Fall 12-21-2019

# Inference of Heterogeneity in Meta-analysis of Rare Binary Events and RSS-structured Cluster Randomized Studies

Chiyu Zhang
*Southern Methodist University*, chiyuz@smu.edu

INFERENCE OF HETEROGENEITY
IN META-ANALYSIS OF RARE BINARY EVENTS
AND RSS-STRUCTURED CLUSTER RANDOMIZED STUDIES

Approved by:

---

Dr. Xinlei (Sherry) Wang
Professor in Department of Statistical Science, SMU

---

Dr. Min Chen
Associate Professor in Department of Mathematical Sciences, UT Dallas

---

Dr. Lynne S. Stokes
Professor in Department of Statistical Science, SMU

---

Dr. Hon Keung "Tony" Ng
Professor in Department of Statistical Science, SMU

# INFERENCE OF HETEROGENEITY IN META-ANALYSIS OF RARE BINARY EVENTS AND RSS-STRUCTURED CLUSTER RANDOMIZED STUDIES

A Dissertation Presented to the Graduate Faculty of the

Dedman College

Southern Methodist University

in

Partial Fulfillment of the Requirements

for the degree of

Doctor of Philosophy

with a

Major in Statistical Science

by

Zhang Chiyu

B.S., Mathematics, Sun Yat-Sen University
M.S., Financial Engineering, University of Illinois, Urbana-Champaign

December 21, 2019

ACKNOWLEDGMENTS

First of all, I would like to express my most sincere gratitude to my advisor, Dr. Xinlei (Sherry) Wang, for her continuous support of my PhD study and research. Dr. Wang is one of the most dedicated, inspiring, and intelligent professors I have ever met. I would not imagine my achievement today without her patient training in every aspect about being a good statistician. I would also like to extend my gratitude to my co-advisor Dr. Min Chen, who has assisted me greatly with the revision of my dissertation. Dr. Chen sacrificed a lot of his personal time, even during holidays, to help me complete my dissertation with high quality in a timely manner. I cannot thank Dr. Wang and Dr. Chen enough for all their help along the way.

I am also grateful for the rest of my committee members: Dr. Lynne S. Stokes and Dr. Hon Keung "Tony" Ng. They have provided insightful comments and suggestions to help improve my dissertation. I also want to thank all the other faculty members in the Department of Statistical Science at SMU for their teaching and help.

I would like to thank Dr. Lie Li and Dr. Mumu Wang, two former students of Dr. Wang's, for their work and advice in the related topics in my dissertation. I also appreciate the company from all of my friends here in Dallas and all over the world. It is your support and care that give me strenghth throughout the journey.

Last but not least, I would like to thank my parents and other family members for their endless love and support.

Chiyu, Zhang B.S., Mathematics, Sun Yat-Sen University
M.S., Financial Engineering, University of Illinois, Urbana-Champaign

Inference of Heterogeneity
in Meta-analysis of Rare Binary Events
and RSS-structured Cluster Randomized Studies

Advisor: Dr. Xinlei (Sherry) Wang
Doctor of Philosophy degree conferred December 21, 2019
Dissertation completed December 3, 2019

This disseration contains two topics: (1) A Comparative Study of Statistical Methods for Quantifying and Testing Between-study Heterogeneity in Meta-analysis with Focus on Rare Binary Events; (2) Estimation of Variances in Cluster Randomized Designs Using Ranked Set Sampling.

Meta-analysis, the statistical procedure for combining results from multiple studies, has been widely used in medical research to evaluate intervention efficacy and safety. In many practical situations, the variation of treatment effects among the collected studies, often measured by the heterogeneity parameter $\tau^2$, may exist and can greatly affect the inference about effect sizes. Comparative studies have been done for only one or two of the heterogeneity-related topics including statistical models used, descriptive measures, estimation, hypothesis testing, and confidence intervals. Also, none of the studies is focused on rare binary events that require special attention. Our goal is to provide a comprehensive review of all the topics and to evaluate the performance of existing methods involved and make recommendations based on simulation studies that examine various realistic scenarios for rare binary events. We summarize 13 models, 11 descriptive measures, 23 estimators, 33 tests, and 16 confidence intervals in total. We not only provide synthesized information but also categorize the methods based on their key features. We find that there is no uniformly "best" estimator or inference method. However, methods with consistently better performance do exist. For the purpose of estimation, we suggest to use the improved Paule-Mandel estimator

in general situations and the Sidik and Jonkman estimator in some specific situations (i.e., extremely rare events coupled with studies of small sample sizes and existence of at least moderate-level heterogeneity) for their relatively low bias and mean squared error. The most commonly used DerSimonian and Laird estimator and its one-step variants tend to perform unsatisfactorily. For the purpose of testing the homogeneity of odds ratios, we recommend the likelihood ratio (LR) test based on the fixed-effect logistic model and the conditional LR test based on the fixed-effects hypergeometric model. For the purpose of interval estimation, we recommend the profile likelihood methods and the approximate Jackson method in general and the Sidik and Jonkman method for the specific situations mentioned above.

We consider the estimation of variance components in cluster randomized designs (CRDs) using ranked set sampling (RSS). Under the hierarchical linear model (HLM), we propose nonparametric estimators for the between and within cluster variances and explore the impact of design parameters on their performance. Simulation studies show that these RSS-based variance estimators are more efficient than the SRS-based estimator even when the ranking is imperfect. We also illustrate our proposed methods with a real data example.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

I dedicate this dissertation to my family.

CHAPTER 1

# A Comparative Study of Statistical Methods for Quantifying and Testing Between-study Heterogeneity in Meta-analysis with Focus on Rare Binary Events

## 1.1. Introduction

Meta-analysis, the statistical procedure synthesizing information from multiple studies, has been widely used in many research areas including social, psychological and especially medical sciences. Meta-analysis is a powerful tool in drug safety evaluation, where the number of cases (adverse events) can be very limited in a single study. The U.S. Food and Drug Administration (FDA) released a draft guidance for industry titled "Meta-Analyses of Randomized Controlled Clinical Trials to Evaluate the Safety of Human Drugs or Biological Products" in November 2018, which demonstrates the importance of meta-analysis in the development of new drugs. Such meta-analysis often involves binary outcomes of rare events, which are the focus of this study. A typical example is the meta-analysis of 48 trials conducted by Nissen and Wolski [57] evaluating the adverse effects of rosiglitazone, where the outcomes are myocardial infarction and death events from cardiovascular causes.

The primary goal of a meta-analysis is usually to estimate and infer the overall effect size while the variability in the effect estimates from component studies should also be properly accounted for. Besides the within-study sampling errors, the variability may come from diverse characteristics of individual studies such as disparities in trial protocols, subjects' conditions, and population characteristics, etc. When the study-wise differences exist, we call these studies heterogeneous and the heterogeneity is typically measured by a between-study variance parameter $\tau^2$. Identifying the existence of such heterogeneity can affect the

choice of model when conducting meta-analysis. Quantifying the level of heterogeneity is also important since it can affect the estimation and inference about the overall effect size. Also, descriptive measures have been widely used by clinicians to provide a more intuitive interpretation about the heterogeneity for ease of understanding.

Heterogeneity related topics not only include point and interval estimation of $\tau^2$ and hypothesis testing for homogeneity of effect sizes, but also cover statistical models used for meta-analysis and descriptive measures quantifying the level of heterogeneity. Although modeling is the foundation of a meta-analysis, very few studies in literature paid attention to this topic. For binary events, besides the traditionally used fixed-effect and random-effects models, quite a few generalized linear mixed-effects (GLMM) models have been developed in the meta-analysis framework. Thus we spend an entire section on the models, especially those for binary outcomes. Descriptive measures are widely used by clinicians to provide an intuitive interpretation about heterogeneity for ease of understanding. Currently, there is no comprehensive and updated review on this topic though there exist various measures and some recently developed ones may have advantages in certain situations over routinely used measures such as $I^2$.

For point estimation of $\tau^2$, the DerSimonian and Laird ($DL$) estimator [21], most widely used in the field, has been frequently challenged for its default use in many software packages, largely due to its sizable negative bias when the heterogeneity level is high [69, 86, 7, 61, 60]. Many modifications over the $DL$ estimator have been suggested based on the method of moments. Other approaches such as likelihood-based and other nonparametric methods can also be applied.

For hypothesis testing of homogeneity of effect sizes, besides the standard test using Cochran's $Q$ statistic [16], many other tests, such as likelihood ratio tests, score tests, and Wald tests, have been developed. Having too many choices actually makes it harder for people to decide which test to employ in solving their problems. Hence it calls for a careful review, classification and benchmarking of these methods before a plausible one can be used

to provide evidence for heterogeneity, especially for low-frequency 0/1 responses.

Confidence intervals (CIs) can be used for hypothesis testing, too, and they provide more information than a yes/no answer. For interval estimation of $\tau^2$, different types of CIs have been constructed to gauge the estimation uncertainty. However, nearly all these methods were constructed without a special consideration of dichotomous data and their performance remains unclear in the context of rare binary events, in which some of these methods may produce large bias or even fail to work.

Comparative studies/review papers exist for all the above topics except for descriptive measures. For example, Veroniki et al. [85], Langan et al. [45], Petropoulou and Mavridis [64] reviewed and compared most of the existing estimators of $\tau^2$, among which only Petropoulou and Mavridis [64] conducted simulation studies to evaluate their performance. Almalik and van den Heuvel [2] compared ten tests specifically designed for testing homogeneity of multiple $2 \times 2$ tables, but no attention was given to rare binary events. Jackson et al. [38] summarized and evaluated seven random-effects models for meta-analysis with odds ratio as effect measure. Previous comparisons about CIs [41, 87, 82] were largely limited to several similar types of CIs. We summarized the relatively comprehensive review papers in Table 1.1. None of these papers covers all the topics for heterogeneity mentioned above, nor do they focus on rare binary events. And most of them are far from being complete, some even outdated, which motivates us to conduct this study to provide useful guidance to clinicians and biostatisticians.

The first part of this dissertation is organized as follows. In Section 1.2, we introduce notation and frequently used terms in meta-analysis. In Section 1.3, we summarize and categorize existing models for meta-analysis that are either general-purpose or binary-specific. Section 1.4 reviews existing descriptive measures quantifying the level of heterogeneity. In Section 1.5, we list estimators for $\tau^2$ and briefly summarize two recently developed ones that are not included in any of the existing review papers. In Section 1.6, we thoroughly summarize over thirty hypothesis testing procedures including both general-purpose tests

and specifically designed tests for meta-analysis of dichotomous outcomes. In Section 1.7, different types of confidence intervals for $\tau^2$ are described and categorized. In Section 1.8, we compare the performance, in terms of bias and mean squared error (MSE) for point estimators, size and power for tests, and empirical coverage probability and width for CIs. Simulation studies are conducted over a large collection of scenarios designed to mimic practical situations. In Section 1.9, we re-analyze the data from a meta-analysis of 20 trials of type 2 diabetes mellitus after gestational diabetes [6] and another cohort of 56 trials of type 2 diabetes patients treated with rosiglitazone to assess risk of cardiovascular side effects [58]. The final section provides recommendations in terms of choosing appropriate estimators and inference procedures in meta-analysis of rare binary events as well as a brief discussion.

| Ref. | Model | Estimator | Hypothesis testing | Confidence interval | Simulation |
|---|---|---|---|---|---|
| Paul and Donner [62] | | | ✓ | | ✓ |
| Takkouche et al. [77] | | | ✓ | | ✓ |
| Reis et al. [66] | | | ✓ | | ✓ |
| Viechtbauer [88] | ✓ | | ✓ | | ✓ |
| Viechtbauer [87] | | | | ✓ | ✓ |
| Sidik and Jonkman [70] | | ✓ | | | ✓ |
| Kontopantelis et al. [43] | | ✓ | | | ✓ |
| Veroniki et al. [85] | | ✓ | | ✓ | |
| Langan et al. [45] | | ✓ | | | |
| Petropoulou and Mavridis [64] | | ✓ | | | ✓ |
| Jackson et al. [38] | ✓ | | | | ✓ |
| Almalik and van den Heuvel [2] | | | ✓ | | ✓ |
| Langan et al. [46] | | ✓ | | | ✓ |

Table 1.1: Existing comparative studies/reviews

## 1.2. Notation & frequently used terms

Suppose a meta-analysis includes $K$ independent studies and the $k$th study contains $n_k$ subjects ($k = 1, ..., K$). In study $k$, let $\theta_k$ be the true but unknown treatment effect and $y_k$ be the observed treatment effect such that $\mathrm{E}[y_k|\theta_k] = \theta_k$ and $\mathrm{Var}[y_k|\theta_k] = \sigma_k^2$, the within-study

variance. Typically $s_k^2$, an estimate of $\sigma_k^2$, is reported along with $y_k$ in published studies and it is often treated as a known quantity in practice (i.e., indistinguishable from $\sigma_k^2$). When the study-specific effects $\theta_k$'s are treated as random variables rather than constants, we assume $\mathrm{E}[\theta_k] = \theta$ and $\mathrm{Var}[\theta_k] = \tau^2$, where $\theta$, a parameter of main interest in the meta-analysis, represents the overall treatment effect across different studies, and $\tau^2$ measures the between-study heterogeneity. Further, for binary responses, we denote the number of events by $x_{k0}\,(x_{k1})$ and the number of subjects by $n_{k0}\,(n_{k1})$ in the control (treatment) group. The probability of having an event in the control (treatment) group is denoted by $p_{k0}\,(p_{k1})$. Effect measures for binary outcomes include risk difference (RD, $p_{k1} - p_{k0}$), risk ratio (RR, $p_{k1}/p_{k0}$) and odds ratio (OR, $[p_{k1}/(1-p_{k1})]/[p_{k0}/(1-p_{k0})]$). For rare binary events, RR$\approx$OR. A logarithm transformation of the odds ratio (LOR) is often used in meta-analysis for a much faster convergence to asymptotic normality, and the within-study variance $\sigma_k^2$ is then estimated by $s_k^2 = \frac{1}{x_{k0}} + \frac{1}{n_{k0}-x_{k0}} + \frac{1}{x_{k1}} + \frac{1}{n_{k1}-x_{k1}}$. Gart [25] added a continuity correction factor of 0.5 to all the cells so that

$$y_k = \log \frac{x_{k1} + 0.5}{n_{k1} - x_{k1} + 0.5} - \log \frac{x_{k0} + 0.5}{n_{k0} - x_{k0} + 0.5},$$

and $\sigma_k^2$ is estimated by

$$s_k^2 = \frac{1}{x_{k0} + 0.5} + \frac{1}{n_{k0} - x_{k0} + 0.5} + \frac{1}{x_{k1} + 0.5} + \frac{1}{n_{k1} - x_{k1} + 0.5},$$

which will be used in our numerical evaluation of rare binary events.

In the literature of meta-analysis, there are two main parametric models, namely $Re$ and $Fe$, to combine results from component studies. The $Re$ model assumes that $y_k = \theta_k + \epsilon_k$, where $\theta_k \sim N(\theta, \tau^2)$ and $\epsilon_k \sim N(0, \sigma_k^2)$. When $\tau^2 = 0$, it is reduced to the $Fe$ model $y_k = \theta + \epsilon_k$, where a common treatment effect $\theta$ is assumed for all component studies (i.e., $\theta_k \equiv \theta$). These models can be used with any effect measure, as long as the assumed normality is (approximately) valid.

Next, we introduce the (generalized) $Q$ statistic [20] and related terms, which will frequently appear in this chapter. For any parameter of interest, we use the corresponding letter/symbol with a hat to denote its estimate. For example, we use $\hat{\theta}$ to denote the estimate of the overall treatment effect $\theta$. The $Q$ statistic is defined as the weighted sum of squared deviations between the estimated overall treatment effect and observed treatment effect in each individual study, namely

$$Q = \sum_{k=1}^{K} w_k (y_k - \hat{\theta})^2, \tag{1.1}$$

where $w_k$ is a positive weight assigned to study $k$, and $\hat{\theta} = \sum_{k=1}^{K} w_k y_k / \sum_{k=1}^{K} w_k$, the weighted average of the estimated study-specific effects. A commonly used weighting scheme is to set $w_k = [\widehat{\text{Var}}(y_k)]^{-1}$, i.e., the inverse of the estimated variance of $y_k$. Under this inverse-variance weighing scheme, the variance of $\hat{\theta}$ can be given by $1/\sum_{k=1}^{K} w_k$ if we treat $w_k$'s as known constants (i.e., indistinguishable from $\text{Var}(y_k)]^{-1}$). Further, this scheme yields $w_k = 1/s_k^2$ for the $Fe$ model, and $w_k = 1/(s_k^2 + \hat{\tau}^2)$ for the $Re$ model, where $\hat{\tau}^2$ can be any estimator discussed in Section 1.5. Under the $Fe$ ($Re$) model with the inverse-variance weights, we denote the corresponding $Q$ statistic by $Q_{Fe}$ ($Q_{Re}$) and the corresponding $\hat{\theta}$ by $\hat{\theta}_{Fe}$ ($\hat{\theta}_{Re}$) with variance $v_{Fe}$ ($v_{Re}$). In fact, the Cochran's $Q$ statistic is $Q_{Fe}$, also known as the DerSimonian and Laird's $Q$ test statistic [21].

DerSimonian and Kacker [20] showed that if the weights $w_k$'s are treated as known constants, the expected value of $Q$ is

$$\text{E}(Q) = \tau^2 \left( \sum_{k=1}^{K} w_k - \frac{\sum_{k=1}^{K} w_k^2}{\sum_{k=1}^{K} w_k} \right) + \left( \sum_{k=1}^{K} w_k \sigma_k^2 - \frac{\sum_{k=1}^{K} w_k^2 \sigma_k^2}{\sum_{k=1}^{K} w_k} \right). \tag{1.2}$$

By equating $Q$ to its expected value, replacing $\sigma_k^2$ by $s_k^2$ in (1.2) , solving for $\tau^2$ and truncating any negative solution to zero , the generalized method of methods (GMM) estimator of $\tau^2$

can be obtained easily:

$$\hat{\tau}^2_{GMM} = \max\left\{\frac{Q - \left(\sum_{k=1}^K w_k s_k^2 - \frac{\sum_{k=1}^K w_k^2 s_k^2}{\sum_{k=1}^K w_k}\right)}{\sum_{k=1}^K w_k - \frac{\sum_{k=1}^K w_k^2}{\sum_{k=1}^K w_k}}, 0\right\}. \tag{1.3}$$

The $DL$ estimator $\hat{\tau}^2_{DL}$ [21] is a special case of $\hat{\tau}^2_{GMM}$, with $w_k = 1/s_k^2$ and $Q = Q_{Fe}$.

Throughout this dissertation, we use $\chi^2_{df}$ to denote a chi-squared distribution with $df$ degrees of freedom, and use $\chi^2_{df,\alpha}$ to denote its $100\alpha$-th percentile.

### 1.3. Statistical models for meta-analysis

We restrict our attention to meta-analysis models without covariates accounting for characteristics of different studies. Such models can be divided into two groups, generic models and binary-specific models. Generic models can be applied to any type of response, continuous or discrete. The conventional fixed-effect ($Fe$) and random-effects ($Re$) models, as described in Section 1.2, belong to this category. Viechtbauer [88] pointed out that they both are special cases of the generalized linear mixed-effects model (GLMM).

Among binary-specific models, as listed in Table 1.2, the simplest is the fixed-effect binomial ($FeB$) model that has been widely used in many earlier papers on statistical analysis of multiple $2\times2$ tables [1]. This model only assumes $x_{ki} \sim \text{Binomial}(n_{ki}, p_{ki})$ for $k = 1, \ldots, K$ and $i = 0, 1$, with all $p_{ki}$'s being unknown constants so that the treatment effect in each study $k$, as a function of $(p_{k0}, p_{k1})$, is a fixed effect. In fact, it is equivalent to the saturated model among all (fixed-effect) logistic regression models with the response variable $O$ and two explanatory variables $S$ and $Z$, where $O$ is the binary outcome of the event of interest (1 for success and 0 for failure), $S$ is a categorical variable indicating which study is involved, and $Z$ is a binary variable indicating which treatment is involved (0 for control and 1 for

treatment). The saturated model is given by

$$\text{logit}\left[P(O=1)\right] \equiv \log\frac{P(O=1)}{P(O=0)} = \mu + \sum_{k=1}^{K-1}\alpha_k I(S=k) + \theta\cdot Z + \sum_{k=1}^{K-1}\beta_k\left[Z\cdot I(S=k)\right], \quad (1.4)$$

where $I(\cdot)$ is the indicator function, $\alpha_k$ represents the effect of study $k$, $\theta$ represents the main effect of the treatment, $\beta_k$ represents the study-treatment interaction for study $k$, $\alpha_K = \beta_K = 0$ for the purpose of identifiability, and all $\mu$, $\alpha_k$, $\theta$, and $\beta_k$ are treated as constants rather than random variables. This fixed-effect logistic ($FeL$) model (1.4) has $2K$ free parameters, and so it is equivalent to $FeB$, which does not assume any reduced structure among $p_{ki}$'s.

For small-sample inference, instead of $FeB$, the fixed-effect hypergeometric ($FeH$) model is often used [48, 1]. Let $\psi_k$ denote the odds ratio in study $k$ and $x_k \equiv x_{k0} + x_{k1}$. In each table $k$, by conditioning on the row total $n_{k1}$ and the column total $x_k$, the distribution of $x_{k1}$ is a (non-central) hypergeometric distribution,

$$P(x_{k1}=t|n_{k1},x_k,n_k,\psi_k) = \frac{\binom{n_{k1}}{t}\binom{n_{k0}}{x_k-t}\psi_k^t}{\sum_{u=a_k}^{b_k}\binom{n_{k1}}{u}\binom{n_{k0}}{x_k-u}\psi_k^u},$$

where $a_k = \max\{0, x_k - n_{k0}\}$ and $b_k = \min\{x_k, n_{k1}\}$. In $FeH$, $\psi_k$ are treated as fixed effects. Liang and Self [48] considered a random-effects hypergeometric ($ReH$) model that assumes $\log\psi_k = \alpha + \tau Z_k$, where $Z_k$ are independent, identically distributed random variables with distribution $F$.

Other binary-specific models are mainly two-stage hierarchical models, among which binomial-normal ($BN$) hierarchical models are the most popular, including $BN_{BA}$ [7], $BN_{SH}$ [71], $BN_{AH}$ [2], $BN_{LW}$ [47] and $BN_{VH}$ [83]. All the $BN$ models use the LOR as the

effect measure and assume that event counts follow binomial distributions, i.e., $x_{ki} \sim Binomial(n_{ki}, p_{ki})$ for $k = 1, \cdots, K$ and $i = 0, 1$. The first two models, $BN_{BA}$ and $BN_{SH}$, assume $\text{logit}(p_{k0}) = \mu_k$ and $\text{logit}(p_{k1}) = \mu_k + \theta_k$, where $\mu_k$ represents the baseline risk of the event in each study, $\theta_k \sim N(\theta, \tau^2)$ is the log odds ratio representing the random treatment effect. The only difference between the two is that $BN_{BA}$ treats the baseline risks as random effects by assuming $\mu_k \sim N(\mu, \sigma^2)$ and $\mu_k \perp \theta_k$, while $BN_{SH}$ treats $\mu_k$'s as fixed effects. Both $BN_{BA}$ and $BN_{SH}$ implicitly assume that the variance of $\text{logit}(p_{k0})$ is not greater than the variance of $\text{logit}(p_{k1})$. This assumption is removed by $BN_{AH}$, which models $(\mu_k, \theta_k)$ by a bivariate normal distribution,

$$\begin{pmatrix} \mu_k \\ \theta_k \end{pmatrix} \sim \text{N}\left( \begin{pmatrix} \mu \\ \theta \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma\tau \\ \rho\sigma\tau & \tau^2 \end{pmatrix} \right).$$

Note that $BN_{AH}$ was recently considered in [2] as a random-effects logistic ($ReL$) regression model, which has the same form as the $FeL$ model (1.4) but assumes both $\alpha_k$ and $\beta_k$ are random with a bivariate normal distribution that has the mean vector $\mathbf{0}$ and the same covariance matrix as $(\mu_k, \theta_k)$. For this reason, we refer to $BN_{AH}$ by $ReL$ in later sections.

The fourth model, $BN_{LW}$, is flexible yet intuitive, as it introduces an additional parameter $\omega \in [0, 1]$ to allow for unequal group variability without assuming any specific direction:

$$\text{logit}(p_{k0}) = \mu_k - \omega\theta_k, \quad \text{logit}(p_{k1}) = \mu_k + (1 - \omega)\theta_k,$$

where $\mu_k \sim N(\mu, \sigma^2)$, $\theta_k \sim N(\theta, \tau^2)$, and $\mu_k \perp \theta_k$. Note that in $BN_{LW}$, $\mu_k$ no longer represents the baseline risk in study $k$. When $\omega = 0$, $BN_{LW}$ becomes $BN_{BA}$, forcing $\text{Var}[\text{logit}(p_{kc})] \leq \text{Var}[\text{logit}(p_{kt})]$. When $\omega = \frac{1}{2}$, $BN_{LW}$ becomes the model used in [74] that assumes the equality of the variances.

The fifth model, $BN_{VH}$, is the most general one, directly describing the joint distribution of the logit transformed probabilities in the treatment and control groups via a bivariate

normal distribution:

$$\begin{pmatrix} \text{logit}(p_{k0}) \\ \\ \text{logit}(p_{k1}) \end{pmatrix} \sim \text{N}\left(\begin{pmatrix} \mu \\ \\ \mu+\theta \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \rho\sigma_0\sigma_1 \\ \\ \rho\sigma_0\sigma_1 & \sigma_1^2 \end{pmatrix}\right).$$

Clearly, $BN_{LW}$ can be expressed as a special case of $BN_{VH}$, yielding

$$\begin{pmatrix} \text{logit}(p_{k0}) \\ \\ \text{logit}(p_{k1}) \end{pmatrix} \sim \text{N}\left(\begin{pmatrix} \mu \\ \\ \mu+\theta \end{pmatrix}, \begin{pmatrix} \sigma^2+\omega^2\cdot\tau^2 & \sigma^2-\omega\,(1-\omega)\,\tau^2 \\ \\ \sigma^2-\omega\,(1-\omega)\,\tau^2 & \sigma^2+(1-\omega)^2\,\tau^2 \end{pmatrix}\right).$$

Jackson et al. [38] compared seven random-effects models for meta-analysis that use the odds ratio as the effect measure, among which Model 1 is the generic *Re* model and Models 2-6 are all $BN$ models. In fact, Model 2 is $BN_{SH}$, Model 3 is $BN_{BA}$, Model 4 is a modified version of $BN_{LW}$ with $\omega = \frac{1}{2}$ and $\mu_k$'s being fixed effects, Model 5 is a special case of $BN_{LW}$ with $\omega = \frac{1}{2}$, and Model 6 is $BN_{VH}$. As mentioned before, Models 2 and 3 are the same except that one treats $\mu_k$'s as random effects and the other as fixed effects. So do Models 4 and 5. Further, Model 6 is a generalization of Models 3 and 5 as it eliminates some independence structures from the two models.

| Type* | Model | Name | Effect Measure | Specification | Ref. |
|---|---|---|---|---|---|
| **1** | Random-effects (Normal-Normal) | $Re$ | any | $y_k = \theta_k + \epsilon_k$; $\theta_k \sim \mathrm{N}(\theta, \tau^2)$, $\epsilon_k \sim \mathrm{N}(0, \sigma_k^2)$ | [21] |
| | Fixed-effect (i.e. common effect) | $Fe$ | any | $y_k = \theta + \epsilon_k$; $\epsilon_k \sim \mathrm{N}(0, \sigma_k^2)$ | |
| **2** | Fixed-effect Binomial | $FeB/FeL$ | RD/RR/OR | $x_{ki} \sim \mathrm{Binomial}(n_{ki}, p_{ki})$ for $i = 0, 1$; all $p_{ki}$'s are unknown constants | [1] |
| | Fixed-effect (non-central) Hypergeometric | $FeH$ | OR | $x_{k1} \sim \mathrm{Hypergeometric}(n_k, x_k, n_{k1}, \psi_k)$; odds ratios $\psi_k$'s are unknown constants | [48] |
| | Random-effects (non-central) Hypergeometric | $ReH$ | OR | $x_{k1} \sim \mathrm{Hypergeometric}(n_k, x_k, n_{k1}, \psi_k)$, where the log odds ratio $\log \psi_k = \alpha + \tau Z_k$, $Z_k \overset{iid}{\sim} F$ | [48] |
| | Binomial-Normal | $BN_{BA}$ | OR | $x_{ki} \sim \mathrm{Binomial}(n_{ki}, p_{ki})$ for $i = 0, 1$; $\mathrm{logit}(p_{k0}) = \mu_k$, $\mathrm{logit}(p_{k1}) = \mu_k + \theta_k$; $\mu_k \sim N(\mu, \sigma^2)$, $\theta_k \sim N(\theta, \tau^2)$ | [7] |
| | | $BN_{SH}$ | OR | $x_{ki} \sim \mathrm{Binomial}(n_{ki}, p_{ki})$ for $i = 0, 1$; $\mathrm{logit}(p_{k0}) = \mu_k$, $\mathrm{logit}(p_{k1}) = \mu_k + \theta_k$; $\mu_k$'s are fixed effects, $\theta_k \sim N(\theta, \tau^2)$ | [71] |
| | | $BN_{AH}/ReL$ | OR | $x_{ki} \sim \mathrm{Binomial}(n_{ki}, p_{ki})$ for $i = 0, 1$; $\mathrm{logit}(p_{k0}) = \mu_k$, $\mathrm{logit}(p_{k1}) = \mu_k + \theta_k$; $\begin{pmatrix} \mu_k \\ \theta_k \end{pmatrix} \sim N\left(\begin{pmatrix} \mu \\ \theta \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma\tau \\ \rho\sigma\tau & \tau^2 \end{pmatrix}\right)$ | [2] |
| | | $BN_{LW}$ | OR | $x_{ki} \sim \mathrm{Binomial}(n_{ki}, p_{ki})$ for $i = 0, 1$; $\mathrm{logit}(p_{k0}) = \mu_k - \omega\theta_k$, $\mathrm{logit}(p_{k1}) = \mu_k + (1-\omega)\theta_k$; $\mu_k \sim N(\mu, \sigma^2)$, $\theta_k \sim N(\theta, \tau^2)$, $\omega \in [0, 1]$ | [47] |
| | | $BN_{VH}$ | OR | $\begin{pmatrix} \mathrm{logit}(p_{k0}) \\ \mathrm{logit}(p_{k1}) \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mu \\ \mu + \theta \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \rho\sigma_0\sigma_1 \\ \rho\sigma_0\sigma_1 & \sigma_1^2 \end{pmatrix}\right)$ | [83] |
| | Hypergeometric-Normal | $HN$ | OR | $P(X_{k1} = x_{k1}\|\theta_k) = \frac{\binom{n_{k1}}{x_{k1}}\binom{n_{k0}}{x_{k0}} exp(\theta_k x_{k1})}{\sum_j \binom{n_{k1}}{j}\binom{n_{k0}}{x_{k0}+x_{k1}-j} exp(\theta_k j)}$, $\theta_i \sim N(\theta, \tau^2)$ | [83] |
| | Beta-Binomial | $BB$ | OR | $x_{ki} \sim \mathrm{BetaBinom}(n_{ki}, p_{ki}, \rho)$ for $i = 0, 1$, where $\rho$ is the intra-class correlation (ICC) | [4] |
| | Poisson-Gamma | $PG$ | RR | $x_{k0} \sim \mathrm{Poisson}(n_{k0}\xi_k)$, $x_{k1} \sim \mathrm{Poisson}(n_{k1}\lambda_{k1})$ $\lambda_{k1} = \xi_k \exp(\theta_k)$, $\xi_k \sim \mathrm{Gamma}(\alpha, \beta)$, $\theta_k \sim \mathrm{N}(\theta, \tau^2)$, $\xi_k$ is the baseline event rate, $\theta_k$ is the log relative risk | [12] |

* Type 1: Generic; Type 2: Binary-Specific

Table 1.2: Overview of statistical models without covariates for meta-analysis

For binary outcomes, several hierarchical models based on other distributions have also been proposed in the literature, as specified in Table 1.2. Van Houwelingen et al. [83]

suggested to use a hypergeometric distribution by conditioning on fixed row and column totals in each $2 \times 2$ table, combined with normally distributed random effects for log-odds. This $HN$ model is actually Model 7 in [38], a mixed-effects conditional logistic regression model. Bakbergenuly and Kulinskaya [4] proposed a beta-binomial ($BB$) model, which assumes a common parameter $\rho$ for both groups that can be interpreted as an intra-class correlation (ICC), to make inference about the odds ratio for over-dispersed count data. Cai et al. [12] proposed a Poisson-Gamma (PG) model to make inference about the risk ratio for meta-analysis of rare events. In Table 1.2, most binary-specific models belong to the GLMM family except for the $BN_{LW}$ and $BB$ models, due to the unknown $\omega$ ($BN_{LW}$) and the pooled $\rho$ ($BB$). Nearly all these GLMMs can be fitted with the R function glmer() in the lme4 package [38].

The above models have not taken into account the influence of moderator variables. In the presence of important study-level characteristics (e.g., measures of research or data quality in each study $k$, selected characteristics of the authors, the sample size, etc.) or differences in model specifications (e.g., whether or not potentially relevant independent variables being omitted from study $k$, differences in functional forms, types of regression, and data definitions or sources, etc.), one can add covariates that reflect such factors in these models, and so meta-analysis becomes meta-regression analysis.

### 1.4. Descriptive measures quantifying between-study heterogeneity

As mentioned in the introduction, (statistical) heterogeneity exists when true effects being evaluated differ among studies in a meta-analysis. Assessing the extent of heterogeneity is essential for model selection between $Fe$ and $Re$ models and decision making. An obvious choice is by estimating the variance parameter $\tau^2$, as is typically done in a random-effects meta-analysis. As pointed out by Higgins and Thompson [32], this measure does not facilitate comparison of heterogeneity across meta-analyses of different types of outcomes (e.g., the

survival time can be either continuous or discrete). Also, its scale is specific to a chosen effect metric and the interpretation can be difficult. For example, odds ratio is a commonly used effect measure for binary data. Still, the variance of log-odds ratio is not easy to understand for many non-statisticians. Alternatively, one may test the existence of the between-study heterogeneity (e.g., through Cochran's $Q$-test [16]), and use the corresponding test statistic or p-value to indicate the extent of heterogeneity. However, such measures depend on the scale of effect sizes or the number of component studies $K$. To overcome these limitations, effort has been devoted to development of various descriptive measures that can provide more intuitive information about the heterogeneity.

Table 1.3 summarizes 11 descriptive heterogeneity measures in the literature. Note that all these measures are general-purpose and none is specifically designed for binary outcomes. Takkouche et al. [77] proposed two measures, $R_I$ and $CV_B$, to quantify the level of heterogeneity in five published meta-analyses. The statistic $R_I$ was developed to estimate $\tau^2/(\tau^2+\sigma^2)$, the proportion of total variation in the effect estimates that is due to between-study heterogeneity. This quantity is also known as the intra-class correlation in the context of cluster sampling. Here, the within-study variances $\sigma_k^2$'s are assumed to be constant, i.e., $\sigma_k^2 \equiv \sigma^2$, which is estimated by $1/\sum_{k=1}^{K} 1/s_k^2$, making $R_I = \frac{\hat{\tau}^2}{\hat{\tau}^2+K/\sum_{k=1}^{K} 1/s_k^2}$. The other statistic $CV_B$ estimates the between-study coefficient of variation $\tau/|\theta|$ by $\sqrt{\hat{\tau}^2}/\left|\hat{\theta}\right|$. Obviously, $CV_B$ is affected by the overall treatment effect $\theta$ and is undefined when $\theta = 0$.

Under the assumption of a common within-study variance $\sigma^2$, Higgins and Thompson [32] formulated a general heterogeneity measure as a function of the overall treatment effect $\theta$, the between-study variance $\tau^2$, the within-study variance $\sigma^2$, and the number of component studies, namely, $f(\theta, \tau^2, \sigma^2, K)$. They proposed three criteria that such a measure should satisfy in general in order to facilitate its comparability and interpretability, including (i) dependence on the extent of heterogeneity, (ii) scale invariance, i.e. $f(\theta, \tau^2, \sigma^2, K) = f(a + b\theta, b^2\tau^2, b^2\sigma^2, K)$ for any $a$ and $b$, and (iii) size invariance, i.e. $f(\theta, \tau^2, \sigma^2, K_1) = f(\theta, \tau^2, \sigma^2, K_2)$ for any positive integers $K_1$ and $K_2$. Criterion (i) implies that the function

$f$ should increase monotonically with $\tau^2$. Criterion (ii) implies that $f$ should be a function of the ratio $\rho \equiv \frac{\tau^2}{\sigma^2}$ and that $\theta$ should not be involved. Criterion (iii) implies that $f$ does not depend on $K$. It can be shown that any monotonically increasing function of $\rho$ satisfies the three criteria. Based on this, three statistics, $H^2$, $R^2$ and $I^2$ were proposed. The first, $H^2$, estimates the quantity $\rho + 1$ by equating the observed value of $Q_{Fe}$ to its expectation so that $H^2 = \frac{Q_{Fe}}{K-1}$ can be interpreted as relative excess in $Q_{Fe}$ over its expected value, the degrees of freedom $K - 1$. The second, $R^2$, attempts to estimate $\rho + 1$ as well; but here, $\rho+1$ is approximated by $v_{Re}/v_{Fe}$ so that $R^2 = \hat{v}_{Re}/\hat{v}_{Fe} = \sum_{k=1}^{K} \frac{1}{s_k^2} / \sum_{k=1}^{K} \frac{1}{s_k^2+\hat{\tau}^2}$, which can be interpreted as the inflation in the confidence interval for $\hat{\theta}_{Re}$ under the $Re$ model compared with $\hat{\theta}_{Fe}$ under the $Fe$ model. Both $H^2$ and $R^2$ should be at least 1, where 1 means perfect homogeneity; and the larger the value, the more heterogeneous the studies. In practice, the authors suggested to use $H$ and $R$ because clinicians may be more familiar with standard deviations than variances. The third statistic, $I^2$, estimates a different function of $\rho$, i.e. $\frac{\rho}{1+\rho} = \frac{\tau^2}{\tau^2+\sigma^2}$, which represents the proportion of total variance that is due to between-study variation. Higgins and Thompson [32] suggested to compute $I^2$ by $I^2_{HT} = 1 - \frac{K-1}{Q_{Fe}}$, which leads to a convenient relationship $I^2_{HT} = 1 - \frac{1}{H^2}$. Jackson et al. [39] suggested to compute $I^2$ by $I^2_R = 1 - \frac{\hat{v}_{Fe}}{\hat{v}_{Re}} = 1 - \sum_{k=1}^{K} \frac{1}{s_k^2+\hat{\tau}^2} / \sum_{k=1}^{K} \frac{1}{s_k^2}$, which leads to another convenient relationship $I^2_R = 1 - \frac{1}{R^2}$. Both $I^2_{HT}$ and $I^2_R$ are usually expressed as percentages between 0% and 100%, where a value of 0% corresponds to no observed heterogeneity, while larger values indicate increasing levels of heterogeneity. They estimate the same quantity as $R_I$ does, but with different within-study variance estimates. Among these measures (i.e. $H^2$, $R^2$, $I^2_{HT}$ or $I^2_R$), $I^2_{HT}$ is most popular and in the literature, $I^2$ typically represents $I^2_{HT}$ as $I^2_R$ is much less known. Higgins et al. [33] empirically provided a rough guide to the interpretation of $I^2$ using overlapping intervals: a value in [0,0.4] suggests that heterogeneity may not be that important; [0.3, 0.6] may represent moderate heterogeneity; [0.5,0.9] may represent substantial heterogeneity; and [0.75,1] implies considerable heterogeneity.

The assumption of a constant within-study variance is probably untrue in many real life data. Thus, Crippa et al. [18] lifted this assumption and proposed a new measure $R_b$,

defined as $R_b = \frac{1}{K}\sum_{k=1}^{K}\frac{\hat{\tau}^2}{s_k^2+\hat{\tau}^2}$, to assess the contribution of the between-study variance $\tau^2$ to $v_{Re}$ (i.e., the variance of the pooled random-effects estimate $\hat{\theta}_{Re}$). It can be viewed as an average of the study-specific proportions of the study-specific variances due to between-study heterogeneity. They showed that the quantity $\tau^2/v_{Re}$ underlying $R_b$ is a strictly increasing function of $\tau^2$ and is scale-invariant. However, this quantity depends on $K$ and so is not size-invariant. They further showed that $R_I \geq \max(R_b, I^2_{HT})$. When $\sigma_k^2 \equiv \sigma^2$ and $\sigma^2$ is estimated by $s^2$, $R_b$, $R_I$, $I^2_{HT}$ and $I^2_R$ all yield the same quantity $\frac{\hat{\tau}^2}{s^2+\hat{\tau}^2}$. The authors conducted a simulation study to examine the performance of $R_I$, $I^2_{HT}$ and $R_b$. Both $R_I$ and $I^2_{HT}$ tend to be positively biased and this overestimation increases as $K$ increases. Confidence intervals based on $R_I$ and $I^2_{HT}$ give lower coverage probabilities compared to those based on $R_b$ and the difference becomes more obvious when the within-study variances vary more and when the heterogeneity level increases.

To reduce the impact of outlying studies, Lin et al. [49] proposed new robust measures $H^2_r$, $H^2_m$, $I^2_r$ and $I^2_m$, which are analogous to and have the same interpretations as $H^2$ and $I^2$, respectively. These methods were developed upon the absolute deviation measures $Q_r$ and $Q_m$ rather than the usual squared deviation measure $Q$, as defined in Table 1.3 and will be described in more detail in Section 1.5.

