

**PERFORMANCE OF RANK-MINIMIZATION UNDER DIFFERENT SCENARIOS: A
SIMULATION STUDY FOCUSING ON BASELINE COVARIATES IMBALANCES IN
CLINICAL TRIALS**

by

Jung-Yi Lin

BS, Department of Public Health, National Taiwan University, Taiwan, 2015

Submitted to the Graduate Faculty of
the Department of Biostatistics
Graduate School of Public Health in partial fulfillment
of the requirements for the degree of
Master of Science

University of Pittsburgh

2017

UNIVERSITY OF PITTSBURGH
GRADUATE SCHOOL OF PUBLIC HEALTH

This thesis was presented

by

Jung-Yi Lin

It was defended on

April 3rd, 2017

and approved by

Thesis Advisor: Douglas P. Landsittel, PhD, Professor of Medicine, Biostatistics, Biomedical Informatics, and Clinical and Translational Science, School of Medicine, University of Pittsburgh

Jeanine M. Buchanich, MEd, PhD, Research Associate Professor, Biostatistics, Graduate School of Public Health, University of Pittsburgh

Stewart J Anderson, PhD, Professor, Biostatistics, Graduate School of Public Health, University of Pittsburgh

Copyright © by Jung-Yi Lin

2017

Douglas P. Landsittel, PhD

**PERFORMANCE OF RANK-MINIMIZATION UNDER DIFFERENT SCENARIOS:
A SIMULATION STUDY FOCUSING ON BASELINE COVARIATES IMBALANCES
IN CLINICAL TRIALS**

Jung-Yi Lin, MS

University of Pittsburgh, 2017

ABSTRACT

Clinical trials are often considered to be the gold standard for assessing effectiveness and safety of medical treatments and public health interventions. The validity of inferences from clinical trials depends on randomizing subjects to different treatment groups. Although simple randomization is the most common approach, and generally prevents differences in baseline covariate imbalances between groups, other approaches may be necessary for balancing covariate distributions within important strata. However, the performance of stratified randomization may be limited when the sample size is small and there are many strata. These scenarios may be better addressed through minimization, or rank-minimization algorithms.

The concept of rank-minimization is straightforward but very little research has been published on the topic. To address this gap in the literature, we conducted a simulation study to investigate how rank-minimization performed, compared to Taves' minimization, with different sample sizes and baseline covariate distributions.

Results indicated that both sample size and covariate distributions influence the performance of rank-minimization and minimization. Overall, rank-minimization yields better properties, and larger sample sizes yield better properties for both methods. However, the performance for both methods decreases when the distribution is more skewed. Results of this

study provide researchers with more information to decide between randomization methods for their specific applications.

Public Health Significance: In clinical trials, the comparability of subjects between different treatment groups is critical to validity of the subsequent inferences. Since clinical trials are often considered the gold standard for assessing medical treatments and public health interventions, and the trials are usually expensive and time-consuming to conduct, optimizing the randomization process represents a highly significant aspect of public health research. Consulting results of the simulation study will provide additional information for researchers to decide the best method for randomization for different size data sets and different covariate distributions encountered in practice.

TABLE OF CONTENTS

1.0	INTRODUCTION	1
2.0	METHOD.....	6
2.1	MINIMIZATION	6
2.2	RANK-MINIMIZATION	8
2.3	A SIMULATION STUDY.....	10
2.4	PERFORMANCE INDICATOR	12
3.0	RESULTS.....	14
3.1	PERFORMANCE FOR N=200 AND STANDARD NORMAL DISTRIBUTIONS	14
3.2	PERFORMANCE UNDER DIFFERENT SAMPLE SIZES	17
3.3	PERFORMANCE UNDER SKEWED DISTRIBUTIONS	19
4.0	DISCUSSION.....	26
	APPENDIX A. R CODE FOR THE SIMULATION	30
	APPENDIX B . ADDITIONAL SIMULATIONS FROM SECTION 3.1	37
	BIBLIOGRAPHY	50

LIST OF TABLES

Table 1. Example of minimization.....	8
Table 2. Rank-matrix	9
Table 3. Example for rank-minimization.....	10
Table 4. Distribution of t-statistics.....	13
Table 5. Rank-minimization distribution of t-statistics (n=200)	15
Table 6. Rank-minimization distribution of diff (n=200).....	17
Table 7.Effect of sample size on the baseline covariate imbalance.....	18
Table 8.Effect of sample size on the allocation imbalance.....	19
Table 9a.Effect of sample size on the baseline covariate imbalance (n=200)	20
Table 9b.Effect of sample size on the baseline covariate imbalance (n=100).....	21
Table 9c.Effect of sample size on the baseline covariate imbalance (n=50).....	21
Table 10a.Effect of sample size on the allocation imbalance (n=200).....	24
Table 10b.Effect of sample size on the allocation imbalance (n=100).....	25
Table 10c.Effect of sample size on the allocation imbalance (n=50).....	25
Table 11.Rank-minimization distribution of t-statistics (n=200, N(0,1)).....	37
Table 12.Rank-minimization distribution of diff (n=200, N(0,1))	37
Table 13.Rank-minimization distribution of t-statistics (n=100, N(0,1)).....	38

Table 14. Rank-minimization distribution of diff (n=100, N(0,1))	38
Table 15. Rank-minimization distribution of t-statistics (n=50, N(0,1)).....	38
Table 16. Rank-minimization distribution of diff (n=50, N(0,1))	39
Table 17. Rank-minimization distribution of t-statistics (n=200, logN(0,0.5)).....	39
Table 18. Rank-minimization distribution of diff (n=200, logN(0,0.5))	39
Table 19. Rank-minimization distribution of t-statistics (n=100, logN(0,0.5)).....	40
Table 20. Rank-minimization distribution of diff (n=100, logN(0,0.5))	40
Table 21. Rank-minimization distribution of t-statistics (n=50, logN(0,0.5)).....	40
Table 22. Rank-minimization distribution of diff (n=50, logN(0,0.5))	41
Table 23. Rank-minimization distribution of t-statistics (n=200, logN(0,1)).....	41
Table 24. Rank-minimization distribution of diff (n=200, logN(0,1))	41
Table 25. Rank-minimization distribution of t-statistics (n=100, logN(0,1)).....	42
Table 26. Rank-minimization distribution of diff (n=100, logN(0,1))	42
Table 27. Rank-minimization distribution of t-statistics (n=50, logN(0,1)).....	42
Table 28. Rank-minimization distribution of diff (n=50, logN(0,1))	43
Table 29. Minimization distribution of t-statistics (n=200, N(0,1))	43
Table 30. Minimization distribution of diff (n=200, N(0,1)).....	43
Table 31. Minimization distribution of t-statistics (n=100, N(0,1))	44
Table 32. Minimization distribution of diff (n=100, N(0,1)).....	44
Table 33. Minimization distribution of t-statistics (n=50, N(0,1))	44
Table 34. Minimization distribution of diff (n=50, N(0,1)).....	45
Table 35. Minimization distribution of t-statistics (n=200, logN(0,0.5))	45
Table 36. Minimization distribution of diff (n=200, logN(0,0.5))	45

Table 37. Minimization distribution of t-statistics (n=100, logN(0,0.5))	46
Table 38. Minimization distribution of diff (n=100, logN(0,0.5))	46
Table 39. Minimization distribution of t-statistics (n=50, logN(0,0.5))	46
Table 40. Minimization distribution of diff (n=50, logN(0,0.5))	47
Table 41. Minimization distribution of t-statistics (n=200, logN(0,1))	47
Table 42. Minimization distribution of diff (n=200, logN(0,1))	47
Table 43. Minimization distribution of t-statistics (n=100, logN(0,1))	48
Table 44. Minimization distribution of diff (n=100, logN(0,1))	48
Table 45. Minimization distribution of t-statistics (n=50, logN(0,1))	48
Table 46. Minimization distribution of diff (n=50, logN(0,1))	49

LIST OF FIGURES

Figure 1. PDF of the distributions	11
Figure 2. Histograms of the t-statistics (variables 1 to 9).....	16
Figure 3. Histograms of the t-statistics (variables 10 to 15).....	16
Figure 4. Histograms of the t-statistics (n=200, logN(0,0.5), variables 1 to 9).....	22
Figure 5. Histograms of the t-statistics (n=200, logN(0,0.5), variables 10 to 15).....	22
Figure 6. Histograms of the t-statistics (n=200, logN(0,1), variables 1 to 9).....	23
Figure 7. Histograms of the t-statistics (n=200, logN(0,1), variables 10 to 15).....	23

1.0 INTRODUCTION

Clinical trials are a common, but expensive and time-consuming approach to assess the effectiveness and safety of public health interventions such as vaccines and medical treatments. The necessity of randomization in clinical trials is well known, including the need to balance the baseline covariates in different groups (Suresh, 2011; Altman & Bland, 1999). Since assigned treatment status (in a randomized controlled trial) is determined entirely by randomization, we can then attribute the result (for the given study population) to either chance (that is controlled at the given α -level) or a true difference in treatment efficacy. In contrast, treatment status in observational studies is determined through self-selection (or assignment by a physician) that depends on many different factors, such as their personal characteristics, insurance status, or other traits of the health system. Findings in observational studies are therefore often subject to substantial bias. To illustrate, in 1980, a paper in the *Lancet* indicated that vitamin supplementation for pregnant women could prevent the neural-tube defect in newborn infants (Altman & Bland, 1999; Smithells & Sheppard). However, results were not convincing since the studies were not randomized, and the treatment and control groups were subsequently not comparable (Altman & Bland, 1999; "Prevention of neural tube defects: results of the Medical Research Council Vitamin Study. MRC Vitamin Study Research Group," 1991). A subsequent randomized study (funded by the Medical Research Council) was conducted to yield more convincing results (Altman & Bland, 1999;). Many other examples in the literature illustrate how

randomized trials are often conducted to provide stronger evidence of medical treatments or other interventions.

The most basic approach to randomization is simple randomization, where each subject is randomly assigned to a treatment without any special consideration of the subsequent distribution of important baseline factors; instead, the distribution of key baseline factors across groups is simply described at the end of the trial. Simple randomization is the most straightforward approach to conduct and, on average, is effective in generating comparable groups. However, simple randomization may lead to chance imbalances that can threaten the interpretation of the trial findings. If, for instance, a trial assessing final pain scores had a chance imbalance in baseline pain scores, any difference in the final pain scores may be difficult to interpret. This is particularly problematic when the sample size is small (Suresh, 2011).

A number of different approaches exist to better avoid the above-described types of chance imbalances. The most common approach to achieve this goal is stratified block randomization (or just stratified randomization). The basic idea of the approach is to divide the targeted sample size into blocks (strata) and the randomization is performed within each block. The first step of stratified randomization is to identify the baseline covariates that are the most critical to balance across treatment groups. The subjects are separated into different blocks by their characteristics and the subjects in the same block all own the same traits. After all subjects are assigned to the blocks, either simple randomization or permuted block randomization is conducted within each block (Pocock & Simon, 1975; Suresh, 2011).

