15] and Qualitative Interaction Trees (QUINT)[16].

Motivated by the methods of G&S and P&G, in this chapter, we propose a non-parametric test to test the sharp null hypothesis that the treatment has null effect on every sub-group in the setting of a large number of sub-groups defined by a set of covariates. We focus on RCTs with discrete covariates in this chapter, though the method can be directly extended to continuous covariates and observational studies (see Discussions). Our strategy for the construction of the test strikes a balance between sufficient sample size for informative accuracy and adequate account of subject characteristics for the detection of HTE. Specifically, the subgroups, defined later in Method section as *cells*, induce a large number of overlapping sub-populations by various combinations. Our strategy is to sample from the pool of sub-populations and then to apply G&S and P&G type approaches. A key innovation of our method is that stochastic search is built into the test statistic to detect signals that may not be detectable through parametric and semi-parametric modeling.

In section 2, we introduce our method and how to construct the test statistics. The test is then applied on a real data example in section 3, followed by two simulations in section 4. We concluded the chapter with a discussion in section 5.

2.2 Method

2.2.1 Definitions and Hypothesis

We focus on randomized clinical trials with one intervention arm and one control arm. Let $x_1, x_2, \dots, x_H$ denote the discrete baseline covariates. If each variable x_h has c_h levels, then the covariate space can be divided in to $\prod_{h=1}^{H} c_h$ unique *cells* that cannot be further divided. In other words, a cell is the smallest sub-population at which treatment effect can be evaluated non-parametrically. For example, if there are three binary baseline variables, then there are eight different cells. In real datasets, some cells may be empty and can be removed. Let L denote the number of cells with at least one subject in each treatment arm.

Let D_l be the true treatment effect in cell $l = 1,2, \dots, L$, where positive values of D_l represent treatment benefit and negative values of D_l represent treatment harm. A two-sided null hypothesis is that there is no treatment effect in any cell, or $H_0^A: D_l = 0, l = 1,2, \dots, L$. The corresponding alternative hypothesis is $H_1^A: D_l \neq 0, \text{for some } l = 1,2, \dots, L$. For one-sided test of either benefit or harm, the null hypotheses are: $H_0^B: D_l \leq 0, l = 1,2, \dots, L$ and $H_0^C: D_l \geq 0, l = 1,2, \dots, L$ respectively. Their corresponding alternative hypotheses are: $H_1^B: D_l > 0, \text{for some } l = 1,2, \dots, L$ and $H_1^C: D_l < 0, \text{for some } l = 1,2, \dots, L$.

2.2.2 Test Statistics

In Gail and Simon[8], the estimated treatment effect $\widehat{D}_l$ was assumed to be independent and normal distributed with mean D_l, and known variance σ_l^2. The two-sided null hypothesis was stated as $H_{02}: \Delta \in 0^+ \cup 0^-$, where $\Delta^T = \{D_1, D_2, \dots, D_L\}$, $0^+ = \{\Delta: D_l \geq 0 \; all \; l\}$ and $0^- = \{\Delta: D_l \leq 0 \; all \; l\}$. Their two-sided null hypothesis is different than the one in this chapter, it is hypothesizing either all cells have benefit or no treatment effect, or all cells have harmful or no treatment effect. In other words, it is testing the existence of QLI. They provided α level critical values $C_{2\alpha}$ for the likelihood ratio test, to reject H_{02} if both $\sum_{l=1}^{L}\{(\widehat{D}_l^2/\sigma_l^2)I(\widehat{D}_l > 0)\} > C_{2\alpha}$ and $\sum_{l=1}^{L}\{(\widehat{D}_l^2/\sigma_l^2)I(\widehat{D}_l < 0)\} > C_{2\alpha}$. Here $I(.)$ is an indicator function, makes the summations over all positive $\widehat{D}_l$'s in the first expression, and over all negative $\widehat{D}_l$'s in the second expression.

The standardized range test by P&G rejects H_{02} at level α if both $max\{\widehat{D}_l/\sigma_l\} > C'_{2\alpha}$ and $min\{\widehat{D}_l/\sigma_l\} < -C'_{2\alpha}$, where $C'_{2\alpha}$ are α level critical values provided by them.

Although in principle we can apply approaches similar to G&S and P&G to test H_0^A, H_0^B or H_0^C, there are some major limitations due to the small sample sizes in each cell. First, conventional asymptotic properties do not apply to cell-specific statistic and the critical values cannot be derived from asymptotic distributions. Second, in the extreme case, some cells may only have one data point in each arm, which makes the calculation of G&S and P&G statistic highly unreliable. Third, the test statistic tends to have high variation leading to reduced power. To address these issues, we propose a non-parametric permutation test that target on stochastic sub-populations that are the union of some cells.

To create a stochastic sub-population, we select each cell with a pre-specified probability p. In the earlier example with eight cells, if p is assumed to be 0.25, then on

average, a total of two cells may be selected to form a sub-population. Choosing the right p is crucial to the proposed test. If p is too large, too many cells will be selected, then different types of subgroups (i.e. beneficial group, harmed group, etc.) may likely to be mixed together causing the dilution effect. If p is too small, then the sample size of the sub-population may not be large enough. Next, k sub-populations can be drawn independently. k is also important since enough variety of sub-populations need to be drawn to be able to detect treatment signals.

For a particular sub-population i, a test statistic Z_i can be calculated depending on the type of outcomes, which represents the magnitude and direction of the treatment effect in sub-population i. In our set-up, positive sign of Z_i corresponds to beneficial effect and the negative sign of Z_i corresponds to harmful effect. For the k sub-populations, if one of them has an extreme Z_i or the average of Z_i's with the same sign are extreme, then the null hypothesis is rejected. Thus, the statistics for beneficial effect can be defined as either

$$S^E = \max(Z_i),$$

or

$$S^A = \sum_{i=1}^{k} \frac{\max(Z_i, 0)}{k}.$$

Similarly, the harm statistics can be defined as

$$T^E = \min(Z_i) \; or \; T^A = \sum_{i=1}^{k} \frac{\min(Z_i, 0)}{k}.$$

We call tests based on S^E or T^E the one-sided extreme value test, and tests based on both of them the two-sided extreme value test. Similarly, one or two-sided average

value test can be constructed using S^A and T^A. The rationale of our test statistics is that if we explore a large number (k) of sub-populations, we will have better chance to detect the existence of treatment benefit/harm in some people.

For binary outcomes, let p be the proportion of patients with favorable outcome. A standard choice of Z_i is

$$Z_i = \frac{\hat{p}_{i1} - \hat{p}_{i0}}{\widehat{SE}_i}$$

$$\widehat{SE}_i = \left\{\hat{p}_i(1-\hat{p}_i)\left[\frac{1}{n_{i1}} + \frac{1}{n_{i0}}\right]\right\}^{0.5}$$

where p_{i1}, p_{i0} are the sample proportions, and n_{i1}, n_{i0} are the sample sizes for treatment and control groups respectively. Standard error (SE_i) can be calculated using a pooled sample proportion (p_i).

In the case of a continuous outcome, the conventional T statistic Z_i can be used:

$$Z_i = \frac{\bar{y}_{i1} - \bar{y}_{i0}}{\sqrt{\frac{s_{i1}^2}{n_{i1}} + \frac{s_{i0}^2}{n_{i0}}}},$$

where $\bar{y}_{i1}$ and $\bar{y}_{i2}$ are the sample means, s_{i1}^2 and s_{i0}^2 are the sample variances of the two groups.

For time-to-event outcome, a cox proportional hazard model can be fitted with treatment as the explanatory variable in the model. For the ith sub-population, the hazard function $h_i(t|X)$ given X at time t can be written as:

$$h_i(t|X) = h_i(t)\exp(Trt * \beta_i^{Trt})$$

where $h_i(t)$ is the baseline hazard function at time t for the ith sub-population.
In this case, Z_i is

$$Z_i = -\frac{\hat{\beta}_i^{Trt}}{\widehat{se}(\hat{\beta}_i^{Trt})}$$

2.2.3 Null distribution of $S(D)$ and $T(D)$

Permutation technique can be used for the construction of the null distribution of the test statistics. Here we are using S^E to illustrate. Under our null hypothesis, the treatment and control do not differ on the outcome (i.e. the outcome is independent of treatment assignment). When we permute the treatment assignment N times, we therefore create N datasets of the possible alternative treatment assignments we could have had, and calculate N possible S^{E*} (i.e. S_1^{E*}, S_2^{E*},…, S_N^{E*}). When N is a fairly large number, we can estimate the empirical null distribution of S^E, and calculate the α level critical values based on this distribution.

For the extreme value tests:

1. Reject H_0^A if either $S^E > C_{2\alpha}^{SE}$ or $T^E < C_{2\alpha}^{TE}$;
2. Reject H_0^B if $S^E > C_{\alpha}^{SE}$;
3. Reject H_0^C if $T^E < C_{\alpha}^{TE}$.

where α level critical values $C_{2\alpha}^{SE}$ and C_{α}^{SE} are defined as $(1-\alpha/2)$ and $(1-\alpha)$ percentile of S^{E*} respectively. Similarly, α level critical values $C_{2\alpha}^{TE}$ and C_{α}^{TE} are defined as $\alpha/2$ and α percentile of T^{E*} correspondingly. The rejection rules are followed in the same way for the average value tests.

2.3. Application to Madit II data

We apply these two tests to Multicenter Automatic Defibrillator Implantation Trial II (MADIT II)[17, 18]. MADITII ran from 1997 to 2001, in which1,232 patients with a prior myocardial infarction and a left ventricular ejection fraction of 0.3 or less were recruited. Patients were randomly assigned in 3:2 ratio to receive an *implantable cardioverter defibrillator* (ICD) (n=742*)* or *conventional medical therapy* (n=490). Patients were followed until death or end of study; the primary outcome is time to all-cause mortality.

Five binary risk factors have been identified as potential effect modifiers in a previous study [18]. They are New York Heart Association functional class (NYHA) > II, age > 70 years, blood urea nitrogen (BUN) > 26 mg/dl, QRS duration > 0.12 s, and atrial fibrillation. 112 patients with missing values in any of the five risk factors were excluded. Summary statistics on these five variables are included in Table 1.

Table 1 Summary Statistics: Five Risk Factors of the 1232 Patients

Characteristic	**Defibrillator Group n=674**	**Conventional-Therapy Group n=446**
Age (yr) > 70, n (%)	195 (29)	137 (31)
NYHA Functional Class >II, n (%)	202 (30)	123 (28)
Blood Urea Nitrogen>26mg/dl, n (%)	181 (27)	133 (30)
Atrial Fibrillation, n (%)	57 (8)	38 (9)
QRS interval≥0.12 sec, n (%)	233 (35)	134 (30)

32 cells were constructed based on these binary covariates; the size of each cell is shown in Table 2. Cell 28 was deleted in the simulation and analysis since both patients were in the treatment arm.

Table 2 The Information of the 32 Cells Based on the Five Risk Factors

Cell Number	**No. of Risk Factors**	**Size**	**Cell Number**	**No. of Risk Factors**	**Size**
1	0	346	17	3	47
2	1	16	18	3	3
3	1	99	19	3	9
4	1	47	20	3	31
5	1	72	21	3	6
6	1	96	22	3	56
7	2	6	23	3	24
8	2	5	24	3	4
9	2	35	25	3	17
10	2	10	26	3	22
11	2	58	27	4	6
12	2	43	28	4	2
13	2	7	29	4	3
14	2	40	30	4	3
15	2	48	31	4	24
16	2	23	32	5	4

The k and p pair was selected to provide the best power in the MaditII data. The explanation is provided in section 4. Specifically, $p = 0.5$ and $k = 100$ were used for the two-sided extreme value test, while $p = 0.1$ and $k = 300$ were used for the two-sided average value test. For one-sided tests, slightly different pairs were chosen. The selected pair of k and p for each test, and the corresponding p-values are shown in Table 3.

H_0^A and H_0^B were both rejected using extreme value tests and average value tests (p<0.0001). However, H_0^C was not rejected with p-values 0.5933 and 0.6483 for the one-sided extreme value test and average value test.

Table 3 The Pair of p and k Selected from the Simulations

	Test Statistics	$\boldsymbol{p}$	$\boldsymbol{k}$	**p-value**
Two-sided	S^E and T^E	0.1	300	<0.0001
	S^A and T^A	0.5	100	<0.0001
One-sided				
Test for beneficial effect	S^E	0.5	500	<0.0001
	S^A	0.1	300	<0.0001
Test for harmful effect	T^E	0.3	100	0.5933
	T^A	0.5	500	0.6483

2.4. Simulation

Two simulations were carried out. Simulation one is to find a set of k and p that has the most power under each simulation setting for each method, and to compare the power using the selected (k^*, p^*) in the proposed methods with some existing tests. Since data were simulated based on the Multicenter Automatic Defibrillator Implantation Trial II (MADIT II)[17, 18], (k^*, p^*) were applied to test the existence of the 'benefit group' and 'harm group' in MADIT II, as described in the previous section.

To better understand the reasons behind the choice of (k^*, p^*), we conducted simulation two. Different factors can affect the selection, so we simulated datasets with equal-sized cells to eliminate the effect from the unequal sample size.

All the parameter values used in the simulations are shown in Table 4.

Table 4 The Parameter Values in the Simulations

	h_0	β_1	β_2	β_3	β_4	β_5	β^{be}	β^{nb}
Simulation One-A	0.5	0.7	0.4	0.5	0.6	0.5	-0.4	NA
Simulation One-B	0.5	0.7	0.4	0.5	0.6	0.5	-0.4	0.4
Simulation Two-A	0.5	0.2	0.4	0.5	0.6	0.5	-0.4	NA
Simulation Two-B	0.5	0.2	0.4	0.5	0.6	0.5	-0.8	0.8
Simulation Two-C	0.5	0.2	0.4	0.5	0.6	0.5	-0.4	0.4

2.4.1 Simulation One

From the past papers [17-19], we believe the patient population in MADIT II consist at least 'benefit group' and 'neutral group', and we suspect the existence of a 'harm group'. As a result, we ran two simulation settings, one is to create a population consist with 'benefit group' and 'neutral group', to apply and examine the one-sided tests. The other setting is to imitate a patient population with all three types of patients ('benefit group', 'harm group' and 'neutral group'), to utilize and assess both the one-sided and two-sided tests. In both simulation settings, 1000 Monte Carlo datasets were generated based on MADIT II data.

2.4.1.1 Null and Alternative Hypotheses

Three sets of hypotheses mentioned in section 2 were used in the simulations:

1. $H_0^A: D_l = 0, l = 1,2, \dots, L$; $H_1^A: D_l \neq 0$, for some $l = 1,2, \dots, L$
2. $H_0^B: D_l \leq 0, l = 1,2, \dots, L$; $H_1^B: D_l > 0$, for some $l = 1,2, \dots, L$
3. $H_0^C: D_l \geq 0, l = 1,2, \dots, L$; $H_1^C: D_l < 0$, for some $l = 1,2, \dots, L$

2.4.1.2 Simulation One-A:

In this simulation, to produce the population with benefit and no effect sub-groups. The simulation setting is as follows:

- Benefit cells: Patients had QRS duration > 0.12s (16 cells, 367 patients);
- No treatment effect cells: the rest of the patients;
- Time to death was assumed to have exponential distribution with rate R_1;
- The five baseline characteristic variables from MADIT II data were used to generate the rate: $R_1 = h_0 \exp(\beta_1 * I(NYHA > II) + \beta_2 * I(age > 70) + \beta_3 * I(BUN > 26) + \beta_4 * I(QRS > 0.12) + \beta_5 * I(atrial\ fibtillation) + \beta^{be} * I(Benefit\ cells) * Trt)$
- Censoring time is uniformly distributed between 2 to 4 years.

Powers were calculated for both one-sided extreme value test and the average value test with $k = 100,200,300,400,500$ and with $p = 0.1,0.2,\ldots,0.5$, to search for the pair that gives the best power. Then compare these two tests using their corresponding optimal pair of k and p with the following existing methods:

1. The overall log-rank test (presented as 'Logrank Test' in the result Table);
2. The true Cox Proportional Hazard models;
3. One-sided likelihood ratio test by G&S[8];
4. One-sided range test by P&G[9].

2.4.1.3 Simulation One-B:

The modification setting one-B from one-A is the harm cells were created, besides the benefit and no effect cells. The results were compared with the same existing methods

as in simulation one-A, with additional comparison with two coxPH models fitted using wrong model assumptions.

The simulation setting is as follows:

- Benefit cells: Patients had QRS duration > 0.12s(16 cells, 367 patients);
- Harm cells: Patients had NYHA ≤ II, age ≤ 70, BUN ≤ 26 mg/dl, QRS duration ≤ 0.12s (2cells, 362 patients);;
- No treatment effect cells: the rest of the patients;
- The time to death was assumed to have exponential distribution with rate R_2;
- Similarly, the five baseline characteristic variables from MADIT II data were used to generate the rate: $R_2 = h_0 \exp(\beta_1 * I(NYHA > II) + \beta_2 * I(age > 70) + \beta_3 * I(BUN > 26) + \beta_4 * I(QRS > 0.12) + \beta_5 * I(atrial\ fibtillation) + \beta^{be} * I(Benefit\ cells) * Trt) + \beta^{nb} * I(harm\ cells) * Trt$;
- Simulated time was censored uniformly between 2 to 3 years.

The two wrong models were fitted as follows:

Wrong Cox Proportional Hazard Model 1:

$W_1 = h_0 \exp(\beta_1 * I(NYHA > II) + \beta_2 * I(age > 70) + \beta_3 * I(BUN > 26) + \beta_4 * I(QRS > 0.12) + \beta_5 * I(atrial\ fibtillation) + \beta^{trt} * Trt + \beta^{tage} * I(age > 70) * Trt)$;

Wrong Proportional Hazard Model 2: Five baseline variables + Trt

$W_2 = h_0 \exp(\beta_1 * I(NYHA > II) + \beta_2 * I(age > 70) + \beta_3 * I(BUN > 26) + \beta_4 * I(QRS > 0.12) + \beta_5 * I(atrial\ fibtillation) + \beta^{trt} * Trt)$.

For the wrong Cox Proportional Hazard models, Wald tests were used to test the hypotheses. The null hypotheses for the Wald tests are $\beta^{trt} = \beta^{tage} = 0$ for wrong model 1, and $\beta^{trt} = 0$ for wrong model 2.

2.4.1.4 Type I error rate

To examine the Type I error, no treatment effect was assumed, simulation data were constructed the same as simulation one-B with $\beta^{be} = \beta^{nb} = 0$. The type I errors were calculate for the two-sided extreme test with $k = 100,200,300,400,500$ and with $p = 0.1,0.2,\ldots,0.5$ using 1000 Monte Carlo datasets. The results are shown in Table 5. Several combination of k and p were used to spot-check the type I error for the rest of the proposed methods. The results were all around 0.05.

Table 5 The Type I Error Results of Two-Sided Test with S^E and T^E

		$\boldsymbol{k}$				
		100	200	300	400	500
$\boldsymbol{p}$	0.1	0.051	0.048	0.052	0.040	0.049
	0.2	0.050	0.046	0.062	0.052	0.053
	0.3	0.056	0.049	0.043	0.062	0.044
	0.4	0.058	0.048	0.050	0.052	0.050
	0.5	0.062	0.050	0.048	0.038	0.045

2.4.1.5 Results

Simulation One-A

The power for each combination of k and p for one-sided extreme value test and average value test was calculated as shown in Tables 6 and 7 respectively. In Table 6, when p increases from 0.1 to 0.5, the power gets larger, and reaches the maximum at 0.694 when $(k^*, p^*) = (400,0.5)$. On the contrary, in Table 7, the power gets smaller

when p increases from 0.1 to 0.5, and $(k^*, p^*) = (500,0.1)$ are found to give the largest power 0.719.

Table 6 The Power to detect existence of sub-populations with treatment benefit under simulation One-A using one-sided extreme value test S^E.

		k				
		100	200	300	400	500
p	0.1	0.442	0.468	0.429	0.458	0.452
	0.2	0.518	0.556	0.557	0.569	0.537
	0.3	0.624	0.616	0.609	0.628	0.647
	0.4	0.633	0.637	0.662	0.664	0.653
	0.5	0.659	0.675	0.67	0.694	0.673

Table 7 The Power to detect existence of sub-population with treatment benefit under simulation One-A using one-sided average value test with S^A

		k				
		100	200	300	400	500
p	0.1	0.648	0.71	0.711	0.709	0.719
	0.2	0.632	0.664	0.691	0.665	0.668
	0.3	0.619	0.61	0.611	0.632	0.649
	0.4	0.57	0.578	0.565	0.551	0.589
	0.5	0.535	0.511	0.537	0.524	0.538

Table 8 Power comparisons for simulation One-A

One-sided Extreme value test k=400 p=0.5	**One-sided Range Test**	**One-sided Average value test k=500 p=0.1**	**One-sided LR Test**	**Logrank Test**	**True Model**
0.694	0.346	0.719	0.669	0.147	0.950

The comparisons of the powers of different methods are shown in Table 8. The one-sided extreme value test is clearly better than its P&G counterpart (power = 0.694 vs 0.346). In addition, the one-sided average value test performs better than the one-sided G&S test (power = 0.719 vs 0.669). Overall, both of the proposed one-sided tests achieved much better power than the logrank test, and the average value test is slightly better than the extreme test.

