Statistical Design and Survival Analysis in Cluster Randomized Trials

by

Zhenzhen Xu

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2011

Doctoral Committee:

Professor John D. Kalbfleisch, Chair
Associate Professor Thomas M. Braun
Associate Professor Ben B. Hansen
Associate Professor Douglas E. Schaubel

To my beloved parents and Bin

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor, Jack Kalbfleisch, for his encouragement, support, and patience throughout my Ph.D studies. He led me into the area of statistical design and survival analysis, and his guidance was essential to the completion of this dissertation. Most importantly, his broad knowledge, insightful ideas, and dedication towards research will serve as an academic role model for me. I would also like to thank my committee members: Bendek Hansen, Douglas Schaubel, and Thomas Braun for their valuable comments and inspiring discussions.

Last, but not least, I would like to thank my parents, Dong Xu and Quan Ren, for their deepest love and understanding, and for many years of support which made this dissertation possible.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

CHAPTER I

Introduction

Cluster randomized trials (CRTs), in which social units are selected as the units of randomization, have been increasingly used in the past three decades to evaluate the effects of intervention operated at a community level. Examples include hospital, workplace, or community-based studies designed to improve the the health care strategy for prevention of disease. Reasons for adopting CRTs are diverse, but mainly include the administrative convenience of community allocation, a desire to avoid treatment contamination as well as maximizing the impact of intervention. The increasing popularity of cluster randomized trials has led to an extensive body of methodology and a growing literature in the design and analysis stage of CRTs (Donner and Klar, 2000; Murray et al., 2004; Donner et al., 2001; Klar and Donner, 2000; Murray, 1998). This thesis is devoted to design and analysis of cluster randomized trials. Regarding design, we propose a new randomization procedure, the balance match weighted (BMW) design, with the general aim of reducing the mean squared error (MSE) of the treatment effect estimator. Regarding analysis, we consider a Cox model with a frailty for regression analysis of correlated failure time data raised from CRTs and develop an approach based on nonparametric likelihood estimation.

Many design methods have been proposed in the literature for cluster randomizations such as the completely randomized design (Abdeljaber et al., 1991), matched-

pair design (Fisher Jr, 1995), stratified design (Graham et al., 1984), etc. On average, randomization eliminates the source of bias in treatment assignment, and achieves balance of both known and unknown confounding factors between intervention groups. In practice, however, investigators can only introduce a small amount of stratification and cannot balance on all the important variables simultaneously. This limitation arises especially when there are many confounding variables in relatively small studies. One common feature of CRTs is that they often involve a modest number of clusters since cluster recruitment is expensive. Such is the case in the *INSTINCT* trial designed to investigate the effectiveness of an education program in enhancing the Tissue Plasminogen Activator (tPA) use in stroke patients. In Chapter II, we introduce a new randomization design, the balance match weighted (BMW) design, which applies the optimal matching with constraints technique to a prospective randomized design and aims to minimize the mean squared error (MSE) of the treatment effect estimator. The key innovation of the BMW design as opposed to other existing randomized designs is that it reduces the chance imbalance in observed covariates by utilizing repeated randomizations and matching on the estimated propensity score.

The method of propensity score matching has been widely used in observational studies to control for bias (Rosenbaum and Rubin, 1984; Gu and Rosenbaum, 1993; Ming and Rosenbaum, 2000; Rosenbaum, 2002; Hansen, 2004). A balancing score, $b(x)$, is a function of observed covariates x such that the conditional distribution of x given $b(x)$ is the same for the treated and control units. Propensity score, $e(x)$, defined as the conditional probability of being assigned to treatment group given the observed covariates x, is the coarsest balancing score. Rosenbaum and Rubin (1984), Theorem 1 proved that treatment assignment z and the observed covariates x are conditionally independent given the propensity score, that is, $x \perp z|e(x)$, which implies that adjustment for the scalar propensity score is sufficient to remove bias due to all observed covariates. Although the true propensity score is typically known from

the randomization scheme in randomized experiments, matching on the estimated score may still have some substantial advantages. Indeed, it has been shown in the literature on observational studies, that matching based on an estimated propensity score has advantage over the use of the true propensity score. This is a decided advantage in observational studies since the true propensity score is typically not known. It is also of help in randomized studies, and this is what the BMW design exploits.

In Chapter II, we introduce the full matching with constrains technique and propose the BMW design which applies this technique. We then conduct a simulation study to evaluate the performance of the proposed design under various confounding scenarios and compare it with a completely randomized or matched-pair design. We also compare the BMW design with a model-based approach adjusting for the estimated propensity score and Robins-Mark-Newey E-estimation procedure in terms of efficiency and robustness of the treatment effect estimator. Finally, we illustrate these methods in proposing a design for the INSTINCT trial.

As is often the case in practice, when evaluating certain treatment programs in randomized experiments or observational studies, a more complex framework appears to be necessary. For example, a drug may be applied in differing dosage levels, a physician may have more than two treatment options to evaluate or all the study subjects may not be available at the onset of the study, instead, arrive sequentially. Therefore, in Chapter III, we aim to extend the BMW design to two directions: first, to clinical trials with more than two arms; and second, to clinical trials with staggered entry. Bo and Rosenbaum (2004) developed an algorithm for the tripartite matching problem, by transforming it to an equivalent optimal nonbipartite matching problem for which good polynomial time algorithms exist. In their approach, three groups are optimally matched into pairs yielding an incomplete block design with blocks of size two. One potential drawback of this design is the loss of efficiency due to the

insertion of the random blocks. We propose two new algorithms for matching in blocks of three with three groups: the two-way tripartite matching with triples, given a predefined reference group, and three-way tripartite matching with triples with an optimally selected reference group, respectively. We adopt these three approaches in the generalization of the BMW design to clinical trials with three or more arms and conduct a simulation study in the three arms case to evaluate the performances. The generalization of the BMW design to clinical trials with staggered entry is also discussed and followed by simulation studies in Chapter III.

Dependencies among cluster members is typical of CRTs and must be considered in the subsequent data analyses. Failure to adjust for this intra-cluster correlation can result in biased covariate-effect estimates or, more usually, inaccurate estimates of standard errors. Chapter IV deals with the regression analysis of correlated failure time data based on a Cox model with a frailty term. There is much literature dealing with the identification and estimation of frailty models using both parametric and semiparametric approaches. In these, parametric models have often been used for the frailty distribution or the baseline hazard or both (McGilchrist and Aisbett, 1991; Ripatti and Palgrem, 2000; Breslow and Clayton, 1993; Clayton, 1978; Vaupel et al., 1979; Nielsen et al., 1992; Therneau et al., 2003). However, the covariate-effect estimates, and thus the inferences one would draw, are sensitive to the parametric form assumed for the hazard and frailty (Heckman and Singer, 1984a; Trussell and Richards, 1985; Nielsen et al., 1992). We consider a frailty model with both the frailty distribution, G, and the cumulative baseline hazard, Λ_0, left nonparametric and propose an approach based on nonparametric maximum likelihood estimation. We then develop a three-step iterative algorithm and investigate its finite sample property for estimating a regression parameter β by using a simulation study. Numerical analysis results show that the proposed nonparametric approach works well for estimating β under various scenarios.

In V, we present conclusions and discuss some possible future work.

CHAPTER II

Propensity Score Matching in Randomized Clinical Trials

2.1 Introduction and motivating example

Cluster randomized trials have been widely used in the past three decades for the evaluation of health care and educational strategies, in which intact social units are selected as the units of randomization. On average, randomized treatment assignment avoids bias, achieves balance of both known and unknown confounding factors between intervention groups, and provides valid comparisons of competing intervention strategies. There is much literature that discusses design methods for cluster randomizations such as the completely randomized design (Abdeljaber et al. (1991)), matched-pair design (Fisher Jr (1995)), stratified design (Graham et al. (1984)) and minimization design (Pocock and Simon (1975)). However, investigators can only introduce a small amount of stratification in practice, which does not ensure balance on all important variables, and post hoc adjustment for many confounders is also problematic. These limitations are particularly important when there are many confounding variables in a small study.

Tissue plasminogen activator (tPA) is a clot-busting drug, which has been found to be an effective treatment for the prevention of post-stroke disability if administered

within a three hour time window of the onset of an ischemic stroke (of Neurological Disorders and rt PA Stroke Study Group (1995)). However, the use of tPA has remained relatively low. A randomized clinical trial, *INSTINCT* trial, was designed in order to investigate the effectiveness of an education program administered to hospital emergency departments in enhancing tPA therapy for stroke patients. Historical data were collected from 24 participating hospitals in Michigan regarding previous stroke volume and demographic variables. Hospitals were the units of randomization and those assigned to the treatment group received educational interventions designed to promote appropriate tPA use, whereas the other hospitals served as controls. The primary outcome is the frequency of appropriate tPA use in each hospital. Stroke volume at baseline (low vs. high), population density (urban vs. rural), age and gender mix are cluster-level factors thought to be strongly associated with outcome. Among these, stroke volume measured as number of stroke discharges and population density were classified as binary. Percentage of female (male) stroke patients who are older than 65 is used as a continuous measure. It is possible to create balance on stroke volume and population density through stratified randomization, however, it is not feasible to balance on all covariates at the same time. As a result, direct estimation of the treatment effect may be subject to bias due to possible imbalance on confounding factors. To resolve this problem, this chapter describes and evaluates a new randomization design based on propensity score matching.

The method of propensity score matching has been widely used in observational studies to control for bias (Rosenbaum and Rubin (1984); Gu and Rosenbaum (1993); Ming and Rosenbaum (2000); Rosenbaum (2002); Hansen (2004)). The propensity score is defined as the conditional probability of a subject being assigned to the treatment group given the observed covariates. Rosenbaum and Rubin (1984) showed that exact matching of treated and control subjects on the propensity score will balance all the observed covariates. In non-randomized experiments, the propensity

score function is always unknown but the sample estimates of the propensity score can be used. On the other hand, in a randomized clinical trial, the true propensity score is often a known function from the randomization scheme. For example, in the simplest randomized trial, subjects are assigned to treatment or control by the flip of a fair coin and the propensity score is equal to one half for all the subjects and the two treatment groups are perfectly matched on the true propensity score (Joffe (1999)). However, especially in small studies, substantial chance imbalances may still exist and yield some (conditional) bias in the direct treatment effect estimator. Although methods based on the estimated propensity score have not been widely used in the randomized studies, it could have some substantial advantages over the methods by using the true scores under certain scenarios. Robins et al. (1992) has shown that there are even theoretical advantages to using estimated propensity scores.

We introduce a new randomization design, the balance match weighted (BMW) design, which applies the optimal full matching with constraints technique (Olsen (1997)) to the given randomization with the general aim of reducing the mean squared error of the treatment effect estimator. In this design, treated and control subjects are matched into subsets based on their estimated propensity score and an overall estimate is constructed using a weighted sum of the subset-specific estimates. In contrast to the existing stratified design, which first stratifies and then randomizes within strata, the BMW design first randomly assign the units to treatments and then stratifies on the randomized sample. In an implementation of the design, this randomization-stratification process is repeated M times in order to choose a randomization that gives a good overall balance. In general, the BMW design has two advantages. First, it reduces the chance imbalance between the treatment groups in observed covariates through optimal matching, and hence decreases the (conditional) bias in the resultant estimator. Second, it controls for the increase in variance due to matching by using the full matching with constraints technique (Olsen, 1997), in

which the choice of the constraint, k, adjusts for the trade-off between the potential gain in bias reduction and possible loss in precision. We examine various strategies for selecting M and k, seeking a good choice which yields good results with respect to mean squared error. It is obvious that MSE performance also depends on the inherent degree of confounding, so we compared the BMW design with the completely randomized design and matched-pair design under different confounding scenarios. If there is no confounding, the three design methods perform equally well. However, if there is considerable confounding, the BMW design can result in a substantial reduction in the MSE of the treatment effect estimator.

The design we propose is appropriate for the situation where all units are available for randomization at the onset, and can't be applied to clinical trials with staggered entry. Pocock and Simon (1975) proposed a sequential strategy, minimization design, which makes the assignment decision one unit at a time, based solely on the covariate information of previously assigned subjects. On the other hand, the minimization design is not well suited for trials where all observational units are available for randomization at the onset.

The rest of the chapter is organized as follows. Notation and models are presented in Section 2. The BMW design is outlined in Section 3 and Section 4 gives results of a simulation study comparing the performance of the BMW design with the completely randomized design, a matched-pair design, the model-based approach by adjusting for the estimated propensity score and the Robins-Mark-Newey E-estimation procedure. Its performance under heterogeneous error is also investigated. Section 5 outlines a case study and the chapter concludes with discussion in Section 6.

2.2 Methods

In this section, we present the notation and problem formulation as well as introduce some optimal matching techniques employed in the proposed design.

2.2.1 Optimal matching

Consider a study with the aim of assessing the effect of treatment. Let N denote the number of subjects available for the study. We assume that N is even and $N/2$ subjects are randomized to each of the treatment and control groups, but the method we propose could allow imbalance in the randomized assignment. Thus, we suppose that a randomization process divides the N subjects into a set T of $N/2$ subjects to be treated and a set C of $N/2$ subjects to receive the control. We also assume that a vector of r covariates, $X = (X_1, X_2, ..., X_r)^{\mathrm{T}}$, is observed for each individual.

Similarity of covariates is measured through an estimated propensity score. Writing Z=1 for the treated subjects, and Z=0 for the control subjects, the (estimated) propensity score distance between the treated unit i and control unit j is given by

$$d_{i,j} = |\widehat{\delta}_i - \widehat{\delta}_j| \tag{2.1}$$

where $\widehat{\delta}_i$ is the estimate of the true propensity score, $\delta_i = Pr(Z = 1 \mid X_i)$, and is obtained from a model such as the logistic regression model

$$\delta_i = Pr(Z = 1 \mid X_i; \alpha) = \exp(\alpha_1 + \sum_{j=2}^{r} \alpha_j X_{ij}) / \{1 + \exp(\alpha_1 + \sum_{j=2}^{r} \alpha_j X_{ij})\} \tag{2.2}$$

In a randomized clinical trial, the true propensity score δ_i is typically determined by the randomization scheme and known. We consider the estimated propensity score $\widehat{\delta}_i$ in defining the distances with the aim of producing a design that reduces the actual observed imbalance between treated and control subjects. Matching assembles treated and control units which are as similar as possible into the same stratum using the overall estimated propensity score distance measure. Given T and C, we consider the collection $P_{C,T}$ of all possible matchings, where a matching corresponds to a collection of S strata comprised of matched subsets $\{(C_1, T_1), (C_2, T_2), ..., (C_S, T_S)\}$, in which,

$C_1, C_2, ..., C_S$ is a partition of C, $T_1, T_2, ..., T_S$ is a partition of T and $1 \leq S \leq N/2$. As is often done (e.g. Rosenbaum (2002)), we measure the quality of a particular matching as

$$\Delta = \sum_{s=1}^{S} w(|T_s|, |C_s|) \bullet \overline{T_s \times C_s} \tag{2.3}$$

where

$$\overline{T_s \times C_s} = \sum_{(i,j) \in T_s \times C_s} |\widehat{\delta}_i - \widehat{\delta}_j| / |T_s \times C_s|$$

is the average distance between the $|T_s \times C_s|$ possible pairs in the s-th strata, and $w(.,.)$ is a weight function. Thus, Δ is a weighted sum of average distances and an optimal matching minimizes Δ over $P_{C,T}$.

A full matching is one in which each stratum is comprised of one treated (or control) subject matched to one or more control (or treated) subjects so that $\min(|T_s|, |C_s|) = 1$, for $s = 1, 2, ..., S$. Rosenbaum (2002) showed that if the weight function in (2) is *neutral* or *favors small subclasses*, then there is always a full matching that is optimal. Among the class of full matchings with the weight function $w(|T_s|, |C_s|) = |T_s| + |C_s| - 1$, equation (2.3) reduces to

$$\Delta = \sum_{s=1}^{S} (|T_s| + |C_s| - 1) \bullet \overline{T_s \times C_s} = \sum_{s=1}^{S} \sum_{(i,j) \in T_s \times C_s} |\widehat{\delta}_i - \widehat{\delta}_j|. \tag{2.4}$$

In this chapter, we use this total distance measure to evaluate the quality of a matching. One potential drawback of the optimal full matching is that some of its matched subsets can be very unbalanced with many controls to one treatment or vice versa. The imbalance among full matching subsets decreases the precision of the estimated treatment effect. One remedy for this is to constrain the full matching so that the ratio of the number of treated versus the number of controls in each stratum is between a lower and upper bound. To accomplish this, we choose an integer

$k \in \{1, 2, ..., N/2 - 1\}$ and consider the optimization problem

$$\text{Minimize } \Delta = \sum_{s=1}^{S} \sum_{(i,j) \in T_s \times C_s} |\widehat{\delta}_i - \widehat{\delta}_j|. \tag{2.5}$$

over the class of full matchings subject to $k^{-1} \leq |T_s|/|C_s| \leq k$. We refer to the solution to this optimization problem as the *optimal full matching with constraint* k. When $k = 1$, we obtain the best matched-pair design with one treated unit and one control unit in each stratum. This assignment leads to a treatment effect estimator with minimum variance in the linear model discussed in the next section, but can result in relatively large bias. When $k = N/2 - 1$, there is no constraint on the balance in the relative numbers of treated and control units in any matched subset, the covariates are optimally balanced so the bias of treatment effect estimator tends, in this case to be small, but the variance is larger. The BMW design we propose searches for the optimal full matching with constraint k. The choice of k represents a trade-off between bias and variance. In the next section, we examine the mean squared error (MSE) as a measure of this trade-off with a class of linear models. For a specific model in this class, we can choose k to generate a BMW design that achieves minimum MSE. It is observed that the choice of k does not depend much on the specific model.

2.2.2 Model

To appreciate the effect of treatment on response in a pooled sample and matched sample, respectively, consider the following model: Let Y_i, i = 1,2,...,N, represents responses of the unit i, conditional on a given treatment assignment T, C and X,

$$Y_i = \alpha + \beta I(i \in T) + \sum_{j=1}^{r} \gamma_j X_{ij} + \varepsilon_i; \tag{2.6}$$

where $I(.)$ is the indicator function, β denotes the true treatment effect, $\gamma_1, \gamma_2, ..., \gamma_r$ are the confounding effects and $\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_N)$ is the vector of the measurement errors with $E[\varepsilon|T, C, X] = 0, \text{Var}[\varepsilon|T, C, X] = \sigma^2 I, \sigma^2 < +\infty$ and I is the $N \times N$ identity matrix.

2.2.2.1 *Pooled sample*

Under model (2.6), the common treatment effect estimator based on the unstratified pooled sample is $\widehat{\beta}_{pool} = \overline{y}_T - \overline{y}_C$, which has conditional expectation

$$E[\widehat{\beta}_{pool}|T, C, X] = \beta + \sum_{j=1}^{r} \gamma_j (\overline{X}_{jT} - \overline{X}_{jC}) \tag{2.7}$$

where the subscripts C and T mean that the averages are computed over the control and treatment groups, respectively. The mean squared error (conditional on T, C and X) is

$$MSE(\widehat{\beta}_{pool}|T, C, X) = \{\sum_{j=1}^{r} \gamma_j (\overline{X}_{jT} - \overline{X}_{jC})\}^2 + 4\sigma^2/N \tag{2.8}$$

2.2.2.2 *Matched sample*

Under model (2.6), estimating the treatment effect for the matched sample involves the computation of a weighted sum. In the s^{th} matched subset (T_s, C_s), the treatment effect estimator is $\widehat{\beta}_{strata,s} = \overline{y}_{T_s} - \overline{y}_{C_s}$, which has conditional expectation

$$E[\widehat{\beta}_{strata,s}|T, C, X] = \beta + \sum_{j=1}^{r} \gamma_j (\overline{X}_{jT_s} - \overline{X}_{jC_s}) \tag{2.9}$$

The overall estimate can be constructed using a weighted sum,

$$\widehat{\beta}_{strata} = \sum_{s=1}^{S} w_s \widehat{\beta}_{strata,s} \tag{2.10}$$

where $\sum_s w_s = 1, w_s \geq 0$. It should be noted that this stratified estimator can be modified to accommodate different weighting methods. Two common choices are weighting in proportion to the number of subjects that each subset contains, $(|T_s| + |C_s|)/N$ (Cochran (1968)), or the inverse variance weighting, $(1/|T_s|+1/|C_s|)^{-1}/\sum_{t=1}^{S}(1/|T_t|+1/|C_t|)^{-1}$. For purpose of this discussion, the former weighting method is considered, but it can be easily modified to handle the latter. It follows that the MSE of the stratified estimator (conditional on T, C and X) can be written as

$$\begin{aligned} MSE(\widehat{\beta}_{strata}|T,C,X) &= \{\sum_{s=1}^{S}\frac{(|T_s|+|C_s|)}{N}\sum_{j=1}^{r}\gamma_j(\overline{X}_{jT_s}-\overline{X}_{jC_s})\}^2 \\ &\quad +\sum_{s=1}^{S}\frac{(|T_s|+|C_s|)^2}{N^2}(\frac{1}{|T_s|}+\frac{1}{|C_s|})\sigma^2 \end{aligned}$$

With no confounding effects or chance imbalance in covariates, the pooled estimator is the unbiased estimate of the treatment effect with minimum variance. In the presence of confounding, stratification reduces the bias but increases the variance. We use the mean squared error to measure the trade-off between bias and variance.

2.3 The BMW design

In a randomized trial with fixed small sample size N and many confounding covariates, it may be impossible to produce balance on all of the variables simultaneously. In order to reduce the actual observed imbalance as well as increase precision of the estimator, we propose the balance match weighted (BMW) design. The design with specified parameter k and M is defined algorithmically as follows:

Step 1. Randomize half of the subjects to the treatment group, and half to control to obtain sets T and C;

Step 2. Compute the estimated propensity scores and create the $|T| \times |C|$ matrix of estimated propensity score distances;

Step 3. Obtain the optimal full matching with constraint k and record the total distance Δ_k.

Step 4. Repeat *Steps* 1 to 3 M times; choose the randomized sample with minimum total distance $\Delta_k^* = \min(\Delta_{1k}, \Delta_{2k}, ..., \Delta_{Mk})$. The choice of M is discussed below.

It is clear that the choice of k represents a trade-off between bias and variance. We use mean squared error (MSE) as a measure of the trade-off. The choice of k ($k \in (1, 2, ..., N/2 - 1)$) which minimizes the MSE of the treatment effect estimator depends on the confounding effect $\gamma = (\gamma_1, \gamma_2, ..., \gamma_r)$. If γ were known and M is fixed, it would be possible to compute the MSE for each k based on the BMW design in *Step 1* to *4* above. It would be possible then to select the k that minimizes the mean squared error. In practice, of course, the true value of γ is unknown; therefore in the next section we use a simulation study to evaluate the effects of k on reducing the MSE under a variety of assumptions about the size of the confounding effects. We find that $k = 2$ is a suitable choice under most of the confounding scenarios considered.

Clearly, the larger M is, the better matching the BMW design attains. In the next section, we explain how the MSE depends on M and find that most of the gain is attained by relatively small M of 10 or so in the cases considered. So we recommend value of M in this range. It should also be noted that, as M increases, the BMW design becomes more deterministic given covariates in the experimental units.

The implementation of *Step* 3 which searches the optimal full matching with constraint k (Olsen, 1997) is conducted using the *Optmodel* Procedure in SAS (Version 9.1.3.2). A similar program *Optmatch* in R has also been developed (Hansen, 2004).

There are alternative ways to adjust for the covariate imbalance resulting from randomization. Since small sample sizes do not allow for control of all variables by model-based method, one possible approach, suggested by an Associate Editor, is to

adjust the estimated propensity score in a regression model such as:

$$Y_i = \alpha + \beta I(i \in T) + \gamma \widehat{\delta}_i + \varepsilon_i. \tag{2.11}$$

Let $\widehat{\beta}_{MB}$ denote the ordinary least squares estimate of β from (12). Our simulations and investigations suggest that the model-based approach seems to work well if the model for the propensity score is *appropriately* specified, where, by appropriately specified, we mean that the regression model for the propensity score includes the same regression parameters and is of the same form as the true model for the outcome variable Y. For example, if the true model is $Y_i = \alpha+\beta I(i \in T)+\gamma_1 X_i+\gamma_2 X_i^2+\varepsilon_i$ and we specify $\text{logit}(\delta_i) = \text{logit}(Pr(Z = 1 \mid X_i; \alpha)) = \alpha_1 + \alpha_2 X_i + \alpha_3 X_i^2$, then regression adjustment using $\widehat{\delta}_i$ will tend to work well. In fact, if the confounding effects are large, $\widehat{\beta}_{MB}$ tends to be somewhat more efficient than the estimator obtained from the BMW approach. On the other hand, the BMW approach is more robust if the propensity score model is *inappropriately* specified as, for example, if the same true model of Y holds and we specify $\text{logit}(\delta_i) = \text{logit}(Pr(Z = 1 \mid X_i; \alpha)) = \alpha_1 + \alpha_2 X_i$. This is examined further in the simulations of Section 4.