Subjects may also be randomized within some block of fixed or random size until half of the subjects in that block are assigned to a given treatment; the remainder of subjects are then assigned to the other treatment. This process, of permuted block randomization, may also be

described as generating many permuted blocks where the size of the blocks is a multiple of the number of treatments (usually 4, 6, 8 for a two-armed trial) and the block sizes may vary within a trial. For example, if there are two treatments (A and B) and the block size is 4, there would be 6 permuted blocks in this scenario (AABB, ABAB, ABBA, BBAA, BABA, BAAB). A random sequence of the permuted blocks would be generated and the subjects would be assigned to a treatment based on the sequence of blocks.

Although stratified randomization has the advantage of being easy to conduct, and is a well-understood process, the performance of controlling baseline covariate imbalances may be poor when the sample size is small and many covariates need to be controlled¹. As one example, a two-arm simulation trial by Therneau found that the performance of balancing baseline factors could be low if the number of cells across the different strata was more than half the sample size (Therneau, 1993). In a number of practical examples, this constraint can be problematic. For instance, in a randomized trial studying spinal manipulation methods, there were three treatments, just over 100 eligible subjects, and three different critical factors that needed to be balanced between treatment groups (baseline pain score, disability, and treatment expectancy) (Schneider, Haas, Glick, Stevans, & Landsittel, 2015). In this example, using stratified randomization would yield very small sample sizes within each stratum and thus produce a design with poor performance characteristics.

To address these concerns, Taves (Taves, 1974) and Pocock and Simon (Pocock & Simon, 1975) proposed a method called minimization. The basic concept of minimization is also straightforward, and offers certain advantages over stratified randomization. The process begins with simple randomization of the initial group of (typically 10-20) subjects. An imbalance score is then calculated to measure how different the treatment groups are specific to the key variables

of interest. The specific details for calculating this score depend on the distribution and relative importance of those variables, and must be determined by the investigators ahead of time. Once the initial pool of subjects is randomized, each new subject is assessed individually. Specifically, the next subject is assumed to be assigned to each group and an imbalance score is calculated specifically for each group. This new subject would be allocated to the group with the lower score (or one might still use a random assignment with a differential probability assignment, with a lower probability of assignment for the treatment having a greater imbalance score). If there is a tie between the scores, this subject would be assigned to a group randomly.

After Taves proposed this minimization algorithm, others proposed different minimization-like approaches; rank-minimization is one such example (Scott, McPherson, Ramsay, & Campbell, 2002). For this paper, we focus on rank-minimization as the most common and intuitive alternative. Unlike other such minimization algorithms, in rank-minimization, there is no need for categorizing continuous variables (Hoehler, 1987; Stigsby & Taves, 2010). In this way, the continuous nature of the distribution is used more efficiently to balance the baseline covariates (Stigsby & Taves, 2010). Rank-minimization has the obvious difference of not depending on the shape or variability of the distribution, although that issue could be a benefit or a limitation, depending on whether the outliers and the shape of the distribution provides critical information which might be lost in ranking the data.

Several publications have proposed variations on rank-minimization methods. In 1987, Hoehler suggested using ranks to address concerns associated with continuous variables (Hoehler, 1987). Stigsby and Taves modified Hoehler's method to propose using rank-sums rather than rank-means (Stigsby & Taves, 2010). The Stigsby's and Taves' paper also conducted a simulation study to compare the performance among rank-minimization, minimization, and

stratified block randomization (Stigsby & Taves, 2010). They simulated a fairly limited scenario of 200 subjects and 15 variables from standard normal distribution.

The objective of this study was to expand the Stigsby's and Taves' simulations to provide more practical information on the statistical properties of the rank-minimization approach as compared to minimization without ranking (Stigsby & Taves, 2010). More specifically, most variables under consideration are unlikely to all be normally distributed, and sample sizes may often be less than 200. For instance, in the previously-mentioned randomized trial of spinal therapy methods (Schneider, Haas, Glick, Stevans, & Landsittel, 2015), there were approximately 100 patients in the complete study. As another illustration, a trial conducted by the International Breast Cancer Study Group (IBCSG) (Chi & Ibrahim, 2006; Tang, Tang, & Zhu, 2017), to compare the overall and disease-free survival between two treatments, four of the baseline covariates followed skewed distributions instead of a normal distribution. Unfortunately, very little research is published on rank-minimization and its associated statistical properties. The current study seeks to conduct a simulation study to assess how sample size and a skewed distribution influences the baseline covariates imbalances after rank-minimization, as well as after minimization, to inform researchers seeking to decide between randomization methods.

2.0 METHOD

This simulation study was conducted to compare the performance of rank-minimization and Taves' minimization method under different sample sizes and different distributions for the baseline covariates used in the randomization procedure. The methods section describes the process of minimization and rank-minimization, the simulation process, and the measures used to assess the performance of the different methods.

2.1 MINIMIZATION

The minimization algorithm for this study is the approach specified by Taves in 1974 (Taves, 1974). For this algorithm, assume there are n subjects, two treatments, A and B, and strata for m baseline covariates. The initial subjects are randomly assigned. In this study, to simplify the problem, only the first subject was randomized into each treatment. After the first subject, the minimization process was conducted to allocate each new subject. More specifically, when a new subject enters the study, the number of people that have the same characteristic with this subject is counted for each variable, denoted as a_i for group A and b_i for group B ($i=1, \dots, m$). This new subject is hypothetically assigned to each treatment and an imbalance score is calculated for each assignment. If this subject is assigned to treatment A, the imbalance score for

treatment A would be $\sum_{i=1}^m |a_i + 1 - b_i|$ and the imbalance score for treatment B would be $\sum_{i=1}^m |b_i + 1 - a_i|$. After calculating the imbalance scores, the subject is then allocated to the treatment with the lower imbalance score. If the imbalance scores are the same in the two treatments, this subject would be assigned to a treatment randomly.

For this process, continuous variables are categorized; specifically continuous variables are separated into 3 groups for this study.

- 1) Group 1: the value is $\geq \mu + \sigma$
- 2) Group 2: the value is between $\mu - \sigma$ and $\mu + \sigma$
- 3) Group 3: the value is $\leq \mu - \sigma$

Table 1 provides an example of the minimization process. Assume there was a two-armed trial and two baseline covariates (BMI & age) need to be balanced between treatment groups. Further, assume 10 subjects were already enrolled in the trial with the given frequencies for each category across BMI and Age (illustrated in the first two rows of the table). A 55-year-old subject with a BMI of 20 then joins the trial. The imbalance score for A would be $|(2 + 1) - 2| + |(1 + 1) - 0| = 3$ and the score for B would be $|2 - (2 + 1)| + |1 - (0 + 1)| = 1$. The imbalance score for A is higher so the new subject is assigned to treatment B.

Table 1. Example of minimization

	BMI			Age				
	<18.5	18.5-24.99	≥25	<40	40-49	50-59	≥60	Sum
Treatment A	0	2	3	2	1	1	1	
Treatment B	1	2	2	2	1	0	2	
Subject 11		1				1		
Assigned to A								
Treatment A		3				2		
Treatment B		2				0		
Difference		1				2		3
Assigned to B								
Treatment A		2				1		
Treatment B		3				1		
Difference		1				0		1
The new subject would be assigned to B								

2.2 RANK-MINIMIZATION

The basic concept of rank-minimization is similar to minimization. The initial subjects are randomly assigned and subsequent subjects are assigned based on an imbalance score that depends on the ranked data. More specifically, when a new subject enters the study, a rank-matrix like Table 2 is generated (Stigsby & Taves, 2010). After hypothetically assigning the subject to one treatment, the sum of ranks is calculated for each variable and each group, and an

average of these sums is calculated. If there are two treatments and two variables, for instance, then there would be $2 \times 2 = 4$ sums of ranks and the mean of these four sums would be calculated. The imbalance score in rank-minimization is the sum of squared deviation from the mean rank sum.

Table 2. Rank-matrix

ID	BMI	BMI Rank	Age	Age Rank	Treatment
1	26	5	61	4	A
2	20	2	63	5	B
3	19	1	43	2	A
4	22	3	39	1	B
5 (New)	25	4	54	3	?

Table 3 illustrates how the imbalance score in rank-minimization is calculated. If the new subject is allocated to treatment A, the sum of ranks is calculated for each variable (BMI & age) and each treatment (A & B). The deviation from the mean rank-sum is then calculated and the imbalance score is defined as the sum of squared deviations from the mean rank-sum. In the following example, the imbalance scores were 17 for treatment A and 9 for treatment B. Therefore, the new subject was assigned to treatment B.

Table 3. Example for rank-minimization

	Assigned to A				Assigned to B			
	BMI		Age		BMI		Age	
	A	B	A	B	A	B	A	B
Sum of ranks	10	5	9	6	6	9	6	9
Mean rank-sum	7.5				7.5			
Deviation from mean	2.5	2.5	1.5	1.5	1.5	1.5	1.5	1.5
Squared deviation	6.25	6.25	2.25	2.25	2.25	2.25	2.25	2.25
Imbalance=sum of squared deviation	17				9			
The new subject would be assigned to B								

2.3 A SIMULATION STUDY

To further assess the statistical properties of these methods, a simulation study was conducted in R. The first set of data was generated according to the simulation in the Stigsby's and Taves' rank-minimization paper (Stigsby & Taves, 2010). Fifteen independent continuous variables were generated for 200 subjects and each variable followed the standard normal distribution. The current study also extends the simulations to different sample sizes and different distributions. More specifically, data was generated with sample sizes of 200, 100 and 50 subjects. For each sample size, the data was also generated from two log-normal distributions with ($\mu=0$, $\sigma=1$) and ($\mu=0$, $\sigma=0.5$) in addition to the normal distribution, thus yielding a total of 9 scenarios in this simulation study. All fifteen variables in a dataset were generated from the same distribution (i.e.

$N(0,1)$, $\log N(0,0.5)$, or $\log N(0,1)$). Figure 1 is the probability density function (PDF) of the different distributions, where the standard normal distribution is perfectly symmetric and the two log-normal distributions are skewed (with the log-normal (0,1) being the most skewed).

