Estimating confidence intervals of proportions and (pooled) risk differences to evaluate human drug safety

Staaden, Jasmijn Irmgard Nine van

Faculteit der Sociale Wetenschappen

Estimating confidence intervals of proportions and (pooled) risk differences to evaluate human drug safety

Jasmijn Irmgard Nine van Staaden

Master's Thesis Psychology
Methodology and Statistics Unit, Institute of Psychology,
Faculty of Social and Behavioral Sciences, Leiden University
Date: 1 July 2021
Student number: s1489437
Supervisors: S.M.H. Huisman (internal) and V. van Paassen (external)

Abstract

This thesis aims to be a contribution to the field of meta-analyses of randomized controlled trials to evaluate the safety of human drugs, which can be assessed by comparing binomial proportions of adverse events between a control and an experimental condition. Confidence intervals of this proportion difference (i.e., "risk difference") allow for statistical inference and give insight into the chance of discovering this difference in the population. The coverage probability of a confidence interval method is the proportion of cases the intervals constructed through this method contain the true parameter of interest. Some of the traditional confidence interval methods for proportions and risk differences are known to have low coverage probabilities, when sample sizes are small and/or point estimates are extreme. Furthermore, confidence intervals derived from the meta-analytical fixed-effect model are known to perform poorly when the proportion or risk difference in at least one study within the meta-analysis is equal to zero. These rare events data are not only highly prevalent in the evaluation of human drug safety since most drug-induced adverse events are rare, but also are the very reason for the demand for meta-analyses in this research field. Firstly, this thesis shows that, for proportions, the Adjusted Wald confidence interval method yields the most accurate coverage probability, whereas for risk differences, the Agresti-Caffo method is most accurate. From the findings of this thesis can further be concluded that the combination of the Tian confidence interval method and the Inverse of Variance weighting strategy is recommended over the fixed-effect and random-effects meta-analytical models. This combination especially handles rare events data well. However, when applying the weighting strategy of Cochran-Mantel-Haenszel, the fixed-effect model is the preferrable choice for obtaining a pooled estimate of the risk difference, along with its confidence interval, within a meta-analysis.

Keywords: proportion, risk difference, confidence interval, coverage probability, meta-analysis

Contents

Section 1. Introduction

Any variable that takes only two outcome values is called binary. The most common example of a binary variable is the coin flip, which can yield only two outcomes. The term "binomial proportion" reflects the number of events – or "successes" – divided by the number of trials in an experiment with a binary outcome variable. In the case of the coin flip, the binominal proportion is the number of observed heads or tails divided by the number of flips.

A study design in which subjects are randomly assigned to either an experimental or a control condition is referred to as a "randomized controlled trial". Within such trials, it is common practice to draw a comparison between the binomial proportions of both conditions. The difference between these two proportions (i.e., "risk difference") acts as an important effect measure in regards to the investigated treatment. When investigating, for instance, the safety of a certain drug through a randomized controlled trial, the risk difference for adverse events between the drug and control condition can be very telling. This thesis is embedded in the contemporary academic literature on binominal proportions and risk differences. It will consider these theoretical concepts in the practical context of human drug safety assessment.

1.1 Structure

This thesis aims to be a contribution to the field of meta-analyses of randomized controlled trials to evaluate the safety of human drugs. It is organized as follows: first, I will introduce the topics of prediction models and confidence intervals for binary data. In the remainder of the Introduction, I will discuss the basic principles of stratified randomization and meta-analyses. At the end of the Introduction, I will present the research hypotheses central to this thesis and formulate their corresponding research questions. In the Methods, I will present the methodology, applicable using R, through which I will investigate the respective hypotheses and answer the corresponding research questions. My final R code can be found in Appendix A. Next, in the Results, I will report the findings of this thesis. Finally, in the Discussion, I will explain how these findings relate to what is already known in the literature as well as to the research questions and hypotheses.

1.2 Prediction models

Binary outcome variables can be predicted through linear regression. The difference between the predicted and true (i.e., actual population) value of the binary outcome variable is called the error. Since the errors of binary outcome variables are not normally distributed, a generalized linear model is required. This model is a generalization of the linear regression model, as it allows for non-normal distributions, by connecting them to the linear model using a link function. Hence, assumptions are not made on the binary data, but instead on the error term within the link function, which is a Gaussian error, just as the error of the linear model is.

Equation 1 shows the linear regression equation, which can be connected to a non-normal outcome variable by link function *f*, as shown in Equation 2. In Equations 1 and 2, $\hat{y}$ is the predicted binary outcome variable, α is the intercept, β is the regression coefficient, and ε is the error term of the linear regression model.

$$\hat{y} = \alpha + \beta x + \varepsilon \quad (1)$$

$$\hat{y} = f(\alpha + \beta x + \varepsilon) \quad (2)$$

Examples of link functions are logit and probit functions. Both link functions connect the non-normal distribution of binary outcome variables to the linear regression function. The values of the outcome variable that needs to be predicted are binary (i.e., either 0 or 1) and have a Bernoulli distribution. On the contrary, binomial proportions can be any number between 0 and 1. Logit and probit models both take any predicted value and rescale it to fall between 0 and 1. Yet they define the link function *f* differently: logit models use a logistic link function, whereas probit models use a cumulative standard normal link function.

1.3 Confidence intervals

In addition to the predicted value (i.e., point estimate) of a proportion or risk difference, its confidence interval is usually reported as it allows for statistical inferences. The confidence interval contains a range of values restricted by a lower and an upper boundary. A confidence interval of a given variable within a random sample with a nominal coverage probability of 0.95 offers the range within which the true value of the parameter of interest can be expected with 95% certainty. In other words, when taking multiple random samples from a single population, 95% of the consequential confidence intervals are expected to contain the population value.

The empirical coverage probability of a confidence interval method is the proportion of cases that the intervals, obtained through this method, contain the true value of interest. Comparing this coverage probability to the nominal coverage probability – which is often set at 0.95 – gives insight into where the method is located on the statistical trade-off between what is called "conservatism" and "liberalism". When the actual coverage probability is less than the nominal coverage probability, the confidence interval method is termed "liberal". A "conservative" confidence interval method has a coverage probability of below the nominal value (Fagerland et al., 2015). A discrepancy between the empirical coverage probability and nominal coverage probability frequently occurs. This is especially so for binomial confidence interval methods (Agresti & Coull, 1998; Brown et al., 2001; Newcombe, 1998).

As a conservative confidence interval method maximizes the probability of containing the true value of interest it would seem more suitable for estimating confidence intervals for proportions or risk differences of drug-induced adverse events in the assessment of drug safety. However, according to the US Food and Drug Administration (2008), statistical approaches for meta-analyses of randomized controlled trials should ensure "confidence intervals have accurate coverage properties" (p. 14). Therefore, the closer the empirical coverage probability of a confidence interval method is to the nominal coverage probability, the more appropriate the method is for investigating proportions or risk differences of drug-induced adverse events, regardless of whether it is more liberal or conservative.

Confidence interval estimation of binomial proportions is one of the most commonly applied analyses in statistical inference. At the same time, it represents one of the most basic problems of statistical practice (Brown et al., 2001, 2002; Guan, 2012; Pires & Amado, 2008). Agresti and Coull (1998) note that the two most common methods for confidence interval estimation of binomial proportions are both plagued by inherent disadvantages. On the one hand, the "Wald" confidence interval method (see Equations 3 – 4 and 9 – 11 in Table 1 on page 7) is prone to having a low coverage probability when its point estimate is extreme (i.e., close to 0 or 1) or when the investigated sample size is small (Fagerland et al., 2015; Kim & Won, 2013). On the other hand, the "Wilson Score" confidence interval method yields a coverage probability close to the nominal value when point estimates are extreme (Wilson, 1927; see Equation 5 in Table 1). Therefore, this method is preferable to the Wald method, although it similarly yields a low coverage probability when sample sizes are small (Agresti & Coull, 1998; Brown et al., 2001).

Due to its low coverage probability, the Wald method might thus not be appropriate for investigating the proportion of drug-induced adverse events. Although the Wilson Score method is recommended, it only yields an acceptable coverage probability when sample sizes are large. This led Agresti and Coull (1998) to a third option: the "Adjusted Wald" confidence interval method has a higher coverage probability than the Wilson Score method when sample sizes are small and is as simple as using the Wald method with two extra successes and failures (e.g., number of observed heads and tails; see Equations 6 – 8 in Table 1). The Adjusted Wald method can be more conservative than the Wilson Score method, but rarely yields a coverage probability of below the nominal value.

Agresti went on to improve the performance of an important confidence interval for risk differences: the "Newcombe Hybrid Score" confidence interval method (see Equations 12 – 14 in Table 1). This interval combines Wilson Score methods for the two proportions to be compared (Newcombe, 1998). It outperforms the Wald method for risk differences, but is disfavored on the same grounds as the Wilson Score method: for the Newcombe Hybrid Score method, an acceptable coverage probability is established when samples consist of 40 or more subjects (Fagerland et al., 2015). Agresti and Caffo (2000) lower this threshold in the same manner as Agresti and Coull (1998) did with the Adjusted Wald method by adding one extra success and one extra failure per sample (see Equations 15 – 19 in Table 1). Comparatively to the Adjusted Wald method, the "Agresti-Caffo" confidence interval method can be more conservative than the Newcombe Hybrid Score method but avoids a coverage probability of below the nominal value.

Table 1.

An overview of the equations of the six confidence interval methods for proportions and risk differences

Proportions		
Wald	*Wilson Score*	*Adjusted Wald*
$\hat{p} \pm z_{\alpha/2}\,\hat{s}$ (3)	$\dfrac{\hat{p} + \frac{{z_{\alpha/2}}^2}{2n} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{{z_{\alpha/2}}^2}{4n^2}}}{1 + \frac{z^2}{n}}$ (5)	$\hat{p}_w \pm z_{\alpha/2}\,\hat{s}_w$ (6)
$\hat{s}^2 = \dfrac{\hat{p}(1-\hat{p})}{n}$ (4)		$\hat{p}_w = \dfrac{x+2}{n+4}$ (7)
		$\hat{s}^2_w = \dfrac{\hat{p}_w(1-\hat{p}_w)}{n+4}$ (8)
Risk differences		
Wald	*Newcombe Hybrid Score*	*Agresti-Caffo*
$\hat{p}_2 - \hat{p}_1 \pm z_{\alpha/2}\sqrt{(\hat{s}_1^2 + \hat{s}_2^2)}$ (9)	Lower bound: $\hat{p}_2 - \hat{p}_1 - \sqrt{(\hat{p}_2 - l_2)^2 + (u_1 - \hat{p}_1)^2}$ (12)	$\hat{p}_{w2} - \hat{p}_{w1} \pm z_{\alpha/2}\sqrt{(\hat{s}^2_{w1} + \hat{s}^2_{w2})}$ (15)
	Upper bound: $\hat{p}_2 - \hat{p}_1 + \sqrt{(\hat{p}_1 - l_1)^2 + (u_2 - \hat{p}_2)^2}$ (13)	
$\hat{s}_1^2 = \dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1}$ (10)	in which each (l_i, u_i) is:	$\hat{p}_{w1} = \dfrac{x_1+1}{n_1+2}$ (16)
$\hat{s}_2^2 = \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}$ (11)	$\hat{p}_i\left(\frac{n_i}{n_i + {z_{\alpha/2}}^2}\right) + \frac{1}{2}\left(\frac{{z_{\alpha/2}}^2}{n_i + {z_{\alpha/2}}^2}\right)$	$\hat{p}_{w2} = \dfrac{x_2+1}{n_2+2}$ (17)
	$\pm {z_{\alpha/2}}^2\sqrt{\frac{1}{n_i + {z_{\alpha/2}}^2}\left[\hat{p}_i(1-\hat{p}_i)\left(\frac{n_i}{n_i + {z_{\alpha/2}}^2}\right) + \frac{1}{4}\left(\frac{{z_{\alpha/2}}^2}{n_i + {z_{\alpha/2}}^2}\right)\right]}$ (14)	$\hat{s}^2_{w1} = \dfrac{\hat{p}_{w1}(1-\hat{p}_{w1})}{n_1+2}$ (18)
	for $i = 1, 2$	$\hat{s}^2_{w2} = \dfrac{\hat{p}_{w2}(1-\hat{p}_{w2})}{n_2+2}$ (19)

Note. $\hat{p}_i$ is the estimated proportion of sample i; $\hat{s}_i$ is the estimated standard deviation of $\hat{p}_i$; $z_{\alpha/2}$ is the critical z-value at the nominal level; $\hat{p}_w$ is the adjusted proportion estimate; $\hat{s}_w$ is the standard deviation of $\hat{p}_w$; x_i is the number of successes in sample i.

