7-18-2012 12:00 AM

Simultaneous Confidence Intervals for Risk Ratios in the Many-to-One Comparisons of Proportions

Jungwon Shin
The University of Western Ontario

Supervisor
Dr. Neil Klar
The University of Western Ontario Joint Supervisor
Dr. Guangyong Zou
The University of Western Ontario

Graduate Program in Epidemiology and Biostatistics
A thesis submitted in partial fulfillment of the requirements for the degree in Master of Science

Simultaneous Confidence Intervals for Risk Ratios in the Many-to-One Comparisons of Proportions

(Spine title: Simultaneous Confidence Intervals for Risk Ratios)

(Thesis format: Monograph)

by

Jungwon Shin, B. Sc.

Graduate Program in Epidemiology & Biostatistics

Submitted in partial fulfillment
of the requirements for the degree of
Master of Science

The School of Graduate and Postdoctoral Studies
The University of Western Ontario
London, Ontario
June 2012

THE UNIVERSITY OF WESTERN ONTARIO

SCHOOL OF GRADUATE AND POSTDOCTORAL STUDIES

CERTIFICATE OF EXAMINATION

Joint Supervisor	Examiners
Dr. Neil Klar	Dr. Yves Bureau
	Dr. Allan Donner
Joint Supervisor	
Dr. Guangyong Zou	Dr. John Koval

The thesis by

Jungwon Shin

entitled

Simultaneous Confidence Intervals for Risk Ratios
in the Many-to-One Comparisons of Proportions

is accepted in partial fulfillment of the
requirements for the degree of
Master of Science

Date: ______________________

Chair of the Thesis Examination Board

ABSTRACT

For many-to-one comparisons of independent binomial proportions using their ratios, we propose the MOVER approach generalizing Fieller's theorem to a ratio of proportions by obtaining variance estimates in the neighbourhood of confidence limits for each proportion. We review two existing methods of inverting Wald and score test statistics and compare their performance with the proposed MOVER approach with score limits and Jeffreys limits for single proportions. As an appropriate multiplicity adjustment incorporating correlations between risk ratios, a Dunnett critical value is computed assuming a common, constant correlation of 0.5 instead of plugging in sample correlation coefficients. The simulation results suggest that the MOVER approach has desirable operating characteristics comparable to those of the method of inverting score test statistics. The MOVER with Jeffreys limits yields the median joint coverage percentage closest to the nominal level but its intervals may be wider than the other intervals in some parameter settings.

KEYWORDS: Bonferroni; Dunnett's adjustment; Jeffreys; multiple comparisons; relative risk; Wilson

ACKNOWLEDGMENTS

First and foremost, I would like to express my sincere gratitude to my supervisors, Dr. Neil Klar and Dr. Guangyong Zou, whose expertise, encouragement, understanding, and patience added greatly to my graduate experience. I am also grateful for their guidance and financial support throughout the program.

I would like to thank those people who have helped me on the path towards my graduate study. It would not have been possible for me to consider returning to school and reconnect to statistics in the field of my interest without the support of Dr. Lang Wu and Dr. Robert Evans and the internship opportunity provided by Mr. Dan Kehler at the Atlantic Service Center of Parks Canada.

I am deeply and forever indebted to my parents and sister for their love, encouragement and support through my entire life. Very special thanks go out to Mr. and Mrs. Kim for welcoming me into their family and providing me with love and encouragement that truly made a difference for the last two years in London, Ontario.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

A confidence interval encompasses hypothesis tests by not only indicating statistical significance but also providing a range of plausible values for the unknown measure of effect at a pre-specified confidence level (Gardner and Altman, 1986). This additional information led to the shift of emphasis from significance testing to interval estimation and editorial policy changes in leading medical journals in the late 1990's (Gardner and Altman, 1986; Walter, 1995; Ludbrook, 1998; Newcombe, 1998*a*). The CONSORT statement for clinical trials recommends that results for each primary and secondary outcome should be reported with the magnitude of an estimated effect and its precision, which essentially necessitates interval estimation as the standard statistical procedure in randomized clinical trials (Schulz *et al.*, 2010). Similar recommendations have been also made in the guidelines for observational studies in the STROBE statement (von Elm *et al.*, 2008).

Clinical trials may employ multiple primary endpoints, multiple treatment arms, repeated analyses over time, or combinations of these features (Cook and Farewell, 1996; Koch and Gansky, 1996; Hothorn, 2007). In fact, such a practice has gained more popularity with growing complexities and multi-faceted nature of investigations that need to address several related questions in evaluating a treatment's overall efficacy or safety in a single trial (Cook and Farewell, 1996; Ludbrook, 1998). For example, different combinations of drug regimens, or different schedules or doses of the same drug can be assigned to multiple treatment arms to compare their relative effectiveness, requiring a joint interpretation of multiple treatment comparisons for the treatment recommendation (Koch and

Gansky, 1996; Freidlin *et al.*, 2008). For this purpose, investigators require a means to construct confidence intervals that maintain the overall confidence level at the nominal level for making not a single but multiple inferences from a single trial.

Consider an investigation by Schiller *et al.* (2002) comparing three experimental chemotherapy regimens with a standard treatment as a control on their therapeutic effects. Several binary endpoints included response rates and adverse event rates. In this study, these binary variables were summarized by testing whether or not any of the treatment group proportions significantly changed compared to the control group proportion. Nevertheless, quantifying the therapeutic effects of the three experimental regimens by constructing confidence intervals is more informative than indicating the absence or presence of any statistically significant treatment effects by hypothesis tests.

When the primary outcome is binary, investigators compare two groups on a binary outcome variable by making inferences on a chosen effect measure. Common effect measures for binary outcomes include the difference of proportions, the ratio of proportions, and the ratio of odds, commonly referred to as the risk difference, the risk ratio and the odds ratio, respectively (Schechtman, 2002). The odds ratio has gained popularity due to the facts that it can be directly estimated from logistic regression models and that it is the only estimable effect measure in case-control studies. However, it exaggerates an effect when interpreted as the risk ratio for common outcomes (Deeks, 1998).

Although several procedures have been proposed and evaluated for risk differences (Piegorsch, 1991; Schaarschmidt *et al.*, 2008; Donner and Zou, 2011; Klingenberg, 2012) or odds ratios (Holford *et al.*, 1989; McCann and Tebbs, 2009), there has been relatively limited attention to simultaneous confidence interval procedures for risk ratios (Klingenberg, 2010). Therefore, this thesis concerns methods for constructing simultaneous confidence intervals for risk ratios in the many-to-one comparisons of binomial proportions.

This introductory chapter consists of six sections. Section 1.1 summarizes some key results from the studies on existing confidence intervals for a proportion as a preliminary to risk ratios. Following Section 1.2 that reviews some basic concepts and methods for multiple comparisons, Section 1.3 presents the problem of this thesis. Sections 1.4 and 1.5 describe the objectives and scope of the thesis, respectively. Finally, Section 1.6 provides a brief description of each chapter of the thesis.

1.1 Inferences on independent proportions and their ratio

1.1.1 Confidence intervals for a proportion

An extensive literature exists for confidence intervals for single proportions acknowledging poor performance of the standard large sample confidence interval based on the normal approximation (Newcombe, 1998*a*,*b*; Brown *et al.*, 2001; Brown and Li, 2005; Agresti and Coull, 1998, see). Wald confidence intervals may have poor, erratic coverage properties due to the discreteness and skewness in the underlying sampling distribution for proportion estimates. Moreover, the Wald method may yield inappropriate confidence limits out of the parameter space or confidence intervals of zero width, which are the two sources of *aberrations*, as termed by Newcombe (1998*a*).

Among alternative large sample methods, a score-based confidence interval for proportions described by Wilson (1927) has been suggested as the method theoretically most appealing but computationally intensive (Miettinen and Nurminen, 1985). A score-based confidence interval is obtained by inverting a score test. The score method yields boundary-respecting confidence limits and coverage probabilities closer to the nominal confidence level than those of Wald intervals (Agresti and Coull, 1998; Newcombe, 1998*a*; Brown *et al.*, 2001, 2002).

Unlike these large sample methods based on asymptotic normality, a Clopper-Pearson interval is constructed by inverting a binomial test rather than its normal approximation

(Clopper and Pearson, 1934). Therefore, this interval estimator is guaranteed to have coverage probability of at least the nominal level for every possible value of proportion (Agresti and Coull, 1998). Unless a minimum coverage greater than the nominal level is demanded, the Clopper-Pearson method is not recommended in practice for its strict conservativeness (Agresti and Coull, 1998; Brown *et al.*, 2001). The conservativeness of Clopper-Pearson confidence intervals reduces slightly using the mid-P value, half the probability of the observed result plus the probability of more extreme results (Lancaster, 1949), but modified Clopper-Pearson confidence intervals tend to be still over conservative (Newcombe, 1998*a*).

1.1.2 Confidence intervals for a risk ratio

The ratio of two independent proportions, also commonly referred to as the risk ratio or relative risk, is an important parameter in both epidemiological and clinical studies. As a relative measure of the effect on average risk due to an exposure or treatment, it provides a natural reference point of unity (Graham *et al.*, 2003).

A large literature exists on the methods for constructing an approximate confidence interval for a risk ratio. Gart and Nam (1988) and Dann and Koch (2005) provide a comprehensive summary of these methods classified into three groups, namely the Wald method (Katz *et al.*, 1978), the Fieller method (Katz *et al.*, 1978) and the score method (Koopman, 1984; Miettinen and Nurminen, 1985). Chapter 2 reviews these methods along with some other methods for a single risk ratio with respect to the coverage properties, aberrations and computational complexity. We also review the extension of the Wald and score methods for a single risk ratio to multiple risk ratios in the many-to-one comparisons of proportions based on the idea of Dunnett (1955) in Chapter 2.

1.2 Simultaneous inferences in multiple comparisons

Performing multiple hypothesis tests from a single trial can be problematic due to multiplicity effects that may inflate the type I error rates (Proschan and Waclawiw, 2000; Ludbrook, 1998; Westfall *et al.*, 1999). Likewise, multiple interval estimates computed without an appropriate multiplicity adjustment may be misleading and fail to attain the nominal joint coverage probability (Koch and Gansky, 1996; Hochberg and Tamhane, 1987). For the sake of clarity, the remaining discussion in this section will be given in terms of hypothesis testing followed by interval estimation.

Multiple comparison procedures account for multiplicity by considering a set of inferences simultaneously as a family. The rationale is that these inferences are related in terms of their content or intended use, and therefore, it is meaningful to consider some combined measure of errors made for the individual inferences in the family (Hochberg and Tamhane, 1987, p.5). The classical multiple hypothesis tests control the probability of a type I error in a given family, which is called *familywise error rate*, to ensure simultaneous correctness in the set of these inferences at the desired level of significance (Hochberg and Tamhane, 1987, pp.8-11). An important issue for multiple comparisons is to determine the appropriate degree of adjustment that controls the type I error and provides adequate power at the same time (Koch and Gansky, 1996; Hothorn, 2007). Similarly, simultaneous confidence interval methods maintain the nominal joint confidence level.

The simplest multiplicity adjustment method is the Bonferroni method which assumes that all comparisons are independent, and sets the type I error rate per comparison to be less than or equal to a given familywise error rate divided by the number of comparisons in the family (Hochberg and Tamhane, 1987). The Bonferroni adjustment is increasingly conservative as the correlation between comparisons or the number of comparisons becomes large (Proschan and Waclawiw, 2000). Less conservative and more powerful multiple comparison methods are available, such as the Tukey-Kramer method for all pairwise comparisons and the Dunnett method for many-to-one comparisons. These methods control the familywise error rate while taking the correlations among comparisons into consideration (Hochberg and Tamhane, 1987). Therefore, simultaneous intervals constructed using an appropriate critical value based on these methods are narrower than those adjusted by the Bonferroni method.

1.3 Statement of the problem

Confidence interval methods for risk ratios mentioned in Section 1.1 can be easily modified to simultaneous interval estimation procedures by adjusting for multiplicity. The simplest Bonferroni method may, however, be too conservative and yield less precise confidence intervals in the many-to-one comparisons commonly involving the control group proportions (Koch and Gansky, 1996; Schaarschmidt *et al.*, 2009; Klingenberg, 2010). Therefore, Klingenberg (2010) proposed computing a lower (upper) confidence limit by inverting the maximum (minimum) score statistics with a Dunnett critical value obtained using the plug-in estimates of correlation coefficients and demonstrated a simultaneous coverage probability close to the nominal level. Nevertheless, a drawback of this method is that the confidence limits must be obtained numerically by an iterative algorithm because no analytical solution exists. Therefore, computationally simpler alternatives are worth exploring, which was remarked by Klingenberg (2010). Also, using the correlation coefficients estimated from a sample to obtain Dunnett's critical value may not be desirable because the variability of the correlation estimates may have a greater impact on the computed critical value for small or moderate sample sizes (Holford *et al.*, 1989).

To address the issue of simplicity, we aim to develop an alternative, non-iterative procedure based on the Method of Variance Estimates Recovery (MOVER) (Zou and Donner, 2008; Zou, 2008). This method entails two steps for constructing simultaneous confidence intervals for multiple risk ratios. First, confidence limits about individual sample proportions are obtained using a critical value from the multivariate normal distribution accounting for correlations among comparisons. A multiplicity adjustment is made using a Dunnett critical value computed assuming a common, constant correlation of 0.5 instead of estimating correlation coefficients from a sample. Second, confidence limits for risk ratios are computed using the variance estimates recovered from the confidence limits for proportions. As the MOVER approach obtains the variance estimates near the confidence limits, it shares the basic idea of the score method that obtains them at the confidence limits (Wilks, 1938).

1.4 Thesis objectives

Specific objectives of this thesis include:

1. To review the existing methods for constructing confidence intervals for a single and multiple risk ratios,

2. To present the MOVER approach in the construction of confidence intervals for one or more risk ratios with Dunnett's critical value that only depends on the number of experimental groups by assuming a common, constant correlation coefficient of 0.5,

3. To evaluate the finite sample properties of the MOVER in comparison to the existing test-inversion methods for constructing simultaneous intervals for risk ratios in many-to-one comparisons,

4. To illustrate the construction of simultaneous confidence intervals applying the MOVER in worked examples.

1.5 Scope of the thesis

We focus our attention to the confidence interval methods applicable to data from completely randomized clinical trials with an equal allocation of individually randomized subjects. However, unequal sample sizes are also considered to examine robustness of the competing methods to unequal sample sizes due to different attrition rates common in practice. We limit attention to two-sided, large sample confidence intervals for a single and multiple risk ratios.

1.6 Organization of the thesis

This thesis consists of six chapters including this introductory chapter. Chapter 2 reviews existing confidence interval procedures for a single and multiple risk ratios based on the multivariate central limit theorem. Chapter 3 provides a general introduction to the MOVER approach for ratios. Chapter 4 presents the design of a simulation study to assess the performance of four competing methods for both single and multiple risk ratios in the many-to-one comparisons. Chapter 5 illustrates the computation of simultaneous confidence intervals for several risk ratios by the MOVER with worked examples. Chapter 6 discusses findings and limitations of the thesis and suggests further research venue.

Chapter 2

LITERATURE REVIEW

This chapter consists of four sections and reviews large sample confidence interval methods for a single and multiple risk ratios. Section 2.1 describes confidence interval methods for a single risk ratio. Section 2.2 presents a methodological extension from single to multiple inferences based on the multivariate central limit theorem. We review the simultaneous confidence interval methods proposed by Klingenberg (2010), implementing the general multiple comparison procedure for the many-to-one comparisons of proportions in Section 2.3. This section also includes a brief review of the literature on the computation of an appropriate critical value in multiple comparisons of proportions. Along with a brief chapter summary in Section 2.4, we propose an alternative, non-iterative method to construct simultaneous confidence intervals based on the method of variance estimates recovery (MOVER) for multiple risk ratios in the many-to-one comparisons of proportions.

2.1 Confidence interval for a risk ratio

Several asymptotic confidence interval methods have been proposed for a single ratio of two independent binomial proportions and evaluated for their finite sample statistical performance in the literature. Gart and Nam (1988) provided the first comprehensive review and evaluation of the existing methods proposed prior to 1988, classifying them into three groups based on the method of derivation. A similar review was subsequently provided by Dann and Koch (2005), including other variants of the existing methods. This section describes the three derivations of existing confidence interval methods for a single risk ratio.

2.1.1 Notation

Let Y_0 and Y_1 denote independent binomial variates with corresponding population proportions of the event of interest π_0 and π_1 and samples sizes n_0 and n_1 for the control group and the treatment group, respectively. We are interested in constructing a $100(1-\alpha)\%$ confidence interval for the ratio of two independent proportions, $\mathrm{RR} = \pi_1/\pi_0$. Let $z_{\alpha/2}$ denote the $\alpha/2$ upper quantile of the standard normal distribution. Given the observed number of events y_0 for the control group and y_1 in the experimental group, the unrestricted maximum likelihood estimates of π_0 and π_1 are $\widehat{\pi}_0 = y_0/n_0$ and $\widehat{\pi}_1 = y_1/n_1$. For the inference under the null hypothesis $\mathrm{H}_0 : \mathrm{RR} = \mathrm{RR}^0$, restricted maximum likelihood estimates, $\widetilde{\pi}_1$ and $\widetilde{\pi}_0$ are obtained by maximizing the reparameterized log-likelihood function of the joint binomial distribution by letting $\widetilde{\pi}_1 = \mathrm{RR}^0\widetilde{\pi}_0$. $\widetilde{\pi}_0$ is the admissible solution to the quadratic equation $a\widetilde{\pi}_0^2 + b\widetilde{\pi}_0 + c = 0$, where $a = (n_0+n_1)\mathrm{RR}^0$, $b = -(y_0+n_1)\mathrm{RR}^0 + y_1 + n_0$ and $c = y_0 + y_1$, resulting in both $\widetilde{\pi}_0$ and $\widetilde{\pi}_1$ in the parameter space [0,1] (Nam, 1995).

The Wald method

Assuming the asymptotic normality of the risk ratio estimator $\widehat{\mathrm{RR}}$, Noether (1957) proposed estimating its variance from the first-order Taylor's series expansion. Subsequently, Katz *et al.* (1978) proposed the analogous Wald method on the logarithmic scale and demonstrated its reasonable finite sample performance. Assuming $\ln\widehat{\mathrm{RR}}$ is normally distributed with mean $\ln\mathrm{RR}$ and variance

$$\mathrm{var}(\ln\widehat{\mathrm{RR}}) = \frac{(1-\pi_1)}{\mathrm{n}_1\pi_1} + \frac{(1-\pi_0)}{\mathrm{n}_0\pi_0},$$

we can construct a $100(1-\alpha)\%$ confidence interval by inverting the Wald test statistic

$$T_W^2 = \frac{(\ln\widehat{\mathrm{RR}} - \ln\mathrm{RR})^2}{\widehat{\mathrm{var}}(\ln\widehat{\mathrm{RR}})} = z_{\alpha/2}^2. \tag{2.1}$$

The resulting symmetric confidence interval is given as

$$\widehat{\mathrm{RR}}\exp\left(\mp z_{\alpha/2}\sqrt{\widehat{\mathrm{var}}(\ln\widehat{\mathrm{RR}})}\right).$$

This method fails when the confidence interval is not computable for $y_1 = 0$ or $y_0 = 0$. The method also yields a degenerate interval of zero width when $y_1 = n_1$ and $y_0 = n_0$. Gart and Nam (1988) evaluated the method by using modified number of events $y_{i'} = y_i + 0.5$ and sample sizes $n_{i'} = n_i + 0.5$ for $i = 0,1$ to avoid incomputable cases. The modified method is always computable, however, it still yields a degenerate interval of zero width when $y_1 = n_1$ and $y_0 = n_0$. An alternative modification to the method was considered by Dann and Koch (2005) adapting Agresti and Coull (1998)'s approach of adding a number of pseudo observations approximating $z^2_{\alpha/2}$ and distributing them to each group according to the population risk ratio assumed under the null hypothesis. For example, 4 pseudo observations are added, 2.67 to the experimental group and 1.33 to the control group for $\mathrm{RR} = 2$, in the estimation of the population risk ratio.