Simulation One-B

The power follows similar trend for the two-sided test when p changes. The power reaches 0.839 when $(k^*, p^*) = (100,0.5)$ for the two-sided extreme value test, and 0.623 when $(k^*, p^*) = (300,0.1)$ for the two-sided average value test. (Table 9 and Table 10)

Table 9 The Power to detect existence of sub-population with treatment benefit under simulation One-B using the two-sided extreme value test

		k				
		100	200	300	400	500
p	0.1	0.72	0.719	0.71	0.708	0.715
	0.2	0.763	0.749	0.76	0.783	0.763
	0.3	0.83	0.795	0.808	0.809	0.822
	0.4	0.823	0.801	0.815	0.828	0.818
	0.5	0.839	0.822	0.829	0.82	0.827

Table 10 The Power to detect existence of sub-population with treatment benefit under simulation One-B using the two-sided average value test

		k				
		100	200	300	400	500
p	0.1	0.542	0.61	0.623	0.597	0.601
	0.2	0.503	0.5	0.537	0.536	0.561
	0.3	0.483	0.479	0.467	0.468	0.488
	0.4	0.384	0.362	0.368	0.394	0.382
	0.5	0.313	0.28	0.274	0.285	0.297

It is interesting that in the one-sided extreme value test and average value test, p were selected the same as the corresponding two-sided tests, i.e. $p = 0.5$ for the one-sided extreme value test and $p = 0.1$ for the one-sided average value test. One the other hand, when test for harmful effect, $p = 0.3$ and $p = 0.5$ were chosen for T^E and T^A respectively (Table 11-14). It may be because there are more cells that have beneficial effect than harmful effect, as a result, p for the two-sided tests are influenced by the one-sided test for beneficial effect.

Table 11 The Power to detect existence of sub-population with treatment benefit under simulation One-B using the one-sided extreme value test with S^E

		k				
		100	200	300	400	500
p	0.1	0.427	0.459	0.451	0.458	0.451
	0.2	0.503	0.529	0.534	0.568	0.544
	0.3	0.614	0.593	0.606	0.61	0.638
	0.4	0.623	0.625	0.641	0.664	0.647
	0.5	0.652	0.655	0.66	0.68	0.684

Table 12 The Power to detect existence of sub-population with treatment benefit under simulation One-B using the one-sided extreme value test with T^E

		k				
		100	200	300	400	500
p	0.1	0.721	0.711	0.722	0.686	0.709
	0.2	0.747	0.729	0.719	0.734	0.728
	0.3	0.751	0.734	0.733	0.738	0.733
	0.4	0.731	0.724	0.73	0.744	0.735
	0.5	0.691	0.735	0.702	0.729	0.73

Table 13 The Power to detect existence of sub-population with treatment benefit under simulation One-B using the one-sided average value test with S^A

		k				
		100	200	300	400	500
p	0.1	0.548	0.616	0.624	0.601	0.614
	0.2	0.487	0.489	0.52	0.528	0.544
	0.3	0.417	0.424	0.428	0.417	0.452
	0.4	0.31	0.291	0.304	0.313	0.295
	0.5	0.203	0.171	0.178	0.176	0.171

Table 14 The Power to detect existence of sub-population with treatment benefit under simulation One-B using the one-sided average value test with T^A

		k				
		100	200	300	400	500
p	0.1	0.175	0.151	0.143	0.136	0.137
	0.2	0.293	0.27	0.316	0.302	0.304
	0.3	0.386	0.368	0.383	0.394	0.383
	0.4	0.377	0.372	0.393	0.388	0.375
	0.5	0.371	0.36	0.371	0.371	0.395

The comparisons of the powers of different methods in this setting are shown in Table 15. Different from Table 8, methods using the extreme values (extreme value test and P&G range test) are in general better than using the average values, although they are all better than the logrank test, and the coxPH models with incorrect assumptions.

Table 15 Power comparisons for simulation One-B

Two-sided Extreme value test k=100 p=0.5	**Range Test**	**Two-sided Average value test k=300 p=0.1**	**LR Test**
0.839	0.726	0.623	0.559
Logrank Test	**Wrong Model 1**	**Wrong Model 2**	**True Model**
0.130	0.297	0.047	Close to 1

2.4.2 Simulation Two

In this simulation section, the focus is to better understand the nature of choosing the set of k and p using the two sets of test statistics. 32 cells with equal number of patients were generated in three different settings. Powers were calculated in these settings for all proposed tests with $k = 100, 200, \ldots, 500$ and $p = 0.1, 0.2, \ldots, 0.5$, to study which pair gives the best power, and to summarize any patterns.

2.4.2.1 Null and Alternative Hypotheses

Two sets of hypotheses were used for this section:

$$H_0^B: D_l \leq 0, l = 1,2, \ldots, L;\ H_1^B: D_l > 0, \text{for some } l = 1,2, \ldots, L$$

$$H_0^C: D_l \geq 0, l = 1,2, \ldots, L;\ H_1^C: D_l < 0, \text{for some } l = 1,2, \ldots, L$$

2.4.2.2 Simulation Two-A

- Five binary baseline characteristic variables I_1, I_2,…, I_5 were generated to create 32 cells with equal number of patients;
- Treatment control ratio is 1:1 in each cell;
- Benefit cells: $n_{benefit}$ out of 32 cells were randomly selected to be cells benefitting from the treatment in each simulated Monte Carlo dataset $(n_{benefit} = 4,8,16)$;
- No treatment effect cells: the rest of the patients;
- Time to death was assumed to have exponential distribution with rate R_3;
- $R_3 = h_0 \exp(\beta_1 * I_1 + \beta_2 * I_2 + \beta_3 * I_3 + \beta_4 * I_4 + \beta_5 * I_5 + \beta^{be} * I(Benefit\ cells) * Trt)$;
- Simulated time was censored uniformly between 1 to 2 years.

2.4.2.3 Simulation Two-B

- Five binary baseline characteristic variables were generated to create 32 cells with equal number of patients;
- Treatment control ratio is 1:1 in each cell;
- Benefit cells: $n_{benefit}$ out of 32 cells were randomly selected to be benefit cells in each Monte Carlo dataset $(n_{benefit} = 4,8,12)$;
- Harm cells: n_{harm} out of 32 cells were randomly selected to be benefit cells in each Monte Carlo dataset $(n_{harm} = 4)$
- No treatment effect cells: the rest of the patients;
- The time to death was assumed to have exponential distribution with rate R_4;

- $R_4 = h_0 \exp(\beta_1 * I_1 + \beta_2 * I_2 + \beta_3 * I_3 + \beta_4 * I_4 + \beta_5 * I_5 + \beta^{be} * I(Benefit\ cells) * Trt) + \beta^{nb} * I(harm\ cells) * Trt$;
- Simulated time was censored uniformly between 1 to 2 years.

2.4.2.4 Simulation Two-C

- Different β^{be}, β^{nb} (with smaller absolute values compare to the fourth simulation) were used to calculate rate R_5;
- Other settings were the same as the fourth simulation setting.

2.4.2.5 Results

Simulation Two-A

The power results of the one-sided tests using S^E and S^A with $n_{benefit} = 4,8,16$ are presented in Table 16 to Table 21 respectively. Similarly to the one-sided test results in simulation one-A, testing using S^A has increasing power when p decreases from 0.5 to 0.1; on the other hand, testing using S^E has increasing power when p increases from 0.1 to 0.5, for $n_{benefit} = 8,16$. For $n_{benefit} = 4$ using S^E, however, power increases and reach the maximum at $p^* = 0.4$, and starts to decrease. There is no noticeable pattern of power change when k changes.

$p^* = 0.1$ was chosen for all the scenarios using S^A, this makes sense since when using the average positive (or average negative), the more sub-populations have only beneficial (or harm) effects, the better the signal can be intensified. So smaller p and larger k are the top pick as in the simulation results.

When using S^E, things are a little bit different, since there is no harm cell to pull the beneficial cell into the opposite direction, and we are using a fairly small k (compare

to the large number of possible combination of subpopulations), we are trying to find a balance between choosing a larger p^* to provide a bigger sub-population with a better chance of more beneficial cells being selected into each sub-population, and choosing a small enough p^*, so not too many no treatment effect cells are included in each sub-population. As a results, $p^* = 0.5$ were chosen for $n_{benefit} = 8$, and $p^* = 0.4$ were chosen for $n_{benefit} = 4$.

Table 16 The Power to detect existence of sub-population with treatment benefit under simulation Two-A with $n_{benefit} = 16$ using the one-sided extreme value test with S^E

		k				
		100	200	300	400	500
p	0.1	0.6	0.626	0.612	0.625	0.629
	0.2	0.747	0.756	0.755	0.75	0.766
	0.3	0.802	0.806	0.833	0.808	0.839
	0.4	0.835	0.845	0.852	0.847	0.855
	0.5	0.855	0.868	0.878	0.856	0.877

Table 17 The Power to detect existence of sub-population with treatment benefit under simulation Two-A with $n_{benefit} = 8$ using the one-sided extreme value test with S^E

		k				
		100	200	300	400	500
p	0.1	0.307	0.297	0.309	0.324	0.314
	0.2	0.333	0.391	0.389	0.393	0.361
	0.3	0.383	0.39	0.404	0.433	0.41
	0.4	0.42	0.412	0.427	0.445	0.453
	0.5	0.437	0.43	0.43	0.454	0.452

Table 18 The Power to detect existence of sub-population with treatment benefit under simulation Two-A with $n_{benefit} = 4$ using the one-sided extreme value test with S^E

		k

		100	200	300	400	500
$\boldsymbol{p}$	0.1	0.144	0.149	0.168	0.14	0.186
	0.2	0.149	0.197	0.175	0.2	0.172
	0.3	0.176	0.183	0.177	0.214	0.176
	0.4	0.168	0.186	0.207	0.214	0.19
	0.5	0.158	0.189	0.205	0.204	0.21

Table 19 The Power to detect existence of sub-population with treatment benefit under simulation Two-A with $n_{benefit} = 16$ using the one-sided extreme value test with S^E

		$\boldsymbol{k}$				
		100	200	300	400	500
$\boldsymbol{p}$	0.1	0.879	0.891	0.896	0.903	0.89
	0.2	0.874	0.903	0.882	0.872	0.89
	0.3	0.855	0.865	0.884	0.89	0.882
	0.4	0.853	0.881	0.885	0.88	0.875
	0.5	0.852	0.879	0.882	0.875	0.873

Table 20 The Power to detect existence of sub-population with treatment benefit under simulation Two-A with $n_{benefit} = 8$ using the one-sided extreme value test with S^E

		$\boldsymbol{k}$				
		100	200	300	400	500
$\boldsymbol{p}$	0.1	0.446	0.456	0.463	0.449	0.447
	0.2	0.384	0.433	0.445	0.447	0.442
	0.3	0.436	0.422	0.42	0.439	0.428
	0.4	0.42	0.434	0.424	0.438	0.443
	0.5	0.409	0.433	0.424	0.431	0.441

Table 21 The Power to detect existence of sub-population with treatment benefit under simulation Two-A with $n_{benefit} = 4$ using the one-sided extreme value test with S^E

		k				
		100	200	300	400	500
p	0.1	0.18	0.195	0.227	0.211	0.206
	0.2	0.186	0.183	0.18	0.194	0.187
	0.3	0.17	0.172	0.171	0.194	0.181
	0.4	0.169	0.17	0.187	0.189	0.187
	0.5	0.159	0.166	0.188	0.183	0.183

Simulation Two-B and Two-C

Results in simulation two-B and two-C have very similar characteristics in terms of p, except the power in two-C is in general smaller due to smaller effects were used, so we are only showing results for simulation two-B here. A table summarizing the selected (p^*, k^*) and power can be found in Table 22.

For one-sided tests using S^A and T^A, and T^E, $p^* = 0.1$ was chosen for all the scenarios ($n_{benefit} = 4,8,12$, $n_{harm} = 4$). It is proven again, for the average value test, smaller p is the better choice.

For one-sided tests using S^E, $p^* = 0.1, 0.3, 0.4$ were chosen for $n_{benefit} = 4,8,12$ respectively (n_{harm} is always 4). It seems that p^* is chosen so that one of the sub-populations can be formed by just the 12 (or 8, or 4) beneficial cells. Similarly, when using T^E, ideally p^* should be selected to form a sub-population with the 4 harm cells, and $p^* = 0.1$ was indeed chosen.

Table 22 The power results of one-sided tests using S^E, T^E, S^A, and T^A and the corresponding chosen p and k

	S^E	T^E	S^A	T^A
Scenario	$n_{benefit} = 4$ $n_{harm} = 4$	$n_{benefit} = 4$ $n_{harm} = 4$	$n_{benefit} = 4$ $n_{harm} = 4$	$n_{benefit} = 4$ $n_{harm} = 4$
p^*	0.1	0.1	0.1	0.1
k^*	500	500	100	100
Power	0.536	0.543	0.237	0.136
Scenario	$n_{benefit} = 8$ $n_{harm} = 4$	$n_{benefit} = 8$ $n_{harm} = 4$	$n_{benefit} = 8$ $n_{harm} = 4$	$n_{benefit} = 8$ $n_{harm} = 4$
p^*	0.3	0.1	0.1	0.1
k^*	500	500	400	200
Power	0.891	0.526	0.789	0.018
Scenario	$n_{benefit} = 12$ $n_{harm} = 4$	$n_{benefit} = 12$ $n_{harm} = 4$	$n_{benefit} = 12$ $n_{harm} = 4$	$n_{benefit} = 12$ $n_{harm} = 4$
p^*	0.4	0.1	0.1	0.1
k^*	400	500	400	100
Power	0.996	0.48	0.991	0.001

2.5. Discussion

Why non-parametric is good? Methods require some assumption of a model. Modeling of interactions between covariates and treatment indicator faces challenges. A major one is choosing the correct covariates in the right forms for the interaction terms, which heavily affects the results. Even if the terms of interactions are linear, the orders of the terms are still need to be determined.[19]

In this chapter, we proposed a test that has a stochastic search built into the test statistic to detect signals that may not be linearly related to the multiple covariates. The one-sided and two-sided extreme value tests are based on the strongest positive and negative signals, while one-sided and two-sided average value tests are based on the

average positive and negative signals. The choice for the pair of k and p in the stochastic search are essential to this test.

The choice of k and p has a major impact on the power of the test. Our simulation studies suggest that larger values of k in general leads to better power or at least not far off the optimal power. The choice of parameter p depends on the method. For the average value tests, smaller p such as 0.1 is recommended. For the extreme value tests, it depends on the number of cells that have benefit (harmful) effect. When both benefit and harm groups exist, the optimal p^* can be roughly estimated by

$$the\ number\ of\ benefit\ (harmful)\ cells\ \div the\ total\ number\ of\ cells.$$

To construct the cells, we need non-missing clinical meaningful covariates. In this paper, we focused on RCTs with discrete covariates. To use the continuous covariates, they can be discretized by a series of the thresholds. The proposed methods can also be directly applied in observational studies, after the standard procedure to adjust for confounding variables before analyzing an observational study.

Although the harmful effect was not detected for ICD using the MADIT-II data, Shen et al(ref) showed that 4-12% may actually be harmed by an ICD in terms of two-year survival after ICD implantation. The difference is due to the fact that the current test cannot detect heterogeneity within a cell. In other words, if some patients in a cell is actually harmed by the ICD, such an effect may be masked if others in the same cell benefit from the ICD.

The proposed tests in this chapter can serve as a first-line procedure to protect against false discoveries of benefit and harm. In next chapter, we apply this test to help calculate power and benefit study design. Our next step is to develop models to identify

who benefit from the ICD and who do not, and the estimated average treatment effect in different sub-populations. Identification of sub-populations with treatment benefit or harm and estimation of treatment effect in sub-populations is an active research area[20-24]. Nevertheless, active search is prone to selection bias and inflation of the probability of false positives. Our test can serve as a gatekeeping procedure to prevent false positives.

CHAPTER 3. POWER CALCULATION FOR STUDY DESIGN

3.1 Background

As a post hoc test, the main purpose of the test is to help guide study design in the future. With some knowledge on the HTE, we can design a study to achieve desired power to detect HTE. In this chapter, the treatment effects are under normal distribution. Normal approximation and central limit theorem were used in the formulation. Some computational methods were used to help deal with some very complex formulation. In section 2, we introduce different theoretical methods to calculate power with different parameter settings using S^E and S^A. Then the simulations are done to discover how the power changes when each parameter changes in the test, and to check the performance of the theoretical results by checking the results between the powers calculated by theory and by simulation in section 3. We concluded this chapter with a discussion in section 4.

3.2 Method

As mentioned in chapter two, there are two sets of test statistics: the maximum, minimum test statistics ($S^E = \max(Z_i)$, $T^E = \min(Z_i)$); and the average positive, negative test statistics ($S^A = \sum_{i=1}^{k} \frac{\max(Z_i,0)}{k}$, $T^A = \sum_{i=1}^{k} \frac{\min(Z_i,0)}{k}$). Since the two test statistics in each set are symmetric, for simplicity, we will focus on using S^E and S^A in this chapter.

First, we introduce some general terms. Let Δ_l be the true mean difference for cell $l, l = 1,2, \dots, L$, and $\widehat{\Delta}_l$ be the estimated differences in sample means. For L cells (L relatively large), assume there are n^* control and rn^* treatment units in each cell. Assume

constant σ^2 within each cell for both treatment and control. For m randomly selected cells $\widetilde{W} = (W_1, W_2, \dots, W_m)$ based on the selection probability $p = m/L$, the Z statistics for testing the equality of the means can be written as

$$Z(\widetilde{W}) = \sqrt{\frac{n^*}{\left(1+\frac{1}{r}\right)m\sigma^2}} \sum_{j=1}^{m} \widehat{\Delta}_{W_j} \triangleq a \sum_{j=1}^{m} \widehat{\Delta}_{W_j}$$

In general, to calculate the power, first we calculate a critical value to control for the type one error, and then use it to calculate the power.

$$\begin{cases} \Pr_{H_0}(S^E(D) > x_0) = \alpha \\ \Pr_{H_A}(S^E(D) > x_0) = 1-\beta \end{cases},$$

and

$$\begin{cases} \Pr_{H_0}(S^A(D) > x_0) = \alpha \\ \Pr_{H_A}(S^A(D) > x_0) = 1-\beta \end{cases}$$

All the power results will be in terms of p and k, we will determine the best pair of p and k to use.

3.2.1 Power Calculation with S^E

First, for a given type I error α, a critical value x_0 can be determined by $\Pr_{H_0}(S^E(D) > x_0) = \alpha$. This equation can be solved using the joint distribution of $Z(\widetilde{W}^{(1)}), Z(\widetilde{W}^{(2)}), \dots, Z(\widetilde{W}^{(k)})$. Under the null,

$$\Pr_{H_0}(S^E(D) < x_0) = 1-\alpha$$

$$= \Pr_{H_0}\left(Z(\widetilde{W}^{(1)}) \le x_0\, Z(\widetilde{W}^{(2)}) \le x_0, \dots, Z(\widetilde{W}^{(k)}) \le x_0\right).$$

Using the critical value x_0, the power of the test can be represented by

$$\Pr_{H_A}(S^E(D) > x_0) = 1 - \beta$$

under the alternative hypothesis, which can be derived into

$$\beta = \Pr_{H_A}\left(Z(\widetilde{W}^{(1)}) \le x_0\, Z(\widetilde{W}^{(2)}) \le x_0\, , \dots , Z(\widetilde{W}^{(k)}) \le x_0\right).$$

The joint distribution of $Z(\widetilde{W}^{(1)}), Z(\widetilde{W}^{(2)}), \dots , Z(\widetilde{W}^{(k)})$ can be represented by k, p, n^* and r. So we can determine k and p by minimizing the type II error β, and also this can help us design a study by choosing the most appropriate sample size in each treatment arm and cell.