Robins and Newey (1992) proposed another procedure based on the propensity score in observational studies. Their approach is designed to provide a consistent estimator, $\widetilde{\beta_E}$, when the model for propensity score $\widehat{\delta}_i$ is *correctly* specified. This estimator is

$$\widetilde{\beta_E} = \sum_{i=1}^{n} Y_i(Z_i - \widehat{\delta}_i) \;/\; \sum_{i=1}^{n} Z_i(Z_i - \widehat{\delta}_i). \tag{2.12}$$

At the suggestion of a reviewer, we also evaluate this approach in the simulations of the next section.

2.4 Simulation Study

In order to assess the performance of the BMW design, we first carried out a simulation study to compare it with a completely randomized design and a matched-pair design. In doing so, we considered a wide variety of settings and, for each setting, estimated the mean squared error based on 1000 replications.

2.4.1 Structure of the simulation

For each of N subjects, we generated a set of r covariates $X_1, X_2, ..., X_r$, where the covariates were drawn independently from various distributions as described below. Given a randomization of subjects to the two treatment groups, the responses were generated conditional on the treatment assignment ($Z_i = 0$ or 1) and the covariates (X_{ij}), where $Pr(Z_i = 1 \mid X_{ij}) = 0.5$. Specifically, the response was obtained from:

$$Y_i = \beta Z_i + \sum_{j=1}^{r} \gamma_j X_{ij} + \varepsilon_i \tag{2.13}$$

where $\varepsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0,1)$ and $i = 1, 2, ..., N$. In the simulations, we considered the following:

- The true treatment effect was taken to be $\beta = 0.7$
- The true confounding effects were $\gamma_j = \gamma,\ j = 1, ..., r$ where $\gamma = 0.5,\ 1.0,\ 1.5$. Note that the results we obtain do not depend on the choice of β. When the covariates follows symmetric distributions, the results do not depend on the signs of the components of γ either.
- For the first three settings, we considered $r = 4$ covariates selected from the following distributions: $(i) X_1, X_2, X_3, X_4 \overset{i.i.d}{\sim}$ Bernoulli(0.5); $(ii) X_1, X_2 \overset{i.i.d}{\sim}$ Bernoulli(0.5); $X_3, X_4 \overset{i.i.d}{\sim} N(0, 0.25)$; $(iii) X_1, X_2 \overset{i.i.d}{\sim}$ Bernoulli(0.5); $X_3, X_4 \overset{i.i.d}{\sim}$ Bernoulli(0.66).

- For the fourth case, we considered $r = 8$ covariates:

  $X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8 \overset{i.i.d}{\sim}$ Bernoulli(0.5)

- We consider sample sizes $N = 30,\ 60$.

The completely randomized design assigns half of the units at random to each of the two treatment groups. For this design, the treatment effect estimator is $\widehat{\beta}_{pooled} = \overline{Y}_T - \overline{Y}_C$ and the corresponding mean squared error (conditional on T, C and X) is given in (2.8). We also consider a matched-pair design in which each unit is matched (so much as possible) to another unit based on the first covariate X_1. One unit in each pair is then randomly assigned to treatment and one to control. The BMW design, as described in the preceding section, creates an optimally matched sample for each constraint k, where k =1,2,...,$N/2-1$, and for each choice of M, this leads to the weighted treatment effect estimator $\widehat{\beta}_{strata}$ in (2.10) along with its mean squared error (2.11). We further consider $\widehat{\beta}_{MB}$, from the model-based approach by adjusting for the estimated propensity score (2.11) and the Robins-Mark-Newey E estimator $\widetilde{\beta_E}$ (2.12). Finally, we examine the possible effects of homoscedastic error on the BMW design by allowing the error variance to depend on the first covariate X_1.

2.4.2 Results

The average mean squared errors based on 1000 replications are summarized in table 2.1. From Cochran (1968), the true unconditional MSE of $\widehat{\beta}_{pool}$ is $4\sigma_y^2/N$, where σ_y^2 refers to the overall variability in outcome Y. In this, one part, $\sum_j \gamma_j^2 Var(\overline{X}_{jT} - \overline{X}_{jC})$, is due to variability in the observed covariates $X_1, X_2, ..., X_r$ and the other to the conditional variations of Y given $X_1, X_2, ..., X_r$. Formally, the unconditional MSE is

(from (2.8))

$$\begin{aligned} MSE(\widehat{\beta}_{pool}) &= E[\{\sum_{k=1}^{K}\gamma_k(\overline{X}_{kT}-\overline{X}_{kC})\}^2+\frac{4}{N}\sigma^2] \\ &= \sum_{k=1}^{K}\gamma_k^2 Var(\overline{X}_{kT}-\overline{X}_{kC})+\frac{4}{N}\sigma^2 \\ &= \frac{4\sigma_y^2}{N} \end{aligned} \tag{2.14}$$

With pre-randomization matching or post-randomization stratification on covariates, the average MSE values are also obtained in the simulation. A similar formula to (2.14) can be obtained for the matched pairs design, but formulas for the BMW design are complicated. For the BMW design, the average MSE for each constraint $k = (1, 2, ..., N/2-1)$ were examined in the simulations, but only those for $k = 1, 2, 3$ are displayed since the MSE changes little when k increases over three. The percent reduction in MSE is $100 \times (MSE - MSE^*_{BMW})/MSE$, where MSE^*_{BMW} corresponds to the minimal value of MSE for each k in the BMW design, and MSE refers to the MSE value for the design to which BMW is being compared (e.g. the completely randomized design or the matched-pair design).

It is interesting to examine how the MSE of the treatment effect estimator is affected by various parameter settings. Overall, the BMW design shows significant reductions in MSE as compared to both the completely randomized and matched-pair designs.

2.4.2.1 Confounding effects γ_j

Table 2.1 reveals that as the confounding effects, measured by $\sum_j \gamma_j$, increase, the average mean squared errors generally increase. However, the MSE in the BMW design increases much more slowly than the MSE in the completely randomized or matched-pair design. This suggests that the BMW design becomes much more effec-

tive in reducing the mean squared error when confounding effects increase. Specifically, as we raise $\sum_{j=1}^{r} \gamma_j$ from 2.0 to 6.0 for Bernoulli distributed covariates (Table 2.1), the MSE reduction of the BMW design with $k = 2$ compared to the matched pair design varies dramatically from 5.96% to 53.77% for $M = 5$, from 7.50% to 54.59% for $M = 10$, and from 9.36% to 56.10% for $M = 20$. An even larger reduction in MSE arises when comparing the BMW design with the completely randomized design.

2.4.2.2 The Choice of the Constraint k

We now examine the MSE as a function of k. When the model contains four covariates of various forms (Table 2.1) and there is relatively little confounding such as $\sum_{j=1}^{r} \gamma_j = 2.0$, then the MSEs corresponding to $k = 1$ are slightly smaller than those corresponding to $k = 2$. As $\sum_{j=1}^{r} \gamma_j$ increases, however, a greater reduction in MSE due to constraint $k = 2$ becomes apparent. Intuitively, for a small sample with strong confounding effects, bias reduction is more important than variance reduction, so the larger value of k ($k = 2$) is more efficient. However, when the number of covariates is $r = 8$, the constraint $k = 2$ minimizes the MSE for all confounding effects considered.

2.4.2.3 Number of Replication M

The MSE is obviously a decreasing function of M for given γ and k. However, when it comes to percent reduction in MSE using the BMW design as compared to the completely randomized design or the matched-pair design, there is an interesting interplay between the number of replications, M, and the confounding effect $\sum_{j} \gamma_j$. The results suggest that if there is little confounding ($\sum_{j=1}^{r} \gamma_j = 2.0$) and the covariates are independently Bernoulli distributed (Table 2.1), the percent reduction in MSE of the BMW design versus the matched-pair design increases from 7.96% to 10.29% to 13.46% for M from 5 to 10 to 20, with $k = 1$. If there is relatively more confounding

($\sum_{j=1}^{r} \gamma_j = 6.0$), the percent reduction in MSE increases more modestly from 53.77% to 54.59% to 56.10% with M, while using matching with constraints $k = 2$. Similar trends are seen in comparing the BMW design with the completely randomized design or using different covariate distributions. We conclude that, when confounding effect are relatively strong, the BMW design even with relatively small M is very effective in reducing MSE. A good compromise value of M is $M = 10$ for the cases considered.

2.4.2.4 Covariate Settings

There are four covariate settings examined in the simulation studies. The results suggest that, in situations where existing designs often fail in producing balance across covariates, the BMW design provides a useful approach. Gains in efficiency are substantial when the covariates are Bernoulli variables with important, but somewhat more modest gains, when the covariates include continuous variables. For given γ, the gains due to the BMW design are similar for symmetric and asymmetric Bernoulli distributions for the covariates. Finally, when the number of Bernoulli covariates increases from four to eight, the BMW design achieves a larger reduction in MSE.

2.4.2.5 Sample Size N

Sample size has an impact on the performance of the BMW design, and as sample size becomes very large, we would expect the relative gains to decrease as randomization itself guarantees substantial balance among the covariate values. Our simulation results reveal, however, that when the sample size increases from 30 to 60, the percent reduction in MSE from the BMW design decreases only very little. This suggests a possible value for this approach even in larger studies. Computational aspects are easily accommodated for the larger sample sizes; for example, the processing time for the simulations with $N = 60$ increases by about 40% over those for $N = 30$.

It is also of interest to compare the BMW design with the model-based approach

Table 2.1: Percent reductions in the MSE of treatment effect estimator for the BMW design compared to a completely randomized design (CR) and matched-pair design (MP). Sample size N=30 subjects. Number of replications=1000.

γ	$\sum_{j=1}^{4}\gamma_j$	M	MSE (CR)	MSE Percent Reduction(%) (BMW vs. CR Design)			MSE (MP)	MSE Percent Reduction(%) (BMW vs. MP Design)		
				$k=1$	$k=2$	$k=3$		$k=1$	$k=2$	$k=3$
$X_1, X_2, X_3, X_4 \overset{i.i.d}{\sim} Bernoulli(0.5)$										
		5		12.21	10.30	6.87		7.96	5.96	2.37
(0.5,0.5,0.5,0.5)	2	10	0.166	14.43	11.77	7.14	0.158	10.29	7.50	2.64
		20		17.45	13.54	8.81		13.46	9.36	4.40
		5		35.61	43.58	39.67		24.57	33.90	29.33
(1.0,1.0,1.0,1.0)	4	10	0.280	40.37	44.45	41.74	0.239	30.15	34.92	31.75
		20		50.39	48.66	46.21		41.87	39.86	36.99
		5		45.39	61.58	57.94		34.29	53.77	49.39
(1.5,1.5,1.5,1.5)	6	10	0.450	52.19	62.26	59.02	0.374	42.47	54.59	50.69
		20		58.43	63.52	60.64		49.97	56.10	52.64
$X_1, X_2 \overset{i.i.d}{\sim} Bernoulli(0.5); X_3, X_4 \overset{i.i.d}{\sim} N(0, 0.25)$										
		5		8.77	5.67	1.09		4.45	1.21	-3.59
(0.5,0.5,0.5,0.5)	2	10	0.155	9.46	5.85	1.66	0.148	5.17	1.40	-2.99
		20		12.17	7.74	3.52		8.01	3.38	-1.05
		5		24.37	30.79	27.29		13.20	20.58	16.56
(1.0,1.0,1.0,1.0)	4	10	0.218	28.89	32.40	29.18	0.190	18.39	22.42	18.73
		20		32.85	33.09	30.13		22.94	23.22	19.82
		5		35.91	50.61	47.45		19.56	38.01	34.04
(1.5,1.5,1.5,1.5)	6	10	0.316	42.98	52.08	48.45	0.252	28.43	39.85	35.29
		20		48.35	51.58	48.30		35.17	39.22	35.10
$X_1, X_2 \overset{i.i.d}{\sim} Bernoulli(0.5); X_3, X_4 \overset{i.i.d}{\sim} Bernoulli(0.66)$										
		5		12.11	12.08	7.31		8.03	8.00	3.01
(0.5,0.5,0.5,0.5)	2	10	0.165	14.93	12.99	8.78	0.158	10.98	8.96	4.55
		20		16.13	12.69	8.72		12.24	8.64	4.48
		5		32.21	40.76	36.77		20.97	30.94	26.29
(1.0,1.0,1.0,1.0)	4	10	0.267	37.92	43.13	39.39	0.229	27.63	33.71	29.34
		20		41.88	44.14	41.22		32.25	34.88	31.48
		5		50.98	61.68	59.36		40.15	53.20	50.37
(1.5,1.5,1.5,1.5)	6	10	0.430	50.63	59.12	55.57	0.352	42.75	52.60	48.48
		20		55.05	59.33	56.08		47.87	52.84	49.07
$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8 \overset{i.i.d}{\sim} Bernoulli(0.5)$										
		5		17.35	23.93	18.68		10.06	17.21	11.49
(0.5,0.5,0.5,0.5,	4	10	0.204	18.63	24.30	19.63	0.187	11.44	17.62	12.53
0.5,0.5,0.5,0.5)		20		22.65	25.22	19.42		15.82	18.62	12.30
		5		28.74	52.41	52.21		23.39	48.83	48.62
(1.0,1.0,1.0,1.0,	8	10	0.390	35.80	56.12	53.11	0.363	30.97	52.82	49.58
1.0,1.0,1.0,1.0)		20		43.23	57.60	54.22		38.96	54.41	50.78
		5		35.07	66.86	68.47		29.12	63.83	65.58
(1.5,1.5,1.5,1.5,	12	10	0.725	46.71	71.55	69.76	0.664	41.83	68.94	66.99
1.5,1.5,1.5,1.5)		20		52.71	73.14	70.29		48.38	70.68	67.57

adjusting for the estimated propensity score and Robins-Mark-Newey E-estimation procedure in terms of efficiency and robustness of the treatment effect estimator. Therefore, we evaluate the MSE property of the three approaches under two scenarios, one where the propensity score model is *appropriately* specified and one where it is not.

2.4.2.6 Propensity score *appropriately* specified

Under this scenario, we specify the true model and propensity score model as follows:

$$Y_i = \alpha + \beta I(i \in T) + \sum_{j=1}^{4} \gamma_j X_{j,i} + \varepsilon_i. \tag{2.15}$$

$$\text{logit}(\delta_i) = \text{logit}\{Pr(Z = 1 \mid X_i; \alpha)\} = \alpha_0 + \sum_{j=1}^{4} \alpha_j X_{j,i} \tag{2.16}$$

From the results summarized in Table 2.2, we see that the MSE obtained by the model-based approach remains relatively constant as the confounding effects increase, provided the terms in the propensity score model mimic that in the true model for Y. If there is relatively little confounding ($\sum_{j=1}^{r} \gamma_j < 6.0$), the MSEs in the BMW design are slightly smaller than those from the model-based approach. As $\sum_{j=1}^{r} \gamma_j$ increases, however, a somewhat greater reduction in MSE is obtained through the model-based approach. Both the BMW design and the model-based estimate perform much better than the E-estimation procedure in the context of these small randomized experiments.

2.4.2.7 Propensity score *inappropriately* specified

In practice, the true model for outcome Y is unknown, and due to the small sample size, it is difficult to determine what model is best; consequently adjustment for many potential confounders may not work well. The simulation studies in Table 2.2 suggest

that when the propensity score model does not mimic the correct regression terms in the true model, the BMW design provides a more robust approach than the model based approach. For illustration purpose, we looked at a true model and propensity score model as follows:

$$Y_i = \alpha + \beta I(i \in T) + \gamma_1 X_i + \gamma_2 X_i^2 + \varepsilon_i. \tag{2.17}$$

$$\text{logit}(\delta_i) = \text{logit}\{Pr(Z = 1 \mid X_i; \alpha)\} = \alpha_1 + \alpha_2 X_i, \tag{2.18}$$

where $X_i \overset{i.i.d}{\sim} Normal(0, 1)$. As the confounding effects γ_j increases from 0.5 to 1.5, the percent reduction in MSE of the BMW design compared to the model-based approach increases from 14.75% to 41.88%, for $M = 10$. Again, the E-estimation procedure does not perform well in this context. This suggests that the BMW design is more robust than the model-based approach when the propensity score model is *inappropriately* specified, as would often be the situation in practice.

2.4.2.8 Heteroscedastic errors

In the clustered randomized trials with few but relatively large clusters, the hemoscedasticity error assumption is unlikely to hold. To investigate the effects of this, we allowed the error distribution of the outcome to vary by the first covariate X_1 in our simulation studies. In particular, in the model (2.6), we specified $\varepsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0, 1)$ if $X_1 = 1$ and $\varepsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0, 0.25)$ if $X_1 = 0$, where $X_1, X_2 \overset{i.i.d}{\sim}$ Bernoulli(0.5) and $X_3, X_4 \overset{i.i.d}{\sim} N(0, 0.25)$. The results in Table 2.3 suggest that the relaxation of the hemoscedasticity error assumption has little impact on the performance of the BMW design. The case study in the next section is a case where such heteroscedasticity may be present.

Table 2.2: Percent reductions in the MSE of treatment effect estimator for the BMW design compared to the model-based adjustment approach adjusting for the estimated propensity score (MB) and E estimation procedure (E-est), where the propensity score model is *appropriately* and *inappropriately* specified, respectively. Number of replications=1000.

γ	M	MSE (MB)	MSE Percent Reduction(%) (BMW vs. MB)			MSE ($E-est$)	MSE Percent Reduction(%) (BMW vs. $E-est$)		
			$k=1$	$k=2$	$k=3$		$k=1$	$k=2$	$k=3$
where propensity score *inappropriately* specified (17) (18)									
$X \overset{i.i.d}{\sim} Normal(0, 0.25)$									
(0.5, 0.5)	10	0.185	0.65	14.75	12.25	0.334	45.06	52.85	51.47
(1.0, 1.0)	10	0.365	-0.15	30.03	32.31	0.964	62.10	73.52	74.39
(1.5, 1.5)	10	0.665	5.80	41.88	46.12	2.013	68.90	80.81	82.21
where propensity score *appropriately* specified (15) (16)									
$X_1, X_2, X_3, X_4 \overset{i.i.d}{\sim} Bernoulli(0.5)$									
(0.5,0.5,0.5,0.5)	10	0.165	15.01	15.74	6.79	0.211	33.41	33.98	26.97
(1.0,1.0,1.0,1.0)	10	0.166	-0.87	6.02	1.44	0.528	68.38	70.54	69.10
(1.5,1.5,1.5,1.5)	10	0.166	-29.84	-2.49	-11.31	0.971	77.85	82.52	81.01
$X_1, X_2 \overset{i.i.d}{\sim} Bernoulli(0.5); X_3, X_4 \overset{i.i.d}{\sim} Bernoulli(0.66)$									
(0.5,0.5,0.5,0.5)	10	0.152	7.19	5.08	0.48	0.247	42.97	41.68	38.85
(1.0,1.0,1.0,1.0)	10	0.152	-8.99	0.16	-6.41	0.492	66.32	69.15	67.12
(1.5,1.5,1.5,1.5)	10	0.153	-32.00	-9.29	-18.78	0.916	77.99	81.78	80.19
$X_1, X_2 \overset{i.i.d}{\sim} Bernoulli(0.5); X_3, X_4 \overset{i.i.d}{\sim} N(0, 0.25)$									
(0.5,0.5,0.5,0.5)	10	0.148	5.41	1.64	-2.74	0.203	30.71	27.95	24.74
(1.0,1.0,1.0,1.0)	10	0.148	-4.52	0.64	-4.09	0.387	59.89	61.88	60.06
(1.5,1.5,1.5,1.5)	10	0.148	-21.56	-2.15	-9.91	0.689	73.82	78.00	76.33

Table 2.3: Percent reductions in the MSE of treatment effect estimator for the BMW design compared to a completely randomized design (CR) and matched-pair design (MP) under the heteroscedastic and hemoscedastic error assumption, respectively. Number of replications=1000. $X_1, X_2 \overset{i.i.d}{\sim} Bernoulli(0.5); X_3, X_4 \overset{i.i.d}{\sim} N(0, 0.25)$

γ	$\sum_{j=1}^{8} \gamma_j$	M	MSE (CR)	MSE Percent Reduction(%) (BMW vs. CR Design)			MSE (MP)	MSE Percent Reduction(%) (BMW vs. MP Design)		
				$k=1$	$k=2$	$k=3$		$k=1$	$k=2$	$k=3$
$\varepsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0,1)$ if $X_1 = 1$ and $\varepsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0, 0.25)$ if $X_1 = 0$										
(0.5,0.5,0.5,0.5)	2	10	0.089	14.19	13.32	8.83	0.075	-0.90	-1.93	-7.20
(1.0,1.0,1.0,1.0)	4	10	0.152	41.40	49.28	44.78	0.121	26.63	36.50	30.87
(1.5,1.5,1.5,1.5)	6	10	0.258	52.57	65.36	65.05	0.166	26.48	46.31	45.83
$\varepsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0,1)$										
(0.5,0.5,0.5,0.5)	2	10	0.155	9.46	5.85	1.66	0.148	5.17	1.40	-2.99
(1.0,1.0,1.0,1.0)	4	10	0.218	28.89	32.40	29.18	0.190	18.39	22.42	18.73
(1.5,1.5,1.5,1.5)	6	10	0.316	42.98	52.08	48.45	0.252	28.43	39.85	35.29

2.5 Planning an Educational Study for tPA Usage in Stroke

In this section, we consider the use of the BMW design in planning an educational study to increase tPA therapy use for stroke patients as described in the Introduction. As noted there, four covariates were measured on participating institutions, and it was impossible to simultaneously obtain a balance in a matched-pair design. The simulation study in Section 4 suggests that design parameter $k = 2$ and the number of replication $M = 10$ give results that are close to optimum over a broad class of covariate distributions and confounding effects. We therefore choose these parameters in proposing a design for the tPA study.

We randomly assigned the 24 hospitals to two treatment groups, and estimated the sample-based propensity score for each hospital. The hospitals were then optimally matched into subsets with $k = 2$ which gave a minimum total distance of 2.5887. We then randomized the hospitals an additional 9 times obtaining distance measures: 2.05, 2.50, 0.20, 1.42, 0.49, 3.00, 1.14, 0.72 and 1.48. The fourth randomization produced the smallest distance. The corresponding BMW design is presented in Table 2.4, where there were 9 matched subsets with treated hospital 1 matched to control 6, treated hospital 2 and 3 jointly matched to control 8, and so on. For comparison, the data were also randomized by using a matched-pair design, where the twenty-four hospitals were matched into twelve pairs based on the two binary covariates, rural versus urban population density and low versus high stroke volume. One hospital in each pair was then randomized to treatment and one to control. `Figure 1` illustrates treatment to control group imbalance in the two continuous covariates, under the BMW and the matched-pair design.

When γ is known, we can determine the constraint k that minimizes the mean squared error when using the BMW design. Preliminary data provided estimates of the regression parameters in a logit model for the proportion of stroke cases receiving tPA as -0.63 (stroke volume), 0.02 (population density), 4.33 (percent female older

Table 2.4: Optimal Matched sample produced by the BMW design with $k = 2$ and $M = 10$ for the case study. X_1: percent of females greater than 65 years old among all females in the census tract (%); X_2: percent of males greater than 65 years old among all males in the census tract (%); X_3: stroke volume (low vs. high); X_4: population density (urban vs. rural). The estimated propensity score ($\widehat{\delta}$) was shown for each subject and the total propensity score distance $\Delta = 0.202$ for the stratum.