The Fieller method

Applying Fieller's theorem for the ratio of two normal means (Fieller, 1944), Katz *et al.* (1978) considered a method for the ratio of two proportions based on the asymptotic normality of the statistic $T = (\widehat{\pi}_1 - \mathrm{RR}\widehat{\pi}_0)$ and its variance estimate given as

$$\widehat{\mathrm{Var}}(T) = \frac{\widehat{\pi}_1(1-\widehat{\pi}_1)}{n_1} + \mathrm{RR}^2\frac{\widehat{\pi}_0(1-\widehat{\pi}_0)}{n_0}.$$

We obtain $100(1-\alpha)\%$ confidence limits, which are the two roots to the quadratic equation in RR

$$T_F^2 = \frac{(\widehat{\pi}_1 - \mathrm{RR}\widehat{\pi}_0)^2}{\widehat{\pi}_1(1-\widehat{\pi}_1)/n_1 + \mathrm{RR}^2\widehat{\pi}_0(1-\widehat{\pi}_0)/n_0} = z^2_{\alpha/2}. \tag{2.2}$$

Fieller's theorem is developed for a ratio of independent normal means having symmetric sampling distributions in unbounded parameter space. Consequently, when applied to the ratio of proportions, this method may yield nonsensical values such as complex, negative, or exclusive limits, typically for small values of y_0 or $n_0\pi_0$ (Katz *et al.*, 1978; Gart and Nam, 1988). To minimize the skewness of the test statistics in (2.2), Bailey (1987) proposed a power transformation of the risk ratio, using the statistic $U = (\widehat{\pi}_1^t - \mathrm{RR}^t\widehat{\pi}_0^t)$ and its variance estimated by the delta method as

$$\widehat{\mathrm{Var}}(U) = t^2\Big[\frac{\widehat{\pi}_1^{2t-1}(1-\widehat{\pi}_1)}{n_1} + \mathrm{RR}^{2t}\frac{\widehat{\pi}_0^{2t-1}(1-\widehat{\pi}_0)}{n_0}\Big],$$

for a known t. Assuming small population proportions in typical cohort studies, Bailey (1987) suggested a cube-root transformation (i.e. $t = 1/3$) as the optimal power transformation to improve coverage probabilities by making the test statistic's sampling distribution more symmetric.

The Score method

Unlike the previous two methods, the score method constructs an interval with the statistic's variance estimated with restricted MLEs $\widetilde{\pi}_1$ and $\widetilde{\pi}_0$. Koopman (1984) heuristically derived the following test statistic consistent with Pearson's chi-square test for $\mathrm{RR} = 1$

$$S_K^2 = \frac{(y_0 - n_0\widetilde{\pi}_0)^2}{n_0\widetilde{\pi}_0(1-\widetilde{\pi}_0)}\Big\{\frac{n_0(1-\mathrm{RR}\widetilde{\pi}_0)}{n_1\mathrm{RR}(1-\widetilde{\pi}_0)} + 1\Big\} = z_{\alpha/2}^2, \tag{2.3}$$

while Miettinen and Nurminen (1985)'s derivation used the statistic of the Fieller's method with its variance estimated using the restricted MLEs as

$$S_M^2 = \frac{(\widehat{\pi}_1 - \mathrm{RR}\widehat{\pi}_0)^2}{\widetilde{\pi}_1(1-\widetilde{\pi}_1)/n_1 + \mathrm{RR}^2\widetilde{\pi}_0(1-\widetilde{\pi}_0)/n_0} = z_{\alpha/2}^2. \tag{2.4}$$

Using the general likelihood method of score statistics (Bartlett, 1953), Gart (1985) derived the statistic in (2.3). Subsequently, Gart and Nam (1988) formally proved equivalence of (2.3) and (2.4). Approximate confidence limits can be obtained by solving a cubic equation iteratively (Koopman, 1984; Miettinen and Nurminen, 1985; Gart, 1985; Gart and Nam, 1988) or non-iteratively by a series of substitutions (Nam, 1995). Compared to the other two methods above, the score method has more desirable properties of confidence intervals. It is computable in all possible outcomes except for the noninformative case and does not yield any aberrant limits. Having the characteristics of efficient score, it is also consistent with Pearson's chi-square test for testing RR $= 1$. In comparison to the Wald confidence interval, the score confidence interval typically shows closer to nominal coverage and generally achieves greater balance in the tail error probabilities (Gart and Nam, 1988). The tail error balance can be further improved by applying Bartlett (1955)'s skewness correction, particularly for a risk ratio far from unity, based on the general theory of the score method. Although the score method results in confidence intervals with superior properties, computing their limits requires an iterative procedure or series of substitutions (Nam, 1995).

2.1.2 Summary

The superior performance of the score method to the other alternative methods have been discussed in the literature (Gart and Nam, 1988; Dann and Koch, 2005; Price and Bonett, 2008). For the Wald method using a log risk ratio, studies have shown its general conservativeness yet providing acceptable performance in terms of attaining the actual coverage probabilities reasonably close to the nominal level for large samples (Katz *et al.*, 1978; Gart and Nam, 1988; Dann and Koch, 2005). Before discussing the extension of the Wald method and the score method for multiple risk ratios, we summarize the construction of simultaneous confidence intervals based on the asymptotic multivariate central limit theorem. The discussion focuses on the multiplicity adjustment by computing an appropriate critical value to maintain the joint coverage probability for several dependent statistics.

2.2 Simultaneous confidence intervals for multiple parameters

We can construct a $100(1-\alpha)\%$ confidence interval for a single parameter of interest θ by inverting the acceptance region of a level α test for the null hypothesis $\mathrm{H}_0 : \theta = \theta_0$. Therefore, a $100(1-\alpha)\%$ confidence interval consists of the plausible values of θ_0 for which the null hypothesis is not rejected at level α given an observed sample statistic. Therefore, such an interval procedure should capture the true parameter θ with a $100(1-\alpha)\%$ probability, preferably for every possible set of data.

When several parameters from a single experiment are of inferential interest, a practical, single-step, multiple comparison procedure constructs a test of the null hypothesis H_0 as an intersection of a family of K hypotheses (Hochberg and Tamhane, 1987, p.28-30). When each $\mathrm{H}_{0k} : \theta_k = \theta_{0k}$ where $\mathrm{H}_0 = \cap_{k=1}^{K} \mathrm{H}_{0k}$ has a suitable test statistic T_k for $k = 1, \ldots, K$, a single global test of H_0 permits multiple inferences on the set of these parameters considered jointly as a family. Therefore, to construct simultaneous confidence intervals, we must determine an appropriate critical value that maintains the joint coverage probability at the nominal level $100(1-\alpha)\%$. Such a critical value may be exactly computed by evaluating the probabilities of the joint distribution of the test statistics. When its distribution is unknown, it is often approximated assuming the multivariate central limit theorem or simply applying various probability inequalities without distributional assumptions such as Bonferroni's inequality. We describe the procedures more formally in the remaining of this section.

Following the notation similar to Westfall *et al.* (1999), we denote estimates of the unknown parameters of interest by $\widehat{\theta} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_J)^T$. Under the multivariate central limit theorem, $\widehat{\theta}$ asymptotically follows a multivariate normal distribution with mean θ and a diagonal variance-covariance matrix having the jth diagonal element equal to $\mathrm{var}(\widehat{\theta}_j)$ where $j = 1, \ldots, J$. The confidence limits for a collection of contrasts $c_k^T \theta$, having a covariance-variance $\mathrm{var}(c_k^T\widehat{\theta})$ can be obtained using the general form

$$\Big(L_k, U_k\Big) = \Big(c_k^T\widehat{\theta} \mp h\sqrt{\mathrm{var}(c_k^T\widehat{\theta})}\Big),$$

where h is an appropriate critical value.

Substituting a sample variance estimate $\widehat{\mathrm{var}}(c_k^T\widehat{\theta})$ for $\mathrm{var}(c_k^T\widehat{\theta})$, simultaneous Wald confidence intervals are constructed with a critical value z_α, chosen to satisfy

$$\Pr\Big(c_k^T\widehat{\theta} - z_\alpha\sqrt{\widehat{\mathrm{var}}(c_k^T\widehat{\theta})} < c_k^T\theta < c_k^T\widehat{\theta} + z_\alpha\sqrt{\widehat{\mathrm{var}}(c_k^T\widehat{\theta})}, \forall k\Big) \cong 1-\alpha. \tag{2.5}$$

Therefore, for a set of Wald test statistics T_k where $k = 1, 2, \ldots, K$ given as

$$T_k = \frac{c_k^T\widehat{\theta} - c_k^T\theta}{\sqrt{c_k^T\widehat{\mathrm{var}}(\widehat{\theta})c_k}},$$

the joint distribution of T_k is asymptotically multivariate normal with mean 0 and correlation matrix

$$R_{K\times K} = D^{-1/2}(C^TVC)D^{-1/2},$$

where $C = (c_1, \ldots, c_K)$, V is a diagonal matrix with kth element given by $\mathrm{var}(\widehat{\theta}_k)$ and D is a diagonal matrix with kth element equal to $\mathrm{var}(c_k^T\widehat{\theta})$. The probability statement (2.5) can be represented equivalently as

$$\Pr\Big(\max|T_k| \leq z_\alpha \Big| R_{K\times K}\Big) \cong 1-\alpha. \tag{2.6}$$

The familiar Bonferroni inequality approximates the probability (2.6), ignoring the correlation $R_{K\times K}$, by its lower bound from the marginal, univariate normal distributions as

$$\Pr\Big(\max|T_k| \leq z_\alpha \Big| R_{K\times K}\Big) \geq 1 - \sum_{k=1}^{K}\Pr\Big(|T_k| \leq z_\alpha\Big).$$

Although this simple method is applicable in general settings, its conservativeness results in simultaneous confidence intervals with joint coverage probability typically greater than the nominal level of $100(1-\alpha)\%$. This is true particularly for highly correlated test statistics and for a large number of comparisons. An improved method, known as the *Sidak procedure*, utilizes the properties of the multivariate normal distribution and yields a critical value sharper that of the Bonferroni method (Sidak, 1968). Assuming the correlation $R_{K\times K}$ is an identity matrix $I_{K\times K}$, the probability (2.6) is estimated as

$$\Pr\Big(\max|T_k| \leq z_\alpha \big| I_{K\times K}\Big) \geq 1 - \prod_{k=1}^{K} \Pr\Big(|T_k| \leq z_\alpha\Big).$$

Unlike these two methods assuming independence among T_k, a critical value may be computed analytically by a complex multidimensional integration, incorporating the correlation matrix for T_k. Genz (1992) proposed a method to obtain an approximate $(1-\alpha)$ quantile of the distribution of $\max|T_k|$ using a consistent estimator for the correlation matrix $R_{K\times K}$ in the inversion algorithm of the multivariate normal distribution. Alternatively, when a critical value cannot be computed analytically, it may be approximated by simulation proposed by (Edwards and Berry, 1987). For common comparisons such as Tukey's all pairwise comparisons and Dunnett's many-to-one comparisons, a critical value may be readily obtained from the SAS IML function `probmc` or the function `qmvnorm` in R package `mvtnorm` (Mi *et al.*, 2009).

Despite these computational tools, obtaining an appropriate critical value for binomial proportions may not be as simple as for normal means from a practical perspective. Since the variance of a sample binomial proportion $\widehat{\pi}$ is a function of the unknown parameter π itself, the correlation matrix required for the computation of the critical value based on the multivariate normal distribution is also unknown and must be estimated from a sample. Consequently, the coverage probability may be affected to a certain degree by the accuracy of the estimated correlation matrix. It may be particularly more problematic for small samples and highly correlated comparisons (Holford *et al.*, 1989). The next section provides a review on the existing approaches to the computation of an appropriate critical value in the multiple comparisons of proportions and the construction of simultaneous confidence intervals for multiple risk ratios.

2.3 Simultaneous confidence intervals for multiple risk ratios

Compared to a large literature for normally distributed data, relatively few multiple comparison procedures of interval estimation have been proposed for binomial proportions (Piegorsch, 1991; Schaarschmidt *et al.*, 2009; Agresti *et al.*, 2008; McCann and Tebbs, 2009; Klingenberg, 2010, 2012) despite their common occurrences in clinical trials and toxicological experiments (Schaarschmidt *et al.*, 2009; Holford *et al.*, 1989). Moreover, a majority of these procedures are intended for inferences based on odds ratios (Holford *et al.*, 1989; McCann and Tebbs, 2009) or risk differences and their linear combinations (Piegorsch, 1991; Agresti *et al.*, 2008; Schaarschmidt *et al.*, 2008, 2009; Donner and Zou, 2011; Klingenberg, 2012). The most useful approximate procedures for multiple risk ratios were provided by Agresti *et al.* (2008) for all pairwise comparisons and Klingenberg (2010) for many-to-one comparisons based on the multivariate central limit theorem. When the critical value is computed based on the asymptotic multivariate normality for multiple risk ratios, these procedures are less conservative than the simple Bonferroni or Sidak procedures, yielding more precise confidence intervals while maintaining the nominal joint coverage probability (Klingenberg, 2010). Klingenberg (2010) extended the Wald and Score confidence interval methods in Section 2.1 by considering a variety of multiplicity adjustments reflected in the computation of the critical value.

2.3.1 Determination of the critical value

A variety of multiplicity adjustments and their impact on coverage probabilities have been discussed in the literature for multiple comparisons of proportions. For comparisons based on log odds ratios, Holford *et al.* (1989) and McCann and Tebbs (2009) investigated how different adjustment methods including the Bonferroni, Sidak, and Dunnett critical values performed in terms of maintaining the correct joint confidence level. Similar investigations are available for the construction of simultaneous confidence intervals of multiple risk differences and contrasts of proportions (Piegorsch, 1991; Schaarschmidt *et al.*, 2009; Donner and Zou, 2011). For the multiplicity adjustment accounting for correlations among comparisons, the critical value was computed using sample correlation estimates. More recently, Klingenberg (2012) considered an alternative method of using a lower bound of correlation under the null hypothesis, requiring only one parameter estimate, $\widehat{\pi}_0$ instead of $K+1$ plug-in proportion estimates in the computation of a Dunnett critical value.

For the many-to-one comparisons of proportions using risk ratios, Klingenberg (2010) explored three different methods to approximate the correlation matrix, namely the Dunnett method with plug-in sample estimates, the Sidak procedure, and the procedure proposed by Berger and Boos (1994). As described in the Section 2.2, the Sidak procedure does not estimate the correlation matrix from the sample but uses an identity matrix assuming independence of many-to-one risk ratios in the computation of the critical value. The procedure proposed by Berger and Boos (1994) uses the lower bound of the correlation matrix for the plausible values of the nuisance parameter π_0. As the Dunnett method with a sample correlation matrix showed a negligible difference from the method by Berger and Boos (1994) on the actual coverage probabilities, Klingenberg (2010) recommended using the Dunnett method with plug-in sample estimates based on its computational simplicity over Berger and Boos (1994)'s method.

The correlation matrix $R_{K\times K}$ required in the computation of a Dunnett critical value satisfying the probability in (2.6) depends on the unknown true proportions $\pi = (\pi_0, \ldots, \pi_K)$. When the sample sizes become large, the correlation estimate would be close to the true correlation value, yielding negligible effect on the computation of the critical value and confidence limits. On the other hand, the variability of the correlation estimates could have a greater impact on the computed critical value for small or moderate samples sizes and for highly correlated comparisons as observed in the case of odds ratios (Holford *et al.*, 1989). However, the findings from Piegorsch (1991) for risk differences confirm that the type of test inversion and its variance estimator affect coverage probabilities more considerably than the type of multiplicity adjustment. Moreover, an evaluation of score confidence intervals with different correlation estimates show a limited improvement on coverage probabilities and precision by incorporating more accurate correlation information (Klingenberg, 2012).

2.3.2 Notation

Let Y_k denote independent binomial variates with corresponding population probabilities of the event of interest and sample sizes n_k in the kth group where $k = 0,\ldots,K$. Given the observed event counts $(y_0, y_1, \ldots, y_K)$, the maximum likelihood estimates for π_0, π_k's and RR_k's are $\widehat{\pi}_0 = y_0/n_0$, $\widehat{\pi}_k = y_k/n_k$ and $\widehat{\mathrm{RR}}_k = \widehat{\pi}_k/\widehat{\pi}_0$, respectively, in the comparison of each of K experimental groups with the control group identified by $k = 0$. For inference under the partial null hypothesis $\mathrm{H}_{0k} : \mathrm{RR}_k = \mathrm{RR}_k^0$, the likelihood function is maximized under the restriction of the given null value RR_k^0 to obtain the restricted MLE of π_0. $\widetilde{\pi}_{0|k}$ is the admissible solution to the quadratic equation $a\widetilde{\pi}_0^2 + b\widetilde{\pi}_0 + c = 0$, where $a = (n_0 + n_k)\mathrm{RR}_k^0$, $b = -(y_0 + n_k)\mathrm{RR}_k^0 + y_k + n_0$ and $c = y0 + y_k$, resulting in $\widetilde{\pi}_0$ and $\widetilde{\pi}_k = \widetilde{\pi}_{0|k}\mathrm{RR}_k^0$ in the parameter space [0,1] (Klingenberg, 2010). Simultaneous confidence intervals for $\mathrm{RR}_k = \pi_k/\pi_0$ can be constructed by inverting the following test statistics.

Simultaneous Wald intervals

Let $\ln\widehat{\mathrm{RR}}_k$ denote a log-transformed risk ratio estimate of the kth treatment group and the control group for $k = 1,\ldots,K$. Based on the multivariate central limit theorem, $\ln\widehat{\mathrm{RR}}_k$ follows a multivariate normal distribution with mean $\ln\mathrm{RR}$ and a variance-covariance matrix having the kth diagonal element $(k = 1,\ldots,K)$ equal to

$$\sigma_k{}^2 = \mathrm{var}(\ln\widehat{\mathrm{RR}}_k) = (1-\pi_0\mathrm{RR}_k)/(n_k\pi_0\mathrm{RR}_k) + (1-\pi_0)/(n_0\pi_0)$$

and off-diagonal elements for $1 \leq i \neq j \leq K$

$$\sigma_{ij} = \operatorname{cov}(\ln \widehat{\mathrm{RR}}_i, \ln \widehat{\mathrm{RR}}_j) = \frac{1-\pi_0}{n_0 \pi_0}.$$

Under the null hypothesis $\mathrm{H}_0 = \cap_{k=1}^{K} \{\mathrm{RR}_k = \mathrm{RR}_k^0\}$,

$$T_k = \frac{\ln \widehat{\mathrm{RR}}_k - \ln \mathrm{RR}_k^0}{\sqrt{(1-\widehat{\pi}_k)/(n_k \widehat{\pi}_k) + (1-\widehat{\pi}_0)/(n_0 \widehat{\pi}_0)}},$$

asymptotically follows a multivariate normal distribution with mean zero and correlation matrix $\Lambda(\pi_0)$ with off-diagonal elements in product correlation form where $1 \leq i \neq j \leq K$

$$\lambda_{ij}(\pi_0) = \frac{\sigma_{ij}}{\sigma_i \sigma_j} = \left(1 + \frac{n_0}{n_i} \frac{1 - \mathrm{RR}_i^0 \pi_0}{\mathrm{RR}_i^0 - \mathrm{RR}_i^0 \pi_0}\right)^{-\frac{1}{2}} \left(1 + \frac{n_0}{n_j} \frac{1 - \mathrm{RR}_j^0 \pi_0}{\mathrm{RR}_j^0 - \mathrm{RR}_j^0 \pi_0}\right)^{-\frac{1}{2}} = \lambda_i(\pi_0)\lambda_j(\pi_0). \tag{2.7}$$

Therefore, the critical value h is selected to satisfy

$$\Pr\Big(\max|T_k| \leq h \big| \Lambda(\pi_0)\Big) = 1 - \alpha. \tag{2.8}$$

The SAS IML function `probmc` (SAS Institute, Inc 2009, p. 976) may be used to compute the critical value h. The syntax using `probmc` is

```
 probmc(distribution, q, prob, df, nparam<,parameters>)
```

where `q` is the quantile from the specified distribution, `prob` is the left probability from the distribution, and `df` is the degrees of freedom. Thus, the critical value for the two-sided Dunnett comparison is computed by

```
probmc("Dunnettt2",., confidence,.,K, parameters),
```

specifying the corresponding sample estimates $\widehat{\lambda}_k$ for `parameters`

$$\widehat{\lambda}_k = \left[1 + \frac{n_0}{n_k} \frac{\widehat{\pi}_0(1-\widehat{\pi}_k)}{\widehat{\pi}_k(1-\widehat{\pi}_0)}\right]^{-\frac{1}{2}}, \text{for } k = 1, \ldots, K.$$

Simultaneous score intervals

The score test statistics can be derived based on the difference $\widehat{\pi}_k - \mathrm{RR}_k^0\widehat{\pi}_0$ with its standard error estimated under the kth partial hypothesis $\mathrm{H}_{0k} : \mathrm{RR}_k = \mathrm{RR}_k^0$. Klingenberg (2010) outlined the steps to derive S_k based on the score method for testing the kth partial hypothesis with π_0 treated as a nuisance parameter. Under the null hypothesis $\mathrm{H}_0 = \cap_{k=1}^{K}\{\mathrm{RR}_k = \mathrm{RR}_k^0\}$, the joint distribution of the score test statistics is asymptotically multivariate normal with mean zero and correlation matrix $\Lambda(\pi_0)$ with off-diagonal elements in product correlation form as given in (2.7). Therefore, simultaneous score confidence intervals are constructed inverting

$$S_k = \frac{(\widehat{\pi}_k - \mathrm{RR}_k^0\widehat{\pi}_0)}{\sqrt{\widetilde{\pi}_k(1-\widetilde{\pi}_k)/n_k + (\mathrm{RR}_k^0)^2\widetilde{\pi}_{0|k}(1-\widetilde{\pi}_{0|k})/n_0}}, \qquad (2.9)$$

using the same critical value h in (2.8) for simultaneous confidence intervals constructed by inverting the Wald test statistics.