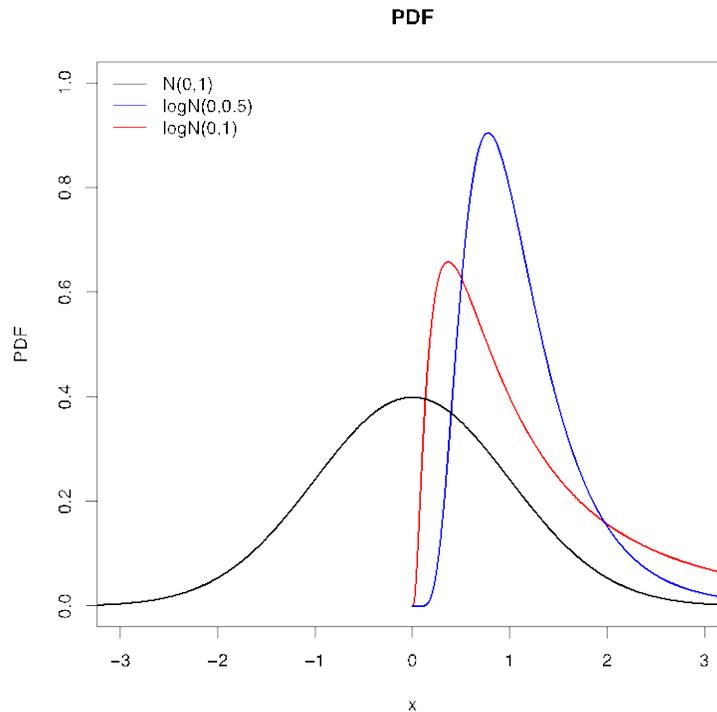


Figure 1. PDF of the distributions

The data was generated using the Lehmer random number generator with modulus (n) = $2^{31}-1$, multiplier (g) = 16807, and increment = 0 ($X_{k+1} = g \cdot X_k \text{ mod } n$) (Park, 1988). The “randtoolbox” package in R was used to generate the data. Each simulation used 1000 runs and the seed was set as 1 to 1000.

2.4 PERFORMANCE INDICATOR

The main focus of this research was the baseline covariate imbalance and t-statistics were used as the indicator of the performance.

$$t_i = \frac{\overline{x_{i1}} - \overline{x_{i2}}}{\sqrt{\frac{s_{i1}^2}{n_1} + \frac{s_{i2}^2}{n_2}}}$$

where $\overline{x_{i1}}$ and $\overline{x_{i2}}$ are the estimated means for variable i in two groups and s_{i1} and s_{i2} are the estimated standard deviation for variable i .

The t-statistics were calculated for each run and each variable. As a result, 15,000 t-statistics were calculated in one simulation. The t-statistics were separated into four groups by their values.

- 1) Group 1: the t-statistics is between 0 and 1
- 2) Group 2: the t-statistics is between 1 and 2
- 3) Group 3: the t-statistics is between 2 and 3
- 4) Group 4: the t-statistics is greater than 3

The number of t-statistics in each group in 1000 runs was counted for each variable. The frequencies were averaged over 15 variables and a standard error was calculated for the mean. For example, in Table 3, there were 900 t-statistics lying between 0 and 1 for BMI and 800 t-statistics lying between 0 and 1 for age. The mean of frequencies between 0 and 1 was 850 and

the standard error (SE) for the mean would be $\sqrt{\frac{(800-850)^2 + (900-850)^2}{\sqrt{2-1}}} / \sqrt{2-1}$. Also, the mean of frequencies between 1 and 2 was 125 and the SE was 35.36.

Table 4. Distribution of t-statistics

Interval of t 	BMI	Age
0-1	900	800
1-2	100	150
2-3	0	50
>3	0	0

As a secondary assessment of the method, we also calculated the balance of sample size allocation across groups. The absolute difference of the number of subjects in each group was the indicator for the balance of allocation, that is,

$$\text{diff} = \frac{|n_1 - n_2|}{n}$$

3.0 RESULTS

The results are described in three sections. The first section is an overview of the performance of rank-minimization when the sample size is 200 and the covariates follow a standard normal distribution. This section essentially reproduces the existing results. This analysis was deemed to be necessary since the previous publication used Excel for the simulation study. The simulation results under different sample sizes and different distributions are showed in sections 3.2 and 3.3.

3.1 PERFORMANCE FOR N=200 AND STANDARD NORMAL DISTRIBUTIONS

Table 5 is the distribution of t-statistics for rank-minimization. To assess whether the number of simulation data sets was sufficient for a given scenario, five simulations were initially run for the current scenario.

The performance of baseline covariate imbalance for rank-minimization was consistent across the five simulations, as the proportions of t-statistics lying between 0 and 1 in 1000 runs were all around 97%. About 3% of t-statistics laid between 1 and 2. No t-statistics were greater than 2.

Table 5. Rank-minimization distribution of t-statistics (n=200)

Interval of t 	1	2	3	4	5
0-1	967±1.97	968±1.95	967±1.95	967±1.92	967±1.98
1-2	33±1.97	32±1.95	33±1.95	33±1.92	33±1.98
2-3	0	0	0	0	0
>3	0	0	0	0	0

The distributions of the t-statistics for the fifteen variables were also examined for this specific scenario. The means of the t-statistics for the fifteen were around 0 and the standard deviations laid around 0.47. Figures 2 and 3 are the histograms for the fifteen variables and the red curves are the PDF of $N(0, sd=0.47)$. The distributions of the t-statistics were similar across the ten variables. All of them distributed symmetrically and seemed to follow the normal distributions. Most of the t-statistics laid between -1 and 1 (>95%).

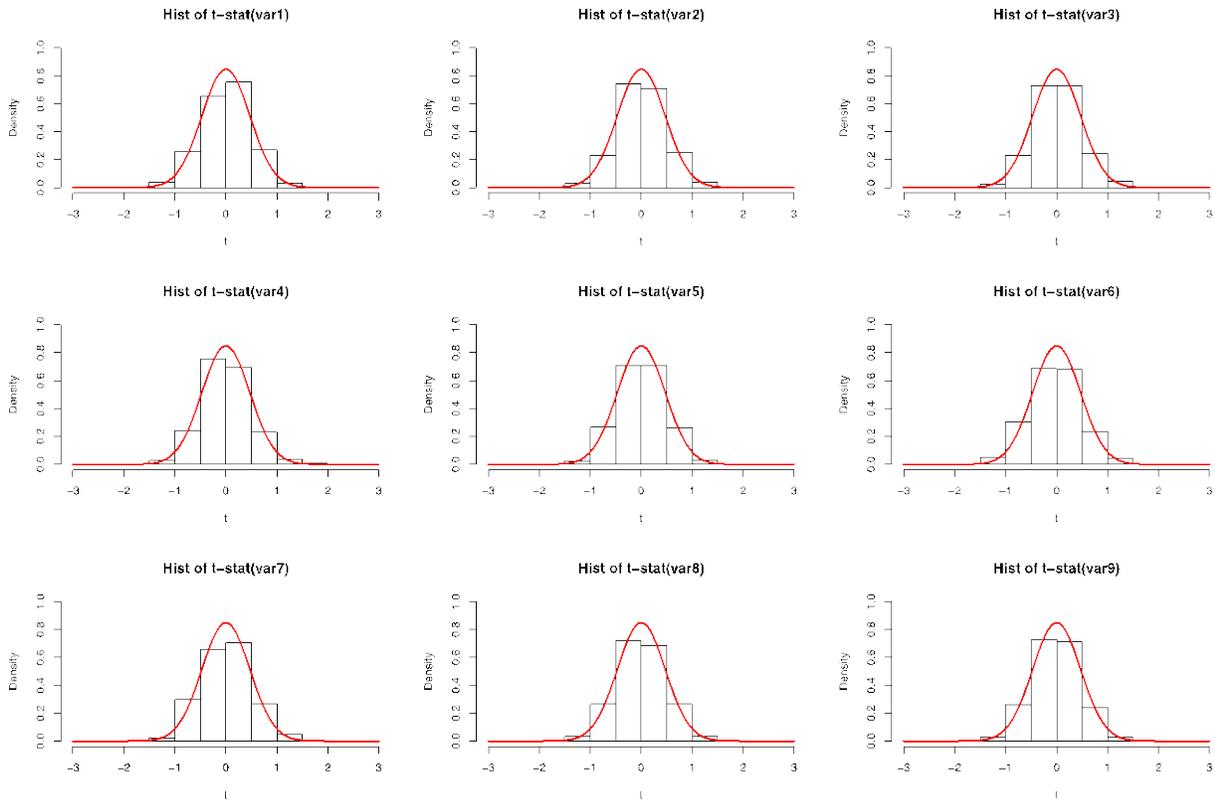


Figure 2. Histograms of the t-statistics (variables 1 to 9)

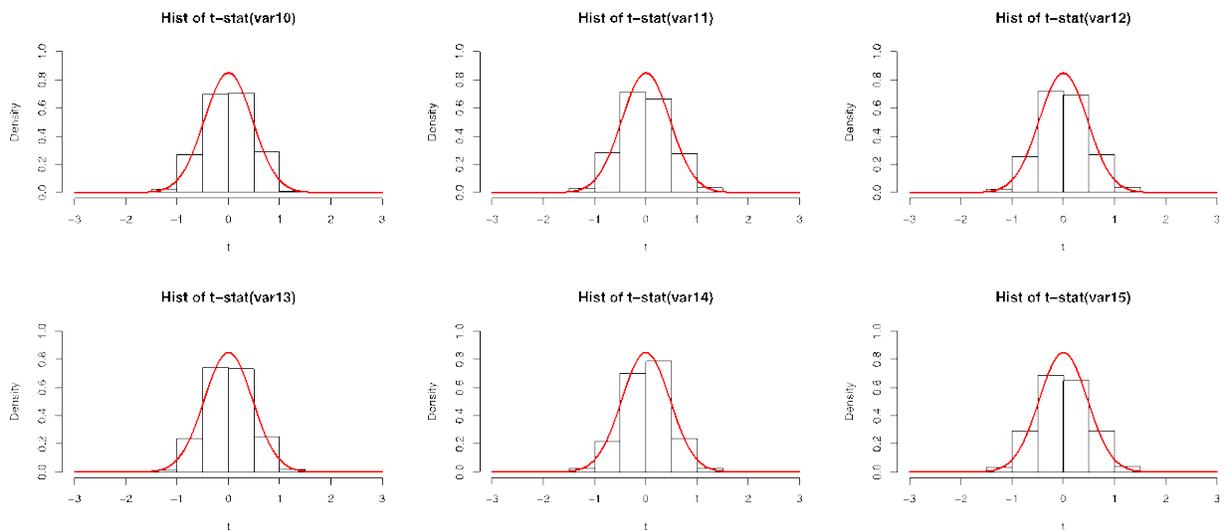


Figure 3. Histograms of the t-statistics (variables 10 to 15)

The results for the absolute differences in the sample size allocation are shown in Table 6. The results were again consistent across the five simulations. Among 1000 runs, 98% of the differences between the number of subjects in each treatment were smaller than or equal to 4. Differences in the sample sizes allocated were as high as 10, but only for 0.1% of the simulations.

Table 6. Rank-minimization distribution of diff (n=200)

Diff	1	2	3	4	5
0	332	331	324	327	325
2	514	514	517	517	519
4	133	135	137	135	134
6	19	18	20	19	20
8	1	1	1	1	1
10	1	1	1	1	1
Total	1000	1000	1000	1000	1000

Given the consistency of simulation results, subsequent simulations are shown for only one set of 1,000 simulated data sets. The additional simulations for the other scenarios are given in Appendix 2.

