4-1-2011 12:00 AM

Confidence Interval Estimation for Continuous Outcomes in Cluster Randomization Trials

Julia Taleban, *The University of Western Ontario*

Supervisor: Dr. Guang Yong Zou, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Biostatistics

CONFIDENCE INTERVAL ESTIMATION FOR CONTINUOUS OUTCOMES IN CLUSTER RANDOMIZATION TRIALS

(Spine title: Confidence Intervals in Cluster Randomization Trials)

(Thesis format: Monograph)

by

Julia Taleban

Graduate Program in Epidemiology & Biostatistics

A thesis submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

School of Graduate and Postdoctoral Studies University of Western Ontario
London, Ontario
April 27, 2011

THE UNIVERSITY OF WESTERN ONTARIO
SCHOOL OF GRADUATE AND POSTDOCTORAL STUDIES

CERTIFICATE OF EXAMINATION

Supervisor

Dr. Guangyong Zou

Co-Supervisor

Dr. John Koval

Committee member

Dr. Allan Donner

Examining Board

Dr. Neil Klar

Dr. Duncan Murdoch

Dr. Serge Provost

Dr. Wendy Lou

The thesis by
Julia Taleban

Entitled
Confidence Interval Estimation for Continuous Outcomes in Cluster Randomization Trials

is accepted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Date: _____

Chair of the Thesis Examination Board

Abstract

Cluster randomization trials are experiments where intact social units (e.g. hospitals, schools, communities, and families) are randomized to the arms of the trial rather than individuals. The popularity of this design among health researchers is partially due to reduced contamination of treatment effects and convenience. However, the advantages of cluster randomization trials come with a price. Due to the dependence of individuals within a cluster, cluster randomization trials suffer reduced statistical efficiency and often require a complex analysis of study outcomes.

The primary purpose of this thesis is to propose new confidence intervals for effect measures commonly of interest for continuous outcomes arising from cluster randomization trials. Specifically, we construct new confidence intervals for the difference between two normal means, the difference between two lognormal means, and the exceedance probability.

The proposed confidence intervals, which use the method of variance estimates recovery (MOVER), do not make certain assumptions that existing procedures make on the data. For instance, symmetry is not forced when the sampling distribution of the parameter estimate is skewed and the assumption of homoscedasticity is not made. Furthermore, the MOVER results in simple confidence interval procedures rather than complex simulation-based methods which currently exist.

Simulation studies are used to investigate the small sample properties of the MOVER as compared with existing procedures. Unbalanced cluster sizes are simulated, with an average range of 50 to 200 individuals per cluster and 6 to 24 clusters per arm. The effects of various degrees of dependence between individuals within the same cluster are also investigated.

When comparing the empirical coverage, tail errors, and median widths of confidence interval procedures, the MOVER has coverage close to the nominal, relatively balanced tail errors, and narrow widths as compared to existing procedure for the

majority of the parameter combinations investigated. Existing data from cluster randomization trials are then used to illustrate each of the methods.

Keywords: cluster randomization trials, confidence intervals, normal mean, lognormal mean, exceedance probability, method of variance estimates recovery, generalized confidence interval procedure, Wald method.

Acknowledgments

The completion of this thesis would not have been possible without the help and support of many people. First, I would like to thank my Ph.D. supervisor, Dr. Guang Yong Zou, who provided me with invaluable guidance and insight. I would also like to thank my co-supervisor, Dr. John Koval, whose encouragement and suggestions have been a great help, and Dr. Allan Donner for the time he took to discuss the drafts of the thesis and his helpful suggestions.

I am grateful to Dr. Thomas Marrie and Dr. Alan Montgomery for the use of their datasets while illustrating various confidence interval methods in Chapter 5.

I am indebted to my parents and friends for their moral support and encouragement. My fiance Luc has shown constant support and endless patience, and for that I am thankful.

This work has been financially supported in part by the Ontario Graduate Scholarship in Science and Technology and by the Ontario Graduate Scholarship from the Ontario Ministry of Training, Colleges, and Universities.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 Randomized controlled trials

A randomized controlled trial is an experiment that randomly assigns units to different intervention groups, usually a control group and an experimental group. Successful randomization can prevent selection bias and balance intervention groups for both known and unknown factors associated with the outcome of interest such that the arms of the trial will differ only according to the intended difference being tested (Julious and Zariffa, 2002). Thus, a well conducted randomized controlled trial is usually considered as the gold standard for evaluating health care interventions. Consequently, many systematic reviews of the effects of health care interventions, such as those of the Cochrane Collaboration, are carried out primarily using data from randomized controlled trials (Clarke *et al.*, 1999, section 13.1.3).

1.2 Cluster randomization trials

Cluster randomization trials are a special case of randomized controlled trials where intact social units, rather than individuals, are randomly allocated to the arms of the trial. These social units are commonly termed "clusters".

Cluster randomization trials have a number of interesting properties. As with individually randomized trials, on average, cluster randomization balances comparative groups according to known and unknown factors associated with the outcome of interest (Donner and Klar, 2000, page 11). Therefore, it is used to eliminate confounding effects at both the cluster level and the individual level.

Cluster sizes are diverse, ranging from as few as one or two individuals to hundreds of thousands. Examples of clusters include families, schools, work sites, physician practices, and entire communities. Another example is given by Daltroy *et al.* (2007), who randomized ferry boats on which entertainment troupes delivered an educational intervention about how to prevent Lyme disease by avoiding contact with ticks.

A distinctive property of cluster randomization trials is that observations within each cluster may be correlated. If inferences are made at the cluster level, then this is not an issue. However, if inferences at the individual level are of interest, clustering may lead to reduced efficiency and a potentially complicated analysis as compared to individual level randomization. Consequently, the effects of clustering must be accounted for in the design and the analysis of the study if the unit of inference (e.g. the individual) differs from the unit of random allocation (e.g. the cluster).

Due to the potential loss in efficiency and the increased complexity of the analysis of cluster randomization trials when the unit of randomization differs from the unit of inference, justification for adopting the clustered design is necessary. Several common reasons warrant the use of a cluster randomization trial. First, interventions may only be delivered at the cluster level. This strategy was adopted by Williamson *et al.* (2007) who compared the effects of an environmental intervention (including changes to the cafeteria menu) to those of an active control for preventing weight gain in children. Since the change to the cafeteria menu is designed to be delivered to an entire school, it would not be feasible to randomize individuals to receive this intervention. Randomization must necessarily be at the cluster level, the school.

Second, given that individuals in the same cluster are likely to interact, randomizing clusters to intervention groups commonly reduces contamination across the arms of a study. For example, Sankaranarayanan *et al.* (2007) evaluated the effect of visual screening as compared to usual care on cervical cancer incidence and mortality in India by randomizing villages to reduce contamination within each village.

Third, a clustered design can enhance the compliance of participants or investi-

gators, improve logsitic convenience, or decrease the cost of the study. For instance, Trevino *et al.* (2004) randomized schools to receive either an educational intervention or usual care for the purpose of decreasing capillary glucose levels in an attempt to control diabetes mellitus. A clustered design was chosen because programs which include social support and peer pressure tend to improve compliance (Perri *et al.*, 1988).

Finally, if the objective of a study is to obtain the total effect of an intervention, a clustered design may be employed. With certain interventions it may be feasible to randomize at either the individual or cluster level, but if the question of interest lies at the cluster level, the unit of randomization must necessarily be the cluster. A common example arises from studies of infectious disease, where both direct and indirect effects are of interest (Hayes and Moulton, 2009, Chapter 3). Direct effects of an intervention such as an immunization may be captured using an individually randomized trial. However, indirect effects include the effects of an intervention which is administered to others, such as herd immunity. Unfortunately, this type of effect cannot be measured in a trial which randomizes individuals. A cluster randomization trial on the other hand would allow the measure of direct, indirect, and total (the combination of direct and indirect) effects of the intervention. Melese *et al.* (2008) were interested in the effect of biannual antibiotic treatment on infectious trachoma as compared to an annual treatment. The total effect of antibiotic treatment on the population was of interest. This was measured by the comparison of village prevalence of occular chlamydial infection, therefore villages were randomized to each arm.

Oftentimes a cluster randomization trial design is used for multiple reasons. For example, a recent cluster randomization trial (Pandey *et al.*, 2007) randomized districts (each containing 5 village clusters) in the Uttar Pradesh state in India to investigate the impact on health and social services delivered by the villages by informing resource-poor rural populations about entitled services. The advantage of randomization by cluster in this case was administrative convenience, as the districts were

obtained from a single state which determined both health and educational services. Furthermore, randomization by cluster was adopted to lower the level of contamination, as travel between intervention and control villages would have been difficult.

When there are acceptable justifications for clustering, there are a number of trial designs to choose from. The most commonly used clustered designs are the completely randomized design, the stratified design, and the matched-pair design.

In a completely randomized cluster trial all of the clusters involved in the study are randomly allocated to the arms of the trial without the statifying or mathing baseline characteristics. For instance, this design was used by Vizcaino *et al.* (2008), where no stratification was used to randomize twenty schools to receive either a physical activity program or usual care to prevent childhood obesity. In this case, random allocation was performed using a computer generated procedure.

An implication of randomizing clusters, however, is that only a small number of clusters may be involved in the study, especially when the cluster sizes are large. As such, randomization may not ensure balance of cluster level confounders between arms. In such cases, stratification or matching may be useful to increase balance between the arms of a trial. For the stratified and matched-pair designs, the clusters in each stratum are assigned according to baseline characteristics or similarities between clusters such that these characteristics are potentially related to the outcome. Examples of common matching characteristics include cluster size and geographical location.

A stratified design assigns multiple clusters to different interventions within each stratum. This design was adopted by Naylor *et al.* (2006) who studied the effect of a physical-activity action plan in schools on physical activity level. With so few clusters, cluster size and geographic location may not have been balanced between trial arms, so the ten schools were first stratified by these variables, then the clusters within each stratum were randomly assigned to the two study arms.

A matched pair design is a special case of a stratified design where there are only

two clusters within a stratum, such that there is a very tight balance of baseline characteristics associated with the outcome between these two clusters. Each cluster in a pair is then randomly allocated to a different arm. As an example, Flannery *et al.* (2003) matched eight schools based on geographic proximity, the percentage of ethnic students, the percentage of students eligible for free or reduced-price lunch, and the percentage of students in English as a Second Language classes. These clusters were randomized to determine the effect of an immediate social peace building intervention on violence prevention, as compared to a delayed intervention. It is important to recognize that the lack of replication of clusters to each arm within each stratum complicates the ability of quantifying the similarity of individuals within each cluster as compared to between clusters, because the variation of responses between clusters is confounded with the effect of intervention. Thus, data from such a design is usually analyzed at the cluster level (Klar and Donner, 1997).

1.3 Notation

Correlated continuous outcomes commonly arise from cluster randomization trials, where these responses may often be approximated with the normal distribution. To allow for a more detailed discussion about useful effect measures, we now digress to introduce some notation by starting with an assumption that data from each arm of the trials is assumed to follow a one-way random effects model. This assumption allows two levels of variance - one at the individual level and one at the cluster level. Furthermore, attention is limited to a completely randomized trial design for the sake of simplicity.

Suppose k_i clusters are randomized to intervention i $(i = 1, 2)$. Let m_{ij} denote the j^{th} cluster size $(j = 1, ...k_i)$ of arm i. Now let

$$Y_{ijl} = \mu_i + A_{ij} + E_{ijl}, \tag{1.1}$$

where Y_{ijl} denotes the l^{th} observed outcome $(l = 1, ..., m_{ij})$ from the j^{th} cluster of

arm i, μ_i is the population mean of arm i, and $A_{ij} \sim N(0, \sigma^2_{A_i})$ is independent of $E_{ijl} \sim N(0, \sigma^2_{E_i})$.

Continuing with the case of two arms, the three parameters in each arm include the overall mean (μ_i), the between-group (or between-cluster) variance component ($\sigma^2_{A_i}$), and the within-group (within-cluster) variance component ($\sigma^2_{E_i}$). Observations can then be used to estimate these three parameters. Attention is limited to Analysis of Variance estimators of variance components, as recommended for low to moderate values of the intraclass correlation coefficient (ICC) (see Section 1.4.2) (Donner and Koval, 1980). This thesis focuses on small to moderate values of the ICC, as discussed in Chapter 4.

The overall mean of arm i, μ_i, may be estimated by

$$\bar{\bar{Y}}_i = M_i^{-1} \sum_j \sum_l Y_{ijl},$$

where $M_i = \sum_j m_{ij}$ denotes the total number of individuals in arm i. The cluster-specific mean may be estimated by

$$\bar{Y}_{ij} = m_{ij}^{-1} \sum_l Y_{ijl}.$$

To estimate the variance components, denote

$$\begin{aligned} \mathrm{MSC}_i &= \frac{\mathrm{SSC}_i}{k_i - 1} \text{ and} \\ \mathrm{MSW}_i &= \frac{\mathrm{SSW}_i}{M_i - k_i} \end{aligned}$$

as the mean squared errors between- and within-clusters, respectively, where

$$\begin{aligned} \mathrm{SSC}_i &= \sum_j m_{ij} \left(\bar{Y}_{ij} - \bar{\bar{Y}}_i \right)^2 \text{ and} \\ \mathrm{SSW}_i &= \sum_j \sum_l \left(Y_{ijl} - \bar{Y}_{ij} \right)^2. \end{aligned}$$

The between-cluster variance component may be estimated by

$$S^2_{A_i} = \frac{\mathrm{MSC}_i - \mathrm{MSW}_i}{m_{oi}}, \tag{1.2}$$

where

$$m_{oi} = \frac{M_i - \sum_j m_{ij}^2/M_i}{k_i - 1},$$

and the within-cluster variance component by

$$S_{E_i}^2 = \text{MSW}_i.$$

An estimate of the total variance in arm i is then given by $S_{T_i}^2 = S_{A_i}^2 + S_{E_i}^2$.

When all cluster sizes are equal (or balanced), the estimated parameters of each arm (assumed to follow a one-way random effects model) of a cluster randomization trial follow familiar distributions. That is,

$$\begin{aligned} \bar{\bar{Y}}_i &\sim N\left(\mu_i, \frac{\sigma_{A_i}^2 + \sigma_{E_i}^2/m_i}{k_i}\right) \text{ (Donner and Klar, 2000, page 8),} \\ \frac{\text{SSC}_i}{\sigma_{E_i}^2 + m_i\sigma_{A_i}^2} &\sim \chi_{k_i-1}^2, \text{ and} \\ \frac{\text{SSW}_i}{\sigma_{E_i}^2} &\sim \chi_{M_i-k_i}^2 \text{ (Graybill, 1976, page 609).} \end{aligned} \tag{1.3}$$

These properties may be used to construct confidence intervals for the overall mean and variance components of the model.

However cluster sizes are rarely balanced in practice. Although $\text{SSW}_i/\sigma_{E_i}^2 \sim \chi_{M_i-k_i}^2$ approximately holds for the unbalanced design, $\text{SSC}_i/(\sigma_{E_i}^2 + m_i\sigma_{A_i}^2)$ no longer follows a chi-squared distribution and the variance of $\bar{\bar{Y}}_i$ must be adjusted (Burdick *et al.*, 2006). Thomas and Hultquist (1978) proposed an unweighted mean squared error which may be used in the unbalanced design,

$$S_{Ui}^2 = \frac{n_{Hi}\sum_j(\bar{Y}_{ij} - \bar{\bar{Y}}_i)^2}{k_i - 1}, \tag{1.4}$$

where

$$\begin{aligned} E(S_{Ui}^2) &= \sigma_{E_i}^2 + n_{Hi}\sigma_{A_i}^2, \\ n_{Hi} &= \frac{k_i}{\sum_j 1/m_{ij}} \end{aligned}$$

is the harmonic mean of the cluster sizes, and

$$\frac{(k_i - 1)S_{Ui}^2}{\sigma_{E_i}^2 + n_{Hi}\sigma_{A_i}^2} \sim \chi_{k_i-1}^2.$$

The variance estimate of the estimated overall mean, $\bar{\bar{Y}}_i$, is then given by $S_{Ui}^2/(k_i n_{Hi})$ (for $i = 1, 2$).

1.4 Brief summary of methods for cluster randomization trials with continuous data

The primary purpose of conducting a cluster randomization trial is to make comparisons between arms. The first step of meeting such an objective is to obtain a useful summary statistic that answers the primary question of interest of the study. The primary question of interest informs whether the analysis will be performed at the cluster level or at the individual level. We start with cluster level analysis as the methods are simpler.

Cluster-level analyses

Although randomization in a cluster randomization trial occurs at the cluster level, investigators usually have a choice between cluster-level or individual-level analyses. Cluster-level analysis begins with creating cluster-specific summary statistics, followed by the application of standard statistical methods which are approximately valid. Each cluster level summary measure is then treated as a single observation, and clusters are independent.

Cluster-level analyses are appropriate when the primary question of interest is directed not to the individual, but to the randomized unit. For example, Lenhart *et al.* (2008) were interested in determining the effect of insecticide treated bednets on the transmission of vector-borne diseases and dengue by mosquitos in Haiti. The treated bednets kill the disease transmitting vectors when mosquitos come into contact

with them. In this study, 9 pairs of sectors containing a total of 1017 houses were randomized to either insecticide treated bednets or bednets which were not treated (usual care). The outcome of primary interest in this case was the number of trap containers with mosquitos with immature stages of disease per 100 households. A paired t-test was used to obtain inferences on this cluster-level outcome, where it was found that insecticide treated bednets had an immediate positive effect on vector diseases and dengue transmission.

As standard statistical methods may be used to analyze data at the cluster level, many valid and efficient procedures already exist. This thesis therefore focuses on analysis procedures at the individual level. Thus, unless otherwise stated, any mention of statistical analyses refers to individual level methods rather than those at the cluster level.

Individual level analyses

Although randomization occurs at the cluster level, inferences at the individual level are commonly of interest. For example, in a prenatal care study, Villar *et al.* (1998) evaluated a new antenatal care program by comparison to standard care through a cluster randomization trial, where the program's effects on birth weight was of primary interest. As another example, Christian *et al.* (2003) randomized 426 communities to five trial arms to determine the effect of micronutrient supplements on birth weight, where comparisons were made using the differences in average birth weights between each treatment arm and the control. To obtain inferences on this difference at the individual level, both between cluster and within cluster variance components must be accounted for.

Clusters in a cluster randomization trial may be assumed to be independent, however the individuals within each cluster may not. Two individuals in the same cluster are likely to be more similar than two individuals from different clusters (Hayes and Moulton, 2009, page 11). Therefore, there exist two levels of variation in a

clustered design: the variation within clusters (σ_E^2) and the variation between clusters (σ_A^2), where the total variation equals $\sigma_T^2 = \sigma_A^2 + \sigma_E^2$. Clustering occurs when the variation between clusters is non-zero. Therefore, these variance components may be used to quantify the similarity of individuals within the same cluster.

Two indices have been used to quantify the similarity of responses within clusters rather than between. The first is the intraclass correlation coefficient (ICC) (Donner *et al.*, 1981), interpreted as the standard Pearson correlation coefficient between two responses in the same cluster. It is given by the ratio of the between cluster variance component to the total variance ($\rho = \sigma_A^2/\sigma_T^2$). When the responses of individuals in the same cluster are no more similar than those of other clusters, the between cluster variation and consequently the ICC both equal zero. At the other extreme, when all of the responses of individuals in the same cluster are identical, the within cluster variance component equals zero, and the ICC equals one. The majority of the time, the ICC falls between these two extremes. In a review by Eldridge *et al.* (2004), the median ICC value for cluster sizes of around 30 was 0.04, with an interquartile range of -0.02 to 0.21. Larger clusters such as communities typically have smaller ICC values of 0.002 to 0.012 (Hannan *et al.*, 1994). Note that negative ICC values are possible, but improbable, typically occurring when individuals within the same cluster are less similar to one another than to individuals in other clusters (Hayes and Moulton, 2009, page 17). Due to their rare occurrence (Donner and Klar, 2000, page 11), negative ICC values will not be investigated further.

The second index used to account for dependent responses in a cluster is the coefficient of variation between clusters ($\text{CV}_A = \sigma_A/\mu$) (Hayes and Moulton, 2009, page 16). As the similarity in responses of two individuals in the same cluster with respect to other clusters increases, the difference in the responses between clusters will also increase. That is, the variance of the cluster-specific means will increase. The between cluster coefficient of variation measures the spread of the cluster-specific means as a proportion of the overall mean. Therefore, when the responses of two individu-

als in the same cluster are no more similar than the responses of two individuals in different clusters, then $\sigma_A^2 = 0$ and $\text{CV}_A = 0$. However, when the responses of two individuals in the same cluster are more similar than the responses of two individuals in different clusters, then $\sigma_A^2 > 0$ and $\text{CV}_A > 0$ (assuming that $\mu > 0$). Shoukri *et al.* (2008) and Quan and Shih (1996) give a description of the within-cluster coefficient of variation ($\text{CV}_E = \sigma_E/\mu$) as a measure of reproducibility. However, this measure (CV_E) is not as relevant to quantifying the effect of clustering as CV_A, because it gives no information about existing differences between clusters.

The common element in both the ICC and the between-cluster coefficient of variation is the between-cluster variance component, σ_A^2. These statistics may further be used to quantify the effect of clustering for analysis at the individual level.

The effect of clustering on data analysis may be quantified by the design effect (Donner and Klar, 2000, page 8). The design effect is interpreted as the ratio of the variance of the estimated effect measure for cluster sampling to that of random sampling, and is given as (Hayes and Moulton, 2009, page 21)

$$\begin{aligned} \text{deff} &= 1 + (m-1)\rho \qquad (1.5) \\ &= 1 + (m-1)\frac{\sigma_A^2}{\sigma_T^2} \\ &= 1 + (m-1)\frac{\sigma_A^2}{\sigma_T^2}\frac{\mu^2}{\mu^2} \\ &= 1 + (m-1)\text{CV}_A\frac{\mu^2}{\sigma_T^2}, \qquad (1.6) \end{aligned}$$

where m is the size of the clusters.

Once the effect of clustering is quantified, there are a variety of analysis procedures to choose from. The choice of analysis procedure depends on the question of interest and the properties of the data.

Current methods for a difference between two normal means

The responses from a cluster randomization trial may often be approximated with the normal distribution. For instance, body weight, height, body temperature, blood pressure, or summary scores from standardized questionnaires frequently follow normal distributions. For example, Kinra *et al.* (2008) randomized villages in India to investigate the effect of protein-calorie supplementation and public health programmes on cardiovascular risk. Outcome measures in this study included height and blood pressure, and were assumed to be normally distributed. As another example, Montgomery *et al.* (2000) randomized twenty-seven general practices to investigate the effect of a computer based clinical decision support system and a risk chart on blood pressure, an outcome which may be assumed to follow a normal distribution. This data is used in Chapter 5 as an illustration of the methods investigated in this thesis.

When sample means and their differences are assumed to follow normal distributions, confidence intervals or hypothesis tests for a difference between means or the equality of means may be used to draw conclusions from a cluster randomization trial. Because they encompass hypothesis testing, confidence intervals are preferred. This claim is detailed in Chapter 2.

A method proposed by Donner and Klar (1993) is to use the design effect (Equation (1.5)) to adjust the variance of the difference for clustering. The inflated variance is then plugged into the usual t-interval, setting degrees of freedom to the number of clusters minus two. This procedure uses a pooled estimate of the standard error, thereby assuming equal variances (homoscedasticity) in the two arms being compared. As in the case of individually randomized designs, variance homogeneity can be assumed in hypothesis testing when the null hypothesis is true, but not in the construction of confidence intervals (Wang and Chow, 2002). Donner and Klar (1993) pointed this out in the context of cluster randomization trials for the assumption of

a common design effect under the null hypothesis.

To avoid the homoscedasticity assumption of many statistical inference procedures, the variance in each arm of a cluster randomization trial may be estimated separately. The usual variance estimate which ignores clustering is biased when ICC > 0, underestimating the total variance (White and Thomas, 2005). Instead, the total variance of the overall mean in each arm of a clustered trial is given by (Donner and Klar, 2000, page 8)

$$\text{var}(\bar{Y}_i) = \frac{\sigma_A^2 + \sigma_E^2/m}{k},$$

where k is the number of clusters of size m. This cluster-adjusted variance may be estimated using the unweighted mean squared error (Thomas and Hultquist, 1978) for variable cluster sizes. El-Bassiouni and Abdelhafez (2000) showed that using the unweighted mean squared error to construct a confidence interval for a single normal mean from a one-way random effects model maintains the desired coverage, although tends to be wide when the ICC is less than 0.2. This confidence interval method may be applied here, because the observations in each arm of a cluster randomization trial may be assumed to follow a one-way random effects model.

Another approach to the above closed-form procedures is the generalized confidence interval method (Weerahandi, 1993). Generalized confidence intervals are based on the simulation of a known generalized pivotal quantity that possess the following two properties:

(i) its probability distribution is not dependent on any unknown parameters and

(ii) the observed value is not dependent on any unknown nuisance parameters.

Property (i) of the generalized pivotal quantity ensures that the confidence region may be found without the knowledge of unknown parameter values. This property is also found in the definition of the usual pivotal quantity - a quantity which is a

function of observations and unknown parameters, but whose distribution does not depend on any unknown parameters. Property (ii) of the generalized pivotal quantity, which is not found in the usual definition of pivotal quantities, further ensures that the confidence region may be obtained with only the use of the observed data.

To construct a generalized confidence interval, the generalized pivotal quantity is simulated a large number of times and the limits are set to the $\alpha/2$ and $1-\alpha/2$ quantiles. However, certain drawbacks of generalized confidence intervals arise as a consequence of property (ii): aside from having to know the generalized pivotal statistic in advance, generalized confidence intervals do not have a closed form, because the observed pivotal is based on simulation. Consequently, two generalized confidence intervals of the same confidence coefficient for the same dataset may differ. This may make a difference particularly in cases where there is a composite parameter of interest.

More complex methods of analysis include mixed regression models (Harville, 1977; Harville and Jeske, 1992) and generalized estimating equations with an exchangeable correlation matrix (Liang and Zeger, 1996). When there are a small number of clusters involved in a trial, there may be chance imbalance of covariates that are predictive of the outcome. These methods have the ability to adjust for unbalanced covariates in different trial arms. Unfortunately, the methods may be invalid when there are fewer than fifteen clusters per arm (Hayes and Moulton, 2009, Chapter 11), precisely when chance imbalance of covariates may occur, although improvements have been suggested (Skene and Kenward, 2010). Additionally, regression methods and generalized estimating equations as applied to cluster randomization trials sacrifice the desired simplicity of many of the confidence interval procedures discussed thus far. Covariate adjustment procedures will therefore be exempt from this thesis. Extensions for covariate adjustment may be considered in future works.

1.5 Alternative effect measures in cluster randomization trials

The standardized mean difference

Some outcomes are not as easily interpreted on the raw scale as outcomes such as height, weight, or blood pressure which have a natural comparison using a difference. For example, quality of life scores are a subjective assignment of values based on a combination of features. Other subjective ratings may also have this problem. In such cases, investigators often compare two groups using Cohen's effect size (Cohen, 1988), defined by the difference in the magnitudes of treatment effects in units of standard deviation. For instance, Jordhoy *et al.* (2001) randomized community health districts and used the effect size as the summary measure to assess the impact of comprehensive palliative care compared with conventional care on the quality of life of cancer patients. Also, Bernstein *et al.* (2005) randomized schools to three arms to compare effectiveness of school-based interventions for anxious children. The primary outcome was the change in composite clinician severity rating from baseline to post-treatment, where the effect size was used as the summary measure when comparing the arms of the trial. Also note that the effect size is often used to summarize results in a meta-analysis. For example, in the meta-analysis by Brunoni *et al.* (2009) the effect size was used to show that placebo responses of major depressive disorders are large regardless of the intervention of the study.

The choice of the effect size expression is more complicated in cluster randomization trials than in individually randomized trials. In an individually randomized trial, the effect size is set to the difference between means divided by the standard deviation. This standard deviation may be set to the pooled sample standard deviation (Hedges and Olkin, 1985), with an interpretation of the difference between means of two arms in terms of the extent to which individual responses vary. Alternatively, the denominator may be set to the control group standard deviation, leading to a measurement of the difference in means between the arms in terms of the extent to

which control group individuals vary amongst one another. However, in cluster randomization trials which have two levels of variation, there are even more possibilities for the denominator. Hedges (2007*b*) discusses three possibilities, each depending on the inference of interest to the researcher. These possibilities include setting the denominator to a function of the within cluster variance component, the between cluster variance component, or the total variance. Consequently, each would have a slightly different interpretation, potentially complicating the summary of results in meta-analyses which often include both cluster randomization trials and individually randomized trials.

Another issue lies with the direct interpretation of the effect size. Cohen (1988) defined "small", "medium", and "large" effect sizes as 0.2, 0.5, and 0.8, respectively. However, this interpretation is meant for application to the behavioral sciences, not necessarily as a generalization to all disciplines. Furthermore, the interpretation in units of standard deviations has more of a statistical meaning rather than a clinical meaning, potentially complicating results to a clinician. Also, although this measure has gained popularity in the behavioral and social sciences, there is no clear interpretation to the measure in a probabilistic sense.

Using the standard normal cumulative distribution function, it can be shown that the interpretation of the effect size is a function of the probability that the observation (Y_1) of a randomly selected individual from one arm of the trial is larger than the mean of another arm (μ_2). That is, assuming $\mu_2 > \mu_1$ and homoscedasticity for simplicity,

$$\begin{aligned} \mathrm{P}(Y_1 > \mu_2) &= \mathrm{P}(Y_1 - \mu_1 > \mu_2 - \mu_1) \\ &= \mathrm{P}\left(\frac{Y_1 - \mu_1}{\sqrt{\sigma^2}} > \frac{\mu_2 - \mu_1}{\sqrt{\sigma^2}}\right) \\ &= \mathrm{P}\left(Z > \frac{\mu_2 - \mu_1}{\sqrt{\sigma^2}}\right) \\ &= \Phi\left(\frac{\mu_1 - \mu_2}{\sqrt{\sigma^2}}\right), \end{aligned}$$

where σ^2 is the pooled variance. Note that if $\mu_1 > \mu_2$ then the effect measure would be $\Phi\left(\frac{\mu_2-\mu_1}{\sqrt{\sigma^2}}\right)$. However, there are two clear concerns with this interpretation. First, the true mean (μ_2) is rarely known in practice. Second, it is illogical to interpret study results by comparing a randomly selected individual in one arm to the mean of the observations from another arm. This comparison does not generally answer typical questions which motivate the conduct of cluster randomization trials.

It would be more logical to find the probability that a randomly selected individual in one arm has a larger outcome than a randomly selected individual in another arm, $\mathrm{P}(Y_1 > Y_2)$. For example, in a medical context comparing the responses of two treatments, where Y_2 represents the response under treatment 1 and Y_2 represents the response under treatment 2, $\mathrm{P}(Y_1 > Y_2) = 0.7$ makes more sense to a clinician and a patient than does $(\mu_1 - \mu_2)/\sigma = 1$. Let us refer to $\mathrm{P}(Y_1 > Y_2)$ as the exceedance probability.

In the past, the exceedance probability has been described as intuitive and simple (McGraw and Wong, 1992; Grissom, 1994; Kraemer *et al.*, 2003) and has shown wide application in the literature. For instance, its application may be found in reliability measurements (Church and Harris, 1970) and clinical equivalence trials (Hauck *et al.*, 2000). It is also closely related to the area under the receiver operator characteristic curve and to non-parametric statistics (Acion *et al.*, 2006). It is its application to cluster randomization trials which is of interest here.

Let $Y_1 \sim N(\mu_1, \sigma^2_{T_1})$ and $Y_2 \sim N(\mu_2, \sigma^2_{T_2})$ represent observations from two arms of a cluster randomization trial, where $\sigma^2_{T_i} = \sigma^2_{A_i} + \sigma^2_{E_i}$. The exceedance probability may

then be manipulated to uncover the effect measure which is of applicable interest,

$$\begin{aligned} \mathrm{P}(Y_1 > Y_2) &= \mathrm{P}(Y_1 - Y_2 > 0) \\ &= \mathrm{P}(Y_1 - Y_2 - (\mu_1 - \mu_2) > -(\mu_1 - \mu_2)) \\ &= \mathrm{P}\left(\frac{Y_1 - Y_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma^2_{T_1} + \sigma^2_{T_2}}} > \frac{-(\mu_1 - \mu_2)}{\sqrt{\sigma^2_{T_1} + \sigma^2_{T_2}}} \right) \\ &= \mathrm{P}\left(Z > \frac{-(\mu_1 - \mu_2)}{\sqrt{\sigma^2_{T_1} + \sigma^2_{T_2}}} \right) \\ &= \Phi\left(\frac{(\mu_1 - \mu_2)}{\sqrt{\sigma^2_{T_1} + \sigma^2_{T_2}}} \right). \end{aligned} \tag{1.7}$$

The effect measure of interest is thus a monotone function of the standardized mean difference,

$$\delta = \frac{\mu_1 - \mu_2}{\sqrt{\sigma^2_{T_1} + \sigma^2_{T_2}}}. \tag{1.8}$$

Although the term 'standardized mean difference' is often used interchangeably with 'Cohen's effect size', here it will refer to δ. Confidence intervals for $\mathrm{P}(Y_1 > Y_2)$ may then be obtained from confidence intervals for δ using the transformation principle (see Section 2.3.2).

In addition to having a logical interpretation, this measure is capable of capturing all of the effects of the treatment as compared to the control. A difference in means is clearly useful when comparing overall results. However, the variability in each arm of the cluster randomization trial reflects the consistency in response. Since δ is a function of the means and variances, it takes into account both the magnitude and consistency of responses.

Once the exceedance probability is estimated, it is important to obtain inferences on this measure. Obtaining inferences on δ is more complicated in cluster randomization trials than in individually randomized trials. In an individually randomized trial,

estimates of the standardized mean difference follows the non-central t-distribution with non-centrality parameter $\mu_1 - \mu_2$ (Owen *et al.*, 1964). However, as a result of the two levels of variation present in clustered designs, this relationship does not hold (Thomas and Hultquist, 1978). The estimates of δ and $\mathrm{P}(Y_1 > Y_2)$ do not follow exact distributions, and therefore exact confidence intervals may not be obtained.

Confidence interval procedures for δ are scarce because the parameter is a function of the normal mean (μ), the between-cluster variance component ($\sigma^2_{A_i}$) and the within-cluster variance component ($\sigma^2_{E_i}$). One option is to apply the multivariate delta method to obtain an expression for the variance of the estimate of δ. Slutsky's theorem (Casella and Berger, 2002, page 239-240) states that

If $X_n \to X$ in distribution and $Y_n \to a$, a constant, in probability, then

i) $Y_n X_n \to aX$ in distribution and

ii) $X_n + Y_n \to X + a$ in distribution.

Using Slutsky's theorem, the estimated variance at the point estimate may then be plugged into Wald type confidence intervals which are constructed by inverting the Wald test (Wald, 1941). That is, the limits are set to the point estimate plus or minus some multiple of this estimated standard error. By using this plug-in principle, the estimated variance used for the upper limit is forced to equal the estimated variance used when constructing the lower limit, thus yielding symmetric limits around the point estimate. In fact, any Wald-type confidence interval procedure which utilizes this plug-in principle will have this restriction. However, as in individually randomized trials, the estimate $\hat{\delta}$ may have a skewed sampling distribution in cluster randomization trials. Therefore, one limit of a symmetric confidence interval may fail to exclude all extreme values while the other limit may exclude too many values. Confidence interval procedures which take into account the shape of the distribution are therefore desirable.

The Fieller method (Fieller, 1954) is not restricted to symmetry and may be used to construct confidence intervals for a ratio, and thus for δ. However, this method

assumes that the sampling distribution of both the numerator and the denominator are normal. Although the estimated numerator of the standardized mean difference may be approximately normal, the denominator is a function of estimated variance components. The unweighted mean squared error (Equation (1.4)) may be used to approximate the denominator, and Thomas and Hultquist (1978) show that this statistic is approximately proportional to the chi-squared distribution. Thus, the denominator of the standardized mean difference is typically skewed in distribution. Therefore, the Fieller method would also fail to capture the underlying distribution of the effect measure.

Alternatively, generalized confidence intervals may be constructed for the standardized mean difference, and thus for $\mathrm{P}(Y_1 > Y_2)$, when the data in each arm follow a one-way random effects model. Recall from Section 1.4.3 that this method requires the existence of a generalized pivotal statistic. Using the results of Thomas and Hultquist (1978), Krishnamoorthy *et al.* (2007) provide generalized pivotal statistics for the normal mean, the between-group variance component, and the within-group variance component when data follow a one-way random effects model,

$$G_\mu = \bar{\bar{Y}} + \frac{Z}{\sqrt{\chi^2_{k-1}}}\sqrt{\frac{\mathrm{SSC}}{k}}, \tag{1.9}$$

$$G_{\sigma^2_A} = \left(\frac{\mathrm{SSC}}{\chi^2_{k-1}} - \frac{\tilde{n}\mathrm{SSW}}{\chi^2_{M-k}}\right)_+, \tag{1.10}$$

$$G_{\sigma^2_E} = \frac{\mathrm{SSW}}{\chi^2_{M-k}}, \tag{1.11}$$

where $Z \sim N(0,1)$, χ^2_{df} is a random variable from the chi-squared distribution with *df* degrees of freedom, SSC is the sum of squares between groups, SSW is the sum of squares within groups, $(x)_+ = \max\{0, x\}_+$, and $\tilde{n} = (1/k)\sum_j^k m_j$. Thus, Equations (1.9) to (1.11) may be obtained by plugging in both summary statistics of the data as well as simulated standard normal and chi-squared random variables. These three generalized pivotal statistics may then be used to obtain a generalized pivotal statistic for the standardized mean difference from a cluster randomization trial. This is

because an attractive feature of the generalized confidence interval procedure is that the generalized pivotal statistic of a function of parameters is simply that function applied to the generalized pivotal statistics of those parameters. Thus, as long as the generalized pivotal quantity of the components of a particular parameter of interest are known, the generalized pivotal quantity of the parameter may be found.

As the standardized mean difference is given by Equation (1.8), the generalized pivotal quantity of this statistic may be expressed as

$$G_\delta = \frac{G_{\mu 1} - G_{\mu 2}}{\sqrt{G_{\sigma^2_{A_1}} + G_{\sigma^2_{E_1}} + G_{\sigma^2_{A_2}} + G_{\sigma^2_{E_2}}}}.$$

To obtain generalized confidence intervals (L_δ, U_δ), an algorithm similar to that found in Krishnamoorthy *et al.* (2007) may be used with G_δ. Generalized confidence intervals for $\mathrm{P}(Y_1 > Y_2)$ are then given by $(\Phi(L_\delta), \Phi(U_\delta))$.

Generalized confidence intervals have not previously been applied to the standardized mean difference or to $\mathrm{P}(Y_1 > Y_2)$ when data arise from a cluster randomization trial. Theoretically, their application would not result in symmetric limits around a parameter with a skewed sampling distribution like those of the Wald interval. However, recall that their application would not result in closed form intervals. A closed form procedure is preferable.

The difference between two lognormal means

In addition to the normal distribution, data found in practice may commonly be approximated by the lognormal distribution. For instance, when a variable is bounded with a concentration of data falling near the lower bound, then the data may be skewed, potentially following a lognormal distribution. Also, according to the multiplicative central limit theorem, the multiple of a large number of independent and positive random variables, each with a finite mean and variance, will approximately follow a lognormal distribution (Limpert *et al.*, 2001). Examples of outcomes which

commonly follow lognormal distributions in practice include hospital wait times and cost data (Thompson and Barber, 2000). Panella *et al.* (2007) randomized 14 hospitals to either treatment according to a clinical pathway or usual care, where non-normal outcomes of interest included length of hospital stay and overall cost. Also, Marrie *et al.* (2000) investigated the efficiency of a critical pathway, randomizing 19 hospitals to either continued conventional management or to implement the critical pathway. One of the outcomes was the length of hospital stay which may be approximated by the lognormal distribution, as we shall see in Chapter 5 where we used this data as an illustration for the methods investigated in this thesis. Similar to outcomes which may follow a normal distribution, two arms of a cluster randomization trial may naturally be compared using a difference of means when outcomes are approximately lognormal.

Obtaining inferences on a single lognormal mean has been a challenge when data follow a one-way random effects model on the log scale (Briggs *et al.*, 2005). This is because the lognormal mean is a function of three parameters: the normal mean (μ), the between-group (or between-cluster in an arm of a cluster randomization trial) variance component (σ_A^2), and the within-group (or within-cluster) variance component (σ_E^2). Existing inference procedures are limited in scope, requiring not only the assumption of homoscedasticity, but symmetric confidence intervals when the lognormal distribution is skewed in shape. An example of this arises when the multivariate delta method is used to estimate the variance of the lognormal mean at the point estimate, which is then plugged into the Wald-type intervals using Slutsky's theorem. Again, the shape of the confidence interval should reflect the shape of the distribution, thereby ensuring that only the extreme values are excluded.

Two types of procedures which are not restricted to symmetry are bootstrap procedures and generalized confidence intervals. Bootstrap confidence intervals have become popular since they were first introduced by Efron (1979), and their use is cautiously encouraged when valid parametric procedures are not available (DiCiccio

and Efron, 1996). However, the performance of the bootstrap has not been ideal for the lognormal mean. Recall that the lognormal mean is a function of the normal mean and variance. DiCiccio and Efron (1996) and Schenker (1985) have shown that the percentile method and the bias-corrected bootstrap procedures for the normal variance have low coverage when sample sizes are small to moderate. These results seem to be consistent throughout the literature. More recently, Dinh and Zhou (2006) and Zou and Donner (2008) have shown that bootstrap confidence interval procedures have low coverage for the lognormal mean when data on the log scale follow a fixed effects model. Also, Flynn and Peters (2004) used Monte Carlo simulations to show that the bias-corrected and accelerated bootstrap procedure lead to confidence intervals with low coverage for a difference between mean costs in cluster randomization trials, where cost data are normally and lognormally distributed.

The poor performance of the bootstrap in these cases stems from the fact that the bootstrap procedure requires the existence of a transformation which can make the sampling distribution of the statistic of interest both normal and pivotal (Schenker, 1985). Such a transformation does not seem to exist for the normal variance. For instance, the log transformation makes it pivotal but not normal, and the cubic root transformation makes it normal but not pivotal (Kendall and Stuart, 1977, page 371).

Alternatively, generalized confidence intervals may be constructed for a difference between two lognormal means when the data are assumed to follow a one-way random effects model on the log scale. The required pivotal statistic may be constructed using Equations (1.9), (1.10), and (1.11). This statistic is given by

$$G_{\text{LN}} = \exp\left(G_{\mu_1} + \frac{G_{\sigma^2_{A_1}} + G_{\sigma^2_{E_1}}}{2}\right) - \exp\left(G_{\mu_2} + \frac{G_{\sigma^2_{A_2}} + G_{\sigma^2_{E_2}}}{2}\right).$$

Similar steps to the generalized confidence interval algorithm for δ (see Section 1.5.1) may be followed to construct confidence intervals for a difference between two lognormal means. However, as previously discussed, this procedure is not closed form. A closed form procedure would be preferred.