2.4 Summary

We have reviewed the Wald, Fieller and score methods to construct large sample confidence intervals for a risk ratio. Among them, the Wald and score methods have reasonable coverage properties and may be extended to simultaneous confidence intervals for multiple risk ratios by adjusting for multiplicity. Klingenberg (2010) demonstrated that inverting the absolute maximum of the score test statistics with Dunnett's critical value yields confidence intervals having good coverage probabilities for multiple risk ratios. Unfortunately, the optimality of the score method comes at a price of computational complexity. Although it is no longer as important an issue as in the past (Gart and Nam, 1988), the merit of computationally simpler alternatives is apparent, especially for a large number of comparisons, as remarked by Klingenberg (2010).

Therefore, we propose a simple, alternative approach based on the method of variance estimates recovery (MOVER) (Zou, 2008; Zou and Donner, 2008) to constructing simultaneous confidence intervals for risk ratios using the confidence limits of the corresponding single proportions. The MOVER approach for a ratio of parameters is a generalized procedure that extends Fieller's theorem (Fieller, 1944) without the normality assumption required for the components of the ratio (Li *et al.*, 2010). The variance for each component's point estimator is obtained separately at its lower and upper confidence limits. As the MOVER incorporates the skewness of the sampling distributions of the point estimators, the confidence intervals obtained by the MOVER reduce to those by Fieller's theorem asymptotically under the assumption of symmetric sampling distributions (Zou, 2008). Using the variance estimates of individual sample proportions separately, the MOVER can be readily applicable to constructing confidence intervals for risk ratios estimated from correlated binary data or for more complicated functions such as ratios of linear combinations of proportions.

To further simplify the simultaneous confidence interval procedures, we compute the Dunnett critical value, assuming a common, constant correlation coefficient between two comparisons to control under the assumption of the multivariate normal distribution of proportions in balanced sample sizes for risk differences or risk ratios. Under complete homogeneity of proportions, $\lambda_{ij}(\pi_0)$ in (2.7) reduces to 0.5, resulting in a critical value that depends only on the desired confidence level α and the number of treatment groups K instead of sample correlation estimates. Therefore, it not only obviates the need for estimating correlation coefficients but also prevents potentially introducing additional variability to the procedure by computing the critical value using sample estimates. This approximation approach was considered by Agresti *et al.* (2008) analogously for constructing simultaneous intervals by inverting a score test for a variety of effect measures in all pairwise comparisons. An appropriate critical value was obtained assuming the asymptotic studentized range distribution with an infinite number of degrees of freedom.

Chapter 3

METHODS

The MOVER is a general approach for constructing confidence intervals for linear functions of parameters and their ratios using the upper and lower confidence limits for each parameter. Reflected in the name of the method, *Method Of Variance Estimates Recovery*, it recovers variance estimates of a simple function of parameters from the lower and upper confidence limits for the component parameters. Contrary to the standard confidence interval method assuming an approximate normal sampling distribution, the MOVER does not require any specific assumptions on the sampling distributions. Following the exposition of the MOVER by Zou and Donner (2008); Zou (2008); Donner and Zou (2011), we describe the method for the simplest case of constructing a confidence interval for a difference of two parameters in Section 3.1. The MOVER approach to confidence intervals for a single and several risk ratios is presented in Sections 3.3 and 3.4, respectively. A summary of this chapter is provided in Section 3.5.

3.1 Confidence interval for a difference between two parameters

Suppose we are interested in constructing a confidence interval for $\theta_1 - \theta_2$ based on independently distributed point estimates $\widehat{\theta}_i$ and confidence limits (l_i, u_i), $j = 1,2$. An approximate two-sided $100(1-\alpha)\%$ confidence interval (L,U) for $\theta_1 - \theta_2$ based on the central limit theorem is given by

$$(L,U) = \widehat{\theta}_1 - \widehat{\theta}_2 \mp z_{\alpha/2}\sqrt{\mathrm{var}(\widehat{\theta}_1) + \mathrm{var}(\widehat{\theta}_2)}, \tag{3.1}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution. The variances $\mathrm{var}(\widehat{\theta}_i)$, $i = 1,2$ remain to be estimated. Asymptotically, the lower and upper confidence limits of a two-sided $100(1-\alpha)\%$ interval for the component parameters θ_i, $i = 1,2$ are also given as

$$(l_i,u_i) = \widehat{\theta}_i \mp z_{\alpha/2}\sqrt{\mathrm{var}(\widehat{\theta}_i)}.$$

In estimating $\mathrm{var}(\widehat{\theta}_i)$, the Wald method assumes that the sampling distributions of $\widehat{\theta}_i$ are close to normal, implying that $\mathrm{var}(\widehat{\theta}_i)$ is constant for all values of θ_i. Therefore, the Wald method yields symmetric confidence intervals (l_i,u_i) for the component parameters θ_i ($i = 1,2$) and thus a symmetric confidence interval (L,U) for a difference $\theta_1 - \theta_2$. Nevertheless, unless the assumption of approximate normality of $\widehat{\theta}_i$ holds, the Wald method may have poor coverage even with moderately large samples (Brown *et al.*, 2001, 2002). A good example is the case of proportions and the difference of two proportions where the variance estimate depends on the true proportion and the sampling distribution may be asymmetrical (Newcombe, 1998*a*,*b*; Brown *et al.*, 2001, 2002).

The performance of interval methods for $\theta_1 - \theta_2$ can be improved by acknowledging the skewness of the sampling distributions of the parameter estimates. The MOVER separately estimates $\mathrm{var}(\widehat{\theta}_1)$ and $\mathrm{var}(\widehat{\theta}_2)$ in the neighbourhood of the confidence limits, L and U, obviating the approximate normality assumption of the Wald method. The variance estimation for $\mathrm{var}(\widehat{\theta}_1 - \widehat{\theta}_2)$ by the MOVER is similar to the principle of the score method estimating the variances at the confidence limits and L and U by an iterative procedure (Zou and Donner, 2008). We may say $100(1-\alpha)\%$ confidence limits L and U for $\theta_1 - \theta_2$ correspond to the minimum and maximum values, respectively, of $\theta_1 - \theta_2$ satisfying

$$\frac{[(\widehat{\theta}_1 - \widehat{\theta}_2) - (\theta_1 - \theta_2)]^2}{\mathrm{var}(\widehat{\theta}_1) + \mathrm{var}(\widehat{\theta}_2)} < z^2_{\alpha/2}.$$

Among plausible parameter values of $\theta_1 - \theta_2$, $l_1 - u_2$ is near the lower limit L and $u_1 - l_2$ near the upper limit U. Therefore, we estimate the lower margin of error of $\widehat{\theta}_1 - \widehat{\theta}_2$ with $\widehat{\mathrm{var}}(\widehat{\theta}_i)$ under the assumptions, $\theta_1 = l_1$ for $\mathrm{var}(\widehat{\theta}_1)$ and $\theta_2 = u_2$ for $\mathrm{var}(\widehat{\theta}_2)$. Similarly, we estimate the upper margin of error of $\widehat{\theta}_1 - \widehat{\theta}_2$ with the variances estimated at $\theta_1 = u_1$, $\theta_2 = l_2$. In fact, it can be shown that the distance between $l_1 - u_2$ and L given by

$$z_{\alpha/2}\left\|\sqrt{\mathrm{var}(\widehat{\theta}_1)+\mathrm{var}(\widehat{\theta}_2)} - \left[\sqrt{\mathrm{var}(\widehat{\theta}_1)} + \sqrt{\mathrm{var}(\widehat{\theta}_2)}\right]\right\|$$

is smaller than the distance between the point estimates $\widehat{\theta}_1 - \widehat{\theta}_2$ and L given by

$$z_{\alpha/2}\left\|\sqrt{\mathrm{var}(\widehat{\theta}_1)+\mathrm{var}(\widehat{\theta}_2)}\right\|$$

Likewise, the distance between $u_1 - l_2$ and U is smaller than that between $\widehat{\theta}_1 - \widehat{\theta}_2$ and U (Li *et al.*, 2010).

By the duality between confidence interval estimation and hypothesis testing (Casella and Berger, 1990),

$$\frac{\widehat{\theta}_i - l_i}{\sqrt{\mathrm{var}(\widehat{\theta}_i)}} \approx z_{\alpha/2} \quad \text{yields} \quad \widehat{\mathrm{var}}(\widehat{\theta}_i) \approx \frac{(\widehat{\theta}_i - l_i)^2}{z^2_{\alpha/2}}$$

at $\theta_i = l_i$ and

$$\frac{u_i - \widehat{\theta}_i}{\sqrt{\mathrm{var}(\widehat{\theta}_i)}} \approx z_{\alpha/2} \quad \text{yields} \quad \widehat{\mathrm{var}}(\widehat{\theta}_i) \approx \frac{(u_i - \widehat{\theta}_i)^2}{z^2_{\alpha/2}}$$

at $\theta_i = u_i$. Substituting these variance estimates in 3.1, the lower limit L for $\theta_1 - \theta_2$ is given by

$$\begin{aligned} L &\approx \widehat{\theta}_1 - \widehat{\theta}_2 - z_{\alpha/2}\sqrt{\widehat{\mathrm{var}}(\widehat{\theta}_1)+\widehat{\mathrm{var}}(\widehat{\theta}_2)} \\ &= \widehat{\theta}_1 - \widehat{\theta}_2 - z_{\alpha/2}\sqrt{(\widehat{\theta}_1 - l_1)^2/z^2_{\alpha/2} + (u_2 - \widehat{\theta}_2)^2/z^2_{\alpha/2}} \\ &= \widehat{\theta}_1 - \widehat{\theta}_2 - \sqrt{(\widehat{\theta}_1 - l_1)^2 + (u_2 - \widehat{\theta}_2)^2}. \end{aligned} \tag{3.2}$$

Similarly, the upper limit U for $\theta_1 - \theta_2$ is given by

$$U \approx \widehat{\theta}_1 - \widehat{\theta}_2 + \sqrt{(u_1 - \widehat{\theta}_1)^2 + (\widehat{\theta}_2 - l_2)^2}. \tag{3.3}$$

By mathematical induction, the MOVER approach can be generalized to the construction of a confidence interval for a contrast of parameters $\sum_{i=1}^{I} c_i\theta_i$ (Zou, 2008; Zou *et al.*, 2009). The resulting limits for a contrast of independent parameter estimates are given by

$$\begin{cases} L^c = \sum_{i=1}^{I} c_i\widehat{\theta}_i - \sqrt{\sum_{i=1}^{I}\left[c_i\widehat{\theta}_i - \min(c_il_i, c_iu_i)\right]^2} \\ U^c = \sum_{i=1}^{K} c_i\widehat{\theta}_i + \sqrt{\sum_{i=1}^{I}\left[c_i\widehat{\theta}_i - \max(c_il_i, c_iu_i)\right]^2} \end{cases} . \tag{3.4}$$

As Zou (2009) pointed out, a similar approach has been applied to constructing a confidence interval for variance components, however, by assuming that the limits are of a certain form and computing them under special conditions (Howe, 1974; Graybill and Wang, 1980). Newcombe (1998*b*) also considered the MOVER approach to confidence interval construction for a difference between two independent proportions, terming it as the *square-and-add* approach, without providing a theoretical justification (Newcombe, 2001, 2011; Zou and Donner, 2008; Zou, 2009).

The MOVER is derived on two fundamental principles (Zou, 2009). First, a confidence interval method using variance estimates obtained at or close to the limits has better coverage properties than the standard Wald method using variance estimates obtained at the point estimates (Efron, 1987). Second, the variance estimates for a linear combination of parameter estimates are contained in and thus may be recovered from the confidence limits about each parameter estimate. Therefore, the MOVER permits more accurate variance estimation for the function of component parameters based on their variance estimates, possibly of different sizes, without any specific distributional assumptions. The performance of the MOVER, therefore, crucially depends on the accuracy of the variance estimators for the component parameters. We thus review various interval methods for a single proportion before considering the MOVER approach for the ratio of independent proportions.

3.2 Confidence interval for a single proportion

Assume that y is the observed number of events out of n trials following the binomial distribution with a true proportion π and its maximum likelihood estimate $\widehat{\pi} = y/n$. The $\alpha/2$ upper quantile of the standard normal distribution is denoted by $z_{\alpha/2}$.

3.2.1 The Clopper-Pearson (exact) method

Inverting the two one-tailed exact tests for binomial proportions, the Clopper-Pearson confidence limits l_e and u_e are obtained for $y = 1, 2, \ldots, n-1$ to satisfy,

$$\sum_{k=y}^{n} \binom{n}{k} l_e^k (1-l_e)^{n-k} \leq \alpha/2$$

and

$$\sum_{k=0}^{y} \binom{n}{k} u_e^k (1-u_e)^{n-k} \leq \alpha/2,$$

where l_e is set to 0 when $y = 0$ and u_e is set to 1 when $y = n$. These limits may be obtained using either quantiles from the F or the beta distribution. The lower limit l_e is the $\alpha/2$ quantile of a beta distribution with parameters y and $n-y+1$, and the upper limit u_e is the $1-\alpha/2$ quantile of a beta distribution with parameters $y+1$ and $n-y$. This method yields no aberrant limits and guarantees a coverage probability strictly at least $100(1-\alpha)\%$ for all parameter values π with $0 < \pi < 1$. Nevertheless, the resulting confidence interval tends to be too conservative due to the discreteness of the binomial distribution (Newcombe, 1998*a*). The method's conservativeness may be reduced by using a continuity correction such as the mid-P enumeration of the tail areas (Brown *et al.*, 2001).

3.2.2 The Wald method

Assuming the asymptotic normality of $\widehat{\pi} = \frac{y}{n}$, the simplest asymptotic method inverts the Wald test statistic to yield confidence limits given as

$$(l_W, u_W) = \frac{y}{n} \mp z_{\alpha/2}\sqrt{\frac{y(n-y)}{n^3}}.$$

The Wald method applies the normal distribution approximation with a constant variance estimate regardless of the value of the true population proportion π. It has been known that ignoring the dependence of the variance on a proportion estimate may result in erratic and poor coverage performance (Brown *et al.*, 2001, 2002; Newcombe, 1998*a*) even for a large sample size or a true proportion away from the boundary values. In finite samples, the assumption of approximate normality of the Wald test is unreasonable due to the correlation between the sample proportion estimator and its variance. Both the bias and skewness of the pivotal result in less accurate coverage probabilities and highly unbalanced noncoverage probabilities (Andersson, 2009).

Moreover, this method has an inherent problem of the violation of the boundaries, also known as *overshoot*. Truncation to 0 or 1 is a quick remedy for such cases but it is not recommendable as it obscures the nature of the problem. For example, when $y = 1$, the lower limit l_W is generally negative and may be truncated to 0. Nevertheless, the lower limit l_W cannot be zero because $\pi = 0$ is ruled out by the data (Newcombe, 1998*a*).

3.2.3 The Wilson method

A refinement of the Wald method is to use the true asymptotic variance of the sampling distribution $n\pi(1-\pi)$ in the asymptotic test inversion by solving the resulting quadratic equation for π. This method is also known as the score method, which is derived from the efficient score approach and has its theoretical advantages (Wilks, 1938). The Wilson limits are given by

$$(l_s, u_s) = \frac{y + z^2_{\alpha/2}/2}{n + z^2_{\alpha/2}} \mp z_{\alpha/2} \frac{\sqrt{y(n-y)/n + z^2_{\alpha/2}/4}}{n + z^2_{\alpha/2}}.$$

The lower limit l_s may be regarded as the mean of the normal distribution with variance $nl_s(1-l_s)$ used to approximate the upper tail of the binomial distribution (n, l_s) such that

$$\Pr\left(\frac{y - nl_s}{\sqrt{nl_s(1-l_s)}} > z_{\alpha/2}\right) \approx \alpha/2.$$

Similarly, the upper limit u_s may be regarded as the mean of the normal distribution with variance $nu_s(1-u_s)$ approximating the lower tail of the binomial distribution (n, u_s) such that

$$\Pr\left(\frac{nu_s - y}{\sqrt{nu_s(1-u_s)}} < z_{\alpha/2}\right) \approx \alpha/2.$$

Wilson (1927) has pointed out that this interval is narrower than (l_w, u_w) when the underlying parameter is in the range of $0.5 \mp 0.5\sqrt{1-(2+z_{\alpha/2}/n)}$. A continuity-corrected version of the Wilson method is also available but its performance is inferior to that without a continuity correction (Brown *et al.*, 2001, 2002).

3.2.4 The Agresti-Coull method

The Agresti-Coull method is an adjusted Wald method using an approximate midpoint of the score interval (l_s, u_s). The method yields confidence limits given by

$$(l_A, u_A) = \tilde{\pi} \mp z_{\alpha/2}\sqrt{\frac{\tilde{\pi}(1-\tilde{\pi})}{n + z^2_{\alpha/2}}}$$

where

$$\tilde{\pi} = \frac{y + z^2_{\alpha/2}/2}{n + z^2_{\alpha/2}}.$$

As this method still uses the normal distribution with a constant variance estimate to approximate two binomial distributions of different shapes, it may yield limits out of the parameter space, inheriting the deficiency of the Wald method. As the score method is asymptotically optimal (Wilks, 1938), this methods is asymptotically inefficient and the resulting confidence intervals are never shorter than the Wilson intervals (Wilson, 1927; Brown *et al.*, 2001, 2002). The Agresti-Coull method is the shrinkage representation of the score method (Agresti and Coull, 1998), which is essentially equivalent to Wilson (1927)'s attempt of plugging in the midpoint of a score interval in the Wald method for computational simplicity. However, the Agresti-Coull method may also be considered as one of similar proportion estimators adding pseudo observations to the number of successes and the number of failures merely on intuitive grounds (Donner and Zou, 2011).

3.2.5 The Jeffreys method

As a convenient prior distribution for proportions to be assumed in the Bayesian framework, the Jeffreys method uses a noninformative Beta (0.5, 0.5) distribution (Brown *et al.*, 2001, 2002). The Jeffreys limits can be numerically obtained from the $\alpha/2$ and $(1-\alpha/2)$ quantiles of the posterior distribution Beta or using its link to the F-distribution,

$$(l_J, u_J) = \Big(B(\alpha/2, x+0.5, n-x+0.5), B(1-\alpha/2, x+0.5, n-x+0.5)\Big),$$

where $B(\alpha; a, b)$ denotes the α quantile of Beta(a,b). Taking the central $1-\alpha$ posterior probability interval results in a $\alpha/2$ non-coverage probability in each tail except for the boundary number of event, $x=0$ or $x=1$, in which case the limits are modified to avoid undesirable coverage results. A closed-form expression for approximate confidence limits obtained by the general approximation to a Beta quantile is also provided by Brown *et al.* (2001).

3.3 Confidence interval for a ratio of proportions

A confidence interval for the ratio of proportions can be constructed on the log scale, relying on the invariance property of interval estimation to a transformation by a monotonic function (Daly, 1998). Therefore, the MOVER approach described in Section 3.1 can be applied to construct a confidence interval for the difference of two log proportions (Zou, 2008; Zou and Donner, 2008; Zou, 2009). However, the MOVER approach generalizing Fieller's theorem is a more direct procedure for a ratio of proportions (Li *et al.*, 2010).

Denote the ratio of two proportions by

$$\mathrm{RR} = \pi_1/\pi_0,$$

which is equivalent to

$$\pi_1 - \mathrm{RR}\pi_0 = 0. \tag{3.5}$$

As $\widehat{\pi}_0$ and $\widehat{\pi}_1$ are independent and $\mathrm{RR} > 0$, we can obtain a confidence interval (l, u) for $\pi_1 - \mathrm{RR}\pi_0$ as,

$$\begin{cases} l = \widehat{\pi}_1 - \mathrm{RR}\widehat{\pi}_0 - \sqrt{(\widehat{\pi}_1 - l_1)^2 + \mathrm{RR}^2(u_0 - \widehat{\pi}_0)^2} \\ u = \widehat{\pi}_1 - \mathrm{RR}\widehat{\pi}_0 + \sqrt{(u_1 - \widehat{\pi}_1)^2 + \mathrm{RR}^2(\widehat{\pi}_0 - l_0)^2}, \end{cases}$$

where the confidence limits for the individual proportions are (l_0, u_0) and (l_1, u_1). To satisfy (3.5), the resulting confidence limits for a constant are constrained to be zero (i.e. $l = 0$ and $u = 0$). Solving the quadratic equations in RR, the smaller root of $l = 0$ and the larger root of $u = 0$ correspond to the lower and upper confidence limits for RR, respectively, yielding confidence limits $(L^{\mathrm{RR}}, U^{\mathrm{RR}})$ given as

$$\begin{cases} L^{\mathrm{RR}} = \dfrac{\widehat{\pi}_1\widehat{\pi}_0 - \sqrt{(\widehat{\pi}_1\widehat{\pi}_0)^2 - l_1 u_0(2\widehat{\pi}_1 - l_1)(2\widehat{\pi}_0 - u_0)}}{u_0(2\widehat{\pi}_0 - u_0)} \\ U^{\mathrm{RR}} = \dfrac{\widehat{\pi}_1\widehat{\pi}_0 + \sqrt{(\widehat{\pi}_1\widehat{\pi}_0)^2 - u_1 l_0(2\widehat{\pi}_1 - u_1)(2\widehat{\pi}_0 - l_0)}}{l_0(2\widehat{\pi}_0 - l_0)}. \end{cases} \tag{3.6}$$

3.4 Simultaneous confidence intervals for multiple risk ratios

As discussed in Chapter 2, a confidence interval procedure for single inference can be extended for multiple inferences by adjusting the critical value to control the joint coverage probability. Suppose Y_k are independently distributed binomial variates, $Y_k \sim \mathrm{Bin}(n_k, \pi_k)$. We denote the numbers of events in treatment groups by y_k and the corresponding sample sizes of the treatment groups by n_k ($k = 0, 1, \ldots, K$). We describe the steps to apply the MOVER generalizing Fieller's Theorem to construct simultaneous confidence intervals for multiple risk ratios in the many-to-one comparisons of proportions.