3.2 PERFORMANCE UNDER DIFFERENT SAMPLE SIZES

Table 7 shows results of the discrepancies between groups (as measured by t-statistics) for different sample sizes. For rank-minimization, no t-statistics were greater than 2 for n=200. In

contrast, however, a small percentage of t-statistics were greater than 2 for n=100 and n=50, and larger percentages of the t-statistics were between 1 and 2 (with means of 3.3%, 8.4% and 15.4% for n=200, 100, and 50, respectively). The same circumstance was also observed for minimization, although smaller discrepancies (between sample sizes) were observed. The overall performance of minimization was worse than rank-minimization for all sample sizes. The mean percentage of t-statistics between 0 and 1 (i.e. having a smaller discrepancy between groups) was between 6% and 9% higher for rank-minimization.

Table 7. Effect of sample size on the baseline covariate imbalance

Interval of t	Rank-minimization mean count \pm SE			Minimization mean count \pm SE		
	n=200	n=100	n=50	n=200	n=100	n=50
0-1	967 \pm 1.84	916 \pm .84	840 \pm 3.29	875 \pm 2.43	826 \pm 2.98	784 \pm 3.60
1-2	33 \pm 1.84	84 \pm 2.85	154 \pm 3.10	121 \pm 2.43	165 \pm 3.08	197 \pm 3.61
2-3	0	0.5 \pm 0.13	5 \pm 0.60	4 \pm 0.35	9 \pm 0.80	19 \pm 0.70
>3	0	0	0	0 \pm 0.07	0 \pm 0.12	1 \pm 0.22

In Table 8, the performance in terms of allocation balance for rank-minimization became worse as the sample size increased. Specifically, there was no difference in sample sizes for only 32% when n=200, but increased to 48% and 65% for n=100 and n=50, respectively. The same trend held for minimization, although differences in allocation were more similar over different sample sizes. For all sample sizes, the allocation balance was better for minimization than rank-minimization.

Table 8. Effect of sample size on the allocation imbalance

diff	Rank-minimization			Minimization		
	n=200	n=100	n=50	n=200	n=100	n=50
0	324	475	652	674	709	758
2	517	480	340	322	290	242
4	137	45	8	4	1	0
6	20	0	0	0	0	0
≥ 8	2	0	0	0	0	0
Total	1000	1000	1000	1000	1000	1000

3.3 PERFORMANCE UNDER SKEWED DISTRIBUTIONS

Table 9 and Table 10 are the results from the skewed distributions, as described in Section 2.3. Table 9a shows the results for skewed distributions for n=200. Table 9b and table 9c show the results for n=100 and for n=50. In Table 9a, the number of t-statistics smaller than 1 for rank-minimization decreased when the baseline covariates were more skewed. Specifically, the mean percentage of t-statistics less than 1 dropped from 97% for N(0,1) to 91% and 80% for logN(0,0.5) and logN(0,1), respectively. The same result was also observed for minimization, although results were more similar across different distributions. When the variables were from logN(0,1), the mean performance for rank-minimization was slightly worse than for minimization.

In Table 9b and Table 9c, when the sample size decreased to 100 and 50, the frequencies of t-statistics between 0 and 1 also decreased as the skewness of the distributions increased (for

rank-minimization and for minimization). The performance for logN(0,1) was similar between rank-minimization and minimization (which was not the case for n=200).

Comparing results across Table 9a, 9b, and 9c reflects less of an effect of skewed distributions for smaller sample sizes. For n=200, the mean percentage of t-statistics less than 1 dropped about 17% from N(0,1) to logN(0,1) for rank-minimization. The proportion of t-statistics less than 1 dropped about 14% for n=100 and dropped about 8.5% for n=50. The same pattern could be found for minimization.

Table 9a. Effect of sample size on the baseline covariate imbalance (n=200)

Interval of t	Rank-minimization mean count ± SE			Minimization mean count ±SE		
	N(0,1)	logN(0,0.5)	logN(0,1)	N(0,1)	logN(0,0.5)	logN(0,1)
0-1	967±1.84	911±5.68	797±9.33	875±2.43	840±7.24	813±10.20
1-2	33±1.84	88±5.68	200±9.56	121±2.43	154±7.18	182±10.36
2-3	0	0±0.09	3±0.64	4±0.35	5±0.65	5±0.94
>3	0	0	0	0±0.07	0	0±0.07

Table 9b. Effect of sample size on the baseline covariate imbalance (n=100)

Interval of t	Rank-minimization mean count \pm SE			Minimization mean count \pm SE		
	N(0,1)	logN(0,0.5)	logN(0,1)	N(0,1)	logN(0,0.5)	logN(0,1)
0-1	916 \pm 2.81	867 \pm 8.15	779 \pm 10.98	826 \pm 2.98	793 \pm 6.25	775 \pm 11.01
1-2	83 \pm 2.81	131 \pm 8.23	216 \pm 11.28	165 \pm 3.08	192 \pm 6.50	215 \pm 11.53
2-3	1 \pm 0.13	1 \pm 0.25	4 \pm 0.69	9 \pm 0.80	14 \pm 0.96	9 \pm 0.86
>3	0	0	0	0 \pm 0.12	0 \pm 0.09	0 \pm 0.07

Table 9c. Effect of sample size on the baseline covariate imbalance (n=50)

Interval of t	Rank-minimization mean count \pm SE			Minimization mean count \pm SE		
	N(0,1)	logN(0,0.5)	logN(0,1)	N(0,1)	logN(0,0.5)	logN(0,1)
0-1	840 \pm 3.29	812 \pm 4.01	755 \pm 3.68	784 \pm 3.60	772 \pm 3.69	760 \pm 3.72
1-2	154 \pm 3.10	182 \pm 3.71	237 \pm 3.43	197 \pm 3.61	208 \pm 3.40	224 \pm 4.00
2-3	5 \pm 0.60	6 \pm 0.64	8 \pm 0.73	19 \pm 0.70	19 \pm 0.71	16 \pm 0.87
>3	0	0	0	1 \pm 0.22	1 \pm 0.22	1 \pm 0.19

The distributions of t-statistics for the 15 variables were also examined for logN(0,0.5) and logN(0,1) with n=200 (Figures 4 to 7). The red curves were PDF of N(0, $\sigma=0.58$) for the data generated from logN(0,0.5) and PDF of N(0, $\sigma=0.76$) for the data from logN(0,1). The t-statistics for all 15 variables from logN(0,0.5) all distributed normally with $\mu=0$ and $\sigma=0.58$. However, the t-statistics for the variables from logN(0,1) did not distributed in a similar way and did not followed a normal distribution.

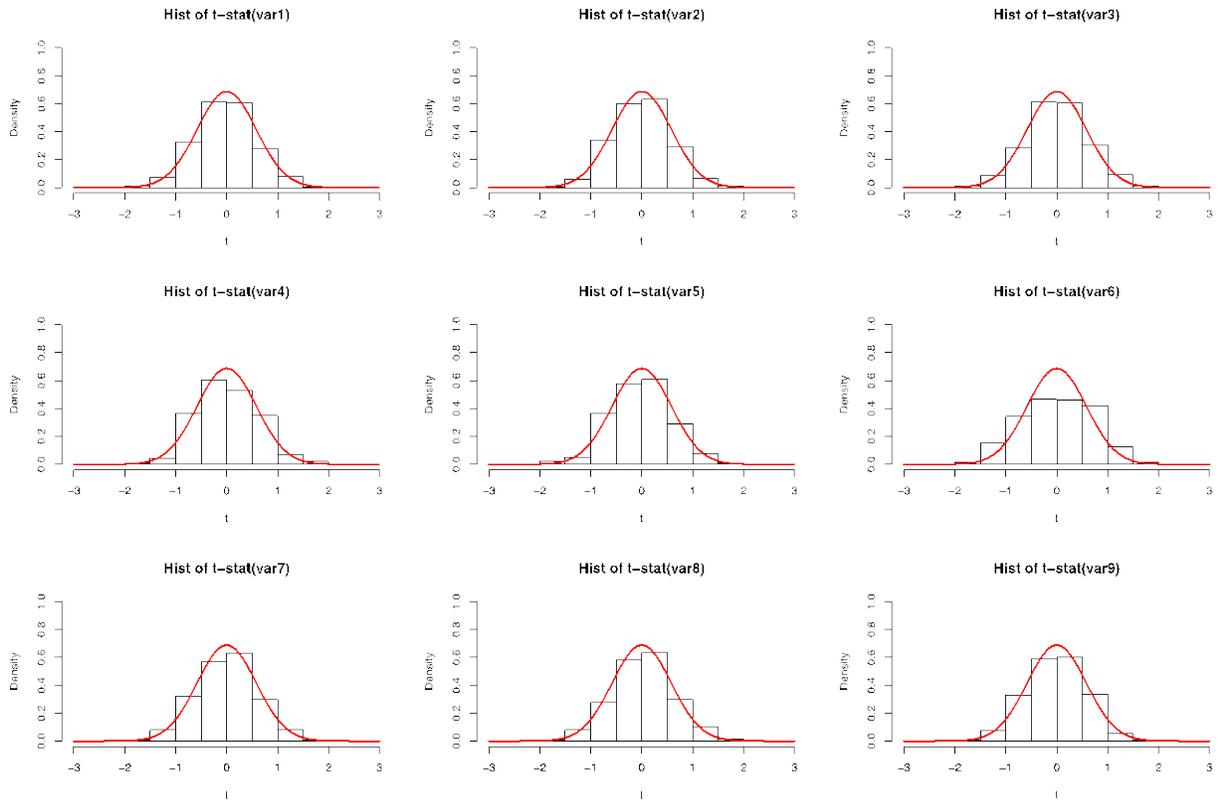


Figure 4. Histograms of the t-statistics ($n=200$, $\log N(0,0.5)$, variables 1 to 9)

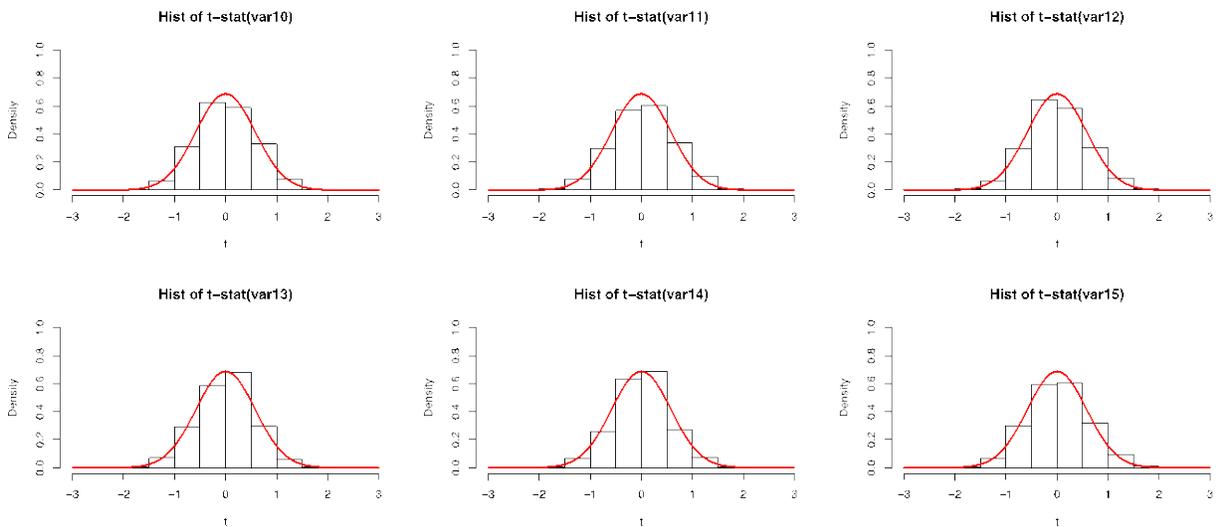


Figure 5. Histograms of the t-statistics ($n=200$, $\log N(0,0.5)$, variables 10 to 15)

