UNIVERSITY OF SOUTHAMPTON

Estimating Heterogeneity Variance under Sparsity

by

Susan Martin

A thesis submitted in partial fulfilment for the degree of Doctor of Philosophy

in the

Faculty of Social Sciences

Mathematical Sciences

March 2020

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SOCIAL SCIENCES
MATHEMATICAL SCIENCES

Doctor of Philosophy

by Susan Martin

Meta-analysis has become the gold standard in medical research analysis. The random-effects model is generally the preferred method to conduct a meta-analysis, as it incorporates between-study heterogeneity - the variability between study estimates as a result of differences in study characteristics. Several methods to estimate the heterogeneity variance parameter in this model have been proposed, including the popular DerSimonian-Laird estimator, which has been shown to produce negatively biased estimates, performing well only in scenarios not seen in real-life data.

Many medical meta-analyses are concerned with rare-event data, where event probabilities are so low that often a small number or zero events are observed in the studies. Examples of this include adverse drug reactions in a clinical trial or very rare diseases in epidemiological studies, where as few as 1 in 1000 people may be affected by the outcome of interest. In such cases, most pre-proposed heterogeneity variance estimators perform poorly, and standard analysis techniques can result in the incorrect estimation of overall treatment effect.

In this thesis, we propose novel methods that we believe are appropriate for the estimation of heterogeneity variance in the case of rare-event data. These are based on generalised linear mixed models (GLMMs), and use the Poisson mixed regression model and the conditional logistic mixed regression model. We conducted a simulation study to compare our novel approaches with a selection of existing heterogeneity variance estimators for use in random-effect binary outcome meta-analyses.

From the results of our simulation study, which agree with results given in previous studies, we found that our novel GLMM-based estimating methods outperform existing methods in terms of the estimation of heterogeneity variance and summary log-risk ratio. This is the case when study sample sizes in the meta-analysis are balanced or unbalanced, and thus we recommend them for use with rare-event data.

Declaration of Authorship

I, **Susan Martin**, declare that the thesis entitled **Estimating Heterogeneity Variance under Sparsity** and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted the work of other, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- none of this work has been published before submission

Signed: .....

Date: .....

Contents

List of Figures

List of Tables

Acknowledgements

I would like to begin by thanking my friends and family, without whom this would never have been possible. To my family for your continued support, including the many care packages and notes of encouragement from Sally and the final read-through by Dad! And to Mike, who has been there through this all (and put up with it all!) This would not have been possible without your amazing support and advice (especially on the medical-related aspects) - I'm sorry you don't get a whole page!

I would also like to thank everyone in Mathematical Sciences and the S3RI who I had the pleasure of working with, for such a friendly working environment, and all the great opportunities in terms of teaching and improving equality within the department. In particular, the girls in the S3RI office, who were so friendly and helpful during our overlap. I'm also grateful for the use of the IRIDIS High Performance Computing Facility, and associated support services at the University of Southampton, in the completion of this work.

My thanks go to Professor Antonello Maruotti for his continued support and advice in relation to all things R-related, particularly his knowledge of GLMM packages, and for being an amazing tour guide and host. I would also like to thank Dr Stefanie Biedermann, Dr Alan Kimber, Dr Antony Overstall and Dr Robin Mitra for their direction via annual review meetings.

Finally, my greatest thanks go to my supervisor Professor Dankmar Böhning for being so supportive and patient with me! Thank you for offering me this research opportunity and for your continued guidance throughout my studies. Your help not only guided me in the right direction during my research, but it also made me a better researcher.

To Grandad

Your collection is now complete

Chapter 1

Introduction

1.1 Systematic reviews and meta-analyses

Clinical studies are essential in the advancement of understanding diseases and development of optimal treatments. Many studies are published that investigate the effect of the same treatment on a particular condition, but use different trial designs or volunteers of differing ethnicity or medical history. In order to maximise the understanding that can be drawn from these individual studies, their results need to be combined together in the form of a systematic review.

The aim of systematic reviews is to collect and synthesise all empirical evidence addressing a given research question using comprehensive procedures. The results found in such a review can be summarised using statistical methods via a meta-analysis to provide an integrative analysis of the quantities of interest. Meta-analyses are seen as the foundation of evidence-based medicine (Uman (2011)), as they can provide a broad overview of a given research question via the accumulation of a diverse base of clinical studies. In addition to providing a much cheaper alternative to a single new large clinical study, meta-analyses also have greater power to detect statistically significant results (Cohn and Becker (2003)).

Medical governing bodies such as the National Institute for Health and Care Excellence (NICE) set guidelines according to the results produced in published studies, with meta-analyses allocated the greatest weighting (National Institute for Clinical Excellence et al. (2005)), demonstrating their importance in healthcare systems. Examples of when meta-analyses of a number of significant trials have had a lasting impact on modern medicine include the implementation of the Bacillus Calmette-Guérin (BCG) vaccines in the prevention of tuberculosis (TB) (Colditz et al. (1994)), and the disproving of a link between the combined measles, mumps and rubella (MMR) vaccine and autism (Taylor et al. (2014)). Over the past few decades, the number of published meta-analyses has increased dramatically, as has the interest in their associated methodology.

1.1.1 Cochrane Database of Systematic Reviews

One of the major contributors to the rise in availability of meta-analyses is Cochrane, previously called the Cochrane Collaboration, which has made over 3000 systematic reviews openly available via the Cochrane Database of Systematic Reviews (CDS) (Handoll et al. (2008)). Their aim is to make information generated from a series of medical clinical studies more readily available to healthcare professionals and the general public. Cochrane also provides a source of advice regarding meta-analysis methodology in the form of the Cochrane handbook for systematic reviews of interventions (Higgins and Green (2011)). This handbook outlines currently approved meta-analysis methodology that can be used by researchers not familiar with meta-analyses or statistics, and as such is regularly updated to represent the most up-to-date guidelines.

1.1.2 Clinical study design

Clinical studies are comprised of both observational studies and clinical trials. They can be further grouped into a number of specific design types, characterised by the length the study, the volunteers used (called the study participants), and the outcome of interest to be measured, e.g. the odds or the risk. Below we shall outline some of the most common study design types used in the medical setting.

Clinical trials

The most common type of clinical trial are randomised controlled trials (RCT), which are generally performed to determine whether a novel treatment or intervention is effective in reducing the risk of developing some outcome or side-effect in individuals with the disease or condition of interest. Here, the participants are randomly assigned into two or more groups, including a 'treatment group' who are administered the new intervention under investigation, and a 'control group' who are given either some pre-existing intervention, a dummy intervention (called a placebo) or no intervention at all. The groups are then followed up over a pre-determined period of time, and the occurrence of the outcome of interest is recorded for each participant (Sibbald and Roland (1998)).

A clinical trial may be blinded (where the participants are not informed which intervention they have been given), or double-blinded (where the researchers are also unaware of this information) in order to reduce any potential bias, especially in the case of patient-reported outcomes (e.g. level of depression). RCTs are the trial of choice for testing the effectiveness of a new treatment with pre-existing treatments, and generally constitute part of phase III and phase IV clinical trials, which involve comparison with standard treatments and investigation of side effects, respectively.

Observational studies

There are a number of different types of observational study. Case-control, or case-referent, studies use participants who have some disease or medical condition of interest (the 'cases') and others who do not (the 'controls'), and measure the number from each group who have experienced some potential causal attribute (also called the exposure). Due to the nature of the study design, the incidence or risk cannot be calculated, and so in the case of binary outcomes, the odds must be used to measure the association between outcome and exposure (Lewallen and Courtright (1998)). Cases may be paired to controls in an effort to improve the efficiency of adjustments made to reduce any potential confounding.

Another form of longitudinal study is the cohort study, where one or more groups (called cohorts) are followed, and the status of some disease or outcome of interest is recorded in order to determine whether it is associated with any cohort-specific characteristic (Song and Chung (2010)). Cohorts are generated by grouping individuals according to characteristics such as age or ethnicity, and if some characteristic is found to be associated with the outcome under investigation, then it is termed as a risk factor. All participants must be free of the outcome of interest at the start of observation, and should be chosen to ensure the cohorts are comparable, especially for the non-exposure cohort. Where this is not possible, confounding factors need to be accounted for. The benefit of such studies is that causal relationships can be detected, which is not possible in cross-sectional studies and more difficult in case-control settings. This type of study can be either prospective or retrospective, however it has a number of disadvantages. For example, cohorts can be difficult to determine as a result of confounding variables, and participant characteristics may be imbalanced due to the lack of randomisation.

Observational studies do not need to be longitudinal in design, as is demonstrated with cross-sectional studies, which simultaneously measure the exposure and outcome of interest from a given population at a single point in time or over a short time interval. As a result of this, this type of study is cost-effective, and has the ability to investigate large populations. However, any associations must be interpreted with caution, as bias may occur from selection of the study population, and the study design can make it difficult to determine cause and effect (Setia (2016)).

1.1.3 Individual participant data vs. summary aggregated data

Meta-analyses can be conducted using either summary aggregated data (AD) or individual participant data (IPD), depending on what is available from the original studies. IPD meta-analyses can be more reliable than AD alternatives, as they use a greater depth of information. However, whereas AD is generally always available from the paper or authors, IPD is rarely given, to avoid potential identification of participants, and

can be difficult or impossible to obtain. In addition to this, both types of meta-analysis produce the same result in the the majority of cases, and so the further time and expense required to conduct IPD meta-analysis is unnecessary (Higgins and Green (2011)). The choice between these two types of meta-analysis can also depend on the type of investigation that is chosen. If the data was IPD, and a regression model was used to conduct the meta-analysis (as described in Section 1.4.3), then parameters specific to the individuals could also be included as covariates in the model.

1.2 Study-level data

The study-level data required for a meta-analysis generally consists of estimates for some parameter and its variance, denoted by $\hat{\theta}_i$ and $\hat{\sigma}_i^2$ respectively, for a given study i. This parameter, θ_i, is often referred to as the effect size, and in a medical setting, it often represents a measure of the difference between two groups, such as an active treatment or experimental group and a control or placebo group. A number of measures can be used to calculate $\hat{\theta}_i$, depending on the nature of the study outcome being investigated. Study outcomes can be grouped into several categories: continuous, binary (or dichotomous), ordinal, survival and repeated measures data. In the study types described in Section 1.1.2, continuous and binary outcomes are the most common, and so we shall focus on these here.

1.2.1 Continuous outcome data

Study-level data is described as being continuous if the variable to be measured is recorded on a numerical scale for each participant. Examples of continuous measures in the medical setting include weight, blood pressure and blood biomarkers. In general, the aim is to determine whether the levels of these variables change significantly between groups or cohorts, or after some intervention or treatment has been administered, and as such the difference in readings is of interest.

Mean difference

In the case of continuous data, results are usually presented as the mean measurement over all participants in each group, and the effect size is given as the difference in means between groups, which is called the mean difference (MD). If the study sample is skewed, then the median difference may instead be used, and if in addition the total number of participants (the sample size) is small (e.g. ≤ 20) then tests that involve ranking the measurements may be recommended (Takeshima et al. (2014)). The MD for a comparison between treatment and control in study i is defined as:

$$\hat{\theta}_i = MD_i = \text{Mean improvement with treatment} - \text{Mean improvement with control}$$

Standardised mean difference

The standardised mean difference (SMD), a variation of the MD that divides by the standard deviation of outcome among study participants, is generally the effect size of choice for continuous outcome data. This is because the pooled standard deviation adjusts for the precision and scale of measurement, as well as the study sample size (Faraone (2008)). The SMD for treatment versus control in study i is defined as:

$$\hat{\theta}_i = SMD_i = \frac{\text{Mean improvement with treatment} - \text{Mean improvement with control}}{\text{Pooled standard deviation}}$$

It should be noted that the pooling of standard deviations in this manner is only appropriate when the population variances are equal. When the inverse is true, we have the Behrens-Fisher problem (Walwyn and Roberts (2017)).

1.2.2 Binary outcome data

In studies with a binary outcome, data can be presented in the form of a contingency table, such as the one given in Table 1.1. From this, one can derive measures that compare the event probability between groups, such as the relative risk, odds ratio or risk difference. The outcomes measures that we shall outline here are the sample estimates from a study. Population versions of these measures also exist, e.g. population attributable risk, however we shall not focus on these as we are primarily interested in clinical study designs.

TABLE 1.1: Example of a 2×2 contingency table for study i with a binary outcome.

	Event	No event	Total
Treatment	a_i	b_i	$n_{it} = a_i + b_i$
Control	c_i	d_i	$n_{ic} = c_i + d_i$
Total	$a_i + c_i$	$b_i + d_i$	$N_i = a_i + b_i + c_i + d_i$

Relative risk

The relative risk (RR), or risk ratio, is the most popular and well-understood binary outcome measure with healthcare professionals and patients. It is defined as the comparison between the risk in the treatment group (a_i/n_{it}), and the risk in the control group (c_i/n_{ic}), and is thus given by $RR_i = \frac{a_i/n_{it}}{c_i/n_{ic}}$ for study i. As such, the RR of an

outcome represents the likelihood that it will occur after exposure to a novel treatment or risk factor, compared with its likelihood in a control setting (Andrade (2015)).

For meta-analyses, the RR is transformed onto the log-scale, as the $\log RR_i$ conforms to an approximately normal sampling distribution, an assumption of the standard meta-analysis model. If the estimate of the parameter of interest θ_i is the $\log RR_i$, then this, along with its variance $(\hat{\sigma}_i^2)$, could be calculated from the data in Table 1.1 as follows:

$$\hat{\theta}_i = \log RR_i = \log\left(\frac{a_i/n_{it}}{c_i/n_{ic}}\right)$$

$$\hat{\sigma}_i^2 = \frac{1}{a_i} - \frac{1}{n_{it}} + \frac{1}{c_i} - \frac{1}{n_{ic}}$$

Using the above equations, it is possible to construct confidence intervals (CI) that are symmetric around the $\log RR_i$. Once these have been calculated, the anti-log of the upper and lower CI bounds and overall $\log RR$ can be taken, in order to allow interpretation of the RR using the following definitions:

1. $RR = 1$ implies that the risk of an event occurring is equal across the treatment and control groups
2. $RR > 1$ implies that the risk of an event occurring in the treatment group is greater than that of the control group
3. $RR < 1$ implies that the risk of an event occurring in the treatment group is lower than that of the control group.

Odds ratio

The other major binary outcome measure used in medical meta-analyses is the odds ratio (OR). The odds of an outcome is the ratio of the likelihood of it occurring to the likelihood of it not occurring. As such, the OR is defined as the odds of the outcome occurring in the treatment group compared to the odds of it occurring in the control group (Andrade (2015)). Similar to the RR, it is transformed onto the log-scale for meta-analyses, and the $\log OR_i$ and its variance $(\hat{\sigma}_i^2)$ are given by:

$$\hat{\theta}_i = \log OR_i = \log\left(\frac{a_i/b_i}{c_i/d_i}\right)$$

$$\hat{\sigma}_i^2 = \frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i}$$

The OR tends to be more difficult to interpret, particularly for those in a healthcare setting, which can sometimes lead to incorrect conclusions and recommendations being made. As a result of this, it is advised that medical meta-analyses are conducted using the more intuitive RR. In addition to this, the OR an inferior measure of association compared to the RR, as it can only be used in case-control settings and logistic regression analyses. However, in the case of rare events, the RR and OR will be very similar (Szumilas (2010)). Interpretation of the OR is similar to that of the RR, with the definitions for OR in relation to 1 being equivalent to those of RR listed above, but with 'risk of event' replaced by 'odds of the outcome'.

Risk difference

An alternative, but less commonly used, binary outcome measure is the risk difference (RD), which is a measure of absolute effect difference that represents the absolute effect of the exposure of interest, or the excess risk of the outcome in the treatment group compared with the control group. The RD_i and its variance ($\hat{\sigma}_i^2$) are defined as:

$$\hat{\theta}_i = RD_i = \frac{a_i}{n_{it}} - \frac{c_i}{n_{ic}}$$

$$\hat{\sigma}_i^2 = \frac{a_i \times b_i}{n_{it}^3} + \frac{c_i \times d_i}{n_{ic}^3}$$

Benefits of the RD include that is easy to interpret and can be calculated when zero events are present, thus not requiring continuity corrections, unlike the RR and OR. However, the RD does not account for the treatment-specific underlying risk, and so this must be considered when interpreting the results (Egger et al. (1997)).

1.3 Between-study heterogeneity

There will always exist some degree of variability between study estimates in a meta-analysis as a result of the within-study sampling error. However, additional variability may occur as a result of differences in study characteristics such as study design, conduct, participants, and other known or unknown factors. This additional variability is called between-study variability (or variability due to heterogeneity), where heterogeneity represents the excess variation in observed treatment effects over that expected from the imprecision of results within each study (within-study variance). When measured as a variance, it is called heterogeneity variance and is denoted by τ^2 (Deeks et al. (2001)). Heterogeneity results from differences in the effects of the populations that the studies represent, and in the meta-analysis setting is the residual heterogeneity among study-specific intervention effects, or the dispersion in study-specific effect sizes.

1.3.1 Sources of heterogeneity

According to the Cochrane handbook, sources of heterogeneity can be split into two categories - those causing clinical heterogeneity and those resulting in methodological heterogeneity (Higgins and Green (2011)). Sources of clinical heterogeneity include variation in participants (e.g. ethnicity), outcomes and interventions (e.g. differences in dose administered), while methodological heterogeneity tends to be the result of variation in study design (e.g. length of follow-up time) and chance of risk (Thompson (1994)).

When conducting a meta-analysis, it is best to group together studies that adhere to certain study characteristic properties (e.g. cohort vs. case-control studies) in order to avoid confounding of the effect size estimate. However, it is not always possible to subset studies, particularly as potential sources of heterogeneity may be unreported and unknown. In these cases, it is essential to adjust for heterogeneity in the analysis itself.

1.3.2 Accounting for between-study heterogeneity

Heterogeneity can either be ignored or incorporated in the analysis via the use of fixed and random-effects models, respectively, and these are discussed in the next section. When it is incorporated in the analysis, then an estimate of τ^2 is required for the calculation of the overall treatment effect θ. If significant heterogeneity is present in a meta-analysis, but is not accounted for, then this can lead to over or under-estimation of the treatment effect, potentially leading to incorrect conclusions. In a medical setting, misleading inferences of drug performance in clinical trial results, in terms of either significance or direction of effect, could have disastrous repercussions.

1.4 Meta-analysis approaches

A number of approaches have been proposed to conduct meta-analyses. The most basic of these approaches, called marginal analysis, involves combining all study results together and calculating the effect size estimate as though the meta-analysis was a single study. For binary outcomes, this would involve merging all study results into a single 2×2 table, such as that seen in Table 1.1, and then calculating the pooled effect size using these total counts. Although this approach can allow for studies with zero counts (but not whole meta-analyses of zero counts) (Sweeting et al. (2004)), it is not recommended for use because it assumes that the event risk is constant across all studies, which is unlikely to be true, and so may produce inaccurate results (Simpson (1951); Altman and Deeks (2002)).

Approaches exist in two main classes: 2-stage (whereby the heterogeneity variance is estimated first, and then used to calculate an estimate of overall effect size measure)

and 1-stage (where the heterogeneity variance and effect size is estimated directly in one step). The majority of existing approaches are 2-stage as they include the use of a τ^2 estimator, however a number of 1-stage methods have recently been proposed, and these generally involve the use of regression models where the estimates of θ and τ^2 are simultaneously extracted as the model's parameters.

The two main approaches suggested for combining the study findings in a meta-analysis are the fixed-effect (FE) model and the random-effects (RE) model. In addition to these two models, meta-analyses can also be conducted by modelling with covariates via a random-effects meta-regression, where the covariates can then be investigated for being the source of any heterogeneity present, or through the use of Bayesian approaches to the above models (Smith et al. (1995); Sutton and Abrams (2001)). We shall outline some of these approaches in this section.

1.4.1 Fixed-effect model

The fixed-effect (FE) model is the simpler of the two main approaches, and assumes that study effects are homogeneous, and so the true effect is the same in all studies in the meta-analysis. As a result, all studies are assumed to be estimating a common effect size θ, and the model does not account for heterogeneity. As mentioned briefly in Section 1.2.2, models for conducting meta-analyses generally assume that the observed study effects in a meta-analysis will follow the normal distribution, as follows:

$$\hat{\theta}_i \sim N(\theta, \sigma_i^2)$$

where θ is the true overall effect size in the meta-analysis and σ_i^2 is the true variance in study i, with $i = 1, \ldots, k$. The FE model is often incorrectly used in meta-analyses by those who wrongly assume that there is no heterogeneity present in their data.

1.4.2 Random-effects model

In contrast to the FE model, the random-effects (RE) model allows the true effect to vary across studies, with the mean true effect being the parameter of interest. As such, analyses based on the RE model incorporate a between-study component of variance for the treatment effect in addition to the within-study variance, encompassing heterogeneity between studies. In this case, the assumed normal distribution is given by:

$$\hat{\theta}_i \sim N(\theta_i, \sigma_i^2)$$

Here the distribution of θ_i is also assumed to be normal, with mean θ and heterogeneity variance τ^2:

$$\theta_i \sim N(\theta, \tau^2)$$

If the heterogeneity variance is estimated to be zero, i.e. $\hat{\tau}^2 = 0$, then the RE model will simplify to a FE model. The RE model is the preferred model, and should be used in favour of the FE model if study results in the meta-analysis are suspected to be heterogeneous. Study heterogeneity is particularly common in medical settings, as even if the study designs of clinical trials are identical, it is likely that the participants may differ in terms of certain non-modifiable risk factors such as age, gender, ethnicity, etc., resulting in variation between the clinical trials in the meta-analysis.

The RE model can also result in wider confidence intervals for the associated overall effect size, leading to more conservative outcomes being produced. However, despite all of these positive attributes, the RE model also has several disadvantages. For example, publication bias is often present (although difficult to detect) in meta-analyses, and such bias can violate the normality assumption regarding the distribution of study effect sizes.

1.4.3 Meta-regression

An alternative approach to conducting meta-analyses is to use a meta-regression. This is similar to simple regression models, but in this case the outcome is the effect size estimate (e.g log odds ratio), and the explanatory variables in the model represent study characteristics that may affect this effect size. These covariates allow for the incorporation of elements that may affect the outcome measure and thus should be incorporated into its estimation, which is one of the major benefits of meta-regression (Thompson and Higgins (2002)).

Studies are weighted using the standard error of their respective outcome measure estimates, giving larger studies greater influence on the overall estimate. Between-study heterogeneity not explained by the covariates can also be incorporated into the regression model, giving random-effects meta-regression (Thompson and Sharp (1999)). A disadvantage of this approach is that it is not recommended for sparse data, as it performs poorly in the case of few studies, and as a result should not be considered when the meta-analysis contains fewer than 10 studies.

1.5 The inverse-variance approach

One of the main aims of a meta-analysis is to combine studies and produce an estimate for θ. Studies typically vary in terms of size, and assuming that all studies are of

equal quality, larger studies tend to estimate the parameter with more precision. The inverse-variance method is commonly used to combine studies in a meta-analysis, as it gives more precise studies a larger weighting and thus more influence on the effect size estimate. Using this method, the effect size θ and its variance can be estimated by:

$$\hat{\theta} = \frac{\sum_{i=1}^{k} w_i \hat{\theta}_i}{\sum_{i=1}^{k} w_i}$$

$$Var(\hat{\theta}) = \frac{1}{\left(\sum_{i=1}^{k} w_i\right)^2} \sum_{i=1}^{k} w_i^2 \sigma_i^2 = \frac{1}{\sum_{i=1}^{k} w_i}$$

where the study weights, denoted by w_i, are calculated by the reciprocal of $Var(\hat{\theta}_i)$, e.g. $w_i = 1/\sigma_i^2$, and k is the number of studies in the meta-analysis. If we assume a common effect, like the FE model, the within-study variance is assumed to account for all the variability of $\hat{\theta}_i$, and therefore $w_{i,FE} = 1/\sigma_i^2$. However, if we allow for random effects, like in the RE model, then $w_{i,RE} = 1/(\sigma_i^2 + \tau^2)$.

1.5.1 Disadvantages of the inverse-variance approach

The inverse-variance method can encounter problems when used alongside the RE model, as if the heterogeneity variance in the meta-analysis is significantly large, then studies with small and large sample sizes may be allocated very similar weights. As a result of this, studies with few participants may be given a relatively large weight, disproportionate to their sample size. This can lead to incorrect inferences of the summary effect size, and the problem will worsen as the difference between study sample sizes within the meta-analysis widens.

1.6 Confidence intervals for the summary effect

When conducting a meta-analysis, in addition to generating a point estimate of the outcome measure of interest, it is also necessary to produce an associated confidence interval to represent the uncertainty in this estimate. Various methods exist for calculating the confidence interval of the summary effect measure, which we shall discuss below. The performances of confidence intervals are generally compared in terms of coverage, where confidence intervals with higher coverage generally produce wide intervals. The performance of the confidence, or credible (if Bayesian approaches were used), intervals can be indicative of the overall performance of the approach used to conduct the meta-analysis and produce the associated summary effect point estimate. As such, it can be useful to compare these intervals when comparing a range of meta-analysis approaches, as specific

pairings of approaches and intervals may perform better than others, while others should not be used together in certain circumstances.