1.6 Scope of the thesis

Cluster randomized trials have been growing in popularity over the last three decades. For example, in the areas of public health and medicine "the number of trials reporting a cluster design has risen exponentially since 1997" (Campbell, 2004). As a result, there have been many advancements in the design and analysis of such studies (Cornfield, 1978; Donner and Klar, 2000; Hayes and Moulton, 2009). The CONSORT statement, a guideline for the reporting of clincal trials, has been extended to improve the reporting of cluster randomization trials (Campbell *et al.*, 2004). This in turn has resulted in some improvement in the design, analysis, and reporting of cluster randomization trials (Bland, 2004; Varnell *et al.*, 2004).

The primary goal of this thesis is to develop statistical methods for quantifying effects in cluster randomization trials with continuous outcomes. Data from cluster randomization trials can often be approximated by the normal distribution on the raw or log scale. Therefore, I focus on normally and lognormally distributed outcomes.

The purpose of conducting a cluster randomization trial is to make comparisons between the results of each arm. A typical question of interest is whether one treatment is better than another. A natural comparison in this case is a difference between the means of the two arms. Alternatively, it may be of interest to find the exceedance probability. Effect measures which directly answer these questions are

$$\Delta = \mu_1 - \mu_2 \tag{1.12}$$

$$\Delta_{LN} = \exp\left(\mu_1 + \frac{\sigma^2_{A_1} + \sigma^2_{E_1}}{2}\right) - \exp\left(\mu_2 + \frac{\sigma^2_{A_2} + \sigma^2_{E_2}}{2}\right) \tag{1.13}$$

$$\mathrm{P}(Y_1 > Y_2) = \Phi\left(\frac{\mu_1 - \mu_2}{\sqrt{\sigma^2_{A_1} + \sigma^2_{E_1} + \sigma^2_{A_2} + \sigma^2_{E_2}}}\right), \tag{1.14}$$

for a difference between two normal means, a difference between two lognormal means, and the exceedance probability for normal outcomes, respectively.

Inference procedures such as hypothesis tests and confidence intervals which put

effect measures of interest into context are important. A rejection of the null hypothesis of equal means provides evidence that the means are in fact different without any information about the magnitude of that difference. Confidence intervals however do provide information about the magnitude. Furthermore, a single confidence interval is equivalent to conducting an infinite number of hypothesis tests (see Chapter 2 for a more thorough comparison of confidence intervals and hypothesis tests). Therefore, this thesis focuses on confidence interval procedures for the three effect measures described above (Equations 1.12, 1.13, and 1.14).

Closed form confidence interval procedures are developed using the method of variance estimates recovery (MOVER) (Zou, 2008). Confidence interval construction for the first two effect measures (Equations (1.12) and (1.13)) begin by defining a set of parameters for which valid confidence interval procedures exist and whose linear combination equals a difference in means. The central limit theorem is invoked to recover the variance estimates from the confidence intervals of each individual parameter. These variance estimates are then used to construct intervals for the linear combination (Zou, 2008).

Confidence intervals for the third effect measure (Equation (1.14)) are developed by first focusing on the standardized mean difference, δ. The standard normal cumulative distribution function (Φ) is a monotone increasing function. Therefore, confidence intervals for $\mathrm{P}(Y_1 > Y_2) = \Phi(\delta)$ may easily be constructed by applying the transformation principal to the intervals of the standardized mean difference. The ratio is first re-parameterized into a linear combination of parameters, then quadratic equations are solved for the upper and lower limits of the standardized mean difference, (L_δ, U_δ) (Zou and Donner, 2010). Finally, confidence limits of $\mathrm{P}(Y_1 > Y_2)$ are given by $(\Phi(L_\delta), \Phi(U_\delta))$.

These new intervals are expected to improve on existing confidence interval procedures by relaxing the assumption of homoscedasticity and avoiding forced symmetry. In fact, Efron and Tibshirani (1993, page 187) stressed that forced symmetry is the

most serious error in confidence interval construction. Homoscedasticity is avoided by estimating the variance separately in each arm, and the proposed confidence intervals are not forced to be symmetric because variances are estimated in the vicinity of the lower limit and upper limit rather than at the point estimate. These improvements are detailed further in Chapter 2, where a review of the development of the MOVER is given, as is a new proof for the application of the MOVER to a linear combination of n parameters using the method of induction.

For simplicity, attention is focused on the completely randomized design. Extensions to the pair-matched and the stratified designs are discussed in the final chapter of the thesis.

1.7 Objectives

The objectives of the thesis are to propose new confidence intervals for the the difference between two normal means, the difference between two lognormal means, and the exceedance probability for normal outcomes, such that these intervals are

1) valid for completely randomized cluster randomization trials with a small number of large clusters,
2) not symmetric for parameter estimates with a skewed sampling distribution,
3) statistically efficient, and
4) simple to apply.

These proposed confidence interval procedures are analytically justified, evaluated, and compared with existing procedures throughout the thesis.

Solutions to these objectives will provide useful confidence interval procedures which may be used to make valid and efficient inferences on data of future trials.

1.8 Organization of the thesis

The thesis consists of six chapters. Chapter 2 includes a detailed description of confidence intervals and their properties, including the principles of confidence interval construction for a single parameter. The general justification of the MOVER is also described.

Chapter 3 outlines the proposed and existing confidence interval procedures with necessary proofs, while Chapter 4 presents a simulation study. The performance of the proposed asymptotic confidence intervals are evaluated and compared to those of existing methods.

The applicability of the three proposed intervals is demonstrated in Chapter 5 using data from two studies. The first study (Montgomery *et al.*, 2000) deals with approximately normal data used to evaluate the effect of a computer based clinical decision support system and risk chart on blood pressure by randomly assigning 27 practices to three intervention groups. The second study (Marrie *et al.*, 2000) deals with data evaluating the use of a critical pathway against conventional management for the treatment of community acquired pneumonia from a randomization of 19 hospitals. The outcome of interest was the length of hospital stay which may be approximated by the lognormal distribution.

General conclusions, limitations of the proposed procedures, with a review of important assumptions, and a discussion of potential future research are given in Chapter 6.

Chapter 2

FUNDAMENTALS OF CONFIDENCE INTERVAL ESTIMATION

2.1 Introduction

Fisher (1925) was one of the first to suggest the use of p-values in judging the significance of a study, where significance testing is a major tool in statistical inference. Significance testing starts with a null hypothesis which itself may or may not be of direct interest. The test statistic then reports the strength of evidence against the null hypothesis using a p-value (Altman, 2005). Alternatively, a hypothesis test (Neyman and Pearson, 1933) is a decision making device, where both a null and an alternative hypothesis are declared. The p-value is not reported, but only the significance level and whether or not one rejects or fails to reject the null. Unfortunately, both significance testing and hypothesis testing may easily be abused. It is the p-value in particular which is often the source of misinterpretation.

Walter (1995) highlights four key points related to the misinterpretation of p-values. First, there may exist some confusion between Fisher's significance test and Neyman-Pearson's hypothesis test. Significance testing refers to a single null value which may not be of direct interest and is intended to measure the strength of evidence suggested by the data. However, the p-value is only based on the null value and does not directly consider other values. Also, the p-value reflects the tail area of the sample space which may include values which are not of interest to the investigator (Walter, 1995). A hypothesis test is a decision making tool, reducing a problem to a yes or no question. This type of test fails to distinguish between p-values of say 0.049 and

< 0.001, both of which would lead to the same conclusion - reject the null.

Second, the p-value is commonly misinterpreted as the probability that the observed effect is an error, thereby ignoring the initial assumption that the null is true. The correct definition of the p-value is the probability of estimating something at least as extreme as the observed had the null hypothesis been true. Fisher suggested that if this probability is very small then there is evidence to reject the null hypothesis or no effect.

Third, the p-value ignores clinical significance by failing to communicate the effect size. This takes the focus away from estimation in its interpretation, despite the fact that the primary goal of a study is to draw conclusions about the magnitude of the effect.

Finally, the p-value combines the magnitude of the effect size and the sample size, potentially leading to misinterpretation. A low p-value may be a result of either a large sample size or a large effect. A large p-value may be a result of a small sample size of no effect.

On the other hand, confidence intervals are useful when interpreting results because they keep the focus on estimation (Neyman, 1937) while providing information about the precision of the point estimate. They are defined as a range of values where the true parameter value is likely to lie according to the sample data. The precision of the estimate is then demonstrated by the width of the interval, with a narrower interval indicating greater precision. This is useful in application when, for example, a wide interval indicates a need for further study.

By focusing on estimation, confidence intervals may be used to judge clinical significance for a range of potential parameter values rather than a single point. Confidence interval construction involves not only the point estimate, but also the limits of the interval, thereby bringing attention to other potential values of the parameter which may be either more or less clinically significant. Therefore, confidence intervals are especially useful when dealing with statistically non-significant results, where the

interval may contain potentially clinically significant values. This is particularly relevant to cluster randomization trials which tend to have lower power than individually randomized trials of the same size due to the similarity of individuals within the same cluster.

In addition to the portrayal of precision and clinical significance, confidence intervals also encompass hypothesis tests thereby providing information on statistical significance. A single interval is equivalent to conducting an infinite number of hypothesis tests, providing information for every possible parameter value. The values within the interval include all the values which would not be rejected had hypothesis tests been conducted at one minus the confidence level, and the values outside the interval would be found statistically significant.

As a result of the advantages of confidence interval construction over hypothesis testing, guidelines for the reporting of randomized trials (such as the CONSORT Statement (Schulz *et al.*, 2010)) recommend the use of confidence intervals in the interpretation of study results.

Unfortunately, investigators tend to have an attachment to the p-value to the point that much of the information communicated with confidence intervals are ignored and confidence intervals are often used to resort back to hypothesis testing. For example, when interest lies in the comparison of sample means, the overlap of the intervals of each mean is used to judge statistical significance (e.g. Djordjevic *et al.*, 2000; Mancuso *et al.*, 2001; Tersmette *et al.*, 2001), despite warnings that overlap does not necessarily mean non-significance (Schenker and Gentleman, 2001; Wolfe and Hanley, 2002; Wilcox, 2003, page 246). Thus, in addition to once again shifting the focus away from estimation, the conclusions reached using this method of hypothesis testing may be fallacious.

The application of the overlap method may also be due to a lack of valid hypothesis tests or confidence intervals for the difference of two parameters. This method is equivalent to naively setting the confidence limits for the difference $\theta_2 - \theta_1$ to $(l_2 -$

$u_1, u_2 - l_1)$, where (l_i, u_i) is the $(1-\alpha)\%$ confidence interval for θ_i. This overestimates the variance, leading to conservative intervals (Schenker and Gentleman, 2001).

2.2 Definition of a confidence interval

There is an inherent connection between hypothesis testing and confidence interval estimation. In fact, confidence limits can be defined based on the Neyman-Pearson hypothesis testing principle (Neyman, 1935, 1937). Let $\theta = \theta_0$ represent the true value of the unknown population parameter. A confidence interval for θ with confidence coefficient α may be given by the random variables (L_θ, U_θ), such that

$$\mathrm{P}(L_\theta \leq \theta \leq U_\theta | \theta = \theta_0) = 1 - \alpha.$$

That is, given the true parameter value, the probability that this fixed value lies between the two random variables L_θ and U_θ is $1 - \alpha$. The probability expression is true for any θ_0, therefore it may be expressed as

$$\mathrm{P}(L_\theta \leq \theta \leq U_\theta) = 1 - \alpha. \tag{2.1}$$

The limits L_θ and U_θ are statistics based upon data. Under repeated random sampling, the estimated $(1-\alpha)100\%$ confidence interval (L_θ, U_θ) will contain the true parameter value $(1 - \alpha)100\%$ of the time. More specifically, if we are interested in equal tail probabilities,

$$\mathrm{P}(\theta < L_\theta) = \alpha/2 \tag{2.2}$$

$$\mathrm{P}(\theta > U_\theta) = \alpha/2, \tag{2.3}$$

so that with repeated random sampling one would expect to exclude the true value from either side of the interval $(\alpha/2)100\%$ of the time.

2.3 Confidence interval estimation for a single parameter

Two general principles have usually been applied when constructing confidence intervals: i) the confidence interval inversion principle, and ii) the confidence interval transformation principle. These two principles are usually applied in conjunction with the delta method.

The inversion principle

The confidence interval inversion principle allows the inversion of test statistics to obtain confidence intervals. Cox and Hinkley (1974) pointed out that "to obtain 'good' $1-\alpha$ upper confidence limits, take all those parameter values not 'rejected' at level α in a 'good' significance test against lower alternatives". According to Casella and Berger (2002, p. 421-422), the inversion principle is as follows: let $A(\theta_0)$ represent the acceptance region of a test of $H_0 : \theta = \theta_0$ at the α level, where $\theta_0 \in \Theta$ and Θ is the parameter space. Then $C(x) = \{\theta_0 : x \in A(\theta_0)\}$ is the random $1-\alpha$ confidence set in the parameter space for each $x \in X$. Conversely, let $C(X)$ represent a $1-\alpha$ confidence set. Then $A(\theta_0) = \{x : \theta_0 \in C(x)\}$ is the acceptance region of the test $H_0 : \theta = \theta_0$ at the α level for any $\theta_0 \in \Theta$. Note that a confidence set is a set of values contained in the confidence interval.

In summary, this means that once a hypothesis test fixes the parameter value at $\theta = \theta_0$, the sample values not rejected by the test at some alpha level make up the confidence interval. On the flip side, once a confidence interval is obtained by fixing a sample value, all of the potential parameter values within that interval would not be rejected by the corresponding hypothesis test, while values outside the interval are statistically significant.

The transformation principle

The confidence interval transformation principle allows confidence intervals for a single parameter to be used to construct confidence intervals for any monotonic transformation of that parameter (Steiger, 2004). Daly (1998) refers to this principle as the 'substitution method'. Let $f(\theta)$ be a monotonic transformation on the single parameter θ, where θ has the $(1-\alpha)100\%$ confidence interval (l, u). If the function is increasing then the confidence interval of $f(\theta)$ is $(f(l), f(u))$. However, if the function is decreasing, then the confidence interval of $f(\theta)$ is $(f(u), f(l))$. As an example, this principle is used in generalized linear models such as logistic regression where confidence intervals for odds ratios are first obtained on the log scale, then transformed back by exponentiating the limits.

Also note that a limitation of this method is that the new measure must be a function of a single parameter rather than of multiple parameters, otherwise the transformation principle will fail (Daly, 1998).

2.4 Wald-type confidence intervals and the delta method

Wald-type confidence intervals are often constructed for a parameter of interest using the delta method and Slutsky's theorem (Casella and Berger, 2002, page 239-240). The Wald test (Wald, 1941) may be inverted (Section 2.3.1) to obtain a two-sided confidence interval expression consisting of the point estimate ($\hat{\theta}$), the variance ($\text{var}(\hat{\theta})$), and some quantile of the standard normal distribution ($z_{\alpha/2}$). That is,

$$\begin{cases} L = \hat{\theta} - z_{\alpha/2}\sqrt{\text{var}(\hat{\theta})} \\ U = \hat{\theta} + z_{\alpha/2}\sqrt{\text{var}(\hat{\theta})}. \end{cases}$$

The delta method may be used to obtain an expression for the variance of $\hat{\theta}$. Slutsky's theorem then allows an estimate of the variance to be plugged into the Wald limits if

the variance expression is a function of parameters. That is,

$$\begin{cases} L = \hat{\theta} - z_{\alpha/2}\sqrt{\widehat{\text{var}}(\hat{\theta})} \\ U = \hat{\theta} + z_{\alpha/2}\sqrt{\widehat{\text{var}}(\hat{\theta})}. \end{cases} \tag{2.4}$$

A single parameter

Let the parameter of interest $g(\theta)$ be a function of a single parameter θ. The delta method uses a first-order Taylor approximation to solve for the variance of the function of a random variable, $\text{var}[g(\hat{\theta})]$,

$$\begin{aligned} g(\hat{\theta}) &\doteq g(\theta) + (\hat{\theta} - \theta)g'(\theta) \\ g(\hat{\theta}) - g(\theta) &\doteq (\hat{\theta} - \theta)g'(\theta) \\ \text{E}\{[g(\hat{\theta}) - g(\theta)]^2\} &\doteq \text{E}[(\hat{\theta} - \theta)]^2[g'(\theta)]^2 \\ \text{var}[g(\hat{\theta})] &\doteq \text{var}(\hat{\theta})[g'(\theta)]^2. \end{aligned} \tag{2.5}$$

This variance approximation is satisfactory only if there is a high probability that the random variable $\hat{\theta}$ is close to θ. Once Slutsky's theorem is applied to plug in the point estimate, the corresponding Wald limits for $g(\theta)$ are

$$\begin{cases} L = g(\hat{\theta}) - z_{\alpha/2}\sqrt{\widehat{\text{var}}(\hat{\theta})(g'(\hat{\theta}))^2} \\ U = g(\hat{\theta}) + z_{\alpha/2}\sqrt{\widehat{\text{var}}(\hat{\theta})(g'(\hat{\theta}))^2}. \end{cases} \tag{2.6}$$

A function of multiple parameters

Let the parameter of interest $g(\theta_1, ..., \theta_n)$ be a function of multiple parameters, $\theta_1, ..., \theta_n$, estimated by $\hat{\theta}_1, ..., \hat{\theta}_n$. Similar to the univariate case, the multivariate delta method may be applied to find an expression for the variance of $g(\hat{\theta}_1, ..., \hat{\theta}_n)$:

$$\text{var}[g(\hat{\theta}_1, ..., \hat{\theta}_n)] = [\mathbf{g'}(\theta_\mathbf{1}, ..., \theta_\mathbf{n})]^T \mathbf{cov}(\hat{\theta}_\mathbf{i}, \hat{\theta}_\mathbf{j})\mathbf{g'}(\theta_\mathbf{1}, ..., \theta_\mathbf{n}), \tag{2.7}$$

where $\mathbf{cov}(\hat{\theta}_\mathbf{i}, \hat{\theta}_\mathbf{j})$ is an n by n variance-covariance matrix, $\mathbf{g'}(\theta_\mathbf{1}, ..., \theta_\mathbf{n})$ is a vector of partial derivatives, and $[\mathbf{M}]^T$ is the transpose of the matrix $\mathbf{M}$. This variance may

then be applied with Wald confidence intervals and Slutsky's theorem to estimate the limits of $g(\theta_1, ..., \theta_n)$,

$$\begin{cases} L = g(\hat{\theta}_1, ..., \hat{\theta}_n) - z_{\alpha/2}\sqrt{\widehat{\text{var}}[g(\hat{\theta}_1, ..., \hat{\theta}_n)]} \\ U = g(\hat{\theta}_1, ..., \hat{\theta}_n) + z_{\alpha/2}\sqrt{\widehat{\text{var}}[g(\hat{\theta}_1, ..., \hat{\theta}_n)]}. \end{cases} \tag{2.8}$$

Properties of Wald-type confidence intervals

The wide application of Wald-type confidence intervals with the delta method is largely due to its simplicity. However, this advantage comes with a price, especially when using the conventional delta method to estimate variances.

The first assumption that the Wald procedure makes is that the sampling distribution of the parameter estimate is approximately normal,

$$\frac{\widehat{\theta} - \theta}{\sqrt{\text{var}(\widehat{\theta})}} \sim N(0, 1).$$

Due to the central limit theorem, the sum of a large number of independent random variables, which each have a finite mean and variance, is approximately normal (Casella and Berger, 2002, page 236). In such cases, Wald-type confidence interval procedures are applicable.

The second assumption of the Wald procedure is made when using the delta method for estimating the variance of a function of a parameter estimate. If the function of the parameter ($g(.)$) is non-linear and there is much variation in the data, the estimated variance of the parameter estimate may not be satisfactory. The delta method uses the first-order Taylor series approximation when estimating the variance (see Equation (2.6)), and as a result, it makes an assumption of a linear transformation over the expected range of the parameter. However, when looking at the function over a narrow enough region, the function may appear somewhat linear. That is, when most of the data fall near the parameter value and the variance is not considerably large, use of the first-order Taylor series approximation may result in a satisfactory

variance expression. To reiterate, when interest lies in the non-linear transformation of a parameter, the widely used delta method with Wald-type confidence intervals should not be applied if the variance is large.

The third assumption of Wald-type confidence intervals when the delta method is used is that the variance estimate is independent of the parameter estimate. Thus, the variance estimates at the limits would be the same as the variance estimate at the point estimate. However, only for the normal distribution is the sample mean independent of the variance (Lukacs, 1942). When the sampling distribution of the parameter estimate is non-normal, this assumption is usually violated. A simple example is the case of a proportion (Anderson, 2009). As a result, Wald-type confidence intervals which make use of the delta method may have poor coverage in practice, because the plugged-in point estimate may not be close to its value at the limits. This is a consequence of fixing the estimated variance when the point estimate is plugged in, potentially posing problems when the sampling distribution of the estimated parameter of interest is skewed.

As another example consider the variance, σ^2, estimated by $s^2 = (n-1)^{-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$, where $X_i \sim N(\mu, \sigma^2)$, $i = 1, ..., n$, are independent and identically distributed. The delta method gives a variance of $2\sigma^4/n$, corresponding to Wald limits of

$$(L, U) = s^2 \pm z_{\alpha/2}\sqrt{2s^4/n}.$$

where $z_{\alpha/2}$ is the upper $(\alpha/2)100\%$ quantile of the standard normal distribution. Alternatively, exact confidence intervals may be constructed for σ^2 because

$$(n-1)s^2/\sigma^2 \sim \chi^2_{n-1},$$

giving the limits

$$\begin{aligned} L &= \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}} \\ U &= \frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} \end{aligned}$$

where $\chi^2_{q,n-1}$ is the q^{th} quantile of the chi-squared distribution with $n-1$ degrees of freedom.

To see the difference between the two procedures, we can use a confidence interval function, which is a graph of confidence intervals at every confidence level for a particular dataset (Poole, 1987). In Figure 2.1 a confidence interval function of the delta method and the exact procedure are superimposed, with the solid vertical line indicating the point estimate. Although the sample variance is skewed, notice how the delta method results in symmetric Wald intervals around the point estimate. Comparison with the exact interval shows the potential of the delta method to result in lopsided tail errors - displayed by the difference between $\|l_d - l_e\|$ and $\|u_d - u_e\|$ at any confidence level $1-\alpha$ ($\alpha \neq 0, 1$), where (l_d, u_d) and (l_e, u_e) are the limits of the Wald interval with the delta method and the exact interval, respectively. Also notice that as the confidence level increases within a practical range, the discrepancy between the delta method and the exact method increase. This is because at low confidence, the intervals are narrower thereby more closely satisfying the assumption of the delta method that the point estimate of the variance is close to the limits.

2.5 Confidence intervals for a function of multiple parameters

Many of the confidence intervals for the effect measures of interest in this thesis (which are functions of other parameters) will be derived using the method of variance estimates recovery (MOVER) and the transformation principle (Section 2.3.2). In fact, the MOVER received its name due to its key step - the recovery of variance estimates for the estimated parameter of interest using the individual confidence intervals of the component parameters (Zou, 2008).

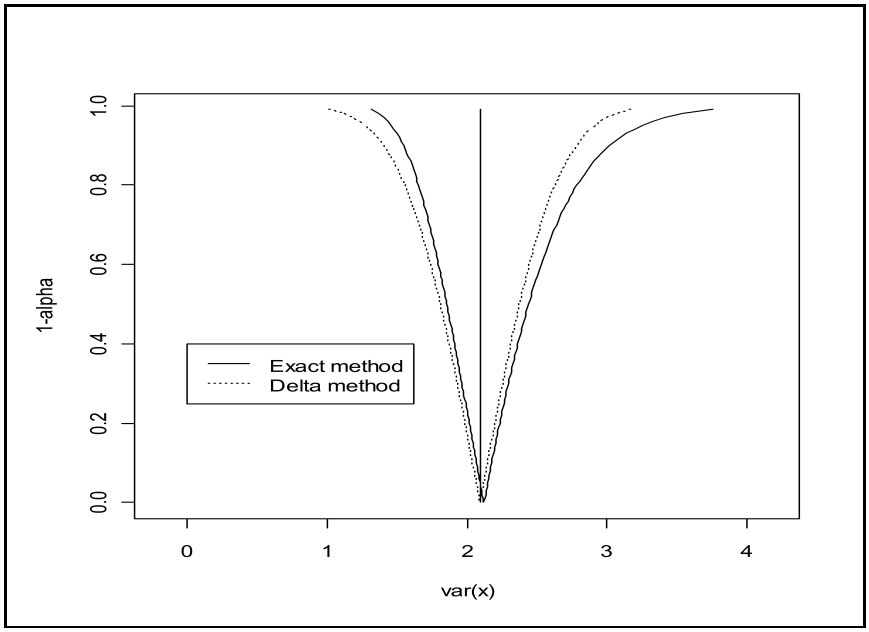

Figure 2.1: A confidence interval function of the delta method and the exact procedure for the normal variance ($\sigma^2 = 2$, $s^2 = 2.1$, $n = 100$)

The MOVER for a linear combination of parameters

Zou and Donner (2008) and Zou (2008) provide a detailed outline of the MOVER. This method was derived as follows: consider two parameters θ_1 and θ_2 which have $(1-\alpha)100\%$ confidence limits (l_1, u_1) and (l_2, u_2), respectively. The individual limits of θ_1 and θ_2 may be used to estimate variances near the limits of the sum of the two parameters ($\Sigma = \theta_1 + \theta_2$). These variance estimates may then be used to obtain a confidence interval for Σ.

To begin, the application of the central limit theorem and standardization results

in the following set of equations:

$$\begin{aligned} z^2_{\alpha/2} &= \frac{\left(\hat{\Sigma} - \Sigma\right)^2}{\widehat{\text{var}}(\hat{\Sigma})} \\ \Rightarrow \quad z^2_{\alpha/2} &= \frac{\left[(\hat{\theta}_1 + \hat{\theta}_2) - (\theta_1 + \theta_2)\right]^2}{\widehat{\text{var}}(\hat{\theta}_1 + \hat{\theta}_2)} \\ \Rightarrow \quad z^2_{\alpha/2} &= \frac{\left[(\hat{\theta}_1 + \hat{\theta}_2) - (\theta_1 + \theta_2)\right]^2}{\widehat{\text{var}}(\hat{\theta}_1) + \widehat{\text{var}}(\hat{\theta}_2) + 2\widehat{\text{cov}}(\hat{\theta}_1, \hat{\theta}_2)}. \end{aligned}$$

This thesis will only focus on the case when $\text{cov}(\hat{\theta}_1, \hat{\theta}_2) = 0$, because $\hat{\theta}_1$ and $\hat{\theta}_2$ are either from separate arms of the trial and are thus independent, or they are functions of the normal mean and normal variance which are independent (Lukacs, 1942). Thus,

$$z^2_{\alpha/2} = \frac{\left[(\hat{\theta}_1 + \hat{\theta}_2) - (\theta_1 + \theta_2)\right]^2}{\widehat{\text{var}}(\hat{\theta}_1) + \widehat{\text{var}}(\hat{\theta}_2)}. \tag{2.9}$$

An expression for the limits of Σ may be obtained by applying the inversion principle (Section 2.3.1) to Equation (2.9):

$$\begin{cases} L_\Sigma = (\hat{\theta}_1 + \hat{\theta}_2) - z_{\alpha/2}\sqrt{\widehat{\text{var}}(\hat{\theta}_1) + \widehat{\text{var}}(\hat{\theta}_2)} \\ U_\Sigma = (\hat{\theta}_1 + \hat{\theta}_2) + z_{\alpha/2}\sqrt{\widehat{\text{var}}(\hat{\theta}_1) + \widehat{\text{var}}(\hat{\theta}_2)}. \end{cases} \tag{2.10}$$

The variances in expression (2.10) are traditionally obtained by estimating them at the point estimate. However, fixing the variance in this way would force the interval to be symmetric around the point estimate. If the sampling distribution of the estimators may be approximated by a normal distribution, this restriction is not a problem. However, when the sampling distribution is skewed, symmetric intervals could lead to intervals with asymmetric tail errors. Consequently, too many potential parameter values may be excluded from one side of the interval and not enough from the other. Thus, Equations 2.2 and 2.3 would not be satisfied.

Figure 2.2a shows the resulting limits of a parameter estimate with a skewed sampling distribution when a single normal curve is used to estimate the variance

at the point estimate. Although balanced tail errors are intended, from the figure it can be seen that $\alpha_1 \neq \alpha_2$. This occurs because the separate limits are constructed by first fixing variance estimates at the point estimate, then plugging them in (using Slutsky's theorem).

Alternatively, the score confidence interval method is obtained by inverting the score test (Rao, 1948). The likelihood may be maximized at the maximum likelihood estimate, where the score equals zero. Therefore, the null hypothesis may then be tested by determining how much the score deviates from zero at the null value, with large deviations providing evidence that the null is untrue (Buse, 1982). This is done using the score statistic, which is the square of the score at the null divided by the variance at the null. This statistic has an approximate chi-squared distribution with one degree of freedom (Buse, 1982). By constantly updating the null, a confidence interval may then be constructed. The confidence interval would not be restricted to symmetry because the variances would be estimated at the limits. That is, the variance at the lower limit may be different from the variance at the upper limit. However, due to the complexity of the score function for clustered designs, this procedure will be excluded in this thesis.

Letting the interval of θ_i be given by (l_i, u_i), the MOVER approximates the score method (which estimates the variance at the lower limit) by using the variance estimate at $l_1 + l_2$ (near the lower limit) rather than the variance estimate at $\hat{\theta}_1 + \hat{\theta}_2$ (used by the delta method with Slutsky's theorem) when constructing the lower limit of $\theta_1 + \theta_2$, because $l_1 + l_2$ is closer to the lower limit than is $\hat{\theta}_1 + \hat{\theta}_2$. The distance

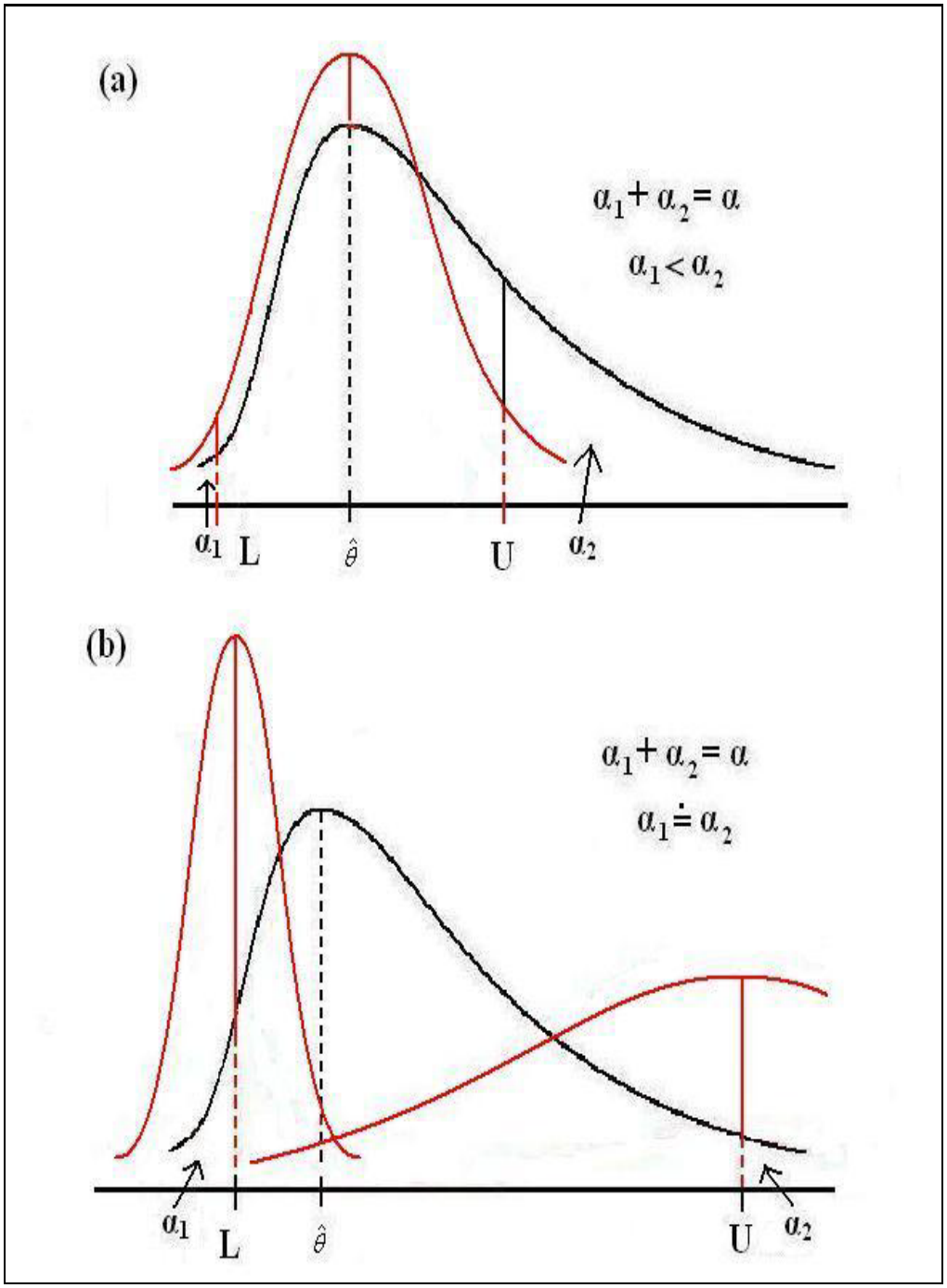

Figure 2.2: a) A symmetric confidence interval (L, U) for a summary measure $(\hat{\theta})$ of data following a skewed distribution using traditional methods. b) An asymmetric confidence interval (L, U) for a summary measure $(\hat{\theta})$ of data following a skewed distribution, by application of the MOVER.

between $l_1 + l_2$ and the lower limit of $\theta_1 + \theta_2$, is

$$\begin{aligned} & ||(l_1 + l_2) - L|| \\ = & \left|\left|\left(\hat{\theta}_1 - z_\alpha\sqrt{\text{var}(\hat{\theta}_1)} + \hat{\theta}_2 - z_\alpha\sqrt{\text{var}(\hat{\theta}_2)}\right) - \left(\hat{\theta}_1 + \hat{\theta}_2 - z_\alpha\sqrt{\text{var}(\hat{\theta}_1) + \text{var}(\hat{\theta}_2)}\right)\right|\right| \\ = & \left|\left|-z_\alpha\sqrt{\text{var}(\hat{\theta}_1)} - z_\alpha\sqrt{\text{var}(\hat{\theta}_1)} + z_\alpha\sqrt{\text{var}(\hat{\theta}_1) + \text{var}(\hat{\theta}_2)}\right|\right| \\ = & z_\alpha\left|\left|\sqrt{\text{var}(\hat{\theta}_1) + \text{var}(\hat{\theta}_2)} - \left(\sqrt{\text{var}(\hat{\theta}_1)} + \sqrt{\text{var}(\hat{\theta}_1)}\right)\right|\right|. \end{aligned}$$

This is smaller than the distance between $\hat{\theta}_1 + \hat{\theta}_2$ and L, expressed as

$$\begin{aligned} & \left|\left|(\hat{\theta}_1 + \hat{\theta}_2) - L\right|\right| \\ = & \left|\left|\hat{\theta}_1 + \hat{\theta}_2 - \left(\hat{\theta}_1 + \hat{\theta}_2 - z_\alpha\sqrt{\text{var}(\hat{\theta}_1) + \text{var}(\hat{\theta}_2)}\right)\right|\right| \\ = & z_\alpha\left|\left|\sqrt{\text{var}(\hat{\theta}_1) + \text{var}(\hat{\theta}_2)}\right|\right|. \end{aligned}$$

Similarly, $u_1 + u_2$ is closer to the upper limit of $\theta_1 + \theta_2$ than is $\hat{\theta}_1 + \hat{\theta}_2$.

Thus, the MOVER improves upon the traditional Wald method by separately estimating the variances *near* the limits. Traditional symmetric intervals are improved upon by removing the symmetry restriction without the use of complex procedures such as the score method. The flexibility of the MOVER stems from the use of two separate normal curves rather than one when estimating variances (see Figure 2.2b). This results in more balanced tail errors ($\alpha_1 \doteq \alpha_2$).

The variance terms in Equation (2.10) may be estimated near the upper and lower confidence limits of Σ using the information already available in the confidence intervals of the individual parameters, θ_1 and θ_2. An application of the central limit theorem ($z \sim (\hat{\theta}_i - \theta_i)/\sqrt{\text{var}(\hat{\theta}_i)}$) and Slutsky's theorem estimates variances near the

lower ($\widehat{\text{var}}_{li}(\hat{\theta}_i)$) and upper ($\widehat{\text{var}}_{ui}(\hat{\theta}_i)$) limits of θ_i as

$$\begin{aligned}\widehat{\text{var}}_{li}(\hat{\theta}_i) &= \frac{\left(\hat{\theta}_i - l_i\right)^2}{z^2_{\alpha/2}} \\ \widehat{\text{var}}_{ui}(\hat{\theta}_i) &= \frac{\left(u_i - \hat{\theta}_i\right)^2}{z^2_{\alpha/2}},\end{aligned}$$

for $i = 1, 2$. Using these estimates with Equation (2.10), two-sided $(1-\alpha)100\%$ confidence limits for Σ may be given as

$$\begin{aligned}L_\Sigma &= \hat{\theta}_1 + \hat{\theta}_2 - z_{\alpha/2}\sqrt{\widehat{\text{var}}_l(\hat{\theta}_1) + \widehat{\text{var}}_l(\hat{\theta}_2)} \\ &= \hat{\theta}_1 + \hat{\theta}_2 - z_{\alpha/2}\sqrt{\frac{(\hat{\theta}_1 - l_1)^2}{z^2_{\alpha/2}} + \frac{(\hat{\theta}_2 - l_2)^2}{z^2_{\alpha/2}}} \\ &= \hat{\theta}_1 + \hat{\theta}_2 - \sqrt{(\hat{\theta}_1 - l_1)^2 + (\hat{\theta}_2 - l_2)^2} \qquad (2.11)\end{aligned}$$

$$\begin{aligned}U_\Sigma &= \hat{\theta}_1 + \hat{\theta}_2 + z_{\alpha/2}\sqrt{\widehat{\text{var}}_u(\hat{\theta}_1) + \widehat{\text{var}}_u(\hat{\theta}_2)} \\ &= \hat{\theta}_1 + \hat{\theta}_2 + z_{\alpha/2}\sqrt{\frac{(u_1 - \hat{\theta}_1)^2}{z^2_{\alpha/2}} + \frac{(u_2 - \hat{\theta}_2)^2}{z^2_{\alpha/2}}} \\ &= \hat{\theta}_1 + \hat{\theta}_2 + \sqrt{(u_1 - \hat{\theta}_1)^2 + (u_2 - \hat{\theta}_2)^2}. \qquad (2.12)\end{aligned}$$

Similarly, confidence intervals may be obtained for a difference between two parameters using the transformation principle (Section 2.3.2). If the confidence limits for θ_2 are (l_2, u_2), then the confidence limits for $-\theta_2$ are $(-u_2, -l_2)$. A two-sided $(1-\alpha)100\%$ confidence interval for $\Delta = \theta_1 + (-\theta_2)$ may then be constructed using the equations above,

$$\begin{cases} L_\Delta = \hat{\theta}_1 - \hat{\theta}_2 - \sqrt{(\hat{\theta}_1 - l_1)^2 + (u_2 - \hat{\theta}_2)^2} \\ U_\Delta = \hat{\theta}_1 - \hat{\theta}_2 + \sqrt{(u_1 - \hat{\theta}_1)^2 + (\hat{\theta}_2 - l_2)^2}. \end{cases} \qquad (2.13)$$

In this thesis, the effect measures of interest may be expressed as linear functions of more than two parameters. The $(1-\alpha)100\%$ confidence interval for $h_1\theta_1 + h_2\theta_2 +$

$\cdots + h_n\theta_n$ is given by

$$L = \sum_{i=1}^{n} h_i\hat{\theta}_i - \sqrt{\sum_{i=1}^{n}\left[h_i\hat{\theta}_i - \min(h_i l_i, h_i u_i)\right]^2} \tag{2.14}$$

$$U = \sum_{i=1}^{n} h_i\hat{\theta}_i + \sqrt{\sum_{i=1}^{n}\left[h_i\hat{\theta}_i - \max(h_i l_i, h_i u_i)\right]^2} \tag{2.15}$$

where $\theta_1, \theta_2, ..., \theta_n$ are independent and $h_1, h_2, ..., h_n$ represent the coefficients of $\theta_1, \theta_2, ..., \theta_n$ which have individual $(1-\alpha)100\%$ limits $(l_1, u_1), (l_2, u_2), ..., (l_n, u_n)$, respectively. This can be proven using the method of induction:

Step 1: prove true for $n = 1$

Note that if the limits of θ_1 are known in practice, then the following steps to obtain the limits of $h_1\theta_1$ would be redundant. Rather, the transformation principle may simply be applied. However, for the purpose of proving Equations (2.14) and (2.15), it will be shown that when $n = 1$, the application of the transformation principle and the central limit theorem give $L = h_1\hat{\theta}_1 - \sqrt{\left[h_1\hat{\theta}_1 - \min(h_1 l_1, h_1 u_1)\right]^2}$ and $U = h_1\hat{\theta}_1 + \sqrt{\left[h_1\hat{\theta}_1 - \max(h_1 l_1, h_1 u_1)\right]^2}$.

Let the $(1-\alpha)\%$ confidence interval for θ_1 be known as (l_1, u_1). According to the transformation principle, if $h_1 > 0$ then the confidence interval for $h_1\theta_1$ is $(h_1 l_1, h_1 u_1)$. Next, the variance of $h_1\theta_1$ may be estimated near the limits using the central limit theorem:

$$\widehat{\text{var}}_{l1}\left[h_1(\hat{\theta}_1)\right] = \frac{\left(h_1\hat{\theta}_1 - h_1 l_1\right)^2}{z^2_{\alpha/2}}$$

$$\widehat{\text{var}}_{u1}\left[h_1(\hat{\theta}_1)\right] = \frac{\left(h_1 u_1 - h_1\hat{\theta}_1\right)^2}{z^2_{\alpha/2}}.$$

Alternatively, if $h_1 < 0$ then the limits of $h_1\theta_1$ are $(h_1 u_1, h_1 l_1)$ and its variance esti-

mates near the limits are

$$\widehat{\text{var}_{l1}}\left[h_1(\hat{\theta}_1)\right] = \frac{\left(h_1u_1 - h_1\hat{\theta}_1\right)^2}{z^2_{\alpha/2}}$$

$$\widehat{\text{var}_{u1}}\left[h_1(\hat{\theta}_1)\right] = \frac{\left(h_1\hat{\theta}_1 - h_1l_1\right)^2}{z^2_{\alpha/2}}.$$

Plugging these estimates into a standard confidence interval formula (Equation (2.10)) gives the desired expressions for the upper and lower limits when $n = 1$.