1.4 Stratified randomization

When predicting a binary outcome variable, or any other dependent variable for that matter, there is always a risk of running into confounding variables. Confounding variables – or "confounders" – are variables that can obscure the observed relation between dependent and independent variables if not controlled for. A confounder can inflate or deflate a studied correlation and thus lead to incorrect results. One cannot control unobserved confounders for the simple reason that one is not aware of them. However, *observed* confounders can be controlled for. One way to do so is by applying "stratified randomization".

Stratified randomization is a sampling method that divides a given sample into subgroups, also known as "strata", with common "baseline characteristics". Baseline characteristics are – often demographic – participant data (e.g., sex and age) collected prior to the investigated treatment. Within strata, participants are randomly assigned to either one of the treatment conditions. If the outcome variable within a randomized controlled trial is expected to be influenced by observed confounding baseline characteristics, stratified randomization reduces their impact, as it enables these confounders – or "stratification factors" – to be more evenly distributed between the conditions than might otherwise be the case (Kendall, 2003).

The estimated treatment effect size obtained through stratified randomization is equal to the risk difference between both conditions, adjusting for stratification factors. This strata-adjusted risk difference is estimated by, firstly, subtracting each condition's proportion within strata, and secondly, calculating the weighted average of these stratum-specific risk differences (Kim & Won, 2013). There are different weighting strategies, most notably the strategy of Cochran-Mantel-Haenszel and the Inverse of Variance strategy (see respectively Equations 20 and 21). The former strategy assigns more weight to larger strata whereas the latter considers the variability of the risk difference within strata by assigning more weight to strata with lower variances, as this reflects higher estimation precision (Kim & Won, 2013).

$$w_j = \frac{\left(\frac{n_{1j}n_{2j}}{n_{1j}+n_{2j}}\right)}{\sum_{j=1}^{s}\left(\frac{n_{1j}n_{2j}}{n_{1j}+n_{2j}}\right)} \tag{20}$$

$$w_j = \frac{\left(\frac{p_{1j}\left(1-p_{1j}\right)}{n_{1j}}+\frac{p_{2j}\left(1-p_{2j}\right)}{n_{2j}}\right)^{-1}}{\sum_{j=1}^{s}\left(\frac{p_{1j}\left(1-p_{1j}\right)}{n_{1j}}+\frac{p_{2j}\left(1-p_{2j}\right)}{n_{2j}}\right)^{-1}} \tag{21}$$

for $s = 1, 2, 3, \ldots, j$ strata, in which w_j is the weight of stratum j, and n_i and p_i are respectively the sample size and proportion estimate of condition i.

1.5 Meta-analyses

In addition to allowing for statistical inference, confidence intervals facilitate comparing studies in meta-analyses. Since most drug-induced adverse events are rare and not evidently drug-related, randomized controlled trials are often insufficiently large to detect the effect of a drug on the occurrence of adverse events. In these cases, meta-analytical techniques can be of use. A "meta-analysis" is defined as the combining – or "pooling" – of information from multiple studies in a statistically appropriate manner, to permit statistical inferences about the population of interest (Arends, 2006). As with stratified randomization, the weighting of study-specific risk differences into a pooled risk difference can be done through the strategy of Cochran-Mantel-Haenszel or the Inverse of Variance strategy. The purpose of a meta-analysis on randomized controlled trials is to estimate the treatment effect size in the population. This results in the combined (i.e., pooled) estimate of the risk difference, along with its confidence interval (US Food and Drug Administration, 2018).

1.5.1 Fixed-effect and random-effects models

In any meta-analysis, the point estimates of the observed effect sizes (e.g., proportions of adverse events) will differ between the investigated studies. This variability can be solely due to sampling error. Should this be the case, the differences between study-specific effect sizes are caused by random variation instead of systematic differences between studies, implying that the true effect size is identical in each study. When sampling error is the only source of variation in observed effect sizes across studies, the true study-specific effect sizes are called "homogeneous" (Arends, 2006). In the case of homogeneity of true effect sizes, it is common statistical practice to implement a "fixed-effect model" on the data.

Nevertheless, in many meta-analyses, the variability in effect sizes across studies is not due to only sampling error. Within these meta-analyses, the true effect size differs – or is "heterogeneous" – across studies. DerSimonian and Laird (1986) introduced the "random-effects model" that incorporates this heterogeneity (Arends, 2006). According to this model, the study-specific true effect sizes have a distribution that is characterized by two parameters: the mean true effect size and the between-studies standard deviation. Both parameters can be estimated from the data. The first describes the average effect size, whereas the latter represents the heterogeneity between the study-specific true effect sizes.

Originally, statisticians assumed that study-specific effect sizes in meta-analyses were homogeneous. However, since DerSimonian's and Laird's (1986) work on heterogeneity in this context, the random-effects model has become a popular meta-analytical method (Arends, 2006). Nevertheless, fixed-effect models are still widely applied due to the narrow (i.e., more precise) confidence intervals they produce. However, fixed-effect models ignore the likely variability in study-specific effect sizes and can, therefore, lead to wrong conclusions (Thompson & Pocock, 1991). The random-effects model would appear to be a more justified option, especially in the evaluation of human drug safety, since studied subjects within this research field typically originate from different countries – and are of different races, ages, etc. – resulting in heterogeneous study-specific effect sizes (US Food and Drug Administration, 2008). Nevertheless, there is an ongoing debate on which models are preferable in which situations (Fleiss, 1993; Arends, 2006).

1.5.2 Confidence intervals

In the case of homogeneity across study-specific effect sizes, the results (i.e., pooled estimate and its confidence interval) from both meta-analytical models are usually the same. However, in the likely instances that heterogeneity is present, the random-effects model yields a higher coverage probability than the fixed-effect model (Arends, 2006). Since it is usually implausible that the true effect sizes of all studies included in a meta-analysis are equal, random-effects meta-analyses are generally recommended (Borenstein et al., 2009).

The US Food and Drug Administration (2008) requires statistical approaches for meta-analyses of randomized controlled trials to construct confidence intervals with "accurate coverage properties" (p. 14). For rare outcomes, such as drug-induced adverse events, meta-analyses may be the only way to obtain reliable effects of the intervention (e.g., the drug). Fixed-effect and random-effects models have two ways for dealing with rare events: applying a continuity correction to studies with zero events (e.g., by adding a fixed value of 0.5 to empty study cells) or simply excluding these studies from the meta-analysis altogether. This means that either a part of the available data is not used or an arbitrary number is assigned to certain data, both making the pooled effect size unreliable (Jiang et al., 2020).

To improve the reliability of the fixed-effect model for risk differences, Tian et al. (2009) propose a simple and effective confidence interval method under the fixed-effect model that deals with the problem of rare events without applying any artificial continuity correction: the "exact fixed-effect confidence interval method" (hereinafter referred to as the "Tian confidence interval method"). When events are rare, this confidence interval method has a higher coverage probability than the fixed-effect model (Tian et al., 2009). On the other hand, supposing probabilities are high (i.e., close to 1), the coverage probability of the fixed-effect model is higher (Jiang et al., 2020). The function of the Tian confidence interval method is assessable in R through the exactmeta package (Yilei & Tian, 2014). As most drug-induced adverse events seldom occur, data from meta-analyses to evaluate the safety of human drugs could be considered rare events data, making the Tian confidence interval method particularly relevant for this thesis.

1.6 Hypotheses

I will define the seven research hypotheses fundamental to this thesis in the following subsection. These hypotheses are based upon the ongoing academic debate on best methods for estimating proportions and risk differences, along with their confidence intervals.

1.6.1 Hypotheses 1 – 2: Confidence interval methods for proportions

The Wald confidence interval method for proportions relies heavily on normal approximation assumptions of the binomial distribution, without applying corrections for when this distribution is not normal. As a result, the coverage probability of the Wald confidence interval method for proportions is often less than the nominal value of 0.95. The Wilson Score method accommodates for this loss of coverage by applying a transformation (see Equation 8 in Table 1). Therefore, I expect the coverage probability of the Wilson Score confidence interval method to be higher than that of the Wald confidence interval method for proportions.

H1. The Wilson Score confidence interval method yields a higher coverage probability than the Wald confidence interval method for proportions.

The Adjusted Wald confidence interval method adds two successes and two failures before using the standard Wald formula (see Equations 5 – 7 in Table 1). By adding these observations, the binomial distribution is "pulled" towards 0.5 and less skewed when the point estimate is extreme. As a result, one would expect the coverage probability of the Adjusted Wald method to be higher when the point estimate approaches 0 or 1 than the coverage probability of the Wald method for proportions. (Agresti & Coull, 1998). I expect the same trend will be visible in the results of this thesis.

H2. The Adjusted Wald confidence interval method has a higher coverage probability than the Wald confidence interval method for proportions.

1.6.2 Hypotheses 3 – 4: Confidence interval methods for risk differences

The Newcombe Hybrid Score confidence interval method is based on the Wilson score confidence interval method, but for risk differences instead of single proportions. As pointed out by Agresti and Caffo (2000), the Newcombe Hybrid Score method has a much higher coverage probability than the Wald method for risk differences. I expect to find similar results in this thesis.

H3. The Newcombe Hybrid Score confidence interval method yields a higher coverage probability than the Wald confidence interval method for risk differences.

The Agresti-Caffo confidence interval method uses a similar correction as the Adjusted Wald interval. In this case, only one success and one failure are added in each treatment condition, equally resulting in two extra successes and two extra failures (see Equations 12 – 16 in Table 1). This transformation solves the low coverage probability when point estimates are extreme in the same way as the Adjusted Wald confidence interval method does. Hence, Agresti and Caffo (2000) demonstrated that the simple adjustment of adding two successes and two failures also works well for comparisons of proportions. Therefore, I expect the coverage probability of the Agresti-Caffo confidence interval method to be higher than the coverage probability of the Wald confidence interval method for risk differences.

H4. The Agresti-Caffo confidence interval method has a higher coverage probability than the Wald confidence interval method for risk differences.

1.6.3 Hypotheses 5 – 7: Meta-analytical methods

In the case of homogeneity across study-specific effect sizes, the results from random-effects and fixed-effect models are usually the same. However, when heterogeneity is present, which likely is often the case in practice, the random-effects model yields a higher coverage probability than the fixed-effect model (Arends, 2006). Since the assumption of all true study-specific effect sizes being equal is implausible in most meta-analyses, the random-effects model is generally recommended (Borenstein et al., 2009). I therefore expect the random-effects model to yield a higher coverage probability than the fixed-effect model.

H5. The random-effects model yields a higher coverage probability than the fixed-effect model.

The Tian confidence interval method deals with the problem of one or both treatment groups within studies having a probability of zero, without using an artificial continuity correction, as the fixed-effect confidence interval does. This results in a higher coverage probability in the case of rare events data (Tian et al., 2009). I expect to see the same pattern in the results of this thesis.

H6. When performing meta-analyses across studies with rare events data, the Tian confidence interval method has a higher coverage probability than the fixed-effect model.

Zhao et al. (2001) compared the weighting strategy of Cochran-Mantel-Haenszel with the Inverse of Variance weighting strategy and concluded that the coverage probability of the confidence intervals constructed using the Cochran-Mantel-Haenszel strategy is closer to the nominal value of .95 (Lu, 2008). Thus, I expect to find a higher coverage probability when the strategy of Cochran-Mantel-Haenszel is used than when the Inverse of Variance strategy is used to weight study-specific risk differences.

H7. When estimating confidence intervals of pooled risk differences in meta-analyses, applying the weighting strategy of Cochran-Mantel-Haenszel yields a higher coverage probability than applying the Inverse of Variance weighting strategy.

1.6.4 Conclusion

In sum, this thesis states the following seven research hypotheses:

H1. The Wilson Score confidence interval method yields a higher coverage probability than the Wald confidence interval method for proportions.

H2. The Adjusted Wald confidence interval method has a higher coverage probability than the Wald confidence interval method for proportions.

H3. The Newcombe Hybrid Score confidence interval method yields a higher coverage probability than the Wald confidence interval method for risk differences.

H4. The Agresti-Caffo confidence interval method has a higher coverage probability than the Wald confidence interval method for risk differences.

H5. The random-effects model yields a higher coverage probability than the fixed-effect model.

H6. When performing meta-analyses across studies with rare events data, the Tian confidence interval method has a higher coverage probability than the fixed-effect model.

H7. When estimating confidence intervals of pooled risk differences in meta-analyses, applying the weighting strategy of Cochran-Mantel-Haenszel yields a higher coverage probability than applying the Inverse of Variance weighting strategy.

