2022-04-18

A Permutation-Based Confidence Distribution for Rare-Event Meta-Analysis

Travis Andersen
Brigham Young University

A Permutation-Based Confidence Distribution for Rare-Event Meta-Analysis

Travis Andersen

A selected thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

Brinley N. Zabriskie, Chair
Gilbert W. Fellingham
Natalie J. Blades

Department of Statistics

Brigham Young University

ABSTRACT

A Permutation-Based Confidence Distribution for Rare-Event Meta-Analysis

Travis Andersen
Department of Statistics, BYU
Master of Science

Confidence distributions (CDs), which provide evidence across all levels of significance, are receiving increasing attention, especially in meta-analysis. Meta-analyses allow independent study results to be combined to produce one overall conclusion and are particularly useful in public health and medicine. For studies with binary outcomes that are rare, many traditional meta-analysis methods often fail (Sutton et al. 2002; Efthimiou 2018; Liu et al. 2018; Liu 2019; Hunter and Schmidt 2000; Kontopantelis et al. 2013). Zabriskie et al. (2021b) develop a permutation-based method to analyze such data when study treatment effects vary beyond what is expected by chance. In this work, we prove that this method can be considered a CD. Additionally, we develop two new metrics to assess a CD's relative performance.

Keywords: confidence curve, confidence interval, small sample sizes, sparse data

ACKNOWLEDGMENTS

First and foremost, I would like to thank and acknowledge my advisor, Dr. Zabriskie. Her continual guidance, direction, and editing support have been invaluable in the creation of this paper. I would also like to acknowledge the contributions of the rest of my committee, whose comments and suggestions have likewise greatly benefited this paper. Finally, I would like to thank my friends and family, in particular my parents and my wife, who have supported me throughout my research as well as my whole life.

CONTENTS

CHAPTER 1

INTRODUCTION

Meta-analyses, statistical procedures used to combine information from independent studies, have become very popular to conduct, especially in public health and medicine (Sutton and Higgins 2008). Meta-analyses are often considered the gold standard for systematic reviews, and they can be especially useful when individual studies provide evidence about binary events that are rare or adverse. In these cases, or when sample sizes are small, traditional meta-analysis methods can perform poorly (Sutton et al. 2002; Efthimiou 2018; Liu et al. 2018; Liu 2019; Hunter and Schmidt 2000; Kontopantelis et al. 2013). The parameter of interest is generally the log odds ratio, which is assumed to follow a normal distribution due to its asymptotic distribution. This can be a faulty assumption, especially when combining information from studies with small samples sizes, rare events, or meta-analyses with few studies (Sutton et al. 2002; Efthimiou 2018). Rare events can also result in zero observed events in one or both treatment arms. Many traditional methods require the use of a continuity correction, a small numerical adjustment made to the data, in order for zero-event studies to be included in the analysis. However, this is an arbitrary stopgap, and many have argued against its use (Liu et al. 2018; Efthimiou 2018; Liu 2019).

Further complicating rare-event meta-analyses is the possibility of non-negligible heterogeneity, differences in the study treatment effects beyond that which is due to chance. There are two general frameworks for performing a meta-analysis: fixed-effect and random-effects. Under a fixed-effect framework, the study-specific effects are assumed to be estimating one common treatment effect. Under a random-effects framework, the study-specific effects are assumed to come from a common distribution of treatment effects. The variance of this distribution of treatment effects is known as the heterogeneity variance, often denoted as τ^2, and is an additional source of variability that must be incorporated into a meta-analysis for accurate results. In practice, heterogeneity is often introduced by the different conditions across studies, such as variation in treatment administration. Fixed-effect models do not account for this additional source of variability, so unless homogeneity is a reasonable assumption, random-effects models should be used.

In summary, both rare events and non-negligible heterogeneity can cause traditional meta-analysis methods to fail (Sutton et al. 2002; Efthimiou 2018; Liu et al. 2018; Liu 2019; Hunter and Schmidt 2000; Kontopantelis et al. 2013). Importantly, these types of datasets occur often in public health and medicine, making it critical to develop methods that can better combine this type of data (Hunter and Schmidt 2000; Kontopantelis et al. 2013). One method developed to better analyze meta-analysis data with rare events and heterogeneity is a method by Zabriskie et al. (2021b). They develop a permutation-based approach that performs well in heterogeneous, rare-event settings. Instead of assuming an asymptotic normal assumption, this method uses the exact, permutation-based distribution of the data. It also allows for studies with zero observed events to be included in the analysis without relying on an artificial continuity correction. They illustrate how this method outperforms other methods in preserving the nominal level of significance.

Other methods that have been developed, in part to address these issues, are methods that take advantage of all available information from each study by using confidence distributions (CDs). While traditional meta-analysis methods combine point estimates from each study to produce one overall p-value and confidence interval (CI) at a certain level of significance, CDs provide this information, and more, for all levels of significance. Recently, there has been an important discussion on the use of p-values and CIs to determine whether to accept or reject the null hypothesis (Wasserstein and Lazar 2016, Ioannidis 2019). p-values and CIs can be used to dichotomize results, which is likely undesireable as the conclusion generated by these approaches depends on the chosen level of significance. CDs can bypass these issues by providing a comprehensive overview of the available inference for a parameter across all levels of significance.

Two functions often used in conjunction with CDs are confidence densities and confidence curves (CVs), see Figure 1.1 (Infanger and Schmidt-Trucksäss 2019). Confidence densities and CVs are both functions of the CD and can be used as alternative ways to display information contained in the CD. As seen in Figure 1.1, a CD is a cumulative density function and a confidence density is a probability density function. A CV is used to easily visualize CIs and extract p-values and it is generally used as the medium for displaying results instead of the CD itself. As illustrated in Figure 1.1, to obtain a $100(1-\alpha)\%$ CI, one could simply draw a horizontal line at α on the y-axis of the CV and then take the two values where the curve intersects that line as the lower and upper bounds of the CI. A CV readily provides a visual of the CIs at all levels of significance. Further, the two-sided p-value is easily seen as the height of the curve at the null value. Another useful metric provided in CV plots is the counternull, displayed on the plot by the open circle, which is the non-null parameter value that is supported by the data just as much as the null value. The counternull can help researchers evaluate practical significance in addition to the statistical significance of results. The CV also displays the statistical power conditional on the true value of the parameter. This can be found by finding the height of the CV at any null value of the parameter, and subtracting it from one.

In meta-analyses, CDs are often created for each individual study and then combined to create one overall CD (Singh et al. 2005). Zabriskie et al. (2021a) summarize and compare various CD approaches in meta-analysis, and we focus on two of those approaches here: one developed by Liu et al. (2014), the other by Cunen and Hjort (2021). The approach by Liu et al. (2014) was developed for binary outcome meta-analyses under the fixed-effect framework with the log odds ratio as the parameter of interest. First, the mid-p adaptation of Fisher's exact test is used to compute a CV for each individual study. The inverse cumulative distribution function of the standard normal distribution is then used to combine the

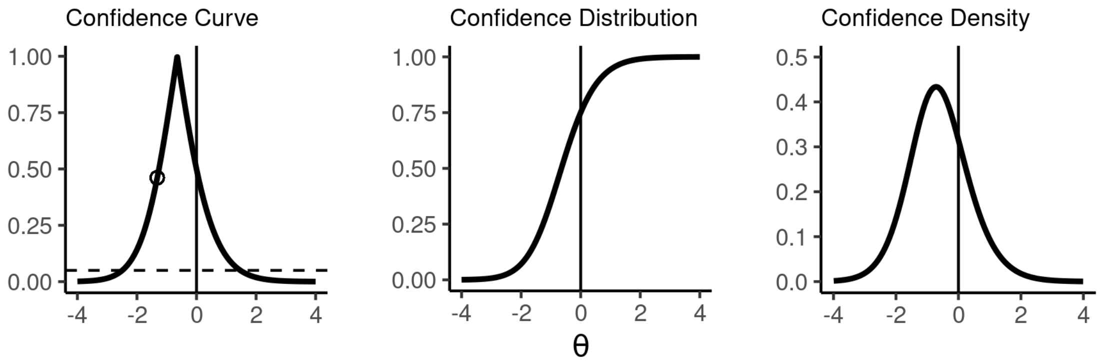

Figure 1.1: CV, CD, and confidence density for an example dataset. The vertical line at 0 indicates the null value of the parameter. The open circle on the CV is the counternull and the dashed horizontal line drawn at 0.05 marks the 95% CI, where the lower and upper bounds are determined by the two points on the x-axis where the dashed line intersects the curve.

individual CVs into one CV, where each study is weighted according to its size and probabilities of an event. This process produces a final combined CD that is only approximately exact since the combined CD may not be exactly normal when the weights are estimated from the data.

The approach developed by Cunen and Hjort (2021) uses a framework denoted by II-CC-FF. This framework has a wide range of varied uses, which extend beyond the meta-analysis settings. Even within the realm of meta-analysis, there are various strategies under this framework for combining the individual studies. The general approach is to start by creating a CD and then obtaining the log-likelihood function for each study. The log-likelihood functions can then be combined under either the fixed-effect or random-effects framework, typically using profiling and the Wilks chi-squared approximation. Under the fixed-effect framework, the log-likelihood functions are summed, and a combined CV is formed. Under the random-effects framework, a function of the log likelihood functions is integrated, from which the combined CV is formed. This function can take on many forms; see Cunen and Hjort (2021) for more details on the function they chose, which incorporates the standard normal probability density function.

When comparing these CD methods via a simulation study, Zabriskie et al. (2021a), and others up until this point, consider traditional metrics, such as the average coverage of 95% CIs. Having to choose a significance level in order to evaluate the performance of a CD, where a main benefit is not having to select a significance level, is less than ideal. How does performance change for 90% or 99% CIs, for example? Clearly, better metrics are needed to summarize a CD's performance without the need to pre-specify a significance level.