	Treatment Group					*Control Group*				
Strata	$ID(\widehat{\delta})$	X_1	X_2	X_3	X_4	$ID(\widehat{\delta})$	X_1	X_2	X_3	X_4
1	1 (0.33)	0.15	0.13	0	0	6 (0.35)	0.19	0.07	0	0
2	2 (0.38)	0.17	0.11	1	0	8 (0.35)	0.22	0.14	0	0
	11 (0.40)	0.22	0.14	1	0					
3	3 (0.63)	0.13	0.06	1	1	9 (0.63)	0.14	0.06	1	1
						19 (0.67)	0.25	0.15	1	1
4	4 (0.58)	0.12	0.06	0	1	12 (0.60)	0.07	0.06	1	1
5	14 (0.32)	0.13	0.07	0	0	13 (0.32)	0.13	0.09	0	0
	15 (0.31)	0.10	0.06	0	0					
6	17 (0.41)	0.24	0.12	1	0	10 (0.41)	0.26	0.18	1	0
	22 (0.43)	0.30	0.17	1	0					
7	20 (0.60)	0.08	0.06	1	1	16 (0.61)	0.10	0.07	1	1
						18 (0.61)	0.09	0.05	1	1
8	21 (0.60)	0.18	0.14	0	1	5 (0.61)	0.19	0.13	0	1
9	24 (0.62)	0.23	0.16	0	1	7 (0.62)	0.24	0.19	0	1
						23 (0.62)	0.11	0.07	1	1

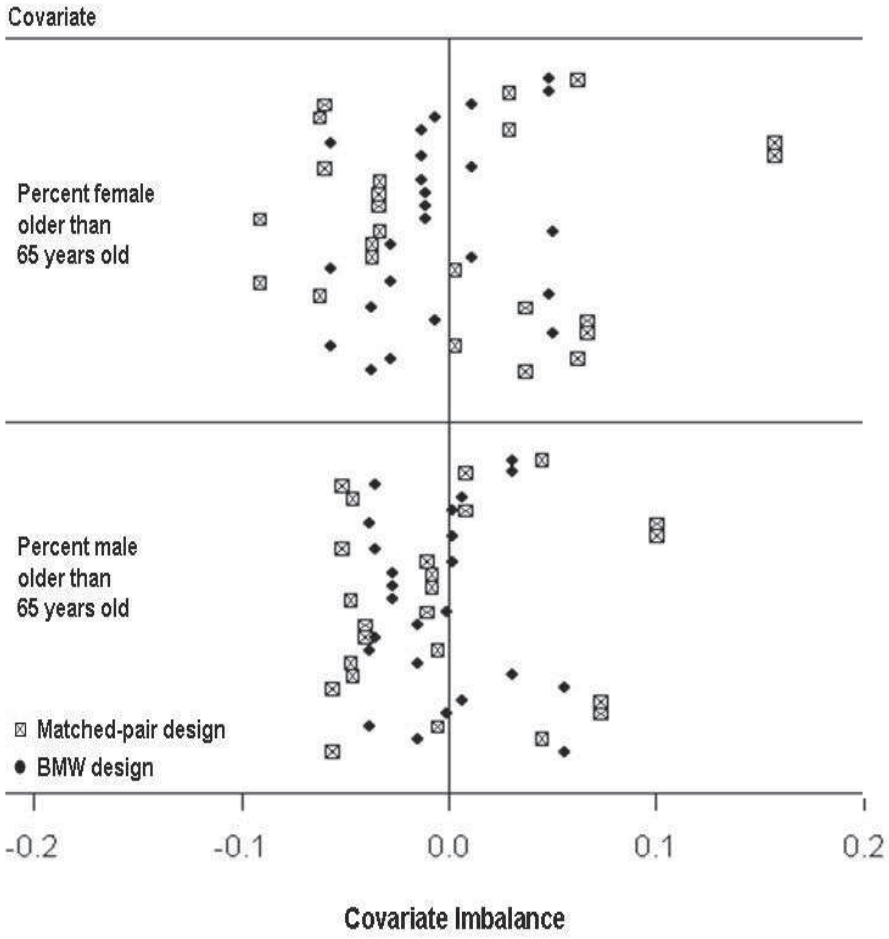


Figure 2.1: Covariate Imbalances from the matched-pair design (matching on the categorical covariates: Population density and stroke volume) and the BMW design. The imbalance value in covariate X for unit i was computed as $Imbalance(X_i) = \sum_{j \in T_s} X_j/|T_s| - \sum_{k \in C_s} X_k/|C_s| = \overline{X}_{T_s} - \overline{X}_{C_s}$ where s is the stratum that unit i belongs to.

than 65) and -1.23 (percent male older than 65). Since there are 24 hospitals, k can take values from 1 to 11. For k=1, $M = 10$ randomizations gave a minimum distance of 0.2936. We then repeated the above process with the same randomized samples but with constraints k=2, 3, ..., 11 and for each k, searched for the optimal sample with minimal distance. Third, based on the approximate value of γ above, we computed the MSE from (2.11) as 0.1076, 0.1045 and 0.1114 for the optimal sample with constraint $k = 1, 2, 3$. This suggests that pair matching and matching with constraint $k = 2$ achieve approximately the same level of optimality in terms of minimizing MSE. Compared with the matched-pair design described above, the BMW design reduced the MSE of the treatment effect estimator by 42%.

2.6 Discussion

The BMW design is, in essence, applying the optimal full matching with constraints technique to randomization in order to achieve overall balance between treatment groups and control the variance of the treatment comparison and so yield good MSE properties. One of the virtues of this design is that it will not only reduce the chance imbalance in observed covariates but also preserve the advantage of traditional randomized designs in balancing the unobserved covariates on average. Although only partial balance on the observed covariates is achieved by the BMW design, it is substantially better than the balance obtained by random assignment of treatments. When there is considerable confounding in small studies, this improvement in balance can result in a substantial decrease of mean squared error in the treatment effect estimator.

The BMW design can be revised to allow the user to select other criteria besides MSE to compromise between bias and variance. If variance of the estimator is not a concern, one can modify this design to achieve optimal balance and so reduce conditional bias (i.e. set $k = N/2 - 1$). On the other hand, if the objective is

to minimize variance, optimal pair matching with constraint k=1 is the best full matching choice.

We recommend use of a super-population model for analysis, and this is the basis of the simulation comparisons that we have made. It is worth noting, however, that the BMW design with a practically reasonable choice of M (e.g. $M = 10$, 20 or 100), can also form the basis of a randomization test. Suppose, for example, that a sample has been collected using the BMW design with given k and M and the value of the test statistic (e.g. t statistic) has been computed. We now repeat the BMW design with the same k and M a large number B of times and each time compute the test statistic based on the fixed outcomes observed. This would lead to a randomization test and confidence intervals following standard methods. Typically, the underlying reference set of this test is reasonably large, for example, say $N = 30$, we will end up with $\binom{30}{15} = 155117520$ possible designs and this number is fairly large even for $M = 100$.

The model-based approach of adjusting for the estimated propensity score and the Robins-Mark-Newey E-estimation procedure could be considered as alternatives to the BMW design. Our simulation studies suggest that, when the propensity score model is *appropriately* specified, the BMW design is more efficient than the model-based approach when the confounding effects are relatively small; the model based approach, however, becomes more efficient than the BMW design when the confounding effects increase. On the other hand, when the propensity score model is *inappropriately* specified, the BMW design achieves substantial gain over the model-based approach. In the context considered in this chapter, the E-estimation procedure is the least efficient and robust.

In practice, investigators may ask which covariates should be adjusted for in matching. This decision is difficult to make before randomization or the outcomes become available. In some instances, investigators may know with certainty which co-

variates will affect the outcomes measured later based on their knowledge or scientific consensus. However, on other circumstances, the prior information is unavailable. We carried out a number of simulations to evaluate the MSE performance of the BMW design under the scenario that there is one true confounder out of the four potential confounding variables used in the propensity score model. Let Y_i, i = 1,2,...,N, represent responses of the unit i, conditional on a given treatment assignment T, C and X,

$$Y_i = \alpha + \beta I(i \in T) + \gamma_1 X_{i1} + \varepsilon_i; \tag{2.19}$$

and

$$\delta_i = Pr(Z = 1 \mid X_i; \alpha) = \exp(\alpha_1 + \sum_{j=1}^{4} \alpha_j X_{ij}) / \{1 + \exp(\alpha_1 + \sum_{j=1}^{4} \alpha_j X_{ij})\} \tag{2.20}$$

The simulations results summarized in Table 2.5 reveal that the BMW design remains effective in reducing the MSE of the treatment effects estimators even though it matches on the propensity scores estimated based on only one true confounding variable out of four potential ones.

Table 2.5: Percent reductions in the MSE of treatment effect estimator for the BMW design compared to a completely randomized design (CR) when the BMW design adjusts for one true confounding variable and three false ones. Sample size N=30 subjects. Number of replications=1000.

γ	$\sum_{j=1}^{4} \gamma_j$	M	MSE (CR)	MSE Percent Reduction(%) (BMW vs. CR Design)		
				$k=1$	$k=2$	$k=3$
$X_1, X_2, X_3, X_4 \overset{i.i.d}{\sim} Bernoulli(0.5)$						
0.5	0.5	10	0.146	12.05	5.88	-0.84
1.0	1.0	10	0.173	20.44	13.30	8.85
1.5	1.5	10	0.219	32.45	35.20	29.43

Greevy et al. (2004) proposed another multivariate matching design based on Mahalanobis distance. This approach searches for the optimal multivariate non-bipartite matching followed by randomization within pairs. We also investigated this in a simulation study presented in Table 2.6. As the confounding effects increase, or the number of covariates increase, the BMW design becomes much more effective in reducing MSE compared to Greevy's design. This may be because the Mahalanobis distance is inferior to propensity scores when there are many covariates.

In general terms, the BMW design appears to provide a viable approach in the context of small studies where adjustment for randomization imbalance may be important. Furthermore, the simplicity of this matching-based design allows researchers to perform simple stratified analyses that adjust for imbalance in the randomization, which is appealing.

Finally, simulation shows that the BMW design can substantially reduce the MSE of the treatment effect estimate, as compared to the existing randomized designs in linear models. These investigations could be extended to other regression models, such as the class of general linear models. It should also be noted that the BMW design can be generalized to clinical trials with more than two treatment arms. Baseline category logit model can be used to estimate the probability of a subject being assigned to each treatment arm, and Euclidean distance can be used to measure the quality of a matching.

Table 2.6: Percent reductions in the MSE of treatment effect estimator for the BMW design compared to multivariate non-bipartite matching design (NB). Number of replications=1000.

γ	$\sum_{j=1}^{8}\gamma_j$	M	MSE (NB Design)	MSE Percent Reduction(%) (BMW vs. NB Design)		
				$k=1$	$k=2$	$k=3$
$X_1, X_2, X_3, X_4 \overset{i.i.d}{\sim} Bernoulli(0.5)$						
(0.5,0.5,0.5,0.5)	2	5	0.146	-0.07	-2.27	-6.11
		10		2.47	-0.55	-5.90
		20		5.91	1.44	-3.91
(1.0,1.0,1.0,1.0)	4	5	0.185	2.42	14.49	8.53
		10		9.62	15.79	11.68
		20		24.78	22.18	18.44
(1.5,1.5,1.5,1.5)	6	5	0.250	1.77	30.92	24.36
		10		14.01	32.12	26.28
		20		25.24	34.40	29.20
$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8 \overset{i.i.d}{\sim} Bernoulli(0.5)$						
(0.5,0.5,0.5,0.5, 0.5,0.5,0.5,0.5)	4	5	0.156	-8.15	0.51	-6.35
		10		-6.41	0.96	-5.13
		20		-1.15	2.18	-5.39
(1.0,1.0,1.0,1.0, 1.0,1.0,1.0,1.0)	8	5	0.222	-25.19	16.39	16.07
		10		-12.76	22.92	17.65
		20		0.26	25.53	19.59
(1.5,1.5,1.5,1.5, 1.5,1.5,1.5,1.5)	12	5	0.338	-39.10	29.01	32.47
		10		-14.16	39.06	35.22
		20		-1.31	42.46	36.37

CHAPTER III

More on Propensity Score Matching in Randomized Clinical Trials

Cluster randomization trials with relatively few clusters have been widely used in recent years for evaluation of health-care strategies. The balance match weighted (BMW) design that was introduced in Chapter II applies the optimal full matching with constraints technique to a prospective randomized design, and aims to minimize the mean squared error (MSE) of the treatment effect estimator. A simulation study shows that, under various confounding scenarios, the BMW design often has superior performance over the completely randomized design, the matched-pair design, the model based approach adjusting for the estimated propensity score, the Robins-Mark-Newey E-estimation procedure and the Greevy et al. (2004) optimal multivariate matched design before randomization, in terms of the MSE reduction of treatment effect estimator. In this chapter, we aim to extend the BMW design to two directions: clinical trials with more than two arms and clinical trials with staggered entry. In addition, we investigate the effects of further increasing M on the MSE performance of the two-arm BMW design.

3.1 The BMW design on trials with three or more arms

The last decade has seen a broad surge of interest in using the method of propensity score matching to estimate the average treatment effect based on observational data. In most observational studies, a treatment group is matched with a single control group. Although costly or difficult, some investigators use a second control group in an effort to detect the hidden biases in the unobserved covariates (Seltser and Sartwell, 1965; Weston and Mansinghka, 1971; Roghmann and Sodeur, 1972; Zabin et al., 1989; Chang et al., 1997; Wells et al., 1997; Bo and Rosenbaum, 2004). As argued by Campbell (2009), although matching can adjust the differences in observed covariates, bias may still exist due to some unobserved covariates and if that is the case, the two control groups may differ from each other substantially on the unobserved covariate. The use of the second control group, if carefully selected, would help reduce possible hidden bias and strengthen the evidence that the observed effects are caused by the treatment.

Although the use of two control groups may not be relevant in randomized experiments, a more complex framework also appears to be useful when evaluating certain treatment programs in randomized clinical trials. For example, a drug may be applied at different dosage levels or a physician may have more than two treatment options to evaluate. In this cases, the BMW propensity score matching method, which only involves two matching groups, appears to be inadequate and extensions become necessary.

The problem of matching with three groups has received some attention in the area of graph theory and observational studies in Epidemiology. It has been shown that the problem of finding an optimal tripartite matching is a NP (nondeterministic polynomial time) complete problem. The most notable characteristic of an NP complete problem is that the time required to solve the problem using any currently available algorithm increases very quickly as the size of the problem grows, and in

many instances, it is not possible to determine the optimal solution with costly computational facilities. For example, consider a small study which contains three arms of m subjects in each arm, the number of comparisons required in search for the optimal matched triples is 36 when $m = 3$, 576 when $m = 4$, 14400 when $m = 5$, 518400 when $m = 6$, and $1.316819e^{13}$ when m gets to 10. This has implications for propensity based matching and an approach such as BMW since the corresponding optimization is NP hard.

In order to circumvent this problem, we can work with some ad hoc approaches which may not lead to the optimal tripartite matching, but to the solutions close to the optimal. Bo and Rosenbaum (2004) propose an algorithm to match three groups into incomplete blocks with disjoint pairs, and develop a search algorithm based on the method of finding the optimal nonbipartite matching. In their approach, three groups are optimally matched into pairs and an incomplete block of size two is formed. One potential drawback of the incomplete block design is that, when comparing treatment A and B, for example, only those pairs that receive treatments A and B are included in the direct comparison. The corresponding treatment effects estimator is $\overline{y}_{\mathcal{A}} - \overline{y}_{\mathcal{B}}$. This leads to a loss of efficiency which becomes even more apparent in small studies. We can reduce the efficiency loss by including the rest of the pairs in the comparison, and propose an estimator as $\frac{2}{3}(\overline{y}_{\mathcal{A}} - \overline{y}_{\mathcal{B}}) + \frac{1}{3}[(\overline{y}_{\mathcal{A}} - \overline{y}_{\mathcal{C}}) + (\overline{y}_{\mathcal{C}} - \overline{y}_{\mathcal{B}})]$. But these estimators are still considerably less efficient compared to those obtained from matching involving triples or blocks of size three. In this section, we propose two new algorithms of matching with three groups: the two-way tripartite matching and the three-way tripartite matching. We suggest the first approach when there is a clear predefined reference group, a control group, to which the other two arms are compared and suggest the second approach when all three groups are to be compared simultaneously. In these designs, three groups are matched into triples. Although the balance achieved by using our proposed matching algorithms may not be optimal,

it would be close to the optimal, and certainly can be substantially better than the balance obtained by random assignment of treatments. We investigate the use of these three approaches in the generalization of the BMW design and use a simulation study to evaluate the performance of the design under various confounding scenarios.

The rest of the section is organized as follows. Notation, the three matching algorithms (incomplete block of size two, two-way and three-way tripartite matching algorithms) and the analysis models are presented in section 3.1.1. The BMW design on clinical trials with three arms based on each of these three matching algorithms is outlined in section 3.1.3 and section 3.1.4 gives results of a simulation study comparing the performance of the BMW design based on different matching algorithms as well as with the completely randomized design. This section concludes with discussion in section 3.1.5.

3.1.1 Methods

In this subsection, we present the notation and problem formulation as well as introduce the three matching algorithms, incomplete block with disjoint pairs (Bo and Rosenbaum, 2004), tripartite matching given predefined reference group and with reference group optimally selected, respectively.

Consider a study with the aim of assessing the effects of three different treatments, A, B and C. Let N denote the number of subjects available for the study. We assume that N is a multiple of 6 and N/3 subjects are randomized to each of the treatment groups. Thus, we suppose that a randomization process divides the N subjects into a set $\mathcal{A}$ of N/3 subjects to be treated with A, a set $\mathcal{B}$ of N/3 subjects to receive treatment B and a set $\mathcal{C}$ of the remaining N/3 subjects for treatment C, where $\mathcal{A} = \{\eta_1^A, ..., \eta_{N/3}^A\}, \mathcal{B} = \{\eta_1^B, ..., \eta_{N/3}^B\}, \mathcal{C} = \{\eta_1^C, ..., \eta_{N/3}^C\}$. We also assume that a vector of r covariates, $\mathbf{X} = (X_1, X_2, ..., X_r)$ with $X_1 = 1$, is observed for each individual.

Similarity of covariates for the given randomization is measured through an estimated probability of being assigned to each group. Writing $Z = 1, 2, 3$ for the subjects who received treatment A, B, C, respectively, the (estimated) Euclidean distance between the subject i in group $\mathcal{A}$ and subject j in group $\mathcal{B}$ is given by,

$$\delta\{(\eta_i^A, \eta_j^B)\} = \sqrt{(\hat{\delta}_{1,i}^A - \hat{\delta}_{1,j}^B)^2 + (\hat{\delta}_{2,i}^A - \hat{\delta}_{2,j}^B)^2 + (\hat{\delta}_{3,i}^A - \hat{\delta}_{3,j}^B)^2} \tag{3.1}$$

where $\hat{\boldsymbol{\delta}}_i^G = (\hat{\delta}_{1,i}^G, \hat{\delta}_{2,i}^G, \hat{\delta}_{3,i}^G)$ with $G = (\mathcal{A}, \mathcal{B}, \mathcal{C})$ is the estimated probability of subject i in group G being assigned to the treatment group $\mathcal{A}, \mathcal{B}, \mathcal{C}$, respectively, which can be obtained from a model such as the baseline category model

$$\delta_{t,i} = Pr(Z = t \mid \mathbf{X}_i; \boldsymbol{\alpha}_t) = \exp\{\boldsymbol{\alpha}_t \mathbf{X}_i^T\}/(1 + \exp\{\boldsymbol{\alpha}_1 \mathbf{X}_i^T\} + \exp\{\boldsymbol{\alpha}_2 \mathbf{X}_i^T\}) \tag{3.2}$$

where $t = 1, 2$ and $\boldsymbol{\alpha}_1 = (\alpha_{11}, ..., \alpha_{1r})$ and $\boldsymbol{\alpha}_2 = (\alpha_{21}, ..., \alpha_{2r})$ are regression coefficients and $\boldsymbol{\alpha}_3 \equiv 0$ since in this parameterization the third group is regarded as the reference.

According to various questions of interest, there are three approaches proposed for the matching problem with three groups.

3.1.1.1 Incomplete block design with disjoint pairs

If one is interested in comparing the treatment effect A with B, A with C and B with C, and the three comparisons are equally important, then Bo and Rosenbaum (2004) argue that the comparisons of three groups not be done in matched triples, but rather as an incomplete block design with matched pairs; thus they propose blocks of size two with one third of the blocks assigned at random to each of the three treatment comparisons. Given sets $\mathcal{A}, \mathcal{B}, \mathcal{C}$ which contains the subjects to be treated by A, B, C, respectively, we consider the collection $P_{\mathcal{A},\mathcal{B},\mathcal{C}}$ of all possible matchings with size (p_{12}, p_{13}, p_{23}), where a matching of size (p_{12}, p_{13}, p_{23}) corresponds to a collection of p_{12} pairs of the form (η_i^A, η_j^B), p_{13} pairs of the form (η_i^A, η_k^C) and p_{23} pairs of the form

(η_j^B, η_k^C), where all the pairs are disjoint and p_{12}, p_{13}, p_{23} are specified constants. Let $\omega \in \mathcal{M}$ be one pair of a particular matching $\mathcal{M} \in P_{\mathcal{A},\mathcal{B},\mathcal{C}}$, we measure the quality of $\mathcal{M}$ as

$$\Delta_{\mathcal{M}} = \sum_{\omega \in \mathcal{M}} \delta(\omega) \tag{3.3}$$

Thus $\Delta_{\mathcal{M}}$ is the total distance within pairs and an optimal tripartite matching minimizes $\Delta_{\mathcal{M}}$ over all $\mathcal{M} \in P_{\mathcal{A},\mathcal{B},\mathcal{C}}$. If there exists such a matching P of size (p_{12}, p_{13}, p_{23}), such that $\Delta(P) < \infty$ then the optimal matching problem is *feasible*; otherwise, it is *infeasible*. In our application of a randomized experiment, $p_{12} = p_{13} = p_{23} = N/6$.

The optimal tripartite matching problem can be transformed to an equivalent nonbipartite matching problem and these two problems were shown to have the same optimal solutions (Bo and Rosenbaum, 2004). Given a single set $\Theta = \mathcal{A} \bigcup \mathcal{B} \bigcup \mathcal{C} = (\eta_1^A, ..., \eta_{N/3}^A, \eta_1^B, ..., \eta_{N/3}^B, \eta_1^C, ...\eta_{N/3}^C)$, and let Φ denote the collection of all possible pair matchings within Θ, where a matching corresponds to $N/2$ matched pairs. The distance can be defined as follows:

$$\delta\{(\eta_i^m, \eta_j^n)\} = \begin{cases} \sqrt{(\hat{\delta}_{1,i}^m - \hat{\delta}_{1,j}^n)^2 + (\hat{\delta}_{2,i}^m - \hat{\delta}_{2,j}^n)^2 + (\hat{\delta}_{3,i}^m - \hat{\delta}_{3,j}^n)^2} & \text{if } m \neq n; \\ +\infty & \text{if } m = n. \end{cases}$$

Where $m, n \in \{A, B, C\}$. Let $\xi \in \mathcal{M}_\Theta$ be one pair formed in a matching $\mathcal{M}_\Theta \in \Phi$, then the total distance of $\mathcal{M}_\Theta$ can be written as

$$\Delta_{\mathcal{M}_\Theta} = \sum_{\xi \in \mathcal{M}_\Theta} \delta(\xi) \tag{3.4}$$

The optimal nonbipartite matching problem is to minimize $\Delta_{\mathcal{M}_\Theta}$ over Φ. A nonbipartite matching $\mathcal{P}$ is called *feasible*, if $\Delta_{\mathcal{P}} < \infty$. Bo and Rosenbaum (2004) (*Claim 1*) proved that P is an optimal nonbipartite matching with $\Delta(P) < +\infty$ if and only if P is also an optimal, feasible tripartite matching into incomplete blocks as described above. Figure 3.1 illustrates how three groups of size 4 each can be matched into in-

complete block design of disjoint pairs through the transformation of a nonbipartite matching.

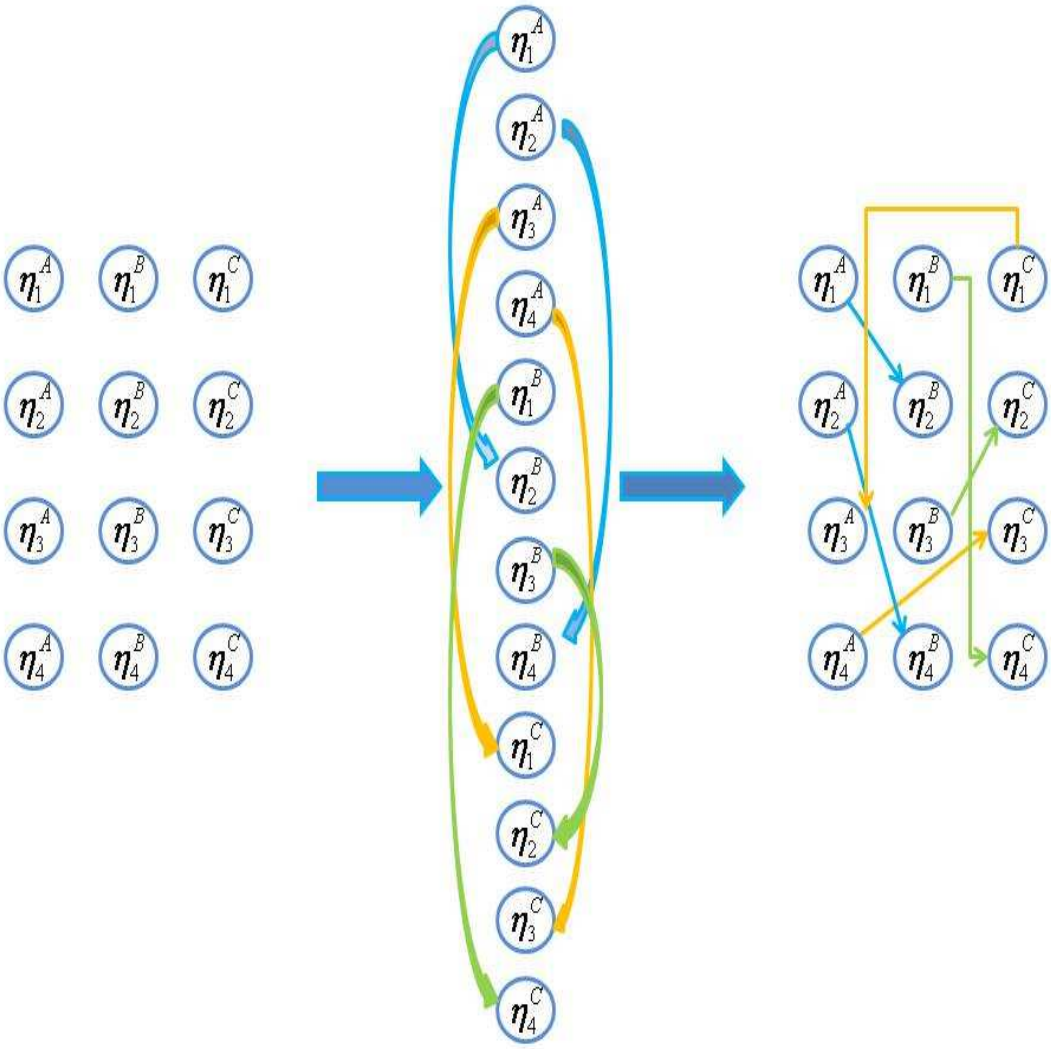


Figure 3.1: The Transformation from a Nonbipartite Matching to an Incomplete Block Design of Size Two.