1.6.1 Wald-type method

The most commonly used confidence interval in meta-analysis is the Wald-type method, which is based on the assumption of the Normal distribution (DerSimonian and Laird (1986)). The endpoints for the summary effect interval are calculated as follows:

$$\hat{\theta} \pm z_{(1-\alpha/2)} \sqrt{\widehat{Var}(\hat{\theta})} \tag{1.6.1}$$

where $\hat{\theta}$ is the estimate of the summary effect, α is the coverage level of the confidence interval, $z_{(1-\alpha/2)}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution, $\widehat{Var}(\hat{\theta}) = 1/\sum_{i=1}^{k} 1/\hat{w}_i$ and $\hat{w}_i$ is the weight estimate appropriate for the type of model used (i.e. fixed or random-effects) as discussed in Section 1.5. For the fixed-effects model, $\hat{w}_{i,FE} = 1/\hat{\sigma}_i^2$, however if the random-effects weight $\hat{w}_{i,RE} = 1/(\hat{\sigma}_i^2 + \hat{\tau}^2)$ is used, then the estimate of $\hat{\tau}^2$ comes from using any of the estimators that we shall discuss later on.

1.6.2 t-distribution method

Another commonly used confidence interval is based on the use of the t-distribution with $k-1$ degrees of freedom (Follmann and Proschan (1999)), where k is the number of studies in the meta-analysis. The confidence interval in this case is given by:

$$\hat{\theta} \pm t_{k-1,(1-\alpha/2)} \sqrt{\widehat{Var}(\hat{\theta})}$$

where $t_{(k-1),(1-\alpha/2)}$ is the $(1-\alpha/2)$-quantile of the Student-t distribution with $(k-1)$ degrees of freedom, and all other values are as given in Equation (1.6.1).

1.6.3 Hartung-Knapp-Sidik-Jonkman method

As the previous two methods had been shown to perform poorly in terms of coverage in a number of meta-analysis scenarios, Hartung and Knapp (2001) and Sidik and Jonkman (2002) proposed an alternate interval that is equivalent to the t-distribution method, however its variance is multiplied by a scaling factor (q) (Wiksten et al. (2016)). This method is known to produce wider intervals than those based on the normal approximation, since the Student-t quantile is greater than the associated normal quantile. However, it can also produce narrower confidence intervals when the scaling factor is less than 1 (Higgins and Thompson (2002)).

$$\hat{\theta} \pm t_{k-1,(1-\alpha/2)} \sqrt{q \times \widehat{Var}(\hat{\theta})} \tag{1.6.2}$$

where $q = \frac{1}{k-1} \sum_{i=1}^{k} (\hat{\theta}_i - \hat{\theta})^2 / (\hat{\sigma}_i^2 + \hat{\tau}^2)$, thus giving a scaled value of the variance of $\hat{\theta}$ that is derived from a non-negative, unbiased estimator of $1/\sum_{i=1}^{k} 1/w_i$. This method was found to perform better than existing confidence intervals, especially when combined with the DerSimonian-Laird heterogeneity variance estimator, particularly when τ^2 was non-zero and there were few studies present. IntHout et al. (2014) noted that confidence intervals based on the normal approximation can easily be converted into HKSJ-adjusted results.

1.6.4 Modified Knapp-Hartung method

The HKSJ confidence interval was observed to be shorter than the original t-distribution approach when q was arbitrarily small (in fact if $\sqrt{q} < \frac{z_{(1-\alpha/2)}}{t_{(k-1,1-\alpha/2)}}$), which contrasts the improvement in coverage this adjustment was proposed to provide. As a result, a modified version of this approach has also been proposed, which involves a simple *ad hoc* modification to the scaling factor q, namely:

$$q^* = max\{1, q\}$$

(Knapp and Hartung (2003)). By using q^* as the scaling factor in Equation (1.6.2) instead of q, this ensures that the value to always be greater than or equal to 1, and thus avoids the consequent narrowing of intervals, ensuring a more conservative approach (Röver et al. (2015)).

1.7 Measuring heterogeneity

The outline of the RE model in Section 1.4.2 demonstrates the importance of estimating the heterogeneity variance in the correct determination of the overall effect size in a meta-analysis based on the RE model. It can be useful to produce a measure of heterogeneity when collating results in a meta-analysis, thus determining what proportion of the variation in the meta-analysis was the result of between-study variability. In this section, we shall discuss several methods used to conduct both of these tasks, along with their associated problems.

1.7.1 Heterogeneity variance estimators

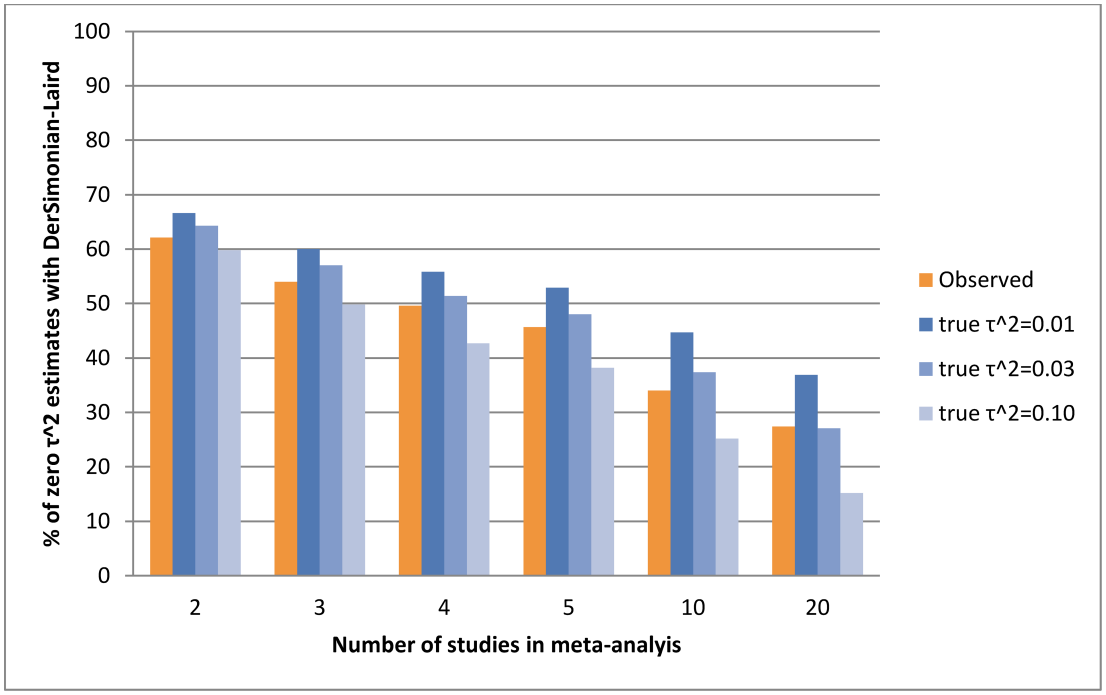
The heterogeneity variance estimate ($\hat{\tau}^2$) is commonly associated with substantial uncertainty, especially in contexts where there are few studies available, such as in small population and rare disease research. For example, simulation studies have shown that estimates of τ^2 are particularly inaccurate when the number of studies (k) in the meta-analysis is small, i.e. < 5 studies (Friede et al. (2017a)). Also, although heterogeneity will always be present to some degree, it tends to be more pronounced in situations where there are few studies or where there are few participants in these studies. Similarly, it it likely to be magnified in situations with large numbers of studies, as there is likely to a higher number of outliers present. If incorrect zero estimates of τ^2 are produced, then this can result in overly optimistic estimates of the summary effect.

A number of methods have been developed to estimate τ^2, including both classical and Bayesian approaches. The methods can be divided into two main groups: closed-form (or non-iterative) methods, and iterative methods. We shall outline some of the main groups of methods below. A wide range of τ^2 estimators are available for use in meta-analyses with STATA's updated *admetan* package.

Method of moments approach

Methods of moments approaches are based on Cochran's Q test statistic for heterogeneity, and can be split into those that are truncated to zero when negative, and those that produce only non-negative estimates by design. An example of a method of moments approach is the non-iterative DerSimonian-Laird (DL) estimator (DerSimonian and Laird (1986)), which is the most commonly used estimator and the default method in many statistical software packages. However, the use of the DL method has been questioned, as the type I error rate (the probability of rejecting a true null hypothesis) of this estimator can be extremely high, unless certain conditions are met. These conditions include the number of studies being large ($k \gg 20$), and no or very little heterogeneity being present - situations that rarely occur in real data (IntHout et al. (2014)).

The DL estimator also has a tendency to generate zero estimates for τ^2, even when the true value differs considerably from zero. Figure 1.1, which is taken from the study by Kontopantelis et al. (2013), demonstrates this poor characteristic of the DL method, for both the Cochrane and simulated data over various values of k. The bar chart shows that the estimator produces inaccurate zero estimates at least 60% of the time when $k = 2$, and although this decreases as k increases, scenarios with small values of τ^2 still results in zero values for over half the estimates when $k = 5$.



*Normal distribution of the effects assumed in the simulations (more extreme distributions produced similar results).

FIGURE 1.1: Bar chart from Kontopantelis et al. (2013) displaying the percentages of zero τ^2 estimates with DerSimonian-Laird method for meta-analyses in the Cochrane library and simulated data, for varying numbers of studies.

Bayesian approach

A number of Bayesian-based approaches have been produced, and are a popular choice for the meta-analysis of few studies ($k < 5$) and rare events, a scenario where most existing estimators perform very poorly and are not recommended (Günhan et al. (2018)). Their advantages involve the choice of priors and prior distributions, which can be crucial in the optimal estimation of parameters, particularly when rare events are involved.

Other estimators

Other frequentist methods include those based on the maximum likelihood-based approach, and the independent Hunter-Schmidt approach. These methods, along with the approaches outlined above, will be described in more detail in Chapter 2.

1.7.2 The Q statistic

As mentioned previously, the method of moments τ^2 estimators, including the DL method, are based on the Q-statistic:

$$Q = \sum_{i=1}^{k} w_{i,FE}(\hat{\theta}_i - \hat{\theta})^2$$

Q itself can also be used as a test statistic for the presence of heterogeneity, where a p-value can be derived by comparing Q to the χ^2-distribution with $k-1$ degrees of freedom. If the sampling variances σ_i^2 adequately account for the total observed variance, and $\hat{\theta}_i$ are normally distributed around θ_i, then $E(Q) = k-1$, i.e. the expected value of χ^2_{k-1}.

DerSimonian-Laird method

To demonstrate the use of Q in the formation of τ^2 estimators, we give the DL method:

$$\hat{\tau}^2_{DL} = max\left\{0, \frac{\sum_{i=1}^{k}\hat{Q} - (k-1)}{\sum_{i=1}^{k}(1/\hat{\sigma}_i^2) - \frac{\sum_{i=1}^{k}(1/\hat{\sigma}_i^2)^2}{\sum_{i=1}^{k}(1/\hat{\sigma}_i^2)}}\right\}$$

where $\hat{Q}$ is the above-defined Q-statistic but with weights generated from estimates of the within-study variances, i.e. $\hat{w}_{i,FE} = 1/\hat{\sigma}_i^2$. In this estimator, the numerator measures the extent to which $\hat{Q}$ exceeds its expected value $(k-1)$ under the assumption of a fixed effect. The denominator then converts the estimator to the same scale as the estimated overall effect size $\hat{\theta}$. When $\hat{Q} < k-1$, the estimator is truncated at zero, as displayed in the above equation.

1.7.3 The I^2 statistic

The I^2-statistic is the most common measure of heterogeneity present in a meta-analysis. It represents the percentage of total variation that is due to heterogeneity between study effects, and is produced by transforming the Q-statistic (Higgins and Thompson (2002)), as follows:

$$I^2 = max\left\{0, \frac{Q-(k-1)}{Q} \times 100\%\right\} \tag{1.7.1}$$

Compared to τ^2, the I^2-statistic is much easier for non-statisticians to interpret, and is also independent of the type of outcome measure used (i.e. I^2 can be compared between meta-analyses with binary and continuous outcomes). An alternative formula for I^2 is given by:

$$I^2 = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \hat{\sigma}^2} \times 100\% \tag{1.7.2}$$

where $\hat{\sigma}^2$ represents the estimate of an overall within-study variance, which would be the case if all studies had equal $\hat{\sigma}_i^2$, and can be calculated using:

$$\hat{\sigma}^2 = \frac{(k-1)\sum_{i=1}^{k} \hat{w}_i}{\left(\sum_{i=1}^{k} \hat{w}_i\right)^2 - \sum_{i=1}^{k} \hat{w}_i^2}$$

where $\hat{w}_i = 1/\hat{\sigma}_i^2$ (Higgins and Thompson (2002)). The issue with this approach for calculating I^2 is that it assumes a common within-study variance, when in reality, the within-study variance σ_i^2 will vary between studies (Borenstein et al. (2010)). As a result of this, Equation (1.7.1) is generally used in preference to Equation (1.7.2) when calculating I^2.

1.8 Forest plots

Forest plots, also called blobbograms, graphically summarise the outcome of a meta-analysis and characteristics of the contributing studies. Figure 1.2 displays a forest plot for a meta-analysis on the effect of the BCG vaccine on incidence of TB, as mentioned in Section 1.1. The upper section of the plot displays information on the studies included in the meta-analysis, giving the number of events and sample size for each intervention group, although this information may not always be displayed. The study-specific effect sizes (in this case the risk ratio) and associated 95% Wald-type confidence intervals are plotted, with the size of the boxes representing the size of the studies (and associated weight if the inverse-variance approach is being used), and given on the right-hand side. In the lower section of the plot, the overall effect size estimates are plotted and displayed, along with their associated p-values testing whether the meta-analysis provides significant evidence of favouring either the control or treatment. In this case, the BCG vaccine is found to be significantly favoured over the control, for both fixed and random-effects approaches.

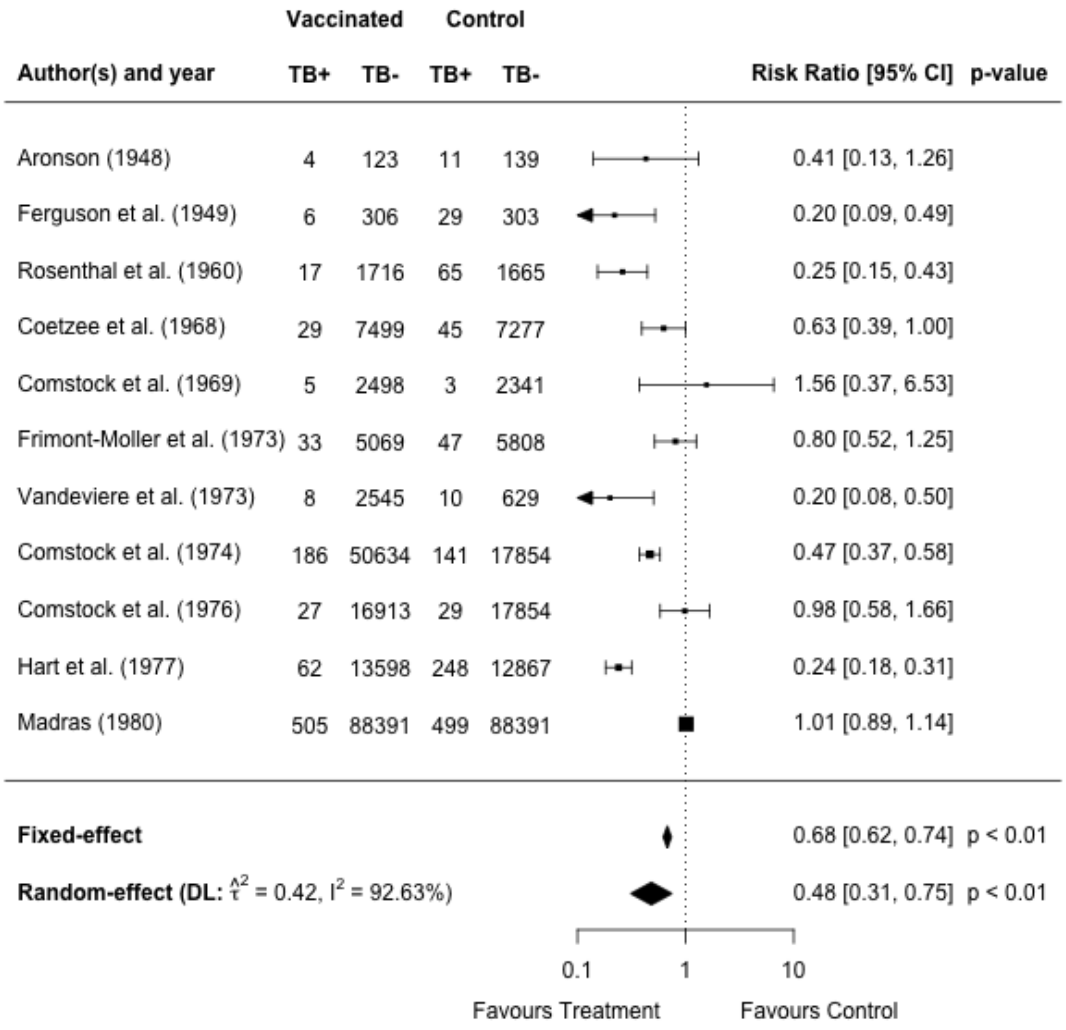


FIGURE 1.2: Forest plot displaying inverse-variance weighted fixed-effect and random-effects meta-analyses of the effect of the BCG vaccine on incidence of TB. For the random-effects approach, the DerSimonian-Laird (DL) τ^2 estimator is used. Wald-type confidence intervals are displayed here.

1.9 Rare-event data

Many medical research datasets tend to involve rare events, where the event occurrence probability is so low that frequently a small number or no events are observed in a clinical trial, despite the fact that either the trial arm sizes or the observation times are not small. This is different from sparse data, where trial sizes are small (often for reasons of patient recruitment) but event probabilities may not necessarily be small (Böhning et al. (2015)).

The Cochrane handbook states in its guidelines that risks of 1 in 1000 would be classed as 'rare events', and that risks as high as 1 in 100, and even 1 in 10, may also be classed as such (Higgins and Green (2011)). They also state that a common feature of rare-event data in a meta-analysis is the presence of zero events of interest in one trial arm (defined as single-zero trials) and the presence of zero events in both trial arms (double-zero trials). Estimators of τ^2 tend to have negative bias in scenarios where both heterogeneity and rare events are present.

1.9.1 Continuity corrections

If there are zero events present in either arm of study i, then for the RR and OR binary outcome measures θ_i and σ_i^2 cannot be estimated. In this scenario, a continuity correction is added to all event counts and sample sizes of studies including zero event counts.

Constant continuity correction

The standard continuity correction involves adding a constant k to all event counts in Table 1.1 (a_i, b_i, c_i, d_i), as proposed by Jewell and Holford (2005). The default choice for this value of k is 0.5, most likely a result of Cox (2018) finding a correction factor of 0.5 to have the least bias when applied to a single arm log-odds analysis. However, constant corrections in general (including $k = 0.5$) have been shown to often add unwanted bias and adversely affect CI coverage, thus causing meta-analysis methods that incorporate such study-level corrections to perform poorly (Bradburn et al. (2007)). As a result, alternative correction factors have been sought.

Reciprocal of the opposing trial arm's sample size

Sweeting et al. (2004) proposed two alternative continuity corrections - the first of these involved using a factor of the reciprocal of the sample size of the alternate treatment group (k/n_{it} and k/n_{ic} for the control and treatment group respectively), where k is the chosen proportionality constant in this case. They believed that this adjustment to the constant correction may result in less bias when the sample sizes of the trial arms were severely unbalanced. In fact, they found this correction to outperform the constant correction for all degrees of sample size imbalance for meta-analyses performed using the Mantel-Haenszel, Peto's, Bayesian and regression-based approaches, but performed similarly poorly for the popular inverse-variance approach.

Empirical continuity correction

The second correction that Sweeting et al. (2004) proposed involves using an empirical estimate of the pooled effect sizes from the remaining non-zero studies in the meta-analysis. It is based on noting that the reciprocal-based correction has a tendency to arbitrarily pull the estimate of the effect size towards that of no effect, and believing that

instead using a correction factor that pulls the estimate towards the pooled effect size using non-zero studies by design would be more preferable. As such, this novel empirical approach uses a correction factor based on the pooled effect estimate from non-zero studies in the meta-analysis, as such acting as prior (that is empirically derived from alternate studies) would in a Bayesian setting. Suppose that an estimate of the pooled odds ratio was derived from the non-zero studies, and is given by $\widehat{OR}$, and that the ratio imbalance between the control and treatment groups is $R = n_c/n_t$, meaning that the control arm sample size can be rewritten as $n_c = n_t \times R$. The empirical correction factors for the treatment and control arms respectively, k_t and k_c, must then satisfy:

$$\frac{k_t(n_t R + k_c)}{k_c(n_t + k_t)} = \widehat{OR} \tag{1.9.1}$$

When the arm-specific study sample sizes are large enough, then the left-hand side of Equation (1.9.1) can be approximated by Rk_t/k_c. Restrictions for the summation of k_t and k_c must be set into place, for example $k_t + k_c = 1$, as is the case with the constant correction of 0.5, in which case we have:

$$\frac{R(1 - k_c)}{k_c} \approx \widehat{OR}$$

and the following empirical corrections are produced:

$$k_c \approx \frac{R}{R + \widehat{OR}}$$

$$k_t \approx \frac{\widehat{OR}}{R + \widehat{OR}}$$

If all studies within the meta-analysis contain at least one zero event trial arm, then the estimate of the pooled effect size is undefined, and an alternative estimate must be defined by the user, using prior information known regarding the intervention under investigation. This method is not recommended for use with risk ratio meta-analyses. As with the reciprocal correction, this method was found to outperform the constant correction factor for all meta-analysis approaches other than the inverse-variance approach. In a simulation study focusing on unbalanced meta-analyses consisting of studies with small and large sample sizes (typical of observational studies), and containing 10 studies with an event rate of 1%, they found that both novel corrections outperformed the constant 0.5 correction. As a result, in sparse-event scenarios it has been recommended to use such alternative corrections, or meta-analysis approaches that can manage zero count data, which we shall discuss later in the chapter.

Both of the novel estimators proposed by Sweeting et al. (2004) are available to use in STATA's *admetan* package for conducting meta-analyses (an update of the popular metan package).

1.9.2 Reported characteristics of meta-analyses

According to Langan (2015), the most prevelant outcome measure recorded in the 2008 version of the CDS was the risk ratio, as they found that this was used for 43% of meta-analyses, excluding those with fewer than 3 studies. They also found that the majority (85%) of meta-analysis contained fewer than 10 studies. Mallett and Clarke (2002) stated that a typical review from the 2001 CDS contained 6 studies, with the maximum number of trials observed being 136. Based on a sample of 258 meta-analyses, they also found that median sample size per trial was 118, with the median total sample size for thew meta-analysis being 945.

1.9.3 Exclusion of double-zero trials

It is crucial to not simply exclude single-zero or double-zero trials when conducting meta-analyses, as they provide a lot of information in terms of the rarity of the outcome in question, and as such should be incorporated into the final summary effect estimate. This argument against exclusion of such data has been discussed previously (Whitehead and Whitehead (1991)). Bhaumik et al. (2012) and Sweeting et al. (2004) both investigated the output of meta-analyses where double-zero studies were excluded, and found that, despite τ^2 estimates being less biased, the summary effect estimates were considerably more biased when the double-zero trials were excluded - providing more evidence for their inclusion.

1.9.4 Rare-event data in a medical setting

In medical data, rare events tend to occur in the form of rare disease occurrences in epidemiological studies and adverse drug reactions in clinical trials. Very rare diseases are often known as orphan diseases, and these are defined as having a prevalence of ≤ 5 in 10,000 (Röver et al. (2015)).

One of the best known examples of rare-event data in a meta-analysis is the effect of rosiglitazone, a drug for the management of type 2 diabetes, with respect to the risk of myocardial infarction and cardiovascular-related death. This dataset contains several single and double-zero trials, and the meta-analysis has been investigated twice by Nissen and Wolski (2007, 2010). Several other studies have also looked at this dataset, and have obtained different results for the effectiveness of this drug depending on the

type of analysis method used (Cai et al. (2010); Diamond et al. (2007)), a worrying outcome in such a serious subject area.

1.9.5 Techniques for the analysis of rare-event data

Analysing rare events requires specialised statistical techniques since common methods such as linear and logistic regression are inappropriate. This is because they can dramatically underestimate the probability of rare events, resulting in the incorrect estimation of the treatment effect (Cai et al. (2010)). Some statistical models have been proposed for use with such sparse-event data, and we shall outline these below.

Mantel-Haenszel approach

The Mantel-Haenszel method is the most popular FE method to produce an estimate of a weighted binary outcome measure, and works by using alternative weights that depend on the outcome measure of the meta-analysis (Mantel (1963)). It has notable beneficial statistical properties in the cases of rare-event data and small sample sizes, where the inverse-variance method performs poorly in estimating the standard errors of the effect sizes (Deeks et al. (2001)). In the Cochrane handbook, the Mantel-Haenszel method is recommended over the inverse-variance method in the case of sparse-event data, however it is noted that the two methods give similar estimates in other scenarios (Higgins and Green (2011)).