Nonetheless, Zabriskie et al. (2021a) show that, when $\alpha = 0.05$, the Liu et al. (2014) method, developed under the fixed-effect framework, predictably performs poorly when meta-analysis data is heterogeneous and events are rare. The Cunen and Hjort (2021) random-effects method, on the other hand, performs well in these settings. While both methods utilize CDs, they were not designed to specifically handle rare events. Other methods have been developed that are designed to analyze rare-event data, but that do not utilize CDs. One of these methods, which we will focus on in this paper, is the permutation-based method of Zabriskie et al. (2021b), introduced previously. Along with evaluating how well these methods preserve the nominal level of significance, we are also interested in how well these methods avoid type II error. We will evaluate the power of these three methods in the simulation study.

There are two main goals of this paper. Our first goal is to develop a CD for the Zabriskie et al. (2021b) method. This will allow us to benefit from a method specifically designed for heterogeneous, rare-event data and one in which a CD is used to provide a wealth of evidence across all levels of significance (and not just at $\alpha = 0.05$, for instance). Our second goal is to propose two new metrics for evaluating the performance of CDs that do not require pre-specifying a significance level.

In Chapter 2, we achieve our first goal by proving the Zabriskie et al. (2021b) method can be considered a CD, enabling researchers to view the permutation-based method results at any level of significance. In this chapter, we also outline how we achieve our second goal of creating new ways of assessing a CD's overall performance across all levels of significance simultaneously. To illustrate the use of these new metrics to assess a CD's relative performance, we apply them in a simulation study in Chapter 3. To highlight the benefits of extending the Zabriskie et al. (2021b) method to become a CD, we present a case study in Chapter 4. Lastly, we end with a discussion in Chapter 5.

CHAPTER 2

METHODS

In this chapter, we begin by providing definitions of a CD. A brief overview of the Zabriskie et al. (2021b) method will then be provided. Next, for the first goal of this work, we prove that this method can be considered a CD. Finally, we present two new metrics for evaluating the overall performance of CDs for the second goal of this work.

2.1 Definition of a CD

Generally when performing statistical inference, a sample of data $\mathbf{z}$ of size n is drawn from a sample set $\mathcal{Z}$ to learn more about some unknown parameter θ with parameter space Θ. Singh et al. (2005) define a CD for θ as a function of the data $\phi(\theta) = \phi(\mathbf{z}, \theta)$ on $\mathcal{Z} \times \Theta \to [0, 1]$ such that,

1. for each $\mathbf{z}$ in the sample set space $\mathcal{Z}$, $\phi(\theta)$ is a continuous cumulative distribution function in the parameter space Θ; and
2. at the true parameter value $\theta = \theta_0$, $\phi(\theta_0) = \phi(\mathbf{z}, \theta_0)$, as a function of the sample set $\mathbf{z}$, has a uniform distribution $U(0, 1)$.

Singh et al. (2005) also provide an alternative method for proving a function is a CD. Let (θ_{-}, ∞) be a one-sided $100(1 - \alpha)\%$ CI for θ, where the lower confidence bound θ_{-} is a function of the level of significance, α. If, for every $\alpha \in (0, 1)$ and $\theta \in \Theta$, $\theta_{-}(\alpha)$ is continuous and increasing in α for each sample $\mathbf{z}$, then $\theta_{-}^{-1}(\cdot) = \phi(\theta)$ is a CD.

2.2 The Permutation-Based Method

Zabriskie et al. (2021b) propose a permutation-based method, founded on conditional logistic regression, that preserves the Type I error rate more effectively than other methods for heterogeneous meta-analyses. Here, we present a high-level overview of their method and focus on inference once the permutational distribution is obtained. As with traditional conditional logistic regression, nuisance parameters are eliminated via conditioning on their sufficient statistics; we denote the sufficient statistics of the nuisance parameters with a dot symbol, $\cdot$, for simplicity. After conditioning, the treatment effect θ is the only remaining unknown parameter, and the distribution of θ's sufficient statistic, $t(\mathbf{z})$ (the total number of observed events in the treatment group across all studies), is obtained via permutations. With this permutational distribution, traditional p-values and CIs can be obtained. Here, we focus on datasets where the conditional maximum likelihood estimate of θ can be used for inference; this happens when the observed value of θ's sufficient statistic $t(\mathbf{z}_{\text{obs}})$ satisfies $\min(t(\mathbf{z})) < t(\mathbf{z}_{\text{obs}}) < \max(t(\mathbf{z}))$.

The permutational distribution consists of possible values of $t(\mathbf{z})$ and their associated probabilities denoted by $C(t(\mathbf{z})|\cdot)$. Let

$$R(\theta) = \frac{\sum_{u^*=t(\mathbf{z}_{\text{obs}})}^{\max(t(\mathbf{z}))} \mathrm{C}(u^*|\cdot)\exp\{\theta u^*\}}{\sum_{u=\min(t(\mathbf{z}))}^{\max(t(\mathbf{z}))} \mathrm{C}(u|\cdot)\exp\{\theta u\}}, \tag{2.1}$$

which denotes the total probability on the right of the observed value of the test statistic $t(\mathbf{z}_{\text{obs}})$ as a function of θ. Then, if (θ_-, θ_+) is a typical two-sided $100(1-\alpha)\%$ CI for θ, letting $R(\theta_-) = \alpha_-$, $R(\theta_+) = 1 - \alpha_+$, and $\alpha_- + \alpha_+ = \alpha$ results in a $100(1-\alpha)\%$ CI for θ for the permutation-based method of $(R^{-1}(\alpha_-), R^{-1}(1-\alpha_+))$. One-sided $100(1-\alpha)\%$ CIs can be given by $(R^{-1}(\alpha), \infty)$ and $(-\infty, R^{-1}(1-\alpha))$.

2.3 The Permutation-Based Method as a CD

The first contribution of this paper is to prove the permutation-based method can be considered a CD. Based on Equation 2.1, we let the lower confidence bound of a one-sided $100(1-\alpha)\%$ CI for θ for the permutation-based method be $R^{-1}(\alpha)$. To prove this lower bound, inverted, is a CD, we need to show that this lower bound is (1) a function of α, (2) continuous, and (3) increasing in α for each sample $\mathbf{z}$, where $\alpha \in (0,1)$ and $\theta \in \Theta$, thus making the inverted lower bound a CD: $(R^{-1}(\alpha))^{-1} = R(\theta) = \phi(\theta)$.

1. Clearly $R^{-1}(\alpha)$ is a function of α for $\alpha \in (0, 1)$.

2. We will start by showing $R(\theta)$ is continuous. Note that the numerator and denominator both contain the exponential function, which is a continuous function. Since the quotient of two continuous functions is also continuous, $R(\theta)$ is continuous. We now must show that $R^{-1}(\alpha)$ is continuous. Since $R(\theta)$ is increasing in α, it is a bijective function. The inverse of a continuous, bijective function is also continuous. Therefore, $R^{-1}(\alpha)$ is continuous.

3. To show $R^{-1}(\alpha)$ is increasing in α, we will first show that $R(\theta)$ being an increasing function in θ implies that $R^{-1}(\alpha)$ is increasing in α. Since the inverse function exists, $R^{-1}(\alpha)$ is bijective. If a bijective function is increasing, then its inverse is also increasing. Therefore, we only need show that $R(\theta)$ is increasing to show that $R^{-1}(\alpha)$ is increasing. We will now show that $R(\theta)$ is increasing in θ.

$$\begin{aligned}
R(\theta) &= \frac{\sum_{u^*=t(\mathbf{z}_{\text{obs}})}^{\max(t(\mathbf{z}))} \mathrm{C}(u^*|\cdot)\exp\{\theta u^*\}}{\sum_{u=\min(t(\mathbf{z}))}^{\max(t(\mathbf{z}))} \mathrm{C}(u|\cdot)\exp\{\theta u\}} \\
&= \frac{\sum_{u^*=t(\mathbf{z}_{\text{obs}})}^{\max(t(\mathbf{z}))} \mathrm{C}(u^*|\cdot)\exp\{\theta u^*\}}{\sum_{u=\min(t(\mathbf{z}))}^{\max(t(\mathbf{z}))} \mathrm{C}(u|\cdot)\exp\{\theta u\}} \times \frac{\exp\{-\theta t(\mathbf{z}_{\text{obs}})\}}{\exp\{-\theta t(\mathbf{z}_{\text{obs}})\}} \\
&= \frac{\sum_{u^*=t(\mathbf{z}_{\text{obs}})}^{\max(t(\mathbf{z}))} \mathrm{C}(u^*|\cdot)\exp\{\theta(u^* - t(\mathbf{z}_{\text{obs}}))\}}{\sum_{u=\min(t(\mathbf{z}))}^{\max(t(\mathbf{z}))} \mathrm{C}(u|\cdot)\exp\{\theta(u - t(\mathbf{z}_{\text{obs}}))\}} \\
&= \left(\frac{\sum_{u=\min(t(\mathbf{z}))}^{t(\mathbf{z}_{\text{obs}})-1} \mathrm{C}(u|\cdot)\exp\{\theta(u - t(\mathbf{z}_{\text{obs}}))\} + \sum_{u^*=t(\mathbf{z}_{\text{obs}})}^{\max(t(\mathbf{z}))} \mathrm{C}(u^*|\cdot)\exp\{\theta(u^* - t(\mathbf{z}_{\text{obs}}))\}}{\sum_{u^*=t(\mathbf{z}_{\text{obs}})}^{\max(t(\mathbf{z}))} \mathrm{C}(u^*|\cdot)\exp\{\theta(u^* - t(\mathbf{z}_{\text{obs}}))\}}\right)^{-1} \\
&= \left(1 + \frac{\sum_{u=\min(t(\mathbf{z}))}^{t(\mathbf{z}_{\text{obs}})-1} \mathrm{C}(u|\cdot)\exp\{\theta(u - t(\mathbf{z}_{\text{obs}}))\}}{\sum_{u^*=t(\mathbf{z}_{\text{obs}})}^{\max(t(\mathbf{z}))} \mathrm{C}(u^*|\cdot)\exp\{\theta(u^* - t(\mathbf{z}_{\text{obs}}))\}}\right)^{-1} \\
&= \left(1 + \frac{f(\theta)}{g(\theta)}\right)^{-1}, \text{ where} \\
f(\theta) &= \sum_{u=\min(t(\mathbf{z}))}^{t(\mathbf{z}_{\text{obs}})-1} \mathrm{C}(u|\cdot)\exp\{\theta(u - t(\mathbf{z}_{\text{obs}}))\} \text{ and} \\
g(\theta) &= \sum_{u^*=t(\mathbf{z}_{\text{obs}})}^{\max(t(\mathbf{z}))} \mathrm{C}(u^*|\cdot)\exp\{\theta(u^* - t(\mathbf{z}_{\text{obs}}))\}.
\end{aligned}$$

Since $f(\theta)$ is a decreasing function in θ and $g(\theta)$ is a non-decreasing function in θ, $R(\theta)$ is increasing in θ.