Now we just need to determine the joint distribution of $Z(\widetilde{W}^{(1)}), Z(\widetilde{W}^{(2)}), \dots , Z(\widetilde{W}^{(k)})$. Under the null hypothesis that there is no mean difference between the treatment and control in each cell, so $Z\left(\widetilde{W}\right)|\widetilde{W}$ is $N(0,1)$. For independently drawn $\widetilde{W}^{(1)}, \widetilde{W}^{(2)}, \dots , \widetilde{W}^{(k)}$, the joint distribution of $Z(\widetilde{W}^{(1)}), Z(\widetilde{W}^{(2)}), \dots$, $Z(\widetilde{W}^{(k)})$ can be described as

$$\left(Z(\widetilde{W}^{(1)}), \dots , Z(\widetilde{W}^{(k)})\right)^T \sim N\left[0 I_{k\times 1},\ \begin{pmatrix} 1 & p & \dots & p \\ p & 1 & \dots & p \\ \vdots & \vdots & \ddots & \vdots \\ p & p & \dots & 1 \end{pmatrix}_{k\times k}\right]$$

according to appendix A.

The joint distribution of $Z(\widetilde{W}^{(1)}), Z(\widetilde{W}^{(2)}), \dots , Z(\widetilde{W}^{(k)})$ under the alternative hypothesis is discussed in two situations, with fixed Δ_l and random Δ_l.

3.2.1.1 $\left(Z(\widetilde{W}^{(1)}), \dots , Z(\widetilde{W}^{(k)})\right)^T$ Distribution With Fixed Δ_l

Assume under the alternative hypothesis, each Δ_l has $N(\mu, \tau^2)$, and they are fixed. This assumption is realistic since the true treatment effect for each subgroup of patients is normally fixed.

Then $Z(\widetilde{W})|\widetilde{W}\sim N\left(a\sum_{j=1}^{m}\Delta_{W_j},1\right)$, and $\sum_{j=1}^{m}\Delta_{W_j}$ is approximately normal with mean $m\mu$ and $cm\tau^2$, where $c=(L-m)/(L-1)$ is the finite sample correction factor, followed by $Z(\widetilde{W})\sim N(am\mu,1+a^2cm\tau^2)$.

According to appendix A, for any $Z(W^{(s)})$, and $Z(W^{(t)})$, $s\neq t$ and $s,t=1,\ldots,k$.

$$Cov\left(Z(\widetilde{W}^{(s)}),Z(\widetilde{W}^{(t)})\right)=E\left(Z(\widetilde{W}^{(s)})Z(\widetilde{W}^{(t)})\right)-E^2\left[Z(\widetilde{W})\right]=p$$

and

$$\left(Z(\widetilde{W}^{(1)}),\ldots,Z(\widetilde{W}^{(k)})\right)^T$$

$$\sim N\left[am\mu I_{k\times 1},\ \begin{pmatrix}1+a^2cm\tau^2 & p & \ldots & p\\ p & 1+a^2cm\tau^2 & \ldots & p\\ \vdots & \vdots & \ddots & \vdots\\ p & p & \ldots & 1+a^2cm\tau^2\end{pmatrix}_{k\times k}\right]$$

3.2.1.2 $\left(Z(\widetilde{W}^{(1)}),\ldots,Z(\widetilde{W}^{(k)})\right)^T$ Distribution With Random Δ_l

In the situation that Δ_ls are random variables, there is no need for the finite sample corrector c. After the derivation in appendix A,

$$Cov\left(Z(\widetilde{W}^{(s)}),Z(\widetilde{W}^{(t)})\right)=E\left(Z(\widetilde{W}^{(s)})Z(\widetilde{W}^{(t)})\right)-E^2\left[Z(\widetilde{W})\right]$$

$$=a^2m^2\mu^2+p+a^2p^2L\tau^2-a^2m^2\mu^2=p+a^2p^2L\tau^2\triangleq p^*$$

$$\left(Z(\widetilde{W}^{(1)}),\ldots,Z(\widetilde{W}^{(k)})\right)^T$$

$$\sim N\left[am\mu I_{k\times 1},\ \begin{pmatrix}1+a^2m\tau^2 & p^* & \ldots & p^*\\ p^* & 1+a^2m\tau^2 & \ldots & p^*\\ \vdots & \vdots & \ddots & \vdots\\ p^* & p^* & \ldots & 1+a^2m\tau^2\end{pmatrix}_{k\times k}\right]$$

3.2.2 Power Calculation with S^A

The idea of calculating the distribution of S^A is to use the central limit theorem. Let $A^{(i)} = max\left(Z\left(\widetilde{W}^{(i)}\right), 0\right), i = 1, \dots, k$, then $A^{(1)}|\hat{\tilde{\Delta}}$, …, $A^{(k)}|\hat{\tilde{\Delta}}$ are independent, and $A^{(i)}|\hat{\tilde{\Delta}}$ has a normal distribution $A^{(i)}|\hat{\tilde{\Delta}} \sim N\left(E\left[A|\hat{\tilde{\Delta}}\right], Var\left[A|\hat{\tilde{\Delta}}\right]\right)$. By central limit theorem, $\bar{A}|\hat{\tilde{\Delta}} = S^A|\hat{\tilde{\Delta}} \sim N\left(E\left[A|\hat{\tilde{\Delta}}\right], Var\left[A|\hat{\tilde{\Delta}}\right]/k\right)$. The distribution of $\bar{A} = S^A$ is hard to determine, but we can use numerical methods to achieve our goal. To control for the type I error to be 0.05, we can calculate a threshold x_0 under the null hypothesis, so that for a large number of given $\hat{\tilde{\Delta}}_i$'s, the average probability of $\left(\bar{A}|\hat{\tilde{\Delta}}\right)_i$ exceeding x_0 is 0.05:

$$\frac{1}{N_{large}} \sum_{ii=1}^{N_{large}} Pr\left[\left(\bar{A}|\hat{\tilde{\Delta}}\right)_{ii} > x_0\right] = 0.05.$$

Since we can calculate the distribution of each $\left(\bar{A}|\hat{\tilde{\Delta}}\right)_{ii}$ given $\left(\hat{\tilde{\Delta}}\right)_{ii}$, the $Pr\left[\left(\bar{A}|\hat{\tilde{\Delta}}\right)_{ii} > x_0\right]$ is easy to write out.

Before determine $E\left[A|\hat{\tilde{\Delta}}\right]$ and $Var\left[A|\hat{\tilde{\Delta}}\right]$, we need to know the distribution of $a\sum_{j=1}^{m} \hat{\Delta}_{W_j}$. Let $c_0 = {(L-m)}/{(L-1)}$, $a\sum_{j=1}^{m} \hat{\Delta}_{W_j}$ has an approximately normal distribution $N(am\hat{\mu}, c_0 a^2 m\hat{\tau}^2)$, where $\hat{\mu} = \frac{\sum_{l=1}^{L} \hat{\Delta}_l}{L}$ and $\hat{\tau}^2 = \frac{\sum_{l=1}^{L} (\hat{\Delta}_l - \hat{\mu})^2}{L}$. Distribution of $\hat{\mu}$ and $\hat{\tau}^2$can be calculated based on the distribution of $\hat{\Delta}_l$. Under the null hypothesis, $\hat{\Delta}_l$ has a normal distribution $N(0, \delta_0^2)$, where $\delta_0^2 = \frac{\sigma^2}{n^*}(1 + 1/r)$. Then, $\hat{\mu} \underset{\sim}{approx} N(0, \delta_0^2/L)$, $\frac{L-1}{\delta_0^2}\hat{\tau}^2 \underset{\sim}{approx} \chi^2(L-1)$, and they are independent.

With the distribution of $\hat{\mu}$ and $\hat{\tau}^2$known, generating a pair of $\hat{\mu}$ and $\hat{\tau}^2$ is equivalent to generating $\widetilde{\Delta}$ (where $\widetilde{\Delta} = \Delta_1, \ldots, \Delta_L$), and then use the estimated $\hat{\widetilde{\Delta}}$ to calculate them. For each generated pair of $\hat{\mu}$ and $\hat{\tau}^2$, assuming $B = am\hat{\mu}$, and $C = \sqrt[2]{(c_0 a^2 m \hat{\tau}^2)}$, then using truncated normal distribution property:

$$E\left[A|\hat{\widetilde{\Delta}}\right] = E\left[\max\left(Z(\widetilde{W}),0\right)|\hat{\widetilde{\Delta}}\right] = Pr\left(Z>0|\hat{\widetilde{\Delta}}\right)E\left[Z|\hat{\widetilde{\Delta}}, Z>0\right]$$

$$= \left[1-\Phi\left(\frac{0-B}{C}\right)\right]\left[B + C\lambda\left(-\frac{B}{C}\right)\right]$$

where $\lambda(x) = \phi(x)/[1-\Phi(x)]$, $\phi(x)$ and $\Phi(x)$ are the density function and the cumulative probability function of standard normal.

$$V\left[A|\hat{\widetilde{\Delta}}\right] = E\left[\max^2\left(Z(\widetilde{W}),0\right)|\hat{\widetilde{\Delta}}\right] - E^2\left[\max\left(Z(\widetilde{W}),0\right)|\hat{\widetilde{\Delta}}\right]$$

$$= \Pr\left(Z>0|\hat{\widetilde{\Delta}}\right) E\left[Z^2(\widetilde{W})|\hat{\widetilde{\Delta}}, Z>0\right] - E^2\left[\max\left(Z(\widetilde{W}),0\right)|\hat{\widetilde{\Delta}}\right]$$

$$= \Pr\left(Z>0|\hat{\widetilde{\Delta}}\right)\left\{E\left[Z^2(\widetilde{W})|\hat{\widetilde{\Delta}}, Z>0\right] - E^2\left[Z(\widetilde{W})|\hat{\widetilde{\Delta}}, Z>0\right]\right\}$$

$$+ \Pr\left(Z>0|\hat{\widetilde{\Delta}}\right)E^2\left[Z(\widetilde{W})|\hat{\widetilde{\Delta}}, Z>0\right] - E^2\left[\max\left(Z(\widetilde{W}),0\right)|\hat{\widetilde{\Delta}}\right]$$

$$= \Pr\left(Z>0|\hat{\widetilde{\Delta}}\right) V\left[Z(\widetilde{W})|\hat{\widetilde{\Delta}}, Z>0\right] + \Pr\left(Z>0|\hat{\widetilde{\Delta}}\right)E^2\left[Z(\widetilde{W})|\hat{\widetilde{\Delta}}, Z>0\right]$$

$$- E^2\left[\max\left(Z(\widetilde{W}),0\right)|\hat{\widetilde{\Delta}}\right]$$

Let $\delta(x) = \lambda(x) * [\lambda(x) - x]$, then

$$V\left[A|\hat{\tilde{\Delta}}\right] = \left[1 - \Phi\left(\frac{0-B}{C}\right)\right] C^2 \left[1 - \delta\left(-\frac{B}{C}\right)\right]$$

$$+ \left[1 - \Phi\left(\frac{0-B}{C}\right)\right]\left[B + C\lambda\left(-\frac{B}{C}\right)\right]^2 + E^2\left[A|\hat{\tilde{\Delta}}\right]$$

Under the alternative distribution, power can be calculated using $\frac{1}{N_{large}}\sum_{ii=1}^{N_{large}} Pr\left[\left(\bar{A}|\hat{\tilde{\Delta}}\right)_{ii} > x_0\right]$. The distribution of each $\left(\bar{A}|\hat{\tilde{\Delta}}\right)_{ii}$ under the alternative distribution can be calculated in two scenarios: with fixed Δ_l and random Δ_l.

3.2.2.1 $A|\hat{\tilde{\Delta}}$ Distribution With Fixed Δ_l

Assume under the alternative hypothesis, each Δ_l has $N(\mu, \tau^2)$, and they are fixed. The distribution of $\hat{\Delta}_l$ is then $N(\mu, \delta_1^2)$, where $\delta_1^2 = \frac{\sigma^2}{n^*}(1 + 1/r)$, and $\hat{\mu} \underset{\sim}{approx} N(\mu, \delta_1^2/L)$, $\frac{L-1}{\delta_1^2}\hat{\tau}^2 \underset{\sim}{approx} \chi^2(L-1)$, and they are independent.

Similar to $E\left[A|\hat{\tilde{\Delta}}\right]$ and $V\left[A|\hat{\tilde{\Delta}}\right]$ calculated earlier under the null hypothesis. For each generated pair of $\hat{\mu}$ and $\hat{\tau}^2$ under the alternative hypothesis and the fixed Δ_l assumption,

$$E\left[A|\hat{\tilde{\Delta}}\right] = \left[1 - \Phi\left(\frac{0-B}{C}\right)\right]\left[B + C\lambda\left(-\frac{B}{C}\right)\right]$$

and

$$V\left[A|\hat{\tilde{\Delta}}\right] = \left[1 - \Phi\left(\frac{0-B}{C}\right)\right] C^2 \left[1 - \delta\left(-\frac{B}{C}\right)\right]$$

$$+ \left[1 - \Phi\left(\frac{0-B}{C}\right)\right]\left[B + C\lambda\left(-\frac{B}{C}\right)\right]^2 + E^2\left[A|\hat{\tilde{\Delta}}\right]$$

where $B = am\hat{\mu}$, and $C = \sqrt[2]{(c_0 a^2 m \hat{\tau}^2)}$, $\lambda(x) = \phi(x)/[1 - \Phi(x)]$, $\phi(x)$ and $\Phi(x)$ are the density function and the cumulative probability function of standard normal.

3.2.2.2 $A|\hat{\tilde{\Delta}}$ Distribution With Random Δ_l

Assume under the alternative hypothesis, each Δ_l has $N(\mu, \tau^2)$, and they are random. $E\left[A|\hat{\tilde{\Delta}}\right]$ and $V\left[A|\hat{\tilde{\Delta}}\right]$ can be calculated using the exact same formula as above, once pairs of $\hat{\mu}$ and $\hat{\tau}^2$are generated under the alternative hypothesis with random Δ_l assunptions. The distribution of $\hat{\Delta}_l$ in this case is $N(\mu, \delta_2^2)$, where $\delta_2^2 = \frac{\sigma^2}{n^*}(1 + 1/r) + \tau^2$. $\hat{\mu} \underset{\sim}{approx} N(\mu, \delta_2^2/L)$, $\frac{L-1}{\delta_2^2}\hat{\tau}^2 \underset{\sim}{approx} \chi^2(L-1)$, and they are independent.

3.2.3 Unequal n^* in Each Cell (Two Values)

In a population of patients, each cell has equal amount of patients is really rare, the more likely situation is most of the cells have a few patients, the rest have more patients. Assume $n^* = \begin{cases} n_1 & p_1 \\ n_2 & p_2 \end{cases}$, for k randomly selected cells, $W = (W_1, W_2, \ldots, W_k)$, $Z(W) = \frac{\sum_{j=1}^{k} V_{w_j}\hat{\Delta}_{w_j}}{Var\left(\sum_{j=1}^{k} V_{w_j}\hat{\Delta}_{w_j}\right)}$ where $V_{w_j} = \frac{n^*_{w_j}}{\sum_{j=1}^{k} n^*_{w_j}}$, then $Var\left(\sum_{j=1}^{k} V_{w_j}\hat{\Delta}_{w_j}\right) = \frac{\sigma^2\left(1+\frac{1}{r}\right)}{\sum_{j=1}^{k} n^*_{w_j}}$

So $Z(W) = \frac{\sum_{j=1}^{k} \frac{n^*_{w_j}}{\sum_{j=1}^{k} n^*_{w_j}}\hat{\Delta}_{w_j}}{\sqrt{\frac{\sigma^2\left(1+\frac{1}{r}\right)}{\sum_{j=1}^{k} n^*_{w_j}}}} = \frac{\sum_{j=1}^{k} n^*_{w_j}\hat{\Delta}_{w_j}}{\sqrt{\sum_{j=1}^{k} n^*_{w_j}\sigma^2\left(1+\frac{1}{r}\right)}}$.

Let $b = 1\Big/\sqrt{\sigma^2\left(1+\frac{1}{r}\right)}$, under the sharp null that the means are the same between the treatment arms and the control arms for all cells, $Z(W)|W \sim N(0,1)$, therefore $Z(W) \sim N(0,1)$.

Given $\hat{\Delta}$, $E\big(Z(W)|\hat{\Delta}\big) = bE\left[\frac{\sum_{j=1}^{k} n_{w_j}^{*}\hat{\Delta}_{w_j}}{\sqrt{\sum_{j=1}^{k} n_{w_j}^{*}}}|\hat{\Delta}\right] = b\sum_{i=1}^{L} V_i\hat{\Delta}_i\, E[\sqrt{\sum_{j=1}^{k} n_{w_j}^{*}}]$. Let $c =$

$E[\sqrt{\sum_{j=1}^{k} n_{w_j}^{*}}] = \sum_{a_1=0}^{k}\sqrt{a_1 n_1 + (k-a_1)n_2}\binom{k}{a_1} p_1^{a_1}\, p_2^{(k-a_1)}$, then

$Cov\big[Z\big(W^{(1)}\big), Z\big(W^{(2)}\big)\big] = E\big[E^2[Z(W)]|\hat{\Delta}\big] = b^2c^2E\left[\big(\sum_{i=1}^{L} V_i\hat{\Delta}_i\big)^2\right] = c^2\frac{1}{\sum_{i=1}^{L} n_i^{*}} \triangleq \rho_1^{*}$

For independently drawn $\widetilde{W}^{(1)}, \widetilde{W}^{(2)}, \ldots, \widetilde{W}^{(k)}$, the joint distribution of

$Z(\widetilde{W}^{(1)}), Z(\widetilde{W}^{(2)}), \ldots, Z(\widetilde{W}^{(k)})$ can be described as

$$\big(Z(\widetilde{W}^{(1)}), \ldots, Z(\widetilde{W}^{(k)})\big)^T \sim N\left[0I_{k\times 1},\ \begin{pmatrix} 1 & \rho_1^{*} & \ldots & \rho_1^{*} \\ \rho_1^{*} & 1 & \ldots & \rho_1^{*} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_1^{*} & \rho_1^{*} & \ldots & 1 \end{pmatrix}_{k\times k}\right].$$

Under the alternative distribution, assume L Δ_is are fixed and follow $N(\mu, \tau^2)$. Then,

$Z(W)|W \sim N(b\sum_{j=1}^{k} V_{w_j}\Delta_{w_j}, 1)$, and $\Delta_{w_j} \sim approx \sim N(\mu, c_0\tau^2), c_0 = \frac{L-k}{L-1}$.

$E[Z(W)] = E\big[E[Z(W)|W]\big] = E\left[b\frac{\sum_{j=1}^{k} n_{w_j}^{*}\Delta_{w_j}}{\sqrt{\sum_{j=1}^{k} n_{w_j}^{*}}}\right] = b\mu E\left[\frac{\sum_{j=1}^{k} n_{w_j}^{*}}{\sqrt{\sum_{j=1}^{k} n_{w_j}^{*}}}\right]$(Assume $\hat{\Delta} \perp n^{*}$)

$= b\mu E\left[\sqrt{\sum_{j=1}^{k} n_{w_j}^{*}}\right] = bc\mu$, where $c = E\left[\sqrt{\sum_{j=1}^{k} n_{w_j}^{*}}\right]$

$Var[Z(W)] = E\big[Var[Z(W)|W]\big] + Var\big[E[Z(W)|W]\big] = 1 + b^2Var\left[\frac{\sum_{j=1}^{k} n_{w_j}^{*}\Delta_{w_j}}{\sqrt{\sum_{j=1}^{k} n_{w_j}^{*}}}\right] \triangleq \delta_3^2.$

$Var\left[\frac{\sum_{j=1}^{k} n_{w_j}^{*}\Delta_{w_j}}{\sqrt{\sum_{j=1}^{k} n_{w_j}^{*}}}\right] = c_0\tau^2E\left[\frac{\sum_{j=1}^{k} {n_{w_j}^{*}}^2}{\sum_{j=1}^{k} n_{w_j}^{*}}\right] + \mu^2(k(p_1n_1 + p_2n_2) - c^2) = 1 + b^2[c_0\tau^2d +$

$\mu^2(k(p_1n_1 + p_2n_2) - c^2)]$, where $d = E\left[\frac{\sum_{j=1}^{k} {n_{w_j}^{*}}^2}{\sum_{j=1}^{k} n_{w_j}^{*}}\right] =$

$\sum_{a_1=0}^{k}\left\{\frac{[a_1n_1^2 + (k-a_1)n_2^2]}{a_1n_1 + (k-a_1)n_2}\binom{k}{a_1} p_1^{a_1}\, p_2^{(k-a_1)}\right\}$.

$$E\left[Z\left(W^{(1)}\right)Z\left(W^{(2)}\right)\right] = E\left[E^2[Z(W)]|\hat{\Delta}\right] = b^2c^2E\left[\left(\sum_{i=1}^{L} V_i\hat{\Delta}_i\right)^2\right] = \frac{c^2}{\sum_{i=1}^{L} n_i^*} + b^2c^2\mu^2$$

$$\Rightarrow Cov\left[Z\left(W^{(1)}\right), Z\left(W^{(2)}\right)\right] = E\left[Z\left(W^{(1)}\right)Z\left(W^{(2)}\right)\right] - E\left[Z\left(W^{(1)}\right)\right]E\left[Z\left(W^{(2)}\right)\right]$$

$$= \frac{c^2}{\sum_{i=1}^{L} n_i^*} + b^2c^2\mu^2 - b^2c^2\mu^2 \triangleq \rho_2^*$$

$$\left(Z(\widetilde{W}^{(1)}), \dots, Z(\widetilde{W}^{(k)})\right)^T \sim N\left[bc\mu I_{k\times 1},\ \begin{pmatrix} \delta_3^2 & \rho_2^* & \dots & \rho_2^* \\ \rho_2^* & \delta_3^2 & \dots & \rho_2^* \\ \vdots & \vdots & \ddots & \vdots \\ \rho_2^* & \rho_2^* & \dots & \delta_3^2 \end{pmatrix}_{k\times k}\right]$$

More detailed derivation can be found in appendix B.