Step 2: assume true for $n = k$

Assume that the $(1-\alpha)100\%$ confidence interval for $h_1\theta_1 + \cdots + h_k\theta_k$ is

$$L_k = \sum_{i=1}^{k} h_i\hat{\theta}_i - \sqrt{\sum_{i=1}^{k}\left[h_i\hat{\theta}_i - \min(h_il_i, h_iu_i)\right]^2}$$

$$U_k = \sum_{i=1}^{k} h_i\hat{\theta}_i + \sqrt{\sum_{i=1}^{k}\left[h_i\hat{\theta}_i - \max(h_il_i, h_iu_i)\right]^2}.$$

Step 3: prove true for $n = k + 1$

The MOVER for a linear combination of two parameters has already been proven above. Thus, if $h_1\theta_1 + \cdots + h_k\theta_k$ is treated as the first parameter and $h_{k+1}\theta_{k+1}$ is treated as the second parameter with respective confidence intervals (L_k, U_k) and (l_{k+1}, u_{k+1}), then the expressions (2.14) and (2.15) hold for $n = k + 1$.

The MOVER for the ratio of two independent parameters

The derivation for the confidence interval of a ratio of two independent parameters using the MOVER may be found in Zou and Donner (2010). This derivation is as follows: let $\theta_1/\theta_2 = R$, where θ_1 and θ_2 are independent and R is some constant. Denote the $(1-\alpha)100\%$ confidence interval of R as (L_R, U_R). By the definition of confidence intervals,

$$\begin{aligned} & \text{P} \quad \left(\frac{\theta_1}{\theta_2} < L_R\right) = \alpha/2 \\ \Rightarrow \quad & \text{P} \quad (\theta_1 - L_R\theta_2 < 0) = \alpha/2 \end{aligned} \tag{2.16}$$

and

$$\begin{aligned} & \mathrm{P} \quad \left(\frac{\theta_1}{\theta_2} > U_R\right) = \alpha/2 \\ \Rightarrow \quad & \mathrm{P} \quad (\theta_1 - U_R\theta_2 > 0) = \alpha/2. \end{aligned} \tag{2.17}$$

Using the transformation principle, if the confidence interval of θ_2 is (l_2, u_2), then the intervals of $-L_R\theta_2$ and $-U_R\theta_2$ are $(-L_R u_2, -L_R l_2)$ and $(-U_R u_2, -U_R l_2)$, respectively. Thus, focusing on $\theta_1 - L_R\theta_2$ and $\theta_1 - U_R\theta_2$ by setting L_Σ and U_Σ to 0 in Equations (2.11) and (2.12) and applying the transformation principle results in the following quadratic equations:

$$(\hat{\theta}_1 - L_R\hat{\theta}_2)^2 = (\hat{\theta}_1 - l_1)^2 + (L_R u_2 - L_R\hat{\theta}_2)^2$$
$$(\hat{\theta}_1 - U_R\hat{\theta}_2)^2 = (u_1 - \hat{\theta}_1)^2 + (U_R\hat{\theta}_2 - U_R l_2)^2.$$

The quadratic formula is then used to solve these quadratic equations, setting

$$L_R = \frac{-b_L - \sqrt{b_L^2 - 4a_L c_L}}{2a_L} \tag{2.18}$$

$$U_R = \frac{-b_U + \sqrt{b_U^2 - 4a_U c_U}}{2a_U} \tag{2.19}$$

to the lower and upper $(1-\alpha)100\%$ limits of R, respectively, where $a_L = u_2(2\hat{\theta}_2 - u_2)$, $b_L = -2\hat{\theta}_1\hat{\theta}_2$, $c_L = l_1(2\hat{\theta}_1 - l_1)$ in Equation (2.18), and $a_U = l_2(2\hat{\theta}_2 - l_2)$, $b_U = -2\hat{\theta}_1\hat{\theta}_2$, $c_U = u_1(2\hat{\theta}_1 - u_1)$ in Equation (2.19).

The MOVER for the ratio of two independent parameters simplifies to Fiellers theorem (Fieller, 1944) when θ_1 and θ_2 are both normally distributed. Blaker and Spjotvoll (2000) show the possibility of three solutions when constructing confidence intervals for the special case of a ratio of two normal means using Fieller's theorem. These solutions include setting the limits to (L_R, U_R), to $(-\infty, U_R)$ and (L_R, ∞) when $U_R < L_R$, and to the whole real line $(-\infty, \infty)$. Figure 2.3 displays a confidence interval curve showing when each of these intervals exist for the general case of a ratio of two parameters. The confidence interval lies between the roots L_R and U_R in

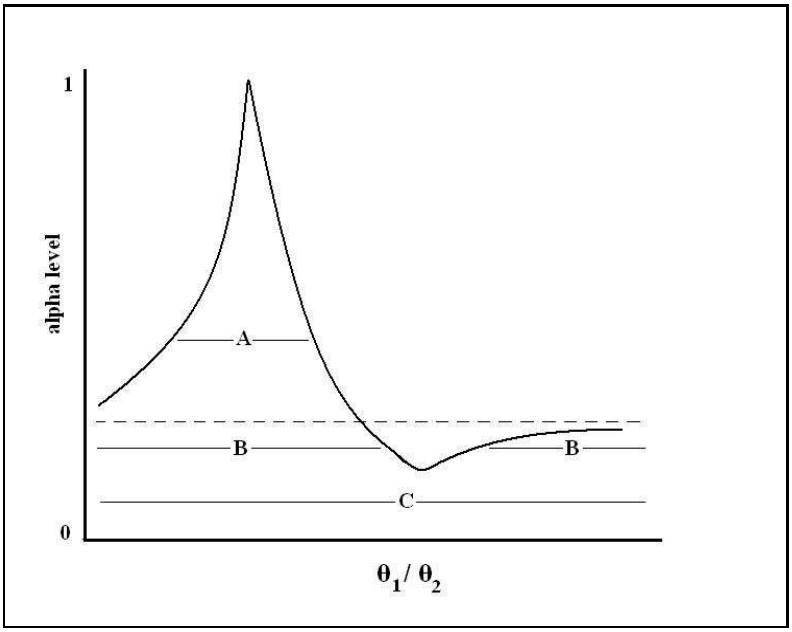

Figure 2.3: Confidence interval curve for a ratio of two parameters

situation 'A' of the figure because $L_R < U_R$ when the denominators of the roots ($2a_L$ and $2a_U$) are greater than zero and the two roots are real. The confidence interval lies outside of the roots in situation 'B' because $L_R > U_R$ when the denominators of the roots are less than zero and the two roots are real. Finally, the confidence interval is set to the whole real line in situation 'C' when the two roots are complex; that is, when $b_L^2 - 4a_L c_L < 0$ and $b_U^2 - 4a_U c_U < 0$. A similar argument is made when solving for quadratic equations in expressions 2.18 and 2.19.

Properties of the MOVER

The MOVER relaxes some assumptions of traditional confidence intervals. First, the MOVER does not restrict the interval to symmetry if the sampling distribution of the parameter estimate is skewed. The key step of the MOVER is the estimation of the variance of a linear combination of parameter estimates near their limits using

the individual confidence intervals of each parameter. Thus, two variance estimates exist for each parameter estimate: one near the upper limit and one near the lower limit. This is consistent with Neyman's definition of the confidence interval (Neyman, 1937), that the lower limit is the smallest estimate of the parameter for which the obtained point estimate would be the largest value that would occur by chance with a probability of $\alpha/2$. The upper limit is defined analogously.

The variance near the upper limit of the linear combination may differ from that near the lower limit, leading to asymmetric confidence intervals. If the variances near the limits are equivalent (this occurs if the individual confidence intervals of each parameter component are symmetric) then the resulting confidence interval of the linear combination will be symmetric. Therefore, it is evident that the application of the MOVER does not force symmetry. A depiction of this property is given in Figure (2.4) and was derived by Zou and Donner (2010) using the Pythagorean theorem. The figure demonstrates that the margins of error on either side of the point estimate (e.g. $\sqrt{(\widehat{\theta}_1 - l_1)^2 + (u_2 - \widehat{\theta}_2)^2}$ and $\sqrt{(u_1 - \widehat{\theta}_1)^2 + (\widehat{\theta}_2 - l_2)^2}$ for the the point estimate $\theta_1 - \theta_2$) are not necessarily restricted to equality. For instance, consider the top left diagram in Figure (2.4) for the difference between two parameters. The diagonal line in quadrant II represents the lower margin of error of the difference $\hat{\theta}_1 - \widehat{\theta}_2$. An estimate of this margin of error is obtained using the margin of error of $\widehat{\theta}_1$ near the lower limit, estimated by $\widehat{\theta}_1 - l_1$, and that of $\widehat{\theta}_2$, estimated by $u_2 - \widehat{\theta}_2$. Therefore, according to the Pythagorean theorem, the margin of error of $\widehat{\theta}_1 - \widehat{\theta}_2$ (depicted by the diagonal line in quadrant II) is given by $\sqrt{(\widehat{\theta}_1 - l_1)^2 + (u_2 - \widehat{\theta}_2)^2}$. Similarly, the diagonal line in quadrant IV represents the upper margin of error of the difference $\widehat{\theta}_1 - \hat{\theta}_2$, estimated by $\sqrt{(u_1 - \widehat{\theta}_1)^2 + (\widehat{\theta}_2 - l_2)^2}$. Note that in this first figure, θ_1 and θ_2 both have symmetric sampling distributions, as indicated by the lengths of the axes. Diagrams in the second row display the margins of error when θ_1 has an asymmetric sampling distribution, while θ_2 has a symmetric sampling distribution. Diagrams in the third row display margins of error when both θ_1 and θ_2 have asymmetric sampling

distributions.

Second, the MOVER does not assume homoscedasticity when constructing confidence intervals for a linear combination of parameters. The MOVER estimates the variance of each parameter estimate separately. These individual variance estimates are then used to obtain the variance estimate for the linear combination of parameter estimates. Since summary measures of each arm of a cluster randomized trail are often compared by their linear combination (e.g. their difference) and the variances of each summary measure are estimated separately, then no assumption of homoscedasticity is made.

A necessary condition for the MOVER to work well is that the confidence limits for each component parameter be valid. Otherwise, the confidence intervals constructed using the MOVER will adopt any handicaps of the confidence intervals of the component parameters. In fact, if confidence intervals for each component parameter are obtained with the Wald method, the MOVER will recover the Wald method for the linear combination of parameters. See below for more details.

Previous applications

The MOVER is a general confidence interval procedure which may be used for a linear combination or a ratio of parameters. It will be used to construct confidence intervals for the parameters of interest in this thesis. The key step of this procedure is estimating variances near the limits of the estimated parameter of interest using the confidence intervals of the component parameters. In fact, this step is what gave the MOVER its name (Zou, 2008).

The MOVER may be applied to recover many previously defined confidence interval procedures. One such confidence interval procedure was proposed by Howe (1974) for the mean of the sum of two random variables. The method was justified using the Cornish-Fisher expansion for the cases of the non-central t-distribution and variance components in the one-way random effects model.

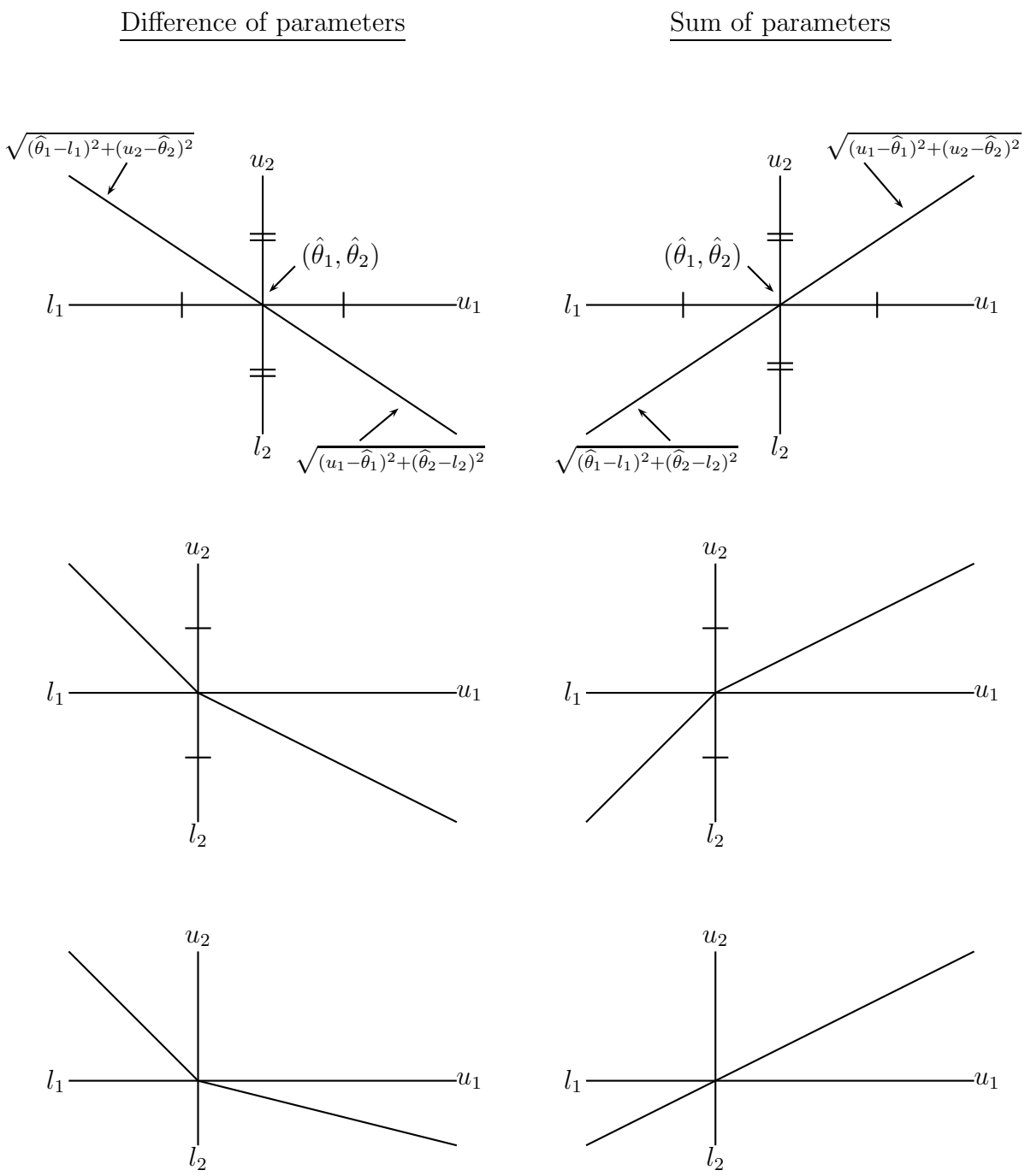

Figure 2.4: The flexibility of the MOVER for differences and sums as shown using margins of errors and the Pythagorean theorem

There has been a number of generalizations and applications of the approximation presented by Howe (1974) which may also be obtained by application of the MOVER. These include those of Graybill and Wang (1980), Ting *et al.* (1990), and Burdick and Graybill (1992) for a linear combination of chi-squared random variables. A generalization for a linear combination of random variables was then provided by Hyslop *et al.* (2000) and was recommended by the Food and Drug Administration (FDA) for evaluating individual bioequivalence (Food and Administration, 1999). Although these resulting confidence interval expressions have appeared previously, the MOVER takes these procedures a step further by providing a new justification while also extending them for a ratio of two random variables (Zou and Donner, 2010).

Other past works which may be justified by application of the MOVER include Burdick and Graybill (1984); Wang and Chow (2002); Ames and Webster (1991); Newcombe (1998); Lee *et al.* (2004); Burdick *et al.* (2006). Under certain situations (specified in each of the papers above), application of the MOVER would result in equivalent confidence interval expressions. For instance, Newcombe (1998) constructed confidence intervals for the difference between two independent proportions using Wilson's score confidence interval for a single proportion (Wilson, 1927) and found that the intervals maintained coverage for a wide range of parameter combinations. Newcomb's method can be justified analytically using the MOVER approach.

We also note that the MOVER may be used to recover standard Wald intervals when the sampling distribution of each individual component parameter has Wald-type confidence limits. For instance, standard Wald-type confidence intervals for $\theta_1 - \theta_2$ are

$$\begin{cases} L = \widehat{\theta}_1 - \widehat{\theta}_2 - z_{\alpha/2}\sqrt{\widehat{\mathrm{var}}(\widehat{\theta}_1) + \widehat{\mathrm{var}}(\widehat{\theta}_2)} \\ U = \widehat{\theta}_1 - \widehat{\theta}_2 + z_{\alpha/2}\sqrt{\widehat{\mathrm{var}}(\widehat{\theta}_1) + \widehat{\mathrm{var}}(\widehat{\theta}_2)}, \end{cases}$$

where θ_1 and θ_2 are independent. Now let θ_i have the limits

$$\begin{cases} l_i = \widehat{\theta}_i - z_{\alpha/2}\sqrt{\widehat{\text{var}}(\widehat{\theta}_i)} \\ u_i = \widehat{\theta}_i + z_{\alpha/2}\sqrt{\widehat{\text{var}}(\widehat{\theta}_i)}. \end{cases}$$

Application of the MOVER in Equations (2.11) and (2.12) then gives the following limits for $\theta_1 - \theta_2$

$$\begin{aligned} L &= \widehat{\theta}_1 - \widehat{\theta}_2 - \sqrt{(\widehat{\theta}_1 - l_1)^2 + (u_2 - \widehat{\theta}_2)^2} \\ &= \widehat{\theta}_1 - \widehat{\theta}_2 - \sqrt{z^2_{\alpha/2}\widehat{\text{var}}(\widehat{\theta}_1) + z^2_{\alpha/2}\widehat{\text{var}}(\widehat{\theta}_2)} \\ &= \widehat{\theta}_1 - \widehat{\theta}_2 - z_{\alpha/2}\sqrt{\widehat{\text{var}}(\widehat{\theta}_1) + \widehat{\text{var}}(\widehat{\theta}_2)} \end{aligned}$$

$$\begin{aligned} U &= \widehat{\theta}_1 - \widehat{\theta}_2 + \sqrt{(\widehat{\theta}_1 - l_1)^2 + (u_2 - \widehat{\theta}_2)^2} \\ &= \widehat{\theta}_1 - \widehat{\theta}_2 + \sqrt{z^2_{\alpha/2}\widehat{\text{var}}(\widehat{\theta}_1) + z^2_{\alpha/2}\widehat{\text{var}}(\widehat{\theta}_2)} \\ &= \widehat{\theta}_1 - \widehat{\theta}_2 + z_{\alpha/2}\sqrt{\widehat{\text{var}}(\widehat{\theta}_1) + \widehat{\text{var}}(\widehat{\theta}_2)}. \end{aligned}$$

This thesis aims to use the MOVER to obtain new confidence interval expressions for common parameters of interest for continuous outcomes, specifically for data arising from cluster randomization trials. These parameters include a difference between two normal means, a difference between two lognormal means, and the exceedance probability. The proposed confidence intervals will not force symmetry or assume homoscedasticity.

Chapter 3

CONFIDENCE INTERVAL ESTIMATION FOR EFFECT MEASURES IN CLUSTER RANDOMIZATION TRIALS

The first chapter of this thesis outlined the notation of a completely randomized cluster randomization trial, with each arm of the trial following a random effects model and two arms additionally differing according to a fixed effect. Next, key outcomes commonly of interest within cluster randomization trials were discussed. These include the difference between two normal means, the difference between two lognormal means, and the exceedance probability. The chapter also introduced existing confidence interval procedures for these effect measures, demonstrating that a simple and valid method for these measures as applied to cluster randomization trials does not currently exist.

Chapter two discussed confidence intervals in more detail. The advantages and challenges of the well known Wald-type confidence intervals with application of the delta method were revealed. The MOVER (Zou and Donner, 2008), a general confidence interval method for the linear combination or a ratio of parameters, was then discussed as a potential remedy to these challenges.

This chapter unites Chapters one and two by applying the MOVER with existing confidence intervals for the components of a random effects model to construct new confidence intervals for the parameters of interest as applied to cluster randomization trials. These closed form intervals are flexible, allowing asymmetric limits around the point estimate for parameter estimates with asymmetric distributions. The proposed and existing confidence interval procedures are asymptotic, suggesting that similar

results will be observed as the number of clusters gets large. Note that for the remainder of the thesis, the proposed confidence intervals will be referred to as the MOVER.

3.1 A difference between two normal means

The MOVER

Clinical trials often compare the mean values of two groups of subjects where each observation has an approximate normal distribution. The confidence interval of the mean of group i may be expressed once the variance of the sample mean is obtained. Donner and Klar (2000, page 8) define the variance of the sample mean as

$$\text{var}(\bar{\bar{Y}}_i) = \frac{\sigma^2_{A_i} + \sigma^2_{E_i}/m_i}{k_i}$$

for the balanced design, where m_i is the size of the clusters in group (arm) i and k_i is the number of clusters in group i. However, unbalanced cluster sizes occur more frequently in practice than balanced cluster sizes (Eldridge *et al.*, 2006). This variance expression may then be extended for the unbalanced design using the unweighted mean squared error expressed in Equation (1.4) (Thomas and Hultquist, 1978), if we use the standard assumption that data from each arm of a cluster randomization trial follows a one-way random effects model. The resulting approximate confidence interval for a single mean is given by

$$\begin{cases} L_{\mu_i} = \bar{\bar{Y}}_i - t_{1-\alpha/2,k_i-1}\sqrt{\dfrac{S^2_{Ui}}{k_i n_{Hi}}} \\ U_{\mu_i} = \bar{\bar{Y}}_i + t_{1-\alpha/2,k_i-1}\sqrt{\dfrac{S^2_{Ui}}{k_i n_{Hi}}}, \end{cases} \tag{3.1}$$

where $t_{1-\alpha/2,k_i-1}$ is the $(1-\alpha/2)100\%$ quantile of the t-distribution with k_i-1 degrees of freedom, S^2_{Ui} is the unweighted mean squared error, and n_{Hi} is the harmonic mean of the cluster sizes for arm i $(i = 1, 2)$, respectively. El-Bassiouni and Abdelhafez

(2000) showed that this approximate interval maintains coverage for the unbalanced one-way random effects model. However, the interval tends to be wide when the ICC < 0.2.

Equation (2.13) may now be applied with the interval of each mean (Equation (3.1)) to obtain the $(1-\alpha)100\%$ confidence interval of the difference between two normal means as

$$\begin{cases} L_\Delta = \bar{\bar{Y}}_1 - \bar{\bar{Y}}_2 - \sqrt{A_1 + A_2} \\ U_\Delta = \bar{\bar{Y}}_1 + \bar{\bar{Y}}_2 + \sqrt{A_1 + A_2}, \end{cases} \tag{3.2}$$

where

$$A_i = t^2_{1-\alpha/2,k_i-1} \frac{S^2_{Ui}}{k_i n_{Hi}}.$$

Note that the variance of each sample mean is estimated separately in Equation (3.2), therefore the procedure does not assume homoscedascity. Also, these limits are symmetric because each mean follows a normal and therefore symmetric distribution, making their difference also normally distributed (Casella and Berger, 2002, page 159-160). The limits in Equation (3.2) may be considered as an extension of those by Wang and Chow (2002), who derived results for independent data rather than clustered data.

Alternative confidence intervals

Wald confidence interval

Wald-type confidence intervals (Equation 2.4) may be constructed using the results of Thomas and Hultquist (1978) for a difference between two normal means, where $\hat{\theta} = \bar{\bar{Y}}_1 - \bar{\bar{Y}}_2$, $\widehat{\text{var}}(\hat{\theta}) = S^2_{U1}/k_1 n_{H1} + S^2_{U2}/k_2 n_{H2}$, S^2_{Ui} is the unweighted mean squared error of arm i, k_i is the number of clusters in arm i, and n_{Hi} is the harmonic mean of

the cluster sizes in arm i. The confidence interval is then given by

$$\begin{cases} L = \bar{\bar{Y}}_1 - \bar{\bar{Y}}_2 - z_{\alpha/2}\sqrt{\dfrac{S_{U1}^2}{k_1 n_{H1}} + \dfrac{S_{U2}^2}{k_2 n_{H2}}} \\ U = \bar{\bar{Y}}_1 + \bar{\bar{Y}}_2 + z_{\alpha/2}\sqrt{\dfrac{S_{U1}^2}{k_1 n_{H1}} + \dfrac{S_{U2}^2}{k_2 n_{H2}}}. \end{cases} \tag{3.3}$$

This interval is similar to the MOVER (Equation (3.2)), except that the MOVER uses the $(1-\alpha/2)100\%$ quantile of the t-distribution while Equation (3.3) uses the $(1-\alpha/2)100\%$ quantile of the standard normal distribution. As a result, confidence intervals obtained using Equation (3.3) will be narrower than those obtained using Equation (3.2).

Cluster-adjusted confidence interval

An alternative confidence interval procedure for the difference between two normal means from a cluster randomization trial is given by Donner and Klar (1993). This procedure is referred to as the cluster-adjusted confidence interval method. This method uses a pooled estimate of the standard error, thereby assuming homoscedasticity. Although a common variance and a common ICC may be assumed under the null hypothesis when hypothesis tests or significance tests are used, these assumptions may be questionable when constructing confidence intervals (Donner and Klar, 2000, page 96). This is because confidence intervals reflect a range of possible values, not just the null value of 0.

Once the pooled variance is adjusted for clustering, it is plugged into the usual t-interval

$$\hat{\theta} \pm t_{\alpha/2,df}\sqrt{\widehat{\text{var}}(\hat{\theta})}$$

and degrees of freedom (df) are set to the total number of clusters minus two. Theoretically this asymptotic method would be better than ignoring the effect of clustering; however, it has not yet been evaluated in a simulation study.

The cluster-adjusted confidence interval (Donner and Klar, 1993) for a difference between two normal means is given by

$$\bar{\bar{Y}}_1 - \bar{\bar{Y}}_2 \pm t_{\alpha/2,df}\widehat{\mathrm{SE}}(\bar{\bar{Y}}_1 - \bar{\bar{Y}}_2), \tag{3.4}$$

where

$$\begin{aligned}
\widehat{\mathrm{SE}}(\bar{\bar{Y}}_1 - \bar{\bar{Y}}_2) &= S_P \left[\frac{C_1}{M_1} + \frac{C_2}{M_2}\right]^{1/2}, \\
S_P &= \sqrt{\frac{(M_1-1)S^2_{T_1} + (M_2-1)S^2_{T_2}}{M-2}}, \\
S^2_{T_i} &= \sum_{j=1}^{k_i}\sum_{l_1}^{m_{ij}} \frac{(Y_{ijl} - \bar{Y}_i)^2}{M_i - 1}, \\
C_i &= \sum_{j=1}^{k_i} m_{ij}\frac{1 + (m_{ij}-1)\hat{\rho}_i}{M_i}, \\
\hat{\rho}_i &= \frac{S^2_{A_i}}{S^2_{A_i} + S^2_{E_i}},
\end{aligned}$$

$df = k_1 + k_2 - 2$, and C_1 and C_2 are estimated separately in the two trial arms. Note that the estimated ICC is truncated at zero, because negative ICC values are not of interest in this thesis. The main difference between this cluster-adjusted procedure and the MOVER is that this method uses a pooled variance estimate while the MOVER does not assume homoscedasticity.

Generalized confidence interval

The generalized confidence interval procedure (Weerahandi, 1993) may be used to construct confidence intervals for a difference between two normal means in a cluster randomization trial when applying the unweighted mean squared error statistic, proposed by Thomas and Hultquist (1978). The generalized confidence interval procedure is based on simulation and requires the existence of a generalized pivotal quantity. Let $G = g(\mathbf{X}; \mathbf{x}, \mathbf{v})$ be a function of $\mathbf{X}$, $\mathbf{x}$, and $\mathbf{v}$, where $\mathbf{X}$ is a random variable, $\mathbf{x}$ is the observed data, $\mathbf{v} = (\theta, \gamma)$ is a vector of unknown parameters, θ is the parameter

of interest, and γ is a vector of nuisance parameters. The statistic G is a generalized pivotal quantity if it satisfies the following two properties (Weerahandi, 1993):

(i) the probability distribution of the pivotal, $G = g(\mathbf{X}; \mathbf{x}, \mathbf{v})$, is free of any unknown parameters, v, and
(ii) the observed value of the pivotal, $G_{obs} = g(\mathbf{x}; \mathbf{v})$, does not depend of γ.

If C_α is a region such that $\mathrm{P}(G \in C_\alpha) = 1 - \alpha$, then the values of θ which satisfy $\{\theta : G(\mathbf{x}; \mathbf{x}, \mathbf{v}) \in C_\alpha\}$ is the $(1-\alpha)100\%$ generalized confidence interval of θ. Note that property (i) guarantees that C_α is independent of θ and γ. Property (ii) guarantees that the generalized confidence interval can be obtained using only the observed values. However, a consequence of property (ii), aside from having to know the generalized pivotal quantity in advance, is that the resulting generalized confidence interval is not closed form. This is because the pivotal quantity is dependent on the simulation of random variables, as shown below.

Krishnamoorthy *et al.* (2007) provide a pivotal quantity for a single normal mean in a one-way random effects model,

$$G_\mu = \bar{\bar{Y}} + \frac{Z}{\sqrt{\chi^2_{k-1}}}\sqrt{\frac{\mathrm{SSC}}{k}},$$

where $\bar{\bar{Y}}$ is the overall mean, $Z \sim N(0,1)$, χ^2_{df} is a random variable from the chi-squared distribution with df degrees of freedom, SSC is the sum of squares among groups (clusters), and k is the number of groups.

Each arm of a cluster randomization trial may follow a one-way random effects model, therefore Equation (3.5) may be used to construct a generalized confidence interval for a normal mean from a cluster randomization trial by including a subscript i to indicate the trial arm,

$$G_{\mu_i} = \bar{\bar{Y}}_i + \frac{Z}{\sqrt{\chi^2_{k_i-1}}}\sqrt{\frac{\mathrm{SSC}_i}{k_i}}. \tag{3.5}$$

Interest lies in constructing generalized confidence intervals for a difference between two normal means. The generalized pivotal quantity of a function of parameters, such as $f(\mu_1, \mu_2) = \mu_1 - \mu_2$, is simply the same function applied to the pivotal quantity of each parameter, $G_{\mu_1} - G_{\mu_2}$ (Krishnamoorthy *et al.*, 2007). Therefore, the following algorithm may be used to construct generalized confidence intervals for a difference between two normal means from a cluster randomization trial:

1. Use the dataset to compute $\bar{\bar{Y}}_i$ and SSC
2. For $i = 1, ..., 1000$

- generate $Z \sim N(0,1)$ and χ^2_{k-1}
- use these generated random variables to compute G_{μ_1} and G_{μ_2}, then $G_{\mu_1} - G_{\mu_2}$

3. Sort $G_{\mu_1} - G_{\mu_2}$ in ascending order
4. Set the lower limit (L) and the upper limit (U) to the $100(\alpha/2)\%$ and $100(1-\alpha/2)\%$ percentiles of the sorted $G_{\mu_1} - G_{\mu_2}$ values, respectively.

A disadvantage of the asymptotic generalized confidence interval procedure is that it is based on simulation, potentially leading to differing results for the same dataset. Another disadvantage is that a generalized pivotal is required, yet there are no general rules for obtaining such a pivotal.

3.2 A difference between two lognormal means

The MOVER for a single mean

Let X_{ijl}, $i = 1, 2$, $j = 1, ..., k_i$, and $l = 1, ..., m_{ij}$ represent lognormal data with parameters μ_i, σ_{A_i}, and σ_{E_i}. Thus, the log-transformed variables Y_{ijl} are normally distributed, $N(\mu_i, \sigma^2_{A_i} + \sigma^2_{E_i})$. The lognormal mean from an arm of a cluster randomized trial may be expressed as

$$\mathrm{E}(X) = \exp\left[\mu_i + \frac{\sigma^2_{iA} + \sigma^2_{iE}}{2}\right]$$

and estimated by

$$\widehat{\mathrm{E}(X)} = \exp\left[\bar{\bar{Y}}_i + \frac{S^2_{A_i} + S^2_{E_i}}{2}\right],$$

where $\bar{\bar{Y}}_i$ is the sample mean of the log-transformed observations of arm i, $S^2_{A_i}$ is the between-cluster sample variance of the log-transformed observations of arm i, and $S^2_{E_i}$ is the within-cluster sample variance of the log-transformed observations of arm i $(i = 1, 2)$.

Let the exponent of the lognormal mean be set to

$$\mu_i + (\sigma^2_{iA} + \sigma^2_{iE})/2 = \theta_{1i} + (\theta_{2i} + \theta_{3i})/(2n_{Hi}),$$

where

$$\begin{aligned} \theta_{1i} &= \mu_i, \\ \theta_{2i} &= \sigma^2_{E_i} + n_{Hi}\sigma^2_{A_i}, \text{and} \\ \theta_{3i} &= (n_{Hi} - 1)\sigma^2_{E_i}. \end{aligned} \tag{3.6}$$

The $(1-\alpha)100\%$ confidence interval of θ_{1i} is given in Equation (3.1) and the intervals of θ_{2i} and θ_{3i} are given by

$$\left[\frac{(k_i - 1)S^2_U}{\chi^2_{1-\alpha/2,k_i-1}}, \frac{(k_i - 1)S^2_U}{\chi^2_{\alpha/2,k_i-1}}\right] \tag{3.7}$$

and

$$\left[\frac{(n_{Hi} - 1)(M - k)S^2_E}{\chi^2_{1-\alpha/2,M_i-k_i}}, \frac{(n_{Hi} - 1)(M - k)S^2_E}{\chi^2_{\alpha/2,M_i-k_i}}\right], \tag{3.8}$$

respectively, since $(k_i-1)S^2_{Ui}/(\sigma^2_{E_i}+n_{Hi}\sigma^2_{A_i}) \sim$ (approx.)$\chi^2_{k_i-1}$ (Thomas and Hultquist, 1978) and $(M_i - k_i)S^2_{E_i}/\sigma^2_{E_i} \sim \chi^2_{M_i-k_i}$. To obtain the confidence interval of a single lognormal mean, Equations (2.11) and (2.12) may be applied with the transformation principle (Section 2.3.2) to the limits of θ_{2i} and θ_{3i} obtained from Equations (3.7) and (3.8) to find the confidence limits of

$$\begin{aligned} \theta_{4i} &= (\theta_{2i} + \theta_{3i})/(2n_{Hi}) \\ &= \frac{\sigma^2_{A_i} + \sigma^2_{E_i}}{2}, \end{aligned}$$

given by

$$\begin{cases} l_{\theta_{4i}} = \dfrac{S_{A_i}^2 + S_{E_i}^2}{2} - \sqrt{\dfrac{B_i}{4} + \dfrac{C_i}{4}} \\ u_{\theta_{4i}} = \dfrac{S_{A_i}^2 + S_{E_i}^2}{2} + \sqrt{\dfrac{D_i}{4} + \dfrac{E_i}{4}}, \end{cases} \tag{3.9}$$

where

$$B_i = \frac{1}{n_{Hi}^2}\left(S_{E_i}^2 + n_{Hi}^2 S_{A_i}^2 - \frac{(k_i-1)S_{Ui}^2}{\chi^2_{1-\alpha/2,k_i-1}}\right)^2,$$

$$C_i = \frac{(n_{Hi}-1)^2 S_{E_i}^4}{n_{Hi}^2}\left(1 - \frac{M_i - k_i}{\chi^2_{1-\alpha/2,M_i-k_i}}\right)^2,$$

$$D_i = \frac{1}{n_{Hi}^2}\left(\frac{(k_i-1)S_{Ui}^2}{\chi^2_{\alpha/2,k_i-1}} - S_{E_i}^2 - n_{Hi}^2 S_{A_i}^2\right)^2,$$

$$E_i = \frac{(n_{Hi}-1)^2 S_{E_i}^4}{n_{Hi}^2}\left(\frac{M_i - k_i}{\chi^2_{\alpha/2,M_i-k_i}} - 1\right)^2.$$

Equivalent limits were evaluated by Burdick and Graybill (1984) and shown to maintain coverage for numerous unbalanced designs, though they may be liberal under extreme unbalance when the ICC is less than 0.2.

Equations (2.11) and (2.12) for a confidence interval of the sum of two parameters may again be applied, this time with the limits of θ_{1i} (Equation (3.1)) and θ_{4i} (Equation (3.9)), to find the limits of $\theta_{1i} + \theta_{4i} = \mu_i + (\sigma_{A_i}^2 + \sigma_{E_i}^2)/2$. Exponentiating the limits of $\mu_i + (\sigma_{A_i}^2 + \sigma_{E_i}^2)/2$ then provides the $(1-\alpha)100\%$ confidence interval of a single lognormal mean,

$$\begin{cases} l_i = \exp\left(\bar{\bar{Y}}_i + \dfrac{S_{A_i}^2 + S_{E_i}^2}{2} - \sqrt{A_i + \dfrac{B_i}{4} + \dfrac{C_i}{4}}\right) \\ u_i = \exp\left(\bar{\bar{Y}}_i + \dfrac{S_{A_i}^2 + S_{E_i}^2}{2} + \sqrt{A_i + \dfrac{D_i}{4} + \dfrac{E_i}{4}}\right). \end{cases} \tag{3.10}$$

Note that these limits are not restricted to symmetry because the limits of the components θ_{2i} and θ_{3i} were not restricted to symmetry.

The MOVER for a difference between two lognormal means

When comparing lognormal means from two groups, researchers are frequently interested in inferences on the difference between means. Let $\mathrm{E}(X_i)$ denote the lognormal mean of arm i $(i = 1, 2)$ and $\widehat{\mathrm{E}(X_i)}$ denote its estimate. Equation (2.13) may then be applied to construct confidence limits for the difference, $\mathrm{E}(X_1) - \mathrm{E}(X_2)$, once a confidence interval for each mean is found using Equation (3.10). The $(1-\alpha)100\%$ confidence interval for the difference between two lognormal means is given by

$$\begin{cases} L_{\Delta_{\mathrm{E}(X)}} = \widehat{\mathrm{E}(X_1)} - \widehat{\mathrm{E}(X_2)} - \sqrt{\widehat{\mathrm{E}(X_1)}^2 F_1^2 + \widehat{\mathrm{E}(X_2)}^2 G_2^2} \\ U_{\Delta_{\mathrm{E}(X)}} = \widehat{\mathrm{E}(X_1)} - \widehat{\mathrm{E}(X_2)} + \sqrt{\widehat{\mathrm{E}(X_1)}^2 G_1^2 + \widehat{\mathrm{E}(X_2)}^2 F_2^2} \end{cases} \tag{3.11}$$

where

$$F_i = 1 - \frac{1}{\exp\left(\sqrt{A_i + B_i/4 + C_i/4}\right)}$$
$$G_i = \exp\left(\sqrt{A_i + D_i/4 + E_i/4}\right) - 1,$$

for $i = 1, 2$. Similar to the confidence interval for a single mean, these limits are not forced to be symmetric because the intervals of the components θ_{2i} and θ_{3i} are not symmetric. Furthermore, they do not assume homoscedasticity because the variance of each sample lognormal mean is estimated separately.

Alternative confidence intervals

Wald confidence interval and the delta method

The multivariate delta method (Section 2.4.2) may be used to construct symmetric Wald-type confidence intervals for the difference between two lognormal means. A single lognormal mean may be expressed as

$$g_i = \exp(\theta_{1i} + \theta_{2i} + \theta_{3i}),$$

where

$$\begin{aligned}\theta_{1i} &= \mu_i \\ \theta_{2i} &= \frac{(n_{Hi}-1)\sigma^2_{E_i}}{2n_{Hi}}, \text{ and} \\ \theta_{3i} &= \frac{\sigma^2_{E_i} + n_{Hi}\sigma^2_{A_i}}{2n_{Hi}}\end{aligned}$$

have sample variances of

$$\begin{aligned}\text{var}(\hat{\theta}_{1i}) &= \frac{S^2_{Ui}}{k_i n_{Hi}}, \\ \text{var}(\hat{\theta}_{2i}) &= \frac{(n_{Hi}-1)^2\sigma^4_{E_i}}{2n^2_{Hi}(M_i-k_i)}, \text{ and} \\ \text{var}(\hat{\theta}_{3i}) &= \frac{S^4_{Ui}}{2n^2_{Hi}(k_i-1)},\end{aligned}$$

respectively for arm $i = 1, 2$. The multivariate delta method may then be used to obtain a variance estimate for the sample lognormal mean of arm i,

$$\widehat{\text{var}}(\hat{g}_i) = H_i + I_i + J_i,$$

where

$$\begin{aligned}H_i &= \frac{\exp(2\bar{\bar{Y}}_i + \sigma^2_{A_i} + \sigma^2_{E_i})S^2_{Ui}}{k_i n_{Hi}} \\ I_i &= \frac{\exp(2\bar{\bar{Y}}_i + \sigma^2_{A_i} + \sigma^2_{E_i})(H_{Hi}-1)^2\sigma^4_{E_i}}{2n^2_{Hi}(M_i-k_i)} \\ J_i &= \frac{\exp(2\bar{\bar{Y}}_i + \sigma^2_{A_i} + \sigma^2_{E_i})S^4_{Ui}}{2n^2_{Hi}(k_i-1)}.\end{aligned}$$

Wald-type confidence intervals for the difference between two lognormal means are then given by

$$\begin{cases} L_W = \widehat{LN}_1 - \widehat{LN}_2 - z_{\alpha/2}\sqrt{\widehat{\text{var}}(\hat{g}_1) + \widehat{\text{var}}(\hat{g}_2)} \\ U_W = \widehat{LN}_1 - \widehat{LN}_2 + z_{\alpha/2}\sqrt{\widehat{\text{var}}(\hat{g}_1) + \widehat{\text{var}}(\hat{g}_2)}. \end{cases} \tag{3.12}$$

It can clearly be seen from Equation (3.12) that these confidence intervals are symmetric, though the sampling distribution for a lognormal mean is skewed. It would therefore be expected that the interval would not have balanced tail errors.

Generalized confidence intervals

Generalized confidence intervals for the difference between two lognormal means in cluster randomization trials may be constructed using the generalized pivotal quantities given by Krishnamoorthy *et al.* (2007) for the three parameters in a one-way random effects model (μ, σ_A^2, and σ_E^2). These generalized pivotal quantities may be constructed using the unweighted mean squared error (Thomas and Hultquist, 1978). The generalized pivotal quantity for μ_i is given by Equation (3.5) in Section 3.1.2.3, while those of $\sigma_{A_i}^2$ and $\sigma_{E_i}^2$ (for $i = 1, 2$) are given by

$$G_{\sigma_{A_i}^2} = \max\left(\frac{\mathrm{SSC_i}}{\chi_{k_i-1}^2} + \frac{\tilde{n}_i \mathrm{SSW_i}}{\chi_{M_i-k_i}^2}, 0\right) \text{ and} \tag{3.13}$$

$$G_{\sigma_{E_i}^2} = \frac{\mathrm{SSW_i}}{\chi_{M_i-k_i}^2}, \tag{3.14}$$

respectively, where χ_{df}^2 is a random variable from the chi-squared distribution with df degrees of freedom, SSC is the sum of squares between groups (or clusters), SSW is the sum of squares within groups, $\tilde{n} = (1/k)\sum_{j=1}^{k} m_j$, and each arm of a cluster randomization trial follows a one-way random effects model.