3.1.1.2 Three-way Tripartite Matching With Triples

We propose an alternative approach of matching three groups when the comparisons between A v.s. B, A v.s. C and B v.s. C are equally important. Given sets $\mathcal{A}, \mathcal{B}, \mathcal{C}$ which contain the subjects to be treated by A, B, C, respectively, we consider

the collection $M_{\mathcal{A},\mathcal{B}}$ of all possible matchings with size m_{12}, $M_{\mathcal{A},\mathcal{C}}$ of matchings with size m_{13}, and $M_{\mathcal{B},\mathcal{C}}$ of matchings with size m_{23}, where a matching of size m_{12}, m_{13}, m_{23} corresponds to a collection of m_{12} pairs of the form (η_i^A, η_j^B), m_{13} pairs of the form (η_i^A, η_k^C) and m_{23} pairs of the form (η_j^B, η_k^C), respectively. The pairs of form (η_i^A, η_j^B) and form (η_i^A, η_k^C) share the common subjects in group A, similarly for the matched pairs between other groups. The size m_{12}, m_{13}, m_{23} is fixed and we consider $m_{12} = m_{13} = m_{23} = N/3$. Figure 3.2 illustrates an example of tripartite matching with three groups of size 4 each and $m_{12} = m_{13} = m_{23} = 4$.

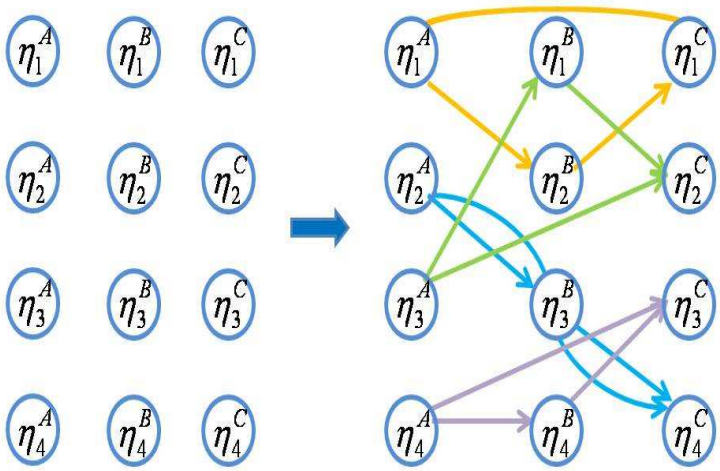


Figure 3.2: Tripartite Matching of triples for three groups of size four each, where $m_{12} = m_{13} = m_{23} = 4$.

Let $\omega \in \mathcal{M}_{\mathcal{A},\mathcal{B}}$ be one pair of a particular matching $\mathcal{M}_{\mathcal{A},\mathcal{B}} \in M_{\mathcal{A},\mathcal{B}}$, we measure the quality of $\mathcal{M}_{\mathcal{A},\mathcal{B}}$ as

$$\Delta_{\mathcal{M}_{\mathcal{A},\mathcal{B}}} = \sum_{\omega \in \mathcal{M}_{\mathcal{A},\mathcal{B}}} \delta(\omega) \tag{3.5}$$

An optimal pair matching corresponds to the minimum distance measure $\Delta^*_{\mathcal{M}_{\mathcal{A},\mathcal{B}}} =$

$\min(\Delta_{\mathcal{M}_{\mathcal{A},\mathcal{B}}})$ over $M_{\mathcal{A},\mathcal{B}}$ with size m_{12}. Similarly, define $\Delta_{\mathcal{M}_{\mathcal{A},\mathcal{C}}}$ and $\Delta_{\mathcal{M}_{\mathcal{B},\mathcal{C}}}$ as follows

$$\Delta_{\mathcal{M}_{\mathcal{A},\mathcal{C}}} = \sum_{\omega \in \mathcal{M}_{\mathcal{A},\mathcal{C}}} \delta(\omega) \tag{3.6}$$

and

$$\Delta_{\mathcal{M}_{\mathcal{B},\mathcal{C}}} = \sum_{\omega \in \mathcal{M}_{\mathcal{B},\mathcal{C}}} \delta(\omega) \tag{3.7}$$

And the optimal pair matched samples which minimize the distance $\Delta_{\mathcal{M}_{\mathcal{A},\mathcal{C}}}$ and $\Delta_{\mathcal{M}_{\mathcal{B},\mathcal{C}}}$ give the minimum distances $\Delta^*_{\mathcal{M}_{\mathcal{A},\mathcal{C}}}$ and $\Delta^*_{\mathcal{M}_{\mathcal{B},\mathcal{C}}}$, respectively.

Once the treatment group A is optimally matched to B, and B to C, the corresponding members of A and C are also paired through their individual matchings with B. Let $\mathcal{M}^+_{\mathcal{A},\mathcal{C}}$ represent this implied matching. It follows that the minimum total distance measure given group B as the reference group is

$$\Delta^*_{\mathcal{M}_{\mathcal{B}}} = \Delta^*_{\mathcal{M}_{\mathcal{A},\mathcal{B}}} + \Delta^*_{\mathcal{M}_{\mathcal{B},\mathcal{C}}} + \sum_{\omega \in \mathcal{M}^+_{\mathcal{A},\mathcal{C}}} \delta(\omega) \tag{3.8}$$

Similarly, the minimum distance measure with groups A and C as the reference groups are,

$$\Delta^*_{\mathcal{M}_{\mathcal{A}}} = \Delta^*_{\mathcal{M}_{\mathcal{A},\mathcal{C}}} + \Delta^*_{\mathcal{M}_{\mathcal{A},\mathcal{B}}} + \sum_{\omega \in \mathcal{M}^+_{\mathcal{B},\mathcal{C}}} \delta(\omega) \tag{3.9}$$

$$\Delta^*_{\mathcal{M}_{\mathcal{C}}} = \Delta^*_{\mathcal{M}_{\mathcal{B},\mathcal{C}}} + \Delta^*_{\mathcal{M}_{\mathcal{A},\mathcal{C}}} + \sum_{\omega \in \mathcal{M}^+_{\mathcal{A},\mathcal{B}}} \delta(\omega) \tag{3.10}$$

The reference group associated with the smallest total distance $\Delta^*_{\mathcal{M}_{\mathcal{A},\mathcal{B},\mathcal{C}}} = \min($ $\Delta^*_{\mathcal{M}_{\mathcal{A}}}, \Delta^*_{\mathcal{M}_{\mathcal{B}}}, \Delta^*_{\mathcal{M}_{\mathcal{C}}})$ is called the *optimal reference group*. The two optimal pair matched samples, with that reference group as their common group, minimize the sum of three pairwise distances. This is the solution for the three-way tripartite matching design.

3.1.1.3 Two-way Tripartite Matching With Triples

As sometimes the case in practice, investigators may have more than two treatment options to compare against a common control. Given the reference group, the matching mechanism can focus on adjusting for the covariate differences between each treatment group and the control rather than the comparison between the two treatment groups themselves.

Given the same sets $\mathcal{A}, \mathcal{B}, \mathcal{C}$ which contain the subjects to be treated by A, B, C, respectively, we assume that the treatment group A is the reference group. The optimal pair matching which minimizes the distance measure $\Delta_{\mathcal{M}_{\mathcal{A},\mathcal{B}}}$ over $M_{\mathcal{A},\mathcal{B}}$ with size m_{12} gives the smallest distance measure between group A and B as $\Delta^*_{\mathcal{M}_{\mathcal{A},\mathcal{B}}}$. Similarly, we can obtain the optimal matched sample corresponding to the minimum distance measure between group A and C as $\Delta^*_{\mathcal{M}_{\mathcal{A},\mathcal{C}}}$.

The combination of the two optimal pair matched samples leads to the optimal solution to the two-way tripartite matching design given the predefined reference group A, and the corresponding total distance measure

$$\Delta^*_{\mathcal{A}} = \Delta^*_{\mathcal{M}_{\mathcal{A},\mathcal{B}}} + \Delta^*_{\mathcal{M}_{\mathcal{A},\mathcal{C}}} \tag{3.11}$$

Worthy of noting, the distance between the corresponding members of B and C, $\sum_{\omega \in \mathcal{M}^+_{\mathcal{B},\mathcal{C}}} \delta(\omega)$, is not included in 3.11 since the adjustment of the covariate imbalance between two treatment groups B and C is not of the primary interest given the predefined reference group A.

3.1.2 Model

To appreciate the effect of treatment on response in a pooled sample and matched sample, respectively, consider the following model: Let Y_i, i = 1,2,...,N, represent

responses of the unit i, conditional on a given treatment assignment A, B, C and X,

$$Y_i = \alpha + \beta_1 I(i \in \mathcal{A}) + \beta_2 I(i \in \mathcal{B}) + \sum_{j=1}^{r} \gamma_j X_{ij} + \varepsilon_i; \tag{3.12}$$

where $I(.)$ is the indicator function, β_1, β_2 denote the true treatment effect A versus C, B versus C, respectively. $\gamma_1, \gamma_2, ..., \gamma_r$ are the confounding effects and $\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_N)$ is the vector of the measurement errors with $E[\varepsilon|\mathcal{A}, \mathcal{B}, \mathcal{C}, X] = 0$, $\text{Var}[\varepsilon|\mathcal{A}, \mathcal{B}, \mathcal{C}, X] = \sigma^2 I$, $\sigma^2 < +\infty$ and I is the $N \times N$ identity matrix.

1. *Pooled Sample.* Under model (3.12), the common treatment effect estimators based on the unstratified pooled sample obtained from the completely randomized design is $\widehat{\beta}_{1,pool} = \overline{y}_{\mathcal{A}} - \overline{y}_{\mathcal{C}}$, $\widehat{\beta}_{2,pool} = \overline{y}_{\mathcal{B}} - \overline{y}_{\mathcal{C}}$ and $\widehat{\beta}_{3,pool} = \overline{y}_{\mathcal{A}} - \overline{y}_{\mathcal{B}}$, respectively, which has conditional expectation

$$E[\widehat{\beta}_{i,pool}|T_1, T_2, X] = \beta_i + \sum_{j=1}^{r} \gamma_j (\overline{X}_{jT_1} - \overline{X}_{jT_2}) \tag{3.13}$$

where $i = 1, 2, 3$ and $T_1, T_2 \in \{\mathcal{A}, \mathcal{B}, \mathcal{C}\}$. If $i = 1$, then $T_1 = \mathcal{A}, T_2 = \mathcal{C}$, and so on. The mean squared error for $\widehat{\beta}_{i,pool}$ where $i = 1, 2$ (conditional on $\mathcal{A}, \mathcal{B}, \mathcal{C}$ and X) is

$$MSE(\widehat{\beta}_{i,pool}|T_1, T_2, X) = \{\sum_{j=1}^{r} \gamma_j (\overline{X}_{jT_1} - \overline{X}_{jT_2})\}^2 + 6\sigma^2/N \tag{3.14}$$

2. *Matched Sample.* Under model (3.12), estimating the treatment effect for the matched sample involves computation of the average of the within-pair differences. For example, the incomplete block design (ICB) with disjoint pairs results in $p_{13} = N/6$ pairs of A-treated subjects matched to C-treated patients, $p_{23} = N/6$ pairs of B-treated subjects matched to C-treated patients and $p_{12} = N/6$ pairs of A-treated subjects matched to B-treated patients, respec-

tively. The corresponding treatment effect estimator of β_1^{ICB} comparing treatment A v.s. C is $\widehat{\beta}_1^{ICB} = \frac{2}{3}\sum_{i=1}^{N/6}\{y_{i,\mathcal{A}} - y_{i,\mathcal{C}}\}/(N/6) + \frac{1}{3}[\sum_{i=1}^{N/6}\{y_{i,\mathcal{A}} - y_{i,\mathcal{B}}\}/(N/6) - \sum_{i=1}^{N/6}\{y_{i,\mathcal{B}} - y_{i,\mathcal{C}}\}/(N/6)] = \frac{2}{3}(\overline{y}_{T_\mathcal{A}} - \overline{y}_{T_\mathcal{C}}) + \frac{1}{3}[(\overline{y}_{T_\mathcal{A}} - \overline{y}_{T_\mathcal{B}}) - (\overline{y}_{T_\mathcal{B}} - \overline{y}_{T_\mathcal{C}})]$, where $T_\mathcal{A}, T_\mathcal{B}$ and $T_\mathcal{C}$ refer to the set of $N/6$ subjects treated by A, B and C, respectively. $\widehat{\beta}_1^{ICB}$ has conditional expectation

$$\begin{aligned} E[\widehat{\beta}_1^{ICB}|T_\mathcal{A}, T_\mathcal{B}, T_\mathcal{C}, X] &= \beta + \frac{2}{3}\sum_{i=1}^{N/6}\{\sum_{j=1}^{r}\gamma_j(X_{jT_{i,\mathcal{A}}} - X_{jT_{i,\mathcal{C}}})\}/(N/6) \\ &\quad + \frac{1}{3}[(\sum_{i=1}^{N/6}\{\sum_{j=1}^{r}\gamma_j(X_{jT_{i,\mathcal{A}}} - X_{jT_{i,\mathcal{B}}})\}/(N/6)) \\ &\quad - (\sum_{i=1}^{N/6}\{\sum_{j=1}^{r}\gamma_j(X_{jT_{i,\mathcal{B}}} - X_{jT_{i,\mathcal{C}}})\}/(N/6))] \end{aligned}$$

The mean squared error for $\widehat{\beta}_{1,AC}$ (conditional on $T_\mathcal{A}, T_\mathcal{B}, T_\mathcal{C}$ and X) is

$$\begin{aligned} MSE(\widehat{\beta}_1^{ICB}|T_\mathcal{A}, T_\mathcal{B}, T_\mathcal{C}, X) &= \{\frac{2}{3}\sum_{i=1}^{N/6}\{\sum_{j=1}^{r}\gamma_j(X_{jT_{i,\mathcal{A}}} - X_{jT_{i,\mathcal{C}}})\}/(N/6) \\ &\quad + \frac{1}{3}[(\sum_{i=1}^{N/6}\{\sum_{j=1}^{r}\gamma_j(X_{jT_{i,\mathcal{A}}} - X_{jT_{i,\mathcal{B}}})\}/(N/6)) \\ &\quad - (\sum_{i=1}^{N/6}\{\sum_{j=1}^{r}\gamma_j(X_{jT_{i,\mathcal{B}}} - X_{jT_{i,\mathcal{C}}})\}/(N/6))]\}^2 + 8\sigma^2/N \end{aligned}$$

The treatment effect estimators of β_2^{ICB}, β_3^{ICB} comparing treatment B v.s. C and A v.s. B can be defined in the similar manner.

On the other hand, the optimal tripartite matching design (TM) with triples leads to $m_{13} = N/3$ pairs of A-treated subjects matched to C-treated patients, $m_{23} = N/3$ pairs of B-treated subjects matched to C-treated patients and $m_{12} = N/3$ pairs of A-treated subjects matched to B-treated patients. The treatment effect estimators comparing A v.s. C from the two-way and three-

way designs are β_1^{2TM} and β_1^{3TM}, respectively, where $\widehat{\beta}_1^{2TM} = \widehat{\beta}_1^{3TM} = \sum_{i=1}^{N/3}\{y_{i,\mathcal{A}} - y_{i,\mathcal{C}}\}/(N/3) = \overline{y}_{S_\mathcal{A}} - \overline{y}_{S_\mathcal{C}}$, and $S_\mathcal{A}, S_\mathcal{C}$ refers to the set of $N/3$ subjects treated by A and C. This estimator has the mean squared error

$$MSE(\widehat{\beta}_1^{kTM}|S_\mathcal{A}, S_\mathcal{C}, X) = [\sum_{i=1}^{N/3}\{\sum_{j=1}^{r}\gamma_j(X_{jS_{i,\mathcal{A}}} - X_{jS_{i,\mathcal{C}}})\}/(N/3)]^2 + 6\sigma^2/N \tag{3.15}$$

where $k = 2, 3$. The conditional bias and MSE for estimators of treatment effects β_2^{kTM} and β_3^{kTM} comparing A v.s. B, B v.s. C can be defined similarly.

3.1.3 The BMW design on trials with three arms

In this section, we propose three balance match weighted (BMW) designs for clinical trials with three arms. These designs with specified parameter M are defined algorithmically as follows:

Step 1. Randomize 1/3, 1/3 and 1/3 of the subjects to the treatment group $\mathcal{A}, \mathcal{B}$ and $\mathcal{C}$, respectively;

Step 2. Compute the estimated probability of being assigned to each treatment group using the baseline category model (3.2) or other multinomial models and create the $|N| \times |N|$ matrix of estimated Euclidean distances;

Step 3. Obtain the optimal matched samples based on a tripartite matching algorithm, (*i*) incomplete block design with disjoint pairs, (*ii*) two-way tripartite matching design or (*iii*) three-way tripartite matching design as described in section 3.1.1, and record the minimum total distance Δ for the given randomization.

Step 4. Repeat *Steps* 1 to 3 M times; choose the randomized sample with minimum total distance $\Delta^* = \min(\Delta_1, \Delta_2, ..., \Delta_M)$.

3.1.4 Simulation Results

In order to assess the performance of the generalized BMW design based on each of these three matching algorithms, we carried out a simulation study to compare each of them with a completely randomized design and the results are presented below. In doing so, we considered a wide variety of settings and, for each setting, estimated the mean squared error based on 1000 replications.

3.1.4.1 Structure of the simulation

For each of N subjects, we generated a set of r covariates $X_1, X_2, ..., X_r$, where the covariates were drawn independently from various distributions as described below. Given a randomization of subjects to the two treatment groups, the responses were generated conditional on the treatment assignment $Z_i \in \{1, 2, 3\}$ and the covariates (X_{ij}), where $Pr(Z_i = 1 \mid X_{ij}) = Pr(Z_i = 2 \mid X_{ij}) = 1/3$. Specifically, the response was obtained from:

$$Y_i = \beta_1 I(Z_i = 1) + \beta_2 I(Z_i = 2) + \sum_{j=1}^{r} \gamma_j X_{ij} + \varepsilon_i \tag{3.16}$$

where $\varepsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0, 1)$ and $i = 1, 2, ..., N$. In the simulations, we considered the following:

- The true treatment effect was taken to be $\beta_1 = \beta_2 = 0.5$
- The true confounding effects were $\gamma_j = \gamma,\ j = 1, ..., r$ where $\gamma = 0.5, 1.0, 1.5$. Note that the results we obtain do not depend on the choice of β. When the covariates follow symmetric distributions, the results do not depend on the signs of the components of γ either.
- For the first three settings, we considered $r = 4$ covariates selected from the following distributions: $(i) X_1, X_2, X_3, X_4 \overset{i.i.d}{\sim}$ Bernoulli(0.5); $(ii) X_1, X_2 \overset{i.i.d}{\sim}$

Bernoulli(0.5); $X_3, X_4 \overset{i.i.d}{\sim} N(0,1)$;

- We consider sample sizes $N = 60$ or 36.

3.1.4.2 Results

Table 3.1 presents the MSE performance of the generalized BMW design based on three tripartite matching algorithms under various confounding scenarios. Since the reduction in bias achieved by the BMW design with incomplete block method and three-way tripartite matching algorithm is symmetrically balanced in all treatment effects estimators, we only show the percent reduction in MSE in one treatment effects estimator for those methods.

In general, the three-arm BMW designs provide important gains in efficiency by reducing the MSE in all three treatment effects estimators simultaneously. This is especially for the design based on the two-way and three-way tripartite matching methods we proposed.

The BMW design based on the incomplete block method or the three-way tripartite matching algorithm could both be applied to situations when the three pairwise comparisons are equally important to investigators. However, the proposed three-way tripartite matching algorithm is substantially more effective in reducing MSE compared to the incomplete block design. This is especially the case when the confounding effects are not too strong (e.g. $\gamma = 0.5$ or 1.0). Specifically, when the common confounding effects $\gamma = 0.5$, the estimated treatment effect, $\hat{\beta}_{AC}$, from the BMW design using the incomplete block method is only 89.7% efficient compared to that from the completely randomized design. Adding random blocks to the sample, as suggested by Bo and Rosenbaum (2004), leads to a loss of efficiency and this limitation arises especially when the confounding effects are not too strong, thus, the control in variance increase is more important than the bias reduction. On the other hand, our proposed three-way tripartite matching algorithm utilizes all the subjects

assigned to each treatment group in the direct comparisons, which minimizes the efficiency loss and reduces the bias by optimally matching the subjects into pairs. Although the balance achieved through optimally matched pairs may not be as good as the optimally matched triples, it is close to the optimum and substantially better than the random assignment of treatments. For example, the BMW design based on the proposed three-way tripartite matching algorithm leads to a reduction in MSE by 12.63% when $\gamma = 0.5$, and this reduction increases to 41.44% as γ increases to 1.5.

When the control group is predefined by the investigators, we proposed the two-way tripartite matching algorithm to primarily reduce the covariate imbalance between each treatment group and the common control. Simulation results suggest that the BMW design based on the two-way matching algorithm is very effective in reducing MSE of the treatment effects estimator compared to the completely randomized design. For example, if group A is set as control, the BMW design can minimize the MSE of $\hat{\beta}_{AC}$ and $\hat{\beta}_{AB}$ by more than 44.79% and 43.16%, respectively, compared to the completely randomized design, when $\gamma = 1.5$. The covariate imbalance between group B and C is also reduced through the individual matching of B and C to the common control A.

The simulation studies also evaluate the effects of sample size N on the performance of these BMW designs. Table 3.1 and 3.2 present the percent MSE reduction given sample size as 60 and 36, respectively. This results reveal that the performance of the BMW design based on the incomplete block design of disjoint pairs decreases when the sample size reduces, however, the designs using our proposed two-way and three-way tripartite matching designs of triples remain approximately unchanged when sample size varies. Therefore, the efficiency gain in MSE due to the two-way and three-way tripartite matching algorithms becomes even more apparent when sample size becomes smaller.

Finally, the effects are more apparent when the covariates include continuous

variables with larger variances but becomes somewhat less when the covariates are all Bernoulli variables.

3.1.5 Conclusions and Discussions

Traditional designs in cluster randomized trials have generally studied the effect of one variable at a time, because it is statistically easier to manipulate. However, in many instances, there may exist two or more factors, and it is impractical or insufficient to analyze each variable individually. The 2×2 factorial design, and of course factorial designs more generally, have been extremely useful in the area of social, medical and agricultural research. This design can highlight the relationships between variables and evaluate the effects of multiple variables simultaneously. In the 2×2 factorial design, there are two factors with each on two levels. The number of different treatment groups is therefore $2 \times 2 = 4$.

The generalized 3-arms BMW design proposed in this chapter can be further extended to optimize the MSE of the resultant treatment effects estimators in the 4-arms or larger trials. Specifically, we can apply the two-way quadripartite matching algorithm, an generalization of the two-way tripartite matching method, when there is a clear predefined reference group; or extend the three-way tripartite matching to four-way quadripartite matching algorithm when these four groups are to be compared simultaneously. An optimal reference group can be found among the four groups. As for the extension of the incomplete blocks of disjoint pairs, however, Bo and Rosenbaum (2004) noted that the nonbipartite algorithm does not extend to four or more treatments and suggest instead using the best pair matching that includes all observations from each group. As a result, our proposed tripartite matching algorithm with triples appears to be more useful in the extensions to four or more groups.

It is clear that the proposed two-way and three-way tripartite matching algorithms can not only be used in the randomized experiments, in an effort to reduce the chance

Table 3.1: Percent reductions in the MSE of treatment effect estimator for the generalized BMW design based on the incomplete block design of disjoint pair (ICB), three-way tripartite matching with triples ($3TM$) and two-way tripartite matching with triples ($2TM$) respectively, compared to a completely randomized design (CR). Sample size N=60 subjects. Number of replications=1000.

γ	$\sum_{j=1}^{4}\gamma_j$	M	MSE (CR)	MSE Percent Reduction(%) (ICB vs. CR Design)	($3TM$ vs. CR Design)	($2TM$ vs. CR Design)		
				$\hat{\beta}_{AC}$	$\hat{\beta}_{AC}$	$\hat{\beta}_{AC}$	$\hat{\beta}_{AB}$	$\hat{\beta}_{BC}$
$X_1, X_2, X_3, X_4 \overset{i.i.d}{\sim} Bernoulli(0.5)$								
(0.5, 0.5, 0.5, 0.5)	2	10	0.126	-10.32	12.63	13.28	12.77	10.99
(1.0, 1.0, 1.0, 1.0)	4	10	0.203	22.17	30.46	32.81	31.78	25.33
(1.5, 1.5, 1.5, 1.5)	6	10	0.331	43.50	41.44	44.79	43.16	35.96
$X_1, X_2 \overset{i.i.d}{\sim} Bernoulli(0.5); X_3, X_4 \overset{i.i.d}{\sim} N(0,1)$								
(0.5, 0.5, 0.5, 0.5)	2	10	0.162	4.32	22.60	23.42	23.78	19.61
(1.0, 1.0, 1.0, 1.0)	4	10	0.348	35.34	42.29	43.98	42.89	36.50
(1.5, 1.5, 1.5, 1.5)	6	10	0.659	47.34	48.63	53.95	49.83	42.73

Table 3.2: Percent reductions in the MSE of treatment effect estimator for the generalized BMW design based on the incomplete block design of disjoint pair (ICB), three-way tripartite matching with triples ($3TM$) and two-way tripartite matching with triples ($2TM$) respectively, compared to a completely randomized design (CR). Sample size N=36 subjects. Number of replications=1000.