3.2.4 Unequal n^* in Each Cell (Shifted Poisson Distribution)

Sometimes, n^* may follow a certain distribution, in this dissertation, shifted Poisson distribution was used as an example. Numerical methods were also applied to help calculate certain values. Detailed derivation can be found in appendix C.

Assume $x \sim poisson(\lambda)$, and let $n^* = x + 1$, then n^* is shifted Poisson distributed.

For k randomly selected cells, $W = (W_1, W_2, \dots, W_k)$, $Z(W) = \frac{\sum_{j=1}^{k} V_{w_j}\hat{\Delta}_{w_j}}{Var\left(\sum_{j=1}^{k} V_{w_j}\hat{\Delta}_{w_j}\right)}$ where

$V_{w_j} = \frac{n_{w_j}^*}{\sum_{j=1}^{k} n_{w_j}^*}$. Since $Var\left(\sum_{j=1}^{k} V_{w_j}\hat{\Delta}_{w_j}\right) = \frac{\sigma^2\left(1+\frac{1}{r}\right)}{\sum_{j=1}^{k} n_{w_j}^*}$, then $Z(W) = \frac{\sum_{j=1}^{k} n_{w_j}^*\hat{\Delta}_{w_j}}{\sqrt{\sum_{j=1}^{k} n_{w_j}^*\sigma^2\left(1+\frac{1}{r}\right)}}$.

Let $b = 1\Big/\sqrt{\sigma^2\left(1+\frac{1}{r}\right)}$, under the sharp null that the means are the same between the treatment arms and the control arms for all cells, $Z(W)|W \sim N(0,1)$, therefore

$Z(W) \sim N(0,1)$. Similarly, given $\hat{\Delta}$, $E\left(Z(W)|\hat{\Delta}\right) = bE\left[\frac{\sum_{j=1}^{k} n_{w_j}^*\hat{\Delta}_{w_j}}{\sqrt{\sum_{j=1}^{k} n_{w_j}^*}}|\hat{\Delta}\right] =$

$b\sum_{i=1}^{L} V_i\hat{\Delta}_i\, E[\sqrt{\sum_{j=1}^{k} n_{w_j}^*}]$. Since $x_{w_j} \sim Poisson(\lambda)$, then $y \triangleq \sum_{j=1}^{k} x_{w_j} \sim Poisson(k\lambda)$.

Let $c \triangleq E\left[\sqrt{\sum_{j=1}^{k} n_{w_j}^*}\right] = E\left[\sqrt{y+k}\right] = \sum_{y=0}^{\infty} \sqrt{y+k} \frac{\lambda^y}{y!} e^{-\lambda}$, which can be calculated by Monte Carlo Method. So $Cov\left[Z\left(W^{(1)}\right), Z\left(W^{(2)}\right)\right] = E\left[E^2[Z(W)]|\hat{\Delta}\right] = c^2 \frac{1}{\sum_{i=1}^{L} n_i^*} \triangleq \rho_3^*$.

Above all, under Null, $\left[Z\left(W^{(1)}\right), Z\left(W^{(2)}\right)\right]^T \sim N\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho_3^*\\ \rho_3^* & 1\end{pmatrix}\right]$.

For independently drawn $\widetilde{W}^{(1)}, \widetilde{W}^{(2)}, \ldots, \widetilde{W}^{(k)}$, the joint distribution of $Z(\widetilde{W}^{(1)}), Z(\widetilde{W}^{(2)}), \ldots, Z(\widetilde{W}^{(k)})$ can be described as

$$\left(Z(\widetilde{W}^{(1)}), \ldots, Z(\widetilde{W}^{(k)})\right)^T \sim N\left[0I_{k\times 1}, \begin{pmatrix}1 & \rho_3^* & \ldots & \rho_3^*\\ \rho_3^* & 1 & \ldots & \rho_3^*\\ \vdots & \vdots & \ddots & \vdots\\ \rho_3^* & \rho_3^* & \ldots & 1\end{pmatrix}_{k\times k}\right]$$

Under alternative hypothesis, $Z(W)|W \sim N(b\sum_{j=1}^{k} V_{w_j}\Delta_{w_j}, 1)$, and $\Delta_{w_j} \sim approx \sim N(\mu, c_0\tau^2), c_0 = \frac{L-k}{L-1}$.

$E[Z(W)] = E\left[E[Z(W)|W]\right] = b\mu E\left[\frac{\sum_{j=1}^{k} n_{w_j}^*}{\sqrt{\sum_{j=1}^{k} n_{w_j}^*}}\right]$ (assume $\hat{\Delta} \perp n^*$)

$= b\mu E\left[\sqrt{\sum_{j=1}^{k} n_{w_j}^*}\right] = bc\mu$ where $c = E\left[\sqrt{\sum_{j=1}^{k} n_{w_j}^*}\right]$, and

$Var[Z(W)] = E\left[Var[Z(W)|W]\right] + Var\left[E[Z(W)|W]\right] = 1 + b^2[c_0\tau^2 d + \mu^2(k\lambda + k - c^2)]$, where $d = E\left[\frac{\sum_{j=1}^{k} {n_{w_j}^*}^2}{\sum_{j=1}^{k} n_{w_j}^*}\right]$, which can be calculated by Monte Carlo

$E\left[Z\left(W^{(1)}\right)Z\left(W^{(2)}\right)\right] = E\left[E^2[Z(W)]|\hat{\Delta}\right] = b^2c^2E\left[\left(\sum_{i=1}^{L} V_i\hat{\Delta}_i\right)^2\right] = \frac{c^2}{\sum_{i=1}^{L} n_i^*} + b^2c^2\mu^2$

$\Rightarrow Cov\left[Z\left(W^{(1)}\right), Z\left(W^{(2)}\right)\right] = E\left[Z\left(W^{(1)}\right)Z\left(W^{(2)}\right)\right] - E\left[Z\left(W^{(1)}\right)\right]E\left[Z\left(W^{(2)}\right)\right] =$

$\frac{c^2}{\sum_{i=1}^{L} n_i^*} + b^2c^2\mu^2 - b^2c^2\mu^2 \triangleq \rho_4^*$

$$\left(Z(\widetilde{W}^{(1)}),\dots,Z(\widetilde{W}^{(k)})\right)^T \sim N\left[bc\mu I_{k\times 1},\ \begin{pmatrix}\delta_3^2 & \rho_4^* & \dots & \rho_4^*\\ \rho_4^* & \delta_3^2 & \dots & \rho_4^*\\ \vdots & \vdots & \ddots & \vdots\\ \rho_4^* & \rho_4^* & \dots & \delta_3^2\end{pmatrix}_{k\times k}\right]$$

3.3 Simulation

There are two objectives of the simulation. One is to discover how the power changes when each parameter changes in the test. The other is to compare the results between the power calculated by theory and by simulation. For each objective, the simulation section will consist two parts, the power simulation using S^E, and using S^A. The following parameter combinations were used:

1. For equal number of patients in each cell $n^* = 3,5,10,15$, or
2. $n^* = \begin{cases}5 & p_1 = 0.8\\ 20 & p_2 = 0.2\end{cases}$, or $n^* = \begin{cases}5 & p_1 = 0.8\\ 30 & p_2 = 0.2\end{cases}$, or
3. $n^* = x + 1$, $x \sim poisson(5)$, $x \sim poisson(10)$
4. $L = 30,100$
5. $\sigma^2 = 1.5, 2, 2.5, 3$
6. $\tau^2 = 0.25, 0.5, 1, 1.5$
7. $r = 1, 2$

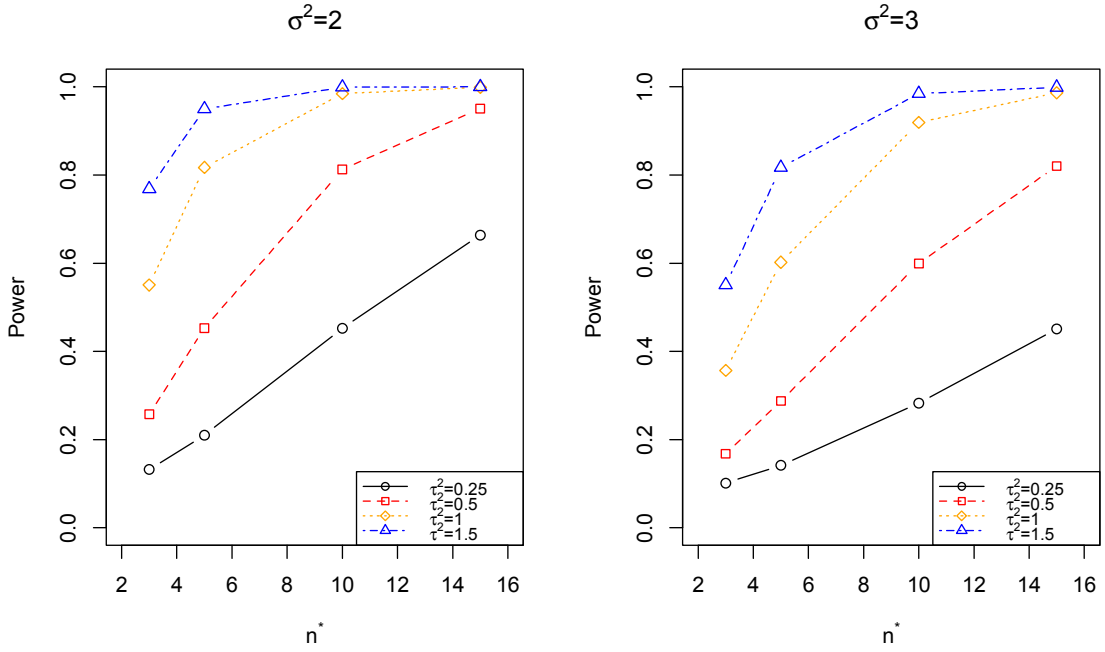
For both section in the second objective, under the null hypothesis, the outcomes for patients under treatment and control for all L cells were generated from the same normal distribution $N(1, \sigma^2)$. Under the alternative hypothesis, the outcomes under the control were generated from $N(0, \sigma^2)$. If using fixed $\tilde{\Delta}$, one set of $\tilde{\Delta}$ was generated from $N(\mu, \tau^2)$, and was used to generate the outcomes under the treatment by each cell. For cell l, the outcomes were generated using $N(\Delta_l, \sigma^2)$, $l = 1,2,\dots,L$.

3.3.1 Objective One

Powers were calculated in different settings, to better show how each parameter affects the power, selected results will be shown in figures. In each setting, the pair of (k, m) corresponding to the best power was used.

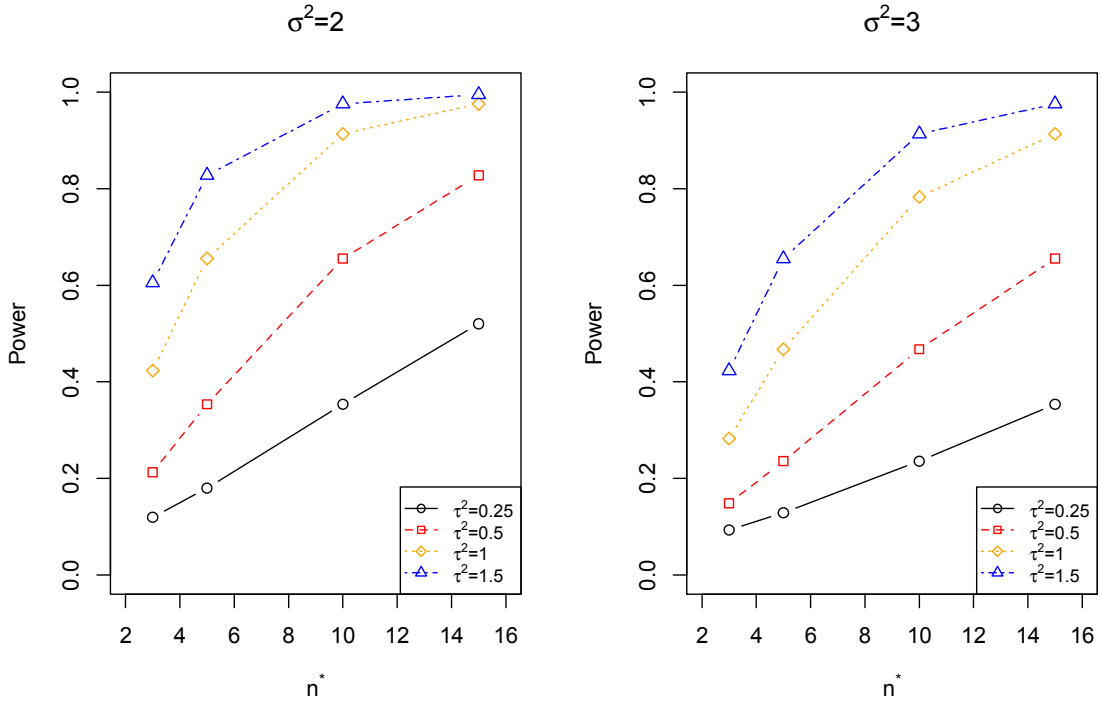
3.3.1.1 Power Simulation Using S^E

Figure 1 Power Plot of $L = 100$, $r = 1$, and Equal Cell Size $n^* = 3,5,10,15$, $\sigma^2 = 2, 3$, $\tau^2 = 0.25, 0.5, 1, 1.5$ using S^E



3.3.1.2 Power Simulation Using S^A

Figure 2 Power Plot of $L = 100$, $r = 1$, and Equal Cell Size $n^* = 3,5,10,15$, $\sigma^2 = 2, 3, \tau^2 = 0.25, 0.5, 1, 1.5$



3.3.2 Objective Two

3.3.2.1 Power Simulation Using S^E

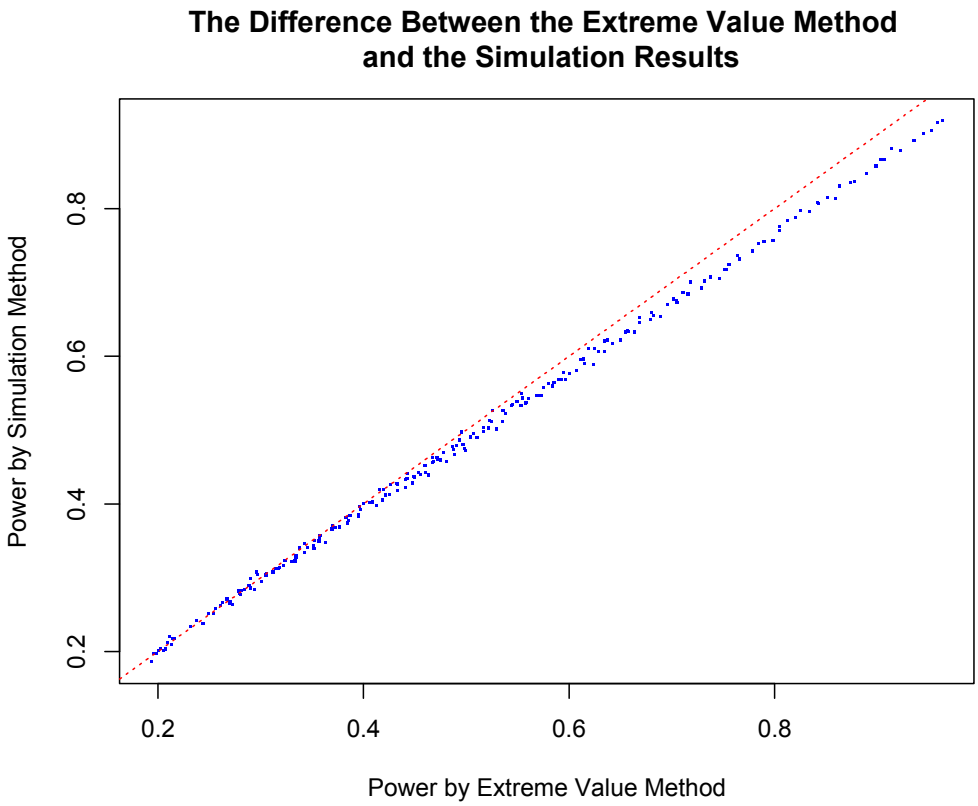
The type I errors were checked for different scenarios by simulations using S^E. The selected results are shown in Table 23. When k is large enough, the type I error is fairly stable around 0.05.

Table 23 Type I error results for $L = 100, r = 1, n^* = 10, \sigma^2 = 1$

		k				
		30	50	100	200	300
m	3	0.046	0.039	0.053	0.053	0.051
	5	0.064	0.05	0.047	0.037	0.057
	10	0.039	0.05	0.043	0.052	0.046
	15	0.045	0.041	0.048	0.041	0.054
	20	0.045	0.048	0.056	0.048	0.053
	30	0.038	0.056	0.036	0.055	0.042
	35	0.035	0.05	0.051	0.049	0.036
	40	0.045	0.052	0.036	0.046	0.048
	50	0.045	0.044	0.059	0.047	0.046
	60	0.06	0.055	0.047	0.037	0.055

The powers calculated by the extreme value method were compared with the simulations. The comparisons were shown in Figure 3. We can see, when the power is small (power<0.5), the two methods have similar results, however, when the power is getting larger, the extreme value method tends to over estimate the power.

Figure 3 Results comparison between the extreme value method and the simulation



3.3.2.2 Power Simulation Using S^A

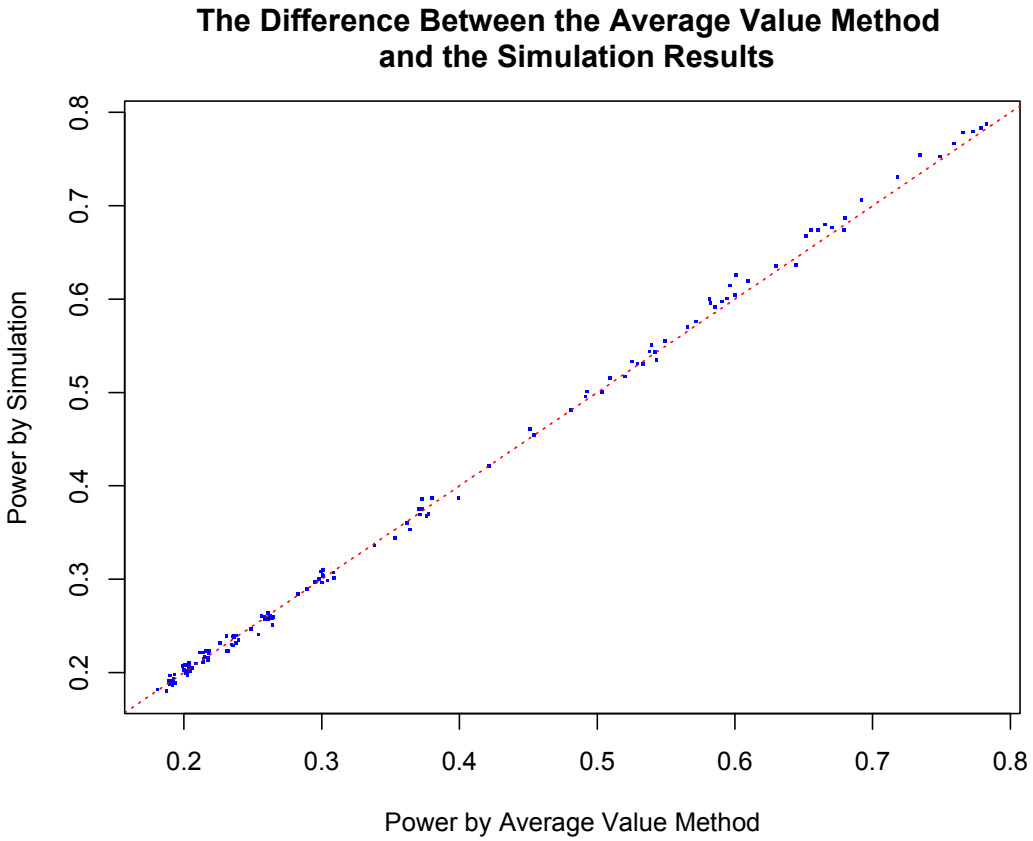
The type I errors were checked for different scenarios by simulations using S^A. The selected results are shown in Table 24. When k is large enough ($k > 300$), the type I error is fairly stable around 0.05.