1.7 Research questions

In the following subsection, I will discuss the four research questions central to this thesis in relation to both their theoretical relevance in the academic debate and their practical potential to improve the safety tests of drug development. Table 2 gives an overview of the research questions and hypotheses. For each type of measure, this overview shows the research question and which method is expected to yield a higher coverage probability in comparison to another method.

Table 2.

An overview of the research questions and corresponding research hypotheses per type of measure

	Type of measure			
	CIs of proportions	*CIs of risk differences*	*Meta-analytical methods*	*Weighting strategies*
Q	What is the best method for estimating confidence intervals of proportions? (*Q1*)	What is the best method for estimating confidence intervals of risk differences? (*Q2*)	Which model is best at estimating a pooled risk difference, along with its confidence interval, in a meta-analysis? (*Q3*)	What is the best strategy for weighting study-specific risk differences in a meta-analysis? (*Q4*)
H	Wilson Score > Wald (*H1*)	Newcombe Hybrid Score > Wald (*H3*)	Random-effects > Fixed-effect (*H5*)	Cochran-Mantel-Haenszel > Inverse of Variance (*H7*)
	Adjusted Wald > Wald (*H2*)	Agresti-Caffo > Wald (*H4*)	Tian > Fixed-effect (*H6*)	

Note. > means "yields a higher coverage probability than"; *Q* = research question; *H* = hypothesis; *CI* = confidence interval. See Table 1 for each confidence interval method's equation.

I have chosen to focus on coverage probabilities, which is in line with the requirements of the US Food and Drug Administration (2008). Moreover, I followed other researchers' footsteps in this regard (Agresti & Coull, 1998; Brown et al., 2001, 2002; Fagerland et al., 2015; Jiang et al., 2020; O'Gorman et al., 1994; Pires & Amado, 2008; Yan & Su, 2010). But besides the reliability of a confidence interval, its meaning is another crucial aspect, which is reflected by its width. Increasing the meaning of a confidence interval parallels reducing the width. Given that a confidence interval offers the range within which the true value of the parameter of

interest can be expected, a narrow confidence interval enables more precise estimates of this parameter. This can be achieved by increasing the sample size, as a smaller sample reduces the variability of the sampling distribution. Confidence interval properties, such as interval width, thus provide useful evaluation criteria for future research. Therefore, I computed the average confidence interval width per method, in addition to each method's coverage probability.

1.7.1 Questions 1 – 2: Confidence intervals methods for proportions and risk differences

Traditional Wald methods are vulnerable to extreme point estimates and small samples resulting in low coverage probabilities. This thesis explores four alternatives to the traditional Wald methods, namely; two confidence interval methods for the estimation of *single* proportions – the Adjusted Wald interval and Wilson Score methods – and two confidence interval methods for the estimation of proportion *differences* (i.e., risk differences) – the Agresti-Caffo and the Newcombe Hybrid Score methods (see Table 1 for their equations). I will evaluate which confidence interval methods have the highest coverage probabilities. This will allow me to answer my first two research questions:

Q1. What is the best method for estimating confidence intervals of proportions?

Q2. What is the best method for estimating confidence intervals of the differences between two proportions (i.e., risk differences)?

Hypotheses 1 and 2 are tested through research question 1 by comparing the coverage probability of the Wald confidence interval method for proportions with coverage probabilities of the Wilson Score and Adjusted Wald confidence interval methods. Research question 2 is focused on testing hypotheses 3 and 4 by comparing the coverage probability of the Wald confidence interval method for risk differences with coverage probabilities of the Newcombe Hybrid Score and Agresti-Caffo confidence interval methods.

1.7.2 Questions 3 – 4: Meta-analytical methods

My final simulation study is directly focused on the need for reliable meta-analyses of randomized controlled trials. To this end, I will estimate pooled risk differences along with their confidence intervals in meta-analyses. I will compare coverage probabilities of the fixed-effect model, the Tian confidence interval method, and the random-effects model. For each method, I will use both the Cochran-Mantel-Haenszel and Inverse of Variance strategies to weight study-specific effect sizes. This final simulation study is aimed at answering my third and fourth research questions:

Q3. Which model is best at estimating a pooled risk difference, along with its confidence interval, in a meta-analysis?

Q4. What is the best strategy for weighting study-specific risk differences in a meta-analysis?

Research question 3 is aimed at testing hypotheses 5 and 6 by comparing coverage probabilities of the fixed-effect model, the Tian confidence interval method, and the random-effects model. My fourth, and final, research question is aimed at testing hypothesis 7 by comparing coverage probabilities as a result of using the Cochran-Mantel-Haenszel strategy or the Inverse of Variance strategy to weight study-specific risk differences.

1.7.3 Conclusion

In sum, this thesis is designed for answering the following four research questions:

Q1. What is the best method for estimating confidence intervals of proportions?

Q2. What is the best method for estimating confidence intervals of risk differences?

Q3. Which model is best at estimating a pooled risk difference, along with its confidence interval, in a meta-analysis?

Q4. What is the best strategy for weighting study-specific risk differences in a meta-analysis?

For my first two research questions, I am especially interested in the vulnerability of the confidence intervals towards small sample sizes and extreme point estimates. My third and fourth research questions are more concerned with rare events data, that typically form the basis of the evaluation of human drug safety through meta-analyses.

Section 2. Methods

The following section covers the intended methodology through which I will test the seven research hypotheses and answer the corresponding four research questions. Table 3 gives an overview of the point estimate, method, and number of output parameters per simulation study, research question, and hypothesis. The three simulation studies were performed using R, and more specifically the R package "meta" (Balduzzi et al., 2019) and "exactmeta" (Yilei & Tian, 2014). The final code of my simulated data, including my functions for confidence interval methods, coverage probabilities, and average widths can be found in Appendix A.

Table 3.

An overview of the point estimates, confidence interval methods, estimation models, weighting strategies, and number of output parameters per simulation study, research question, and research hypothesis

Simulation study	*Q*	*H*	*Point estimate*	*CI-methods*	*Weighting strategies*	*Output parameters*
1	1	1-2	Proportion ($\hat{p}$)	- Wald for proportions - Adjusted Wald - Wilson Score		7
2	2	3-4	Risk difference ($\hat{p}_1 - \hat{p}_2$)	- Wald for risk differences - Agresti-Caffo - Newcombe hybrid Score		7
3	3-4	5-7	Pooled risk difference ($\hat{p}_1 - \hat{p}_2$)	- Fixed-effect - Tian - Random-effects	- Cochran-Mantel-Haenszel - Inverse of Variance	18

Note. *Q* = research question; *H* = research hypothesis; *CI-methods* = confidence interval methods. See Table 1 for each confidence interval method's equation.

2.1 Questions 1 – 2 (hypotheses 1 – 4)

2.1.2 Parameters

What is the best method for estimating confidence interval of proportions and risk differences? To determine which confidence interval methods for proportions and risk differences yield the most accurate coverage probabilities, I compared the coverage probabilities of the three confidence interval methods per research question (see the first two rows of Table 3). As I am particularly interested in the robustness of these confidence interval methods towards small sample sizes and extreme proportion estimates, I varied two simulation parameters: the sample size and true proportion or true risk difference.

I performed two simulation studies to this end: one for the first and one for the second research question. For my first research question, the sample size took two values (i.e., 100 and 5,000) and the true proportion took three values (i.e., 0.01, 0.50, and 0.98). In total, this resulted in 6 (i.e., 2 x 3) combinations of parameter values. For my second research question, each sample consisted of two conditions, of which the sizes and true proportions took the same values as for research question 1. In total, this resulted in 36 (i.e., 2^2 x 3^2) simulation parameter combinations.

Per combination, I simulated 1,000 replications, each with a different random seed used to initialize generating 100 or 5,000 random values of a proportion of 0.01, 0.50, or 0.98. For each of the resulting 6,000 or 36,000 simulations, I produced, in addition to the simulation parameters, the point estimate of the proportion or risk difference, and the lower and upper confidence interval boundaries of each method. This resulted in 7 output parameters per simulation (see the first two rows of Table 3).

2.1.3 Procedure

For each simulation parameter combination, I computed each method's confidence interval (see Equations 3 – 19 in Table 1). Thereafter, I calculated the coverage probability of each confidence interval method as the proportion of confidence intervals per method that embed the true proportion or risk difference. Additionally, I calculated each method's average confidence interval width as the mean difference between the lower and the upper boundary of all confidence intervals created with that method.

2.2 Questions 3 – 4 (hypotheses 5 – 7)

2.2.1 Parameters

Which model is best at estimating a pooled risk difference, along with its confidence interval, in a meta-analysis? And what is the best strategy for weighting study-specific risk differences in a meta-analysis? To answer my third and fourth research questions, I performed a third simulation study in which I compared the coverage probabilities of the fixed-effect model, Tian confidence interval method, and random-effects model, using both the Cochran-Mantel-Haenszel and Inverse of Variance weighting strategies.

I simulated two studies, both consisting of two conditions. As I am interested in the robustness of the meta-analytical methods towards rare events data, I kept each condition's

sample size constant (i.e., 5,000), but varied the true proportions, in which I included a proportion of zero (i.e., 0.00 and 0.01). In total, this resulted in 16 (i.e., 2^4) simulation parameter combinations.

Per combination, I simulated 10 replications, each with a different random seed used to initialize generating 5,000 random values of a proportion of either 0.00 or 0.01. For each of the resulting 160 simulations, I produced, in addition to the simulation parameters, the model estimates of the pooled risk differences, and the lower and upper confidence interval boundaries of these estimates. This resulted in 18 parameters per simulation, besides the simulation parameters (see the third row of Table 3).

2.2.2 Procedure

I used the function "metabin", from the meta package (Balduzzi et al., 2019), to perform the fixed-effect and random-effects models on each parameter combination. Within this function, I specified either the Inverse of Variance weighting strategy or the Cochran-Mantel-Haenszel weighting strategy. This function produces the pooled risk difference estimate and its confidence interval for both the fixed-effect and random-effects model. To compare the fixed-effect and random-effects models with the Tian confidence interval method, I computed the Tian risk difference estimate and confidence interval through the function "meta.exact" from the exactmeta package (Yilei & Tian, 2014). This function allows for specifying the same weighting strategies as the metabin function. In the end, I calculated the coverage probability and average width of each weighting strategy – meta-analytical model combination in the same way as for research questions 1 and 2.

Section 3. Results

In the following section, I will present the findings of this thesis based upon the information gathered as a result of the applied methodology. I will relate these findings to their respective hypotheses and research questions, to be able to choose the best methods for estimating proportions and risk differences, along with their confidence intervals, for either single studies, stratified samples, or meta-analyses.

3.1 Questions 1 – 2 (hypotheses 1 – 4)

What is the best method for estimating confidence intervals of proportions and risk differences? After performing my first two simulation studies, which correspond to my first two hypotheses, I computed the proportion of cases for which the true proportion or risk difference lies within the confidence intervals of each confidence interval method (i.e., the coverage probability; see the first column of Table 4). Additionally, I computed the average width of each method (see the second column of Table 4).

Table 4.

An overview of the coverage probability and average width for the investigated confidence interval methods for proportions and risk differences

Confidence interval method	*Coverage probability*	*Average width*
For proportions		
Wald	0.8465	0.0491
Wilson Score	0.9392	0.0554
Adjusted Wald	0.9585	0.0596
For risk differences		
Wald	0.8603	0.0851
Newcombe Hybrid Score	0.9999	0.1664
Agresti-Caffo	0.9104	0.0892

Note. See Table 1 for each confidence interval method's equation.

Among the confidence interval methods for proportions and risk differences, the Wald methods for proportions and risk differences had the lowest coverage probabilities (i.e., respectively 0.8465 and 0.8603; see the first column of Table 4), as compared to the other methods. Therefore, my first four research hypotheses, that respectively the Wilson Score, Adjusted Wald, Newcombe Hybrid Score, and Agresti-Caffo confidence interval methods yield higher coverage probabilities than Wald confidence interval methods, can be supported by these results.

The confidence interval method for proportions with the highest – and most accurate (i.e., closest to 0.95) – coverage probability was the Adjusted Wald method, with a coverage probability of 0.9585 (see the first column of Table 4). For risk differences, the Newcombe Hybrid Score confidence interval method had the highest – and most conservative – coverage probability (i.e., 0.9999). Although the coverage probability of the Agresti-Caffo method was lower (i.e., 0.9104), it was more accurate.

To visualize the performance of the confidence interval methods under different simulation parameter combinations, I have created confidence interval plots for all six methods, in which the true parameter (i.e., true proportion or risk difference) and the sample size are either small or large (see Figure 1). In the upper-left plot, the true proportion (i.e., see the black lines in Figure 1) is 0.01 and the sample size is 100, whereas in the upper-right plot, the true proportion is 0.99 and the sample size is 5000. In the lower-left plot, the true risk difference is 0.00 and the sample sizes are both 100, while in the lower-right plot, the true risk difference is 0.98 and the sample sizes are both 5000. For all four simulation parameter combinations, I chose to plot solely the first 10 (instead of all 1000) simulations, for the sake of clarity.