3.4.1 Computation of a critical value

In the many-to-one comparisons of binomial proportions, a critical value may be computed incorporating the correlation information among test statistics using plug-in estimates as described in Chapter 2 (Piegorsch, 1991; Donner and Zou, 2011; Schaarschmidt *et al.*, 2009; Klingenberg, 2010) or those estimated under the null hypotheses (Klingenberg, 2012). Alternatively, instead of estimating correlation coefficients from a sample, we may assume a common correlation of 0.5 for all comparisons to control, giving out the correlation matrix $R^*_{K\times K}$ as

$$R^*_{K\times K} = \begin{bmatrix} 1 & 0.5 & 0.5 & 0.5 & \cdots & 0.5 \\ 0.5 & 1 & 0.5 & 0.5 & \cdots & 0.5 \\ 0.5 & 0.5 & 1 & 0.5 & \cdots & 0.5 \\ 0.5 & 0.5 & 0.5 & 1 & 0.5 & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & 0.5 \\ 0.5 & \cdots & \cdots & \cdots & 0.5 & 1 \end{bmatrix}.$$

The Dunnett critical value for constructing simultaneous confidence intervals with a specific level of confidence is obtained from the SAS procedure `probmc` without passing the parameter estimates as `probmc ("Dunnett2", . , confidence, . ,K)`, or the R function `qmvnorm(confidence, corr, "both")` in the package `mvtnorm`, specifying corr $= R^*_{K \times K}$ as the arbitrary correlation matrix required for computation. Therefore, the resulting critical value depends only on the level of confidence and the number of experimental groups K compared with the control group.

Assuming a common correlation for all comparisons might seem too simple; however, it may be a pragmatic approach to determine an approximate critical value quite close to the exact Dunnett critical value. In fact, for normal means with unequal sample sizes and unequal variances, the mean of $K(K-1)/2$ correlation coefficients results in quite a close approximate value of the exact Dunnett critical value (Hochberg and Tamhane, 1987, pp.144-146).

The chosen correlation value of 0.5 is justifiable under the assumption of the multivariate normal distribution with equal sample sizes and equal variances for many-to-one comparisons of means, analogously under the complete homogeneity of proportions for many-to-one comparisons of proportions with the large sample assumptions. If the control group proportion is much greater than either of the two experimental group proportions in comparison, the correlation between the two risk ratios is close to 0; it is close to 1 for the converse (Holford *et al.*, 1989). However, the true correlation between two risk ratios is unknown and unobservable.

Therefore, the use of a common correlation value of 0.5 not only obviates estimating sample correlation coefficients but also prevents introducing additional variability in the confidence interval procedure due to the computation of the critical value. With correlation being inherent in the trial design for many-to-one comparisons, the exchangeable correlation matrix $R^*_{K \times K}$ must be a better choice than the identity matrix of the Sidak method. Moreover, unless the mean correlation deviates substantially from 0.5, the assumed value is reasonable and does not substantially affect the coverage probabilities of simultaneous confidence intervals (Julious and McIntyre, 2012).

3.4.2 Computation of simultaneous confidence limits

In any of the closed-form of expression of the MOVER, the critical value does not appear because it is cancelled as shown in equation (3.2). It implies that the variance estimates for a function of parameter estimates are contained in the confidence limits about each parameter estimate. Therefore, the multiplicity adjustment in the construction of simultaneous confidence intervals based on the MOVER is made on the confidence limits for the component parameters. We use a multiplicity-adjusted critical value to recover the variance estimates for correlated risk ratios and obtain their confidence limits in two steps:

1. Obtain the upper and lower confidence limits for π_i where $k = 0, 1, \ldots, K$ using a critical value z_α obtained from the multivariate normal distribution with mean 0 and the exchangeable correlation matrix $R^*_{K \times K}$.

2. Compute simultaneous confidence limits for the kth risk ratio using the MOVER generalizing Fieller's Theorem in Section 3.3.

3.5 Summary

The MOVER is applicable to constructing an approximate confidence interval for a ratio of proportions, assuming no specific sampling distributions for individual proportions but using their point estimates and confidence limits. When the sampling distributions of proportion estimates are substantially skewed, the skewness can be incorporated in the variance estimation for the ratio of proportions. The MOVER recovers variance estimates near the confidence limits, adopting the idea of the score method, but requiring no iterative algorithms. Therefore, the performance of the confidence interval constructed by the MOVER for a ratio of proportions depends crucially on the confidence interval method for the two individual proportions.

The MOVER approach for a ratio of proportions may be extended to correlated multiple ratios of proportions by adjusting the critical value required to compute the confidence limits for proportions. For the many-to-one comparisons, a Dunnett critical value may be obtained from a multivariate normal distribution assuming a constant, common correlation coefficient instead of the sample correlation coefficients. Compared to the latter, the former multiplicity adjustment accounts for the correlations among comparisons without introducing additional variability. Therefore, approximate simultaneous confidence intervals for multiple ratios of proportions constructed by the MOVER with such a critical value have good coverage properties in finite samples, provided that the chosen confidence interval method for single proportions performs well.

Among the confidence interval methods described in Section 3.2, it has been demonstrated that approximate confidence intervals by Wilson, Jeffreys or Agresti-Coull perform better than the Wald or the Clopper-Pearson intervals in terms of having coverage probabilities close to the nominal level (Agresti and Coull, 1998; Newcombe, 1998*a*; Brown *et al.*, 2001, 2002). Specifically, score intervals or Jeffreys intervals are recommended for small sample sizes and Agresti-Coull intervals for large sample sizes (Brown *et al.*, 2001). Nevertheless, the Agresti-Coull method produces unnecessarily wide intervals for single proportions (Donner and Zou, 2011; Brown *et al.*, 2002), thus we consider the MOVER with score and Jeffreys limits for single proportions in our simulation study in Chapter 4.

Chapter 4

SIMULATION STUDY

Both the test-inversion methods in Chapter 2 and the MOVER approach in Chapter 3 require the assumption of large samples, which may not be always satisfied in practice. Therefore, we evaluate the performance of these methods with different sample sizes in a simulation study. As discussed in the previous chapters, simultaneous confidence interval procedures for several parameters are extended from the procedures for a single parameter by adjusting for multiplicity. Consequently, simultaneous confidence interval procedures presuppose the validity of the corresponding procedures for a single parameter. However, no formal evaluations of the MOVER approach compared to the test-inversion methods have been done specifically in the context of the ratio of binomial proportions. Therefore, this chapter evaluates the four competing methods for a single risk ratio and multiple risk ratios in a simulation study. For the confidence limits for proportions in the MOVER approach, we consider those obtained by the Wilson method and the Jeffreys method as recommended for their desirable coverage properties in small sample sizes (Newcombe, 1998*a*; Donner and Zou, 2011; Brown *et al.*, 2001, 2002).

We evaluate the four competing confidence intervals by computing the performance measures for accuracy and precision, addressing the important issues of coverage probability, conservatism and interval width for binomial proportions (Newcombe, 1998*a*). For single risk ratios, the proximity of empirical percentages of coverage and left and right tail errors to their nominal levels is the evaluation criterion for accuracy. For comparable confidence intervals in terms of accuracy, the precision of intervals is evaluated by comparing their median interval widths. The analogous performance measures for multiple risk ratios include joint coverage percentages and median total interval widths as the tail error balance is irrelevant in case of multiple comparisons.

Due to the discreteness of the binomial distribution, the coverage probability for binomial proportions oscillates with both sample sizes and unknown population proportions (Newcombe, 1998*a*; Brown *et al.*, 2001). Therefore, we evaluate the performance of the interval methods in two parts. We compare the performance measures computed at selected parameter settings to examine how the operating characteristics are affected by the changes in the true proportions and sample sizes. We attempt to evaluate the overall performance of the methods by considering different combinations of population risk ratios, formed by proportions spanning the whole parameter space at a given set of sample sizes. By taking a large number of different combinations of proportion values, we intend to cover a large region of the entire parameter space for proportions. Therefore, the results from the second part serve as the basis to determine a recommended method for use in practice.

Following the recommendations in Burton *et al.* (2006), the design and evaluation criteria of the simulation study are described in Section 4.1. The simulation study results are presented in Section 4.2 and followed by a discussion in Section 4.4. This chapter ends with a summary of the results in Section 4.5.

4.1 Design of simulation studies

4.1.1 Data generation

For selected parameter settings, we consider three different control group proportions $\pi_0 = 0.10, 0.20, 0.30$ and two risk ratios, $\mathrm{RR} = 1, 3$ for a single risk ratio. For multiple risk ratios $\{\mathrm{RR}_1, \ldots, \mathrm{RR}_K\}$ comparing K experimental groups to the control group, the experimental group proportions are determined by a vector of equally spaced population risk ratios

$\mathrm{RR} \in [1,3]$. Therefore, the experimental group proportions are determined by the specific choices of risk ratios and control group proportions, given as $\pi_k = \mathrm{RR}_k\pi_0$, $k = 1,\ldots,K$. We consider $K = 2,3,4$ comparisons to the control group. Table 4.1 shows population proportions $\pi = \{\pi_0,\pi_1,\ldots,\pi_K\}$ for the control and K experimental groups. For unrestricted parameter settings, 1000 sets of $K+1$ group proportions will be independently sampled from the uniform distribution $U(0,1)$ such that $\pi_k \sim U(0,1)$ for $k = 0,\ldots,K$. The population proportions are randomly chosen and intended to cover a large region of the whole parameter space $\pi \in (0,1)^{K+1}$.

We consider equal sample sizes $n = 10,20,30,50,100$ and slightly unequal sample sizes where two adjacent groups differ by 10 observations (i.e. $n_k = n_{k-1} + 10$ for $k = 1,\ldots,K$). For each simulation scenario, we obtain a set of $K+1$ independent binomial variates $Y_k \sim Bin(\pi_k,n_k)$ for $k = 0,1,\ldots,K$ as the number of events in each group. All simulations are performed using exact computations in SAS 9.2 PROC IML (SAS Institute, NC).

4.1.2 Computation of confidence intervals

As discussed in Chapter 2, the Wald method yields incomputable or yield aberrant confidence intervals for extreme outcomes $y_k = 0$ or $y_k = n_k$. When both treatment and control groups have zero events, none of the methods are computable. To avoid such incomputable cases and prevent aberrant confidence limits, we replace all extreme counts $y_k = 0$ and $y_k = n_k$ with $y_k\prime = 0.5$ and $y_k\prime = n_k - 0.5$ and compute confidence limits using these modified counts $y_k\prime$ for all these four methods. Although it is not necessary for the score method or the MOVER approach, it also avoids undefined point estimates of risk ratios and ensures consistent comparison across all methods.

The existing methods of inverting the Wald and score test statistics are referred to as the Wald method and the score method, respectively. We examine two additional intervals

Table 4.1: Selected proportions considered in the simulation study

K	$(\mathrm{RR}_1, \ldots, \mathrm{RR}_K)$	$(\pi_0, \ldots, \pi_K)$
1	1.00	0.10, 0.10
		0.20, 0.20
		0.30, 0.30
	3.00	0.10, 0.30
		0.20, 0.60
		0.30, 0.90
2	1.00, 3.00	0.10, 0.10, 0.30
		0.20, 0.20, 0.60
		0.30, 0.10, 0.90
3	1.00, 2.00, 3.00	0.10, 0.10, 0.20, 0.30
		0.20, 0.20, 0.40, 0.60
		0.30, 0.10, 0.60, 0.90
4	1.00, 1.67, 2.33, 3.00	0.10, 0.10, 0.17, 0.23, 0.30
		0.20, 0.20, 0.34, 0.46, 0.60
		0.30, 0.30, 0.51, 0.75, 0.90

constructed by applying the MOVER for a ratio (Li *et al.*, 2010) with the confidence limits obtained by the Wilson method and the Jeffreys method for single proportions. These methods will be referred to as the MOVER with score limits and the MOVER with Jeffreys limits. Both the score method and the MOVER with score confidence limits are referred to as a score-based method throughout the thesis.

4.1.3 Performance measures and evaluation criteria

Single risk ratio

Coverage For a selected parameter setting, an empirical coverage percentage is obtained by computing the percentage of the confidence intervals containing the true risk ratio from 10,000 simulation runs. The number of simulation is chosen to yield a desired simulation margin of error of 0.4%. Therefore, we consider the methods yielding empirical coverage percentages between 94.6% and 95.4% appropriate for the 95% nominal confidence level. Similarly, for unrestricted parameter settings, an empirical coverage probability for each of 1000 parameter combinations is evaluated on 1000 simulation runs. The numbers of parameter combinations and simulation runs are chosen for practical reasons, considering the required computational time. We compare the methods in terms of the proximity of the median coverage percentage to the nominal confidence level. In addition, we examine the distribution of the 1000 empirical coverage percentages and prefer a method yielding more coverage percentages within 0.2α, equivalently, between 94% and 96% (Schaarschmidt *et al.*, 2009). An empirical coverage percentage exceeding the nominal level indicates overcoverage, implying a method's conservativeness that may lead to a loss of precision. To the contrary, undercoverage with an empirical coverage percentage lower than the appropriate level implies that a method is too liberal.

Tail error balance The left and right tail error percentages estimate the probability that the true risk ratio is missed from the left (ML) and from the right (MR) of a confidence interval. Related to the interval location, the criterion is also important for two-sided confidence intervals as they are intended to provide accurate inference in both directions (Efron, 2003; Newcombe, 1998*a*). For the selected parameter settings, we examine whether the left and right tail error percentages are close to the 2.5% nominal level. For unrestricted parameter settings, the imbalance of tail errors is measured by the relative bias percentage given as

$$100\frac{|\mathrm{ML}-\mathrm{MR}|}{\mathrm{ML}+\mathrm{MR}}(\%).$$

Interval width For a selected parameter setting, we consider the median confidence width among the 10,000 computed confidence interval widths to compare the precision of competing confidence interval methods. For unrestricted parameter settings, we examine the distribution of 1000 computed median interval widths, each of which is obtained from 1000 simulation runs. When methods perform well in terms of the first two criteria, we prefer a method yielding shorter intervals.

Multiple risk ratios

Joint coverage For a selected parameter setting, an empirical joint coverage percentage is obtained from 10,000 simulation runs by computing the percentage of the 10,000 sets of confidence intervals simultaneously containing all K true risk ratios RR_k for $k = 1, \ldots, k$. For unrestricted parameter settings, we take 1000 sets of $K+1$ population proportions and obtain empirical joint coverage percentages from 1,000 simulation runs. The evaluation criteria for coverage percentages of the single confidence intervals are applicable analogously to joint coverage percentages of the simultaneous confidence interval methods.

Total interval width The widths of simultaneous confidence intervals for K risk ratios are summed to yield a total confidence interval width. For a selected parameter setting, we compare the median total interval width among the 10,000 sets of K simultaneous confidence intervals. For unrestricted parameter settings, we examine the distribution of 1000 median total interval widths. A method yielding shorter total interval width is preferred when two methods have comparable coverage properties.

4.2 Results for a single risk ratio

4.2.1 Selected parameters

For the selected values of a single risk ratio, the coverage and tail error percentages of the confidence intervals constructed by each method are shown in Tables 4.2, 4.3, 4.5 and 4.6 and their corresponding median widths in Tables 4.4 and 4.7.

For the risk ratio $\mathrm{RR} = 1$, the Wald method is generally quite conservative for most of the values of π_0 with small to moderate sample sizes. However, its coverage percentages improve for $\mathrm{RR} = 3$ with moderate to large sample sizes. The coverage properties improve as proportions π_k increase, yielding coverage percentages within the margin of error for all sample sizes when $\pi_0 = 0.3$ and $\mathrm{RR} = 3$. However, the tail errors of Wald confidence intervals are extremely disparate for small to moderate sample sizes, mostly missing the true risk ratio $\mathrm{RR} = 3$ from the right. The resulting coverage probabilities are substantially greater than the nominal level.

For moderate to large sample sizes of $n = 20$ or more, the score method performs well and yields coverage percentages close to the nominal confidence level for both $\mathrm{RR} = 1,3$. For a small π_0, the nominal coverage probability is achieved with large sample sizes over $n = 50$. The left and right tail error percentages are similar in most cases for $\mathrm{RR} = 1$ but some are disparate for $\mathrm{RR} = 3$. Similar to the Wald method, the score method has typically greater right tail errors.

The MOVER with score limits performs similarly as the iterative score test-inversion method in terms of the coverage and tail errors in most of parameter combinations and sample sizes. However, when $\pi_0 = \pi_1 = 0.1$, the MOVER with score limits yields slightly wider confidence intervals than the score method even for moderate to large samples. The coverage percentages for a unity risk ratio seem too conservative for small proportions unless the sample sizes are large.

Compared to the score-based methods, the MOVER with Jeffreys confidence limits tends to be less conservative but yields wider confidence intervals, particularly more so for a risk ratio $\mathrm{RR} = 3$. However, their coverage percentages are much closer to the nominal level even for small proportions and sample sizes.

4.2.2 Unrestricted parameters

Figures 4.1 and 4.2 display the boxplots of percentages of coverage and relative bias of tail errors, and median interval widths of the confidence intervals for equal sample sizes and unequal sample sizes, respectively.

Coverage percentages Having the lower quartiles around the nominal level, the Wald intervals are uniformly conservative for all sample sizes although their conservativeness reduces as sample sizes increase in both balanced and unbalanced cases. All the other confidence intervals tend to be slightly conservative but attain a median coverage percentage close to the nominal level with sample sizes as small as $n = 20$. Among them, the median coverage percentage for the MOVER approach with Jeffreys limits is closest to the nominal level even for small sample sizes $n = 10$. All four methods have good coverage properties, yielding an approximate lower quartile of 94.5% and an upper quartile of coverage percentages of 96.1% with moderate to large sample sizes of $n = 50, 100$.

Relative bias percentages Similar to the results for the selected parameters, the left and right tail errors of Wald confidence intervals are quite disparate even with large sample sizes in both balanced and unbalanced cases. For equal sample sizes $n = 10$, the left and right tail error percentages are extremely disparate for all methods except the MOVER approach with Jeffreys limits. The median relative bias percentage is above 80% for the Wald and score-based methods. With smaller variability in relative bias percentages, the MOVER approach with Jeffreys limits achieves the nominal, balanced tail errors in more parameter settings than the other methods, even for small sample sizes such as $n = 20$, for both balanced and unbalanced cases.

Interval width Unlike the discrepancies observed in coverage probabilities of the competing methods, they have comparable median interval widths for moderate and large sample sizes. Nevertheless, the MOVER with Jeffreys limits may yield wider intervals than the Wald method, which is the most conservative in terms of coverage percentages with small sample sizes. The MOVER approach with score limits tends to yield shorter median interval widths than the other methods.

4.3 Results for multiple risk ratios

4.3.1 Selected Parameters

The joint coverage percentage and median total width of the simultaneous confidence intervals constructed for a selected set of multiple risk ratios are shown in Tables 4.8 - 4.13. Similar to the results for single risk ratios, the Wald method is consistently more conservative than the other methods for small to moderate values of π_k for $k = 0, \ldots, K$. The score-based methods have similar operating characteristics both in terms of joint coverage percentages and total interval widths. For many parameter combinations, the MOVER with Jeffreys confidence limits on average tends to yield less conservative intervals than the score-based intervals with small sample sizes. The degree of overcoverage of the Wald confidence intervals reduces quickly as π_k's for $k = 0,\ldots,K$ increase. There is a direct correspondence between the conservativeness and precision of the Wald and score-based intervals. The more prominent overcoverage of the Wald intervals than the score-based intervals corresponds to their wider median total interval widths in all parameter settings. The MOVER approach with Jeffreys limits, demonstrating the best coverage properties among all methods, often yields wider median total interval widths particularly with small to moderate sample sizes. The four competing methods have similar performances both with equal and unequal sample sizes.