A generalized pivotal quantity for the lognormal mean, $\exp\left(\mu_i + (\sigma_{A_i}^2 + \sigma_{E_i}^2)/2\right)$, of arm i may be obtained using the generalized pivotal quantities of the three parameter components of the lognormal mean (μ_i, $\sigma_{A_i}^2$, and $\sigma_{E_i}^2$) by substituting G_{μ_i}, $G_{\sigma_{A_i}^2}$, and $G_{\sigma_{E_i}^2}$ for μ_i, $\sigma_{A_i}^2$, and $\sigma_{E_i}^2$, respectively. A generalized pivotal quantity for the lognormal mean is then given by

$$G_{\mathrm{E}(X)_i} = \exp\left(G_{\mu_i} + \frac{G_{\sigma_{A_i}^2} + G_{\sigma_{E_i}^2}}{2}\right),$$

$(i = 1, 2)$. A generalized pivotal quantity for a difference between two lognormal means may be expressed as

$$G_{\mathrm{E}(X)_1} - G_{\mathrm{E}(X)_2}. \tag{3.15}$$

Generalized confidence intervals for the difference between two lognormal means may be obtained using the algorithm in Section 3.1.2.3, by substituting $G_{\mathrm{E}(X)_i}$ for G_{μ_i}.

Similar to the MOVER (Equation (3.11)), these intervals make use of the unweighted mean squared error (Equation (1.4)), proposed by Thomas and Hultquist (1978). Therefore, they will have similar properties to the MOVER for the difference between two lognormal means which also apply the unweighted mean squared error.

The major differences between these generalized confidence intervals and the MOVER are that generalized confidence intervals are based on simulation, whereas the MOVER is easier to obtain and are closed form. This is especially important when constructing confidence intervals for the lognormal mean, because the limits are first estimated on the log scale, then exponentiated to obtain the limits of the mean. Without a closed form solution where different limits may be obtained at separate occasions, even a small difference in the limits on the log scale may translate into a clinically significant difference once exponentiated.

3.3 The exceedance probability

The MOVER

To obtain confidence intervals for

$$\mathrm{P}\left(Y_1 > Y_2\right) = \Phi\left(\frac{\mu_1 - \mu_2}{\sqrt{\sigma_{T_1}^2 + \sigma_{T_2}^2}}\right),$$

where $\sigma_{T_i}^2 = \sigma_{A_i}^2 + \sigma_{E_i}^2$, the limits of the standardized mean difference must first be obtained. The expression for the standardized mean difference is given by

$$\delta = \frac{(\mu_1 - \mu_2)}{\sqrt{\sigma_{T_1}^2 + \sigma_{T_2}^2}}.$$

Confidence intervals of the numerator of δ are given in Equation (3.2). The confidence limits of

$$\begin{aligned} \sigma_{T_i}^2 &= \frac{1}{n_{Hi}}\left[\left(\sigma_{E_i}^2 + n_{Hi}\sigma_{A_i}^2\right) + \left((n_{Hi}-1)\sigma_{E_i}^2\right)\right] \\ &= \sigma_{A_i}^2 + \sigma_{E_i}^2 \end{aligned}$$

may be found using the transformation principle (Section 2.3.2) with Equations (2.11) and (2.12), such that $(k_i - 1)S_{Ui}^2/(\sigma_{E_i}^2 + n_{Hi}\sigma_{A_i}^2) \sim$ (approx.)$\chi_{k_i-1}^2$ (Thomas and Hultquist, 1978) and $(M_i - k_i)S_{E_i}^2/\sigma_{E_i}^2 \sim \chi_{M_i-k_i}^2$ (Graybill, 1976, page 613). The limits of $\sigma_{T_1}^2 + \sigma_{T_2}^2$ are given by

$$\begin{aligned} L &= \left(S_{A_1}^2 + S_{E_1}^2 + S_{A_2}^2 + S_{E_2}^2\right) - \sqrt{B_1 + C_1 + B_2 + C_2} \\ U &= \left(S_{A_1}^2 + S_{E_1}^2 + S_{A_2}^2 + S_{E_2}^2\right) + \sqrt{D_1 + E_1 + D_2 + E_2}. \end{aligned}$$

The limits of the denominator of the standardized mean difference, $\sqrt{\sigma_{T_1}^2 + \sigma_{T_2}^2}$, are then set to $(\sqrt{L}, \sqrt{U})$ using the transformation principle. This interval, denoted here as (l_2, u_2), and the interval for the numerator (the difference between two normal means), denoted by (l_1, u_1), may be used to find the $(1-\alpha)100\%$ confidence interval of the standardized mean difference with Equation (2.18),

$$\begin{cases} L_\delta = \dfrac{\hat{\theta}_1\hat{\theta}_2 - \sqrt{\hat{\theta}_1^2\hat{\theta}_2^2 - u_2 l_1(2\theta_2 - u_2)(2\hat{\theta}_1 - l_1)}}{u_2(2\hat{\theta}_2 - u_2)} \\ U_\delta = \dfrac{\hat{\theta}_1\hat{\theta}_2 + \sqrt{\hat{\theta}_1^2\hat{\theta}_2^2 - l_2 u_1(2\hat{\theta}_2 - l_2)(2\hat{\theta}_1 - u_1)}}{l_2(2\hat{\theta}_2 - l_2)}, \end{cases} \tag{3.16}$$

where $\hat{\theta}_1 = \bar{\bar{Y}}_1 - \bar{\bar{Y}}_2$ and $\hat{\theta}_2 = S_T = \sqrt{S_{A_1}^2 + S_{E_1}^2 + S_{A_2}^2 + S_{E_2}^2}$.

According to the transformation principle, the $(1-\alpha)100\%$ confidence interval for $\mathrm{P}(Y_1 > Y_2)$ is then given by $(\Phi(L_\delta), \Phi(U_\delta))$. This confidence interval is not restricted to symmetry because the intervals of the denominator are asymmetric.

Alternative confidence intervals

Wald confidence interval and the delta method

Symmetric Wald-type confidence intervals for the standardized mean difference may be constructed using the multivariate delta method (Section 2.4.2). The standardized

mean difference may be expressed as

$$\begin{aligned}\delta &= \frac{\mu_1 - \mu_2}{\sqrt{\sigma^2_{T_1} + \sigma^2_{T_2}}} \\ &= \frac{\theta_1 - \theta_2}{\sqrt{\theta_3 + \theta_4 + \theta_5 + \theta_6}},\end{aligned}$$

where the estimates of

$$\begin{aligned}\theta_1 &= \mu_1 \\ \theta_2 &= \mu_2 \\ \theta_3 &= \frac{n_{H1} - 1}{n_{H1}}\sigma^2_{E_1} \\ \theta_4 &= \frac{\sigma^2_{E_1} + n_{H1}\sigma^2_{A_1}}{n_{H1}} \\ \theta_5 &= \frac{n_{H2} - 1}{n_{H2}}\sigma^2_{E_2} \\ \theta_6 &= \frac{\sigma^2_{E_2} + n_{H2}\sigma^2_{A_2}}{n_{H2}}\end{aligned}$$

have sample variances

$$\begin{aligned}\mathrm{var}(\hat{\theta}_1) &= \frac{S^2_{U1}}{k_1 n_{H1}} \\ \mathrm{var}(\hat{\theta}_2) &= \frac{S^2_{U2}}{k_2 n_{H2}} \\ \mathrm{var}(\hat{\theta}_3) &= \frac{2(n_{H1} - 1)^2\sigma^4_{E_1}}{n^2_{H1}(M_1 - k_1)} \\ \mathrm{var}(\hat{\theta}_4) &= \frac{2S^2_{U1}}{n_{H1}(k_1 - 1)} \\ \mathrm{var}(\hat{\theta}_5) &= \frac{2(n_{H2} - 1)^2\sigma^4_{E_2}}{n^2_{H2}(M_2 - k_2)} \\ \mathrm{var}(\hat{\theta}_6) &= \frac{2S^2_{U2}}{n_{H2}(k_2 - 1)},\end{aligned}$$

respectively. Application of the multivariate delta method gives the variance of the estimated standardized mean difference,

$$\widehat{\mathrm{var}}(\hat{\delta}) = O + P + Q + R + S + T,$$

where

$$
\begin{aligned}
O &= \frac{S_{U1}^2}{\left[\frac{(n_{H1}-1)}{n_{H1}}\sigma_{E_1}^2 + \frac{S_{U1}^2}{n_{H1}} + \frac{(n_{H2}-1)}{n_{H2}}\sigma_{E_2}^2 + \frac{S_{U2}^2}{n_{H2}}\right] k_1 n_{H1}} \\
P &= \frac{S_{U2}^2}{\left[\frac{(n_{H1}-1)}{n_{H1}}\sigma_{E_1}^2 + \frac{S_{U1}^2}{n_{H1}} + \frac{(n_{H2}-1)}{n_{H2}}\sigma_{E_2}^2 + \frac{S_{U2}^2}{n_{H2}}\right] k_2 n_{H2}} \\
Q &= \frac{(\mu_1-\mu_2)^2(n_{H1-1})^2\sigma_{E_1}^4}{2\left[\frac{(n_{H1}-1)}{n_{H1}}\sigma_{E_1}^2 + \frac{S_{U1}^2}{n_{H1}} + \frac{(n_{H2}-1)}{n_{H2}}\sigma_{E_2}^2 + \frac{S_{U2}^2}{n_{H2}}\right]^3 n_{H1}^2(M_1-k_1)} \\
R &= \frac{(\mu_1-\mu_2)^2 S_{U1}^4}{2\left[\frac{(n_{H1}-1)}{n_{H1}}\sigma_{E_1}^2 + \frac{S_{U1}^2}{n_{H1}} + \frac{(n_{H2}-1)}{n_{H2}}\sigma_{E_2}^2 + \frac{S_{U2}^2}{n_{H2}}\right]^3 n_{H1}^2(k_1-1)} \\
S &= \frac{(\mu_1-\mu_2)^2(n_{H2-1})^2\sigma_{E_2}^4}{2\left[\frac{(n_{H1}-1)}{n_{H1}}\sigma_{E_1}^2 + \frac{S_{U1}^2}{n_{H1}} + \frac{(n_{H2}-1)}{n_{H2}}\sigma_{E_2}^2 + \frac{S_{U2}^2}{n_{H2}}\right]^3 n_{H2}^2(M_2-k_2)} \\
T &= \frac{(\mu_1-\mu_2)^2 S_{U2}^4}{2\left[\frac{(n_{H1}-1)}{n_{H1}}\sigma_{E_1}^2 + \frac{S_{U1}^2}{n_{H1}} + \frac{(n_{H2}-1)}{n_{H2}}\sigma_{E_2}^2 + \frac{S_{U2}^2}{n_{H2}}\right]^3 n_{H2}^2(k_2-1)}.
\end{aligned}
$$

Plugging in the variance estimate using Slutsky's theorem (Casella and Berger, 2002, page 239), Wald-type confidence intervals for the standardized mean difference are obtained,

$$
\begin{cases}
L = \hat{\delta} - z_{\alpha/2}\sqrt{\widehat{\text{var}}(\hat{\delta})} \\
U = \hat{\delta} + z_{\alpha/2}\sqrt{\widehat{\text{var}}(\hat{\delta})}.
\end{cases}
\tag{3.17}
$$

Using the transformation principle (Section 2.3.2), the confidence interval for $\text{P}(Y_1 > Y_2)$ is then given by $(\Phi(L), \Phi(U))$. Although the standardized mean difference can be skewed in distribution and thus so can $\text{P}(Y_1 > Y_2)$, these limits are symmetric around the point estimate, potentially leading to unbalanced tail errors.

Generalized confidence interval

Generalized confidence intervals may be constructed for the standardized mean difference using a generalized pivotal quantity for the standardize mean difference. This

statistic may be obtained using the generalized pivotal quantity for μ_i (Equation (3.5)), $\sigma^2_{A_i}$ (Equation (3.13)), and $\sigma^2_{E_i}$ (Equation (3.14)).

The generalized pivotal quantity for the standardized mean difference may be expressed as

$$G_{SMD} = \frac{G_{\mu_1 1} - G_{\mu_2}}{\sqrt{G_{\sigma^2_{A_1}} + G_{\sigma^2_{E_1}} + G_{\sigma^2_{A_2}} + G_{\sigma^2_{E_2}}}}. \tag{3.18}$$

To obtain generalized confidence intervals, the algorithm in Section 3.1.2.3 may be applied while changing step 2 to compute G_{μ_i}, $G_{\sigma^2_{A_i}}$, $G_{\sigma^2_{E_i}}$, and G_{SMD} instead of $G_{\mu_1} - G_{\mu_2}$. Letting (L, U) represent the $(1-\alpha)100\%$ confidence interval of the standardized mean difference, the generalized confidence interval for $\mathrm{P}(Y_1 > Y_2)$ is given by $[\Phi(L), \Phi(U)]$.

Chapter 4

SIMULATION STUDY OF CONFIDENCE INTERVAL PROCEDURES

4.1 Introduction

The discussion in Chapter 3 was based on an algebraic comparison of various existing confidence interval procedures as compared to the MOVER for each of the three effect measures, 1) the difference between two normal means, 2) the difference between two lognormal means, and 3) the exceedance probability. The validity of each procedure is based on large sample theory. Consequently, we expect coverage to approach nominal levels as the number of clusters gets large. It is therefore important to evaluate the methods to identify conditions under which the procedures perform well, in addition to when and how they fail.

In this chapter, the confidence interval procedures shown in Chapter 3 are compared under a variety of common parameter combinations using Monte Carlo simulations, following three steps. First, data are generated according to chosen parameter values to resemble random variables found in practice. Second, the generated data are analyzed using all methods considered, and third the analysis is evaluated by comparing estimates with the truth (the known parameter values). Empirical coverage rates, tail errors and median widths are used to identify the finite sample properties of procedures.

4.2 Objectives

In general, a simulation study allows the finite properties of asymptotic statistics to be examined and various methods to be compared. This simulation study evaluates confidence intervals for a difference between two normal means, the difference between two lognormal means, and the exceedance probability when data arise from cluster randomization trials. The simulation study evaluates these confidence intervals with three main objectives:

1) to determine the parameters required per trial arm to maintain confidence internal coverage rates, which indicate the procedure's overall validity,
2) to compare confidence interval tail error rates, which show the validity of each confidence limit, and
3) to compare confidence interval widths, which reflect efficiency.

Achieving these objectives will allow recommendations to be made for confidence interval procedure as applied to cluster randomization trials with normal or lognormal outcomes.

For all three objectives 95% confidence intervals will be investigated. Empirical coverage is set to the percentage of times the confidence intervals obtained from the generated data contain the true parameter value, where the aim is to have the empirical coverage fall close to the nominal coverage of 95%. With 1000 simulation runs, we regard empirical coverage between $(93.6\%, 96.4\%)$ as acceptable. This range was determined by $0.95 \pm 1.96\sqrt{(0.95)(0.05)/1000}$. Note that this confidence interval for the nominal coverage is symmetric, because it is expected that 1000 replicate data sets will likely lead to a symmetric sampling distribution for the coverage.

If empirical coverage falls within the interval given above, left and right tail errors will then be compared. Left and right tail errors are the percentage of times the confidence interval lies above or below the parameter value. We emphasize tail errors because we define a 95% confidence interval for a parameter θ as given by (L, U),

where $P(L \geq \theta) = 0.025$ and $P(U \leq \theta) = 0.025$. On average, tail errors demonstrate how often the interval misses the truth from each side. Also, balance between tail errors is desirable to ensure that only extreme values are excluded from the interval. Ideally, the interval should miss the true parameter value 2.5% of the time from each side for a 95% confidence interval. If tail errors are not balanced, then the interval may exclude likely parameter values from the side with the larger tail and may include extreme values which are unlikely to be true from the side with the smaller tail.

If more than one procedure has acceptable coverage and comparable tail errors, the procedures will be compared based on confidence interval widths. The median confidence interval width will be compared from the 1000 runs of each procedure, where a narrower width, indicating greater precision, is desirable. Median widths are of interest rather than mean widths, because the sampling distribution of confidence interval width is skewed for the confidence intervals of a difference between two lognormal means.

4.3 Methods

Parameter combinations

Many parameter values were varied within each simulation study according to previously published cluster randomization trials to make recommendations about the procedures and to learn about their shortcomings. For simplicity, the design of the experiment was a completely randomized cluster randomization trial with two arms. Extensions to other clustered trial designs and to more than two arms may be investigated in future work. The parameters which varied within each simulation study included the number of clusters in each arm, the average cluster size, the ratio of variances in the two arms, and the ICC.

Many cluster randomization trials occur at the practice or even the community level, demonstrating a need for statistical inference procedures when there are a small

number of large clusters. Such methods would also be applicable when there are a large number of small clusters, because increasing the number of clusters (k) typically contributes more to the precision of the parameter estimates than increasing the cluster size (m), as can be seen from the variance expression in Donner and Klar (2000, page 8) for a single normal mean,

$$\text{var}(\bar{Y}_i) = \frac{\sigma^2}{km}[1 + (m-1)\rho],$$

when σ^2 is the unknown variance of Y and ρ is the ICC. Thus, a balanced and unbalanced number of clusters in the two arms (control, experimental) will be set to $(6, 6)$, $(12, 6)$, $(12, 12)$, $(24, 12)$, and $(24, 24)$ with an average of 50, 100, and 200 observations per cluster. These parameters are consistent with the simulation study performed by Flynn and Peters (2004), where the performance of the Huber-White robust variance estimator (Huber, 1981; White, 1980) was shown to have closer coverage to the nominal than bias-corrected and accelerated bootstrap confidence intervals (Efron, 1987) for normal and lognormal data from cluster randomization trials. The above sample sizes often occur in community randomized trials and trials randomizing physician practices (Feng *et al.*, 1996; Donner and Klar, 1996) and will therefore be used in this simulation study.

Balanced cluster sizes rarely occur in practice. The imbalance of cluster sizes may be described by the imbalance parameter,

$$v = \frac{1}{1 + k^2}, \tag{4.1}$$

where $k = \sigma/\mu$ denotes coefficient of variation (Ahrens and Pincus, 1981), which is a normalized measure of the dispersion of data points around the mean, σ is the standard deviation of the cluster sizes, and μ is the mean cluster size. Imbalance parameters range from 0 to 1, with 1 denoting complete balance between cluster sizes. Cluster sizes of 50, often seen in general practices (Eldridge *et al.*, 2004), typically have an imbalance parameter of roughly 0.8 (Eldridge *et al.*, 2006). Several trials (e.g.

Marrie *et al.*, 2000; Burns and Kendrick, 1997) and previous simulation studies (e.g. Zou, 2002; Klar, 1993; Donner *et al.*, 1994) examined had estimates of the imbalance parameter approximately equal to 0.8. Thus, cluster sizes will be generated from the uniform distribution with the imbalance parameter set to 0.8. A detailed description is provided in Section 4.3.2.1.

Individuals within any one of these clusters typically have a positive correlation with others in the same cluster. The degree of their similarity as compared to individuals in other clusters for some outcome may be measured using the ICC. The ICC is important because it is used to estimate the design effect (Chapter 1), which may then be used for the proper design and analysis of cluster randomization trials when interest lies at the individual level. Although negative values of the ICC are theoretically possible, we limit the discussion to positive values in the context of cluster randomization trials. Typical ICC values for the sample sizes above range from 0.005 to 0.2 (Hedges, 2007*a*; Donner and Klar, 2004; Feng *et al.*, 1996). The simulation studies will therefore investigate data with ICC values of 0.005, 0.01, 0.1, and 0.2. Note that the ICC value for lognormal data is not of direct interest as it is common practice for investigators to transform the data onto the log scale. Also, this thesis is not focused on the direct interpretation of the ICC. Rather, the ICC is used here to potentially correct for the effect of clustering when inferences are at the individual level, or to quantify the effect of clustering on the raw or log scale.

Without loss of generality, the normal mean was set to $\mu_T = 1.0$ for the experimental arm and to $\mu_C = 0$ for the control arm for the difference between two normal means and the difference between two lognormal means according to the simulation study performed by Flynn and Peters (2004).

Although variance homogeneity may be assumed in hypothesis testing under the null, this assumption may not hold for confidence interval construction. Heteroscedasticity may arise in practice in the presence of an intervention effect, where the intervention may also have an effect on the sample variance of the outcome in the

intervention arm. The effects of heteroscedasticity are therefore investigated by the simulation study. Following the notation in Section 1.3, the ratio of variances (experimental arm to control arm, $\sigma^2_{T_i}/\sigma^2_{C_i}$) is set to 1.0 and 1.4, while keeping the variance of the control arm constant at 5.0 units^2. Note that the maximum ratio of normal variances was not set larger than 1.4 particularly due to the exponentiated lognormal data. A larger variance of roughly exp(10) is not realistic or of practical value.

For the exceedance probability, $\mathrm{P}(Y_1 > Y_2) = \Phi(\mathrm{SMD})$, the values of μ_T and μ_C are altered such that $\mathrm{P}(Y_1 > Y_2) = 0.5$ and 0.9. These values are changed to $\mu_T = 0$ and $\mu_C = 0$ for $\mathrm{P}(Y_1 > Y_2) = 0.5$. When $\mathrm{P}(Y_1 > Y_2) = 0.9$, $\mu_T = 4.0$ and $\mu_C = 0$ when $\sigma^2_{T_1}/\sigma^2_{T_2} = 1.0$ ($\sigma^2_{T_2} = 5$), and $\mu_T = 4.4$ and $\mu_C = 0$ when $\sigma^2_{T_1}/\sigma^2_{T_2} = 1.4$ ($\sigma^2_{T_2} = 5$).

A summary of the parameters investigated in the simulation study is given in Table 4.1. A factorial design is followed with a total of 120 parameter combinations for the difference between two normal means and the difference between two lognormal means, and 240 parameter combinations for the exceedance probability, $\mathrm{P}(Y_1 > Y_2)$.

Burton *et al.* (2006) defines two types of datasets generated in simulation studies - fully independent datasets and moderately independent datasets. Fully independent datasets are defined as different sets of independent datasets for each method and each parameter combination in each of the 1000 runs, while moderately independent datasets are defined as the same simulated dataset for each method within a scenario, but different and independent datasets for different scenarios (or different parameter combinations). This study uses moderately independent simulations to more easily detect any differences between the procedures.

Data generation

Cluster sizes

Cluster sizes were simulated using the discrete uniform distribution such that the average cluster size and degree of imbalance may easily be controlled. Existing trials

Table 4.1: Parameter combinations used for Monte Carlo simulations

Parameter	value
Runs per parameter combination	1000
α	0.05
Clusters/arm (Control, Experimental)	(6,6), (12,6), (12,12), (24,12), (24,24)
Average cluster size	50, 100, 200
ICC	0.005, 0.01, 0.1, 0.2
Imbalance parameter (v)	0.8
$\mu_2 - \mu_1$, ($\mu_1 = 0$)	1
$\sigma^2_{T_1}/\sigma^2_{T_2}$, ($\sigma^2_{T_2} = 5.0$)	1.0, 1.4
P($Y_1 > Y_2$), when $\sigma^2_{T_1}/\sigma^2_{T_2} = 1.0, 1.4$ ($\sigma^2_{T_2} = 5.0$)	0.5, 0.9

typically have imbalance parameters of 0.8 (Eldridge *et al.*, 2006; Marrie *et al.*, 2000; Burns and Kendrick, 1997) and previous simulation studies have generated clustered data using an imbalance parameter of 0.8 for cluster sizes (e.g. Zou, 2002; Klar, 1993; Donner *et al.*, 1994). Therefore, an imbalance parameter of 0.8 and average cluster sizes of 50, 100, and 200 individuals were generated. These cluster sizes were chosen to reflect typical sizes in existing trials and past simulation studies (Eldridge *et al.*, 2006; Flynn and Peters, 2004).

To sample data from the uniform distribution, the mean and variance are first required. Equation (4.1) may be used by setting the value of v to 0.8 for unbalanced cluster sizes and solving for k ($k = 0.5$). Using the desired mean (50, 100, or 200), the standard deviation of the uniform distribution may be obtained using the expression for the coefficient of variation, $\text{CV} = k = \sigma/\mu$ (Eldridge *et al.*, 2006). The variance of the discrete uniform distribution is given by

$$\sigma^2 = \frac{(b-a+1)^2 - 1}{12},$$

where a and b are the endpoints. The width of the uniform distribution is then given by $b - a + 1$. By using the variance expression to solve for the width, the endpoints of the uniform distribution may be expressed as

$$\left(\mu - \frac{b-a+1}{2}, \mu + \frac{b-a+1}{2}\right). \tag{4.2}$$

Table 4.2 gives the end points of each uniform distribution for each average cluster size when $v = 0.8$.

Correlated normal data

Once the cluster sizes have been determined, the observations must be generated for the simulation study. As discussed in Chapter 1, an extensively used distribution for the generation of data is the normal distribution because it commonly approximates many types of data found in practice, including continuous and relatively symmetric

Table 4.2: Imbalance parameter and the corresponding endpoints of the discrete uniform distribution used to sample unbalanced cluster sizes ($v = 0.8$)

Average cluster size	Endpoints
50	(7, 93)
100	(13, 187)
200	(27, 373)

health-related data such as blood pressure and weight. The normal distribution was used to generate observations for the simulation study when interest lay in inferences on a difference between two normal means and the exceedance probability.

Data were generated according to the one-way random effects model. Specifically, the l^{th} observation, Y_{ijl}, from the j^{th} cluster in the i^{th} arm is given by

$$Y_{ijl} = \mu_i + A_{ij} + E_{ijl}$$

where μ_i is the population mean of arm i, $A_{ij} \sim N(0, \sigma^2_{A_i})$ is independent of $E_{ijl} \sim N(0, \sigma^2_{E_i})$, and two observations within a cluster have correlation ρ (ρ is the value of the ICC).

Correlated lognormal data

A multivariate lognormal distribution may be used to approximate positively skewed data which commonly occur in cluster randomization trials with outcomes such as hospital wait times and health care costs (see Chapter 1). Following the notation in Section 1.3, multivariate lognormal data were generated by exponentiating the multivariate normal (MVN) observations. That is,

$$X_{ijl} = \exp(Y_{ijl})$$

where $Y_{ijl} \sim \text{MVN}(\mu, \mathbf{\Sigma})$ with coerrelation coefficient ρ.

Computer software for data generation

All data were generated using SAS (Statistical Analysis Systems) software. For instance, cluster sizes were then sampled in SAS IML using the `UNIFORM` function using the endpoints specified in Table 4.2. Also, correlated normal data were generated in Proc IML in SAS with the `NORMAL` function. Correlated lognormal data were generated by simply exponentiating the correlated normal observations.

Methods of comparison

The algebraic expressions of the methods of comparison for each of the parameters of interest are given in Chapter 3. These methods are organized in Table 4.3 for each of the three parameters.

For the difference between two normal means, the MOVER is evaluated and compared to the Wald method, the cluster-adjusted confidence interval procedure (Donner and Klar, 1993), and the generalized confidence interval procedure (Weerahandi, 1993; Krishnamoorthy *et al.*, 2007).

For the difference between two lognormal means, the MOVER is evaluated and compared to the Wald method, and the generalized confidence interval procedure (Weerahandi, 1993; Krishnamoorthy *et al.*, 2007).

For the exceedance probability, the MOVER is evaluated and compared to the Wald method, and the generalized confidence interval procedure (Weerahandi, 1993; Krishnamoorthy *et al.*, 2007). Evaluations and comparisons are consistent with those described in Section 4.2.

All of these confidence interval procedures adjust for clustering and heteroscedasticity. However, only the MOVER and the generalized confidence interval procedure allow asymmetric limits around the parameter estimate when its sampling distribution is skewed.

Table 4.3: Methods of comparison for the difference between two normal means ($\mathrm{E}(Y_1) - \mathrm{E}(Y_2)$), the difference between two lognormal means ($\mathrm{E}(X_1) - \mathrm{E}(X_2)$), and the exceedance probability ($\mathrm{P}(Y_1 > Y_2)$).

$\mathrm{E}(Y_1) - \mathrm{E}(Y_2)$	$\mathrm{E}(X_1) - \mathrm{E}(X_2)$	$\mathrm{P}(Y_1 > Y_2)$
MOVER, Equation (3.2)	MOVER, Equation (3.11)	MOVER, Equation (3.16)
Wald, Equation (3.3)	Wald, Equation (3.12)	Wald, Equation (3.17)
Cluster-adjusted, Equation (3.4)	GCI, Equation (3.15)	GCI, Equation (3.18)
GCI, Equation (3.5)		

4.4 Results

The simulation results for each of the parameters of interest are presented in tabular form, each table differing by the number of clusters per arm. This presentation design was chosen because the number of clusters appeared to have the greatest impact on empirical coverage for each of the three parameters investigated. Also, each table displays the results of all three objectives of the study (Section 4.2). For each parameter investigated, empirical coverage is first discussed, followed by balance between tail errors, and finally median interval width.

The difference between two normal means

Empirical coverage results ($\alpha = 0.05$), tail errors, and median widths for the Wald method, the cluster-adjusted confidence interval procedure, the generalized confidence interval procedure and the MOVER as applied to unbalanced, completely randomized cluster randomization trials for a difference between two normal means are presented in Tables 4.4 to 4.8.

Confidence interval coverage

Overall, all of the procedures, except the cluster-adjusted confidence interval (Donner and Klar, 1993), show great improvement in empirical coverage as the number of clusters per arm increase. Other parameters such as the average cluster size, the ICC, and variance homogeneity/heterogeneity do not greatly influence coverage rates.

When the number of clusters per arm is small, e.g. at least one arm with 6 clusters, the Wald method has low coverage rates, with an average empirical coverage of 91.8% in Tables 4.4 and 4.5. As the number of clusters increases to 24 (Table 4.8) the method's performance improves greatly to coverage rates of almost all parameter combinations falling within the desired range, i.e. 93.6% to 96.4%.

The cluster-adjusted confidence interval procedure shows consistently high coverage throughout all of the parameter combinations (Tables 4.4-4.8), showing only very slight improvements as the number of clusters increase. An average empirical coverage of 99.4% (nominal coverage of 95%) was obtained for the 120 parameter combinations investigated due to overestimated variances.

The generalized confidence interval procedure performed reasonably well overall, falling within the desired range 73.6% of the time. The other 26.4% of the time, the procedure had an average empirical coverage of 97%. With only 6 clusters per arm, the method showed coverage outside the desired range 83.3% of the time.

The MOVER had similar empirical coverage performance to the simulation intensive generalized confidence interval procedure. When there are only 6 clusters per group the method has coverage outside the desired range 75% of the time, exceeding the nominal coverage by an average of 2.2%. However, when the number of clusters increases to (12, 6) (Table 4.5), coverage rates fall closer to the nominal 95%. This improvement continues as the number of clusters increase to 24 clusters per arm (Table 4.8).

Table 4.4: Empirical coverage (%), tail errors ((<, >)%), and median widths (WD) for the difference between two normal means when the number of clusters per arm equal 6 (control) and 6 (experimental), $\alpha = 0.05$, imbalance parameter=0.8, and runs=1000

ICC	m_{ij}	$\frac{\sigma^2_{T_1}}{\sigma^2_{T_2}}$	Wald	ClusterAdj	GCI	MOVER
			Cov, (<, >)%, WD	Cov (<, >)% WD	Cov (<, >)% Width	Cov (<, >)% WD
0.005	50	1.0	92.6 (3.4, 4.0) 0.9	99.9 (0.1, 0.0) 1.6	97.6 (0.9, 1.5) 1.2	97.5 (1.0, 1.5) 1.2
		1.4	92.8 (3.4, 3.8) 1.0	99.9 (0.1, 0.0) 1.8	97.3 (1.0, 1.7) 1.3	97.8 (0.9, 1.3) 1.3
	100	1.0	91.9 (3.8, 4.3) 0.7	99.7 (0.1, 0.2) 1.2	96.8 (0.9, 2.3) 0.9	96.8 (1.2, 2.0) 0.9
		1.4	91.9 (3.9, 4.2) 0.7	99.7 (0.1, 0.2) 1.4	97.3 (0.9, 1.8) 1.0	96.9 (1.2, 1.9) 1.0
	200	1.0	91.8 (4.5, 3.7) 0.5	100.0 (0.0, 0.0) 1.0	97.4 (1.2, 1.4) 0.7	97.7 (1.1, 1.2) 0.7
		1.4	91.0 (4.8, 4.2) 0.6	100.0 (0.0, 0.0) 1.1	97.6 (1.0, 1.4) 0.8	97.7 (1.1, 1.2) 0.8
0.01	50	1.0	92.1 (3.9, 4.0) 1.0	99.9 (0.1, 0.0) 1.8	97.3 (1.0, 1.7) 1.3	97.5 (1.0, 1.5) 1.3
		1.4	92.1 (3.6, 4.3) 1.1	99.9 (0.1, 0.0) 1.9	97.6 (0.9, 1.5) 1.4	97.7 (0.9, 1.4) 1.4
	100	1.0	91.3 (4.2, 4.5) 0.8	99.6 (0.2, 0.2) 1.4	97.0 (1.1, 1.9) 1.0	96.8 (1.3, 1.9) 1.0
		1.4	91.5 (4.0, 4.5) 0.8	99.6 (0.2, 0.2) 1.6	97.0 (1.2, 1.8) 1.1	96.8 (1.3, 1.9) 1.1
	200	1.0	91.5 (4.6, 3.9) 0.6	99.9 (0.1, 0.0) 1.2	97.1 (1.1, 1.8) 0.8	96.9 (1.3, 1.8) 0.8
		1.4	91.2 (4.5, 4.3) 0.7	99.7 (0.2, 0.1) 1.3	97.1 (1.2, 1.7) 0.9	97.1 (1.2, 1.7) 0.9
0.1	50	1.0	91.7 (3.9, 4.4) 1.8	99.5 (0.1, 0.4) 3.5	96.6 (1.4, 2.0) 2.3	96.6 (1.5, 1.9) 2.3
		1.4	91.7 (3.8, 4.5) 1.9	99.5 (0.1, 0.4) 3.8	96.5 (1.4, 2.1) 2.5	96.4 (1.6, 2.0) 2.5
	100	1.0	91.3 (4.1, 4.6) 1.6	99.7 (0.2, 0.1) 3.3	95.7 (2.2, 2.1) 2.1	96.1 (2.1, 1.8) 2.2
		1.4	91.6 (3.9, 4.5) 1.8	99.7 (0.2, 0.1) 3.6	95.9 (2.0, 2.1) 2.3	95.8 (2.2, 2.0) 2.4
	200	1.0	92.2 (3.7, 4.1) 1.6	99.6 (0.3, 0.1) 3.2	97.1 (1.3, 1.6) 2.1	97.1 (1.4, 1.5) 2.1
		1.4	91.8 (4.0, 4.2) 1.7	99.6 (0.3, 0.1) 3.6	96.8 (1.5, 1.7) 2.3	97.2 (1.4, 1.4) 2.3
0.2	50	1.0	92.7 (3.5, 3.8) 2.4	99.6 (0.1, 0.3) 4.7	96.6 (1.7, 1.7) 3.1	96.8 (1.7, 1.5) 3.1
		1.4	91.8 (4.1, 4.1) 2.6	99.5 (0.2, 0.3) 5.1	96.6 (1.7, 1.7) 3.4	96.4 (1.8, 1.8) 3.4
	100	1.0	91.3 (4.1, 4.6) 2.2	99.6 (0.2, 0.2) 4.5	95.6 (2.1, 2.3) 2.9	95.8 (2.0, 2.2) 2.9
		1.4	91.2 (3.9, 4.9) 2.5	99.7 (0.2, 0.1) 5.0	95.8 (1.9, 2.3) 3.2	95.8 (2.1, 2.1), 3.2
	200	1.0	91.9 (4.0, 4.1) 2.2	99.8 (0.1, 0.1) 4.5	97.2 (1.3, 1.5) 2.9	97.1 (1.5, 1.4) 2.9
		1.4	91.7 (4.2, 4.1) 2.4	99.7 (0.2, 0.1) 4.9	97.0 (1.3, 1.7) 3.2	97.3 (1.3, 1.4) 3.2

Table 4.5: Empirical coverage (%), tail errors ((<, >)%), and median widths (WD) for the difference between two normal means when the number of clusters per arm equal 12 (control) and 6 (experimental), $\alpha = 0.05$, imbalance parameter=0.8, and runs=1000

ICC	m_{ij}	$\frac{\sigma^2_{T_1}}{\sigma^2_{T_2}}$	Wald	ClusterAdj	GCI	MOVER
			Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% Width	Cov (<, >)% WD
0.005	50	1.0	92.7 (3.4, 3.9) 0.8	99.6 (0.1, 0.3) 1.2	97.5 (1.2, 1.3) 1.0	97.3 (1.4, 1.3) 1.0
		1.4	92.2 (3.7, 4.1) 0.9	99.5 (0.1, 0.4) 1.3	97.0 (1.3, 1.7) 1.1	97.3 (1.3, 1.4) 1.1
	100	1.0	92.9 (4.2, 2.9) 0.6	99.5 (0.3, 0.2) 0.9	96.2 (2.4, 1.4) 0.7	96.9 (2.1, 1.0) 0.7
		1.4	92.7 (4.4, 2.9) 0.7	99.5 (0.3, 0.2) 1.0	96.2 (2.4, 1.4) 0.8	96.7 (2.1, 1.2) 0.8
	200	1.0	92.0 (3.6, 4.4) 0.5	99.3 (0.4, 0.3) 0.7	95.6 (1.5, 2.9) 0.6	96.0 (1.6, 2.4) 0.6
		1.4	91.6 (3.7, 4.7) 0.5	98.8 (0.5, 0.7) 0.8	95.4 (1.5, 3.1) 0.6	95.7 (1.5, 2.8) 0.6
0.01	50	1.0	91.9 (3.9, 4.2) 0.8	99.3 (0.2, 0.5) 1.3	96.6 (1.6, 1.8) 1.0	96.6 (1.7, 1.7) 1.0
		1.4	91.6 (4.1, 4.3) 0.9	99.2 (0.2, 0.6) 1.4	96.3 (1.9, 1.8) 1.2	96.3 (1.9, 1.8) 1.2
	100	1.0	92.1 (4.5, 3.4) 0.7	99.3 (0.5, 0.2) 1.0	95.9 (2.6, 1.5) 0.8	96.7 (2.1, 1.2) 0.8
		1.4	92.0 (4.5, 3.5) 0.7	99.3 (0.5, 0.2) 1.1	96.1 (2.4, 1.5) 0.9	96.3 (2.3, 1.4) 0.9
	200	1.0	92.0 (3.7, 4.3) 0.5	99.1 (0.5, 0.4) 0.9	95.7 (1.6, 2.7) 0.7	95.8 (1.7, 2.5) 0.7
		1.4	90.9 (4.2, 4.9) 0.6	98.9 (0.6, 0.5) 1.0	95.3 (1.6, 3.1) 0.8	95.8 (1.7, 2.5) 0.8
0.1	50	1.0	91.9 (3.9, 4.2) 0.8	99.3 (0.2, 0.5) 1.3	96.6 (1.6, 1.8) 1.0	96.6 (1.7, 1.7) 1.0
		1.4	91.6 (4.1, 4.3) 0.9	99.2 (0.2, 0.6) 1.4	96.3 (1.9, 1.8) 1.2	96.3 (1.9, 1.8) 1.2
	100	1.0	92.1 (4.5, 3.4) 0.7	99.3 (0.5, 0.2) 1.0	95.9 (2.6, 1.5) 0.8	96.7 (2.1, 1.2) 0.8
		1.4	92.0 (4.5, 3.5) 0.7	99.3 (0.5, 0.2) 1.1	96.1 (2.4, 1.5) 0.9	96.3 (2.3, 1.4) 0.9
	200	1.0	92.0 (3.7, 4.3) 0.5	99.1 (0.5, 0.4) 0.9	95.7 (1.6, 2.7) 0.7	95.8 (1.7, 2.5) 0.7
		1.4	90.9 (4.2, 4.9) 0.6	98.9 (0.6, 0.5) 1.0	95.3 (1.6, 3.1) 0.8	95.8 (1.7, 2.5) 0.8
0.2	50	1.0	91.3 (5.2, 3.5) 2.0	99.1 (0.6, 0.3) 3.4	95.4 (2.6, 2.0) 2.4	95.4 (2.6, 2.0) 2.5
		1.4	90.5 (5.4, 4.1) 2.2	99.1 (0.6, 0.3) 3.8	95.1 (2.8, 2.1) 2.8	95.2 (2.9, 1.9) 2.8
	100	1.0	91.7 (4.9, 3.4) 1.9	98.5 (1.0, 0.5) 3.4	96.1 (2.2, 1.7) 2.4	95.9 (2.3, 1.8) 2.4
		1.4	91.0 (5.6, 3.4) 2.2	98.4 (1.1, 0.5) 3.7	96.1 (2.1, 1.8) 2.7	95.8 (2.4, 1.8) 2.7
	200	1.0	92.2 (4.0, 3.8) 1.9	99.2 (0.2, 0.6) 3.3	96.7 (1.7, 1.6) 2.3	96.7 (1.9, 1.4) 2.4
		1.4	91.4 (4.2, 4.4) 2.1	99.1 (0.3, 0.6) 3.7	96.7 (1.8, 1.5) 2.6	96.6 (1.9, 1.5) 2.7

Table 4.6: Empirical coverage (%), tail errors ((<, >)%), and median widths (WD) for the difference between two normal means when the number of clusters per arm equal 12 (control) and 12 (experimental), $\alpha = 0.05$, imbalance parameter=0.8, and runs=1000