The Mantel-Haenszel summary relative risk (RR_{MH}) and odds ratio (OR_{MH}) estimates are calculated as follows:

$$\widehat{RR}_{MH} = \frac{\sum_{i=1}^{k}(a_i n_{ic})/N_i}{\sum_{i=1}^{k}(c_i n_{it})/N_i} \tag{1.9.2}$$

$$\widehat{OR}_{MH} = \frac{\sum_{i=1}^{k}(a_i d_i)/N_i}{\sum_{i=1}^{k}(c_i b_i)/N_i}$$

The major downfall of the Mantel-Haenszel method is that it is a FE-based method, and so does not account for heterogeneity, which is likely to be more pronounced when sparsity is present in the data. In addition to this, this approach is undefined for $\log RR_{MH}$ and $\log OR_{MH}$ when events in all study treatment arms are zero or events in all study control arms are zero. In this case, a continuity correction can be added to all event counts and sample sizes across the meta-analysis, such as a constant of 0.5, however there is little literature to support the use of a correction for the Mantel-Haenszel method, and so its appropriateness is unknown.

Peto's method

Peto's method (or Peto's odds ratio) is designed specifically for use with the odds ratio, as its name suggests (Peto and Peto (1972)). It is based on the inverse-variance approach, but uses different weights to estimate the odds ratio. The approach can be viewed as a sum of the observed − expected statistics, where in this case 'observed' represents the observed count of events in the treatment group of each study, and 'expected' is the respective expected count of events. The formula for Peto's odds ratio is given by:

$$\widehat{OR}_{i,Peto} = \exp\left(\frac{O_i - E_i}{V_i}\right)$$

where $O_i = a_i$, $E_i = \frac{(a_i+b_i)(a_i+c_i)}{N_i}$, $V_i = \frac{(a_i+b_i)(c_i+d_i)(a_i+c_i)(b_i+d_i)}{N_i^2(N_i-1)}$ and $i = 1, \ldots, k$. When using this method, continuity corrections for single-zero studies are not required, and so their associated bias is avoided. However, for a double-zero study i, the values O_i, E_i and V_i are all zero by construction, and so such studies would not contribute to the pooled odds ratio estimate or variance.

As with the Mantel-Haenszel method, this approach is based on the FE model, and so does not account for the heterogeneity that is likely to be present. In addition to this, Peto's method has been shown to cause bias itself, particularly in the case where study treatment and control group sizes differ significantly, and when the true odds ratio is far from one (Greenland and Salvan (1990)). However, Deeks et al. (1999) found that Peto's method performed best in terms of bias and power, when compared to alternate approaches, in a simulation study focusing on rare-event data with little sample size imbalance (typical of RCTs).

Regression modelling techniques

One of the most recently proposed, and as a result, least investigated, approaches to dealing with rare-event data in a meta-analysis involves the use of mixed regression models (Stijnen et al. (2010)). This method allows for the inclusion of covariates of interest as fixed or random-effects, and provides a more satisfying approach for dealing with study heterogeneity. It also allows for the easy estimation of the outcome measure of interest and the heterogeneity variance through parameters in the associated model.

Currently proposed suggestions for the models in this approach include the Poisson mixed regression model (Böhning et al. (2015)), originally considered by Deeks et al. (1999), and the conditional logistic mixed regression model (Stijnen et al. (2010)). A more in-depth discussion of these models and their application shall be presented in Chapter 4.

1.10 Techniques used in the Cochrane library

Figure 1.3 (taken from the study by Kontopantelis et al. (2013)) displays the types of meta-analysis methods used within the Cochrane library for various k. They have divided the methods into the rare-event Peto's and Mantel-Haenszel methods discussed in Section 1.9.5 (including a random-effects variation of the Mantel-Haneszel approach), and the FE and RE standard inverse-variance methods. It can be seen that the method most frequently used is the FE Mantel-Haenszel approach, with the FE inverse-variance method coming second. This demonstrates how the FE approaches are much more widely used, potentially inappropriately at times, and this is likely due to its default application in many statistical packages. This plots also demonstrates that a considerable number of meta-analyses in the Cochrane review contain few studies, with approximately 19,000 containing only 2 studies, while only around 7000 meta-analyses have 10 or more.

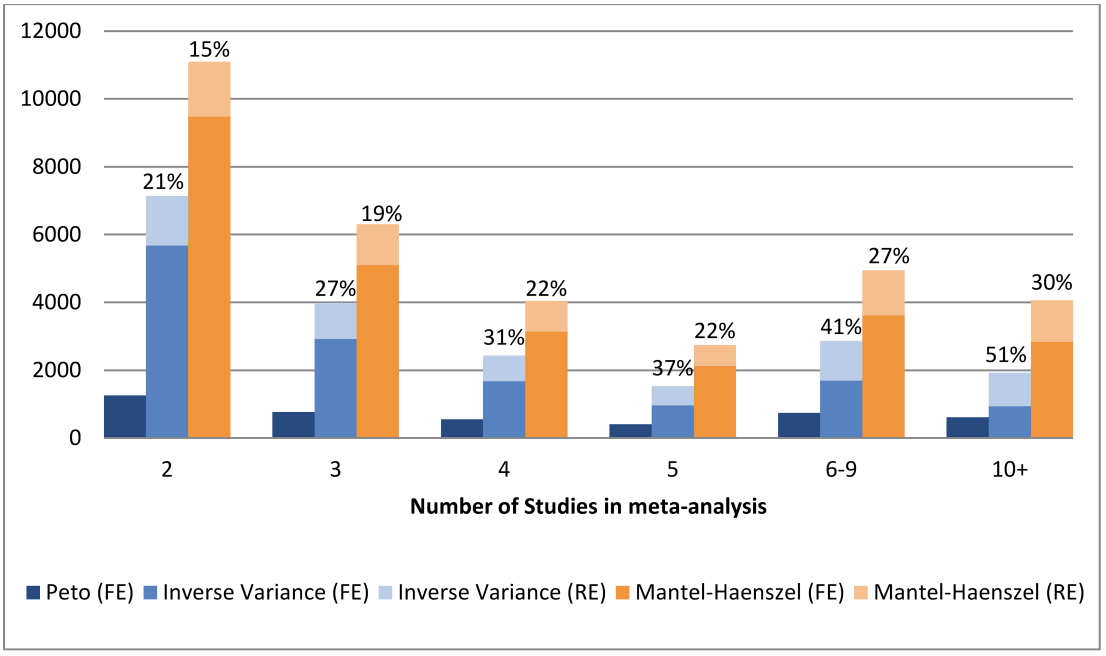


FIGURE 1.3: Bar chart from Kontopantelis et al. (2013) displaying the counts and percentages of types of methods used for meta-analyses in the Cochrane library over varying numbers of studies.

1.11 Overview of thesis

Recent studies have focused on comparing the available τ^2 estimators under a number of scenarios, including sparse data (Friede et al. (2017a)), however none have yet concentrated on the case of rare events. For our project, we aim to determine the effectiveness of currently available heterogeneity variance estimators and other statistical methods for the meta-analysis of rare-event data, and develop new techniques that we believe to be appropriate. Below we shall list the aims of our research in detail, as well as provide an outline of the structure of the remainder of this thesis.

1.11.1 Aims

The aims of our research are to investigate the performance of τ^2 estimators under sparsity in binary outcome meta-analyses, focusing on situations where: (1) the probability of an event in either study intervention arm is low (e.g. 1 in 1000), (2) we have few studies in the meta-analysis, and (3) the individual studies have small sample sizes. We shall use the risk ratio as our outcome measure because this is the recommended choice for clinicians, and was found to be the most popular in the CDS. We will propose new methods based on mixed regression models, as well as others based on conditional approaches and mixture models, that we believe to be appropriate for use with rare-event data.

Once we have developed our new approaches, we will assess their performance against that of existing methods under various scenarios of sparsity. In order to do this, we will use empirical data that meets our sparsity requirements, as well as design and run simulations to recreate these situations, using the statistical software package R (R Core Team (2019)). The design of these simulations will be based on simulations conducted in studies focusing on similar problems, and according to meta-analyses that we conduct in practice. However, we shall be extending them to fewer numbers of studies and smaller sample sizes, as well as more pronounced between-study heterogeneity, to better reflect the circumstances commonly encountered in rare-event data. We will generate the data according to a RE model, and vary the following parameters:

- Number of studies
- Study sample sizes (looking at both balanced and unbalanced cases)
- True heterogeneity variance
- Probability of events
- Variance of study-specific baseline risk.

In addition to investigating the performance of τ^2 estimators, we shall construct our simulation study to also compare existing continuity corrections and confidence intervals, simulating meta-analyses from a range of realistic rare-event scenarios. For completeness, we will also compare the point estimates of the risk ratio produced using these τ^2 estimates with that produced using the fixed-effect Mantel-Haenszel approach, in order to determine the impact of accounting for heterogeneity.

Using our results, we will then be able to investigate the performance of estimation of the heterogeneity τ^2 and effect measure θ for each of the methods considered, in terms of performance measures such as bias and mean squared error. We will then construct confidence and credibility intervals, and compare all combinations of estimators and confidence intervals in terms of coverage. From this, we will be able to determine the τ^2 estimators, and associated confidence intervals, that perform best for a range of given scenarios, and thus produce recommendations and guidelines as to the appropriate methods to use given the structure and characteristics of the meta-analysis data.

1.11.2 Structure of thesis

In Chapter 2, we will be describing the current methods available to estimate heterogeneity variance, and will outline their performance for sparse-event data using results from previous simulation studies. To obtain a feeling of the differences in τ^2 estimating methods and how they respond to common scenarios present in rare-event data, we shall then look at several empirical datasets containing low-probability events or small number of studies and apply all mentioned estimators to these cases in Chapter 3. We shall compare the results of these estimators, plotting the respective forest plots for each case study investigated.

In Chapter 4, we will also propose a new idea for the analysis of rare-event data (conditional logistic mixed regression modelling), as well as describe the previously suggested Poisson mixed regression models. In addition to these, we will also propose two additional methods, one based on a conditional approach, that uses modified versions of estimating equations proposed by Böhning and Sarol (2000) that have been adapted for the risk ratio in Chapter 5. Our second additional approach, discussed in Chapter 6, involves using mixture models, and applying the EM algorithm to determine the model of best fit and extract parameter estimates.

The latter portion of this thesis will focus on a description of our methods for comparing these estimators through the use of a new simulation study, focusing largely on rare-event data, where we shall produce a comprehensive list of all the scenarios that we are interested in investigating in Chapter 7. Having conducted these simulations, we shall then discuss certain results of interest regarding the performance of estimators that have been applied to the range of simulated meta-analyses, determining which methods

perform best in certain scenarios in Chapter 8. From this, we will be able to make recommendations and construct guidelines based on the results obtained, and discuss what implications these have in terms of meta-analysis methodology for rare-event data in Chapter 9.

Chapter 2

Methods for estimating heterogeneity variance

2.1 Introduction

Heterogeneity variance (τ^2) estimates can provide a measure of disparity between study treatment effects and are an essential component of random-effects (RE) meta-analyses. In this chapter, we will present a comprehensive review of the methods available to estimate heterogeneity variance, focusing only on those that are used in two-step meta-analyses (as described in Section 1.4), as well as Bayesian approaches. Although we shall concentrate primarily on methods suitable for use with binary endpoint effect measures, as this is our area of interest, we will also briefly mention other methods that do not meet this requirement but which are well-cited and demonstrate the use of alternative methodology. The heterogeneity variance estimators we discuss fall into a number of distinct approaches, including the method of moments, maximum likelihood and Bayesian approaches. A summary table of all those estimators discussed here is presented later in this chapter.

The methods available to estimate τ^2 vary in terms of performance for varying meta-analysis scenarios, e.g. some estimators only perform well with high numbers of studies or large heterogeneity. As a result of this, a number of simulation studies have been conducted to compare heterogeneity variance estimators under a wide range of these realistic scenarios. Most have conducted these comparisons by measuring the bias or mean squared error (MSE) of the estimators, and have focused on cases where the number of studies k and the true value of the heterogeneity variance τ^2 vary in magnitude. In the closing sections of this chapter, we will outline the general findings of such simulation studies for the estimators we describe here, paying particular attention to their performance in the case of sparse-event data. We shall also review the performance of these methods when combined with confidence intervals for the summary effect, again for both

common and rare-event binary data scenarios. Finally, we will conclude on the methods available for rare-event meta-analyses, and how our findings provide motivation for the construction of novel approaches.

2.2 Method of moments approach

The method of moments approach was proposed by Kacker (2004), and is based on Cochran's generalised Q-statistic, Q_{MM} (defined as the weighted sum of squared differences between individual study effects and the pooled effect across studies):

$$Q_{MM} = \sum_{i=1}^{k} w_i(\hat{\theta}_i - \hat{\theta})^2 \tag{2.2.1}$$

where w_i is the weight of study i, and k is the number of studies in the meta-analysis. For the general method of moments approach, we assume that w_i does not take a specific form, and may be known or estimated using the study data. As a result of this, $\hat{\theta}$ is a generic weighted average of study effects $\hat{\theta}_i$. Q_{MM} becomes the Q-statistic introduced in Chapter 1 when $w_{i,FE} = 1/\sigma_i^2$ is substituted in for w_i in Equation (2.2.1).

The approach involves equating Q_{MM} to its expected value, giving the general method of moments (MM) estimator:

$$\hat{\tau}^2_{MM} = \frac{\left(\sum_{i=1}^{k} w_i(\hat{\theta}_i - \hat{\theta})^2\right) - \left(\sum_{i=1}^{k} w_i\sigma_i^2 - \frac{\sum_{i=1}^{k} w_i^2\sigma_i^2}{\sum_{i=1}^{k} w_i}\right)}{\sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i}} \tag{2.2.2}$$

A detailed derivation of this estimator can be seen in Appendix A.1. The approach assumes that σ_i^2 are known but these are generally replaced by estimates $\hat{\sigma}_i^2$ in Formula (2.2.2). Every method of moments estimator can be derived from Formula (2.2.2), by replacing w_i with its associated functional form. Because the proposed study weights in this equation do not ensure that $\hat{\tau}^2 > 0$, all method of moments estimators are truncated to zero whenever they would otherwise produce negative values.

2.2.1 DerSimonian-Laird

As mentioned in Chapter 1, the most commonly used method to estimate heterogeneity variance is the non-iterative DerSimonian-Laird (DL) method (DerSimonian and Laird (1986)). This estimator is derived by equating the observed value of Q_{MM} with FE weights $\hat{w}_{i,FE} = 1/\hat{\sigma}_i^2$ (which we denote $\hat{Q}$), with its expected value $E(\hat{Q})$, where $E(\hat{Q}) = \tau^2\left(\sum_{i=1}^{k}(1/\hat{\sigma}_i^2) - \frac{\sum_{i=1}^{k}(1/\hat{\sigma}_i^2)^2}{\sum_{i=1}^{k}(1/\hat{\sigma}_i^2)}\right) + (k-1)$, yielding:

$$\hat{\tau}^2_{DL} = max\left\{0, \frac{\sum_{i=1}^{k}(1/\hat{\sigma}_i^2)(\hat{\theta}_i - \hat{\theta}_{DL})^2 - (k-1)}{\sum_{i=1}^{k}(1/\hat{\sigma}_i^2) - \frac{\sum_{i=1}^{k}(1/\hat{\sigma}_i^2)^2}{\sum_{i=1}^{k}(1/\hat{\sigma}_i^2)}}\right\}$$

where $\hat{\theta}_{DL} = \sum_{i=1}^{k}(\hat{w}_{i,FE}\hat{\theta}_i)/\sum_{i=1}^{k}\hat{w}_{i,FE}$. We can also obtain this estimator by directly substituting the FE weights $w_{i,FE} = 1/\hat{\sigma}_i^2$ into Equation (2.2.2).

A number of variations of the DL method were proposed by Kontopantelis et al. (2013). These include the positive DL (DLp) method, which uses truncation to an arbitrary positive constant (e.g. 0.01) rather than zero, to ensure a positive estimate. A non-parametric bootstrap DL (DLb) method was also proposed, which aims to minimise the proportion of zero estimates. This bootstrap approach is conducted by first randomly sampling k studies with replacement, and then calculating $\hat{\tau}^2_{DL}$ for the sample. These steps are then repeated B times, where say $B = 10,000$, and then the estimate $\hat{\tau}^2_{DLb}$ is calculated as the mean of these B estimates. This bootstrap method could theoretically be employed for any τ^2 estimator.

2.2.2 Hedges-Olkin

Another non-iterative estimator from the method of moments approach is the Hedges-Olkin (HO) method, also known as Cochran's ANOVA or variance component type estimator (Hedges and Olkin (1985); Cochran (1954)). It is obtained by setting the sample variance, $S^2_\theta = \frac{1}{k-1}\sum_{i=1}^{k}(\hat{\theta}_i - \hat{\theta}_{HO})^2$, equal to its expected value and solving for τ^2:

$$\hat{\tau}^2_{HO} = max\left\{0, \frac{1}{k-1}\sum_{i=1}^{k}\left(\hat{\theta}_i - \hat{\theta}_{HO}\right)^2 - \frac{1}{k}\sum_{i=1}^{k}\hat{\sigma}_i^2\right\}$$

where $\hat{\theta}_{HO}$ is the unweighted average of $\hat{\theta}_i$, given by:

$$\hat{\theta}_{HO} = \sum_{i=1}^{k}\hat{\theta}_i/k \tag{2.2.3}$$

This heterogeneity variance estimator can equivalently be produced by substituting weights $w_i = 1/k$ into Equation (2.2.2):

$$\begin{aligned}\hat{\tau}^2_{HO} &= \frac{\left(\sum_{i=1}^k (1/k)(\hat{\theta}_i - \hat{\theta}_{HO})^2\right) - \left(\sum_{i=1}^k (1/k)\hat{\sigma}_i^2 - \frac{\sum_{i=1}^k (1/k)^2\hat{\sigma}_i^2}{\sum_{i=1}^k (1/k)}\right)}{\sum_{i=1}^k (1/k) - \frac{\sum_{i=1}^k (1/k)^2}{\sum_{i=1}^k (1/k)}} \\ &= \frac{(1/k)\left(\sum_{i=1}^k (\hat{\theta}_i - \hat{\theta}_{HO})^2\right) - \left((1/k)\sum_{i=1}^k \hat{\sigma}_i^2 - (1/k)^2\sum_{i=1}^k \hat{\sigma}_i^2\right)}{1 - (1/k)} \\ &= \frac{(1/k)\sum_{i=1}^k (\hat{\theta}_i - \hat{\theta}_{HO})^2 - ((1/k) - (1/k)^2)\sum_{i=1}^k \hat{\sigma}_i^2}{(k-1)/k} \\ &= \frac{1}{k-1}\sum_{i=1}^k \left(\hat{\theta}_i - \hat{\theta}_{HO}\right)^2 - \frac{1}{k}\sum_{i=1}^k \hat{\sigma}_i^2\end{aligned}$$

2.2.3 Mandel-Paule

The Mandel-Paule (MP), or empirical Bayes, estimator is an iterative example of the method of moments approach (Paule and Mandel (1982)). In this case, the estimate $\hat{\tau}^2$ is calculated as the unique solution to the MP estimating equation, $F(\hat{\tau}^2_{MP})$, by means of fixed-point iteration. This equation is obtained by equating Q_{MM}, with weights $\hat{w}_{i,RE} = 1/(\hat{\sigma}_i^2 + \hat{\tau}^2_{MP})$, to its expected value $k-1$, giving:

$$F(\hat{\tau}^2_{MP}) = \sum_{i=1}^k \frac{\left(\hat{\theta}_i - \hat{\theta}_{MP}\right)^2}{\hat{\sigma}_i^2 + \hat{\tau}^2_{MP}} - (k-1) = 0$$

where $\hat{\theta}_{MP} = \sum_{i=1}^k (\hat{w}_{i,RE}\hat{\theta}_i)/\sum_{i=1}^k \hat{w}_{i,RE}$. The solution, $\hat{\tau}^2_{MP}$, is determined through a process of numerical iteration until convergence, starting with an initial estimate of $\hat{\tau}^2_0 = 0$. If $F(\hat{\tau}^2_{MP})$ is negative for all $\hat{\tau}^2_{MP} \geq 0$, the estimate is set to zero. An alternative method to derive this estimate involves substituting the above RE weights into Equation (2.2.2), giving:

$$\hat{\tau}^2_{MP} = \frac{\sum_{i=1}^k \hat{w}_{i,RE}(\hat{\theta}_i - \hat{\theta}_{MP})^2 - (\sum_{i=1}^k \hat{w}_{i,RE}\hat{\sigma}_i^2 - (\sum_{i=1}^k \hat{w}_{i,RE}\hat{\sigma}_i^2)/(\sum_{i=1}^k \hat{w}_{i,RE}))}{\sum_{i=1}^k \hat{w}_{i,RE} - (\sum_{i=1}^k \hat{w}^2_{i,RE})/(\sum_{i=1}^k \hat{w}_{i,RE})}$$

As before, a process of iteration is needed to determine the $\hat{\tau}^2_{MP}$ estimate here, however this second approach is seen to be more intuitive.

2.2.4 Improved Mandel-Paule

For the meta-analysis of odds ratio outcome data, Bhaumik et al. (2012) suggested an improved, stabilising MP (IMP) estimator for rare events. Their approach provides an

alternative method of calculating sampling variance, and involves borrowing strength from other studies when estimating each within-study variance. Instead of using the standard $\hat{\sigma}_i^2$, they suggest:

$$\hat{\sigma}_i^2(*) = \frac{1}{n_{it}+1}\left[\exp\left(-CGR_i - \hat{\theta}_{HO} + \frac{\hat{\tau}^2}{2}\right) + 2 + \exp\left(CGR_i + \hat{\theta}_{HO} + \frac{\hat{\tau}^2}{2}\right)\right] + \frac{1}{n_{ic}+1}\left[\exp(-CGR_i) + 2 + \exp(CGR_i)\right]$$

where n_{it} and n_{ic} are the sample sizes in the treatment and control group respectively, of study i, CGR_i is the estimated risk of an event in the control group of study i, and $\hat{\theta}_{HO}$ is the equally-weighted combined effect estimate in Equation (2.2.3), but with a continuity correction for zero events. The risk CGR_i can be calculated as $CGR_i = c_i/n_{ic}$, where c_i is the number with the event of interest in the control group, and the values of c_i and n_{ic} can be taken from a 2×2 contingency table, such as the one given in Table 1.1.

To obtain the improved MP estimator, $\hat{\tau}^2_{IMP}$, the same process is used as in MP, but this time using shared-strength weights $w^*_{i,RE} = 1/(\hat{\tau}^2 + \hat{\sigma}_i^2(*))$. These alternative estimates of the within-study variance could in principle be applied to every τ^2 estimator. Since this estimator is solely for use with the odds ratio outcome measure, it will not be of use for our risk ratio meta-analysis simulation study, however we include it here for completeness and to provide an example of estimators proposed specifically for use with rare-event data.

2.2.5 Two-step estimators

Non-iterative, two-step versions of the MP method have been proposed by DerSimonian and Kacker (2007). Similar to the MP method, these estimators require RE study weights, but their iteration is restricted to only two steps rather than until convergence. Two forms of this two-step approach exist, one with initial estimate $\hat{\tau}_0^2 = \hat{\tau}^2_{DL}$ (DL2), the other with initial estimate $\hat{\tau}_0^2 = \hat{\tau}^2_{HO}$ (HO2). If we substitute $\hat{w}_{i,RE} = 1/(\hat{\tau}^2_{DL} + \hat{\sigma}_i^2)$ in Equation (2.2.2), and use $\hat{\theta}_{DL}$ as defined above, we get the two-step estimate $\hat{\tau}^2_{DL2}$, where:

$$\hat{\tau}^2_{DL2} = \frac{\sum_{i=1}^k \hat{w}_{i,RE}(\hat{\theta}_i - \hat{\theta}_{DL})^2 - (\sum_{i=1}^k \hat{w}_{i,RE}\hat{\sigma}_i^2 - (\sum_{i=1}^k \hat{w}^2_{i,RE}\hat{\sigma}_i^2)/(\sum_{i=1}^k \hat{w}_{i,RE}))}{\sum_{i=1}^k \hat{w}_{i,RE} - (\sum_{i=1}^k \hat{w}^2_{i,RE})/(\sum_{i=1}^k \hat{w}_{i,RE})}$$

The HO two-step estimator, $\hat{\tau}^2_{HO2}$, can be constructed in a similar manner, by substituting $\hat{w}_{i,RE} = 1/(\hat{\sigma}_i^2 + \hat{\tau}^2_{HO})$ and $\hat{\theta}_{HO}$ into Equation (2.2.2), thus giving:

$$\hat{\tau}^2_{HO2} = \frac{\sum_{i=1}^k \hat{w}_{i,RE}(\hat{\theta}_i - \hat{\theta}_{HO})^2 - (\sum_{i=1}^k \hat{w}_{i,RE}\hat{\sigma}_i^2 - (\sum_{i=1}^k \hat{w}_{i,RE}^2\hat{\sigma}_i^2)/(\sum_{i=1}^k \hat{w}_{i,RE}))}{\sum_{i=1}^k \hat{w}_{i,RE} - (\sum_{i=1}^k \hat{w}_{i,RE}^2)/(\sum_{i=1}^k \hat{w}_{i,RE})}$$

These two-step estimators are included here for completeness, however they shall not be used in our simulation study as they only advance their respective base estimators by one further step of iteration, and so are not seen as significantly improving the estimation.