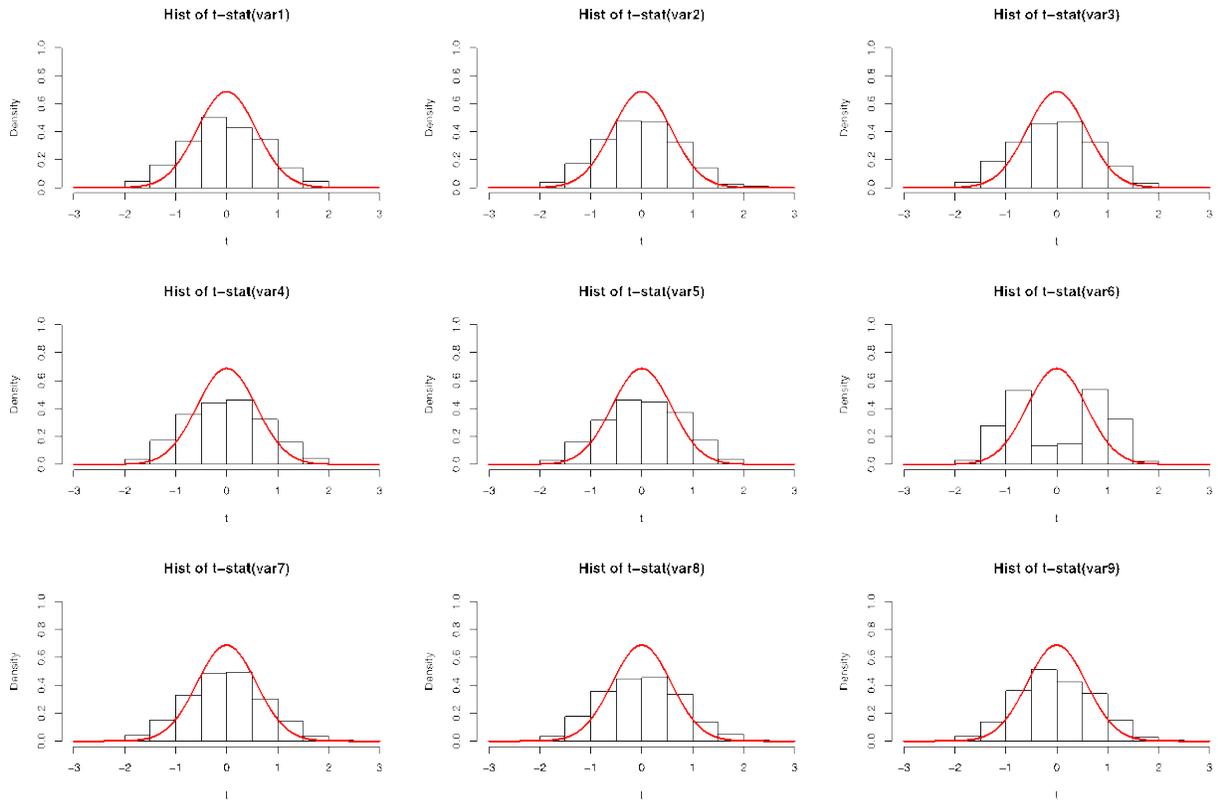


Figure 6. Histograms of the t-statistics (n=200, logN(0,1), variables 1 to 9)

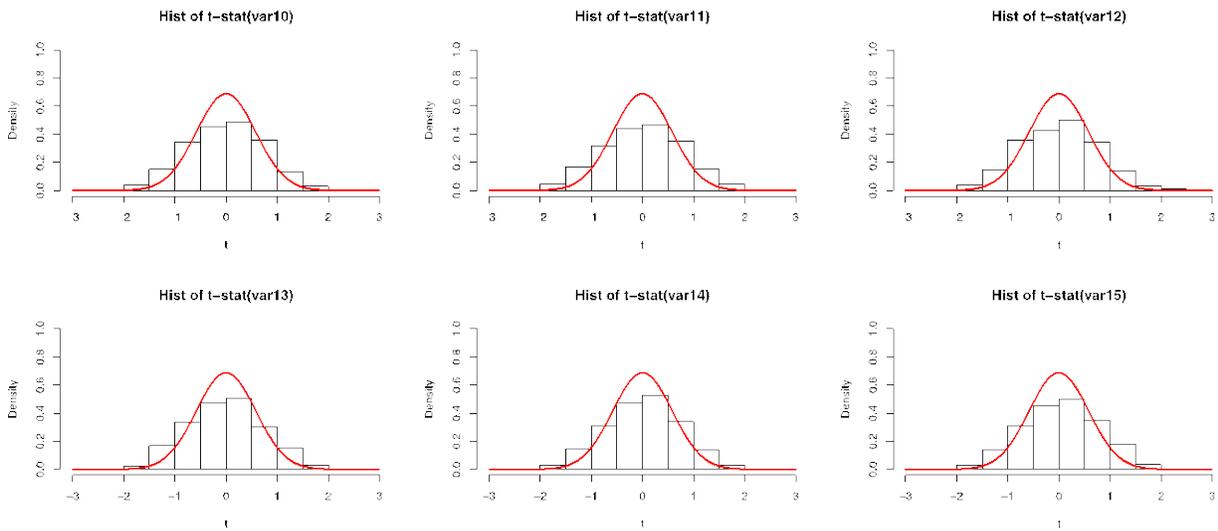


Figure 7. Histograms of the t-statistics (n=200, logN(0,1), variables 10 to 15)

Tables 10a, 10b, and 10c show the effect of skewed distributions on allocation imbalance for $n=200$, $n=100$, and $n=50$. In Table 10a, the distribution had little to no effect on allocation balance for rank-minimization when the sample size was 200, with 32-33% having no difference in sample size, 52% having a difference of 2, and 13-14% having a difference of 4. More skewed distributions did however show a better balance with minimization. There was no difference in sample size for 67% of simulations with the standard normal distribution versus 70% and 89% for $\log N(0,0.5)$ and $\log N(0,1)$, respectively. A difference of 2 in the sample size was observed for 32% with a standard normal distribution versus 30% and 11% for $\log N(0,0.5)$ and $\log N(0,1)$, respectively. The same pattern could also be observed for $n=100$ and for $n=50$. The performance of allocation balance for rank-minimization did not vary a lot when the distributions became more and more skewed. However, the performance for minimization was better when the distributions were more skewed.

Table 10a. Effect of sample size on the allocation imbalance (n=200)

diff	Rank-minimization			Minimization		
	N(0,1)	logN(0,0.5)	logN(0,1)	N(0,1)	logN(0,0.5)	logN(0,1)
0	324	330	327	674	696	888
2	517	516	518	322	301	112
4	137	134	133	4	3	0
6	20	18	20	0	0	0
≥ 8	2	1	1	0	0	0
Total	1000	1000	1000	1000	1000	1000

Table 10b. Effect of sample size on the allocation imbalance (n=100)

diff	Rank-minimization			Minimization		
	N(0,1)	logN(0,0.5)	logN(0,1)	N(0,1)	logN(0,0.5)	logN(0,1)
0	478	474	475	709	740	919
2	479	482	481	290	259	81
4	43	44	44	1	1	0
6	0	0	0	0	0	0
≥ 8	0	0	0	0	0	0
Total	1000	1000	1000	1000	1000	1000

Table 10c. Effect of sample size on the allocation imbalance (n=50)

diff	Rank-minimization			Minimization		
	N(0,1)	logN(0,0.5)	logN(0,1)	N(0,1)	logN(0,0.5)	logN(0,1)
0	652	648	649	758	776	927
2	340	344	343	242	223	73
4	8	8	8	0	1	0
6	0	0	0	0	0	0
≥ 8	0	0	0	0	0	0
Total	1000	1000	1000	1000	1000	1000

4.0 DISCUSSION

In this simulation study, we found out that both sample size and distribution of baseline covariates would have impact on the performance of rank-minimization. When the distribution is more skewed or the sample size decreases, the performance of baseline covariate balances for rank-minimization and minimization become worse. Previous research showed that the performance of minimization was worse when the proportion of variables to sample size increases (Chi & Ibrahim, 2006). The allocation balance of rank-minimization was better when the sample size was smaller. This is due to one characteristic of binomial distribution. The number of people in treatment A should follow a binomial distribution with n =sample size and $p=0.5$. If the number of people in A is the mean of the binomial distribution, which is half the sample size, the $\text{diff} = |n_A - n_B|$ would be 0. When the sample size becomes larger, $P(x = \mu)$ under binomial distribution is smaller and it leads to a larger difference.

Overall, rank-minimization works better than minimization when trying to control continuous variables at baseline, which was the same conclusion from Stigsby and Taves (Stigsby & Taves, 2010). However, if the variables are very skewed, the rank-minimization is not the preference over minimization anymore. In fact, the performance for rank-minimization on baseline covariate balances is similar to the one for minimization. The researchers in the clinical trials should consider other factors to choose the suitable randomization methods.

On the other hand, rank-minimization yielded worse performance for allocation balance as compared to minimization. These results did not seem to be affected by sample size or the distribution of baseline covariates. Regardless of the symmetry or skewness of the covariate distributions, rank-minimization leads to worse balance in the sample size allocation.

In addition to the factors we considered, there are other potential limitations of rank-minimization and minimization. Although simulations indicate that minimization and rank-minimization effectively balance the groups, the assignment may be predictable for a given subject. However, a review of minimization algorithms indicated that this limitation is also a disadvantage of other randomization methods, including stratified and block randomization (Scott, McPherson, Ramsay, & Campbell, 2002).