4.3.2 Unrestricted Parameters

As the performance measures are distributed similarly for the multiple comparisons with $K = 2,3$, the results are shown for $K = 2,4$ in Figures 4.3 and 4.4 for the equal sample sizes, and in Figures 4.5 and 4.6 for the unequal sample sizes. The MOVER with Jeffreys confidence limits outperforms, in terms of achieving the nominal joint coverage percentages, the other methods typically showing conservativeness in more parameter settings. In general, for all methods, the degree of conservativeness reduces more quickly for a small number of comparisons with equal sample sizes than for a large number of comparisons with unequal sample sizes. For example, with $K = 4$ and unbalanced sample sizes, coverage percentages tend to be farther from the nominal level than with $K = 2$ and balanced sample sizes in the conservative direction.

4.4 Discussion

The Wald method generally results in higher coverage percentages than the other methods although the degree of conservativeness reduces as sample sizes increase or proportions are farther from the boundary of (0,1). Therefore, overcoverage appears to be related to inadequate variance estimates of $\ln\widehat{\pi}_k - \ln\widehat{\pi}_0$ under the assumption of symmetric sampling distributions of log proportion estimates, and particularly for proportions near the boundary of the parameter space. Moreover, when a true risk ratio departures from the unity, the Wald method yields more disparate tail error probabilities than the other methods.

To the contrary, the MOVER approach acknowledges asymmetric sampling distributions of binomial proportions. Therefore, the variances for sample proportions are estimated separately near the confidence limits, resulting in more accurate variance estimates and consequently better coverage properties. Regarded as being in the spirit of the score method, the MOVER approach has similar operating characteristics as the score method. The score method yields more disparate tail errors for a risk ratio farther from unity, even with moderate sample sizes, particularly for small proportions. Consequently, the MOVER approach with score limits inherits the same disadvantages of the score method that may yield highly skewed intervals for a large value of risk ratios (Gart and Nam, 1988). Nevertheless, the MOVER approach with Jeffreys limits has superior performance, yielding coverage percentages closest to the nominal level and maintaining the tail error balance. Therefore, the performance the MOVER with two different confidence intervals for proportion corroborates the fact that the validity of the MOVER for a ratio of proportions crucially relies on the validity of confidence intervals for proportions.

For moderate or large sample sizes, the confidence interval widths tend to be similar across all methods despite slight overcoverage of Wald intervals than the other intervals. For small or moderate sample sizes, the score-based methods often yield shorter intervals than the Wald method whereas the MOVER with Jeffreys limits yields some wider intervals than the other methods. However, between the MOVER with Jeffreys limits and the MOVER with score limits, the more frequent occurrences of wider intervals by the MOVER with Jeffreys limits appear to be a direct consequence of the location of the confidence limits for proportions and their range yielding shorter interval widths. For $\alpha = 5\%$, score confidence intervals for proportions have shortest widths when $0.201 \leq \pi \leq 0.799$ whereas

Jeffreys confidence intervals are the shortest when $0.137 \leq \pi \leq 0.201$ or $0.799 \leq \pi \leq 0.863$ (Brown *et al.*, 2002). As our simulation study considers the whole range of the set of proportions randomly sampled from uniform distributions, the MOVER approach with Jeffreys limits may yield wider interval widths more frequently compared to the score-based methods. Based on the simulation results for the MOVER approach and the score method, a method yielding wider confidence intervals does not necessarily imply their general tendency to overcoverage. Rather, there is a trade-off between the desirable coverage and noncoverage properties and the interval widths.

For equal sample sizes, as the number of comparisons K increases, the variability of the median coverage percentages increases (i.e. the interquartile range for $K = 4$ is greater than that for $K = 2$) but their median coverage percentage remains unaffected by the change in the number of comparisons. For unequal sample sizes, the distribution of the median coverage percentages for each method shows a slight upward shift, suggesting a tendency of modest overcoverage compared to the case of equal sample sizes.

4.5 Summary

Both score-based intervals, constructed by either inverting score test statistics or by applying the MOVER with score limits for single proportions, have similar coverage properties. Therefore, the MOVER with score limits can construct confidence intervals similar to score intervals without an iterative algorithm. The Wald method is generally more conservative, resulting in joint coverage percentages above the nominal level more often than the other methods, particularly with small and moderate sample sizes. The MOVER approach with Jeffreys limits typically results in superior coverage properties than the two score-based intervals, yielding coverage percentages closet to the nominal level even with small sample sizes for a wide range of parameter space. Nevertheless, there is a cost to the more desirable small sample properties of the MOVER with Jeffreys limits because it may yield confidence intervals substantially greater than the other confidence intervals under some parameter settings.

Table 4.2: Estimated percentages of coverage, and left and right tail errors (ML, MR) of two-sided 95% confidence intervals constructed for risk ratio RR = 1 with control group proportion π_0 and equal sample sizes of $n_0 = n_1 = n$ by the four methods based on 10,000 simulation runs.

		Coverage (ML, MR)			
π_0	n	Wald	Score	Ms[a]	Mj[b]
0.1	10	100.0 (0.0, 0.0)	100.0 (0.0, 0.0)	100.0 (0.0, 0.0)	99.0 (0.5, 0.5)
	20	100.0 (0.0, 0.0)	98.4 (0.8, 0.8)	98.4 (0.8, 0.8)	96.2 (1.9, 1.9)
	30	99.6 (0.2, 0.2)	96.1 (2.0, 1.9)	98.2 (1.0, 0.8)	**94.8 (2.7, 2.5)**
	50	97.7 (1.1, 1.3)	**95.0 (2.5, 2.5)**	96.4 (1.8, 1.8)	93.9 (3.0, 3.1)
	100	95.9 (2.1, 2.1)	**94.9 (2.6, 2.5)**	**95.6 (2.2, 2.3)**	**94.8 (2.6, 2.6)**
0.2	10	100.0 (0.0, 0.0)	99.0 (0.5, 0.5)	99.0 (0.5, 0.5)	95.8 (2.1, 2.2)
	20	99.0 (0.5, 0.5)	**95.1 (2.5, 2.4)**	**95.1 (2.5, 2.4)**	**94.5 (2.8, 2.7)**
	30	97.4 (1.3, 1.3)	**95.3 (2.4, 2.2)**	95.6 (2.3, 2.1)	**95.3 (2.4, 2.2)**
	50	96.2 (1.9, 2.0)	**95.1 (2.5, 2.5)**	**95.4 (2.3, 2.3)**	**95.1 (2.5, 2.5)**
	100	**95.5 (2.3, 2.2)**	**94.8 (2.6, 2.6)**	**95.0 (2.5, 2.5)**	**94.7 (2.7, 2.6)**
0.3	10	99.7 (0.1, 0.2)	97.3 (1.4, 1.3)	97.3 (1.4, 1.3)	93.7 (3.2, 3.1)
	20	97.6 (1.3, 1.1)	**94.6 (2.7, 2.8)**	**94.6 (2.7, 2.8)**	**94.5 (2.7, 2.8)**
	30	96.5 (2.0, 1.6)	**95.5 (2.5, 2.1)**	**95.5 (2.5, 2.1)**	**95.5 (2.5, 2.1)**
	50	96.0 (2.1, 1.9)	**95.4 (2.5, 2.2)**	**95.4 (2.5, 2.2)**	**95.4 (2.5, 2.2)**
	100	**95.2 (2.6, 2.2)**	**94.9 (2.7, 2.4)**	**95.0 (2.7, 2.4)**	**94.9 (2.8, 2.4)**

[a] MOVER with the Score confidence limits

[b] MOVER with the Jeffreys confidence limits

Table 4.3: Estimated percentages of coverage, and left and right tail errors (ML, MR) of two-sided 95% confidence intervals constructed for risk ratio RR $= 3$ with control group proportion π_0 and equal sample sizes of $n_0 = n_1 = n$ by the four methods based on 10,000 simulation runs.

		Coverage (ML, MR)			
π_0	n	Wald	Score	Ms[a]	Mj[b]
0.1	10	98.2 (0.0, 1.9)	96.2 (0.0, 3.8)	96.2 (0.0, 3.8)	96.1 (0.0, 3.8)
	20	96.8 (0.0, 3.2)	96.8 (0.0, 3.2)	96.8 (0.0, 3.2)	95.5 (1.3, 3.2)
	30	97.6 (0.0, 2.4)	96.0 (0.3, 3.7)	96.1 (0.2, 3.7)	**94.8 (2.8, 2.4)**
	50	96.8 (0.1, 3.1)	**95.4 (1.6, 3.1)**	**95.1 (1.5, 3.3)**	94.3 (2.7, 3.0)
	100	95.7 (1.4, 2.9)	**95.0 (2.0, 3.1)**	**95.2 (1.8, 3.1)**	**94.8 (2.6, 2.6)**
0.2	10	95.9 (0.0, 4.1)	96.2 (0.0, 3.8)	96.2 (0.0, 3.8)	94.3 (1.9, 3.8)
	20	96.2 (0.0, 3.8)	**95.4 (1.4, 3.2)**	**94.8 (1.4, 3.8)**	94.0 (2.9, 3.2)
	30	96.2 (0.2, 3.6)	**94.7 (1.9, 3.4)**	**95.1 (1.5, 3.4)**	**94.5 (2.7, 2.9)**
	50	**95.3 (1.2, 3.5)**	**94.7 (2.0, 3.3)**	**94.6 (2.0, 3.4)**	94.4 (2.5, 3.0)
	100	**95.4 (1.5, 3.1)**	**95.0 (2.1, 2.9)**	**94.8 (2.0, 3.1)**	**94.9 (2.4, 2.7)**
0.3	10	**94.6 (0.0, 5.4)**	96.4 (0.0, 3.6)	96.4 (0.0, 3.6)	93.7 (2.7, 3.6)
	20	**95.2 (0.0, 4.8)**	**95.4 (1.7, 3.0)**	**95.2 (1.7, 3.2)**	**94.7 (2.4, 3.0)**
	30	**95.1 (0.6, 4.3)**	**95.6 (1.8, 2.6)**	**95.0 (1.8, 3.2)**	**95.2 (2.3, 2.5)**
	50	**95.2 (1.1, 3.7)**	**95.0 (2.1, 2.9)**	**95.0 (2.1, 3.0)**	**94.8 (2.2, 2.9)**
	100	**95.1 (1.6, 3.4)**	**94.8 (2.3, 2.9)**	**94.8 (2.3, 2.9)**	**94.7 (2.4, 2.9)**

[a] MOVER with the Score confidence limits

[b] MOVER with the Jeffreys confidence limits

Table 4.4: Median width of two-sided 95% confidence intervals constructed for risk ratio RR $= 1$ and risk ratio RR $= 3$ with control group proportion π_0 and equal sample sizes of $n_0 = n_1 = n$ by the four methods based on 10,000 simulation runs.

		Median width for RR = 1				Median width for RR = 3			
p_0	n	Wald	Score	Ms[a]	Mj[b]	Wald	Score	Ms[a]	Mj[b]
0.1	10	13.8	8.7	8.9	12.2	23.8	18.8	17.9	28.4
	20	6.3	5.1	5.2	6.1	14.0	12.8	12.5	16.2
	30	5.1	3.8	4.0	4.3	9.9	9.4	9.2	11.0
	50	3.0	2.7	2.8	2.9	6.6	6.4	6.3	7.1
	100	1.9	1.8	1.8	1.9	4.4	4.3	4.3	4.5
0.2	10	5.6	4.7	4.5	5.5	10.7	10.4	9.9	13.1
	20	3.2	3.0	2.9	3.2	7.0	7.0	6.8	7.9
	30	2.4	2.3	2.3	2.4	5.1	5.2	5.1	5.6
	50	1.7	1.7	1.7	1.7	3.8	3.8	3.8	4.0
	100	1.2	1.2	1.2	1.2	2.6	2.6	2.6	2.7
0.3	10	3.6	3.3	3.1	3.6	6.8	7.1	7.1	8.4
	20	2.2	2.1	2.1	2.2	4.4	4.6	4.6	5.0
	30	1.7	1.7	1.6	1.7	3.5	3.6	3.6	3.8
	50	1.3	1.3	1.2	1.3	2.7	2.7	2.7	2.8
	100	0.9	0.9	0.9	0.9	1.9	1.9	1.9	1.9

[a] MOVER with the Score confidence limits

[b] MOVER with the Jeffreys confidence limits

Table 4.5: Estimated percentages of coverage, and left and right tail errors (ML, MR) of two-sided 95% confidence intervals constructed for risk ratio RR $= 1$ with control group proportion π_0 and sample sizes of $n_1 = n_0 + 10$ by the four methods based on 10,000 simulation runs.

		Coverage (ML, MR)			
π_0	n_0	Wald	Score	Ms[a]	Mj[b]
0.1	10	99.5 (0.0, 0.5)	98.6 (0.0, 1.4)	98.6 (0.0, 1.4)	98.1 (0.5, 1.4)
	20	99.6 (0.0, 0.4)	98.3 (0.3, 1.4)	98.5 (0.2, 1.4)	95.7 (2.1, 2.2)
	30	98.9 (0.1, 1.0)	96.7 (1.1, 2.3)	96.9 (1.1, 2.0)	**95.1 (2.3, 2.6)**
	50	97.9 (0.5, 1.6)	95.8 (1.9, 2.4)	96.3 (1.7, 1.9)	94.4 (3.2, 2.4)
	100	96.0 (1.9, 2.2)	**94.8 (2.5, 2.7)**	**95.0 (2.5, 2.5)**	**94.5 (2.8, 2.7)**
0.2	10	98.6 (0.0, 1.4)	97.5 (0.4, 2.2)	97.5 (0.4, 2.2)	95.6 (2.3, 2.2)
	20	97.6 (0.3, 2.2)	96.1 (1.5, 2.4)	96.3 (1.4, 2.4)	**95.0 (2.6, 2.4)**
	30	96.4 (0.9, 2.7)	**94.9 (2.4, 2.7)**	**94.9 (2.4, 2.7)**	**94.9 (2.4, 2.7)**
	50	96.1 (1.8, 2.1)	**95.1 (2.5, 2.5)**	**95.4 (2.3, 2.3)**	**95.0 (2.5, 2.5)**
	100	**95.1 (2.4, 2.5)**	**94.7 (2.7, 2.7)**	**94.7 (2.7, 2.7)**	**94.7 (2.7, 2.7)**
0.3	10	98.0 (0.0, 2.0)	96.1 (1.2, 2.7)	96.1 (1.2, 2.7)	94.4 (3.0, 2.7)
	20	96.6 (1.0, 2.5)	**95.4 (2.1, 2.5)**	**95.5 (2.1, 2.5)**	**95.3 (2.3, 2.5)**
	30	95.9 (1.6, 2.6)	**94.9 (2.5, 2.6)**	**94.9 (2.5, 2.6)**	**94.9 (2.5, 2.6)**
	50	95.9 (1.8, 2.3)	**95.4 (2.3, 2.3)**	**95.4 (2.3, 2.3)**	**95.3 (2.4, 2.3)**
	100	95.7 (2.1, 2.1)	**95.0 (2.3, 2.6)**	**95.0 (2.3, 2.6)**	**95.0 (2.3, 2.6)**

[a] MOVER with the Score confidence limits

[b] MOVER with the Jeffreys confidence limits

Table 4.6: Estimated percentages of coverage, and left and right tail errors (ML,MR) of two-sided 95% confidence intervals constructed for risk ratio RR $= 3$ with control group proportion π_0 and unequal sample sizes of $n_1 = n_0 + 10$ by the four methods based on 10,000 simulation runs.

		Coverage (ML, MR)			
p_0	n_0	Wald	Score	Ms[a]	Mj[b]
0.1	10	96.3 (0.0, 3.7)	95.7 (0.0, 4.3)	95.7 (0.0, 4.3)	97.5 (0.0, 2.5)
	20	96.7 (0.0, 3.3)	96.8 (0.0, 3.2)	95.9 (0.0, 4.1)	**95.0 (1.8, 3.2)**
	30	96.8 (0.0, 3.2)	96.7 (0.1, 3.2)	96.2 (0.1, 3.8)	94.1 (3.0, 2.9)
	50	96.8 (0.1, 3.2)	**95.2 (1.4, 3.5)**	**95.4 (1.2, 3.5)**	94.3 (2.7, 3.0)
	100	95.7 (1.3, 3.0)	**95.0 (2.0, 3.1)**	**95.1 (1.8, 3.1)**	**94.9 (2.4, 2.7)**
0.2	10	95.6 (0.0, 4.5)	96.1 (0.0, 3.9)	96.1 (0.0, 3.94)	94.1 (2.6, 3.3)
	20	96.4 (0.0, 3.6)	95.7 (1.0, 3.3)	95.7 (0.7, 3.59)	94.3 (2.8, 2.9)
	30	96.0 (0.1, 3.9)	**95.2 (1.7, 3.1)**	**94.9 (1.6, 3.5)**	**94.3 (2.7, 2.9)**
	50	**95.4 (1.0, 3.6)**	**95.2 (1.8, 3.0)**	**94.8 (1.6, 3.6)**	**94.9 (2.2, 2.9)**
	100	**95.0 (1.7, 3.3)**	**95.1 (2.1, 2.9)**	**95.2 (2.0, 2.9)**	**95.0 (2.3, 2.7)**
0.3	10	**95.2 (0.0, 4.8)**	97.0 (0.0, 3.0)	97.02 (0.0, 2.98)	94.2 (2.9, 3.0)
	20	95.5 (0.0, 4.6)	**95.4 (1.8, 2.8)**	**95.4 (1.8, 2.8)**	**94.8 (2.5, 2.8)**
	30	95.5 (0.6, 3.9)	**94.9 (2.2, 2.9)**	**95.0 (1.8, 3.3)**	**95.0 (2.2, 2.7)**
	50	**94.8 (1.1, 4.2)**	**94.9 (2.2, 2.9)**	**95.4 (1.7, 3.0)**	**95.1 (2.2, 2.7)**
	100	**95.0 (1.4, 3.6)**	**95.0 (2.1, 2.9)**	**95.1 (2.0, 2.9)**	**95.1 (2.2, 2.7)**

[a] MOVER with the Score confidence limits

[b] MOVER with the Jeffreys confidence limits

Table 4.7: Median width of two-sided 95% confidence intervals constructed for risk ratio RR = 1 and risk ratio RR = 3 with control group proportion π_0 and unequal sample sizes of $n_1 = n_0 + 10$ by the four methods based on 10,000 simulation runs.

		Median width for RR = 1				Median width for RR = 3			
p_0	n_0	Wald	Score	Ms[a]	Mj[b]	Wald	Score	Ms[a]	Mj[b]
0.1	10	9.7	7.1	7.3	10.6	24.2	17.5	19.6	31.8
	20	5.3	4.5	4.6	5.5	12.8	11.9	11.6	15.2
	30	4.0	3.5	3.6	4.0	9.4	9.0	8.8	10.6
	50	2.8	2.6	2.7	2.9	6.6	6.4	6.4	7.1
	100	1.8	1.8	1.8	1.8	4.2	4.2	4.1	4.4
0.2	10	4.3	4.0	3.9	4.9	10.8	10.7	10.5	13.7
	20	2.8	2.7	2.6	2.9	6.7	6.7	6.5	7.6
	30	2.2	2.1	2.1	2.2	5.2	5.2	5.1	5.6
	50	1.7	1.6	1.6	1.7	3.8	3.8	3.8	4.0
	100	1.1	1.1	1.1	1.1	2.6	2.6	2.6	2.7
0.3	10	2.9	2.9	2.8	3.2	6.7	7.0	7.0	8.3
	20	2.0	2.0	1.9	2.0	4.4	4.5	4.5	5.0
	30	1.6	1.6	1.5	1.6	3.6	3.7	3.7	3.9
	50	1.2	1.2	1.2	1.2	2.7	2.7	2.7	2.8
	100	0.9	0.9	0.8	0.9	1.9	1.9	1.9	1.9

[a] MOVER with the Score confidence limits

[b] MOVER with the Jeffreys confidence limits

Table 4.8: Estimated joint coverage (%) and median total interval width for two-sided 95% simultaneous confidence intervals constructed by the four methods for risk ratios comparing 2 experimental groups to a control group. Each entry in the table is based on the 10,000 simulation replicates of binomial variates given true proportions and equal sample sizes.

	Coverage				Median total width			
n	Wald	Score	Ms[a]	Mj[b]	Wald	Score	Ms[a]	Mj[b]
				$(\pi_0, \pi_1, \pi_2) = (0.1, 0.1, 0.3)$				
10	99.1	98.1	98.1	97.0	56.1	32.8	35.9	66.4
20	98.3	97.9	97.9	96.9	25.0	21.3	20.8	29.5
30	98.5	96.7	97.1	95.4	17.2	15.5	15.3	19.3
50	97.9	95.7	96.5	94.8	11.5	10.9	10.8	12.3
100	96.9	95.5	95.4	95.4	7.2	7.0	7.0	7.4
				$(\pi_0, \pi_1, \pi_2) = (0.2, 0.2, 0.6)$				
10	97.6	97.4	96.9	95.2	21.4	18.0	17.3	24.8
20	97.7	96.4	95.8	94.6	12.0	11.6	11.3	13.5
30	97.9	95.9	96.0	94.9	9.0	8.8	8.7	9.7
50	96.5	95.3	95.2	95.0	6.5	6.4	6.4	6.8
100	96.0	95.6	95.6	95.5	4.4	4.4	4.3	4.5
				$(\pi_0, \pi_1, \pi_2) = (0.3, 0.3, 0.9)$				
10	95.1	96.1	96.0	94.2	12.3	12.4	11.9	15.1
20	96.8	95.6	95.6	95.1	7.7	7.8	7.7	8.6
30	96.2	95.7	95.7	95.6	6.0	6.1	6.0	6.5
50	95.6	95.3	95.1	95.2	4.6	4.6	4.6	4.8
100	95.1	95.2	94.9	95.0	3.1	3.1	3.1	3.2

[a] MOVER with the Score confidence limits

[b] MOVER with the Jeffreys confidence limits

Table 4.9: Estimated coverage (%) and median total interval width for two-sided 95% simultaneous confidence intervals constructed by the four methods for risk ratios comparing 3 experimental groups to a control group. Each entry in the table is based on the 10,000 simulation replicates of binomial variates given true proportions and equal sample sizes.