2.3 Non-truncated moments-based approaches

As mentioned in Section 2.2, the majority of method of moments estimators require truncation to zero as they allow for the production of negative estimates. However, there are some estimators based on this approach that do not require such truncation.

2.3.1 Hartung-Makambi

One of these is the Hartung-Makambi (HM) method, which is a non-iterative modification of the DL method that always produces positive results (Hartung and Makambi (2003)). Recall that:

$$\hat{\tau}^2_{DL} = max\left\{0, \frac{\hat{Q} - (k-1)}{c}\right\}$$

where $\hat{Q}$ is the value of Q_{MM} given in Section 2.2.1 and $c = \sum_{i=1}^k \hat{w}_{i,FE} - (\sum_{i=1}^k \hat{w}_{i,FE}^2 / \sum_{i=1}^k \hat{w}_{i,FE})$.

The HM method involves multiplying the (always positive) first term of $\hat{\tau}^2_{DL}$ $(\hat{Q}/c)$, by a positive correction factor, denoted by ϵ, which accounts for the bias resulting from the exclusion of the term $(k-1)/c$. The HM estimator therefore takes the form $\hat{\tau}^2_{HM} = \epsilon \cdot \hat{Q}/c$, where:

$$\epsilon = \frac{\hat{Q}}{2(k-1) + \hat{Q}}$$

giving the HM estimator:

$$\hat{\tau}^2_{HM} = \frac{\hat{Q}^2}{(2(k-1) + \hat{Q})\left(\sum_{i=1}^k \hat{w}_{i,FE} - \frac{\sum_{i=1}^k \hat{w}_{i,FE}^2}{\sum_{i=1}^k \hat{w}_{i,FE}}\right)}$$

2.3.2 Sidik-Jonkman

Another estimator that always yields positive results is the non-iterative Sidik-Jonkman (SJ) method, or model error variance estimator, which is based on weighted least squares (Sidik and Jonkman (2005)). This estimator is methodologically similar to the MP method, as the weights are equivalent to the RE study weights, multiplied by a constant $\hat{\tau}^2$ to ensure positivity. To obtain the SJ estimator, we first substitute the following new study weights $\hat{w}_{i,SJ} = \frac{1}{(\hat{\sigma}_i^2/\hat{\tau}^2)+1} = \frac{\hat{\tau}^2}{\hat{\sigma}_i^2+\hat{\tau}^2}$ into the standard formula for $Var(\hat{\theta})$ introduced in Chapter 1 $\left(Var(\hat{\theta}) = \frac{1}{\left(\sum_{i=1}^k w_i\right)^2} \sum_{i=1}^k w_i^2\sigma_i^2\right)$, giving:

$$\widehat{Var}(\hat{\theta}) = \frac{\sum_{i=1}^k \left(\frac{\hat{\tau}^2}{\hat{\sigma}_i^2+\hat{\tau}^2}\right)^2 (\hat{\sigma}_i^2+\hat{\tau}^2)}{\left(\sum_{i=1}^k \frac{\hat{\tau}^2}{\hat{\sigma}_i^2+\hat{\tau}^2}\right)^2} = \frac{\hat{\tau}^4 \sum_{i=1}^k \frac{1}{\hat{\sigma}_i^2+\hat{\tau}^2}}{\left(\sum_{i=1}^k \frac{\hat{\tau}^2}{\hat{\sigma}_i^2+\hat{\tau}^2}\right)^2} = \frac{\hat{\tau}^2}{\sum_{i=1}^k \frac{\hat{\tau}^2}{\hat{\sigma}_i^2+\hat{\tau}^2}} = \frac{\hat{\tau}^2}{\sum_{i=1}^k \hat{w}_{i,SJ}}$$

This variance is then equated to an alternative, weighted estimate of $Var(\hat{\theta})$, proposed by Hartung (1998):

$$\widehat{Var}_{HK}(\hat{\theta}) = \frac{\sum_{i=1}^k \hat{w}_{i,SJ}(\hat{\theta}_i - \hat{\theta})^2}{(k-1)\sum_{i=1}^k \hat{w}_{i,SJ}}$$

These two values for the variance are equated, and then rearranged in terms of τ^2, giving us the SJ estimator:

$$\hat{\tau}_{SJ}^2 = \frac{1}{k-1}\sum_{i=1}^k \frac{1}{(\hat{\sigma}_i^2/\hat{\tau}_0^2)+1}(\hat{\theta}_i - \hat{\theta}_{SJ})^2$$

where $\hat{\theta}_{HO}$ is the unweighted estimate of θ defined earlier, and $\hat{\theta}_{SJ}$ is the weighted least squares estimate of θ with weights $\hat{w}_{i,SJ} = 1/((\hat{\sigma}_i^2/\hat{\tau}^2)+1)$. Since this formula is iterative, Sidik and Jonkman (2005) proposed a two-step approach with initial estimate $\hat{\tau}_0^2 = \frac{1}{k}\sum_{i=1}^k(\hat{\theta}_i - \hat{\theta}_{HO})^2$. We set $\hat{\tau}_{SJ}^2 = 0$ in the unlikely event that $\hat{\tau}_0^2 = 0$ (all $\hat{\theta}_i$ are equal), and so the study weights $\hat{w}_{i,SJ}$ are undefined.

The idea behind this estimator comes from replacing the standard random-effects weights $\hat{w}_i = \frac{1}{\hat{\sigma}_i^2+\hat{\tau}^2}$ with an alternative involving a ratio of the variances $r_i = \hat{\sigma_i}^2/\hat{\tau}^2$, giving $\hat{w}_{i,SJ} = \frac{1}{(\hat{\sigma}_i^2/\hat{\tau}^2)+1}$, which we take from noting that $Var(\theta_i) = \sigma_i^2 + \tau^2 = \tau^2(r_i+1)$ where they assume that the within-study variances are known, and thus the ratios are also assumed to be known (despite usually being unknown in practice), and also assuming that $\tau^2 \neq 0$. By viewing the effect measure as a linear model, the best linear unbiased estimate for the summary outcome measure is given by the weighted least

squares estimators, where the rates are given as above. If both variances are known, then this estimator is equivalent to using the standard random-effects weights. This is a two-step approach since a crude initial estimate of τ^2 is obtained through the above equation for $\hat{\tau}_0^2$, and this is then input into the alternative weights, which is in turn input into the weighted residual sum of squares to provide an estimate for τ^2.

Sidik and Jonkman (Sidik and Jonkman (2005)) noted that alternative $\hat{\tau}_0^2$ estimates may lead to an estimator with better properties. Therefore, Sidik and Jonkman (Sidik and Jonkman (2007)) proposed $\hat{\tau}_0^2 = max(0.01, \hat{\tau}_{HO}^2)$ in a follow-up paper; I denote the resulting estimator as SJ2. As with the original estimator, $\hat{\tau}_{SJ2}^2$ is a two-step estimator that is simple to compute and always results in a positive estimate of the heterogeneity variance.

2.4 Likelihood-based approach

The maximum likelihood (ML) and restricted maximum likelihood (REML) estimators are both computationally intensive, iterative methods based on the marginal distribution $\hat{\theta}_i \sim N(\theta, \sigma_i^2 + \tau^2)$. These approaches differ from the method of methods approach as the parameter θ_i and associated estimate $\hat{\theta}_i$ are assumed to be normally distributed around the central parameter, θ.

2.4.1 Maximum likelihood

The classical maximum likelihood (ML) estimate is obtained by maximising the log-likelihood function (Hardy and Thompson (1996)), given by:

$$\log L_{ML}(\theta, \tau^2) = -\frac{k}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{k}\log(\sigma_i^2 + \tau^2) - \frac{1}{2}\sum_{i=1}^{k}\frac{(\theta_i - \theta)^2}{\sigma_i^2 + \tau^2}$$

ML estimators for θ and τ^2 can be obtained by partially differentiating $\log L_{ML}$ with respect to θ and τ^2, respectively, and equating the resulting equations to zero. This gives the following ML estimates:

$$\hat{\theta}_{ML} = \frac{\sum_{i=1}^{k} \hat{w}_{i,RE}\hat{\theta}_i}{\sum_{i=1}^{k} \hat{w}_{i,RE}} \tag{2.4.1}$$

$$\hat{\tau}_{ML}^2 = max\left\{0, \frac{\sum_{i=1}^{k} \hat{w}_{i,RE}^2((\hat{\theta}_i - \hat{\theta}_{ML})^2 - \hat{\sigma}_i^2)}{\sum_{i=1}^{k} \hat{w}_{i,RE}^2}\right\} \tag{2.4.2}$$

where $\hat{w}_{i,RE} = 1/(\hat{\sigma}_i^2 + \hat{\tau}_{ML}^2)$ and $\hat{\theta}_{ML}$ is the maximum likelihood estimate of θ. The ML estimates are then calculated by solving the above two equations simultaneously and iteratively, beginning with an initial estimate $\hat{\tau}_0^2$. As such, these values could be estimated using an iterative scheme, where you start with some value $\hat{\tau}_0^2$, and substitute this into Equation (2.4.1) to give a value for $\hat{\theta}_{ML}$. This value for $\hat{\theta}_{ML}$ would then be substituted into Equation (5.4.4) to generate a new value for $\hat{\tau}_{ML}^2$, and this process would be continued until the solutions converge.

2.4.2 Restricted maximum likelihood

The restricted maximum likelihood (REML) method can be used to correct for the negative bias that is associated with the above ML method (Raudenbush (2009)). In this case, the estimate is produced by maximising the restricted log-likelihood function:

$$\log L_{REML}(\tau^2) = -\frac{k}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{k}\log(\sigma_i^2+\tau^2) - \frac{1}{2}\sum_{i=1}^{k}\frac{(\hat{\theta}_i - \hat{\theta})^2}{(\sigma_i^2+\tau^2)} - \frac{1}{2}\log\left(\sum_{i=1}^{k}\frac{1}{(\sigma_i^2+\tau^2)}\right)$$

The estimate for τ^2 is derived by partially differentiating $\log L_{REML}$ with respect to τ^2, and then setting this equal to zero and solving the resulting equation, giving:

$$\hat{\tau}_{REML}^2 = max\left\{0, \frac{\sum_{i=1}^{k}\hat{w}_{i,RE}^2((\hat{\theta}_i - \hat{\theta}_{REML})^2 - \hat{\sigma}_i^2)}{\sum_{i=1}^{k}\hat{w}_{i,RE}^2} + \frac{1}{\sum_{i=1}^{k}\hat{w}_{i,RE}}\right\}$$

where $\hat{\theta}_{REML} = \sum_{i=1}^{k}(\hat{w}_{i,RE}\hat{\theta}_i)/\sum_{i=1}^{k}\hat{w}_{i,RE}$ and $\hat{w}_{i,RE} = 1/(\hat{\sigma}_i^2 + \hat{\tau}_{REML}^2)$. As before, the estimate is calculated by a process of iteration with an initial estimate of $\hat{\tau}_0^2 \geq 0$, where each iteration step requires non-negativity. The difference between the REML and ML estimators for τ^2 is the addition of the term $1/\sum_{i=1}^{k}\hat{w}_{i,RE}$ in the REML formula.

2.4.3 Approximate restricted maximum likelihood

An approximate restricted maximum likelihood (AREML) method has also been proposed by Morris (1983), which uses a direct adjustment for the loss of degrees of freedom. The AREML estimate for τ^2 is the iterative solution to:

$$\hat{\tau}_{AREML}^2 = max\left\{0, \frac{\sum_{i=1}^{k}\hat{w}_{i,RE}^2((\frac{k}{k-1})(\hat{\theta}_i - \hat{\theta}_{AREML})^2 - \hat{\sigma_i}^2)}{\sum_{i=1}^{k}\hat{w}_{i,RE}^2}\right\}$$

where the weights are defined as $\hat{w}_{i,RE} = 1/(\hat{\sigma}_i^2 + \hat{\tau}^2_{AREML})$. This method yields similar estimates to REML, and in the case where all sampling variances are equal, AREML and REML estimates are identical.

2.5 Hunter-Schmidt

Other frequentist estimators have been proposed that do not fit into any of the grouped methodology types. One of these is the direct variance component approach Hunter-Schmidt (HS) estimator (Schmidt and Hunter (2014)), which is obtained by writing the variance components for $\hat{\theta}$ as $Var(\hat{\theta}) = \tau^2 + \sigma^2$, and then substituting the weighted unbiased estimates of $Var(\hat{\theta})$ and σ^2 into the variance components. It is produced by setting the Q-statistic, $Q = \sum_{i=1}^{k} w_{i,FE}(\hat{\theta}_i - \hat{\theta})^2$, equal to its expected value and solving for τ^2:

$$E(Q) = \sum_{i=1}^{k} w_{i,FE} E(\hat{\theta}_i - \hat{\theta})^2 \approx \sum_{i=1}^{k} w_{i,FE}(\sigma_i^2 + \tau^2) = k + \sum_{i=1}^{k} w_{i,FE}\tau^2$$

$$\Rightarrow \hat{\tau}^2_{HS} = max\left\{0, \frac{\sum_{i=1}^{k} \hat{w}_{i,FE}(\hat{\theta}_i - \hat{\theta})^2 - k}{\sum_{i=1}^{k} \hat{w}_{i,FE}}\right\}$$

2.6 Bayesian approach

As mentioned in Chapter 1, meta-analyses can be conducted in a Bayesian, rather than classical, framework. In these cases, the heterogeneity variance can also be estimated by means of a Bayesian approach, and these allow for the incorporation of prior beliefs of model parameters to be included with the meta-analysis data. Here we will discuss a fully Bayesian approach, as well as some semi-Bayesian estimators.

2.6.1 Full Bayesian

The full Bayesian (FB) method allows estimates to be obtained simultaneously with all other parameters of interest, incorporating any uncertainty in these estimates. In a Bayesian RE model with no covariates, the prior distributions for θ and τ^2 can be defined as:

$$\theta \sim \pi_1(\phi_1)$$

$$\tau^2 \sim \pi_2(\phi_2)$$

where π_1 and π_2 are chosen probability distributions with fixed parameter vectors ϕ_1 and ϕ_2. The prior distributions and their parameter vectors are chosen to reflect the prior knowledge, or set as vague if none is available. Possible prior distributions for τ^2 include inverse-gamma and uniform, and recently half-normal priors have been proposed (Günhan et al. (2018)), while θ_i will generally be assumed a normal distribution. Estimates for these two parameters are extracted from the joint posterior distribution for θ and τ^2, which is itself calculated by combining the prior distributions with the meta-analysis data using Markov chain Monte Carlo (MCMC) methods. We could also make θ and τ^2 dependent by using joint priors for these parameters, where the prior for one parameter is set to be conditional on the other.

2.6.2 Rukhin Bayes

As mentioned above, a series of semi-Bayesian estimators exist, one of which is the Rukhin Bayes (RB) method (Rukhin (2013)). This estimator is simpler to compute than the FB method described above, and only requires a fixed prior estimate for τ^2, denoted by $\hat{\tau}_0^2$. It is based on the generalised method of moments approach, and involves deriving the formula for $Var(\hat{\tau}^2)$, and then choosing $\hat{\tau}^2$ such that $Var(\hat{\tau}^2)$ is locally minimised around the prior estimate of τ^2. The general form of the RB estimator is:

$$\hat{\tau}^2_{RB} = max\left\{0, \frac{\sum_{i=1}^k(\hat{\theta}_i - \hat{\theta})^2}{k+1} + \frac{(\sum_{i=1}^k(n_{it}+n_{ic}) - k)(2k\hat{\tau}_0^2 - (k-1)\sum_{i=1}^k \hat{\sigma}_i^2)}{\sum_{i=1}^k((n_{it}+n_{ic}) - k + 2)k(k+1)}\right\}$$

where $\hat{\theta} = \sum_{i=1}^k(\hat{\theta}_i\hat{w}_{i,RE})/\sum_{i=1}^k \hat{w}_{i,RE}$, $\hat{w}_{i,RE} = 1/(\hat{\sigma}_i^2 + \hat{\tau}_0^2)$, and n_{it} and n_{ic} are the sample sizes in the treatment and control groups, respectively. Rukhin recommended estimating $\hat{\tau}^2_{RB}$ using the prior estimate $\hat{\tau}_0^2 = 0$ (RB0).

2.6.3 Bayes Modal

Another semi-Bayesian approach is the Bayes Modal (BM) method, which can estimate τ^2 without the need for MCMC methods (Chung et al. (2013, 2014)). To derive the BM estimator, you need to use the profile likelihood, given by:

$$\log L_p(\tau) = -\frac{k}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^k \log(\sigma_i^2 + \tau^2) - \frac{1}{2}\sum_{i=1}^k \frac{\left(\hat{\theta}_i - \frac{\sum_{i=1}^k(\sigma_i^2+\tau^2)^{-1}\hat{\theta}_i}{\sum_{i=1}^k(\sigma_i^2+\tau^2)^{-1}}\right)^2}{(\sigma_i^2+\tau^2)}$$

The BM estimate, $\hat{\tau}^2_{BM}$, is then obtained by approximating $\log L_p(\tau)$ using the ML estimator $\hat{\tau}^2_{ML}$ and a Taylor expansion, giving the closed-form solution:

$$\hat{\tau}^2_{BM} = \begin{cases} Var(\hat{\tau}^2_{ML}) & , \ \hat{\tau}_{ML} = 0 \\ \left(\frac{\hat{\tau}_{ML}}{2} + \frac{\hat{\tau}_{ML}}{2} \sqrt{1 + \frac{4Var(\hat{\tau}^2_{ML})}{\hat{\tau}^2_{ML}}} \right)^2 & , \ \hat{\tau}_{ML} > 0 \end{cases}$$

2.7 Summary of binary-outcome heterogeneity variance estimators

In Sections 2.2 to 2.6, we gave a comprehensive review of the methods proposed to estimate heterogeneity variance in a meta-analysis with binary outcome measures. We included all those estimators of interest at the time of publication. Table 2.1 lists all of these estimators, along with their location in the chapter, and the abbreviations that we shall use to refer to them throughout the remainder of this thesis.

TABLE 2.1: A summary of the heterogeneity variance estimators along with their respective abbreviations used in this thesis.

Estimator	Abbreviation	Section
Method of moments approach		2.2
DerSimonian-Laird	DL	2.2.1
Positive DerSimonian-Laird	DLp	2.2.1
Bootstrap DerSimonian-Laird	DLb	2.2.1
Hedges-Olkin	HO	2.2.2
Mandel-Paule	MP	2.2.3
Improved Mandel-Paule	IMP	2.2.4
Two-step DerSimonian-Laird	DL2	2.2.5
Two-step Hedges-Olkin	HO2	2.2.5
Non-truncated moments-based approaches		2.3
Hartung-Makambi	HM	2.3.1
Sidik-Jonkman	SJ	2.3.2
Sidik-Jonkman (HO initial estimate)	SJ2	2.3.2
Likelihood-based approach		2.4
Maximum likelihood	ML	2.4.1
Restricted maximum maximum likelihood	REML	2.4.2
Approximate restricted maximum likelihood	AREML	2.4.3
Independent approach		2.5
Hunter-Schmidt	HS	2.5
Bayesian approach		2.6
Full Bayesian	FB	2.6.1
Rukhin Bayes	RB	2.6.2
Rukhin Bayes with prior estimate $\hat{\tau}_0^2 = 0$	RB0	2.6.2
Bayes Modal	BM	2.6.3

2.8 Other approaches

The estimators that we have discussed up until now in this chapter are appropriate for binary outcome meta-analyses, and which are well-cited and compared in the literature. We shall now briefly describe some alternative estimators that either cannot be applied to binary outcome meta-analyses, or we do not believe are appropriate for inclusion in our simulation study, but nevertheless we feel are important to mention here for completeness. The estimators considered up until this point are unspecific and can be applied to any generic estimate - any quantity of interest, effect measure or proportion.

2.8.1 Malzahn, Böhning and Holling

Malzahn et al. (2000) proposed a τ^2 estimator that can only be used with standardised mean difference meta-analyses (the continuous outcome measure described in Section 1.2.1). The key feature of this estimator is that it makes no assumptions regarding the distribution of $\{\theta_1, \ldots, \theta_k\}$, a significant advantage over other estimators. The estimate of τ^2 ($\hat{\tau}^2_{MBH}$) is derived by calculating the difference between the variance of θ_i under the fixed-effect model and under the random-effects model, giving:

$$\hat{\tau}^2_{MBH} = \left(\frac{1}{k-1}\right)\sum_{i=1}^{k}(1-K_i)(\hat{\theta}_i - \hat{\theta}_{HO}) - \frac{1}{k}\sum_{i=1}^{k}\left(\frac{n_{it}+n_{ic}}{n_{it}n_{ic}}\right) - \frac{1}{k}\sum_{i=1}^{k}\left(K_i\hat{\theta}_i^2\right)$$

where $K_i = 1 - ((N_i - 2)/N_i J_i^2)$, $N_i = n_{it} + n_{ic} - 2$, $J_i = 1 - 3/(4N_i - 1)$ (suggested by Hedges (1981) to correct for bias), and $\hat{\theta}_{HO}$ is the equally-weighted outcome measure estimate in Equation (2.2.3).

2.8.2 Böhning-Sarol

Another estimator that makes no assumptions regarding the distribution of the study-specific effect sizes is that proposed by Böhning and Sarol (2000). This estimator can only be applied to meta-analyses with the standardised mortality ratio outcome - the ratio of the observed and expected count of mortality cases. The estimate, $\hat{\tau}^2_{BS}$, is calculated using the following formula:

$$\hat{\tau}^2_{BS} = \frac{1}{k}\left[\sum_{i=1}^{k}(O_i - e_i\mu)^2/e_i^2 - \mu\sum_{i=1}^{k}\frac{1}{e_i}\right]$$

where O_i is the observed number of mortality cases in study i, e_i is the expected (non-random) number of mortality cases in study i, k is the number of studies in the meta-analysis, and μ is an unknown parameter which must be estimated. We shall further expand on this estimator, and a risk ratio variation of it that we are proposing, in Chapter 5. The estimator is quite general as it can be applied to any rate data where it is assumed that, conditional upon the study, the observed count is of Poisson type where the rate might involve some study-specific parameter.

2.8.3 Within-study variance estimators

The IMP approach described in Section 2.2.4 is an example of a within-study variance estimator - a method based on using an alternative estimate of the within-study variance to the standard $\hat{\sigma}_i^2$ appropriate for the outcome measure of choice (examples of $\hat{\sigma}_i^2$ were

given in Section 1.2.2). This alternate estimate of σ_i^2 is then used to calculate an estimate of the heterogeneity variance. Other within-study variance estimators exist, including that proposed by Berkey et al. (1995), who suggested an alternative estimate of σ_i^2 for risk ratio meta-analyses:

$$\hat{\sigma}_i^2(*) = \frac{1}{kn_{1i}} \sum_{i=1}^{k} \left(\frac{n_{it} - a_i}{a_i} \right) + \frac{1}{kn_{2i}} \sum_{i=1}^{k} \left(\frac{n_{ic} - c_i}{c_i} \right)$$

where a_i and c_i are the count of events in the treatment and control groups, respectively. This approach was proposed to minimise correlation between the estimates of effect size and within-study variances. An altered version of this estimator that includes a continuity correction, and again can only be used for risk ratio meta-analyses, was proposed by Knapp and Hartung (2003):

$$\hat{\sigma}_i^2(*) = \frac{1}{kn_{1i}} \sum_{i=1}^{k} \left(\frac{n_{it} - a_i + C}{a_i + C} \right) + \frac{1}{kn_{2i}} \sum_{i=1}^{k} \left(\frac{n_{ic} - c_i + C}{c_i + C} \right)$$

where they set the correction C to be 0.5, in line with the constant continuity correction discussed in Section 1.9.1.

These within-study variance estimators differ from the methods discussed previously as they do not directly estimate τ^2, instead estimating the value of σ_i^2 that is used later in the calculation of $\hat{\tau}^2$. As a result of this, once one of these $\hat{\sigma}_i^2$ methods has been applied, the problem of deciding on an appropriate estimator for τ^2 still exists. As such, we shall not include these approaches in our simulation study, as there is no evidence that they will significantly improve the performance of the paired τ^2 estimator.

2.9 Performance of heterogeneity variance estimators

The τ^2 estimators included in Table 2.1 have been included in a number of simulation studies in order to analyse and compare their performance in terms of measures such as bias and mean squared error (MSE). In this section, we shall provide an overview of the results of these simulation studies for each method, grouping the estimators according to their associated methodology.