γ	$\sum_{j=1}^{4}\gamma_j$	M	MSE (CR)	MSE Percent Reduction(%) (ICB vs. CR Design)	($3TM$ vs. CR Design)	($2TM$ vs. CR Design)		
				$\hat{\beta}_{AC}$	$\hat{\beta}_{AC}$	$\hat{\beta}_{AC}$	$\hat{\beta}_{AB}$	$\hat{\beta}_{BC}$
$X_1, X_2, X_3, X_4 \overset{i.i.d}{\sim} Bernoulli(0.5)$								
(0.5, 0.5, 0.5, 0.5)	2	10	0.209	-12.10	12.03	13.10	12.44	10.54
(1.0, 1.0, 1.0, 1.0)	4	10	0.337	18.07	30.97	33.13	31.60	27.31
(1.5, 1.5, 1.5, 1.5)	6	10	0.550	39.02	42.15	42.90	44.83	35.93
$X_1, X_2 \overset{i.i.d}{\sim} Bernoulli(0.5); X_3, X_4 \overset{i.i.d}{\sim} N(0,1)$								
(0.5, 0.5, 0.5, 0.5)	2	10	0.273	3.00	23.39	23.71	24.53	20.48
(1.0, 1.0, 1.0, 1.0)	4	10	0.592	33.13	42.41	43.19	43.94	35.61
(1.5, 1.5, 1.5, 1.5)	6	10	1.123	43.93	50.45	53.46	52.45	42.16

imbalance in observed covariates between treatment groups, but also be applied to observational studies with two control groups as were the context that motivated the incomplete block design of Bo and Rosenbaum (2004). One limitation arises when there are unequal number of observations between the three groups. We suggest, in that case, either to select a subset of observations so that each treatment group would has equal size, then search for the optimal two-way tripartite matching, or form a matching close to optimal by including all observations. This is an interesting and important problem that deserves further study.

The proposed BMW designs based on the propensity score matching have one disadvantage. It may not perform well in the studies with very small sample size (e.g. group size is less than 10) but relatively large number of confounding variables (e.g. 4 or more). In that case, the model used to estimate the propensity scores may not work well due to the complete separation of cases and controls by covariates. One may use some ad hoc method such as reducing the number of independent variables in the model to estimate the propensity score when that occurs. Further studies on proposing some appropriate ad hoc methods will also be conducted.

3.2 The BMW design on trials with staggered entry

One limitation of the BMW design discussed so far is that it requires that all units are available for randomization at the onset of the study. So the next generalization of the BMW design is to extend it to clinical trials with staggered entry. Traditional methods of restricted randomization includes covariate adjustment, "permuted block design" and "permuted block design with strata". However, each of these has drawbacks in terms of minimizing the covariance imbalance for several prognostic factors simultaneously in sequential treatment assignment, and the limitation is most serious for small studies.

The disadvantage of covariate adjustment method is that the number of variables

which can be treated in this manner is limited, and in small studies, it is not always clear what covariate model is appropriate. If an incorrect model is chosen, covariate adjustment can result in biased estimation of the true overall treatment effect. The "permuted block design" first forms patients in blocks of fixed, equal size, bN, where b is the number of treatments and N denotes the number of new patients in each treatment arm, as they enter the trial and a random permutation of treatments is determined for each block. This method can ensure equal number of treatments are achieved in every block of patients but Efron (1971) points out a big limitation of this design. Within each block, after $bN - 1$ treatments have been assigned and last one is predetermined. This limitation may cause considerable bias especially when bN is small in unblinded single center studies. Another method of achieving balance with respect to prognostic factors is the "permuted block design within strata". This approach divides each factor into several levels and each patient is assigned to a stratum according to his/her particular combination of factor levels. Treatments are then assigned at random and typically with balance within each stratum. The major difficulty in this approach is the number of strata (i.e. the number of combination of factor levels) can be very large if the number of prognostic factors increases. In an extreme case, the number of strata can be as large as the number of patients accrued, then some strata will have no patients or only one patient and this constrained randomization becomes equivalent to complete randomization.

Pocock and Simon (1975) proposed the minimization design, which is a sequential strategy by making the assignment decision one unit at a time, based solely on the covariate information of previously assigned subjects. This procedure requires that the prognostic factors are categorical variables, or categorized into categories. For each new patient, we calculate the total covariate imbalance, based on the covariate information of previously assigned subjects, for each possible treatment assignment of this new patient, and choose the one which minimizes the total imbalance. The

computation of the total imbalance for all possible treatment assignments becomes very intensive when the number of prognostic variables increases.

The BMW design can be generalized to account for sequential treatment assignment. Suppose in a randomized trial with two arms, a new block of patients with size N_1 arrives and there are N_2 patients already randomized to treatments. We then randomize the N_1 patients M times and retain the assignment for those N_2 patients. For each randomized sample, compute the estimated propensity score and search for the optimal full matching with constraints based on all patients. The randomization which leads to the minimal total distance in propensity score would determine the treatment assignment for the N_1 new patients. One potential limitation of this approach proposed is that the block size N_1 may need to be reasonably large to get good properties. We carry out a simulation study to evaluate the MSE performance of the proposed approach.

For each of N subjects, we generated a set of r covariates $X_1, X_2, ..., X_r$, where the covariates were drawn independently from the distributions as described below. Given a randomization of subjects to the two treatment groups, the responses were generated conditional on the treatment assignment ($Z_i = 0$ or 1) and the covariates (X_{ij}) (where $Pr(Z_i = 1 \mid X_{ij}) = 0.5$). Specifically, the response was obtained from:

$$Y_i = \beta Z_i + \sum_{j=1}^{r} \gamma_j X_{ij} + \varepsilon_i \tag{3.17}$$

where $\varepsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0, 1)$ and $i = 1, 2, ..., N$. In the simulations, we considered the following:

- The true treatment effect was taken to be $\beta = 0.7$
- The true confounding effects were $\gamma_j = \gamma,\ j = 1, ..., r$ where $\gamma = 0.5,\ 1.0,\ 1.5$.
- For the covariate setting, we considered $r = 4$ covariates selected from the

following distribution: $(ii) X_1, X_2 \overset{i.i.d}{\sim}$ Bernoulli(0.5); $X_3, X_4 \overset{i.i.d}{\sim} N(0, 0.25)$;

- We consider sample sizes $N = 36$.

There are three scenarios of staggered entry under investigation, first, all 36 patients are available at study onset; second, the patients arrive in two blocks of size 18 each; third, all these patients arrive at three different times of size 12 each. The BMW design, as described in the preceding section, is applied repeatedly each time when the new block of subjects enter the study. The completely randomized design assigns half of the units at random to each of the two treatment groups and the assignment takes place independently whenever the new patients enter the study. For this design, the treatment effect estimator is $\widehat{\beta}_{pooled} = \overline{Y}_T - \overline{Y}_C$ and the corresponding mean squared error (conditional on T, C and X) is given in (2.8). We also consider a permuted block design within strata in which subjects were first divided into blocks based on the first two binary covariates X_1 and X_2. All the patients in the same strata share the common covariate values for the first two covariates, and the treatment assignments are performed separately for each stratum.

Simulation results in table 3.3 suggest that the generalized BMW design is effective in reducing the MSE of the treatment effect estimator in clinical trials with staggered entry, compared to the completely randomized design and permuted block design within strata. The improvement in MSE becomes more apparent as the confounding effects become stronger. It appears that the percent of MSE reduction may increase slightly as the number of blocks with staggered entry increases. For example, the percent reduction in MSE is generally higher when patients arrive sequentially in three groups of 12 each, compared to the case that subjects are all available at the study onset. When patients arrive in three blocks, the BMW design are repeatedly performed three times and each time with the number of replication M. Under the scenario of no staggered entry, however, the BMW design is only applied once at the study onset with the same M. As discussed in Chapter II, increasing the number

of replication M in BMW design can produce greater reduction in MSE. Further investigations find that if we increase M in the scenarios with staggered entry, the difference in MSE reduction among the three scenarios considered disappears.

3.3 The BMW design with a large M

In Chapter II, we introduced the two-arm BMW design with a prespecified parameter k and M. Clearly, the choices of parameter k and M have an important impact on the MSE performance of the BMW design. Specifically, the value of k represents a trade-off between bias reduction and precision loss, whereas, the value of M controls the level of balance that the BMW design can attain. Through simulation studies, we found that $k = 2$ and $M = 10$ or so (out of the candidates values of 5, 10 and 20) are usually suitable choices under most of the confounding scenarios considered. In Chapter II, we recommended a full matching with constraint $k = 2$ with the number of replication of $M = 10$ for implementing the BMW design in practice. However, further investigation revealed that, as M increases, the advantage of $k = 1$ (pair matching) becomes more apparent in optimizing the MSE of treatment effects estimator as compared to other choices of k. For example, when M is as small as 5 or 10, $k = 2$ appears to be optimal with the confounding effect $\gamma = 1.0$ or 1.5; however, as M increases to 20 or even 100, $k = 1$ becomes as effective as or even more effective than $k = 2$ in terms of reducing the MSE (Table 3.4). These results suggest that in practice, the BMW design with a large number of replications of $M = 100$ or more would give some gains in efficiency and that pair matching would then be the optimal choice.

Table 3.3: Percent reductions in the MSE of treatment effect estimator for the BMW design compared to completely randomized design (CR) and permuted block design within strata (BL) under three scenarios of staggered entry. Number of replications=1000. $X_1, X_2 \overset{i.i.d}{\sim} Bernoulli(0.5); X_3, X_4 \overset{i.i.d}{\sim} N(0, 0.25)$.

γ_j	$\sum_{j=1}^{4} \gamma_j$	M	MSE (CR)	MSE Reduction(%) (BMW vs. CR Design)			MSE (BL)	MSE Reduction(%) (BMW vs. BL Design)		
				1X36	2X18	3X12		1X36	2X18	3X12
0.5	2	10	0.129	15.17	17.87	18.32	0.115	-0.38	2.72	3.23
1.0	4	10	0.180	37.24	38.82	40.36	0.127	9.60	12.39	12.85
1.5	6	10	0.266	56.69	58.03	58.25	0.145	20.26	22.72	23.13

Table 3.4: Percent reductions in the MSE of treatment effect estimator for the BMW design with $M = 100$ compared to a completely randomized design (CR) and matched-pair design (MP). Sample size N=30 subjects. Number of replications=1000.

γ	M	MSE (CR)	MSE Percent Reduction(%) (BMW vs. CR Design)			MSE (MP)	MSE Percent Reduction(%) (BMW vs. MP Design)		
			$k=1$	$k=2$	$k=3$		$k=1$	$k=2$	$k=3$
$X_1, X_2, X_3, X_4 \overset{i.i.d}{\sim} Bernoulli(0.5)$									
(0.5,0.5,0.5,0.5)	5	0.166	12.21	10.30	6.87	0.158	7.96	5.96	2.37
	10		14.43	11.77	7.14		10.29	7.50	2.64
	20		17.45	13.54	8.81		13.46	9.36	4.40
	100		17.49	13.18	10.62		13.95	9.46	6.78
(1.0,1.0,1.0,1.0)	5	0.280	35.61	43.58	39.67	0.239	24.57	33.90	29.33
	10		40.37	44.45	41.74		30.15	34.92	31.75
	20		50.39	48.66	46.21		41.87	39.86	36.99
	100		52.05	50.44	47.96		43.82	41.93	39.04
(1.5,1.5,1.5,1.5)	5	0.450	45.39	61.58	57.94	0.374	34.29	53.77	49.39
	10		52.19	62.26	59.02		42.47	54.59	50.69
	20		58.43	63.52	60.64		49.97	56.10	52.64
	100		64.07	64.66	62.38		56.77	57.47	54.73

CHAPTER IV

A Non-parametric maximum likelihood estimation approach to frailty model

4.1 Introduction and motivating example

Survival data are sometimes clustered into groups. Observations sampled from the same group often share certain unmeasured characteristics and as a result tend to be correlated. Failure to adjust for this intra-class correlation may bias the covariate-effect estimates or lead to inaccurate estimates of standard errors. In addition, the lack of independence is not just a nuisance that must be taken into account in the analysis, since quantifying the heterogeneity may itself be of interest. Medical examples of clustered survival data include studies of the time to occurrence of a genetic disease among siblings, the onset of visual loss in left and right eyes, and the failure times of individuals within center in a multi-center study. Such failure times are often subject to right censoring. The presence of censoring and intra-class dependence poses serious challenges in the regression analysis of clustered failure time data. Note that such data also arise in the context of cluster randomized trials.

There are two major modeling approaches for analysis of clustered or multivariate failure time data: conditional models (Vaupel et al., 1979; Clayton, 1978; Clayton and Cuzick, 1985; Cai et al., 2002) and marginal models (Zeng et al., 2008). The

former provides a flexible way for directly modeling the relationship and dependence structure between correlated failure times, whereas the latter focuses on covariate effects on individual failure times and does not attempt to model the dependence structure of related failure times. In this article, we develop a conditional approach through a frailty model.

A Cox proportional hazards model in which a multiplicative frailty factor is included to account for intra-cluster correlation has often been considered in the literature. In this context, there is a substantial literature dealing with the identification and estimation of frailty models using both parametric and semiparametric approaches. In such approaches, parametric models have often been used for the frailty distribution or the baseline hazard, or both. For example, some authors have worked with a parametric baseline hazard (e.g. piecewise constant) and nonparametric frailty distribution (Heckman and Singer, 1984a; Trussell and Richards, 1985; Guo, 1992) as well as many other parametric options such as gamma, Gaussian, log-normal or stable law (McGilchrist and Aisbett, 1991; Ripatti and Palgrem, 2000; Breslow and Clayton, 1993; Clayton, 1978; Vaupel et al., 1979; Nielsen et al., 1992). On the other hand, Therneau et al. (2003) assumed a parametric model for the frailty distribution while estimating the baseline hazard nonparametrically, and obtained a solution via a penalized regression. It is obvious that the covariate-effect estimates and, thus, the inferences one would draw, could be sensitive to the parametric form assumed for the hazard and frailty. Furthermore, the parametric assumption on frailty certainly induces a restrictive form of dependence.

Heckman and Singer (1984a) studied the sensitivity of parameter estimates to the choice of distribution for the unobservable heterogeneity while assuming a piecewise constant distribution for the baseline hazard, and recommended a nonparametric approach for frailty distribution estimation. Further, Trussell and Richards (1985) found that, even with a nonparametric representation of heterogeneity, results can still

be sensitive to choice of a model for the baseline hazard. Without very refined theory upon which to support hazard and mixing distribution specifications, the results suggest that one should seek a strategy that is somewhat agnostic of Trussell and Richards (1985) regarding both the hazard and the mixing frailty distribution. Nielsen et al. (1992) also commented that "It has been shown in the regression context that the model with arbitrary frailty distribution with finite mean is identifiable, so one could in principle allow both the frailty distribution and the underlying hazard to vary freely". However, there is little literature dealing with the identification and estimation of frailty models using a purely nonparametric approach. Heckman and Singer (1984c) and Heckman and Singer (1984b) established the identifiability of the hazard function and the mixing distribution pairs by introducing minimal moment restrictions on the mixing distributions and allowing the hazards to be represented by Box-Cox transformations. On one hand, such an approach provides a model specification approach that makes fewer assumptions about the hazards than most previous analysis. But on the other hand, constraints are imposed on the mixing distributions, as opposed to simply allowing them to have arbitrary structure.

Our approach to these models is through nonparametric maximum likelihood. In particular, we consider a frailty model with both the frailty distribution, G, and the cumulative baseline hazard, Λ_0, left nonparametric. We propose an approach based on nonparametric maximum likelihood estimation. For implementation, a three-step iterative algorithm is developed. First, assuming initial estimates for $\Lambda_0(\cdot)$ and β, we use a fast converging algorithm, such as the Intra-simplex Direction Method (ISDM) of Lesperance and Kalbfleisch (1992), or the Constrained Newton Method with multiple support points inclusion (CNM) of Wang (2007) to estimate the frailty distribution G. Second, for frailty G and β as given, Λ_0 is estimated using a variation of the Breslow (1972) cumulative baseline hazard estimator. At the third step, the regression parameter, β, is estimated given the current estimate of the frailty distribution

and baseline hazard. Different from the parametric or semiparametric approaches proposed previously, our "purely" nonparametric approach relaxes any distributional assumption on the baseline hazard and frailty distribution, allowing both of them to vary freely. Therefore, it has potential advantages of flexibility and robustness.

The rest of the chapter is organized as follows. Section 4.2 describes the basic multiplicative frailty model and introduces nonparametric specifications of the distribution of the unobserved random effect. Section 4.3.1 shows how the frailty distribution can be estimated using ISDM or CNM. Section 4.3.2 deals with estimating the baseline hazard function by using a variation on Breslow's hazard estimator. Section 4.4 describes the three-step iterative algorithm and Section 4.5 introduces a competing semiparametric method proposed by Therneau et al. (2003). In section 4.6, some simulation studies are also given to investigate the performance of our proposed algorithm and to compare it with the Therneau et al. (2003) method in terms of efficiency and robustness. Section 4.8 discusses some potential extension of future work.

4.2 Model

Consider a survival study that involves N right-censored failure time data clustered into M small groups. Let T_{ij} denote the failure time of interest from subject j in cluster i with the associated treatment assignment Z_{ij}, where $i = 1, 2, ..., M$, $j = 1, 2, ..., n_i$. We also assume that subjects in the same cluster share common treatment assignment, i.e. $Z_{ij} = Z_i$, as the treatment assignment in the clustered randomized trials where the cluster is the unit of randomization. We assume that each cluster has a cluster-specific latent variable U_i, where $U_1, ..., U_M$ are independently and identically distributed from an underlying frailty distribution G, and that given U_i, the survival times T_{ij}, $j = 1, 2, ..., n_i$, are identically and mutually independently distributed. Also let C_{ij} denote the censoring time and $\delta_{ij} = I(T_{ij} \leq C_{ij})$.

Given $U_i = u_i$, the conditional hazard function $\lambda(t_{ij})$ satisfies the multiplicative frailty model

$$\lambda(t_{ij}) = \lambda_0(t_{ij}) u_i e^{Z_i\beta} \tag{4.1}$$

where $\lambda_0(t), t > 0$ denotes an arbitrary baseline hazard, and β represents the treatment effect estimator. We impose the restriction $\Lambda_0(1) = 1$ for identifiability. The restriction of $E(G) = 1$ is a nice alternative.

The model 4.1 is often referred to as the proportional hazards frailty model and has been extensively used for regression analysis of clustered right-censored failure time data with parametric models for G and $\lambda_0(\cdot)$. (Clayton and Cuzick, 1985; Lee et al., 1992; Cai and Prentice, 1997; Hougaard, 2000). It is easy to see that if the variance of the U_i reduces to zero, model 4.1 becomes the regular proportional hazards model (Cox, 1972; Kalbfleisch and Prentice, 2002). In a cluster randomized trial, Z_{ij} is reduced to Z_i since the subjects in the same cluster share the common treatment assignment. In this article, we assume that given Z_i, the frailty follows a completely unknown distribution $G(\cdot)$ and T_{ij} is independent of C_{ij} given (U_i, Z_i). The marginal likelihood function is proportional to

$$\begin{aligned} L(\beta; G(u), \Lambda_0(t)) &= \prod_{i=1}^{M} \int_u \prod_{j=1}^{n_i} \lambda_0(t_{ij})^{\delta_{ij}} u^{\delta_{ij}} e^{\delta_{ij} Z_i\beta} e^{-u\Lambda_0(t_{ij}) e^{Z_i\beta}} \, dG(u) \\ &= \prod_{i=1}^{M} \int_u f_i(y_i; \beta, \Lambda_0, u) \, dG(u) \end{aligned} \tag{4.2}$$

where $y_i = (t_{ij}, \delta_{ij})$ and $f_i(\cdot)$ is being implicitly defined. In the next section, we consider maximizing likelihood 4.2 with respect to β, G and Λ_0.

4.3 Methods

In this section, we introduce the methods used to derive the NPMLE of the frailty distribution as well as the baseline hazard estimator.

4.3.1 Estimation of Frailty Distribution

We consider estimating the distribution of the unobserved random frailty effects through nonparametric maximum likelihood. As has been shown (Heckman and Singer, 1984a; Laird, 1978; Lindsay, 1983), the nonparametric approach leads to a nonparametric maximum likelihood estimator (NPMLE), $\hat{G}$, that is discrete with mass $\hat{\pi}_1, ..., \hat{\pi}_J$ on a fixed number J of support points, $\hat{u}_1, ..., \hat{u}_J$, respectively. The support set can contain no more points than the number of distinct values in the sample.

There are several methods in the literature for computing the nonparametric MLE of a mixture distribution G, for example, the expectation-maximization (EM) algorithm (Laird, 1978), the vertex direction method (VDM) (Fedorov, 1972), the vertex exchange method (VEM) (Bohning, 1985), the semi-infinite programming method (SIP) (Susko et al., 1998) and the quadratic method (Atwood, 1976). All these methods have the disadvantage of slow convergence.

The problem of convergence, however, can now be solved efficiently due to the availability of two fast algorithms. Lesperance and Kalbfleisch (1992) modified the VDM and proposed the Intra-simplex Direction Method (ISDM); Wang (2007) extended ISDM using Atwood's method and developed the constrained Newton method with multiple support points inclusion (CNM). Both methods dramatically improved the efficiency and shared somewhat similar spirit: first, they are both based on the directional derivative; second, they both consider all local maxima of the directional derivative in each iteration step instead of just one as in the previous methods.

4.3.1.1 The Geometry of Mixture Likelihoods

Lindsay (1983) provided a geometric interpretation of mixture likelihood which enables a clear description of the algorithms such as ISDM and CNM used in computing the NPMLE of a mixture distribution G, We summarize it in this section.

We will analyze the likelihood 4.2 but with β and $\Lambda_0(\cdot)$ suppressed. We let

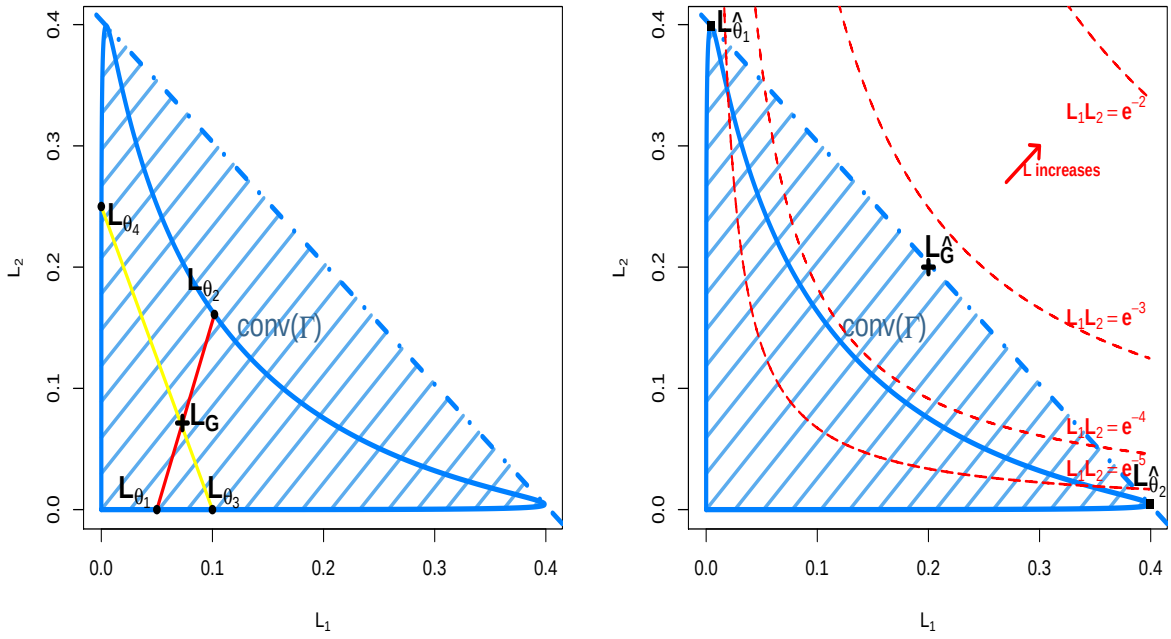
$$L_i(G) = \int_u f_i(y_i; u) dG(u) \tag{4.3}$$

and consider $L_G = (L_1(G), ..., L_M(G))$ as a point in R^m. The log-likelihood 4.2 can be written as

$$l(G) = \sum_{i=1}^{M} \log \{L_i(G)\} = \sum_{i=1}^{M} \log \left\{ \int_u f_i(y_i; u) dG(u) \right\} \tag{4.4}$$

If $G = \delta_\theta$, where δ_θ places mass one on a specific $\theta \in \Omega$, then the likelihood vector $L_\theta = (L_1(\delta_\theta), ..., L_M(\delta_\theta)) = (f_1(y_i; \theta), ..., f_M(y_i; \theta))$ is also called the atomic likelihood vector. Let $\Gamma = \{L_\theta : \theta \in \Omega\}$ represent the set of all possible atomic likelihood vectors. Note that Γ traces out a trajectory in R^m. The convex hull of Γ, which is the set of all convex combinations of Γ, is written as $conv(\Gamma)$=$\{L_G : G \in \mathcal{G}$, G has finite support$\}$. Any point in $conv(\Gamma)$ corresponds to the likelihood attainable under a mixture model for one or more distributions $G \in \mathcal{G}$. Figure 4.1a provides a simple example to illustrate the quantities introduced above.