ICC	m_{ij}	$\frac{\sigma^2_{T_1}}{\sigma^2_{T_2}}$	Wald	ClusterAdj	GCI	MOVER
			Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% Width	Cov (<, >)% WD
0.005	50	1.0	93.9 (3.2, 2.9) 0.7	99.5 (0.3, 0.2) 1.0	95.5 (2.4, 2.1) 0.7	96.1 (2.2, 1.7) 0.7
		1.4	93.4 (3.7, 2.9) 0.7	99.6 (0.4, 0.0) 1.1	95.7 (2.4, 1.9) 0.8	96.0 (2.3, 1.7) 0.8
	100	1.0	92.6 (3.7, 3.7) 0.5	99.5 (0.2, 0.3) 0.7	95.6 (2.1, 2.3) 0.5	95.4 (2.1, 2.5) 0.6
		1.4	92.8 (3.4, 3.8) 0.5	99.3 (0.3, 0.4) 0.8	95.2 (2.2, 2.6) 0.6	95.1 (2.4, 2.5) 0.6
	200	1.0	93.6 (3.3 , 3.1) 0.4	99.4 (0.4, 0.2) 0.6	96.6 (1.5, 1.9) 0.4	96.7 (1.8, 1.5) 0.4
		1.4	93.3 (3.4, 3.3) 0.4	99.6 (0.2, 0.2) 0.7	96.1 (1.7, 2.2) 0.5	96.5 (1.8, 1.7) 0.5
0.01	50	1.0	93.7 (3.4, 2.9) 0.7	99.4 (0.5, 0.1) 1.0	95.9 (2.4, 1.7) 0.8	96.3 (2.2, 1.5) 0.8
		1.4	93.6 (3.7, 2.7) 0.8	99.4 (0.5, 0.1) 1.1	95.8 (2.5, 1.7) 0.9	96.0 (2.4, 1.6) 0.9
	100	1.0	92.7 (3.4, 3.9) 0.6	98.8 (0.6, 0.6) 0.9	95.4 (2.1, 2.5) 0.6	95.7 (2.1, 2.2) 0.6
		1.4	93.0 (3.0, 4.0) 0.6	98.7 (0.7, 0.6) 0.9	94.9 (2.2, 2.9) 0.7	95.4 (2.1, 2.5) 0.7
	200	1.0	93.9 (3.1, 3.0) 0.5	99.7 (0.1, 0.2) 0.7	96.3 (1.6, 2.1) 0.5	96.6 (1.4, 2.0) 0.5
		1.4	93.3 (3.6, 3.1) 0.5	99.7 (0.1, 0.2) 0.8	96.5 (1.4, 2.1) 0.6	96.5 (1.3, 2.2) 0.6
0.1	50	1.0	93.3 (4.2, 2.5) 1.3	99.6 (0.3, 0.1) 2.1	95.1 (3.3, 1.6) 1.4	95.0 (3.3, 1.7) 1.4
		1.4	93.7 (4.0, 2.3) 1.4	99.6 (0.3, 0.1) 2.3	95.2 (3.1, 1.7) 1.5	95.1 (3.2, 1.7) 1.6
	100	1.0	93.4 (3.1, 3.5) 1.2	99.6 (0.3, 0.1) 2.0	94.7 (2.3, 3.0) 1.3	95.2 (2.2, 2.6) 1.3
		1.4	93.5 (3.0, 3.5) 1.3	99.6 (0.3, 0.1) 2.2	95.0 (2.1, 2.9) 1.4	95.5 (1.8, 2.7) 1.5
	200	1.0	92.4 (3.7, 3.9) 1.1	99.9 (0.1, 0.0) 2.0	96.1 (1.5, 2.4) 1.3	96.6 (1.3, 2.1) 1.3
		1.4	92.6 (3.5, 3.9) 1.2	100.0 (0.0, 0.0) 2.2	96.1 (1.6, 2.3) 1.4	96.9 (1.1, 2.0) 1.4
0.2	50	1.0	92.6 (4.5, 2.9) 1.7	99.7 (0.2, 0.1) 2.9	95.4 (2.8, 1.8) 1.9	95.5 (2.9, 1.6) 1.9
		1.4	93.6 (4.0, 2.4) 1.8	99.6 (0.2, 0.2) 3.2	95.3 (2.8, 1.9) 2.0	95.3 (3.0, 1.7) 2.1
	100	1.0	93.6 (2.9, 3.5) 1.6	99.6 (0.3, 0.1) 2.8	95.6 (1.8, 2.6) 1.8	95.7 (1.7, 2.6) 1.8
		1.4	93.8 (2.9, 3.3) 1.8	99.6 (0.3, 0.1) 3.1	95.4 (1.8, 2.8) 2.0	95.6 (1.8, 2.6) 2.0
	200	1.0	93.2 (3.2, 3.6) 1.6	100.0 (0.0, 0.0) 2.8	96.1 (1.5, 2.4) 1.8	96.2 (1.4, 2.4) 1.8
		1.4	93.0 (3.2, 3.8) 1.7	100.0 (0.0, 0.0) 3.0	95.9 (1.9, 2.2) 1.9	96.6 (1.5, 1.9) 2.0

Table 4.7: Empirical coverage (%), tail errors ((<, >)%), and median widths (WD) for the difference between two normal means when the number of clusters per arm equal 24 (control) and 12 (experimental), $\alpha = 0.05$, imbalance parameter=0.8, and runs=1000

ICC	m_{ij}	$\frac{\sigma^2_{T_1}}{\sigma^2_{T_2}}$	Wald	ClusterAdj	GCI	MOVER
			Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% Width	Cov (<, >)% WD
0.005	50	1.0	94.3 (3.2, 2.5) 0.6	98.9 (0.7, 0.4) 0.8	95.9 (2.7, 1.4) 0.6	95.8 (2.7, 1.5) 0.6
		1.4	93.7 (3.6, 2.7) 0.6	98.8 (0.8, 0.4) 0.8	95.7 (2.6, 1.7) 0.7	95.7 (2.6, 1.7) 0.7
	100	1.0	93.5 (3.4, 3.1) 0.4	98.8 (0.7, 0.5) 0.6	95.5 (2.2, 2.3) 0.5	95.3 (2.3, 2.4) 0.5
		1.4	93.1 (3.5, 3.4) 0.5	98.5 (0.7, 0.8) 0.6	95.3 (2.0, 2.7) 0.5	95.6 (1.9, 2.5) 0.5
	200	1.0	94.6 (3.0, 2.4) 0.3	99.1 (0.5, 0.4) 0.5	95.8 (2.4, 1.8) 0.4	95.9 (2.3, 1.8) 0.4
		1.4	94.0 (3.3, 2.7) 0.4	98.9 (0.7, 0.4) 0.5	95.9 (2.5, 1.6) 0.4	96.1 (2.4, 1.5) 0.4
0.01	50	1.0	94.1 (3.7, 2.2) 0.6	99.1 (0.6, 0.3) 0.8	96.1 (2.6, 1.3) 0.7	96.2 (2.7, 1.1) 0.7
		1.4	93.9 (3.6, 2.5) 0.7	98.9 (0.8, 0.3) 0.9	96.2 (2.7, 1.1) 0.7	96.2 (2.6, 1.2) 0.7
	100	1.0	92.9 (3.2, 3.9) 0.5	98.6 (0.7, 0.7) 0.7	94.9 (2.4, 2.7) 0.5	95.1 (2.3, 2.6) 0.5
		1.4	92.8 (3.1, 4.1) 0.5	98.5 (0.7, 0.8) 0.8	94.3 (2.4, 3.3) 0.6	94.6 (2.3, 3.1) 0.6
	200	1.0	93.8 (3.3, 2.9) 0.4	99.2 (0.4, 0.4) 0.6	95.7 (2.1, 2.2) 0.4	96.4 (1.8, 1.8) 0.4
		1.4	93.7 (3.2, 3.1) 0.4	99.1 (0.5, 0.4) 0.7	95.7 (2.1, 2.2) 0.5	96.3 (1.7, 2.0) 0.5
0.1	50	1.0	94.7 (2.8, 2.5) 1.1	99.9 (0.1, 0.0) 1.7	96.0 (1.8, 2.2) 1.2	96.2 (1.8, 2.0) 1.2
		1.4	94.3 (3.0, 2.7) 1.2	99.7 (0.2, 0.1) 1.9	95.8 (1.9, 2.3) 1.3	96.1 (1.9, 2.0) 1.4
	100	1.0	93.9 (2.7, 3.4) 1.0	99.3 (0.2, 0.5) 1.7	95.1 (2.2, 2.7) 1.1	95.3 (2.2, 2.5) 1.1
		1.4	93.3 (3.0, 3.7) 1.1	99.3 (0.2, 0.5) 1.8	94.8 (2.4, 2.8) 1.3	95.1 (2.2, 2.7) 1.3
	200	1.0	92.9 (3.0, 4.1) 1.0	99.8 (0.2, 0.0) 1.6	95.0 (1.9, 3.1) 1.1	95.7 (1.8, 2.5) 1.1
		1.4	93.2 (3.0, 3.8) 1.1	99.7 (0.2, 0.1) 1.8	94.6 (2.2, 3.2) 1.2	95.2 (1.8, 3.0) 1.2
0.2	50	1.0	94.1 (3.1, 2.8) 1.4	99.8 (0.1, 0.1) 2.3	95.6 (2.2, 2.2) 1.6	96.3 (1.7, 2.0) 1.6
		1.4	93.7 (3.6, 2.7) 1.6	99.7 (0.2, 0.1) 2.6	95.5 (2.2, 2.3) 1.8	96.1 (1.8, 2.1) 1.8
	100	1.0	94.2 (2.8, 3.0) 1.4	99.6 (0.0, 0.4) 2.3	95.9 (1.7, 2.4) 1.5	96.3 (1.7, 2.0) 1.5
		1.4	93.4 (3.0, 3.6) 1.6	99.4 (0.0, 0.6) 2.5	95.5 (2.0, 2.5) 1.7	95.7 (2.1, 2.2) 1.7
	200	1.0	93.0 (3.1, 3.9) 1.4	99.7 (0.2, 0.1) 2.3	95.1 (1.9, 3.0) 1.5	95.5 (1.9, 2.6) 1.5
		1.4	93.3 (3.0, 3.7) 1.5	99.5 (0.2, 0.3) 2.5	94.9 (2.0, 3.1) 1.7	95.1 (2.0, 2.9) 1.7

Table 4.8: Empirical coverage (%), tail errors ((<, >)%), and median widths (WD) for the difference between two normal means when the number of clusters per arm equal 24 (control) and 24 (experimental), $\alpha = 0.05$, imbalance parameter=0.8, and runs=1000

ICC	m_{ij}	$\frac{\sigma^2_{T_1}}{\sigma^2_{T_2}}$	Wald	ClusterAdj	GCI	MOVER
			Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% Width	Cov (<, >)% WD
0.005	50	1.0	95.8 (1.6, 2.6) 0.5	99.4 (0.3, 0.3) 0.6	96.9 (1.1, 2.0) 0.5	96.9 (1.2, 1.9) 0.5
		1.4	95.6 (1.7, 2.7) 0.5	99.3 (0.3, 0.4) 0.7	96.8 (1.4, 1.8) 0.5	96.7 (1.3, 2.0) 0.5
	100	1.0	93.9 (3.1, 3.0) 0.4	98.6 (0.5, 0.9) 0.5	94.8 (2.8, 2.4) 0.4	94.8 (2.8, 2.4) 0.4
		1.4	93.8 (3.3, 2.9) 0.4	98.6 (0.5, 0.9) 0.5	95.1 (2.4, 2.5) 0.4	95.3 (2.5, 2.2) 0.4
	200	1.0	93.5 (3.3, 3.2) 0.3	99.0 (0.5, 0.5) 0.4	94.8 (2.6, 2.6) 0.3	95.0 (2.7, 2.3) 0.3
		1.4	93.9 (3.0, 3.1) 0.3	99.1 (0.5, 0.4) 0.4	94.5 (2.8, 2.7) 0.3	95.0 (2.7, 2.3) 0.3
0.01	50	1.0	96.1 (1.4, 2.5) 0.5	99.0 (0.5, 0.5) 0.7	96.6 (1.0, 2.4) 0.5	96.7 (1.1, 2.2) 0.5
		1.4	96.0 (1.4, 2.6) 0.5	99.0 (0.5, 0.5) 0.7	96.7 (1.2, 2.1) 0.6	96.8 (1.2, 2.0) 0.6
	100	1.0	93.9 (3.4, 2.7) 0.4	99.0 (0.4, 0.6) 0.6	94.7 (2.9, 2.4) 0.4	95.1 (2.6, 2.3) 0.4
		1.4	94.1 (3.2, 2.7) 0.4	98.9 (0.4, 0.7) 0.6	95.1 (2.6, 2.3) 0.5	95.2 (2.8, 2.0) 0.5
	200	1.0	93.6 (3.3 , 3.1) 0.3	99.4 (0.4, 0.2) 0.5	94.4 (3.0, 2.6) 0.3	94.6 (2.8, 2.6) 0.3
		1.4	93.8 (3.0, 3.2) 0.4	99.3 (0.5, 0.2) 0.5	94.7 (2.9, 2.4) 0.4	94.9 (2.5, 2.6) 0.4
0.1	50	1.0	94.9 (2.5, 2.6) 0.9	99.4 (0.3, 0.3) 1.4	96.1 (1.7, 2.2) 0.9	96.2 (1.8, 2.0) 0.9
		1.4	95.0 (2.4, 2.6) 1.0	99.3 (0.4, 0.3) 1.5	96.0 (1.8, 2.2) 1.0	96.4 (1.6, 2.0) 1.0
	100	1.0	94.4 (3.1, 2.5) 0.8	99.8 (0.2, 0.0) 1.4	95.3 (2.7, 2.0) 0.9	95.4 (2.7, 1.9) 0.9
		1.4	94.1 (3.0, 2.9) 0.9	99.8 (0.2, 0.0) 1.5	95.3 (2.6, 2.1) 1.0	95.4 (2.7, 1.9) 1.0
	200	1.0	94.0 (2.6, 3.4) 0.8	99.7 (0.2, 0.1) 1.3	94.8 (2.4, 2.8) 0.9	95.3 (2.1, 2.6) 0.9
		1.4	93.8 (2.8, 3.4) 0.9	99.7 (0.2, 0.1) 1.5	94.8 (2.4, 2.8) 0.9	95.2 (2.2, 2.6) 0.9
0.2	50	1.0	95.0 (2.5, 2.5) 1.2	99.4 (0.3, 0.3) 1.9	95.7 (2.0, 2.3) 1.2	95.6 (2.0, 2.4) 1.2
		1.4	95.1 (2.2, 2.7) 1.3	99.3 (0.4, 0.3) 2.1	96.1 (1.6, 2.3) 1.4	96.1 (1.6, 2.3) 1.4
	100	1.0	94.7 (2.9, 2.4) 1.2	99.8 (0.2, 0.0) 1.9	95.3 (2.4, 2.3) 1.2	95.6 (2.5, 1.9) 1.2
		1.4	93.9 (3.4, 2.7) 1.3	99.8 (0.2, 0.0) 2.1	95.3 (2.6, 2.1) 1.3	95.7 (2.4, 1.9) 1.3
	200	1.0	93.9 (2.7, 3.4) 1.1	99.7 (0.2, 0.1) 1.9	94.9 (2.3, 2.8) 1.2	95.2 (1.9, 2.9) 1.2
		1.4	93.9 (2.6, 3.5) 1.2	99.8 (0.1, 0.1) 2.1	95.1 (2.3, 2.6) 1.3	95.4 (2.2, 2.4) 1.3

Tail errors

The Wald method and the cluster-adjusted method both result in symmetric limits around the point estimate, while the generalized confidence interval method and the MOVER do not. This enforced symmetry is not an issue however when the sampling distribution of the parameter estimate is symmetric. The difference between two normal means follows a normal distribution, therefore it was expected that all of the procedure would perform similarly in terms of balance between tail errors. This is clearly seen in the simulation results found in Tables 4.4-4.8. There is no need to compare tail errors of the cluster-adjusted procedure because its coverage results are not valid.

Median width

Procedures with empirical coverage close to the nominal are compared based on their median confidence interval widths, where a narrower width is translated into greater precision. For all of the methods, the median width increases with the ICC value and under heteroscedasticity, whereas an increase in both the number of clusters per arm and the average cluster size lead to narrower confidence interval widths, and therefore greater precision.

The Wald method is comparable to the generalized confidence interval procedure and the MOVER when there are a large number of clusters per arm (Table 4.8). The method shows slightly narrower median widths than the other procedures, especially as the average cluster sizes increase.

Discussion of the confidence interval width of the cluster adjusted procedure is excluded because the method failed to satisfy the coverage requirements for all of the parameter combinations. Confidence interval width is therefore irrelevant if the interval itself cannot maintain nominal coverage.

The most meaningful comparison lies between the generalized confidence interval

procedure and the MOVER. These two methods had similar empirical coverage and tail error performances throughout all of the parameter combinations. The generalized confidence interval procedure showed slightly narrower widths than the MOVER, but only occasionally. With 24 clusters per arm there were no differences between the median widths of the two procedures.

The difference between two lognormal means

Empirical coverage results, tail errors, and median widths for the Wald method, the generalized confidence interval procedure and the MOVER as applied to unbalanced, completely randomized cluster randomization trials for a difference between two lognormal means are presented in Tables 4.9 to 4.13.

Confidence interval coverage

The three confidence interval procedures are investigated for 120 parameter combinations. Overall, each procedure shows improved coverage performance as the number of clusters per arm increase.

Even with only 6 clusters per arm, the empirical coverage of the Wald method falls within the desired range ($93.6\% - 96.4\%$) when the ICC is less than 0.1 (see Table 4.9). With larger ICC values, the Wald method shows anti-conservative results, particularly when the variances in the two arms differ. Fortunately, as the number of clusters increase to 12 or 24, the Wald method shows improved coverage results with larger ICC values. Under homoscedasticity with a larger effective sample size, empirical coverage is close to the nominal. However, under heteroscedasticity, results remain anti-conservative.

The generalized confidence interval procedure often has high coverage when there are only 6 clusters per group, exceeding the desired coverage range ($93.6\% - 96.4\%$) 46% of the time. When exceeding this acceptable range, empirical coverage falls an average of 1.9-percentage points above the nominal. However, when the number of clus-

ters per arm increases to (24,12) and (24,24), the empirical coverage results fall closely around the nominal. The average cluster size, homoscedasticity/heteroscedasticity, and the ICC value do not greatly impact coverage results for this method.

The MOVER performs best out of all three procedures in terms of confidence interval coverage. With only 6 clusters per arm, the empirical coverage exceeds the desired range only 17% of the time by an average of 1.8-percentage points above the nominal. Furthermore, similar to the generalized confidence interval procedure, the empirical coverage is consistently close to the nominal when there are (24,12) and (24,24) clusters per arm. The method appears to be somewhat sensitive to high ICC values (ICC=0.2) with a small number of clusters per arm, however average cluster size and homoscedasticity/heteroscedasticity do not influence coverage results.

Tail errors

The lognormal mean is a function of the normal mean and variance, making it the sum of a normal random variable and a chi-squared random variable. Consequently, the sampling distribution of the lognormal mean is skewed. Symmetric confidence interval procedures would therefore fail to balance tail errors, resulting in excluding potentially plausible parameter values from one side of the interval while failing to exclude extreme values from the other. Alternatively, confidence interval procedures which capture the underlying distribution of the lognormal mean would improve tail error performance.

The Wald method is the only symmetric confidence interval procedure investigated for the difference between two lognormal means to demonstrate the flaw in such a restriction. The lognormal distribution is skewed to the right, suggesting that a symmetric interval would miss plausible parameter values on its right. This is clearly seen in the simulation results in Tables 4.9 to 4.13, where the method consistently misses less than 2.5% from the left and more than 2.5% from the right.

The generalized confidence interval procedure and the MOVER both have rela-

Table 4.9: Empirical coverage (%), tail errors ((<, >)%), and median widths (WD) for the difference between two lognormal means when the number of clusters per arm equal 6 (control) and 6 (experimental), $\alpha = 0.05$, imbalance parameter=0.8, and runs=1000

ICC	m_{ij}	$\frac{\sigma^2_{T_1}}{\sigma^2_{T_2}}$	Wald	GCI	MOVER
			Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% Width
0.005	50	1.0	94.7 (0.1, 5.2) 36.8	96.8, (2.1, 1.1) 59.6	96.1, (2.6, 1.3) 52.8
	50	1.4	92.5 (0.0, 7.5) 121.0	96 (3.1, 0.9) 201.4	95.7 (2.8, 1.5) 176.1
	100	1.0	94.8 (0.5, 4.7) 26.3	97.1 (1.5, 1.4) 36.4	96.0 (1.9, 2.1) 31.9
	100	1.4	93.8 (0.2, 6.0) 88.1	96.5 (2.4, 1.1) 121.6	96.0 (2.3, 1.7) 108.2
	200	1.0	94.3 (1.0, 4.7) 19.3	96.3 (1.7, 2.0) 24.6	94.6 (2.7, 2.7) 21.6
	200	1.4	93.4 (0.6, 6.0) 65.1	95.7 (2.4 1.9) 82.6	95.1 (2.7 2.2) 73.1
0.01	50	1.0	94.8 (0.2, 5.0) 38.0	9.0 (2.0, 1.0) 64.4	96.4 (2.3, 1.3) 56.9
	50	1.4	91.6 (0.0, 8.4) 124.2	96.4 (2.8, 0.8) 216.8	95.7 (2.8, 1.5) 193.8
	100	1.0	94.7 (0.4, 4.9) 28.1	96.8 (1.5, 1.7) 41.4	95.8 (1.8, 2.4) 36.3
	100	1.4	92.8 (0.1, 7.1) 91.7	96.3 (2.3, 1.4) 137.7	95.7 (2.6, 1.7) 120.6
	200	1.0	93.3 (1.4, 5.3) 21.3	96.4 (1.9, 1.7) 28.9	94.3 (2.8, 2.9) 25.1
	200	1.4	92.9 (0.3, 6.8) 69.9	95.7 (2.4, 1.9) 94.3	94.8 (2.6, 2.6) 83.6
0.1	50	1.0	92.4 (0.2, 7.4) 59.5	96.3 (1.6, 2.1) 286.8	96.0 (1.6, 2.4) 251.0
	50	1.4	87.8 (0.0, 12.2) 176.1	95.9 (2.4, 1.7) 1075.7	95.6 (2.1, 2.3) 912.7
	100	1.0	94.0 (0.0, 6.0) 50.4	96.8 (0.9, 2.3) 208.8	96.7 (0.8, 2.5) 177.8
	100	1.4	88.3 (0.1, 11.6) 151.4	96.7 (1.4, 1.9) 730.9	96.3 (1.3, 2.4) 618.1
	200	1.0	91.8 (0.2, 8.0) 46.9	97.1 (1.0, 1.9) 181.9	96.2 (1.1, 2.7) 152.9
	200	1.4	87.5 (0.0, 12.5) 139.4	96.4 (1.5, 2.1) 588.0	95.3 (1.8, 2.9) 500.1
0.2	50	1.0	92.2 (0.2, 7.6) 83.9	96.0 (1.1, 2.9) 1359.0	96 (1.2, 2.8) 1098.8
	50	1.4	84.0 (0.0, 16.0) 230.0	96.0 (1.7, 2.3) 5683.5	96.3 (1.4, 2.3) 4516.3
	100	1.0	93.3 (0.0, 6.7) 74.8	97.0 (0.8, 2.2) 948.9	97.1 (0.6, 2.3) 827.5
	100	1.4	84.6 (0.0, 15.4) 205.7	96.2 (1.6, 2.2) 4243.3	96.1 (1.3, 2.6) 3553.8
	200	1.0	92.8 (0.0, 7.2) 73.5	97.3 (0.8, 1.9) 881.1	96.9 (0.7, 2.4) 741.1
	200	1.4	84.2 (0.0, 15.8) 206.8	96.6 (1.3, 2.1) 3575.9	96.5 (1.0, 2.5) 3052.0

Table 4.10: Empirical coverage (%), tail errors ((<, >)%), and median widths (WD) for the difference between two lognormal means when the number of clusters per arm equal 12 (control) and 6 (experimental), $\alpha = 0.05$, imbalance parameter=0.8, and runs=1000

ICC	m_{ij}	$\frac{\sigma^2_{T_1}}{\sigma^2_{T_2}}$	Wald	GCI	MOVER
			Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% Width
0.005	50	1.0	93.8 (0.1 , 6.1) 34.8	96.2 (2.6, 1.2) 51.7	95.6 (2.5, 1.9) 45.6
	50	1.4	92.3 (0.0, 7.7) 119.8	96.3 (2.7, 1.0) 198.3	96.3 (2.0, 1.7) 171.7
	100	1.0	94.6 (0.2, 5.2) 25.5	97.4 (1.6, 1.0) 33.5	96.2 (1.9, 1.9) 29.7
	100	1.4	92.3 (0.0, 7.7) 89.2	97.1 (1.9, 1.0) 123.0	96.5 (1.8, 1.7) 109.1
	200	1.0	94.6 (0.6, 4.8) 19.0	96.6 (2.1, 1.3) 23.4	95.0 (2.5, 2.5) 20.8
	200	1.4	94.8 (0.2, 5.0) 65.2	96.9 (2.2, 0.9) 82.4	95.9 (2.2, 1.9) 74.1
0.01	50	1.0	92.7 (0.2, 7.1) 35.7	96.1 (2.4, 1.5) 55.1	96.0 (2.1, 1.9) 48.9
	50	1.4	91.8 (0.0, 8.2) 122.4	96.2 (2.4, 1.4) 212.9	96.3 (2.0, 1.7) 186.1
	100	1.0	94.3 (0.2, 5.5) 27.1	97.2 (1.7, 1.1) 37.1	96.1 (1.9, 2.0) 32.8
	100	1.4	92.5, (0.1, 7.4) 93.1	96.7 (2.1, 1.2) 136.4	96.4 (1.8, 1.8) 120.4
	200	1.0	94.4 (0.7, 4.9) 20.8	96.7 (2.0, 1.3) 27.4	95.0 (2.4, 2.6) 23.9
	200	1.4	94.1 (0.4, 5.5) 70.5	97.1 (2.1, 0.8) 96.7	95.6 (2.2, 2.2) 84.8
0.1	50	1.0	90.6 (0.1, 9.3) 52.9	96.8 (1.8, 1.4) 168.6	96.3 (1.8, 1.9) 144.8
	50	1.4	86.9 (0.0, 13.1) 170.8	96.0 (2.6, 1.4) 792.7	96.3 (1.8, 1.9) 671.4
	100	1.0	90.1 (0.3, 9.6) 48.3	96.4 (1.6, 2.0) 146.8	95.7 (1.8, 2.5) 124.8
	100	1.4	86.5 (0.1, 13.4) 155.6	95.5 (2.4, 2.1) 673	95.7 (1.9, 2.4) 571.9
	200	1.0	91.5 (0.2, 8.3) 43.3	96.5 (1.4, 2.1) 120.6	96.4 (1.1, 2.5) 104.6
	200	1.4	88.5 (0.1, 11.4) 139.8	96.4 (1.9, 1.7) 517.1	96.2 (1.4, 2.4) 443.3
0.2	50	1.0	88.2 (0.1, 11.7) 70.6	96.4 (1.9, 1.7) 500.5	96.5 (1.2, 2.3) 437.5
	50	1.4	82.8 (0.1, 17.1) 223.3	96.0 (2.6, 1.4) 3132.7	96.4 (1.5, 2.1) 2631.5
	100	1.0	89.3 (0.1, 10.6) 69.3	95.3 (1.9, 2.8) 449.0	96.0 (1.2, 2.8) 385.1
	100	1.4	83.6 (0, 16.4) 217.2	95.4 (2.2, 2.4) 2706.8	95.7 (1.6, 2.7) 2445.5
	200	1.0	90.0 (0.1, 9.9) 62.8	96.3 (1.6, 2.1) 384.2	96.9 (0.6, 2.5) 334.7
	200	1.4	84.4 (0.0, 15.6) 195.6	95.7 (2.2, 2.1) 2229.5	96.0 (1.2, 2.8) 2052.2

Table 4.11: Empirical coverage (%), tail errors ((<, >)%), and median widths (WD) for the difference between two lognormal means when the number of clusters per arm equal 12 (control) and 12 (experimental), $\alpha = 0.05$, imbalance parameter=0.8, and runs=1000

ICC	m_{ij}	$\frac{\sigma^2_{T_1}}{\sigma^2_{T_2}}$	Wald	GCI	MOVER
			Cov, (<, >)%, WD	Cov, (<, >)%, WD	Cov, (<, >)%, Width
0.005	50	1.0	96.0 (0.4, 3.6) 25.7	97.1 (1.6, 1.3) 30.4	96.6, (1.8, 1.6) 28.7
	50	1.4	95.0 (0.0, 5.0) 85.3	96.9 (2.1, 1.0) 102.7	96.5 (2.2, 1.3) 96.6
	100	1.0	94.5 (0.4, 5.1) 18.7	96.0 (2.1, 1.9) 20.9	95.0 (2.0, 3.0) 19.9
	100	1.4	93.5 (0.1, 6.4) 61.9	96.4 (2.2, 1.4) 70.2	96.1 (1.8, 2.1) 66.9
	200	1.0	95.1 (1.0, 3.9) 14.1	96.7 (1.5, 1.8) 15.5	95.4 (2.0, 2.6) 14.6
	200	1.4	94.8 (0.5, 4.7) 47.2	95.7 (2.8, 1.5) 51.5	95.2 (2.6, 2.2) 49.1
0.01	50	1.0	95.4 (0.4, 4.2) 26.4	97.4 (1.4, 1.2) 31.8	97.1 (1.5, 1.4) 29.9
	50	1.4	94.5 (0.0, 5.5) 86.9	96.7 (2.0, 1.3) 106.9	96.8 (1.6, 1.6) 100.5
	100	1.0	94.4 (0.5, 5.1) 19.7	95.6 (2.3, 2.1) 22.3	95.2 (2.1, 2.7) 21.2
	100	1.4	93.8 (0.2, 6.0) 65.0	96.4 (2.0, 1.6) 74.6	95.9 (1.8, 2.3) 70.7
	200	1.0	94.9 (0.8, 4.3) 15.5	96.3 (1.6, 2.1) 17.4	95.3 (1.9, 2.8) 16.2
	200	1.4	94.6 (0.4, 5.0) 50.9	95.9 (2.6, 1.5) 57.3	94.9 (2.6, 2.5) 54.1
0.1	50	1.0	94.5 (0.2, 5.3) 40.5	96.6 (1.8, 1.6) 67.0	95.9 (1.6, 2.5) 61.9
	50	1.4	90.1 (0.0, 9.9) 126.9	96.2 (2.2, 1.6) 220.4	96.0 (1.9, 2.1) 203.8
	100	1.0	93.5 (0.2, 6.3) 36.3	95.6 (2.1, 2.3) 57.7	95.8 (1.5, 2.7) 52.5
	100	1.4	89.9 (0, 10.1) 114.3	95.2 (2.5, 2.3) 189.6	95.4 (2.2, 2.4) 173.8
	200	1.0	93.1 (0.5, 6.4) 34.8	95.4 (1.7, 2.9) 54.4	94.9 (1.6, 3.5) 49.7
	200	1.4	89.6 (0.2, 10.2) 108.5	94.7 (2.4, 2.9) 175.9	94.6 (1.7, 3.7) 161.3
0.2	50	1.0	93.4 (0.0, 6.6) 57.2	95.6 (1.7, 2.7) 132.2	95.9 (1.4, 2.7) 124.8
	50	1.4	87.0 (0.0, 13.0) 171.5	95.0 (2.5, 2.5) 469.3	95.6 (1.7, 2.7) 430.9
	100	1.0	93.7 (0.0, 6.3) 53.7	95.9 (1.6, 2.5) 119.8	95.9 (1.3, 2.8) 110.1
	100	1.4	87.8 (0.0, 12.2) 164.3	95.3 (2.5 2.2) 424.4	95.3 (2.1, 2.6) 378.1
	200	1.0	93.0 (0.0, 7.0) 52.8	94.9 (2.0, 3.1) 116.0	95.1 (1.4, 3.5) 106.4
	200	1.0	87.2 (0.0, 12.8) 161.2	94.9 (2.3, 2.8) 394.6	95.2 (1.5, 3.3) 367.5

Table 4.12: Empirical coverage (%), tail errors ((<, >)%), and median widths (WD) for the difference between two lognormal means when the number of clusters per arm equal 24 (control) and 12 (experimental), $\alpha = 0.05$, imbalance parameter=0.8, and runs=1000

ICC	m_{ij}	$\frac{\sigma^2_{T_1}}{\sigma^2_{T_2}}$	Wald	GCI	MOVER
			Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% Width
0.005	50	1.0	94.9 (0.0, 5.1), 24.7	95.7 (2.5, 1.8) 29.0	94.9 (2.6, 2.5) 27.3
	50	1.4	93.5 (0.0, 6.5) 85.7	95.1 (3.3, 1.6) 104.4	95.2 (2.6, 2.2) 97.5
	100	1.0	93.5 (1.2, 5.3) 18.1	94.9 (2.6, 2.5) 20.1	94.2 (2.5, 3.3) 19.0
	100	1.4	93.3 (0.6, 6.1) 62.9	95.0 (2.7, 2.3) 70.5	94.3 (2.9, 2.8) 67.3
	200	1.0	94.1 (1.5, 4.4) 13.5	94.7 (3.0, 2.3) 14.7	94.4 (2.8, 2.8) 13.9
	200	1.4	94.1 (1.1, 4.8) 46.4	94.7 (3.1, 2.2) 50.6	94.3 (2.8, 2.9) 48.2
0.01	50	1.0	94.5 (0.1, 5.4) 25.4	95.4 (2.7, 1.9) 30.2	95.0 (2.5, 2.5) 28.4
	50	1.4	92.9 (0.0, 7.1) 87.8	95.2 (3.2, 1.6) 108.1	95.4 (2.4, 2.2) 101.1
	100	1.0	93.4 (1.2, 5.4) 19.1	95.3 (2.4, 2.3) 21.5	93.9 (2.7, 3.4) 20.3
	100	1.4	93.6 (0.7, 5.7) 65.5	94.9 (2.8, 2.3) 75.2	94.8 (2.5, 2.7) 71.0
	200	1.0	94.1 (1.9, 4.0) 15.0	94.4 (3.2, 2.4) 16.6	94.2 (3.0, 2.8) 15.6
	200	1.4	94.9 (0.9, 4.2) 50.6	94.1 (3.5, 2.4) 57.0	94.6 (2.8, 2.6) 53.3
0.1	50	1.0	92.7 (0.2, 7.1) 38.1	95.2 (3.0, 1.8) 58.7	95.8 (2.3, 1.9) 54.2
	50	1.4	90.7 (0.0, 9.3) 126.6	95.3 (2.9, 1.8) 225.3	95.8 (2.2, 2.0) 202.7
	100	1.0	92.2 (0.2, 7.6) 34.8	95.6 (2.6, 1.8) 51.3	95.7 (1.6, 2.7) 47.4
	100	1.4	91.2 (0.1, 8.7) 115.2	95.3 (2.9, 1.8) 188.5	95.1 (2.5, 2.4) 175.7
	200	1.0	93.3 (0.2, 6.5) 33.1	94.0 (3.6, 2.4) 48.3	94.5 (2.3, 3.2) 44.7
	200	1.4	92.1 (0.1, 7.8) 108.3	94.1 (3.8, 2.1) 176.3	94.4 (2.3, 3.3) 163.5
0.2	50	1.0	92.2 (0.0, 7.8) 51.4	95.0 (2.6, 2.4) 105.5	95.2 (2.1, 2.7) 96.1
	50	1.4	89.4 (0.0, 10.6) 169.9	94.2 (3.5, 2.3) 440.2	95.0 (2.5, 2.5) 403.6
	100	1.0	91.5 (0.1, 8.4) 50.1	95.4 (2.3, 2.3) 97.2	95.7 (1.8, 2.5) 89.1
	100	1.4	88.7 (0.0, 11.3) 163.8	94.8 (3.1, 2.1) 399.5	96.2 (1.5, 2.3) 360.9
	200	1.0	91.6 (0.1, 8.3) 50.0	94.0 (3.3 , 2.7) 96.0	94.9 (1.9, 3.2) 88.5
	200	1.4	88.4 (0.0, 11.6) 164.1	94.2 (3.6, 2.2) 398.6	95.0 (2.2, 2.8) 359.9

Table 4.13: Empirical coverage (%), tail errors ((<, >)%), and median widths (WD) for the difference between two lognormal means when the number of clusters per arm equal 24 (control) and 24 (experimental), $\alpha = 0.05$, imbalance parameter=0.8, and runs=1000

ICC	m_{ij}	$\frac{\sigma^2_{T_1}}{\sigma^2_{T_2}}$	Wald	GCI	MOVER
			Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% Width
0.005	50	1.0	95.3 (0.6, 4.1) 18.3	95.8 (2.6, 1.6) 19.8	95.5 (2.6, 1.9) 19.1
	50	1.4	94.3 (0.3, 5.4) 61.4	95.3 (3.1, 1.6) 66.8	95.0 (2.9, 2.1) 64.9
	100	1.0	94.4 (1.4, 4.2) 13.3	95.2 (1.8, 3.0) 13.9	95.0 (2.0, 3.0) 13.6
	100	1.4	93.9 (1.2, 4.9) 44.4	95.3 (2.4, 2.3) 46.9	94.7 (2.3, 3.0) 45.7
	200	1.0	95.6 (1.6, 2.8) 9.9	96.5 (2.0, 1.5) 10.4	95.8 (2.5, 1.7) 10.1
	200	1.4	95.0 (1.3, 3.7) 32.9	95.7 (2.3, 2.0) 34.3	95.7 (2.2, 2.1) 33.4
0.01	50	1.0	95.7 (0.5, 3.8) 18.9	95.5 (2.9, 1.6) 20.4	95.4 (2.6, 2.0) 19.8
	50	1.4	94.2 (0.3, 5.5) 62.5	95.5 (2.7, 1.8) 69.2	95.2 (2.6, 2.2) 67.0
	100	1.0	94.2 (1.4, 4.4) 14.1	95.4 (2.2, 2.4) 14.9	95.1 (2.2, 2.7) 14.5
	100	1.4	93.7 (0.9, 5.4) 46.8	95.2 (2.4, 2.4) 49.6	94.8 (2.2, 3.0) 48.4
	200	1.0	95.1 (1.9, 3.0) 10.9	96.1 (2.3, 1.6) 11.5	96.0 (2.3, 1.7) 11.1
	200	1.4	94.5 (1.4, 4.1) 35.8	95.8 (2.1, 2.1) 37.7	95.6 (2.3, 2.1) 36.6
0.1	50	1.0	94.1 (0.1, 5.8) 28.5	95.4 (1.7, 2.9) 35.1	95.4 (1.5, 3.1) 33.5
	50	1.4	92.3 (0.0, 7.7) 92.3	95.0 (1.9, 3.1) 115.6	95.1 (1.5, 3.4) 111.3
	100	1.0	94.3 (0.4, 5.3) 25.2	94.7 (2.2, 3.1) 30.5	94.6 (1.9, 3.5) 29.1
	100	1.4	92.0 (0.1, 7.9) 80.2	93.7 (3.2, 3.1) 98.6	93.8 (2.5, 3.7) 95.0
	200	1.0	95.2 (0.4, 4.4) 23.9	95.7 (2.3, 2.0) 28.5	95.4 (1.9, 2.7) 27.3
	200	1.4	92.6 (0.2, 7.2) 75.4	95.6 (2.5, 1.9) 92.7	95.8 (1.6, 2.6) 88.8
0.2	50	1.0	93.6 (0.0, 6.4) 39.2	95.7 (1.4, 2.9) 55.0	95.6 (1.3, 3.1) 53.0
	50	1.4	90.5 (0.0, 9.5) 124.6	94.9 (2.0, 3.1) 184.2	95.0 (1.4, 3.6) 177.9
	100	1.0	94.4 (0.0, 5.6) 36.5	95.2 (1.8, 3.0) 51.0	95.0 (1.6, 3.4) 48.4
	100	1.4	90.5 (0.0, 9.5) 113.8	94.0 (3.2, 2.8) 167.5	94.0 (2.7, 3.3) 157.6
	200	1.0	94.5 (0.1, 5.4) 35.9	95.3 (1.8, 2.9) 49.3	95.6 (1.6, 2.8) 47.1
	200	1.4	90.1 (0.0, 9.9) 113.4	95.7 (2.1, 2.2) 162.8	95.4 (2.0, 2.6) 155.9

tively balanced tail errors, although not perfect. Tail error imbalance typically occurs more so with the generalized confidence interval procedure when empirical coverage is high.

Median width

The Wald method generally has low median interval widths. However, the low coverage with high ICC values and heteroscedasticity, acompanied by the unbalanced tail errors makes this method less desirable than its alternatives.

The generalized confidence interval procedure is consistently wider than the MOVER, suggesting that the MOVER demonstrates greater precision. However, both methods demonstrate increased width as the ICC value increases. Heteroscedasticity also increases confidence interval widths. An increase in the number of clusters per arm, and to a lesser extent the average cluster size, result in greater precision for these confidence interval procedures.

The exceedance probability

Empirical coverage results ($\alpha = 0.05$), tail errors, and median widths for the Wald method, the generalized confidence interval procedure and the MOVER as applied to unbalanced, completely randomized cluster randomization trials for the exceedance probability are presented in Tables 4.14 to 4.23. For each objective, the results of $\mathrm{P}(Y_1 > Y_2) = 0.5$ (Tables 4.14 to 4.18) are first discussed, followed by those of $\mathrm{P}(Y_1 > Y_2) = 0.9$ (Tables 4.19 to 4.23).

Confidence interval coverage

When $\mathrm{P}(Y_1 > Y_2) = 0.5$, the Wald method has low coverage when there are a small number of clusters per arm, although this method shows some evidence of improvements as the ICC increases (Tables 4.14 and 4.15). When the number of clusters

per arm increases to 12, the Wald method greatly improves, with 95.8% of empirical coverage results falling within the desired range of $93.6\% - 96.4\%$. Tables 4.16 to 4.18 show only slight, but consistent, coverage below 95%. For all 120 parameter combinations, the Wald method maintains coverage 61.8% of the time. A similar pattern is seen when $P(Y_1 > Y_2) = 0.9$ - empirical coverage is low when there are a small number of clusters per arm (Tables 4.19 and 4.20), but improves as the number of clusters increase (Tables 4.21 to 4.23). The ICC value, the average cluster size, and homoscedasticity/heteroscedasticity do not influence empirical coverage greatly. The Wald method maintains coverage 58.1% of the time for all parameter combinations.

For $P(Y_1 > Y_2) = 0.5$, The generalized confidence interval procedure shows conservative coverage results when there are (6,6) or (6,12) clusters in the arms, but improves in performance as the number of clusters increase, though not as quickly as the Wald method. The generalized confidence interval procedure shows comparable performance to the Wald method when the number of clusters are set to (24,12). In Tables 4.17 and 4.18, empirical coverage lies close to the nominal 95% level. The procedure shows coverage closer to the nominal as the ICC increases to 0.2, but cluster sizes and homoscedasticity/heteroscedasticity do not greatly influence coverage results. Overall, the generalized confidence interval procedure maintains coverage for 61.7% of the parameter combinations. For $P(Y_1 > Y_2) = 0.9$, a similar general pattern of results were observed - with empirical coverage falling closer to the nominal as the number of clusters increases. The procedure maintains reasonable coverage within the range $93.6\% - 96.4\%$ for 85% of the 120 parameter combinations.