2.9.1 Method of moments approach

As mentioned in Chapter 1, the most widely used heterogeneity variance estimator is the DL method, however its frequent use has been questioned because of the significant bias it can have in particular scenarios, leading to unreliable τ^2 estimates (Veroniki et al.

(2016)). The estimator has been shown to be positively biased and over-estimate τ^2 on average (Viechtbauer (2005)), and it is only acceptable when the true τ^2 is small or close to 0, and k is large (Bowden et al. (2011); Novianti et al. (2014)). When τ^2 is large, DL can produce estimates with signifiant negative bias, particularly when the effect size measure is binary (Veroniki et al. (2016)). The alternative versions of the DL method also have their issues, with DLb performing well only with a large number of studies. In fact, DLb has been shown to have greater bias than DL, with the bias being more profound in small meta-analyses (Kontopantelis et al. (2013)). Similarly, Bhaumik et al. (2012) showed the DL2 method to be downwardly biased for rare events.

The MP method is mostly unbiased for large sample sizes (Panityakul et al. (2013)), however it has been shown to have upward bias for small k and τ^2, and downward bias for large k and τ^2 (Sidik and Jonkman (2007)). Despite this, it has been found to be generally less biased than alternative estimators, for both binary and continuous outcomes, and has been recommended for use with such data (Bowden et al. (2011)). The IMP estimator, proposed by Bhaumik et al. (2012) for use with rare-event data, has less bias than both the DL and MP estimators. The HO estimator performs well in the presence of substantial between-study variance, especially when k is large (≥ 30 studies), but has greater MSE than the other method of moments estimators (Chung et al. (2014); Panityakul et al. (2013); Sidik and Jonkman (2007)).

2.9.2 Non-truncated moments-based approaches

With regards to the non-truncated moments-based approaches, Thorlund et al. (2011) found that the HM method tends to over-estimate for small to moderate τ^2. The SJ estimator, which is methodologically similar to the MP method, has larger MSE and greater bias than the DL estimator for small τ^2 and few studies (Sidik and Jonkman (2005)). It also has smaller MSE compared with the HO method, irrespective of the value of τ^2 or the number of studies (Sidik and Jonkman (2007)). However, it has large bias compared with other methods when τ^2 is small, but this bias decreases as τ^2 increases (Novianti et al. (2014)), and for large τ^2, SJ and MP methods have been suggested by Sidik and Jonkman (2007) as the best estimators in terms of bias.

2.9.3 Hunter-Schmidt and likelihood based approaches

The HS and ML estimators have similar MSEs, which in turn are lower than the MSEs of the DL and HO methods. However, the HS estimator has been shown to be significantly negatively biased, and so this method should be avoided (Viechtbauer (2005)). The performance of the ML estimator depends on the choice of maximisation method, and may fail to converge, especially if there is a flat likelihood, which may be the case with a small number of studies. Although this estimator has small MSE, it exhibits large

downward bias for large τ^2 when k is small to moderate and the sample sizes are small, and so is not recommended (Kontopantelis et al. (2013); Panityakul et al. (2013)). For binary outcome data, the REML method is less downwardly biased than DL but has greater MSE (Chung et al. (2014)), and underestimates τ^2 when the data are sparse (Goldstein and Rasbash (1996)).

2.9.4 Bayesian approach

For the full Bayesian approach, the choice of prior distribution is important when k is small, and can have a substantial impact on the estimates of τ^2 and the mean treatment effect θ (Lambert et al. (2005)). In particular, inverse-gamma, uniform and Wishart prior distributions for τ^2 perform poorly for small k, and produce estimates with large bias (Gelman et al. (2006)). The semi-Bayesian RB estimator has an inherent positive bias, but is recommended for small to moderate k (Rukhin (2013)). A simulation study by Kontopantelis et al. (2013) with $k < 5$ showed that RB0 had less bias than DL (all variations), HO, REML and SJ estimators. In contrast, the BM method overestimates τ^2 and has large bias when the true $\tau^2 = 0$, especially when the sample sizes and k are small (Chung et al. (2014); Veroniki et al. (2016)).

2.10 Performance of heterogeneity variance estimators with rare events

As shown in the previous section, a number of simulation studies have been conducted to compare τ^2 estimators under realistic meta-analysis scenarios, however very few have studied the case of rare-event data. One of these is the study by Langan et al. (2018), who recently looked at odds ratio meta-analyses with an event probability of 0.05. They considered the DL, HO, MP, DL2, HO2, HM, SJ, SJ2 and REML estimators, and found that all methods performed poorly in terms of bias, MSE and coverage (when combined with summary effect confidence intervals). In particular, they noted that all estimators had considerable negative bias under this rare-event scenario, for both balanced and unbalanced study sample sizes, and for a range of true heterogeneity values and numbers of studies (k). Although they recommended the REML estimate overall, they emphasised its inappropriateness for sparse-event data, and concluded that alternate methods should be sought for this particular data type.

As mentioned in Section 2.2.4, Bhaumik et al. (2012) proposed their IMP estimator for use with odds ratio rare-event meta-analyses. To demonstrate its suitability, they conducted a simulation study, comparing its performance against that of existing estimators - namely the DL, DL2 and associated MP methods. They concentrated on scenarios with zero treatment effect (i.e. a summary log-odds ratio of 0), 20 studies and

highly unbalanced study sample sizes. For the case of extremely rare events, they found their novel IMP estimator to have the least bias, consistently outperforming the pre-existing estimators for all levels of heterogeneity investigated (τ^2 from 0 to 1.2). When the true heterogeneity was considerably large (e.g. $\tau^2 = 0.8$), the existing estimators were observed to significantly underestimate τ^2, thereby failing to detect the variation in treatment effect size across studies. As mentioned previously, the disadvantage of the IMP estimator is its limitation for use with only the odds ratio, and as such it is limited in applicability to a binary outcome that is difficult to interpret and not recommended for use by clinicians.

2.10.1 Performance of heterogeneity variance estimators with rare events and few studies

Some of the simulation studies investigating $\hat{\tau}^2$ performance for rare events also focus on the case of few studies in the meta-analysis (e.g. $k < 5$), a combined scenario that raises more challenges in terms of estimating τ^2, as a result of the consequent high proportion of zero or very low counts. For example, Friede et al. (2017a) compared DL, MP, REML and BM methods for the meta-analysis of few small studies in rare diseases. The plot in Figure 2.1 was taken from their paper, and shows the bias of these estimators given a range of values of τ (the heterogeneity standard deviation) and k. It shows that the extent of bias strongly depends on k for all estimators, with small bias for large k and substantial bias for small k. It can also be seen that the direction and size of bias heavily depends on the estimator and the value of τ, and also on the choice of prior distribution in the case of the BM method.

Another plot from the paper by Friede et al. (2017a) is shown in Figure 2.2, this time showing the proportion of τ estimates equal to zero depending on k for those estimators that are not strictly positive by construction (i.e. DL, MP and REML). It shows that for small k the proportion of estimates being equal to 0 is substantial for small to moderate τ, but this effect lessens with increasing k.

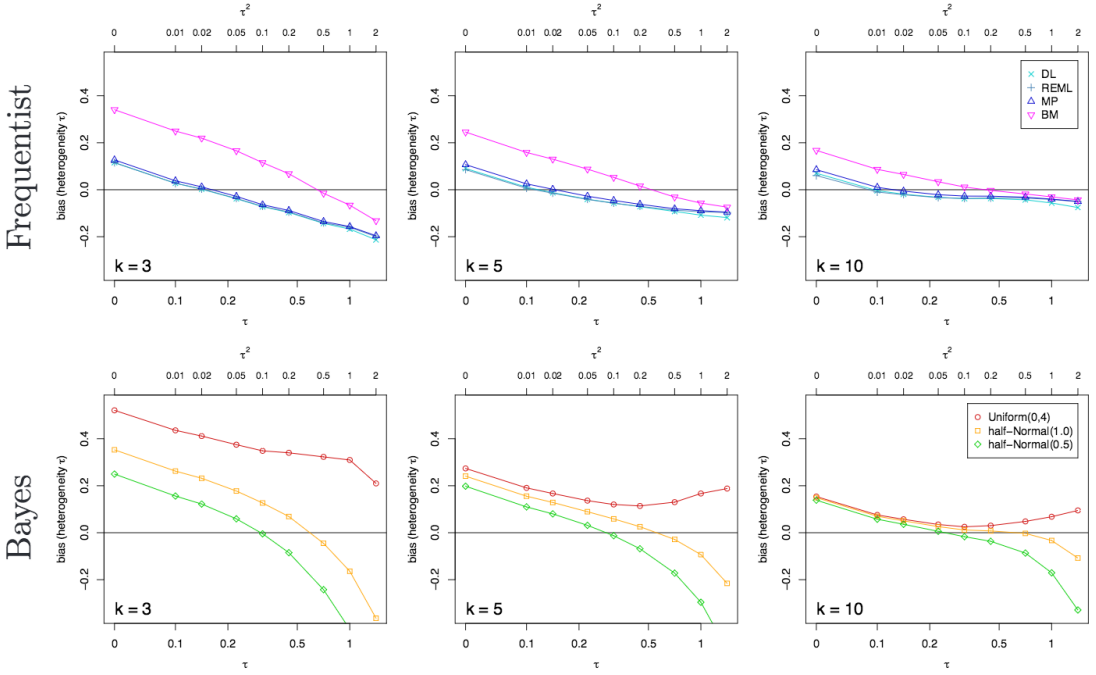


FIGURE 2.1: Plot from Friede et al. (2017a) showing the bias in estimating the between-study heterogeneity τ^2 for DL, REML, MP and BM estimators, and for several numbers k of studies included in the meta-analyses.

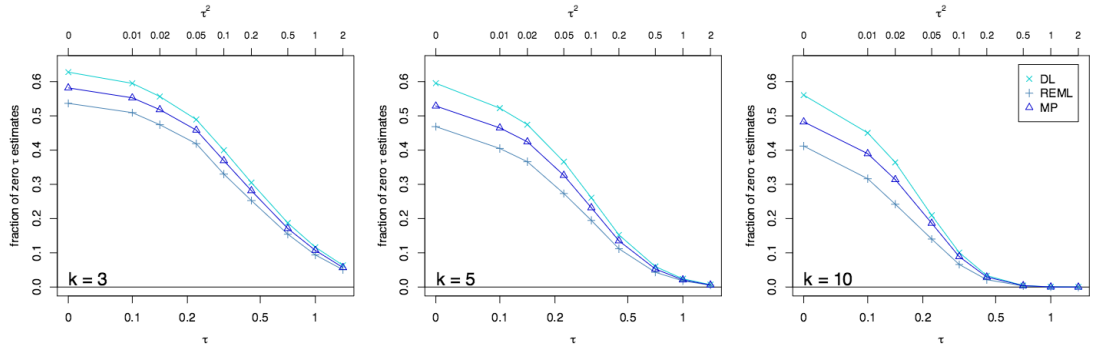


FIGURE 2.2: Plot from Friede et al. (2017a) showing the proportion of estimates of the between-study heterogeneity τ equal to zero for those estimators that are not strictly positive by construction depending on the number k of studies.

Bayesian approaches are a popular choice for this particular problem, and Günhan et al. (2018) recently tested two priors against the existing ML estimator. They used vague and weakly informative priors, both with a half-normal distribution and scale parameter of 0.5, and point estimates of τ^2 were taken as the posterior medians. For $\tau^2 = 0.28$, they found that the ML estimator consistently underestimated τ^2, whereas the Bayesian

approaches overestimated it, but to a lesser degree. As such, they concluded that the Bayesian approaches outperformed the ML.

The fewest number of studies that can be included in a meta-analysis is two, and Friede et al. (2017b) focused on this particular scenario (which occurs frequently in medical meta-analyses), specifically looking at two-study meta-analyses with rare events. They compared the DL estimator with two fully Bayesian methods under this scenario, with the Bayesian approaches given half-normal prior distributions with scale parameters 0.5 and 1, similar to the previous study mentioned. From the results of their simulation study, they found that the DL estimator was biased in all scenarios considered, overestimating τ^2 when the true heterogeneity was small ($\tau^2 = 0, 0.1$) and underestimating it when it was large ($\tau^2 = 0.5, 1$). In addition to this, a number of zero DL estimates were produced when substantial heterogeneity was present, agreeing with previous findings. Posterior median estimates obtained from the Bayesian approaches showed a similar pattern of bias, however underestimation of τ^2 only occurred when heterogeneity was very large, and the degree of bias depended heavily on the prior used. For example, the half-normal (1) prior had the greater bias of the two priors for small heterogeneity, as a result of it favouring larger values of τ^2. However, they point out that less importance can be placed on Bayesian estimates of τ^2, as the posterior distribution incorporates the uncertainty of $\hat{\tau}^2$ in the subsequent estimation of the summary effect.

2.11 Performance of summary effect confidence intervals

In Section 1.6 we outlined the main confidence intervals proposed for use with meta-analysis outcome measures: Wald-type, t-distribution, Hartung-Knapp-Sidik-Jonkman (HKSJ) and modified Knapp-Hartung (mKH) confidence intervals. A number of simulation studies comparing heterogeneity variance estimators have also investigated their performance when combined with a range of confidence intervals for the summary effect size.

Recently, Langan et al. (2018) compared the Wald-type, t-distribution and HKSJ confidence intervals for summary effect after applying a range of τ^2 estimators (DL, HO, MP, DL2, HO2, HM, SJ, SJ2 and REML). They investigated all combinations of these τ^2 estimators and confidence intervals (using the 95% level), measuring performance in terms of coverage, for both standardised mean difference and odds ratio meta-analyses with event probabilities ranging from 0.1 to 0.5. For the Wald-type method, they found that all τ^2 estimators behaved similarly (with the coverage between them differing by up to 5%). Coverage was very low, at only 65%, for small numbers of studies (e.g. $k < 5$) when significant heterogeneity was present ($I^2 = 90\%$, $\tau^2 = 0.194$), but tended towards 95% as k increased. In homogeneous situations, however, the coverage remained between 96% and 100% for all scenarios considered.

In the same paper, they found the t-distribution confidence interval to be far more robust to changes in heterogeneity. For all τ^2 estimators considered, the method was conservative when $k < 5$, producing coverages close to 100%. However, as with the Wald-type method, there was no notable difference between the τ^2 estimators, and as a result, no single τ^2 estimator could be deemed the optimum choice. Similarly, for the HKSJ confidence interval, no τ^2 estimator outperformed the others when combined with this method. However, in contrast to the previous two approaches, the HKSJ method had a very small range of 94% to 96% when the outcome measure was the standardised mean difference, for all scenarios considered. As such, they recommended this confidence interval overall, combined with either the REML, MP, or DL2 estimators in the case of balanced study samples sizes (basing the choice of τ^2 estimators on their individual merits in terms of bias and MSE), and combined with either REML or DL2 for highly unbalanced samples sizes.

2.11.1 Performance of summary effect confidence intervals with rare events

As we have discussed the results of simulation studies that focus on rare events and those that compare summary effect confidence intervals, it is only logical that we now review those that have looked at these two subjects together. In addition to looking at confidence intervals under common events, Langan et al. (2018) also compared the Wald-type, t-distribution and HKSJ methods for rare-event odds ratio meta-analyses (with an average event probability of 0.05). In contrast to the results discussed in Section 2.11, they found that the otherwise recommended HKSJ method performed poorly in terms of coverage with rare-event data when $k > 20$. The Wald-type and t-distribution confidence intervals also had extreme levels of coverage, giving similar results to those observed with common events. These results imply one of two possibilities: either all of these confidence intervals perform poorly in the case of zero counts, or they are unable to produce appropriate intervals as a result of using biased τ^2 estimators.

In their analysis of two-study meta-analyses, Friede et al. (2017b) also compared Wald-type, HKSJ and mKH confidence intervals (combined with the DL estimator) against credible intervals produced via the Bayesian approaches discussed in Section 2.10.1. They measured the coverage and interval length for each method, under both simulated and empirical data. When little heterogeneity was present, they found that all methods performed well in terms of coverage except for the Wald-type method with DL, where coverage was below the nominal 95%. Both the HKSJ and mKH methods had good coverage when study sample sizes were balanced, however the HKSJ approach performed poorly in unbalanced scenarios. In addition, these two methods produced very wide and, in some instances, implausible intervals. In comparison, the Bayesian credible intervals produced much shorter, and thus realistic, 95% intervals, and performed well in terms of

coverage regardless of sample size balance, however this was only the case when τ^2 was less than the parameter of the chosen half-normal prior distribution and so was a-priori more likely.

2.12 Conclusions

In this chapter, we have outlined the most common heterogeneity variance estimators appropriate for two-step meta-analyses, as taken from the current literature, as well as some recently proposed methods. Both frequentist and Bayesian estimators exist, and these fit into a number of distinct approaches, each with their own advantages and disadvantages. Some methods are designed specifically for use with certain data types, while others can be applied to a variety of outcome measures.

Results from previous simulation studies show that certain methods perform better than others in terms of bias and MSE over a range of scenarios. Reviewing these studies, we found that the majority of estimators considered only perform well when there are large numbers of studies, little heterogeneity, common events and balanced sample sizes. However, these scenarios are not characteristic of empirical meta-analysis data, particularly that from clinical trials.

A more common feature of medical research data is high proportions of zero or low event counts. However, we found that very little has been investigated in terms of comparing the performance of τ^2 estimators under such rare-event scenarios, with the work by Friede et al. (2017a) being one of the few examples available. As such, no or little information is available in the form of recommendations and guidelines to follow when conducting meta-analyses on this type of data. From the results that are available however, it can be deduced that all pre-existing estimators either perform poorly in this scenario or are not appropriate for use with the risk ratio, our outcome measure of interest. Given this, there is motivation to design and propose alternate methods that are better designed to work with sparse-event data in risk ratio meta-analyses.

Heterogeneity variance estimators are not the only component that needs to be carefully chosen according to the characteristics of the meta-analysis data, as there are also a variety of confidence intervals available for the summary effect. Similar to τ^2 estimators, these confidence intervals are sensitive to the probability of events. Some studies have looked at the performance of combinations of τ^2 estimators and summary effect confidence intervals, and found that all confidence intervals perform poorly in terms of coverage and interval length under rare-event scenarios, for each of the τ^2 methods considered. As such, an alternate method to estimate the heterogeneity variance, that performs well in the case of rare-event data, may promote better results in terms of coverage of these confidence intervals, in particular the HKSJ method, which has promising results in all other scenarios.

Chapter 3

Meta-analysis case studies

3.1 Introduction

In this chapter, we shall present the results of several meta-analyses of empirical rare-event data to illustrate the inconsistency in outcomes that result from using fixed versus random-effects approaches, and within the latter different heterogeneity variance (τ^2) estimators. The case-study datasets we analyse are all obtained from published medical clinical trials and contain a number of zero and very sparse events, but vary in terms of sample size and the number of trials included in the meta-analysis. While most of the cases we look at have between 20 and 60 studies, we also look at a meta-analysis of fewer than 5 studies (a more problematic and therefore very specific scenario). This diversity in sample size and structure provides us with a wide range of meta-analysis scenarios to investigate, which we shall later be mimicking in our simulation study.

As we are interested in the performance of τ^2 estimators only in the case of log-risk ratio meta-analyses, we have used this as our outcome measure for all cases considered here. For each case study, we present the τ^2 estimates produced using some of the estimators discussed in Chapter 2, as well as the associated log-risk ratio estimate produced when $\hat{\tau}^2$ is applied to the two-step inverse-variance approach (as discussed in Section 1.5). All heterogeneity variance estimates given in this chapter were calculated using the methods described in Chapter 2, with abbreviations for these estimators listed in Table 2.1, and using code we developed in the statical software package R (R Core Team (2019)), which can be seen in Appendix C. Using each of these estimates for τ^2, we were also able to provide an estimate for the heterogeneity measure I^2, which is calculated using Equation (1.7.2). For each meta-analysis, we also present the value of the log-risk ratio generated using the fixed-effect Mantel-Haneszel approach (outlined in Section 1.9.5). All analyses presented in this chapter were conducted using the standard constant continuity correction method (adding 0.5 to all event counts and sample sizes in studies containing a zero count in either arm), as outlined in Section 1.9.1.

To present the uncertainty of the summary effect measures, we used three alternate confidence intervals: Wald-type, t-distribution and Hartung-Knapp-Sidik-Jonkman, all of which are described in Section 1.6. Calculating various confidence intervals allows us to visualise how the results produced using these methods also vary with alternate estimators of τ^2. For each meta-analysis we display the associated forest plot, created using the *metafor* package in R. Within each forest plot, in addition to presenting the random-effects results produced using some of the τ^2 estimators of significant interest to us, we also display the standard fixed-effect result (assuming $\tau^2 = 0$ when using the inverse-variance approach). While we plot all 3 of the above confidence intervals for each of the random-effects approaches included, we only give the Wald-type confidence interval with the fixed-effect approach. This is because we are more interested in comparing the summary effect confidence intervals in respect to the τ^2 estimators than between fixed and random-effects. While we shall record the log-risk ratio estimates in the results tables within this chapter, we shall plot the exponentiated result of this (the risk ratio) in our forest plots, as this is generally how it would be visually presented in publications to aid easy interpretation of the results.

3.2 Rare-event meta-analyses

3.2.1 Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes

As mentioned in Chapter 1, the most well-explored medical dataset in the area of rare-event meta-analysis methodology is that looking at the effect of the drug rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. Rosiglitazone is an anti-diabetic therapy drug used to lower blood glucose and glycated haemoglobin levels in patients with type 2 diabetes (Shen et al. (2008)). Since over 65% of deaths in patients with this condition are from cardiovascular-related causes, it is of utmost importance to identify if a treatment drug contributes to any microvascular or macrovascular complications, and thus the risks associated with taking such a treatment (Deshpande et al. (2008)). The data for this meta-analysis is taken from the original paper published by Nissen and Wolski (2007), and is displayed in Table 3.1.

As can be seen from the data in Table 3.1, this meta-analysis contains a considerable number of single and double-zero trials, displaying the rare occurrence of the event of myocardial infarctions, and the even rarer event of death from cardiovascular complications. We conducted two meta-analyses: one looking at the risk of myocardial infarctions and the other looking at the risk of death from cardiovascular causes, and their associated forest plots can be seen below in Figures 3.1 and 3.2, respectively. The abbreviations used to detect the estimators in these figures are explained in Table 2.1.

TABLE 3.1: Study data for the meta-analysis on the effect of rosiglitazone; MI refers to myocardial infarctions, Death refers to death from cardiovascular causes, n is the size of the respective study arm.