All the measures except for $CV_B$ depend on the precision of the study-specific effects. As the sample sizes of the component studies increase, $\sigma_k^2$'s would decrease to zero so that $R_I$, $R_B$ and all $I^2$'s would increase to 1 and all $H^2$'s and $R^2$ would become arbitrarily large, even when there is little between-study heterogeneity. The measure $CV_B$ avoids this drawback but has its own limitation: it would approach $+\infty$ as $\theta$ goes to 0. Finally, we mention that some of the measures involve the estimated value $\hat{\tau}^2$. In principle, $\hat{\tau}^2$ can be any estimator of $\tau^2$, but most software uses the $DL$ estimator $\hat{\tau}^2_{DL}$ as the default choice.

| Name | $f(\theta, \tau^2, \sigma^2, K)$ | Formula | Ref. | Interpretation | Assume $\sigma_k^2 \equiv \sigma^2$? |
|---|---|---|---|---|---|
| $R_I$ | $\frac{\tau^2}{\tau^2+\sigma^2}$ | $\frac{\hat{\tau}^2}{\hat{\tau}^2+K/\sum_{k=1}^K 1/s_k^2}$ | [77] | Proportion of total variation in the estimates of treatment effect due to between-study heterogeneity | Yes |
| $CV_B$ | $\frac{\tau}{\lvert\theta\rvert}$ | $\frac{\sqrt{\hat{\tau}^2}}{\lvert\hat{\theta}\rvert}$ | [77] | Between-study coefficient of variation | No |
| $H^2$ | $\frac{\tau^2+\sigma^2}{\sigma^2}$ | $\frac{Q_{Fe}}{K-1}$ | [32] | Relative excess in $Q_{Fe}$ over its degrees of freedom | Yes, but can be used for different $\sigma_k^2$. |
| $R^2$ | $\frac{\tau^2+\sigma^2}{\sigma^2} \approx \frac{v_{Re}}{v_{Fe}}$ | $\frac{\sum_{k=1}^K \frac{1}{s_k^2}}{\sum_{k=1}^K \frac{1}{s_k^2+\hat{\tau}^2}}$ | [32] | Inflation in the confidence interval for a single summary estimate under $Re$ model compared with $Fe$ model | Yes, but can be used for different $\sigma_k^2$. |
| $I^2_{HT}$ | $\frac{\tau^2}{\tau^2+\sigma^2}$ | $1-\frac{K-1}{Q_{Fe}}$ | [32] | Same as $R_I$ | Yes |
| $I^2_R$ | $\frac{\tau^2}{\tau^2+\sigma^2}$ | $1-\frac{\sum_{k=1}^K \frac{1}{s_k^2+\hat{\tau}^2}}{\sum_{k=1}^K \frac{1}{s_k^2}}$ | [39] | Same as $R_I$ | Yes |
| $R_b$ | $\frac{\tau^2}{v_{Re}} \approx \frac{1}{K}\sum_{k=1}^K \frac{\tau^2}{\sigma_k^2+\tau^2}$ | $\frac{1}{K}\sum_{k=1}^K \frac{\hat{\tau}^2}{s_k^2+\hat{\tau}^2}$ | [18] | Proportion of the between-study heterogeneity $\tau^2$ relative to $v_{Re}$, the variance of $\hat{\theta}_{Re}$. | No |
| $H_r^2$ | $\frac{\tau^2+\sigma^2}{\sigma^2}$ | $\frac{\pi Q_r^2}{2K(K-1)}$, $Q_r = \sum_{k=1}^K \frac{1}{s_k}\left\lvert y_k - \hat{\theta}_{Fe}\right\rvert$ | [49] | Same as $H^2$ | Yes |
| $I_r^2$ | $\frac{\tau^2}{\tau^2+\sigma^2}$ | $1-\frac{2K(K-1)}{\pi Q_r^2}$ | [49] | Same as $R_I$ | Yes |
| $H_m^2$ | $\frac{\tau^2+\sigma^2}{\sigma^2}$ | $\frac{\pi Q_m^2}{2K^2}$, $Q_m = \sum_{k=1}^K \frac{1}{s_k}\left\lvert y_k - \hat{\theta}_m\right\rvert$, $\hat{\theta}_m$ is weighted median estimate | [49] | Same as $H^2$ | Yes |
| $I_m^2$ | $\frac{\tau^2}{\tau^2+\sigma^2}$ | $\frac{Q_m^2-2K^2/\pi}{Q_m^2}$ | [49] | Same as $R_I$ | Yes |

Table 1.3: Descriptive measures quantifying the between-study heterogeneity

## 1.5. Estimators

We summarize 23 estimators for $\tau^2$ in Table 1.4, among which most can be applied to all kinds of effect measures except for the improved Paule and Mandel estimator ($IPM$, [7]) and Malzahn, Böhning, and Holling ($MBH$, [54]). $IPM$ is specifically designed to work with OR for binary outcomes, and $MBH$ can be only used for standardized mean difference (SMD).

| Estimators | Abbreviation | Reference | Iterative? | Sign | Effect Measure |
|---|---|---|---|---|---|
| ***Method of Moments*** | | | | | |
| Hedges and Olkin | $HO$ | Hedges and Olkin [31] | No | $>= 0$ | |
| Two-step Hedges and Olkin | $HO_2$ | DerSimonian and Kacker [20] | No | $>= 0$ | |
| DerSimonian and Laird | $DL$ | DerSimonian and Laird [21] | No | $>= 0$ | |
| Positive DerSimonian and Laird | $DL_p$ | Kontopantelis et al. [43] | No | $> 0$ | |
| Two-step DerSimonian and Laird | $DL_2$ | DerSimonian and Kacker [20] | No | $>= 0$ | |
| **Multi-step Dersimonian and Laird** | $DL_k$ | van Aert and Jackson [81] | No | $>= 0$ | |
| Paule and Mandel | $PM$ | Paule and Mandel [63] | Yes | $>= 0$ | |
| Improved Paule and Mandel | $IPM$ | Bhaumik et al. [7] | Yes | $>= 0$ | OR |
| Hartung and Makambi | $HM$ | Hartung and Makambi [30] | No | $> 0$ | |
| Hunter and Schmidt | $HS$ | Hunter and Schmidt [34] | No | $>= 0$ | |
| **Lin, Chu and Hodges** | $LCH$ | Lin et al. [49] | No | $>= 0$ | |
| ***Likelihood-based*** | | | | | |
| Maximum Likelihood | $ML$ | Hardy and Thompson [28] | Yes | $>= 0$ | |
| Restricted maximum likelihood | $REML$ | Viechtbauer [86] | Yes | $>= 0$ | |
| Approximate restricted maximum likelihood | $AREML$ | Morris [56] | Yes | $>= 0$ | |
| ***Model error variance (Least squared)*** | | | | | |
| Sidik and Jonkman | $SJ$ | Sidik and Jonkman [69] | No | $> 0$ | |
| Sidik and Jonkman ($HO$ prior) | $SJ_{HO}$ | Sidik and Jonkman [70] | No | $> 0$ | |
| ***Bayesian*** | | | | | |
| Rukhin Bayes | $RB_0$ | Rukhin [67] | Yes | $>= 0$ | |
| Positive Rukhin Bayes | $RB_p$ | Rukhin [67] | Yes | $> 0$ | |
| Empirical Bayes (Equivalent to PM) | $EB$ | Morris [56] | Yes | $>= 0$ | |
| Fully Bayes | $FB$ | Smith et al. [74] | Yes | $> 0$ | |
| Bayes Modal | $BM$ | Chung et al. [15, 14] | Yes | $> 0$ | |
| ***Other nonparametric*** | | | | | |
| Malzahn, Böhning, and Holling | $MBH$ | Malzahn et al. [54] | No | $>= 0$ | SMD |
| Non-parametric bootstrap DerSimonian and Laird | $DL_b$ | Kontopantelis et al. [43] | No | $>= 0$ | |

Table 1.4: Overview of 23 estimators for the between-study variance $\tau^2$

Table 1.5 shows previous studies that reviewed and compared (large) subsets of these estimators. Recommendations were made either based on their own simulations or conclusions from the literature. Among them, Veroniki et al. [85], Langan et al. [45] and Petropoulou and Mavridis [64] are more comprehensive. Veroniki et al. [85] reviewed 17 estimators as listed in Table 1.5, including Hedges and Olkin ($HO$, [31]), two-step $HO$ ($HO_2$, [20]), $DL$ [21], two-step $DL$ ($DL_2$, [20]), positive $DL$ ($DL_p$, [43]), nonparametric bootstrap $DL$ ($DL_b$, [43]), Paule and Mandel ($PM$, [63]), Hartung and Makambi ($HM$, [30]), Hunter and Schmidt ($HS$, [34]), Maximum Likelihood ($ML$, [28]), Restricted Maximum Likelihood ($REML$, [86]), Ap-

proximate $REML$ ($AREML$, [56]), Sidik and Jonkman ($SJ$, [69]), Rukhin Bayes ($RB$) with the default (zero-mean) prior ($RB_0$, [67]), positive $RB$ ($RB_p$, [67]), Fully Bayes ($FB$, [74]), and Bayes Modal ($BM$, [14, 15]). Langan et al. [45] and Petropoulou and Mavridis [64] added $IPM$, $MBH$, and $SJ$ with the $HO$ prior ($SJ_{HO}$, [70]) into comparison. Note that $IPM$ was briefly summarized but not compared with other estimators in Veroniki et al. [85]. Also, the Empirical Bayes method ($EB$, [56]) mentioned in [64] has been shown to be equivalent to $PM$. Langan et al. [45] also added $RB$ estimators with different priors, $RB_u$ and $RB_a$.

| Review paper | Estimators compared | Effect measure | Recommendations |
|---|---|---|---|
| Viechtbauer [86] | $HO$, $DL$, $HS$, $ML$, $REML$ | SMD and MD | $REML$ |
| Sidik and Jonkman [70] | $HO$, $DL$, $SJ$, $SJ_{HO}$, $ML$, $REML$, $EB$ | OR | $SJ_{HO}$ when $\tau^2$ is expected to be small or moderate; $SJ$ when $\tau^2$ is expected to be large. |
| Kontopantelis et al. [43] | $HO$, $HO_2$, $DL$, $DL_2$, $DL_b$, $DL_p$, $SJ$, $SJ_{HO}$, $ML$, $RB$, $RB_p$ | Generic | $DL_b$ |
| Veroniki et al. [85] | $HO$, $HO_2$, $DL$, $DL_2$, $DL_p$, $DL_b$, $PM$, $HM$, $HS$, $ML$, $REML$, $AREML$, $SJ$, $RB$, $RB_p$, $FB$, $BM$ | Generic | $PM$ |
| Langan et al. [45] | Estimators in [85] except for $FB$ plus $IPM$, $SJ_{HO}$, $RB_u$, $RB_a$, $MBH$ | RR, OR, SMD, MD and Generic | $PM$ |
| Petropoulou and Mavridis [64] | Estimators in [45] except for $RB_u$, $RB_a$ | OR and MD | $DL_b$ and $DL_p$ |
| Langan et al. [46] | $DL$, $HO$, $PM$, $PM_{HO}$, $PM_{DL}$, $HM$, $SJ$, $SJ_{HO}$, $REML$ | OR and Generic | $REML$, $PM$ and $PM_{DL}$ for continuous outcomes and non-rare binary events |

Table 1.5: Existing comparative studies for various estimators of the between-study variance $\tau^2$.

All the estimators can be divided into five groups: method of moments, likelihood-based, model error variance (least square), Bayes, and other nonparametric estimators. Some of these estimators have closed form expressions while the others require iterative solutions. Some methods provide positive estimates naturally while the others require truncation to zero when a negative value occurs. All these properties of the estimators are summarized in Table 1.4. Two newly proposed estimators, the multi-step $DL$ estimator $DL_m$ [81] and the $LCH$ estimators [49], are included in our pool. We mark them in bold in Table 1.4 and provide a brief description for each in the paragraph below. The $IPM$ estimator [7] is described as well because it is the only method specifically designed for rare binary events. More details about other estimators can be found in [85] and references therein.

**Lin, Chu and Hodges (LCH)** Lin et al. [49] proposed two alternative estimators, $\hat{\tau}_r^2$ and $\hat{\tau}_m^2$, designed to be less affected by outliers than conventional estimators based on the $Q$ statistics in (1.1). For the purpose of robustness, they are based on $Q_r$ and $Q_m$, defined as the weighted sums of absolute differences between the study-specific treatment effects and the overall treatment effect, namely

$$Q_r = \sum_{k=1}^{K} \frac{1}{s_k} \left| y_k - \hat{\theta}_{Fe} \right|, \quad Q_m = \sum_{k=1}^{K} \frac{1}{s_k} \left| y_k - \hat{\theta}_m \right|.$$

Here, $\hat{\theta}_{Fe} = \sum_{k=1}^{K} \frac{y_k}{s_k^2} / \sum_{k=1}^{K} \frac{1}{s_k^2}$ is the fixed-effect estimate of $\theta$ as defined in Section 1.2, and $\hat{\theta}_m$ is the weighted median estimator that is the solution to the equation $\sum_{k=1}^{K} w_k \left[ I(\theta \geq y_i) - 0.5 \right] = 0$, where $I(\cdot)$ is the indicator function. The estimators $\hat{\tau}_r^2$ and $\hat{\tau}_m^2$ , based on $Q_r$ and $Q_m$, respectively, can be derived similarly as the $DL$ estimator $\hat{\tau}_{DL}^2$ by equating observed $Q_r$ and $Q_m$ to their corresponding expected values.

**Multistep** $DL$ As discussed in Section 1.2, the inverse-variance weighing scheme yields $w_k = 1/(s_k^2 + \hat{\tau}^2)$ when calculating the (generalized) $Q$ statistic (1.1) under the $Re$ model. Recall that the original $DL$ estimator $\hat{\tau}_{DL}^2$ can be obtained by specifying $w_k = 1/s_k^2$ in (1.3), which is equivalent to setting $\hat{\tau}^2 = 0$ in the $Re$ weights. The two-step $DL$ method [20] first obtains $\hat{\tau}_{DL}^2$ and then sets $\hat{\tau}^2 = \hat{\tau}_{DL}^2$ in the $Re$ weights to obtain $\hat{\tau}_{DL_2}^2$ from (1.3).

van Aert and Jackson [81] proposed the multistep $DL$ estimator as a natural extension of the two-step $DL$ estimator. The $m$-step $DL$ estimator $\hat{\tau}_{DL_m}^2$ can be obtained recursively by computing $\hat{\tau}_{DL}^2$, $\hat{\tau}_{DL_2}^2$, ..., $\hat{\tau}_{DL_m}^2$ using (1.3). It has been shown that the limit of the multistep $DL$ estimator, $\hat{\tau}_{DL_\infty}^2$, when it exists, is equivalent to the $PM$ estimator. As further suggested by the authors, divergence problems seldom happen in practice and the convergence is usually achieved quickly.

**Improved Paule and Mandel ($IPM$)** For meta-analysis of rare binary events, Bhaumik et al. [7] adopted a standard binomial-normal random-effects model (labeled $BN_{BA}$), which can be specified by

$$
\begin{aligned}
&x_{ki} \sim \text{Binomial}(n_{ki}, p_{ki}) \text{ for } i = 0, 1; \\
&\text{logit}(p_{k0}) = \mu_k, \text{logit}(p_{k1}) = \mu_k + \theta_k; \\
&\mu_k \sim N(\mu, \sigma^2), \theta_k \sim N(\theta, \tau^2), \mu_k \perp \theta_k \text{ for } k = 0, \ldots, K.
\end{aligned}
$$

They proposed a simple average estimator, $\hat{\theta}_{sa}$, for the overall treatment effect $\theta$ and then developed the $IPM$ estimator for $\tau^2$ based on $\hat{\theta}_{sa}$ and the iterative $PM$ method. The treatment effect $\theta_k$ (measured by log-odds ratio) in study $k$ is estimated with a correction factor $a$ added to each cell count, namely, $y_{ka} = \log\left[(x_{k1} + a)/(n_{k1} - x_{k1} + a)\right] - \log\left[(x_{k0} + a)/(n_{k0} - x_{k0} + a)\right]$ . The simple average estimator for $\theta$ is then given by $\hat{\theta}_{sa} = \sum_{k=1}^{K} y_{ka}/K$. The authors further proved that $a$ should be $\frac{1}{2}$ in order for $\hat{\theta}_{sa}$ to be the least biased for large samples. They noticed that the $PM$ estimator for $\tau^2$ depends on $s_k^2$ and proposed to improve $PM$ by borrowing strength from all component studies when estimating each within-study variance,

$$
s_k^2(*) = \frac{1}{n_{k1} + 1}\left[\exp\left(-\hat{\mu} - \hat{\theta}_{s_{\frac{1}{2}}} + \frac{\tau^2}{2}\right) + 2 + \exp\left(\hat{\mu} + \hat{\theta}_{s_{\frac{1}{2}}} + \frac{\tau^2}{2}\right)\right] + \frac{1}{n_{k0} + 1}\left[\exp(-\hat{\mu}) + 2 + \exp(\hat{\mu})\right].
$$

Denote the corresponding weights by $w_k(*) \equiv 1/[s_k^2(*) + \tau^2]$ and $\hat{\tau}_{IPM}^2$ can be obtained by solving $Q - (K - 1) = 0$ iteratively with weights $w_k(*)$ in the calculation of $Q$.

### 1.6. Hypothesis testing procedures

A central issue in meta-analysis is the selection of an appropriate statistical model to characterize individual effects of component studies. Different model assumptions can lead to different or even contrary conclusions about the overall treatment effect. For example,

contrary conclusions were made when conducting meta-analysis of clinical trials in [57] about the side effect of rosiglitazone on myocardial infarction (MI). By assuming homogeneous treatment effects, the exact approach by [51] gives a $p$-value of 0.029 while the simple average (SA) method proposed by [7] under the model $BN_{BA}$ provides a $p$-value of 0.463, reported in [3], when testing whether there exists any effect of rosiglitazone on MI.

| Ref. | Test statistics compared | Effect measure | Recommendations |
|---|---|---|---|
| Jones et al. [40] | $LR_{U.FeL}$, $US_{FeL}$, $CS$, $BD$, $MBD$, $Z_{CS.FeH}$, $CS_{ReH}$ | OR | Only cases of $n_{k0} = n_{k1}$ examined; use $BD$ when $n_{k1} \equiv n$; use $Z_{CS.FeH}$ and $CS_{ReH}$ when $n_{k1}$ varies. |
| Paul and Donner [62] | $LR_{U.FeL}$, $LR_C$, $US_{FeL}$, $AUS_{FeL}$, $CS$, $ACS$, $MDB$, $AMDB$, $Q_G$ | OR | $AMDB$ in all cases and $Q_G$ for the balanced design ($n_{k0} = n_{k1} = n$) with large samples. |
| Takkouche et al. [77] | $Q_{Fe}$, $Z^2_{WLS}$, $Z^2_{WLS,R}$, $Z^2_K$, $LR_{ML}$, parametric bootstrap versions of these tests and $\tau^2_{DL}$-bootstrap | OR | Cochran's $Q$-test |
| Reis et al. [66] | $BD$, $US_{FeL}$, $CS_{FeH}$, $LR_{U.FeL}$, $LR_C$, $Peto$ | OR | $US_{FeL}$ and $BD$ |
| Viechtbauer [88] | $Q_{Fe}$, $W_{ML}$, $W_{REML}$, $LR_{ML}$, $LR_{REML}$, $S_{ML}$, $S_{REML}$ | (standardized) MD, (Fisher transformed) correlation | $Q$-test in terms of Type I error rate (requiring each $n_{ki}$ to be large) |
| Almalik and van den Heuvel [2] | $LR_{U.FeL}$, $LR_{U.ReL}$, $Q_{Fe}$, $Q_B$, $MBD$, $US_{FeL}$, $US_{ReL}$, $CS$, $Peto$ | OR | Use $MBD$ and avoid $Q_{Fe}$, $Q_B$, $Peto$, $LR_{U.ReL}$ and $US_{ReL}$. |

Table 1.6: Existing comparative studies for hypothesis testing procedures. Test statistics are defined in Table 1.7.

The above example emphasizes the need for a test on homogeneity of treatment effects, a topic with abundant research in the past. Two types of approaches have been commonly used for this purpose. One is to test $\mathcal{H}_0 : \theta_1 = \theta_2 = \cdots = \theta_K$ when the model treats $\theta_k$s as fixed effects (e.g., $FeB$, $FeH$), and the other is to test $\mathcal{H}_0 : \tau^2 = 0$, when the model treats $\theta_k$s as random effects with variance $\tau^2$ (e.g., $Re$, $ReH$, $BN_{BA}$, $BN_{AH}$). We summarize existing review papers about testing the homogeneity in Table 1.6 in terms of the test statistics compared, effect measures and recommendations. Clearly, each of the six review papers only includes a limited number of tests. So far, no comprehensive review has been done in the literature about different testing procedures, and their performance in meta-analysis of rare binary data has not been systematically evaluated.

| Method | Test Statistic | Asymptotic Null Dist. | Conditions | Response/Effect Measure | Model | Ref. |
|---|---|---|---|---|---|---|
| ***(Modified) Q-test*** | | | | | | |
| (Cochran's) $Q$-test | $Q_{Fe}$ | $\chi^2_{K-1}$ | $n_{ki} \to \infty$, $K$ fixed | Generic/Any | $Re$ | [16] |
| Modified $Q$-test | $\tilde{Q}_r$ | $\chi^2_{K-1}$ | $n_{ki} \to \infty$, $K$ fixed | Generic/Any | $Re$ | [41] |
| Jackson | $Q_J$ | Lin. comb. of $\chi^2_1$s | $n_{ki} \to \infty$, $K$ fixed | Generic/Any | $Re$ | [35] |
| Gart | $Q_G$ | $\chi^2_{K-1}$ | $n_{ki} \to \infty$, $K$ fixed | Binary/OR | $FeB$ | [25] |
| Bliss | $Q_B$ | $\chi^2_{K-1}$ | $n_{ki} \to \infty$, $K$ fixed | Binary/OR | $FeB$ | [10, 2] |
| Bhaumik's $T_3$ based on $Q_{Re}$ | $Bh.T_3$ | $N(0,1)$ | $n_{ki} \to \infty$, $K \to \infty$ | Binary/OR | $BN_{BA}$ | [7] |
| Kulinskaya's $Q_\gamma$ | $Q_G$ | $Ga(\alpha, \beta)$ | $n_{ki} \to \infty$, $K$ fixed | Binary/OR | $FeB$ | [44] |
| ***Likelihood ratio tests*** | | | | | | |
| MLE under $Re$ | $LR_{ML}$ | $0.5\chi^2_0 + 0.5\chi^2_1$ | $n_{ki} \to \infty$, $K \to \infty$ | Generic/Any | $Re$ | [84, 88] |
| REML under $Re$ | $LR_{REML}$ | $0.5\chi^2_0 + 0.5\chi^2_1$ | $n_{ki} \to \infty$, $K \to \infty$ | Generic/Any | $Re$ | [84, 88] |
| Unconditional under $FeL$ | $LR_{U.FeL}$ | $\chi^2_{K-1}$ | $n_{ki} \to \infty$, $K$ fixed | Binary/OR | $FeB$ | [62] |
| Unconditional under $ReL$ | $LR_{U.ReL}$ | $0.5\chi^2_0 + 0.5\chi^2_2$ | $n_{ki} \to \infty$, $K \to \infty$ | Binary/OR | $BN_{AH}$ | [2] |
| Conditional likelihood | $LR_C$ | $\chi^2_{K-1}$ | $n_{ki} \to \infty$, $K$ fixed | Binary/OR | $FeH$ | [62] |
| ***(Modified)Score tests*** | | | | | | |
| MLE under $Re$ | $S_{ML}$ | $N(0,1)$ | $n_{ki} \to \infty$, $K \to \infty$ | Generic/Any | $Re$ | [84, 88] |
| REML under $Re$ | $S_{REML}$ | $N(0,1)$ | $n_{ki} \to \infty$, $K \to \infty$ | Generic/Any | $Re$ | [84, 88] |
| (Approx.)Unconditional under $FeL$ | $(A)US_{FeL}$ | $\chi^2_{K-1}$ | $n_{ki} \to \infty$, $K$ fixed | Binary/OR | $FeB$ | [5, 59, 62] |
| (Approx.) Unconditional under $ReL$ | $(A)US_{ReL}$ | $\chi^2_{K-1}$ | $n_{ki} \to \infty$, $K$ fixed | Binary/OR | $BN_{AH}$ | [2] |
| (Approx.) Conditional under $FeH$ | $(A)CS_{FeH}$ | $\chi^2_{K-1}$ | $n_{ki} \to \infty$, $K$ fixed | Binary/OR | $FeH$ | [48, 62] |
| (Normal approx.) Conditional under $FeH$ | $Z_{CS.FeH}$ | $N(0,1)$ | $K \to \infty$, $n_{ki}$s fixed | Binary/OR | $FeH$ | [48, 40] |
| Conditional under $ReH$ | $CS_{ReH}$ | $N(0,1)$ | $K \to \infty$, $n_{ki}$s fixed | Binary/OR | $ReH$ | [48, 40] |
| Breslow-Day test | $BD$ | $\chi^2_{K-1}$ | $n_{ki} \to \infty$, $K$ fixed | Binary/OR | $FeB$ | [11] |
| (Approx.) Modified Breslow-Day | $(A)MBD$ | $\chi^2_{K-1}$ | $n_{ki} \to \infty$, $K$ fixed | Binary/OR | $FeB$ | [78] |
| ***Wald tests*** | | | | | | |
| MLE under $Re$ | $W_{ML}$ | $N(0,1)$ | $n_{ki} \to \infty$, $K \to \infty$ | Generic/Any | $Re$ | [88] |
| REML under $Re$ | $W_{REML}$ | $N(0,1)$ | $n_{ki} \to \infty$, $K \to \infty$ | Generic/Any | $Re$ | [88] |
| $FeL$ | $W^2_{FeL}$ | $\chi^2_{K-1}$ | $n_{ki} \to \infty$, $K$ fixed | Binary/OR | $FeB$ | |
| ***Others*** | | | | | | |
| Peto | $Peto$ | $\chi^2_{K-1}$ | $n_{ki} \to \infty$, $K$ fixed | Binary/OR | $FeB$ | [92] |
| Lipsitz tests | $Z^2_{WLS}$, $Z^2_{WLS,R}$ | $F_{1,K-1}$ | $n_{ki} \to \infty$, $K \to \infty$ | Binary/RD,RR,OR | $FeB$ | [50, 77] |
| Lipsitz test | $Z^2_V$ | $F_{1,K-1}$ | $n_{ki} \to \infty$, $K \to \infty$ | Binary/RD | $FeB$ | [50] |
| Lipsitz test | $Z^2_K$ | $F_{1,K-1}$ | $n_{ki} \to \infty$, $K$ fixed | Binary/RD,RR,OR | $FeB$ | [50, 77] |
| Bhaumik's $T_4$ based on SA | $Bh.T_4$ | $N(0,1)$ | $n_{ki} \to \infty$, $K \to \infty$ | Binary/OR | $BN_{BA}$ | [7] |
| Bayesian by DIC | — | — | — | Binary/OR | $BN_{BA}$ | [3] |
| Bootstrap | Any | — | — | Generic/Any | Any | [23, 43] |
| Generalized variable approach | $R_{\tau^2}$ | — | — | Normal/MD | $Re$ | [80] |

Table 1.7: Tests of between-study homogeneity in meta-analysis.

Table 1.7 summarizes over thirty methods for testing the homogeneity that can be applied to meta-analysis of dichotomous events except for the generalized variable approach in [80] that is designed for the measure mean difference (MD) but is nonetheless included herein for completeness. These tests can be grouped into five categories as discussed below. For each test, Table 1.7 presents key information including the test statistic, asymptotic null distribution and conditions, the data type and effect measure, as well as the model under which the the test is derived.

### 1.6.1. Tests based on (modified) $Q$ statistics

One of the most commonly used testing procedures is based on Cochran's $Q$ statistic $Q_{Fe} = \sum_{k=1}^{K} \frac{1}{s_k^2}(y_k - \hat{\theta}_{Fe})^2$. According to [16], $Q_{Fe}$ follows a $\chi^2_{K-1}$ distribution (asymptotically) under the null hypothesis $\mathcal{H}_0 : \tau^2 = 0$. One rejects $\mathcal{H}_0$ if $Q_{Fe} > \chi^2_{K-1,1-\alpha}$. This procedure is referred to as the (standard) $Q$-test. Many later proposed testing procedures were developed by modifying a $Q$ statistic taking the form of formula (1.1) such as $Q_{Fe}$ and $Q_{Re}$ to make the null distribution better approximated by the $\chi^2_{K-1}$ distribution or by applying different asymptotic theories about the null distribution.

**Generic tests** Among random-effects meta-analysis using the generic $Re$ model described in Section 1.2, Knapp et al. [41] considered $Q$ in (1.1) with weights $w_k = 1/(\tau^2 + s_k^2)$, and proposed a regularization function for $w_k$s in order to better approximate the distribution of (regularized) $Q$ by $\chi^2_{K-1}$. The regularization term $r_k$ is derived through a moment matching approach based on approximating the distribution of $\tau^2 + s_k^2$ by a scaled $\chi^2$ distribution [29] and the new regularized weight is $w_{rk} = r_k w_k$. The resulting sum of weighted squared deviation based on $w_{rk}$s, denoted by$\tilde{Q}_r(\tau^2)$, is referred to as the modified $Q$ statistic, which can be inverted to produce confidence intervals for $\tau^2$ as described in [41]. When the resulting $100(1-\alpha)\%$ confidence intervals do not include zero, reject the null at the significance level $\alpha$.

Jackson [35] proved that under the $Re$ model, the generalized $Q$ in (1.1) is distributed as a positive linear combination of $\chi^2_1$ random variables, whose cumulative distribution function can be numerically obtained using Farebrother's algorithm [24] via the CompQuadForm R package. Then like $\tilde{Q}_r(\tau^2)$, the test statistic $Q$, as a function of $\tau^2$, can be inverted to produce confidence intervals for $\tau^2$ and perform hypothesis testing. Based on simulation, Jackson [35] further suggested that weighting component studies by the reciprocal of their within-study standard errors (i.e. $1/s_k$), rather than by their variances (i.e. $1/s_k^2$) as the convention dictates, appears to provide a sensible and viable option when there is little a priori knowledge about the extent of heterogeneity. This version of $Q$ is referred to as $Q_J$.

**Binary-specific tests** For meta-analysis of multiple $2 \times 2$ tables, Woolf [91] applied the $Q$-test to examine the homogeneity of LORs, where $y_k = \log \frac{x_{k1}}{n_{k1}-x_{k1}} - \log \frac{x_{k0}}{n_{k0}-x_{k0}}$ and $s_k^2 = \frac{1}{x_{k0}} + \frac{1}{n_{k0}-x_{k0}} + \frac{1}{x_{k1}} + \frac{1}{n_{k1}-x_{k1}}$ when computing $Q_{Fe}$. Bliss [10] applied the $Q$-test in a slightly different way and the test statistic is given by $Q_B = K - 1 + \sqrt{\frac{\bar{n}-4}{\bar{n}-1}}\left[\frac{\bar{n}-2}{\bar{n}}Q_{Fe} - (K-1)\right]$, where $\bar{n} = \sum_{k=1}^{K}(n_k - 2)/K$ is the average of the degrees of freedom over all studies (two parameters estimated per study under a fixed-effects logistic regression, thus 2 degrees of freedom are lost). The distribution of $Q_B$ may be approximated by $\chi^2_{K-1}$ under the null of homogeneity [10]. For infrequent dichotomous events, Gart [25] added the continuity correction (CC) factor of 0.5 to all the four cells so that

$$y_k = \log \frac{x_{k1} + 0.5}{n_{k1} - x_{k1} + 0.5} - \log \frac{x_{k0} + 0.5}{n_{k0} - x_{k0} + 0.5}, \tag{1.5}$$

and the within-study variance $\sigma_k^2$ is estimated by

$$s_k^2 = \frac{1}{x_{k0} + 0.5} + \frac{1}{n_{k0} - x_{k0} + 0.5} + \frac{1}{x_{k1} + 0.5} + \frac{1}{n_{k1} - x_{k1} + 0.5} \tag{1.6}$$

when computing $Q_{Fe}$. This version of the $Q$-test is denoted by $Q_G$. In our numerical experiments, we perform the $Q$-test using $Q_G$ in the meta-analysis of rare binary events (with LOR as the effect measure) as there are often studies with zero event counts.

Bhaumik et al. [7] proposed to test $\mathcal{H}_0 : \tau^2 = 0$ in the $BN_{BA}$ model using $Q_{Re}$ for rare event data based on the observation that the normal distribution approximation is better for a standardized log-transformed gamma distribution than for the standardized gamma distribution. Here, $Q_{Re}$ is computed with $y_k$ in (1.5), $s_k^2$ in (1.6) and $w_k = 1/(s_k^2 + \hat{\tau}^2)$. The test statistic

$$Bh.T_3 = \frac{(K-1)[\log Q_{Re}(\hat{\tau}^2) - \log(K-1)]}{\sqrt{\widehat{\mathrm{Var}}[\log Q_{Re}(0)]}}$$

follows the standard normal distribution asymptotically, where $\widehat{\mathrm{Var}}[\log Q_{Re}]$ can be derived by the delta method under $\mathcal{H}_0$.

Kulinskaya and Dollinger [44] argued that when LOR is used as the effect measure, the $\chi^2_{K-1}$ distribution of Gart's $Q_G$ (i.e., $Q_{Fe}$ with the CC factor 0.5 applied to each cell) is too conservative for moderate-size studies, though asymptotically correct. Using a mixture of theoretical results and simulations, they derived formulas to estimate the mean and variance of $Q_G$,

$$\begin{aligned}
(K-1) - \mathrm{E}[Q_G] &= 0.678\,\{(K-1) - \mathrm{E}_{th}[Q_G\},\\
\mathrm{Var}[Q_G] &= 4.74(K-1) - 12.17\mathrm{E}[Q_G] + 9.42\mathrm{E}[Q_G]^2/(K-1),
\end{aligned}$$

where $\mathrm{E}_{th}[Q_G]$ is the approximation to the mean of $Q_G$ purely based on the Taylor expansion, and the constants like 0.678 and so on were obtained from simulation. Then the null distribution of $Q_G$ can be better approximated by matching these moments to those of a gamma distribution $\mathrm{Ga}(\alpha, \beta)$, where the shape parameter $\alpha$ is set to $\mathrm{E}[Q_G]^2/\mathrm{Var}[Q_G]$ and the scale parameter $\beta$ is set to $\mathrm{Var}[Q_G]/\mathrm{E}[Q_G]$. The test based on this gamma distribution, denoted $Q_\gamma$, was compared with $Q_G$ and the Breslow-Day (BD) test in [44]. They found that $Q_\gamma$ was not superior to the BD test either in accuracy or in power except for sparse data (i.e., small within-study sample sizes) where BD often performs poorly.

### 1.6.2. Likelihood ratio (LR) tests

**Generic tests** Under the generic *Re* model, likelihood ratio tests can be carried out with both maximum likelihood (ML) and restricted maximum likelihood (REML) estimation [84, 88].

Since $y_k \sim \mathrm{N}(\theta, \tau^2 + \sigma_k^2)$ in the *Re* model, the log-likelihood function of $(\theta, \tau^2)$ is given by

$$l(\theta, \tau^2) = -\frac{K}{2}\ln 2\pi - \frac{1}{2}\sum_{k=1}^{K}\ln(\tau^2 + \sigma_k^2) - \frac{1}{2}\sum_{k=1}^{K}\frac{(y_k - \theta)^2}{\tau^2 + \sigma_k^2}. \tag{1.7}$$

Given the value of $\tau^2$, it can be shown that the ML estimator of $\theta$ can be obtained by

$$\hat{\theta}_{ML}(\tau^2) = \sum_{k=1}^{K}\frac{y_k}{\tau^2 + \sigma_k^2} \Big/ \sum_{k=1}^{K}\frac{1}{\tau^2 + \sigma_k^2}, \tag{1.8}$$

where $\sigma_k^2$ is often replaced by $s_k^2$ in practice. Denote the log-likelihood with ML estimates $\hat{\theta}_{ML}$ and $\hat{\tau}^2_{ML}$ by $l(\hat{\theta}_{ML}(\hat{\tau}^2_{ML}), \hat{\tau}^2_{ML})$ and the log-likelihood assuming $\tau^2 = 0$ by $l(\hat{\theta}_{ML}(0), 0)$. Then the LR test statistic $LR_{ML} = -2\left(l\left(\hat{\theta}_{ML}(0), 0\right) - l\left(\hat{\theta}_{ML}(\hat{\tau}^2_{ML}), \hat{\tau}^2_{ML}\right)\right)$ is asymptotically distributed as a 50:50 mixture of a degenerate random variable $\chi_0^2$ at zero and a $\chi_1^2$ random variable under the null hypothesis $\mathcal{H}_0 : \tau^2 = 0$. Therefore, when $P(\chi_1^2 > LR_{ML}) < 2\alpha$, one rejects $\mathcal{H}_0$ at the significance level $\alpha$.

Alternatively, when using REML estimation, the restricted log-likelihood function of $\tau^2$ is given by

$$l_R(\tau^2) = -\frac{K}{2}\ln 2\pi - \frac{1}{2}\sum_{k=1}^{K}\ln(\tau^2 + \sigma_k^2) - \frac{1}{2}\ln\sum_{k=1}^{K}\frac{1}{\tau^2 + \sigma_k^2} - \frac{1}{2}\sum_{k=1}^{K}\frac{\left(y_k - \hat{\theta}_{ML}(\tau^2)\right)^2}{\tau^2 + \sigma_k^2}. \tag{1.9}$$

The corresponding test statistic is $LR_{REML} = -2\left(l_R(0) - l_R(\hat{\tau}^2_{REML})\right)$, where $\hat{\tau}^2_{REML}$ is the REML or approximate REML (AREML) estimate of $\tau^2$ (by maximizing $l_R$), and one rejects $\mathcal{H}_0$ when $P(\chi_1^2 > LR_{REML}) < 2\alpha$. Note that AREML has a direct adjustment for the loss

of degrees of freedom due to estimating $\theta$ in (1.9). Since the REML and AREML estimates are very similar [70, 85], we only examine the performance of the tests based on ML and REML in our numerical experiments.

**Binary-specific tests** For meta-analysis of binary event data, there have been three LR tests reported in the literature [62, 2], including two unconditional tests and one conditional test, with test statistics denoted by $LR_{U.FeL}$, $LR_{U.ReL}$ and $LR_C$, respectively. The major difference between the unconditional and conditional tests is that $LR_C$ is derived from the likelihood based on (non-central) hypergeometric distributions, where the row and column totals in each study are both treated as fixed, while $LR_U$s are derived from the likelihood based on binomial distributions, where only row totals are treated as fixed. The difference between the two unconditional tests is that $LR_{U.FeL}$ is related to fixed-effects logistic regression and $LR_{U.ReL}$ is related to random-effects logistic regression, as described in Section 1.3.

When assuming the event counts $x_{k0}$ and $x_{k1}$ follow separate binomial distributions and using LOR as the effect measure, the meta-analysis can be put under a framework of fixed-effects logistic regression and testing between-study homogeneity can be done via model comparison. Denote the (unconditional) log-likelihood under the null hypothesis of homogeneous odds ratios by $l^0_{U.FeL}$ and that under the alternative by $l^1_{U.FeL}$, which correspond to the model with main effects only (i.e.., $\text{logit}\,[P(O=1)] = \mu + \sum_{k=1}^{K-1} \alpha_k I(S=k) + \theta Z$) and the saturated model (1.4) , respectively. Then $LR_{U.FeL} = 2(l^1_{U.FeL} - l^0_{U.FeL})$ follows a $\chi^2_{K-1}$ distribution asymptotically. This test is equivalent to the goodness of fit test via model deviance for the model with main effects only. Under the hypergeometric distribution assumption for event counts (conditional on fixed margins), the meta-analysis can be done via conditional logistic regression and $LR_C$ can be derived in the same spirit, which also follows a $\chi^2_{K-1}$ distribution asymptotically.

Alternatively, the meta-analysis can be put under a framework of random-effects logistic regression, where $(\alpha_k, \beta_k)$ in (1.4) follows a bivariate normal distribution with zero means,

variances $\sigma^2$ and $\tau^2$, and correlation $\rho$, as described by the model $ReL$ (or equivalently $BN_{AH}$) in Section 1.3. Then testing homogeneity becomes testing $\mathcal{H}_0 : \tau^2 = 0$. Denote the (unconditional) log-likelihood under $\mathcal{H}_0$ by $l^0_{U.ReL}$, and that under the alternative by $l^1_{U.ReL}$. When $\mathcal{H}_0$ holds, $LR_{U.ReL} = 2(l^1_{U.ReL} - l^0_{U.ReL})$ approximately follows a 50:50 mixture of a degenerate random variable $\chi^2_0$ with all probability mass concentrated at 0 and a $\chi^2_2$ random variable. Thus, when $P(\chi^2_2 > LR_{U.ReL}) < 2\alpha$, one rejects $\mathcal{H}_0$ at the significance level $\alpha$ [84].

### 1.6.3. (Modified) Score tests

Both likelihood ratio and score tests are generally used to conduct hypothesis testing about parameters estimated with maximum likelihood methods. Rather than using the ratio of the likelihoods of the null model vs. the alternative model, score tests rely on the score function, defined as the first derivative of the log-likelihood evaluated under the null hypothesis, to derive the score statistics. It is generally believed that for binary data, score test statistics require fewer observations to reach the convergence to asymptotic distributions and are less sensitive to data sparsity than LR tests [1].

**Generic tests** Under the $Re$ model, score tests can be used with both ML and REML estimation [84, 88]. Starting from the log-likelihood or restricted log-likelihood function, obtaining the corresponding score function, dividing it by the square root of the corresponding Fisher information, and finally substituting all unknown parameters by their estimates yields

$$S_{ML} = \frac{\sum_{k=1}^{K} s_k^{-4} \left[ \left( y_k - \hat{\theta}_{ML}(0) \right)^2 - s_k^2 \right]}{\sqrt{2 \sum_{k=1}^{K} s_k^{-4}}},$$

$$S_{REML} = \frac{\sum_{k=1}^{K} s_k^{-4} \left[ \left( y_k - \hat{\theta}_{ML}(0) \right)^2 - s_k^2 + \left( \sum_{k=1}^{K} s_k^{-2} \right)^{-1} \right]}{\sqrt{2 \sum_{k=1}^{K} s_k^{-4} - 4 \sum_{k=1}^{K} s_k^{-6} / \sum_{k=1}^{K} s_k^{-2} + 2 \left( \sum_{k=1}^{K} s_k^{-4} / \sum_{k=1}^{K} s_k^{-2} \right)^2}},$$

where $\hat{\theta}_{ML}(0)$ is the ML estimate of $\theta$ under the $FE$ model. Because $\tau^2 \geq 0$, one-sided score tests are performed, where $\mathcal{H}_0 : \tau^2 = 0$ is rejected at the significance level $\alpha$ when $S_{ML}$ or $S_{REML}$ exceeds $z_{1-\alpha}$, the $100(1-\alpha)$th percentile of a standard normal distribution.

**Binary-specific tests** For meta-analysis of binary event data, there are two unconditional and two conditional score tests, with test statistics denoted by $US_{FeL}$ [5, 59, 62], $US_{ReL}$ [2], $CS_{FeH}$ [48], and $CS_{ReH}$ [48], respectively, plus several modified score tests.

Bartlett [5] proposed an (unconditional) score test to test homogeneity of odds ratios in two $2 \times 2$ contingency tables and Norton [59] extended it to the more general case of multiple $2 \times 2$ tables under $FeL$ (or equivalently $FeB$). Recently, Almalik and van den Heuvel [2] proposed an (unconditional) score test based on the random-effects logistic regression model $ReL$ for testing zero between-study variance. The two unconditional test statistics $US_{FeL}$ and $US_{ReL}$ have the same form

$$US = \sum_{k=1}^{K} \frac{\left[x_{k1} - \mathrm{E}(x_{k1} \mid x_k, \psi_k = \hat{\psi}_U)\right]^2}{\mathrm{Var}(x_{k1} \mid x_k, \psi_k = \hat{\psi}_U)},$$

where $x_k = x_{k0} + x_{k1}$ is the total number of events in study $k$. For $US_{FeL}(US_{ReL})$, $\hat{\psi}_U$ is substituted by $\hat{\psi}_{U.FeL}(\hat{\psi}_{U.ReL})$, the (unconditional) ML estimator based on $FeL(ReL)$; the expectation and variance are computed under the null hypothesis of homogeneity based on $FeL(ReL)$; and for fixed $K$ and large $n_{ki}$s, the null distribution can be approximated by $\chi^2_{K-1}$. Note that $US_{FeL}$ has been shown to be equivalent to the Pearson Chi-squared test statistic developed by [93] and [27]. When computing either $US$ test, obtaining the exact moments of $x_{k1}$ can be computationally difficult, and they are usually replaced by (estimated) asymptotic moments, which leads to $AUS$ [26, 62]. For $AUS$, $\mathrm{E}(x_{k1} \mid x_k, \psi_k = \hat{\psi}_U)$ is replaced by $e_k(\hat{\psi}_U)$ and $\mathrm{Var}(x_{k1} \mid x_k, \psi_k = \hat{\psi}_U)$ is replaced by $v_k(\hat{\psi}_U)$, where $e_k$ is the solution to the

equation $\frac{e_k(n_{k0}-x_k+e_k)}{(x_k-e_k)(n_{k1}-e_k)} = \hat{\psi}_U$, and $v_k$ can be calculated by

$$v_k = \left(\frac{1}{e_k} + \frac{1}{x_k - e_k} + \frac{1}{n_{k1} - e_k} + \frac{1}{n_{k0} - x_k + e_k}\right)^{-1}. \tag{1.10}$$

Liang and Self [48] proposed a conditional score test for homogeneity of odds ratios based on the fixed-effects hypergeometric model $FeH$, which is asymptotically equivalent to $US$ with a similar form

$$CS_{FeH} = \sum_{k=1}^{K} \frac{\left[x_{k1} - \mathrm{E}_C(x_{k1} \mid x_k, \psi_k = \hat{\psi}_C)\right]^2}{\mathrm{Var}_C(x_{k1} \mid x_k, \psi_k = \hat{\psi}_C)}.$$

Here, $\hat{\psi}_C$ is the conditional ML estimator based on $FeH$, and $\mathrm{E}_C(\cdot)$ and $\mathrm{Var}_C(\cdot)$ are the conditional expectation and variance under the null hypothesis of homogeneous odds ratios, conditioning on the fixed margins $n_{k0}$, $n_{k1}$ and $x_k$ in each $2 \times 2$ table. Like $US$, an approximate version of $CS_{FeH}$, denoted by $ACS_{FeH}$, can be obtained using the (estimated) asymptotic mean and variance $e_k(\hat{\psi}_C)$ and $v_k(\hat{\psi}_C)$, where $e_k(\hat{\psi}_C)$ is the solution to the equation $\frac{e_k(n_{k0}-x_k+e_k)}{(x_k-e_k)(n_{k1}-e_k)} = \hat{\psi}_C$ and $v_k(\hat{\psi}_C)$ is then computed from (1.10) based on $e_k(\hat{\psi}_C)$.