Table 24 Type I error results for $L = 100, r = 1, n^* = 10, \sigma^2 = 1.5$

$\boldsymbol{m}$		$\boldsymbol{k = 50}$	$\boldsymbol{k = 100}$	$\boldsymbol{k = 150}$	$\boldsymbol{k = 200}$	$\boldsymbol{k = 250}$
	5	0.0295	0.0352	0.0373	0.0406	0.0384
	10	0.0355	0.0366	0.0417	0.0456	0.048
	15	0.0369	0.0435	0.0467	0.0493	0.0484
	20	0.0401	0.0444	0.0446	0.0451	0.048
	25	0.0428	0.0482	0.046	0.0477	0.0467
	30	0.0408	0.0462	0.0493	0.0493	0.049
	35	0.0419	0.0426	0.0484	0.0505	0.047
	40	0.0451	0.0444	0.0479	0.0469	0.0485
	45	0.0446	0.0461	0.049	0.047	0.0486
	50	0.0463	0.0446	0.0467	0.0472	0.0548
	55	0.0449	0.0494	0.0495	0.049	0.0501
	60	0.0487	0.0474	0.0454	0.0515	0.0503
		$\boldsymbol{k = 300}$	$\boldsymbol{k = 350}$	$\boldsymbol{k = 400}$	$\boldsymbol{k = 450}$	$\boldsymbol{k = 500}$
	5	0.0453	0.0463	0.0455	0.0471	0.0471
	10	0.0503	0.0457	0.0504	0.0453	0.0463
	15	0.0453	0.0485	0.0489	0.0468	0.0486
	20	0.0456	0.0491	0.0461	0.045	0.0486
	25	0.0479	0.0486	0.0498	0.0452	0.0521
	30	0.0489	0.045	0.0502	0.0487	0.0503
	35	0.0504	0.0493	0.0523	0.0501	0.0467
	40	0.0505	0.0505	0.0454	0.0479	0.0508
	45	0.0492	0.0483	0.049	0.0486	0.0513
	50	0.0516	0.0464	0.053	0.0512	0.0489
	55	0.0478	0.0488	0.0481	0.049	0.0488
	60	0.0513	0.0488	0.0475	0.0492	0.0498

The powers calculated by the average value method were compared with the simulations. The comparisons were shown in Figure 4. The figure shows the two methods have similar results.

Figure 4 Results comparison between the average value method and the simulation



3.4 Discussion

In this chapter, some theoretical methods were proposed to help determine the most appropriate study design, such as how many subjects in each cell to achieve certain desired power.

Normal approximation was used in the extreme value method, and hence induced some bias when reach high power. The central limit theorem was used in the average value method, and no obvious bias detected comparing to the simulation results.

For both methods, the larger the constant σ^2 within each cell (for both treatment and control), the lower the power of the tests; while the greater the sample size n^*in each cell, and the greater the variance (τ^2) among the true mean difference (Δ_l) for different cells, the higher the power of the tests. In other words, smaller σ^2, larger n^* and τ^2 make the tests easier to reject the null hypothesis that there is no treatment difference.

All the simulation results shown in this chapter were under the assumption that each cell has equal sample size, however, in the real circumstances, there could be two or more different sample sizes in different cells, sometimes the sample sizes may follow a specific distribution. The study design can be extended to accommodate these situations, which were mentioned in the method section.

CHAPTER 4. LOGISTIC-COX PROPORTIONAL HAZARDS MIXTURE MODEL

4.1 Background

Mixture models have been applied to different settings in survival analysis. Farewell[25, 26] used the combination of logistic regression with proportional hazards regression to distinguish the individuals who will eventually experience the event and the ones who will never experience the event in the model. Kuk et al.[27] proposed a semiparametric mixture model based on Farewell's parametric model. Ng and McLachlan[28] developed a semiparametic mixture model approach to analyze the competing-risks data. Peng and Dear[29] studied a general nonparametric mixture model to estimate the cure rate of the patients. In addition, Corbière et al. [30]proposed a penalized likelihood approach in the mixture cure model to allow flexible hazard function assumption and direct way to calculate the variance of parameters.

In this chapter, the mixture model was used to model the heterogeneity of the treatment effect. The treatment effect on any given patient can be conceptually defined, but estimation is impossible in most studies as we only observe one outcome (either under control or intervention). Let $A = 1$ represents the patient under the treatment, Y_1 be the corresponding outcome, and $A = 0$ represents the patient under the control, Y_0 be the corresponding outcome. Then the outcome Y can be described as $Y = AY_1 + (1 - A)Y_0$. One can only observe either Y_1 or Y_0, not both, so the treatment effect $\Delta = Y_1 - Y_0$ can not be directly calculated for each patient. Although we reply on average treatment effect, we can assume different groups to have different average treatment effect. In our proposed mixture model, we define there are two groups, one with the benefit averaged treatment effect, and the other one without. The probability of a patient in either group

can be calculated with the given information about each patient, and used as a weight in the proposed model. Different survival models can be constructed for the two groups separately to analyze the characteristic and treatment effect of each group.

In section 2, we introduce the mixture model, EM algorithm and other techniques used in the simulations, which are demonstrated in section 3. In section 4, the mixture model is applied to the MADITII data again, followed by discussion in section 5.

4.2 The mixture model

4.2.1 Cox proportional hazard (CoxPH) model

Given data $(t_i, \delta_i, \tilde{x}_i), i = 1, \dots, n$, where t_i is the event time when $\delta_i = 1$, and t_i is the censored time when $\delta_i = 0$, $\tilde{x}_i$ is the covariate vector. In the CoxPH model, for the ith individual, it is assumed that

$$\lambda(t; \tilde{x}_i) = \lambda_0(t)\exp(\tilde{x}_i^T \tilde{b} + Ab_A)$$

where $\lambda(t; \tilde{x}_i)$ is the hazard for individual i at time t, given covariate values $\tilde{x}_i$, and $\lambda_0(t)$ is the baseline hazard function of t. b_A is the coefficient of the treatment, when $b_A < 0$, the treatment reduces the hazard, and vice versa. If we order the event time as D^* distinct time: $t_1 < t_2 < \cdots < t_{D^*}$ with no tied events, then $\tilde{b}$ can be estimated by maximizing the partial likelihood

$$\prod_{j=1}^{D^*} \frac{\exp[\tilde{x}_{(j)}^T \tilde{b} + Ab_A]}{\sum_{k \in R(t_j)} \exp[\tilde{x}_k^T \tilde{b} + Ab_A]}$$

where $R(t_j)$ represents all individuals at risk at a time just prior to t_j, and $\tilde{x}_{(j)}$ is the covariate vector of the individual whose failure time is t_j.

4.2.2 Mixture Model and EM Algorithm

To set up the mixture model, we introduce a latent variable Z, the benefit indicator, i.e. $Z = 1$ indicates the individual is in the group that assumed to have benefit average treatment effect ($b_A < 0$), and $Z = 0$ means the individual is in the other group ($b_A \geq 0$). For simplicity, we call the individuals who have $Z = 1$ in the 'benefit group', and the rest who have $Z = 0$ in the 'not benefit group'. Z is assumed to have a Bernoulli distribution with probability p, which is modeled as a logistic model:

$$\log\left(\frac{p}{1-p}\right) = \widetilde{H}^T\tilde{\beta}$$

where $\widetilde{H}$ are covariates, such as baseline characteristic variables. $\widetilde{H}$ can be the same or different as $\tilde{X}$.

Our mixture model consists of two parts, the logistic regression part, and the survival part. The logistic regression part describes the probability of an individual in the 'benefit group', and the survival part is the corresponding survival probability. Let g_i be the survival likelihood function if individual i is in the 'benefit group', and h_i be the function if he/she is in the 'not benefit group'. If we have the complete set of data (X, T, Z), the likelihood for individual i can be represent as

$$L_i^c = c(x_i)p_i^{z_i}(1-p_i)^{1-z_i}g_i^{z_i}h_i^{(1-z_i)}$$

g_i and h_i are assumed in the form of $\left[\lambda_0(t)\exp\left(\tilde{x}^T\tilde{b} + Ab_A\right)\right]^{\delta}S_0(T)^{\exp\left(\tilde{x}^T\tilde{b}+Ab_A\right)}$ with the proportional hazard assumption. The only difference between the two functions is the assumption about the coefficients for treatment.

Since Z cannot be directly observed, the dataset (X, T, Z) is not complete; EM algorithm can be used to solve the problem. The observed likelihood for individual i can be written as

$$L_i^o = c(x_i)(\Pr(Z = 1)\, g_i + \Pr(Z = 0)\, h_i).$$

Let $\tilde{\theta}$ represents the unknown parameters in the model. To start the process of estimating $\tilde{\theta}$, assume initial values $\tilde{\theta}^{(0)}$. The process will be illustrated at the kth iteration, $k = 0,1, \ldots K$. $\tilde{\theta}^{(k)}$ is used to denote the current value of $\tilde{\theta}$, and $lnL_i^{o\,(k)}$ can be estimated by plugging in $\tilde{\theta}^{(k)}$. E step is performed to calculate:

$$E_{\tilde{\theta}^{(k)}}\left[lnL^c(\theta)\middle|\tilde{\theta}^{(k)}, X, T\right] = \sum_{i=1}^{n} u_i ln p_i + (1 - u_i) ln(1 - p_i) + u_i ln g_i + (1 - u_i) ln h_i$$

where $u_i = \Pr[Z_i = 1|\tilde{\theta}^{(k)}, X, T]$. Then $\tilde{\theta}^{(k+1)}$ is estimated and updated in the M step by solving the weighted logistic model and the weighted CoxPH model. At the end of the M step, calculate $lnL_i^{o\,(k+1)}$. The EM algorithm improves the observed log-likelihood, i.e. $lnL_i^{o\,(k+1)} - lnL_i^{o\,(k)}$ is always positive. Repeat the E step and the M step, until the Kth iteration where $lnL_i^{o\,(K+1)} - lnL_i^{o\,(K)}$ is sufficiently small.

4.2.3 Baseline Hazard $\lambda_0(t)$ Estimation

After $\widehat{\widehat{b}}$ and $\widehat{b_A}$ were estimated using partial likelihood, they can be plug into the likelihood function, and the likelihood function can be written in a function of the baseline hazard $\lambda_0(t)$:

$$L_\beta(\lambda_0(t)) \propto \prod_{i=1}^{D^*}\left[\lambda_0(t_i)\exp\left[-\lambda_0(t_i)\sum_{j\in R(t_i)} e_j\right]\right]$$

where $e_j \triangleq z_j \exp(\beta^{be} A_j) + (1-z_j)\exp(\beta^{nb} A_j)$.

$$LL_\beta(\lambda_0(t)) = \sum_{i=1}^{D}\left[\log(\lambda_0(t_i)) - \lambda_0(t_i)\sum_{j\in R(t_i)} e_j\right] \Rightarrow \frac{\partial}{\partial\lambda_0(t_i)} = \frac{1}{\lambda_0(t_i)} - \sum_{j\in R(t_i)} e_j = 0$$

$$\Rightarrow \hat{\lambda}_0(t) = \frac{1}{\sum_{j\in R(t_i)} e_j} = \left(\sum_{j\in R(t_i)} z_j \exp(\beta^{be} A_j) + (1-z_j)\exp(\beta^{nb} A_j)\right)^{-1}$$

The more detailed derivation can be found in the appendix.

4.2.4 Louis's Method

Louis's method is used to calculate the standard errors of the estimators estimated by the EM algorithm. At the last (Kth) iteration of the M step, $\widehat{\widehat{\theta}}$ is estimated to maximize $E_{\tilde{\theta}^{(k)}}\left[lnL^c(\theta)\middle|\tilde{\theta}^{(K)}, X, T\right]$. Define $S_C(\theta; X, T, Z)$ to be the score function of $\sum_{i=1}^{n} lnL_i^c$, then the information matrix $I\left(\widehat{\widehat{\theta}}; X, T\right)$ can be calculated by:

$$I\left(\widehat{\widehat{\theta}}; X, T\right) = I_C\left(\widehat{\widehat{\theta}}; X, T\right) - I_m\left(\widehat{\widehat{\theta}}; X, T\right)$$

where $I_C\left(\tilde{\theta}^{(K)}; X, T\right) = E_\theta\left\{-\frac{\partial^2}{\partial\theta\partial\theta^T} lnL^C(\theta)|X, T\right\}_{\theta=\widehat{\widehat{\theta}}}$ and $I_m\left(\tilde{\theta}^{(K)}; X, T\right) = [Cov_\theta\{S_C(\theta; X, T, Z)|X, T\}]_{\theta=\widehat{\widehat{\theta}}}$.

4.2.5 Calculate Treatment Benefit and Treatment Harm Rate

Treatment benefit rate (TBR) and treatment harm rate (THR) can be used to describe the proportion of the patients that benefit or harmed by the treatment as compared with the control respectively[31]. They are also important measurements to describe the heterogeneity within the given group of patients, although they describe things slightly different than Z. In the setting of survival outcomes, $Z = 1$ indicates that an individual survives longer under the treatment than under the control, regardless the actual survival time. Whereas in the calculation of TBR, 'benefit' means an individual survives beyond a fixed time threshold under treatment and dies before the threshold under control.

According to the paper[31], in the time-to-event outcome case, assume

$$S^0(t) = \Pr(T > t|X, A = 0),$$

$$S^1(t) = \Pr(T > t|X, A = 1),$$

then TBR and THR can be expressed as

$$TBR(t) = E[\big(1 - S^0(t)\big)S^1(t)],$$

$$THR(t) = E\big[S^0(t)\big(1 - S^1(t)\big)\big].$$

In our situation with the un-observed parameter Z, the expressions become

$$TBR(t) = E[Z\left(1 - S^0_{benefit}(t)\right)S^1_{benefit}(t) +$$

$$(1 - Z)\left(1 - S^0_{NOTbenefit}(t)\right)S^1_{NOTbenefit}(t)],$$

$$THR(t) = E[ZS^0_{benefit}(t)\left(1 - S^1_{benefit}(t)\right) +$$

$$(1 - Z)S^0_{NOTbenefit}(t)\left(1 - S^1_{NOTbenefit}(t)\right)].$$

4.3 Simulation

4.3.1 Objective

There are two objectives for the simulation. The first one is to examine the performance of the mixture model with simulated patient data. The second one is to compare the Louis's method estimated standard error with the true standard error.

4.3.2 Data and Setting

Data were simulated based on the MADIT II data, the same dataset as in Chapter2.

For objective one, n =1,000 and 10,000 patients were simulated with patients' age, sex, health index, treatment assignment. Each patient's age was generated from a normal distribution with mean 50 and standard deviation of 2. Probability of a male patient is 0.5, and the probability of being assigned to treatment arm is also 0.5. Health indexes were generated from the standard normal distribution. The probability of a patient benefiting from the treatment is determined by health index (H) – the logistic regression part. The logit of the probability for each patient to benefit is a linear function of his/her health index:

$$p_i^{true} = \exp(\beta_1 H) / (1 + \exp(\beta_1 H))$$

The benefit indicator (Z_i) for each patient was then generated using p_i^{true}.

Two different sets of coefficients were assumed for the 'benefit group' and the 'not benefit group'. Particularly, b_A was assumed to be a negative number for the 'benefit group' and a non-negative number for the 'not benefit group'. The time to event was then generated with the rate based on a patient's age, sex, and treatment assignment for each group respectively using exponential distribution:

$$rate = Z \times \left[\lambda_0(t)\exp\left(\tilde{x}^T\tilde{b}^{be} + Ab_A^{be}\right)\right] + (1-Z) \times \left[\lambda_0(t)\exp\left(\tilde{x}^T\tilde{b}^{nb} + Ab_A^{nb}\right)\right].$$

For objective two, n =1000 patients were simulated in a simple setting with only health index (H) and treatment assignment (A). The coefficients for treatment assignment b_A in the 'benefit group' and 'not benefit group' were assumed to be negative and non-negative respectively. Monte Carlo simulations for 200, 500, 1000, and 5000 times were completed to calculate both the bootstraping and Louis's method estimation of standard error of each parameter. The difference between the two sets of standard errors was then compared.

4.3.3 Selected Results for Objective One

Different coefficients were used to discover the characteristic of the mixture model. In this dissertation, two scenarios are presented. Both scenarios have the same beneficial treatment effect in the 'benefit group'. While first scenario has harmful treatment effect in the 'not benefit group', the second scenario has no treatment effect in that group. In addition, the second scenario added a intercept parameter into the model, The true value, the estimated value, and the standard error of each coefficient, are shown in the tables below for n =1,000 and n =10,000.

Table 25 First Scenario n =1,000

n=1,000	Health Index (β_1)	BG-age (b_1^{be})	BG-sex (b_2^{be})	BG-trt (b_A^{be})	NBG-age (b_1^{nb})	NBG-sex (b_2^{nb})	NBG-trt (b_A^{nb})
Mixture Model Results	1.232	1.038	-0.649	-0.991	0.275	-0.330	0.537
True Value	1	1	-0.8	-1	0.3	-0.3	0.5
Standard Error	0.182	0.081	0.213	0.211	0.033	0.141	0.142

Table 26 First Scenario n =10,000

n=10,000	Health Index (β_1)	BG-age (b_1^{be})	BG-sex (b_2^{be})	BG-trt (b_A^{be})	NBG-age (b_1^{nb})	NBG-sex (b_2^{nb})	NBG-trt (b_A^{nb})
Mixture Model Results	0.904	0.984	-0.808	-0.943	0.291	-0.309	0.468
True Value	1	1	-0.8	-1	0.3	-0.3	0.5
Standard Error	0.078	0.027	0.068	0.073	0.011	0.045	0.052

The comparison of the computed and the true baseline cumulative hazard over time are shown in the figures.