Study	Treatment arm			Control arm		
	n	MI	Death	n	MI	Death
49653/011	357	2	1	176	0	0
49653/020	391	2	0	207	1	0
49653/024	774	1	0	185	1	0
49653/093	213	0	0	109	1	0
49653/094	232	1	1	116	0	0
100684	43	0	0	47	1	0
49653/143	121	1	0	124	0	0
49653/211	110	5	3	114	2	2
49653/284	382	1	0	384	0	0
712753/008	284	1	0	135	0	0
AVM100264	294	0	2	302	1	1
BRL 49653C/185	563	2	0	142	0	0
BRL 49653/334	278	2	0	279	1	1
BRL 49653/347	418	2	0	212	0	0
49653/015	395	2	2	198	1	0
49653/079	203	1	1	106	1	1
49653/080	104	1	0	99	2	0
49653/082	212	2	1	107	0	0
49653/085	138	3	1	139	1	0
49653/095	196	0	1	96	0	0
49653/097	122	0	0	120	1	0
49653/125	175	0	0	173	1	0
49653/127	56	1	0	58	0	0
49653/128	39	1	0	38	0	0
49653/134	561	0	1	276	2	0
49653/135	116	2	2	111	3	1
49653/136	148	1	2	143	0	0
49653/145	231	1	1	242	0	0
49653/147	89	1	0	88	0	0
49653/162	168	1	1	172	0	0
49653/234	116	0	0	61	0	0
49653/330	1172	1	1	377	0	0
49653/331	706	0	1	325	0	0
49653/137	204	1	0	185	2	1
SB-712753/002	288	1	1	280	0	0
SB-712753/003	254	1	0	272	0	0
SB-712753/007	314	1	0	154	0	0
SB-712753/009	162	0	0	160	0	0
49653/132	442	1	1	112	0	0
AVA100193	394	1	1	124	0	0
DREAM	2635	15	12	2634	9	10
ADOPT	1456	27	2	2895	41	5

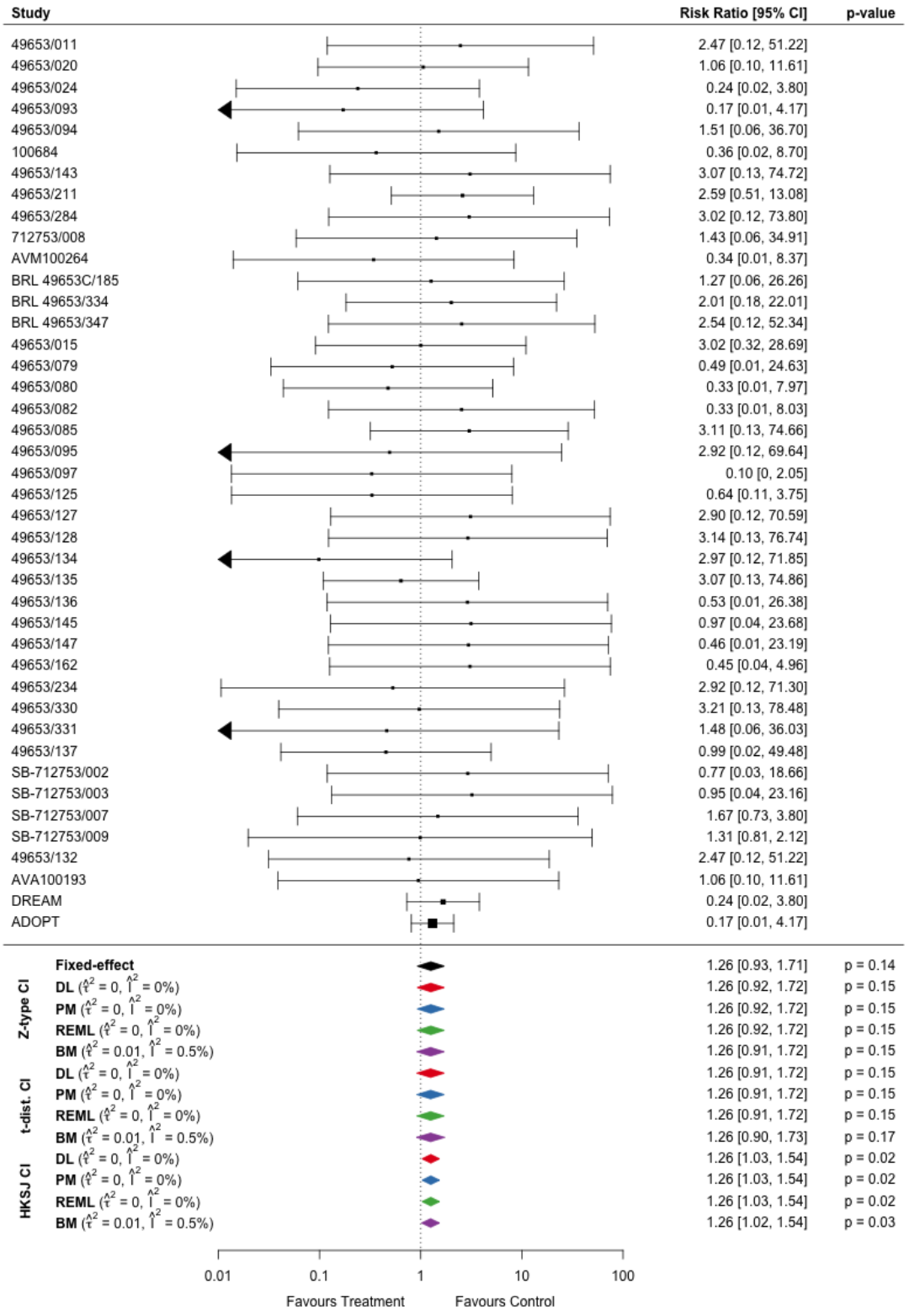


FIGURE 3.1: Forest plot of the risk ratio for myocardial infarctions.

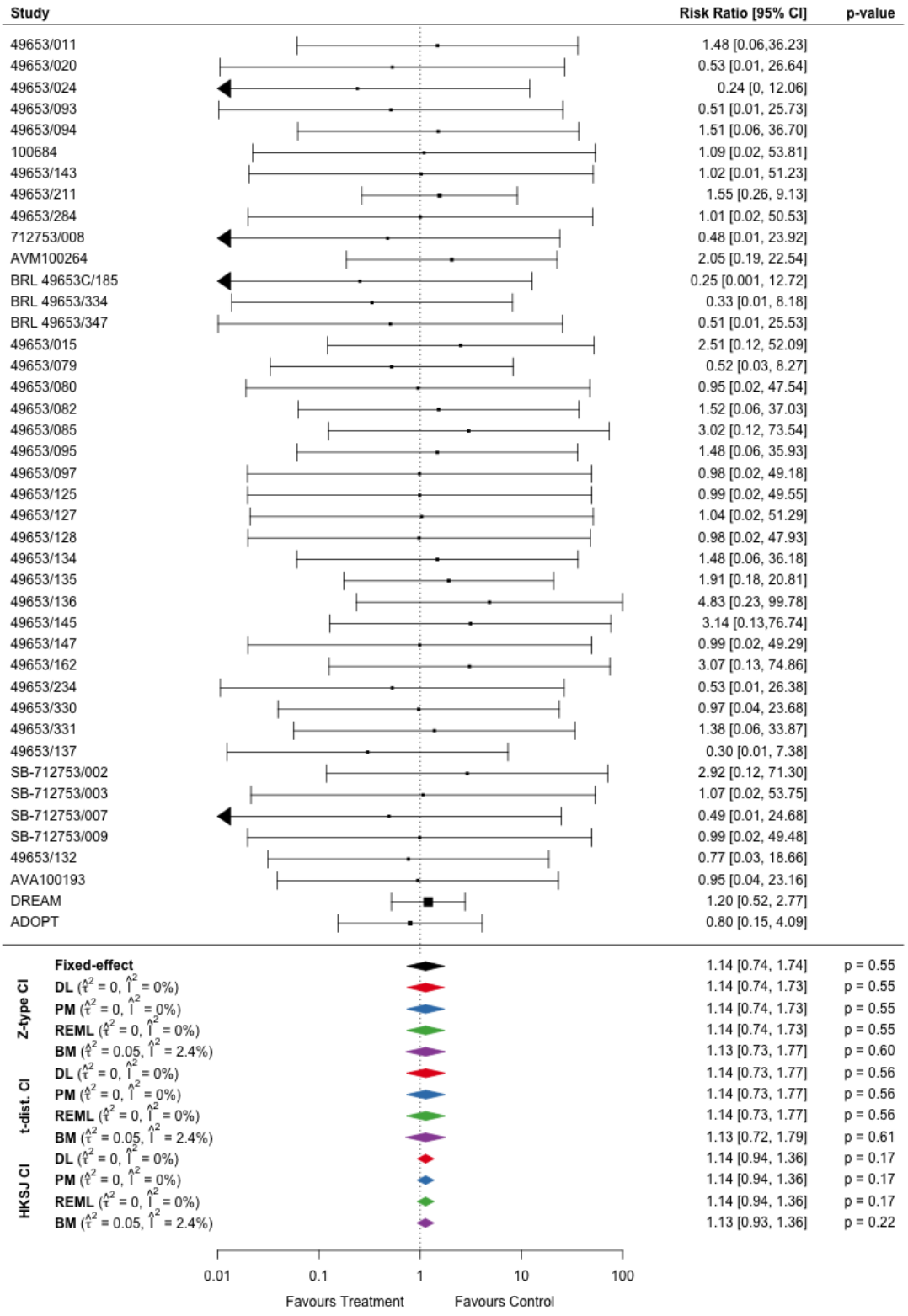


FIGURE 3.2: Forest plot of the risk ratio for death from cardiovascular causes.

These forest plots indicate that both the risk of myocardial infarction and the risk of death by cardiovascular causes are greater in the rosiglitazone treatment group than the control group, as the risk ratio is consistently greater than 1, although this result is not significant for most of the approaches applied here. It can be seen that both meta-analyses appear to have very little heterogeneity present, with the τ^2 and I^2 estimates being zero for 3 out of the 4 τ^2 estimators displayed here (DL, PM, and REML), and with the non-zero Bayes Modal estimates resulting in I^2 estimates of only 0.5% and 2.4% for the two meta-analyses. As a result, the frequentist estimators produce almost identical results to the fixed-effect approach with the Wald (Z)-type interval, and this similarity in results is also seen with the t-distribution confidence interval and the random-effects approaches. The Hartung-Knapp-Sidik-Jonkman (HKSJ) method, however, produces consistently narrower confidence intervals for each of the τ^2 estimators considered in both meta-analyses. This narrowing of intervals is accompanied by smaller p-values, with the associated p-values in the myocardial infarction analysis being significant (< 0.05) for all τ^2 estimators included here.

Finally, we applied all of the heterogeneity variance estimators from Chapter 2 that we deemed appropriate and are interested in comparing, as well as the fixed-effect Mantel-Haenszel approach, to each of these meta-analyses, and the results can be seen below in Tables 3.2 and 3.3. As before, the abbreviations used for the estimators in these tables are given in Table 2.1. As the Mantel-Haenszel approach is fixed-effect, no estimates for τ^2 and I^2 are produced. It should be noted that the Mantel-Haenszel (MH) estimates in these tables are generated using the approach in Section 1.9.5, while the 'Fixed-effect' estimates in the previous figures are produced using the standard FE inverse-variance approach detailed in Section 1.5. As such, while both of these methods are fixed-effect, they represent different methodology.

TABLE 3.2: Heterogeneity variance estimates for the meta-analysis on the effect of rosiglitazone on myocardial infarctions.

Estimator	$\hat{\tau}^2$	$\hat{I}^2$	$\log \widehat{RR}$	Confidence Interval		
				Z-type CI	t-type CI	HKSJ CI
DL	0.00	0.00	0.23	(-0.08, 0.54)	(-0.09, 0.54)	(0.03, 0.43)
DLp	0.01	0.81	0.22	(-0.09, 0.54)	(-0.10, 0.55)	(0.02, 0.43)
DLb	0.00	0.00	0.23	(-0.08, 0.54)	(-0.09, 0.54)	(0.03, 0.43)
HO	0.00	0.00	0.23	(-0.08, 0.54)	(-0.09, 0.54)	(0.03, 0.43)
PM	0.00	0.00	0.23	(-0.08, 0.54)	(-0.09, 0.54)	(0.03, 0.43)
HM	0.09	6.56	0.20	(-0.17, 0.57)	(-0.18, 0.58)	(-0.04, 0.43)
HS	0.00	0.00	0.23	(-0.08, 0.54)	(-0.09, 0.54)	(0.03, 0.43)
SJ	0.24	16.54	0.17	(-0.25, 0.59)	(-0.26, 0.60)	(-0.09, 0.43)
ML	0.00	0.00	0.23	(-0.08, 0.54)	(-0.09, 0.54)	(0.03, 0.43)
REML	0.00	0.00	0.23	(-0.08, 0.54)	(-0.09, 0.54)	(0.03, 0.43)
AREML	0.00	0.00	0.23	(-0.08, 0.54)	(-0.09, 0.54)	(0.03, 0.43)
AB	0.00	0.00	0.23	(-0.08, 0.54)	(-0.09, 0.54)	(0.03, 0.43)
RB	0.00	0.00	0.23	(-0.08, 0.54)	(-0.09, 0.54)	(0.03, 0.43)
RB0	0.00	0.00	0.23	(-0.08, 0.54)	(-0.09, 0.54)	(0.03, 0.43)
BM	0.01	0.51	0.23	(-0.09, 0.54)	(-0.10, 0.55)	(0.02, 0.43)
MH	-	-	0.35	(0.05, 0.66)	(0.04, 0.67)	(0.15, 0.56)

These two tables display relatively similar results, as those estimators that produce a zero heterogeneity variance estimate in one meta-analysis also do so in the other. A total of 11 of the 15 estimators included here provide a zero estimate for τ^2, with all of these also then giving a zero estimate for I^2. These zero estimates appear to largely stem from those estimators based on the truncated method of moments and likelihood-based approaches, thus displaying their inability to work well in the case of rare events, as they are generally truncated to zero in such situations. It can also be noted that in both cases, the SJ estimator produces much higher estimates for τ^2 and lower estimates for $\log RR$ than the alternate approaches. The results of the confidence intervals displayed here is consistent with those included in the associated forest plots, as the Wald and t-type methods produce very similar intervals for all τ^2 estimators, while the HKSJ method produces consistently narrower intervals. The fixed-effect Mantel-Haenszel approach produces much larger estimates for the log-risk ratio and associated confidence intervals that do not agree with any of the random-effects approaches, perhaps indicating poor performance as a result of not accounting for heterogeneity.

TABLE 3.3: Heterogeneity variance estimates for the meta-analysis on the effect of rosiglitazone on death from cardiovascular causes.

Estimator	$\hat{\tau}^2$	$\hat{I}^2$	$\log \widehat{RR}$	Confidence Interval		
				Z-type CI	t-type CI	HKSJ CI
DL	0.00	0.00	0.13	(-0.30, 0.55)	(-0.31, 0.57)	(-0.06, 0.31)
DLp	0.01	0.47	0.13	(-0.30, 0.56)	(-0.32, 0.57)	(-0.06, 0.31)
DLb	0.00	0.00	0.13	(-0.30, 0.55)	(-0.31, 0.57)	(-0.06, 0.31)
HO	0.00	0.00	0.13	(-0.30, 0.55)	(-0.31, 0.57)	(-0.06, 0.31)
PM	0.00	0.00	0.13	(-0.30, 0.55)	(-0.31, 0.57)	(-0.06, 0.31)
HM	0.03	1.43	0.12	(-0.31, 0.56)	(-0.32, 0.57)	(-0.06, 0.31)
HS	0.00	0.00	0.13	(-0.30, 0.55)	(-0.31, 0.57)	(-0.06, 0.31)
SJ	0.07	3.29	0.12	(-0.33, 0.57)	(-0.34, 0.58)	(-0.07, 0.31)
ML	0.00	0.00	0.13	(-0.30, 0.55)	(-0.31, 0.57)	(-0.06, 0.31)
REML	0.00	0.00	0.13	(-0.30, 0.55)	(-0.31, 0.57)	(-0.06, 0.31)
AREML	0.00	0.00	0.13	(-0.30, 0.55)	(-0.31, 0.57)	(-0.06, 0.31)
AB	0.00	0.00	0.13	(-0.30, 0.55)	(-0.31, 0.57)	(-0.06, 0.31)
RB	0.00	0.00	0.13	(-0.30, 0.55)	(-0.31, 0.57)	(-0.06, 0.31)
RB0	0.00	0.00	0.13	(-0.30, 0.55)	(-0.31, 0.57)	(-0.06, 0.31)
BM	0.05	2.41	0.12	(-0.32, 0.57)	(-0.33, 0.58)	(-0.07, 0.31)
MH	-	-	0.53	(0.10, 0.95)	(0.09, 0.97)	(0.30, 0.75)

3.2.2 Catheter-related bloodstream infection

For our next example, we will focus on the scenario where we have fewer studies but similarly sparse events, using data from a meta-analysis conducted by Niel-Weise et al. (2007) on the effect of anti-infective-treated central venous catheters versus standard catheters on catheter-related bloodstream infection (CRBSI) in the acute care setting. This meta-analysis consists of 18 clinical trials, and again includes both single and double-zero trials. The full list of data from this study is reproduced in Table 3.4.

TABLE 3.4: Study data for the meta-analysis on the effect of anti-infective-treated catheter in comparison to standard catheter; CRBSI refers to catheter-related bloodstream infection events, n is the size of the respective study arm.

Study	Treatment arm		Control arm	
	CRBSI	n	CRBSI	n
Bach et al. 1996	0	116	3	117
George et al. 1997	1	44	3	35
Maki et al. 1997	2	208	9	195
Raad et al. 1997	0	130	7	136
Heard et al. 1998	5	151	6	157
Collin 1999	1	98	4	139
Hannan et al. 1999	1	174	3	177
Marik et al. 1999	1	74	2	39
Pierce et al. 2000	1	97	19	103
Sheng et al. 2000	1	113	2	122
Chatzinikolaou et al. 2003	0	66	7	64
Corral et al. 2003	0	70	1	58
Brun-Buisson et al. 2004	3	188	5	175
Leon et al. 2004	6	187	11	180
Yücel et al. 2004	0	118	0	105
Moretti et al. 2005	0	252	1	262
Rupp et al. 2005	1	345	3	362
Osma et al. 2006	4	64	1	69

A forest plot for CRBSI events is provided in Figure 3.3, displaying a visual impression of the distribution of the risk ratio across included studies. This plot tells us that the risk of a CRBSI event in the anti-infective-treated catheter group is lower than that of the group assigned the standard catheter, implying that the treated catheter has a positive impact on patients by reducing the occurrence of infections. This benefit of the treatment is highly significant in this case, as the p-values are < 0.01 for all approaches included here. As before, all frequentist estimators produce zero estimates for τ^2, and consequently I^2, resulting in their risk ratio estimates being very similar to that produced from the fixed-effect approach. The Bayes Model estimator again produces the only non-zero estimate for τ^2 (0.05), with an I^2 estimate of 5.2%, however the resulting risk ratio estimate does not differ much from the alternate methods. In this case, all confidence intervals produce almost identical results for all estimators applied.

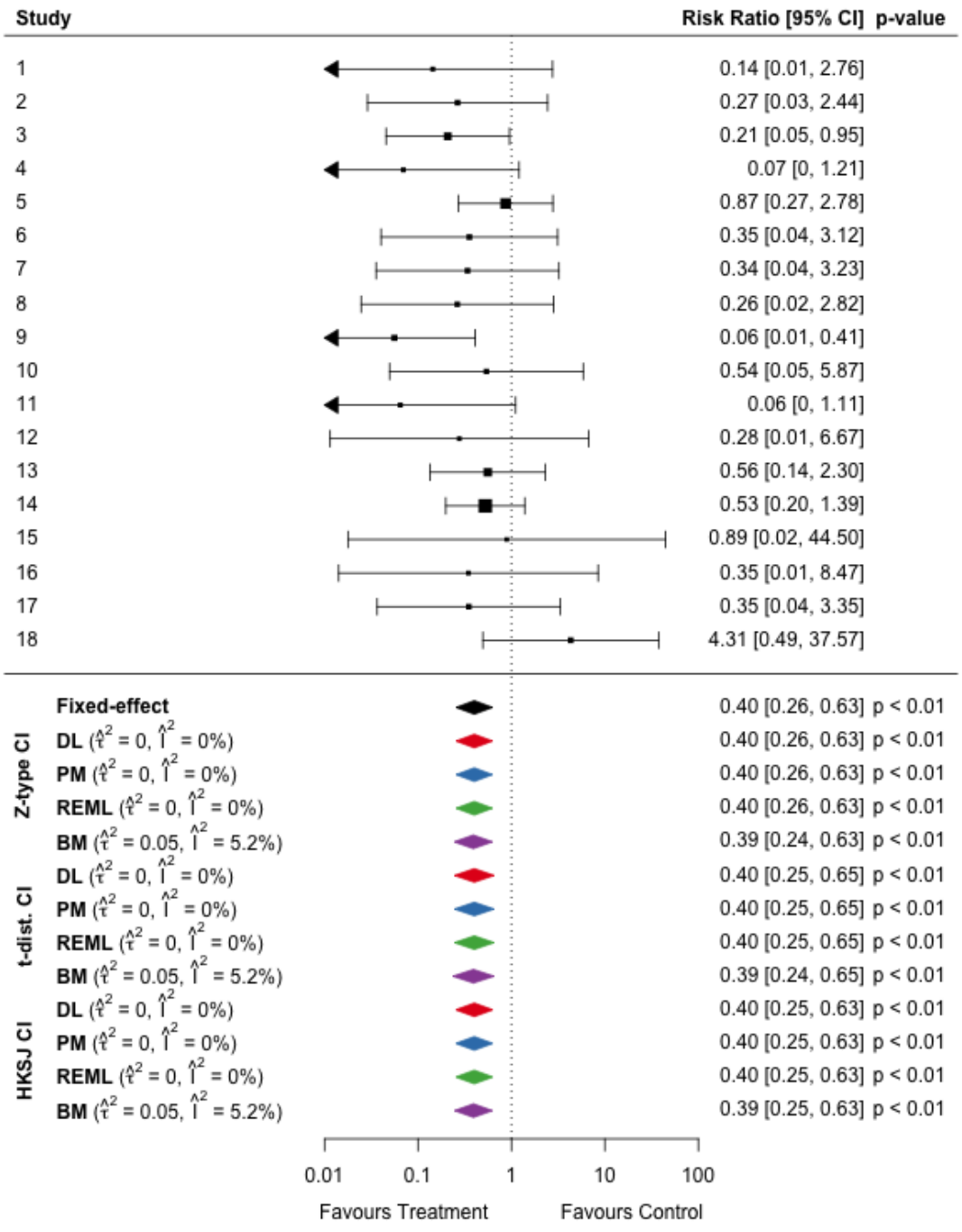


FIGURE 3.3: Forest plot of the risk ratio for CRBSI events.

As with the previous case studies, we calculated many of the appropriate τ^2 estimators from Chapter 2 for this meta-analysis, and have listed a summary of their estimates and associated summary effect results in Table 3.5. In this case, 10 of the 15 approaches have

produced zero estimates for τ^2, and subsequently I^2. Again it can be noted that a number of the truncated estimators have produced zero estimates, again from the truncated method of moments and likelihood-based approaches. The estimate of I^2 is considerably larger for the HM and SJ estimators, at 21.9% and 32.3% respectively, indicating that significant heterogeneity may be present but simply not being detected. For this meta-analysis, all confidence intervals produced very similar results per τ^2 estimator, with no one method producing wider or narrower intervals. As before, the fixed-effect Mantel-Haenszel approach produced a log-risk ratio estimate that was remarkably different from those produced using the random-effects approaches.

TABLE 3.5: Heterogeneity variance estimates for the meta-analysis on the effect of anti-infective-treated catheter in comparison to standard catheter.

Estimator	$\hat{\tau}^2$	$\hat{I}^2$	$\log \widehat{RR}$	Confidence Interval		
				Z-type CI	t-type CI	HKSJ CI
DL	0.00	0.00	-0.92	(-1.36, -0.47)	(-1.40, -0.43)	(-1.37, -0.46)
DLp	0.01	1.00	-0.92	(-1.37, -0.47)	(-1.41, -0.43)	(-1.38, -0.46)
DLb	0.07	6.69	-0.94	(-1.42, -0.47)	(-1.45, -0.43)	(-1.41, -0.47)
HO	0.00	0.00	-0.92	(-1.36, -0.47)	(-1.40, -0.43)	(-1.37, -0.46)
PM	0.00	0.00	-0.92	(-1.36, -0.47)	(-1.40, -0.43)	(-1.37, -0.46)
HM	0.28	21.93	-0.99	(-1.53, -0.45)	(-1.57, -0.41)	(-1.48, -0.50)
HS	0.00	0.00	-0.92	(-1.36, -0.47)	(-1.40, -0.43)	(-1.37, -0.46)
SJ	0.47	32.20	-1.02	(-1.61, -0.43)	(-1.65, -0.39)	(-1.51, -0.52)
ML	0.00	0.00	-0.92	(-1.36, -0.47)	(-1.40, -0.43)	(-1.37, -0.46)
REML	0.00	0.00	-0.92	(-1.36, -0.47)	(-1.40, -0.43)	(-1.37, -0.46)
AREML	0.00	0.00	-0.92	(-1.36, -0.47)	(-1.40, -0.43)	(-1.37, -0.46)
AB	0.00	0.00	-0.92	(-1.36, -0.47)	(-1.40, -0.43)	(-1.37, -0.46)
RB	0.00	0.00	-0.92	(-1.36, -0.47)	(-1.40, -0.43)	(-1.37, -0.46)
RB0	0.00	0.00	-0.92	(-1.36, -0.47)	(-1.40, -0.43)	(-1.37, -0.46)
BM	0.05	5.21	-0.94	(-1.41, -0.47)	(-1.44, -0.43)	(-1.40, -0.47)
MH	-	-	-1.18	(-1.63, -0.73)	(-1.66, -0.70)	(-1.66, -0.70)

3.2.3 Prophylactic antibiotics in caesarean section

Our next example is based on data obtained from the study by Hofmeyr and Smaill (2002), detailing the effects of prophylactic antibiotic treatment on the incidence of wound infection in women undergoing caesarean delivery, via the comparison of an antibiotic treatment group with a placebo group. This meta-analysis comprises of a much larger total of 61 studies, but again includes very rare events of infections in both treatment and control arms of the included studies, with the presence of both single and double-zero studies. The data for these 61 trials can be seen in Table 3.6.

TABLE 3.6: Study data for the meta-analysis on the effect of antibiotic prophylaxis for caesarean section; n is the size of the respective study arm.