Another issue that needs to be considered in the context of minimization methods is the corresponding statistical analysis method. Taves pointed out that the usual statistical analysis might not be suitable for the trials using minimization to assign patients (Suresh, 2011; Taves, 1974). Most analyses assumed that treatment assignment was purely random; however, minimization does not meet this criterion. The tables describing the distributions of t-statistics in this study for data generated from the standard normal distribution could also help us to have a rough idea how randomly the sample was separated into two groups. In theory, if a variable from the standard normal distribution is split into two groups totally randomly, both groups should follow the standard normal distribution. The t-statistics in one run should also be distributed normally with mean=0 and standard error (SE) = 1. Therefore, if the subjects are totally randomly assigned, there should be 68% t-statistics lying between 0 and 1. However, about 97% of t-statistics for rank-minimization lied between 0 and 1 when the sample size equaled 200. Also, 88% of t-statistics for minimization lied between 0 and 1 when the sample size equaled

200. Therefore, the patients assigned by rank-minimization and minimization are not purely randomly assigned.

Although the limitation may cause some concern, stratified randomization and other adaptive methods also suffer from the same limitation (Scott, McPherson, Ramsay, & Campbell, 2002). The distributions of t-statistics described in the last paragraph also describe the trade-off between baseline covariate balances and complete randomization. Using simple randomization may be problematic for the scenarios described, where the number of subjects is small relative to the total number of strata. These tradeoffs need to be considered in selecting the optimal randomization approach for a given scenario.

There are some limitations of this study. First, the variables considered here are all continuous variables. In practice however, minimization methods may need to consider a mix of categorical, ordinal, and continuous variables. Future research could consider the performance of rank-minimization on ordinal data and on the mix of different types of variables. In addition, this study only compared rank-minimization with Taves' minimization; comparisons with stratified blocked randomization and Pocock's and Simon's minimization should also be investigated to evaluate the subsequent balance in key baseline covariates. Taves and Stigsby pointed out that, in some scenarios, rank-minimization and minimization yield better performance as compared to stratified blocked randomization (Stigsby & Taves, 2010). Other literature showed that the performance of Pocock's and Simon's minimization was similar to Taves' minimization (Zielhuis et al., 1990). Our study indicated that the performance of rank-minimization might suffer substantially for highly skewed distributions.

To conclude, there are many factors need to be considered to choose the best allocation method for a given clinical trial. When trying to balance continuous variables, overall, rank-

minimization showed improved covariate balance as compared to minimization and stratified minimization. However, when the data is not normally distributed, other factors need to be considered make a final decision.

APPENDIX A: R CODE FOR THE SIMULATION

A.1 RANK-MINIMIZATION

```
rm(list=ls())
path<-" /Users/linjungyi/Documents/thesis"
setwd(path)
library(randtoolbox) #for generate data

## rank minimization function
rank_min<-function(data,num.var,n,n_trt){#n for sample size
trt<-rep(NA,n)
data<-cbind(trt,data)
for(subj in 1:n){
  #add one subject each time
  dat<-matrix(data[1:subj,],nrow=subj,ncol=1+num.var)
  dat<-cbind(dat,rep(NA,subj)) #the last col is for rank
  total_imbalance<-cbind(c(1:n_trt),rep(0,n_trt),rep(0,n_trt))
  #total_imb: rows->if assign to 1st trt, 2nd trt..
  #total_imb: col2--imb, col3--cal ties
  for(i in 2:(num.var+1)){
    # cal imb of each var once a time and add to total imb
    x<-dat[,i] #ith var in data, 1st col is trt
    dat[, (1+num.var+1)]<-rank(x) #sum rank in each trt
    rank<-rank(x)
    ntrtfunc<-n_trt
    imbalance<-NA
    while(ntrtfunc>0){
      #Assume i subj assign to A or B or ...
      #Assign to ntrtth trt, then ntrt-1 trt, until 1st trt
      dat[subj,1]<-ntrtfunc
      ntrtfunc2<-n_trt
      sum_rank<-NA
      while(ntrtfunc2>0){
        #sum the rank in each group
        sum_rank[ntrtfunc2]<-
          sum(dat[, (1+num.var+1)][dat[,1]==ntrtfunc2])
        ntrtfunc2<-ntrtfunc2-1
      }
      mean_rank<-mean(sum_rank,na.rm=T)
      sd_rank<-(sum_rank-mean_rank)^2
      imbalance[ntrtfunc]<-sum(sd_rank)
      ntrtfunc<-ntrtfunc-1
    }
  }
}
```

```

    }
    for(t in 1:n_trt){
      total_imbalance[t,2]<-
      total_imbalance[t,2]+imbalance[t]
    }
  }#end for(i)
  ## if tie, randomize the subject
  for(t in 1:n_trt){
    if(total_imbalance[t,2]==
    min(total_imbalance[,2],na.rm=T)){
      total_imbalance[t,3]=1
    }
  }
  if(sum(total_imbalance[,3])>=2){
    data[subj,1]<-
    sample(total_imbalance[,1][total_imbalance[,3]==1])[1]
  }else{
    data[subj,1]<-
    total_imbalance[,1][total_imbalance[,3]==1]
  }
}
return(data)
}

```

```

##### easy try
GFR<-c(6,9,1,5,3.5,7,8,2,3.5)
DIA<-c(5,8,9,1,4,3,6,7,2)
try<-cbind(GFR,DIA)
n_var<-2
sam_size<-9
num_trt<-2
final<-rank_min(try,n_var,sam_size,num_trt)

```

```

#####
## generate variable ##
#####
run<-1000 #1000 final
n<-50 #total sample size
num.var<-15 # number of variables
n_trt<-2 # number of trt
path2<-"/output/skew/rank/"
filetype<-"_50skew05"

```

```

#### Calculate t statistics
t<-matrix(NA,nrow=run,ncol=num.var)
#### Calculate diff
d<-matrix(NA,nrow=run,ncol=5)
for(j in 1:5){

for(i in 1:run){
## generate data
setSeed(i)
c<-congruRand(n, dim = num.var)
#var<-qnorm(c,0,1)

```

```

#var<-qlnorm(c,0,1)
var<-qlnorm(c,0,0.5)
final<-rank_min(var,num.var,n,n_trt)
write.csv(final,paste0(path,path2,j,"finalrank",filetype,".csv"))
for(v in 1:num.var){
  x1bar<-mean(final[,1+v][final[,1]==1],na.rm=T)
  x2bar<-mean(final[,1+v][final[,1]==2],na.rm=T)
  s1<-sd(final[,1+v][final[,1]==1],na.rm=T)
  s2<-sd(final[,1+v][final[,1]==2],na.rm=T)
  n1<-sum(final[,1]==1,na.rm=T)
  n2<-sum(final[,1]==2,na.rm=T)
  t[i,v]<-abs(x1bar-x2bar)/sqrt(s1^2/n1+s2^2/n2)
}
#### calculate the difference
d[i,j]<-abs(sum(final[,1]==1)-sum(final[,1]==2))
cat("i=",i)
}#end for(i)

#### sum how many t in interval
t01<-NA
t12<-NA
t23<-NA
tg3<-NA

for(v in 1:num.var){
  t01[v]<-sum(t[,v]>=0&t[,v]<1)
  t12[v]<-sum(t[,v]>=1&t[,v]<2)
  t23[v]<-sum(t[,v]>=2&t[,v]<3)
  tg3[v]<-sum(t[,v]>=3)
}

ttt<-cbind(t01,t12,t23,tg3)
write.csv(ttt,paste0(path,path2,j,"tt",filetype,".csv"))
cat("j=",j)
}#end for(j)
write.csv(d,paste0(path,path2,"diff",filetype,".csv"))

#####
meansd<-matrix(NA,nrow=10,ncol=8)
for(j in 1:5){
  tt<-read.csv(paste0(path,path2,j,"tt",filetype,".csv"))
  for(i in 1:4){
    meansd[j,2*i-1]<-mean(tt[,i+1])
    meansd[j,2*i]<-sd(tt[,i+1])/sqrt(num.var)
  }
}
colnames(meansd)<-
c("t01mean","t01sd","t12mean","t12sd","t23mean","t23sd","tg3mean","tg3sd")
write.csv(meansd,paste0(path,path2,"meansd",filetype,".csv"))

#####
d<-read.csv(paste0(path,path2,"diff",filetype,".csv"))
diff<-NA
for(j in 1:5){
  diff<-rbind(diff,table(d[,j])) #1st col is X
}
diff<-na.omit(diff)

```

```
rownames(diff)<-c(1,2,3,4,5);
write.csv(diff,paste0(path,path2,"diff_table",filetype,".csv"))
```

A.2 MINIMIZATION

```
rm(list=ls())
path<-" /Users/linjungyi/Documents/thesis"
setwd(path)
library(randtoolbox) #for generate data

## minimization function
minimize<-function(data,num.var,n,n_trt){#n_trt have to be 2 here
  trt<-rep(NA,n)
  data<-cbind(trt,data)
  for(subj in 1:n){
    #add one subject each time
    dat<-matrix(data[1:subj,],nrow=subj,ncol=1+num.var)
    ##dat<-cbind(dat,rep(NA,subj))
    total_imbalance<-cbind(c(1:n_trt),rep(0,n_trt),rep(0,n_trt))
    colnames(total_imbalance)<-c("trt","imbalance","min_imbalance")
    #total_imb: rows->if assign to 1st trt, 2nd trt..
    #total_imb:col2--imb, col3--cal ties
    for(i in 2:(num.var+1)){
      # cal imb of each var once a time and add to total imb

      #freq<-table(x) #count how many subj with that char in the trt

      ntrtfunc<-n_trt
      imbalance<-NA
      while(ntrtfunc>0){
        #Assume i subj assign to A or B or ...
        #Assign to ntrtth trt, then ntrt-1 trt, until 1 trt
        dat[subj,1]<-ntrtfunc
        ntrtfunc2<-n_trt
        sum_freq<-NA
        while(ntrtfunc2>0){
          #sum the freq in each group
          sum_freq[ntrtfunc2]<-
            sum(dat[,i][dat[,1]==ntrtfunc2]==dat[subj,i])
          ntrtfunc2<-ntrtfunc2-1
        }
        imbalance[ntrtfunc]<-abs(sum_freq[2]-sum_freq[1])
        ntrtfunc<-ntrtfunc-1
      } #end of while(ntrtfunc>0)
      for(t in 1:n_trt){
        total_imbalance[t,2]<-
          total_imbalance[t,2]+imbalance[t]
      }
    }#end for(i in 2:(num.var+1))
  }#end for(t in 1:n_trt){
```

```

        if(total_imbalance[t,2]==
min(total_imbalance[,2],na.rm=T)){
            total_imbalance[t,3]=1
        }
    }
    if(sum(total_imbalance[,3])>=2){
        data[subj,1]<-
sample(total_imbalance[,1][total_imbalance[,3]==1],
size=1,prob=rep(1,sum(total_imbalance[,3]==1)))
    }else{
        data[subj,1]<-
total_imbalance[,1][total_imbalance[,3]==1]
    }
} #end for(subj)
return(data)
}

#easy try & debug
num.var<-2
n<-10
n_trt<-2

set.seed(3)
(v1<-sample(x=c(0,1,2),size=n,replace=T,prob=c(0.2,0.6,0.2)))
(v2<-sample(x=c(0,1,2),size=n,replace=T,prob=c(0.2,0.6,0.2)))
v<-cbind(v1,v2)

data<-v

(result<-minimize(data=v,num.var=2,n=10,n_trt=2))

#####
## generate variable ##
#####
# each variable with 3 categories
run<-1000 #1000 final
n<-50 #total sample size
num_var<-15 # number of variables
n_trt<-2 # number of trt
path2<-"/output/skew/min2/"
filetype<-"_50skew05"

#### Calculate t statistics
t<-matrix(NA,nrow=run,ncol=num_var)
#### Calculate diff
d<-matrix(NA,nrow=run,ncol=5)

for(j in 1:5){

for(i in 1:run){
#generate data
setSeed(i)
c<-congruRand(n, dim = num_var)
#var<-qnorm(c,0,1)
#var<-qlnorm(c,0,1)
var<-qlnorm(c,0,0.5)