Liang and Self [48] proposed two other statistics $CS_{ReH}$ and $Z_{CS.FeH}$, which were specifically designed for the "sparse-data" situation. Here, sparsity means that within-study sample sizes can be small so that the assumption each $n_{ki} \to +\infty$ is not appropriate. However, it may be plausible to assume the number of studies $K \to +\infty$ in such situations. The conditional score test $CS_{ReH}$ was derived for testing $\mathcal{H}_0 : \tau^2 = 0$ under the random-effects hypergeometric model $ReH$, wherein $\log \psi_k = \alpha + \tau Z_k$ is assumed and $Z_k$s are independent and identically distributed random variables. The statistic $Z_{CS.FeH}$ was derived based on a normal approximation to $CS_{FeH}$. Both $CS_{ReH}$ and $Z_{CS.FeH}$ are asymptotically normal as $K \to +\infty$ for fixed $n_{ki}$. Their (lengthy) technical detail can be found in [48, 40].

The Breslow-Day test statistic $BD$ [11] has the same form as $US$ except for replacing $\hat{\psi}_U$ by the Mantel-Haenszel estimator $\hat{\psi}_{MH} = \frac{\sum_{k=1}^{K} x_{k1}(n_{k0}-x_{k0})/n_k}{\sum_{k=1}^{K} x_{k0}(n_{k1}-x_{k1})/n_k}$. The $BD$ can be computed

conveniently using $e_k(\hat{\psi}_{MH})$ and $v_k(\hat{\psi}_{MH})$, where $e_k$ and $v_k$ can be obtained as in $AUS$. Tarone [78] demonstrated that a score test statistic that substitutes a consistent but inefficient estimator (e.g. $\hat{\psi}_{MH}$) for a ML estimator would be stochastically larger than a $\chi^2_{K-1}$ random variable under $\mathcal{H}_0$. He then proposed a modified test statistic $MBD$ to improve the Chi-squared approximation under $\mathcal{H}_0$, given by

$$MBD = \sum_{k=1}^{K} \frac{\left[x_{k1} - \mathrm{E}(x_{k1} \mid x_k, \psi_k = \hat{\psi}_{MH})\right]^2}{\mathrm{Var}(x_{k1} \mid x_k, \psi_k = \hat{\psi}_{MH})} - \frac{\sum_{k=1}^{K} x_{k1} - \sum_{k=1}^{K} \mathrm{E}(x_{k1} \mid x_k, \hat{\psi}_{MH})}{\sum_{k=1}^{K} \mathrm{Var}(x_{k1} \mid x_k, \hat{\psi}_{MH})},$$

where $\mathrm{E}(x_{k1} \mid x_k, \hat{\psi}_{MH})$ and $\mathrm{Var}(x_{k1} \mid x_k, \hat{\psi}_{MH})$ can be approximated by $e_k(\hat{\psi}_{MH})$ and $v_k(\hat{\psi}_{MH})$ as in $BD$, leading to $AMBD$. Note that $MBD$ differs from $BD$ only by the correction term. Though $MBD$ seems to be more sound theoretically, Kulinskaya and Dollinger [44] found that there was a minimal difference between the BD and MBD tests, which is also observed in the simulation results of [40].

### 1.6.4. Wald tests

For likelihood-based inference, Wald, likelihood ratio, and score tests are three commonly used approaches to hypothesis testing. Among the three, Wald tests are believed to be the most sensitive to violations of regularity conditions required for asymptotic theories and of sample size requirements; even when the conditions are met, their convergence to the asymptotic distribution tends to be the slowest. Nevertheless, Wald tests can be easily derived under the generic $RE$ model to test $\mathcal{H}_0 : \tau^2 = 0$ [88]. The test statistics have the form $W = \hat{\tau}^2/SE(\hat{\tau}^2)$, where $\hat{\tau}^2$ can be $\hat{\tau}^2_{ML}$ or $\hat{\tau}^2_{REML}$, and the standard error is estimated

by

$$\widehat{SE}(\hat{\tau}^2_{ML}) = \sqrt{2\left[\sum_{k=1}^{K} w^2_{ML.k}\right]^{-1}},$$

$$\widehat{SE}(\hat{\tau}^2_{REML}) = \sqrt{2\left[\sum_{k=1}^{K} w^2_{REML.k} - 2\frac{\sum_{k=1}^{K} w^3_{REML.k}}{\sum_{k=1}^{K} w_{REML.k}} + \left(\frac{\sum_{k=1}^{K} w^2_{REML.k}}{\sum_{k=1}^{K} w_{REML.k}}\right)^2\right]^{-1}}$$

with $w_{ML.k} = 1/(\hat{\tau}^2_{ML} + s^2_k)$ and $w_{REML.k} = 1/(\hat{\tau}^2_{REML} + s^2_k)$. We label the Wald statistics based on ML and REML estimation by $W_{ML}$ and $W_{REML}$, respectively. Like score tests, one-sided Wald tests are performed, where $\mathcal{H}_0 : \tau^2 = 0$ is rejected when $W_{ML}$ or $W_{REML}$ exceeds $z_{1-\alpha}$.

We should also mention that based on the fixed-effects logistic regression model (1.4), the test of homogeneity of odds ratios can be done by testing whether the interaction terms $\beta_k$s can be dropped. Let $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_{K-1})$. The multivariate Wald statistic is given by $W^2_{FeL} = \hat{\beta}_{ML}[\widehat{Var}(\hat{\beta}_{ML})]^{-1}\hat{\beta}^T_{ML}$, and the testing is done via a $\chi^2_{K-1}$ distribution.

Because $\tau^2$ falls on the boundary of the parameter space $[0, +\infty)$ under $\mathcal{H}_0$, the standard likelihood theories are not valid [68] and so we no longer expect the Wald tests to control the Type I error rate adequately. However, it has been reported by [88] that $W_{ML}$ and $W_{REML}$ are still used by practitioners, and so we include these Wald tests in our numerical experiments.

### 1.6.5. Other tests

**Peto Chi-squared test** Yusuf et al. [92] proposed the *Peto* odds ratio as a modified Mantel-Haenszel estimator of the common odds ratio $\psi$, and derived the *Peto* statistic for

testing the homogeneity of odds ratios, namely

$$Peto = \sum_{k=1}^{K} \frac{(x_{k1} - E_{k1})^2}{V_k} - \frac{\left(\sum_{k=1}^{K}(x_{k1} - E_{k1})\right)^2}{\sum_{k=1}^{K} V_k},$$

where $E_{k1} = \frac{x_k n_{k1}}{n_k}$, $x_k = x_{k1} + x_{k0}$, $n_k = n_{k1} + n_{k0}$, and $V_k = \frac{x_k(n_k - x_k)n_{k1}n_{k0}}{n_k^2(n_k - 1)}$. This $Peto$ statistic has an approximate null distribution of $\chi^2_{K-1}$.

**Lipsitz et al.'s $Z^2_{WLS}$, $Z^2_{WLS,R}$, $Z^2_V$ and $Z^2_K$** For meta-analysis of multiple $2 \times 2$ tables with sparse data, Lipsitz et al. [50] proposed several test statistics for testing homogeneous risk differences, i.e., $\mathcal{H}_0 : \theta_1 = \theta_2 = \cdots = \theta_K = \theta$, where $\theta_k = p_{k1} - p_{k0}$ is the risk difference in study $k$ and is treated as a fixed effect such that $\text{Var}[y_k] \equiv \text{Var}[y_k|\theta_k] = \sigma_k^2$. The first statistic, $Z^2_{WLS}$, is a transformation of $Q_{Fe}$,

$$Z^2_{WLS} = \frac{[Q_{Fe} - (K-1)]^2}{2(K-1)},$$

whose distribution can be approximated by $\chi^2_1$ when the number of studies $K$ and each within-study sample size $n_{ki}$ are all large. Lipsitz et al. [50] proposed three other test statistics, $Z^2_{WLS,R}$, $Z^2_V$ and $Z^2_K$, derived based on a general quantity $Z(a, b)$,

$$Z(a,b) = \frac{\sum_{k=1}^{K} a_k \left[\left(y_k - \hat{\theta}(b)\right)^2 - s_k^2\right]}{\sqrt{\text{Var}\left(\sum_{k=1}^{K} a_k \left[\left(y_k - \hat{\theta}(b)\right)^2 - s_k^2\right]\right)}},$$

where $\hat{\theta}(b) = \sum_{k=1}^{K} b_k y_k / \sum_{k=1}^{K} b_k$, and $a = \{a_k\}_{k=1}^{K}$ and $b = \{b_k\}_{k=1}^{K}$ are different choices of weights. Lipsitz et al. [50] showed that under $\mathcal{H}_0$, the variance term in the numerator of $Z(a, b)$ can be approximated by $\sum_{k=1}^{K} a_k^2 \left[\left(y_k - \hat{\theta}(b)\right)^2 - s_k^2\right]^2$. Then $Z^2_{WLS,R}$ was obtained

by setting $a_k = b_k = 1/s_k^2$, yielding

$$Z^2_{WLS,R} = \frac{(Q_{Fe} - K)^2}{\sum_{k=1}^K \left[(y_k - \hat{\theta}_{Fe})^2/s_k^2 - 1\right]^2}.$$

The $Z_V^2$ was obtained by specifying $a_k = 1/\text{Var}[\left(y_k - \hat{\theta}_{Fe}\right)^2 - s_k^2] \approx 1/\text{Var}[(y_k - \theta)^2]$ and $b_k = 1/s_k^2$, where the large-sample approximation $\text{Var}[(y_k - \theta)^2]$ is estimated under $\mathcal{H}_0$. The $Z_K^2$ uses the same $a_k$ and $b_k$ as $Z_V^2$, but these weights are estimated by assuming no treatment effect (i.e., $p_{kt} = p_{kc} = p$) so that $a_k$ and $b_k$ can be specified by $(\frac{1}{n_{k0}} + \frac{1}{n_{k1}})^{-2}$ and $(\frac{1}{n_{k0}} + \frac{1}{n_{k1}})^{-1}$, respectively.

As with $Z^2_{WLS}$, the distributions of $Z^2_{WLS,R}$ and $Z_V^2$ can be approximated by $\chi_1^2$ when $K$ and each $n_{ki}$ are all large. The $Z_K^2$ also follows the $\chi_1^2$ distribution asymptotically but it only requires large $K$ and so is less sensitive to small within-study sample sizes. Actually, Lipsitz et al. [50] used $\text{F}_{1,K-1}$ (an F-distribution with 1 and $K-1$ degrees of freedom) as a better approximation than $\chi_1^2$ for finite-sample distributions of $Z^2_{WLS}$, $Z^2_{WLS,R}$, $Z_V^2$ and $Z_K^2$. They recommended to use one of the four $Z^2$ tests instead of the conventional $Q$-test for testing the homogeneity of risk difference when data are sparse.

Though originally designed for risk difference, $Z^2_{WLS}$, $Z^2_{WLS,R}$, and $Z_K^2$ have been generalized by [77] to other measures such as odds ratio and risk ratio for binary outcomes. The difference lies in the specification of $a_k$, $b_k$ and $s_k^2$ for the different measures. Based on simulation results using OR as the effect measure, they compared these tests with several other test statistics, including $Q_{Fe}$ and $LR_{ML}$, and concluded that in terms of validity, power, and computational ease, the $Q$-test was clearly the best choice.

**SA estimator-based test ($Bh.T_4$ in [7])** Under the $BN_{BA}$ model, Bhaumik et al. [7] proposed a simple average (SA) estimator $\hat{\theta}_{S\frac{1}{2}}$ of the overall treatment effect measured by the log odds ratio $\theta$, $\hat{\theta}_{S\frac{1}{2}} = \frac{1}{K}\sum_{k=1}^K y_k$, where $y_k$ is given by (1.5) with the CC factor 0.5 used by [25]. Let $z_k = (y_k - \hat{\theta}_{S\frac{1}{2}})^2$. Then under $\mathcal{H}_0 : \tau^2 = 0$, $B_k = \frac{K-2}{K}s_k^2 + \frac{\sum_{k=1}^K s_k^2}{K^2}$ is an

estimate for the mean of $z_k$ and $2B_k$ is an estimate for the variance of $z_k$, where $s_k^2$ is given by (1.6). The test statistic for testing $\mathcal{H}_0$ based on the SA estimator is defined as $Bh.T_4 = \sum_{k=1}^{K}(z_k - B_k)/\sqrt{\sum_{k=1}^{K} 2B_k}$. This test statistic approximately follows a standard normal distribution when $K$ and each $n_{ki}$ are all large. Bhaumik et al. [7] showed via simulation that $Bh.T_4$ is very conservative for finite samples and thus the parametric bootstrap technique is used instead to determine the critical value.

**Bayesian testing** As described in Section 1.3, most of the models are hierarchical and thus can be fit under a Bayesian framework by assigning prior distributions to parameters involved and then computing the joint posterior distribution via Markov chain Monte Carlo (MCMC). For meta analysis of sparse dichotomous events, Bai et al. [3] developed a Bayesian model selection method using Deviance Information Criterion (DIC) for simultaneously testing the existence of the treatment effect and between-study heterogeneity. The basic idea is to select the best model (model with the smallest DIC) among all four possible models in terms of the overall treatment effect $\theta = 0$ vs. $\theta \neq 0$ and the between-study variance $\tau^2 = 0$ vs. $\tau^2 \neq 0$. They concluded that for rare event data, the Bayesian method based on DIC performed better in most cases than classical methods (e.g., AIC and BIC) in terms of the percentage of selecting correct models.

**Bootstrap** Hypothesis testing can be performed using bootstrap techniques, which relax certain distributional assumptions and require no analytical derivation of the reference distribution of the test statistic. For nonparametric bootstrap, we can sample $K$ studies with replacement from the observed set of studies $B$ times to get $B$ bootstrap samples. For each sample we calculate the corresponding estimate for $\tau^2$. If the $\alpha$th percentile of the $B$ estimates is greater than 0, then we reject the null hypothesis of $H_0 : \tau^2 = 0$ at the significance level $\alpha$. Most of the $\tau^2$ estimators summarized in [85] can be used with this nonparametric procedure, but we only perform it for the most widely used DerSimonian and Laird (DL)

estimator [21], $\hat{\tau}^2_{DL}$, for illustration(labeled $BS_{DL}$):

$$\hat{\tau}^2_{DL} = \max\left\{\frac{Q-(K-1)}{\sum_k s_k^{-2} - \sum_k s_k^{-4}/\sum_k s_k^{-2}}\right\}.$$

For parametric bootstrap, we first obtain the parameter estimates and then generate samples from the assumed distributions with these estimates. This parametric procedure was used in [7] to determine the critical value of $Bh.T_4$, which performs much better than the corresponding asymptotic test in terms of the power to detect heterogeneity. We perform the parametric bootstrap procedure based on $Bh.T4$ in our numerical experiments (labeled $BS_{Bh.T4}$).

**The generalized variable (GV) approach** For meta-analysis of normally distributed outcomes, Tian [80] proposed inference procedures based on the generalized pivotal quantity for $\tau^2$. A pivotal quantity is a function of observations and parameters such that the distribution of the function does not depend on the parameters including nuisance parameters. Let $\sigma^2_{k0}$ ($\sigma^2_{k1}$) be the population variance of the control (treatment) group in study $k$; let $s^2_{k0}$ ($s^2_{k1}$) be the corresponding sample variance. For normally distributed outcomes, it is well known that $V_{ki} \equiv (n_{ki}-1)s^2_{ki}/\sigma^2_{ki} \sim \chi^2_{n_{ki}-1}$ for $k = 1, \ldots, K$ and $i = 0, 1$. Denote $Q$ in (1.1) with weight $w_k = 1/(\sigma^2_{k0}/n_{k0}+\sigma^2_{k1}/n_{k1}+\tau^2)$ by $Q(\tau^2)$, which follows $\chi^2_{K-1}$ and is a monotonic decreasing function of $\tau^2$. Thus, given a real number $\eta \geq 0$, there exists a unique $\tau^2_\eta \geq 0$ such that $Q(\tau^2_\eta) = \eta$. Based on this, Tian [80] defined the generalized pivotal quantity for $\tau^2$ by $R_{\tau^2} = \tau^2_\eta$ if $\eta \leq Q(0)$ and $R_{\tau^2} = 0$ otherwise. Given the observed treatment effects $y_k$s and sample variances $s^2_{ki}$s, the distribution of $R_{\tau^2}$ does not depend on any nuisance parameters. A series of $R_{\tau^2}$ values can be obtained by first simulating $V_{ki} \sim \chi^2_{n_{ki}-1}$ and $\eta \sim \chi^2_{K-1}$ and setting $\sigma^2_{ki} = (n_{ki}-1)s^2_{ki}/V_{ki}$ in $Q(\tau^2)$ for $k = 1, \ldots, K$ and $i = 0, 1$, and then solving for $\tau^2_\eta$. The p-value for testing $H_0 : \tau^2 = 0$ can be approximated by the proportion of $R_{\tau^2} = 0$.

## 1.7. Confidence intervals

Table 1.8 reports 16 existing methods for constructing CIs for $\tau^2$ in terms of abbreviation used and key features including whether the algorithm for computing a CI is iterative, whether truncation for non-negativity is needed, which distribution is used for construction, and whether the CI is exact under the *Re* model. All the methods are general-purpose and so can be applied to meta-analysis of binary events except for the generalized variable approach [80], which is specifically designed for the mean difference (MD) metric based on normally distributed outcomes. Some of the CIs are obtained via a test-inverting process based on different statistics for testing $\mathcal{H}_0 : \tau^2 = 0$.

In Table 1.9, we list existing review papers on constructing confidence intervals for $\tau^2$. Clearly, none of these reviews is comprehensive.

| Method | Abbreviation | Iterative? (Y/N) | Truncation to 0? (Y/N) | Distribution Used | Exact Method for *Re* ?(Y/N) | Reference |
|---|---|---|---|---|---|---|
| ***CIs based on (modified) Q statistics*** | | | | | | |
| $Q$-Profile | QP | Y | Y | $\chi^2_{K-1}$ | Y | [29, 41] |
| Modified $Q$-Profile | MQP | Y | Y | $\chi^2_{K-1}$ | N | [29, 41] |
| Biggerstaff and Tweedie | BT | Y | Y | $Ga(r, \lambda)$ | N | [8] |
| Biggerstaff and Jackson | BJ | Y | Y | A positive linear combination of $\chi^2_1$ | Y | [9] |
| Jackson | J | Y | Y | A positive linear combination of $\chi^2_1$ | Y | [35] |
| Approximate Jackson | AJ | N | Y | Normal | N | [37] |
| Unequal-tail $Q$-profile | UTQ | Y | Y | $\chi^2_{K-1}$ | Y | [36] |
| ***Profile likelihood CIs*** | | | | | | |
| PL based on ML estimation | $\text{PL}_{\text{ML}}$ | Y | Y | $\chi^2_1$ | N | [28] |
| PL based on REML estimation | $\text{PL}_{\text{REML}}$ | Y | Y | $\chi^2_1$ | N | [87] |
| ***Wald CIs*** | | | | | | |
| Wald based on ML estimation | $\text{W}_{\text{ML}}$ | N | Y | $N(0, 1)$ | N | [8, 88] |
| Wald based on REML estimation | $\text{W}_{\text{REML}}$ | N | Y | $N(0, 1)$ | N | [88] |
| ***Others*** | | | | | | |
| Sidik and Jonkman | SJ | N | N | $\chi^2_{K-1}$ | N | [69] |
| Sidik and Jonkman with *HO* priori | $\text{SJ}_{\text{HO}}$ | N | N | $\chi^2_{K-1}$ | N | [70] |
| Bayesian credible intervals | — | Y | N | — | N | [85] |
| Bootstrap | $\text{BS}_{\text{P}}/\text{BS}_{\text{NP}}$ | Y | Y | — | N | [23, 43] |
| Generalized variable approach | GV | Y | Y | — | N | [80] |

Table 1.8: CI methods for $\tau^2$ in random-effects meta-analysis.

| Review paper | CI methods reviewed/compared | Effect measure | Recommendations |
|---|---|---|---|
| Knapp et al. [41] | QP, MQP, BT, $PL_{ML}$, $W_{ML}$ | MD/OR | QP and MQP |
| Viechtbauer [87] | QP, BT, PL, W, SJ, BS | OR | QP |
| Veroniki et al. [85] | PL, W, BT, BJ, J, QP, SJ, BS, BC | Generic | — |
| van Aert et al. [82] | QP, BJ, J | OR | None recommended when $p_{ki} < 0.1$ in combination with either $K \geq 80$ or ($K \geq 40$ and $n_{ki} < 30$) |

Table 1.9: Existing comparative studies on constructing CIs for $\tau^2$ in random-effects meta-analysis

### 1.7.1. Confidence intervals based on (modified) $Q$ statistics

**$Q$-profile and modified $Q$-profile CIs** Knapp et al. [41] and Viechtbauer [87] considered the $Q$-profile CIs based on the generalized $Q$ statistic in (1.1) with weights $w_k = 1/(\tau^2 + s_k^2)$, denoted by $Q(\tau^2)$, which depends on $\tau^2$ and treats $s_k^2$s as known constants. It can be shown that $Q(\tau^2)$ follows a $\chi^2_{K-1}$ distribution under the *Re* model for any $\tau^2$. It follows that $P(\chi^2_{K-1,\alpha/2} < Q(\tau^2) < \chi^2_{K-1,1-\alpha/2}) = 1-\alpha$. Based on the test-inversion principle, a $100(1-\alpha)\%$ confidence interval for $\tau^2$ can be obtained as the interval $(\tilde{\tau}_l^2, \tilde{\tau}_u^2)$ satisfying $Q(\tilde{\tau}_l^2) = \chi^2_{K-1,1-\alpha/2}$ and $Q(\tilde{\tau}_u^2) = \chi^2_{K-1,\alpha/2}$. Since $\tau^2$ is a non-negative quantity, $\tilde{\tau}_l^2$ is truncated to 0 if $Q(0) < \chi^2_{K-1,1-\alpha/2}$ (meaning that $\tilde{\tau}_l^2$ is negative); and the CI is set to $[0, 0]$ (or $\{0\}$, the set containing only zero) if $Q(0) < \chi^2_{K-1,\alpha/2}$ (meaning that $\tilde{\tau}_u^2$ is also negative). This type of CIs is referred to as the $Q$-profile (QP) CIs as we are profiling $Q(\tau^2)$ with different $\tau^2$ values when solving the above equations for $\tilde{\tau}_l^2$ and $\tilde{\tau}_u^2$ iteratively.

Knapp et al. [41] considered the fact that $s_k^2$'s are only estimates and so have error variability, and constructed CIs using the test statistic $\tilde{Q}_r$ that replaces the weights in $Q(\tau^2)$ with regularized variants $w_{rk} = r_k/(\tau^2 + s_k^2)$ to achieve a closer approximation to $\chi^2_{K-1}$, where the regularization factor $r_k$ is derived through a moment matching approach based on approximating the distribution of $\tau^2 + s_k^2$ by a scaled $\chi^2$ distribution [29]. The lower bound $\tilde{\tau}_l^2$ is obtained by profiling $\tilde{Q}_r(\tau^2)$ while the upper bound $\tilde{\tau}_u^2$ is still obtained by profiling

$Q(\tau^2)$, satisfying $\tilde{Q}_r(\tilde{\tau}_l^2) = \chi^2_{K-1,1-\alpha/2}$ and $Q(\tilde{\tau}_u^2) = \chi^2_{K-1,\alpha/2}$. We refer to this type of CIs as the modified $Q$-profile (MQP) CIs.

Like the $Q$-profile CIs, the MQP CIs need left truncation to zero if the lower bound $\tilde{\tau}_l^2$ turns out to be negative, and they are set to $\{0\}$ if the upper bound $\tilde{\tau}_u^2$ is also negative. The same rule applies to all other types of CIs based on (modified) $Q$ statistics in Section 1.7.1, as discussed below.

**BT and BJ CIs based on Cochran's $Q$ statistic** Biggerstaff and Tweedie [8] proposed to approximate the distribution of the Cochran's $Q$ statistic $Q_{Fe}$ by a gamma distribution with a shape parameter $r(\tau^2) \equiv \mathrm{E}^2(Q_{Fe})/\mathrm{Var}(Q_{Fe})$ and a scale parameter $\lambda(\tau^2) \equiv \mathrm{Var}(Q_{Fe})/\mathrm{E}(Q_{Fe})$. The mean and variance of $Q_{Fe}$ under the $Re$ model are given by $\mathrm{E}(Q_{Fe}) = (K-1)+(S_1 - S_2/S_1)\,\tau^2$ and $\mathrm{Var}(Q_{Fe}) = 2(K-1)+4\,(S_1 - S_2/S_1)\,\tau^2+2\,(S_2 + S_2^2/S_1^2 - 2S_3/S_1)\,\tau^4$, where $S_r \equiv \sum_{k=1}^K [1/s_k^2]^r$. CIs for $\tau^2$ can be obtained similarly based on this gamma approximation instead of $\chi^2_{K-1}$ using the above profiling approach, which we refer to as the BT intervals.

Biggerstaff and Jackson [9] derived the exact CDF of $Q_{Fe}$ under the $Re$ model, denoted by $F_Q(q;\tau^2)$, as a positive linear combination of $\chi_1^2$ random variables, whose cumulative distribution function can be obtained using Farebrother's algorithm [24] via the CompQuadForm R package. They then obtained $(\tilde{\tau}_l^2, \tilde{\tau}_u^2)$ by solving the two equations numerically, $F_Q(c\hat{\tau}^2_{uDL} + K - 1; \tilde{\tau}_l^2) = 1 - \alpha/2$ and $F_Q(c\hat{\tau}^2_{uDL} + K - 1; \tilde{\tau}_u^2) = \alpha/2$, where $c = S_1 - S_2/S_1$ and $\hat{\tau}^2_{uDL} = [Q_{Fe} - (K-1)]/c$ is the untruncated version of the DL estimator of $\tau^2$. This type of CIs is referred to as the BJ intervals.

**Jackson and approximate Jackson CIs** Following the numerical approach in [9], Jackson [35] proposed CIs by test inversion based on the generalized $Q$ in (1.1), which is also distributed as a positive linear combination of $\chi_1^2$ random variables under the $Re$ model. Jackson et al. [37] further proposed to apply the arcsinh transformation to the untruncated

version of $\hat{\tau}^2_{GMM}$ for variance stabilization and then constructed CIs for $\tau^2$ based on a normal approximation. These types of CIs are referred to as the Jackson (J) and approximate Jackson (AJ) CIs, respectively. Based on simulation, Jackson [35] further commented that weighting component studies by the reciprocal of their within-study standard errors (i.e. $1/s_k$), rather than by their variances (i.e. $1/s_k^2$) as the convention dictates, appears to provide a sensible and viable option when there is little a priori knowledge about the extent of heterogeneity.

**Unequal-tail $Q$ profile CIs** Jackson and Bowden [36] advocated to use unequal tail probabilities to obtain shorter intervals whenever such methods are justifiable. For example, when constructing a $100(1-\alpha)\%$ unequal-tail $Q$-profile (UTQ) confidence interval, the lower and upper bounds, $\tilde{\tau}^2_l$ and $\tilde{\tau}^2_u$, are obtained by solving $Q(\tilde{\tau}^2_l) = \chi^2_{K-1,1-\alpha_1}$ and $Q(\tilde{\tau}^2_u) = \chi^2_{K-1,\alpha_2}$, respectively, where $\alpha_2 > \alpha_1$ and $\alpha_1 + \alpha_2 = \alpha$. They further suggested to use a pre-specified $\alpha$-split with $\alpha_1 = 0.01$ and $\alpha_2 = 0.04$ for a 95% CI, which was shown to be able to retain the nominal coverage and reduce the width under the *Re* model. Obviously, the idea of unequal tails can be applied to all kinds of confidence intervals. In our numerical evaluation, we examine the performance of the $Q$-profile CIs with $\alpha_1 = 0.01$ and $\alpha_2 = 0.04$ as a representative.

### 1.7.2. Profile likelihood confidence intervals

Under the *Re* model, Hardy and Thompson [28] proposed the profile likelihood CIs based on maximum likelihood (ML) estimation, referred to as $\text{PL}_{\text{ML}}$. The profile log-likelihood for $\tau^2$ takes into account the fact that $\theta$ is also unknown and must be estimated, given by $l(\hat{\theta}_{ML}(\tau^2), \tau^2)$, where the log-likelihood function of $(\theta, \tau^2)$ is given by

$$l(\theta, \tau^2) = -\frac{K}{2}\ln 2\pi - \frac{1}{2}\sum_{k=1}^{K}\ln(\tau^2 + \sigma_k^2) - \frac{1}{2}\sum_{k=1}^{K}\frac{(y_k - \theta)^2}{\tau^2 + \sigma_k^2},$$

and given the value of $\tau^2$, the ML estimator of $\theta$ can be obtained by

$$\hat{\theta}_{ML}(\tau^2) = \sum_{k=1}^{K} \frac{1}{\tau^2 + \sigma_k^2} y_k / \sum_{k=1}^{K} \frac{1}{\tau^2 + \sigma_k^2}.$$

Then the $100(1-\alpha)\%$ CI for $\tau^2$ is given by the set of $\tau^2$ values satisfying $l(\hat{\theta}_{ML}(\tau^2), \tau^2) > l(\hat{\theta}_{ML}(\hat{\tau}^2_{ML}), \hat{\tau}^2_{ML}) - \chi^2_{1,1-\alpha}/2$.

Viechtbauer [87] proposed to construct profile likelihood CIs based on restricted maximum likelihood (REML) estimation, referred to as $\text{PL}_{\text{REML}}$. The $100(1-\alpha)\%$ CI for $\tau^2$ is given by the set of $\tau^2$ values satisfying $l_R(\tau^2) > l_R(\hat{\tau}^2_{REML}) - \chi^2_{1,1-\alpha}/2$, where the restricted log-likelihood function of $\tau^2$ is given by

$$l_R(\tau^2) = -\frac{K}{2} \ln 2\pi - \frac{1}{2} \sum_{k=1}^{K} \ln(\tau^2 + \sigma_k^2) - \frac{1}{2} \sum_{k=1}^{K} \ln \frac{1}{\tau^2 + \sigma_k^2} - \frac{1}{2} \sum_{k=1}^{K} \frac{\left(y_k - \hat{\theta}_{ML}(\tau^2)\right)^2}{\tau^2 + \sigma_k^2},$$

and $\hat{\tau}^2_{REML}$ is the REML estimate of $\tau^2$ (by maximizing $l_R$). Viechtbauer [87] found that the REML-based CIs were slightly more accurate than the ML-based CIs in terms of coverage probability, especially for small $K$.

Because ML and REML estimates of $\tau^2$ require non-negativity, the lower bounds of profile likelihood (PL) intervals are always non-negative and the upper bounds are strictly positive after applying the same truncation for $Q$-profile CIs.

### 1.7.3. Wald confidence intervals

The Wald test statistics for testing $\mathcal{H}_0 : \tau^2 = 0$ under the *Re* model have the form $W = \hat{\tau}^2/SE(\hat{\tau}^2)$, where $\hat{\tau}^2$ can be $\hat{\tau}^2_{ML}$ or $\hat{\tau}^2_{REML}$, and the standard error is estimated by

$$\widehat{SE}(\hat{\tau}^2_{ML}) = \sqrt{2\left[\sum_{k=1}^{K} w^2_{ML.k}\right]^{-1}},$$

$$\widehat{SE}(\hat{\tau}^2_{REML}) = \sqrt{2\left[\sum_{k=1}^{K} w^2_{REML.k} - 2\frac{\sum_{k=1}^{K} w^3_{REML.k}}{\sum_{k=1}^{K} w_{REML.k}} + \left(\frac{\sum_{k=1}^{K} w^2_{REML.k}}{\sum_{k=1}^{K} w_{REML.k}}\right)^2\right]^{-1}}$$

with $w_{ML.k} = 1/(\hat{\tau}^2_{ML} + s^2_k)$ and $w_{REML.k} = 1/(\hat{\tau}^2_{REML} + s^2_k)$. We label the Wald statistics based on ML and REML estimation by $W_{ML}$ and $W_{REML}$, respectively. The corresponding $100(1-\alpha)\%$ Wald (W) CI for $\tau^2$ can be easily obtained by $\hat{\tau}^2_{ML} \pm z_{1-\alpha/2}\widehat{SE}(\hat{\tau}^2_{ML})$ or $\hat{\tau}^2_{REML} \pm z_{1-\alpha/2}\widehat{SE}(\hat{\tau}^2_{REML})$ [8, 87], where $z_\alpha$ is the $100\alpha$-th percentile of the standard normal distribution. Negative lower bounds of the Wald CIs should be truncated to 0 since both ML and REML estimates of $\tau^2$ are constrained to be non-negative.

### 1.7.4. Other confidence intervals

**Sidik and Jonkman (SJ) CIs** Sidik and Jonkman [69] proposed confidence intervals based on the SJ estimator of $\tau^2$, which is derived from the weighted residual sum of squares in the framework of a linear regression model. Let the crude estimate $\hat{\tau}_0 = \sum_{k=1}^{K}(y_k - \bar{y})^2/K$ be an a priori value for $\tau^2$. Then the SJ estimator is given by $\hat{\tau}^2_{SJ} = \frac{\hat{\tau}^2_0}{K-1}\sum_{k=1}^{K} \hat{w}_k(y_k - \hat{\theta}_0)^2$, where $\hat{w}_k = 1/(s^2_k + \hat{\tau}^2_0)$, and $\hat{\theta}_0 = \sum_{k=1}^{K} \hat{w}_k y_k / \sum_{k=1}^{K} \hat{w}_k$. It follows that $(K-1)\hat{\tau}^2_{SJ}/\tau^2$ has an asymptotic distribution of $\chi^2_{K-1}$. Thus an approximate $100(1-\alpha)\%$ confidence interval can be calculated by

$$\frac{(K-1)\hat{\tau}^2_{SJ}}{\chi^2_{K-1,1-\alpha/2}} \leq \tau^2 \leq \frac{(K-1)\hat{\tau}^2_{SJ}}{\chi^2_{K-1,\alpha/2}}.$$

Since $\hat{\tau}^2_{SJ}$ is always positive, the SJ confidence intervals have positive lower and upper bounds. Sidik and Jonkman [70] later proposed an improved estimator $\hat{\tau}^2_{SJ_{HO}}$ by using $\hat{\tau}^2_{HO}$ as the a priori value. Then improved confidence intervals can be constructed correspondingly.

**Bayesian credible intervals** Bayesian credible (BC) intervals can be obtained when a Bayesian approach is employed and posterior samples are drawn from the (joint) posterior distribution of all parameters involved using an MCMC algorithm. The lower and upper points of a $100(1-\alpha)\%$ CI can be the $100(\alpha/2)$ and $100(1-\alpha/2)$ percentiles of the posterior sample of $\tau^2$'s, or determined by the region that gives the highest posterior density. Such intervals may be heavily affected by the prior selection when the number of studies $K$ is small.

**Bootstrap CIs** Bootstrap techniques can be used to obtain confidence intervals for nearly all $\tau^2$ estimators. For nonparametric bootstrap (denoted by $\text{BS}_{\text{NP}}$), we sample $K$ studies with replacement from the observed set of studies $B$ times to get $B$ bootstrap samples. For parametric bootstrap (denoted by $\text{BS}_{\text{P}}$), we first obtain the parameter estimates and then generate $B$ samples from the assumed distributions with these estimates. For each (parametric or nonparametric) sample, we calculate the corresponding estimate $\hat{\tau}^2$. Then the $100(\alpha/2)$th and $100(1-\alpha/2)$th percentiles of the $B$ estimates of $\tau^2$ are respectively the lower and upper bounds of a $100(1-\alpha)\%$ bootstrap confidence interval. In our numerical experiment, we only perform the nonparametric bootstrap procedure for the $DL$ estimator for illustration.

**The generalized variable (GV) approach** For meta-analysis of normally distributed outcomes, Tian [80] proposed inference procedures based on the generalized pivotal quantity for $\tau^2$. A pivotal quantity is a function of observations and parameters such that the distribution of the function does not depend on the parameters including nuisance parameters. Let $\sigma^2_{k0}$ ($\sigma^2_{k1}$) be the population variance of the control (treatment) group in study $k$; let $s^2_{k0}$

$(s_{k1}^2)$ be the corresponding sample variance. For normally distributed outcomes, it is well known that $V_{ki} \equiv (n_{ki}-1)s_{ki}^2/\sigma_{ki}^2 \sim \chi_{n_{ki}-1}^2$ for $k = 1, \ldots, K$ and $i = 0, 1$. Denote $Q$ in (1.1) with weight $w_k = 1/(\sigma_{k0}^2/n_{k0}+\sigma_{k1}^2/n_{k1}+\tau^2)$ by $Q(\tau^2)$, which follows $\chi_{K-1}^2$ and is a monotonic decreasing function of $\tau^2$. Thus, given a real number $\eta \geq 0$, there exists a unique $\tau_\eta^2 \geq 0$ such that $Q(\tau_\eta^2) = \eta$. Based on this, Tian [80] defined the generalized pivotal quantity $R_{\tau^2}$ for $\tau^2$ as $R_{\tau^2} = \tau_\eta^2$ if $\eta \leq Q(0)$ and $R_{\tau^2} = 0$ otherwise. Given the observed treatment effects $y_k$'s and sample variances $s_{ki}^2$'s, the distribution of $R_{\tau^2}$ does not depend on any nuisance parameters. A series of $R_{\tau^2}$ values can be obtained by first simulating $V_{ki} \sim \chi_{n_{ki}-1}^2$ and $\eta \sim \chi_{K-1}^2$ and setting $\sigma_{ki}^2 = (n_{ki}-1)s_{ki}^2/V_{ki}$ in $Q(\tau^2)$ for $k = 1, \ldots, K$ and $i = 0, 1$, and then solving for $\tau_\eta^2$. A $100(1-\alpha)\%$ confidence interval is given by $(R_{\tau^2,\alpha/2}, R_{\tau^2,1-\alpha/2})$, where the lower and upper bounds are the $100(\alpha/2)$th and $100(1-\alpha/2)$th percentile of the generated $R_{\tau^2}$'s.

### 1.8. Simulation focusing on rare binary events

For meta-analysis of rare binary events, Li and Wang [47] conducted a comprehensive simulation study to compare the performance of various estimators of the overall treatment effect $\theta$, where the $BN_{LW}$ model was used for data generation to accommodate treatment groups with unequal variability. In this section, we adopt the same simulation setup, to examine the performance of methods for estimating and testing the between-study heterogeneity, as summarized in previous sections. Here, bias and MSE are reported for point estimation, the actual type I error rate and power are reported for hypothesis testing, and the actual coverage probability and width of confidence intervals reported for interval estimation. To be specific, we set the number of studies $K$ to 10, 20 and 50 to reflect different sizes of meta-analysis. We generated the number of events $x_{ki}$ from Binomial$(n_{ki}, p_{ki})$ for $k = 1, \ldots, K$ and $i = 0, 1$, where $n_{k0}$s were generated from Uniform$[2000, 3000]$ to examine large-sample performance and from Uniform$[20, 1000]$ to examine small-sample performance, and then rounded to the nearest integers. For small sample sizes, as noted in [47],

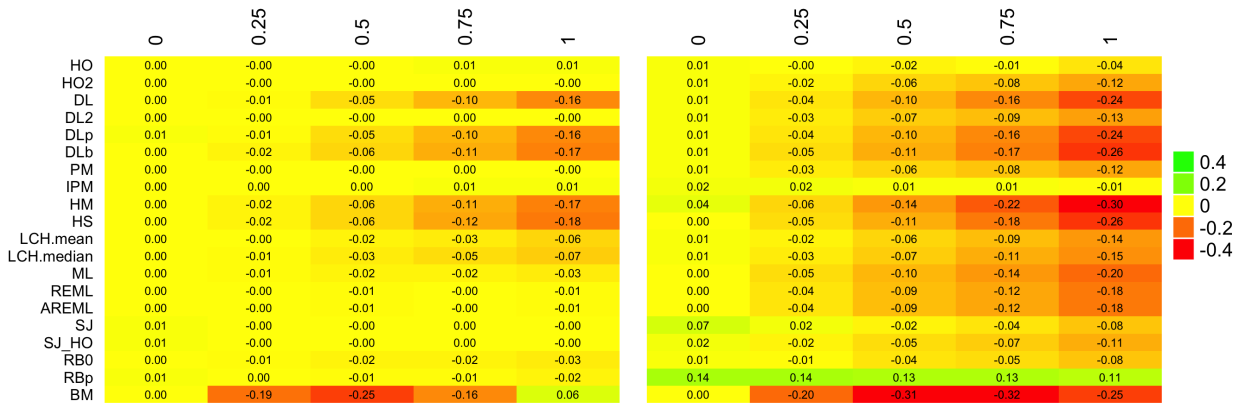
the range $[20, 1000]$ was chosen so that the empirical means of $\min(n_{k0}p_{k0}, n_{k0}p_{k0})_{k=1}^{K}$ in all the settings are below one while it still allows for cases where most component studies have small sample sizes but a few can have sample sizes close to 1000. To allow varying allocation ratios across studies, the within-study sample sizes were set to follow the relationship $n_{k1} = R_k n_{k0}$, where $\log_2 R_k \sim N(\log_2 R, \sigma_R^2)$, $R \in \{1, 2, 4\}$ and $\sigma_R^2 = 0.5$. To generate $p_{ki}$s, we fixed $\sigma^2$ at 0.5, and set $\tau^2 \in \{0, 0.25, 0.5, 0.75, 1\}$ for evaluating different estimators and $\tau^2 \in \{0, 0.1, 0.2, \cdots, 0.9, 1\}$ for evaluating different tests and CIs. We further set $\theta \in \{-1, 0, 1\}$ to reflect different directions of the overall treatment effect, set $\mu \in \{-2.5, -5\}$ to represent low and very low incidence rates of the binary event (i.e., 0.076 and 0.0067 in the probability magnitude), and set $\omega \in \{0, 0.5, 1\}$ to represent smaller/equal/larger variability in the control group, compared to the treatment group. For each setting, 1000 datasets were simulated to compute empirical values of the performance measures by taking the average.