When $P(Y_1 > Y_2) = 0.5$, the MOVER has similar performance to the Wald method - with anti-conservative coverage when there are (6,6) or (12,6) clusters and coverage consistently close to the nominal as the number of clusters increase. Additionally, this procedure shows improvements in validity as the ICC increases, with slight evidence of improvements when the cluster sizes increase. As evident in the tables, the MOVER is not sensitive to homoscedasticity/heteroscedasticity. Overall, this method shows the

best performance out of all three methods, maintaining coverage 68.3% of the time. When $P(Y_1 > Y_2)$ was set to 0.9, the MOVER shows considerable improvements in empirical coverage results, even with a small number of clusters per arm (Tables 4.19 and 4.20). For all 120 parameter combinations, this method has coverage close to the nominal 90% of the time - better than either of the other two procedures investigated.

Tail errors

Although some imbalance is observed with the confidence intervals derived using the Wald method when $P(Y_1 > Y_2) = 0.5$ the method shows balanced tail errors for many of the parameter combinations, especially when the ICC is low. When the ICC=0.2, the confidence intervals miss the true parameter value from the left more often than the right. This pattern lessens when there are 24 clusters per arm. When $P(Y_1 > Y_2) = 0.9$, the Wald method has unbalanced tail errors more frequently than when $P(Y_1 > Y_2) = 0.5$. A similar pattern of increased imbalance occurs as the ICC increases to 0.2, as well as increased balance as the number of clusters per arm reaches 24.

The balance of tail errors for the generalized confidence interval procedure is better than that of the Wald method, however this procedure still experiences unbalanced tail errors when the ICC is high (ICC=0.2) - with the interval missing the true parameter value from the right more often than the left. When $P(Y_1 > Y_2) = 0.9$ there is more imbalance than when $P(Y_1 > Y_2) = 0.5$, though in both cases tail error performance improves as the number of clusters per arm increase.

The MOVER maintains balanced tail errors more often than either the Wald method or the generalized confidence interval procedure. When $P(Y_1 > Y_2) = 0.5$ and there are only 6 clusters per arm some imbalance is observed. Above 6 clusters per arm, tail error imbalance for the MOVER is not an issue. When $P(Y_1 > Y_2) = 0.9$, tail errors of the MOVER are relatively balanced for the 120 parameter combinations, even at high ICC values and 6 clusters per arm.

Table 4.14: Empirical coverage (%), tail errors ((<, >)%), and median widths (WD) for $P(Y_1 > Y_2) = 0.5$ when the number of clusters per arm equal 6 (control) and 6 (experimental), $\alpha = 0.05$, imbalance parameter=0.8, and runs=1000.

ICC	m_{ij}	$\frac{\sigma^2_{T_1}}{\sigma^2_{T_2}}$	Wald	GCI	MOVER
			Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% Width
0.005	50	1.0	93.0 (4.1, 2.9) 0.1	98.0 (1.1, 0.9) 0.1	92.9 (4.2, 2.9) 0.1
		1.4	93.1 (3.8, 3.1) 0.1	97.6 (1.3, 1.1) 0.1	93.1 (3.8, 3.1) 0.1
	100	1.0	92.7 (3.6, 3.7) 0.1	97.7 (0.9, 1.4) 0.1	92.6 (3.6, 3.8) 0.1
		1.4	93.0 (3.2, 3.8) 0.1	97.7 (1.0, 1.3) 0.1	92.9 (3.3, 3.8) 0.1
	200	1.0	91.5 (4.2, 4.3) 0.1	97.3 (1.5, 1.2) 0.1	91.5 (4.2, 4.3) 0.1
		1.4	91.4 (4.1, 4.5) 0.1	97.1 (1.4, 1.5) 0.1	91.4 (4.1, 4.5) 0.1
0.01	50	1.0	92.9 (3.7, 3.4) 0.1	98.1 (1.1, 0.8) 0.2	92.9 (3.7, 3.4) 0.1
		1.4	92.9 (3.7, 3.4) 0.1	97.8 (1.2, 1.0) 0.2	92.9 (3.7, 3.4) 0.1
	100	1.0	93.1 (3.1, 3.8) 0.1	97.0 (1.2, 1.8) 0.1	93.1 (3.1, 3.8) 0.1
		1.4	92.9 (3.3, 3.8) 0.1	97.3 (1.3, 1.4) 0.1	92.8 (3.3, 3.9) 0.1
	200	1.0	90.9 (4.4, 4.7) 0.1	97.2 (1.7, 1.1) 0.1	90.8 (4.4, 4.8) 0.1
		1.4	91.0 (3.9, 5.1) 0.1	97.0 (1.5, 1.5) 0.1	91.0 (3.9, 5.1) 0.1
0.1	50	1.0	92.4 (2.9, 4.7) 0.2	97.6 (1.1, 1.3) 0.3	92.4 (2.9, 4.7) 0.2
		1.4	92.4 (3.0, 4.6) 0.2	97.2 (1.1, 1.7) 0.3	92.4 (3.0, 4.6) 0.2
	100	1.0	93.1 (2.8, 4.1) 0.2	97.6 (1.1, 1.3) 0.2	93.0 (2.8, 4.2) 0.2
		1.4	92.8 (3.0, 4.2) 0.2	97.7 (0.9, 1.4) 0.2	92.8 (3.0, 4.2) 0.2
	200	1.0	90.9 (4.8, 4.3) 0.2	97.0 (1.9, 1.1) 0.2	90.7 (4.8, 4.5) 0.2
		1.4	90.9 (4.8, 4.3) 0.2	96.2 (2.2, 1.6) 0.2	90.8 (4.8, 4.4) 0.2
0.2	50	1.0	94.6 (3.8, 1.6) 0.1	96.5 (0.7, 2.8) 0.2	97.1 (1.2, 1.7) 0.1
		1.4	93.8 (3.6, 2.6) 0.2	96.6 (0.4, 3.0) 0.2	95.9 (1.5, 2.6) 0.2
	100	1.0	92.4 (4.9, 2.7) 0.1	96.3 (0.5, 3.2) 0.2	95.2 (1.9, 2.9) 0.2
		1.4	92.5 (4.7, 2.8) 0.2	96.9 (0.5, 2.6) 0.2	95.2 (1.9, 2.9) 0.2
	200	1.0	91.5 (5.7, 2.8) 0.1	95.9 (1.5, 2.6) 0.2	93.7 (3.4, 2.9) 0.2
		1.4	91.6 (5.2, 3.2) 0.1	95.5 (1.6, 2.9) 0.2	93.5 (3.3, 3.2) 0.2

Table 4.15: Empirical coverage (%), tail errors ((<, >)%), and median widths (WD) for $P(Y_1 > Y_2) = 0.5$ when the number of clusters per arm equal 12 (control) and 6 (experimental), $\alpha = 0.05$, imbalance parameter=0.8, and runs=1000.

ICC	m_{ij}	$\frac{\sigma^2_{T_1}}{\sigma^2_{T_2}}$	Wald	GCI	MOVER
			Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% Width
0.005	50	1.0	93.4 (3.3, 3.3) 0.1	97.6 (1.3, 1.1) 0.1	93.3 (3.4, 3.3) 0.1
		1.4	92.9 (3.5, 3.6) 0.1	97.3 (1.4, 1.3) 0.1	92.9 (3.5, 3.6) 0.1
	100	1.0	94.0 (2.9, 3.1) 0.1	96.8 (1.5, 1.7) 0.1	94.0 (2.9, 3.1) 0.1
		1.4	93.3 (3.1, 3.6) 0.1	96.9 (1.5, 1.6) 0.1	93.3 (3.1, 3.6) 0.1
	200	1.0	94.9 (2.5, 2.6) 0.1	98.3 (0.9, 0.8) 0.1	94.9 (2.5, 2.6) 0.1
		1.4	94.8 (2.4, 2.8) 0.1	98.2 (1.2, 0.6) 0.1	94.8 (2.4, 2.8) 0.1
0.01	50	1.0	93.3 (3.4, 3.3) 0.1	97.6 (1.1, 1.3) 0.1	93.3 (3.4, 3.3) 0.1
		1.4	92.8 (3.5, 3.7) 0.1	97.4 (1.3, 1.3) 0.1	92.8 (3.5, 3.7) 0.1
	100	1.0	92.4 (3.8, 3.8) 0.1	96.9 (1.2, 1.9) 0.1	92.4 (3.8, 3.8) 0.1
		1.4	92.5 (3.6, 3.9) 0.1	96.8 (1.4, 1.8) 0.1	92.5 (3.6, 3.9) 0.1
	200	1.0	94.7 (2.6, 2.7) 0.1	98.3 (0.9, 0.8) 0.1	94.7 (2.6, 2.7) 0.1
		1.4	94.3 (2.8, 2.9) 0.1	98.3 (1.0, 0.7) 0.1	94.3 (2.8, 2.9) 0.1
0.1	50	1.0	92.4 (3.5, 4.1) 0.2	96.9 (1.7, 1.4) 0.2	92.4 (3.5, 4.1) 0.2
		1.4	92.3 (3.7, 4.0) 0.2	96.6 (1.7, 1.7) 0.2	92.2 (3.8, 4) 0.2
	100	1.0	92.9 (3.6, 3.5) 0.2	96.0 (2.0, 2.0) 0.2	92.9 (3.6, 3.5) 0.2
		1.4	92.5 (3.8, 3.7) 0.2	95.8 (2.1, 2.1) 0.2	92.5 (3.8, 3.7) 0.2
	200	1.0	94.2 (2.7, 3.1) 0.2	96.9 (1.4, 1.7) 0.2	94.2 (2.7, 3.1) 0.2
		1.4	93.9 (2.8, 3.3), 0.2	97.0 (1.4, 1.6) 0.2	93.9 (2.8, 3.3) 0.2
0.2	50	1.0	93.4 (4.3, 2.3) 0.1	95.9 (0.8, 3.3) 0.1	96.3 (1.2, 2.5) 0.1
		1.4	91.6 (5.3, 3.1) 0.1	95.9 (1.0, 3.1) 0.2	94.2 (2.6, 3.2) 0.2
	100	1.0	92.7 (4.7, 2.6) 0.1	96.0 (1.4, 2.6) 0.2	94.1 (3.2, 2.7) 0.1
		1.4	92.1 (4.8, 3.1) 0.1	95.5 (1.6, 2.9) 0.2	93.8 (3.1, 3.1) 0.1
	200	1.0	94.4 (3.9, 1.7) 0.1	96.5 (1.3, 2.2) 0.1	96.1 (2.2, 1.7) 0.1
		1.4	94.1 (4.0, 1.9) 0.1	96.5 (1.3, 2.2) 0.2	95.6 (2.5, 1.9) 0.1

Table 4.16: Empirical coverage (%), tail errors ((<, >)%), and median widths (WD) for $P(Y_1 > Y_2) = 0.5$ when the number of clusters per arm equal 12 (control) and 12 (experimental), $\alpha = 0.05$, imbalance parameter=0.8, and runs=1000.

ICC	m_{ij}	$\frac{\sigma^2_{T_1}}{\sigma^2_{T_2}}$	Wald	GCI	MOVER
			Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% Width
0.005	50	1.0	94.0 (3.2, 2.8) 0.1	96.0 (2.2, 1.8) 0.1	94.0 (3.2, 2.8) 0.1
		1.4	94.0 (3.2, 2.8) 0.1	96.1 (2.2, 1.7) 0.1	94.0 (3.2, 2.8) 0.1
	100	1.0	94.7 (3.0, 2.3) 0.1	97.5 (1.4, 1.1) 0.1	94.7 (3.0, 2.3) 0.1
		1.4	94.8 (3.2, 2.0) 0.1	97.2 (1.6, 1.2) 0.1	94.8 (3.2, 2.0) 0.1
	200	1.0	94.7 (2.6, 2.7) < 0.1	96.7 (1.7, 1.6) 0.1	94.7 (2.6, 2.7) < 0.1
		1.4	94.5 (2.7, 2.8) < 0.1	96.5 (1.7, 1.8) 0.1	94.5 (2.7, 2.8) < 0.1
0.01	50	1.0	94.1 (3.1, 2.8) 0.1	96.2 (2.0, 1.8) 0.1	94.0 (3.1, 2.9) 0.1
		1.4	93.7 (3.2, 3.1) 0.1	95.8 (2.3, 1.9) 0.1	93.7 (3.2, 3.1) 0.1
	100	1.0	94.6 (2.9, 2.5) 0.1	96.8 (1.5, 1.7) 0.1	94.6 (2.9, 2.5) 0.1
		1.4	94.4 (3.0, 2.6) 0.1	96.7 (1.6, 1.7) 0.1	94.4 (3.0, 2.6) 0.1
	200	1.0	94.1 (3.0, 2.9) 0.1	96.1 (1.8, 2.1) 0.1	94.1 (3.0, 2.9) 0.1
		1.4	93.8 (3.0, 3.2) 0.1	95.9 (1.9, 2.2) 0.1	93.8 (3.0, 3.2) 0.1
0.1	50	1.0	94.1 (2.7, 3.2) 0.2	96.0 (1.8, 2.2) 0.2	94.1 (2.7, 3.2) 0.2
		1.4	93.8 (2.7, 3.5) 0.2	95.8 (1.8, 2.4) 0.2	93.8 (2.7, 3.5) 0.2
	100	1.0	94.0 (2.9, 3.1) 0.1	95.9 (1.8, 2.3) 0.2	94.0 (2.9, 3.1) 0.1
		1.4	94.1 (3.0, 2.9) 0.1	95.6 (2.0, 2.4) 0.2	94.1 (3.0, 2.9) 0.1
	200	1.0	93.8 (2.8, 3.4) 0.1	95.8 (1.9, 2.3) 0.2	93.8 (2.8, 3.4) 0.1
		1.4	93.6 (3.1, 3.3) 0.1	95.7 (2.0, 2.3) 0.2	93.6 (3.1, 3.3) 0.1
0.2	50	1.0	95.7 (2.8, 1.5) 0.1	96.5 (0.8, 2.7) 0.1	96.8 (1.5, 1.7) 0.1
		1.4	94.6 (3.4, 2.0) 0.1	96.0 (1.0, 3.0) 0.1	95.6 (2.3, 2.1) 0.1
	100	1.0	94.2 (3.6, 2.2) 0.1	96.2 (0.9, 2.9) 0.1	95.8 (1.9, 2.3) 0.1
		1.4	94.5 (3.2, 2.3) 0.1	96.1 (1.0, 2.9) 0.1	95.2 (2.3, 2.5) 0.1
	200	1.0	93.5 (4.0, 2.5) 0.1	96.0 (1.3, 2.7) 0.1	94.3 (3.1, 2.6) 0.1
		1.4	93.6 (4.0, 2.4) 0.1	95.4 (1.6, 3.0) 0.1	94.4 (3.2, 2.4) 0.1

Table 4.17: Empirical coverage (%), tail errors ((<, >)%), and median widths (WD) for $P(Y_1 > Y_2) = 0.5$ when the number of clusters per arm equal 24 (control) and 12 (experimental), $\alpha = 0.05$, imbalance parameter=0.8, and runs=1000.

ICC	m_{ij}	$\frac{\sigma^2_{T_1}}{\sigma^2_{T_2}}$	Wald	GCI	MOVER
			Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% Width
0.005	50	1.0	94.8 (2.2, 3.0) 0.1	96.3 (1.8, 1.9) 0.1	94.8 (2.2, 3.0) 0.1
		1.4	94.4 (2.5, 3.1) 0.1	96.3 (2.0, 1.7) 0.1	94.4 (2.5, 3.1) 0.1
	100	1.0	93.7 (3.2, 3.1) 0.1	95.7 (2.0, 2.3) 0.1	93.7 (3.2, 3.1) 0.1
		1.4	94.2 (3.1, 2.7) 0.1	95.6 (2.2, 2.2) 0.1	94.2 (3.1, 2.7) 0.1
	200	1.0	93.8 (2.9, 3.3) < 0.1	95.6 (2.1, 2.3) < 0.1	93.8 (2.9, 3.3) < 0.1
		1.4	93.7 (3.1, 3.2) < 0.1	95.7 (2.0, 2.3) < 0.1	93.7 (3.1, 3.2) < 0.1
0.01	50	1.0	94.5 (2.7, 2.8) 0.1	96.3 (2.0, 1.7) 0.1	94.5 (2.7, 2.8) 0.1
		1.4	94.5 (2.7, 2.8) 0.1	96.0 (2.1, 1.9) 0.1	94.5 (2.7, 2.8) 0.1
	100	1.0	93.6 (3.2, 3.2) 0.1	95.7 (2.3, 2.0) 0.1	93.6 (3.2, 3.2) 0.1
		1.4	93.6 (3.2, 3.2) 0.1	95.5 (2.4, 2.1) 0.1	93.6 (3.2, 3.2) 0.1
	200	1.0	94.3 (2.6, 3.1) 0.1	95.7 (2.1, 2.2) 0.1	94.3 (2.6, 3.1) 0.1
		1.4	94.1 (2.6, 3.3) 0.1	96.2 (2.0, 1.8) 0.1	94.1 (2.6, 3.3) 0.1
0.1	50	1.0	93.6 (3.5, 2.9) 0.1	96.1 (2.1, 1.8) 0.1	93.6 (3.5, 2.9) 0.1
		1.4	93.6 (3.7, 2.7) 0.1	96.2 (2.2, 1.6) 0.1	93.5 (3.7, 2.8) 0.1
	100	1.0	93.1 (3.5, 3.4) 0.1	95.1 (2.6, 2.3) 0.1	93.1 (3.5, 3.4) 0.1
		1.4	92.9 (3.8, 3.3) 0.1	94.9 (2.9, 2.2) 0.1	92.9 (3.8, 3.3) 0.1
	200	1.0	94.6 (2.6, 2.8) 0.1	96.1 (1.8, 2.1) 0.1	94.4 (2.8, 2.8) 0.1
		1.4	94.8 (2.5, 2.7) 0.1	96.5 (1.6, 1.9) 0.1	94.8 (2.5, 2.7) 0.1
0.2	50	1.0	93.7 (4.2, 2.1) 0.1	94.4 (2.5, 3.1) 0.1	94.5 (3.1, 2.4) 0.1
		1.4	92.9 (5.1, 2.0) 0.1	95.4 (1.9, 2.7) 0.1	94.0 (4.0, 2.0) 0.1
	100	1.0	93.5 (4.2, 2.3) 0.1	95.2 (1.9, 2.9) 0.1	94.3 (3.3, 2.4) 0.1
		1.4	93.1 (4.4, 2.5) 0.1	95.1 (2.1, 2.8) 0.1	93.8 (3.7, 2.5) 0.1
	200	1.0	94.5 (3.4, 2.1) 0.1	96.4 (0.9, 2.7) 0.1	95.4 (2.4, 2.2) 0.1
		1.4	94.5 (3.5, 2.0) 0.1	96.2 (1.3, 2.5) 0.1	95.7 (2.2, 2.1) 0.1

Table 4.18: Empirical coverage (%), tail errors ((<, >)%), and median widths (WD) for $P(Y_1 > Y_2) = 0.5$ when the number of clusters per arm equal 24 (control) and 24 (experimental), $\alpha = 0.05$, imbalance parameter=0.8, and runs=1000.

ICC	m_{ij}	$\frac{\sigma^2_{T_1}}{\sigma^2_{T_2}}$	Wald	GCI	MOVER
			Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% Width
0.005	50	1.0	94.3 (2.7, 3.0) 0.1	95.2 (2.2, 2.6) 0.1	94.3 (2.7, 3.0) 0.1
		1.4	94.3 (2.7, 3.0) 0.1	95.3 (2.0, 2.7) 0.1	94.2 (2.7, 3.1) 0.1
	100	1.0	94.9 (2.1, 3.0) < 0.1	95.7 (1.7, 2.6) < 0.1	94.9 (2.1, 3.0) < 0.1
		1.4	95.0 (2.0, 3.0) < 0.1	95.7 (1.6, 2.7) < 0.1	95.0 (2.0, 3.0) < 0.1
	200	1.0	94.1 (3.1, 2.8) < 0.1	95.3 (2.4, 2.3) < 0.1	94.1 (3.1, 2.8) < 0.1
		1.4	93.5 (3.3, 3.2) < 0.1	94.8 (2.4, 2.8) < 0.1	93.5 (3.3, 3.2) < 0.1
0.01	50	1.0	94.7 (2.3, 3.0) 0.1	95.3 (2.0, 2.7) 0.1	94.7 (2.3, 3.0) 0.1
		1.4	95.3 (2.0, 2.7) 0.1	95.7 (1.9, 2.4) 0.1	95.3 (2.0, 2.7), 0.1
	100	1.0	94.0 (2.4, 3.6) < 0.1	95.5 (1.5, 3.0) 0.1	94.0 (2.4, 3.6) < 0.1
		1.4	94.1 (2.4, 3.5) < 0.1	94.6 (2.1, 3.3) 0.1	94.1 (2.4, 3.5) < 0.1
	200	1.0	93.2 (3.2, 3.6) < 0.1	94.7 (2.5, 2.8) < 0.1	93.2 (3.2, 3.6) < 0.1
		1.4	93.2 (3.5, 3.3) < 0.1	94.0 (2.9, 3.1) < 0.1	93.2 (3.5, 3.3) < 0.1
0.1	50	1.0	94.1 (3.3, 2.6) 0.1	94.4 (3.2, 2.4) 0.1	94.1 (3.3, 2.6) 0.1
		1.4	94.1 (3.1, 2.8) 0.1	94.8 (2.5, 2.7) 0.1	94.1 (3.1, 2.8) 0.1
	100	1.0	94.9 (2.7, 2.4) 0.1	95.4 (2.3, 2.3) 0.1	94.9 (2.7, 2.4) 0.1
		1.4	94.8 (2.6, 2.6) 0.1	95.4 (2.5, 2.1) 0.1	94.8 (2.6, 2.6) 0.1
	200	1.0	94.6 (2.8, 2.6) 0.1	95.9 (2.1, 2.0) 0.1	94.6 (2.8, 2.6) 0.1
		1.4	94.4 (2.9, 2.7) 0.1	95.3 (2.3, 2.4), 0.1	94.4 (2.9, 2.7) 0.1
0.2	50	1.0	94.6 (3.2, 2.2) 0.1	94.7 (1.8, 3.5) 0.1	95.5 (2.0, 2.5) 0.1
		1.4	94.0 (3.7, 2.3) 0.1	94.7 (2.1, 3.2) 0.1	94.6 (2.9, 2.5) 0.1
	100	1.0	94.7 (3.0 , 2.3) 0.1	95.5 (1.9, 2.6) 0.1	95.1 (2.6, 2.3) 0.1
		1.4	94.8 (2.9, 2.3) 0.1	95.0 (2.1, 2.9) 0.1	94.9 (2.7, 2.4) 0.1
	200	1.0	94.5 (3.9, 1.6) 0.1	95.9 (1.7, 2.4) 0.1	95.1 (3.2, 1.7) 0.1
		1.4	94.0 (3.9, 2.1) 0.1	95.3 (1.6, 3.1) 0.1	94.1 (3.6, 2.3) 0.1

Table 4.19: Empirical coverage (%), tail errors ((<, >)%), and median widths (WD) for $P(Y_1 > Y_2) = 0.9$ when the number of clusters per arm equal 6 (control) and 6 (experimental), $\alpha = 0.05$, imbalance parameter=0.8, and runs=1000.

ICC	m_{ij}	$\frac{\sigma^2_{T_1}}{\sigma^2_{T_2}}$	Wald	GCI	MOVER
			Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% Width
0.005	50	1.0	93.8 (3.5, 2.7) 0.1	97.3 (0.4, 2.3) 0.1	94.6 (2.7, 2.7) 0.1
	50	1.4	93.3 (3.7, 3.0) 0.1	97.2 (0.5, 2.3) 0.1	94.4 (2.6, 3.0) 0.1
	100	1.0	93.5 (2.9, 3.6) < 0.1	97.3 (0.6, 2.1) 0.1	93.9 (2.5, 3.6) < 0.1
	100	1.4	93.4 (2.9, 3.7) < 0.1	97.3 (0.6, 2.1) 0.1	93.9 (2.5, 3.6) < 0.1
	200	1.0	92.8 (3.9, 3.3) < 0.1	97.7 (0.9, 1.4) < 0.1	93.2 (3.5, 3.3) < 0.1
	200	1.4	92.9 (3.7, 3.4) < 0.1	97.5 (0.8, 1.7) < 0.1	93.3 (3.3, 3.4) < 0.1
0.01	50	1.0	93.6 (3.6, 2.8) 0.1	97.5 (0.5, 2.0) 0.1	94.8 (2.4, 2.8) 0.1
	50	1.4	93.4 (3.5, 3.1) 0.1	97.0 (0.6, 2.4) 0.1	94.5 (2.5, 3.0) 0.1
	100	1.0	93.3 (2.8, 3.9) < 0.1	97.2 (0.8, 2.0) 0.1	93.9 (2.2, 3.9) < 0.1
	100	1.4	93.2 (2.7, 4.1) < 0.1	96.8 (1.0, 2.2) 0.1	93.7 (2.3, 4.0) < 0.1
	200	1.0	92.2 (4.3, 3.5) < 0.1	97.2 (1.2, 1.6) < 0.1	92.7 (4.0, 3.3) < 0.1
	200	1.4	91.9 (4.7, 3.4) < 0.1	96.8 (1.3, 1.9) < 0.1	92.2 (4.4, 3.4) < 0.1
0.1	50	1.0	93.9 (3.3, 2.8) 0.1	97.0 (0.4, 2.6) 0.1	95.6 (1.6, 2.8) 0.1
	50	1.4	93.7 (3.3, 3.0) 0.1	96.9 (0.4, 2.7) 0.1	95.1 (1.9, 3.0) 0.1
	100	1.0	92.5 (4.2, 3.3) 0.1	96.7 (0.5, 2.8) 0.1	94.1 (2.6, 3.3) 0.1
	100	1.4	92.9 (4.2, 2.9) 0.1	97.0 (0.3, 2.7) 0.1	94.9 (2.1, 3.0) 0.1
	200	1.0	91.5 (5.6, 2.9) 0.1	95.9 (1.7, 2.4) 0.1	93.2 (3.8, 3.0) 0.1
	200	1.4	91.0 (5.9, 3.1) 0.1	95.6 (1.7, 2.7) 0.1	93.3 (3.4, 3.3) 0.1
0.2	50	1.0	94.1 (3.8, 2.1) 0.1	96.8 (0.3, 2.9) 0.2	96.2 (1.5, 2.3) 0.2
	50	1.4	93.7 (3.8, 2.5) 0.1	96.3 (0.4, 3.3) 0.2	96.1 (1.3, 2.6) 0.2
	100	1.0	92.4 (4.9, 2.7) 0.1	96.3 (0.5, 3.2) 0.2	95.2 (1.9, 2.9) 0.2
	100	1.4	92.8 (4.6, 2.6) 0.1	96.7 (0.5, 2.8) 0.2	95.4 (1.9, 2.7) 0.2
	200	1.0	91.5 (5.7, 2.8) 0.1	95.9 (1.5, 2.6) 0.2	93.7 (3.4, 2.9) 0.2
	200	1.4	91.8 (5.2, 3.0) 0.1	95.5 (1.6, 2.9) 0.2	93.9 (3.1, 3.0) 0.1

Table 4.20: Empirical coverage (%), tail errors ((<, >)%), and median widths (WD) for $P(Y_1 > Y_2) = 0.9$ when the number of clusters per arm equal 12 (control) and 6 (experimental), $\alpha = 0.05$, imbalance parameter=0.8, and runs=1000.

ICC	m_{ij}	$\frac{\sigma^2_{T_1}}{\sigma^2_{T_2}}$	Wald	GCI	MOVER
			Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% Width
0.005	50	1.0	93.3 (4.2, 2.5) < 0.1	96.3 (1.3, 2.4) 0.1	93.9 (3.6, 2.5) 0.1
	50	1.4	92.9 (4.2, 2.9) 0.1	96.1 (1.6, 2.3) 0.1	93.7 (3.3, 3.0) 0.1
	100	1.0	93.3 (3.6, 3.1) < 0.1	96.5 (1.1, 2.4) < 0.1	93.8 (3.1, 3.1) < 0.1
	100	1.4	93.7 (2.9, 3.4) < 0.1	96.3 (1.3, 2.4) < 0.1	94.0 (2.7, 3.3) < 0.1
	200	1.0	95.9 (2.3, 1.8) < 0.1	98.1 (0.7, 1.2) < 0.1	95.9 (2.3, 1.8) < 0.1
	200	1.4	95.5 (2.7, 1.8) < 0.1	98.1 (0.8, 1.1) < 0.1	95.9 (2.4, 1.7) < 0.1
0.01	50	1.0	93.1 (4.1, 2.8) 0.1	96.4 (1.3, 2.3) 0.1	93.6 (3.6, 2.8) 0.1
	50	1.4	92.9 (4.0, 3.1) 0.1	96.5 (1.5, 2.0) 0.1	93.4 (3.5, 3.1) 0.1
	100	1.0	93.8 (3.4, 2.8) < 0.1	96.5 (0.9, 2.6) < 0.1	94.2 (3.0, 2.8) < 0.1
	100	1.4	93.3 (3.5, 3.2) < 0.1	96.4 (1.1, 2.5) 0.1	93.9 (2.9, 3.2) < 0.1
	200	1.0	95.5 (2.5, 2.0) < 0.1	98.1 (0.9, 1.0) < 0.1	95.7 (2.3, 2.0) < 0.1
	200	1.4	95.3 (2.8, 1.9) < 0.1	98.1 (0.9, 1.0) < 0.1	95.8 (2.4, 1.8) < 0.1
0.1	50	1.0	92.5 (4.6, 2.9) 0.1	96.5, (0.8, 2.7) 0.1	94.0 (3.1, 2.9) 0.1
	50	1.4	92.1 (5.2, 2.7) 0.1	95.8, (1.3, 2.9) 0.1	93.8 (3.4, 2.8) 0.1
	100	1.0	93.0 (4.2, 2.8) 0.1	96.2, (1.5, 2.3) 0.1	94.1 (3.1, 2.8) 0.1
	100	1.4	92.6 (4.3, 3.1) 0.1	95.8, (1.5, 2.7) 0.1	93.5 (3.3, 3.2) 0.1
	200	1.0	94.5 (3.4, 2.1) 0.1	96.7, (1.4, 1.9) 0.1	95.8 (2.1, 2.1) 0.1
	200	1.4	94.1 (3.8, 2.1) 0.1	96.8, (1.2, 2.0) 0.1	95.6 (2.2, 2.2) 0.1
0.2	50	1.0	93.1 (4.5, 2.4) 0.1	96.2, (0.8, 3.0) 0.2	95.3 (2.2, 2.5) 0.1
	50	1.4	91.6 (5.4, 3.0) 0.1	96.0, (0.9, 3.1) 0.2	94.8 (2.2, 3.0) 0.1
	100	1.0	92.7 (4.7, 2.6) 0.1	96.0, (1.4, 2.6) 0.2	94.1 (3.2, 2.7) 0.1
	100	1.4	92.2 (4.9, 2.9) 0.1	95.6, (1.5, 2.9) 0.2	94.1 (3.0, 2.9) 0.1
	200	1.0	94.4 (3.9, 1.7) 0.1	96.5, (1.3, 2.2) 0.1	96.1 (2.2, 1.7) 0.1
	200	1.4	94.1 (4.1, 1.8) 0.1	96.5, (1.3, 2.2) 0.2	96.0 (2.2, 1.8) 0.1

Table 4.21: Empirical coverage (%), tail errors ((<, >)%), and median widths (WD) for $P(Y_1 > Y_2) = 0.9$ when the number of clusters per arm equal 12 (control) and 12 (experimental), $\alpha = 0.05$, imbalance parameter=0.8, and runs=1000.

ICC	m_{ij}	$\frac{\sigma^2_{T_1}}{\sigma^2_{T_2}}$	Wald	GCI	MOVER
			Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% Width
0.005	50	1.0	94.3 (3.2, 2.5) < 0.1	95.8 (1.3, 2.9) < 0.1	94.7 (2.9, 2.4) < 0.1
	50	1.4	93.7 (3.3, 3.0) < 0.1	95.2 (1.5, 3.3) < 0.1	94.0 (3.2, 2.8) < 0.1
	100	1.0	95.3 (2.5, 2.2) < 0.1	96.8 (1.3, 1.9) < 0.1	95.3 (2.5, 2.2), < 0.1
	100	1.4	95.0 (2.8, 2.2) < 0.1	96.1 (1.7, 2.2) < 0.1	95.1 (2.7, 2.2) < 0.1
	200	1.0	95.8 (2.5, 1.7) < 0.1	97.1 (1.7, 1.2) < 0.1	95.9 (2.4, 1.7) < 0.1
	200	1.4	95.4 (2.8, 1.8) < 0.1	96.7 (1.7, 1.6) < 0.1	95.5 (2.7, 1.8) < 0.1
0.01	50	1.0	94.1 (3.3, 2.6) < 0.1	95.7 (1.4, 2.9) < 0.1	94.3 (3.1, 2.6) < 0.1
	50	1.4	93.6 (3.5, 2.9) < 0.1	95.2 (1.7, 3.1) < 0.1	93.8 (3.3, 2.9) < 0.1
	100	1.0	95.4 (2.9, 1.7) < 0.1	96.6 (1.5, 1.9) < 0.1	95.4 (2.9, 1.7) < 0.1
	100	1.4	94.8 (3.0, 2.2) < 0.1	96.6 (1.5, 1.9) < 0.1	94.9 (2.9, 2.2) < 0.1
	200	1.0	95.2 (2.9, 1.9) < 0.1	96.4 (1.9, 1.7) < 0.1	95.2 (2.9, 1.9) < 0.1
	200	1.4	94.8 (3.1, 2.1) < 0.1	96.5 (2.0, 1.5) < 0.1	94.9 (3.0, 2.1) < 0.1
0.1	50	1.0	94.1 (3.2, 2.7) 0.1	95.6 (1.1, 3.3) 0.1	94.3 (2.9, 2.8) 0.1
	50	1.4	94.2 (3.2, 2.6) 0.1	95.4 (1.4, 3.2) 0.1	94.5 (2.8, 2.7) 0.1
	100	1.0	94.6 (3.1, 2.3) 0.1	96.2 (1.2, 2.6) 0.1	95.5 (2.2, 2.3) 0.1
	100	1.4	94.8 (3.0, 2.2) 0.1	95.9 (1.1, 3.0) 0.1	95.4 (2.4, 2.2) 0.1
	200	1.0	93.6 (3.5, 2.9) 0.1	95.8 (1.4, 2.8) 0.1	94.0 (2.9, 3.1) 0.1
	200	1.4	93.5 (3.8, 2.7) 0.1	95.8 (1.6, 2.6) 0.1	94.1 (3.2, 2.7) 0.1
0.2	50	1.0	94.6 (3.3, 2.1) 0.1	96.1 (0.9, 3.0) 0.1	95.2 (2.7, 2.1) 0.1
	50	1.4	94.6 (3.4, 2.0) 0.1	95.9 (1.0, 3.1) 0.1	95.6 (2.3, 2.1) 0.1
	100	1.0	94.2 (3.6, 2.2) 0.1	96.2 (0.9, 2.9) 0.1	95.8 (1.9, 2.3) 0.1
	100	1.4	94.7 (3.2, 2.1) 0.1	96.4 (1.0, 2.6) 0.1	95.2 (2.3, 2.5) 0.1
	200	1.0	93.5 (4.0, 2.5) 0.1	96.0 (1.3, 2.7) 0.1	94.3 (3.1, 2.6) 0.1
	200	1.4	93.6 (4.1, 2.3) 0.1	95.7 (1.4, 2.9) 0.1	94.4 (3.2, 2.4) 0.1

Table 4.22: Empirical coverage (%), tail errors ((<, >)%), and median widths (WD) for $P(Y_1 > Y_2) = 0.9$ when the number of clusters per arm equal 24 (control) and 12 (experimental), $\alpha = 0.05$, imbalance parameter=0.8, and runs=1000.

ICC	m_{ij}	$\frac{\sigma^2_{T_1}}{\sigma^2_{T_2}}$	Wald	GCI	MOVER
			Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% Width
0.005	50	1.0	94.3 (2.7, 3.0) < 0.1	95.5 (1.1, 3.4) < 0.1	94.5, (2.5, 3.0) < 0.1
	50	1.4	94.7 (2.5, 2.8) < 0.1	95.6 (1.1, 3.3) < 0.1	94.8, (2.4, 2.8) < 0.1
	100	1.0	94.2 (3.4, 2.4) < 0.1	95.6 (2.1, 2.3) < 0.1	94.4, (3.3, 2.3) < 0.1
	100	1.4	94.0 (3.5, 2.5) < 0.1	95.5 (2.0, 2.5) < 0.1	94.2, (3.3, 2.5) < 0.1
	200	1.0	94.6 (2.8, 2.6) < 0.1	95.9 (2.0, 2.1) < 0.1	94.6, (2.8, 2.6) < 0.1
	200	1.4	94.9 (2.7, 2.4) < 0.1	95.8 (2.0, 2.2) < 0.1	94.9, (2.7, 2.4) < 0.1
0.01	50	1.0	94.3 (2.8, 2.9) < 0.1	95.6 (1.1, 3.3) < 0.1	94.6, (2.5, 2.9) < 0.1
	50	1.4	94.6 (2.6, 2.8) < 0.1	95.5 (1.2, 3.3) < 0.1	94.7, (2.5, 2.8) < 0.1
	100	1.0	94.1 (3.4, 2.5) < 0.1	95.5 (2.1, 2.4) < 0.1	94.2, (3.3, 2.5) < 0.1
	100	1.4	93.8 (3.6, 2.6) < 0.1	95.4 (2.0, 2.6) < 0.1	94.0, (3.4, 2.6) < 0.1
	200	1.0	94.9 (2.9, 2.2) < 0.1	96.0 (2.0, 2.0) < 0.1	95.1, (2.7, 2.2) < 0.1
	200	1.4	94.9 (2.8, 2.3) < 0.1	95.6 (2.0, 2.4) < 0.1	94.9, (2.8, 2.3) < 0.1
0.1	50	1.0	93.6 (4.0, 2.4) 0.1	95.1 (1.9, 3.0), 0.1	94.0 (3.6, 2.4) 0.1
	50	1.4	93.4 (4.4, 2.2) 0.1	95.3 (2.0, 2.7), 0.1	94.0 (3.7, 2.3) 0.1
	100	1.0	93.3 (4.2, 2.5) 0.1	95.0 (2.4, 2.6), 0.1	93.8 (3.7, 2.5) 0.1
	100	1.4	93.3 (4.2, 2.5) 0.1	95.0 (2.3, 2.7), 0.1	93.6 (3.9, 2.5) 0.1
	200	1.0	94.4 (3.2, 2.4) 0.1	95.5 (1.6, 2.9), 0.1	94.8 (2.7, 2.5) 0.1
	200	1.4	94.6 (3.3, 2.1) 0.1	95.9 (1.5, 2.6), 0.1	94.8 (2.9, 2.3) 0.1
0.2	50	1.0	92.7 (4.9, 2.4) 0.1	95.2 (1.9, 2.9), 0.1	93.9 (3.7, 2.4) 0.1
	50	1.4	92.9 (5.1, 2.0) 0.1	95.3 (1.9, 2.8), 0.1	94.0 (3.9, 2.1) 0.1
	100	1.0	93.5 (4.2, 2.3) 0.1	95.2 (1.9, 2.9), 0.1	94.3 (3.3, 2.4) 0.1
	100	1.4	93.2 (4.4, 2.4) 0.1	95.3 (1.9, 2.8), 0.1	94.0 (3.5, 2.5) 0.1
	200	1.0	94.5 (3.4, 2.1) 0.1	96.4 (0.9, 2.7), 0.1	95.4 (2.4, 2.2) 0.1
	200	1.4	94.5 (3.5, 2.0) 0.1	95.8 (1.4, 2.8), 0.1	95.6 (2.3, 2.1) 0.1

Table 4.23: Empirical coverage (%), tail errors ((<, >)%), and median widths (WD) for $P(Y_1 > Y_2) = 0.9$ when the number of clusters per arm equal 24 (control) and 24 (experimental), $\alpha = 0.05$, imbalance parameter=0.8, and runs=1000.

ICC	m_{ij}	$\frac{\sigma^2_{T_1}}{\sigma^2_{T_2}}$	Wald	GCI	MOVER
			Cov (<, >)% WD	Cov (<, >)% WD	Cov (<, >)% Width
0.005	50	1.0	93.9 (3.3, 2.8) < 0.1	94.2 (2.6, 3.2) < 0.1	94.0 (3.2, 2.8) < 0.1
	50	1.4	93.9 (3.1, 3.0) < 0.1	94.1 (2.4, 3.5) < 0.1	93.9 (3.1, 3.0) < 0.1
	100	1.0	94.7 (2.4, 2.9) < 0.1	95.5 (1.5, 3.0) < 0.1	94.8 (2.3, 2.9) < 0.1
	100	1.4	94.6 (2.6, 2.8) < 0.1	95.8 (1.6, 2.6) < 0.1	94.6 (2.6, 2.8) < 0.1
	200	1.0	93.4 (3.1, 3.5) < 0.1	93.8 (2.7, 3.5) < 0.1	93.4 (3.1, 3.5) < 0.1
	200	1.4	93.2 (3.0, 3.8) < 0.1	94.2 (2.4, 3.4) < 0.1	93.2 (3.0, 3.8) < 0.1
0.01	50	1.0	93.6 (3.7, 2.7) < 0.1	93.9 (2.8, 3.3) < 0.1	93.7 (3.7, 2.6) < 0.1
	50	1.4	94.0 (3.3, 2.7) < 0.1	93.9 (2.6, 3.5) < 0.1	94.1 (3.2, 2.7) < 0.1
	100	1.0	94.9 (2.7, 2.4) < 0.1	95.3 (1.9, 2.8) < 0.1	95.0 (2.6, 2.4) < 0.1
	100	1.4	94.5 (2.9, 2.6) < 0.1	95.2 (2.2, 2.6) < 0.1	94.5 (2.9, 2.6) < 0.1
	200	1.0	93.2 (3.5, 3.3) < 0.1	93.8 (2.9, 3.3) < 0.1	93.3 (3.4, 3.3) < 0.1
	200	1.4	92.8 (3.6, 3.6) < 0.1	93.9 (2.7, 3.4) < 0.1	92.9 (3.5, 3.6) < 0.1
0.1	50	1.0	93.6 (3.9, 2.5) 0.1	94.1 (2.7, 3.2) 0.1	93.8 (3.6, 2.6) 0.1
	50	1.4	93.9 (3.6, 2.5) 0.1	94.5 (2.2, 3.3) 0.1	94.2 (3.1, 2.7) 0.1
	100	1.0	94.6 (3.1, 2.3) < 0.1	95.4 (2.0, 2.6) 0.1	95.0 (2.7, 2.3) 0.1
	100	1.4	94.7 (2.9, 2.4) < 0.1	94.9 (2.4, 2.7) 0.1	94.8 (2.8, 2.4) < 0.1
	200	1.0	94.1 (3.8, 2.1) < 0.1	95.1 (1.7, 3.2) 0.1	94.3 (3.5, 2.2) < 0.1
	200	1.4	93.4 (4.2, 2.4) < 0.1	94.7 (1.9, 3.4) 0.1	93.6 (3.8, 2.6) < 0.1
0.2	50	1.0	93.5 (4.1, 2.4), 0.1	94.6 (2.4, 3.0) 0.1	94.4 (3.2, 2.4) 0.1
	50	1.4	94.3 (3.4, 2.3), 0.1	94.7 (2.1, 3.2) 0.1	94.7 (2.9, 2.4) 0.1
	100	1.0	94.7 (3.0, 2.3), 0.1	95.5 (1.9, 2.6) 0.1	95.1 (2.6, 2.3) 0.1
	100	1.4	94.8 (2.9, 2.3) 0.1	95.0 (2.1, 2.9) 0.1	94.9 (2.6, 2.5) 0.1
	200	1.0	94.5 (3.9, 1.6) 0.1	95.9 (1.7, 2.4) 0.1	95.1, (3.2 1.7) 0.1
	200	1.4	94.2 (3.9, 1.9) 0.1	95.4 (1.6, 3.0) 0.1	94.2 (3.6, 2.2) 0.1

Median width

All three confidence interval procedures show comparable widths, as shown in Tables 4.14 to 4.23. There is a general pattern of increased median width when the ICC increases to 0.1 or 0.2, however this increase is very slight. Furthermore, the generalized confidence interval procedure shows slightly larger widths than the other two procedures when there are 6 or 12 clusters per arm.