2.4 Evaluating the Overall Performance of CDs

The second contribution of this paper is to supply metrics for comparing CDs, computed using any method. There are no metrics currently available for evaluating the results of CDs from simulation studies aside from comparing their information at a single significance level (universally chosen to be 0.05) using traditional metrics (e.g. average CI coverage). Generally, when evaluating the performance of CIs, the interval coverage and width are reported. When evaluating the performance of CDs, researchers typically report the interval coverage and width at a certain level of significance, typically $\alpha = 0.05$. We find this approach lacking as the key benefit of using a CD is ignored when a set level of significance is chosen from which to evaluate performance. This, in effect, results in evaluating the performance of a $100(1-\alpha)\%$ CI instead of evaluating the performance of the entire CD. To better characterize a CD's overall performance, we develop two metrics which are roughly parallel to measuring the coverage and width of a CI. We utilize the CV for these two metrics.

First, to get an overall measure of the width of a CV, across all levels of α, we calculate the area under the CV (AUCV). Wider CVs with a relatively larger AUCV result in CIs being wider, on average, compared with narrower CVs, which equate to narrower CIs, on average. All else being equal, namely for CVs that maintain proper coverage, we prefer CVs that result in a smaller AUCV since these CVs reflect more precise estimates of the parameter of interest.

Second, to get an overall measure of the coverage of a CV, across all levels of α, we compute the height of the confidence curve at the true value of the parameter θ. This is equivalent to computing the proportion of all CIs from 0% to 100% that cover the true parameter since every CI with a level of significance smaller than the height will include the true parameter, and every CI with a level of significance larger than the height will not include the true parameter. Therefore, a bigger height indicates that the confidence curve has a higher measure of coverage overall. In theory, this height metric should follow a Unif(0, 1) distribution. Therefore, we would expect the mean height taken across many CVs to be close to 0.5.

CHAPTER 3

SIMULATION STUDY

We now present a simulation study in order to illustrate the use of the two new metrics developed in Chapter 2.4 to assess a CD's relative performance. For illustration, we will use the Liu et al. (2014) CD method, Cunen and Hjort (2021) CD method, and the newly created Zabriskie et al. (2021b) CD method for comparison. We generate meta-analysis data based on a method reviewed by Pateras et al. (2018), denoted by *pRandom*, which allows the baseline probability of an event to be set directly and assumes equal variances for both treatment arms. We use this data generating method since it was used in Zabriskie et al. (2021a), but we note that many other data generating methods are available, and research shows that the choice of data generating method can impact the results of meta-analyses when events are rare (Kulinskaya et al. 2021; Pateras et al. 2018).

Let n_{ij} and y_{ij} represent the total number of participants and the total number of participants with an observed event, respectively, in the i^{th} study ($i = 1, 2, \ldots, k$) with the j^{th} treatment ($j = 0$ for the control group, $j = 1$ for the treatment group). We set the true log odds ratio to be $\theta = -1, -0.5, 0, 0.5$, and 1, with corresponding average treatment event rates of 0.02, 0.03, 0.05, 0.08, and 0.13. Note that as θ increases, the probability of an event in the treatment arm also increases. We set the heterogeneity variance to be $\tau^2 = 0, 0.2, 0.4$, and 0.8. When $\tau^2 = 0$, datasets are homogeneous, and a fixed-effect framework could be appropriately used; whereas, when $\tau^2 > 0$, datasets are heterogeneous, and a random-effects framework is more appropriate. Additionally, we let the baseline probability of an event in the control group be $p_c = 0.05$, and the number of studies in each meta-analysis dataset be $k = 10$. Then, we use the following procedure to generate the meta-analysis datasets, based on the *pRandom* data generating method:

1. Calculate the baseline probability of an event in the treatment group: $p_t = \frac{p_c \exp(\theta)}{1-p_c+p_c \exp(\theta)}$.

2. Sample n_{ij} from a discrete Uniform$(10, 50)$ distribution, resulting in unbalanced treatment arms.

3. Generate the study-specific log odds θ_{ij} by sampling from a Normal $\left(\log\left(\frac{p_j}{1-p_j}\right), \frac{\tau^2}{2}\right)$ distribution.

4. Calculate the number of observed events y_{ij} by sampling from a Binomial $\left(n_{ij}, \frac{1}{1+\exp(-\theta_{ij})}\right)$ distribution.

We exclude datasets which do not satisfy the following conditions:

1. There are at least two studies in the meta-analysis dataset that have observed events in at least one arm.

2. There is at least one event across all studies in the treatment group in the meta-analysis dataset.

3. There is at least one event across all studies in the control group in the meta-analysis dataset.

4. Based on results from the Zabriskie et al. (2021b) method, the observed value of θ's sufficient statistic is not on the extreme of its distribution.

5. The derivative of the CV changes sign only at the point estimate, or cusp of the CV.

The first three requirements ensure the meta-analysis dataset contains enough information to provide meaningful results, especially in the random-effects framework where at least two studies are needed to estimate the heterogeneity variance parameter. Requirement four is needed so that conditional maximum likelihood estimation can be used for inference, and since the Zabriskie et al. (2021b) method would produce one-sided CIs, instead of the desired two-sided intervals, if this requirement was not enforced. This requirement resulted in 18.7%

of the simulated datasets being discarded. The last condition is needed since 0.6% of the CVs produced by the Cunen and Hjort (2021) method had a bump in the CV, making them invalid CVs since this signifies having multiple CIs at the same level of significance. After simulating datasets and excluding those that did not meet these criteria, we obtained 1000 meta-analysis datasets for each combination of θ and τ^2.

For each meta-analysis dataset, we calculate the CV for the Zabriskie et al. (2021b) method, the method of Liu et al. (2014), and the random-effects version of the Cunen and Hjort (2021) method. The Zabriskie et al. (2021b) method is implemented using information provided by the *rema* R package (Zabriskie et al. 2021c). The CV for their method is not implemented in the package, but we provide code in Appendix A, which shows how to use the provided information to then obtain the CV. We implement the Liu et al. (2014) method with the *gmeta* R package (Yang et al. 2021), and we implement the Cunen and Hjort (2021) method using code kindly provided by the authors. We create the CVs over values of θ ranging from -15 to 15 in increments of 0.01. For 1.28% of the datasets, the endpoints of the CV (at θ values of -15 and 15) for the Cunen and Hjort (2021) method summed up to be greater than 0.05. This indicates that the CV was not fully captured within the specified θ grid. Without the full curve, the true AUCV will be underestimated. Expanding the grid to contain more values of θ is too computationally expensive, especially since the Cunen and Hjort (2021) method takes markedly longer to run than the other two methods. Thus, to get an approximate AUCV in these situations, we need to evaluate the area outside of the grid, and so we apply an approximation to the CV using the numerical first derivative at each ends of the curve to draw a straight line outward until the CV intersects the x-axis. This is a conservative approximation method since CVs are a convex function on either side of the point estimate. This approximation is applied to both tails of all such curves before calculating the AUCV. Code for these procedures is provided in Appendix A.

3.1 Simulation Results

We now present the results of the simulation in order to demonstrate the utility of the two new metrics of Chapter 2.4 and to compare the Zabriskie et al. (2021b) CD method of Chapter 2.3 with other CD methods. While we will use these metrics to compare the performance of the Zabriskie et al. (2021b), Cunen and Hjort (2021), and Liu et al. (2014) methods, we emphasize that these metrics can be used to evaluate any CD.

Figure 3.1 shows the average AUCV plotted against the true value of θ on the x-axis and the true value τ^2 divided by facets. Note that the variability of these estimates is very small, with the largest standard error being 0.05. The Zabriskie et al. (2021b) method results in the largest average AUCV, regardless of the true underlying parameters, indicating wider overall CVs than the other methods, on average. This is expected, as the main advantage of this method is not in producing narrow intervals but in preserving the nominal level of significance. Thus, we expect rather conservative results. The Liu et al. (2014) and Cunen and Hjort (2021) methods result in similar AUCVs, with the Cunen and Hjort (2021) method resulting in slightly wider CVs overall, on average.