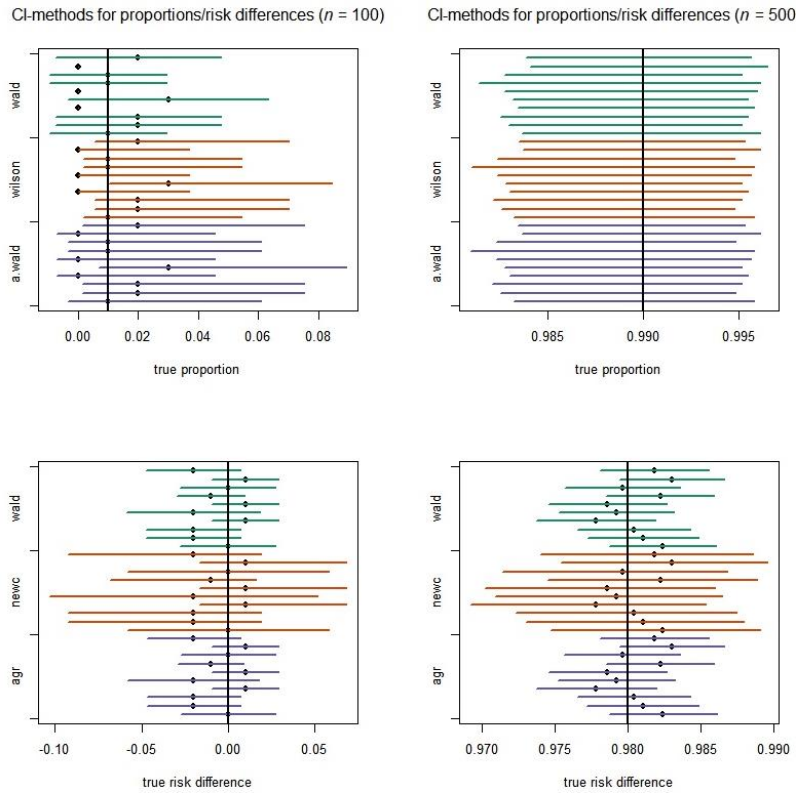

Figure 1. The distribution of point estimates (i.e., black dots) and confidence intervals (i.e., colored lines) of confidence interval methods for true proportions (i.e., first simulation study; upper two plots) and risk differences (i.e., second simulation study; lower two plots) around either the true proportion or the true risk difference (i.e., black lines)

The upper-left plot of Figure 1 demonstrates that when the true proportion and sample size are low, three out of ten Wald confidence intervals had both lower and upper boundaries of zero. This resulted in a low average width and low coverage probability of the Wald method for proportions, as compared to the Wilson Score and Adjusted Wald methods. The coverage probability of the Wald method was lower, since the chance of a parameter being covered by a confidence interval without width is, naturally, small.

3.2 Questions 3 – 4 (hypotheses 5 – 7)

Which model is best at estimating a pooled risk difference, along with its confidence interval, and what is the best strategy for weighting study-specific risk differences in a meta-analysis? Table 5 gives an overview of the coverage probabilities and average widths of the confidence intervals derived from the fixed-effect model, Tian confidence interval method, and random-effects model, when using either the weighting strategy of Cochran-Mantel-Haenszel or the Inverse of Variance weighting strategy.

Table 5.

An overview of the coverage probability and average width of the investigated meta-analytical methods, using Cochran-Mantel-Haenszel and Inverse of Variance weighting strategies

Meta-analytical method	*Weighting strategy*			
	CMH		*IOV*	
	Coverage probability	*Average width*	*Coverage probability*	*Average width*
Fixed-effect model	0.9625	0.0038	0.7188	0.0028
Tian confidence interval method	0.8813	0.0037	0.9313	0.0101
Random-effects model	1.000	0.0319	1.000	0.0161

Note. CMH = Cochran-Mantel-Haenszel; IOV = Inverse of Variance.

My fifth and sixth hypotheses stated that respectively the random-effects model and Tian confidence interval method yield higher coverage probabilities than the fixed-effect model. The random-effects model had the highest coverage probability (i.e., 1.000; see the first and third columns of Table 5), confirming my fifth hypothesis. My sixth hypothesis can, however, only partly be confirmed by my results, since the coverage probability of the Tian confidence interval method was only higher than the coverage probability of the fixed-effect model when the Inverse of Variance weighting strategy is applied.

Based on only the coverage probability, it does not seem to matter which weighting strategy is applied when running a random-effects model. Nonetheless, the lower average confidence interval width of the random-effects model after applying the Inverse of Variance

weighting strategy makes this strategy preferable (see the third column of Table 5). For the fixed-effect model, applying the strategy of Cochran-Mantel-Haenszel yielded a higher coverage probability, whereas the opposite conclusion can be made for the Tian confidence interval method (see Table 5). These results partly confirm my seventh hypothesis – which states that the weighting strategy of Cochran-Mantel-Haenszel yields a higher coverage probability than the Inverse of Variance weighting strategy – as this expectation was only met for the fixed-effect model.

To visualize the performance of the methods and weighting strategies, I have created confidence interval plots for all six method-strategy combinations, in which the true risk difference is either 0.00 or 0.01 (see Figure 2). In the left plots, the weighting strategy of Cochran-Mantel-Haenszel is applied, whereas in the right plots the Inverse of Variance weighting strategy is used. In the upper plots, the true risk difference (i.e., see the black lines in Figure 2) is equal to 0.00, while in the lower plots, this value is equal to 0.01.

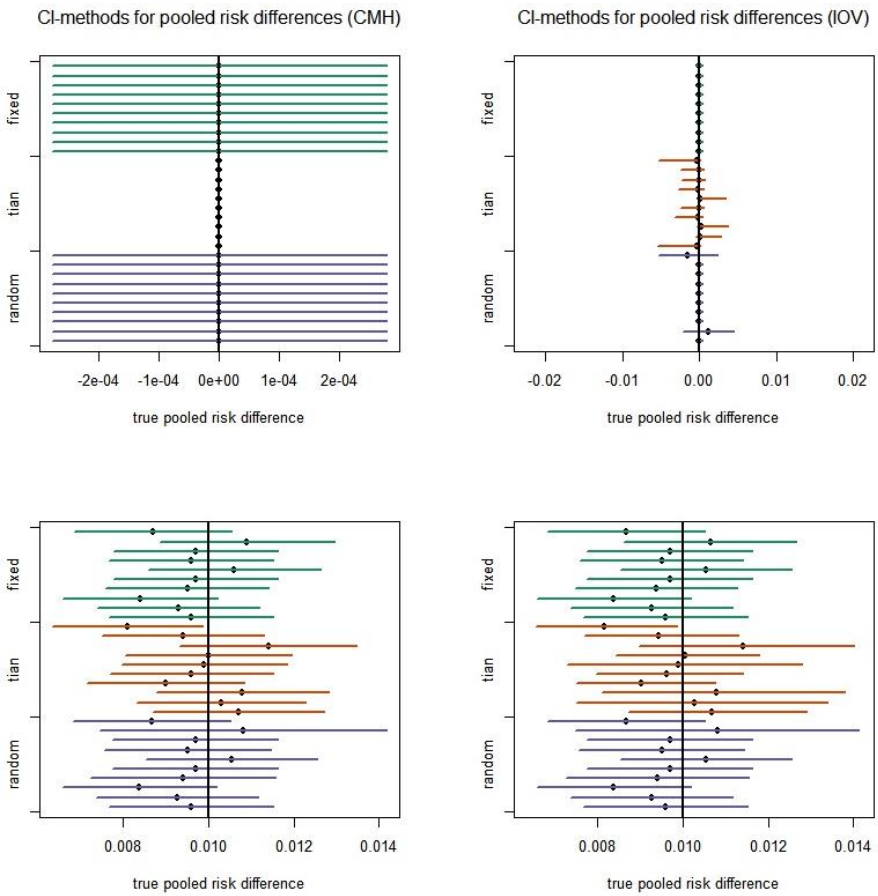

Figure 2. The distribution of point estimates (i.e., black dots) and confidence intervals (i.e., colored lines) around the pooled risk difference (i.e., black lines) of meta-analytical methods, using either the weighting strategy of Cochran-Mantel-Haenszel (i.e., left plots) or the Inverse of Variance weighting strategy (i.e., right plots)

The upper-right plot of Figure 2 shows that, when the Inverse of Variance weighting strategy was applied, the Tian confidence interval method dealt better with true pooled risk differences of 0.00 as compared to the fixed-effect and random-effects models. In this plot, most of the intervals produced with the latter two models had no width, decreasing the chance of covering the true parameter, whereas the Tian method produced intervals with a certain width, increasing its coverage probability. On the contrary, when the weighting strategy of Cochran-Mantel-Haenszel was used and the true pooled risk difference was 0.00 (see the upper-left plot Figure 2), the confidence intervals produced by the Tian method had no width at all, resulting in a coverage probability of zero for these simulations. This pattern was also reflected in the results of Table 5: the coverage probability of the Tian method was higher when using the Inverse of Variance strategy than when using the strategy of Cochran-Mantel-Haenszel.

When the true pooled risk difference was 0.01 (see the lower plots of Figure 2), there is no visible difference in confidence intervals produced through applying either the weighting strategy of Cochran-Mantel-Haenszel or the Inverse of Variance weighting strategy. Nonetheless, the Tian method showed less coverage than the other methods, as for each weighting strategy, one confidence interval of the Tian method did not cover the true parameter.

Section 4. Discussion

In this section, I will present the Discussion on the basis of the results which have been presented in the previous section. The aim is to interpret the observed results in order to answer my research questions. I have divided this section into four subsections in which I will describe the findings of this thesis and their implications, along with limitations and suggestions for future research. At first, I will summarize the problem definition. After that, I will answer the research questions posed in the Introduction and confirm or reject the corresponding research hypotheses. Thereafter, I will explain the implications of the answers to the research questions and how these answers match the existing knowledge in the field. Finally, I will consider methodological limitations and discuss suggestions for future research.

4.1 Problem definition

This thesis intends to contribute to the field of human drug safety evaluation. The safety of human drugs can be evaluated by comparing binomial proportions of drug-induced adverse events between a control and an experimental condition. Confidence intervals of this proportion difference (i.e., risk difference) not only allow for statistical inference but also give information about the expectation of finding this difference in the population. Confidence interval estimation of binomial proportions is one of the most commonly applied analyses in statistical inference. Nevertheless, it constitutes one of the elemental problems of statistical practice as there is no consensus on which confidence interval methods are recommended in which situations (Brown et al., 2001, 2002; Guan, 2012; Pires & Amado, 2008).

When estimating the risk difference of drug-induced adverse events between two treatment conditions, stratified randomization can facilitate treatment balance by controlling for observable confounders. Different strategies can be applied to weight stratum-specific risk differences, most notably the Cochran-Mantel-Haenszel and Inverse of Variance weighting strategies. The same strategies apply to the estimation of pooled risk differences within meta-analyses. As both strategies take different aspects into account in order to reliably estimate pooled risk differences, which strategy is most reliable might be open for debate (Kim & Won, 2013).

The goal of meta-analyses of randomized controlled trials is to estimate the treatment effect size in the population, resulting in the combined (i.e., pooled) estimate of the proportion or risk difference, along with its confidence interval (US Food and Drug Administration, 2018). As stated by the US Food and Drug Administration (2008), statistical approaches for meta-

analyses of randomized controlled trials should ensure "confidence intervals have accurate coverage properties" (p. 14). However, the traditional Wald confidence interval methods for proportions and risk differences are infamous for their low coverage probabilities when sample sizes are small or point estimates are extreme.

Since meta-analyses became popular in medicine, there has been a growing interest and involvement in the development of statistical meta-analytic methods amongst biostatisticians (Arends, 2006). At first, attention was paid to obtaining a common effect across studies within a meta-analysis, whereas later on, the focus moved to quantifying and reporting the heterogeneity between those studies. This shift in focus parallels the transition in popularity from fixed-effect to random-effects models.

Confidence intervals derived from the fixed-effect model are known to perform poorly when the proportion or risk difference in at least one study within the meta-analysis is equal to zero. These rare events data are particularly common in the evaluation of the safety of human drugs since most drug-induced adverse events are rare and not evidently drug-related. In these cases, the investigated effect can be too small to uncover. The occurrence of rare events data is exactly the reason for the demand for meta-analyses in this research field. Even so, one of the two most widely applied meta-analytical models does not seem appropriate for these kinds of data.