4.5 Discussion

Asymptotic procedures such as the ones investigated above improve in performance (coverage, tail errors, and width) as the effective sample size of a study increases. An increase in the number of clusters contributes more to the effective sample size than does the cluster size (Donner, 1998). However, many cluster randomization trials have a small number of large clusters (Donner and Klar, 2001). Inference procedures must therefore be evaluated when the number of clusters is small.

The difference between two normal means

If clustering is ignored when making inferences on a cluster randomization trial, then the results of the trial will be invalid, with elevated type I errors or low confidence interval coverage (Donner and Klar, 1996). To avoid this mistake, the cluster adjusted confidence interval procedure (Donner and Klar, 1993) was introduced. This method inflates the variance estimate using the ICC value and the cluster sizes. Another commonly used procedure is the standard Wald method after adjusting the variance for the effect of clustering. However, the empirical coverage of the Wald method, using the delta method with unweighted mean squared error (Thomas and Hultquist, 1978), and the cluster adjusted confidence interval method have not been previously evaluated in a simulation study for the difference between two normal means. Finally, simulation based procedures, such as the generalized confidence interval procedure

(Weerahandi, 1993) may be applied to the difference between two normal means. Unfortunately, such a procedure is computationally intensive and is not closed form. The MOVER is in closed form and simple to apply, requiring only the limits of each individual mean, which are readily available.

Confidence interval coverage

Simulation results evaluating and comparing the four confidence interval procedures show comparable coverage results of the generalized confidence interval method and the MOVER. Although the coverage results of the Wald procedure improve as the sample size increases, it does not do so as quickly as those of the generalized confidence interval method and the MOVER. The fourth method, the cluster adjusted confidence interval, consistently has very conservative results with no signs of improvement with increasing sample size.

The often anticonservative performance of the Wald method is explained by its use of the standard normal critical value. Although the variance of the MOVER is the same as that of the Wald, the larger critical value of the t-distribution is used rather than that of the standard normal distribution. The MOVER therefore has larger margins of error than the Wald method, in this case leading to less liberal coverage results. The works of El-Bassiouni and Abdelhafez (2000) show that that the t-interval with the variance estimated using the unweighted mean squared error for a single normal mean from a random effects model has coverage close to the nominal for a range of unbalanced designs using a simulation study. Although El-Bassiouni and Abdelhafez (2000) initially used Satterthwaite degrees of freedom (Satterthwaite, 1946), they show that degrees of freedom may be approximated by the number of clusters minus one. The MOVER (setting degrees of freedom to the number of clusters minus one) showed somewhat conservative results when the number of clusters per arm was as small as 6, however improvements were observed as these numbers increased.

It is expected that the MOVER would have similar results to the generalized

confidence intervals described in Krishnamoorthy *et al.* (2007), because both procedures use the unweighted mean squared error (Thomas and Hultquist, 1978) and the t-distribution in their corresponding estimated confidence limits. The key difference between these two procedures is that the MOVER is closed form, while the generalized confidence interval procedure is based on simulation. This renders the MOVER as more easily applicable.

Not only is the MOVER a valid confidence interval procedure and easily applicable, but it solves the common misconceptions of using overlapping confidence intervals (Wolfe and Hanley, 2002; Schenker and Gentleman, 2001; Wilcox, 2003, page 246) to judge statistical significance. The overlap method is both a backwards step to reducing confidence intervals to "yes" or "no" hypothesis tests, and may be incorrect when no statistical significance is declared with two overlapping intervals. With the MOVER, if the $(1-\alpha)\%$ confidence intervals for two separate means are obtained using the unweighted mean squared error (Thomas and Hultquist, 1978), a confidence interval for their difference may easily be obtained, using Equation 2.13.

Confidence interval tail errors

Tail errors for all four confidence interval procedures are relatively balanced. Each confidence interval procedure is symmetric, because the sampling distribution of a difference between two normal means is symmetric. Therefore, none of the procedures stand out as "better" than the other based on confidence interval tail error equality.

The tail errors of the cluster adjusted procedure were smaller than the desired 2.5% on each side, while those of the Wald method were larger. This is explained by the conservative coverage results of the cluster adjusted confidence interval procedure and the liberal results of the Wald procedure.

Confidence interval widths

When there are as many as 24 clusters per arm, the Wald method is comparable to the MOVER in terms of coverage, shifting the focus of comparisons to confidence interval widths. As expected, the median widths of the Wald method are narrower than those of the MOVER. This is explained by the standard normal critical value of the Wald versus the larger critical value of the MOVER from the t-distribution.

The MOVER has comparable widths to those of the generalized confidence interval procedure, which happens to also have comparable coverage for all parameter combinations investigated. Note that the slight difference between median widths is not alarming, but only a difference of a tenth of a decimal place. There is likely to be very little difference from both a numerical and a clinical perspective between the two procedures. The more meaningful difference between the procedures lies in the simplicity of the MOVER.

It should be noted that the increase in the confidence interval widths of all three procedures with heteroscedasticity is at least partially due to the simulation parameters. The overall variance for the difference between two means was increased from 5 units^2 per arm under homoscedasticity to 5 units^2 (in arm 1) and 7 units^2 (in arm 2) under heteroscedasticity when simulating data.

The difference between two lognormal means

The Wald method is an asymptotic confidence interval procedure commonly applied due to its simplicity. However, its assumption of normality forces the limits to be symmetrically placed around the point estimate. Forced symmetry is seen as one of the most serious errors in confidence interval construction (Efron and Tibshirani, 1993, page 180). A remedy for this problem is found in the generalized confidence interval procedure (Weerahandi, 1993), which takes into account the skewness of the parameter estimate, in particular the skewness of the sampling distribution of the

normal variance estimate. However, this method is based on simulation and is thus more complex than the Wald method.

The MOVER has the simplicity advantage of the Wald method, while also taking into account the asymmetric of the sample lognormal mean. This is done using the unweighted mean squared error, presented by Thomas and Hultquist (1978).

Confidence interval coverage

An asymptotic confidence interval procedure for the difference between two lognormal means in cluster randomization trials, utilizing the MOVER and the transformation principle, has been presented. Previous procedures have been symmetric in nature, while the MOVER takes into account the skewness of the underlying lognormal distribution. Note that a further advantage of the MOVER is that it solves the problem of overlapping confidence intervals. Further details have been provided in Section 4.5.1.1.

The MOVER makes the assumption that the confidence intervals of the components are valid. Fortunately, the confidence intervals constructed for each of the components (the normal mean and variance components) have previously been evaluated and shown to have coverage close to nominal (El-Bassiouni and Abdelhafez, 2000; Burdick and Graybill, 1984).

In general, the simulation study suggests that the MOVER performs better than existing procedures. Caution must be taken when there are as few as 6 clusters per arm, as simulations show conservative behavior and wide interval widths for high ICC values. These issues tend to subside when the number of clusters increases to 12 per arm.

As the number of clusters per arm increases, the empirical coverage of the Wald method improves. The results of Flynn and Peters (2004) suggest the Wald method to be valid with the robust sandwich variance estimator (Huber, 1981; White, 1980). However, these results pool the performance of the Wald method for both the dif-

ference between two normal means and the difference between two lognormal means, making it difficult to draw conclusions on any one of these parameters separately. Although Flynn and Peters (2004) used a different variance estimate, their use of the Wald method imposed an assumption of normality and symmetry on the confidence intervals. Our results show that the Wald method is anticonservative for the difference between two lognormal means as the ICC increases to 0.1 or 0.2, particularly when the variances in the two group differ. However the procedure improves as the number of clusters increases to 24 per arm, although low empirical coverage is still observed with high ICC values. Despite its improved validity as the number of clusters increases, the symmetry of the Wald interval consistently causes unbalanced tail errors, rendering the procedure inferior to the generalized confidence interval procedure and the MOVER.

Confidence interval tail errors

As expected, the symmetric Wald method has unbalanced tail errors for all parameter combinations investigated, with the upper tail consistently exceeding the lower. This is an intuitive result, because the lognormal mean (a function of the normal mean and the normal variance) is skewed to the right.

Even with valid coverage results, the unbalanced tail errors of the Wald method suggest that the interval should not be used. One advantage of confidence intervals lies in eliminating extreme values from either side of the point estimate. If the right tail is larger than $\alpha/2$ and the left smaller than $\alpha/2$, then this suggests that unlikely larger values than the point estimate were included within the interval while potentially true values less than the point estimate were excluded.

Furthermore, the upper (U) and lower limits (L) are defined by the conditions that $\mathrm{P}(L \geq \theta) = \alpha/2$ and $\mathrm{P}(U \leq \theta) = \alpha/2$, respectively (Neyman, 1937). That is, confidence intervals set the upper and lower limits such that values above and below these limits would be rejected by hypothesis tests, while values within the

limits would not. If tail errors are unbalanced, that is if the probability conditions above are not satisfied and the corresponding one-sided $\alpha/2$ hypothesis tests are either liberal or conservative, then the limits have failed regardless of whether or not the overall coverage is valid. Confidence interval procedures with valid overall coverage and balanced tail errors will always be preferred.

The results also show that the generalized confidence interval procedure and the MOVER have comparably balanced tail errors, unless the number of clusters in both arms is fewer than 12. Under such conditions, the prevalence of imbalance is rare for the generalized confidence interval procedure and even rarer for the MOVER.

Confidence interval widths

The generalized confidence interval procedure and the MOVER both have generally valid confidence interval methods for the difference between two lognormal means, with similar tail error performance. Inspection of the results shows that the MOVER has slightly narrower widths, indicating greater precision. Again, an increase in median width is observed under heterscedasticity, but this is at least partially due to the increase in overall variance when simulating the data.

The exceedance probability

Although the Wald method is simple to apply, it makes the assumption of normality on the exceedance probability. However, the exceedance probability is a standard normal cumulative distribution function applied to the standardized mean difference, a ratio of the mean and standard deviation. The generalized confidence interval procedure should be better able to account for the shape of the parameter estimate, however it is based on simulation and therefore more complex apply. The MOVER is a closed form procedure which carries the simplicity of the Wald method while able to account for the underlying distribution of the sample exceedance probability.

Confidence interval coverage

The Wald-type confidence intervals with the Wald method and the generalized confidence interval procedure have not previously been evaluated for the exceedance probability. Our results show that when there are only 6 clusters per arm the generalized confidence interval procedure has conservative results and the Wald method and MOVER have liberal results. When the number of clusters increases to 12 in one arm and 6 in the other, the generalized confidence interval procedure and the MOVER both perform well, while the Wald method requires at least 12 clusters per arm to have coverage close to the nominal 95% level.

The improvement of the Wald method with a larger effective sample size is a result of the central limit theorem. The improvements of the generalized confidence interval procedure and the MOVER with an increasing effective sample size is explained by the asymptotic relationship of the unweighted mean squared error to the chi-squared distribution (Thomas and Hultquist, 1978).

A further advantage of the MOVER includes avoidance of the box method for the ratio of two parameters (Briggs *et al.*, 1999). The box method crudely sets the limits of the ratio of θ_1 (l_1, u_1) and θ_2 (l_2, u_2) to $(l_1/u_1, u_2/l_2)$, resulting in conservative limits and should thus be avoided.

Confidence interval tail errors

All three procedures have relatively balanced tail errors when the exceedance probability is set to 0.5, except for the Wald method and the generalized confidence interval procedure when the ICC is as large as 0.2 with less than 24 clusters per arm. The MOVER rarely has unbalanced tail errors, even when the ICC is large. Similar results are observed when the exceedance probability is set to 0.9, except that the tail errors of the generalized confidence interval procedure are unbalanced when there are only 6 clusters per arm, even when the ICC is as low as 0.005.

The results for $P(Y_1 > Y_2) = 0.9$ may by explained by the fact that the tail error performance of the exceedance probability is dependent on the tail error performance of the standardized mean difference, which is a ratio of the normal mean and variance. Such a parameter has a sampling distribution that is skewed to the left.

Confidence interval widths

All three procedures show similar confidence interval widths, with the generalized confidence interval procedure rarely exceeding the width of the MOVER by 0.1. This results in a 10% difference in the exceedance probability. It is up to the investigator to decide whether this difference is clinically significant. Note that the width of the MOVER also exceeded the width of the Wald method by 0.1 when the ICC was set to 0.2. However this occurs when there are a few clusters per arm, when the Wald method has invalid coverage.

4.6 Overall conclusions

The overall conclusion is that the MOVER should be used for all three effect measures of interest. For a difference between two normal means from a cluster randomization trial, the MOVER and the generalized confidence interval procedure performed best, with empirical coverage closest to the nominal. However, the MOVER procedures are simpler to obtain and are closed form, making them more favorable. For the exceedance probability, $P(Y_1 > Y_2)$, the MOVER again has coverage closer to the nominal than the alternative procedures as the number of clusters increases. Caution must be practiced for the difference between two normal means and the exceedance probability when there are only 6 clusters per arm. The MOVER shows some conservative behavior in the former and can be liberal in the latter.

A further advantage of the MOVER is that it may be used to avoid the practice of overlapping confidence intervals when intervals for each individual mean exist and

inferences are desired for their difference. If the unweighted mean squared error is used with the t-distribution for the confidence limits of each individual mean (either normal or lognormal) then the MOVER may be recovered for the difference between two means. Constructing confidence intervals for the difference not only avoids the common misconception that two overlapping confidence intervals suggest statistical insignificance (Schenker and Gentleman, 2001), but also avoids ignoring useful information within confidence intervals by simply treating them as hypothesis tests, as is done by Cumming (2009) amd Maghsoodloo and Huang (2010).

Chapter 5

EXAMPLES

Chapters 1 and 2 introduced the confidence interval procedures investigated in this thesis. Chapter 3 then presented the proposed procedure (the MOVER) for each of the three parameters of interest as well as some existing confidence interval methods. These methods were evaluated for finite sample sizes using Monte Carlo simulations in Chapter 4, showing that the MOVER and the generalized confidence interval procedure were the most reliable in terms of confidence interval coverage with relatively balanced tail errors for each of the parameters. Furthermore, the MOVER often demonstrated greater precision than the generalized confidence interval procedure with narrower widths. This chapter illustrates the methods for each of the effect measures by applying them to datasets arising from published cluster randomization trials.

5.1 The difference between two normal means

Introduction

Data from the cluster randomization trial by Montgomery *et al.* (2000) are used in this section to illustrate confidence interval contruction of a difference between two normal means. Montgomery *et al.* (2000) investigated the effect of a computer based clinical decision support system and risk chart on patient blood pressure. Multiple risk factors exist for cardiovascular disease, including high blood pressure, total body mass index, total cholesterol, smoking, and diabetes (Jackson *et al.*, 1993). It may therefore be difficult for health professionals to estimate cardiovascular risk without

the use of risk charts or computer based clinical decision support systems, which have the advantages of organizing patient information, performing complex evaluations, and presenting results quickly. Thus, the study by Montgomery *et al.* (2000) evaluated the effect of computer decision support systems and risk charts on cardiovascular risk and blood pressure.

Methods

The clinical decision support system uses a patient's sex, age, diabetes status, smoking habits, blood pressure, cholesterol, body mass index, symptomatic cardiovascular disease, family history of ischaemia heart disease, and familial hypercholesterolaemia to calculate the five-year risk of a cardiovascular event. A risk chart provides the same information about cardiovascular risk as that of a computer based clinical decision support system. A cardiovascular event includes a new diagnosis of angina, myocardial infarction, coronary heart disease, stroke, or transient ischaemic attack.

After the practices were stratified by computer system, an independent researcher blinded to the practice identity used a random number generator to randomize twenty-seven general practices to one of three arms: computer based clinical decision support system with cardiovascular risk chart, cardiovascular risk chart alone, or usual care. For the sake of simplicity, only two of the arms will be investigated in this illustration: the computer based clinical decision support system plus risk chart arm and the usual care arm. Also, the trials will be treated as a completely randomized trial for the sake of this illustration.

Eligible patients included those who were 60-80 years of age and who were prescribed anti-hypertensive drugs within the last year. Measurements were taken at baseline and at 12 months follow up. This illustration will use data from the follow-up period to find confidence intervals for a difference between two normal means.

The primary outcome was the percentage of patients in each arm with five year cardiovascular risk $\geq$ 10%. Secondary outcomes included systolic blood pressure,

diastolic blood pressure and amount of cardiovascular drugs prescribed. Confidence intervals will be constructed for the difference between mean systolic blood pressure (SBP) in this illustration because preliminary analyses show that systolic blood pressure in each cluster appears to be approximately normally distributed (see the quantile-quantile plots in Figures 5.1 and 5.2). Note that it may be of interest to also a single quantile-quantile plot of the residuals (see e.g. SAS Institute Inc, 2009, page 530-533), however for the sake of this illustration, Figures 5.1 and 5.2 will suffice.

Results and recommendations

Twenty seven clusters were randomized to the three arms of the study. The two arms of interest in this example contained 17 of those clusters, with 10 clusters in the computer decision support system arm and 7 clusters in the usual care arm. Descriptive statistics at follow-up for these two arms (computer and usual care) are given in Table 5.1. The imbalance statistic for number of participants per practice was close to one, because the trial was originally designed to have balanced practice sizes. However due to eligibility criteria and some loss to follow-up, the imbalance statistics fell slightly below one.

Confidence intervals for the difference in mean systolic blood pressure between the two arms are constructed using the Wald method (and the delta method), the cluster adjusted method (Donner and Klar, 1993), the generalized confidence interval method (Weerahandi, 1993) and the MOVER. Table 5.2 shows the estimated difference in means of the two arms with the four confidence intervals and their corresponding widths.

Although the Wald method has the narrowest width, the empirical coverage results in Chapter 4 suggests the use of either the MOVER or the generalized confidence interval procedure for the analysis of this data. Consistent with the simulation results, the MOVER has a narrower width than the more computationally intensive generalized confidence interval procedure. Therefore, the MOVER is recommended

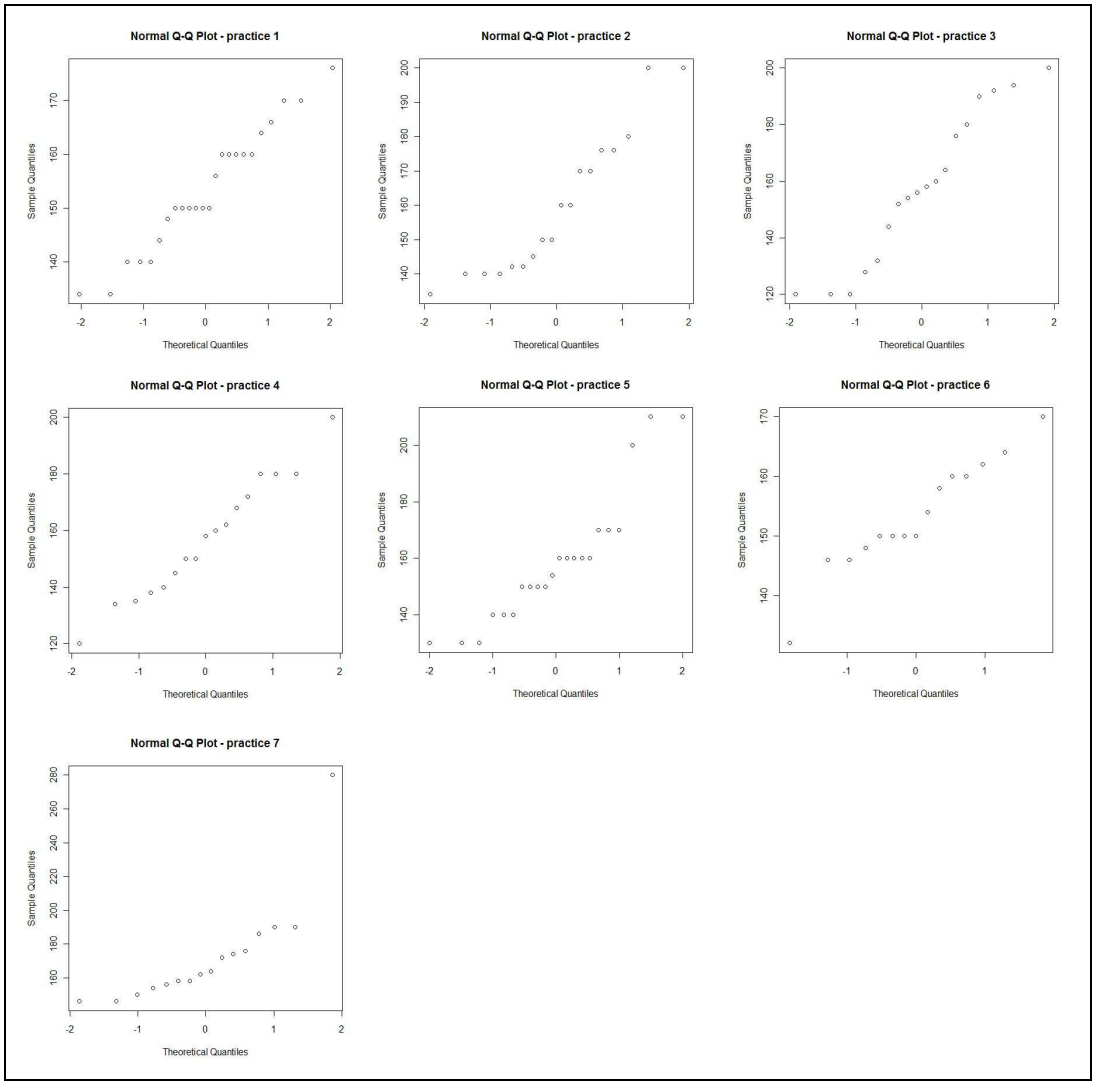

Figure 5.1: Q-Q plots of SBP by practice (7 clusters) in the usual care arm

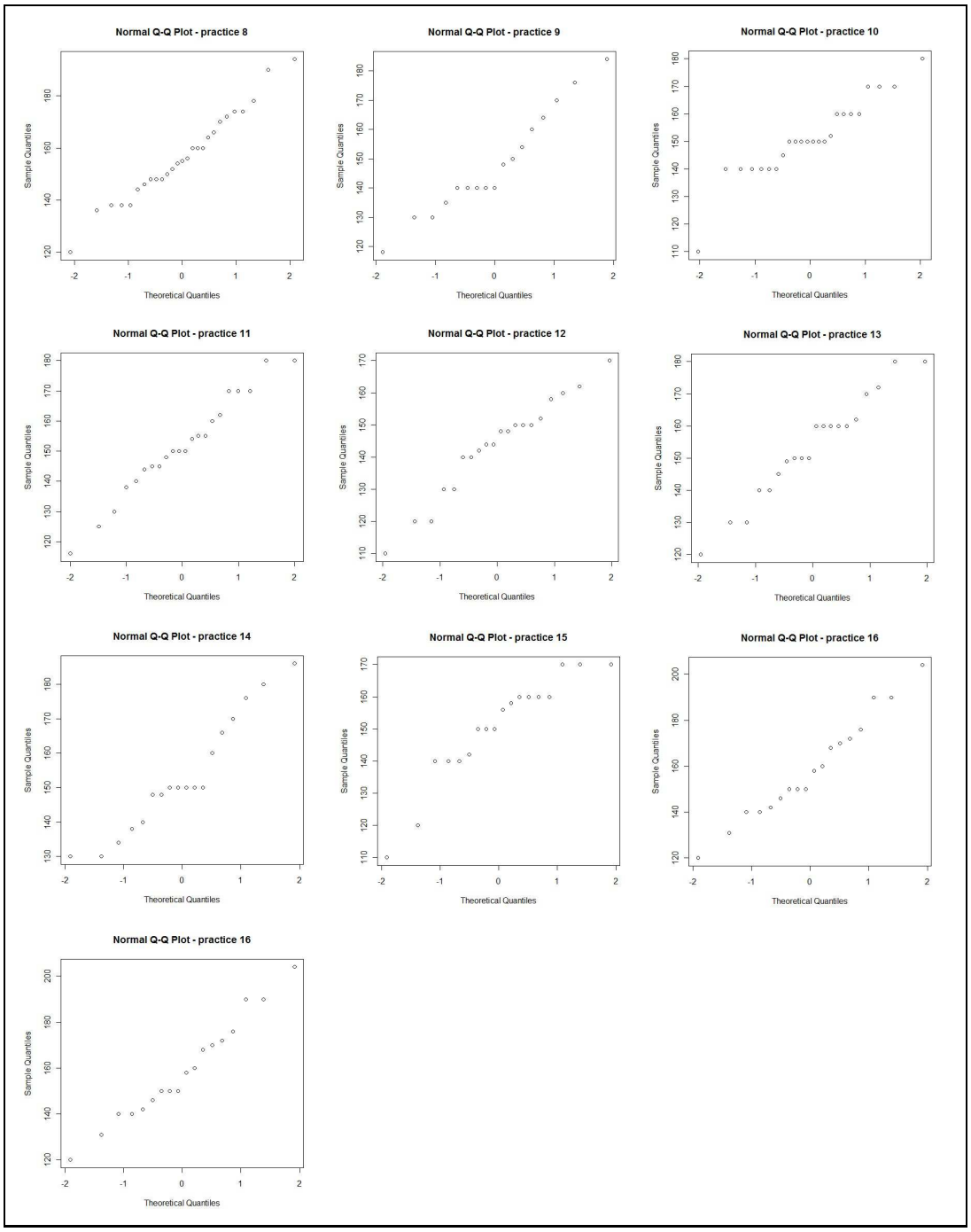

Figure 5.2: Q-Q plots of SBP by practice (10 clusters) in the clinical decision support system with cardiovascular risk chart arm

Table 5.1: Descriptive statistics for the computer clinical decision support system arm and usual care arm of the hypertension study

	Computer support plus chart	Usual care
No. practices (clusters)	10	7
No. patients at follow-up	202	130
Mean no. participants per practice	20	19
Imbalance statistic for practice size	0.98	0.97
ICC	0.08	0.04
Mean SBP (std dev)	153 (17)	159 (22)

for the analysis of this data if interest lies in the difference between the mean systolic blood pressure in the two groups. The data suggest that the true difference in means may lie anywhere from -13 mm Hg to 2 mm Hg with 95% confidence, with an estimated difference of -5.5 mm Hg. Although the confidence interval contains zero, corresponding to a statistically non-significant result, the lower limit of the interval suggests a possible mean systolic blood pressure of 13 mm Hg lower in the computer plus risk chart arm as compared to the usual care arm. Also, note that it is not surprising that the cluster-adjusted procedure has the widest width, as coverage results for this procedure show conservative behavior. The results provided in Table 5.2 are consistent with those found by Montgomery *et al.* (2000).

Collins and Peto (1994) found that a reduction of 10 mm Hg in systolic blood pressure was associated with a $35-40\%$ reduction in stroke and a $20-25\%$ reduction in coronary heart disease. Further research investigating the effect of a computer based clinical decision support system is therefore recommended.

Table 5.2: The Wald-type confidence interval (Wald), cluster-adjusted confidence interval, generalized confidence interval (GCI) and the MOVER for the difference between mean systolic blood pressure (mm Hg) in the treatment arm vs. the control arm.

	Wald	Cluster adjusted	GCI	MOVER
Estimated difference	-5.5	-5.5	-5.5	-5.5
95% CI	$(-11.8, 0.6)$	$(-16.1, 5.0)$	$(-14.0, 2.0)$	$(-13.0, 2.0)$
Width	12.4	21.1	16.0	15.0

5.2 The difference between two lognormal means

Introduction

Each year approximately 600,000 people are admitted to hospital for community acquired pneumonia in the United States, of which approximately 15% die (Bartlett and Mundy, 1995). However, a large amount of variation exists among hospitals in their use of treatment resources.

The analysis of administrative data has revealed a large amount of variation in admission rates, length of hospital stay, and use of institutional resources (Fine *et al.*, 1993; Gilbert *et al.*, 1998). These resources have a high cost to society. Therefore, interventions which improve the care and efficiency of treatment of community acquired pneumonia are desirable.

A critical pathway is a strategy used to define the necessary steps of complex processes. In this case, the complex process is the treatment of community acquired pneumonia. Critical pathways are used to improve the quality of care and/or reduce the cost of care by ensuring that the necessary steps are followed in an efficient manner.

The study by Marrie *et al.* (2000) used a stratified cluster randomization trial

to evaluate the use of a critical pathway for the treatment of community acquired pneumonia in 19 hospitals in the United States. One outcome of the study was length of hospital stay, which was used as a surrogate for resource utilization. This outcome is used here to illustrate the confidence interval procedures for the difference between two lognormal means, since such data usually follow an approximately lognormal distribution (Thompson and Barber, 2000). Note that although a stratified design (based on the type of hospital and historical length of stay) was used, the study will be treated as a completely randomized cluster randomization trial in the analysis, as it was in Marrie *et al.* (2000) .

Methods

The primary hypothesis of the study by Marrie *et al.* (2000) was that institutional resources would be reduced if a critical pathway was implemented for the treatment of community acquired pneumonia, without impairing the safety and efficacy of the therapy. For this investigation of resource utilization, an outcome of interest was the length of hospital stay.

Computer-generated random numbers were used to allocate the 19 participating hospitals to either implement the critical pathway or to continue with conventional management. Hospitals were matched prior to random assignment by type of institution (teaching hospital or community hospital) and historical length of stay. The components of the critical pathway included the use of clinical prediction rule to assist in admission decisions, treatment with antibiotics, and criteria for switching from intravenous to oral antibiotics and hospital discharge.

Hospital charts were used to collect data for resource utilization. The length of stay for patients who died in the hospital was calculated as the admission time to the date of death. Furthermore, the length of stay was arbitrarily censored at 42 days to avoid a large amount of skewness due to patients with extended hospital stays. This cut-off point will not be implemented in this example, because the confidence interval

Table 5.3: Descriptive statistics for the critical pathway versus usual care for the treatment of community acquired pneumonia.

	Critical pathway	Usual care
No. hospitals (clusters)	9	10
No. patients at follow-up	351	587
Mean no. participants per hospital	39	58.7
Imbalance statistic for hospital size	0.73	0.83
ICC (log-scale)	0.01	0.07
Mean length of stay (std dev)	9.41 (13.65)	10.28 (13.20)

procedures do not all assume symmetry.

The mean length of hospital stay is compared between the two arms and confidence intervals for the difference between these two means are constructed. Preliminary analyses show that the log length of stay in each cluster appears to be approximately normally distributed (see the quantile-quantile plots in Figures 5.3 and 5.4). Therefore, the Wald interval, the MOVER, and the generalized confidence interval are constructed at the $\alpha = 5\%$ level for the difference between lognormal means.

Results and recommendations

Nineteen hospitals (clusters) were randomized to two arms, with 9 receiving the critical pathway intervention and 10 receiving usual care for the treatment of community acquired pneumonia. Descriptive statistics of the two arms may be found in Table 5.3. The descriptive statistics found in the table are similar to the parameters investigated in the simulation study (Tables 4.9 to 4.11 in Appendix B).

Confidence intervals for the difference between mean length of hospital stay in the two arms are constructed using the Wald method (and the delta method for variance

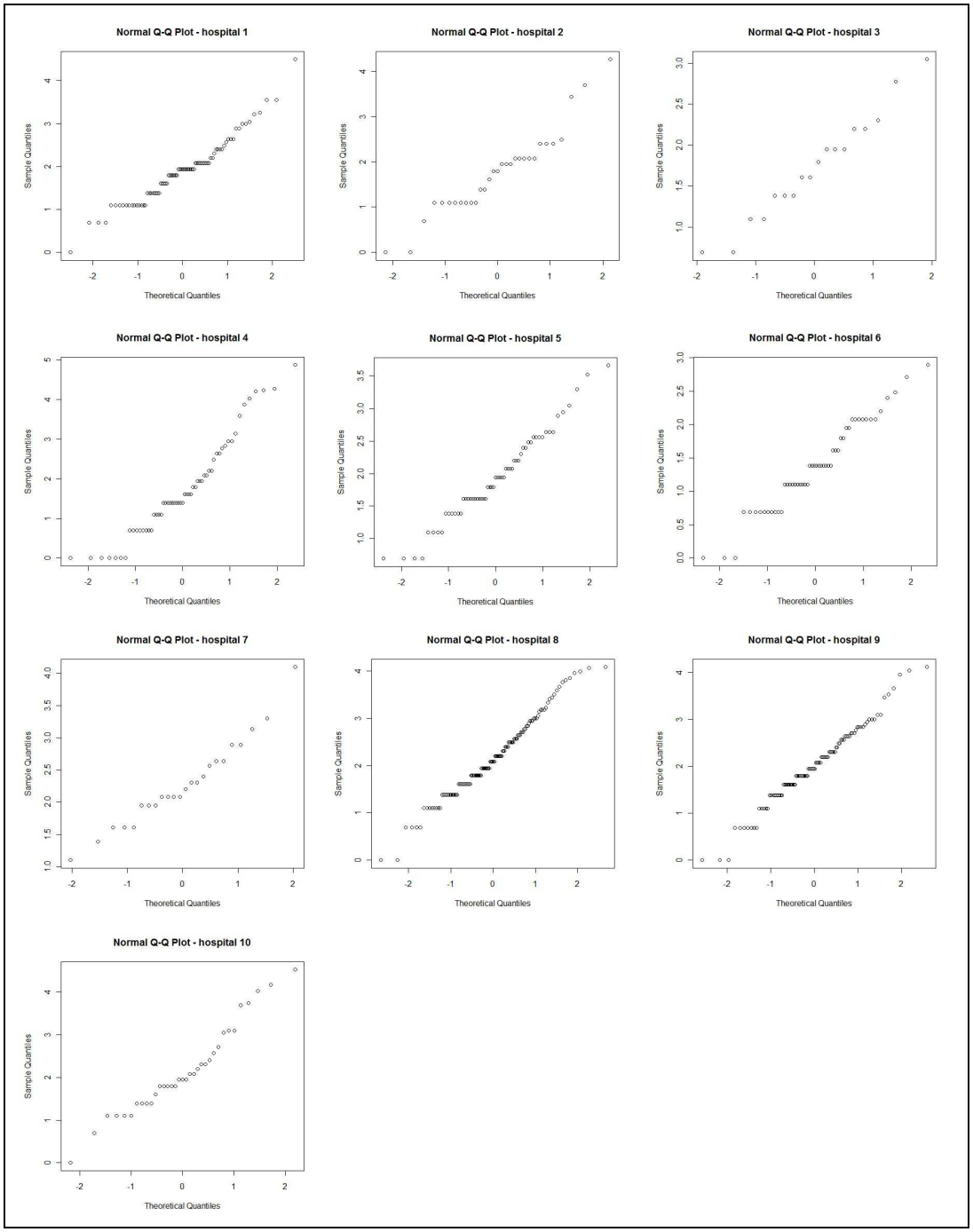

Figure 5.3: Q-Q plots of log length of stay (10 clusters) in the usual care arm

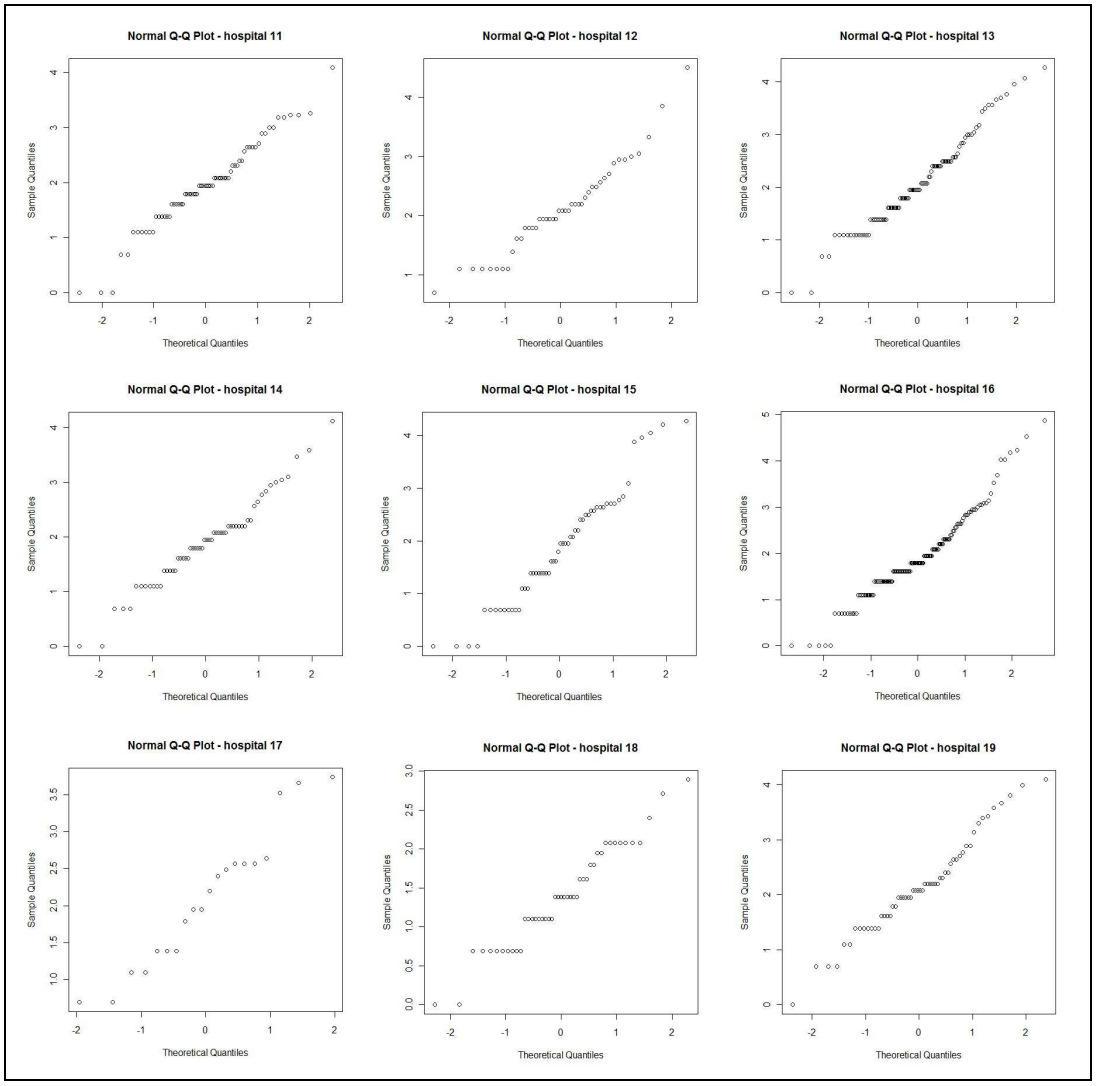

Figure 5.4: Q-Q plots of log length of stay (9 clusters) in the critical pathway arm

Table 5.4: The Wald-type confidence interval (Wald), generalized confidence interval (GCI) and the MOVER for the difference between mean length of stay (in days) in the critical pathway arm and the usual care arm.

	Wald	GCI	MOVER
Estimate	-0.87	-0.87	-0.87
CI	$(-3.23, 1.50)$	$(-3.88, 2.02)$	$(-3.44, 1.59)$
Width	4.73	5.09	5.03

estimation), the generalized confidence interval procedure, and the MOVER. Point estimates and confidence intervals with their corresponding widths for each of the three procedures may be found in Table 5.4.

Table 5.4 shows that the Wald confidence interval has the narrowest width, followed by the MOVER, and finally the generalized confidence interval. Simulation results in Chapter 4 (tables in Appendix B) for the difference between two lognormal means show that both the Wald method and the MOVER have coverage close to the nominal 95% when parameters are similar to those estimated in this example. That is, when the ICC is 0.01 or less, when variances in the two arms are equal (Table 5.4 shows similar standard deviations for the length of stay in the two arms), when the effective sample size (number of clusters per arm and average cluster size) is low, and when the imbalance statistic of the cluster size is roughly 0.8. However, the Wald method imposes symmetry on the interval, thereby resulting in unbalanced tail errors (Appendix B). The generalized confidence interval procedure is not recommended over the other two procedures due to the somewhat conservative coverage behaviour found in the simulation results. Furthermore, this procedure is not as precise as the Wald or the MOVER. The MOVER is therefore recommended for the analysis of this data.

The findings using the confidence interval procedures above are difficult to compare

to those of Marrie *et al.* (2000), because the original study made inferences on the median length of hospital stay rather than the mean due to the skewness of the data. The mean length of stay in the critical pathway arm was found to be 0.87 days lower than that of the usual care arm (9.41 days vs. 10.28 days), with a 95% confidence interval of 3.44 fewer days to 1.59 more days (MOVER). Although an average of 0.87 fewer days seems like a minor difference, when multiplied by the size of the hospital, the monetary savings for each hospital can be quite major, especially over an extended period of time. Note the lack of statistical significance when inferences are performed on the arithmetic mean, as well as a difference potentially as large as 3.44 days between the length of stay in the two arms. Further study is thus recommended.

The original study found that the median length of stay was 1.7 days lower in the intervention arm than the control (5.0 days vs. 6.7 days, p-value= 0.01). The difference between the mean length of stay and median length of stay in each arm is expected when data are lognormally distributed, as the arithmetic mean exceeds the geometric mean (or median). Note that the lognormal mean is a function of both μ and σ^2, while the lognormal median is only a function of μ. This if interest lies in the mean, then inferences on the median are irrelavant and could lead to a misinterpretation of results, as seen in the original study which provided inferences on median length of stay.

The original study found a reduction in the rate of admission in the critical pathway arm as compared to the usual care arm, which combined with the lower length of stay in the treatment arm can potentially save approximately $1700 US per patient admitted to hospital (Niederman *et al.*, 1998; Guest and Morirs, 1997).