```

```

var1<-matrix(NA,nrow=n,ncol=num_var)
for(nv in 1:num_var){
  var1[,nv]<-ifelse(var[,nv]>=(mean(var[,nv])+sd(var[,nv])),2,
                    ifelse(var[,nv]<mean(var[,nv])-sd(var[,nv]),0,1))
}
final<-minimize(data=var1,num.var=num_var,n=n,n_trt=n_trt)
final<-cbind(final[,1],var)
write.csv(final,paste0(path,path2,j,"finalmin",filetype,".csv"))
for(v in 1:num_var){
  x1bar<-mean(final[,1+v][final[,1]==1],na.rm=T)
  x2bar<-mean(final[,1+v][final[,1]==2],na.rm=T)
  s1<-sd(final[,1+v][final[,1]==1],na.rm=T)
  s2<-sd(final[,1+v][final[,1]==2],na.rm=T)
  n1<-sum(final[,1]==1,na.rm=T)
  n2<-sum(final[,1]==2,na.rm=T)
  t[i,v]<-abs(x1bar-x2bar)/sqrt(s1^2/n1+s2^2/n2)
}
#### calculate the difference
d[i,j]<-abs(sum(final[,1]==1)-sum(final[,1]==2))
cat("i=",i)

}#end for(i in 1:run)

#### sum how many t in interval
t01<-NA
t12<-NA
t23<-NA
tg3<-NA

for(v in 1:num_var){
  t01[v]<-sum(t[,v]>=0&t[,v]<1)
  t12[v]<-sum(t[,v]>=1&t[,v]<2)
  t23[v]<-sum(t[,v]>=2&t[,v]<3)
  tg3[v]<-sum(t[,v]>=3)
}

ttt<-cbind(t01,t12,t23,tg3)
write.csv(ttt,paste0(path,path2,j,"tstat_min",filetype,".csv"))
cat("j=",j,"\n")
}#end for(j)

write.csv(d,paste0(path,path2,"diff_min",filetype,".csv"))

#####
meansd<-matrix(NA,nrow=10,ncol=8)
for(j in 1:5){
  tt<-read.csv(paste0(path,path2,j,"tstat_min",filetype,".csv"))
  for(i in 1:4){#t01 t12 t23 tg3
    meansd[j,2*i-1]<-mean(tt[,i+1])
    meansd[j,2*i]<-sd(tt[,i+1])/sqrt(num_var)
  }
}
colnames(meansd)<-
c("t01mean","t01sd","t12mean","t12sd","t23mean","t23sd","tg3mean","tg3sd")
write.csv(meansd,paste0(path,path2,"meansd_min",filetype,".csv"))

```

```
#####  
d<-read.csv(paste0(path,path2,"diff_min",filetype,".csv"))  
diff<-NA  
for(j in 1:5){  
  diff<-rbind(diff,table(d[, (j+1)]))  
}  
diff<-na.omit(diff)  
rownames(diff)<-c(1,2,3,4,5);  
write.csv(diff,paste0(path,path2,"diffmin_table",filetype,".csv"))
```

APPENDIX B: ADDITIONAL SIMULATIONS FROM SECTION 3.1

Table 11. Rank-minimization distribution of t-statistics (n=200, N(0,1))

Rank-minimization mean count \pm SE (n=200, N(0,1))					
Interval of t	1	2	3	4	5
0-1	967 1.97	968 1.95	967 1.95	967 1.92	967 1.98
1-2	33 1.97	32 1.95	33 1.95	3 1.92	33 1.98
2-3	0	0	0	0	0
>3	0	0	0	0	0

Table 12. Rank-minimization distribution of diff (n=200, N(0,1))

Rank-minimization diff (n=200, N(0,1))					
d	1	2	3	4	5
0	332	331	324	327	325
2	514	514	517	517	519
4	133	135	137	135	134
6	19	18	20	19	20
≥ 8	2	2	2	2	2
Total	1000	1000	1000	1000	1000

Table 13. Rank-minimization distribution of t-statistics (n=100, N(0,1))

Rank-minimization mean count \pm SE (n=100, N(0,1))					
Interval of t	1	2	3	4	5
0-1	916 .81	916	916 3.14	916 3.06	916
1-2	83 2.81	83 2.68	84 3.14	84 3.06	83 2.95
2-3	1	0	0	0	0
>3	0	0	0	0	0

Table 14. Rank-minimization distribution of diff (n=100, N(0,1))

Rank-minimization diff (n=100, N(0,1))					
d	1	2	3	4	5
0	478	477	475	471	472
2	479	479	481	487	487
4	43	44	44	42	41
6	0	0	0	0	0
≥ 8	0	0	0	0	0
Total	1000	1000	1000	1000	1000

Table 15. Rank-minimization distribution of t-statistics (n=50, N(0,1))

Rank-minimization mean count \pm SE (n=50, N(0,1))					
Interval of t	1	2	3	4	5
0-1	840 3.29	840 3.21	840 3.64	840 3.35	840 3.38
1-2	154 3.10	155 3.01	154 3.40	154 3.16	154 3.10
2-3	5 0.60	5 0.61	6 0.65	5 0.61	5 0.64
>3	0	0	0	0	0

Table 16. Rank-minimization distribution of diff (n=50, N(0,1))

Rank-minimization diff (n=50, N(0,1))					
d	1	2	3	4	5
0	652	646	648	647	650
2	340	346	344	345	342
4	8	8	8	8	8
6	0	0	0	0	0
≥ 8	0	0	0	0	0
Total	1000	1000	1000	1000	1000

Table 17. Rank-minimization distribution of t-statistics (n=200, logN(0,0.5))

Rank-minimization mean count \pm SE (n=200, logN(0,0.5))					
Interval of t	1	2	3	4	5
0-1	911 5.68	911 5.59	911 5.62	912 5.63	911 5.71
1-2	88 5.68	89 5.59	89 5.62	89 5.63	89 5.71
2-3	0	0	0	0	0
>3	0	0	0	0	0

Table 18. Rank-minimization distribution of diff (n=200, logN(0,0.5))

Rank-minimization diff (n=200, logN(0,0.5))					
d	1	2	3	4	5
0	330	331	330	329	333
2	516	517	513	517	513
4	134	132	133	134	131
6	18	18	22	18	21
≥ 8	1	1	1	1	1
Total	1000	1000	1000	1000	1000

Table 19. Rank-minimization distribution of t-statistics (n=100, logN(0,0.5))

Rank-minimization mean count ± SE (n=100, logN(0,0.5))					
Interval of t	1	2	3	4	5
0-1	867±8.15	867±7.91	867±8.37	867±8.07	867±7.91
1-2	131±8.23	131±8.00	132±8.39	132±8.16	132±8.01
2-3	1±0.25	1±0.25	1±0.24	1±0.25	1±0.26
>3	0	0	0	0	0

Table 20. Rank-minimization distribution of diff (n=100, logN(0,0.5))

Rank-minimization diff (n=100, logN(0,0.5))					
d	1	2	3	4	5
0	474	475	476	478	471
2	482	483	478	478	485
4	44	42	46	44	44
6	0	0	0	0	0
≥8	0	0	0	0	0
Total	1000	1000	1000	1000	1000

Table 21. Rank-minimization distribution of t-statistics (n=50, logN(0,0.5))

Rank-minimization mean count ± SE (n=50, logN(0,0.5))					
Interval of t	1	2	3	4	5
0-1	812 4.01	811 4.15	813 4.28	814 3.99	813 4.12
1-2	182 3.71	183 3.82	181 3.88	180 3.60	181 3.79
2-3	6 0.64	6 0.60	6 0.66	6 0.68	6 0.68
>3	0	0	0	0	0

Table 22. Rank-minimization distribution of diff (n=50, logN(0,0.5))

Rank-minimization diff (n=50, logN(0,0.5))					
d	1	2	3	4	5
0	648	644	653	650	652
2	344	348	339	342	340
4	8	8	8	8	8
6	0	0	0	0	0
≥8	0	0	0	0	0
Total	1000	1000	1000	1000	1000

Table 23. Rank-minimization distribution of t-statistics (n=200, logN(0,1))

Rank-minimization mean count ± SE (n=200, logN(0,1))					
Interval of t	1	2	3	4	5
0-1	797±9.33	796±9.54	795±9.37	797±9.40	797±9.42
1-2	200±9.56	200±9.79	202±9.60	200±9.64	200±9.65
2-3	3±0.64	4±0.65	4±0.63	3±0.63	3±0.62
>3	0	0	0	0	0

Table 24. Rank-minimization distribution of diff (n=200, logN(0,1))

Rank-minimization diff (n=200, logN(0,1))					
d	1	2	3	4	5
0	327	330	323	324	327
2	518	516	517	515	516
4	133	132	138	137	136
6	20	20	20	22	19
8	1	1	1	1	1
Total	1000	1000	1000	1000	1000

Table 25. Rank-minimization distribution of t-statistics (n=100, logN(0,1))

Rank-minimization mean count \pm SE (n=100, logN(0,1))					
Interval of t	1	2	3	4	5
0-1	779 \pm 10.98	780 \pm 10.71	780 \pm 10.48	780 \pm 10.80	780 \pm 10.80
1-2	216 \pm 11.28	216 \pm 10.99	215 \pm 10.79	216 \pm 11.09	215 \pm 11.11
2-3	4 \pm 0.69	4 \pm 0.63	5 \pm 0.60	4 \pm 0.61	4 \pm 0.62
>3	0	0	0	0	0

Table 26. Rank-minimization distribution of diff (n=100, logN(0,1))

Rank-minimization diff (n=100, logN(0,1))					
d	1	2	3	4	5
0	475	473	472	476	478
2	481	483	484	480	478
4	44	44	44	44	44
6	0	0	0	0	0
8	0	0	0	0	0
Total	1000	1000	1000	1000	1000

Table 27. Rank-minimization distribution of t-statistics (n=50, logN(0,1))

Rank-minimization mean count \pm SE (n=50, logN(0,1))					
Interval of t	1	2	3	4	5
0-1	755 \pm 3.68	756 \pm 3.50	756 \pm 3.69	756 \pm 3.44	756 \pm 3.76
1-2	237 \pm 3.43	236 \pm 3.33	236 \pm 3.40	236 \pm 3.27	236 \pm 3.53
2-3	8 \pm 0.73	8 \pm 0.60	8 \pm 0.70	8 \pm 0.68	8 \pm 0.77
>3	0	0	0	0	0