4.2 Findings

4.2.1 Questions 1 – 2 (hypotheses 1 – 4)

What is the best method for estimating confidence intervals of proportions or risk differences? Conforming to my expectations, the Wilson Score, Adjusted Wald, Newcombe Hybrid Score, and Agresti-Caffo confidence interval methods yielded higher coverage probabilities than Wald confidence interval methods (see Table 4). Therefore, my first four research hypotheses are supported by my results. For proportions, the Adjusted Wald confidence interval method had the highest – and most accurate (i.e., closest to 0.95) – coverage probability, whereas, for risk differences, the Newcombe Hybrid Score method yielded the highest coverage probability. However, the coverage probability of the Agresti-Caffo method was, although lower, more accurate. Plots of the confidence intervals created with the Wald method for proportions visualize that this method showed an especially low average width when the true proportion and sample size were small, causing a low coverage probability (see Figure 1).

4.2.3 Questions 3 – 4 (hypotheses 5 – 7)

Which model is best at estimating a pooled risk difference, along with its confidence interval, in a meta-analysis? When the Inverse of Variance weighting strategy was applied, hypotheses 5 and 6 are both confirmed, as respectively the Tian confidence interval method and the random-effects model yielded higher coverage probabilities than the fixed-effect model. Nevertheless, the coverage probability of the Tian confidence interval method was lower than that of the fixed-effect model when the weighting strategy of Cochran-Mantel-Haenszel was applied (see Table 5), rejecting my sixth hypothesis. Hence, although my fifth hypothesis can be confirmed, my sixth hypothesis can only *partly* be confirmed.

The random-effects model yielded the highest coverage probability, regardless of which weighting strategy was applied. However, this coverage probability was equal to 1.000, and, therefore, highly conservative. When the weighting strategy of Cochran-Mantel-Haenszel was applied, the most accurate coverage probability (i.e., 0.9625) belonged to the fixed-effect model. The Tian confidence interval method was most accurate (i.e., had a coverage probability of 0.9313) when the Inverse of Variance weighting strategy was used (see Table 5).

The upper plots in Figure 2 reveal that, when the Inverse of Variance weighting strategy was applied and the true risk difference was equal to 0.00, the Tian confidence interval method had a higher coverage probability than the other methods, since it produced intervals with a certain width, instead of intervals with mostly no width at all. Yet, if the weighting strategy of Cochran-Mantel-Haenszel was applied, the reversed pattern is visible. Furthermore, the lower plots in Figure 2 show that the Tian method yielded a lower coverage probability than the other methods when the estimated risk difference was not equal to zero, regardless of which weighting strategy was applied.

What is the best strategy for weighting study-specific risk differences in a meta-analysis? Under the random-effects model, the use of either the strategy of Cochran-Mantel-Haenszel or the Inverse of Variance strategy to weight study-specific risk differences did not affect this model's coverage probability, but *did* influence its average width: the average width was lower after applying the Inverse of Variance weighting strategy (see Table 5).

Under the fixed-effect model, applying the Cochran-Mantel-Haenszel weighting strategy yielded the highest, and most accurate, coverage probability. On the contrary, applying this strategy within the Tian confidence interval method yielded a lower, and less accurate, coverage probability than applying the strategy of Inverse of Variance. The findings of my third simulation study (see Table 5 and Figure 2), therefore, partly support my seventh hypothesis

that applying the weighting strategy of Cochran-Mantel-Haenszel weighting yields a higher coverage probability than applying the Inverse of Variance weighting strategy, as this only applies to the fixed-effect model.

4.3 Implications

This thesis has led to conclude that the alternative confidence interval methods for proportions and risk differences perform better (i.e., yield more accurate coverage probabilities) than the Wald methods. As expected, based on research by Agresti and Caffo (2000), the transformation within the Wilson Score and Newcombe Hybrid Score confidence interval methods caused an increase in coverage probabilities, as compared to the Wald methods. Furthermore, adding two successes and failures before using the Wald equation, as done by the Adjusted Wald and Agresti-Caffo confidence interval methods, increased the coverage probabilities of these method, conforming to findings of respectively Agresti and Coull (1998), and Agresti and Caffo (2000).

In accordance with previous findings by Arends (2006) and Borenstein et al. (2009), the random-effects model showed a higher coverage probability than the fixed-effect model. Nevertheless, the coverage probability of the former model was highly conservative, and therefore this model is not recommended. The Tian confidence interval method, as proposed by Tian et al. (2009), indeed, showed a higher, and more accurate, coverage probability than the fixed-effect model in case of rare events data, when the Inverse of Variance weighting strategy was used. However, this did not hold when the weighting strategy of Cochran-Mantel-Haenszel was applied. Furthermore, the Tian method yielded a lower coverage probability than the fixed-effect and random-effects models when the estimated risk difference was not equal to zero, regardless of which weighting strategy was applied.

Additionally, this thesis partly ties well with previous studies wherein the two aforementioned weighting strategies are compared: in line with my expectations and previous findings of Zhao et al. (2001), more accurate confidence intervals were constructed by the fixed-effect model when applying the strategy of Cochran-Mantel-Haenszel strategy than the Inverse of Variance strategy. On the contrary, the Inverse of Variance weighting strategy is a better choice when applying the Tian confidence interval method. Which weighting strategy was applied under the random-effects model did not influence this model's coverage probability, but *did* affect its average confidence interval width: the Inverse of Variance weighting strategy resulted in the lowest average width. Therefore, under the random-effects

model, the use of this weighting strategy is recommended, since it produces more meaningful (i.e., more precise) confidence intervals. This is in line with research by Arends (2006), stating that a random-effects model takes the between-study variance into account, which leads to more variance of the effect size, and thus, a wider confidence interval of the pooled effect size estimate.

The most important implication of this thesis is that the combination of the Tian confidence interval method and Inverse of Variance weighting strategy is preferable when estimating pooled risk differences in meta-analyses consisting of rare events data. When applying the weighting strategy of Cochran-Mantel-Haenszel, the fixed-effect model was, however, more accurate. This goes against the expectation that the random-effects model is a more justified option than the fixed-effect model when study-specific effect sizes are heterogeneous, which is typically the case in human drug safety evaluation (US Food and Drug Administration, 2008).

4.4 Future research

This thesis aims to be a contribution to the field of meta-analyses of randomized controlled trials to evaluate the safety of human drugs. As the problem of awaiting confounders is also present within this field, controlling for these confounders is a useful tool for estimating pooled risk differences more reliably. Through research questions 3 and 4, this thesis has compared different meta-analytical techniques to compare pooled safety data from multiple studies, through the last research question. However, estimating pooled risk differences whilst adjusting for baseline characteristics is beyond the scope of this thesis. Therefore, this thesis has only touched upon the estimation of pooled risk differences in meta-analyses to evaluate drug-induced adverse events. Thus, future research could aim at investigating how to properly deal with confounders when estimating pooled risk differences of controlled trials in meta-analyses to assess drug safety.

Future research could additionally explore coverage probabilities of the investigated confidence interval methods per parameter combination. This could help explain low coverage probabilities and raise insight into the robustness of the investigated methods towards extreme true parameter values and small sample sizes. Because of the time frame of this thesis, I decided not to investigate this further than the plots in Figure 1 and Figure 2. Therefore, it remains partly unclear how the coverage probabilities of the investigated confidence interval methods are

related to the simulation parameters, and which methods should be preferred in which situations.

As explained in the Methods, I have chosen to use linear models (i.e., fixed-effect and random-effects models) to estimate pooled risk differences. However, there is no apparent reason, except perhaps for the sake of simplicity, to favor these models over others. Furthermore, the confidence interval methods I investigated are parametric and give analytical solutions under certain assumptions and approximations. There are modern methods, such as bootstrap and permutation methods, that may give more reliable results. Future research could explore alternative ways of modeling proportions or risk differences and estimating their confidence intervals.

Another apparent limitation of this thesis is the limited number of cells in the data-generating design of the third simulation study. The fitting of the meta-analytical simulations was computationally intensive due to the slow meta.exact function. Given the limited number of simulations (i.e., 10) for the third simulation study, the conclusions drawn from the third and fourth research questions should consequently be treated with considerable caution. As more replications would allow for a more reliable evaluation, future research should attempt to replicate this thesis with more computer power. Additionally, the simulative nature of this thesis is a possible limitation, as "real" safety data might differ from the created data. This limits the generalizability of the results of this thesis.

Despite the limitations mentioned, this thesis has identified the most appropriate methods for estimating confidence intervals of proportions, risk differences, and pooled risk differences in meta-analyses of randomized controlled trials. Therefore, it contributes to both the academic debate on binomial confidence interval methods and the current research field of human drug safety evaluation.

References

Agresti, A., & Coull, B. A. (1998). Approximate is better than "Exact" for interval estimation of binomial proportions. *American Statistician*, *52*(2), 119–126.

Arends, L. R. (2006). *Multivariate meta-analysis: modelling the heterogeneity; dangerous or delicious?* [Doctoral dissertation, Erasmus University Rotterdam]. https://repub.eur.nl/pub/7845/.

Balduzzi, S., Rücker, G., & Schwarzer, G. (2019). How to perform a meta-analysis with R: A practical tutorial. *Evidence-Based Mental Health*, *22*, 153–160.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. P. (2009). *Introduction to Meta-Analysis* (Issue January).

Brown, L. D., Cai, T. T., & DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, *16*(2), 101–133.

Brown, L. D., Cai, T. T., & DasGupta, A. (2002). Confidence intervals for a binomial proportion and asymptotic expansions. *Annals of Statistics*, *30*(1), 160–201.

Fagerland, M. W., Lydersen, S., & Laake, P. (2015). Recommended confidence intervals for two independent binomial proportions. *Statistical Methods in Medical Research*, *24*(2), 224–254.

Fleiss, J. L. (1993). Review papers: The statistical basis of meta-analysis. *Statistical Methods in Medical Research*, *2*(2), 121–145.

Guan, Y. (2012). A generalized score confidence interval for a binomial proportion. *Journal of Statistical Planning and Inference*, *142*(4), 785–793.

Jiang, T., Cao, B., & Shan, G. (2020). Accurate confidence intervals for risk difference in meta-analysis with rare events. *BMC Medical Research Methodology*, *20*(1), 1–10.

Kendall, J. M. (2003). Designing a research project: Randomised controlled trials and their principles. *Emergency Medicine Journal*, *20*(2), 164–168.

Kim, Y., & Won, S. (2013). Adjusted proportion difference and confidence interval in stratified randomized trials. *PharmaSUG*, 1–8.

Newcombe, R. G. (1998). Interval estimation for the difference between independent proportions: Comparison of eleven methods. *Statistics in Medicine*, *17*(8), 873–890.

O'Gorman, T. W., Woolson, R. F., & Jones, M. P. (1994). A comparison of two methods of estimating a common risk difference in a stratified analysis of a multicenter clinical trial. *Controlled Clinical Trials*, *15*(2), 135–153.

Pires, A., & Amado, C. (2008). Interval estimators for a binomial proportion: Comparison of

twenty methods. *REVSTAT–Statistical Journal*, *6*(2), 165–197.

Thompson, S. G., & Pocock, S. J. (1991). Can meta-analysis be trusted? *Lanset*, *338*(8775), 1127–1130.

Tian, L., Cai, T., Pfeffer, M. A., Piankov, N., Cremieux, P. Y., & Wei, L. J. (2009). Exact and efficient inference procedure for meta-analysis and its application to the analysis of independent 2 × 2 tables with all available data but without artificial continuity correction. *Biostatistics*, *10*(2), 275–281.

US Food and Drug Administration. (2018). *Meta-Analyses of Randomized, Controlled, Clinical Trials (RCTs) to Evaluate the Safety of Human Drugs or Biologic Products Guidance for Industry*. *November*, 1–29.

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, *22*(158), 209–212.

Yan, X., & Su, X. G. (2010). Stratified Wilson and Newcombe confidence intervals for multiple binomial proportions. *Statistics in Biopharmaceutical Research*, *2*(3), 329–335.