5.3 The exceedance probability

Introduction

The confidence intervals for the exceedance probability investigated in Chapter 4 are applied to data from a study investigating the effect of clinical decision support system and risk chart on cardiovascular disease and blood pressure (Montgomery *et al.*, 2000). Computer clinical decision support systems and risk charts help organize patient information, perform complex evaluations, and present results quickly when assessing the risk of cardiovascular disease and high blood pressure. It is desired that the intervention will lead to lower the systolc blood pressure of patients. Background about the study may be found in Section 5.1.1.

Methods

The methods of the study by Montgomery *et al.* (2000) are summarized in Section 5.1.2. Confidence intervals for the exceedance probability are of interest for the secondary outcome, systolic blood pressure. That is, confidence intervals will be constructed for the probability that a randomly selected individual from the usual care arm has a higher systolic blood pressure than that from the computer decision support system and risk chart arm. Quantile-quantile plots in Figures 5.1 and 5.2 show systolic blood pressure outcomes to follow an approximate normal distribution in each clusters. Therefore the procedures outlined in Section 3.3 may be applied to obtain confidence intervals for the exceedance probability.

Results and recommendations

Descriptive statistics for systolic blood pressure in the computer based clinical management and risk chart arm and the usual care arm are presented in Table 5.1. With the number of clusters per arm and the estimated ICC value, empirical results in

Table 5.5: The Wald-type confidence interval (Wald), generalized confidence interval (GCI) and the MOVER (MOVER) for the exceedance probability of systolic blood pressure for the control arm vs. the treatment arm

	Wald	GCI	MOVER
Estimate	0.579	0.579	0.579
CI	$(0.490, 0.664)$	$(0.469, 0.684)$	$(0.491, 0.665)$
Width	0.174	0.215	0.174

Chapter 4 suggest the use of the MOVER or the Wald method when constructing confidence intervals for the exceedance probability.

Table 5.5 gives similar results using the Wald-type interval and the MOVER, with the intervals containing the un-informative exceedance probability of 50%. The Wald-type interval suggests that the true exceedance probability could lie anywhere from 49% to 66.4%, whereas the MOVER suggests that the truth could be anywhere from 49.1% to 66.5% with 95% confidence, a negligible difference. It is not surprising that the width of the generalized confidence interval procedure is larger than that of the other two procedures, as this was also observed in the simulation results in Chapter 4.

Chapter 6

SUMMARY

6.1 Introduction

The primary objective of this thesis was to develop and evaluate confidence interval procedures using the MOVER for three common effect measures occurring in cluster randomization trials, the difference between two normal means, the difference between two lognormal means, and the exceedance probability. As a starting point and for the sake of simplicity, attention was given to the completely randomized design. Extensions to more complex designs are discussed at the end of this chapter. The main purpose of this chapter is thus to summarize results, to make recommendations, and to propose areas of future works by identifying existing limitations.

6.2 Overall findings and recommendations

The finite sample properties of four confidence interval procedures were compared for a difference between two normal means. More specifically, we compared the Wald method, the cluster adjusted method, the generalized confidence interval method, and the MOVER, based on the empirical coverage results, the tail error results, and the median interval widths. The Wald method, the generalized confidence interval procedure, and the MOVER are all derived using the expression for the unweighted mean squared error, introduced by Thomas and Hultquist (1978). The results showed that the MOVER and the generalized confidence interval method had empirical coverage closer to the nominal 95% coverage than the alternatives. Between these two procedures, the MOVER is recommended due to its simplicity and closed-form re-

sults. Although when there are more than 12 clusters per arm, the Wald method has comparable coverage results and greater precision than the MOVER and generalized confidence interval procedures. Thus, this method may be applied as the effective sample size increases.

For the difference between two lognormal means arising from a cluster randomization trial, we compared three confidence interval procedures based on their finite sample properties. The Wald method (with the variance estimated using the delta method), the generalized confidence interval method, and the MOVER were all derived using the results of Thomas and Hultquist (1978). The MOVER is recommended primarily due to it's valid coverage performance as compared to the alternatives, and secondarily to its balanced tail error performance, precision, and simplicity of application.

We compared the finite sample properties of three confidence interval procedures for the exceedence probability. These procedures were the Wald method (with the use of the delta method to estimate the variance), the generalized confidence interval method, and the MOVER. Again, all three confidence intervals were derived using the unweighted mean squared error, introduced by Thomas and Hultquist (1978). We recommend the MOVER for all of the parameter combinations investigated due to it's overall empirical coverage performance. However, we also recommend caution when there are less than 12 clusters per arm, as simulation results showed anti-conservative behavior when $\mathrm{P}(Y_1 > Y_2) = 0.5$. Alternatively, the more complex and somewhat conservative generalized confidence interval procedure may be a useful option.

To apply the MOVER for the parameters of interest, we re-write each of them into components for which valid confidence limits already exist, such as a single normal mean and the variance components (Thomas and Hultquist, 1978, page 613). The crucial step in each MOVER procedure is estimation of the variance near the lower limit and near the upper limit separately using the existing and valid confidence intervals of the components. Alternatively, the Wald-type confidence interval estimates

the variance of the parameter estimate at the point estimate, thereby fixing the variance and potentially failing to estimate valid limits. Furthermore, estimating the variance at the point estimate imposes symmetry restrictions on the interval when the parameter estimate may have a skewed sampling distribution. However, although the MOVER for each of the three parameters of interest has shown many desirable finite sample properties, a number of limitations exist.

6.3 Limitations

The previous section made confidence interval recommendations for each of the three parameters investigated. These recommendations included the MOVER for each parameter presented in Chapter 3. However, the limitations of these MOVER procedures are important to note before their application.

The thesis focused on the analysis of a two-armed cluster randomization trial, however many trials contain more than two arms. For instance, in a trial with three arms (intervention 1, intervention 2, and usual care) the results of each intervention arm may be compared with the usual care arm using either differences between the means or exceedance probabilities. This would result in two simultaneous comparisons, (intervention 1 v.s. usual care) and (intervention 2 v.s. usual care). Here, the issue with multiple comparisons needs to be taken into account. Usual confidence interval procedures for the comparison parameter of interest, such as the ones outlined in this thesis, would lead to anti-conservative confidence intervals for the overall experiment (Westfall *et al.*, 1999, page 18). Adjustments than the MOVER may be made to allow for multiple comparisons, as outlined in the next section.

Another limitation of the proposed procedures obtained using the MOVER is that they are specifically for data following an approximately normal distribution on the raw or log scale. Although each MOVER-based procedure in Chapter 3 is derived using the MOVER (Zou, 2008), which itself makes no parametric assumptions,

but only requires valid confidence limits for each component, the derived procedures assume that the data approximately follow a one-way random effects model on the raw or log scale. That is, the mean and variance components of the one-way random effects model are used to construct confidence intervals for the difference between two normal means, the difference between two lognormal means, and the exceedance probability. Alternatively, non-parametric randomization methods may be used (e.g. The Commit Research Group, 1995; Gail *et al.*, 1996), however note that these methods lack the computational simplicity of the MOVER.

Although tests of normality exist for independent data, they generally do not apply to clustered data. Quantile-quantile plots for each cluster may be used to check the parametric assumption of normality, as illustrated in Chapter 5. Each plot will be based on the individual cluster sizes, not on the overall sample size. When the plots show an apparent deviation from normality, the proposed confidence interval procedures are not recommended. Similarly, the assumption of lognormally distributed outcomes (when estimating the difference between two lognormal means) may be checked by taking the logarithm of the data and observing the quantile-quantile plot separately for each cluster.

A third limitation of the MOVER is that the variance of the sampling distribution of the parameter estimate is estimated near the limits, not at the limits. Alternatively, the Score confidence interval method is a likelihood-based method which estimates the variance at the limits, leading to narrower confidence interval widths. However, the required likelihood function is complex for clustered data. Even so, the validity of the score is still based on the central limit theorem. Furthermore, it would be infeasible to parameterize a composite parameter, such as the lognormal mean or the standardized mean difference, into a function of a single parameter. Instead, the key step in each of the MOVER intervals uses existing and valid confidence intervals for the components of the parameter to estimate the variance of that parameter *near* its lower limit and *near* its upper limit, thereby simplifying the procedure. In

other words, the MOVER sacrifices some precision for the sake of simplicity. The simulation study in Chapter 4 has shown the MOVER to have valid coverage results, balanced tail errors, and narrow widths for the majority of the small-sample parameter combinations investigated, therefore this limitation should not be a major concern.

As a final limitation, we recognize that the MOVER does not allow for covariate adjustments. Although on average a completely randomized cluster randomization trial design adjusts for known and unknown confounders related to the outcome, any one cluster randomization trial may have chance imbalance of baseline covariates. This is particularly true when there are fewer clusters randomized, and thus a smaller effective sample size, precisely the condition investigated in this thesis. Possible remedies include changing the design of the study (consider a stratified cluster randomization trial or a pair matched cluster randomization trial) or adjusting the analysis of the study (e.g. by considering adjusted means obtained from analysis of covariance). Adjustments to the MOVER are considered at the end of the following section.

6.4 Future research

The simulation study in Chapter 4 indicated the empirical coverage, tail errors, and median widths of each of the procedures investigated when data arose from a completely randomized cluster randomization trial with either normally distributed or lognromally distributed outcomes. With the MOVER commonly recommended for each parameters of interest, it is also important to recognize the parametric assumptions of these confidence interval procedures, that is the assumption of normal or lognormal data. It would therefore be interesting for future works to investigate the performance of the three MOVER intervals when parametric assumptions are violated, particularly when the number of clusters is small.

As previously mentioned, the MOVER intervals were developed specifically for a

completely randomized cluster randomization trial design. However, other common designs include the stratified and the matched-pair cluster randomization trial designs. The MOVER may easily be extended to accommodate a stratified design using two steps. First, the point estimate may be set to the average or weighted average (weighted by the amount of information in each stratum) of the point estimates in all the strata. Second, a confidence interval may be constructed for this average using the MOVER for a linear combination of parameters (Zou, 2008) and the transformation principle. That is, the MOVER will be applied to obtain the confidence limit of the sum of the stratum-specific means, followed by the application of the transformation principle to find the limits of the average (or weighted average) mean of the strata. Note that if a weighted average is used, the overall variance estimate of the weighted average will also have to be weighted (Donner and Klar, 2000, page 126).

Obtaining confidence intervals for the parameters of interest in a pair matched study is more complex, because the estimation of the between-cluster variance is confounded by the possible effect of the intervention. One option is to extend the MOVER to account for the similarities of the matched pairs. We can use the Pearson correlation as computed over the paired clusters to estimate the covariance between the two arms (Freedman *et al.*, 1997). This covariance may then be applied by inserting a covariance term within the MOVER, as shown in the appendix of Zou (2008). However, it is important to realize that the potential gain in efficiency due to pairing may be overshadowed by the smaller effective sample size. Half as many observations are available when dealing with paired data as compared to unpaired data. Martin *et al.* (1993) showed that with a maximum of 10 pairs, the Pearson correlation between pairs must be at least 0.2 for an efficient use of matching. This thesis investigated confidence interval procedures for cluster randomization trials with as few as 6 clusters per arm. Therefore, a pair matched cluster randomization trial with only 6 pairs would require a relatively large Pearson correlation coefficient over matched pairs to be considered efficient.

Another option is to break the matches and treat the trial as if it were a completely randomized trial. However, this would lead to conservative confidence limits with coverage above the nominal alpha level if matching is strong (Donner *et al.*, 2007). The validity of the MOVER for stratified or pair matched cluster randomization trial designs must first be examined through a simulation study before their application can be implemented in practice.

An additional open reserach question concerns simultaneous confidence intervals. As mentioned in the previous section, the MOVER intervals are intended for cluster randomization trials comparing only two arms. Extensions to more than two arms, where comparisons are made based on the same control group, may be made by accounting for multiple comparisons to control empirical coverage values. One simple option is to translate the Bonferroni correction for multiple hypothesis testing (Perneger, 1998) to confidence interval construction by adjusting the critical value used from the upper $\alpha/2^{\text{th}}$ quantile to the upper $\alpha/2m^{\text{th}}$ quantile, where m is the number of comparisons conducted. However, this over-controls for the false positive error rate, resulting in rather conservative limits, and also comes at the cost of an elevated false negative error rate, or lower precision. Alternative options include using the simultaneous confidence interval procedure by Dunnett (1980) which does not assume equal variances, the Tukey-Kramer procedure (Tukey, 1953; Kramer, 1956) which does assume equal variances, or using a multivariate normal distribution rather than a univariate normal distribution when constructing simultaneous confidence intervals (Hasler and Hothorn, 2008; Donner and Zou, 2010).

Another area for future research is the performance of the MOVER for the difference between two normal or lognormal means and the exceedance probability when the groups being compared are correlated. This would not occur in a completely randomized cluster randomization trial due to the random assignment of independent clusters to trial arms, but is common in pair matched studies and in cross-over trials. Possible extensions for pair matched cluster randomization trials have been discussed

above. In cross-over trials, the same subjects receive an intervention and also act as their own controls. This is done by evaluating subjects at different times (Rothman *et al.*, 2008). Therefore, the variance for the estimated difference must be adjusted to factor in a correlation term. This may be done using the MOVER, as outlined in the appendix of Zou (2008). Again, the validity of these altered asymptotic confidence interval procedures would require verification through a simulation study.

As a final discussion of future research, we consider an extension to adjust for covariates which are related to the outcome. Although baseline covariate imbalance may be tackled in the design of the study by adopting a stratified or pair matched cluster randomization trial, an alternative option is to adjust the analysis of the clustered data. Such adjustments may be used to avoid bias due to covariate imbalances or to improve the precision of the estimates (Hauck *et al.*, 1998; Donner and Klar, 2000, page 121). The mixed-effects regression models may be used to estimate the mean difference in treatment effects on the raw or log scale, as well as the variance components of this outcome after the data have been adjusted for baseline covariates. This may be done in SAS using the `MIXED` procedure with the option `TYPE=VC` in the random statement. However, aside from the more complex analysis, Hayes and Moulton (2009, Chapter 11) caution the use of regression models for the adjustment of covariates when there are fewer than 15 clusters per arm, as results may be unreliable.

BIBLIOGRAPHY

Acion, L., Peterson, J. J., Temple, S. and Arndt, S. (2006). Probabilistic index: an intuitive non-parametric apprach to measuring the size of treatment effects. *Statistics in Medicine* **25**, 591–602.

Ahrens, H. and Pincus, R. (1981). On two measures of unbalancedness in a one-way model and their relation to efficiency. *Biometrical Journal* **23**, 227–235.

Altman, D. G. (2005). Why we need confidence intervals. *World Journal of Surgery* **29**, 554–556.

Ames, M. H. and Webster, J. T. (1991). On estimating approximate degrees of freedom. *The American Statistician* **45**, 45–50.

Anderson, P. G. (2009). A simple correlation adjustment procedure applied to confidence interval construction. *The American Statistician* **63**, 258–262.

Bartlett, J. G. and Mundy, L. M. (1995). Community-acquired pneumonia. *The New England Journal of Medicine* **333**, 1618–1624.

Bernstein, G. A., Layne, E., Egan, E. A. and Tennison, D. M. (2005). School-based interventions for anxious children. *Journal of American Academy of Child and Adolescent Psychiatry* **44**, 1118–1127.

Blaker, H. and Spjotvoll, E. (2000). Paradoxes and improvements in interval estimation. *The American Statistician* **54**, 242–247.

Bland, J. M. (2004). Cluster randomised trials in the medical literature: two bibliometric surverys. *BMC Medical Research Methodology* **4:21**.

Briggs, A., Nixon, N., Dixon, S. and Thompson, S. (2005). Parametric modelling of cost data: some simulation evidence. *Health Economics* **14**, 421–428.

Briggs, A. H., Mooney, C. Z. and Wonderling, D. E. (1999). Constructing confidence intervals for cost-effectiveness ratios: an evaluation of parametric and non-parametric techniques using Monte Carlo simulation. *Statistics in Medicine* **18**, 3245–3262.

Brunoni, A. R., Lopes, M., Kaptchukm, T. J. and Fregni, F. (2009). Placebo response of non-pharmacological and pharmacological trials in major depression: a systematic review and meta-analysis. *PLoS ONE* **4**, e4824.

Burdick, R. K. and Graybill, F. A. (1984). Confidence intervals on linear combinations of variance components in the unblanced one-way classification. *Technometrics* **26**, 131–136.

Burdick, R. K. and Graybill, F. A. (1992). *Confidence Intervals on Variance Components*. Marcel Dekker, Inc.

Burdick, R. K., Quiroz, J. and Iyer, H. K. (2006). The present status of confidence interval estimation for one-factor random models. *Journal of Statistical Planning and Inference* **136**, 4307–4325.

Burns, T. and Kendrick, T. (1997). Care of long-term mentally ill patients by british general practitioners. *Psychiatric Services* **48**, 1586–1588.

Burton, A., Altman, D. G., Royston, P. and Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine* **25**, 4279–4292.

Buse, A. (1982). The Likelihood ratio, Wald, and Lagrange multiplier tests: an expository note. *The American Statistician* **36**, 153–157.

Campbell, M. J. (2004). Extending CONSORT to include cluster trials. *British Medical Journal* **328**, 654–655.

Campbell, M. J., Elbourne, D. R. and Altman, D. G. (2004). The CONSORT statement: extension to cluster randomised trials. *British Medical Journal* **328**, 702–708.

Casella, G. and Berger, R. L. (2002). *Statistical Inference (2nd ed.)*. Belmont, CA: Duxbury Press.

Christian, P., Khatry, S. K., Katz, J., Pradhan, E. K., LeClerq, S. C., Shrestha, S. R., Adhikari, R. K., Sommer, A. and West, K. P. J. (2003). Effects of alternative maternal micronutrient supplements on low birth weight in rural nepal: double blind randomised community trial. *British Medical Journal* **326**, 1–6.

Church, J. and Harris, B. (1970). The estimation of reliability from stress-strength relationships. *Technometrics* **12**, 49–54.

Clarke, M., Oxman, A. D. and eds. (1999). *Cochrane Library: Cochrane reviewers' handbook 4.0*. Oxford: Update Software.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates Inc.

Collins, R. and Peto, R. (1994). *Antihypertensive drug therapy: effects on stroke and coronary heart disease*. In: Swales JD, ed. Textbook of hypertension. Oxford: Blackwell Scientific.

Cornfield, J. (1978). Randomization by group: a formal analysis. *American Journal of Epidemiology* **108**, 100–102.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman and Hall.

Cumming, G. (2009). Inference by eye: reading the overlap of independent confidence intervals. *Statistics in Medicine* **28**, 205–220.

Daltroy, L. H., Phillips, C., Lew, R., Wright, E., Shadick, N. A. and Liang, M. H. (2007). A controlled trial of a novel primary prevention program for lyme disease and other tick-borne illnesses. *Health Education and Behavior* **34**, 531–542.

Daly, L. E. (1998). Confidence limits made easy: interval estimation using a substitution method. *American Journal of Epidemiolgy* **147**, 783–790.

DiCiccio, T. J. and Efron, B. (1996). Rejoinder of "bootstrap confidence intervals". *Statistical Science* **11**, 223–228.

Dinh, P. and Zhou, X. H. (2006). Nonparametric statistical methods for cost-effectiveness analysis. *Biometrics* **62**, 576–588.

Djordjevic, M. V., Stellman, S. D. and Zang, E. (2000). Doses of nicotine and lung carcinogens delivered to cigarette smokers. *Journal of the National Cancer Institute* **92**, 106–111.

Donner, A. (1998). Some aspects of the design and analysis of cluster randomization trials. *Applied Statistics* **47**, 95–113.

Donner, A., Birkett, N. and Buck, C. (1981). Randomization by cluster: sample size requirements and analysis. *American Journal of Epidemiology* **114**, 906–914.

Donner, A., Eliasziw, M. and Klar, N. (1994). A comparison of methods for testing homogeneity of proportions in teratologic studies. *Statistics in Medicine* **13**, 1253–1264.

Donner, A. and Klar, N. (1993). Confidence interval construction for effect measures arising from cluster randomization trials. *Journal of Clinical Epidemiology* **46**, 123–131.

Donner, A. and Klar, N. (1996). Statistical considerations in the design and analysis of community intervention trials. *Journal of Clinical Epidemiology* **49**, 435–439.

Donner, A. and Klar, N. (2000). *Design and Analysis of Cluster Randomization Trials in Health Research*. London: Arnold Publishing Company.

Donner, A. and Klar, N. (2001). Current and future challenges in the design and analysis of cluster randomization trials. *Statistics in Medicine* **20**, 3729–3740.

Donner, A. and Klar, N. (2004). Pitfalls of and controversies in cluster randomization trials. *American Journal of Public Health* **94**, 416–422.

Donner, A. and Koval, J. (1980). The estimation of intraclass correlation in the analysis of family data. *Biometrics* **36**, 19–25.

Donner, A., Taljaard, M. and Klar, N. (2007). The merits of breaking the matches: a cautionary tale. *Statistics in Medicine* **26**, 2036–2051.

Donner, A. and Zou, G. (2010). Estimating simultaneous confidence intervals for multiple contrasts of proportions by the method of variance estimates recovery. *Statistics in Biopharmaceutical Research* **?**, DOI:10.1198/sbr.2010.09050.

Dunnett, C. W. (1980). Pairwise multiple comparisons in the unequal variance case. *Journal of the American Statistical Association* **75**, 796–800.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics* **7**, 1–26.

Efron, B. (1987). Better bootstrap confidence intervals (with discussion.). *Journal of the American Statistical Association* **82**, 171–200.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman and Hall.

El-Bassiouni, M. Y. and Abdelhafez, M. E. M. (2000). Interval estimation of the mean in a two-stage nested model. *Journal of Statistical Computation and Simulation* **67**, 333–350.

Eldridge, S. M. nd Ashby, D., , Feder, G. S., Rudnicka, A. R. and Ukoumunne, O. C. (2004). Lessons for cluster randomized trials in the twenty-first century: a systematic review of trials in primary care. *Clinical Trials* **1**, 80–90.

Eldridge, S. M., Ashby, D. and Kerry, S. (2006). Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis methods. *International Journal of Epidemiology* **35**, 1292–1300.

Feng, Z., McLerran, D. and Grizzle, J. (1996). A comparison of statistical methods for clustered data analysis with Gaussian error. *Statistics in Medicine* **15**, 1793–1806.

Fieller, E. C. (1944). A fundamental formula in the statistics of biological assay, and some applications. *Quarterly Journal of Pharmacy and Pharmacology* **17**, 117–123.

Fieller, E. C. (1954). Some problems in interval estimation. *Journal of the Royal Statistical Society, Series B* **16**, 175–185.

Fine, M. J., Singer, D. E., Phelps, A. L., Hanusa, B. H. and Kapoor, W. N. (1993). Differences in length of hospital stay in patients with community-acquired pneumonia: a prospective four-hospital study. *Medical Care* **31**, 371–380.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Edinburg: Oliver and Boyd.

Flannery, D. J., Liau, A. K., Powell, K. E., Vesterdal, W., Vazsonyi, A. T., Guo, S., Atha, H. and Embry, D. (2003). Initial behavior outcomes for the peachbuilders universal school-based violence prevention program. *Developmental Psychology* **39**, 292–308.

Flynn, T. N. and Peters, T. J. (2004). Use of the bootstrap in analsing cost data from cluster randomised trials: some simulation results. *BMC Health Services Research* **4:33**.

Food and Administration, D. (1999). Average, population, and individual approaches to establishing bioequivalence. *U.S. Food and Drug Administration, Rockville, Maryland* **0**.

Freedman, L. S., Gail, M. H., Green, S. B. and Corle, D. K. (1997). The efficiency of the matched-pairs design of the community intervention trial for smoking cessation (COMMIT). *Controlled Clinical Trials* **18**, 131–139.

Gail, M. H., Mark, S. D., Carroll, R. J., Green, S. B. and Pee, D. (1996). On design considerations and randomization-based inference for community intervention trials. *Statistics in Medicine* **15**, 1069–1092.

Gilbert, K., Gleason, P. P., Singer, D. E., Marrie, T. J., Coley, C. M., Obrosky, D. S., Lave, J. R., Kapoor, W. N. and Fine, M. J. (1998). Variations in antimicrobial use and cost in more than 2000 patients with community-acquired pneumonia. *American Journal of Medicine* **104**, 17–27.

Graybill, F. and Wang, C. M. (1980). Confidence intervals on nonnegative linear combination of variances. *Journal of the American Statistical Association* **75**, 869–873.

Graybill, F. A. (1976). *Theory and Application of the Linear Model*. Belmont, CA: Duxbury Press.

Grissom, R. (1994). Probability of the superior outcome of one treatment over another. *Journal of Applied Psychology* **79**, 314–316.

Guest, J. F. and Morirs, A. (1997). Community-acquired pneumonia. *European Respiratory Journal* **10**, 1530–1534.

Hannan, P. J., Murray, D. M., Jacobs, D. R. J. and McGovern, P. G. (1994). Parameters to aid in the design and analysis of community trials: intraclass correlations from the minnesota heart health program. *Epidemiology* **5**, 88–95.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association* **72**, 320–338.

Harville, D. A. and Jeske, D. R. (1992). Mean square error of estimation or prediction under a general linear model. *Journal of the American Statistical Association* **87**, 724–731.

Hasler, M. and Hothorn, L. A. (2008). Multiple contrast tests in the presence of heteroscedasticity. *Biometrical Journal* **50**, 793–800.

Hauck, W., Hyslop, T. and Anderson, S. (2000). Generalized treatment effects for clinical trials. *Statistics in Medicine* **19**, 887–899.

Hauck, W. W., Anderson, S. and Marcus, S. M. (1998). Should we adjust for covariates in nonlinear regression analysis of randomized trials? *Controlled Clinical Trials* **19**, 249–256.

Hayes, R. J. and Moulton, L. H. (2009). *Cluster Randomised Trials*. Chapman & Hall.

Hedges, L. V. (2007*a*). Correcting a significance test for clustering. *Journal of Educational and Behavioral Statistics* **32**, 151–179.

Hedges, L. V. (2007*b*). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics* **32**, 341–370.

Hedges, L. V. and Olkin, I. (1985). *Statistical Methods for Meta-analysis*. San Diago: Academic Press.

Howe, W. G. (1974). Approximate confidence limits on the mean of X+Y where X and Y are two tabled independent random variables. *Journal of the American Statistical Association* **69**, 789–794.

Huber, P. J. (1981). *Robust Statistics*. New York, Wiley.

Hyslop, T., Hsuan, F. and Holder, D. J. (2000). A small sample confidence interval approach to assess individual bioequivalence. *Statistics in Medicine* **19**, 2885–2897.

Jackson, R., Barham, P., Bills, J., McLennan, L., MacMahon, S. and maling, T. (1993). Guidelines for the management of mildly raised blood pressure in new zealand: a discussion document. *British Medical Journal* **307**, 107–110.

Jordhoy, M. S., Fayers, P., Loge, J. H., Ahlner-Elmqvist, M. and Kaasa, S. (2001). Quality of life in palliative cancer care: results from a cluster randomized trial. *Journal of Clinical Oncology* **19**, 3884–3894.

Julious, S. and Zariffa, N. (2002). The ABC of pharmaceutical trial design: some basic principles. *Pharmaceutical statistics* **1**, 45–53.

Kendall, M. G. and Stuart, A. (1977). *The Advanced Theory of Statistics*. New York: Macmillan.

Kinra, S., Sarma, K. V. R., Ghafoorunissa, Mendu, V. V. R., Ravikumar, R., Mohan, V., Wilinson, I. B., Cockcroft, J. R., Smith, G. D. and Ben-Shlomo, Y. (2008). Effect of integration of supplemental nutrition with public health programmes in pregnancy and early childhood on cardiovascular risk in rural indian adolescents: long term follow-up of hyderabad nutrition trial. *British Medical Journal* **337**, 1–10.

Klar, N. (1993). *Tests of the Effect of Treatment in Stratified Cluster Randomization Trials*. Ph.D. Thesis. University of Western Ontario. London, Ontario, Canada.

Klar, N. and Donner, A. (1997). The merits of matching in community intervention trials: a cautionary tale. *Statistics in Medicine* **16**, 1753–1764.

Kraemer, H., Morgan, G., Leech, N., Gliner, J., Vaske, J. and Harmon, R. (2003). Measures of clinical significance. *Journal of the American Academy of Child and Adolescent Psychiatry* **42**, 1524–1529.

Kramer, C. Y. (1956). Extension of multiple range tests to group means with unequal numbers of replications. *Biometrics* **12**, 307–310.

Krishnamoorthy, K., Mathew, T. and Ramachandran, G. (2007). Upper limits for exceedance probabilities under the one-way random effects model. *The Annals of Occupational Hygiene* **51**, 397–406.

Lee, Y., Shao, J. and Chow, S. C. (2004). Modified large-sample confidence intervals for linear combinations of variance components: extension, theory, and application. *Journal of the American Statistical Association* **99**, 467–478.

Lenhart, A., Orelus, N., Maskill, R., Alexander, N., Streit, T. and McCall, P. J. (2008). Insecticide-treated bednets to control dengue vectors: preliminary evidence from a controlled trial in haiti. *Tropical Medicine and International Health* **13**, 56–67.

Liang, K. Y. and Zeger, S. L. (1996). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

Limpert, E., Stahel, W. A. and Markus, A. (2001). Log-normal distributions across the sciences: keys and clues. *BioScience* **51**, 342–352.

Lukacs, E. (1942). A characterization of the normal distribution. *The Annals of Mathematical Statistics* **13**, 91–93.

Maghsoodloo, S. and Huang, C. Y. (2010). Comparing the overlapping of two independent confidence intervals with a single confidence interval for two normal population paramters. *Journal of Statistical Planning and Inference* **140**, 3295–3305.

Mancuso, C. A., E., P. M. G. and Charlson, M. E. (2001). Comparing discriminative validity between a diesease-specific and a general health scale in patients with miderate asthma. *Journal of Clinical Epidemiology* **54**, 263–274.

Marrie, T. J., Lau, C. Y., Wheeler, S. L., Wong, C. J., Vandervoort, M. K. and Feagan, B. G. (2000). A controlled trial of a critical pathway for treatment of community-acquired pneumonia. *Journal of the American Medical Association* **283**, 749–755.

Martin, D. C., Diehr, P., Perrin, E. B. and Koepsell, T. D. (1993). The effect of matching on the power of randomized community intervention studies. *Statistics in Medicine* **12**, 329–338.

McGraw, K. O. and Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin* **111**, 361–365.

Melese, M., Alemayehu, W., Lakew, T., Yi, E., House, J., Chidambaram, J. D., Zhou, Z., Cevallos, V., Ray, K., Hong, K. C., Porco, T. C., Phan, I., Zaidi, A., Gaynor, B. D., Whitcher, J. P. and Lietman, T. M. (2008). Comparison of annual and biannual mass antibiotic administration for elimination of infectious trachoma. *Journal of the American Medical Association* **299**, 778–784.

Montgomery, A. A., Fahey, T., Peters, T. J., MacIntosh, C. and Sharp, D. (2000). Evaluation of computer based clinical decision support system and risk chart for management of hypertention in primary care: randomised controlled trial. *British Medical Journal* **320**, 686–690.

Naylor, P. J., Macdonald, H. M., Zebedee, J. A., Reed, K. E. and McKay, H. A. (2006). Lessons learned from action schools! BC - An 'active school' model to promote physical activity in elementary schools. *Journal of Science and Medicine in Sport* **9**, 413–423.

Newcombe, R. G. (1998). Interval estimation for the difference between independent proportions: comparison of eleven methods. *Statistics in Medicine* **17**, 873–890.

Neyman, J. (1935). On the problem of confidence intervals. *Annals of Mathematical Statistics* **6**, 111–116.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of hte Royal Society of London. Series A, Mathematical and Physical Sciences* **236**, 333–380.

Neyman, J. and Pearson, E. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society A* **231**, 289–337.

Niederman, M. S., McCombs, J., Unger, A., Kumar, A. and Popovian, R. (1998). The cost of treating community-acquired pneumonia. *Clinical Therapeutics* **20**, 820–836.

Owen, D. B., Craswell, K. J. and Hanson, D. L. (1964). Nonparametric upper confidence bounds for $\Pr\{Y < X\}$ and confidence limits for $\Pr\{Y < X\}$ when X and Y are noraml. *Journal of the American Statistical Association* **59**, 906–924.

Pandey, P., Sehgal, A. R., Riboud, M., Levine, D. and Goyal, M. (2007). Informing resource-poor populations and the delivery of entitled health and social services in rural india: a cluster randomized controlled trial. *The Journal of the American Medical Association* **298**, 1867–1875.

Panella, M., Marchisio, S., Gardini, A. and Di Stanislao, F. (2007). A cluster randomized controlled trial of clinical pathway for hospital treatment of heart failure: study design and population. *BMC Health Services Research* **7:179**.

Perneger, T. V. (1998). What's wrong with Bonferroni adjustments. *British Medical Journal* **316**, 1236–1238.

Perri, M. G., McAllister, D. A., Gange, J. J. and Nezu, A. M. (1988). Effects of four maintenance programs on the long-term management of obesity. *Journal of Consulting and CLinical Psychology* **56**, 529–534.

Poole, C. (1987). Beyond the confidence interval. *American Journal of Public health* **77**, 195–199.

Quan, H. and Shih, W. J. (1996). Assessing reproducibility by the within-subject coefficient of variation with random effects models. *Biometrics* **52**, 1195–1203.

Rao, C. R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Proceedings of the Cambridge Philosophocal Society* **44**, 50–57.

Rothman, K. J., Greenland, S. and Lash, T. L. (2008). *Modern Epidemiology (3rd ed.)*. Lippincott Williams & Wilkins.

Sankaranarayanan, R., Esmy, P. O., Rajkumar, R., Muwong, R., Swaminathan, R., Shanthakumari, S., Fayette, J. M. and Cherian, J. (2007). Effects of visual screening on cervical cancer incidence and mortality in tamil nadu, india: a cluster-randomised trial. *Lancet* **370**, 398–406.

SAS Institute Inc (2009). *SAS/STAT* 9.2 *User's Guide: Mixed MOdeling (Book Excerpt)*. Cary, NY: SAS Institute Inc.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin* **2**, 110–114.

Schenker, N. (1985). Qualms about bootstrap confidence intervals. *Journal of the American Statistical Association* **80**, 360–361.

Schenker, N. and Gentleman, J. F. (2001). On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician* **55**, 182–186.

Schulz, K. F., Altman, D. G. and Moher, D. (2010). CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *British Medical Journal* **340**, 698–702.

Shoukri, M. M., Colak, D., Kaya, N. and Donner, A. (2008). Comparison of two dependent within subject coefficients of variation to evaluate the reproducibility of measurement devices. *BMC Medical Research Methodology* **8:24**.

Skene, S. S. and Kenward, M. G. (2010). The analysis of very small samples of repeated measurements i: an adjusted sandwich estimator. *Statistics in Medicine* **29**, 2825–2837.

Steiger, J. H. (2004). Beyond the F test: effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods* **9**, 164–182.

Tersmette, A. C., Petersen, G. M., Offerhaus, G. J. A., Falatko, F. C., Brune, K. A., Goggins, M., Rozenblum, E., Wilentz, R. E., Yeo, C. J., Cameron, J. L., Kern, S. E. and Hruban, R. H. (2001). Increased risk of incident pancreatic cancer among first-degree relatives of patients with familial pancreatic cancer. *Clinical Cancer Research* **7**, 738–744.

The Commit Research Group (1995). Community intervention trial for smoking cessation (commit): I and II. *American Journal of Public Health* **85**, 183–200.

Thomas, J. D. and Hultquist, R. A. (1978). Interval estimation for the unbalanced case of the one-way random effects model. *The Annals of Statistics* **6**, 582–587.

Thompson, S. G. and Barber, J. A. (2000). How should cost data in pragmatic randomised trials be analysed? *British Medical Journal* **320**, 1197–1200.

Ting, N., Burdick, R. K., Graybill, F. A., Jeyaratnum, S. and Lu, T. C. (1990). Confidence intervals on linear combination of variance components that are unrestricted in sign. *Journal of Statistical Computation and Simulation* **35**, 135–143.

Trevino, R. P., Z, Y., Hernandez, A., Hale, D. E., Garcia, O. A. and Mobley, C. (2004). Impact of the bienestar school-bsed diabetes mellitus prevention program on fasting capillary glucose levels. *Archives of Pediatrics and Adolescent Medicine* **158**, 911–918.

Tukey, J. W. (1953). *The problem of multiple comparisons.* Unpublished report, Prineton University.

Varnell, S. P., Murray, D. M., Janega, J. B. and Blitstein, J. L. (2004). Design and analysis of group-randomized trials: a review of recent practices. *American Journal of Public Health* **94**, 393–399.

Villar, J., Bakketeig, L., Donner, A., Al-Mazrou, Y., Ba'aqeel, H., Belizan, J. M., Carroli, G., Farnot, U., Lumbiganon, P., Piagio, G. and Berendes, H. (1998). The who antenatal care randomised controlled trial: rationale and study design. *Paediatric and Perinatal Epidemiology* **12**, 27–58.

Vizcaino, V. M., Aguilar, F. S., Gutierrez, R. F., Martinez, M. S., Lopez, M. S., Martinez, S. S., Garcia, E. L. and Artalejo, F. R. (2008). Assessment of an after-school physical activity program to prevent obesity among 9- to 10-year-old children: a cluster randomized trial. *International Journal of Obesity* **32**, 12–22.

Wald, A. (1941). Asymptotically most powerful tests of statistical hypotheses. *Institute of Mathematical Statisitcs* **12**, 1–19.

Walter, S. D. (1995). Methods of reporting statistical results from medical research studies. *American Journal of Epidemiology* **141**, 896–906.

Wang, H. and Chow, S. C. (2002). A practical approach for comparing means of two groups without equal variance assumption. *Statistics in Medicine* **21**, 3137–3151.

Weerahandi, S. (1993). Generalized confidence intervals. *Journal of the American Statistical Association* **88**, 899–905.

Westfall, P., Tobias, R. D. and Rom, D. (1999). *Multiple Comparisons and Multple Tests Using SAS*. Cary, NC: SAS.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* **48**, 817–830.

White, I. R. and Thomas, J. (2005). Standardized mean differences in individually-randomized and cluster-randomized trials, with applications to meta-analysis. *Clincal Trials* **2**, 141–151.

Wilcox, R. R. (2003). *Applying Contemporary Statistical Techniques*. Elsevier Science (USA).

Williamson, D. A., Copeland, A. L., Anton, S. D., Champagne, C., Han, H., Lewis, L., Martin, C., Newton, R. L., Sothern, M., Stewart, T. and Ryan, D. (2007). Wise mind project: a school-based environmental approach for preventing weight gain in children. *Obesity* **15**, 906–917.

Wilson, E. B. (1927). Probable inference, the law of succession, and staitstical inference. *Journal of the American Statistical Assiciation* **22**, 209–212.

Wolfe, R. and Hanley, J. (2002). If we're so different, why do we keep overlapping? when 1 plus 2 doesn't make 2. *Canadian Medical Association Journal* **166**, 65–66.

Zou, G. (2002). *Interim Analyses for Cluster Randomization Trials with Binary Outcomes*. Ph.D. Thesis. University of Western Ontario. London, Ontario, Canada.

Zou, G. Y. (2008). On the estimation of additive interaction using the four-by-two table and beyond. *American Journal of Epidemiology* **168**, 212–224.

Zou, G. Y. and Donner, A. (2008). Construction of confidence limits about effect measures: a general approach. *Statistics in Medicine* **27**, 1693–1702.

Zou, G. Y. and Donner, A. (2010). *A generalization of Fieller's Theorem for ratios of non-normal variables and some practical applications. In [Schuster, H. and Metzger, W.] Biometrics: Methods, Applications and Analysis*. NOVA Publishers. Pages 197-216.

Curriculum Vitae

NAME	Julia G. Taleban
PLACE OF BIRTH	Shiraz, Iran
YEAR OF BIRTH	1982
EDUCATION	Department of Biology Department of Mathematics and Statistics Queen's University, Kingston, Ontario 2001-2005, B.Sc.(H.) Biology and Mathematics Department of Mathematics and Statistics Queen's University, Kingston, Ontario 2005-2006, M.Sc. Statistics Department of Epidemiology and Biostatistics University of Western Ontario, London, Ontario 2006-2011 Ph.D. Biostatistics
HONORS AND AWARDS	Queen's University Entrance Scholarship, 2001-2002. Toronto General Hospital University Scholarship, 2001-2002. Mildred K. Walters Award, 2002-2003. Ontario Graduate Scholarship in Science and Technology, 2008-2009. Ontario Graduate Scholarship, 2009-2011.
EXPERIENCE	Teaching Assistant Department of Mathematics and Statistics Queen's University Kingston, Ontario, 2005-2006 Research Assistant Department of Epidemiology and Biostatistics University of Western Ontario London, Ontario, 2006-2011

Teaching Assistant
University of Western Ontario
London, Ontario, 2009-2010

Biostatistician
Mount Sinai Hospital
Toronto, Ontario, 2011-present

Conference Presentations

Taleban J, Rotondi M, Chen S F. Student case study. Oral presentation at the Statistical Society of Canada meeting in St. John's, Newfoundland, 2007.

Taleban J, Zou G Y. Confidence intervals for cost data in cluster randomization trials: does it make a difference? Oral presentation at the Statistical Society of Canada meeting in Vancouver, British Columbia, 2009

Zou, G Y, Donner A, Taleban J. Confidence interval estimation for 12 parameters in one-way random effects models. Oral Presentation at the Joint Statistical meetings in Vancouver, British Columbia, 2011.

Taljaard M, McRae A, Weijer C, Bennett C, Dixon S, Taleban J, Skea Z, Brehaut J, Eccles M, Donner A, Saginur R, Boruch R, Grimshaw J. Reporting of research ethics review and informed consent practices in cluster randomized trials. Oral presentation at the 32nd Annual Society for Clinical Trials meeting in Vancouver, British Columbia, 2011.

Ivers N, Taljaard M, Grimshaw J, Bennett C, Dixon S, Taleban J, McRae A, Skea Z, Boruch R, Brehaut J, Eccles M, Weijer C, Donner A. Did the extension of CONSORT to cluster randomized trials result in improved quality of reporting and study methodology? Oral presentation at the 32nd Annual Society for Clinical Trials meeting in Vancouver, British Columbia, 2011.

Refereed Publications

Zou G Y, Huo C, Taleban J. (2009). Simple confidence intervals for lognormal means and their differences with environmental applications. Environmetrics 20, 172-180.

Zou G Y, Taleban J, Huo C. (2009). Confidence interval estimation for lognormal data with application to health economics. Computational Statistics and Data Analysis, 53, 3755-3764.

Taljaard M, McRae A, Weijer C, Bennett C, Dixon S, Taleban J, Skea Z, Brehaut J, Eccles M, Donner A, Saginur R, Grimshaw J. Inadequate reporting of research ethics review and informed consent in cluster randomised trials: review of a random sample of published trials. British Medical Journal (In Press).