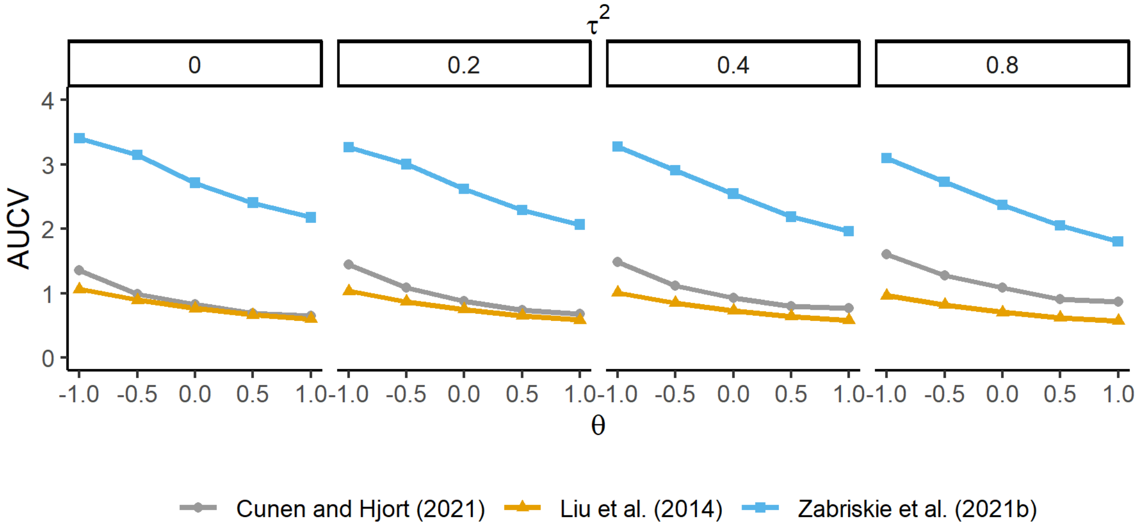

Figure 3.1: Average AUCV for the three considered methods for each combination of the true log odds ratio θ and the true heterogeneity variance τ^2.

Figure 3.2 shows the mean height of the CV at the true value of θ, with the true θ value on the x-axis, and true τ^2 values broken up by facets. Again, the variability of these estimates is small, with the largest standard error being 0.01. The mean height is equivalent to the proportion of CIs that cover the true parameter value. We expect the mean height to be close to 0.5, indicated by the dashed, black line. Since the methods generally have heights above 0.5, especially for smaller values of τ^2, the methods tend to be overly conservative. The Cunen and Hjort (2021) method generally results in the average height being the closest to 0.5. The Zabriskie et al. (2021b) method has slightly higher height at most parameter values. This indicates that there is a higher proportion of CIs that cover the true parameter. This reinforces the idea that the permutation-based method provides conservative results.

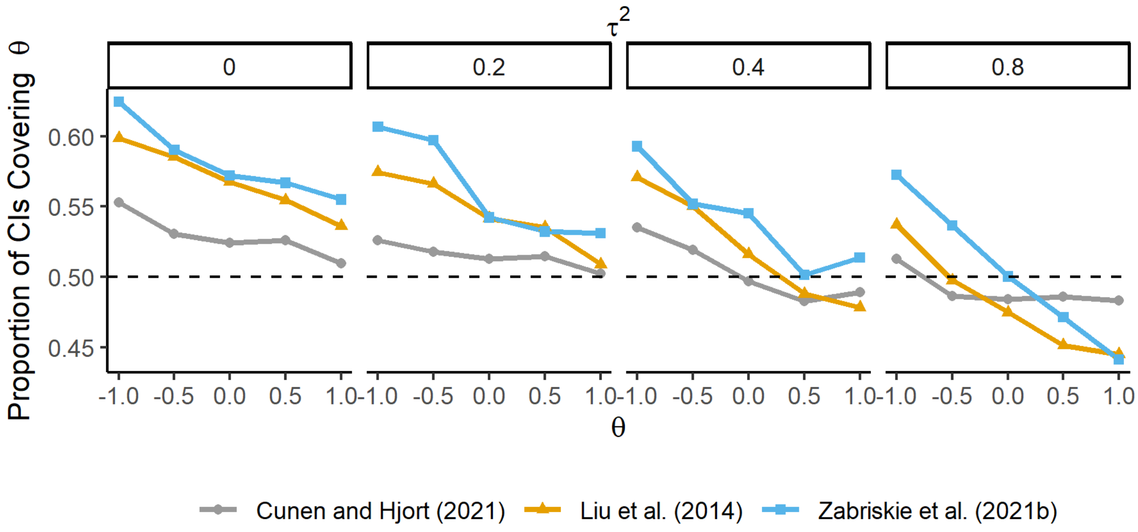

Figure 3.2: Mean height of the CV at the true parameter value for the three considered methods for each combination of the true log odds ratio θ and the true heterogeneity variance τ^2. The height is equivalent to computing the proportion of CIs that cover θ. The dashed, black line marks the expected proportion of 0.50.

Interestingly, the proportion of CIs covering θ decreases as θ increases. This is because of the impact of θ on the underlying probability of an event in the treatment arm. Smaller values of θ produce datasets with very rare events. Due to the presence of very rare events, there is little information available in these datasets for estimating θ, which results in wider

CIs when θ is small. However, when θ is larger, the treatment event rate is also larger, and there is more information for estimation, which results in more narrow CIs.

Additionally, we see that the proportion of CIs covering θ tends to decrease as τ^2 or θ increases. For the Cunen and Hjort (2021) method, particularly, it is interesting to note that as τ^2 increases, the AUCV increases, but the proportion of CIs covering the true parameter value decreases. The decrease in θ is due to the same reasons described above for the AUCV metric.

The height metric can also be used to calculate coverage as a function of α, rather than averaging over α. This can be done by finding the proportion of the heights of all datasets that are greater than some α. These results, where each function is transformed to be the difference between empirical coverage and expected coverage, are displayed in Figure 3.3. This transformation was done to make the distinctions between the methods more easily discernible visually. We can see that all three methods tend to have greater coverage than expected for most values of α, except when τ^2 and θ are large. Figure 3.2 tells a similar story as these plots, but summarizes across all values of α. We can see that even after summarizing over α, the two figures give similar results.

To further compare the three methods, we can utilize the height metric concept by finding the average power across all levels of α. This can be calculated as one minus the CV evaluated at the null value of the parameter. These results are presented in Figure 3.4. We see that the Liu et al. (2014) and Cunen and Hjort (2021) methods achieve their lowest average power at $\theta = 0$ (the Type I error rate), which then increases as θ moves away from zero. The Zabriskie et al. (2021b) method has lower average power than these two methods for all θ values, with lower power for $\theta < 0$ than $\theta = 0$. This is due to the greater uncertainty at low θ values. This is also the reason why power is not symmetric around $\theta = 0$.

To provide context for the average power plot, we provide the results of relative power for when α is 0.05 in Figure 3.5. As we can see, the same general trends appear in both Figure 3.5 and Figure 3.4. This demonstrates how averaging over α values retains enough

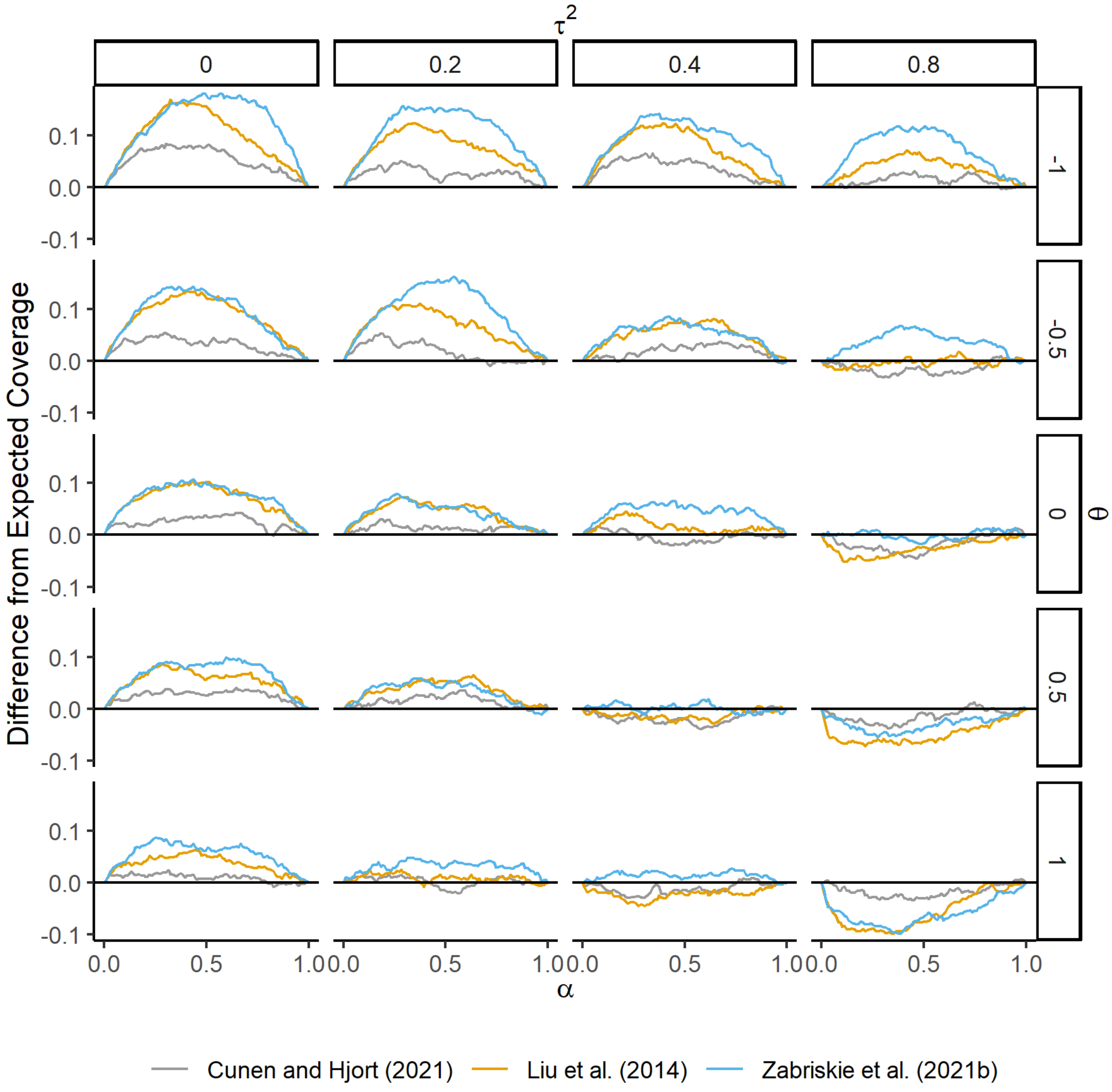

Figure 3.3: The difference between empirical coverage and expected coverage for all levels of α.

information to still portray the important general, relative trends of the methods being compared.