	Coverage				Median total width			
n	Wald	Score	Ms[a]	Mj[b]	Wald	Score	Ms[a]	Mj[b]
				$(\pi_0,\pi_1,\pi_2,\pi_3)=(0.1,0.1,0.2,0.3)$				
10	99.5	98.1	98.1	97.5	96.4	49.5	58.8	132.1
20	98.8	96.4	97.0	96.2	41.3	34.0	33.0	50.2
30	98.6	97.4	97.4	95.5	28.7	24.9	24.5	32.2
50	98.1	96.4	96.8	95.6	18.6	17.4	17.2	20.2
100	96.8	95.2	95.7	95.0	11.6	11.3	11.2	12.1
				$(\pi_0,\pi_1,\pi_2,\pi_3)=(0.2,0.2,0.4,0.6)$				
10	98.6	97.4	96.8	96.7	35.1	31.2	29.4	46.0
20	97.6	95.3	96.2	94.7	19.3	18.6	17.9	22.3
30	97.7	96.0	95.8	95.0	14.5	14.2	13.8	15.9
50	97.0	95.6	95.7	95.4	10.5	10.5	10.3	11.2
100	96.5	95.9	96.0	95.8	7.0	7.0	6.9	7.2
				$(\pi_0,\pi_1,\pi_2,\pi_3)=(0.3,0.3,0.6,0.9)$				
10	95.7	97.6	96.4	95.0	20.4	20.3	19.4	25.8
20	97.3	95.6	95.7	95.6	12.6	12.7	12.5	14.3
30	97.3	96.1	95.8	95.5	9.8	10.0	9.8	10.7
50	96.5	96.2	96.0	95.8	7.3	7.4	7.3	7.6
100	96.0	95.6	95.6	95.6	5.0	5.0	5.0	5.1

[a] MOVER with the Score confidence limits

[b] MOVER with the Jeffreys confidence limits

Table 4.10: Estimated coverage (%) and median total interval width for two-sided 95% simultaneous confidence intervals constructed by the four methods for risk ratios comparing 4 experimental groups to a control group. Each entry in the table is based on the 10,000 simulation replicates of binomial variates given true proportions and equal sample sizes.

	Coverage				Median total width			
n	Wald	Score	Ms[a]	Mj[b]	Wald	Score	Ms[a]	Mj[b]
				$(\pi_0,\pi_1,\pi_2,\pi_3,\pi_4) = (0.10, 0.10, 0.17, 0.23, 0.30)$				
10	99.9	97.9	97.6	97.8	142.3	68.9	81.7	201.5
20	99.1	97.4	97.0	96.2	60.0	47.4	46.3	73.8
30	98.7	96.7	97.4	95.5	40.3	34.7	34.1	46.3
50	98.5	96.1	97.0	95.7	26.2	24.3	24.0	28.6
100	97.2	95.7	96.0	95.3	16.3	15.8	15.7	17.1
				$(\pi_0,\pi_1,\pi_2,\pi_3,\pi_4) = (0.20, 0.20, 0.34, 0.46, 0.60)$				
10	98.4	98.0	96.6	97.3	49.8	43.6	40.9	67.4
20	98.2	96.7	96.2	95.2	27.2	26.0	25.0	31.8
30	97.8	95.6	95.8	95.2	20.5	20.0	19.5	22.8
50	97.0	95.7	95.7	95.9	14.6	14.4	14.2	15.6
100	96.5	96.1	96.0	95.9	9.7	9.7	9.6	10.0
				$(\pi_0,\pi_1,\pi_2,\pi_3,\pi_4) = (0.30, 0.30, 0.51, 0.69, 0.90)$				
10	96.1	97.1	96.9	95.3	28.9	28.8	27.5	38.0
20	97.3	96.3	95.7	95.8	17.5	17.8	17.3	20.1
30	97.1	96.0	96.1	95.8	13.7	13.9	13.6	15.1
50	96.5	95.7	95.7	95.6	10.1	10.2	10.1	10.7
100	96.5	96.1	95.9	95.9	7.0	7.0	7.0	7.2

[a] MOVER with the Score confidence limits

[b] MOVER with the Jeffreys confidence limits

Table 4.11: Estimated coverage (%) and median total interval width for two-sided 95% simultaneous confidence intervals constructed by the four methods for risk ratios comparing 2 experimental groups to a control group. Each entry in the table is based on the 10,000 simulation replicates of binomial variates given true proportions and unequal sample sizes $n_k = n_{k-1} + 10$, $k = 1, \ldots, K$.

	Coverage				Median total width			
n_0	Wald	Score	Ms[a]	Mj[b]	Wald	Score	Ms[a]	Mj[b]
				$(\pi_0, \pi_1, \pi_2) = (0.1, 0.1, 0.3)$				
10	97.3	97.2	97.2	97.3	42.0	29.7	31.2	62.2
20	98.0	96.7	97.5	96.3	21.5	18.9	18.7	26.9
30	98.0	96.8	97.3	95.2	15.7	14.3	14.2	18.1
50	98.1	96.5	96.7	95.5	10.9	10.4	10.3	11.9
100	96.3	95.3	95.5	94.9	7.0	6.9	6.8	7.3
				$(\pi_0, \pi_1, \pi_2) = (0.2, 0.2, 0.6)$				
10	96.4	96.6	96.6	95.9	17.8	16.9	16.5	24.3
20	97.1	96.5	96.8	95.8	11.0	10.8	10.6	12.8
30	96.6	95.2	96.0	95.1	8.5	8.4	8.3	9.4
50	96.1	95.3	95.4	95.1	6.3	6.2	6.2	6.6
100	96.2	95.7	95.7	95.7	4.3	4.3	4.2	4.4
				$(\pi_0, \pi_1, \pi_2) = (0.3, 0.3, 0.9)$				
10	94.8	96.7	96.9	94.3	11.3	11.7	11.5	14.8
20	95.9	96.0	95.8	95.5	7.4	7.6	7.5	8.4
30	95.9	95.4	95.4	95.2	5.8	6.0	5.9	6.4
50	95.6	95.5	95.3	95.4	4.5	4.6	4.5	4.7
100	95.3	95.4	95.3	95.3	3.1	3.1	3.1	3.1

[a] MOVER with the Score confidence limits

[b] MOVER with the Jeffreys confidence limits

Table 4.12: Estimated joint coverage (%) and median total interval width for two-sided 95% simultaneous confidence intervals constructed by the four methods for risk ratios comparing 3 experimental groups to a control group. Each entry in the table is based on the 10,000 simulation replicates of binomial variates given true proportions and unequal sample sizes $n_k = n_{k-1} + 10$, $k = 1, \ldots, K$.

	Coverage				Median total width			
n_0	Wald	Score	Ms[a]	Mj[b]	Wald	Score	Ms[a]	Mj[b]
				$(\pi_0, \pi_1, \pi_2, \pi_3) = (0.1, 0.1, 0.2, 0.3)$				
10	98.0	97.8	97.8	97.4	71.7	45.3	51.2	119.1
20	99.0	98.9	98.9	97.8	35.1	30.3	29.9	46.6
30	98.3	96.5	97.1	94.7	26.2	23.9	23.7	31.7
50	97.5	96.3	96.7	95.6	18.0	17.1	17.0	20.1
100	96.7	95.2	95.3	95.4	10.9	10.7	10.6	11.5
				$(\pi_0, \pi_1, \pi_2, \pi_3) = (0.2, 0.2, 0.4, 0.6)$				
10	95.9	95.9	96.1	95.8	28.9	27.1	26.6	42.1
20	97.4	96.6	96.7	96.2	17.4	17.1	16.8	21.0
30	97.1	95.7	95.8	95.7	13.7	13.6	13.4	15.5
50	96.8	96.2	95.9	96.0	10.3	10.3	10.1	11.0
100	96.8	96.1	96.5	96.3	6.9	6.9	6.9	7.2
				$(\pi_0, \pi_1, \pi_2, \pi_3) = (0.3, 0.3, 0.6, 0.9)$				
10	95.4	97.6	97.6	95.7	18.5	19.1	18.8	25.3
20	97.9	97.1	96.9	96.8	11.8	12.1	11.9	13.7
30	96.8	96.0	95.9	95.9	9.2	9.4	9.3	10.1
50	97.2	97.4	97.1	97.1	7.0	7.1	7.0	7.4
100	95.4	96.0	95.8	95.9	4.9	4.9	4.9	5.0

[a] MOVER with the Score confidence limits

[b] MOVER with the Jeffreys confidence limits

Table 4.13: Estimated coverage (%) and median total interval width for two-sided 95% simultaneous confidence intervals constructed by the four methods for risk ratios comparing 4 experimental groups to a control group. Each entry in the table is based on the 10,000 simulation replicates of binomial variates given true proportions and unequal sample sizes $n_k = n_{k-1} + 10$, $k = 1, \ldots, K$.

	Coverage				Median total width			
n_0	Wald	Score	Ms[a]	Mj[b]	Wald	Score	Ms[a]	Mj[b]
$(\pi_0, \pi_1, \pi_2, \pi_3, \pi_4) = (0.10, 0.10, 0.17, 0.23, 0.30)$								
10	97.0	96.4	96.4	97.3	99.9	58.3	69.1	179.3
20	97.2	96.6	96.5	96.9	49.4	41.9	41.5	68.3
30	97.4	96.9	96.7	96.0	34.9	31.5	31.2	43.3
50	97.8	96.3	96.8	95.8	24.3	23.0	22.8	27.5
100	97.0	95.8	96.1	95.5	15.6	15.2	15.1	16.5
$(\pi_0, \pi_1, \pi_2, \pi_3, \pi_4) = (0.20, 0.20, 0.34, 0.46, 0.60)$								
10	96.3	96.9	96.9	97.4	40.6	37.9	37.2	62.2
20	97.2	96.8	96.7	96.0	24.4	23.9	23.5	30.2
30	97.2	96.1	96.3	95.8	18.5	18.4	18.1	21.2
50	97.1	96.1	96.2	95.9	13.9	13.9	13.7	15.1
100	96.3	95.6	95.7	95.6	9.4	9.4	9.4	9.8
$(\pi_0, \pi_1, \pi_2, \pi_3, \pi_4) = (0.30, 0.30, 0.51, 0.69, 0.90)$								
10	95.6	97.5	97.9	95.5	25.6	26.5	26.0	36.4
20	96.6	96.3	96.1	96.3	16.4	16.9	16.6	19.5
30	96.9	96.8	96.6	96.5	13.0	13.3	13.1	14.5
50	96.8	96.5	96.5	96.5	9.8	10.0	9.9	10.5
100	96.3	96.0	96.0	96.1	6.9	6.9	6.9	7.1

[a] MOVER with the Score confidence limits

[b] MOVER with the Jeffreys confidence limits

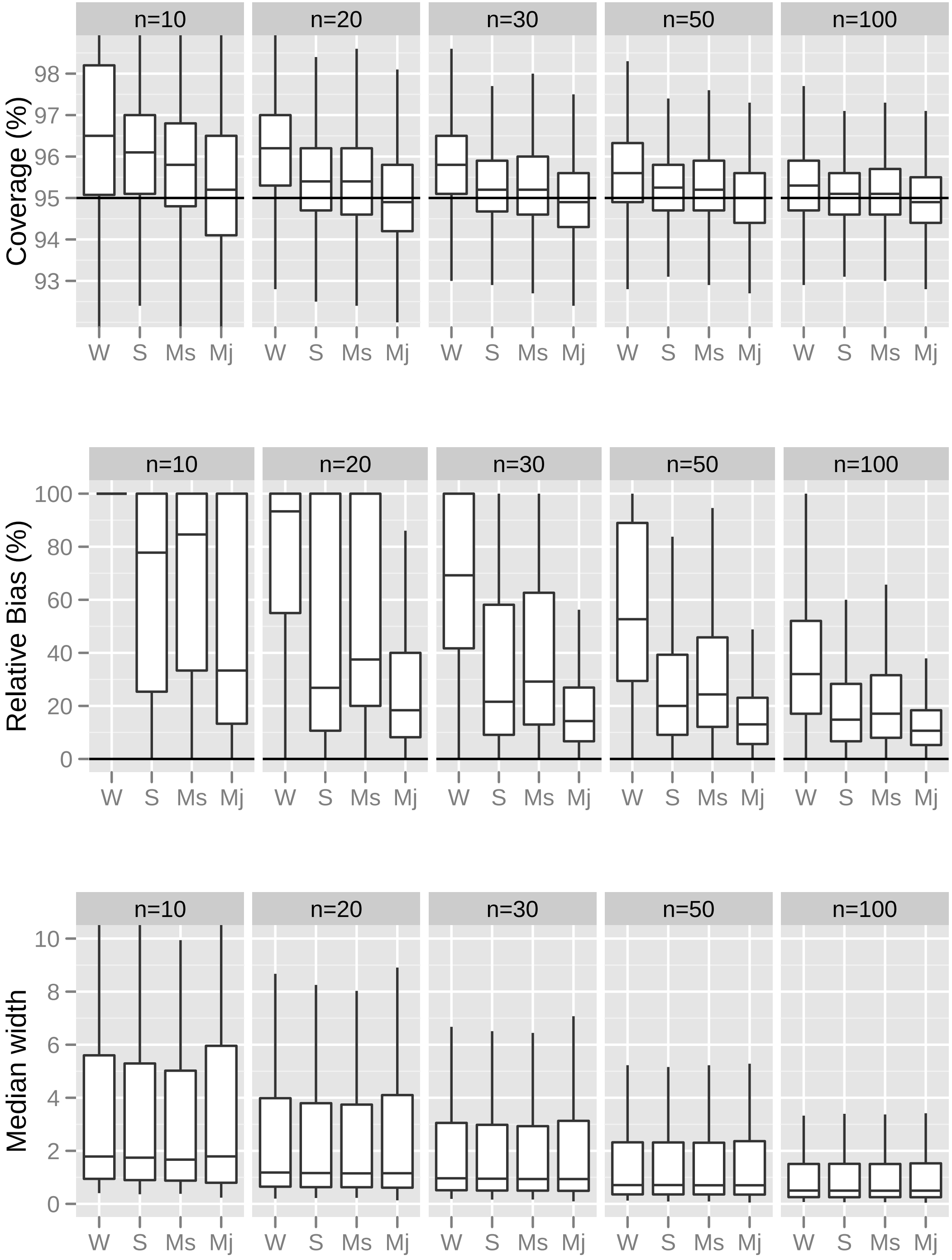

Figure 4.1: Comparative performance of the four methods for constructing a confidence interval for a single risk ratio: percentages (%) of empirical coverage and relative bias in tail errors and median width. Each boxplot displays the performance measures of 1000 population risk ratios for each of which is estimated by 1000 simulation runs.

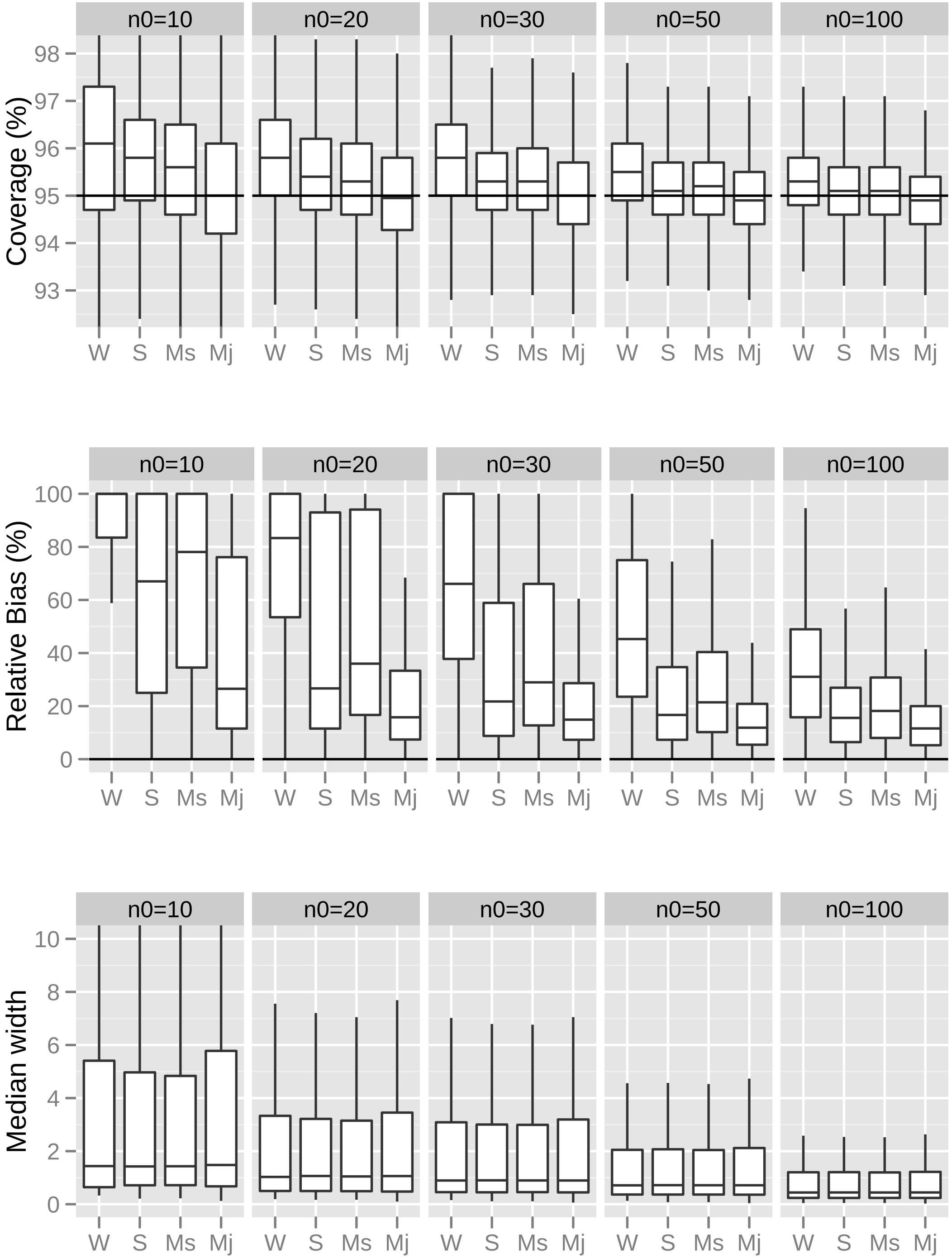

Figure 4.2: Comparative performance of the four methods for constructing a confidence interval for a single risk ratio: percentages (%) of empirical coverage and relative bias in tail errors and median width. Each boxplot displays the performance measures of 1000 population risk ratios for each of which is estimated by 1000 simulation runs. The sample size for the experimental group is increased by 10 (i.e. $n_1 = n_0 + 10$).