### 1.8.1. Comparison of different heterogeneity estimators

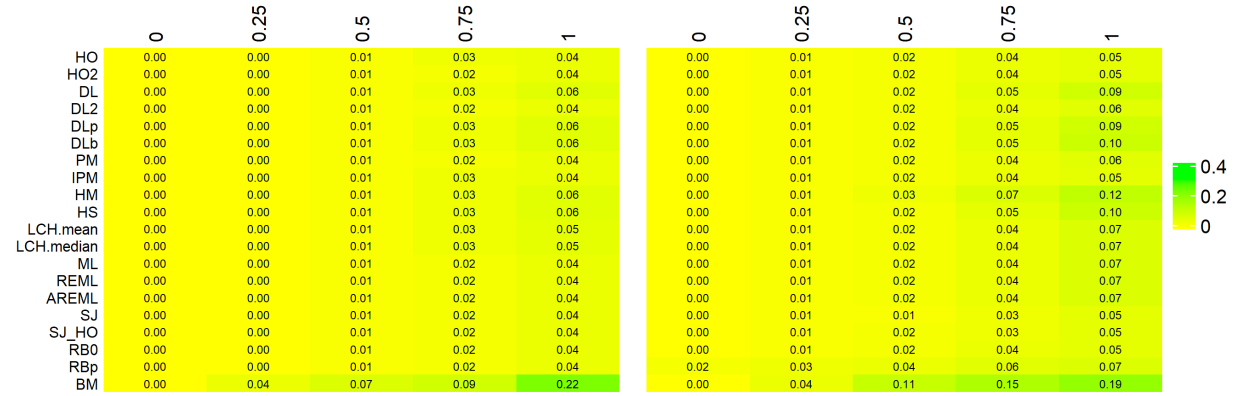
We compared all the methods listed in Table 1.4 except for $FB$, and $MBH$. Since the full Bayesian method can be greatly affected by the prior choice and other factors (such as convergence), we eliminated $FB$ from our simulation. The $MBH$ method is designed specifically for standard mean difference thus not suitable for binary events. In addition, the empirical Bayes method $EB$ is equivalent to $PM$ and the multistep $DL$ method has the property that $DL_\infty$ converges to $PM$. Thus, we include $PM$ in the comparison and leave $EB$ and $DL_k$ out. We use heat maps to visualize the bias and MSE results where the rows of each map represent different methods and columns represents different $\tau^2$ values in [0,1].

**Large-sample results** Figure 1.1 presents the bias and MSE results of different estimators for $\mu = -2.5$ and $\mu = -5$ based on large-sample settings with $R = 1$, $K = 50$, $\theta = 0$, and $w = 0$. As shown in Figure 1.1(a), as the event of interest becomes rarer, all methods seem to produce more bias when estimating the heterogeneity parameter $\tau^2$. Almost all methods

underestimate the between-study heterogeneity when $\tau^2 > 0$. The $RBp$ estimator, however, consistently overestimates $\tau^2$ when the event is very rare ($\mu = -5$). As $\tau^2$ increases, most estimators produce more bias except for $BM$ and $RB_p$; the bias from $BM$ first increases then decreases, and the bias from $RB_p$ decreases for very rare events ($\mu = -5$). When the events are not that rare ($\mu = -2.5$), most estimators have similarly low bias except for the one-step $DL$ estimators ($DL$, $DLp$, $DL_b$), $HM$, $HS$, and $BM$. However, $IPM$ stands out with the lowest bias when the incidence rate becomes very low, especially when $\tau^2 \geq 0.5$. The $HS$, $HM$, $BM$ and one-step $DL$ family methods remain the worst and should be avoided in terms of bias. All three likelihood-based methods, $ML$, $REML$ and $AREML$, produce similar results with a moderate level of bias. In terms of MSE, most methods have similar performance except for $HM$ and $BM$, which are the most inefficient according to Figure 1.1(b). Those with relatively large magnitude of bias tend to have relatively large MSE.




(a) Comparison of estimation bias. Left panel: $\mu = -2.5$; Right panel: $\mu = -5$.



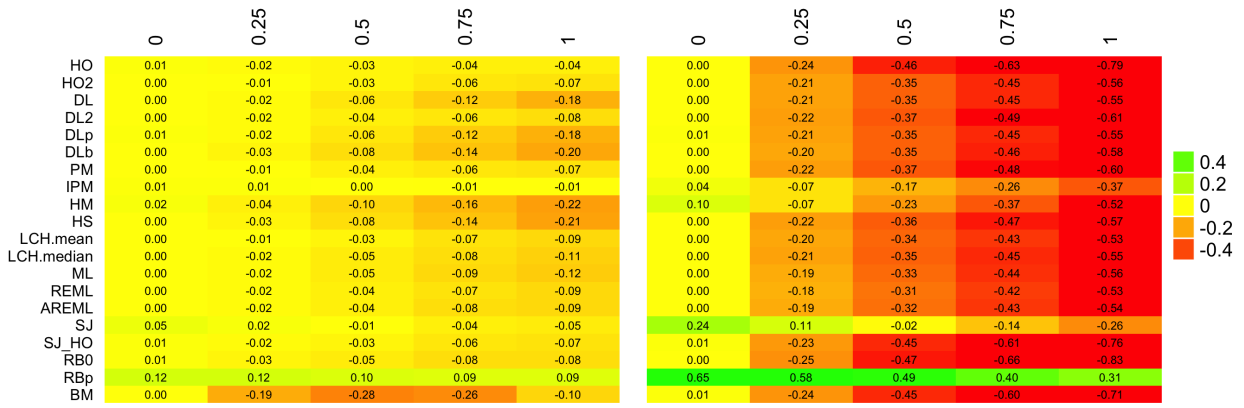

(b) Comparison of MSE. Left panel: $\mu = -2.5$; Right panel: $\mu = -5$.

Figure 1.1: Large-sample performance of different $\tau^2$ estimators based on settings with $R = 1$, $K = 50$, $\theta = 0$, and $w = 0$.

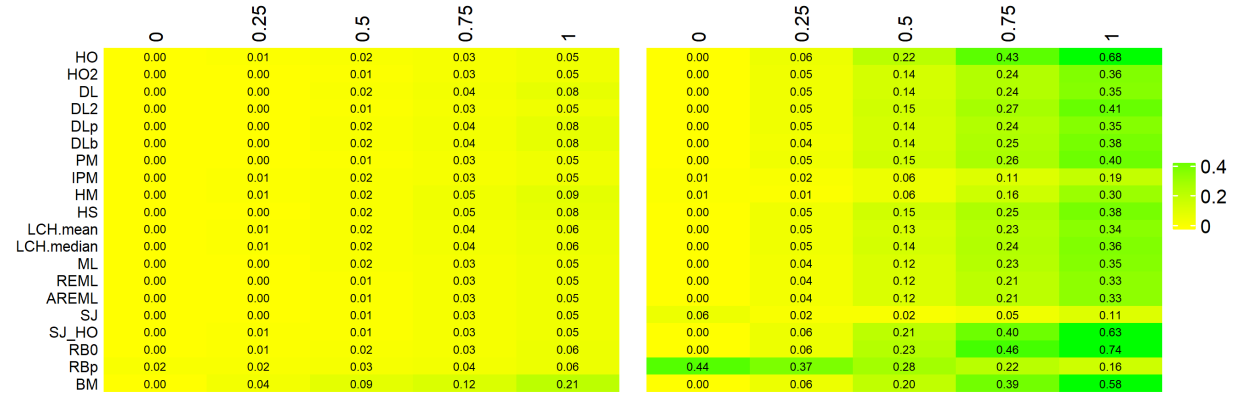
We next discuss the potential impacts of $R$, $K$, $\theta$, and $w$ on the estimation performance for the large-sample case. Figures A.1 and A.2 in Appendix A show the bias and MSE results for different $R$ and $K$ values, respectively, based on settings with $\mu = -2.5$, $\theta = 0$ and $w = 0$. We can see that when $\tau^2 < 0.5$, regardless of $R$ and $K$, all the methods perform somewhat similarly and have both bias and MSE close to zero except for $BM$ which has much larger bias. As $K$ increases, MSE decreases significantly for every estimator when $\tau^2 \geq 0.5$ but bias for a few estimators seems not to get closer to zero (e.g., $DL$ for $\tau^2 = 1$, $BM$ for $\tau^2 = 0.5$, and 0.75). However, the heat maps show very similar color patterns both vertically and horizontally, indicating that the impact of $R$ and $K$ on the *relative* performance of these methods is merely marginal. Figures A.3 and A.4 in Appendix A show the bias and MSE

results for different $\theta$ and $w$ values, respectively, based on settings with $R = 1$, $K = 50$ and $\mu = -5$. When $\theta = -1$, bias decreases as $w$ increases while this trend reverses when $\theta = 1$. This effect of $w$ is minimal when there is no treatment effect ($\theta = 0$). Similar trends are observed but less obvious for MSE. Also, we find that $IPM$ maintains the best performance in terms of both bias and MSE while $DL$, $DLp$, $DL_b$, $HS$, $HM$, and $BM$ are among the worst in nearly all the settings considered.

**Small-sample results** Figure 1.2 presents the bias and MSE results of different estimators for $\mu = -2.5$ and $\mu = -5$ based on small-sample settings with $R = 1$, $K = 50$, $\theta = 0$, and $w = 0$. From Figure 1.2(a), we can see that when $\tau^2 > 0$, the underestimation observed in the large-sample results for all the estimators but $RB_p$ is much more severe for small samples, where the magnitude of bias increases substantially for very rare events ($\mu = -5$). Note that $RB_p$ consistently overestimates $\tau^2$ for both $\mu = -2.5$ and $\mu = -5$, and unlike most other estimators, the bias decreases as $\tau^2$ increases. When events are not that rare ($\mu = -2.5$), $IPM$ is still the least biased. However, for very rare events ($\mu = -5$), $SJ$ becomes the least biased estimator for $\tau^2 \geq 0.5$. The problem of $SJ$ is that it significantly overestimates $\tau^2$ when there is no or little heterogeneity, due to its positive nature. From Figure 1.2(b) we can see that MSE does not change much when $\mu = -2.5$ but dramatically increases when $\mu = -5$ compared to results from large samples. For very rare events ($\mu = -5$), $SJ$ is the most efficient method except for $\tau^2 = 0$ and $IPM$ seems to be the second best in terms of MSE. Note that when $\tau^2 = 1$, $RB_p$ has smaller MSE than $IPM$ for very rare events, but it does not perform as well as $IPM$ for smaller $\tau^2$ values.




(a) Comparison of estimation bias. Left panel: $\mu = -2.5$; Right panel: $\mu = -5$.




(b) Comparison of MSE. Left panel: $\mu = -2.5$; Right panel: $\mu = -5$.

Figure 1.2: Small-sample performance of different $\tau^2$ estimators based on settings with $R = 1$, $K = 50$, $\theta = 0$, and $w = 0$.

The impacts of $R$, $K$, $\theta$, and $w$ on the estimation bias and MSE for the small-sample case are shown in Figure A.5-A.8 of Appendix A. Since several methods (e.g., the likelihood-based methods) failed in some small-sample settings for very rare events ($\mu = -5$), we show results for $\mu = -2.5$ in these figures. Although the effect of $K$ on MSE becomes more significant for small samples (i.e., MSE decreases more as $K$ increases), it is still the case that both $R$ and $K$ have little impact on the *relative* performance of the different methods. Also, similar trends for both bias and MSE occur when $w$ and $\theta$ change as in the large-sample case. For these $\mu = -2.5$ settings, $IPM$ seems to be the best estimator due to its consistently top-level performance across various settings. This also agrees with the results in the left panels of Figure 1.2. On the other hand, $DL$, $DL_p$, $DL_b$, $HM$, $HS$, and $BM$ should be avoided due

to their generally large bias.

### 1.8.2. Comparison of methods for testing homogeneity of odds ratios

Among those summarized in Table 1.7, we compared 29 testing methods at the significance level $\alpha = 0.05$, as listed in Table 1.10. We excluded Bayesian methods for the same reason as mentioned for $FB$ in Section 1.8.1, the Cochran's $Q$ test based on $Q_{Fe}$ that is not applicable to rare binary data with zero counts without continuity correction, one of the Lipsitz tests based on $Z_V^2$, which only works for the risk difference measure, the multivariate Wald test based on $W_{FeL}^2$ that failed to work in many of our simulation settings, and the generalized variable approach that is only applicable to normal data. Among all the $Q$-statistic based tests, $\tilde{Q}_r$, $Q_B$ and $Q_G$ have similar performance thus we only report power results from $Q_G$. Recall that $Q_G$ is actually $Q_{Fe}$ applied to binary outcomes with the continuity correction factor 0.5. The power values of $Bh.T_3$, two Wald tests $W_{ML}$ and $W_{REML}$, and two Lipsitz tests $Z_{WLS,R}^2$ and $Z_K^2$ are much lower than those of the other tests thus are also not shown in the comparison figures. According to [7], the (asymptotic) test using $Bh.T_4$ is very conservative, which is confirmed by our simulation results in Table 1.10. Thus we only include the bootstrap version of $Bh.T_4$ for power comparison. With all above mentioned methods excluded, there are 21 left in our final comparison using Figures 1.3 and 1.4. From our simulation results (not reported due to the space limit), we find that the influences of $R$, $\theta$, and $w$ on the performance of different testing procedures are marginal, and so we only report results for settings with $R = 1$, $\theta = 0$, and $w = 0$.

Table 1.10 presents the test sizes of different methods for both large- and small-sample cases and different incidence rates ($\mu = -2.5$, -5) based on settings with $R = 1$, $K = 20$, $\theta = 0$, $w = 0$ and $\tau^2 = 0$. The Wald tests ($W_{ML}$ and $W_{REML}$), $Bh.T_3$, and $Bh.T_4$ are very conservative in all the settings, which was previously reported in [7, 88]. Also, $LR_{ML}$, $LR_{REML}$, $Z_{WLS}^2$, and $BS_{DL}$ seem to be quite conservative, too. On the other hand, the tests

based on $Z^2_{WLS,R}$ and $Z^2_K$ have severely inflated Type I error rates. When $\mu = -2.5$ and the sample sizes are large, most tests except for those mentioned above seem to maintain the Type I error rates close to the target level 0.05. As events get rarer (i.e., $\mu$ decreases), nearly all tests become more conservative with smaller test sizes except for the conditional and unconditional $LR$ tests ($LR_{U.FeL}$, $LR_{U.ReL}$ and $LR_C$), $Z^2_{WLS,R}$ and $Z^2_K$. And for very rare events ($\mu = -5$), most of these tests become more conservative with test sizes getting very close to zero when the sample sizes become smaller, but they appear to be less impacted by sample size when $\mu = -2.5$.

Figure 1.3 shows power curves of different heterogeneity tests for different $K$ values based on large-sample settings with $R = 1$, $\mu = -5$, $\theta = 0$ and $w = 0$. For all the methods, the power increases when $\tau^2$ increases as we expect, and as the number of studies $K$ increases, the curves increase more rapidly. The biggest improvement from increasing $K$ is obtained by the test $BS_{DL}$ based on the nonparametric bootstrap procedure combined with the $DL$ estimator. The relative performance of different tests does not vary much in different $K$ settings. Note that each curve starts at $\tau^2 = 0$, where the power becomes the empirical type I error rate. It appears that the effect of $K$ on the test size is marginal for the different methods.

Figure 1.4 shows power curves of different homogeneity tests for both large- and small-sample cases and different $\mu$ values based on settings with $R = 1$, $K = 20$, $\theta = 0$, and $w = 0$. When events are not that rare ($\mu = -2.5$) and sample sizes are large, most tests achieve very high power (close to one) even when the true heterogeneity level is low. However, as events become rarer or the sample sizes are smaller, the power decreases for each method, starting from smaller $\tau^2$ values. The unconditional $LR$ test based on the $ReL$ model ($LR_{U.ReL}$) has the highest power in all the settings. The power of $LR_{U.ReL}$ is especially higher than the other tests when sample sizes are small and events are very rare, but it comes with inflated test sizes, as shown in Table 1.10. To better compare the power of different tests, we focus on the case of $\mu = -5$ and small sample sizes, where the differences become most obvious. The $LR$

test based on the $FeL$ model and the conditional $LR$ test based on the $FeH$ model ($LR_{U.FeL}$ and $LR_C$) have close performance and are the second most powerful tests. The test based on $Q_\gamma$ has the highest power among all the (modified) $Q$-statistic based tests, which we believe is due to the more accurate gamma approximation achieved with extra computational burden. Among all the score tests, the conditional score test based on the $ReH$ model ($CS_{ReH}$) has the highest power and performs slightly better than $Q_\gamma$. Another observation is that tests based on the $REML$ estimates performed better than the corresponding tests based on the $ML$ estimates.

Overall, we recommend $LR_{U.FeL}$ and $LR_C$ for testing the homogeneity of treatment effects when dealing with rare binary events as these two tests achieve high power while maintaining the nominal Type I error rate roughly. The widely used Cochran's $Q$-test with continuity correction $Q_G$ is not recommended due to its lackluster performance especially when the events are very rare and the sample sizes are small.

| ($\mu$, size) | $Q_G$ | $\tilde{Q}_r$ | $Q_J$ | $Q_B$ | $Bh.T_3$ | $Q_\gamma$ | $LR_{ML}$ | $LR_{REML}$ | $LR_{U.FeL}$ | $LR_{U.ReL}$ | $LR_C$ | $S_{ML}$ | $S_{REML}$ | $US_{FeL}$ | $US_{ReL}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (−2.5 LS) | 0.047 | 0.046 | 0.048 | 0.046 | 0.019 | 0.049 | 0.027 | 0.036 | 0.048 | 0.054 | 0.028 | 0.04 | 0.053 | 0.049 | 0.049 |
| (−5, LS) | 0.027 | 0.027 | 0.029 | 0.027 | 0 | 0.042 | 0.011 | 0.019 | 0.062 | 0.113 | 0.06 | 0.018 | 0.03 | 0.036 | 0.036 |
| (−2.5, SS) | 0.04 | 0.037 | 0.025 | 0.039 | 0.004 | 0.055 | 0.024 | 0.031 | 0.066 | 0.084 | 0.04 | 0.046 | 0.064 | 0.049 | 0.049 |
| (−5, SS) | 0 | 0 | 0 | 0 | 0 | 0.025 | 0.001 | 0.003 | 0.072 | 0.156 | 0.069 | 0.001 | 0.004 | 0.002 | 0.003 |

| ($\mu$, size) | $CS_{FeH}$ | $Z_{CS.FeH}$ | $CS_{ReH}$ | $BD$ | $MBD$ | $W_{ML}$ | $W_{REML}$ | $Peto$ | $Z^2_{WLS}$ | $Z^2_{WLS,R}$ | $Z^2_K$ | $Bh.T_4$ | $BS_{Bh.T_4}$ | $BS_{DL}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (−2.5, LS) | 0.049 | 0.045 | 0.042 | 0.049 | 0.049 | 0.001 | 0.002 | 0.048 | 0.03 | 0.095 | 0.127 | 0 | 0.060 | 0.014 |
| (−5, LS) | 0.036 | 0.036 | 0.041 | 0.036 | 0.036 | 0.001 | 0.001 | 0.049 | 0.017 | 0.119 | 0.143 | 0.003 | 0.041 | 0.016 |
| (−2.5, SS) | 0.047 | 0.046 | 0.054 | 0.049 | 0.049 | 0 | 0 | 0.059 | 0.026 | 0.11 | 0.142 | 0.006 | 0.058 | 0.011 |
| (−5, SS) | 0.002 | 0.005 | 0.016 | 0.002 | 0.002 | 0 | 0 | 0.019 | 0.014 | 0.375 | 0.225 | 0 | 0.068 | 0 |

Table 1.10: Actual Type I error rates of different homogeneity tests for settings with $R = 1$, $K = 20$, $\theta = 0$, $w = 0$, and $\tau^2 = 0$. Here, LS represents large sample and SS represents small sample.

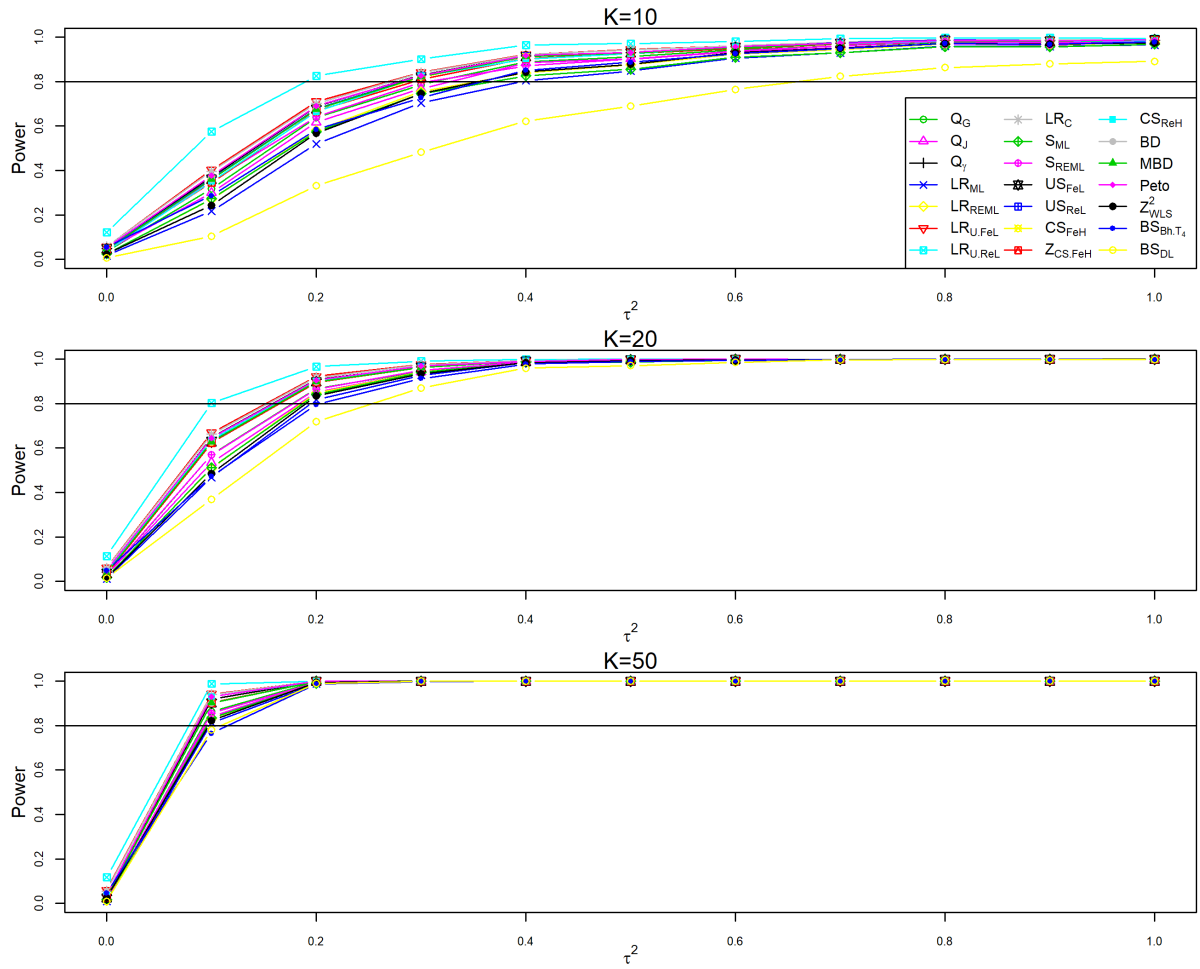



Figure 1.3: Power curves of different homogeneity tests for different $K$ values based on large-sample settings with $R = 1$, $\mu = -5$, $\theta = 0$, and $w = 0$.

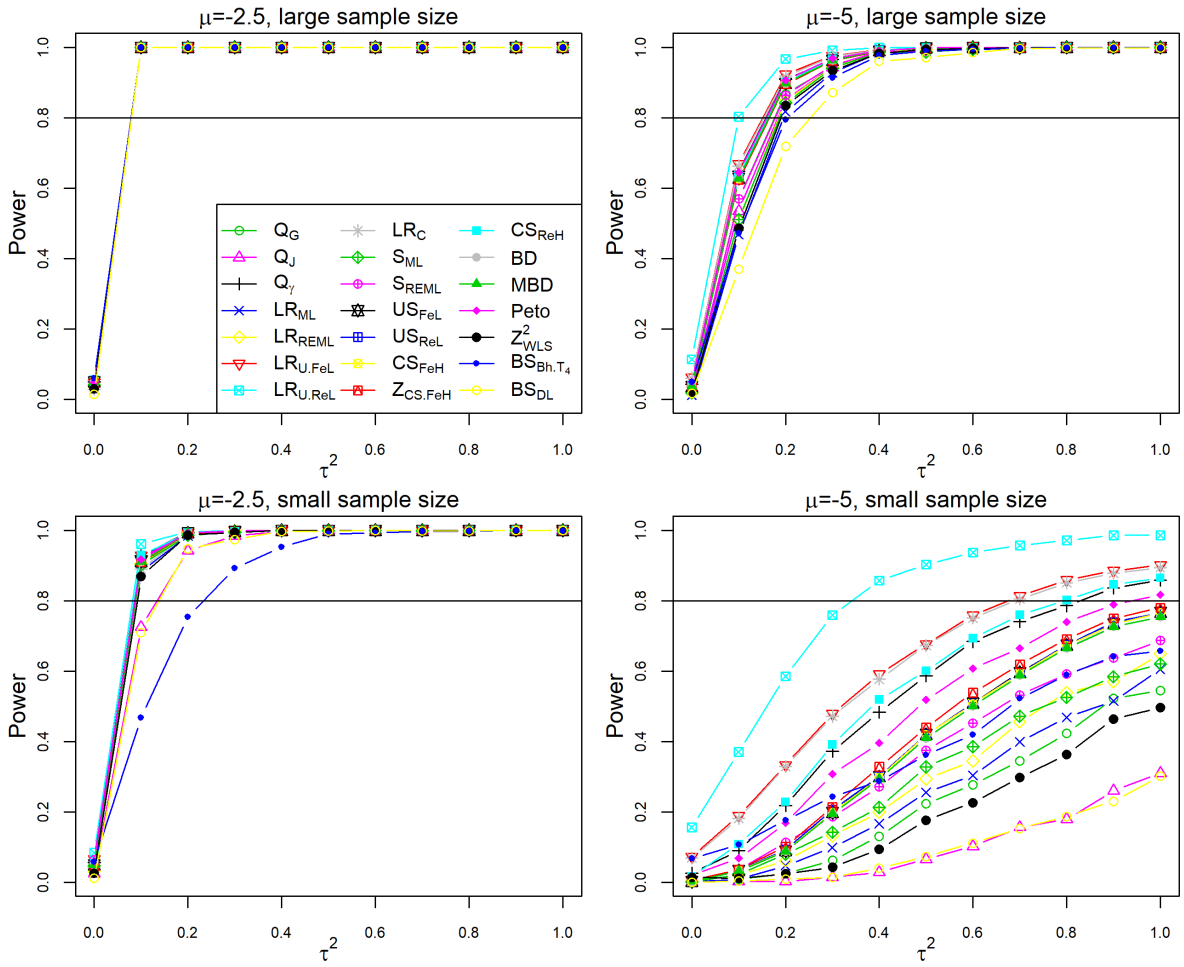



Figure 1.4: Power curves of different homogeneity tests for both large- and small-sample cases and different $\mu$ values based on settings with $R = 1$, $K = 20$, $\theta = 0$, and $w = 0$.

### 1.8.3. Comparison of different types of CIs

Among those summarized in Table 1.8, we compared 14 different types of 95% CIs for the heterogeneity parameter $\tau^2$ in Figures 1.5 and 1.6, excluding Bayesian credible intervals and the GV method as before. As mentioned in Section 1.7, $\text{BS}_{\text{NP}}$ represents the nonparametric bootstrap procedure combined with the $DL$ estimator and UTQ represents the unequal-tail $Q$-profile CI with $\alpha_1 = 0.01$ and $\alpha_2 = 0.04$. Again, from our (unreported) simulation results, we find that the influences of $R$, $\theta$, and $w$ on the empirical coverage probability are marginal.

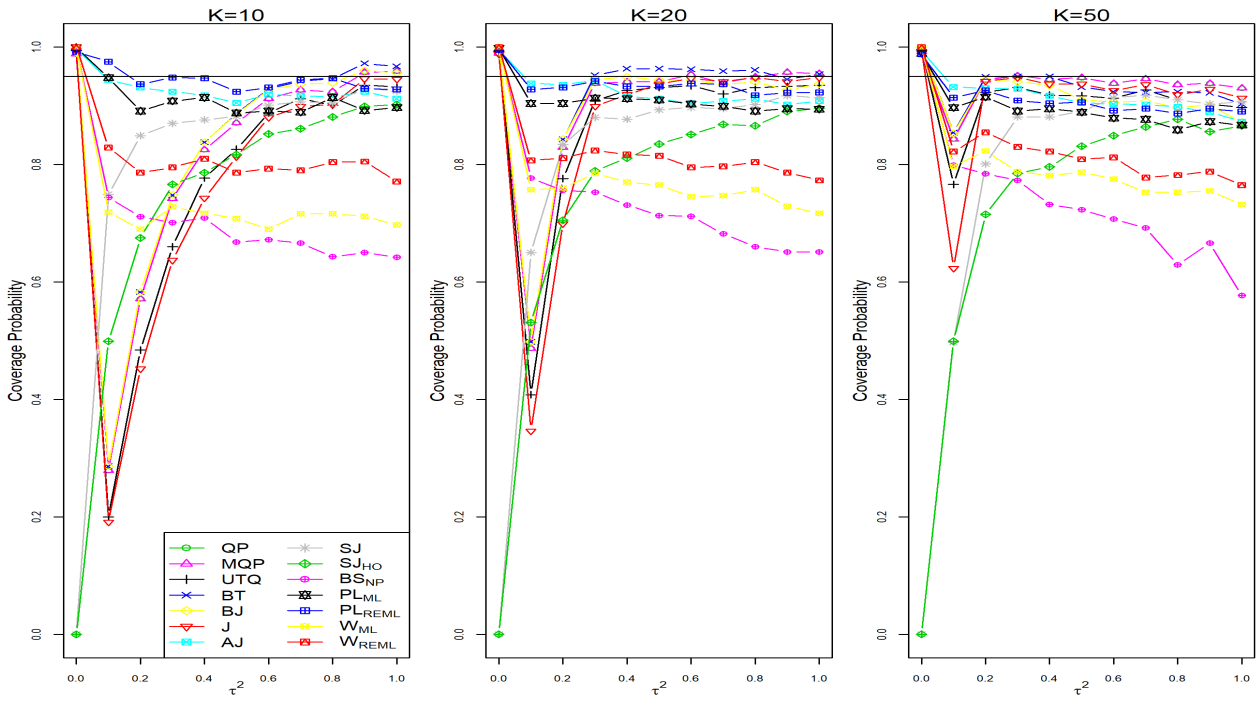



Figure 1.5: Actual coverage probabilities of different types of 95% CIs for different $K$ values based on large-sample settings with $R = 1$, $\mu = -5$, $\theta = 0$, and $w = 0$.

Figure 1.5 shows actual coverage probabilities of different types of CIs for different $K$ values based on large-sample settings with $R = 1$, $\mu = -5$, $\theta = 0$, and $w = 0$. When there is no between-study heterogeneity ($\tau^2 = 0$), all the methods provide 100% coverage except for SJ and $\text{SJ}_{\text{HO}}$ that produce strictly positive intervals and so have zero coverage. When $\tau^2$ is small, as $K$ increases, the methods based on (modified) $Q$ statistics gain some improvement in coverage except for AJ, which achieves relatively high coverage for all $K$ and $\tau^2$ values. As $\tau^2$ gets larger, most methods do not improve their coverage by increasing $K$.

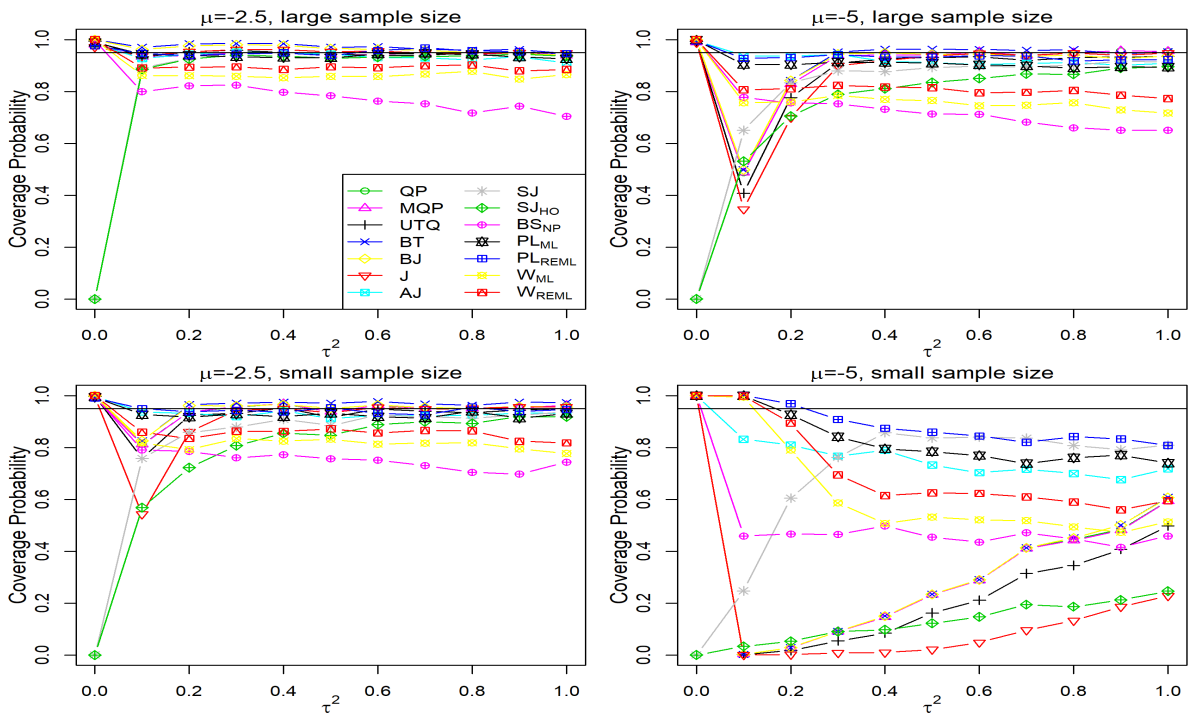



Figure 1.6: Actual coverage probabilities of different types of 95% CIs for both large- and small-sample cases and different $\mu$ values based on settings with $R = 1$, $K = 20$, $\theta = 0$, and $w = 0$.

Figure 1.6 presents actual coverage probabilities of different types of CIs for both large- and small-sample cases and different $\mu$ values based on settings with $R = 1$, $K = 20$, $\theta = 0$, and $w = 0$. When $\mu = -2.5$, most methods have actual coverage close to the nominal level 0.95. Among all, the nonparametric bootstrap CI has the lowest coverage, followed by the two Wald CIs when $\tau^2 > 0$. The influence of sample sizes is not obvious except for J, SJ and $\text{SJ}_{\text{HO}}$ that improve their coverage for large sample sizes when $\tau^2$ is small. For very rare events ($\mu = -5$), the impact of sample sizes is much more severe and some of the CIs (e.g., $\text{SJ}_{\text{HO}}$, J, UTQ) do not even achieve 50% coverage in most small-sample settings. In the large-sample settings, $\text{PL}_{\text{ML}}$, $\text{PL}_{\text{REML}}$, and AJ maintain the nominal 95% coverage quite well at all positive levels of $\tau^2$. As the sample sizes become small, all methods fail to do so for very rare events when $\tau^2 \geq 0.3$. Still, $\text{PL}_{\text{ML}}$ and $\text{PL}_{\text{REML}}$, and AJ are among those with the highest coverage. We also find that when $\tau^2 \geq 0.4$, SJ joins the top-performing group with the following order SJ $\approx \text{PL}_{\text{REML}} > \text{PL}_{\text{ML}} >$ AJ. This matches with the estimation

results reported in Section 1.8.1 that for very rare events coupled with small samples, the $SJ$ estimator is the least biased and has the smallest MSE when $\tau^2 \geq 0.5$. In such situations, the $Q$-statistic based CIs have generally low coverage and thus should be avoided; meanwhile the Wald and nonparametric bootstrap CIs have moderate coverage instead of being the worst in the other three cases.

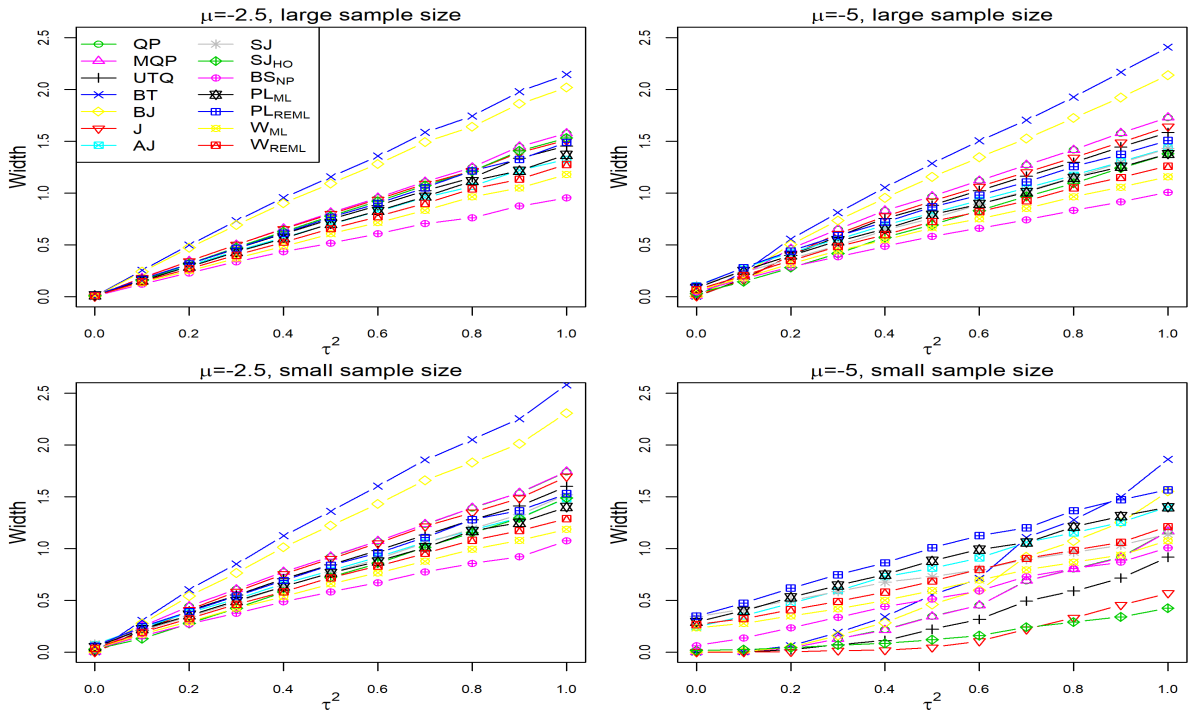



Figure 1.7: Width curves of different types of 95% CIs for both large- and small-sample cases and different $\mu$ values based on settings with $R = 1$, $K = 20$, $\theta = 0$, and $w = 0$.

Figure 1.7 shows width curves of different types of CIs under the same settings of Figure 1.6, where for all estimators, the width shows an increasing pattern as $\tau^2$ increases. The influence of sample sizes on the CI width is only obvious when $\mu = -5$, where all the CIs become narrower when sample sizes decrease. This is anti-intuitive. A closer examination reveals that when events are very rare and sample sizes are small, many simulation iterations produce confidence intervals of a point $\{0\}$, which makes the average width become smaller. In the first three situations (either $\mu = -2.5$ or large samples), BT and BJ produce the widest intervals, and PL and AJ intervals, which offer higher coverage than most other

methods, have moderate widths among all CIs. Unsurprisingly, the nonparametric bootstrap procedure produces the narrowest CIs. In the last situation (very rare events coupled with small samples), PL and AJ intervals are among the widest. Here, CIs with shorter widths are not necessarily desirable as they may reflect more $\{0\}$ intervals due to sparsity. SJ produces intervals with moderate widths though it also provides higher coverage when $\tau^2$ is large. Overall, we recommend PL and AJ intervals in general meta-analysis of rare binary events for their high coverage. For very rare events with small samples, we recommend SJ intervals if we know there exists at least moderate-level heterogeneity. Besides, AJ and SJ intervals are much easier to obtain than PL intervals.

## 1.9. Example

### 1.9.1. Type 2 diabetes mellitus after gestational diabetes

Women with gestational diabetes are believed to have a higher chance to develop type 2 diabetes. Bellamy et al. [6] performed a comprehensive systematic review and meta-analysis to assess the strength of this association. They selected 20 cohort studies that included 675,455 women with/without gestational diabetes and 10,859 type 2 diabetic events from 205 reports between Jan 1, 1960, and Jan 31, 2009, from Embase and Medline (see Table A.1 in Appendix A). We reanalyzed the data focusing on inference about the heterogeneity parameter $\tau^2$. We note that the overall event rate is $\sim$1.61% and many studies have very small sample sizes with zero event counts. So this data example fits in the scenario of very rare events coupled with small sample sizes. Recall that in this scenario, $SJ$ gives the least bias and most efficient estimator when there exists a moderate or large level of heterogeneity and $IPM$ is the second best which tends to underestimate $\tau^2$.