The nonparametric maximization problem can be formulated as: find $\hat{G}$ such that $L_{\hat{G}} \in conv(\Gamma)$ maximizes the log-likelihood $l(G) = \sum_{i=1}^{M} logL_i(G)$. If Γ is compact, then there exists an unique optimal vector $L(\hat{G})$ on the boundary of $conv(\Gamma)$. As shown in Figure 4.1b, the likelihood contours $L_1 * L_2 = c$, where the constant $c = e^{(-2)}, .., e^{(-5)}$, indicates that the likelihood increases as it moves further from the origin, the unique maximum point $L_{\hat{G}}$ is attained on the boundary of $conv(\Gamma)$. In



(a) Γ and *conv*(Γ)

(b) Finding NPMLE

Figure 4.1: The Geometry of Mixture Likelihoods. The heavy blue curve: Γ, the 2-dimensional solid figure: *conv*(Γ) and the red dashed curves: likelihood contours (b) for two normal observations. $\Gamma = [\phi(1-\theta), \phi(4-\theta) : \theta \in R]$, where $\phi(\cdot)$ is the standard normal probability density function. The log-likelihood function $log(L_1) + log(L_2)$, where $(L_1, L_2) \in R^2$. In (a): For any point $L_G(L_1, L_2) \in conv(\Gamma)$, then $L_1 = \int_u \phi(1-u)dG(u)$ and $L_2 = \int_u \phi(4-u)dG(u)$ for some distribution G and L_G can be attained by the convex combination of atomic points L_{θ_1} and L_{θ_2}, or L_{θ_3} and L_{θ_4}, etc.

this example, the corresponding mixing distribution $\hat{G}$ can be written as an unique convex combination of two support points $L_{\hat{\theta}1}$ and $L_{\hat{\theta}2}$ with equal masses. Typically, the optimal mixing distribution would be unique although in some cases, it could be multiple maxims. Caratheodorey's theorem guarantees that the optimal mixture distribution $\hat{G}$ associated with $L(\hat{G})$ has M or fewer support points (Lindsay, 1983).

Directional derivative can be used to characterize the optimal vector $L_{\hat{G}} \in conv(\Gamma)$ and the corresponding NPMLE of the mixing distribution $\hat{G}$. Consider two mixture distributions, $G_1(u)$ and $G_2(u)$, $u \in \Omega$. The directional derivative of the likelihood from the point L_{G_1} towards L_{G_2} is defined as

$$\begin{aligned} D(L_{G_2}; L_{G_1}) &= \lim_{\epsilon \downarrow 0} \frac{l\{(1-\epsilon)G_1 + \epsilon G_2\} - l(G_1)}{\epsilon} \\ &= \sum_{i=1}^{M} [\frac{L_i(G_1)}{L_i(G_2)} - 1] \end{aligned} \tag{4.5}$$

If $G_2 = \delta_\theta$, then we write $g(\theta; L_{G_1}) = D(L_{G_2}; L_{G_1})$ and refer to $g(\theta; G_1)$ as the gradient function from L_{G1} to the points L_θ on Γ. (Figure 4.2).

The famous *General Equivalence Theorem* in Lindsay (1983) guarantees that the NPMLE $\hat{G}$ can be characterized by the gradient function:

Theorem 1.

- A. The measure $\hat{G}$ that maximizes $l(G)$ can be equivalently characterized by three conditions: (1) $\hat{G}$ maximizes $l(G)$; (2) $\hat{G}$ minimizes $\sup_{\theta \in \Omega} g(\theta, G)$; (3) $\sup_{\theta \in \Omega} g(\theta, \hat{G}) = 0$.

- B. The support of $\hat{G}$ is contained in the set of θ for which $g(\theta, \hat{G}) = 0$.

Furthermore, the following result also holds and provides an ideal stopping criterion.

$$\sup_{\theta} \{g(\theta; G)\} \geq l(\hat{G}) - l(G) \tag{4.6}$$

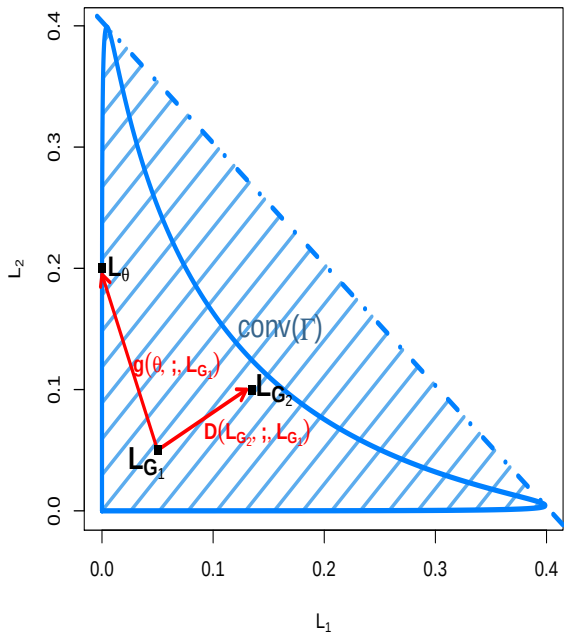


Figure 4.2: Directional Derivative of the example in Figure 4.1.

4.3.1.2 Intra-simplex Direction Method (ISDM)

The ISDM method expanded the support points set by adding all new local maxima of the gradient function with non-negative directional derivative and re-distributing the mass among them. One beneficial feature of ISDM is that it does not require that one keeps track of the accumulated support points set, but only the current likelihood point $L_G = (L_1(G), ..., L_M(G))$. The MLE $\hat{G}$ is found at the final iteration.

Algorithm ISDM: Set s=0. From an initial estimate G_0, obtain $L_0 = (L_1(G_0), ..., L_M(G_0))$ where we assume $l_i(G_0) > -\infty$, $i = 1, 2, ..., M$.

- *Step 1: Expand the support points sets:*

  Denote by $L_s = (L_1(G_s), ..., L_M(G_s))$ the current point in $conv(\Gamma)$. Compute all local maxima θ^*_{s1}, ..., $\theta^*_{sp_s}$ of $g(\theta; L_s) \geq 0$, $\theta \in \Omega$. If $\max_j\{g(\theta^*_{sj}; L_s)\} = 0$, stop.

- *Step 2: Find the optimal weights to determine the new likelihood points*

  Compute ϵ^*_{s0}, ..., $\epsilon^*_{sp_s}$, to maximize the log-likelihood

$$K(\epsilon_0, \epsilon_1, ..., \epsilon_{p_s}) = \sum_{i=1}^{M} log\{\epsilon_0 L_{si} + \sum_{j=1}^{p_s} \epsilon_j L_i(\theta^*_{sj})\} \tag{4.7}$$

  subject to $\sum_{j=0}^{p_s} \epsilon_{sj} = 1$ and $\epsilon_{sj} \geq 0$ where $L_i(\theta) = f_i(y_i; \theta) = L_i(\delta_\theta)$.

- *Step 3: Obtain the new point:* Set $L_{s+1} = \epsilon^*_{s0} L_s + \sum_{j=1}^{p_s} \epsilon^*_{sj} L(\theta^*_{sj})$. Set $s = s + 1$ and go to *Step 1*.

After the final step, we obtain $\hat{L} = (\hat{L}_1, ..., \hat{L}_M)$, the likelihood point corresponding to the NPMLE $\hat{G}$. If $\hat{G}$ is also of interest, then we can find the support points, $\hat{u}$, of $\hat{G}$ as the local maxima obtained at the final iteration, and the appropriate weights $\hat{\pi}$ by an extra run of *Step 2* with $\epsilon_0 = 0$.

The key innovation of the ISDM method is that the choice of direction at each iteration is determined by maximizing the likelihood within a probability simplex, which is chosen to approximate the important part of the convex hull of Γ. It is this innovation that contributes a substantial improvement in efficiency over previous methods. ISDM is guaranteed to converge monotonically to the NPMLE $\hat{G}$ if Γ is compact (Theorem 2 in Lesperance and Kalbfleisch (1992)).

4.3.1.3 CNM

The CNM method parallels with ISDM, except in the following ways: First, it requires that one keep track of the set of support points, instead of only the likelihood point, L, at each iteration step. Second, the set of support points is expanded by adding all the local maxima of the gradient function rather than only those with non-negative direction derivative, and on the other hand, contracted by discarding those "bad" points with zero weights after the weights are updated. Third, at each stage,

the optimal weights are obtained by solving a quadratic programming subproblem via a linear regression formulation.

Suppose that G has mass points at $\mathbf{u} = (u_1, ..., u_J)$ with masses $\boldsymbol{\pi} = (\pi_1, ..., \pi_J)$ so that

$$G(u) = \sum_{j=1}^{J} \pi_j I(u \geq u_j) = \sum_{j=1}^{J} \pi_j \delta_{u_j} \tag{4.8}$$

Then, again with $\Lambda_0(\cdot)$ and β suppressed, the log-likelihood becomes

$$l(\boldsymbol{\pi}, \mathbf{u}) = \sum_{i=1}^{M} l_i(\boldsymbol{\pi}, \mathbf{u}) = \sum_{i=1}^{M} \log \left\{ \sum_{j=1}^{J} \pi_j f_i(y_i; u_j) \right\} \tag{4.9}$$

Given the support points, $\mathbf{u}$, the updating of $\boldsymbol{\pi}$ utilizes a second-order Taylor series expansion of the log-likelihood function in the neighborhood of $\boldsymbol{\pi}$. Denote

$$\mathbf{s}_i(\boldsymbol{\pi}, \mathbf{u}) = \frac{\partial l_i(\boldsymbol{\pi}, \mathbf{u})}{\partial \boldsymbol{\pi}} = \left(\frac{f_i(y_i; u_1)}{\sum_{j=1}^{J} \pi_j f_i(y_i; u_j)}, ..., \frac{f_i(y_i; u_J)}{\sum_{j=1}^{J} \pi_j f_i(y_i; u_j)} \right) \tag{4.10}$$

where $i = 1, ..., M$ and define the $J \times M$ matrix $\mathbf{S} = \mathbf{S}(\boldsymbol{\pi}, \mathbf{u}) = (\mathbf{s}_1(\boldsymbol{\pi}, \mathbf{u})^T, ..., \mathbf{s}_M(\boldsymbol{\pi}, \mathbf{u})^T)$. The gradient and Hessian function $l = l(\boldsymbol{\pi}, \mathbf{u})$ are

$$\nabla l = \mathbf{S}\mathbf{1}^T \tag{4.11}$$

$$\nabla^2 l = -\mathbf{S}\mathbf{S}^T \tag{4.12}$$

where $\mathbf{1} = (1, ..., 1)$. It follows that $l(\boldsymbol{\pi}, \mathbf{u}) - l(\boldsymbol{\pi'}, \mathbf{u})$ can be approximated by the second order Taylor Series expansion:

$$l(\boldsymbol{\pi}, \mathbf{u}) - l(\boldsymbol{\pi'}, \mathbf{u}) \cong -\mathbf{1}\mathbf{S}^T(\boldsymbol{\pi'} - \boldsymbol{\pi})^T + \frac{1}{2}(\boldsymbol{\pi'} - \boldsymbol{\pi})\mathbf{S}\mathbf{S}^T(\boldsymbol{\pi'} - \boldsymbol{\pi})^T \tag{4.13}$$

$$
\begin{aligned}
Q(\boldsymbol{\pi'}|\boldsymbol{\pi}, \mathbf{u}) &\equiv -\mathbf{1}\mathbf{S}^T(\boldsymbol{\pi'} - \boldsymbol{\pi})^T + \frac{1}{2}(\boldsymbol{\pi'} - \boldsymbol{\pi})\mathbf{S}\mathbf{S}^T(\boldsymbol{\pi'} - \boldsymbol{\pi})^T \\
&= \frac{1}{2}||\mathbf{S}^T(\boldsymbol{\pi'} - \boldsymbol{\pi})^T - \mathbf{1}^T||^2 - \frac{J}{2} \\
&= \frac{1}{2}||\mathbf{S}^T\boldsymbol{\pi'}^T - \mathbf{2}^T||^2 - \frac{J}{2} \qquad (4.14)
\end{aligned}
$$

where $\mathbf{2} = (2, ..., 2)$ and $||\cdot||$ refers to the L_2-norm. With known support points $\mathbf{u}$, $\boldsymbol{\pi'}$ can then be obtained by solving the least square linear regression problem:

$$\text{Minimize } ||\mathbf{S}^T\boldsymbol{\pi'}^T - \mathbf{2}^T||^2 \qquad (4.15)$$

s.t. $\boldsymbol{\pi'}^T\mathbf{1} = 1$, $\boldsymbol{\pi'} \geq \mathbf{0}$.

Algorithm CNM: Set s=0. From an initial estimate G_0 with finite support and $l(G_0) > -\infty$, repeat the following steps.

- *Step 1:* Denote by $G_s = \sum_{j=1}^{J} \pi_{sj}\delta_{\theta_{sj}}$ the current mixing distribution, where $\sum_{j=1}^{J} \pi_{sj} = 1$ and $\theta_{sj} \in \Omega$, where $j = 1, ..., J$. Compute all local maxima θ^*_{s1}, ..., $\theta^*_{sp_s}$ of $g(\theta; G_s)$, $\theta \in \Omega$. If $\max_j\{g(\theta^*_{sj}; G_s)\} = 0$, stop.

- *Step 2:* Set $\theta^+_s = (\theta^T_s, \theta^*_{s1}, ..., \theta^*_{sp_s})$ and $\pi^+_s = (\pi^T_s, 0, ..., 0)$. Find π^-_{s+1}, the constrained solution of minimizing $Q(\pi'|\pi^+_s, \theta^+_s)$ (4.15). Define G^-_{s+1} that consists of π^-_{s+1} and θ^+_s.

- *Step 3:* Use step-halving or optimization to find $\epsilon_s \in [0, 1]$, respectively, to increase or maximize $l(G_s + \epsilon(G^-_{s+1} - G_s))$.

- *Step 4:* Set $G_{s+1} = G_s + \epsilon_s(G^-_{s+1} - G_s)$ to update π^-_{s+1}.

- *Step 5:* Discard all support points with zero entries in π^-_{s+1}, which gives θ_{s+1} and π_{s+1} of G_{s+1}. Set s=s+1 and go to *Step 1.*

Theorem 1 in Wang (2007) provides the convergence proof which guarantees the CNM method to converge to a nonparametric MLE.

4.3.2 Estimation of baseline hazard function

We now turn our attention to the estimation of $\lambda(t)$ in (4.2) when β and $G(\cdot)$ are known. Since $\hat{G}$ is a step function, we consider a discrete G as in (4.8). For nonparametric estimation of the baseline hazard $\lambda(t)$, we developed a variant of the Breslow (1972) cumulative baseline hazard estimator. Breslow (1972) proposed a non-parametric estimator for the cumulative baseline hazard function in univariate survival analysis, by treating the $\lambda_0(t)$ as piecewise constant between distinct uncensored failure times. Following his idea, we proposed our estimator of $\Lambda_0(t)$ for the multivariate case. For a sample which contains M clusters, we rank all the failure times to N distinct times, $t_{(1)}$, ..., $t_{(N)}$, where $0 \equiv t_{(0)} < t_{(1)} < ... < t_{(N)}$ and the corresponding multiplicities are $\Delta_0 \equiv 0, \Delta_1, ..., \Delta_N$, then assume a constant hazard λ_l between failure time $t_{(l-1)}$ and $t_{(l)}$, where $l = 1, ..., N$. Note that $\Delta_1 = ... = \Delta_N = 1$ corresponds to the case of no ties. The likelihood in model (4.2) therefore becomes

$$\begin{aligned} L(\beta; G(u), \lambda_0(t)) &= \prod_{i=1}^{M}[\prod_{j=1}^{n_i} \lambda_0(t_{ij})^{\delta_{ij}} (e^{Z_i\beta})^{\delta_{ij}}][\sum_{j=1}^{J} \pi_j u_j^{d_i} e^{-u_j e^{Z_i\beta} \sum_{j=1}^{n_i} \Lambda_0(t_{ij})}] \\ &= \prod_{i=1}^{M}[(\prod_{l=1}^{N} {\lambda_l}^{\Delta_{li}})(e^{Z_i\beta})^{d_i}][\sum_{j=1}^{J} \pi_j u_j^{d_i} A_{ij}] \\ &= (\prod_{l=1}^{N} \lambda_l^{\Delta_l}) \prod_{i=1}^{M} e^{d_i Z_i\beta}[\sum_{j=1}^{J} \pi_j u_j^{d_i} A_{ij}] \end{aligned} \tag{4.16}$$

where

$$A_{ij} = e^{-u_j e^{Z_i\beta}\{\sum_{l=1}^{N} m_{li}\lambda_l\}}$$

and $\Delta_{li} = \sum_{j=1}^{n_i} \delta_{ij} I(t_{ij} = t_{(l)})$, which is the number of failures which occurred at failure time $t_{(l)}$ from cluster i; d_i represents the total number of failures in cluster i; and $m_{li} = \sum_{j=1}^{n_i} (\min\{t_{ij}, t_{(l)}\} - t_{(l-1)}) I(t_{ij} \geq t_{(l-1)})$, i.e., the sum of at risk (censoring or failure) times of the subjects from cluster i in the interval $(t_{(l-1)}, t_{(l)})$. The likelihood $L(\beta, G(\cdot), \boldsymbol{\lambda}_0(\cdot))$ thus becomes the function of $\beta, \boldsymbol{u}, \boldsymbol{\pi}$ and $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_N)$.

Suppose the estimates of G and β are known, the cumulative baseline hazard can be estimated by maximizing the log-likelihood function (4.16) directly subject to the constraint $\Lambda_0(1) = 1$ which is specified for identifiability. This is equivalent to solving the constrained optimization problem

$$\text{Maximize } \sum_{l=1}^{N} \ln \lambda_l^{\Delta_l} + \sum_{i=1}^{M} [d_i Z_i \beta + \ln(\sum_{j=1}^{J} \pi_j u_j^{d_i} A_{ij})] \tag{4.17}$$

$$\text{s.t. } \sum_{l=1}^{L_1} \lambda_l (t_{(l)} - t_{(l-1)}) + \lambda_{L_1+1}(1 - t_{(L_1)}) = 1$$

where $L_1 = max\{l : t_{(l)} \leq 1\}$.

There are many algorithms for solving the optimization problem subject to constraints. Our implementation uses a combination of Lagrange Multiplier method and a fixed point algorithm. Given the estimate of G and β, our numerical approach to solve (4.17) for $\boldsymbol{\lambda}$ converges quickly and can be summarized as below.

Algorithm LM-FP: Set $s = 0$. Choose an initial estimate $\lambda_l^{(0)}$, where $l = 1, ..., N$, repeat the following steps.

- *Step 1.* Update $\lambda_l^{(s+1)}$ based on the formulas

$$\lambda_l^{(s+1)} = \frac{\Delta_l}{h_l(\boldsymbol{\lambda}^{(s)}) - \gamma(t_{(l)} - t_{(l-1)})}, \ l = 1, ..., L_1 \tag{4.18}$$

$$\lambda_l^{(s+1)} = \frac{\Delta_l}{h_l(\boldsymbol{\lambda}^{(s)}) - \gamma(1 - t_{(l)})}, \ l = L_1 + 1 \tag{4.19}$$

$$\lambda_l^{(s+1)} = \frac{\Delta_l}{h_l(\boldsymbol{\lambda}^{(s)})}, \ l = L_1 + 2, ..., N \tag{4.20}$$

where

$$h_l(\boldsymbol{\lambda}) = \sum_{i=1}^{M} \frac{\sum_{j=1}^{J} \pi_j u_j^{d_i+1} m_{li} e^{Z_i \beta} A_{ij}}{\sum_{j=1}^{J} \pi_j u_j^{d_i} A_{ij}}$$

and γ can be obtained by solving

$$\sum_{l=1}^{L_1} \frac{\Delta_l(t_{(l)} - t_{(l-1)})}{h_l(\boldsymbol{\lambda}) - \gamma(t_{(l)} - t_{(l-1)})} + \frac{\Delta_{L_1+1}(1 - t_{(L_1)})}{h_{(L_1+1)}(\boldsymbol{\lambda}) - \gamma(1 - t_{(L_1)})} = 1$$

where $\gamma < \min\{\frac{h_l(\boldsymbol{\lambda})}{(t_{(l)} - t_{(l-1)})}, \frac{h_{L_1+1}(\boldsymbol{\lambda})}{(1 - t_{(L_1)})},\ l = 1, ..., L_1 + 1\}$.

- *Step 2.* If $\sum_{l=1}^{N} \mid \lambda_l^{(s+1)} - \lambda_l^{(s)} \mid \leq \epsilon$ for prespecified ϵ, then stop; otherwise, go back to *Step 1.*

4.4 Algorithm

To maximize the likelihood function with respect to $\beta, G(u, \pi)$ and λ, we represent a three-step iterative algorithm below that iterates between the estimation of frailty distribution G, baseline hazard λ and the estimation of β while, at each step, fixing others.

In general, the three-step algorithm can be summarized as follows:

- *Step 0.* Set $k = 0$. Choose initial estimates of G, $\boldsymbol{\lambda}$ and β as $G^{(0)}$, $\boldsymbol{\lambda}^{(0)}$ and $\beta^{(0)}$, respectively. Typically, we can set $G^{(0)} = \delta_1$, $\boldsymbol{\lambda}^{(0)} = (1, ..., 1)^T$ such that $\Lambda^{(0)}(1) = 1$, and $\beta^{(0)} = 0$. Repeat the following steps.

- *Step 1.* Update $G^{(k+1)}$ by using Algorithm ISDM or Algorithm CNM, based on the estimates of $\boldsymbol{\lambda}^{(k)}$, $\beta^{(k)}$ and $G^{(k)}$.

- *Step 2.* Update $\boldsymbol{\lambda}^{(k+1)}$ from Algorithm LM-FP, given the current estimates of $\boldsymbol{\lambda}^{(k)}$, $\beta^{(k)}$ and $G^{(k+1)}$.

- *Step 3.* Update $\beta^{(k+1)}$ by optimizing the logarithm of likelihood function 4.16 directly, conditional on the updated estimates $\boldsymbol{\lambda}^{(k+1)}$ and $G^{(k+1)}$. Repeat *Step 1* to *Step 3* until the estimates of β converges, i.e. $|\beta^{(k+1)} - \beta^{(k)}| \leq \epsilon$ for prespecified ϵ.

The estimate of β obtained at the final step is the NPMLE of β, $\hat{\beta}_n$. In the algorithm above, the updating of parameter β involves maximization of a non-linear function. We adopted function *optim()* in R in our implementation.

It is worthy noting that the computation of local maximum of the gradient function in **Algorithm ISDM** or **Algorithm CNM** is based on a combined Newton and bisection method. Newton's method requires that the derivative of the function be calculated directly. If the initial value is outside the region of convergence, Newton's method may fail to converge. For this reason, we used a combined Newton and bisection method to avoid the scenario when an iterate is generated that lies outside the interval.

Algorithm Newton-Bisection: Set s=0. Consider the function $g'(\theta; G_s) = \partial g(\theta; G_s)/\partial\theta$, and set an initial interval $a^{(0)} = a$ and $b^{(0)} = b$ with $g'(a; G_s) \geq 0$ and $g'(b; G_s) \leq 0$. Let the initial estimate $\theta^{(0)} = (a + b)/2$, repeat the following steps.

- *Step 1.* If $g'(a^{(s)}; G_s) \times g'(\theta^{(s)}; G_s) > 0$ then set $a^{(s+1)} = \theta^{(s)}$; otherwise, if $g'(b^{(s)}; G_s) \times g'(\theta^{(s)}; G_s) > 0$ then set $b^{(s+1)} = \theta^{(s)}$.

- *Step 2.* Let $\theta_*^{(s+1)} = \theta^{(s)} - g'(\theta^{(s)}; G_s)/g''(\theta^{(s)}; G_s)$; if $\theta_*^{(s+1)} \in [a^{(s+1)}, b^{(s+1)}]$ then let $\theta^{(s+1)} = \theta_*^{(s+1)}$; otherwise, let $\theta^{(s+1)} = (a^{(s+1)} + b^{(s+1)})/2$.

- *Step 3.* Repeat *Step 1* and *Step 2* until the estimates of θ converges, i.e. $|\theta^{(s+1)} - \theta^{(s)}| \leq 1 \times e^{-6}$.

According to our experience, the three-step algorithm we proposed above generally converges to the global maximum NPMLE. However, this is not guaranteed and the proof of the convergence does not seem to be easy. In fact, the mixing distribution $\hat{G}$ found by using ISDM or CNM methods at *Step 1* is proved to be the global maximum NPMLE given $\boldsymbol{\lambda}$ and β. However, finding the baseline hazard $\boldsymbol{\lambda}$ and β in *Step 2 and 3* by maximizing the log-likelihood function may not be convex optimization problems. It should be noted, however, that even if all these algorithms were convex optimization

problems, the combination of the three algorithms is not guaranteed to be a convex optimization problem, nor can we assure that this algorithm would be guaranteed to converge to the overall MLE. In practice, we suggest that users try different starting points to check for convergence to the same estimates.