Study	Treatment arm		Placebo arm		Study	Treatment arm		Placebo arm	
	Infection	n	Infection	n		Infection	n	Infection	n
Adeleye et al. 1981	11	58	14	48	Leonetti et al. 1989	0	100	1	50
Bibi et al. 1994	4	133	28	136	Levin et al. 1983	0	85	3	43
Chan et al. 1989	27	299	12	101	Lewis et al. 1990	1	36	1	25
Conover et al. 1984	2	68	1	56	Lewis et al. 1990	2	76	4	75
Cormier et al. 1989	5	55	8	55	Mahomed et al. 1988	12	115	15	117
Dashow et al. 1986	3	100	0	33	Mallaret et al. 1990	6	136	16	130
Dashow et al. 1986	4	183	3	44	McCowan et al. 1980	9	35	7	38
De Boer et al. 1989	1	11	5	17	Miller et al. 1968	13	150	23	150
De Boer et al. 1989	10	80	21	74	Moodley et al. 1981	2	40	4	20
Dillon et al. 1981	0	46	4	55	Moro et al. 1974	0	74	2	74
Duff et al. 1980	0	26	1	31	Padilla et al. 1983	0	34	5	37
Duff et al. 1982	0	42	0	40	Phelan et al. 1979	2	61	2	61
Elliot et al. 1986	0	119	1	39	Polk et al. 1982	3	146	9	132
Engel et al. 1986	1	50	9	50	Rehu et al. 1980	4	88	4	40
Fugere et al. 1983	2	60	6	30	Roex et al. 1986	1	64	7	65
Gall 1979	1	46	1	49	Ross et al. 1984	7	57	7	58
Gerstner et al. 1980	3	53	9	50	Rothbard et al. 1975	0	16	1	16
Gibbs et al. 1972	0	33	4	28	Rothbard et al. 1975	2	31	6	37
Gibbs et al. 1973	0	34	6	34	Ruiz-Moreno et al. 1991	1	50	4	50
Gibbs et al. 1981	0	50	2	50	Saltzman et al. 1985	1	50	2	49
Gordon et al. 1979	0	78	1	36	Schedvins et al. 1986	2	26	0	27
Hager et al. 1983	1	43	1	47	Stage et al. 1983	3	133	12	66
Hagglund et al. 1989	0	80	3	80	Stiver et al. 1983	6	244	17	117
Harger et al. 1981	2	196	14	190	Tully et al. 1983	1	52	2	61
Hawrylyshyn et al. 1983	2	124	2	58	Tzingounis et al. 1982	2	46	4	50
Ismail et al. 1990	2	74	8	78	Weissberg et al. 1971	0	40	3	40
Jakobi et al. 1994	4	167	5	140	Wong et al. 1978	2	48	3	45
Karhunen et al. 1985	2	75	9	77	Work et al. 1977	3	40	1	40
Kreutner et al. 1978	0	48	2	49	Yip et al. 1997	1	160	1	160
Kristensen et al. 1990	0	102	1	99	Young et al. 1983	1	50	4	50
Lapas et al. 1989	1	50	10	50	-	-	-	-	-

The forest plot produced for this meta-analysis can be seen in Figure 3.4. Although our forest plots contain double-zero studies, it is important to note that other packages such as *metan* in STATA eliminate those double-zero trials by default, demonstrating the protocol of some statistical software packages to ignore such studies. Here it can be seen that the use of prophylactic antibiotics is favoured highly significantly (all $p < 0.01$), as the risk of contracting a wound infection after having a caesarean section is lower in the treatment group than the placebo group. In this case, all 4 τ^2 estimators considered have produced non-zero estimates, thus all are in agreement that heterogeneity is present to some degree. However, the degree of this heterogeneity varies between estimators, as the MM-based DL and PM methods result in I^2 estimates of only 3.7% and 2.4%, whereas the REML and BM methods are associated with $\hat{I}^2$ of 20.9% and 22.2% respectively. As such, the risk ratio estimates of the two MM-based estimators are similar to that of the fixed-effect approach, but differ slightly from that of the others. The confidence interval results mirror those seen in the previous example, with all methods producing very similar width intervals.

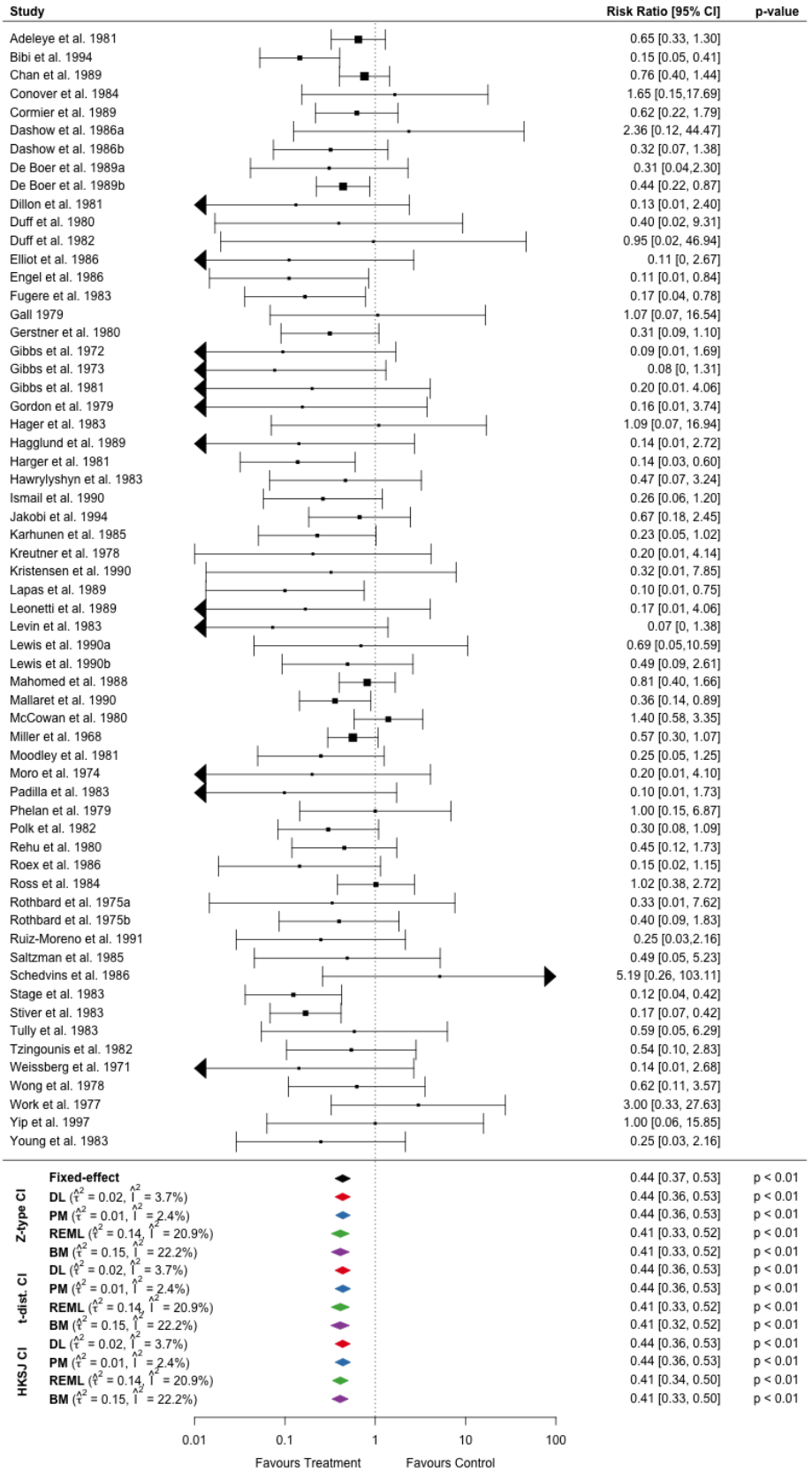


FIGURE 3.4: Forest plot of the risk ratio for infection after caesarean section.

As before, τ^2 estimates were calculated using a larger selection of methods than those plotted, and these can be seen in Table 3.7. These results are more diverse than those of the previous examples, with only 3 estimators producing zero τ^2 and I^2 estimates - the HO, RB and RB0 methods. The reason for the increase in the number of non-zero estimates is likely to be related to the large number of studies, the relatively large sample sizes of these studies, and the fact that only one double-zero trial is present in this dataset. With regards to those estimators producing non-zero τ^2 estimates, it can be seen that there is some agreement between different types of estimators, with the truncated MM-based approaches all resulting in I^2 estimates of $2.4 - 6.7\%$ and the ML-based approaches all producing respective estimates of $19.6 - 20.9\%$. The non-truncated SJ estimator, however, produces a considerably larger estimate for τ^2, resulting in a $\hat{I}^2$ of 39.7%. For all approaches tabulated here, the confidence intervals produced using the three alternate methods are almost identical. As before, the Mantel-Haenszel log-risk ratio estimate is fairly low compared to the others, however in this case it is not that different to the estimate resulting from the SJ estimator, despite the high heterogeneity estimated by the former approach.

TABLE 3.7: Heterogeneity variance estimates for the meta-analysis on the effect of antibiotic prophylaxis for caesarean section.

Estimator	$\hat{\tau}^2$	$\hat{I}^2$	$\log \widehat{RR}$	Confidence Interval		
				Z-type CI	t-type CI	HKSJ CI
DL	0.02	3.67	-0.83	(-1.02, -0.64)	(-1.02, -0.64)	(-1.02, -0.64)
DLp	0.02	3.67	-0.83	(-1.02, -0.64)	(-1.02, -0.64)	(-1.02, -0.64)
DLb	0.04	6.71	-0.84	(-1.04, -0.65)	(-1.04, -0.64)	(-1.04, -0.65)
HO	0.00	0.00	-0.81	(-0.99, -0.63)	(-1.00, -0.63)	(-1.00, -0.62)
PM	0.01	2.39	-0.82	(-1.01, -0.64)	(-1.02, -0.63)	(-1.02, -0.63)
HM	0.19	26.18	-0.90	(-1.14, -0.67)	(-1.15, -0.66)	(-1.11, -0.70)
HS	0.01	2.01	-0.82	(-1.01, -0.64)	(-1.01, -0.63)	(-1.01, -0.63)
SJ	0.36	39.73	-0.93	(-1.20, -0.67)	(-1.21, -0.66)	(-1.14, -0.72)
ML	0.13	19.57	-0.89	(-1.11, -0.66)	(-1.11, -0.66)	(-1.09, -0.68)
REML	0.14	20.86	-0.89	(-1.12, -0.66)	(-1.12, -0.66)	(-1.09, -0.69)
AREML	0.14	20.32	-0.89	(-1.11, -0.66)	(-1.12, -0.66)	(-1.09, -0.69)
AB	0.02	3.67	-0.83	(-1.02, -0.64)	(-1.02, -0.64)	(-1.02, -0.64)
RB	0.00	0.00	-0.81	(-0.99, -0.63)	(-1.00, -0.63)	(-1.00, -0.62)
RB0	0.00	0.00	-0.81	(-0.99, -0.63)	(-1.00, -0.63)	(-1.00, -0.62)
BM	0.15	22.20	-0.89	(-1.12, -0.66)	(-1.13, -0.66)	(-1.10, -0.69)
MH	-	-	-0.95	(-1.13, -0.77)	(-1.14, -0.76)	(-1.14, -0.76)

3.2.4 Mortality of human albumin solution

Finally, we look at the effect of administering human albumin, or plasma protein fraction, on increasing the number of deaths of patients that are critically ill. Human albumin solution is administered to patients who have suffered an acute loss of plasma volume as is the case following burn injuries, or to individuals with severe hypoalbuminaemia as a supplemental therapy in liver disease (Alderson et al. (2002)). Reviewers (1998) conducted a systematic review investigating the relationship between the use of this solution and the risk of death, after a number of conflicting studies had been published on the matter (Goldwasser and Feldman (1997)). The data from the meta-analysis by Reviewers (1998) can be seen in Table 3.8 below.

TABLE 3.8: Study data for the meta-analysis on mortality of human albumin solution for resuscitation in critically ill patients; n is the size of the respective study arm.

Study ID	Treatment arm		Control arm		Study ID	Treatment arm		Control arm	
	Death	n	Death	n		Death	n	Death	n
Lowe et al. 1977	3	57	3	84	Woittiez 1998	8	15	4	16
Shah et al. 1977	2	9	3	11	Jelenko et al. 1979	1	7	2	7
Lucas et al. 1978	7	27	0	25	Goodwin et al. 1983	11	40	3	39
Virgilio et al. 1979	1	15	1	14	Greenhalgh et al. 1995	7	34	3	36
Boutros et al. 1979	0	7	2	17	Bland et al. 1976	4	14	1	13
Zetterstrom et al. 1981	0	15	1	15	Nilsson et al. 1980	1	29	0	30
Zetterstrom 1981	2	9	0	9	Brown et al. 1988	6	34	4	33
Grundmann et al. 1982	1	14	0	6	Foley et al. 1990	7	18	6	22
Rackow et al. 1983	6	9	6	8	Kanarek et al. 1992	3	12	2	12
Woods et al. 1993	1	37	0	32	Greenough et al. 1993	6	20	4	20
Tølløfsrud et al. 1995	0	10	1	10	Golub et al. 1994	12	116	6	103
So et al. 1997	7	32	5	31	Rubin et al. 1997	2	16	1	15

It can be seen from the table that this dataset contains several single-zero trials (from both trial arms), but no double-zero ones. The associated forest plot for this meta-analysis is shown in Figure 3.6. The varying conclusions regarding the effect of albumin solution on mortality reached in different studies is very apparent in this plot, however the overall conclusion from the meta-analysis is that albumin treatment does not reduce that risk of death over control, and in fact the control performs significantly better than albumin ($p \leq 0.01$ for all methods). All log-risk ratio estimates across fixed and random-effects approaches are very similar, as the only τ^2 estimator to produce a non-zero estimate in this case was BM, which lead to an I^2 estimate of only 1.8%. The confidence intervals produced form the Wald and t-type methods were very similar as before, however the HKSJ method had consistently narrower intervals, resulting in a lower p-values for all τ^2 estimators displayed (p dropped from 0.01 to < 0.01).

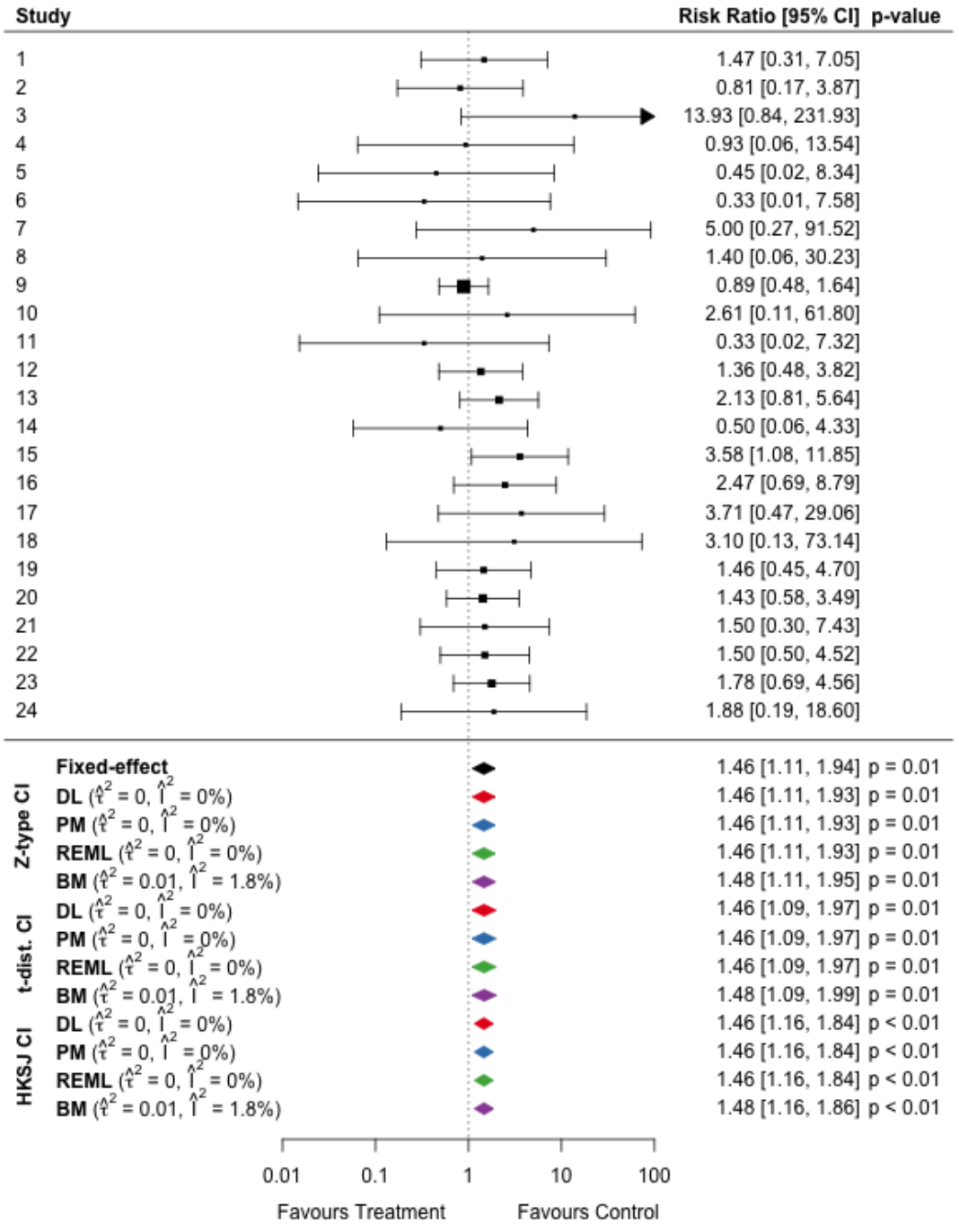


FIGURE 3.5: Forest plot of the risk ratio for mortality in albumin treatment vs. placebo.

Table 3.9 displays the τ^2 and associated log-risk ratio estimates produced for this case study. The τ^2 estimates calculated for this meta-analysis show similar patterns to those seen in the examples discussed previously, with regards to an observable large number

of zero estimates and associated high rates of truncation taking place in the truncated methods. In this case, only 4 of the 15 estimators have produced non-zero τ^2 estimates - the DLp, HM, SJ and BM methods. The level of heterogeneity estimated by these 4 methods, with BM resulting in an I^2 estimate of 1.8%, while the SJ approach generated a significantly greater estimate of 30.6%. As a result, the log-risk ratio estimates also varied accordingly, however the degree to this variation was rather small ($\log \widehat{RR}$ took values between 0.38 and 0.43 for random-effects approaches). The fixed-effect Mantel-Haenszel approach again produced a log-risk ratio estimate of greater magnitude than the random-effects approaches, with a corresponding estimate of 0.55. The confidence intervals generated using the Wald and t-type methods were relatively similar for all meta-analysis approaches considered, however the HKSJ method resulted in consistently narrower intervals, as was seen in the forest plot.

TABLE 3.9: Heterogeneity variance estimates for the meta-analysis on mortality of human albumin solution for resuscitation in critically ill patients.

Estimator	$\hat{\tau}^2$	$\hat{I}^2$	$\log \widehat{RR}$	Confidence Interval		
				Z-type CI	t-type CI	HKSJ CI
DL	0.00	0.00	0.38	(0.10, 0.66)	(0.09, 0.68)	(0.15, 0.61)
DLp	0.01	1.90	0.39	(0.10, 0.67)	(0.09, 0.69)	(0.15, 0.62)
DLb	0.00	0.04	0.38	(0.10, 0.66)	(0.09, 0.68)	(0.15, 0.61)
HO	0.00	0.00	0.38	(0.10, 0.66)	(0.09, 0.68)	(0.15, 0.61)
PM	0.00	0.00	0.38	(0.10, 0.66)	(0.09, 0.68)	(0.15, 0.61)
HM	0.08	12.97	0.41	(0.09, 0.73)	(0.08, 0.75)	(0.17, 0.66)
HS	0.00	0.00	0.38	(0.10, 0.66)	(0.09, 0.68)	(0.15, 0.61)
SJ	0.23	30.59	0.43	(0.06, 0.81)	(0.04, 0.83)	(0.17, 0.69)
ML	0.00	0.00	0.38	(0.10, 0.66)	(0.09, 0.68)	(0.15, 0.61)
REML	0.00	0.00	0.38	(0.10, 0.66)	(0.09, 0.68)	(0.15, 0.61)
AREML	0.00	0.00	0.38	(0.10, 0.66)	(0.09, 0.68)	(0.15, 0.61)
AB	0.00	0.00	0.38	(0.10, 0.66)	(0.09, 0.68)	(0.15, 0.61)
RB	0.00	0.00	0.38	(0.10, 0.66)	(0.09, 0.68)	(0.15, 0.61)
RB0	0.00	0.00	0.38	(0.10, 0.66)	(0.09, 0.68)	(0.15, 0.61)
BM	0.01	1.75	0.39	(0.10, 0.67)	(0.09, 0.69)	(0.15, 0.62)
MH	-	-	0.55	(0.27, 0.83)	(0.25, 0.84)	(0.30, 0.79)

3.3 Meta-analyses with rare events and few studies

In this section, we shall look at the case where there are very few studies in the meta-analysis, i.e. $k \leq 5$, as well as sparse-event data occurring. This is likely to be the most problematic of the scenarios that we investigate, as there is very little data to base

the heterogeneity variance estimates on, thus increasing the likelihood and magnitude of potential bias of the estimators.

3.3.1 Post-transplant lymphoproliferative disease in paediatric liver transplantation

In this example, we shall look at the use of novel biological drugs used to improve success rates and reduce detrimental side-effects following paediatric liver transplantation. Crins et al. (2014) conducted a meta-analysis of clinical trials (both randomised and not) investigating the effect of the interleukin-2 receptor antagonists (IL-2RA) together with the standard baseline therapy of concomitant immunosuppression compared against the baseline treatment alone in children. In particular, they concentrate on studies using the IL-2RA drugs basiliximab and daclizumab. These drugs work as targeted immunosuppressive agents, aiming to decrease the risk of acute rejection, and have already been incorporated into the standard treatment regime for adults. In their analysis, Crins et al. (2014) looked at a number of outcomes, including acute and chronic rejection of the transplanted organ. Here, we shall focus on their example investigating the experimental drugs' effect on post-transplant lymphoproliferative disease (PTLD) - a potentially fatal disorder that can develop in the recipient after transplantation.

TABLE 3.10: Study data for the meta-analysis on post-transplant lymphoproliferative disease in experimental paediatric transplantation vs. control; n is the size of the respective study arm.

Study	Experimental arm		Control arm	
	PTLD	n	PTLD	n
Schuller et al.	0	18	0	12
Ganschow et al.	1	54	0	54
Spada et al.	1	36	1	36

The data from this meta-analysis is displayed in Table 3.10. As can be seen from this table, there were only three studies in this meta-analysis, and they consist of one single-zero trial and one double-zero trial. The sample sizes of these studies are also fairly small (ranging from 12 to 54), and are well-balanced across treatment arms. The associated forest plot for this analysis is displayed in Figure 3.6, and indicates that the risk of post-transplant lymphoproliferative disease is slightly greater in those individuals in the experimental arm compared to those in the control arm, although only slightly, as the risk ratio is close to 1 for all methods. Similar to the examples with $k > 5$, 3 of the 4 estimators plotted here produce zero estimates for τ^2 and I^2, with BM being the only one not to, and in fact resulting in an I^2 estimate of 60.9%. This marked difference between the frequentist and semi-Bayesian approaches included in this plot may correlate

with the fact that Bayesian approaches are the preferred and most researched types of method for meta-analyses with $k < 5$. In contrast to the larger case studies however, the confidence intervals form the Wald-type and HKSJ methods appear to be very similar, while the t-distribution method has produced exceedingly wide intervals in comparison, particularly when combined with the BM estimator.

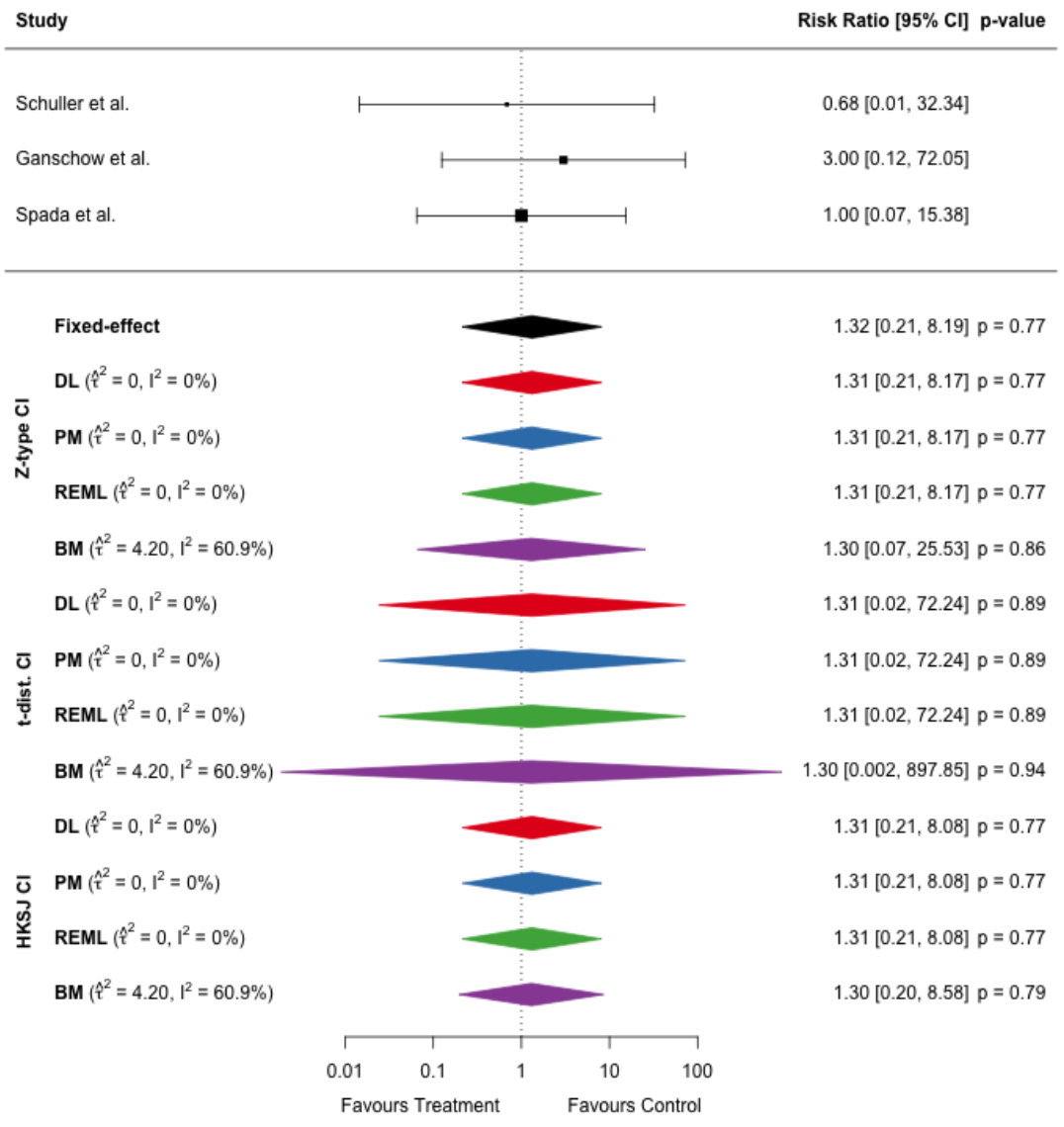


FIGURE 3.6: Forest plot of the risk ratio for post-transplant lymphoproliferative disease in experimental paediatric transplantation vs. control.