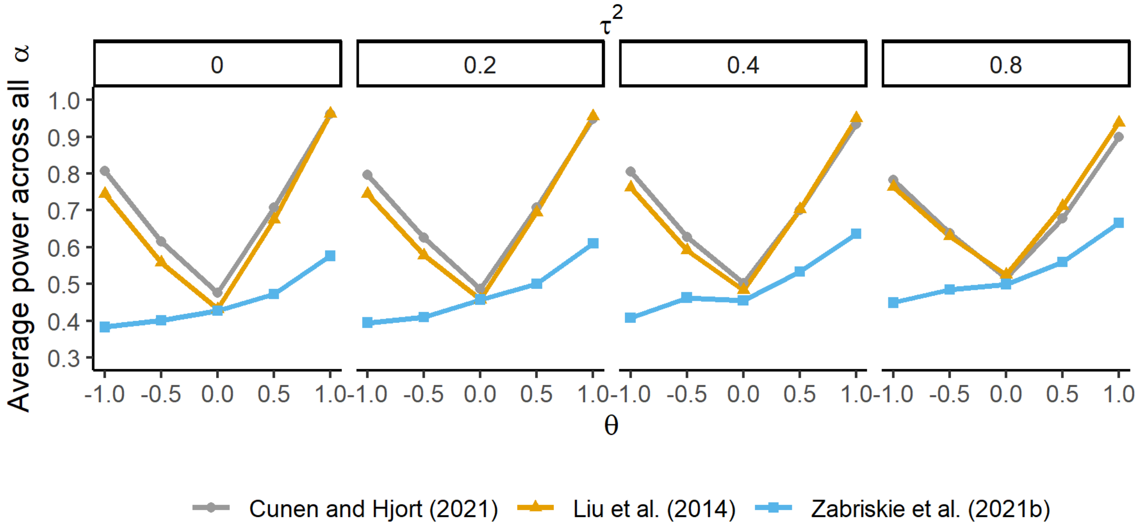

Figure 3.4: Average relative power across all levels of α. The trends of this plot are similar to those in Figure 3.5.

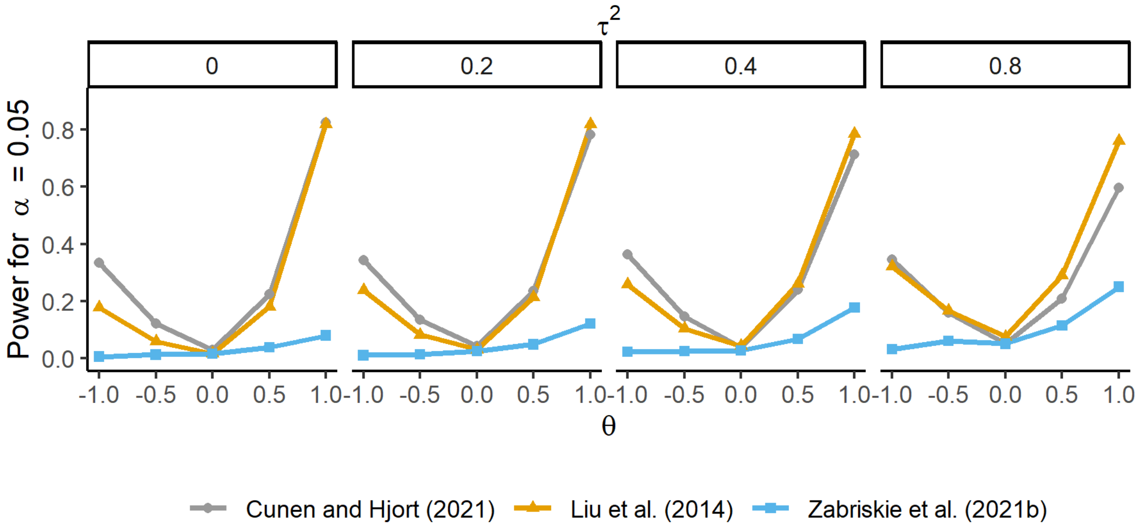

Figure 3.5: Relative power for $\alpha = 0.05$.

CHAPTER 4

CASE STUDY

In order to highlight the benefits of extending the Zabriskie et al. (2021b) method to become a CD, we present a case study using a meta-analysis performed by Tsivgoulis and Georgios (2016) to determine the association of cerebral microbleeds (CMBs) with the risk of symptomatic intracerebral hemorrhage (sICH) in patients with acute ischemic stroke treated with intravenous thrombolysis. This has an important application because CMBs have been found to be an independent predictor of cerebral bleeding (Tsivgoulis and Georgios 2016). Table 4.1 shows the results from the eight studies we consider for this meta-analysis. Note that for our analysis we removed a ninth study by Turc et al. due to its large sample size so that the data set would reflect meta-analyses that consist only of small studies. Additionally, Tsivgoulis and Georgios (2016) use the risk ratio as the effect size, while we apply the same methods they did (DerSimonian and Laird random-effects) but use the log odds ratio as the effect size for ease of comparison. Note that with rare events, the risk ratio and the odds ratio are often comparable.

The observed average event rates are 0.065 for the treatment group and 0.044 for the control group. The estimate for the heterogeneity parameter (τ^2), measuring the variation of the treatment effect, is 0.42. These characteristics roughly parallel the simulation setting in Chapter 3, where datasets were simulated with ten studies and an average event rate of 0.05. Table 4.2 provides the results from the methods used in the original analysis by Tsivgoulis and Georgios (2016) along with the results from the three CD methods for $\alpha = 0.05$. The estimates are roughly comparable, with all indicating greater odds of sICH for patients with CMBs compared to patients without CMBs. However, the Zabriskie et al. (2021b) method results in a wider CI and a non-significant p-value. Prior to this work, the interpretation

Table 4.1: This table contains the hemorrhage meta-analysis dataset: author of each study, the number of patients who had CMBs with sICH, the total number of patients with CMBs, the number of patients who did not have CMBs with siCH, and the total number of patients without CMBs. Estimates of the log odds ratio and corresponding 95% CI and p-value are also provided for each study.

Study	CMB Event	CMB Total	CTRL Event	CTRL Total	Log Odds Ratio	95% CI	p-value
Dannenberg et al.	7	81	3	245	2.03	(0.65, 3.41)	0.004
Derex et al.	1	8	2	36	0.89	(-1.65, 3.42)	0.493
Fiehler et al.	5	86	13	484	0.80	(-0.25, 1.86)	0.136
Goyal et al.	1	3	0	18	3.10	(-0.36, 6.56)	0.079
Gratz et al.	2	38	4	136	0.61	(-1.13, 2.34)	0.494
Kakuda et al.	0	11	5	59	-0.84	(-3.81, 2.12)	0.578
Kimura et al.	4	72	2	152	1.48	(-0.24, 3.21)	0.091
Yan et al.	6	132	2	201	1.56	(-0.06, 3.17)	0.059

of these results would stop here. However, now that we have extended the Zabriskie et al. (2021b) to be a CD, we can further analyze this dataset by considering the evidence provided across all levels of α, as seen in Figure 4.1.

Table 4.2: This table contains meta-analysis results of the hemorrhage dataset at the $\alpha = 0.05$ level. Results from the original analysis (using the DerSimonian and Laird random-effects method) are provided along with conclusions from the three CD methods when $\alpha = 0.05$.

Method	Estimate	95% CI	p-value
Original Method from Tsivgoulis and Georgios (2016)	1.20	(0.60, 1.79)	0.00008
Liu et al. (2014)	1.19	(0.60, 1.78)	0.00011
Cunen and Hjort (2021)	1.13	(0.50, 1.76)	0.00595
Zabriskie et al. (2021b)	1.37	(-0.02, 2.57)	0.05950

From Figure 4.1, we can see the results from Table 4.2 (namely, the estimates, which are at the cusps, and the 95% CIs) along with so much more. We see that the Liu et al. (2014) and Cunen and Hjort (2021) methods produce similar results, as did their results for a fixed α of 0.05, both agreeing with the authors of this study that there is a significant association between CMBs and sICH (Tsivgoulis and Georgios 2016). By comparison, the Zabriskie et al. (2021b) method gives results that are more conservative, which is expected given the simulation study results of Chapter 3 (when $\theta = 1$, which is close to this dataset's