Yilei, Y., & Tian, L. (2014). *Exactmeta: Exact fixed effect meta analysis* (1.0-2). https://cran.r-project.org/package=exactmeta

Appendix A – R code

```
################################################################################
#                    CODE THESIS - JASMIJN VAN STAADEN                         #
################################################################################

library(meta) # necessary for fixed-effect and random-effects CIs
library(exactmeta) # necessary for exact fixed-effect CI
library(scales) # necessary for plots

######## 1. FUNCTIONS ########
# Wald CI for a Proportion
waldInterval.prop <- function(p, n, conf.level = 0.95){
  sd <- sqrt(p*((1 - p)/n))
  z <- qnorm(c((1 - conf.level)/2, 1 - (1 - conf.level)/2))
  ci <- p + z*sd
  return(ci)
}

# Adjusted Wald CI
a.waldInterval <- function(p, n, conf.level = 0.95){
  x <- p*n
  pw <- (x + 2)/(n + 4)
  sd <- sqrt((pw*(1 - pw))/(n + 4))
  z <- qnorm(c((1 - conf.level)/2, 1 - (1 - conf.level)/2))
  ci <- pw + z*sd
  return(ci)
}

# Wilson Score CI
wilsonInterval <- function(p, n, conf.level = 0.95){
  z <- qnorm(c((1 - conf.level)/2, 1 - (1 - conf.level)/2))
  var <- p*((1 - p)/n)
  ci <- (p + z^2/(2*n) + z*sqrt(var + z^2/(4*n^2)))/(1 + z^2/n)
  return(ci)
}

# Wald CI for a Proportion Difference
waldInterval.diff <- function(p1, n1, p2, n2, conf.level = 0.95){
  var1 <- p1*((1 - p1)/n1)
  var2 <- p2*((1 - p2)/n2)
  z <- qnorm(c((1 - conf.level)/2, 1 - (1 - conf.level)/2))
  ci <- p2 - p1 + z*sqrt(var1 + var2)
  return(ci)
}

# Agresti-Caffo CI
agresticaffoInterval <- function(p1, n1, p2, n2, conf.level = 0.95){
  x1 <- p1*n1
  x2 <- p2*n2
  pw1 <- (x1 + 1)/(n1 + 2)
  pw2 <- (x2 + 1)/(n2 + 2)
  var1 <- p1*((1 - pw1)/(n1 + 2))
  var2 <- p2*((1 - pw2)/(n2 + 2))
  z <- qnorm(c((1 - conf.level)/2, 1 - (1 - conf.level)/2))
  ci <- p2 - p1 + z*sqrt(var1 + var2)
}


# Newcombe Hybrid Score CI
newcombehybridInterval <- function(p1, n1, p2, n2, conf.level = 0.95){
  z <- qnorm(1 - (1 - conf.level)/2)
  p1.u <- p1*(n1/(n1 + z^2)) + 0.5*(z^2/(n1 + z^2)) +
    z^2*(sqrt((1/(n1 + z^2)*((p1*(1 - p1)*(n1/(n1 + z^2) + 0.25*(z^2/(n1 + z^2))))))))
  p1.l <- p1*(n1/(n1 + z^2)) + 0.5*(z^2/(n1 + z^2)) -
    z^2*(sqrt((1/(n1 + z^2)*((p1*(1 - p1)*(n1/(n1 + z^2) + 0.25*(z^2/(n1 + z^2))))))))
  p2.u <- p2*(n2/(n2 + z^2)) + 0.5*(z^2/(n2 + z^2)) +
    z^2*(sqrt((1/(n2 + z^2)*((p2*(1-p2)*(n2/(n2 + z^2)+0.25*(z^2/(n2 + z^2))))))))
  p2.l <- p2*(n2/(n2 + z^2)) + 0.5*(z^2/(n2 + z^2)) -
    z^2*(sqrt((1/(n2 + z^2)*((p2*(1 - p2)*(n2/(n2 + z^2) + 0.25*(z^2/(n2 + z^2))))))))
  ci <- c((p2 - p1 - sqrt((p2 - p2.l)^2 + (p1.u - p1)^2)), (p2 - p1 + sqrt((p1 - p1.l)^2 +
(p2.u - p2)^2)))
  return(ci)
}

# Proportion CI Estimation (Q1)
```

```
estimate.prop <- function(p.x, p.n, p.p){
  # sample
  y <- rbinom(n = p.n, size = 1, prob = p.p)
  p <- mean(y)
  n <- p.n

  # output
  wald <- waldInterval.prop(p, n, conf.level = 0.95)
  wilson <- wilsonInterval(p, n, conf.level = 0.95)
  a.wald <- a.waldInterval(p, n, conf.level = 0.95)
  output <- c(p, wald, wilson, a.wald)

  # overall output
  outmirror <- c(p.x, p.n, p.p)
  output <- c(outmirror, output)
  return(output)
}

# Risk Difference CI Estimation (Q2)
estimate.diff <- function(p.x1, p.n1, p.x2, p.n2, p.p1, p.p2){
  # sample
  y1 <- rbinom(n = p.n1, size = 1, prob = p.p1)
  y2 <- rbinom(n = p.n2, size = 1, prob = p.p2)
  p1 <- mean(y1)
  p2 <- mean(y2)
  n1 <- p.n1
  n2 <- p.n2

  # output
  wald <- waldInterval.diff(p1, n1, p2, n2, conf.level = 0.95)
  newcombe <- newcombehybridInterval(p1, n1, p2, n2, conf.level = 0.95)
  agresti <- agresticaffoInterval(p1, n1, p2, n2, conf.level = 0.95)
  output <- c((p2 - p1), wald, newcombe, agresti)

  # overall output
  p.pdiff <- p.p2 - p.p1
  outmirror <- c(p.x1, p.n1, p.x2, p.n2, p.pdiff)
  output <- c(outmirror, output)
  return(output)
}

# Pooled Risk Difference CI Estimation in Meta-analyses (Q3 & Q4)
estimate.meta <- function(p.x1.st1, p.n1.st1, p.x1.st2, p.n1.st2,
                          p.x2.st1, p.n2.st1, p.x2.st2, p.n2.st2,
                          p.p1.st1, p.p1.st2, p.p2.st1, p.p2.st2){
  # samples
  y1.st1 <- rbinom(n = p.n1.st1, size = 1, prob = p.p1.st1)
  y1.st2 <- rbinom(n = p.n1.st2, size = 1, prob = p.p1.st2)
  y2.st1 <- rbinom(n = p.n2.st1, size = 1, prob = p.p2.st1)
  y2.st2 <- rbinom(n = p.n2.st2, size = 1, prob = p.p2.st2)

  x1.st1 <- sum(y1.st1)
  n1.st1 <- length(y1.st1)
  x1.st2 <- sum(y1.st2)
  n1.st2 <- length(y1.st2)
  x2.st1 <- sum(y2.st1)
  n2.st1 <- length(y2.st1)
  x2.st2 <- sum(y2.st2)
  n2.st2 <- length(y2.st2)

  # meta-analysis fixed-effect and random-effects IOV
  meta.IOV <- metabin(c(x2.st1, x2.st2), # nr events group 2
                      c(n2.st1, n2.st2), # nr observations group 2
                      c(x1.st1, x1.st2), # nr events group 1
                      c(n1.st1, n1.st2), # nr observations group 1
                      sm = "RD", method = "Inverse") # methods: risk difference and IOV

  # meta-analysis fixed-effect and random-effects CMH
  meta.CMH <- metabin(c(x2.st1, x2.st2), # nr events group 2
                      c(n2.st1, n2.st2), # nr observations group 2
                      c(x1.st1, x1.st2), # nr events group 1
                      c(n1.st1, n1.st2), # nr observations group 1
                      sm = "RD", method = "MH") # methods: risk difference and CMH

  # meta-analysis with exact fixed-effect model
  st1 <- cbind(x1.st1, x2.st1, n1.st1, n2.st1)
  st2 <- cbind(x1.st2, x2.st2, n1.st2, n2.st2)
```

```
  exacttable <- as.data.frame.matrix(rbind(st1, st2))
  exactInterval <- meta.exact(data = exacttable, type = "risk difference")

  # output
  CMH.output <- c(meta.CMH$TE.fixed, meta.CMH$lower.fixed, meta.CMH$upper.fixed,
                  exactInterval$ci.fixed[1,4], exactInterval$ci.fixed[2,4],
exactInterval$ci.fixed[3,4],
                  meta.CMH$TE.random, meta.CMH$lower.random, meta.CMH$upper.random)
  IOV.output <- c(meta.IOV$TE.fixed, meta.IOV$lower.fixed, meta.IOV$upper.fixed,
                  exactInterval$ci.fixed[1,2], exactInterval$ci.fixed[2,2],
exactInterval$ci.fixed[3,2],
                  meta.IOV$TE.random, meta.IOV$lower.random, meta.IOV$upper.random)
  output <- c(CMH.output, IOV.output)

  # overall output
  p.pdiff <- ((p.p2.st1 - p.p1.st1) + (p.p2.st2 - p.p1.st2))/2
  outmirror <- c(p.x1.st1, p.n1.st1, p.x2.st1, p.n2.st1,
                 p.x1.st2, p.n1.st2, p.x2.st2, p.n2.st2,
                 p.pdiff)
  output <- c(outmirror, output)
  return(output)
}

# Coverage Probability and Average Width Proportion CIs (Q1)
coverage.width.prop <- function(){
  p.p <- thesisoutput.prop$p.p
  is.covered.wald <- (thesisoutput.prop$wald.prop.L<p.p) & (p.p<thesisoutput.prop$wald.prop.U)
# if the true proportion is covered within the CI 1 is returned, else 0
  width.wald <- thesisoutput.prop$wald.prop.U - thesisoutput.prop$wald.prop.L
  is.covered.wilson <- (thesisoutput.prop$wilson.prop.L<p.p) &
(p.p<thesisoutput.prop$wilson.prop.U)
  width.wilson <- thesisoutput.prop$wilson.prop.U - thesisoutput.prop$wilson.prop.L
  is.covered.a.wald <- (thesisoutput.prop$a.wald.prop.L<p.p) &
(p.p<thesisoutput.prop$a.wald.prop.U)
  width.a.wald <- thesisoutput.prop$a.wald.prop.U - thesisoutput.prop$a.wald.prop.L

  cov.wald <- mean(is.covered.wald)
  mean.width.wald <- mean(width.wald)
  cov.wilson <- mean(is.covered.wilson)
  mean.width.wilson <- mean(width.wilson)
  cov.a.wald <- mean(is.covered.a.wald)
  mean.width.a.wald <- mean(width.a.wald)

  output <- list("Coverage Probability Wald CI for Proportions" = cov.wald,
                 "Average width Wald CI for Proportions" = mean.width.wald,
                 "Coverage Probability Wilson CI" = cov.wilson,
                 "Average width Wilson CI" = mean.width.wilson,
                 "Coverage Probability Adjusted Wald CI" = cov.a.wald,
                 "Average width Adjusted Wald CI" = mean.width.a.wald)
  return(output)
}

# Coverage Probability and Average Width Risk Difference CIs (Q2)
coverage.width.diff <- function(){
  p.pdiff <- thesisoutput.diff$p.pdiff
  is.covered.wald <- (thesisoutput.diff$wald.diff.L<p.pdiff) &
(p.pdiff<thesisoutput.diff$wald.diff.U)
  width.wald <- thesisoutput.diff$wald.diff.U - thesisoutput.diff$wald.diff.L
  is.covered.newcombe <- (thesisoutput.diff$newcombe.diff.L<p.pdiff) &
(p.pdiff<thesisoutput.diff$newcombe.diff.U)
  width.newcombe <- thesisoutput.diff$newcombe.diff.U - thesisoutput.diff$newcombe.diff.L
  is.covered.agresti <- (thesisoutput.diff$agresti.diff.L<p.pdiff) &
(p.pdiff<thesisoutput.diff$agresti.diff.U)
  width.agresti <- thesisoutput.diff$agresti.diff.U - thesisoutput.diff$agresti.diff.L

  cov.wald <- mean(is.covered.wald)
  mean.width.wald <- mean(width.wald)
  cov.newcombe <- mean(is.covered.newcombe)
  mean.width.newcombe <- mean(width.newcombe)
  cov.agresti <- mean(is.covered.agresti)
  mean.width.agresti <- mean(width.agresti)

  output <- list("Coverage Probability Wald CI for Risk Differences" = cov.wald,
                 "Average width Wald CI for Risk Differences" = mean.width.wald,
                 "Coverage Probability Newcombe Hybrid CI" = cov.newcombe,
                 "Average width Newcombe Hybrid CI" = mean.width.newcombe,
                 "Coverage Probability Agresti-Caffo CI" = cov.agresti,
```

```
                "Average width Agresti-Caffo CI" = mean.width.agresti)
  return(output)
}

# Coverage Probability and Average Width Pooled Risk Difference CIs in Meta-Analyses (Q3 & Q4)
coverage.width.meta <- function(){
  p.pdiff <- thesisoutput.meta$p.pdiff
  is.covered.fixed.CMH <- (thesisoutput.meta$fixed.CMH.L<p.pdiff) &
(p.pdiff<thesisoutput.meta$fixed.CMH.U)
  width.fixed.CMH <- thesisoutput.meta$fixed.CMH.U - thesisoutput.meta$fixed.CMH.L
  is.covered.fixed.IOV <- (thesisoutput.meta$fixed.IOV.L<p.pdiff) &