Figure 5 Baseline Cumulative Hazard over Time n =1,000

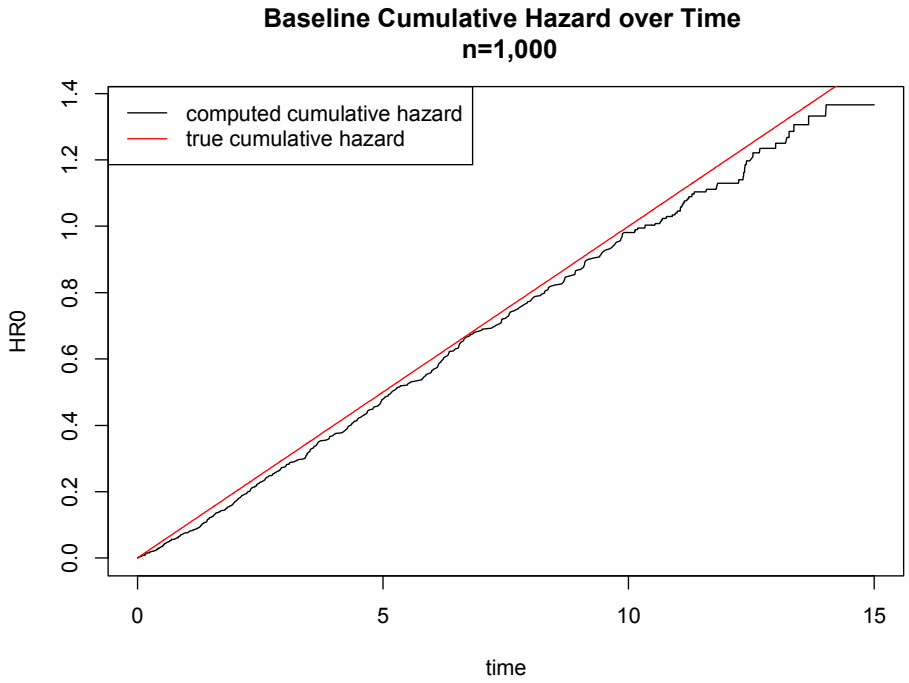


Figure 6 Baseline Cumulative Hazard over Time n =10,000

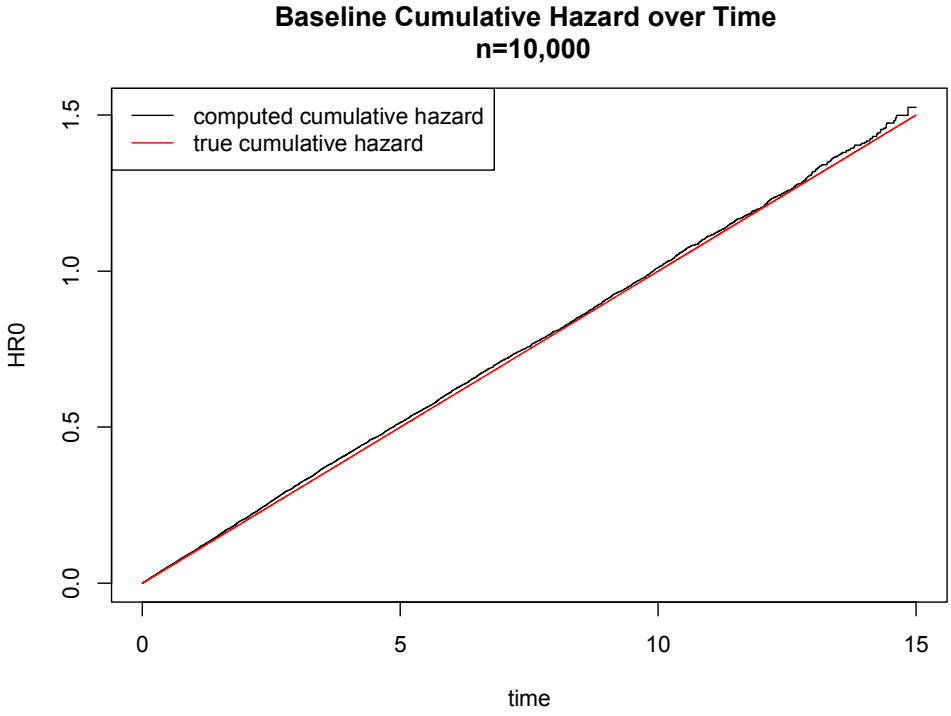


Table 27 Second Scenario n =1,000

n=1,000	Intercept	Health Index	BG-age	BG-sex	BG-trt	NBG-age	NBG-sex	NBG-trt
Mixture Model Results	-0.231	1.315	0.895	-0.831	-0.716	0.247	-0.268	4.941 E-06
True Value	-0.5	1	1	-0.8	-1	0.3	-0.3	0
Standard Error	0.476	0.329	0.167	0.285	0.198	0.045	0.128	0.077

Table 28 Second Scenario n =10,000

n=10,000	Intercept	Health Index	BG-age	BG-sex	BG-trt	NBG-age	NBG-sex	NBG-trt
Mixture Model Results	-0.559	0.955	0.986	-0.730	-0.991	0.300	-0.334	1.204 E-07
True Value	-0.5	1	1	-0.8	-1	0.3	-0.3	0
Standard Error	0.116	0.082	0.037	0.074	0.082	0.013	0.038	0.024

Table 29 Results for Objective Two n =1,000, B =200

n=1,000	B=200	seed=123	
	$\boldsymbol{\beta_1}$	$\boldsymbol{b_A^{be}}$	$\boldsymbol{b_A^{nb}}$
True SE	0.131	0.198	0.159
Louis's Method SE	0.121	0.212	0.145
Difference	0.010	-0.013	0.014

Table 30 Results for Objective Two n =1,000, B =500

n=1,000	B=500	seed=736	
	$\boldsymbol{\beta_1}$	$\boldsymbol{b_A^{be}}$	$\boldsymbol{b_A^{nb}}$
True SE	0.116	0.212	0.155
Louis's Method SE	0.121	0.211	0.145
Difference	-0.005	0.001	0.010

Table 31 Results for Objective Two n =1,000, B =1,000

n=1,000	B=1000	seed=3294	
	$\boldsymbol{\beta_1}$	$\boldsymbol{b_A^{be}}$	$\boldsymbol{b_A^{nb}}$
True SE	0.118	0.211	0.151
Louis's Method SE	0.121	0.212	0.145
Difference	-0.003	-0.001	0.006

Table 32 Results for Objective Two n =1,000, B =500

n=1,000	B=5000	seed=123	
	$\boldsymbol{\beta_1}$	$\boldsymbol{b_A^{be}}$	$\boldsymbol{b_A^{nb}}$
True SE	0.121	0.213	0.151
Louis's Method SE	0.121	0.213	0.145
Difference	0.000	0.001	0.006

4.4 Application to MADITII Data

The mixture model was applied to the MADITII data using, the five clinical characteristic variables in the logistic regression part, and the five clinical characteristic variables in the survival part. Two year TBR and THR were calculated and compared with the results in the paper of Shen et al[31].

Table 33 Results of TBR , THR, and the Mean Posterior Probability of Z_i

TBR	**THR**	**the mean of posterior probability of Zi**
0.176	0.098	0.962

From the results in the table, we can see that 17.6% of the patients who can survive beyond two years under treatment would die before two years under control (compare to 18% in Shen et al.[31]); while 9.8% patients would die within two years in the treatment arm but survive in the control arm (compare to 10% in Shen et al.[31]). The average probability of an individual survives longer under the treatment than under the control, regardless the actual survival time is 0.962.

4.5 Discussion

n=1000 seems sufficient to estimate the coefficients well. Clearly, when there are more patients, the mixture model gives better estimation.

As mentioned earlier, in the logistic part of the model, baseline characteristic variables ($\widetilde{H}$) are used to determine the model of Z. In some cases, there are maybe hundreds even thousands baseline variables for us to use. To select the important variables for the model, some variable selection method, such as the LASSO selection procedure, can be added in the M step when solving the weighted logistic regression.

In this dissertation, we simply used 0 as the threshold to separate the two groups, i.e. the 'benefit group' is assumed to have $b_A < 0$, and the 'not benefit group' is assumed to have $b_A \geq 0$. This means on average in the 'benefit group', the treatment reduces the hazard, and in the 'not benefit group', the treatment is the same as the control or even increase the hazard.

Other threshold values can be chosen depending on the nature of the treatment. Sometimes, we may want to choose a negative value instead of 0 as the threshold, for example, when both sub groups are beneficial from the treatment, with one group benefit more than the other.

We need to try to confirm the existence of two groups with (quantitatively or qualitatively) different treatment effect by either testing or by experience before using the mixture model. Since there is a constraint in the model, it forces b_As to separate by the threshold. Our simulation shows, if both b_As are actually on the same side, and the threshold assumption is wrong, then the estimation of the one $b_A$that assumed to be on the wrong side of the threshold, will be really close to the threshold, and the other coefficients for all the other parameters will be estimated poorly.

CHAPTER 5 CONCLUSIONS AND DISCUSSIONS

Heterogeneity of the treatment effect in the treated population is frequently encountered in medical research. The literature of testing is dominated by either improved subgroup analysis with fixed subgroups, or by post hoc subgroup search with certain model assumptions. In this dissertation, we proposed a non-parametric test adopting G&S's likelihood ratio test, and P&G's range test. We extended their tests with a built-in stochastic search. This extension allows us to detect signals that may not be linearly related to the multiple covariates. The one-sided and two-sided extreme value tests are based on the strongest positive and negative signals, while one-sided and two-sided average value tests are based on the average positive and negative signals. The choice for the pair of k and p in the stochastic search are essential to this test. We evaluated our methods through simulation study and applied this method to the MADIT II data. Nevertheless, active search is prone to selection bias and inflation of the probability of false positives. Our test can serve as a gatekeeping procedure to prevent false positives.

Using our proposed test, with some knowledge of the targeted patient population, we can calculate power under different settings to help guide future study designs. We focused on normal distributed outcomes in this chapter. It seems like the 'extreme value test' works better than the 'average value test' when the true treatment effects are from a normal distribution. Without surprise, changing from 3 patients in each cell, to 10 patients in each cell (in each treatment arm) can considerably increase the power of our tests. In addition, larger variance between the true treatment effects, and less variation of the treatment effect for patients within each cell, improve the power as well. Although

formulation maybe hard to derive in other outcomes, such as survival outcomes, one can certainly use numerical methods to calculate powers for study design. Faster computational methods may need to be developed to be able to handle large calculations.

Finally, we developed a mixture model to model the patients who benefit from the treatment and who do not, and to estimate the average treatment effect in different sub-populations. In our proposed mixture model, we defined two groups, with and without the benefit averaged treatment effect. We calculated the posterior probability of each patient in either group with the given clinical knowledge about each patient, and used it as a weight in the proposed model. Different survival models were constructed for the two groups separately to analyze the characteristic and treatment effect of each group. The mixture model can be easily extend to serve more than two sub-populations, however, one should bear in mind that more available patient samples may be required to estimate all parameters in a more complex model. We evaluated our methods through simulation study and applied this method to the MADIT II data. The use of this method allows us to discover the treatment effects when qualitative treatment interactions exist, to separate patients into different groups, and to estimate the average treatment effect in different sub-populations.

APPENDIX A DISTRIBUTION OF $\left(Z(\widetilde{W}^{(1)}),\dots,Z(\widetilde{W}^{(k)})\right)^T$ UNDER NULL AND ALTERNATIVE HYPOTHESES

For m randomly selected cells, $p = m/L$, $\widetilde{W} = (W_1,\dots,W_m)$, the Z statistic for testing the equality of the means is of the form

$$Z(\widetilde{W}) = \sqrt{\frac{n^*}{\left(1+\frac{1}{r}\right)m\sigma^2}}\sum_{j=1}^{m}\hat{\Delta}_{w_j} \triangleq a\sum_{j=1}^{m}\hat{\Delta}_{w_j}.$$

Under the sharp null that the means are the same between the treatment arms for all cells, $Z(\widetilde{W})|\widetilde{W}\sim N(0,1)$, and therefore $Z(\widetilde{W})\sim N(0,1)$. For two independent draws, $\widetilde{W}^{(1)}$ and $\widetilde{W}^{(2)}$, we have $Cov\left(Z(\widetilde{W}^{(1)}),Z(\widetilde{W}^{(2)})\right) = E\left(Z(\widetilde{W}^{(1)}),Z(\widetilde{W}^{(2)})\right) =$

$E\left[E\left(Z(\widetilde{W}^{(1)}),Z(\widetilde{W}^{(2)})\right)|\hat{\Delta}\right] = E\left[E\left(Z(\widetilde{W}^{(1)})|\hat{\Delta}\right)E\left(Z(\widetilde{W}^{(2)})|\hat{\Delta}\right)\right] =$

$E\left[E^2\left(Z(\widetilde{W}^{(2)})|\hat{\Delta}\right)\right] = a^2E\left[\left(p\sum_{i=1}^{L}\hat{\Delta}_i\right)^2\right] = \frac{a^2p^2L\left(1+\frac{1}{r}\right)\sigma^2}{n^*} = p$. (Note that

$E\left[\sum_{j=1}^{m}\hat{\Delta}_{w_j}|\hat{\Delta}\right] = p\sum_{i=1}^{L}\hat{\Delta}_i$). Since for any $Z(\widetilde{W}^{(s)})$ and $Z(\widetilde{W}^{(t)})$, $s\neq t$ and $s,t = 1,\dots,k$, $Cov\left(Z(\widetilde{W}^{(s)}),Z(\widetilde{W}^{(t)})\right)$ is the same value, the distribution of $Z(\widetilde{W}^{(1)}),Z(\widetilde{W}^{(2)}),\dots,Z(\widetilde{W}^{(k)})$ can be described as

$$\left(Z(\widetilde{W}^{(1)}),\dots,Z(\widetilde{W}^{(k)})\right)^T\sim N\left[0_{I_{k\times1}},\ \begin{pmatrix}1 & p & \dots & p\\ p & 1 & \dots & p\\ \vdots & \vdots & \ddots & \vdots\\ p & p & \dots & 1\end{pmatrix}\right].$$

Under the alternative hypothesis, $\Delta_l\sim N(\mu,\tau^2)$, $Z(\widetilde{W})|\widetilde{W}\sim N\left(a\sum_{j=1}^{m}\Delta_{W_j},1\right)$, and $\sum_{j=1}^{m}\Delta_{W_j}$ is approximately normal with mean $m\mu$ and $cm\tau^2$, where $c = (L-m)/(L-1)$ is the finite sample correction factor. Then, when Δ_ls are fixed, $Z(\widetilde{W})\sim N(am\mu,1+a^2cm\tau^2)$.

For any $Z\left(\widetilde{W}^{(s)}\right)$, and $Z(\widetilde{W}^{(t)})$, $s \neq t$ and $s, t = 1, \dots, k$,

$$E\left(Z\left(\widetilde{W}^{(s)}\right)Z(\widetilde{W}^{(t)})\right) = E\left[E^2\left(Z\left(\widetilde{W}\right)|\hat{\Delta}\right)\right] = a^2p^2E\left[\left(\sum_{i=1}^{L}\hat{\Delta}_i\right)^2\right]$$

$$= a^2p^2\left\{E^2\left(\sum_{i=1}^{L}\hat{\Delta}_i\right) + V\left(\sum_{i=1}^{L}\hat{\Delta}_i\right)\right\}$$

$$= a^2p^2L^2\mu^2 + a^2p^2\left[L\frac{(1+1/r)\sigma^2}{n^*}\right] = a^2m^2\mu^2 + p$$

$$Cov\left(Z\left(\widetilde{W}^{(s)}\right), Z(\widetilde{W}^{(t)})\right) = E\left(Z\left(\widetilde{W}^{(s)}\right)Z(\widetilde{W}^{(t)})\right) - E^2\left[Z\left(\widetilde{W}\right)\right]$$

$$= a^2m^2\mu^2 + p - a^2m^2\mu^2 = p$$

Then, $\left(Z(\widetilde{W}^{(1)}), \dots, Z(\widetilde{W}^{(k)})\right)^T$

$$\sim N\left[am\mu_{I_{k\times 1}},\ \begin{pmatrix} 1+a^2cm\tau^2 & p & \dots & p \\ p & 1+a^2cm\tau^2 & \dots & p \\ \vdots & \vdots & \ddots & \vdots \\ p & p & \dots & 1+a^2cm\tau^2 \end{pmatrix}\right]$$

Then, when Δ_is are random, $Z\left(\widetilde{W}\right) \sim N(am\mu, 1 + a^2m\tau^2)$. For any $Z\left(\widetilde{W}^{(s)}\right)$, and $Z(\widetilde{W}^{(t)})$, $s \neq t$ and $s, t = 1, \dots, k$,

$$E\left(Z\left(\widetilde{W}^{(s)}\right)Z(\widetilde{W}^{(t)})\right) = a^2p^2\left\{E^2\left(\sum_{i=1}^{L}\hat{\Delta}_i\right) + V\left(\sum_{i=1}^{L}\hat{\Delta}_i\right)\right\}$$

$$= a^2p^2L^2\mu^2 + a^2p^2\left[L\frac{(1+1/r)\sigma^2}{n^*} + L\tau^2\right] = a^2m^2\mu^2 + p + a^2p^2L\tau^2$$

$$Cov\left(Z(\widetilde{W}^{(s)}), Z(\widetilde{W}^{(t)})\right) = E\left(Z(\widetilde{W}^{(s)})Z(\widetilde{W}^{(t)})\right) - E^2[Z(\widetilde{W})]$$

$$= a^2m^2\mu^2 + p + a^2p^2L\tau^2 - a^2m^2\mu^2 = p + a^2p^2L\tau^2 \triangleq p^*$$

Then, $\left(Z(\widetilde{W}^{(1)}), \dots, Z(\widetilde{W}^{(k)})\right)^T$

$$\sim N\left[am\mu_{I_{k\times 1}}, \begin{pmatrix} 1+a^2m\tau^2 & p^* & \dots & p^* \\ p^* & 1+a^2m\tau^2 & \dots & p^* \\ \vdots & \vdots & \ddots & \vdots \\ p^* & p^* & \dots & 1+a^2m\tau^2 \end{pmatrix}\right].$$

APPENDIX B FORMULATION OF POWER CALCULATION FOR UNEQUAL PATIENT SAMPLE SIZE CASE (TWO VALUES)

Assume $n^* = \begin{cases} n_1 & p_1 \\ n_2 & p_2 \end{cases}$,

For k randomly selected cells, $W = (W_1, W_2, \dots, W_k)$

$$Z(W) = \frac{\sum_{j=1}^{k} V_{w_j} \hat{\Delta}_{w_j}}{Var\left(\sum_{j=1}^{k} V_{w_j} \hat{\Delta}_{w_j}\right)} \text{ where } V_{w_j} = \frac{n^*_{w_j}}{\sum_{j=1}^{k} n^*_{w_j}}$$

$$Var\left(\sum_{j=1}^{k} V_{w_j} \hat{\Delta}_{w_j}\right) = \sum_{j=1}^{k} V_{w_j}^2 Var\left(\hat{\Delta}_{w_j}\right) = \sum_{j=1}^{k} V_{w_j}^2 \frac{\sigma^2}{n^*_{w_j}}\left(1 + \frac{1}{r}\right) = \frac{\sum_{j=1}^{k} \frac{\boxed{n^*_{w_j}}^2 \sigma^2}{\boxed{n^*_{w_j}}}\left(1+\frac{1}{r}\right)}{\left(\sum_{j=1}^{k} n^*_{w_j}\right)^2} =$$

$$\frac{\sigma^2\left(1+\frac{1}{r}\right)\boxed{\sum_{j=1}^{k} n^*_{w_j}}}{\boxed{\left(\sum_{j=1}^{k} n^*_{w_j}\right)}^2} = \frac{\sigma^2\left(1+\frac{1}{r}\right)}{\sum_{j=1}^{k} n^*_{w_j}}$$

$$Z(W) = \frac{\sum_{j=1}^{k} \frac{n^*_{w_j}}{\boxed{\sum_{j=1}^{k} n^*_{w_j}}} \hat{\Delta}_{w_j}}{\sqrt{\frac{\sigma^2\left(1+\frac{1}{r}\right)}{\boxed{\sum_{j=1}^{k} n^*_{w_j}}}}} = \frac{\sum_{j=1}^{k} n^*_{w_j} \hat{\Delta}_{w_j}}{\sqrt{\sum_{j=1}^{k} n^*_{w_j} \sigma^2\left(1+\frac{1}{r}\right)}}$$

Let $b = 1 \Big/ \sqrt{\sigma^2\left(1 + \frac{1}{r}\right)}$

Under the sharp null that the means are the same between the treatment arms and the control arms for all cells, $Z(W)|W \sim N(0,1)$, the therefore $Z(W) \sim N(0,1)$.

$$E\left(Z(W)|\hat{\Delta}\right) = bE\left[\frac{\sum_{j=1}^{k} n^*_{w_j} \hat{\Delta}_{w_j}}{\sqrt{\sum_{j=1}^{k} n^*_{w_j}}} |\hat{\Delta}\right]$$

$$= b\sum_{j=1}^{k}\left[E\left[\frac{n^*_{w_j}}{\sqrt{\sum_{j=1}^{k} n^*_{w_j}}}|\hat{\Delta}\right]E\left[\hat{\Delta}_{w_j}|\hat{\Delta}\right]\right] \quad \text{Assume } \hat{\Delta} \perp n^*$$

$$= b\sum_{i=1}^{L} V_i\hat{\Delta}_i\, E[\frac{\sum_{j=1}^{k} n^*_{w_j}}{\sqrt{\sum_{j=1}^{k} n^*_{w_j}}}|\hat{\Delta}]$$

$$= b\sum_{i=1}^{L} V_i\hat{\Delta}_i\, E[\sqrt{\sum_{j=1}^{k} n^*_{w_j}}]$$

Let $c = E[\sqrt{\sum_{j=1}^{k} n^*_{w_j}}] = \sum_{a_1=0}^{k} \sqrt{a_1 n_1 + (k-a_1)n_2}\binom{k}{a_1} p_1^{a_1}\, p_2^{(k-a_1)}$

$$Cov\left[Z\left(W^{(1)}\right), Z\left(W^{(2)}\right)\right] = E\left[Z\left(W^{(1)}\right)Z\left(W^{(2)}\right)\right] = E\left[E^2[Z(W)]|\hat{\Delta}\right]$$

$$= b^2c^2E\left[\left(\sum_{i=1}^{L} V_i\hat{\Delta}_i\right)^2\right]$$

$$E\left[\left(\sum_{i=1}^{L} V_i\hat{\Delta}_i\right)^2\right] = Var\left[\sum_{i=1}^{L} V_i\hat{\Delta}_i\right] = \sum_{i=1}^{L} (V_i)^2\, Var\left[\hat{\Delta}_i\right]$$

$$= \left(1+\frac{1}{r}\right)\sigma^2\sum_{i=1}^{L}\left(\frac{1}{\boxed{n^*i}}\right)\cdot\left(\frac{\boxed{n_i^*}}{\sum_{i=1}^{L} n_i^*}\right)^2 = \left(1+\frac{1}{r}\right)\sigma^2\frac{\sum_{i=1}^{L} n_i^*}{\left(\sum_{i=1}^{L} n_i^*\right)^2} = \left(1+\frac{1}{r}\right)\sigma^2\frac{1}{\sum_{i=1}^{L} n_i^*}$$

$$\Rightarrow\ Cov\left[Z\left(W^{(1)}\right), Z\left(W^{(2)}\right)\right] = \boxed{b^2}c^2\boxed{\left(1+\frac{1}{r}\right)\sigma^2}\frac{1}{\sum_{i=1}^{L} n_i^*} = c^2\frac{1}{\sum_{i=1}^{L} n_i^*} \triangleq \rho_1^*$$