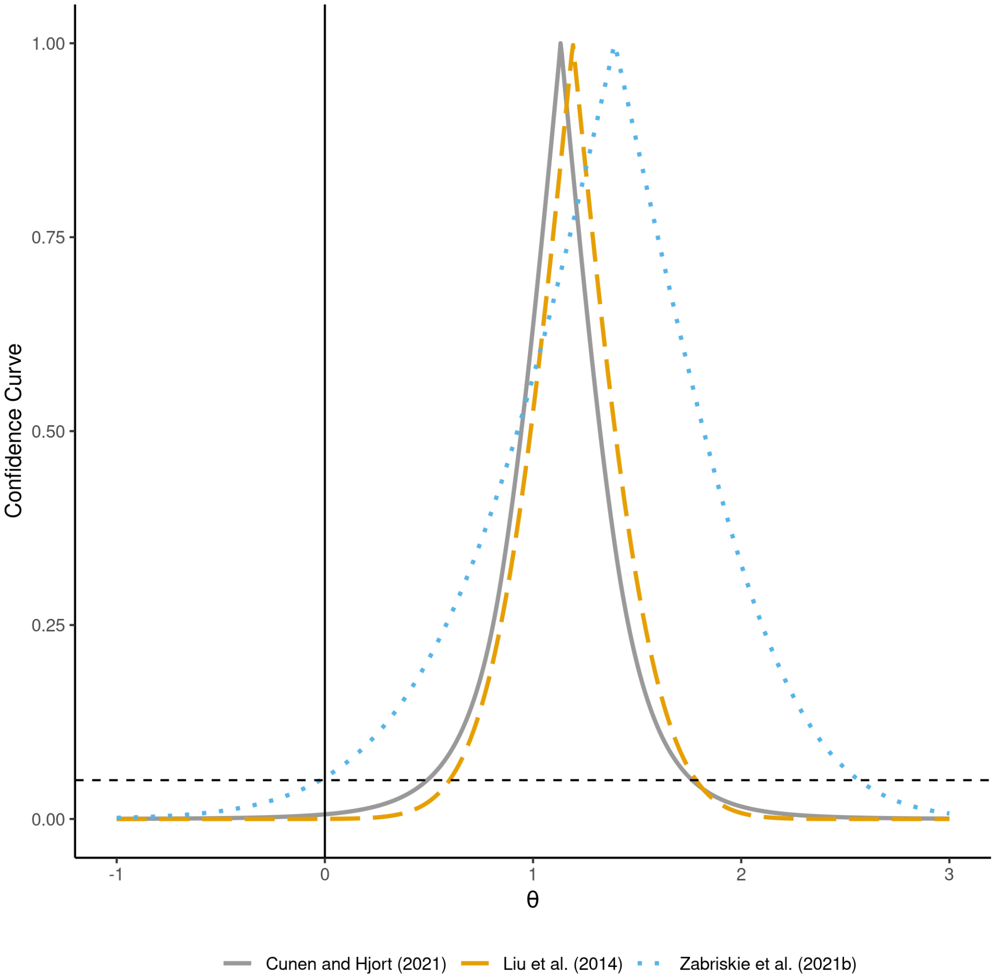

Figure 4.1: CVs for the Zabriskie et al. (2021b), Liu et al. (2014), and Cunen and Hjort (2021) methods for the hemorrhage meta-analysis dataset.

estimated combined treatment effect, and $\tau^2 = 0.4$, the Zabriskie et al. (2021b) method has a larger AUCV and a higher proportion of CIs that cover θ than the other two methods). Specifically, as we have seen at the $\alpha = 0.05$ level, the Zabriskie et al. (2021b) method does not agree that there is a statistically significant association. However, considering the Zabriskie et al. (2021b) CV as a whole indicates the results could possibly be practically significant, since its counternull is far from the null value of zero.

Thus, these three methods give a fairly consistent message, which demonstrates the importance of considering CV's when interpreting study results. Prior to our work, the Zabriskie et al. (2021b) method would appear to result in a different conclusion than the other methods, but now that we have created a CD from this method, we see that the conclusions are actually fairly uniform. CV's offer clarity to the results of a study that a binary judgement of "significant" or "not significant" cannot replicate. Even results that at first appear contradictory according to this binary system may actually be compatible when taking their respective CV's into account.

CHAPTER 5

DISCUSSION

Meta-analysis provides a way to combine individual studies to produce one overall result and is increasingly becoming the gold standard for presenting evidence in public health and medicine. CDs can be used to combine results for a meta-analysis, and they do not require selecting a single level of significance. In this work, we present two main contributions to the meta-analysis field. First, we extend the Zabriskie et al. (2021b) meta-analysis method by proving it can be considered a CD. Second, we develop two metrics for evaluating the relative performance of a CD across all levels of significance.

The first goal of showing the Zabriskie et al. (2021b) method can be a CD heightens the usefulness of this method by lending it all the advantages of CDs. Often in academic literature, scientific results are designated “significant” or “not significant” based upon some pre-chosen, rather arbitrary level of significance, which is almost uniformly selected to be 0.05. This binary system provides only a small window into the evidence provided by the data. This practice can draw arbitrary lines between otherwise similar findings. Furthermore, given the bias within the academic community towards publishing papers that find significant effects, the emphasis on limited metrics where a significance level must be chosen can contribute to p-hacking, as researchers respond to the incentive to procure a p-value less than 0.05.

In contrast, CDs do not require the selection of a single level of significance, and they provide a wealth of information on the parameter of interest at all levels of significance. As a result, they create a more complete picture of the evidence provided by the data, and they can reduce the incentive to p-hack by shifting the focus away from achieving some arbitrary p-value threshold. Furthermore, this more detailed rendering of model results can clarify different study's findings in relation to one another. Study results that may seem contradictory under the dichotomous classification system of "significant" versus "not significant" may prove to be similar when considering CDs as a whole. This circumstance was demonstrated in Chapter 4, where the Zabriskie et al. (2021b) CD was compared against two other methods. Under traditional analysis methods (e.g. a 95% CI) the three methods display different results; however, the CDs all convey similar messages.

The second main contribution of this paper, developing two new metrics to evaluate any CD, provide researchers a way to compare CDs without having to select a level of significance. These metrics, which relate to calculating the coverage and width of a single CI, allow for the comparison of any two CDs—not just those described in the paper. In light of the more nuanced picture that CDs convey compared to traditional methods, the development of such metrics is critical to understanding and comparing the results of CDs. These metrics can help clarify results and allow for better synthesis of the evidence provided by data.

Another result of this paper, discussed in Chapter 3, is understanding the performance of the permutation-based meta-analysis method when it is a CD. Here, we saw the Zabriskie et al. (2021b) method CD is conservative, as the AUCV tends to be larger than those of the other two methods. The average height of the CV at the true value of θ is also generally slightly higher for the permutation-based CV, indicating that the proportion of all confidence intervals that cover the true parameter is slightly higher than this proportion from the other methods. The permutation-based CD was also more conservative in the case study. While the Zabriskie et al. (2021b) method was shown to outperform other *traditional* meta-analysis methods at the $\alpha = 0.05$ level (in their paper), we find that this method, in its CD form, does not outperform other *CD* meta-analysis methods. As noted in Chapter 3, the method used to generate data can greatly impact the results of rare-event meta-analyses. Future work would involve applying many data generating methods to better understand the performance of all methods considered here.

Additional future work could be to improve this method as a CD. In the R package implementing the Zabriskie et al. (2021b) method (Zabriskie et al. 2021c), a more powerful test, using their method, is provided that helps to reduce conservatism. They use the conditional probabilities test rather than the conditional scores test. We used the conditional scores test here, so future work could involve creating a CD for this method based on the conditional probabilities test to reduce conservatism.

Another area of future work would be further development of the metric for evaluating a CD's coverage. In theory, the height of the CV at the true parameter value follows a Uniform(0, 1) distribution. Therefore, the mean height across all simulations should be around 0.5. It could be useful to have a metric which provides more insight on what is happening at different levels of α. Another reason for further analysis is that CIs are generally only of interest for α values less than 0.2. It could be ideal to give additional weight to common significance levels.

BIBLIOGRAPHY

Cunen, C., and Hjort, N. L. (2021), "Combining information across diverse sources: The II-CC-FF paradigm," *Scandinavian Journal of Statistics*, Early View, DOI: 10.1111/sjos.12530.

Efthimiou, O. (2018), "Practical guide to the meta-analysis of rare events," *Evidence-Based Mental Health*, 21(2), 72–76, DOI: 10.1136/eb-2018-102911.

Hunter, J. E., and Schmidt, F. L. (2000), "Fixed Effects vs. Random Effects Meta-Analysis Models: Implications for Cumulative Research Knowledge," *International Journal of Selection and Assessment*, 8(4), 275–292, DOI: 10.1111/1468-2389.00156.

Infanger, D., and Schmidt-Trucksäss, A. (2019), "*P* value functions: An underused method to present research results and to promote quantitative reasoning," *Statistics in Medicine*, 38(21), 4189–4197, DOI: 10.1002/sim.8293.

Ioannidis, J. P. A. (2019), "What Have We (Not) Learnt from Millions of Scientific Papers with *P* Values?" *The American Statistician*, 73(Sup1), 20–25, DOI: 10.1080/00031305.2018.1447512.

Kontopantelis, E., Springate, D. A., and Reeves, D. (2013), "A Re-Analysis of the Cochrane Library Data: The Dangers of Unobserved Heterogeneity in Meta-Analyses," *PLoS ONE*, 8(7): e69930, DOI: 10.1371/journal.pone.0069930.

Kulinskaya, E., Hoaglin, D. C., and Bakbergenuly, I. (2021), "Exploring consequences of simulation design for apparent performance of methods of meta-analysis," *Statistical Methods in Medical Research*, 30(7), 1667–1690, DOI: 10.1177/09622802211013065.

Liu, D. (2019), “Meta-Analysis of Rare Events,” *Wiley StatsRef: Statistics Reference Online*, DOI: 10.1002/9781118445112.stat08167.

Liu, D., Liu, R., and Xie, M. (2018), “Exact Inference Methods for Rare Events,” *Wiley StatsRef: Statistics Reference Online*, DOI: 10.1002/9781118445112.stat08065.

Liu, D., Liu, R. Y., and Xie, M. (2014), “Exact Meta-Analysis Approach for Discrete Data and its Application to 2 x 2 Tables With Rare Events,” *Journal of the American Statistical Association*, 109(508), 1450–1465, DOI: 10.1080/01621459.2014.946318.