Point estimates for the heterogeneity parameter $\tau^2$ and corresponding inverse-variance weighted estimates for the treatment effect $\theta$ are summarized in Table 1.11(a). Here, most methods give an estimate between 0.4 and 0.7 for $\tau^2$, where the estimate from $IPM$ is 0.563 and that from $SJ$ is 0.679. This seems to suggest a moderate to large level of heterogeneity, especially after accounting for the underestimation from $IPM$. The $RB_p$ method, which has been shown to severely overestimate $\tau^2$ for very rare events, gives the largest estimate of 1.162 as we expect. On the other hand, the $HS$ estimate is much smaller than the others. The resulting estimated odds ratios do not vary as much except for the one from $RB_p$. The p-values from testing $\mathcal{H}_0 : \tau^2 = 0$ are presented in Table 1.11(b). We find that except for $Bh.T_3$, $Bh.T_4$, $Z^2_{WLS,R}$ and $Z^2_K$, all other methods reject the null hypothesis of homogeneity at the significance level 0.05. This is not surprising as the four tests have low power in detecting the existence of heterogeneity, as mentioned in Section 1.8.2. Table 1.11(c) shows the confidence intervals from all the compared methods. BT gives a very large upper bound, which seems to be odd. All CIs except for those from BT, BJ, and Wald methods exclude zero, among which SJ yields the shortest interval with the largest lower bound and the upper bound in line with that from PL and AJ methods. Recall that SJ tends to produce the best interval with higher coverage and relatively shorter width when there exists at least moderate-level heterogeneity, as reported in Section 1.8.3. In this example, we lean toward reporting the SJ interval, among the top performing methods PL, AJ and SJ. Based on the estimation and inference results above, we believe that these studies are heterogeneous.

| Estimator | $HO$ | $HO_2$ | $DL$ | $DL_2$ | $DL_p$ | $DL_b$ | $PM$ |
|---|---|---|---|---|---|---|---|
| $\hat{\tau}^2$ | 0.220 | 0.418 | 0.466 | 0.411 | 0.466 | 0.265 | 0.413 |
| $\hat{\theta}$ | 2.093 | 2.136 | 2.146 | 2.135 | 2.146 | 2.104 | 2.135 |
| OR | 8.112 | 8.469 | 8.547 | 8.457 | 8.547 | 8.197 | 8.461 |
| 95% CI for OR | (5.658, 11.630) | (5.435, 13.197) | (5.395, 13.540) | (5.442, 13.141) | (5.395, 13.540) | (5.552, 12.293) | (5.439, 13.162) |

| Estimator | $IPM$ | $HM$ | $HS$ | $LCH_{mean}$ | $LCH_{median}$ | $ML$ | $REML$ |
|---|---|---|---|---|---|---|---|
| $\hat{\tau}^2$ | 0.563 | 0.419 | **0.046** | 0.519 | 0.298 | 0.396 | 0.449 |
| $\hat{\theta}$ | 2.162 | 2.137 | 2.092 | 2.155 | 2.111 | 2.132 | 2.142 |
| OR | 8.691 | 8.470 | 8.099 | 8.626 | 8.260 | 8.432 | 8.520 |
| 95% CI for OR | (5.321, 14.194) | (5.435, 13.200) | (6.424, 10.210) | (5.354, 13.897) | (5.553, 12.285) | (5.455, 13.034) | (5.409, 13.419) |

| Estimator | $AREML$ | $SJ$ | $SJ_{HO}$ | $RB_0$ | $RB_p$ | $BM$ |
|---|---|---|---|---|---|---|
| $\hat{\tau}^2$ | 0.433 | 0.679 | 0.290 | 0.198 | **1.162** | 0.195 |
| $\hat{\theta}$ | 2.139 | 2.180 | 2.110 | 2.088 | 2.235 | 2.088 |
| OR | 8.493 | 8.846 | 8.245 | 8.072 | 9.345 | 8.067 |
| 95% CI for OR | (5.432, 13.302) | (5.241, 14.932) | (5.562, 12.223) | (5.694, 11.443) | (4.953, 17.631) | (5.698, 11.419) |

(a) Estimates of the heterogeneity parameter $\tau^2$ and treatment effect $\theta$ from different methods

| Test | $Q_G$ | $\tilde{Q}_r$ | $Q_J$ | $Q_B$ | $Bh.T_3$ | $Q_\gamma$ | $LR_{ML}$ | $LR_{REML}$ | $LR_{U.FeL}$ | $LR_{U.ReL}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| P-value | 0.000 | 0.000 | 0.011 | 0.000 | 0.571 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |

| Test | $LR_C$ | $S_{ML}$ | $S_{REML}$ | $US_{FeL}$ | $US_{ReL}$ | $CS_{FeH}$ | $Z_{CS.FeH}$ | $CS_{ReH}$ | $BD$ | $MBD$ |
|---|---|---|---|---|---|---|---|---|---|---|
| P-value | 0.000 | 0.014 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |

| Test | $W_{ML}$ | $W_{REML}$ | $W^2_{FeL}$ | $Peto$ | $Z^2_{WLS}$ | $Z^2_{WLS,R}$ | $Z^2_K$ | $Bh.T_4$ | $BS_{Bh.T_4}$ | $BS_{DL}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| P-value | 0.041 | 0.044 | 0.000 | 0.000 | 0.000 | 0.055 | 0.184 | 0.256 | $<0.05$ | $<0.05$ |

(b) P-values from different methods for testing homogeneity of odds ratios

| Method | QP | MQP | UTQ | BT | BJ | J | AJ |
|---|---|---|---|---|---|---|---|
| CI | (0.109, 1.603) | (0.106, 1.603) | (0.083, 1.403) | [0, 8.610) | [0, 2.660) | (0.048, 1.540) | (0.004, 1.396) |

| Method | SJ | $SJ_{HO}$ | $BS_{NP}$ | $PL_{ML}$ | $PL_{REML}$ | $W_{ML}$ | $W_{REML}$ |
|---|---|---|---|---|---|---|---|
| CI | (0.393,1.449) | (0.168, 0.620) | (0.012, 0.670) | (0.113, 1.285) | (0.129, 1.458) | [0, 0.841) | [0, 0.966) |

(c) Confidence intervals for the heterogeneity parameter $\tau^2$ from different methods

Table 1.11: Data example of gestational diabetes meta-analysis

### 1.9.2. Rosiglitazone meta-analysis

The side effect of rosiglitazone on cardiovascular (CV) safety has been evaluated by [57] with a meta-analysis of 48 trials, which concluded a significantly elevated risk for myocardial infarction (MI) and a borderline significant increased risk for cardiovascular death (CVD). However, debate over the CV safety of rosiglitazone remained as some subsequent

meta-analyses reported inconclusive results [72, 22]. To address the controversy, Nissen and Wolski [58] conducted an updated meta-analysis of 56 trials including 35,531 patients: 19,509 assigned to treatment groups (rosiglitazone) and 16,022 assigned to control group (see Table A.2 in Appendix A). Here we reanalyzed the data focusing on inference about the heterogeneity parameter $\tau^2$. Among all the 56 trials, only three have sample sizes larger than 2000 in either groups. In the rosiglitazone groups, there are 159 MI and 103 CVD cases reported while in the control groups, there are 136 MI and 98 CVD cases. The overall event rate is 0.83% for MI and 0.57% for CVD. So this data example fits in the scenario of very rare events coupled with small sample sizes.

Table 1.12 summarizes results for the MI data from different aspects and methods. Point estimates for $\tau^2$ from different estimators and corresponding weighted average estimates for $\theta$ and odds ratio are presented in Table 1.12(a). A 95% t-confidence interval is also given for OR. All the methods except for those positive ones give zero estimates for $\tau^2$. All the CIs for OR include 1, indicating no treatment effect of rosiglitazone. Here we favor the conclusion of $\tau^2 = 0$ instead of $0 < \tau^2 \leq 0.25$ for the following reason. Based on the small-sample simulation results for estimation bias with very rare events (Figure 1.2), all the nonnegative estimator are unbiased when $\tau^2 = 0$. The positive bias of 0.659 from $RB_p$ here is very close to that of 0.65 when $\tau^2 = 0$ in the simulation. On the other hand, if $0 < \tau^2 \leq 0.25$, $IPM$, $HM$, $SJ$, and $RB_p$ are expected to have (much) larger estimates than what was observed here.

Results from hypothesis testing and confidence intervals for $\tau^2$ presented in Table 1.12(b) and 1.12(c) further consolidate our conclusion. All the tests, except for $Z^2_{WLS}$, $Z^2_{WLS,R}$, and $Z^2_K$, fail to reject the null hypothesis of homogeneous effects. As shown in Table 1.10, when $\mu = -5$ and sample sizes are small, $Z^2_{WLS,R}$, and $Z^2_K$ have largely inflated Type I error rate of 0.375 and 0.225 for a test at the significance level of $\alpha = 0.05$ while most of the other tests maintain sizes well below $\alpha$. Thus we believe that those three tests falsely reject the null hypothesis and conclude that the treatment effects in different studies are homogeneous. All

the CIs for $\tau^2$ include zero except for SJ and $\text{SJ}_{\text{HO}}$ as presented in Table 1.12(c). Due to their positive nature, the SJ and $\text{SJ}_{\text{HO}}$ CIs have zero coverage when $\tau^2 = 0$, which is mostly likely to be true in this example. Combining all the evidences from estimation and testing, we conclude that $\tau^2 = 0$ and rosiglitazone has no significant effect on MI.

| Estimator | $HO$ | $HO_2$ | $DL$ | $DL_2$ | $DL_p$ | $DL_b$ | $PM$ |
|---|---|---|---|---|---|---|---|
| $\hat{\tau}^2$ | 0.000 | 0.000 | 0.000 | 0.000 | 0.010 | 0.000 | 0.000 |
| $\hat{\theta}$ | 0.159 | 0.159 | 0.159 | 0.159 | 0.156 | 0.159 | 0.159 |
| OR | 1.172 | 1.172 | 1.172 | 1.172 | 1.168 | 1.172 | 1.172 |
| 95% CI for OR | (0.937, 1.467) | (0.937, 1.467) | (0.937, 1.467) | (0.937, 1.467) | (0.920, 1.484) | (0.937, 1.467) | (0.937, 1.467) |
| Estimator | $IPM$ | $HM$ | $HS$ | $LCH_{mean}$ | $LCH_{median}$ | $ML$ | $REML$ |
| $\hat{\tau}^2$ | 0.000 | 0.040 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| $\hat{\theta}$ | 0.159 | 0.145 | 0.159 | 0.159 | 0.159 | 0.159 | 0.159 |
| OR | 1.172 | 1.156 | 1.172 | 1.172 | 1.172 | 1.172 | 1.172 |
| 95% CI for OR | (0.937, 1.467) | (0.883, 1.514) | (0.937, 1.467) | (0.937, 1.467) | (0.937, 1.467) | (0.937, 1.467) | (0.937, 1.467) |
| Estimator | $AREML$ | $SJ$ | $SJ_{HO}$ | $RB_0$ | $RB_p$ | $BM$ | |
| $\hat{\tau}^2$ | 0.000 | 0.156 | 0.003 | 0.000 | **0.659** | 0.002 | |
| $\hat{\theta}$ | 0.159 | 0.116 | 0.158 | 0.159 | 0.073 | 0.158 | |
| OR | 1.172 | 1.123 | 1.171 | 1.172 | 1.076 | 1.172 | |
| 95% CI for OR | (0.937, 1.467) | (0.810, 1.558) | (0.931, 1.473) | (0.937, 1.467) | (0.706, 1.639) | (0.933, 1.470) | |

(a) Estimates of the heterogeneity parameter $\tau^2$ and treatment effect $\theta$ from different methods

| Test | $Q_G$ | $\tilde{Q}_r$ | $Q_J$ | $Q_B$ | $Bh.T_3$ | $Q_\gamma$ | $LR_{ML}$ | $LR_{REML}$ | $LR_{U.FeL}$ | $LR_{U.ReL}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| P-value | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 0.983 | 0.5 | 0.5 | 0.935 | 0.879 |
| Test | $LR_C$ | $S_{ML}$ | $S_{REML}$ | $US_{FeL}$ | $US_{ReL}$ | $CS_{FeH}$ | $Z_{CS.FeH}$ | $CS_{ReH}$ | $BD$ | $MBD$ |
| P-value | 0.937 | 0.898 | 0.901 | 1.000 | 1.000 | 1.000 | 1.000 | 0.879 | 1.000 | 1.000 |
| Test | $W_{ML}$ | $W_{REML}$ | $W^2_{FeL}$ | $Peto$ | $Z^2_{WLS}$ | $Z^2_{WLS,R}$ | $Z^2_K$ | $Bh.T_4$ | $BS_{Bh.T_4}$ | $BS_{DL}$ |
| P-value | 0.5 | 0.5 | NA | 1.000 | 0.001 | 0.000 | 0.003 | 1.000 | >0.05 | >0.05 |

(b) P-values from different methods for testing homogeneity of odds ratios

| Method | QP | MQP | UTQ | BT | BJ | J | AJ |
|---|---|---|---|---|---|---|---|
| CI | [0, 0] | [0, 0] | [0, 0] | [0, 0] | [0, 0] | [0, 0] | [0, 0] |
| Method | SJ | $\text{SJ}_{\text{HO}}$ | $\text{BS}_{\text{NP}}$ | $\text{PL}_{\text{ML}}$ | $\text{PL}_{\text{REML}}$ | $\text{W}_{\text{ML}}$ | $\text{W}_{\text{REML}}$ |
| CI | (0.111, 0.235) | (0.002, 0.005) | [0, 0] | [0, 0.091) | [0, 0.111) | [0, 0.080) | [0, 0.108) |

(c) Confidence intervals for the heterogeneity parameter $\tau^2$ from different methods

Table 1.12: Results for MI data in rosiglitazone meta-analysis

Table 1.13 summarizes results for the CVD data. Similar observations can be made from Table 1.13: all non-negative $\tau^2$ estimates are zero and the positive ones have values even smaller than those for the MI data, all estimated ORs are close to 1, and all their CIs include one. Thus, we conclude that $\tau^2 = 0$ and rosiglitazone has no significant effect on CVD, too.

| Estimator | $HO$ | $HO_2$ | $DL$ | $DL_2$ | $DL_p$ | $DL_b$ | $PM$ |
|---|---|---|---|---|---|---|---|
| $\hat{\tau}^2$ | 0.000 | 0.000 | 0.000 | 0.000 | 0.010 | 0.000 | 0.000 |
| $\hat{\theta}$ | -0.059 | -0.059 | -0.059 | -0.059 | -0.045 | -0.059 | -0.059 |
| OR | 0.943 | 0.943 | 0.943 | 0.943 | 0.956 | 0.943 | 0.943 |
| 95% CI for OR | (0.727, 1.223) | (0.727, 1.223) | (0.727, 1.223) | (0.727, 1.223) | (0.723, 1.256) | (0.727, 1.223) | (0.727, 1.223) |
| Estimator | $IPM$ | $HM$ | $HS$ | $LCH_{mean}$ | $LCH_{median}$ | $ML$ | $REML$ |
| $\hat{\tau}^2$ | 0.000 | 0.020 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| $\hat{\theta}$ | -0.059 | -0.034 | -0.059 | -0.059 | -0.059 | -0.059 | -0.059 |
| OR | 0.943 | 0.966 | 0.943 | 0.943 | 0.943 | 0.943 | 0.943 |
| 95% CI for OR | (0.727, 1.223) | (0.720, 1.297) | (0.727, 1.223) | (0.727, 1.223) | (0.727, 1.223) | (0.727, 1.223) | (0.727, 1.223) |
| Estimator | $AREML$ | $SJ$ | $SJ_{HO}$ | $RB_0$ | $RB_p$ | $BM$ | |
| $\hat{\tau}^2$ | 0.000 | 0.061 | 0.002 | 0.000 | 0.429 | 0.002 | |
| $\hat{\theta}$ | -0.059 | -0.009 | -0.056 | -0.059 | 0.011 | -0.056 | |
| OR | 0.943 | 0.991 | 0.945 | 0.943 | 1.011 | 0.946 | |
| 95% CI for OR | (0.727, 1.223) | (0.709, 1.383) | (0.726, 1.231) | (0.727, 1.223) | (0.654, 1.562) | (0.726, 1.232) | |

(a) Estimates of the heterogeneity parameter $\tau^2$ and treatment effect $\theta$ from different methods

| Test | $Q_G$ | $\tilde{Q}_r$ | $Q_J$ | $Q_B$ | $Bh.T_3$ | $Q_\gamma$ | $LR_{ML}$ | $LR_{REML}$ | $LR_{U.FeL}$ | $LR_{U.ReL}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| P-value | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 0.993 | 0.500 | 0.500 | 1.000 | 0.967 |
| Test | $LR_C$ | $S_{ML}$ | $S_{REML}$ | $US_{FeL}$ | $US_{ReL}$ | $CS_{FeH}$ | $Z_{CS.FeH}$ | $CS_{ReH}$ | $BD$ | $MBD$ |
| P-value | 1.000 | 0.814 | 0.835 | 1.000 | 1.000 | 1.000 | 1.000 | 0.645 | 1.000 | 1.000 |
| Test | $W_{ML}$ | $W_{REML}$ | $W^2_{FeL}$ | $Peto$ | $Z^2_{WLS}$ | $Z^2_{WLS,R}$ | $Z^2_K$ | $Bh.T_4$ | $BS_{Bh.T_4}$ | $BS_{DL}$ |
| P-value | 0.500 | 0.500 | NA | 1.000 | 0.000 | 0.000 | 0.002 | 1.000 | >0.05 | >0.05 |

(b) P-values from different methods for testing homogeneity of odds ratios

| Method | QP | MQP | UTQ | BT | BJ | J | AJ |
|---|---|---|---|---|---|---|---|
| CI | [0, 0] | [0, 0] | [0, 0] | [0, 0] | [0, 0] | [0, 0] | [0, 0] |
| Method | SJ | $SJ_{HO}$ | $BS_{NP}$ | $PL_{ML}$ | $PL_{REML}$ | $W_{ML}$ | $W_{REML}$ |
| CI | (0.043, 0.092) | (0.001, 0.003) | [0, 0] | [0, 0.130) | [0, 0.161) | [0, 0.085) | [0, 0.164) |

(c) Confidence intervals for the heterogeneity parameter $\tau^2$ from different methods

Table 1.13: Results for CVD data in rosiglitazone meta-analysis

## 1.10. Discussion and recommendations

In the development of reliable statistical techniques for research synthesis, much effort has been put on inference about the effect sizes. However, identifying and quantifying between-study heterogeneity is important as well. Inspired by the wide application of meta-analysis in medical research that focuses on the odds ratio between two conditions in the presence of a binary response of disease status, treatment efficacy or an adverse reaction, we have made our best effort to fully review models applicable to meta-analysis with binary outcomes and heterogeneity measures, and thoroughly evaluated inference procedures about the between-study variance $\tau^2$ covering point estimation, interval estimation, and hypothesis testing. Our simulation studies focus on *rare* binary events, which have the greatest practical importance as most diseases or adverse events have (very) low incidence rates but are often ignored in previous comparative studies. Unlike published reviews that focus on either estimation or testing and only include a limited subset of methods, we attempt to provide a systematic and updated review that includes all applicable methods covering both estimation and testing for rare binary events.

Based on our comprehensive simulation studies for large-sample meta-analysis of rare binary events, we recommend the $IPM$ method for estimating the heterogeneity parameter $\tau^2$ if reducing estimation bias is of high priority, especially when the events are extremely rare. Most of the methods do not differ much in terms of MSE. We suggest to avoid using $HM$, $HS$ and $BM$ since they have relatively large bias and MSE compared with other estimators. The most widely used $DL$ estimator and its one-step variants $DL_p$ and $DL_b$ do not perform satisfactorily and should also be avoided. For small-sample meta-analysis of rare events, $IPM$ is still recommended and $SJ$ also performs much better than the other estimators in terms of both bias and MSE when $\tau^2 \geq 0.5$ and the events are extremely rare. In terms of hypothesis testing of homogeneity of odds ratios, we recommend the $LR$ test based on the $FeL$ model and the conditional $LR$ test ( $LR_{U.FeL}$ and $LR_C$). Regardless of sample sizes, we

should avoid $Bh.T_3$, $W_{ML}$, $W_{REML}$, $Z^2_{WLS,R}$ and $Z^2_K$ for their largely biased test sizes and low power. In terms of interval estimation, we recommend the profile likelihood methods ($\text{PL}_{\text{ML}}$ and $\text{PL}_{\text{REML}}$) and the approximate Jackson method AJ in general situations. Among the three, $\text{PL}_{\text{REML}}$ usually produces higher coverage but with wider intervals. The SJ method is a good candidate when events are extremely rare, sample sizes are small, and $\tau^2 \geq 0.4$. We did not examine the performance of Bayesian methods in estimation and inference of the heterogeneity parameter because of the computation burden, convergence detection issue, and potential sensitivity to prior choices. However, as presented in [3], Bayesian hierarchical modeling can be a good alternative in eliminating estimation bias and improving testing power.

We also notice that most estimators for $\tau^2$ are negatively biased in our simulation, an interesting phenomenon observed in other simulation studies with binary outcomes [42, 69, 70, 7] as well. In simulation studies with continuous outcomes [43], most of the estimators show positive bias when $\tau^2$ is small ($< 0.1$) and the magnitude of bias of $RB_p$ is much larger than the other estimators; for larger $\tau^2$ values, the $HS$ and $ML$ estimators are negatively biased and the magnitude increases as $\tau^2$ increases [86]. Viechtbauer [86] provides some analytical results for the bias of estimators $HO$, $DL$, $HS$, $ML$, and $REML$. Most of these results were derived based on the homogeneous within-study variance assumption ($\sigma^2_k = \sigma^2$). Under this assumption, the bias due to truncation is always positive for $DL$, $HO$ and $REML$ with all levels of heterogeneity and is negative for $HS$ and $ML$ when $\tau^2 \geq 0.5$. However, we believe that in the rare events context, it is the sparsity (caused by zero counts) and lack of accuracy in estimating the within-study variances that cause the large magnitude of underestimation for many methods. This underestimation is much reduced by the $IPM$ estimator where the within-study variance estimates are improved by pooling information from all the studies.

All testing procedures, except for Bayesian, bootstrap, and generalized variable methods, use asymptotic null distributions to obtain p-values. Reis et al. [66] examined the exact

version of several tests, including $US_{FeL}$, $CS_{FeH}$, and $LR_{FeL}$, where the exact p-value is given by the null conditional probability of attaining a statistic value not less than the observed one. These exact tests were shown to maintain the test size better in small-sample or sparse data situations and the power is only slightly higher than the corresponding asymptotic tests under certain settings. However, the exact p-value can be much more computationally expensive to obtain and may not worth the effort in common meta-analysis with a moderate number of studies and sample sizes.

We use two data examples to illustrate inference and interpretation of between-study heterogeneity in meta-analysis of rare binary events. These two applications are somewhat representative: the diabetes example shows heterogeneous studies and a significant overall effect; the rosiglitazone example shows no heterogeneity and no significant effect. Sometimes, unanimous conclusions are hard to be drawn when different methods provide different indications and then simulation results can provide us useful guidance. It is worth mentioning that our conclusion about the effect of rosiglitazone on MI is different from those in [58], which used Peto and Mantel-Haenszel methods in estimating ORs and 95% CIs. Both methods assume a fixed treatment effect, i.e., a common OR, which is also the assumption in [51] mentioned in Section 1.6. Thus different assumptions about the treatment effects in different studies can lead to contrary conclusions about the inference of the overall effect. The potential heterogeneity should be carefully accounted for in a meta-analysis. Our conclusion matches with those in [47] and supports the actions that FDA lifted its earlier restrictions on rosiglitazone in 2013 and further eliminated the Risk Evaluation and Mitigation Strategy (REMS) in 2015.

Finally, we should mention that, when synthesizing information from multiple studies to get more reliable conclusions, one should not simply rely on one point estimate or one p-value (especially those from the default methods in software packages) without considering the rich selection of statistical tools offered in the literature. Each of the above reviewed models or methods has its own limitations. In practice, all kinds of evidence should be combined and

evaluated together with the specific characteristics of component studies included in the meta-analysis.

CHAPTER 2

Estimation of Variances in Cluster Randomized Designs Using Ranked Set Sampling

## 2.1. Introduction

Cluster randomized design (CRD) is a commonly used statistical design with a hierarchical structure in many agricultural and educational studies. A typical CRD involves a two-stage sampling, at the cluster level and at the individual level. Simple random sampling (SRS) is usually applied in both stages to obtain samples. Standard textbooks, such as [79] and [52] present details of such design with SRS. An example of the application of CRD in educational studies is the evaluation of schools' performance where schools are nested within their corresponding school districts. Here school districts are clusters and schools are individuals. In agricultural applications, for example, when evaluating the crop production in the United States, we treat states as clusters and counties within each state as individuals. Such hierarchical design can better address the potential heterogeneity among different clusters while individuals within a cluster are homogeneous.

In order to improve cost-efficiency in the sampling process, McIntyre [55] proposed ranked set sampling (RSS) in the estimation of pasture yields. This sampling method can reduce the amount of measurement, and thus reduce the cost, by incorporating the ranking information, which is presumably cheaper to acquire than to take actual measures. In general, one can perform RSS as follows. Let $H$ be a predetermined size of a ranked set. In each iteration $i$, $i = 1, \ldots, H$, select a set of random samples containing $H$ sampling units from the population using SRS. Order the $H$ units within that set from the smallest to the largest based on some ranking variable $X$ that is easy to measure and is related the the parameter of interest.

For example, to estimate crop production of farms, the farm size can be used as a ranking variable which is easily obtainable and is highly related to the crop yield. Then, measure the $i$th unit in the set and discard the remaining $H-1$ units. We refer to the above process of $H$ iterations as one cycle. We can obtain a total of $N = Hm$ observations after $m$ cycles. Although $H^2m$ sampling units are randomly selected from the population, only $Hm$ ones are actually measured. This is only an illustration of the process of RSS in a one stage sampling. In a two-stage CRD sampling, the RSS procedure can be applied at the cluster level only, at the individual level only or at both levels. For sampling at the cluster level, clusters are the sampling units, and for sampling at the individual level, RSS is performed for each cluster and the individuals within the corresponding cluster become the population.

Most of the existing research on RSS and the application of RSS with CRDs focus on the inference about the mean. For example, for one level sampling, Takahasi and Wakimoto [76] first proved that under perfect ranking, the RSS mean is an unbiased estimator of the population mean and is more efficient than the SRS sample average in terms of mean squared error (MSE). Dell and Clutter [19] derived closed form formula for the relative precision (RP) of the RSS mean versus the SRS mean under perfect ranking. For the application of RSS with CRD on inference about population mean, Wang et al. [90] developed a nonparametric estimator for the treatment effect, studied its theoretical properties, and quantified the magnitude of improvement over the corresponding SRS-based estimator. They also proposed a new test to detect treatment effects. The variance components can be important in finding an efficient design. For example, Chen and Lim [13] proposed estimators of the variances in ranked set sampling, which can help obtain an optimal design for unbalanced RSS. However, related topics were much less studied compared to the inference of the mean. In terms of the estimation and inference on the variance using RSS in a one-stage sampling, Stokes [75] proposed an asymptotically unbiased estimator. Sinha et al. [73] and Philip et al. [65] developed improved variance estimators for the Normal distribution with perfect judgment ranking. MacEachern et al. [53] proposed an alternative unbiased estimator which is more efficient than Stokes's estimator and applies it to non-normal distributions and imperfect

ranking. However, all above mentioned RSS-based variance estimators were designed for one-level models. To the best of our knowledge, no study has focused on estimating variances in CRDs using RSS, which is the topic of the second part of this thesis. Inference about the variances, especially the between-cluster variance, can be very informative as it can assist the statistical design suited for applications with differing population characteristics. If the between-study variance is relatively small, the benefit may be marginal to conduct a cluster randomized design.

In the second part of the thesis, we explore the use of RSS with CRDs in the estimation of the variance components. We first introduce the population model and RSS-structured data in Section 2.2. In Section 2.3, we derive the nonparametric method of moment (MOM) estimators for the between and within cluster variances. In Section 2.4, we evaluate the impact of design parameters, ranking schemes, and the presence of imperfect ranking on the performance of the newly proposed estimators via simulation studies. In Section 2.5, we propose the Panle and Mandel (PM) estimator for the between cluster variance, which is inspired by the PM estimator for the heterogeneity parameter in meta-analysis. In Section 2.6, we present the improvement in efficiency of the PM estimator over the MOM estimator. In Section 2.7, we provide an example using the California API data. Section 2.8 concludes with discussions and possible extensions to the proposed work.

## 2.2. Data, model and notation

### 2.2.1. The population model

We consider a typical CRD with two-stage sampling [90], using SRS to sample clusters and individuals within each selected cluster. Let $Y_{k(ij)}$ be the response of the $k$th subject of the $j$th cluster in the $i$th treatment group, $k = 1, \ldots K_{j(i)}$, $j = 1, \ldots J_i$, $i = 0, 1$, where $K_{j(i)}$

is the sample size in cluster $j$ under treatment $i$, and $J_i$ is the number of selected clusters for treatment $i$. The HLM reflecting the nested structure under the traditional CRD can be written as

$$Y_{k(ij)} = \mu + a_i + b_{j(i)} + r_{k(ij)}, \tag{2.1}$$

where $\mu$ is the mean response of the control group; $a_i$ is the fixed effect of treatment $i$ with $a_0 \equiv 0$; $b_{j(i)}$ is the random effect of cluster $j$ nested in treatment $i$; $r_{k(ij)}$ is the random error term. We assume $b_{j(i)}$'s follow a common but unknown distribution with mean 0 and variance $\sigma_b^2$, and $r_{k(ij)}$'s are independent and identically distributed random variables from an unspecified distribution with mean 0 and variance $\sigma_r^2$. We also assume that $b_{j(i)}$'s and $r_{k(ij)}$'s are independent. Under model (2.1), responses from subjects in the same cluster are dependent and those from different clusters are independent. Define the intra-class correlation (ICC), a measure of the variation among different clusters, as $\sigma_b^2/(\sigma_b^2 + \sigma_r^2)$. Here we are interested in estimating $\sigma_b^2$ and $\sigma_r^2$.

### 2.2.2. RSS-based data and notation

Besides individual outcomes, ranking information becomes available for data from RSS-structured CRDs. As mentioned previously, there are three possible ranking schemes in a two-stage CRD: (1) ranking at the cluster level only, (2) ranking at the individual level only, and (3) ranking at both levels. For ranking at the cluster level only, let $H_i^c$ denote the set size for treatment group $i$, $m_i^c$ denote the number of cycles for group $i$, $O_{j(i)}^c$ denote the (judgmental) order of cluster $j$ in group $i$ among its own comparison set, and $K_{j(i)}$ denote the number of individuals in cluster $j$ in group $i$. Then the data can be expressed by $\boldsymbol{D}^c = \{Y_{k(ij)}, O_{j(i)}^c\}$ given the design parameters $\{H_i^c, m_i^c, K_{j(i)}\}$ for $k = 1, \ldots, K_{j(i)}$, $j = 1, \ldots, J_i$ ($J_i = H_i^c \times m_i^c$), and $i = 0, 1$. For ranking at the individual level only, let $J_i$ be the number of clusters under treatment $i$, $H_{j(i)}^{id}$ be the set size for cluster $j$ in treatment group $i$, $m_{j(i)}^{id}$ be the number of cycles for cluster $j$ in treatment group $i$, and $O_{k(ij)}^{id}$ be the

(judgmental) order of individual $k$ within cluster $j$ under treatment group $i$ among its own comparison set. Then the data can be expressed by $\boldsymbol{D}^{id} = \{Y_{k(ij)}, O^{id}_{k(ij)}\}$ given the design parameters $\{J_i, H^{id}_{j(i)}, m^{id}_{j(i)}\}$ for $k = 1, \ldots, K_{j(i)}$ ($K_{j(i)} = H^{id}_{j(i)} \times m^{id}_{j(i)}$), $j = 1, \ldots, J_i$, and $i = 0, 1$. For ranking at both levels, the data can be expressed by $\boldsymbol{D}^{b} = \{Y_{k(ij)}, O^{c}_{j(i)}, O^{id}_{k(ij)}\}$ given the design parameters $\{H^{c}_{i}, m^{c}_{i}, H^{id}_{j(i)}, m^{id}_{j(i)}\}$ for $k = 1, \ldots, K_{j(i)}$, $j = 1, \ldots, J_i$, $i = 0, 1$. Note single level ranking is a special case of ranking at both levels: $H^{id}_{j(i)} = 1$ and $m^{id}_{j(i)} = K_{j(i)}$ for ranking at the cluster level only; and $H^{c}_{i} = 1$ and $m^{c}_{i} = J_i$ for ranking at the individual level only. As in [90], we drop the superscripts "c", "id", and "b" when no ambiguity exists. In the HLM of (2.1), the term $b_{j(i)}$ represents the average effect of cluster $j$ in treatment $i$, and the random error $r_{k(ij)}$ reflects the difference among subjects from cluster $j$ under treatment $i$. Thus, ranking clusters is equivalent to ranking based on values of $b_{j(i)}$, and ranking individuals is equivalent to ranking based on values of $r_{k(ij)}$. However, the values of $b_{j(i)}$ and $r_{k(ij)}$ are not directly observable. In real application we usually rank clusters or individuals based on some latent variables that are correlated with $b_{j(i)}$ and $r_{k(ij)}$, respectively.

### 2.3. Method of moments (MOM) estimators

Under model (2.1), we use the same notations as those in [90]. Denote the treatment effect by $\Delta \equiv \mu_1 - \mu_0 = a_1$, where $\mu_i = \mu + a_i$, the mean of responses in treatment group $i$. We denote the RSS-based estimator for $\sigma^2_b$ ($\sigma^2_r$) by $\hat{\sigma}^2_b$ ($\hat{\sigma}^2_r$) and the SRS-based estimator by $\tilde{\sigma}^2_b$ ($\tilde{\sigma}^2_r$). For ranking at the cluster level only and both levels, we define the index set $\mathcal{J}_i(h) = \{j : \text{cluster } j \text{ in treatment } i \text{ has rank } h\}$ for $h = 1, \ldots, H_i$ and $i = 0, 1$; denote the number of clusters in $\mathcal{J}_i(h)$ by $J_{ih} = m^{c}_{i}$; further let $\mu_{b.ih} \equiv E[b_{j(i)} \mid O_{j(i)} = h]$ and $\sigma^2_{b.ih} \equiv var[b_{j(i)} \mid O_{j(i)} = h]$ be the mean and variance, respectively, of the $h$th judgment order statistic of the cluster effect $b$. For ranking at the individual level only and ranking at both levels, we define the index set $\mathcal{K}_{j(i)}(h') = \{k : \text{ individual } k \text{ within cluster } j \text{ under treatment } i \text{ has rank } h'\}$, where $h' = 1, \ldots, H_{j(i)}$; further let $\mu_{r.ijh'} \equiv E[r_{k(ij)} \mid O_{k(ij)} = h']$ and $\sigma^2_{r.ijh'} \equiv var[r_{k(ij)} \mid O_{k(ij)} = h']$ be the mean and variance, respectively, of the $h'$th judgment order statistic of

the individual effect $r$. Denote the number of subjects in $\mathcal{K}_{j(i)}(h')$ by $K_{ijh} = m^{id}_{j(i)}$.

### 2.3.1. Ranking at the cluster level

#### *2.3.1.1. Estimating $\sigma_r^2$.*

For ranking at the cluster level only, ranking information does not contribute to estimating $\sigma_r^2$. Let $SSW^c \equiv \sum_{i=0}^{1}\sum_{h=1}^{H_i}\sum_{j\in\mathcal{J}_i(h)}\sum_{k=1}^{K_{j(i)}}\left(Y_{k(ij)} - \bar{Y}_{j(i)}\right)^2$, where $\bar{Y}_{j(i)} = \frac{1}{K_{j(i)}}\sum_{k=1}^{K_{j(i)}} Y_{k(ij)}$. Note that $\bar{Y}_{j(i)} = \mu + a_i + b_{j(i)} + \bar{r}_{j(i)}$, where $\bar{r}_{j(i)} = \frac{1}{K_{j(i)}}\sum_{k=1}^{K_{j(i)}} r_{k(ij)}$. So we have $SSW^c = \sum_{i=0}^{1}\sum_{h=1}^{H_i}\sum_{j\in\mathcal{J}_i(h)}\sum_{k=1}^{K_{j(i)}}\left(r_{k(ij)} - \bar{r}_{j(i)}\right)^2$and it follows that

$$E[SSW^c] = \sum_{i=0}^{1}\sum_{h=1}^{H_i}\sum_{j\in\mathcal{J}_i(h)}\left(K_{j(i)} - 1\right)\sigma_r^2 = \left(K_{\cdot\cdot} - J_{\cdot}\right)\sigma_r^2,$$

where $K_{\cdot\cdot}$ is the total number of subjects and $J_{\cdot} = J_0 + J_1$, the total number of clusters. Thus, an unbiased estimator for $\sigma_r^2$ can be written as

$$\hat{\sigma}_r^2 = \frac{SSW^c}{K_{\cdot\cdot} - J_{\cdot}}. \tag{2.2}$$

#### *2.3.1.2. Estimating $\sigma_b^2$.*

To estimate $\sigma_b^2$, we apply a relationship derived from equation (2) in [19]:

$$\sigma_{b.i}^2 = \frac{1}{H_i}\sum_{h=1}^{H_i}\left(\mu_{b.ih}^2 + \sigma_{b.ih}^2\right) \tag{2.3}$$

for $i = 0, 1$, where $\sigma_{b.i}^2$ is the variance of cluster effect $b$ in treatment group $i$. In order

to obtain an unbiased estimator for $\sigma^2_{b.i}$, we first derive unbiased estimators for $\sum_{h=1}^{H_i}\mu^2_{b.ih}$ and $\sum_{h=1}^{H_i}\sigma^2_{b.ih}$. Let $SSBR^c(i) \equiv \sum_{h=1}^{H_i}(\hat{\mu}_{ih}-\hat{\mu}_i)^2$, where $\hat{\mu}_{ih} = \sum_{j\in\mathcal{J}_i(h)}\bar{Y}_{j(i)}/J_{ih}$ and $\hat{\mu}_i = \sum_{h=1}^{H_i}\hat{\mu}_{ih}/H_i$. Then we have

$$\begin{aligned} SSBR^c(i) &= \sum_{h=1}^{H_i}\left[\left(\mu+a_i+\hat{b}_{ih}+\hat{r}_{ih}\right)-\left(\mu+a_i+\hat{b}_i+\hat{r}_i\right)\right]^2 \\ &= \sum_{h=1}^{H_i}\left[\left(\hat{b}_{ih}-\hat{b}_i\right)-(\hat{r}_{ih}-\hat{r}_i)\right]^2, \end{aligned}$$

where $\hat{b}_{ih} = \sum_{j\in\mathcal{J}_i(h)} b_{j(i)}/J_{ih}$, $\hat{r}_{ih} = \sum_{j\in\mathcal{J}_i(h)}\bar{r}_{j(i)}/J_{ih}$, $\hat{b}_i = \sum_{h=1}^{H_i}\hat{b}_{ih}/H_i$, and $\hat{r}_i = \sum_{h=1}^{H_i}\hat{r}_{ih}/H_i$. Let $SSB^c(i,h) \equiv \sum_{j\in\mathcal{J}_i(h)}(\bar{Y}_{j(i)}-\hat{\mu}_{ih})^2$. From the derivation of $\hat{V}\left(\hat{\Delta}_{RSS}\right)$ in Section 4 of [90], we have $E\left[\frac{SSB^c(i,h)}{(J_{ih}-1)J_{ih}}\right] = \frac{\sigma^2_{b.ih}}{J_{ih}} + \frac{\sum_{j\in\mathcal{J}_i(h)}\sigma^2_r}{J^2_{ih}}$. We can obtain the expected value of $SSBR^c(i)$ as

$$E\left[SSBR^c(i)\right] = \sum_{h=1}^{H_i}\mu^2_{b.ih} + (1-\frac{1}{H_i})\sum_{h=1}^{H_i}E\left[\frac{SSB^c(i,h)}{(J_{ih}-1)J_{ih}}\right],$$

and thus an unbiased estimator for $\sum_{h=1}^{H_i}\mu^2_{b.ih}$ is given by

$$SSBR^c(i) - (1-\frac{1}{H_i})\sum_{h=1}^{H_i}\frac{SSB^c(i,h)}{(J_{ih}-1)J_{ih}}. \tag{2.4}$$

Also, combining (2.4) with the unbiased estimator of $\sigma^2_r$ in (2.2), we have an unbiased estimator for each $\sigma^2_{b.ih}$ for $h = 1,\ldots,H_i$ and $i = 0,1$ as

$$\sigma^2_{b.ih} = \frac{SSB^c(i,h)}{J_{ih}-1} - \frac{1}{J_{ih}}\sum_{j\in\mathcal{J}_i(h)}\frac{1}{K_{j(i)}}\frac{SSW^c}{K_{..}-J_.}.$$

It then follows from (2.3) that an unbiased estimator for $\sigma^2_{b.i}$ can be expressed as

$$\hat{\sigma}^2_{b.i} = \frac{SSBR^c(i)}{H_i} + \frac{1}{H_i}\sum_{i=1}^{H_i}\left(1 - \frac{1}{J_{ih}} + \frac{1}{H_i J_{ih}}\right)\frac{SSB^c(i,h)}{J_{ih}-1} - \frac{1}{H_i}\sum_{h=1}^{H_i}\frac{1}{J_{ih}}\sum_{j\in\mathcal{J}_i(h)}\frac{1}{K_{j(i)}}\cdot\frac{SSW^c}{K_{..}-J_{.}}.$$

Finally, an unbiased estimator for $\sigma^2_b$ is given by $\hat{\sigma}^2_b = \frac{\hat{\sigma}^2_{b.0}+\hat{\sigma}^2_{b.1}}{2}$. We truncate negative estimates to 0, which makes the estimator no longer unbiased.

When the set sizes in the two treatment groups are equal, we can further improve our estimation by pooling information from both groups. When $H_i = H^c$, we can express (2.3) as $\sigma^2_b = \frac{1}{H^c}\sum_{h=1}^{H^c}(\mu^2_{b.h} + \sigma^2_{b.h})$, where $\mu^2_{b.h} = \frac{1}{2}\sum_{i=0}^{1}\mu^2_{b.ih}$ and $\sigma^2_{b.h} \equiv \sigma^2_{b,0h} \equiv \sigma^2_{b,1h}$. This implies that we no longer need to estimate the variances for the two treatment groups separately. Instead, we can directly estimate $\sigma^2_b$. The corresponding estimator for $\sigma^2_b$ can be obtained following the above derivations and is shown in Table 2.1. For a completely balanced design ($m_i = m^c$, $K_{j(i)} = K$), the estimator of $\sigma^2_b$ can be further simplified (see Table 2.1).

### 2.3.2. Ranking at the individual level

#### *2.3.2.1. Estimating $\sigma^2_b$.*

Let $SSW^{id}(i,j,h) \equiv \sum_{k\in\mathcal{K}_{j(i)}(h)}(Y_{k(ij)} - \bar{Y}_{ijh})^2$, where $\bar{Y}_{ijh} = \frac{1}{K_{ijh}}\sum_{k\in\mathcal{K}_{j(i)}(h)} Y_{k(ij)}$. Note that $\bar{Y}_{ijh} = \mu + a_i + b_{j(i)} + \bar{r}_{ijh}$, where $\bar{r}_{ijh} = \sum_{k\in\mathcal{K}_{j(i)}(h)} r_{k(ij)}/K_{ijh}$. Thus we have $SSW^{id}(i,j,h) = \sum_{k\in\mathcal{K}_{j(i)}(h)}\left[r_{k(ij)} - \bar{r}_{ijh}\right]^2$. It follows that

$$E\left[SSW^{id}(i,j,h)\right] = (K_{ijh}-1)\sigma^2_{r.ijh}.$$

Therefore, an unbiased estimator for $\sigma^2_{r.ijh}$ is given by

$$\hat{\sigma}^2_{r.ijh} = \frac{SSW^{id}(i,j,h)}{K_{ijh}-1}.$$

Let $SSB^{id}(i) \equiv \sum_{j=1}^{J_i}(\hat{\mu}_{j(i)} - \hat{\mu}_i)^2$, where $\hat{\mu}_{j(i)} = \frac{1}{H_{j(i)}}\sum_{h=1}^{H_{j(i)}} \bar{Y}_{ijh}$, and $\hat{\mu}_i = \frac{1}{J_i}\sum_{j=1}^{J_i}\hat{\mu}_{j(i)}$. As mentioned above, from the derivation of $\hat{V}\left(\hat{\Delta}_{RSS}\right)$ in Section 4 of [90], we have $E\left[\frac{SSB^{id}(i)}{J_i-1}\right] = \sigma^2_{b.i} + \frac{1}{J_i}\sum_{j=1}^{J_i}\frac{1}{H^2_{j(i)}}\sum_{h=1}^{H_{j(i)}}\frac{\sigma^2_{r.ijh}}{K_{ijh}}$. Then an unbiased estimator for $\sigma^2_{b.i}$ is given by

$$\hat{\sigma}^2_{b.i} = \frac{SSB^{id}(i)}{J_i-1} - \frac{1}{J_i}\sum_{j=1}^{J_i}\frac{1}{H^2_{j(i)}}\sum_{h=1}^{H_{j(i)}}\frac{\hat{\sigma}^2_{r.ijh}}{K_{ijh}},$$

and $\sigma^2_b$ can be estimated by $\frac{\hat{\sigma}^2_{b.0}+\hat{\sigma}^2_{b.1}}{2}$.

When the set sizes are equal in all clusters of both groups, i.e. $H_{j(i)} = H^{id}$, we can further improve the estimation of $\sigma^2_{r.ijh}$ by pooling information from all subjects within one ranking stratum where $O_{k(ij)} = h$ for $j = 1,\ldots,J_i$ and $i = 0,1$. Let $SSW^{id}(h) \equiv \sum_{i=0}^{1}\sum_{j=1}^{J_i}\sum_{k\in\mathcal{K}_{j(i)}(h)}\left[Y_{k(ij)} - \bar{Y}_{ijh}\right]^2 = \sum_{i=0}^{1}\sum_{j=1}^{J_i} SSW^{id}(i,j,h)$. Then we have $E\left[SSW^{id}(h)\right] = \sum_{i=0}^{1}\sum_{j=1}^{J_i}(K_{ijh}-1)\sigma^2_{r.h}$, where $\sigma^2_{r.h}$ is the variance of the $h$th judgement order statistic of the individual effect $r$. Then an unbiased estimator of $\sigma^2_{r.h}$ is given by

$$\hat{\sigma}^2_{r.h} = \frac{SSW^{id}(h)}{K_{..h}-J.},$$

where $K_{..h}$ is the total number of subjects in the $h$th ranking stratum and $J. = J_0 + J_1$. An improved estimator for $\sigma^2_b$ based on $\hat{\sigma}^2_{r.h}$ is shown in Table 2.1 along with the simplified estimator for $\sigma^2_b$ when the design is completely balanced.