(p.pdiff<thesisoutput.meta$fixed.IOV.U)
  width.fixed.IOV <- thesisoutput.meta$fixed.IOV.U - thesisoutput.meta$fixed.IOV.L
  is.covered.exact.CMH <- (thesisoutput.meta$exact.CMH.L<p.pdiff)  &
(p.pdiff<thesisoutput.meta$exact.CMH.U)
  width.exact.CMH <- thesisoutput.meta$exact.CMH.U - thesisoutput.meta$exact.CMH.L
  is.covered.exact.IOV <- (thesisoutput.meta$exact.IOV.L<p.pdiff)  &
(p.pdiff<thesisoutput.meta$exact.IOV.U)
  width.exact.IOV <- thesisoutput.meta$exact.IOV.U - thesisoutput.meta$exact.IOV.L
  is.covered.random.CMH <- (thesisoutput.meta$random.CMH.L<p.pdiff)  &
(p.pdiff<thesisoutput.meta$random.CMH.U)
  width.random.CMH <- thesisoutput.meta$random.CMH.U - thesisoutput.meta$random.CMH.L
  is.covered.random.IOV <- (thesisoutput.meta$random.IOV.L<p.pdiff) &
(p.pdiff<thesisoutput.meta$random.IOV.U)
  width.random.IOV <- thesisoutput.meta$random.IOV.U - thesisoutput.meta$random.IOV.L

  cov.fixed.CMH <- mean(is.covered.fixed.CMH)
  mean.width.fixed.CMH <- mean(width.fixed.CMH)
  cov.fixed.IOV <- mean(is.covered.fixed.IOV)
  mean.width.fixed.IOV <- mean(width.fixed.IOV)
  cov.exact.CMH <- mean(is.covered.exact.CMH)
  mean.width.exact.CMH <- mean(width.exact.CMH)
  cov.exact.IOV <- mean(is.covered.exact.IOV)
  mean.width.exact.IOV <- mean(width.exact.IOV)
  cov.random.CMH <- mean(is.covered.random.CMH)
  mean.width.random.CMH <- mean(width.random.CMH)
  cov.random.IOV <- mean(is.covered.random.IOV)
  mean.width.random.IOV <- mean(width.random.IOV)

  output <- list("Coverage Probability Fixed Effect CI - CMH" = cov.fixed.CMH,
                 "Average width Fixed Effect CI - CMH" = mean.width.fixed.CMH,
                 "Coverage Probability Fixed Effect CI - IOV" = cov.fixed.IOV,
                 "Average width Fixed Effect CI - IOV" = mean.width.fixed.IOV,
                 "Coverage Probability Tian CI - CMH" = cov.exact.CMH,
                 "Average width Tian CI - CMH"= mean.width.exact.CMH,
                 "Coverage Probability Tian CI - IOV" = cov.exact.IOV,
                 "Average width Tian CI - IOV" = mean.width.exact.IOV,
                 "Coverage Probability Random Effects - CMH" = cov.random.CMH,
                 "Average width Random Effects CI - CMH" = mean.width.random.CMH,
                 "Coverage Probability Random Effects - IOV" = cov.random.IOV,
                 "Average width Random Effects CI - IOV" = mean.width.random.IOV)
  return(output)
}

######## 2. SIMULATIONS ########
# Simulation Parameters
p.rep <- 1000 # number of simulations Q1 & Q2
p.rep.meta <- 10 # number of simulations Q3 & Q4

p.p <- c(.01, .50, .99) # true proportions
p.n <- c(100, 5000) # true number of observations
design.prop <- expand.grid(p.p = p.p, p.n = p.n)
design.prop$p.x <- design.prop$p.p*design.prop$p.n # add true number of events
design.prop # design matrix of simulation parameters for proportions (Q1)

p.p1 <- c(.01, .50, .99) # true proportions group 1
p.p2 <- c(.01, .50, .99) # true proportions group 2
p.n1 <- c(100, 5000) # true number of observations group 1
p.n2 <- c(100, 5000) # true number of observations group 2
design.diff <- expand.grid(p.p1 = p.p1, p.p2 = p.p2,
                           p.n1 = p.n1, p.n2 = p.n2)
design.diff$p.x1 <- design.diff$p.p1*design.diff$p.n1 # add true number of events group 1
design.diff$p.x2 <- design.diff$p.p2*design.diff$p.n2 # add true number of events group 2
design.diff # design matrix of simulation parameters for risk differences (Q2)

p.p1.st1 <- c(.00, .01) # true number of events group 1 study 1
p.p1.st2 <- c(.00, .01) # true proportions group 1 study 2
```

```
p.p2.st1 <- c(.00, .01) # true proportions group 2 study 1
p.p2.st2 <- c(.00, .01) # true proportions group 2 study 2
p.n1.st1 <- c(5000) # true number of observations group 1 study 1
p.n1.st2 <- c(5000) # true number of observations group 1 study 2
p.n2.st1 <- c(5000) # true number of observations group 2 study 1
p.n2.st2 <- c(5000) # true number of observations group 2 study 2
design.meta <- expand.grid(p.p1.st1 = p.p1.st1, p.p1.st2 = p.p1.st2, p.p2.st1 = p.p2.st1,
p.p2.st2 = p.p2.st2,
                           p.n1.st1 = p.n1.st1, p.n1.st2 = p.n1.st2, p.n2.st1 = p.n2.st1,
p.n2.st2 = p.n2.st2)
design.meta$p.x1.st1 <- design.meta$p.p1.st1*design.meta$p.n1.st1 # add true number of events
group 1 study 1
design.meta$p.x1.st2 <- design.meta$p.p1.st2*design.meta$p.n1.st2 # add true number of events
group 1 study 2
design.meta$p.x2.st1 <- design.meta$p.p2.st1*design.meta$p.n2.st1 # add true number of events
group 2 study 1
design.meta$p.x2.st2 <- design.meta$p.p2.st2*design.meta$p.n2.st2 # add true number of events
group 2 study 2
design.meta # design matrix of simulation parameters for pooled risk differences in meta-
analyses (Q3 & Q4)

# Simulations Loop for Proportions (Q1)
t1.prop <- Sys.time()
for(r in 1:nrow(design.prop)){
  cellData.prop <- NULL
  for(k in 1:p.rep){
    set.seed(1000*r + k)
    repData.prop <- estimate.prop(design.prop[r,3], design.prop[r,2], design.prop[r,1])
    cellData.prop <- rbind(cellData.prop, repData.prop)
  }
  rownames(cellData.prop) <- NULL
  colnames(cellData.prop) <- c("p.x", "p.n", "p.p", "p",
                               "wald.prop.L", "wald.prop.U",
                               "wilson.prop.L", "wilson.prop.U",
                               "a.wald.prop.L", "a.wald.prop.U")
  outfile <- paste("thesisoutput.prop",r,".csv",sep="")
  write.csv(cellData.prop, file=outfile)
}
duration.prop <- Sys.time() - t1.prop
duration.prop

# Simulations Loop for Risk Differences (Q2)
t1.diff <- Sys.time()
for(r in 1:nrow(design.diff)){
  cellData.diff <- NULL
  for(k in 1:p.rep){
    set.seed(1000*r + k)
    repData.diff <- estimate.diff(design.diff[r,5], design.diff[r,3], design.diff[r,6],
design.diff[r,4], design.diff[r,1], design.diff[r,2])
    cellData.diff <- rbind(cellData.diff, repData.diff)
  }
  rownames(cellData.diff) <- NULL
  colnames(cellData.diff) <- c("p.x1", "p.n1","p.x2", "p.n2", "p.pdiff", "pdiff",
                               "wald.diff.L", "wald.diff.U",
                               "newcombe.diff.L", "newcombe.diff.U",
                               "agresti.diff.L", "agresti.diff.U")
  outfile <- paste("thesisoutput.diff", r, ".csv", sep = "")
  write.csv(cellData.diff, file = outfile)
}
duration.diff <- Sys.time() - t1.diff
duration.diff

# Simulations Loop for Pooled Risk Differences in Meta-analyses (Q3 & Q4)
t1.meta <- Sys.time()
for(r in 1:nrow(design.meta)){
  cellData.meta <- NULL
  for(k in 1:p.rep.meta){
    set.seed(1000*r + k)
    repData.meta <- estimate.meta(design.meta[r,9], design.meta[r,5], design.meta[r,10],
design.meta[r,6],
                                  design.meta[r,11], design.meta[r,7], design.meta[r,12],
design.meta[r,8],
                                  design.meta[r,1], design.meta[r,2], design.meta[r,3],
design.meta[r,4])
    cellData.meta <- rbind(cellData.meta, repData.meta)
  }
  rownames(cellData.meta) <- NULL
```

```
  colnames(cellData.meta) <- c("p.x1.st1", "p.n1.st1", "p.x2.st1", "p.n2.st1",
                               "p.x1.st2", "p.n1.st2", "p.x2.st2", "p.n2.st2", "p.pdiff",
                               "fixed.pdiff.CMH", "fixed.CMH.L", "fixed.CMH.U",
                               "exact.pdiff.CMH", "exact.CMH.L", "exact.CMH.U",
                               "random.pdiff.CMH", "random.CMH.L", "random.CMH.U",
                               "fixed.pdiff.IOV", "fixed.IOV.L", "fixed.IOV.U",
                               "exact.pdiff.IOV", "exact.IOV.L", "exact.IOV.U",
                               "random.pdiff.IOV", "random.IOV.L", "random.IOV.U")
  outfile <- paste("thesisoutput.meta", r, ".csv",sep = "")
  write.csv(cellData.meta, file = outfile)
}
duration.meta <- Sys.time() - t1.meta
duration.meta

######## 3. WRITE SIMULTATIONS TO FILE ########
# Proportions (Q1)
totoutput <- NULL
for(r in 1:nrow(design.prop)){
  filename <- paste("thesisoutput.prop", r, ".csv", sep = "")
  rownr <- rep(r, p.rep)
  temp <- cbind(rownr, read.csv(filename))
  totoutput <- rbind(totoutput, temp)
}
write.csv(totoutput, file = "thesisoutput.prop.csv")

# Risk Differences (Q2)
totoutput <- NULL
for(r in 1:nrow(design.diff)){
  filename <- paste("thesisoutput.diff", r, ".csv", sep = "")
  rownr <- rep(r, p.rep)
  temp <- cbind(rownr, read.csv(filename))
  totoutput <- rbind(totoutput, temp)
}
write.csv(totoutput, file = "thesisoutput.diff.csv")

# Pooled Risk Differences in Meta-analyses (Q3 & Q4)
totoutput <- NULL
for(r in 1:nrow(design.meta)){
  filename <- paste("thesisoutput.meta", r, ".csv", sep = "")
  rownr <- rep(r, p.rep.meta)
  temp <- cbind(rownr, read.csv(filename))
  totoutput <- rbind(totoutput, temp)
}
write.csv(totoutput, file = "thesisoutput.meta.csv")

######## 4. COVERAGE PROBABILITIES AND AVERAGE WIDTHS ########
thesisoutput.prop <- read.csv("thesisoutput.prop.csv", row.names = 1, header = T, sep = ",")
thesisoutput.diff <- read.csv("thesisoutput.diff.csv", row.names = 1, header = T, sep = ",")
thesisoutput.meta <- read.csv("thesisoutput.meta.csv", row.names = 1, header = T, sep = ",")
results.prop <- coverage.width.prop() # Q1
results.diff <- coverage.width.diff() # Q2
results.meta <- coverage.width.meta() # Q3 & Q4

######## 5. PLOTS ########
# Proportions (Q1)
par(mfrow = c(2,2))

conditions <- unique(thesisoutput.prop[, c(1,3:5)]) # set conditions
i.condition <- 1 # set condition here: in thesis 1 and 6
cond.sel <- conditions[i.condition, ]
data.sel <- thesisoutput.prop[thesisoutput.prop$rownr == cond.sel$rownr, ]
data.sel <- data.sel[1:10, ] # only pick first 10 simulations
nm <- 3 # number of methods
nreps <- nrow(data.sel)
colvec <- brewer_pal("qual", 2)(nm) # pretty colours
y.pos <- seq(nreps, (1/nm), - (1/nm)) # y positions for plotting
y.pos.list <- split(y.pos, rep(1:nm, each = nreps))