One can obtain the confidence interval for β by approximating the likelihood ratio distribution with the χ^2 distribution with one degree of freedom. Cox and Hinkley (1979), for instance, consider interval estimation such as likelihood ratio based confidence interval for single and multiple parameters, and the case of a single parameter of interest with several nuisance parameters in the parametric model. Murphy et al. (1997), Murphy and van der Vaart (2000) and Van der Vaart (1996) show that results associated with the profile likelihood can be extended to certain semiparametric problems. These problems would include inferences in semiparametric models where only G or only Λ_0 was nonparametric. Our situation is an extension of this. So we state a hypothesized result that we expect to hold in the doubly nonparametric model. Specifically, define the profile loglikelihood function of β as

$$pl(\beta) = l(\beta, (\hat{\Lambda}_{0,\beta}, \hat{G}_\beta)) \tag{4.21}$$

and treat both Λ_0 and G as the nuisance parameters. Consider the hypothesis H_0 : $\beta = \beta_0$. The likelihood ratio criterion to test H_0 is then

$$\lambda(\beta_0) = L(\beta_0, (\hat{\Lambda}_{0,\beta_0}, \hat{G}_{\beta_0}))/L(\hat{\beta}, (\hat{\Lambda}_{0,\hat{\beta}}, \hat{G}_{\hat{\beta}})) \tag{4.22}$$

If the semiparametric results carry over to this case, we would expect that $-2log\lambda(\beta_0)$ has an asymptotic χ^2 distribution with one degree of freedom. Hence, the P-value is obtained as $Pr\{\chi^2_{(1)} \geq -2log\lambda(\beta_0)\}$. The $100(1-\alpha)\%$ confidence set for $\hat{\beta}$ can be obtained by inverting this test. In this case, we obtain $\{\beta : -2log\lambda(\beta) \leq \chi^2_{(1),\alpha}\}$. This yields an interval and the upper and lower confidence bounds for $\hat{\beta}$ can be given

by points β_L and β_U on the profile likelihood curve for β for which $-2log\lambda(\beta_L) = -2log\lambda(\beta_U) = \chi^2_{(1),\alpha}$.

In practice, there exist a number of numerical techniques which allow β_L and β_U to be determined in a simple one-dimensional search (Gill et al., 1981; Richard, 1988). In our case, we adopt a bisection method to search for the upper and lower confidence limits of the ML estimate of β.

4.5 Comparison

Therneau et al. (2003) considered a semiparametric approach and obtained the solution via penalized models. In their approach, they assumed a parametric distribution (e.g. Gamma, Gaussian) with unit mean and unknown variance θ for the frailty parameter. For any fixed θ, the frailty term is treated as an additional regression coefficient in the usual Cox partial log-likelihood function, but the values of the frailty was restricted by a penalty function that depends on θ. Typically, θ was chosen to control the amount of restrictions that "shrink" the frailty towards zero.

Consider an alternative version of the hazard (4.1),

$$\lambda(t_{ij}) = \lambda_0(t_{ij})u_i e^{Z_i\beta} = \lambda_0(t_{ij})e^{Z_i\beta + w_i} \tag{4.23}$$

where $w_i = exp(u_i)$. The estimation is done by maximizing a penalized log partial likelihood function

$$PPL(\beta, w, \theta) = PL(\beta, w) - g(w; \theta) \tag{4.24}$$

where $PL(\beta, w)$ is the usual log partial likelihood

$$PL(\beta, w) = \sum_{i=1}^{M} \int_0^{\infty} [Y_i(t)(Z_i\beta + w_i) - log\{\sum_k Y_k(t)e^{Z_k\beta + w_k}\}]dN_i(t) \tag{4.25}$$

and g is the penalty function. For gramma frailties, Therneau et al. (2003) showed

that, with penalty function $g(w;\theta) = -1/\theta \sum_{i=1}^{M}[w_i - exp(w_i)]$, the solution to the penalized partial likelihood model coincides with the solution obtained from the EM algorithm for any fixed value of θ. For Gaussian frailties with variance θ, the penalty function was suggested as $g(w;\theta) = -\sum_{i=1}^{M} w_i^2/(2\theta)$.

Therneau et al. (2003) provided a fitting algorithm based on an inner and outer loop. For any fixed θ, the Newton-Raphson algorithm was used to fit the penalized likelihood by solving the score function $\partial PPL/\partial\beta = 0$ and $\partial PPL/\partial w = 0$. The outer loop evaluated θ by maximizing the profile likelihood of β. The baseline hazard was identified by using the Breslow estimator after β, θ and w were estimated. In the next section, we compare our proposed approach with this semiparametric method in terms of efficiency and robustness.

4.6 Simulation Study

Simulations were conducted to evaluate the properties of our proposed nonparametric method in finite samples and to evaluate the asymptotic approximation. In addition, we compare our approach with the semiparametric method of Therneau et al. (2003) with the focus on estimation of β. One single binary covariate, Z, was generated taking values 1 or 0 with probability 0.5. Subjects in the same cluster were assumed to receive a common treatment assignment. The censoring time was taken to follow a continuous uniform distribution on [0,5] or [4,9]. Given the frailty U and the covariate Z, a subject's event time was generated from an exponential distribution with rate $Ue^{Z\beta}$. Thus $\Lambda_0(t) = t$.

Simulation settings varied with respect to number of clusters ($M = 30, 60, 120$), cluster size ($n = 3, 15$), magnitude of the treatment effects ($\beta_0 = 0, 0.7, 1.0$) and the frailty distribution. We examined two different models for the true underlying frailty distributions. First, U followed a gamma distribution with unit mean and variance

0.11, so that the frailty distribution was correctly specified in the semiparametric model. Second, U followed a beta distribution $Beta(0.5, 0.5)$ with mean 0.5 and variance 0.125; this distribution for the frailty corresponded to a substantial misspecification in the semiparametric model where the frailty was taken to be Gaussian. For each setting, we compared the performances of our proposed nonparametric method and semi-parametric approach by Therneau et al. (2003) with respect to bias, mean squared error (MSE) and empirical coverage of 95% confidence intervals. When we applied the semiparametric method, we chose the "corrected AIC" method to select a solution for θ, the variance of the frailty distribution as suggested by Verweij and Van Houwelingen (1994) and Hurvich et al. (1998).

4.6.1 Frailty distribution correctly specified in the semiparametric method

Table 4.1 presents the results from the nonparametric and semiparametric approach when the frailty follows a $Gamma(9, 9)$ distribution with mean 1 and variance $\theta = 1/9$. In this setting, the frailty distribution is correctly specified in the semiparametric method. The goal is to examine the relative efficiency loss of the nonparametric approach when the parametric assumption holds in the semiparametric model. The simulation study is based on 1000 simulated samples. It can be seen that the nonparametric approach generally performs well compared with the semiparametric method. There is virtually no evidence of bias in the estimation of β from the nonparametric approach in any of the cases considered. However, the estimation of β from the semiparametric approach seems to have some bias towards the null. As expected, the empirical standard errors for the estimates from the nonparametric approach are relatively larger than those from the semiparametric method, which leads to a slight efficiency loss, ranging from 10% to 29%. However, the loss in efficiency decreases as the number of clusters increases, suggesting that the nonparametric approach becomes as efficient as the semiparametric methods when the

number of clusters is large. The converge probabilities are close to the nominal level from both approaches. It is worthy noting that increasing the number of clusters or cluster size generally reduces the mean squared error of treatment effect estimates, but the coverage performance of the nonparametric approach seems to depend more on the number of clusters rather than cluster size, while the semiparametric method seems to perform better given larger cluster size. For example, the nonparametric method seems to perform equally well with respect to the coverage probability in a sample of 360 subjects with $M = 120$ and $n = 3$ and one of 900 subjects with $M = 60$ and $n = 15$.

4.6.2 Frailty distribution misspecified in the semiparametric method

Table 4.2 summarizes the results from the nonparametric and semiparametric approaches, respectively, when U follows a $Beta(0.5, 0.5)$ distribution. In this setting, the frailty distribution is misspecified as Gaussian in the semiparametric model. We aim to access the robustness performance of the two methods when the parametric assumption for U is violated. There is substantial bias in the estimation of β in the semiparametric approach towards the null, and the bias grows larger as $|\beta|$ increases. In contrast, the nonparametric approach remains approximately unbiased for all the cases considered. It can be seen that the nonparametric estimate is generally more efficient as compared to the semiparametric estimate for $\beta \neq 0$ and the efficiency gain becomes more apparent as β grows from zero and sample size increases. For example, when $\beta = 1.0$ and $M = 120$, the semiparametric estimate is only 67.8% efficient relative to the nonparametric estimate. There is an interesting contrast between the two methods in terms of coverage probability. If the sample size is small ($M = 30$, $n = 3$), both approaches cover the true parameter almost equally well (95% C.I. $\approx$ 89%). However, as the number of clusters increases, the nonparametric approach demonstrates a substantial improvement in the coverage probability, whereas

Table 4.1: Comparison of the nonparametric approach (NP) with the semiparametric method (SP) with the focus on estimation of β. Frailty follows a gamma distribution with shape and rate parameter of 9 and the frailty distribution is correctly specified in the semiparametric model. Censoring time follows a uniform distribution on [0,5]. Number of replications=1000.

M	n	% censoring	β	Bias		ESE		MSE		RE	95% CI	
				NP	SP	NP	SP	NP	SP		NP	SP
120	15	16.48		-0.003	0.001	0.092	0.084	0.008	0.007	1.199	0.931	0.947
120	3	16.41		-0.002	-0.024	0.148	0.136	0.022	0.019	1.141	0.940	0.928
60	15	16.37	0.7	-0.006	-0.006	0.125	0.112	0.016	0.013	1.240	0.940	0.947
60	3	16.40		0.007	-0.022	0.209	0.191	0.044	0.037	1.181	0.943	0.921
30	15	16.36		-0.012	-0.014	0.179	0.157	0.032	0.025	1.295	0.927	0.947
30	3	16.41		0.006	-0.021	0.316	0.281	0.100	0.079	1.254	0.940	0.924
120	15	21.78		0.002	0.003	0.090	0.083	0.008	0.007	1.169	0.921	0.950
120	3	21.71		-0.002	0.001	0.140	0.134	0.020	0.018	1.096	0.947	0.941
60	15	21.70	0	-0.004	-0.003	0.125	0.113	0.016	0.013	1.220	0.937	0.949
60	3	21.72		-0.001	-0.002	0.205	0.191	0.042	0.036	1.152	0.950	0.934
30	15	21.71		-0.012	-0.011	0.178	0.157	0.032	0.025	1.294	0.934	0.948
30	3	21.71		-0.011	-0.009	0.308	0.282	0.095	0.080	1.188	0.938	0.923

M: Number of clusters. n: Cluster size. ESE: Empirical standard error. MSE: Mean squared error. RE: Relative efficiency. 95% CI: 95% confidence interval.

the performance of the semiparametric method deteriorates given $\beta \neq 0$. These results suggest that the nonparametric framework would need relative large number of clusters to capture the shape of the frailty distribution. As we expect, the proposed nonparametric approach appears to be more robust than the semiparametric method under the misspecification of the frailty distribution.

4.7 Application

In this section, we consider the use of the proposed nonparametric model to the *INSTINCT* study. As introduced in Chapter II, the *INSTINCT* study is a cluster randomized trial, designed to investigate the effectiveness of an educational intervention administered to 24 Michigan hospitals in enhancing tPA therapy use in stroke patients. Previous analysis revealed that there was a marginally significant difference between the intervention and control sites with respect to the proportion of stroke patients that were treated with tPA therapy. This was the primary purpose of the *INSTINCT* study. Stroke is a time sensitive disease and the acute stroke treatment should be accomplished as quickly as possible. The current target for initiation of tPA treatment is within 60 minutes of hospital arrival. One secondary hypothesis of the *INSTINCT* trial was that the educational intervention would improve door to treatment times (DTT). In this section, we fit the proposed nonparametric model to evaluate this secondary hypothesis. The semiparametric model proposed by Therneau et al. (2003) was also applied for comparison.

A total of 557 stroke patients who received tPA treatment during the INSTINCT trial were included in this analysis; 54% were treated in the 12 treatment hospitals. No observation was censored or dropped out the study. The outcome of interest was door to treatment time in hours. DTT per patient ranged from 0.13 hours to 13.42 hours with an median of 1.33 hours. The covariate considered was the education intervention operated on the hospital level. Patients were clustered in hospitals and

Table 4.2: Comparison of the nonparametric approach (NP) with the semiparametric method (SP) with the focus on estimation of β. Frailty follows a beta distribution with both shape parameters as 0.5 and the frailty distribution is misspecified in the semiparametric model. Censoring time follows a uniform distribution on [4,9]. Number of replications=1000.

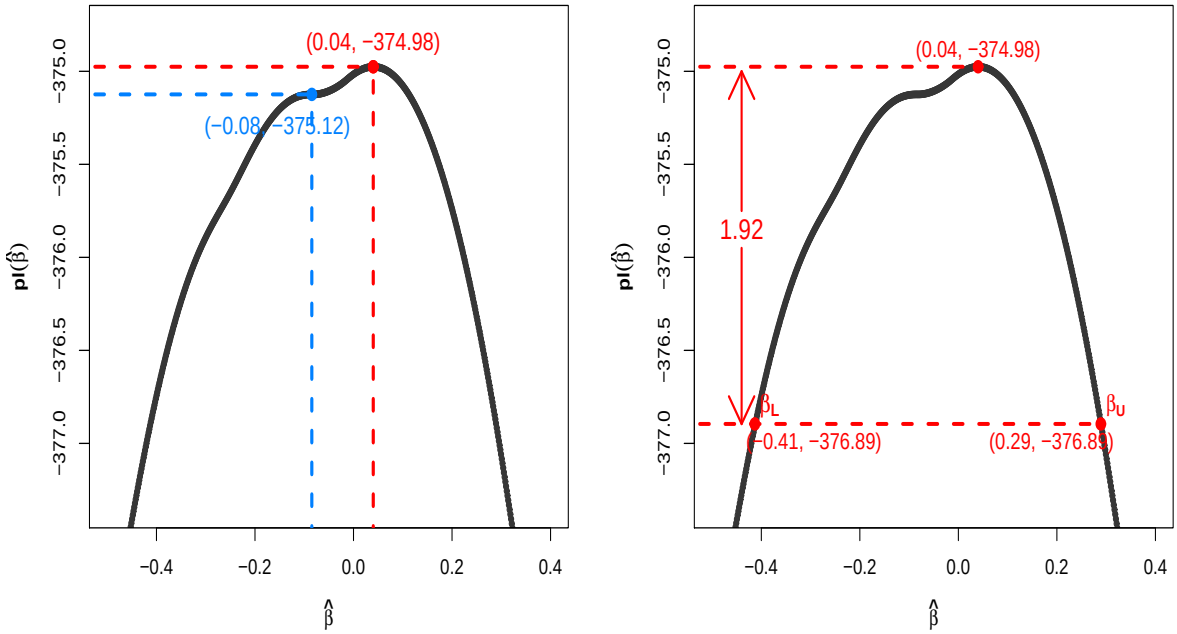
M	n	% censoring	β	Bias		ESE		MSE		RE	95% CI	
				NP	SP	NP	SP	NP	SP		NP	SP
120	3	18.87		0.007	-0.173	0.234	0.227	0.055	0.081	0.675	0.936	0.831
60	3	18.97	1.0	0.001	-0.184	0.353	0.318	0.125	0.135	0.925	0.930	0.859
30	3	19.15		-0.009	-0.149	0.561	0.517	0.314	0.289	1.087	0.897	0.893
120	3	20.06		0.011	-0.125	0.229	0.220	0.052	0.064	0.820	0.930	0.867
60	3	20.16	0.7	0.012	-0.134	0.353	0.311	0.125	0.115	1.085	0.924	0.890
30	3	20.37		-0.004	-0.092	0.564	0.563	0.318	0.326	0.977	0.888	0.896
120	3	23.81		0.001	0.001	0.227	0.210	0.051	0.044	1.165	0.927	0.911
60	3	23.94	0.0	-0.006	-0.007	0.348	0.293	0.121	0.086	1.408	0.921	0.919
30	3	24.13		-0.002	0.022	0.572	0.476	0.326	0.227	1.438	0.886	0.899

M: Number of clusters. n: Cluster size. ESE: Empirical standard error. MSE: Mean squared error. RE: Relative efficiency. 95% CI: 95% confidence interval.

the intra-cluster correlation were taken into account by the frailty term.

In the application of the nonparametric approach, we tried a number of initial values for $G^{(0)} = \delta_1$, $\delta_{0.5}$ or $\delta_{1.5}$, which are the points on the Γ curve, and obtained an initial estimates of the cumulative baseline hazard $\mathbf{\Lambda}^{*(0)}$ by fitting the Cox proportional hazard model and ignoring the correlation. We then let $\mathbf{\Lambda}^{(0)} = \mathbf{\Lambda}^{*(0)}/\Lambda^*(1)$. Thus, $\mathbf{\Lambda}^{(0)}$ satisfied the constraint $\mathbf{\Lambda}^{(0)}(1) = 1$. An alternative choice for $\mathbf{\Lambda}^{(0)}$ was the observed distinct time to treatment. We also tried several different starting values for $\beta^{(0)} = -0.5, -0.1, -0.05, 0, 0.1, 0.5$, respectively, for the nonparametric approach and plotted the profile likelihood (Figure 4.3). With the starting values $\beta^{(0)}$ at -0.05, 0, 0.1 and 0.5 and any of the starting points considered for $G^{(0)}$ or $\mathbf{\Lambda}^{(0)}$, the nonparametric algorithm converged to the global maximum $\hat{\beta}$=0.04 as shown in Figure (4.3)(a). However, if we started at $\beta^{(0)} = -0.5, -0.1$ and various initial values for $G^{(0)}$ or $\mathbf{\Lambda}^{(0)}$, this algorithm led to a local maximum of -0.08. We obtained 95% confidence set for $\hat{\beta}$ by inverting the likelihood ratio test. As shown in Figure 4.3(b), the upper and lower confidence bounds for $\hat{\beta}$ can be given by points $\beta_L = -0.41$ and $\beta_U = 0.29$ on the profile likelihood curve for β for which $-2log\lambda(\beta_L) = -2log\lambda(\beta_U) = 3.84$. According to this result, the hazard ratio between the predicted hazard rate for a patient who was treated in the intervention hospital and that for one admitted in the control hospital was $e^{0.04} = 1.04$, although the effects of intervention was not significant.

In contrast, the semiparametric model with a specified Gamma frailty of Therneau et al. (2003) gave an estimate of β at -0.08 with a 95% confidence interval of (-2.22, 2.05). This result indicated that the semiparametric approach converged to the local maximum and the interval estimation based on the curvature of the local maximum led to a fairly wide estimate of the confidence set. Trying different starting values may solve this problem. However, the *frailty* routine in R that implements the Therneau et al. (2003) approach sets the default starting value and does not allow users to change the default. We also carried out a naive analysis by fitting a Cox proportional



(a) Finding local and global maximums on $pl(\hat{\beta})$ (b) Finding 95% C.I. based on Likelihood Ratio Test

Figure 4.3: The Profile Likelihoods.

hazard model which ignores the correlation, and obtained $\hat{\beta} = -0.234$ with a P-value as 0.006. Without taking into account the intra-cluster correlation, the naive analysis appears to lead to incorrect parameter estimates and inference.

4.8 Discussion and Future Work

In this section, we have considered a Cox frailty model for clustered survival data with both the frailty distribution, G, and cumulative baseline hazard, Λ_0, left non-parametric. We propose an approach based on nonparametric maximum likelihood estimation. The key improvement of our proposed method is that the distributional assumptions on both frailty and baseline hazard are completely dropped. Therefore, our approach should be robust and applicable to whatever distribution holds for the frailty or baseline hazard. This is the main advantage of the analysis compared to

other approaches, in which, parametric models are used for the frailty distribution or the baseline hazard or both.

The relaxation of the parametric assumptions may result in efficiency loss compared to other parametric or semiparametric approaches when the corresponding assumptions hold. Simulation studies suggest that, at least under the situations considered, the relative efficiency loss of our proposed approach is relatively small when compared to the popular semiparametric approach proposed by Therneau et al. (2003). As the number of clusters increases (e.g. $M = 120$ and $n = 3$), the nonparametric approach becomes almost as efficient as the semiparametric method (Efficiency Loss = 9.6%). On the other hand, simulation studies also confirm the robustness advantage of the nonparametric approach. When the frailty distribution is misspecified in the semiparametric model, the nonparametric approach appears to be more efficient and avoid bias as compared to the semiparametric approach for the estimation of β when β is not near zero. In addition, the nonparametric approach, with a moderate number of clusters (e.g. $M = 60, 120$), demonstrates higher coverage probability compared to the semiparametric approach. The gain in efficiency and coverage probability becomes more apparent as the treatment effect grows from zero and the sample size increases.

Another virtue of the proposed method is its computational efficiency. The step of finding the NPMLE $L_{\hat{G}}$ plays an important role in the three-step algorithm. There are several existing methods in the literature which also involve the computation of a frailty distribution nonparametrically by using the EM algorithm, the vertex direction method, etc. As is often the case, however, those algorithms are slow to converge and the corresponding methods are much more computationally expensive than our proposed method based on ISDM or CNM for the estimation of NPMLE $L_{\hat{G}}$. Although both ISDM and CNM methods are very efficient, there are still areas to further improve the computational efficiency. One advantage of ISDM is that

it does not require that one keeps track of the accumulated set of support points, but only the current likelihood point $L_G = (L_1(G), ..., L_M(G))$. CNM includes a nice approach to update the weights of support points accurately at each step but requires that one keeps track of the set of support points at each iteration. Therefore, a new algorithm based on ISDM and CNM method, which avoids the book keeping of the set of support points in each step but adopts a variation to the approach in CNM for updating the weights, will be proposed in the future work.

We have assumed nonparametric forms of the frailty U and cumulative baseline hazard Λ_0. The asymptotic properties are therefore difficult to verify fully. Murphy and van der Vaart (2000) provide the asymptotic properties of the semiparametric profile likelihood for a shared gamma frailty model, and their results are conjectured to extend to our more general approach. The simulation studies in the previous section suggest that they do. We can also consider the particular semiparametric model by assuming a piecewise constant distribution for the baseline hazard and leaving the frailty distribution, G, completely unspecified. Murphy and van der Vaart (2000) results would apply to this case. By increasing the number of components in the piecewise constant formulation, we can approach to the fully nonparametric model, and again these considerations suggest that the asymptotic results should extend to the nonparametric case. A detailed proof of the asymptotic results in the nonparametric case is a matter for future work. Simulation studies indicate, however, the proposed estimators are empirically unbiased with coverage probability close to the nominal level and should be valid in many practical settings as expected.

The proposed method assumes common treatment assignment within cluster and can be applied to data with this characteristic. This approach can be extended to handle data where the covariates as treatments vary among subjects within cluster. As suggested by Neuhaus and Kalbfleisch (1998), when the between- and within-cluster covariate effects are different, models that assume that these effects are the

same would be misleading.

Finally, for identifiability, we place a constraint on the baseline hazard such that $\Lambda_0(1) = 1$. There is an alternative option which is to restrict the mean of frailty distribution by requiring $E(G) = 1$. The latter restriction would require the use of a constrained ISDM algorithm as developed by Susko (1996) in finding NPMLE of $L_{\hat{G}}$. The proposed algorithm can be modified to accommodate the constraint on the frailty distribution.

CHAPTER V

Conclusions and Future Work

5.1 Conclusions

In this dissertation we have studied the design and analysis of cluster randomized trials (CRTs). Regarding design, we proposed a new randomization procedure, with the general aim of reducing the mean squared error of the treatment effect estimator. The key innovation of this design is that it reduces the chance imbalance in observed covariates remained after randomization by matching on the estimated propensity score and choosing the best of several randomizations. Regarding analysis, we studied the nonparametric regression for correlated failure time data as often arise in a CRT. We extended the Cox proportional hazards model with a frailty term to allow for flexible structure for both frailty and hazard and developed an approach based on the nonparametric maximum likelihood estimation. Therefore, the main advantage of this approach is its robustness to the misspecification of distributional assumption on either the frailty or baseline hazard or both.

In Chapter II, we proposed the BMW design which, in essence, applied the optimal full matching with constraints technique to randomization in order to achieve overall balance between treatment groups and control the variance of the treatment comparison, and so yield good MSE properties. In such studies, there are typically rather few participating units and often several variables that describe the properties

of these units. It is then important to attempt to balance across these factors and one approach to this is through the use of matched pairs or other blocks defined by a subset of the covariates. The BMW design involved considering several (M) randomizations of the participating units into the two treatment groups. With each randomization, the technique of optimal full matching with constraint k was used in order to identify the best blocking for that randomization. The distance measure for the matching was based on estimated propensity scores, as has been proposed in observational studies. We then chose the randomization and corresponding full matching that led to the smallest total distance. The parameters M and k were examined and recommendations made. A simulation study showed that, under various confounding scenarios, the BMW design had good properties and can yield substantial reductions in the MSE of the estimates of the treatment effect as compared to various designs and analysis methods that have been proposed in the literature. The design was also seen to be robust against heterogeneous error. We illustrated these methods in proposing a design for the INSTINCT trial.