*2.3.2.2. Estimating $\sigma_r^2$.*

To estimate $\sigma_r^2$ when the data are obtained using RSS at the individual level only, we first use a similar relationship as (2.3) to estimate $\sigma_{r.ij}^2$:

$$\sigma_{r.ij}^2 = \frac{1}{H_{j(i)}} \sum_{h=1}^{H_{j(i)}} \left(\mu_{r.ijh}^2 + \sigma_{r.ijh}^2\right) \tag{2.5}$$

for $j = 1, \ldots, J_i$ and $i = 0, 1$. Let $SSBR^{id}(i,j) \equiv \sum_{h=1}^{H_{j(i)}} (\bar{Y}_{ijh} - \hat{\mu}_{j(i)})^2$ and an unbiased estimator for $\sum_{h=1}^{H_{j(i)}} \mu_{r.ijh}^2$ is given by

$$SSBR^{id}(i,j) - \left(1 - \frac{1}{H_{j(i)}}\right) \sum_{h=1}^{H_{j(i)}} \frac{SSW^{id}(i,j,h)}{(K_{ijh}-1)K_{ijh}}.$$

Thus an unbiased estimator for $\sigma_{r.ij}^2$ is given by

$$\hat{\sigma}_{r.ij}^2 \equiv \frac{SSBR^{id}(i,j)}{H_{j(i)}} + \frac{1}{H_{j(i)}} \sum_{h=1}^{H_{j(i)}} \left(1 - \frac{1}{K_{ijh}} + \frac{1}{H_{j(i)}K_{ijh}}\right) \frac{SSW^{id}(i,j,h)}{K_{ijh}-1}.$$

Further, $\sigma_r^2$ can be estimated by averaging $\hat{\sigma}_{r.ij}^2$ over all $i$ and $j$ as $\hat{\sigma}_r^2 = \frac{1}{J.} \sum_{i=0}^{1} \sum_{j=1}^{J_i} \hat{\sigma}_{r.ij}^2$. As mentioned above, when $H_{j(i)} = H^{id}$, we can improve our estimation by pooling information from clusters within the same ranking stratum in both treatment groups.

### 2.3.3. Ranking at both levels

#### *2.3.3.1. Estimating $\sigma_b^2$.*

Let $SSW^B(i,j,h) \equiv \sum_{k\in\mathcal{K}_{j(i)}(h)}(Y_{k(ij)} - \bar{Y}_{ijh})^2$ and we have $E\left[SSW^B(i,j,h)\right] = (K_{ijh} - 1)\sigma^2_{r.ijh}$ as the case for ranking at the individual level only. Let $SSBR1(i) \equiv \sum_{h=1}^{H_i}(\hat{\mu}_{ih} - \hat{\mu}_i)^2$, which has the same expression as $SSBR^c(i)$. Similarly, we can have the expected value of $SSBR1(i)$ as

$$E[SSBR1(i)] = \sum_{h=1}^{H_i}\mu^2_{b.ih} + \left(1 - \frac{1}{H_i}\right)\sum_{h=1}^{H_i}\left[\frac{1}{J_{ih}}\sigma^2_{b.ih} + \frac{1}{J^2_{ih}}\sum_{j\in\mathcal{J}_i(h)}\frac{1}{H_i^2}\sum_{h'=1}^{H_{j(i)}}\frac{\sigma^2_{r.ijh'}}{K_{ijh'}}\right].$$

Let $SSB^B(i,h) \equiv \sum_{j\in\mathcal{J}_i(h)}\left(\hat{\mu}_{j(i)} - \hat{\mu}_{ih}\right)^2$and following the derivation in [90] we have $E\left[\frac{SSB^B(i,h)}{(J_{ih}-1)J_{ih}}\right] = \frac{1}{J_{ih}}\sigma^2_{b.ih} + \frac{1}{J^2_{ih}}\sum_{j\in\mathcal{J}_i(h)}\frac{1}{H^2_{j(i)}}\sum_{h'=1}^{H_{j(i)}}\frac{\sigma^2_{r.ijh'}}{K_{ijh'}}$. We can further obtain an unbiased estimator of $\sum_{h=1}^{H_i}\mu^2_{b.ih}$ as

$$SSBR1(i) - \left(1 - \frac{1}{H_i}\right)\sum_{h=1}^{H_i}\frac{SSB^B(i,h)}{(J_{ih}-1)J_{ih}}$$

and an unbiased estimator for $\sigma^2_{b.ih}$ as

$$\frac{SSB^B(i,h)}{J_{ih}-1} - \frac{1}{J_{ih}}\sum_{j\in\mathcal{J}_i(h)}\frac{1}{H^2_{j(i)}}\sum_{h'=1}^{H_{j(i)}}\frac{SSW^B(i,j,h)}{(K_{ijh'}-1)K_{ijh'}}.$$

With the relationship in (2.3), we can get estimate $\sigma^2_{b.i}$ as

$$\hat{\sigma}^2_{b.i} = \frac{SSBR1^B(i)}{H_i} + \frac{1}{H_i}\sum_{i=1}^{H_i}\left(1 - \frac{1}{J_{ih}} + \frac{1}{H_iJ_{ih}}\right)\frac{SSB^B(i,h)}{J_{ih}-1} - \frac{1}{H_i}\sum_{h=1}^{H_i}\frac{1}{J_{ih}}\sum_{j\in\mathcal{J}_i(h)}\sum_{h'=1}^{H_{j(i)}}\frac{SSW^B(i,j,h)}{(K_{ijh'}-1)K_{ijh'}}$$

When the design becomes balanced in both levels, an improved estimator for $\sigma^2_b$ can be

achieved by applying the same information-pooling technique specified above for both levels. This improved estimator can be further simplified in completely balanced designs and all are shown in Table 2.1.

*2.3.3.2. Estimating $\sigma_r^2$.*

Let $SSBR2(i,j) \equiv \sum_{h=1}^{H_{j(i)}} \left[\bar{Y}_{ijh} - \hat{\mu}_{j(i)}\right]^2$, which has the same expression as $SSBR^{id}(i,j)$. Similarly, we can have the expectation of $SSBR2(i,j)$ expressed as

$$E[SSBR2(i,j)] = \sum_{h=1}^{H_{j(i)}} \mu_{r.ijh}^2 + \left(1 - \frac{1}{H_{j(i)}}\right) \sum_{h=1}^{H_{j(i)}} \frac{\sigma_{r.ijh}^2}{K_{ijh}}.$$

Thus, an unbiased estimator for $\sigma_{r.ij}^2$ can be written as

$$\hat{\sigma}_{r.ij}^2 = \frac{SSBR2(i,j)}{H_{j(i)}} + \frac{1}{H_{j(i)}} \left(1 - \frac{1}{K_{ijh}} + \frac{1}{H_{j(i)} K_{ijh}}\right) \frac{SSW^B(i,j,h)}{K_{ijh} - 1}$$

and $\sigma_r^2$ can be estimated by $\frac{\sum_{i=0}^{1} \sum_{j=1}^{J_i} \hat{\sigma}_{r.ij}^2}{J_0 + J_1}$.

| | Design | Estimator |
|---|---|---|
| Ranking at cluster level | General | $\frac{1}{2}\sum_{i=0}^{1}\left[\frac{SSBR^c(i)}{H_i}+\frac{1}{H_i}\sum_{h=1}^{H_i}(1-\frac{1}{J_{ih}}+\frac{1}{H_iJ_{ih}})\frac{SSB^c(i,h)}{J_{ih}-1}-\frac{1}{H_i}\sum_{h=1}^{H_i}\frac{1}{J_{ih}}\sum_{j\in\mathcal{J}_i(h)}\frac{1}{K_{j(i)}}\frac{SSW^c}{K_{..}-J_.}\right]$ |
| | $H_i=H^c$ | $\frac{1}{2H^c}\sum_{i=0}^{1}SSBR^c(i)+\frac{1}{H^c}\sum_{i=0}^{1}\sum_{h=1}^{H^c}\left[\frac{1}{J_.-2}-\frac{1-H^{-c}}{2J_{ih}(J_{ih}-1)}\right]SSB^c(i,h)-\frac{1}{H^c}\sum_{i=0}^{1}\sum_{h=1}^{H^c}\frac{1-J_{ih}^{-1}}{J_.-2}\sum_{j\in\mathcal{J}_i(h)}\frac{1}{K_{j(i)}}\frac{SSW^c}{K_{..}-J_.}$ |
| | $H_i=H^c$, $J_{ih}=m^c$ | $\frac{1}{2H^c}\sum_{i=0}^{1}SSBR^c(i)+\frac{1}{2H^c(m^c-1)}\left[1-\frac{1}{m^c}+\frac{1}{H^cm^c}\right]\sum_{i=0}^{1}\sum_{h=1}^{H^c}SSB^c(i,h)-\frac{1}{2m^cH^c}\sum_{i=0}^{1}\sum_{h=1}^{H^c}\sum_{j\in\mathcal{J}_i(h)}\frac{1}{K_{j(i)}}\frac{SSW^c}{K_{..}-J_.}$ |
| | $H_i=H^c$, $J_{ih}=m^c$, $K_{j(i)}=K$ | $\frac{\sum_{i=0}^{1}SSBR^c(i)}{2H^c}+\left(1-\frac{1}{m^c}+\frac{1}{H^cm^c}\right)\frac{\sum_{i=0}^{1}SSB^c(i,h)}{2H^c(m^c-1)}-\frac{SSW^c}{2H^cm^cK(K-1)}$ |
| Ranking at individual level | General | $\frac{1}{2}\sum_{i=0}^{1}\left[\frac{SSB^I(i)}{J_i-1}-\frac{1}{J_i}\sum_{j=1}^{J_i}\frac{1}{H_{j(i)}^2}\sum_{h=1}^{H_i}\frac{SSW^I(i,j,h)}{K_{ijh}(K_{ijh}-1)}\right]$ |
| | $H_{j(i)}=H^{id}$ | $\frac{1}{J_.-2}\sum_{i=0}^{1}\left[SSB^I(i)-(1-\frac{1}{J_i})\frac{1}{(H^{id})^2}\sum_{j=1}^{J_i}\sum_{h=1}^{H^{id}}\frac{SSW^{id}(h)}{K_{\cdot h}-J_.}\right]$ |
| | $H_{j(i)}=H^{id}$, $K_{ijh}=m^{id}$ | $\frac{1}{J_.-2}\sum_{i=0}^{1}SSB^{id}(i)-\frac{1}{J_.(H^{id})^2m^{id}(m^{id}-1)}\sum_{h=1}^{H^{id}}SSW^{id}(h)$ |
| Ranking at both levels | General | $\frac{1}{2}\sum_{i=0}^{1}\left[\frac{SSBR1(i)}{H_i}+\frac{1}{H_i}\sum_{h=1}^{H_i}(1-\frac{1}{J_{ih}}+\frac{1}{H_iJ_{ih}})\frac{SSB^B(i,h)}{J_{ih}-1}-\frac{1}{H_i}\sum_{h=1}^{H_i}\frac{1}{J_{ih}}\sum_{j\in\mathcal{J}_i(h)}\sum_{h'=1}^{H_{j(i)}}\frac{SSW^B(i,j,h')}{K_{ijh'}(K_{ijh'}-1)}\right]$ |
| | $H_i=H^c$, $H_{j(i)}=H^{id}$ | $\frac{1}{2H^c}\sum_{i=0}^{1}\left[SSBR1(i)-\left(1-\frac{1}{H^c}\right)\sum_{h=1}^{H^c}\frac{1}{J_{ih}}\hat{\sigma}^2_{b.h}-\left(1-\frac{1}{H^c}\right)\sum_{h=1}^{H^c}\frac{1}{J_{ih}^2}\sum_{j\in\mathcal{J}_i(h)}\frac{1}{(H^{id})^2}\sum_{h'=1}^{H^{id}}\frac{\hat{\sigma}^2_{r.h'}}{K_{ijh'}}+\frac{1}{H^c}\sum_{h=1}^{H^c}\hat{\sigma}^2_{b.h}\right]$ |
| | $H_i=H^c$, $m_i=m^c$, $H_{j(i)}=H^{id}$, $m_{j(i)}=m^{id}$ | $\frac{\sum_{i=0}^{1}SSBR1(i)}{2H^c}+\frac{\sum_{i=0}^{1}\sum_{h=1}^{H^c}(i,h)}{2H^c(m^c-1)}\left(1-\frac{1}{m^c}+\frac{1}{H^cm^c}\right)-\frac{\sum_{i=0}^{1}\sum_{j=1}^{J_i}\sum_{h=1}^{H^{id}}SSW^B(i,j,h)}{2H^cm^c(H^{id})^2m^{id}(m^{id}-1)}$ |

Table 2.1: Method of moment estimators for $\sigma_b^2$

| | Design | Estimator |
|---|---|---|
| Ranking at cluster level | General | $\frac{SSW^c}{K_{..}-J_.}$ |
| Ranking at individual level | General | $\frac{1}{J_.}\sum_{i=0}^{1}\sum_{j=1}^{J_i}\frac{1}{H_{j(i)}}\left[SSBR^{id}(i,j)+\sum_{h=1}^{H_{j(i)}}\left(1-\frac{1}{K_{ijh}}+\frac{1}{H_{j(i)}K_{ijh}}\right)\frac{SSW^{id}(i,j,h)}{K_{ijh}-1}\right]$ |
| | $H_{j(i)}=H^{id}$ | $\frac{1}{H^{id}J_.}\sum_{i=0}^{1}\sum_{j=1}^{J_i}SSBR^{id}(i,j)+\frac{1}{H^{id}}\sum_{h=1}^{H^{id}}\left[1-\frac{1}{J_.}(1-\frac{1}{H^{id}})\sum_{i=0}^{1}\sum_{j=1}^{J_i}\frac{1}{K_{ijh}}\right]\frac{SSW^{id}(h)}{K_{\cdot h}-J_.}$ |
| | $H_{j(i)}=H^{id}$, $K_{ijh}=m^{id}$ | $\frac{1}{J_.H^{id}}\sum_{i=0}^{1}\sum_{j=1}^{J_i}SSBR^{id}(i,j)+\frac{\sum_{h=1}^{H^{id}}SSW^{id}(h)}{J_.H^{id}(m^{id}-1)}\left(1-\frac{1}{m^{id}}+\frac{1}{H^{id}m^{id}}\right)$ |
| Ranking at both levels | General | $\frac{SSBR2(i,j)}{H_{j(i)}}+\frac{1}{H_{j(i)}}\sum_{h=1}^{H_{j(i)}}\left(1-\frac{1}{K_{ijh}}+\frac{1}{H_{j(i)K_{ijh}}}\right)\frac{SSW^B(i,j,h)}{K_{ijh}-1}$ |
| | $H_i=H^c$, $H_{j(i)}=H^{id}$ | $\frac{1}{J_.H^{id}}\left[\sum_{i=0}^{1}\sum_{j=1}^{J_i}SSBR2(i,j)-\left(1-\frac{1}{H^{id}}\right)\sum_{i=0}^{1}\sum_{j=1}^{J_i}\sum_{h=1}^{H^{id}}\frac{\hat{\sigma}^2_{r.h}}{K_{ijh}}\right]+\frac{1}{H^{id}}\sum_{h=1}^{H^{id}}\hat{\sigma}^2_{r.h}$ |
| | $H_i=H^c$, $m_i=m^c$, $H_{j(i)}=H^{id}$, $m_{j(i)}=m^{id}$ | $\frac{1}{2H^cH^{id}m^c}\sum_{i=0}^{1}\sum_{j=1}^{J_i}SSBR2(i,j)+\frac{\sum_{i=0}^{1}\sum_{j=1}^{J_i}\sum_{h=1}^{H^{id}}SSW^B(i,j,h)}{2H^cm^cH^{id}(m^{id}-1)}\left(1-\frac{1}{m^{id}}+\frac{1}{H^{id}m^{id}}\right)$ |

Table 2.2: Method of moment estimators for $\sigma_r^2$

## 2.4. Impact of design parameters and ranking schemes

We examined the impact of design parameters on the estimation of $\sigma_b$ and $\sigma_r$ based on the relative efficiency (RE), defined as the ratio of the Mean Squared Error (MSE) of RSS versus that of SRS. Unlike the estimation of the treatment effect, there is no closed form formula for the RE. To estimate the RE, we conducted simulations under two ranking schemes – ranking at the cluster level only and ranking at the individual level only – with completely balanced designs. For ranking at the cluster level only, a completely balanced design means $H_i \equiv H^c$, $m_i \equiv m^c$ and $K_{j(i)} \equiv K$, where $(H^c, m^c, K)$ are the design parameters, for all $i$ and $j$. For ranking at the individual level only, that means $J_i \equiv J$, $H_{j(i)} \equiv H^{id}$ and $m_{j(i)} \equiv m^{id}$, where $(H^{id}, m^{id}, J)$ are the design parameters, for all $i$ and $j$. In our simulations, we assumed $\mu = 0$ and $\Delta = 0$ since they do not affect the RE of the estimation of the variance components. We used the same settings as used by [90], namely $\sigma_b^2 = 1$ and $\sigma_r^2 = 4$, with ICC=0.2, a typical value for many applications in educational studies. We examined the performance of the proposed MOM estimators with different underlying distributions, including the Normal, Uniform and Lognormal, of the cluster and individual effects. For each combination of the design parameters we simulated 50,000 samples.

### 2.4.1. Estimating $\sigma_b$

#### *2.4.1.1. Perfect ranking*

When estimating $\sigma_b$, first we assumed the ranking was perfect at both the cluster and individual level. Imperfect ranking will be discussed later.

**Ranking at the cluster level only.** Considering completely balanced CRDs with design parameters $(H^c, m^c, K)$, we performed simulations for ranking at the cluster level only, in which the RE of estimating $\sigma_b$ was denoted by $\text{RE}^c$.

The top panel of Figure 2.1 displays the estimated $\text{RE}^c$ for various values of $(H^c, K)$ when $m^c$ is fixed at 6. The five numbered lines in every subplot correspond to $H^c = 2, 3, 4, 6, 8$, respectively. Figure 2.1(a) shows that the $\text{RE}^c$ of all situations are greater than one, suggesting that RSS is more efficient than SRS, although the gain in efficiency largely depends on the distributions, where the Uniform has the largest estimated $\text{RE}^c$, followed by the Normal distribution, whereas the Lognormal only has marginal improvement in the efficiency. In addition, the $\text{RE}^c$ increases as $H^c$ increases when other parameters are equal for all three distributions, although the increments in the Lognormal case are rather small. When fixing $H^c$ and $m^c$, the $\text{RE}^c$ seems to increase as $K$ increases, at least for the Uniform and Normal cases, but the upward trend is visible only when $H^c$ is large. For the Lognormal distribution, an increasing trend is barely noticeable, if at all. The bottom panel of Figure 2.1 shows the estimated $\text{RE}^c$ of different values of $(H^c, m^c)$ when $K$ is 30. Unlike the case of estimating the treatment effect when $m^c$ has no effect on $RE^c$, for the Uniform and Normal distributions, the estimated $\text{RE}^c$ decreases as $m^c$ increases with the remaining parameters fixed. Again, in the Lognormal case, a similar trend, if any, is barely visible. Finally we held $H^c$ constant at 2 and 8, and examined the influence of $(m^c, K)$. The two panels in Figure 2.2 display the estimated $RE^c$ are above one for various values of $(m^c, K)$ with $H^c$ fixed at 2 and 8 for all three distributions. The Lognormal distribution, however, shows no conclusive patterns in any setting. As a result, in the remaining part of this paragraph, all observations we made are limited to the Normal and Uniform cases. It is noticeable that the $\text{RE}^c$, when $H^c$ and $K$ are unchanged, becomes smaller as $m^c$ increases. Figure 2.2(b), when $H^c = 8$ and $m^c$ is held constant, displays a notable increasing trend in the $\text{RE}^c$ as $K$ increases. A similar pattern is hardly observable in Figure 2.2(a) when $H^c = 2$.

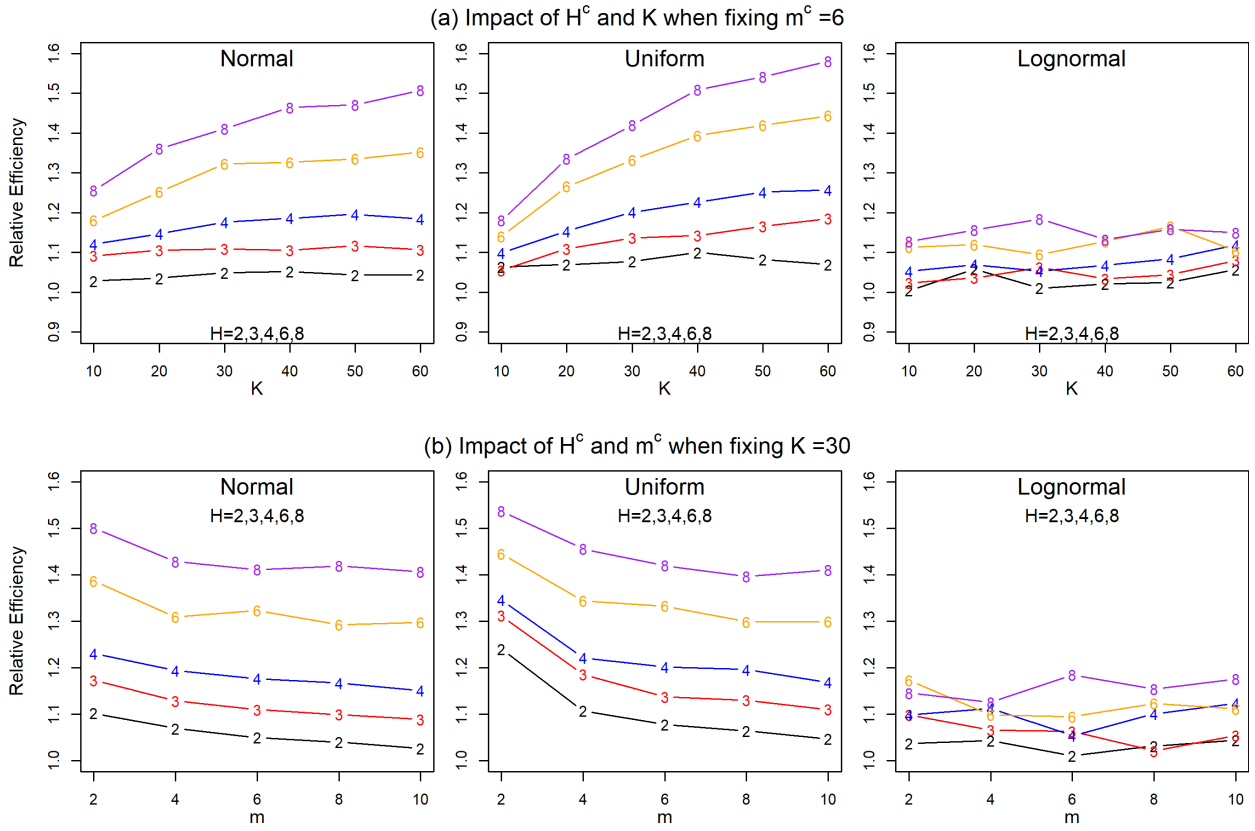



Figure 2.1: The impact of $(H^c, K)$ and $(H^c, m^c)$ on relative efficiencies of $\hat{\sigma}_b$ under completely balanced CRDs, with perfect ranking at the cluster level only

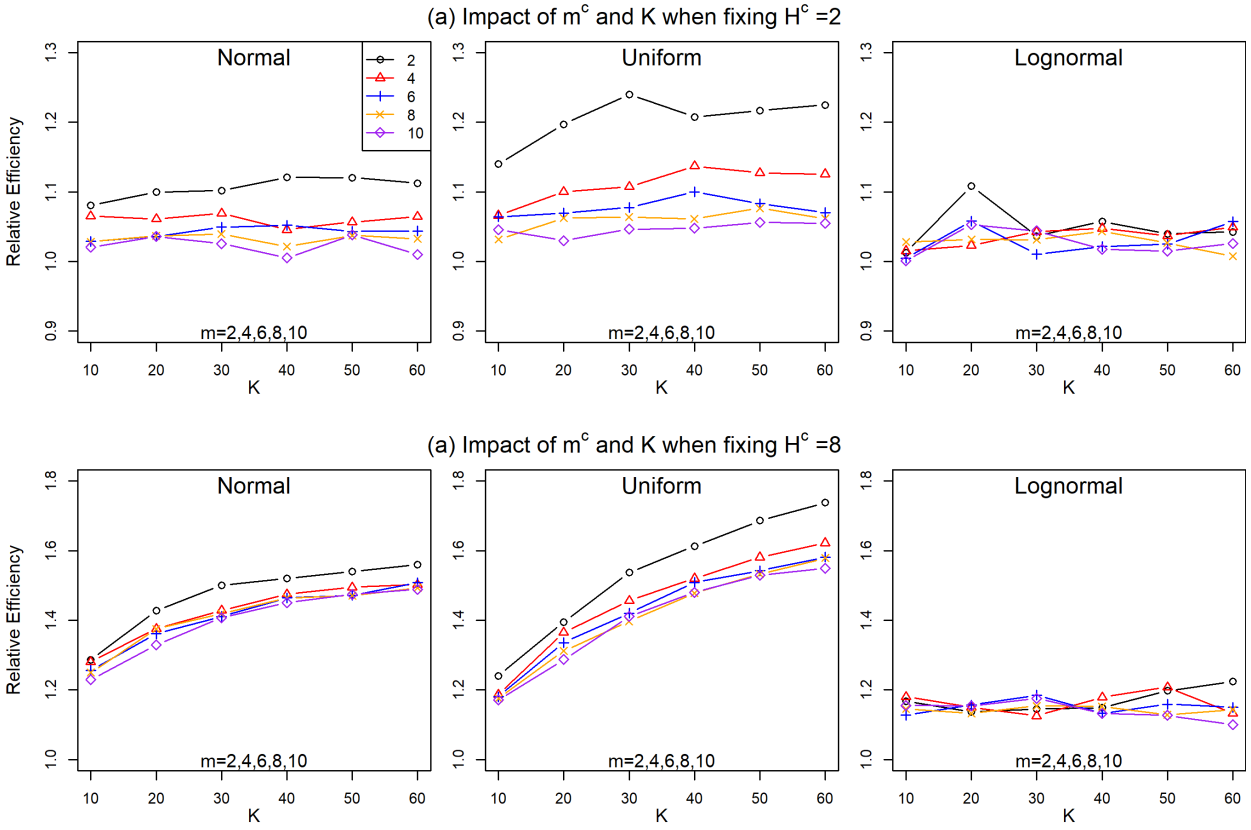



Figure 2.2: The impact of $(m^c, K)$ on relative efficiencies of $\hat{\sigma}_b$ under completely balanced CRDs, with perfect ranking at the cluster level only

**Ranking at the individual level only.** For ranking at the individual level only, we conducted simulations to estimate the RE of $\hat{\sigma}_b$, denoted by $RE^{id}$, in completely balanced CRDs with design parameters $(H^{id}, m^{id}, J)$.

Figure 2.3 displays the estimated values of $RE^{id}$ with respect to various values of $(H^{id}, m^{id}, J)$. The five numbered lines in all subplots correspond to $H^{id} = 2, 3, 4, 6, 8$, respectively. First we note that all estimated values of $RE^{id}$ are greater than one, suggesting that RSS is more efficient than SRS in estimating $\sigma_b$ when ranking is performed at the individual level only. We can see from Figure 2.3(a) that, when $J = 20$, for each $H^{id}$, the estimated $\text{RE}^{id}$ decreases as $m^{id}$ increases (or equivalently as $K$ increases). This trend is clear for the Normal and

Uniform but not so for the Lognormal distribution. The influence of $H^{id}$ on $\text{RE}^{id}$ does not have a clear pattern with fixed $m^{id}$, especially when $m^{id}$ is large. This is probably because when $m^{id}$ is sufficiently large, so is the sample size, making $\hat{\sigma}_b$ almost as efficient as $\tilde{\sigma}_b$. For this reason we fixed $m^{id}$ fixed at 2 and plotted the estimated $\text{RE}^{id}$ against values of $(H^{id}, J)$ in Panel (b) of Figure 2.3. For the Normal and Uniform distributions , when $H^{id} > 2$, $\text{RE}^{id}$ seems to decrease as $H^{id}$ increases for fixed $J$. For the Lognormal distribution, at any fixed $J$, the estimated $\text{RE}^{id}$s are similar for different set sizes except $H^{id} = 2$, which has smaller $\text{RE}^{id}$ than the others. Figures 2.4(a) and 2.4(b) plot the estimated $\text{RE}^{id}$ versus different values of $(m^{id}, J)$ when holding $H^{id}$ at 2 and 8, respectively. From both panels we can see that, for any value of $J$, the $\text{RE}^{id}$ decreases as $m^{id}$ increases. The impact of $J$, however, when fixing $m^{id}$, seems to be negligible as the curves in Figure 2.4 (b) overlap largely with each other.

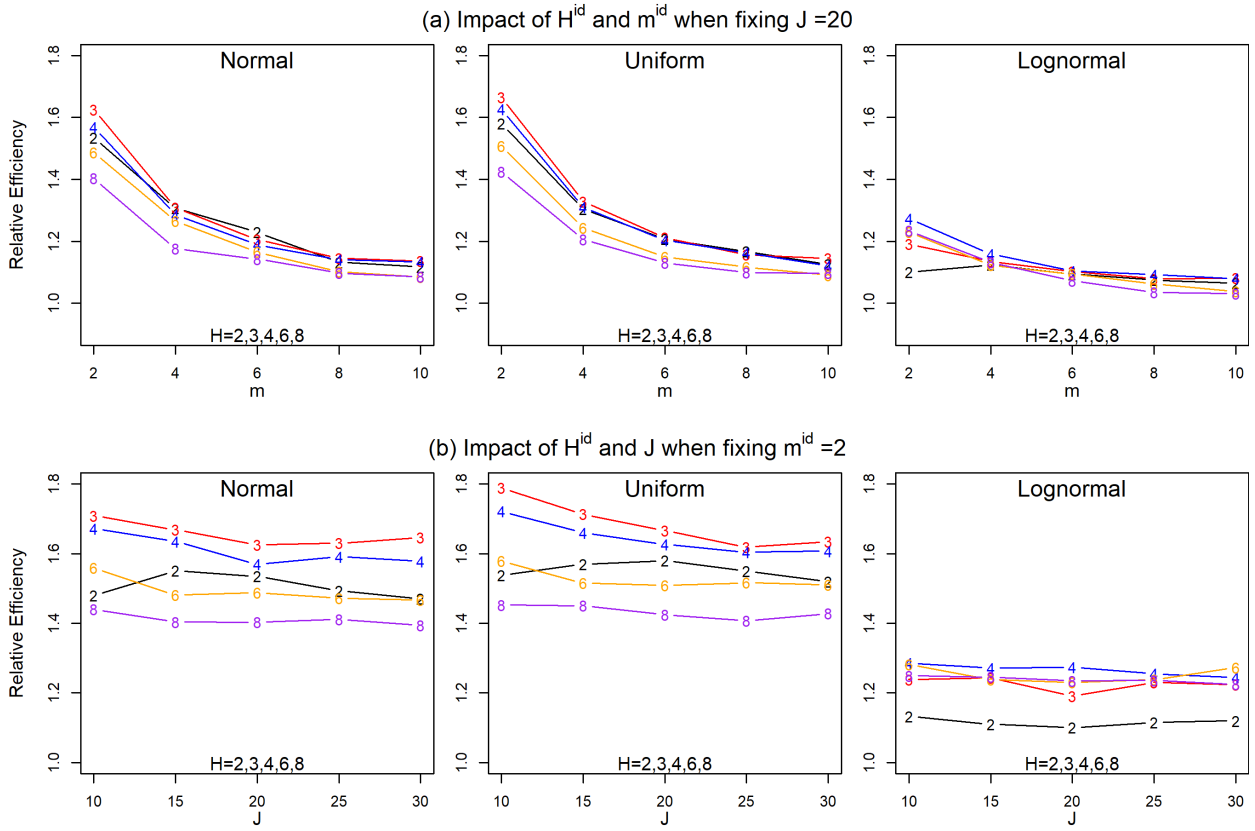



Figure 2.3: The impact of $(H^{id}, m^{id})$ and $(H^{id}, J)$ on relative efficiencies of $\hat{\sigma}_b$ under completely balanced CRDs, when ranking is conducted at the individual level only and is perfect.

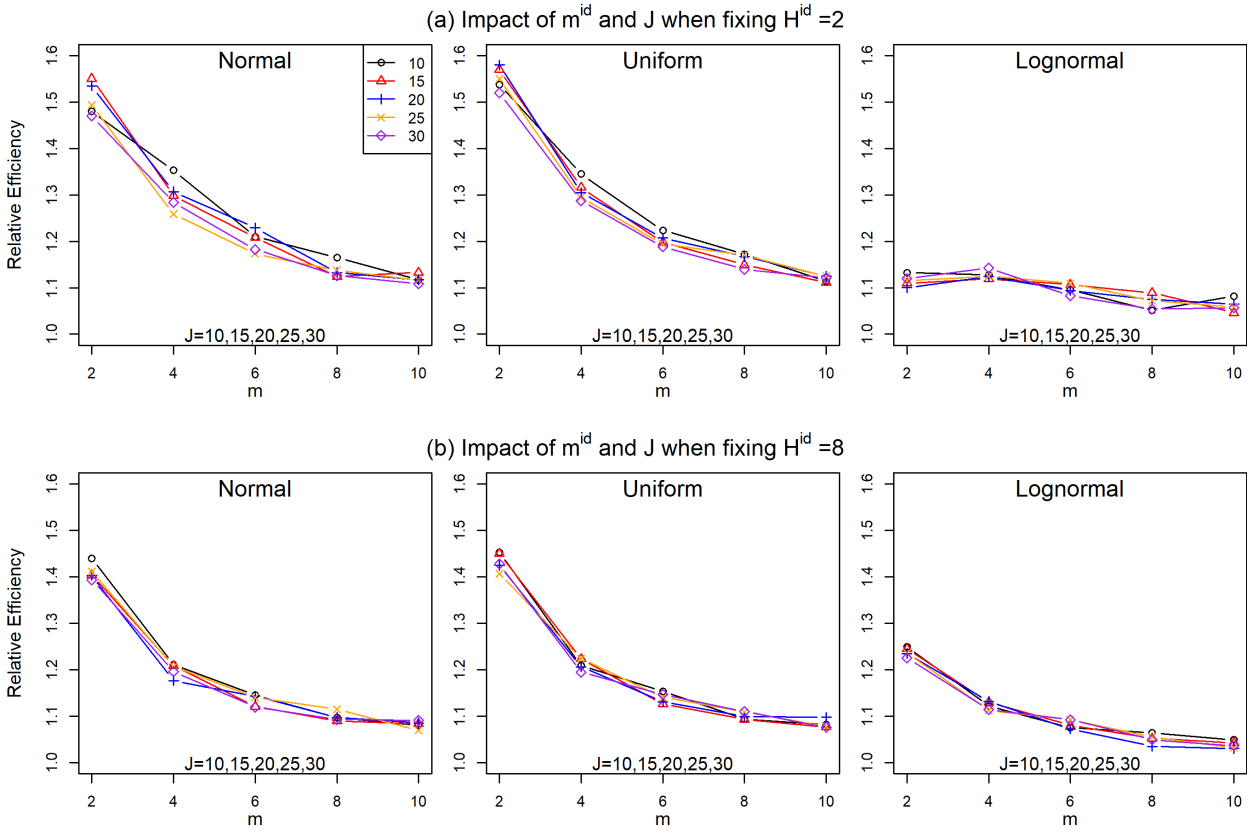



Figure 2.4: The impact of $(m^{id}, J)$ on relative efficiencies of $\hat{\sigma}_b$ under completely balanced CRDs, when ranking is conducted at the individual level only and is perfect.

### *2.4.1.2. Imperfect ranking*

When ranking is imperfect, we used a linear ranking error model as that in [90]. For cluster-level ranking only, ranking is performed via a latent variable $X$, which is related to $b$ through $b = X + \epsilon_x$, where $\epsilon_x$ is the error term, independent of $X$, with mean 0 and variance $\sigma^2_{\epsilon_x}$. Denote the correlation between $b$ and $X$ by $\rho^c$ and we have $\rho^c = \sqrt{\sigma^2_b - \sigma^2_{\epsilon_x}}/\sigma_b$. Similarly, for ranking at the individual level only, ranking is performed via an individual-level latent variable $Z$ and $r = Z + \epsilon_z$, where $\epsilon_z$ is the error term, independent of $Z$, with mean 0 and

variance $\sigma^2_{\epsilon_z}$. Denote the correlation between $r$ and $Z$ by $\rho^{id}$ and we have $\rho^{id} = \sqrt{\sigma^2_r - \sigma^2_{\epsilon_z}}/\sigma_r$. We plotted the estimated RE values against $\rho$ in Figure 2.5 for the three distributions under completely balanced CRDs. The left panel shows RE for ranking at the cluster level only, with the parameters $H^c = 4$, $m^c = 3$, $K = 20$, $\sigma^2_b = 1$, and $\sigma^2_r = 4$, whilst the right panel shows RE for ranking at the individual level only, with the parameters $H^{id} = 4$, $m^{id} = 5$, $J = 12$, $\sigma^2_b = 1$, and $\sigma^2_r = 4$. We can observe that in general RE increases when the ranking quality increases, and this trend becomes more obvious after $\rho > 0.5$. In terms of the three distributions, the improvement follows the order: Uniform>Normal>Lognormal when $\rho > 0.6$. When $\rho < 0.6$, all three distributions seem to have similar RE for ranking at the cluster level only; and for ranking at the individual level only, the Lognormal distribution has the least improvement in RE. .

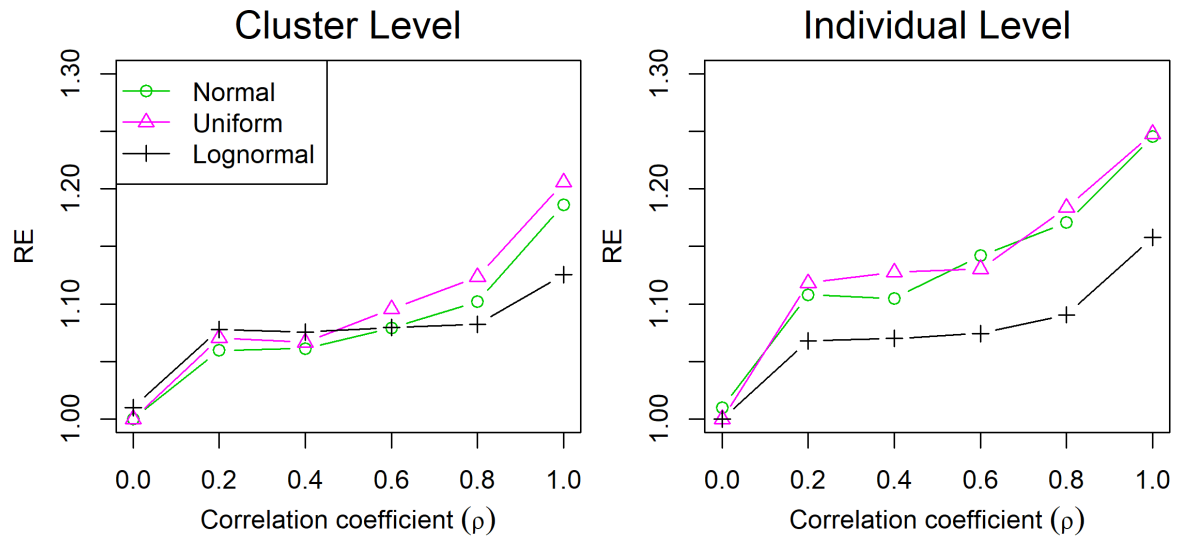



Figure 2.5: Relative efficiency of $\hat{\sigma}_b$ versus $\rho$ under completely balanced CRDs for three distributions of the ranking variables, including $N(0,1)$, $U(-1.74, 1.74)$, $\text{LN}(0, 0.481) - 1.27$ for ranking at the cluster level and $N(0, 2^2)$, $U(-3.47, 3.47)$, $\text{LN}(0, 0.941) - 1.6$ for ranking at the individual level.

### 2.4.2. Estimating $\sigma_r$

#### *2.4.2.1. Ranking at the cluster level only.*

As outlined in Section 2.3.1, ranking at the cluster level does not affect the estimation of $\sigma_r$, as shown in Figure 2.6 where all estimated RE values are centered around 1 in all the settings for all three distributions.

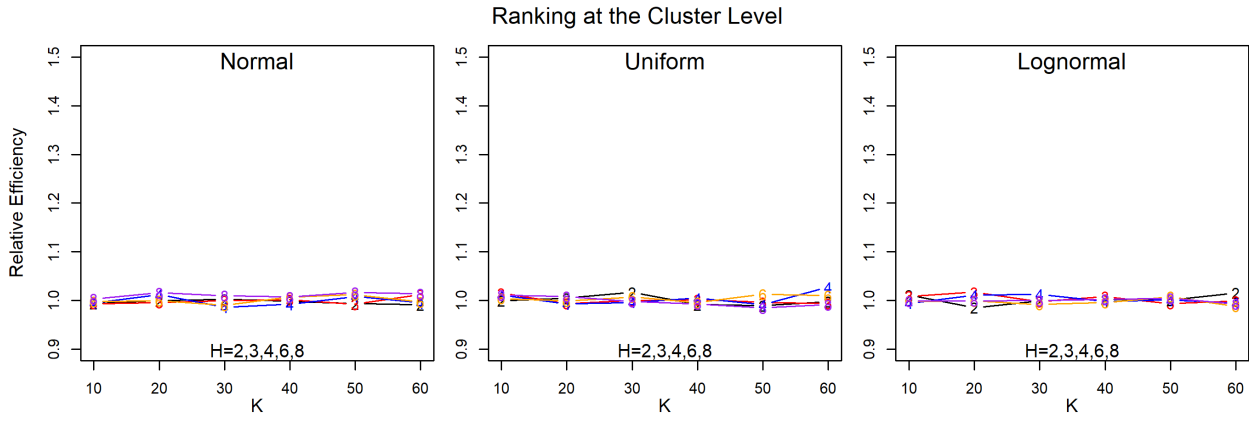



Figure 2.6: Relative efficiencies of $\hat{\sigma}_r$ versus $\tilde{\sigma}_r$ under completely balanced CRDs, with perfect ranking at the cluster level only

#### *2.4.2.2. Ranking at the individual level only.*

Figure 2.7 and Figure 2.8 display estimated $\text{RE}^{id}$ of $\hat{\sigma}_r$ with various settings of the design parameters when ranking is at the individual level only. The figures show that $\hat{\sigma}_r$ is notably more efficient than $\tilde{\sigma}_r$ for the Normal and Uniform distributions but not for the Lognormal distribution, whose estimated $\text{RE}^{id}$ is marginally greater than one. From Figure 2.7(a) we can observe that for normally and uniformly distributed individual effect $r$, $\text{RE}^{id}$ increases

as $H^{id}$ increases when $m^{id}$ and $J$ are held constant. There is little impact of $m^{id}$ on $\text{RE}^{id}$ for the Normal distribution while the $RE^{id}$ decreases as $m^{id}$ increases for uniform distribution when fixing other parameters. Panel $(b)$ in Figure 2.7 shows that the number of clusters $J$ does not seem to influence $\text{RE}^{id}$. Similar to the case of estimating $\sigma_b$, Figure 2.8 shows that the $RE^{id}$ decreases as $m^{id}$ increase for fixed $J$, while the influence of $J$ on $RE^{id}$ is negligible for all values of $m^{id}$. In addition, we also examined the impact of imperfect ranking on estimating $\sigma_r$. Figure 2.9 shows that RE increases as $\rho$ increases for both Normal and Uniform distributions. The influence of $\rho$ for the Lognormal distribution does not show a noticeable pattern since RSS seems to have little improvement in estimating $\sigma_r$ over SRS in all the settings. Among all three distributions, the Uniform has the largest RE while the Lognormal has the smallest RE.