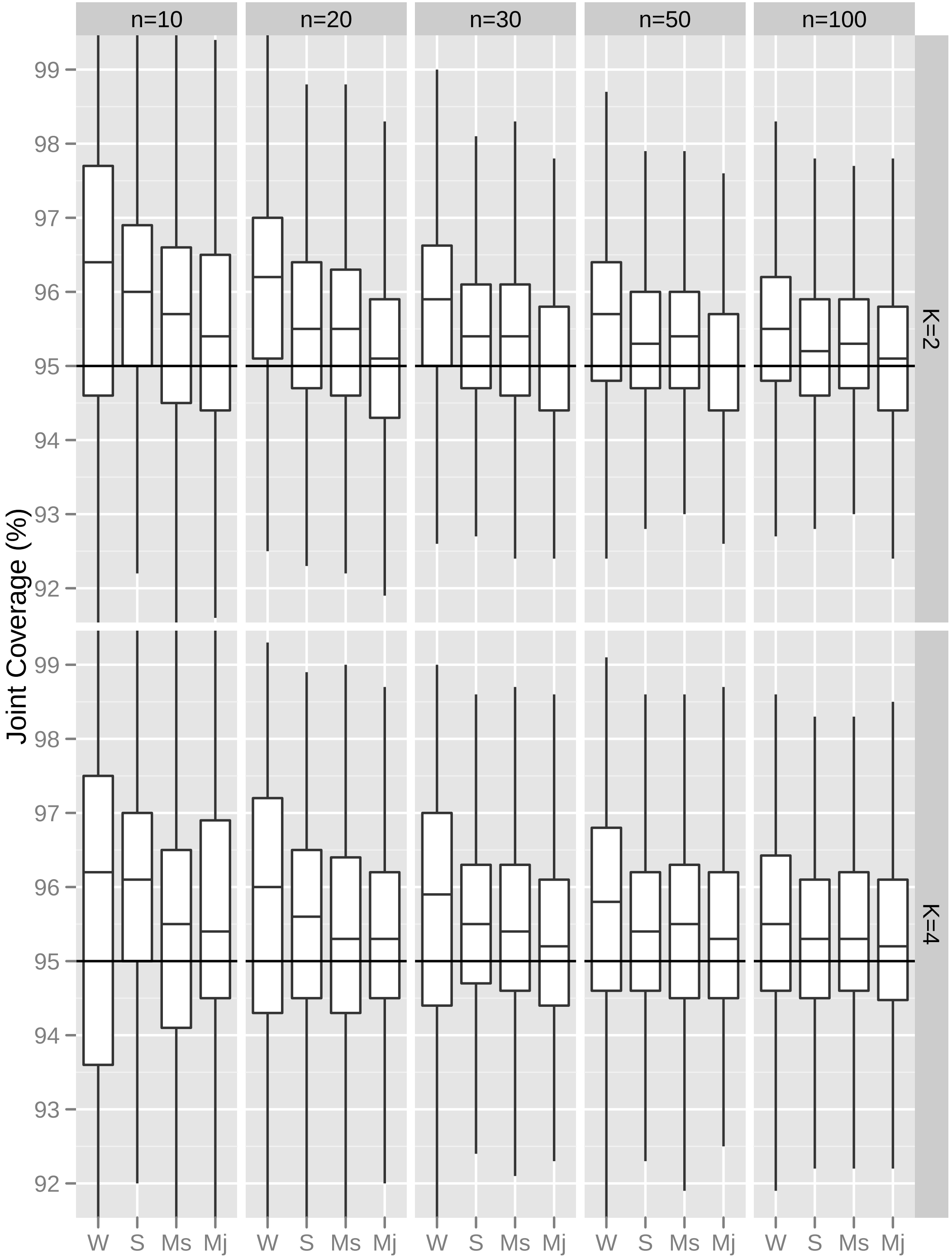

Figure 4.3: Estimated joint coverage (%) of two-sided 95% simultaneous confidence intervals constructed by the four methods for K risk ratios with equal sample sizes $n_k = n$ ($k = 0, \ldots, K$). Each boxplot displays median joint coverage percentages of 1000 combinations of K risk ratios, where each combination is estimated by 1000 simulation runs.

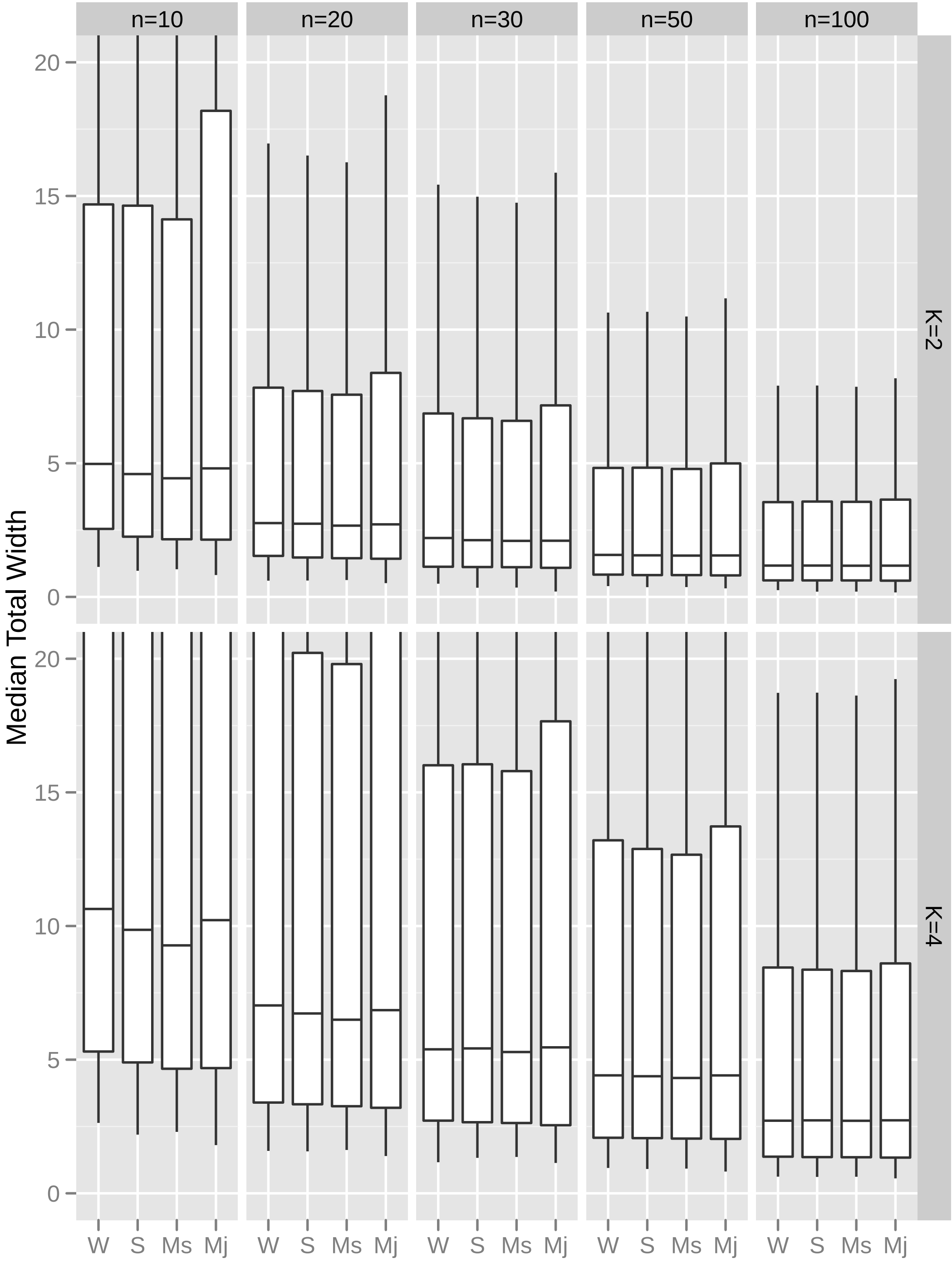

Figure 4.4: Median total widths of two-sided 95% simultaneous confidence intervals constructed by the four methods for risk ratios comparing K experimental groups to a control group of equal sample sizes $n_k = n$, $k = 0, \ldots, K$. Each boxplot displays median total interval widths of 1000 combinations of K risk ratios, where each combination is estimated by 1000 simulation runs.

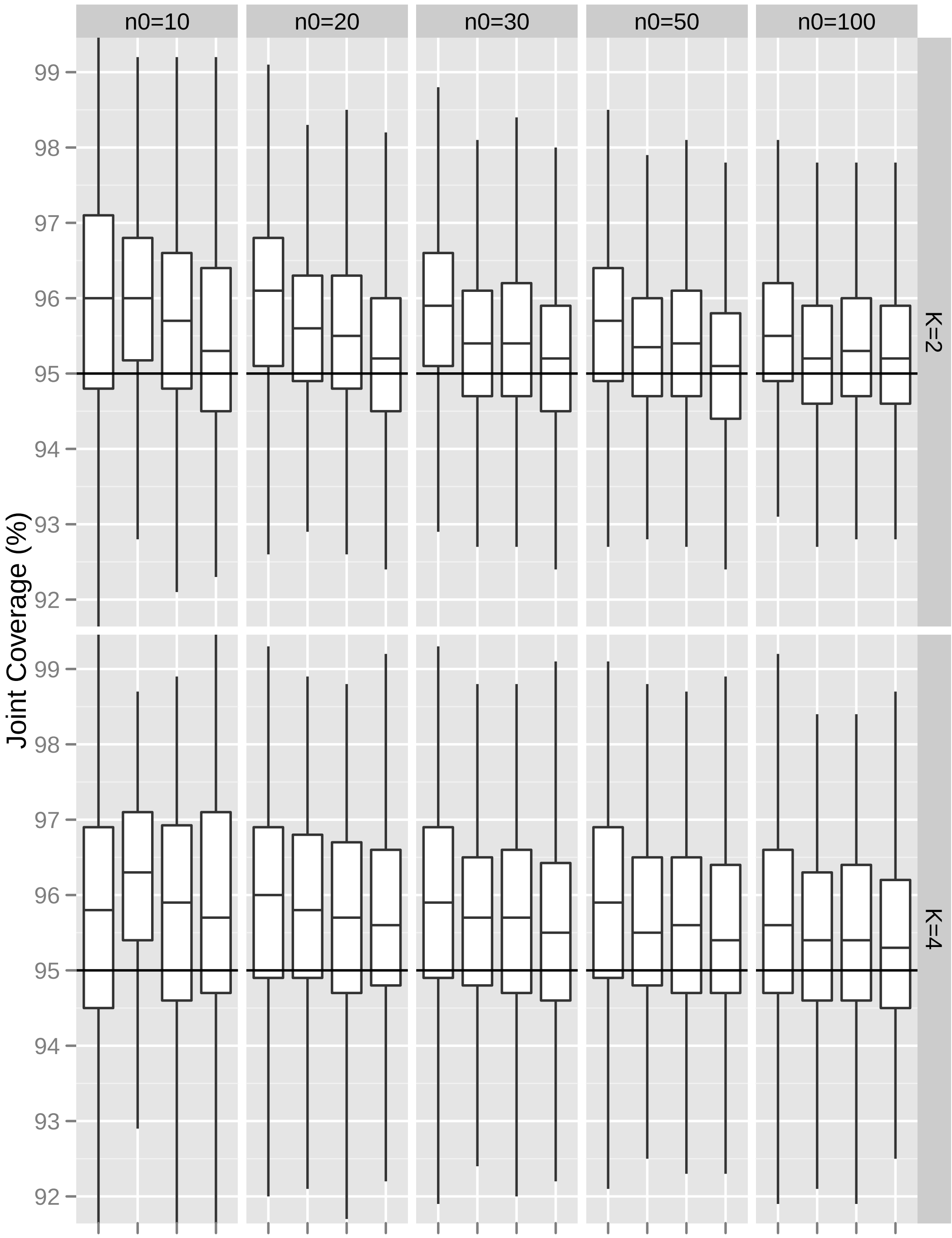

Figure 4.5: Estimated joint coverage (%) of two-sided 95% simultaneous confidence intervals constructed by the four methods for risk ratios comparing K experimental groups to a control group of unequal sample sizes $n_k = n_{k-1} + 10$, $k = 1, \dots, K$. Each boxplot displays median joint coverage percentages of 1000 combinations of K risk ratios, where each combination is estimated by 1000 simulation runs.

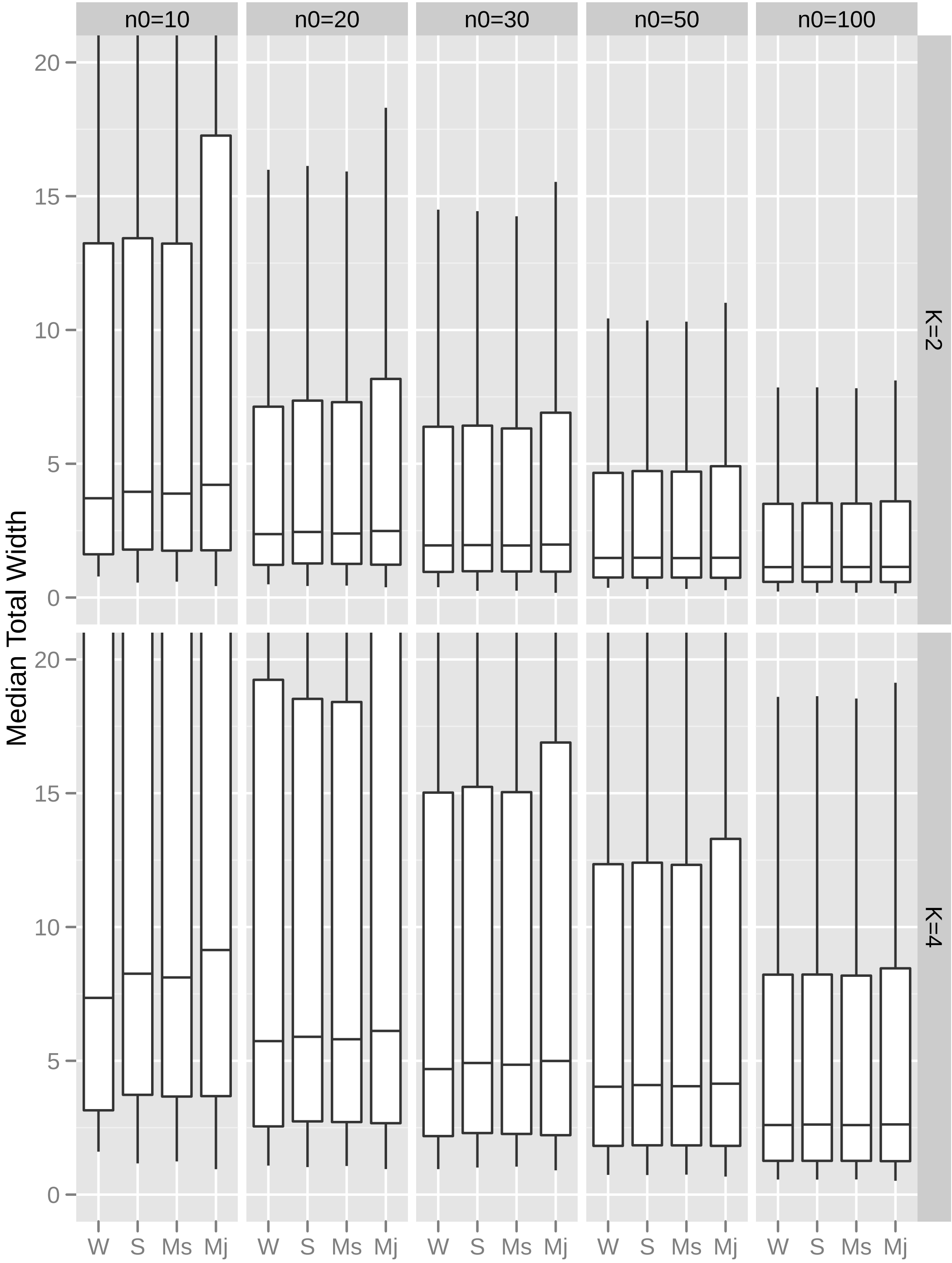

Figure 4.6: Median total widths of two-sided 95% simultaneous confidence intervals constructed by the four methods for risk ratios comparing K experimental groups to a control group of unequal sample sizes $n_k = n_{k-1} + 10$, $k = 1, \ldots, K$. Each boxplot displays median total interval widths of 1000 combinations of K risk ratios, where each combination is estimated by 1000 simulation runs.

Chapter 5

WORKED EXAMPLES

We illustrate how the MOVER approach can be applied to constructing simultaneous confidence intervals for multiple risk ratios in the many-to-one comparisons of proportions using the data from two randomized clinical trials. These data sets have also been chosen to illustrate other simultaneous confidence interval methods proposed for multiple comparisons of proportions (Agresti *et al.*, 2008; Donner and Zou, 2011; Schaarschmidt *et al.*, 2009).

5.1 A factorial trial of coenzyme Q10 and remacemide in Huntingtons disease

Kieburtz and Huntington Study Group (2001) investigated the efficacy of coenzyme Q10, remacemide hydrochloride, and their combination in potentially slowing the functional decline of early Huntington's disease. Three hundred and forty-seven participants with early Huntington's disease were equally randomized to the four treatments and evaluated every 4 to 5 months for a total of 30 months. The primary measure of efficacy was the decline in the total functional capacity score between the baseline and the end of the 30-month study period. Safety of the treatments was measured by the frequency of common adverse events, such as depression, headache, dizziness and nausea, experienced by the patients. Assuming no statistical interaction between coenzyme Q10 and remacemide, the main effects of these two factors were of interest for all outcome and safety measures in Kieburtz and Huntington Study Group (2001). However, to illustrate the MOVER method for ratios, we construct simultaneous confidence intervals for the risk ratios comparing the three experimental groups to the placebo group. Table 5.1 provides the number of patients and those reporting nausea in each treatment group.

Table 5.1: Occurrences of nausea in the treatment of early Huntington's disease

Treatment	Placebo	Coenzyme	Remacemide	Combination
Cases with nausea	9	13	27	22
Sample size	87	87	86	87

To obtain the multiplicity-adjusted confidence limits for proportions required in the MOVER approach, we compute the Dunnett critical value assuming a common correlation of 0.5 as described in Chapter 3. We can use either the R function `qmvnorm` or the SAS function `probmc` without specifying the degrees of freedom and correlation parameter estimates, given as

$$\begin{aligned} z_\alpha &= \texttt{probmc}(\texttt{distribution}, \texttt{q}, \texttt{prob}, \texttt{df}, \texttt{nparam} <, \texttt{parameters} >) \\ &= \texttt{probmc}(\text{“}\texttt{Dunnett2}\text{”}, \texttt{.}, \texttt{0.95}, \texttt{.}, \texttt{3}) \\ &= 2.349. \end{aligned}$$

A confidence interval about an estimated proportion is obtained with the given critical value to maintain the joint coverage probability of the resulting three risk ratios in the many-to-one comparisons of proportions. The score limits are obtained as,

$$\begin{aligned} (l_{s0}, u_{s0}) &= \frac{9+(2.349)^2/2}{87+(2.349)^2} \mp 2.349 \frac{\sqrt{9(87-9)/87+(2.349)^2/4}}{87+(2.349)^2} \\ &= (0.049, 0.205), \end{aligned}$$

for the placebo group, and

$$\begin{aligned} (l_{s1}, u_{s1}) &= \frac{13+(2.349)^2/2}{87+(2.349)^2} \mp 2.349 \frac{\sqrt{13(87-13)/87+(2.349)^2/4}}{87+(2.349)^2} \\ &= (0.081, 0.260), \end{aligned}$$

Table 5.2: Confidence intervals for the proportion of patients experiencing nausea

Treatment	Placebo	Coenzyme	Remacemide	Combination
Sample proportion	**0.10**	**0.15**	**0.31**	**0.25**
Score limits	$(0.05, 0.21)$	$(0.08, 0.26)$	$(0.21, 0.44)$	$(0.16, 0.38)$
Jeffreys limits	$(0.05, 0.20)$	$(0.08, 0.25)$	$(0.21, 0.44)$	$(0.16, 0.37)$

for the coenzyme group. Table 5.2 provides the estimated proportions and their confidence limits for all four groups.

Using these confidence limits for the proportions, we can construct simultaneous confidence intervals for the risk ratios comparing the three experimental groups to the placebo by applying the MOVER equation generalizing Fieller's theorem (3.7) in Chapter 3. For example, the confidence interval for the risk ratio comparing the coenzyme group with the placebo group is constructed with the estimated proportions of nausea and their confidence limits. The values to two decimal places are shown for simplification but more precise numbers were used in the actual calculation. The lower and upper confidence limits of the risk ratio are computed as

$$\begin{aligned} L^{\text{RR}} &= \frac{\widehat{\pi}_1\widehat{\pi}_0 - \sqrt{(\widehat{\pi}_1\widehat{\pi}_0)^2 - l_1u_0(2\widehat{\pi}_1 - l_1)(2\widehat{\pi}_0 - u_0)}}{u_0(2\widehat{\pi}_0 - u_0)} \\ &= \frac{(0.15)(0.10) - \sqrt{((0.15)(0.10))^2 - (0.08)(0.21)(2(0.15) - 0.18)(2(0.10) - 0.21)}}{0.21(2(0.10) - 0.21)} \\ &= 0.58, \end{aligned}$$

$$\begin{aligned} U^{\text{RR}} &= \frac{\widehat{\pi}_1\widehat{\pi}_0 + \sqrt{(\widehat{\pi}_1\widehat{\pi}_0)^2 - u_1l_0(2\widehat{\pi}_1 - u_1)(2\widehat{\pi}_0 - l_0)}}{l_0(2\widehat{\pi}_0 - l_0)} \\ &= \frac{(0.15)(0.10) + \sqrt{((0.15)(0.10))^2 - (0.25)(0.05)(2(0.15) - 0.26)(2(0.10) - 0.05)}}{0.05(2(0.10) - 0.05)} \\ &= 3.65. \end{aligned}$$

Table 5.3: Simultaneous two-sided 95% confidence intervals for risk ratios comparing the occurrences of nausea in patients treated with coenzyme, remacemide and their combinations to placebo

Comparison	Coenzyme	Remacemide	Combination
Sample RR	**1.45**	**3.05**	**2.46**
Wald	(0.56, 3.75)	(1.32, 7.00)	(1.04, 5.77)
Score	(0.58, 3.65)	(1.38, 6.91)	(1.08, 5.70)
Ms[a]	(0.57, 3.63)	(1.36, 6.82)	(1.07, 5.63)
Mj[b]	(0.57, 3.86)	(1.40, 7.42)	(1.08, 6.09)

[a] MOVER with the Score confidence limits

[b] MOVER with the Jeffreys confidence limits

Table 5.3 summarizes the estimated risk ratios and corresponding confidence intervals constructed by the four methods. As the sample sizes are moderately large, the confidence limits obtained by the four methods are similar, particularly more so for the coenzyme group having a sample risk ratio close to unity. For the other groups having a greater risk ratio estimate, the values of confidence limits are slightly different. As suggested in the simulation study, the intervals constructed by the score-based methods are shorter than those by the Wald method or the MOVER with Jeffreys limits. Moreover, the Wald intervals are shifted slightly leftwards relative to the other intervals. We can conclude that patients treated with either remacemide or the combination of remacemide and coenzyme have a greater risk of nausea compared with the patients in the placebo group.