ylims <- c((1/nm), nreps) # range for y-axis
cols.results <- 5:12 # define the columns for plotting
xlims <- range(data.sel[, cols.results], na.rm = T) # range x-axis = observed full range
plot(0, xlim = xlims, ylim = ylims, type = "n", xlab = "true proportion", ylab = "", yaxt =
"n", main = expression('CI-methods for proportions/risk differences ('*italic(n)*' = 100)'))
axis(2, at = c(sapply(y.pos.list, min) - .5/nm, nreps + .5/nm), labels = NA)
axis(2, at = sapply(y.pos.list, mean), labels = c("wald", "wilson", "a.wald"), tick = F, col =
colvec) # adjust manually (labels of the methods)
points(rep(data.sel$p, times = nm), y.pos, pc = 16)
```

```
segments(data.sel$wald.prop.L, y.pos.list[[1]], data.sel$wald.prop.U, y.pos.list[[1]], col =
colvec[1], lwd = 2)
segments(data.sel$wilson.prop.L, y.pos.list[[2]], data.sel$wilson.prop.U, y.pos.list[[2]],col
= colvec[2],lwd = 2)
segments(data.sel$a.wald.prop.L, y.pos.list[[3]], data.sel$a.wald.prop.U, y.pos.list[[3]], col
= colvec[3], lwd = 2)
abline(v = cond.sel$p.p, lwd = 2, lty = 1) # add true proportion

i.condition <- 6 # set condition here: in thesis 1 and 6
cond.sel <- conditions[i.condition, ]
data.sel <- thesisoutput.prop[thesisoutput.prop$rownr == cond.sel$rownr, ]
xlims <- range(data.sel[, cols.results], na.rm = T) # range x-axis = observed full range
plot(0, xlim = xlims, ylim = ylims, type = "n", xlab = "true proportion", ylab = "", yaxt =
"n", main = expression('CI-methods for proportions/risk differences ('*italic(n)*' = 5000)'))
axis(2, at = c(sapply(y.pos.list, min) - .5/nm, nreps + .5/nm), labels = NA)
axis(2, at = sapply(y.pos.list, mean), labels = c("wald", "wilson", "a.wald"), tick = F, col =
colvec) # adjust manually (labels of the methods)
points(rep(data.sel$p, times = nm), y.pos, pc = 16)
segments(data.sel$wald.prop.L, y.pos.list[[1]], data.sel$wald.prop.U, y.pos.list[[1]], col =
colvec[1], lwd = 2)
segments(data.sel$wilson.prop.L, y.pos.list[[2]], data.sel$wilson.prop.U, y.pos.list[[2]],col
= colvec[2],lwd = 2)
segments(data.sel$a.wald.prop.L, y.pos.list[[3]], data.sel$a.wald.prop.U, y.pos.list[[3]], col
= colvec[3], lwd = 2)
abline(v = cond.sel$p.p, lwd = 2, lty = 1) # add true proportion

# Risk Differences (Q2)
conditions <- unique(thesisoutput.diff[, c(1,3:7)])
i.condition <- 1 # set condition here: in thesis 1 and 34
cond.sel <- conditions[i.condition, ]
data.sel <- thesisoutput.diff[thesisoutput.diff$rownr == cond.sel$rownr, ]
data.sel <- data.sel[1:10, ] # only pick first 10 simulations
cols.results <- 7:14 # define the columns for plotting
xlims <- range(data.sel[, cols.results], na.rm = T) # range x-axis = observed full range
plot(0, xlim = xlims, ylim = ylims, type = "n", xlab = "true risk difference", ylab = "", yaxt
= "n")
axis(2, at = c(sapply(y.pos.list, min) - .5/nm, nreps + .5/nm), labels = NA)
axis(2, at = sapply(y.pos.list, mean), labels = c("wald", "newc", "agr"), tick = F, col =
colvec)
points(rep(data.sel$pdiff, times = nm), y.pos, pc = 16)
segments(data.sel$wald.diff.L, y.pos.list[[1]], data.sel$wald.diff.U, y.pos.list[[1]], col =
colvec[1], lwd = 2)
segments(data.sel$newcombe.diff.L, y.pos.list[[2]], data.sel$newcombe.diff.U,
y.pos.list[[2]],col = colvec[2], lwd = 2)
segments(data.sel$agresti.diff.L, y.pos.list[[3]], data.sel$agresti.diff.U, y.pos.list[[3]],
col = colvec[3], lwd = 2)
abline(v = cond.sel$p.pdiff, lwd = 2, lty = 1) # add true risk difference

i.condition <- 34 # set condition here: in thesis 1 and 34
cond.sel <- conditions[i.condition, ]
data.sel <- thesisoutput.diff[thesisoutput.diff$rownr == cond.sel$rownr, ]
data.sel <- data.sel[1:10, ] # only pick first 10 simulations
xlims <- range(data.sel[, cols.results], na.rm = T) # range x-axis = observed full range
plot(0, xlim = xlims, ylim = ylims, type = "n", xlab = "true risk difference", ylab = "", yaxt
= "n")
axis(2, at = c(sapply(y.pos.list, min) - .5/nm, nreps + .5/nm), labels = NA)
axis(2, at = sapply(y.pos.list, mean), labels = c("wald", "newc", "agr"), tick = F, col =
colvec)
points(rep(data.sel$pdiff, times = nm), y.pos, pc = 16)
segments(data.sel$wald.diff.L, y.pos.list[[1]], data.sel$wald.diff.U, y.pos.list[[1]], col =
colvec[1], lwd = 2)
segments(data.sel$newcombe.diff.L, y.pos.list[[2]], data.sel$newcombe.diff.U,
y.pos.list[[2]],col = colvec[2], lwd = 2)
segments(data.sel$agresti.diff.L, y.pos.list[[3]], data.sel$agresti.diff.U, y.pos.list[[3]],
col = colvec[3], lwd = 2)
abline(v = cond.sel$p.pdiff, lwd = 2, lty = 1) # add true risk difference

# Pooled Risk Differences in Meta-analyses (Q3 & Q4)
par(mfrow = c(2,2))

conditions <- unique(thesisoutput.meta[, c(1,3:11)])
i.condition <- 1 # set condition here: in thesis 1 and 13
cond.sel <- conditions[i.condition, ]
data.sel <- thesisoutput.meta[thesisoutput.meta$rownr == cond.sel$rownr, ]
data.sel <- data.sel[1:10, ] # only pick first 10 simulations
nm <- 3 # number of methods
nreps <- nrow(data.sel)
```

```
colvec <- brewer_pal("qual", 2)(nm) # pretty colours
y.pos <- seq(nreps, (1/nm), - (1/nm)) # y-positions for plotting
y.pos.list <- split(y.pos, rep(1:nm, each = nreps))
ylims <- c((1/nm), nreps) # range for y-axis
cols.results <- 11:29 # define the columns for plotting
xlims <- range(data.sel[, cols.results], na.rm = T) # range x-axis = observed full range
plot(0, xlim = xlims, ylim = ylims, type = "n", xlab = "true pooled risk difference", ylab =
"", yaxt = "n", main = expression('CI-methods for pooled risk differences (CMH)'))
axis(2, at = c(sapply(y.pos.list, min) - .5/nm, nreps + .5/nm), labels = NA)
axis(2, at = sapply(y.pos.list, mean), labels = c("fixed", "tian", "random"), tick = F, col =
colvec)
points(data.sel$fixed.pdiff.CMH, y.pos.list[[1]], pc = 16)
points(data.sel$exact.pdiff.CMH, y.pos.list[[2]], pc = 16)
points(data.sel$random.pdiff.CMH, y.pos.list[[3]], pc = 16)
segments(data.sel$fixed.CMH.L, y.pos.list[[1]], data.sel$fixed.CMH.U, y.pos.list[[1]], col =
colvec[1], lwd = 2)
segments(data.sel$exact.CMH.L, y.pos.list[[2]], data.sel$exact.CMH.U, y.pos.list[[2]], col =
colvec[2], lwd = 2)
segments(data.sel$random.CMH.L, y.pos.list[[3]], data.sel$random.CMH.U, y.pos.list[[3]],col =
colvec[3], lwd = 2)
abline(v = cond.sel$p.pdiff, lwd = 2, lty = 1) # add true pooled risk difference

i.condition <- 1 # set condition here: in thesis 1 and 13
cond.sel <- conditions[i.condition, ]
data.sel <- thesisoutput.meta[thesisoutput.meta$rownr == cond.sel$rownr, ]
data.sel <- data.sel[1:10, ] # only pick first 10 simulations
xlims <- range(data.sel[, cols.results], na.rm = T) # range x-axis = observed full range
plot(0, xlim = xlims, ylim = ylims, type = "n", xlab = "true pooled risk difference", ylab =
"", yaxt = "n", main = expression('CI-methods for pooled risk differences (IOV)'))
axis(2, at = c(sapply(y.pos.list, min) - .5/nm, nreps + .5/nm), labels = NA)
axis(2, at = sapply(y.pos.list, mean), labels = c("fixed", "tian", "random"), tick = F, col =
colvec)
points(data.sel$fixed.pdiff.IOV, y.pos.list[[1]], pc = 16)
points(data.sel$exact.pdiff.IOV, y.pos.list[[2]], pc = 16)
points(data.sel$random.pdiff.IOV, y.pos.list[[3]], pc = 16)
segments(data.sel$fixed.IOV.L, y.pos.list[[1]], data.sel$fixed.IOV.U, y.pos.list[[1]], col =
colvec[1], lwd = 2)
segments(data.sel$exact.IOV.L, y.pos.list[[2]], data.sel$exact.IOV.U, y.pos.list[[2]], col =
colvec[2], lwd = 2)
segments(data.sel$random.IOV.L, y.pos.list[[3]], data.sel$random.IOV.U, y.pos.list[[3]],col =
colvec[3], lwd = 2)
abline(v = cond.sel$p.pdiff, lwd = 2, lty = 1) # add true pooled risk difference

i.condition <- 13 # set condition here: in thesis 1 and 13
cond.sel <- conditions[i.condition, ]
data.sel <- thesisoutput.meta[thesisoutput.meta$rownr == cond.sel$rownr, ]
xlims <- range(data.sel[, cols.results], na.rm = T) # range x-axis = observed full range
plot(0, xlim = xlims, ylim = ylims, type = "n", xlab = "true pooled risk difference", ylab =
"", yaxt = "n")
axis(2, at = c(sapply(y.pos.list, min) - .5/nm, nreps + .5/nm), labels = NA)
axis(2, at = sapply(y.pos.list, mean), labels = c("fixed", "tian", "random"), tick = F, col =
colvec)
points(data.sel$fixed.pdiff.CMH, y.pos.list[[1]], pc = 16)
points(data.sel$exact.pdiff.CMH, y.pos.list[[2]], pc = 16)
points(data.sel$random.pdiff.CMH, y.pos.list[[3]], pc = 16)
segments(data.sel$fixed.CMH.L, y.pos.list[[1]], data.sel$fixed.CMH.U, y.pos.list[[1]], col =
colvec[1], lwd = 2)
segments(data.sel$exact.CMH.L, y.pos.list[[2]], data.sel$exact.CMH.U, y.pos.list[[2]], col =
colvec[2], lwd = 2)
segments(data.sel$random.CMH.L, y.pos.list[[3]], data.sel$random.CMH.U, y.pos.list[[3]],col =
colvec[3], lwd = 2)
abline(v = cond.sel$p.pdiff, lwd = 2, lty = 1) # add true pooled risk difference

i.condition <- 13 # set condition here: in thesis 1 and 13
cond.sel <- conditions[i.condition, ]
data.sel <- thesisoutput.meta[thesisoutput.meta$rownr == cond.sel$rownr, ]
xlims <- range(data.sel[, cols.results], na.rm = T) # range x-axis = observed full range
plot(0, xlim = xlims, ylim = ylims, type = "n", xlab = "true pooled risk difference", ylab =
"", yaxt = "n")
axis(2, at = c(sapply(y.pos.list, min) - .5/nm, nreps + .5/nm), labels = NA)
axis(2, at = sapply(y.pos.list, mean), labels = c("fixed", "tian", "random"), tick = F, col =
colvec)
points(data.sel$fixed.pdiff.IOV, y.pos.list[[1]], pc = 16)
points(data.sel$exact.pdiff.IOV, y.pos.list[[2]], pc = 16)
points(data.sel$random.pdiff.IOV, y.pos.list[[3]], pc = 16)
segments(data.sel$fixed.IOV.L, y.pos.list[[1]], data.sel$fixed.IOV.U, y.pos.list[[1]], col =
colvec[1], lwd = 2)
```

```
segments(data.sel$exact.IOV.L, y.pos.list[[2]], data.sel$exact.IOV.U, y.pos.list[[2]], col =
colvec[2], lwd = 2)
segments(data.sel$random.IOV.L, y.pos.list[[3]], data.sel$random.IOV.U, y.pos.list[[3]],col =
colvec[3], lwd = 2)
abline(v = cond.sel$p.pdiff, lwd = 2, lty = 1) # add true pooled risk difference
```