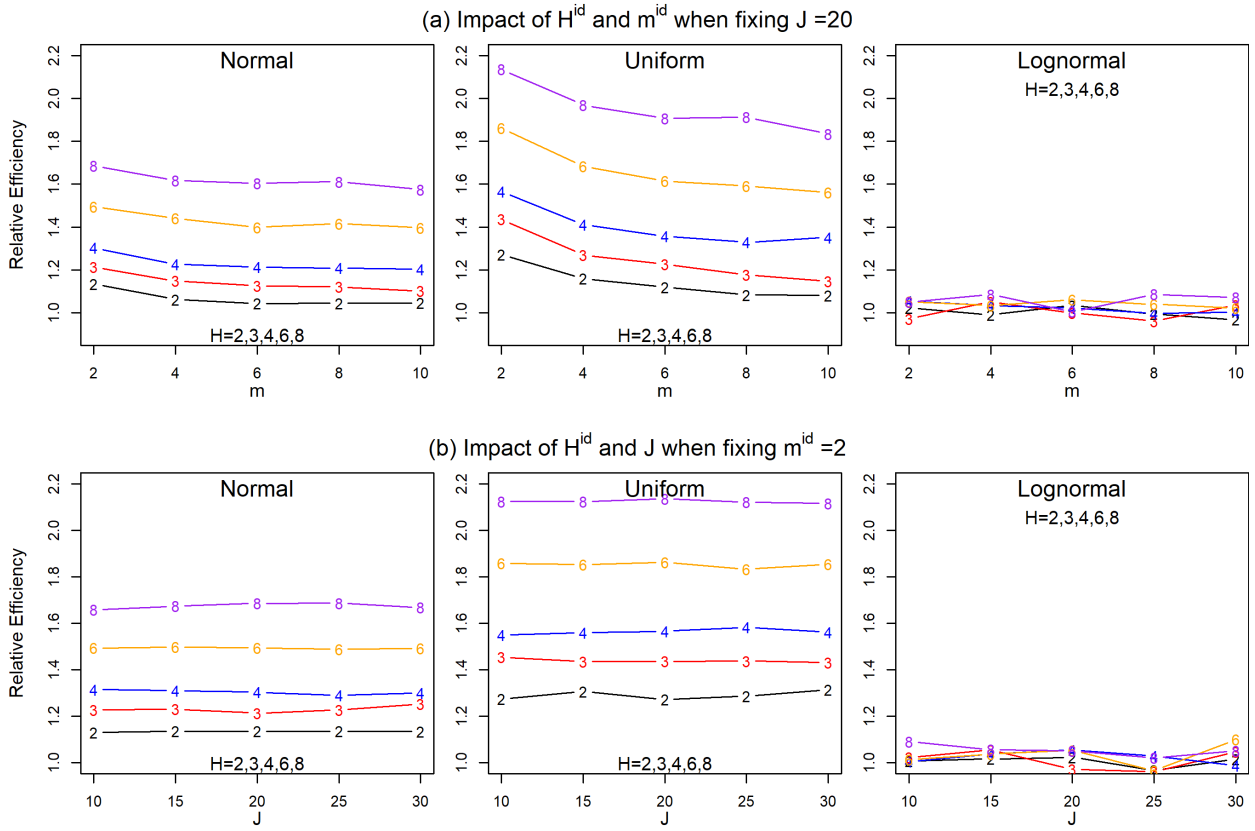



Figure 2.7: The impact of $(H^{id}, m^{id})$ and $(H^{id}, J)$ on relative efficiencies of $\hat{\sigma}_r$ versus $\tilde{\sigma}_r$ under completely balanced CRDs, with perfect ranking at the individual level only

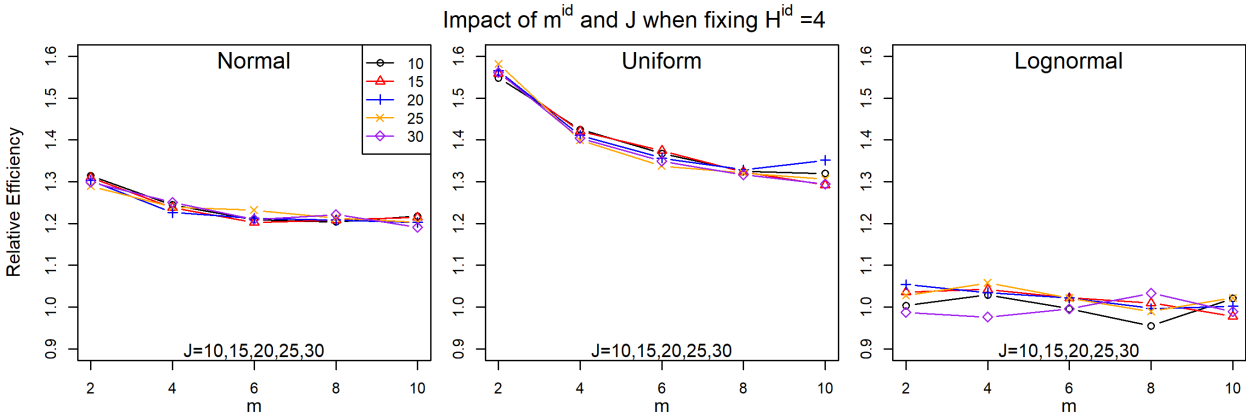



Figure 2.8: The impact of $(m^{id}, J)$ on relative efficiencies of $\hat{\sigma}_r$ versus $\tilde{\sigma}_r$ under completely balanced CRDs, with perfect ranking at the individual level only

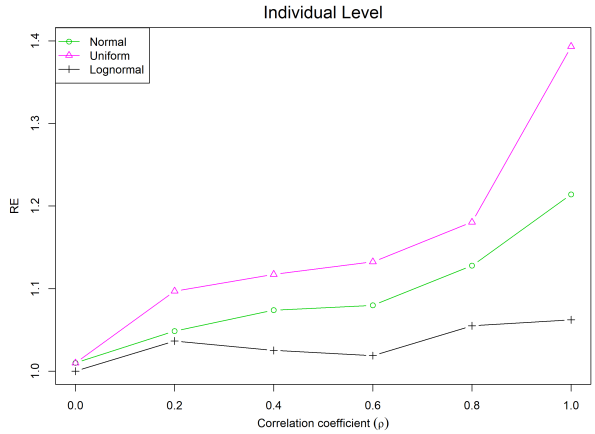



Figure 2.9: Relative efficiency of $\hat{\sigma}_r$ versus $\rho$ under completely balanced CRDs for three distributions of the ranking variables, including $N(0, 2^2)$, $U(-3.47, 3.47)$, $\text{LN}(0, 0.941) - 1.6$.

## 2.5. Paule and Mandel estimator

Meta-analysis is a statistical procedure to combine data from multiple studies targeting at making inferences on a common treatment effect. Often researchers assign different weights to different studies in estimating the overall treatment effect and one of the most commonly used weighting scheme is inverse-variance weighting. There is a structural similarity between a CRD and a meta-analysis. We can view the studies in a meta-analysis as the clusters in a CRD. Thus, a widely used method in meta-analysis for estimating the between-study heterogeneity, proposed by Paule and Mandel [63], can be applied here in combination with the ranking information to estimate the between cluster variance. We denote this method by PM and the corresponding estimator of $\sigma_b^2$ by $\hat{\sigma}_{b.PM}^2$. Under model (2.1), let $\hat{\Delta}_j$ be the estimated treatment effect in cluster $j$ for $j = 1, \ldots, J$. Under the inverse-variance weighting scheme, the overall treatment effect can be estimated by $\hat{\Delta} = \frac{\sum_{j=1}^{J} w_j \hat{\Delta}_j}{\sum_{j=1}^{J} w_j}$, where $w_j = \left[\widehat{Var}(\hat{\Delta}_j)\right]^{-1} = \frac{1}{s_j^2 + \sigma_b^2}$ and $s_j^2$ is the estimated within-cluster variance of cluster $j$. A useful quantity in estimating $\sigma_b^2$ is the $Q$-statistic defined as

$$Q \equiv \sum_{j=1}^{J} w_j \left(\hat{\Delta}_j - \hat{\Delta}\right)^2. \tag{2.6}$$

The expected value of $Q$ is $J-1$. The estimator $\hat{\sigma}_{b.PM}^2$ can be obtained by iteratively solving the equation $Q \equiv \sum_{j=1}^{J} w_j \left(\hat{\Delta}_j - \hat{\Delta}\right)^2 = J - 1$. Ranking information from RSS can be integrated into this procedure to further improve the estimation efficiency.

### 2.5.1. Ranking at the cluster level

For ranking at the cluster level only, we use (2.3) to estimate $\sigma_{b.i}^2$ first. The PM approach can be applied here to estimate each $\sigma_{b.ih}^2$ for $h = 1, \ldots, H_i$ and $i = 0, 1$ . Define the

$Q$-statistic for the ranking stratum with $O_{j(i)} = h$ within treatment group $i$ as

$$Q_{ih}(\sigma^2_{b.ih}) \equiv \sum_{j \in \mathcal{J}_i(h)} w^*_{j(i)} \left(\hat{\mu}_{j(i)} - \mu^*_{ih}\right)^2,$$

where $w^*_{j(i)} = \left[Var\left(\hat{\mu}_{j(i)}\right)\right]^{-1} = \left[\sigma^2_{b.ih} + \frac{\hat{\sigma}^2_r}{K_{j(i)}}\right]^{-1}$, $\hat{\mu}^*_{ih} = \frac{\sum_{j \in \mathcal{J}_i(h)} w^*_{j(i)} \hat{\mu}_{j(i)}}{\sum_{j \in \mathcal{J}_i(h)} w^*_{j(i)}}$ and $\hat{\sigma}^2_r = \frac{SSW^c}{K_{\cdot\cdot} - J_{\cdot}}$ as derived in Section 2.3. The expected value of $Q_{ih}(\sigma^2_{b.ih})$ is $J_{ih} - 1$, and thus we can obtain the PM estimator for $\sigma^2_{b.ih}$, denoted by $\hat{\sigma}^2_{b.ih.PM}$, by iteratively solving the following equation

$$Q_{ih}(\sigma^2_{b.ih}) = \sum_{j \in \mathcal{J}_i(h)} w^*_{j(i)} \left(\hat{\mu}_{j(i)} - \hat{\mu}^*_{ih}\right)^2 = J_{ih} - 1.$$

If $\sum_{j \in \mathcal{J}_i(h)} w^*_{j(i)} \left(\hat{\mu}_{j(i)} - \hat{\mu}^*_{ih}\right)^2 > J_{ih} - 1$ for all $\sigma^2_{b.ih} \geq 0$, we set $\hat{\sigma}^2_{b.ih.PM} = 0$. Combining the unbiased estimator for $\sum_{h=1}^{H_i} \mu^2_{ih}$ in (2.4) and $\hat{\sigma}^2_{b.ih.PM}$, we can get an PM estimator for $\sigma^2_{b.i}$ as

$$\hat{\sigma}^2_{b.i.PM} = \frac{1}{H_i} SSBR^c(i) - \frac{1}{H_i}(1 - \frac{1}{H_i}) \sum_{h=1}^{H_i} \frac{SSB^c(i,h)}{(J_{ih} - 1)J_{ih}} + \frac{1}{H_i} \sum_{h=1}^{H_i} \hat{\sigma}^2_{b.ih.PM}.$$

Thus, the PM estimator for $\sigma^2_b$ can be easily obtained by taking the average of the estimates from the two treatment groups as $\hat{\sigma}^2_{b.PM} = \frac{\hat{\sigma}^2_{b.0.PM} + \hat{\sigma}^2_{b.1.PM}}{2}$. When $H_i \equiv H^c$, we can aggregate information from both treatment groups to improve our estimation as we did previously in the MOM method.

### 2.5.2. Ranking at the individual level

For ranking at the individual level only, we no longer need to perform the $PM$ procedure for each ranking stratum separately. Instead, we can obtain a $Q$-statistic for the entire data

set. We denote the $Q$-statistic for treatment group $i$ by $Q_i(\sigma_b^2)$, defined as

$$Q_i(\sigma_b^2) \equiv \sum_{j=1}^{J_i} w^*_{j(i)} \left(\hat{\mu}_{j(i)} - \mu^*_i\right)^2,$$

where $\mu^*_i = \sum_{j=i}^{J_i} w^*_{j(i)}\hat{\mu}_{j(i)} / \sum_{j=1}^{J_i} w^*_{j(i)}$ and $w^*_{j(i)} = \left[Var(\hat{\mu}_{j(i)})\right]^{-1} = \left[\sigma_b^2 + \frac{1}{H^2_{j(i)}} \sum_{h=1}^{H_{j(i)}} \frac{\hat{\sigma}^2_{r.ijh}}{K_{ijh}}\right]^{-1}$. Let $Q(\sigma_b^2) \equiv \sum_{i=0}^{1} Q_i(\sigma_b^2)$ and we have $E\left[Q(\sigma_b^2)\right] = J_0 + J_1 - 2$. By setting $Q(\sigma_b^2) = \sum_{i=0}^{1}\sum_{j=1}^{J_i} w^*_{j(i)} \left(\hat{\mu}_{j(i)} - \hat{\mu}^*_i\right) = J_0 + J_1 - 2$ and iteratively solving for $\sigma_b^2$, we can obtain the $PM$ estimator $\hat{\sigma}^2_{b.PM}$. Similarly, if $Q(\sigma_b^2) < J_0 + J_1 - 2$ for all $\sigma_b^2 \geq 0$, we set $\hat{\sigma}^2_{b.PM} = 0$. When $H_{j(i)} \equiv H^{id}$, $\hat{\sigma}^2_{b.PM}$ can be improved by pooling information to better estimate $\sigma^2_{r.ijh}$ by $\hat{\sigma}^2_{r.h} = \frac{\sum_{i=0}^{1}\sum_{j=1}^{J_i} SSW^{id}(i,j,h)}{K_{..h} - J_.}$.

### 2.5.3. Ranking at both levels

When ranking is performed at both levels, the $PM$ method is similar to the case for ranking at the cluster level only expect for the specification of weights $w^*_{j(i)}$'s. Since we are ranking at both levels, the weight $w^*_{j(i)}$ for cluster $j$ within the $h$th ranking stratum in treatment group $i$ is $\left[Var\left(\hat{\mu}_{j(i)}\right)\right]^{-1} = \left[\sigma^2_{b.ih} + \frac{1}{H^2_{j(i)}} \sum_{h=1}^{H_{j(i)}} \frac{\hat{\sigma}^2_{r.ijh}}{K_{ijh}}\right]^{-1}$. The $PM$ procedure is performed for each ranking stratum in the two treatment groups as for ranking at the cluster level only.

We list all expressions and weight specifications for the PM method for all three ranking schemes in Table 2.3, including cases where information can be pooled to improve the estimation efficiency. Note that those expressions involve variance estimates, such as $\hat{\sigma}^2_{b.ih.PM}$, which should be solved numerically.

| | Design | Estimator |
|---|---|---|
| Ranking at cluster level | General | $\frac{1}{2}\sum_{i=0}^{1}\left[\frac{1}{H_i}SSBR^c(i)-\frac{1}{H_i}(1-\frac{1}{H_i})\sum_{h=1}^{H_i}\frac{SSB^c(i,h)}{(J_{ih}-1)J_{ih}}+\frac{1}{H_i}\sum_{h=1}^{H_i}\hat{\sigma}^2_{b.ih.PM}\right]$ |
| | $H_i=H^c$ | $\frac{1}{2H^c}\sum_{i=0}^{1}\left[SSBR^c(i)-(1-\frac{1}{H^c})\sum_{h=1}^{H^c}\frac{1}{J_{ih}}\hat{\sigma}^2_{b.h.PM}-(1-\frac{1}{H^c})\sum_{h=1}^{H^c}\frac{1}{J^2_{ih}}\sum_{j\in\mathcal{J}_i(h)}\frac{1}{K_{j(i)}}\frac{SSW^c}{K_{..}-J_.}\right]+\frac{1}{H^c}\sum_{h=1}^{H^c}\hat{\sigma}^2_{b.h.PM}$ |
| | $H_i=H^c$, $J_{ih}=m^c$ | $\frac{1}{2H^c}\sum_{i=0}^{1}\left[SSBR^c(i)-\frac{1}{(m^c)^2}(1-\frac{1}{H^c})\sum_{h=1}^{H^c}\sum_{j\in\mathcal{J}_i(h)}\frac{SSW}{2H^cm^cK_{j(i)}(K_{j(i)}-1)}\right]+\left(\frac{1}{H^c}-\frac{1}{H^cm^c}+\frac{1}{(H^c)^2m^c}\right)\sum_{h=1}^{H^c}\hat{\sigma}^2_{b.h.PM}$ |
| | $H_i=H^c$, $J_{ih}=m^c$, $K_{j(i)}=K$ | $\frac{1}{2H^c}\sum_{i=0}^{1}SSBR^c(i)-\frac{(H^c-1)SSW}{2(H^cm^c)^2K(K-1)}+\left(\frac{1}{H^c}-\frac{1}{H^cm^c}+\frac{1}{(H^c)^2m^c}\right)\sum_{h=1}^{H^c}\hat{\sigma}^2_{b.h.PM}$ |
| Ranking at individual level | General | $\sum_{i=0}^{1}\sum_{j=1}^{J_i}\hat{w}^*_{j(i)}(\hat{\mu}_{j(i)}-\hat{\mu}^*_i)^2=J_.-2$Solve for $\sigma^2_b$, where $\hat{w}^*_{j(i)}=\left[\sigma^2_b+\frac{1}{H^2_{j(i)}}\sum_{h=1}^{H_{j(i)}}\frac{\hat{\sigma}^2_{r.ijh}}{K_{ijh}}\right]^{-1}$ |
| | $H_{j(i)}=H^{id}$ | $\hat{w}^*_{j(i)}=\left[\sigma^2_b+\frac{1}{(H^{id})^2}\sum_{h=1}^{H^{id}}\frac{\hat{\sigma}^2_{r.h}}{K_{ijh}}\right]^{-1}$ |
| | $H_{j(i)}=H^{id}$, $K_{ijh}=m^{id}$ | $\hat{w}^*_{j(i)}=\left[\sigma^2_b+\frac{1}{(H^{id})^2m^{id}}\sum_{h=1}^{H^{id}}\hat{\sigma}^2_{r.h}\right]^{-1}$ |
| Ranking at both levels | General | $\hat{\sigma}^2_b=\frac{1}{2}\sum_{i=0}^{1}\left[\frac{SSBR1(i)}{H_i}-\frac{1}{H_i}(1-\frac{1}{H_i})\sum_{h=1}^{H_i}\frac{SSB^B(i,h)}{(J_{ih}-1)J_{ih}}+\frac{1}{H_i}\sum_{h=1}^{H_i}\hat{\sigma}^2_{b.ih.PM}\right]$ |
| | $H_i=H^c$, $H_{j(i)}=H^{id}$ | $\frac{1}{2H^c}\sum_{i=0}^{1}\left[SSBR1(i)-(1-\frac{1}{H^c})\sum_{h=1}^{H^c}\frac{1}{J_{ih}}\hat{\sigma}^{2(PM)}_{b.h}-(1-\frac{1}{H^c})\sum_{h=1}^{H^c}\frac{1}{J^2_{ih}}\sum_{j\in\mathcal{J}_i(h)}\frac{1}{(H^{id})^2}\sum_{h'=1}^{H^{id}}\frac{\hat{\sigma}^2_{r.h'}}{K_{ijh'}}\right]+\frac{1}{H^c}\sum_{h=1}^{H^c}\hat{\sigma}^2_{b.h.PM}$ |
| | $H_i=H^c$, $m_i=m^c$, $H_{j(i)}=H^{id}$, $m_{j(i)}=m^{id}$ | $\frac{1}{2H^c}\sum_{i=0}^{1}SSBR1(i)+\left(\frac{1}{H^c}-\frac{1}{H^cm^c}+\frac{1}{(H^c)^2m^c}\right)\sum_{h=1}^{H^c}\hat{\sigma}^2_{b.h.PM}-\frac{(H^c-1)\sum_{i=0}^{1}\sum_{j=1}^{J_i}\sum_{h'=1}^{H^{id}}SSW^B(i,j,h')}{2(H^cm^c)^2m^{id}(m^{id}-1)}$ |

Table 2.3: PM estimators for $\sigma^2_b$

## 2.6. Improvement of PM over MOM

We ran simulations to estimate the RE of the PM estimator. The most significant improvement of the PM estimator is observed when $K$ is small for ranking at the cluster level only as shown in Figure 2.10. The improvement of the PM method in estimating the heterogeneity parameter in meta-analysis is also found to be significant when sample sizes are small [61, 60]. As we can observe from Panel (a) of Figure 2.10, the PM estimator has much higher RE compared to the MOM estimator when $K=3$ and the difference diminishes as $K$ increases. The PM method does not improve the estimation of $\sigma_b$ when RSS is applied at the individual level only as shown in Figure 2.10(b).

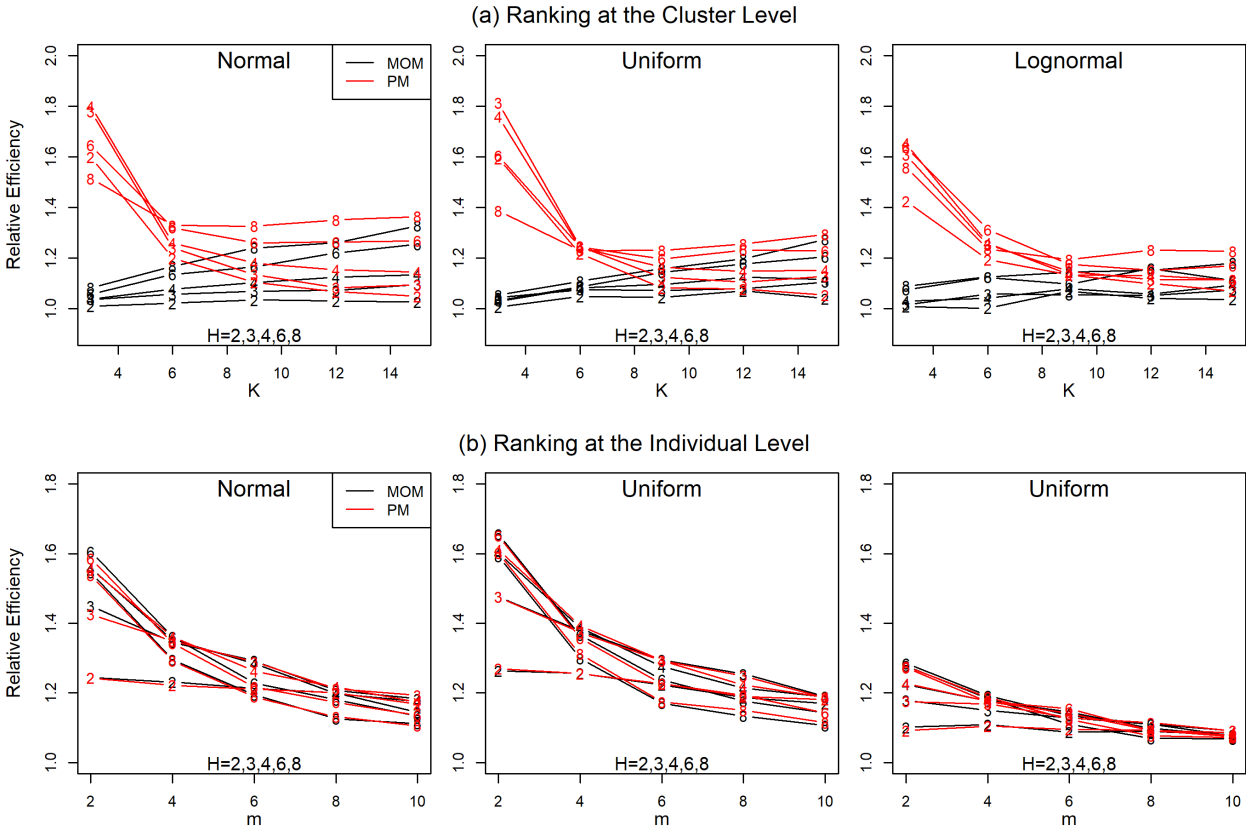



Figure 2.10: The improvement in relative efficiency of PM estimator over MOM estimator.

We also examined the influence of imperfect ranking to the PM estimator under the same settings as specified in Section 2.4.1.2. In general, RE increases as $\rho$ increases for both ranking schemes. For ranking at the cluster level only, the improvement of the Uniform distribution is greater than the Normal distribution in general as is the case for the MOM estimator. However, for ranking at the individual level only, The Uniform and the Normal distributions have similar RE and both are greater than the RE of the Lognormal case.

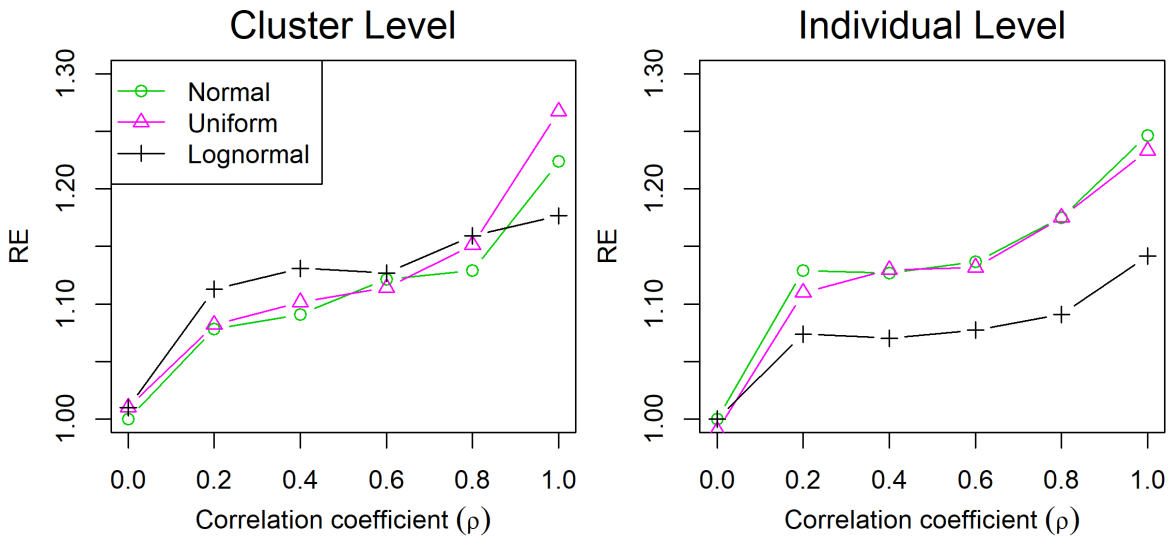



Figure 2.11: Relative efficiency of $\hat{\sigma}_{b.PM}$ versus $\rho$ under completely balanced CRDs for three distributions of the ranking variables, including $N(0,1)$, $U(-1.74, 1.74)$, $\text{LN}(0, 0.481) - 1.27$ for ranking at the cluster level and $N(0, 2^2)$, $U(-3.47, 3.47)$, $\text{LN}(0, 0.941) - 1.6$ for ranking at the individual level.

## 2.7. Data example

We conducted an empirical study to illustrate the proposed methods. We applied RSS-structured CRDs to the Academic Performance Index (API) Data from the California Department of Education. The API is a single number on a scale of 200-1000 indicating the performance of public schools based on students' test scores in the spring from the previous year. The APIs were calculated based on test results of the Standardized Testing and Reporting (STAR) Program, the California High School Exit Examination (CAHSEE), and the California Alternate Performance Assessment (CAPA). All public schools with at least 100 students are included. For comprehensive information about API, please refer to https://www.cde.ca.gov/re/pr/api.asp.

We used the 2000 API scores as the response, which are publicly available in the R package "survey". The data present a hierarchical structure where schools are nested within school districts. There are 6194 schools from 757 school districts in total. For illustrative purposes we removed districts with number of schools above 100 or below 20 and treated the remaining 60 school districts, with a total of 2077 schools, as our population, from which samples were drawn using either SRS or RSS. We generated rankings using the 1999 API scores as they are highly correlated with the 2000 API scores with a correlation of 0.99 at the district level, and 0.98 at the school level. The average district-level scores were used for ranking at the cluster level. The ICC is about 0.55 for the preprocessed data.

We are interested in estimating the two standard deviations, $\sigma_b$ and $\sigma_r$. We first examined the performance of the two proposed estimators under different (completely balanced) designs. We fixed the number of school districts at $J = 15$ and the number of schools from each selected district at $K = 6$. Eight RSS-structured (completely balanced) CRDs were considered: two for scheme (i) ranking at the cluster level only, two for scheme (ii) ranking at the individual level only, and four for scheme (iii) ranking at both levels, as listed in Table 2.4. For each design, we generated 100,000 RSS samples and 100,000 SRS samples from the population. School districts were randomly assigned to either the treatment or the control group since the actual treatment effect is zero. Then we computed MSE for RSS-based estimators $\hat{\sigma}_b$, $\hat{\sigma}_{b.PM}$, $\hat{\sigma}_r$ and SRS-based estimators $\tilde{\sigma}_b$, $\tilde{\sigma}_r$. The empirical RE was recorded in Table 2.4 as the ratios of the corresponding MSE. The true standard deviations were estimated from the population.

The results are consistent with most of the properties observed in our simulation studies in Section 2.4. In terms of estimating $\sigma_b$, Design i-2 has a higher RE than i-1, suggesting a larger $H^c$ may lead to a higher efficiency, in agreement with the conclusions summarized based on Figure 2.1 (b). For ranking at the individual level only, Design ii-2 is slightly more efficient than ii-1, suggesting the RE may decrease as $m^{id}$ increases. We conjecture that the small improvement of RSS over SRS when ranking is conducted at the individual level only

may result from the relatively large ICC. Ranking at both levels (Design iii) improves the performance of RSS-based estimators, and Design iii-2-2 is the best for the MOM method. The PM method provides higher RE in general compared to the MOM method since we have a small sample size ($K = 6$). In terms of estimating $\sigma_r$, ranking at the cluster level does not contribute to the estimation efficiency, again in agreement with the observation from Figure 2.6. Similarly, due to the large ICC, the reduction in MSE of RSS-based estimator for $\sigma_r$ is limited when ranking is performed at the individual level only.

| | | | | | MOM | | PM |
|---|---|---|---|---|---|---|---|
| DesignID | $H^c$ | $m^c$ | $H^{id}$ | $m^{id}$ | $\hat{\sigma}_b$ | $\hat{\sigma}_r$ | $\hat{\sigma}_{b.PM}$ |
| i-1 | 3 | 5 | 1 | 6 | 1.08 | 1.01 | 1.12 |
| i-2 | 5 | 3 | 1 | 6 | 1.22 | 1.00 | 1.49 |
| ii-1 | 1 | 15 | 2 | 3 | 1.01 | 1.01 | 1.06 |
| ii-2 | 1 | 15 | 3 | 2 | 1.04 | 1.08 | 1.06 |
| iii-1-1 | 3 | 5 | 2 | 3 | 1.17 | 1.00 | 1.16 |
| iii-1-2 | 3 | 5 | 3 | 2 | 1.14 | 1.03 | 1.13 |
| iii-2-1 | 5 | 3 | 2 | 3 | 1.23 | 1.00 | 1.55 |
| iii-2-2 | 5 | 3 | 3 | 2 | 1.28 | 1.04 | 1.37 |

Table 2.4: California API example: comparing performance of different (completely balanced) designs in estimating $\sigma_b$ and $\sigma_r$ . Estimators are compared via empirical RE.

## 2.8. Conclusion and discussion

Among all the research on RSS, most focused on the inference of the mean. However, the variance components, especially in CRDs, can provide extra information about the heterogeneity of different clusters. In this work we proposed two RSS-based nonparametric methods for estimating $\sigma_b^2$ and $\sigma_r^2$. Both methods were shown through simulations to improve estimation efficiency over the SRS-based estimators. We explored the impact of design parameters and imperfect ranking on the RE of $\hat{\sigma}_b$ and $\hat{\sigma}_r$ . It was shown that using RSS is better than SRS in estimating both variance components even when ranking is imper-

fect. There do exist some general properties in terms of the impact of design parameters on RE. However, due to the truncation in the estimation and the unstable performance under certain settings, these trends are not as consistent and clear as those in the estimation of treatment effects. Whether to apply RSS in a real application, especially for the inference on variances and what schemes to apply, can be complicated and should be addressed carefully by considering the characteristic of the population, such as ICC, and the choices of potential cost-efficient ranking methods.

We also found that some widely used methods in meta-analysis can be applied to further improve RSS under certain settings. Though we only applied the widely used and fairly straightforward method, to avoid the potential complexity when ranking information is available, we believe other estimators can possibly be modified to achieve better performance.

We note that the idea of RSS can be easily generalized to data with multiple levels. Another potential extension is to evaluate the proposed methods under unbalanced designs as what [89] did for estimating and testing treatment effects.

# APPENDIX A

# APPENDIX OF CHAPTER 1

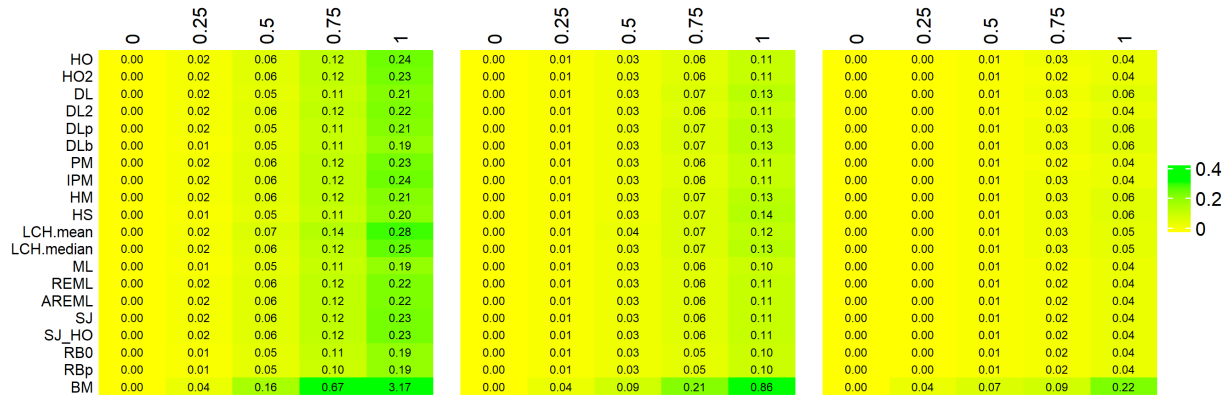
## A.1. Evaluation of the impact of $R$, $K$, $\theta$, and $w$ on estimation bias and MSE


(a) R=1

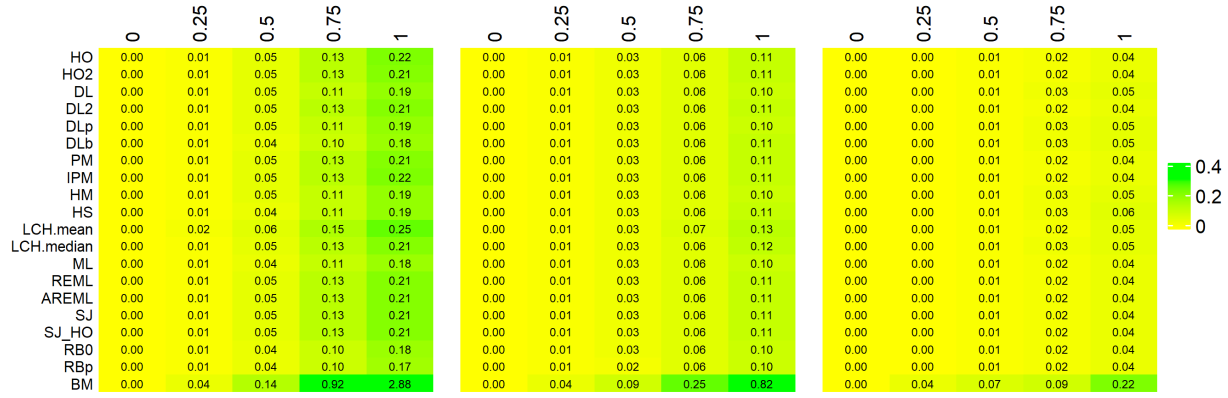

(b) R=2


(c) R=4

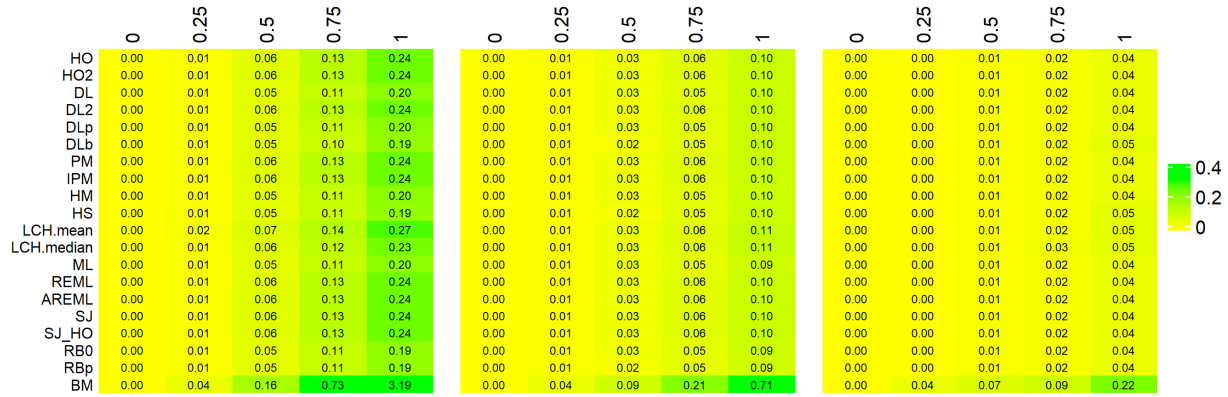
Figure A.1: Large-sample performance of different $\tau^2$ estimators in terms of estimation bias for different $R$ and $K$ values based on settings with $\mu = -2.5$, $\theta = 0$ and $w = 0$. In the $3 \times 3$ matrix of heat maps, the rows correspond to $R = 1, 2, 4$ from top to bottom and the columns correspond to $K = 10, 20, 50$ from left to right.




(a) R=1

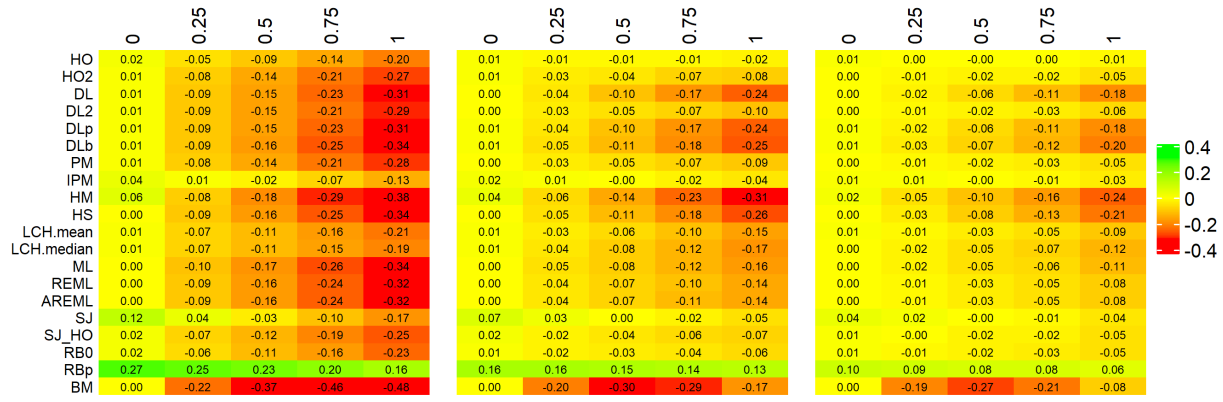



(b) R=2

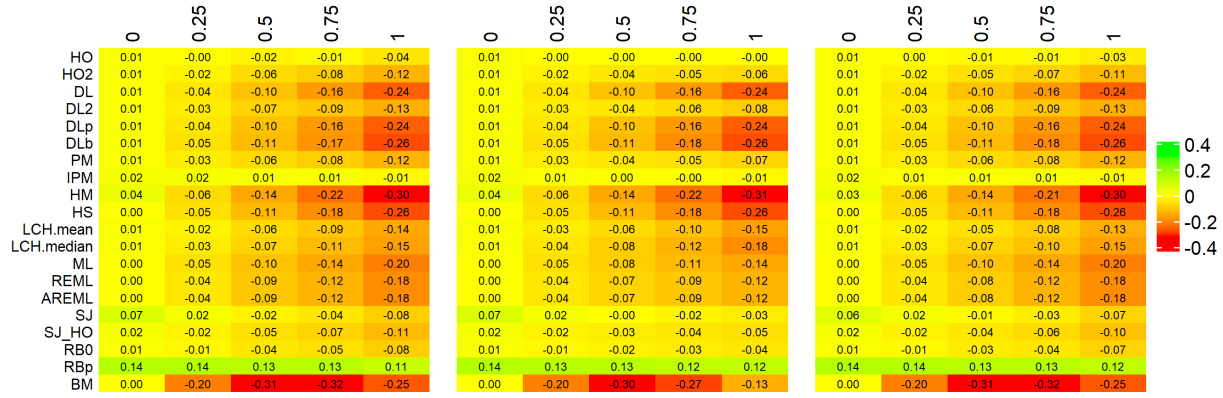



(c) R=4

Figure A.2: Large-sample performance of different $\tau^2$ estimators in terms of MSE for different $R$ and $K$ values based on settings with $\mu = -2.5$, $\theta = 0$ and $w = 0$. In the $3 \times 3$ matrix of heat maps, the rows correspond to $R = 1, 2, 4$ from top to bottom and the columns correspond to $K = 10, 20, 50$ from left to right.