Applying the typical analysis for factorial trials, statistical significance of the effect of each factor, coenzyme or remacemide, was investigated in Kieburtz and Huntington Study Group (2001) using Fisher's exact test. The effect of coenzyme was assessed by pooling across all patients who received coenzyme and comparing them with all patients who did not, regardless of whether or not they received remacemide. The same analysis for the effect of remacemide indicated that its observed adverse effect on nausea was statistically significant ($P < 0.0003$). Although the results in the original study agreed with the results of the simultaneous confidence intervals constructed by the four methods in Table 5.3, the multiplicity issue was not addressed in Kieburtz and Huntington Study Group (2001).

5.2 A randomized, dose-ranging clinical trial of liarozole in psoriasis

Berth-Jones *et al.* (2000) reported a study investigating the lowest effective dose of liarozole in the treatment of psoriasis in a dose-ranging clinical trial. The trial was conducted with a total of 139 subjects equally randomized to receive placebo or liarozole at doses of 50 mg, 75 mg or 150 mg. Excluding two subjects who failed to attend a post-randomization visit, response to treatment was assessed based on an eight-point global scale representing the degree of improvement from 1 (cleared, 100% improvement except for residual discolouration) to 8 (worse). The primary endpoint was the proportion of the subjects in each treatment group having scores 3 or lower at the end of a 12-week treatment period. Table 5.4 summarizes the observed number of patients with marked improvement and sample size of each treatment group.

Table 5.4: Number of patients with marked improvement of psoriasis

Treatment	Placebo	50 mg	75 mg	150 mg
Cases of improvement	2	6	4	13
Sample size	34	33	36	34

As described in the previous example, the confidence limits for single proportions are computed using the same critical value $z_\alpha = 2.349$ and provided in Table 5.5. Table 5.6 summarizes the sample estimates of the risk ratios and their simultaneous 95% confidence intervals constructed by the four methods. Although the sample risk ratios of comparing all

Table 5.5: Proportion estimates and confidence intervals for proportions of improvement

Treatment	Placebo	50 mg	75 mg	150 mg
Sample proportion	**0.06**	**0.18**	**0.11**	**0.38**
Score limits	(0.01, 0.23)	(0.07, 0.38)	(0.04, 0.29)	(0.22, 0.58)
Jeffreys limits	(0.01, 0.21)	(0.07, 0.37)	(0.03, 0.27)	(0.21, 0.58)

Table 5.6: Simultaneous two-sided 95% confidence intervals for the ratio of proportions of improvement in the treatment of psoriasis with different doses of liarozole to placebo

Comparison	50 mg	75 mg	150 mg
Sample RR	**3.00**	**1.83**	**6.33**
Wald	(0.50, 19.27)	(0.27, 13.35)	(1.20, 35.25)
Score	(0.62, 16.23)	(0.34, 10.81)	(1.50, 30.96)
Ms[a]	(0.59, 15.94)	(0.32, 10.85)	(1.46, 30.34)
Mj[b]	(0.60, 24.26)	(0.31, 15.75)	(1.58, 48.28)

[a] MOVER with the Score confidence limits

[b] MOVER with the Jeffreys confidence limits

three dose groups to the placebo group are greater than the null value of unity, the treatment effect only at the highest dose level is statistically significant. Therefore, we can conclude that 150 mg of liarozole is at least 1.6 times more effective than placebo in the treatment of psoriasis based on the simultaneous confidence intervals constructed by the MOVER with Jeffreys limits.

In Berth-Jones *et al.* (2000), multiple testing was performed with a Dunnett critical value obtained by assuming the mean of sample correlation coefficients as the common correlation coefficient (Piegorsch, 1991). It was also found that only the treatment effect at

150 mg was statistically significant ($P < 0.001$) in Berth-Jones *et al.* (2000). As an illustrative example, Schaarschmidt *et al.* (2009) computed 95% simultaneous lower limits for risk differences using a Dunnett critical value incorporating sample correlation coefficients. Both results from the multiple testing in Berth-Jones *et al.* (2000) and the 95% simultaneous confidence intervals for risk differences in Schaarschmidt *et al.* (2009) suggest that the treatment effect of 150 mg of liarozole is statistically significant, which are also consistent with the results in Table 5.6. However, simultaneous confidence intervals for a chosen effect measure provide each group's estimated effect size with its margin of error (Walter, 1995).

Chapter 6

SUMMARY AND DISCUSSION

6.1 Summary

This thesis evaluated four approximate methods for constructing simultaneous confidence intervals for multiple risk ratios in the many-to-one comparisons of binomial proportions. For a single risk ratio, the score interval is superior to inverting the Wald interval in terms of coverage probability, interval width and tail error balance (Gart and Nam, 1988). Extending these methods to multiple risk ratios, Klingenberg (2010) recommends using a Dunnett critical value incorporating their sample correlation coefficients with the score test statistics. Despite its outstanding performance in terms of coverage and interval width, it has two practical drawbacks. First, inverting score tests requires an iterative algorithm or a series of substitutions, which may hinder widespread applications in practice. Second, using plug-in estimates of correlation coefficients to compute a Dunnett critical value may introduce additional variability for a limited improvement on the coverage properties. Especially, the effect of using estimates rather than true values may not be negligible if sample sizes are not large or correlations are greater than 0.5 (Holford *et al.*, 1989).

The MOVER approach generalizing Fieller's theorem for a ratio of parameters (Li *et al.*, 2010) may be an alternative, non-iterative procedure to construct confidence intervals for risk ratios. Given accurate confidence limits for proportions and their point estimates, the MOVER approach allows the variances of sample risk ratios to be estimated separately at points closer to the lower and upper limits of risk ratios. Acknowledging the potential skewness of the sampling distributions of proportion estimates, the MOVER approach obtains variance estimates similar to the score method in principle but without an iterative algorithm (Zou, 2008; Donner and Zou, 2011). As the variance estimates are recovered from the confidence limits for single proportions, the confidence limits for component proportions should be computed with an appropriate adjustment for multiplicity. A reasonable Dunnett critical value may be approximately obtained assuming a common correlation coefficient without substantially affecting the coverage probabilities of simultaneous confidence intervals.

The simulation results suggest that the MOVER approach with score or Jeffreys confidence limits have desirable operating characteristics, comparable to those of the score method with a Dunnett critical value computed using a constant correlation coefficient of 0.5 assumed under the complete homogeneity of proportions. Especially with small samples sizes, the MOVER approach with Jeffreys confidence limits for single proportions have superior coverage properties than the score-based methods, either inverting score test statistics or applying the MOVER approach with score limits. The Wald method tends to be quite conservative for small to moderate sample sizes. All four competing confidence intervals tend to have comparable median total confidence interval widths. Nevertheless, the MOVER with Jeffreys limits may yield intervals somewhat wider than the other methods, especially with small sample sizes.

6.2 Limitations

Confidence interval methods relying on the central limit theorem are known to be robust against mild deviations from normality, yet substantial skewness of the underlying distribution may be problematic, yielding less accurate coverage probabilities even for large samples sizes (Pocock, 1982; Andersson, 2009). Acknowledging disparate tail errors for large values of risk ratios, Gart and Nam (1988) proposed an improved score interval corrected for skewness (Bartlett, 1953, 1955). Therefore, a more comprehensive evaluation including additional methods and simulation scenarios (e.g. a larger number of comparisons K) may be beneficial.

Although the effect of unequal sample sizes on the performance of the methods was considered, the evaluation did not include unequal sample sizes resulting from unequal randomization. For example, unequal allocation adopting a popular recommendation $n_0 \approx n_i\sqrt{K}$ for normal means (Dunnett, 1955) may be demanded for financial reasons. The evaluation of the methods under unequal allocation may be beneficial. If the total sample size is kept equal as with the balanced design, the performance measures will be more directly comparable.

The Dunnett critical value for binomial data is commonly computed using plug-in sample estimates, relying on the assumption of large samples. Nevertheless, we speculate that using plug-in sample proportion estimates may result in unstable critical values due to the variability of the estimated correlation coefficients, and consequently less accurate coverage probabilities. However, it would be interesting to directly compare the effect of variability of estimated correlation coefficients and a common, constant correlation coefficient on coverage probabilities as the number of comparisons K, range of proportions π_k, and sample sizes n_k vary. Furthermore, a sensitivity analysis using correlation coefficient values other than 0.5 would be also useful.

6.3 Future Research

The MOVER approach for risk ratios for many-to-one comparisons of proportions can be readily extended to all pairwise comparison by using a critical value from the studentized range distribution. Moreover, we can consider the proposed approach in other contexts of comparing several proportions. For instance, it would be useful to develop simultaneous confidence intervals corresponding to the Williams trend test (Williams, 1971, 1972) as ratios, extending the method by Hothorn and Djira (2011) to binomial proportions. Commonly occurring in toxicological studies to determine the minimal effective dose, an analogous evaluation based on multiple differences of contrasts was also considered in Donner and Zou (2011). The simultaneous confidence intervals for multiple ratios of contrasts are constructed by applying the MOVER for ratios with the confidence limits for corresponding contrasts, which are also obtained by applying the MOVER with an appropriate multiplicity-adjusted confidence limits for single proportions.

The proposed method may be readily extended to the analysis of clustered binary data commonly arising from teratological experiments involving multiple treatment groups. For the simple comparison between two treatment groups, the MOVER approach is applied with the confidence limits about each proportion estimate incorporating the variance inflation factor. For the case of multiple comparisons to control, the proposed method of computing a Dunnett critical value assuming a common correlation coefficient of 0.5 would be more favourable than using sample correlation coefficients because of additional variability to the estimated sample proportions in correlated binary data. Preliminary simulation results indicate a satisfactory performance of the proposed MOVER approach with an exchangeable correlation matrix for the case of correlated binary data.

BIBLIOGRAPHY

Agresti, A., Bini, M., Bertaccini, B. and Ryu, E. (2008). Simultaneous confidence intervals for comparing binomial parameters. *Biometrics* **64**, 1270–1275.

Agresti, A. and Coull, A. (1998). Approximate is better than ‘exact’ for interval estimation of binomial proportions. *American Statistician* **52**, 119–126.

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

Bailey, B. (1987). Confidence limits to the risk ratio. *Biometrics* **43**, 201–205.

Bartlett, M. (1953). Approximate confidence intervals: II. more than one unknown parameter. *Biometrika* **40**, 306–317.

Bartlett, M. (1955). Approximate confidence intervals: III. a bias correction. *Biometrika* **42**, 201–204.

Berger, R. and Boos, D. (1994). P values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association* **89**, 1012–1016.

Berth-Jones, J., Todd, G., Hutchinson, P., Thestrup-Pedersen, K. and Vanhoutte, F. (2000). Treatment of psoriasis with oral liarozole: a dose-ranging study. *British Journal of Dermatology* **143**, 1170–1176.

Brown, L., Cai, T. and DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science* **16**, 101–117.

Brown, L., Cai, T. and DasGupta, A. (2002). Confidence intervals for a binomial proportion and asymptotic expansions. *The Annals of Statistics* **30**, 160–201.

Brown, L. and Li, X. (2005). Confidence intervals for two sample binomial distributions. *Journal of Statistical Planning and Inference* **130**, 359–375.

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

Casella, G. and Berger, R. (1990). *Statistical Inference*. Wadsworth &Brooks, California.

Clopper, C. and Pearson, E. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* **26**, 404–413.

Cook, R. J. and Farewell, V. T. (1996). Multiplicity considerations in the design and analysis of clinical trials. *Journal of the Royal Statistical Society* **159**, 93–110.

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

Dann, R. and Koch, G. (2005). Reveiw and evaluation of methods for computing confidence intervals for the ratio of two proportions and considerations for non-inferiority clinical trials. *Journal of Biopharmaceutical Statistics* **15**, 85–107.

Deeks, J. (1998). When can odds ratios mislead? *British Journal of Medicine* **317**, 1155–1156.

Donner, A. and Zou, G. Y. (2011). Estimating simultaneous confidence intervals for multiple contrasts of proportions by the method of variance estimates recovery. *Statistics in Biopharmaceutical Research* **3**, 320–335.

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association* **50**, 1096 –1121.

Edwards, D. and Berry, J. (1987). The efficiency of simulation-based multiple comparisons. *Biometrics* **43**, 913–928.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association* **82**, 171–185.

Efron, B. (2003). Second thoughts on the bootstrap. *Statistical Science* **18**, 135–140.

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

Freidlin, B., Korn, E., Gray, R. and Martin, A. (2008). Multi-arm clinical trials of new agents: some design considerations. *Statistics in Clinical Cancer Research* **14**, 4368–4371.

Gardner, M. and Altman, D. (1986). Confidence intervals rather than p values: estimation rather than hypothesis testing. *British Medical Journal* **292**, 746–750.

Gart, J. (1985). Approximate tests and interval estimation of the common relative risk in the combination of 2 x 2 tables. *Biometrika* **72**, 673–677.

Gart, J. and Nam, J. (1988). Approximate interval estimation of the ratio of binomial parameters: A review and correction for skewness. *Biometrics* **44**, 323–338.

Genz, A. (1992). Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics* **1**, 141–149.

Graham, P., Mengersen, K. and Morton, A. (2003). Confidence limits for the ratio of two rates based on likelihood scores: non-iterative method. *Statistics in Medicine* **22**, 2071–2083.

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

Hochberg, Y. and Tamhane, A. (1987). *Multiple Comparison Procedures*. John Wiley & Sons, New York.

Holford, T., Walter, S. and Dunnett, C. (1989). Simultaneous interval estimation of the odds ratio in studies with two or more comparisons. *Journal of Clinical Epidemiology* **42**, 427–434.

Hothorn, L. A. (2007). How to deal with multiple treatment or dose groups in randomized clinical trials? *Fundamental & Clinical Pharmacology* **21**, 137–154.

Hothorn, L. A. and Djira, G. D. (2011). A ratio-to-control Williams-type test for trend. *Pharmaceutical Statistics* **10**, 289–292.

Howe, W. (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.

Julious, S. and McIntyre, N. (2012). Sample sizes for trials involving multiple correlated must-win comparisons. *Pharmaceutical Statistics* **11**, 177–185.

Katz, D., Baptista, J., Azen, S. and Pike, M. (1978). Obtaining confidence intervals for the risk ratio in cohort studies. *Biometrics* **34**, 469–474.

Kieburtz, K. and Huntington Study Group (2001). A randomized, placebo-controlled trial of coenzyme Q10 and remacemide in Huntingtons disease. *Neurology* **57**, 397–404.

Klingenberg, B. (2010). Simultaneous confidence bounds for relative risks in multiple comparisons to control. *Statistics in Medicine* **29**, 3232–3244.

Klingenberg, B. (2012). Simultaneous score confidence bounds for risk differences in multiple comparisons to a control. *Computational Statistics and Data Analysis* **56**, 1079–1089.

Koch, G. and Gansky, S. (1996). Statistical considerations for multiplicity in confirmatory protocols. *Drug Information Journal* **30**, 523–533.

Koopman, P. (1984). Confidence intervals for the ratio of two binomial proportions. *Biometrics* **40**, 513–517.

Lancaster, H. (1949). The combination of probabilities arising from data in discrete distributions. *Biometrika* **36**, 370–382.

Li, Y., Koval, J. J., Donner, A. and Zou, G. Y. (2010). Interval estimation for the area under the receiver operating characteristic curve when data are subject to error. *Statistics in Medicine* **29**, 2521–2531.

Ludbrook, J. (1998). Multiple comparison procedures updated. *Clinical and Experimental Pharmacology and Physiology* **25**, 1034.

McCann, M. and Tebbs, J. (2009). Simultaneous logit-based confidence intervals for odds ratios in $2 \times k$ classification tables with a fixed reference level. *Communications in Statistics - Simulation and Computation* **38**, 961–975.

Mi, X., Miwa, T. and Hothorn, T. (2009). `mvtnorm`: new numerical algorithm for multivariate normal probabilities. *The R Journal* **1**, 37–39.

Miettinen, O. and Nurminen, M. (1985). Comparative analysis of two rates. *Statistics in Medicine* **4**, 213–226.

Nam, J. (1995). Confidence limits for the ratio of two binomial proportions based on likelihood scores: non-iterative method. *Biometrical Journal* **3**, 375–379.

Newcombe, R. (1998*a*). Two-sided confidence intervals for the single proportion; comparison of seven methods. *Statistics in Medicine* **17**, 857–872.

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

Newcombe, R. (2001). Estimating the difference between differences: measurement of additive scale interaction for proportions. *Statistics in Medicine* **20**, 2885–2893.

Newcombe, R. (2011). Propagating imprecision: combining confidence intervals from independent sources. *Communications in Statistics - Theory and Methods* **40**, 3154–3180.

Noether, G. (1957). Two confidence intervals for the ratio of two probabilities and some measures of effectiveness. *Journal of the American Statistical Association* **52**, 36–45.

Piegorsch, W. (1991). Multiple comparisons for analyzing dichotomous response. *Biometrics* **47**, 45–52.

Pocock, S. J. (1982). When not to rely on the central limit theorem - an example from absenteeism data. *Communications in Statistics - Theory and Methods* **11**, 2169–2179.

Price, R. and Bonett, D. (2008). Confidence intervals for a ratio of two independent binomial proportions. *Statistics in Medicine* **27**, 5497–5508.

Proschan, M. A. and Waclawiw, M. A. (2000). Practical guidelines for multiplicity adjustment in clinical trials. *Controlled Clinical Trials* **21**, 527–539.

Schaarschmidt, F., Biesheuvel, E. and Hothorn, L. A. (2009). Approximate simultaneous confidence intervals for many-to-one comparisons of binary proportions in randomized clinical trials. *Journal of Biopharmaceutical Statistics* **19**, 292–310.

Schaarschmidt, F., Sill, M. and Hothorn, L. A. (2008). Approximate simultaneous confidence intervals for multiple contrasts of binomial proportions. *Biometrical Journal* **50**, 782–792.

Schechtman, E. (2002). Odds ratio, relative risk, absolute risk reduction, and the number needed to treat - which of these should we use? *Value in Health* **5**, 431–436.

Schiller, J., Harrington, D., Belani, C., Langer, C., Sandler, A. and Krook, J. (2002). Comparison of four chemotherapy regimens for advanced non-small-cell lung cancer. *New England Journal of Medicine* **346**, 92–98.

Schulz, K., Altman, D., Moher, D. and the CONSORT Group (2010). Consort 2010 statement: updated guidelines for reporting parallel group randomized trials. *BMC Medicine* **8**, 18.

Sidak, Z. (1968). On multivariate normal probabilities of rectangles: their dependence on the correlations. *Annals of Mathematical Statistics* **39**, 1425–1434.

von Elm, E., Altman, D., Egger, M., Pocock, S., Gøtzsche, P. and Vandenbroucke, J. (2008). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Journal of Clinical Epidemiology* **61**, 344–349.

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

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

Wilks, S. (1938). Shortest average confidence intervals from large samples. *Annals of Mathematical Statistics* **9**, 166–175.

Williams, D. (1971). A test for differences between treatment means when several dose levels are compared with a zero dose control. *Biometrics* **27**, 103–117.

Williams, D. (1972). The comparison of several dose levels with a zero dose control. *Biometrics* **28**, 519–531.

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

Zou, G., Huang, W. and Zhang, X. (2009). A note on confidence interval estimation for a linear function of binomial proportions. *Computational Statistics and Data Analysis* **53**, 1080–1085.

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

Zou, G. Y. (2009). Assessment of risks by predicting counterfactuals. *Statistics in Medicine* **28**, 3761–3781.

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

VITA

NAME	Jungwon Shin
EDUCATION	M.Sc. in Biostatistics University of Western Ontario London, Ontario 2010–2012
	B.Sc. in Mathematics and Statistics University of British Columbia Vancouver, British Columbia 1998–2003
WORK EXPERIENCE	Research Assistant University of Western Ontario London, Ontario 2010–2012
AWARDS	University of Western Ontario Schulich Graduate Scholarship 2010–2012
	British Columbia Provicial Government Scholarship University of British Columbia Outstanding Student Initiative Scholarship Norman MacKenzie Alumni Scholarship 1998–1999