Pateras, K., Nikolakopoulos, S., and Roes, K. (2018), “Data-generating models of dichotomous outcomes: Heterogeneity in simulation studies for a random-effects meta-analysis,” *Statistics in Medicine*, 37(7), 1115–1124, DOI: 10.1002/sim.7569.

Singh, K., Xie, M., and Strawderman, W. E. (2005), “Combining information from independent sources through confidence distributions,” *The Annals of Statistics*, 33(1), 159–183, DOI: 10.1214/009053604000001084.

Sutton, A. J., Cooper, N. J., Lambert, P. C., Jones, D. R., Abrams, K. R., and Sweeting, M. J. (2002), “Meta-analysis of rare and adverse event data,” *Expert Review in Pharmacoeconomics Outcomes Research*, 2(4), 367–379, DOI: 10.1586/14737167.2.4.367.

Sutton, A. J., and Higgins, J. P. T. (2008), “Recent developments in meta–analysis,” *Statistics in Medicine*, 27(5), 625–650, DOI: 10.1002/sim.2934.

Tsivgoulis, and Georgios (2016), “Risk of Symptomatic Intracerebral Hemorrhage After Intravenous Thrombolysis in Patients With Acute Ischemic Stroke and High Cerebral Microbleed Burden: A Meta-analysis.” *JAMA neurology*, 73(6), 675–683, DOI: 10.1001/jamaneurol.2016.0292.

Wasserstein, R. L., and Lazar, N. A. (2016), “The ASA’s Statement on p-Values: Context, Process, and Purpose,” *The American Statistician*, 70(2), 129–133, DOI: 10.1080/00031305.2016.1154108.

Yang, G., Cheng, J. Q., Xie, M., and Qian, W. (2021), *gmeta: Meta-Analysis via a Unified Framework of Confidence Distribution*, R package version 2.3-1, https://CRAN.R-project.org/package=gmeta.

Zabriskie, B. N., Corcoran, C., and Senchaudhuri, P. (2021a), "A comparison of confidence distribution approaches for rare event meta-analysis," *Statistics in Medicine*, 40(24), 5276–5297, DOI: 10.1002/sim.9125.

—— (2021b), "A permutation-based approach for heterogeneous meta-analyses of rare events," *Statistics in Medicine*, 40(25), 5587–5604, DOI: 10.1002/sim.9142.

Zabriskie, B. N., Kinard, B., Sypherd, C., and Whetten, R. (2021c), *rema: Rare Event Meta Analysis*, r package version 0.0.1.

APPENDICES

Appendix A: R Code

Following is a general overview of the code files included in this appendix. The files relating to the Cunen and Hjort (2021) method were not included as they are proprietary.

- sim_data.R: Generates synthetic datasets and writes them to data/datasets.
- model_data.R: Calculates a point estimate, CI, *p*-value, and CV for each dataset in data/datasets, and writes the corresponding R workspace to data/workspaces.
- summarize_data.R: Loads workspaces from data/workspaces and calculates the AUCV and height for each dataset, which are written to data/sim_results.csv.
- sim_results.R: Reads data/sim_results.csv, and makes two graphs summarizing the AUCV and height results, which are written to data/area.png and data/height.png
- hemorrhage.R: Reads hemorrhage dataset from data/hemorrhage.csv, and generates a graph of the Zabriskie et al. (2021b), Liu et al. (2014), and Cunen and Hjort (2021) CVs which is written to graphs/hemorrhage.png.
- pRandom_sameVar.R: Function for simulating datasets.
- get_cd.R: Function for calculating the CD for the Zabriskie et al. (2021b) method.

sim_data.R

```
library(tidyverse)
source('code/source/pRandom_sameVar.R')

#' Simulates a meta-analysis dataset and writes it to raw_data folder
#'
#' @param rep repetition number
#' @param theta true theta
#' @param k number of studies
#' @param p_ic_init true control arm probability
#' @param tau2 true tau^2
#' @param min.n Minimum number of subjects
#' @param max.n Maximum number of subjects
sim_data <- function(rep, theta, k, p_ic_init, tau2, min.n, max.n){
  sim <- pRandom(theta, k, p_ic_init, tau2, min.n, max.n) %>%
    as_tibble
  ident_str <- str_c(theta, '_', tau2, '_', rep)
  write_csv(sim, str_c('data/datasets/', ident_str, '.csv'))
}

# Simulate all datasets, and write to raw_data
future::plan('multisession', workers = 100)

expand_grid(
  rep = 1:2500,
  theta = c(-1, -0.5, 0, 0.5, 1),
  k = 10,
  p_ic_init = 0.05,
  tau2 = c(0, 0.2, 0.4, 0.8),
  min.n = 10,
  max.n = 50
) %>%
  filter(!(rep > 1200 & theta >= 0), !(rep > 2000 & theta == -0.5)) %>%
  furrr::future_pmap(sim_data, .options = furrr::furrr_options(seed =
     1234))
```

model_data.R

```
library(tidyverse)
library(rema)
source('code/source/get_cd.R')
source('code/source/ii_cc_ff.R')
source('code/source/0_icf_code.R')

model_data <- function(filepath){
  # Read data
  sim <- read_csv(filepath)

  # Get parameters
  params <- filepath %>%
    str_remove('data/datasets/') %>%
    str_remove('.csv') %>%
    str_split('_') %>%
    unlist %>%
    as.numeric
  theta <- params[1]
  tau2 <- params[2]
  rep <- params[3]

  # See whether dataset is extreme or only one in distribution
  perm <- with(sim, rema(TRT_event, TRT_n, CTRL_event, CTRL_n, alpha =
     0.05))
  extreme <- perm$tstat %in% range(perm$dist$test.stat)
  only_one <- nrow(perm$dist) == 1

  if(only_one){
    ci_df <- NULL
    cd_df <- NULL
  } else{
    # Calculate CD for each method
    thetas <- seq(-15, 15, by = 0.01)
    perm_cd <- get_cd(thetas, perm$dist$test.stat, perm$dist$norm.probs,
       perm$tstat)
    liu <- gmeta::gmeta(
      with(sim, cbind(TRT_event, TRT_n, CTRL_event, CTRL_n)),
      gmi.type = ’2x2’,
      method = ’exact1’,
      gmo.xgrid = thetas
    )
    dyn.load(TMB::dynlib(’code/source/meta2x2_re_full_new’))
    m_ii_cc_ff <- with(sim, ii_cc_ff(TRT_event, TRT_n, CTRL_event, CTRL_n,
        thetas))

        # Calculate point estimate, 95% CI, and p-value for each method
    ci_df <- tibble(
      names = c(’estimate’, ’lower’, ’upper’, ’p_value’),
      ii_cc_ff = with(m_ii_cc_ff, c(estimate, lower, upper, NA)),
      liu = with(liu, c(combined.mean, combined.ci[1], combined.ci[2],
         pvalue)),
      perm = with(perm, log(c(TE, CI[1], CI[2], exp(pval))))
    ) %>%
      pivot_longer(-names, ’method’) %>%
      pivot_wider(method, names)

    # Combine CDs for each method
        cd_df <- tibble(thetas, perm = perm_cd, liu = liu$combined.cd, ii_
           cc_ff = 1-m_ii_cc_ff$cd) %>%
          mutate_at(c(’perm’, ’liu’), function(x) 1-abs(1-2*x))
  }
  file_out <- filepath %>%
    str_replace(’.csv’, ’.RData’) %>%
    str_replace(’datasets’, ’workspaces’)
  save(ci_df, cd_df, theta, tau2, rep, extreme, only_one, file = file_out)
}

# Model data for every dataset in raw_data directory, and write to results
    directory
future::plan(’multisession’, workers = 75)

dir(’data/datasets’, full.names = TRUE) %>%
  furrr::future_map(model_data, .options = furrr::furrr_options(seed = 1))
```

summarize_data.R

```
library(tidyverse)

count_invalid <- function(cd_df){
  cd_df %>%
    pivot_longer(-thetas, ’method’) %>%
    group_by(method) %>%
    mutate(
      estimate = thetas[which.max(value)],
      side = ifelse(thetas < estimate, ’left’, ’right’),
      d1 = c(diff(value), NA),
      invalid = (side == 'left' & d1 < -1e-10) | (side == 'right' & d1 > 1
         e-10)
    ) %>%
    filter(abs(thetas - estimate) >= 0.01) %>%
    ungroup %>%
    summarize(invalid = sum(invalid, na.rm = TRUE)) %>%
    pull(invalid)
}

tail_approx <- function(cd_df){
  df <- cd_df %>%
    pivot_longer(-thetas, 'method', values_to = 'cv') %>%
    mutate_at('thetas', round, 2) %>%
    group_by(method) %>%
    mutate(
      a0 = first(cv),
      a1 = first(diff(cv))/0.01,
      b0 = last(cv),
      b1 = last(diff(cv))/0.01
    )

  df %>%
    summarize(
      left = ifelse(a1 > 0, round(-a0/a1, 2) - 15, -15),
      right = ifelse(b1 < 0, round(-b0/b1, 2) + 15, 15)
    ) %>%
    slice(1) %>%
    mutate(left = replace_na(left, -15), right = replace_na(right, 15))
       %>%
    ungroup %>%
    rowwise %>%
    do(tibble(method = .$method, thetas = seq(.$left, .$right, 0.01))) %>%
    ungroup %>%
    mutate_at('thetas', round, 2) %>%
    full_join(df, c('thetas', 'method')) %>%
    fill(c('a0', 'a1', 'b0', 'b1'), .direction = 'updown') %>%
    mutate(
      cv_hat = case_when(
        thetas < -15 ~ a1*(thetas + 15) + a0,
        thetas > 15 ~ b1*(thetas - 15) + b0,
        TRUE ~ 0
      ),
      cv = coalesce(cv, cv_hat)
    )
}

summarize_data <- function(filepath){
  load(filepath)
  if(extreme || count_invalid(cd_df) != 0){
    df <- tibble(method = c('ii_cc_ff', 'liu', 'perm'))
  } else{
    df <- tail_approx(cd_df) %>%
      group_by(method) %>%
      summarize(
        area = bayestestR::area_under_curve(thetas, cv, 'trapezoid'),
        height = cv[thetas == theta],
        height2 = bayestestR::area_under_curve(
          thetas[cv <= height],
          cv[cv <= height],
          method = 'trapezoid'
        )/area
      )
  }
  df %>%
    mutate(extreme, only_one, theta, tau2, rep)
}

# Calculate metrics for each CV from results directory, and write to sim_
    results.csv
future::plan('multisession', workers = 75)

dir('data/workspaces', full.names = TRUE) %>%
  furrr::future_map(summarize_data) %>%
  bind_rows %>%
  write_csv('data/sim_results.csv')
```