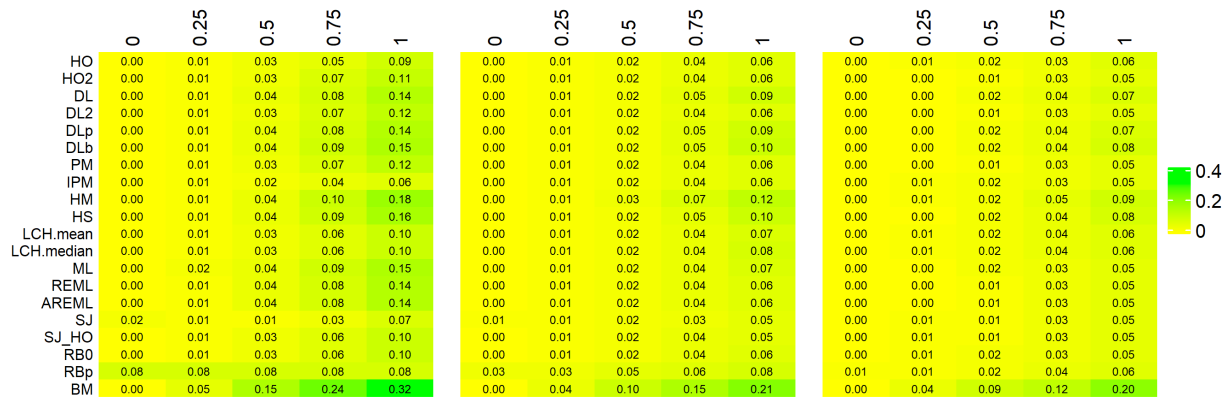

(a) $\theta$=-1



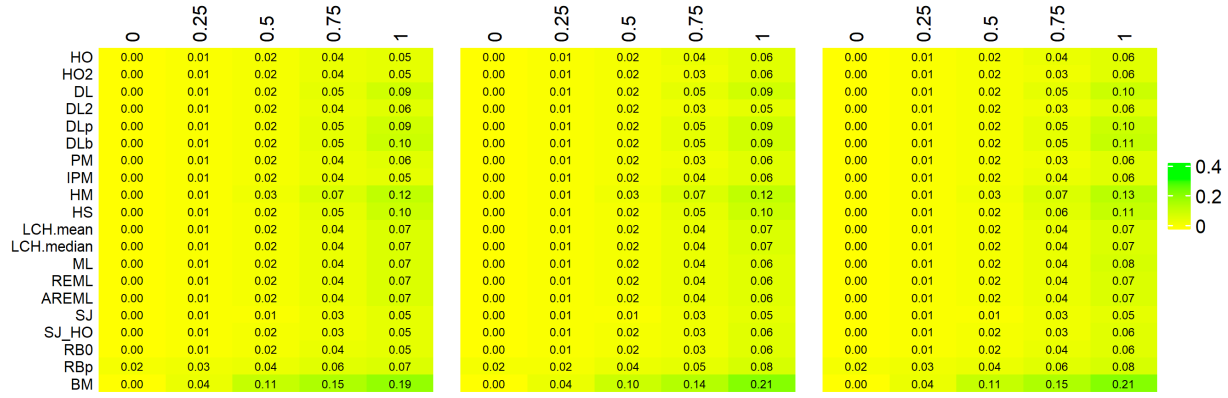

(b) $\theta = 0$


(c) $\theta = 1$

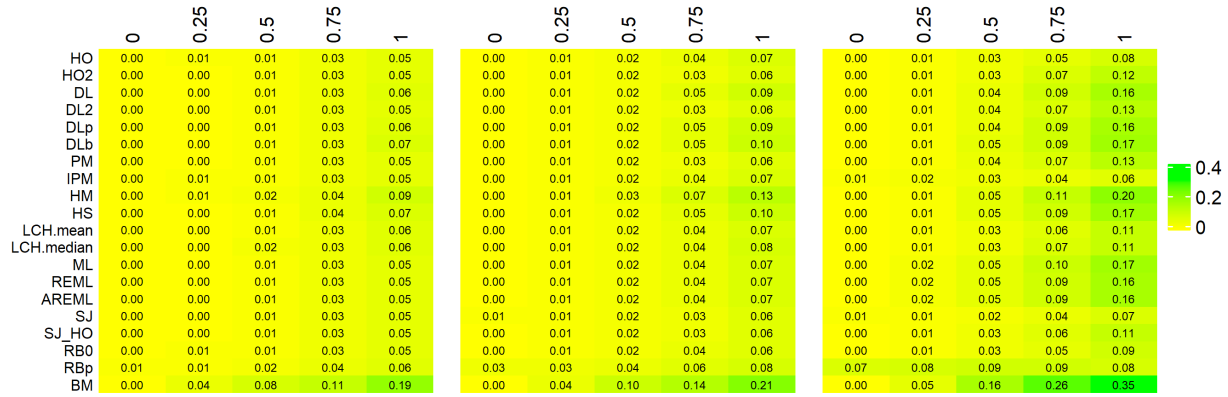
Figure A.3: Large-sample performance of different $\tau^2$ estimators in terms of estimation bias for different $\theta$ and $w$ values based on settings with $R = 1$, $K = 50$ and $\mu = -5$. In the $3 \times 3$ matrix of heat maps, the rows correspond to $\theta = -1, 0, 1$ from top to bottom and the columns correspond to $w = 0, 0.5, 1$ from left to right.




(a) $\theta$=-1

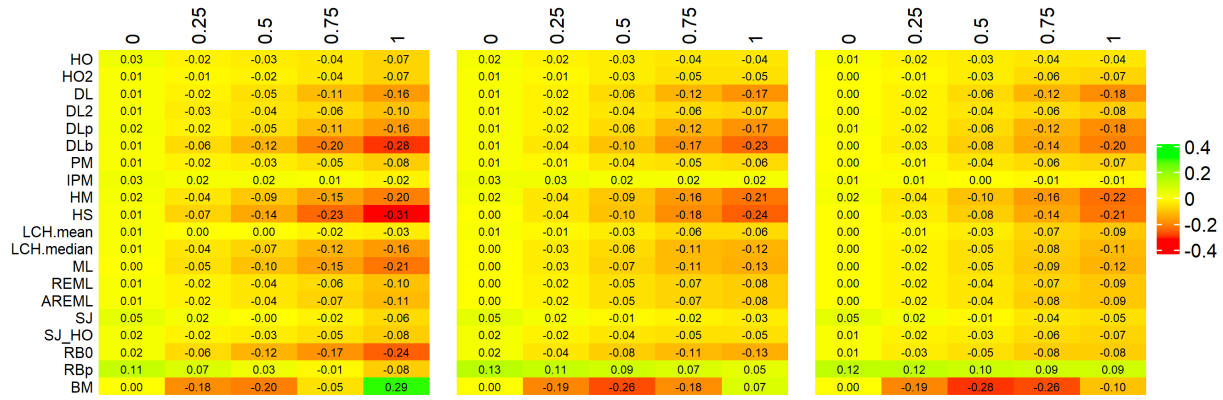



(b) $\theta = 0$

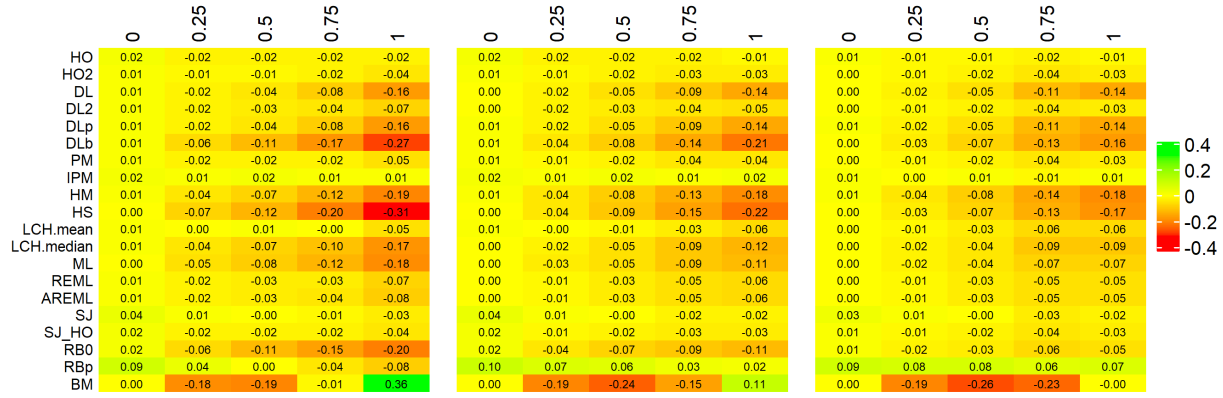



(c) $\theta = 1$

Figure A.4: Large-sample performance of different $\tau^2$ estimators in terms of MSE for different $\theta$ and $w$ values based on settings with $R = 1$, $K = 50$ and $\mu = -5$. In the $3 \times 3$ matrix of heat maps, the rows correspond to $\theta = -1, 0, 1$ from top to bottom and the columns correspond to $w = 0, 0.5, 1$ from left to right.



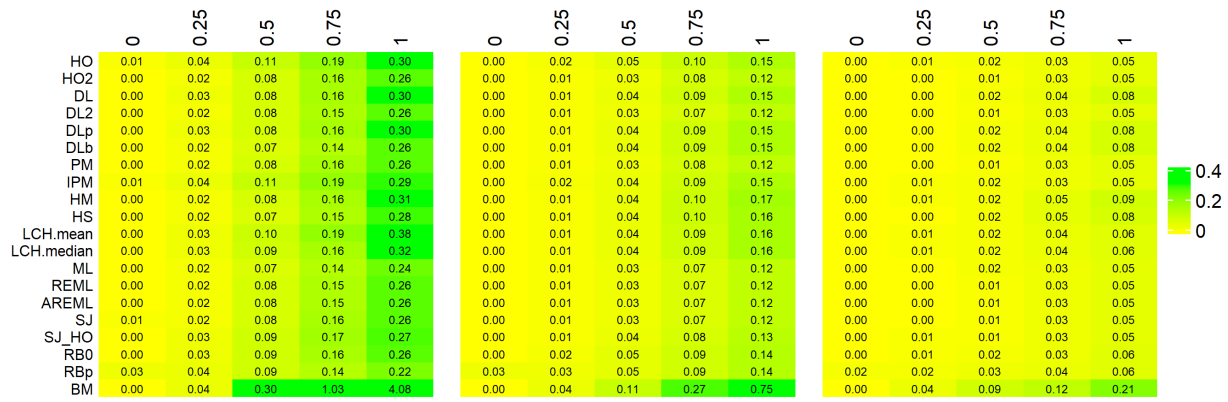

(a) R=1



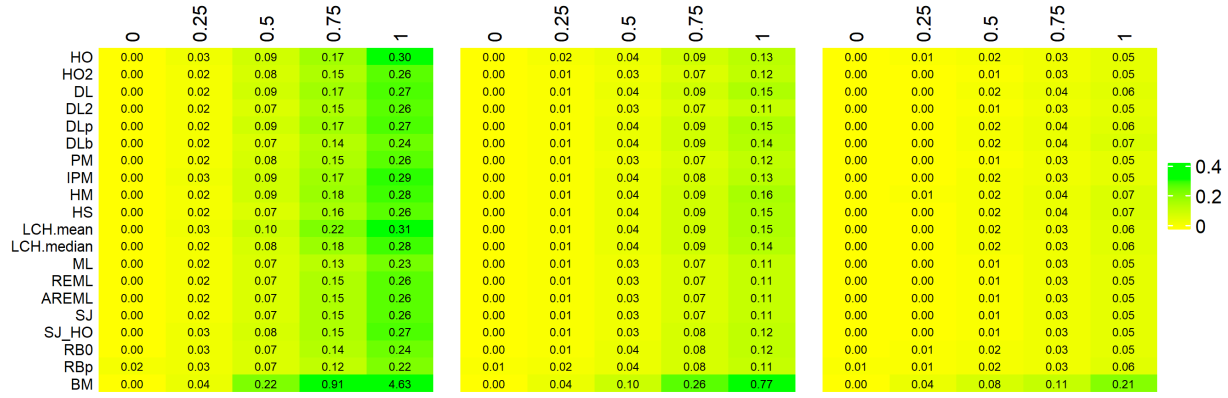

(b) R=2


(c) R=4

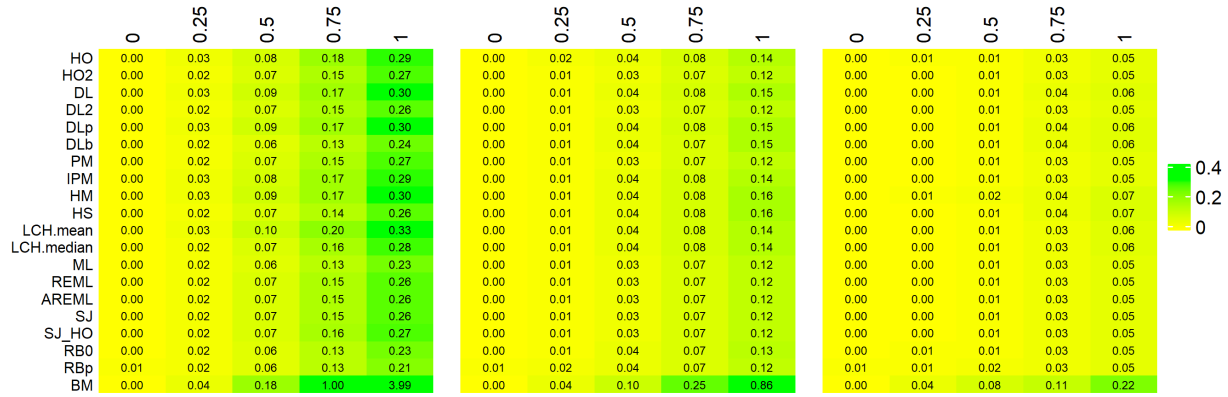
Figure A.5: Small-sample performance of different $\tau^2$ estimators in terms of estimation bias for different $R$ and $K$ values based on settings with $\mu = -2.5$, $\theta = 0$ and $w = 0$. In the $3 \times 3$ matrix of heat maps, the rows correspond to $R = 1, 2, 4$ from top to bottom and the columns correspond to $K = 10, 20, 50$ from left to right.




(a) R=1

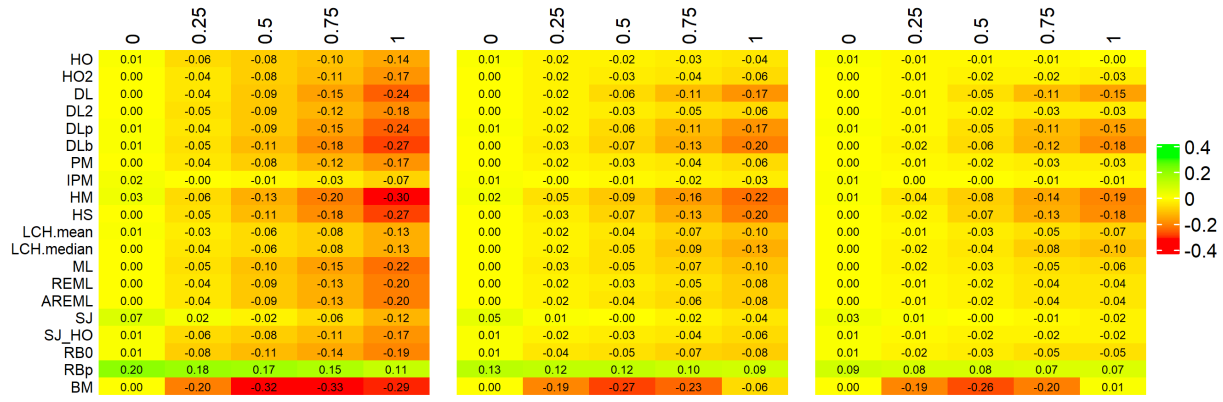



(b) R=2

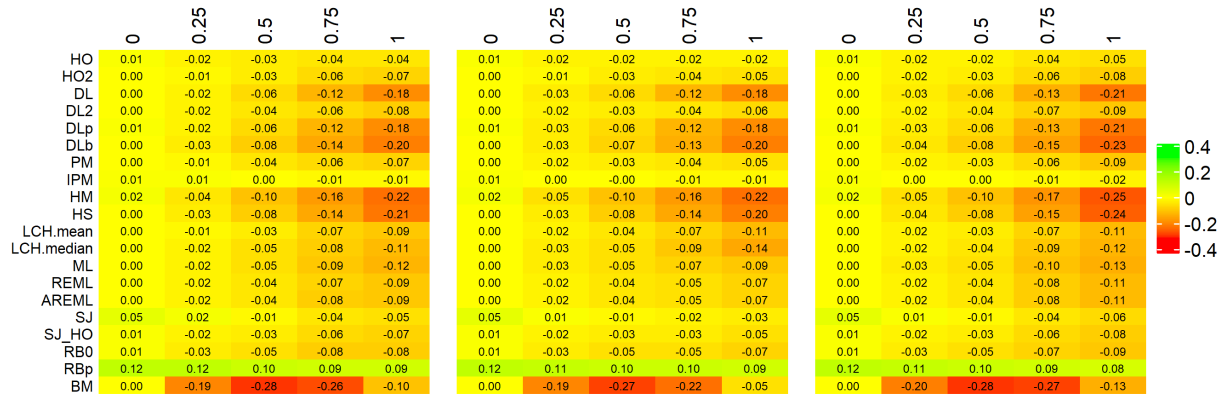



(c) R=4

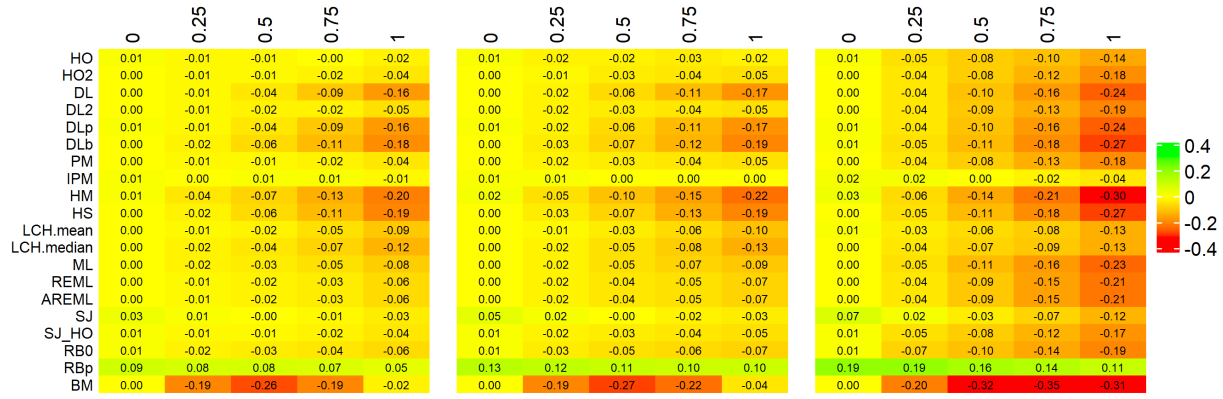
Figure A.6: Small-sample performance of different $\tau^2$ estimators in terms of MSE for different $R$ and $K$ values based on settings with $\mu = -2.5$, $\theta = 0$ and $w = 0$. In the $3 \times 3$ matrix of heat maps, the rows correspond to $R = 1, 2, 4$ from top to bottom and the columns correspond to $K = 10, 20, 50$ from left to right.




(a) $\theta$=-1

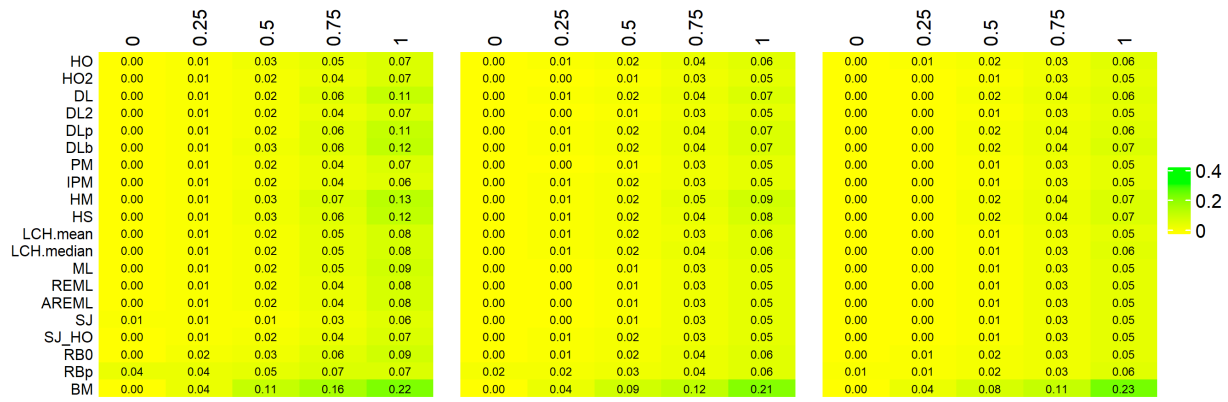



(b) $\theta = 0$

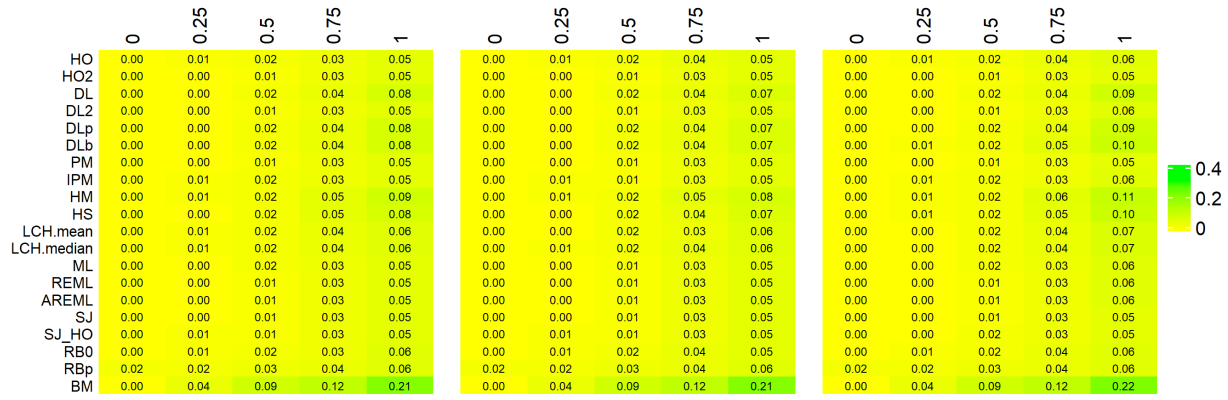



(c) $\theta = 1$

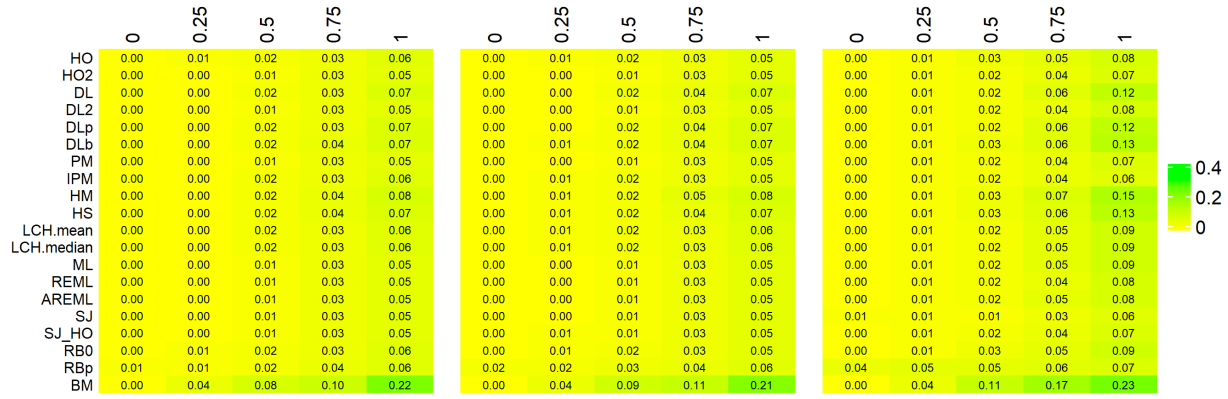
Figure A.7: Small-sample performance of different $\tau^2$ estimators in terms of estimation bias for different $\theta$ and $w$ values based on settings with $R = 1$, $K = 50$ and $\mu = -2.5$. In the $3 \times 3$ matrix of heat maps, the rows correspond to $\theta = -1, 0, 1$ from top to bottom and the columns correspond to $w = 0, 0.5, 1$ from left to right.




(a) $\theta$=-1




(b) $\theta = 0$




(c) $\theta = 1$

Figure A.8: Small-sample performance of different $\tau^2$ estimators in terms of MSE for different $\theta$ and $w$ values based on settings with $R = 1$, $K = 50$ and $\mu = -2.5$. In the $3 \times 3$ matrix of heat maps, the rows correspond to $\theta = -1, 0, 1$ from top to bottom and the columns correspond to $w = 0, 0.5, 1$ from left to right.

## A.2. Data

| | Type 2 diabetes mellitus with gestational diabetes | | Type 2 diabetes mellitus without gestational diabetes | |
|---|---|---|---|---|
| Study | # events | # observations | # events | # observations |
| 1 | 2847 | 21823 | 6628 | 637341 |
| 2 | 71 | 620 | 22 | 868 |
| 3 | 21 | 68 | 0 | 39 |
| 4 | 43 | 166 | 150 | 2242 |
| 5 | 53 | 295 | 1 | 111 |
| 6 | 405 | 5470 | 16 | 783 |
| 7 | 6 | 70 | 7 | 108 |
| 8 | 13 | 35 | 8 | 489 |
| 9 | 7 | 23 | 0 | 11 |
| 10 | 23 | 435 | 0 | 435 |
| 11 | 44 | 696 | 0 | 70 |
| 12 | 21 | 229 | 1 | 61 |
| 13 | 10 | 28 | 0 | 52 |
| 14 | 15 | 45 | 1 | 39 |
| 15 | 105 | 801 | 7 | 431 |
| 16 | 10 | 15 | 0 | 35 |
| 17 | 33 | 241 | 0 | 57 |
| 18 | 14 | 47 | 3 | 47 |
| 19 | 224 | 615 | 18 | 328 |
| 20 | 5 | 145 | 0 | 41 |

Table A.1: Data from a meta-analysis of studies on type 2 diabetes mellitus after gestational diabetes [6].

| | Rosiglitazone | | | Control | | | | Rosiglitazone | | | Control | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Study | N | # MI | # CVD | N | # MI | # CVD | Study | N | # MI | # CVD | N | # MI | # CVD |
| 1 | 357 | 2 | 1 | 176 | 0 | 0 | 29 | 89 | 1 | 0 | 88 | 0 | 0 |
| 2 | 391 | 2 | 0 | 207 | 1 | 0 | 30 | 168 | 1 | 1 | 172 | 0 | 0 |
| 3 | 774 | 1 | 0 | 185 | 1 | 0 | 31 | 116 | 0 | 0 | 61 | 0 | 0 |
| 4 | 213 | 0 | 0 | 109 | 1 | 0 | 32 | 1172 | 1 | 1 | 377 | 0 | 0 |
| 5 | 232 | 1 | 1 | 116 | 0 | 0 | 33 | 706 | 0 | 1 | 325 | 0 | 0 |
| 6 | 43 | 0 | 0 | 47 | 1 | 0 | 34 | 204 | 1 | 0 | 185 | 2 | 1 |
| 7 | 121 | 1 | 0 | 124 | 0 | 0 | 35 | 288 | 1 | 1 | 280 | 0 | 0 |
| 8 | 110 | 5 | 3 | 114 | 2 | 2 | 36 | 254 | 1 | 0 | 272 | 0 | 0 |
| 9 | 382 | 1 | 0 | 384 | 0 | 0 | 37 | 314 | 1 | 0 | 154 | 0 | 0 |
| 10 | 284 | 1 | 0 | 135 | 0 | 0 | 38 | 162 | 0 | 0 | 160 | 0 | 0 |
| 11 | 294 | 0 | 2 | 302 | 1 | 1 | 39 | 442 | 1 | 1 | 112 | 0 | 0 |
| 12 | 563 | 2 | 0 | 142 | 0 | 0 | 40 | 394 | 1 | 1 | 124 | 0 | 0 |
| 13 | 278 | 2 | 0 | 279 | 1 | 1 | 41 | 132 | 0 | 0 | 131 | 0 | 1 |
| 14 | 418 | 2 | 0 | 212 | 0 | 0 | 42 | 160 | 0 | 0 | 213 | 0 | 0 |
| 15 | 395 | 2 | 2 | 198 | 1 | 0 | 43 | 331 | 8 | 4 | 337 | 7 | 3 |
| 16 | 203 | 1 | 1 | 106 | 1 | 1 | 44 | 331 | 1 | 0 | 250 | 1 | 1 |
| 17 | 104 | 1 | 0 | 99 | 2 | 0 | 45 | 49 | 0 | 0 | 49 | 0 | 0 |
| 18 | 212 | 2 | 1 | 107 | 0 | 0 | 46 | 101 | 0 | 0 | 51 | 0 | 0 |
| 19 | 138 | 3 | 1 | 139 | 1 | 0 | 47 | 232 | 0 | 0 | 115 | 0 | 0 |
| 20 | 196 | 0 | 1 | 96 | 0 | 0 | 48 | 52 | 0 | 0 | 25 | 0 | 0 |
| 21 | 122 | 0 | 0 | 120 | 1 | 0 | 49 | 196 | 0 | 0 | 195 | 0 | 0 |
| 22 | 175 | 0 | 0 | 173 | 1 | 0 | 50 | 70 | 0 | 0 | 75 | 0 | 0 |
| 23 | 56 | 1 | 0 | 58 | 0 | 0 | 51 | 28 | 0 | 0 | 29 | 0 | 0 |
| 24 | 39 | 1 | 0 | 38 | 0 | 0 | 52 | 25 | 0 | 0 | 24 | 0 | 0 |
| 25 | 561 | 0 | 1 | 276 | 2 | 0 | 53 | 26 | 0 | 0 | 24 | 0 | 0 |
| 26 | 116 | 2 | 2 | 111 | 3 | 1 | 54 | 2635 | 15 | 12 | 2634 | 9 | 10 |
| 27 | 148 | 1 | 2 | 143 | 0 | 0 | 55 | 1456 | 27 | 2 | 2895 | 41 | 5 |
| 28 | 231 | 1 | 1 | 242 | 0 | 0 | 56 | 2220 | 64 | 60 | 2227 | 56 | 71 |

Table A.2: Data from a meta-analysis of 56 studies on the cardiovascular side effects of rosiglitazone [58].

# BIBLIOGRAPHY


[1] Agresti, A. (2013). *Categorical Data Analysis*. John Wiley & Sons., 3rd edition.

[2] Almalik, O. and van den Heuvel, E. R. (2018). Testing homogeneity of effect sizes in pooling 2x2 contingency tables from multiple studies: A comparison of methods. *Cogent Mathematics & Statistics*, page 1478698.

[3] Bai, O., Chen, M., and Wang, X. (2016). Bayesian estimation and testing in random effects meta-analysis of rare binary adverse events. *Statistics in Biopharmaceutical Research*, 8(1):49–59.

[4] Bakbergenuly, I. and Kulinskaya, E. (2017). Beta-binomial model for meta-analysis of odds ratios. *Statistics in medicine*, 36(11):1715–1734.

[5] Bartlett, M. S. (1935). Contingency table interactions. *Supplement to the Journal of the Royal Statistical Society*, 2(2):248–252.

[6] Bellamy, L., Casas, J.-P., Hingorani, A. D., and Williams, D. (2009). Type 2 diabetes mellitus after gestational diabetes: a systematic review and meta-analysis. *The Lancet*, 373(9677):1773–1779.

[7] Bhaumik, D. K., Amatya, A., Normand, S.-L. T., Greenhouse, J., Kaizar, E., Neelon, B., and Gibbons, R. D. (2012). Meta-analysis of rare binary adverse event data. *Journal of the American Statistical Association*, 107(498):555–567.

[8] Biggerstaff, B. and Tweedie, R. (1997). Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. *Statistics in medicine*, 16(7):753–768.

[9] Biggerstaff, B. J. and Jackson, D. (2008). The exact distribution of cochran's heterogeneity statistic in one-way random effects meta-analysis. *Statistics in medicine*, 27(29):6093–6110.

[10] Bliss, C. I. (1952). *The statistics of bioassay: with special reference to the vitamins*. Academic Press Inc.

[11] Breslow, N. E., Day, N. E., et al. (1980). *Statistical methods in cancer research. Vol. 1. The analysis of case-control studies.*, volume 1. Distributed for IARC by WHO, Geneva, Switzerland.

[12] Cai, T., Parast, L., and Ryan, L. (2010). Meta-analysis for rare events. *Statistics in medicine*, 29(20):2078–2089.

[13] Chen, M. and Lim, J. (2011). Estimating variances of strata in ranked set sampling. *Journal of Statistical Planning and Inference*, 141(8):2513–2518.

[14] Chung, Y., Rabe-Hesketh, S., and Choi, I.-H. (2013a). Avoiding zero between-study variance estimates in random-effects meta-analysis. *Statistics in medicine*, 32(23):4071–4089.

[15] Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A., and Liu, J. (2013b). A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. *Psychometrika*, 78(4):685–709.

[16] Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics*, 10(1):101–129.

[17] Crainiceanu, C. M. and Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):165–185.

[18] Crippa, A., Khudyakov, P., Wang, M., Orsini, N., and Spiegelman, D. (2016). A new measure of between-studies heterogeneity in meta-analysis. *Statistics in medicine*, 35(21):3661–3675.

[19] Dell, T. and Clutter, J. (1972). Ranked set sampling theory with order statistics background. *Biometrics*, pages 545–555.

[20] DerSimonian, R. and Kacker, R. (2007). Random-effects model for meta-analysis of clinical trials: an update. *Contemporary clinical trials*, 28(2):105–114.

[21] DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled clinical trials*, 7(3):177–188.

[22] Diamond, G. A., Bax, L., and Kaul, S. (2007). Uncertain effects of rosiglitazone on the risk for myocardial infarction and cardiovascular death. *Annals of Internal Medicine*, 147(8):578–581.

[23] Efron, B. and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical science*, pages 54–75.

[24] Farebrother, R. (1984). Algorithm as 204: the distribution of a positive linear combination of $\chi$ 2 random variables. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 33(3):332–339.

[25] Gart, J. J. (1966). Alternative analyses of contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 164–179.

[26] Gart, J. J. (1970). Point and interval estimation of the common odds ratio in the combination of 2$\times$ 2 tables with fixed marginals. *Biometrika*, 57(3):471–475.

[27] Halperin, M., Ware, J. H., Byar, D. P., Mantel, N., Brown, C. C., Koziol, J., Gail, M., and SYLVAN B, G. (1977). Testing for interaction in an i$\times$ j$\times$ k contingency table. *Biometrika*, 64(2):271–275.

[28] Hardy, R. J. and Thompson, S. G. (1996). A likelihood approach to meta-analysis with random effects. *Statistics in medicine*, 15(6):619–629.

[29] Hartung, J. and Knapp, G. (2005). On confidence intervals for the among-group variance in the one-way random effects model with unequal error variances. *Journal of statistical planning and inference*, 127(1-2):157–177.

[30] Hartung, J. and Makambi, K. (2002). Positive estimation of the between-study variance in meta-analysis: theory and methods. *South African Statistical Journal*, 36(1):55–76.

[31] Hedges, L. and Olkin, I. (1985). Statistical methods for meta-analysis. book, Orlando: Academic Press.

[32] Higgins, J. and Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in medicine*, 21(11):1539–1558.

[33] Higgins, J. P., Green, S., et al. (2008). Cochrane handbook for systematic reviews of interventions.

[34] Hunter, J. and Schmidt, F. (2004). Methods of meta-analysis: Correcting error and bias in research findings,(2004). *Mol Psychiatry*.

[35] Jackson, D. (2013). Confidence intervals for the between-study variance in random effects meta-analysis using generalised cochran heterogeneity statistics. *Research Synthesis Methods*, 4(3):220–229.

[36] Jackson, D. and Bowden, J. (2016). Confidence intervals for the between-study variance in random-effects meta-analysis using generalised heterogeneity statistics: should we use unequal tails? *BMC medical research methodology*, 16(1):118.

[37] Jackson, D., Bowden, J., and Baker, R. (2015). Approximate confidence intervals for moment-based estimators of the between-study variance in random effects meta-analysis. *Research synthesis methods*, 6(4):372–382.

[38] Jackson, D., Law, M., Stijnen, T., Viechtbauer, W., and White, I. R. (2018). A comparison of 7 random-effects models for meta-analyses that estimate the summary odds ratio. *Statistics in medicine*.

[39] Jackson, D., White, I. R., and Riley, R. D. (2012). Quantifying the impact of between-study heterogeneity in multivariate meta-analyses. *Statistics in medicine*, 31(29):3805–3820.

[40] Jones, M. P., O'Gorman, T. W., Lemke, J. H., and Woolson, R. F. (1989). A monte carlo investigation of homogeneity tests of the odds ratio under various sample size configurations. *Biometrics*, pages 171–181.

[41] Knapp, G., Biggerstaff, B. J., and Hartung, J. (2006). Assessing the amount of heterogeneity in random-effects meta-analysis. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 48(2):271–285.

[42] Knapp, G. and Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in medicine*, 22(17):2693–2710.

[43] 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.

[44] Kulinskaya, E. and Dollinger, M. B. (2015). An accurate test for homogeneity of odds ratios based on cochran's q-statistic. *BMC medical research methodology*, 15(1):49.

[45] Langan, D., Higgins, J., and Simmonds, M. (2017). Comparative performance of heterogeneity variance estimators in meta-analysis: a review of simulation studies. *Research synthesis methods*, 8(2):181–198.

[46] Langan, D., Higgins, J. P., Jackson, D., Bowden, J., Veroniki, A. A., Kontopantelis, E., Viechtbauer, W., and Simmonds, M. (2019). A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. *Research synthesis methods*, 10(1):83–98.

[47] Li, L. and Wang, X. (2019). Meta-analysis of rare binary events in treatment groups with unequal variability. *Statistical Methods in Medical Research*, 28(1):263–274.

[48] Liang, K. Y. and Self, S. G. (1985). Tests for homogeneity of odds ratio when the data are sparse. *Biometrika*, 72(2):353–358.

[49] Lin, L., Chu, H., and Hodges, J. S. (2017). Alternative measures of between-study heterogeneity in meta-analysis: Reducing the impact of outlying studies. *Biometrics*, 73(1):156–166.

[50] Lipsitz, S. R., Dear, K. B., Laird, N. M., and Molenberghs, G. (1998). Tests for homogeneity of the risk difference when data are sparse. *Biometrics*, pages 148–160.

[51] Liu, D., Liu, R. Y., and Xie, M.-g. (2014). Exact meta-analysis approach for discrete data and its application to 2x2 tables with rare events. *Journal of the American Statistical Association*, 109(508):1450–1465.

[52] Lohr, S. L. (2019). *Sampling: Design and Analysis: Design and Analysis*. Chapman and Hall/CRC.

[53] MacEachern, S. N., Öztürk, Ö., Wolfe, D. A., and Stark, G. V. (2002). A new ranked set sample estimator of variance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(2):177–188.

[54] Malzahn, U., Böhning, D., and Holling, H. (2000). Nonparametric estimation of heterogeneity variance for the standardised difference used in meta-analysis. *Biometrika*, 87(3):619–632.

[55] McIntyre, G. (1952). A method for unbiased selective sampling, using ranked sets.

[56] Morris, C. N. (1983). Parametric empirical bayes inference: theory and applications. *Journal of the American Statistical Association*, 78(381):47–55.

[57] Nissen, S. E. and Wolski, K. (2007). Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. *New England Journal of Medicine*, 356(24):2457–2471.

[58] Nissen, S. E. and Wolski, K. (2010). Rosiglitazone revisited: an updated meta-analysis of risk for myocardial infarction and cardiovascular mortality. *Archives of internal medicine*, 170(14):1191–1201.

[59] Norton, H. W. (1945). Calculation of chi-square for complex contingency tables. *Journal of the American Statistical Association*, 40(230):251–258.

[60] Novianti, P. W., Roes, K. C., and van der Tweel, I. (2014). Estimation of between-trial variance in sequential meta-analyses: a simulation study. *Contemporary clinical trials*, 37(1):129–138.

[61] Panityakul, T., Bumrungsup, C., and Knapp, G. (2013). On estimating residual heterogeneity in random-effects meta-regression: a comparative study. *J Stat Theory Appl*, 12(3):253.

[62] Paul, S. and Donner, A. (1989). A comparison of tests of homogeneity of odds ratios in k 2× 2 tables. *Statistics in Medicine*, 8(12):1455–1468.

[63] Paule, R. C. and Mandel, J. (1982). Consensus values and weighting factors. *Journal of Research of the National Bureau of Standards*, 87(5):377–385.

[64] Petropoulou, M. and Mavridis, D. (2017). A comparison of 20 heterogeneity variance estimators in statistical synthesis of results from studies: a simulation study. *Statistics in medicine*, 36(27):4266–4280.

[65] Philip, L., Lam, K., and Sinha, B. M. K. (1999). Estimation of normal variance based on balanced and unbalanced ranked set samples. *Environmental and Ecological Statistics*, 6(1):23–46.

[66] Reis, I. M., Hirji, K. F., and Afifi, A. A. (1999). Exact and asymptotic tests for homogeneity in several 2× 2 tables. *Statistics in medicine*, 18(8):893–906.

[67] Rukhin, A. L. (2013). Estimating heterogeneity variance in meta-analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(3):451–469.

[68] Self, S. G. and Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398):605–610.

[69] Sidik, K. and Jonkman, J. N. (2005). Simple heterogeneity variance estimation for meta-analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(2):367–384.

[70] Sidik, K. and Jonkman, J. N. (2007). A comparison of heterogeneity variance estimators in combining results of studies. *Statistics in medicine*, 26(9):1964–1981.

[71] Simmonds, M. C. and Higgins, J. P. (2016). A general framework for the use of logistic regression models in meta-analysis. *Statistical methods in medical research*, 25(6):2858–2877.

[72] Singh, S., Loke, Y. K., and Furberg, C. D. (2007). Long-term risk of cardiovascular events with rosiglitazone: a meta-analysis. *Jama*, 298(10):1189–1195.

[73] Sinha, B. K., Sinha, B. K., and Purkayastha, S. (1996). On some aspects of ranked set sampling for estimation of normal and exponential parameters. *Statistics & Risk Modeling*, 14(3):223–240.

[74] Smith, T. C., Spiegelhalter, D. J., and Thomas, A. (1995). Bayesian approaches to random-effects meta-analysis: A comparative study. *Statistics in medicine*, 14(24):2685–2699.

[75] Stokes, S. L. (1980). Estimation of variance using judgment ordered ranked set samples. *Biometrics*, pages 35–42.

[76] Takahasi, K. and Wakimoto, K. (1968). On unbiased estimates of the population mean based on the sample stratified by means of ordering. *Annals of the Institute of Statistical Mathematics*, 20(1):1–31.

[77] Takkouche, B., Cadarso-Suarez, C., and Spiegelman, D. (1999). Evaluation of old and new tests of heterogeneity in epidemiologic meta-analysis. *American journal of epidemiology*, 150(2):206–215.

[78] Tarone, R. E. (1985). On heterogeneity tests based on efficient scores. *Biometrika*, 72(1):91–95.

[79] Thompson, S. (2002). *Sampling*. New York: Wiley.

[80] Tian, L. (2008). Inferences about the between-study variance in meta-analysis with normally distributed outcomes. *Biometrical Journal*, 50(2):248–256.

[81] van Aert, R. C. and Jackson, D. (2018). Multistep estimators of the between-study variance: The relationship with the paule-mandel estimator. *Statistics in Medicine*.

[82] van Aert, R. C., van Assen, M. A., and Viechtbauer, W. (2018). Statistical properties of methods based on the q-statistic for constructing a confidence interval for the between-study variance in meta-analysis. *Research synthesis methods*.

[83] Van Houwelingen, H. C., Zwinderman, K. H., and Stijnen, T. (1993). A bivariate approach to meta-analysis. *Statistics in medicine*, 12(24):2273–2284.

[84] Verbeke, G. and Molenberghs, G. (2003). The use of score tests for inference on variance components. *Biometrics*, 59(2):254–262.

[85] Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., Kuss, O., Higgins, J., Langan, D., and Salanti, G. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research synthesis methods*, 7(1):55–79.

[86] Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics*, 30(3):261–293.

[87] Viechtbauer, W. (2007a). Confidence intervals for the amount of heterogeneity in meta-analysis. *Statistics in medicine*, 26(1):37–52.

[88] Viechtbauer, W. (2007b). Hypothesis tests for population heterogeneity in meta-analysis. *British Journal of Mathematical and Statistical Psychology*, 60(1):29–60.

[89] Wang, X., Ahn, S., and Lim, J. (2017). Unbalanced ranked set sampling in cluster randomized studies. *Journal of Statistical Planning and Inference*, 187:1–16.

[90] Wang, X., Lim, J., and Stokes, L. (2016). Using ranked set sampling with cluster randomized designs for improved inference on treatment effects. *Journal of the American Statistical Association*, 111(516):1576–1590.

[91] Woolf, B. (1955). On estimating the relation between blood group and disease. *Annals of human genetics*, 19(4):251–253.

[92] Yusuf, S., Peto, R., Lewis, J., Collins, R., and Sleight, P. (1985). Beta blockade during and after myocardial infarction: an overview of the randomized trials. *Progress in cardiovascular diseases*, 27(5):335–371.

[93] Zelen, M. (1971). The analysis of several 2× 2 contingency tables. *Biometrika*, 58(1):129–137.