Table 28. Rank-minimization distribution of diff (n=50, logN(0,1))

Rank-minimization diff (n=50, logN(0,1))					
d	1	2	3	4	5
0	649	644	647	649	648
2	343	348	345	343	344
4	8	8	8	8	8
6	0	0	0	0	0
8	0	0	0	0	0
Total	1000	1000	1000	1000	1000

Table 29. Minimization distribution of t-statistics (n=200, N(0,1))

Minimization mean count \pm SE (n=200, N(0,1))					
Interval of t	1	2	3	4	5
0-1	875 2.43	870 2.59	876 2.15	872 2.98	875 2.44
1-2	121	126 2.61	121 2.04	124 2.87	120 2.37
2-3	4 0.35	4 0.71	4 0.69	3 0.49	4 0.33
>3	0 0.07	0	0	0 0.07	0

Table 30. Minimization distribution of diff (n=200, N(0,1))

Minimization diff (n=200, N(0,1))					
d	1	2	3	4	5
0	674	673	679	676	657
2	322	326	318	316	342
4	4	1	3	8	1
6	0	0	0	0	0
≥ 8	0	0	0	0	0
Total	1000	1000	1000	1000	1000

Table 31. Minimization distribution of t-statistics (n=100, N(0,1))

Minimization mean count \pm SE (n=100, N(0,1))					
Interval of t	1	2	3	4	5
0-1	826 2.98	828 4.07	827 2.87	830 3.57	829 2.93
1-2	165 3.08	163 3.84	165 3.16	162 3.38	162
2-3	9 0.80	8 0.68	8 0.76	8 0.76	9 0.71
>3	0 0.12	0 0.13	0 0.11	0 0.11	0 0.16

Table 32. Minimization distribution of diff (n=100, N(0,1))

Minimization diff (n=100, N(0,1))					
d	1	2	3	4	5
0	709	745	705	724	703
2	290	255	294	276	297
4	1	0	1	0	0
6	0	0	0	0	0
≥ 8	0	0	0	0	0
Total	1000	1000	1000	1000	1000

Table 33. Minimization distribution of t-statistics (n=50, N(0,1))

Minimization mean count \pm SE (n=50, N(0,1))					
Interval of t	1	2	3	4	5
0-1	784 3.60	781 2.85	788 3.59	789 3.25	787 2.53
1-2	197 3.61	200 2.75	194 3.90	193 3.03	196 2.03
2-3	19 0.70	18 1.11	18 0.73	16 0.90	16 1.07
>3	1 0.22	1 0.17	0 0.17	1 0.33	1 0.17

Table 34. Minimization distribution of diff (n=50, N(0,1))

Minimization diff (n=50, N(0,1))					
d	1	2	3	4	5
0	758	769	752	745	760
2	242	231	248	255	240
4	0	0	0	0	0
6	0	0	0	0	0
≥ 8	0	0	0	0	0
Total	1000	1000	1000	1000	1000

Table 35. Minimization distribution of t-statistics (n=200, logN(0,0.5))

Minimization mean count \pm SE (n=200, logN(0,0.5))					
Interval of t	1	2	3	4	5
0-1	840 7.24	841 4.44	844 4.63	840 6.46	837 5.45
1-2	154 7.18	153 4.22	149 4.41	153 6.17	157 5.31
2-3	5	6	6 0.64	7 0.74	6 0.55
>3	0	0	0	0	0

Table 36. Minimization distribution of diff (n=200, logN(0,0.5))

Minimization diff (n=200, logN(0,0.5))					
d	1	2	3	4	5
0	696	691	685	694	685
2	301	308	311	305	311
4	3	1	4	1	4
6	0	0	0	0	0
≥ 8	0	0	0	0	0
Total	1000	1000	1000	1000	1000

Table 37. Minimization distribution of t-statistics (n=100, logN(0,0.5))

Minimization mean count ± SE (n=100, logN(0,0.5))					
Interval of t	1	2	3	4	5
0-1	793 6.25	795 4.83	798 5.61	802 5.18	802 5.75
1-2	192 6.50	193 4.73	190 5.50	185 5.02	187 5.62
2-3	14 0.96	11 0.88	12 0.83	12 1.08	11 1.04
>3	0 0.09	0 0.19	0 0.12	0 0.13	0 0.16

Table 38. Minimization distribution of diff (n=100, logN(0,0.5))

Minimization diff (n=100, logN(0,0.5))					
d	1	2	3	4	5
0	740	732	748	732	726
2	259	268	250	268	273
4	1	0	2	0	1
6	0	0	0	0	0
≥8	0	0	0	0	0
Total	1000	1000	1000	1000	1000

Table 39. Minimization distribution of t-statistics (n=50, logN(0,0.5))

Minimization mean count ± SE (n=50, logN(0,0.5))					
Interval of t	1	2	3	4	5
0-1	772 3.69	773 2.70	771 3.00	766 3.30	764 2.88
1-2	208 3.40	207 3.10	208 3.01	214 2.94	215 3.30
2-3	19 0.71	19 1.68	19 1.49	19 1.31	20 1.18
>3	1 0.22	1 0.28	1 0.33	1 0.24	1 0.21

Table 40. Minimization distribution of diff (n=50, logN(0,0.5))

Minimization diff (n=50, logN(0,0.5))					
d	1	2	3	4	5
0	776	785	784	782	795
2	223	215	216	218	204
4	1	0	0	0	1
6	0	0	0	0	0
≥8	0	0	0	0	0
Total	1000	1000	1000	1000	1000

Table 41. Minimization distribution of t-statistics (n=200, logN(0,1))

Minimization mean count ± SE (n=200, logN(0,1))					
Interval of t	1	2	3	4	5
0-1	813 10.20	809 11.39	808 11.25	817 11.79	813 10.94
1-2	182 10.36	185 11.72	187 11.48	179 10.90	181 11.09
2-3	5	6	5 0.51	4 0.49	5 0.54
>3	0	0	0	0	0

Table 42. Minimization distribution of diff (n=200, logN(0,1))

Minimization diff (n=200, logN(0,1))					
d	1	2	3	4	5
0	888	860	886	867	866
2	112	140	114	133	134
4	0	0	0	0	0
6	0	0	0	0	0
8	0	0	0	0	0
Total	1000	1000	1000	1000	1000

Table 43. Minimization distribution of t-statistics (n=100, logN(0,1))

Minimization mean count \pm SE (n=100, logN(0,1))					
Interval of t	1	2	3	4	5
0-1	775 11.01	776 12.19	771 11.16	770 12.72	774 14.05
1-2	215 11.53	213 12.86	218 11.79	220 13.28	216 14.68
2-3	9 0.86	10 1.08	10 0.97	9 1.00	9 0.88
>3	0 0.07	0 0.12	0 0.07	0 0.13	0 0.07

Table 44. Minimization distribution of diff (n=100, logN(0,1))

Minimization diff (n=100, logN(0,1))					
d	1	2	3	4	5
0	919	909	910	914	920
2	81	91	90	86	80
4	0	0	0	0	0
6	0	0	0	0	0
8	0	0	0	0	0
Total	1000	1000	1000	1000	1000

Table 45. Minimization distribution of t-statistics (n=50, logN(0,1))

Minimization mean count \pm SE (n=50, logN(0,1))					
Interval of t	1	2	3	4	5
0-1	760 3.72	758 4.08	756 4.73	761 5.07	751 4.34
1-2	224 4.00	226 3.93	230 4.69	224 4.45	234 4.07
2-3	16 0.87	16 0.33	14 0.20	15 0.40	15 0.47
>3	1 0.19	0 0.16	0 0.11	0 0.16	0 0.19

Table 46. Minimization distribution of diff (n=50, logN(0,1))

Minimization diff (n=50, logN(0,1))					
d	1	2	3	4	5
0	927	917	915	917	933
2	73	83	85	83	67
4	0	0	0	0	0
6	0	0	0	0	0
8	0	0	0	0	0
Total	1000	1000	1000	1000	1000

BIBLIOGRAPHY

- Altman, D. G., & Bland, J. M. (1999). Statistics notes. Treatment allocation in controlled trials: why randomise? *BMJ*, 318(7192), 1209.
- Chi, Y. Y., & Ibrahim, J. G. (2006). Joint models for multivariate longitudinal and multivariate survival data. *Biometrics*, 62(2), 432-445. doi:10.1111/j.1541-0420.2005.00448.x
- Hoehler, F. K. (1987). Balancing allocation of subjects in biomedical research: a minimization strategy based on ranks. *Comput Biomed Res*, 20(3), 209-213.
- Park, S. K., Miller, K. W. . (1988). Random number generators: good ones are hard to find. *Commun. ACM*, 31(10), 1192-1201. doi:10.1145/63039.63042
- Pocock, S. J., & Simon, R. (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics*, 31(1), 103-115.
- Prevention of neural tube defects: results of the Medical Research Council Vitamin Study. MRC Vitamin Study Research Group. (1991). *Lancet*, 338(8760), 131-137.
- Schneider, M., Haas, M., Glick, R., Stevans, J., & Landsittel, D. (2015). Comparison of spinal manipulation methods and usual medical care for acute and subacute low back pain: a randomized clinical trial. *Spine (Phila Pa 1976)*, 40(4), 209-217. doi:10.1097/BRS.0000000000000724
- Scott, N. W., McPherson, G. C., Ramsay, C. R., & Campbell, M. K. (2002). The method of minimization for allocation to clinical trials. a review. *Control Clin Trials*, 23(6), 662-674.
- Smithells, R. W., & Sheppard, S. (1980). Possible prevention of neural-tube defects by periconceptional vitamin supplementation. *Lancet*, 1(8169), 647.
- Stigsby, B., & Taves, D. R. (2010). Rank-Minimization for balanced assignment of subjects in clinical trials. *Contemp Clin Trials*, 31(2), 147-150. doi:10.1016/j.cct.2009.12.001
- Suresh, K. (2011). An overview of randomization techniques: An unbiased assessment of outcome in clinical research. *J Hum Reprod Sci*, 4(1), 8-11. doi:10.4103/0974-1208.82352

- Tang, A. M., Tang, N. S., & Zhu, H. (2017). Influence analysis for skew-normal semiparametric joint models of multivariate longitudinal and multivariate survival data. *Stat Med*. doi:10.1002/sim.7211
- Taves, D. R. (1974). Minimization: a new method of assigning patients to treatment and control groups. *Clin Pharmacol Ther*, 15(5), 443-453.
- Therneau, T. (1993). How many stratification factors are "too many" to use in a randomization plan? *Control Clin Trials*, 14(2), 98-108. doi:10.1016/0197-2456(93)90013-4
- Zielhuis, G. A., Straatman, H., van 't Hof-Grootenboer, A. E., van Lier, H. J., Rach, G. H., & van den Broek, P. (1990). The choice of a balanced allocation method for a clinical trial in otitis media with effusion. *Stat Med*, 9(3), 237-246.