$\Rightarrow$ Under Null: $\left[Z\left(W^{(1)}\right), Z\left(W^{(2)}\right)\right]^T \sim N\left[\binom{0}{0}, \begin{pmatrix} 1 & \rho_1^* \\ \rho_1^* & 1 \end{pmatrix}\right]$

Under alternative distribution:

$Z(W)|W \sim N(b\sum_{j=1}^{k} V_{w_j}\Delta_{w_j}, 1)$, and $\Delta_{w_j} \sim approx \sim N(\mu, c\tau^2), c = \frac{L-k}{L-1}$.

$$E[Z(W)] = E\left[E[Z(W)|W]\right] = E\left[b\frac{\sum_{j=1}^{k} n^*_{w_j}\Delta_{w_j}}{\sqrt{\sum_{j=1}^{k} n^*_{w_j}}}\right] = b\mu E\left[\frac{\sum_{j=1}^{k} n^*_{w_j}}{\sqrt{\sum_{j=1}^{k} n^*_{w_j}}}\right] \text{ Assume } \hat{\Delta} \perp n^*$$

$$= b\mu E\left[\sqrt{\sum_{j=1}^{k} n^*_{w_j}}\right] = bc\mu \text{ where } c = E\left[\sqrt{\sum_{j=1}^{k} n^*_{w_j}}\right]$$

$$Var[Z(W)] = E\left[Var[Z(W)|W]\right] + Var\left[E[Z(W)|W]\right] = 1 + b^2Var\left[\frac{\sum_{j=1}^{k} n^*_{w_j}\Delta_{w_j}}{\sqrt{\sum_{j=1}^{k} n^*_{w_j}}}\right].$$

$$Var\left[\frac{\sum_{j=1}^{k} n_{w_j}^{*}\Delta_{w_j}}{\sqrt{\sum_{j=1}^{k} n_{w_j}^{*}}}\right] = \sum_{j=1}^{k}\left[Var\left[\frac{n_{w_j}^{*}\Delta_{w_j}}{\sqrt{\sum_{j=1}^{k} n_{w_j}^{*}}}\right]\right] = \sum_{j=1}^{k}\left\{E\left[\left(\frac{n_{w_j}^{*}\Delta_{w_j}}{\sqrt{\sum_{j=1}^{k} n_{w_j}^{*}}}\right)^{2}\right] - E^{2}\left[\frac{n_{w_j}^{*}\Delta_{w_j}}{\sqrt{\sum_{j=1}^{k} n_{w_j}^{*}}}\right]\right\}$$

$$= \sum_{j=1}^{k}\left\{E\left[\frac{{n_{w_j}^{*}}^{2}}{\sum_{j=1}^{k} n_{w_j}^{*}}\right]\cdot E\left[\Delta_{w_j}^{2}\right] - E^{2}\left[\frac{n_{w_j}^{*}}{\sqrt{\sum_{j=1}^{k} n_{w_j}^{*}}}\right]\cdot E^{2}\left[\Delta_{w_j}\right]\right\}$$

$$= \sum_{j=1}^{k}\left\{E\left[\frac{{n_{w_j}^{*}}^{2}}{\sum_{j=1}^{k} n_{w_j}^{*}}\right]\cdot(\mu^{2}+c_0\tau^{2}) - \left(E\left[\frac{{n_{w_j}^{*}}^{2}}{\sum_{j=1}^{k} n_{w_j}^{*}}\right] - Var\left[\frac{n_{w_j}^{*}}{\sqrt{\sum_{j=1}^{k} n_{w_j}^{*}}}\right]\right)\cdot\mu^{2}\right\}$$

$$= \sum_{j=1}^{k}\left\{\boxed{E\left[\frac{{n_{w_j}^{*}}^{2}}{\sum_{j=1}^{k} n_{w_j}^{*}}\right]\cdot\mu^{2}} + E\left[\frac{{n_{w_j}^{*}}^{2}}{\sum_{j=1}^{k} n_{w_j}^{*}}\right]c_0\tau^{2} - \boxed{E\left[\frac{{n_{w_j}^{*}}^{2}}{\sum_{j=1}^{k} n_{w_j}^{*}}\right]\mu^{2}} + Var\left[\frac{n_{w_j}^{*}}{\sqrt{\sum_{j=1}^{k} n_{w_j}^{*}}}\right]\cdot\mu^{2}\right\}$$

$$= \sum_{j=1}^{k}\left\{E\left[\frac{{n_{w_j}^{*}}^{2}}{\sum_{j=1}^{k} n_{w_j}^{*}}\right]c_0\tau^{2} + Var\left[\frac{n_{w_j}^{*}}{\sqrt{\sum_{j=1}^{k} n_{w_j}^{*}}}\right]\cdot\mu^{2}\right\}$$

$$= c_0\tau^{2}E\left[\frac{\sum_{j=1}^{k} {n_{w_j}^{*}}^{2}}{\sum_{j=1}^{k} n_{w_j}^{*}}\right] + \mu^{2}\left(Var\left[\frac{\sum_{j=1}^{k} n_{w_j}^{*}}{\sqrt{\sum_{j=1}^{k} n_{w_j}^{*}}}\right]\right) = c_0\tau^{2}E\left[\frac{\sum_{j=1}^{k} {n_{w_j}^{*}}^{2}}{\sum_{j=1}^{k} n_{w_j}^{*}}\right] +$$

$$\mu^{2}\left(Var\left[\sqrt{\sum_{j=1}^{k} n_{w_j}^{*}}\right]\right)$$

$$= c_0\tau^{2}E\left[\frac{\sum_{j=1}^{k} {n_{w_j}^{*}}^{2}}{\sum_{j=1}^{k} n_{w_j}^{*}}\right] + \mu^{2}\left(E\left[\sum_{j=1}^{k} n_{w_j}^{*}\right] - E^{2}\left[\sqrt{\sum_{j=1}^{k} n_{w_j}^{*}}\right]\right)$$

$$= c_0\tau^{2}E\left[\frac{\sum_{j=1}^{k} {n_{w_j}^{*}}^{2}}{\sum_{j=1}^{k} n_{w_j}^{*}}\right] + \mu^{2}(k(p_1n_1+p_2n_2) - c^{2})$$

$$d = E\left[\frac{\sum_{j=1}^{k} {n_{w_j}^*}^2}{\sum_{j=1}^{k} n_{w_j}^*}\right] = \sum_{a_1=0}^{k}\left\{\frac{[a_1 n_1^2 + (k-a_1)n_2^2]}{a_1 n_1 + (k-a_1)n_2}\binom{k}{a_1} p_1^{a_1} p_2^{(k-a_1)}\right\}$$

$$Var[Z(W)] = 1 + b^2[c_0\tau^2 d + \mu^2(k(p_1 n_1 + p_2 n_2) - c^2)].$$

$$E\left[Z\left(W^{(1)}\right)Z\left(W^{(2)}\right)\right] = E\left[E^2[Z(W)]|\hat{\Delta}\right] = b^2c^2E\left[\left(\sum_{i=1}^{L} V_i\hat{\Delta}_i\right)^2\right]$$

$$E\left[\left(\sum_{i=1}^{L} V_i\hat{\Delta}_i\right)^2\right] = Var\left[\sum_{i=1}^{L} V_i\hat{\Delta}_i\right] + E^2\left[\sum_{i=1}^{L} V_i\hat{\Delta}_i\right] = \sum_{i=1}^{L} V_i^2 Var[\hat{\Delta}_i] +$$

$$\left(\sum_{i=1}^{L} V_i E\left[\hat{\Delta}_i\right]\right)^2$$

$$= \left(1 + \frac{1}{r}\right)\sigma^2 \sum_{i=1}^{L}\left(\frac{1}{\boxed{n^*_i}}\right)\cdot\left(\frac{\boxed{n_i^*}}{\sum_{i=1}^{L} n_i^*}\right)^2 + \left(\sum_{i=1}^{L} V_i\Delta_i\right)^2$$

$$= \left(1 + \frac{1}{r}\right)\sigma^2 \frac{1}{\sum_{i=1}^{L} n_i^*} + \mu^2$$

$$\Rightarrow E\left[Z\left(W^{(1)}\right)Z\left(W^{(2)}\right)\right] = \boxed{b^2}c^2\left[\boxed{\left(1 + \frac{1}{r}\right)\sigma^2}\frac{1}{\sum_{i=1}^{L} n_i^*} + \mu^2\right] = \frac{c^2}{\sum_{i=1}^{L} n_i^*} + b^2c^2\mu^2$$

$$E\left[Z\left(W^{(1)}\right)\right]E\left[Z\left(W^{(2)}\right)\right] = E^2[Z(W)] = b^2c^2\mu^2$$

$$\Rightarrow\ Cov\left[Z\left(W^{(1)}\right), Z\left(W^{(2)}\right)\right] = E\left[Z\left(W^{(1)}\right)Z\left(W^{(2)}\right)\right] - E\left[Z\left(W^{(1)}\right)\right]E\left[Z\left(W^{(2)}\right)\right]$$

$$= \frac{c^2}{\sum_{i=1}^{L} n_i^*} + b^2c^2\mu^2 - b^2c^2\mu^2 = \rho_1^*$$

APPENDIX C FORMULATION OF POWER CALCULATION FOR UNEQUAL PATIENT SAMPLE SIZE CASE (POISSON DISTRIBUTION)

L cells, within each cell there are n^* control units and rn^* treatment units.

Assume $x \sim poisson(\lambda), n^* = x + 1$ shifted Possion (or truncated)

Δ_i : true mean difference for cell $i, i = 1,2, \dots, L$, and $\hat{\Delta}_i$: the difference of sample means.

σ^2: constant variance within each cell.

For k randomly selected cells, $W = (W_1, W_2, \dots, W_k)$

$$Z(W) = \frac{\sum_{j=1}^{k} V_{w_j} \hat{\Delta}_{w_j}}{Var\left(\sum_{j=1}^{k} V_{w_j} \hat{\Delta}_{w_j}\right)} \text{ where } V_{w_j} = \frac{n^*_{w_j}}{\sum_{j=1}^{k} n^*_{w_j}}$$

$$Var\left(\sum_{j=1}^{k} V_{w_j} \hat{\Delta}_{w_j}\right) = \sum_{j=1}^{k} V_{w_j}^2 Var\left(\hat{\Delta}_{w_j}\right) = \sum_{j=1}^{k} V_{w_j}^2 \frac{\sigma^2}{n^*_{w_j}}\left(1 + \frac{1}{r}\right) = \frac{\sum_{j=1}^{k} \boxed{n^*_{w_j}}^2 \sigma^2 \left(1+\frac{1}{r}\right)}{\boxed{n^*_{w_j}}} = \frac{\sigma^2\left(1+\frac{1}{r}\right)\boxed{\sum_{j=1}^{k} n^*_{w_j}}}{\boxed{\left(\sum_{j=1}^{k} n^*_{w_j}\right)}^2} = \frac{\sigma^2\left(1+\frac{1}{r}\right)}{\sum_{j=1}^{k} n^*_{w_j}}$$

$$Z(W) = \frac{\sum_{j=1}^{k} \frac{n^*_{w_j}}{\boxed{\sum_{j=1}^{k} n^*_{w_j}}} \hat{\Delta}_{w_j}}{\sqrt{\frac{\sigma^2\left(1+\frac{1}{r}\right)}{\boxed{\sum_{j=1}^{k} n^*_{w_j}}}}} = \frac{\sum_{j=1}^{k} n^*_{w_j} \hat{\Delta}_{w_j}}{\sqrt{\sum_{j=1}^{k} n^*_{w_j} \sigma^2\left(1+\frac{1}{r}\right)}}$$

Let $b = 1 \Big/ \sqrt{\sigma^2\left(1 + \frac{1}{r}\right)}$

Under the sharp null that the means are the same between the treatment arms and the control arms for all cells, $Z(W)|W \sim N(0,1)$, the therefore $Z(W) \sim N(0,1)$.

$$E\left(Z(W)|\hat{\Delta}\right) = bE\left[\frac{\sum_{j=1}^{k} n^*_{w_j} \hat{\Delta}_{w_j}}{\sqrt{\sum_{j=1}^{k} n^*_{w_j}}} |\hat{\Delta}\right]$$

$$= b\sum_{j=1}^{k}\left[E[\frac{n_{w_j}^*}{\sqrt{\sum_{j=1}^{k} n_{w_j}^*}}|\hat{\Delta}]E\left[\hat{\Delta}_{w_j}|\hat{\Delta}\right]\right] \quad \text{Assume } \hat{\Delta}\perp n^*$$

$$= b\sum_{i=1}^{L} V_i\hat{\Delta}_i\, E[\frac{\sum_{j=1}^{k} n_{w_j}^*}{\sqrt{\sum_{j=1}^{k} n_{w_j}^*}}|\hat{\Delta}]$$

$$= b\sum_{i=1}^{L} V_i\hat{\Delta}_i\, E[\sqrt{\sum_{j=1}^{k} n_{w_j}^*}]$$

Since $x_{w_j}\sim Poisson(\lambda)$, then $y \triangleq \sum_{j=1}^{k} x_{w_j} \sim Poisson(k\lambda)$

$c \triangleq E\left[\sqrt{\sum_{j=1}^{k} n_{w_j}^*}\right] = E\left[\sqrt{y+k}\right] = \sum_{y=0}^{\infty}\sqrt{y+k}\,\frac{\lambda^y}{y!}e^{-\lambda}$ (Can be calculated by Monte Carlo Method)

$$Cov\left[Z\left(W^{(1)}\right), Z\left(W^{(2)}\right)\right] = E\left[Z\left(W^{(1)}\right)Z\left(W^{(2)}\right)\right] = E\left[E^2[Z(W)]|\hat{\Delta}\right]$$

$$= b^2c^2E\left[\left(\sum_{i=1}^{L} V_i\hat{\Delta}_i\right)^2\right]$$

$$E\left[\left(\sum_{i=1}^{L} V_i\hat{\Delta}_i\right)^2\right] = Var\left[\sum_{i=1}^{L} V_i\hat{\Delta}_i\right] = \sum_{i=1}^{L}(V_i)^2\, Var\left[\hat{\Delta}_i\right]$$

$$= \left(1+\frac{1}{r}\right)\sigma^2\sum_{i=1}^{L}\left(\frac{1}{\boxed{n^*i}}\right)\cdot\left(\frac{\boxed{n_i^*}}{\sum_{i=1}^{L} n_i^*}\right)^2 = \left(1+\frac{1}{r}\right)\sigma^2\frac{\sum_{i=1}^{L} n_i^*}{\left(\sum_{i=1}^{L} n_i^*\right)^2} = \left(1+\frac{1}{r}\right)\sigma^2\frac{1}{\sum_{i=1}^{L} n_i^*}$$

$$\Rightarrow\ Cov\left[Z\left(W^{(1)}\right), Z\left(W^{(2)}\right)\right] = \boxed{b^2}c^2\boxed{\left(1+\frac{1}{r}\right)\sigma^2}\frac{1}{\sum_{i=1}^{L} n_i^*} = c^2\frac{1}{\sum_{i=1}^{L} n_i^*} \triangleq \rho_1^*$$

$\Rightarrow$ Under Null: $\left[Z\left(W^{(1)}\right), Z\left(W^{(2)}\right)\right]^T \sim N\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho_1^*\\ \rho_1^* & 1\end{pmatrix}\right]$

$Z(W)|W\sim N(b\sum_{j=1}^{k} V_{w_j}\Delta_{w_j}, 1)$, and $\Delta_{w_j}\sim approx\sim N(\mu, c_0\tau^2), c_0 = \frac{L-k}{L-1}$.

$$E[Z(W)] = E\left[E[Z(W)|W]\right] = E\left[b\frac{\sum_{j=1}^{k} n_{w_j}^*\Delta_{w_j}}{\sqrt{\sum_{j=1}^{k} n_{w_j}^*}}\right] = b\mu E\left[\frac{\sum_{j=1}^{k} n_{w_j}^*}{\sqrt{\sum_{j=1}^{k} n_{w_j}^*}}\right] \text{ Assume } \hat{\Delta}\perp n^*$$

$$= b\mu E\left[\sqrt{\sum_{j=1}^{k} n_{w_j}^*}\right] = bc\mu \text{ where } c = E\left[\sqrt{\sum_{j=1}^{k} n_{w_j}^*}\right]$$

$$Var[Z(W)] = E\big[Var[Z(W)|W]\big] + Var\big[E[Z(W)|W]\big] = 1 + b^2 Var\left[\frac{\sum_{j=1}^{k} n^*_{w_j}\Delta_{w_j}}{\sqrt{\sum_{j=1}^{k} n^*_{w_j}}}\right].$$

$$Var\left[\frac{\sum_{j=1}^{k} n^*_{w_j}\Delta_{w_j}}{\sqrt{\sum_{j=1}^{k} n^*_{w_j}}}\right] = \sum_{j=1}^{k}\left[Var\left[\frac{n^*_{w_j}\Delta_{w_j}}{\sqrt{\sum_{j=1}^{k} n^*_{w_j}}}\right]\right] = \sum_{j=1}^{k}\left\{E\left[\left(\frac{n^*_{w_j}\Delta_{w_j}}{\sqrt{\sum_{j=1}^{k} n^*_{w_j}}}\right)^2\right] - E^2\left[\frac{n^*_{w_j}\Delta_{w_j}}{\sqrt{\sum_{j=1}^{k} n^*_{w_j}}}\right]\right\}$$

$$= \sum_{j=1}^{k}\left\{E\left[\frac{{n^*_{w_j}}^2}{\sum_{j=1}^{k} n^*_{w_j}}\right]\cdot E\left[\Delta^2_{w_j}\right] - E^2\left[\frac{n^*_{w_j}}{\sqrt{\sum_{j=1}^{k} n^*_{w_j}}}\right]\cdot E^2\left[\Delta_{w_j}\right]\right\}$$

$$= \sum_{j=1}^{k}\left\{E\left[\frac{{n^*_{w_j}}^2}{\sum_{j=1}^{k} n^*_{w_j}}\right]\cdot(\mu^2 + c_0\tau^2) - \left(E\left[\frac{{n^*_{w_j}}^2}{\sum_{j=1}^{k} n^*_{w_j}}\right] - Var\left[\frac{n^*_{w_j}}{\sqrt{\sum_{j=1}^{k} n^*_{w_j}}}\right]\right)\cdot\mu^2\right\}$$

$$= \sum_{j=1}^{k}\left\{\boxed{E\left[\frac{{n^*_{w_j}}^2}{\sum_{j=1}^{k} n^*_{w_j}}\right]\cdot\mu^2} + E\left[\frac{{n^*_{w_j}}^2}{\sum_{j=1}^{k} n^*_{w_j}}\right]c_0\tau^2 - \boxed{E\left[\frac{{n^*_{w_j}}^2}{\sum_{j=1}^{k} n^*_{w_j}}\right]\mu^2} + Var\left[\frac{n^*_{w_j}}{\sqrt{\sum_{j=1}^{k} n^*_{w_j}}}\right]\cdot\mu^2\right\}$$

$$= \sum_{j=1}^{k}\left\{E\left[\frac{{n^*_{w_j}}^2}{\sum_{j=1}^{k} n^*_{w_j}}\right]c_0\tau^2 + Var\left[\frac{n^*_{w_j}}{\sqrt{\sum_{j=1}^{k} n^*_{w_j}}}\right]\cdot\mu^2\right\}$$

$$= c_0\tau^2 E\left[\frac{\sum_{j=1}^{k}{n^*_{w_j}}^2}{\sum_{j=1}^{k} n^*_{w_j}}\right] + \mu^2\left(Var\left[\frac{\sum_{j=1}^{k} n^*_{w_j}}{\sqrt{\sum_{j=1}^{k} n^*_{w_j}}}\right]\right) = c_0\tau^2 E\left[\frac{\sum_{j=1}^{k}{n^*_{w_j}}^2}{\sum_{j=1}^{k} n^*_{w_j}}\right] +$$

$$\mu^2\left(Var\left[\sqrt{\sum_{j=1}^{k} n^*_{w_j}}\right]\right)$$

$$= c_0\tau^2 E\left[\frac{\sum_{j=1}^{k}{n^*_{w_j}}^2}{\sum_{j=1}^{k} n^*_{w_j}}\right] + \mu^2\left(E\left[\sum_{j=1}^{k} n^*_{w_j}\right] - E^2\left[\sqrt{\sum_{j=1}^{k} n^*_{w_j}}\right]\right)$$

$$= c_0\tau^2 E\left[\frac{\sum_{j=1}^{k} {n_{w_j}^*}^2}{\sum_{j=1}^{k} n_{w_j}^*}\right] + \mu^2\left(E[y+k\,] - E^2\left[\sqrt{\sum_{j=1}^{k} n_{w_j}^*}\right]\right)$$

$$= c_0\tau^2 E\left[\frac{\sum_{j=1}^{k} {n_{w_j}^*}^2}{\sum_{j=1}^{k} n_{w_j}^*}\right] + \mu^2(k\lambda + k - c^2)$$

$d = E\left[\frac{\sum_{j=1}^{k} {n_{w_j}^*}^2}{\sum_{j=1}^{k} n_{w_j}^*}\right]$ Can be calculated by Monte Carlo

$Var[Z(W)] = 1 + b^2[c_0\tau^2 d + \mu^2(k\lambda + k - c^2)]$.