sim_results.R

```
library(tidyverse)

########## Prepare data ##########
df <- read_csv('data/sim_results.csv') %>%
  drop_na %>% # Remove extreme datasets and those with invalid slopes
  filter( # These two datasets had a negative area for the Liu method
    !(theta == -1 & tau2 == 0.2 & rep == 233),
    !(theta == -0.5 & tau2 == 0.8 & rep == 875)
  ) %>%
  group_by(theta, tau2) %>%
  mutate(row = ceiling(row_number()/3)) %>%
  filter(row <= 1000) %>% # Keep first 1000 datasets for each combination
      of parameters
  group_by(method, theta, tau2) %>%
  summarize_at(c('area', 'height'), mean) %>%
  mutate(
    method = recode(
      method,
      perm = 'Zabriskie␣et␣al.␣(2021b)',
      liu = 'Liu␣et␣al.␣(2014)',
      ii_cc_ff = 'Cunen␣and␣Hjort␣(2021)'
    )
  )

########## Set plot ##########
plot <- ggplot(df, aes(theta, col = method)) +
  facet_grid(cols = vars(tau2)) +
  theme_classic() +
  xlab(expression(theta)) +
  scale_x_continuous(
    breaks = c(-1, -0.5, 0, 0.5, 1),
    sec.axis = sec_axis(~ ., name = expression(tau^2), breaks = NULL,
       labels = NULL)
  ) +
  scale_colour_manual(
    values = c('#999999', '#E69F00', '#56B4E9')
  ) +
  theme(
    legend.position = 'bottom',
    legend.title = element_blank()
  )

########## Area plot ##########
plot +
  geom_line(aes(y = area), size = 1) +
  geom_point(aes(y = area, shape = method)) +
  scale_y_continuous('Area␣Under␣the␣CV', breaks = 0:4, limits = c(0, 4))
ggsave(filename = 'graphs/area.png', width = 6, height = 3)

########## Height plot ##########
plot +
  geom_line(aes(y = height), size = 1) +
  geom_point(aes(y = height, shape = method)) +
  geom_hline(yintercept = 0.5, linetype = 'dashed') +
  scale_y_continuous(
    expression('Proportion␣of␣CIs␣Covering␣' ~ theta),
    breaks = seq(0.4, 0.65, 0.05),
    limits = c(0.4, 0.65)
  )
ggsave(filename = 'graphs/height.png', width = 6, height = 3)
```

hemorrhage.R

```
library(tidyverse)
library(rema)
source('code/source/get_cd.R')
source('code/source/ii_cc_ff.R')
source('code/source/0_icf_code.R')

df <- read_csv('data/hemorrhage.csv') %>%
  slice(-8)

thetas <- seq(-3, 3, length = 1000)
perm <- with(df,
  rema(CMB_Event, CMB_Total, CTRL_Event, CTRL_Total, alpha = 0.05)
)
perm_cd <- get_cd(
  thetas,
  perm$dist$test.stat,
  perm$dist$norm.probs,
  perm$tstat
)

liu <- gmeta::gmeta(
  with(df, cbind(CMB_Event, CMB_Total, CTRL_Event, CTRL_Total)),
  gmi.type = ’2x2’,
  method = ’exact1’,
  gmo.xgrid = thetas
)
dyn.load(’code/source/meta2x2_re_full_new.so’)
m_ii_cc_ff <- with(df,
  ii_cc_ff(CMB_Event, CMB_Total, CTRL_Event, CTRL_Total, thetas)
)

cd_df <- tibble(
  thetas,
  perm = perm_cd,
  liu = liu$combined.cd,
  ii_cc_ff = 1-m_ii_cc_ff$cd
) %>%
     mutate_at(c(’perm’, ’liu’), function(x) 1-abs(1-2*x))

cd_df %>%
  pivot_longer(-thetas, ’method’) %>%
  mutate(
    method = recode(
      method,
      perm = ’Zabriskie␣et␣al.␣(2021b)’,
      liu = ’Liu␣et␣al.␣(2014)’,
      ii_cc_ff = ’Cunen␣and␣Hjort␣(2021)’
    )
  ) %>%
  ggplot(aes(thetas, value, col = method, linetype = method)) +
#   geom_point(aes(counternull, perm$pval), shape = ’circle open’, size =
    2) +
  geom_line(size = 1) +
  geom_hline(yintercept = 0.05, linetype = ’dashed’) +
  geom_vline(xintercept = 0) +
  xlim(-1, 3) +
  xlab(expression(theta)) +
  ylab(’Confidence␣Curve’) +
  scale_colour_manual(
    values = c(’#999999’, ’#E69F00’, ’#56B4E9’)
  ) +
  scale_linetype_manual(
    values = c(’solid’, ’longdash’, ’dotted’)
  ) +
  theme_classic() +
  theme(
    legend.position = ’bottom’,
    legend.title = element_blank()
  )
ggsave(’graphs/hemorrhage.png’)

# counternull = thetas[which.min(abs(curve - m$pval))]
```

pRandom_sameVar.R

```r
# based on the incorporation of the between-study variance in both
# treatment arms via the use of logits

# theta=1; k=10; p_ic_init=0.05; tau2=0.2; min.n=10; max.n=50; set.seed
   (1234)

pRandom <- function(theta, k, p_ic_init, tau2, min.n, max.n) {

  # 2
  p_it_init <- p_ic_init * exp(theta) / (1 - p_ic_init + p_ic_init * exp(
     theta))

  # For balanced designs:
     ------------------------------------------------------
  # # 3
  # n_i <- round(runif(n = k, min = min.n, max = max.n))
  #
  # # 4
  # n_ic <- n_i
  # n_it <- n_i
  #
     ----------------------------------------------------------------------------


  # For unbalanced designs:
     ----------------------------------------------------
  # 3 & 4
  n_ic <- round(runif(n = k, min = min.n, max = max.n))
  n_it <- round(runif(n = k, min = min.n, max = max.n))
  #
     ----------------------------------------------------------------------------



  # 5
  mu_ic <- log(p_ic_init / (1 - p_ic_init))
  mu_it <- log(p_it_init / (1 - p_it_init))

  # For equal variances:
     -------------------------------------------------------
  # 6
  theta_ic <- rnorm(n = k, mean = mu_ic, sd = sqrt(tau2) / sqrt(2))
  theta_it <- rnorm(n = k, mean = mu_it, sd = sqrt(tau2) / sqrt(2))
  #
     ----------------------------------------------------------------------------


  # For unequal variances:
     -----------------------------------------------------
  # 6
  # theta_ic <- rnorm(n = k, mean = mu_ic, sd = sqrt(0.5))
  # theta_it <- rnorm(n = k, mean = mu_it, sd = sqrt(tau2))
  # ------------------------------------------------------------------------------

  # 7
  p_ic <- 1 / (1 + exp(-theta_ic))
  p_it <- 1 / (1 + exp(-theta_it))

  # 8
  r_ic <- rbinom(n = k, size = n_ic, prob = p_ic)
  r_it <- rbinom(n = k, size = n_it, prob = p_it)

  # first check: make sure there are at least two studies with non-zeros in at
  # least one arm (if not, then estimating heterogeneity is pointless since
  # only one study is used in the analysis)
  # second check: make sure there is at least one event across studies in the
  # treatment group
  # third check: make sure there is at least one event across studies in the
  # control group
  while(sum((r_ic + r_it) > 0) < 2 ||
        sum(r_it) == 0 ||
        sum(r_ic) == 0) {
    r_ic <- rbinom(n = k, size = n_ic, prob = p_ic)
    r_it <- rbinom(n = k, size = n_it, prob = p_it)
  }

  r.list <- list("CTRL_n" = n_ic, "TRT_n" = n_it,
                 "CTRL_event" = r_ic, "TRT_event" = r_it)
  return(r.list)

}
```

get_cd.R

```
#' Compute a CD for the perm method
#'
#' @param thetas An x-grid of theta values.
#' @param u Possible values of distribution of test statistic.
#' @param prob Probabilities of each test statistic in distribution.
#' @param t_obs Observed test statistic.
#' @return A vector containing the corresponding CD for thetas.
#' @examples
#' m <- rema::rema(TRT_event, TRT_n, CTRL_event, CTRL_n)
#' theta_grid <- seq(-1, 1, length = 1000)
#' get_cd(theta_grid, m$dist$norm.probs, m$dist$test.stat, m$tstat)
get_cd <- function(thetas, u, prob, t_obs){
  # rescale test statistic
  t_obs <- t_obs - min(u)
  u <- u - min(u)
  # get normalized probabilites for each theta
  P_unnorm <- prob*exp(u %x% t(thetas))
  P_norm <- t(t(P_unnorm)*colSums(P_unnorm)^(-1))
  # sum up extreme values
  extreme <- as.numeric(u > t_obs) + 0.5*(u == t_obs)
  colSums(P_norm*extreme)
}
```