In Chapter III, we extended the BMW design to clinical trials with more than two arms or with staggered entry. First, We investigated the use of the three tripartite matching algorithms, incomplete block with disjoint pairs (Bo and Rosenbaum, 2004), two-way and three-way tripartite matching with triples, in the generalization of the BMW design to clinical trials with three arms and used a simulation study to compare the performance of the design under various confounding scenarios. The numerical analysis suggested that in general, the three-arms BMW designs led to important gains in efficiency by reducing the MSE in all three treatment effect estimators simultaneously, especially for the design based on the two-way and three-way tripartite matching methods we proposed. It is worthy noting that, the BMW design based on the incomplete block with disjoint pairs had the drawback of efficiency loss especially when the confounding effects were not too strong (e.g. $\gamma = 0.5$ or 1.0).

The addition of random blocks to the data led to a loss of efficiency which became even more apparent in small studies. On the other hand, our proposed tripartite matching algorithms with triples minimized the efficiency loss, hence, appeared to be substantially more effective in reducing the MSE. In an effort of extending the BMW design to accommodate the staggered entry, we found that the generalized design was effective in reducing MSE of the treatment effect estimator compared to the completely randomized design and permuted block design within strata through simulation studies.

Once data are collected, then there is a problem of analysis. One most notable characteristic of the data raised from CRTs is the dependencies among cluster members. Chapter IV considers the nonparametric regression analysis of correlated failure time data based on a Cox model with a frailty term. As discussed in the literature, Heckman and Singer (1984a) found that the parameter estimates are sensitive to the choice of distribution for the unobservable heterogeneity while assuming a piecewise constant distribution for the baseline hazard, and Trussell and Richards (1985) further discovered that even with a nonparametric representation of heterogeneity, results can still be sensitive to choice of a model for the baseline hazard. Therefore, in this chapter, we extended a frailty model by allowing both the frailty distribution, G, and the cumulative baseline hazard, Λ_0, left nonparametric and proposed an approach based on nonparametric maximum likelihood estimation.

For implementation, we developed a three-step iterative algorithm. First, estimate the frailty distribution G nonparametrically based on **Algorithm ISDM** or **Algorithm CNM**, given the current estimates of Λ_0 and β; second, update Λ_0 nonparametrically by using the **Algorithm LM-FP** and the current estimates of G, β. Finally, compute β by optimizing the logarithm of likelihood function directly, conditional on the updated estimates λ and G.

To evaluate the finite sample property of the proposed iterative algorithm, we

carried out a simulation study by comparing it with the semiparametric approach by Therneau et al. (2003). The results showed that our proposed fully nonparametric approach provided important gains in robustness by resulting in reasonable small loss in efficiency compared to the semiparametric approach. Specifically, the point estimator from our approach was unbiased, meanwhile, the estimator for the confidence interval based on the converted likelihood ratio test worked well and the empirical coverage probabilities of the 95% confidence intervals were quite close to the nominal value given moderate number of clusters, say 60 or more.

5.2 Futurework

5.2.1 Randomization Test based on the BMW design

The proposed BMW design appears to be very effective in reducing the MSE of the treatment effects estimator. However, one might ask how to use this design to do a randomization test. We argued that the BMW design with a practically reasonable choice of M, can also form the basis of a randomization test.

To illustrate the basic idea of the test based on the BMW design, suppose, for example, that a study sample of size N has been collected using the BMW design with given k and M which contains two groups A and B whose sample group means of the observed outcomes are $\bar{x}_A$ and $\bar{x}_B$, respectively. Without loss of generality, we assume that the two groups have the same size of $N/2$ each. And that we aim to test, the null hypothesis $H_0 : u_1 = u_2$, where u_1 and u_2 refer to the true mean of the two groups. The test proceeds as follows. First, the sample test statistic is calculated as the difference in means between the two samples, $t_{obs} = \bar{x}_A - \bar{x}_B$. Next, the BMW design with the same k and M is repeated for a large number of B times and each time, the test statistic based on the fixed outcomes observed is computed. The set of these calculated test statistics is the exact reference distribution of possible

differences under the null hypothesis that group label does not matter. The one-sided (two-sided) p-value of the test is calculated as the proportion of the reference set where the (absolute) difference in means is greater than or equal to t_{obs}.

To construct the approximate 95% confidence interval based on the BMW design, the approach can be illustrated as follows. The lower confidence limit for t_{obs} is obtained by finding the value of C_L such that if we subtract the value of C_L from each member of group A, then the randomization test described above based on the BMW design using the adjusted data gives a new value of $t^L_{obs} = \bar{x}_A - C_L - \bar{x}_B$ corresponding to the upper 2.5% point of the new randomization distribution. Likewise, the upper limit of the 95% confidence interval is obtaining by finding the value of C_U that if we add the value to group A and repeat the randomization test in the same manner, the new $t^U_{obs} = \bar{x}_A + C_U - \bar{x}_B$ corresponds to the lower 2.5% point of the adjusted randomization distribution. In this process, the BMW design is involved in obtaining the reference distribution based on the adjusted data.

For a practical value of M (e.g. $M = 10$, 20 or 100), this would typically yield a reasonably large reference set as the basis of the test. For example, say $N = 20$, then we will end up with $\binom{20}{10} = 184756$ possible designs. Even M is as large as 100, it is still very small compared to the number of randomizations possible and so the reference set would still be quite large. An illustrative example will be given in the future.

5.2.2 ISDM-CNM

One advantage of ISDM is that it does not require that one keeps track of the accumulated set of support points, but only the current likelihood point $L_G = (L_1(G), ..., L_M(G))$. CNM includes a nice approach to update the weights of support points at each step, by solving a quadratic programming subproblem via a least square linear regression formulation. This method can find exact zero weights and

compute small positive masses accurately. However, CNM requires that one keeps track of the set of support points at each iteration. Therefore, a new algorithm based on ISDM and CNM method, which avoids the book keeping of the set of support points in each step but adopts a variation to the approach in CNM for updating the weights, can be proposed. More specifically, for the s^{th} iteration, in the *step 2* of *Algorithm ISDM*, we need solve the constrained optimization problem

$$\text{Maximize } K(\epsilon_0, \epsilon_1, ..., \epsilon_{p_s}) = \sum_{i=1}^{M} log\{\epsilon_0 L_{si} + \sum_{j=1}^{p_s} \epsilon_j L_i(\theta^*_{sj})\} \tag{5.1}$$

s.t. $\sum_{j=0}^{p_s} \epsilon_{sj} = 1$ and $\epsilon_{sj} \geq 0$

with respect to $\boldsymbol{\epsilon} = (\epsilon_0, ..., \epsilon_{p_s})$. We can solve this via a linear regression formulation in a similar manner as CNM. Denote

$$K_i(\boldsymbol{\epsilon}) = log\{\epsilon_0 L_{si} + \sum_{j=1}^{p_s} \epsilon_j L_i(\theta^*_{sj})\}$$

$$\mathbf{v}_i(\boldsymbol{\epsilon}) = \frac{\partial K_i(\boldsymbol{\epsilon})}{\partial \boldsymbol{\epsilon}} = \left(L_{si}, L_i(\theta^*_{s1}), ..., L_i(\theta^*_{sp_s})\right) / (\epsilon_0 L_{si} + \sum_{j=1}^{p_s} \epsilon_j L_i(\theta^*_{sj})) \tag{5.2}$$

where $i = 1, ..., M$ and define the $(p_s+1)\times M$ matrix $\mathbf{V} = \mathbf{V}(\boldsymbol{\epsilon}) = (\mathbf{v}_1(\boldsymbol{\epsilon})^T, ..., \mathbf{v}_M(\boldsymbol{\epsilon})^T)$. The gradient and Hessian function $K = K(\boldsymbol{\epsilon})$ are

$$\nabla K = \mathbf{V}\mathbf{1}^T \tag{5.3}$$

$$\nabla^2 K = -\mathbf{V}\mathbf{V}^T \tag{5.4}$$

where $\mathbf{1} = (1, ..., 1)$. It follows that $K(\boldsymbol{\epsilon}) - K(\boldsymbol{\epsilon'})$ can be approximated by the second order Taylor Series expansion:

$$K(\boldsymbol{\epsilon}) - K(\boldsymbol{\epsilon'}) \cong -\mathbf{1}\mathbf{V}^T(\boldsymbol{\epsilon'} - \boldsymbol{\epsilon})^T + \frac{1}{2}(\boldsymbol{\epsilon'} - \boldsymbol{\epsilon})\mathbf{V}\mathbf{V}^T(\boldsymbol{\epsilon'} - \boldsymbol{\epsilon})^T \tag{5.5}$$

$$
\begin{aligned}
Q(\boldsymbol{\epsilon'}|\boldsymbol{\epsilon}) &\equiv -\mathbf{1}\mathbf{V}^T(\boldsymbol{\epsilon'}-\boldsymbol{\epsilon})^T + \frac{1}{2}(\boldsymbol{\epsilon'}-\boldsymbol{\epsilon})\mathbf{V}\mathbf{V}^T(\boldsymbol{\epsilon'}-\boldsymbol{\epsilon})^T \\
&= \frac{1}{2}||\mathbf{V}^T(\boldsymbol{\epsilon'}-\boldsymbol{\epsilon})^T - \mathbf{1}^T||^2 - \frac{(p_s+1)}{2} \\
&= \frac{1}{2}||\mathbf{V}^T\boldsymbol{\epsilon'}^T - \mathbf{2}^T||^2 - \frac{(p_s+1)}{2} \qquad (5.6)
\end{aligned}
$$

where $\mathbf{2} = (2, ..., 2)$ and $||\cdot||$ refers to the L_2-norm. $\boldsymbol{\epsilon'}$ can then be obtained by solving the least square linear regression problem:

$$\text{Maximize } ||\mathbf{V}^T\boldsymbol{\epsilon'}^T - \mathbf{2}^T||^2 \qquad (5.7)$$

s.t. $\boldsymbol{\epsilon'}^T\mathbf{1} = 1$, $\boldsymbol{\epsilon'} \geq \mathbf{0}$

A new algorithm based on ISDM and CNM can be summarized as below:

Algorithm ISDM-CNM: Set s=0. From an initial estimate G_0, obtain $L_0 = (L_1(G_0), ..., L_M(G_0))$ where we assume $l_i(G_0) > -\infty$, $i = 1, 2, ..., M$.

- *Step 1: Expand the support points sets:*
  Denote by $L_s = (L_1(G_s), ..., L_M(G_s))$ the current point in $conv(\Gamma)$. Compute all local maxima θ^*_{s1}, ..., $\theta^*_{sp_s}$ of $g(\theta; L_s) \geq 0$, $\theta \in \Omega$. If $\max_j\{g(\theta^*_{sj}; L_s)\} = 0$, stop.

- *Step 2:* Set $\epsilon^+_s = (1, 0, ..., 0)$. Find ϵ^-_{s+1}, the constrained solution of minimizing $Q(\epsilon'|\epsilon^+_s)$ (5.7). Define L^-_{s+1} the new likelihood point.

- *Step 3:* Use step halving or optimization to find $\eta^*_s \in [0, 1]$, respectively, to increase or maximize $log\{L_s + \eta^*_s(L^-_{s+1} - L_s)\}$.

- *Step 4:* Set $L_{s+1} = L_s + \eta^*_s(L^-_{s+1} - L_s)$ to find the likelihood point for the next step. Set $s = s + 1$ and go back to *Step 1.*

This new algorithm would accommodate desirable features of both methods and should further improve the efficiency.

5.2.3 Asymptotic Properties

There exists a growing literature dealing with the asymptotic properties of a semiparametric maximum likelihood estimator on frailty models. Nielsen et al. (1992) proposes an EM algorithm to estimate the cumulative baseline hazard with an assumed shared-gamma frailty distribution. The consistency and the asymptotic distribution of the estimates for this model have been rigorously studied by Murphy (1994, 1995) for the case with no covariates, and by Parner (1998) for the case with covariates. However, statistical inference based on the non-standard asymptotic properties of the NPMLE remains an interesting and challenging problem.

Murphy and van der Vaart (2000) study the asymptotic properties of the semiparametric profile likelihood for a frailty model. In this model, the unobserved frailty G follows a gamma distribution with unit mean and variance σ and the cumulative baseline hazard Λ_0 is left completely unspecified. β are regression parameters. So the unknown parameters φ contains the parameters of interest $\theta = (\beta, \sigma)$ and the nuisance parameter Λ_0. The profile likelihood for θ can be written as

$$pl(\theta) = l(\theta, \hat{\Lambda}_{0,\theta}) \tag{5.8}$$

where $\hat{\Lambda}_{0,\theta}$ is the maximizer of $l(\theta, \Lambda_0)$ for given θ. Under regularity conditions, Murphy and Van Der Vaart prove that the curvature of the profile likelihood at the MLE, $\hat{\theta}$, of θ provides consistent estimate of the asymptotic variance of $\hat{\theta}$.

Murphy and van der Vaart's results can be conjectured to generalize to a number of results for our approach. First, we can consider the semiparametric model by assuming a piecewise constant distribution for the baseline hazard and leaving frailty distribution, G, completely unspecified. That is, $\Lambda_0(t) = \sum_{k=1}^{K} \lambda_k I(\tau_{k-1} < t \leq \tau_k)$, where $0 \equiv \tau_1 < ... < \tau_K < t_{(N)}$ are pre-determined cutoff points of sub-periods and $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_K)$ are constant hazard during each sub-period. Let $\theta = (\beta, \lambda)$ be

the parameters of interest and G, the nuisance parameter. We consider the profile likelihood

$$pl(\theta) = l(\theta, \hat{G}_\theta) \tag{5.9}$$

and denote the maximizer of (5.9) as $\hat{\theta}_s = (\hat{\beta}_s, \hat{\lambda}_s)$. Van der Vaart (1996) gave the asymptotic properties of the exponential frailty model where the mixing distribution G is completely unspecified. These arguments should also apply to the piecewise constant model outlined above.

The nonparametric approach considered in Chapter IV is obtained by further relaxing the restriction on $\Lambda_0(\cdot)$ from the semiparametric model above. We can focus on the parameter of interests, β, and consider both Λ_0 and G as the nuisance parameters. The profile likelihood for β becomes

$$pl(\beta) = l(\beta, (\hat{\Lambda}_{0,\beta}, \hat{G}_\beta)) \tag{5.10}$$

and the NPMLE $\hat{\beta}_n$ can be obtained by maximizing (5.10). We expect that the estimate of the asymptotic variance for the NPMLE $\hat{\beta}_n$ would be obtained from the Hessian matrix of the profile likelihood (5.10) at $\hat{\beta}_n$. We also expect the asymptotic variance for the NPMLE $\hat{\beta}_n$ will be close to that for $\hat{\beta}_s$ (5.9) when the number of sub-periods , K, defined in the piecewise constant distribution of Λ_0 in (5.9) becomes very large and the width of each sub-period in which $\tau_{k-1} < t \leq \tau_k$, gets very small (Efron, 1977; Oakes, 1977). In fact, our nonparametric approach adopted to estimate Λ_0, by assuming a jump function with jumps at the observed event times can be regarded as the most refined piecewise constant distribution with varying cutoff points.

There is a very large class of substantial mathematical problems associated with understanding the asymptotics of the fully nonparametric frailty model. We have concentrated above on issues related to estimating β, there are problems also related to understanding the asymptotics of the estimates of Λ_0 and G. Future work will be

devoted to this area.

BIBLIOGRAPHY

BIBLIOGRAPHY

Abdeljaber, M. H., Monto, A. S., Tilden, R. L., Schork, M. A., and Tarwotjo, I. (1991). The impact of vitamin a supplementation on morbidity: a randomized community intervention trial. *American Journal of Public Health* **81,** 1654–1656.

Atwood, C. (1976). Convergent design sequences, for sufficiently regular optimality criteria. *The Annals of Statistics* **4,** 1124–1138.

Bo, L. and Rosenbaum, P. (2004). Optimal pair matching with two control groups. *Journal of Computational and Graphical Statistics* **13,** 422–434.

Bohning, D. (1985). Numerical estimation of a probability measure. *Journal of Statistical Planning and Inference.* **11,** 57–69.

Breslow, N. E. (1972). Discussion of the paper by d. r. cox. *Journal Of The Royal Statistical Society Series B* **34,** 216–217.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. journal american statistical association. *Journal American Statistical Association* **88,** 9–25.

Cai, J. and Prentice, R. (1997). Regression estimation using multivariate failure time data and a common baseline hazard function model. *Lifetime Data Analysis* **3,** 197–213.

Cai, T., Cheng, S., and Wei, L. (2002). Semiparametric Mixed-Effects Models for Clustered Failure Time Data. *Journal of the American Statistical Association* **97,** 514–523.

Campbell, D. (2009). Prospective: Artifact and Control1. *Artifacts in Behavioral Research: Robert Rosenthal and Ralph L. Rosnow's Classic Books* page 264.

Chang, W., Hwang, B., Wang, D., and Wang, J. (1997). Cytogenetic effect of chronic low-dose, low-dose-rate [gamma]-radiation in residents of irradiated buildings. *The Lancet* **350,** 330–333.

Clayton, D. and Cuzick, J. (1985). Multivariate generalizations of the proportional hazards model. *Journal of the Royal Statistical Society. Series A (General)* **148,** 82–117.

Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* **65,** 141–151.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassifications in removing bias in observational studies. *Biometrics* **24,** 295–313.

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

Cox, D. and Hinkley, D. (1979). *Theoretical statistics*. Chapman & Hall/CRC.

Donner, A. and Klar, N. (2000). *Design and Analysis of Cluster Randomization Trials in Health Research* . Hodder Arnold Publication.

Donner, A., Piaggio, G., and Villar, J. (2001). Statistical methods for the meta-analysis of cluster randomization trials. *Statistical methods in medical research* **10,** 325.

Efron, B. (1971). Forcing a sequential experiment to be balanced. *Biometrika* **58,** 403.

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

Fedorov, V. (1972). Theory of optimal experiments. *New York* .

Fisher Jr, E. (1995). The results of the COMMIT trial. Community Intervention Trial for Smoking Cessation. *American Journal of Public Health* **85,** 159.

Gill, P., Murray, W., and Wright, M. (1981). Practical optimization.

Graham, J., Flay, B., Johnson, C., Hansen, W., and Collins, L. (1984). A multiattribute utility measurement approach to the use of random assignment with small numbers of aggregated units. *Evaluation Review* **8,** 247–260.

Greevy, R., Lu, B., Silber, J., and Rosenbaum, P. (2004). Optimal multivariate matching before randomization. *Biostatistics* **5,** 263–275.

Gu, X. S. and Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics* **2,** 405–420.

Guo, G. (1992). Estimating a multivariate proportional hazards model for clustered data using the EM algorithm, with an application to child survival in Guatemala. *Journal of the American Statistical Association* **87,** 969–976.

Hansen, B. B. (2004). Full matching in an observational study of coaching for the sat. *Journal of the American Statistical Association* **99,** 609–618.

Heckman, J. and Singer, B. (1984a). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica: Journal of the Econometric Society* pages 271–320.

Heckman, J. and Singer, B. (1984b). Econometric duration analysis. *Journal of Econometrics* **24,** 63–132.

Heckman, J. and Singer, B. (1984c). The identifiability of the proportional hazard model. *The Review of Economic Studies* **51,** 231–241.

Hougaard, P. (2000). *Analysis of multivariate survival data*. Springer Verlag.

Hurvich, C., Simonoff, J., and Tsai, C. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **60,** 271–293.

Joffe, M. M. (1999). Propensity scores. *American Journal of Epidemiology* **150,** 327–333.

Kalbfleisch, J. and Prentice, R. (2002). *The statistical analysis of failure time data*. Wiley New York.

Klar, N. and Donner, A. (2000). Current and future challenges in the design and analysis of cluster randomization trials. *Statistics in Medicine* **20,** 1729–1740.

Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association* pages 805–811.

Lee, E., Wei, L., and Amato, D. (1992). Cox-type regression analysis for large numbers of small groups of correlated failure time observations. *Survival Analysis: State of the Art* **211,** 237–247.

Lesperance, M. L. and Kalbfleisch, J. D. (1992). An algorithm for computing the nonparametric mle of a mixing distribution. *Journal of the American Statistical Association* **87,** 120–126.

Lindsay, B. (1983). The geometry of mixture likelihoods: a general theory. *The Annals of Statistics* pages 86–94.

McGilchrist, C. A. and Aisbett, C. W. (1991). Regression with frailty in survival analysis. *Biometrics* **47,** 461–466.

Ming, K. and Rosenbaum, P. R. (2000). Substantial gains in bias reduction from matching with a variable number of controls. *Biometrics* **56,** 118–124.

Murphy, S., Rossini, A., and Van der Waart, A. (1997). Maximum Likelihood Estimation in the Proportional Odds Model. *Journal of the American Statistical Association* **92,**.

Murphy, S. A. (1994). Consistency in a proportional hazards model incorporating a random effect. *The Annals of Statistics* **22,** 712–731.

Murphy, S. A. (1995). Asymptotic theory for the frailty model. *The Annals of Statistics* **23,** 182–198.

Murray, D. (1998). *Design and analysis of group-randomized trials*. Oxford University Press, USA.

Murray, D., Varnell, S., and Blitstein, J. (2004). Design and analysis of group-randomized trials: a review of recent methodological developments. *American Journal of Public Health* **94,** 423.

Neuhaus, J. and Kalbfleisch, J. (1998). Between-and within-cluster covariate effects in the analysis of clustered data. *Biometrics* **54,** 638–645.

Nielsen, G. G., Gill, R. D., Andersen, P. K., and Sorensen, T. I. (1992). A counting process approach to maximum likelihood estimation of frailty models. *Scandinavian Journal of Statistics* **19,** 25–43.

Oakes, D. (1977). The asymptotic information in censored survival data. *Biometrika* **64,** 441.

of Neurological Disorders, N. T. N. I. and rt PA Stroke Study Group, S. (1995). Tissue plasminogen activator for acute ischemic stroke. *The New England Journal of Medicine* **333,** 1581–1588.

Olsen, S. P. (1997). *Multivariate matching with non-normal covariates in observational studies*. Ph.D. Thesis, University of Pennsylvania.

Parner, E. (1998). Asymptotic theory for the correlated gamma-frailty model. *The Annals of Statistics* **26,** 183–214.

Pocock, S. and Simon, R. (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics* **31,** 103–115.

Richard, L. (1988). Burden, J. Douglas Faires, Numerical analysis.

Ripatti, S. and Palgrem, J. (2000). Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics* **56,** 1016–1022.

Robins, J., Mark, S., and Newey, W. (1992). Estimating exposure effects by modeling the expectation of exposure conditional on confounders. *Biometrics* **48,**.

Roghmann, K. and Sodeur, W. (1972). The impact of military service on authoritarian attitudes: Evidence from West Germany. *The American Journal of Sociology* **78,** 418–433.

Rosenbaum, P. R. (2002). *Observational Studies*. Springer.

Rosenbaum, P. R. and Rubin, D. B. (1984). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70,** 41–55.

Seltser, R. and Sartwell, P. (1965). The influence of occupational exposure to radiation on the mortality of American radiologists and other medical specialists. *Am J Epidemiol* **81,** 2–22.

Susko, E. (1996). *Nonparametric maximum likelihood estimation for mixture models.* Ph.D. Thesis, University of Waterloo.

Susko, E., Kalbfleisch, J., and Chen, J. (1998). Constrained nonparametric maximum-likelihood estimation for mixture models. *Canadian Journal of Statistics* **26,** 601–617.

Therneau, T., Grambsch, P., and Pankratz, V. (2003). Penalized survival models and frailty. *Journal of Computational and Graphical Statistics* **12,** 156–175.

Trussell, J. and Richards, T. (1985). Correcting for unmeasured heterogeneity in hazard models using the Heckman-Singer procedure. *Sociological methodology* **15,** 242–276.

Van der Vaart, A. (1996). Efficient maximum likelihood estimation in semiparametric mixture models. *The Annals of Statistics* **24,** 862–878.

Vaupel, J. W., Manton, K. G., and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* **16,** 439–454.

Verweij, P. and Van Houwelingen, H. (1994). Penalized likelihood in cox regression. *Statistics in Medicine* **13,** 2427–2436.

Wang, Y. (2007). On fast computation of the non-parametric maximum likelihood estimate of a mixing distribution. *Journal Of The Royal Statistical Society Series B* **69,** 185–198.

Wells, K., Roberts, C., Daniels, S., Hann, D., Clement, V., Reintgen, D., and Cox, C. (1997). Comparison of psychological symptoms of women requesting removal of breast implants with those of breast cancer patients and healthy controls. *Plastic and Reconstructive Surgery* **99,** 680.

Weston, J. and Mansinghka, S. (1971). Tests of the efficiency performance of conglomerate firms. *The Journal of Finance* **26,** 919–936.

Zabin, L., Hirsch, M., and Emerson, M. (1989). When urban adolescents choose abortion: effects on education, psychological status and subsequent pregnancy. *Family Planning Perspectives* **21,** 248–255.

Zeng, D., Lin, D., and Lin, X. (2008). Semiparametric transformation models with random effects for clustered failure time data. *Statistica Sinica* **18,** 355.