$E\left[Z\left(W^{(1)}\right)Z\left(W^{(2)}\right)\right] = E\left[E^2[Z(W)]|\hat{\Delta}\right] = b^2c^2E\left[\left(\sum_{i=1}^{L} V_i\hat{\Delta}_i\right)^2\right]$

$E\left[\left(\sum_{i=1}^{L} V_i\hat{\Delta}_i\right)^2\right] = Var\left[\sum_{i=1}^{L} V_i\hat{\Delta}_i\right] + E^2\left[\sum_{i=1}^{L} V_i\hat{\Delta}_i\right] = \sum_{i=1}^{L} V_i^2 Var[\hat{\Delta}_i] +$

$\left(\sum_{i=1}^{L} V_i E\left[\hat{\Delta}_i\right]\right)^2$

$$= \left(1+\frac{1}{r}\right)\sigma^2 \sum_{i=1}^{L}\left(\frac{1}{\boxed{n^*i}}\right)\cdot\left(\frac{\boxed{n_i^*}}{\sum_{i=1}^{L} n_i^*}\right)^2 + \left(\sum_{i=1}^{L} V_i\Delta_i\right)^2$$

$$= \left(1+\frac{1}{r}\right)\sigma^2 \frac{1}{\sum_{i=1}^{L} n_i^*} + \mu^2$$

$$\Rightarrow E\left[Z\left(W^{(1)}\right)Z\left(W^{(2)}\right)\right] = \boxed{b^2}c^2\left[\boxed{\left(1+\frac{1}{r}\right)\sigma^2}\frac{1}{\sum_{i=1}^{L} n_i^*} + \mu^2\right] = \frac{c^2}{\sum_{i=1}^{L} n_i^*} + b^2c^2\mu^2$$

$E\left[Z\left(W^{(1)}\right)\right]E\left[Z\left(W^{(2)}\right)\right] = E^2[Z(W)] = b^2c^2\mu^2$

$$\Rightarrow\ Cov\left[Z\left(W^{(1)}\right), Z\left(W^{(2)}\right)\right] = E\left[Z\left(W^{(1)}\right)Z\left(W^{(2)}\right)\right] - E\left[Z\left(W^{(1)}\right)\right]E\left[Z\left(W^{(2)}\right)\right]$$

$$= \frac{c^2}{\sum_{i=1}^{L} n_i^*} + b^2c^2\mu^2 - b^2c^2\mu^2 = \rho_1^*$$

APPENDIX D PARTIAL LIKELIHOOD ESTIMATION

Partial Likelihood (no tied events):

Data: $\left(T_j, \delta_j, X_j\right), j = 1,2,...,n.$

Ordered event times: $t_1 < t_2 < \cdots < t_D.$

$R(t_i)$: All individuals still under study at a time just prior to t_i.

$X_{(i)}$: Covariate of the individual whose failure time is t_i.

$$P\left(\text{the particular person in } R(t_i) \text{ died at } t_i \mid \text{one death at } t_i\right) = \frac{\exp[\beta^T X_{(i)}]}{\sum_{j \in R(t_i)} \exp[\beta^T X_j]}$$

$$LL(\beta) = \sum_{i=1}^{D} X_{(i)}^T \beta - \sum_{i=1}^{D} \log\left[\sum_{j \in R(t_i)} \exp(\beta^T X_j)\right]$$

$$U_l(\beta) = \sum_{i=1}^{D} x_{(i)l} - \sum_{i=1}^{D} \frac{\sum_{j \in R(t_i)} z_{jl} \exp(\beta^T X_j)}{\sum_{j \in R(t_i)} \exp(\beta^T X_j)}$$

In our case:

$$P\left(\text{the particular person in } R(t_i) \text{ died at } t_i \mid \text{one death at } t_i\right) =$$

$$\left\{\frac{\exp[\beta^{be} A_{(i)}]}{\sum_{j \in R(t_i)} \left\{z_j \exp[\beta^{be} A_j]\right\}}\right\}^{z_i} \left\{\frac{\exp[\beta^{nb} A_{(i)}]}{\sum_{j \in R(t_i)} \left\{\left(1 - z_j\right) \exp[\beta^{nb} A_j]\right\}}\right\}^{1-z_i}$$

$$LL\left(\beta^{be}, \beta^{nb}\right) = \sum_{i=1}^{D} \left\{z_i * \beta^{be} A_{(i)} - z_i \log\left(\sum_{j \in R(t_i)} \left\{z_j \exp[\beta^{be} A_j]\right\}\right)\right.$$

$$\left. + (1 - z_i) * \beta^{nb} A_{(i)} - (1 - z_i) \log\left(\sum_{j \in R(t_i)} \left\{\left(1 - z_j\right) \exp[\beta^{nb} A_j]\right\}\right)\right\}$$

$$U\left(\beta^{be}\right) = \sum_{i=1}^{D} z_i \left\{A_{(i)} - \frac{\sum_{j \in R(t_i)} \left[z_j A_j \exp[\beta^{be} A_j]\right]}{\sum_{j \in R(t_i)} \left\{z_j \exp[\beta^{be} A_j]\right\}}\right\} \square\ 0$$

$$U\left(\beta^{nb}\right)=\sum_{i=1}^{D}\left(1-z_i\right)\left\{A_{(i)}-\frac{\sum_{j\in R(t_i)}\left[\left(1-z_j\right)A_j\exp[\beta^{nb}A_j]\right]}{\sum_{j\in R(t_i)}\left\{\left(1-z_j\right)\exp[\beta^{nb}A_j]\right\}}\right\}\square\ 0$$

The likelihood:

$$g_j=f(T_j\mid A_j)^{\delta_j}S(T_j\mid A_j)^{1-\delta_j}=\left[\lambda(T_j\mid A_j)S(T_j\mid A_j)\right]^{\delta_j}S(T_j\mid A_j)^{1-\delta_j}$$
$$=\left[\lambda(T_j\mid A_j)\right]^{\delta_j}S(T_j\mid A_j)=\left[\lambda_0(T_j)\exp(\beta^{be}A_j)\right]^{\delta_j}S_0(T_j)^{\exp(\beta^{be}A_j)}$$

$$h_j=\left[\lambda_0(T_j)\exp(\beta^{nb}A_j)\right]^{\delta_j}S_0(T_j)^{\exp(\beta^{nb}A_j)}$$

Breslow's estimator of $H_0(T_j)=\sum_{t_{i\le T_j}}\lambda_0(t_i)$

And $S_0(T_j)=\sum_{t_{i\le T_j}}\left[1-\lambda_0(t_i)\right]$

$$\Rightarrow L_\beta(\lambda_0(t))=\prod_{j=1}^{n}g_j^{z_j}h_j^{1-z_j}$$
$$=\prod_{j=1}^{n}\left\{\left[\lambda_0(T_j)\exp(\beta^{be}A_j)\right]^{\delta_j}S_0(T_j)^{\exp(\beta^{be}A_j)}\right\}^{z_j}\cdot\left\{\left[\lambda_0(T_j)\exp(\beta^{nb}A_j)\right]^{\delta_j}S_0(T_j)^{\exp(\beta^{nb}A_j)}\right\}^{1-z_j}$$
$$=\prod_{j=1}^{n}\left\{\left[\lambda_0(T_j)\exp(\beta^{be}A_j)\right]^{\delta_j}\exp\left[-H_0(T_j)\exp(\beta^{be}A_j)\right]\right\}^{z_j}$$
$$\cdot\left\{\left[\lambda_0(T_j)\exp(\beta^{nb}A_j)\right]^{\delta_j}\exp\left[-H_0(T_j)\exp(\beta^{nb}A_j)\right]\right\}^{1-z_j}$$
$$=\prod_{j=1}^{n}\left[\lambda_0(T_j)\exp(z_j\beta^{be}A_j+(1-z_j)\beta^{nb}A_j)\right]^{\delta_j}$$
$$\cdot\exp\left[-H_0(T_j)\left[z_j\exp(\beta^{be}A_j)+(1-z_j)\exp(\beta^{nb}A_j)\right]\right]$$

Let $d_i\ \square\ z_i\beta^{be}A_i+(1-z_i)\beta^{nb}A_i$ and $e_j\ \square\ z_j\exp(\beta^{be}A_j)+(1-z_j)\exp(\beta^{nb}A_j)$

$$L_\beta(\lambda_0(t)) = \left[\prod_{i=1}^{D} \lambda_0(t_i)\exp(d_i)\right]\exp\left[-\sum_{j=1}^{n}\sum_{i=1}^{D}\left(I_{\left\{t_{i\le T_j}\right\}}\lambda_0(t_i)e_j\right)\right]$$

$$= \left[\prod_{i=1}^{D} \lambda_0(t_i)\exp(d_i)\right]\exp\left[\sum_{i=1}^{D} -\lambda_0(t_i)\sum_{j\in R(t_i)} e_j\right]$$

$$= \prod_{i=1}^{D}\left[\lambda_0(t_i)\underbrace{\exp(d_i)}_{\text{constant}}\exp\left[-\lambda_0(t_i)\sum_{j\in R(t_i)} e_j\right]\right]$$

$$\propto \prod_{i=1}^{D}\left[\lambda_0(t_i)\exp\left[-\lambda_0(t_i)\sum_{j\in R(t_i)} e_j\right]\right]$$

$$LL_\beta\left(\lambda_0(t)\right) = \sum_{i=1}^{D}\left[\log\left(\lambda_0(t_i)\right) - \lambda_0(t_i)\sum_{j\in R(t_i)} e_j\right] \Rightarrow \frac{\partial}{\partial\lambda_0(t_i)} = \frac{1}{\lambda_0(t_i)} - \sum_{j\in R(t_i)} e_j = 0$$

$$\Rightarrow \hat{\lambda}_0(t) = \frac{1}{\sum_{j\in R(t_i)} e_j} = \left(\sum_{j\in R(t_i)} z_j\exp(\beta^{be}A_j) + (1-z_j)\exp(\beta^{nb}A_j)\right)^{-1}$$

BIBLIOGRAPHY

1. Sørensen, T.I., *Which patients may be harmed by good treatments?* The Lancet 1996. **348**(9024): p. 351-352.
2. R., S., *Patient heterogeneity in clinical trials.* Cancer Treatment Reports, 1980. **64**: p. 405-410.
3. Kravitz, R.L., Duan, Naihua and Braslow, Joel . , *Evidence-Based Medicine, Heterogeneity of Treatment Effects, and the Trouble with Averages.* Milbank Quarterly 2004. **82**(4): p. 661-687.
4. Peto, R., *Statistical aspects of cancer trials.* Treatment of Cancer 1982: p. 867-871.
5. Lagakos, S.W., *The challenge of subgroup analyses-reporting without distorting.* New England Journal of Medicine, 2006. **354**(16): p. 1667-1669.
6. Rothwell, P.M., *Subgroup analysis in randomised controlled trials: importance, indications, and interpretation.* The Lancet. **365**(9454): p. 176-186.
7. Rui Wang, S.L., James Ware, David Hunter, and Jeffrey Drazen., *Statistics in medicine—reporting of subgroup analyses in clinical trials.* New England Journal of Medicine, 2007. **357**(21): p. 2189-2194.
8. Gail, M., and Simon, R. , *Testing for qualitative interactions between treatment effects and patient subsets.* Biometrics, 1985. **41**: p. 361-372.
9. Piantadosi, S., and Gail, M., *A comparison of the power of two tests for qualitative interactions.* Statistics in Medicine, 1993. **12**(13): p. 1239-1248.
10. Li, J. and I.S. Chan, *Detecting Qualitative Interactions in Clinical Trials: An Extension of Range Test.* Journal of Biopharmaceutical Statistics, 2006. **16**(6): p. 831-841.
11. Bayman, E.Ö., K. Chaloner, and M.K. Cowles., *Detecting Qualitative Interactions: A Bayesian approach.* Statistics in Medicine, 2010. **29**(4): p. 455-463.
12. Gelber., M.B.a.R., *Patterns of treatment effects in subsets of patients in clinical trials.* Biostatistics, 2004. **5**(3): p. 465-481.
13. Wei Chen, D.G., Trivellore Raghunathan, Maxim Norkin, Daniel Sargent, and Gerold Bepler., *On Bayesian methods of exploring qualitative interactions for targeted treatment.* Statistics in Medicine, 2012. **31**(28): p. 3693-3707.
14. Tian, L., Ash Alizadeh, Andrew Gentles, and Robert Tibshirani. , *A simple method for detecting interactions between a treatment and a large number of covariates.* arXiv preprint, 2012. **1212**(2995).
15. Foster, J., Taylor, J., and Ruberg, S., *Subgroup identification from randomized clinical trial data.* Statistics in Medicine, 2011. **30**(24): p. 2867-2880.
16. Dusseldorp, E.a.M., Iven Van., *Qualitative interaction trees: a tool to identify qualitative treatment–subgroup interactions.* Statistics in Medicine, 2014. **33**(2): p. 219-237.
17. Moss, A.J., et al., *Prophylactic implantation of a defibrillator in patients with myocardial infarction and reduced ejection fraction.* New England Journal of Medicine 2002. **346**(12): p. 877-883.

18. Goldenberg, I., et al. , *Risk stratification for primary implantation of a cardioverter-defibrillator in patients with ischemic left ventricular dysfunction.* Journal of the American College of Cardiology, 2008. **51**(3): p. 288-296.
19. Foster, J.C., Jeremy MG Taylor, and Stephen J. Ruberg., *Subgroup identification from randomized clinical trial data.* Statistics in Medicine, 2011. **30**(24): p. 2867-2880.
20. Xu, Y., Yu, Menggang , Zhao, Ying-Qi, Li, Quefeng, Wang, Sijian, and Shao, Jun., *Regularized outcome weighted subgroup identification for differential treatment effects.* Biometrics, 2015.
21. Marks-Konczalik, J., Costa,Maria, Robertson, Jon, McKie, Elizabeth, Yang, Shuying , and Pascoe, Steven., *A post-hoc subgroup analysis of data from a six month clinical trial comparing the efficacy and safety of losmapimod in moderate-severe COPD patients with ≤2% and >2% blood eosinophils.* Respiratory Medicine, 2015. **109**(7): p. 860-869.
22. Schou, I.M., and C Marschner, Ian., *Methods for exploring treatment effect heterogeneity in subgroup analysis: an application to global clinical trials.* Pharmaceutical statistics, 2015. **14**(1): p. 44-55.
23. Freedman, M.S., De Stefano, Nicola, Barkhof, Frederik, Polman, Chris H., Comi, Giancarlo, Uitdehaag, Bernard MJ, Casset-Semanaz, Florence et al., *Patient subgroup analyses of the treatment effect of subcutaneous interferon β-1a on development of multiple sclerosis in the randomized controlled REFLEX study.* Journal of neurology, 2014. **261**(3): p. 490-499.
24. Berger, J.O., Wang, Xiaojing, and Shen, Lei., *A Bayesian approach to subgroup identification.* Journal of biopharmaceutical statistics, 2014. **24**(1): p. 110-129.
25. Farewell, V.T., *The use of mixture models for the analysis of survival data with long-term survivors.* Biometrics, 1982. **38**: p. 1041-1046.
26. Farewell, V.T., *Mixture models in survival analysis: Are they worth the risk?* Canadian Journal of Statistics, 1986. **14**(3): p. 257-262.
27. Kuk, A.Y., and Chen-Hsin Chen., *A mixture model combining logistic regression with proportional hazards regression.* Biometrika, 1992. **79**(3): p. 531-541.
28. Ng, S.K., and G. J. McLachlan., *An EM-based semi-parametric mixture model approach to the regression analysis of competing-risks data.* Statistics in Medicine, 2003. **22**(7): p. 1097-1111.
29. Peng, Y., and Keith BG Dear., *A nonparametric mixture model for cure rate estimation.* Biometrics, 2000. **56**(1): p. 237-243.
30. Corbière, F., et al., *A penalized likelihood approach for mixture cure models.* Statistics in Medicine, 2009. **28**(3): p. 510-524.
31. Shen, C., Jeong, Jaesik, Li, Xiaochun, Chen, Peng-Sheng, *Treatment Benefit and Treatment Harm Rate to Characterize Heterogeneity in Treatment Effect.* Biometrics, 2013. **69**(3): p. 724-731.

CURRICULUM VITAE

Lin H. Taft

EDUCATION **Ph.D. in Biostatistics**(Epidemiology minor) August 2016

Department of Biostatistics

Indiana University, Indianapolis, IN

M.S. in Statistics July 2010

University of Minnesota, Duluth, MN

B.S. in Applied Mathematics, and B.S. in Economics July 2008

Xiamen University, Xiamen, China

WORK **Principal Statistician** September 2015 to present

EXPERIENCE GlaxoSmithKline, Upper Providence, PA

Research Assistant July 2013 to June 2015

School of Medicine, Indiana University, Indianapolis, IN

Graduate Assistant August 2011 to June 2013

Statistical Computer Lab/Statistical Consulting Center,

IUPUI, Indianapolis, IN

PUBLICATIONS [1] Doytchinova, A., Patel, J., Zhou, S., Chen, L. S., **Lin, H**., Shen, C., ... & Chen, P. S. (2015). Subcutaneous nerve activity and spontaneous ventricular arrhythmias in ambulatory dogs. *Heart Rhythm, 12*(3), 612-620.

[2] Hellman, Y., Malik, A. S., **Lin, H**., Shen, C., Wang, I. W., Wozniak, T. C., ... & Hadi, A. (2014). B-Type Natriuretic Peptide

Guided Therapy and Length of Hospital Stay Post Left Ventricular Assist Device Implantation. *ASAIO Journal, 61*(2), 156-160.

[3] Steenburg, S. D., Petersen, M. J., Shen, C., & **Lin, H**. (2014). Multi-detector CT of blunt mesenteric injuries: usefulness of imaging findings for predicting surgically significant bowel injuries. *Abdominal imaging, 40*(5), 1026-1033.

[4] Balint, B. J., Steenburg, S. D., **Lin, H**., Shen, C., Steele, J. L., & Gunderman, R. B. (2014). Do Telephone Call Interruptions Have an Impact on Radiology Resident Diagnostic Accuracy? *Academic radiology 21*(12), 1623-1628.

[5] Whitesell, R., Steenburg, S., Shen, C., & **Lin, H**. Facial Fracture in the Setting of Whole Body Computed Tomography for Trauma Incidence and Clinical Predictors. *American Journal of Roentgenology, 205*(1), W4-W10.