When the heterogeneity variance was estimated for this meta-analysis, a total of 11

out of the 15 available methods considered produced a τ^2 estimate of zero, as can be seen in Table 3.11. Similar to the previous examples, only the DLp, HM, SJ and BM estimators produced non-zero estimates for τ^2, but these varied considerably (from 0.01 to 4.20), leading to high variation in their corresponding I^2 estimates (0.37-60.87%). The BM method was responsible this extremely high estimate of $I^2 = 60.87\%$, and this may reflect the recommendation of Bayesian approaches when $k < 5$, as only this semi-Bayesian approach could detect the heterogeneity to this level, or the method could simply be seriously over-estimating τ^2 in this case. Either way, these results reflect the extreme difficulty that all τ^2 estimators encounter when analysing rare-event data where few studies are available. As before, the Mantel-Haenszel approach produced a notably different estimate of the log-risk ratio, with the value being much higher than those produced by the random-effects approaches. In terms of the summary effect confidence intervals, it can be seen that the Wald-type and HKSJ methods produce relatively similar intervals, with the HKSJ producing slightly narrower intervals for all approaches. Meanwhile, the t-distribution method consistently produced significantly wider intervals for this scenario, with the combination of t-distribution and BM methods producing the widest intervals of all. This could represent the inability of the t-distribution method to perform well when there are few studies in the meta-analysis, but may also throw into question the assumed appropriateness of the BM method for such a scenario.

TABLE 3.11: Heterogeneity variance estimates for the meta-analysis on post-transplant lymphoproliferative disease in experimental paediatric transplantation vs. control.

Estimator	$\hat{\tau}^2$	$\hat{I}^2$	$\log \widehat{RR}$	Confidence Interval		
				Z-type CI	t-type CI	HKSJ CI
DL	0.00	0.00	0.27	(-1.55, 2.10)	(-3.73, 4.28)	(-1.54, 2.09)
DLp	0.01	0.37	0.27	(-1.55, 2.10)	(-3.74, 4.29)	(-1.54, 2.09)
DLb	0.00	0.00	0.27	(-1.55, 2.10)	(-3.73, 4.28)	(-1.54, 2.09)
HO	0.00	0.00	0.27	(-1.55, 2.10)	(-3.73, 4.28)	(-1.54, 2.09)
PM	0.00	0.00	0.27	(-1.55, 2.10)	(-3.73, 4.28)	(-1.54, 2.09)
HM	0.05	1.89	0.27	(-1.57, 2.12)	(-3.78, 4.32)	(-1.55, 2.10)
HS	0.00	0.00	0.27	(-1.55, 2.10)	(-3.73, 4.28)	(-1.54, 2.09)
SJ	0.07	2.59	0.27	(-1.58, 2.13)	(-3.79, 4.34)	(-1.55, 2.10)
ML	0.00	0.00	0.27	(-1.55, 2.10)	(-3.73, 4.28)	(-1.54, 2.09)
REML	0.00	0.00	0.27	(-1.55, 2.10)	(-3.73, 4.28)	(-1.54, 2.09)
AREML	0.00	0.00	0.27	(-1.55, 2.10)	(-3.73, 4.28)	(-1.54, 2.09)
AB	0.00	0.00	0.27	(-1.55, 2.10)	(-3.73, 4.28)	(-1.54, 2.09)
RB	0.00	0.00	0.27	(-1.55, 2.10)	(-3.73, 4.28)	(-1.54, 2.09)
RB0	0.00	0.00	0.27	(-1.55, 2.10)	(-3.73, 4.28)	(-1.54, 2.09)
BM	4.20	60.87	0.26	(-2.72, 3.24)	(-6.28, 6.80)	(-1.63, 2.15)
MH	-	-	0.69	(-1.13, 2.52)	(-3.31, 4.70)	(-1.53, 2.91)

3.4 Summary characteristics of rare-event meta-analysis case studies

In this section we shall present a summary of the characteristics of the meta-analysis case studies that we have discussed thus far in this chapter. We are interested in aspects such as the sample sizes for each trial, the number of single-zero and double-zero trials, and the probability of the event of interest in the control and treatment arms. Table 3.12 lists the summary characteristics of the case study meta-analyses, which we shall take into account when deciding on appropriate parameter ranges in the design of our own simulation study. While all meta-analyses listed in this table contain at least one single-zero study (a requirement for them to be used in this chapter), not all include double-zero trials. The mean sample sizes of the studies included in these meta-analyses do not differ significantly between treatment arms, indicating that the examples we have looked at are fairly balanced in regards to this aspect. The mean event probability ranges form 0.001 to 0.2, and we have included both cases where the event probability is greater in the treatment arm, and that when it is greater in the control arm.

TABLE 3.12: Summary characteristics of case study meta-analyses; k is the number of studies, n_0 and n_1 are the sample sizes, and p_0 and p_1 are the event probabilities in the control and treatment arms respectively (presented as mean (SD)).

Meta-analysis	k	SZ	DZ	n_0	n_1	p_0	p_1
Rosig. (MI)	42	26 (62%)	4 (10%)	292.3 (566.6)	370.4 (456.4)	0.004 (0.01)	0.01 (0.01)
Rosig. (death)	42	17 (41%)	19 (45%)	292.3 (566.6)	370.4 (456.4)	0.001 (0.003)	0.003 (0.01)
CRBSI	18	5 (28%)	1 (6%)	138.6 (81.8)	138.6 (76.8)	0.04 (0.1)	0.01 (0.02)
C-section	61	18 (30%)	1 (2%)	63.1 (79.8)	79.8 (38.0)	0.10(0.1)	0.04 (0.05)
Albumin	24	8 (33%)	0 (0%)	25.3 (23.3)	24.8 (23.2)	0.1 (0.2)	0.2 (0.2)
Transplant	3	1 (33%)	1 (33%)	94.3 (61.9)	70.0 (21.5)	0.01 (0.01)	0.02 (0.01)

3.5 Conclusions

From conducting meta-analyses on a range of rare-event medical datasets, we have demonstrated the varying performance of the heterogeneity estimators presently available. In addition to this, we have seen the high number of zero estimates that are produced in these scenarios, particularly by a number of MM-based estimators. The method used to calculate the confidence interval for the summary effect can have a considerable impact on its width and the associated significance of the result. We focused only on clinical trials because our interest is solely in the log-risk ratio outcome measure, for which case-control studies are not appropriate.

In all cases, we observed that the SJ estimator behaved very differently to the other methods - producing considerably larger estimates of τ^2, sometimes with considerably larger estimates of τ^2, sometimes with correspondingly lower estimates of $\log RR$. This pattern relates to the difference in methodology of this approach, and agrees with some previously published results regarding its detection of high heterogeneity (Sidik and Jonkman (2005)). It would be interesting to see whether this pattern is repeated for scenarios other than those represented here, and so we shall ensure to include this estimator in our simulation study.

For the meta-analyses containing greater than 5 studies, we found that in most cases the HKSJ method produced the shortest interval of those investigated for all τ^2 estimators considered, and that the Wald-type and t-distribution methods produced similar results. In the case study involving fewer than 5 studies, this pattern of results was dramatically different. In this specialist scenario, the HKSJ approach produced the shortest intervals for all estimators considered, with the t-distribution approach resulting in the widest intervals. In particular, the Bayes Modal τ^2 estimator combined with the t-distribution produced exceedingly large intervals, with associated large p-value.

Finally, we have also summarised the characteristics of the case studies that we analysed in this chapter. This allowed us to generate a feel of the average probability of events, number of zero events, sample size and number of studies for real rare-event meta-analyses. As such, we shall be able to incorporate these findings when designing our own simulation study in order to best represent the type and variety of data that can be encountered.

Chapter 4

Generalised linear mixed models to estimate heterogeneity variance

4.1 Introduction

The aim of our research is to design and compare methods to estimate heterogeneity variance when conducting a meta-analysis of rare-event trials. So far in this thesis, we have focused only on meta-analysis using a Normal-based random-effects model, and the different methods to estimate the heterogeneity variance (τ^2) required for this model. In the case of rare events in binary outcome data, special care is needed when using the standard inverse-variance approach described in Section 1.5, as the degree of bias in the τ^2 estimation is proportional to the rarity of the study events. As a result of this, it has been suggested that the inverse-variance method should be avoided altogether (Veroniki et al. (2016)).

There exist several alternative methods that have been proposed to analyse such sparse-event data. These include using the fixed-effect model with either the Mantel-Haenszel method for unbalanced treatment and control group sizes, or Peto's method for balanced group sizes, as discussed in Section 1.9.5. However, these methods have their own drawbacks, namely that they do not incorporate the potential heterogeneity, and Peto's method can only be applied for the odds ratio outcome measure. In Chapter 3, we saw how estimates from the Mantel-Haenszel approach differed sometimes quite considerably from those produced using the random-effects approach when applied to our rare-event meta-analysis case studies. Although the results differed, these differences were not to the extent of changing the overall conclusion, and may have occurred as a result of the Mantel-Haenszel approach not requiring continuity corrections for zero event counts (except for the case of all-study single-arm zeros). In general, however, our

empirical analysis demonstrated the high proportion of zero estimates produced with sparse data, and the significant differences in results produced between fixed-effect and random-effects analysis with certain heterogeneity variance (τ^2) estimators, potentially indicative of their inability to perform well in such scenarios.

As a result of the lack of confidence and inconsistency in existing meta-analysis approaches, it has been suggested to use methods or models based on exact distributional assumptions (Böhning et al. (2015)). Generalised linear mixed models (GLMMs) have a number of benefits over existing approaches, including their use of a binomial-normal likelihood, and their potential to produce more accurate inference in a single 1-step approach. A considerable amount of work has recently been conducted in regards to comparing the use of a variety of GLMMs in the case of odds ratio meta-analysis, however little has been done for the risk ratio or its log, and the range of models under investigation has been limited to a select few (Bakbergenuly and Kulinskaya (2018); Jackson et al. (2018)). In addition, little has been done to specifically compare these models in terms of their performance in estimating the τ^2 parameter - a key component of random-effects meta-analyses.

In this chapter, we shall discuss the general use of GLMMs to conduct meta-analyses, and demonstrate the application of GLMMs with fixed and random intercepts. We will then outline two novel GLMM-based approaches that we are proposing to estimate τ^2 for the case of rare-event log-risk ratio outcomes - Poisson mixed regression models with a random effect on the treatment parameter, and conditional logistic mixed regression models. Alternative GLMMs that have been previously proposed elsewhere for alternative settings will then briefly be discussed. As GLMMs can face difficulties in terms of convergence, particularly with sparse data, we will outline the model-based options we used when applying our two chosen models, and list the scenarios where these models cannot be applied by definition. Finally, we shall look at the results produced when applying our chosen models to the case studies portrayed in Chapter 3, looking for any differences between statistical software packages and integration methods, and comparing with the results produced using the estimators from Chapter 2, before summarising on the use of the models in general.

4.2 The use of generalised linear mixed models for meta-analysis

Generalised linear mixed models (GLMMs) are an alternative to standard meta-analysis approaches, that theoretically have the potential to overcome the disadvantages of the latter through their use of a binomial-normal likelihood. They are an extension of generalised linear models that include random effects in addition to fixed effects, and where the likelihood is used for the inference. Despite their potential positive attributes,

these models are rarely used in practice as a result of their complex nature, and the associated difficulty in applying such models in standard statistical software packages. In addition, there exists little in the way of published recommendations regarding the types of model suitable for specific data types such as rare-event scenarios.

However, they have recently gained growing attention in this field, in an attempt to determine whether their theoretical benefits have standing in real-life data analysis. In particular, Bakbergenuly and Kulinskaya (2018) and Jackson et al. (2018) have looked at the performance of several GLMMs in conducting odds ratio meta-analyses, compared with the standard and fixed-effect and random-effects inverse-variance approaches. In addition, there is now scope for their application, as GLMMs have recently become much easier to implement in known software packages, giving the opportunity for recommendations to be made regarding these now simple-to-use models. For example, the R package *metafor*, used for applying meta-analyses, now has four GLMM methods available for application.

As the availability of individual-level data for studies is continuously increasing, and this is our area of interest, we shall focus on modelling individual-level binary outcome data, although summary-data can also be incorporated into such models. We shall be concentrating on log-risk ratio meta-analyses here, unless otherwise specified, as this is an easily interpretable measure, however other outcome measures can be used with the models with some modifications. For all the models we propose, we shall also assume that the true effect sizes, θ_i, are normally distributed between the k studies in the meta-analysis, however alternate mixing distributions can be used for the random effects, including the beta-binomial model.

4.2.1 Benefits of generalised linear mixed models

The use of GLMMs for the estimation of τ^2 in meta-analyses has several benefits over the alternative estimators described in Chapter 2, particularly for the case of sparse-event data with zero counts. Firstly, GLMMs can account for the standard errors of the outcome measure (or the square root of the within-study variances), denoted σ_i, that need to be estimated from the data, and which other methods incorrectly assume to be known. In addition to this, these models can account for any correlation between the outcome measure estimates ($\hat{\theta}_i$) and their associated standard error estimates ($\hat{\sigma}_i$), which can have a profound effect on the overall outcome of the analysis if significant correlation is present but not accounted for. The majority of existing methods require continuity corrections, such as those outlined in Section 1.9.1, when dealing with single and double-zero studies, and these have been shown to be associated with significant bias (Sweeting et al. (2004)). Based on their structure, GLMMs have the ability to avoid the use of such (arbitrary) continuity corrections in the case of zero events, thereby avoiding the corresponding bias. Finally, these models can avoid the use of the normal

approximation, which is generally violated in the case of studies with few events, by assuming alternate, more suitable distributions for the data.

4.2.2 Theory behind the general case

We shall now describe the theory behind the general case of applying GLMMs to conduct meta-analyses, as discussed in Bakbergenuly and Kulinskaya (2018). Let y_i be the univariate observation from study i, and the covariate vectors x_i and z_i (with dimensions p and q) represent the fixed and random effects of study i respectively, where $i = 1, \ldots, k$. Assume that the observations y_i are independent with conditional means and variances given by:

$$E(y_i|b_i) = \mu_i(b_i)$$

$$Var(y_i|b_i) = \Delta a_i \nu(\mu_i(b_i))$$

where b_i is a random effect, Δ is the dispersion parameter, a_i is some known constant, and $\nu(\cdot)$ is a variance function (Breslow and Clayton (1993)). These paired conditional means and variances have a mean-variance relationship, and it can be seen that both depend on the random effect b_i. Now, given the q-dimensional vector of random effects, $b = (b_1, \ldots, b_q)$, the general GLMM in this case has the form:

$$\eta_i^b(b) = x_i^t \beta + z_i^t b \tag{4.2.1}$$

where t is the matrix transpose and β is the regression parameter vector. Similar to the case of the more simplified generalised linear model, the conditional mean $E(y_i|b_i)$ is associated with a linear predictor via some link function, which is given by $g(\mu_i(b_i)) = \eta_i(b_i))$. Inverting this link function, we have $H = g^{-1}$, and using X and Z to represent the design matrices with rows x_i^t and z_i^t, respectively, the overall conditional mean can be written as:

$$E(y|b) = H(X\beta + Zb)$$

where $y = (y_1, \ldots, y_k)$. The random effect b will follow some distribution, generally a multivariate normal, with a mean of 0 and variance-covariance matrix $D = D(\zeta)$, for some unknown variance component vector ζ. Breslow and Clayton (1993) considered Poisson, binomial and hypergeometric models for the conditional distribution of y_i, with the conditional variance containing a dispersion parameter of $\Delta = 1$. Over-dispersion

can be modelled by using $\Delta > 1$, where Δ is jointly estimated with the variance component vector ζ via $D = D(\zeta)$.

With GLMMs, the maximum likelihood approach is used to estimate the parameters. However, as the model in this case is nonlinear and contains random effects, the maximum likelihood approach would involve a marginal distribution requiring integration with respect to the unobservable random effects. As a result, this integration generally does not have a closed form, and so an analytic solution is not possible. In order to evaluate this integral, alternative estimation techniques must be applied, which are then also used in the approximation of the corresponding log-likelihood function, information matrices and score equations. Such alternative numerical methods include adaptive Hermite quadrature, Laplace's approach and higher order approximations, and penalised quasi-likelihood and equivalent pseudo-likelihood methods (Breslow and Clayton (1993); Demidenko (2013)). Alternative approaches to fit GLMMs include Bayesian methods involving Markov chain Monte Carlo (MCMC) stochastic integration, and Gibbs sampling, with a number of hybrid methods also being available (Capanu et al. (2013)). Methods based on the moment-based generalised estimation equation are also available for the estimation of population-average parameters in marginal models.

If x_i represent binary outcomes and $g(\cdot)$ is the logit link function, then the model given in Equation (4.2.1) represents a random-intercept logistic regression model. When this is applied to meta-analysis data, then the intercept represents the study-specific effects, and the slope of the treatment arm indicator variable corresponds to the treatment effect. For these GLMMs applied to meta-analyses, the treatment effect is generally set to be random, and the intercepts (study-specific effects) can either be assumed to be fixed or random.

For binary outcomes, conditional on the fixed effects the outcomes are generally assumed to follow a binomial or non-central hypergeometric distribution, whereas the random effects are assumed to be normally distributed, corresponding to a binomial-normal or hypergeometric-normal likelihood, respectively. In comparison, the standard random-effects meta-analysis approach discussed in Section 1.4.2 assumes the distribution of log-risk ratios to be normally distributed, resulting in the normal-normal model. Alternative models have been proposed for varying outcome measures, for example with incidence rates, where the Poisson-normal model has been suggested.

4.2.3 GLMMs with fixed intercept

As mentioned in the previous section, meta-analysis GLMMs with a fixed intercept fall into the category of mixed-effects logistic regression models, and are able to account for the heterogeneity between studies (Turner et al. (2000)). Here we shall base these models on the log-odds ratio outcome measure, as this provides the basic structure from

which the corresponding log-risk ratio models can be built upon, as we shall discuss later in the chapter with our proposed methods. In this case, the fixed-intercept model for the log-odds ratio can be written as:

$$y_{ij}|p_{ij} \sim Binomial(n_{ij}, p_{ij})$$

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha_i + (\theta + \beta_i) \times j$$

where p_{ij} are the treatment arm-specific event probabilities, α_i is the log-odds of the control arm (i.e. $\alpha_i = \log(p_{i0}/(1-p_{i0}))$), θ is the overall effect size (log-odds ratio), and β_i are the deviations of the study-specific treatment effect from θ, with $j = 1$ representing the active treatment arm and $j = 0$ otherwise, for $i = 1, \ldots, k$. Here, the fixed intercept is represented by α_i, while $\beta_i \sim N(0, \tau^2)$ corresponds to the random effects, where τ^2 is the between-study variance. As such, the above model can be rewritten for the treatment and control groups separately as follows:

$$\log\left(\frac{p_{i1}}{1-p_{i1}}\right) = \alpha_i + \theta + \beta_i$$

and

$$\log\left(\frac{p_{i0}}{1-p_{i0}}\right) = \alpha_i$$

so that

$$\begin{pmatrix} \log\left(\frac{p_{i0}}{1-p_{i0}}\right) \\ \log\left(\frac{p_{i1}}{1-p_{i1}}\right) \end{pmatrix} \sim N\left(\begin{pmatrix} \alpha_i \\ \alpha_i + \theta \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 0 & \tau^2 \end{pmatrix}\right)$$

This model makes the assumption that the treatment arms have higher variability than the associated control arms. To avoid the asymmetry that results from this assumption, an adjustment of $+0.5$ and -0.5 was proposed as a treatment-arm dummy variable by Turner et al. (2000), instead of simply using $j = 0, 1$ as above. If we replace the whole parameter j with the dummy variable $z_{ij} = \pm 0.5$, then the above model can be rewritten for the treatment and control arms as:

$$\log\left(\frac{p_{i1}}{1-p_{i1}}\right) = \alpha_i^* + \theta + 0.5\beta_i$$

and

$$\log\left(\frac{p_{i0}}{1-p_{i0}}\right) = \alpha_i^* - 0.5\beta_i$$

where $\alpha_i^* = \alpha_i - 0.5\theta$ after re-parametrisation, giving

$$\begin{pmatrix} \log\left(\frac{p_{i0}}{1-p_{i0}}\right) \\ \log\left(\frac{p_{i1}}{1-p_{i1}}\right) \end{pmatrix} \sim N\left(\begin{pmatrix} \alpha_i^* \\ \alpha_i^* + \theta \end{pmatrix}, \begin{pmatrix} 0.25\tau^2 & -0.25\tau^2 \\ -0.25\tau^2 & 0.25\tau^2 \end{pmatrix}\right)$$

Both of the models we have discussed here are logistic regression models with unknown parameters α_i, θ and τ^2 that need to be estimated. This can be done iteratively using methods such as penalised quasi-likelihood, marginal quasi-likelihood, or first- or second-order Taylor expansions. If using penalised quasi-likelihood methods, the associated bias of the τ^2 estimates can be reduced by applying a two-step bootstrap approach (Turner et al. (2000)). Recently, (Jackson et al. (2018)) conducted a simulation study that showed the model with adjusted group dummy variables to be superior than that without, in terms of consistently underestimating τ^2, and suggested a theoretical explanation for this observation.

4.2.4 GLMMs with random intercept

Now we shall discuss the use of GLMMs with a random intercept for conducting meta-analyses, again for the base case of the log-odds ratio. This type of GLMM is equivalent to a mixed-effects logistic regression model with random intercept and treatment effects (Turner et al. (2000)), and can be written as:

$$y_{ij} \sim Binomial(n_{ij}, p_{ij})$$

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha + \gamma_i + (\theta + \beta_i) \times j$$

where p_{ij} are the treatment arm-specific event probabilities, α is the baseline log-odds, θ is the overall effect size (log-odds ratio), and β_i are the deviations of the study-specific treatment effect from θ, with $j = 1$ representing the active treatment arm and $j = 0$ otherwise, for $i = 1, \ldots, k$. In this case, both the intercept and slope are random effects, assumed to follow a bivariate normal distribution with $\gamma_i \sim N(0, \sigma^2)$, $\beta_i \sim N(0, \tau^2)$ and $Cov(\gamma_i, \beta_i) = \omega\sigma\tau$ (Van Houwelingen et al. (1993); Stijnen et al. (2010)). When it is assumed that $Cov(\gamma_i, \beta_i) = 0$, then the above model can be rewritten for the treatment and control groups separately as follows:

$$\log\left(\frac{p_{i1}}{1-p_{i1}}\right) = \alpha + \gamma_i + \theta + \beta_i$$

and

$$\log\left(\frac{p_{i0}}{1-p_{i0}}\right) = \alpha + \gamma_i$$

so that

$$\begin{pmatrix} \log\left(\frac{p_{i0}}{1-\pi_{i0}}\right) \\ \log\left(\frac{\pi_{i1}}{1-\pi_{i1}}\right) \end{pmatrix} \sim N\left(\begin{pmatrix} \alpha \\ \alpha+\theta \end{pmatrix}, \begin{pmatrix} \sigma^2 & \sigma^2 \\ \sigma^2 & \sigma^2+\tau^2 \end{pmatrix}\right)$$

Similar to Section 4.2.3, if we replace the treatment-arm dummy variable $j = 0, 1$ by the adjusted variable $z_{ij} = \pm 0.5$, and again assume that $Cov(\gamma_i, \beta_i) = 0$, then the above model can be re-parametrised for the treatment and control arms as follows:

$$\log\left(\frac{p_{i1}}{1-p_{i1}}\right) = \alpha^* + \gamma_i + \theta + 0.5\beta_i$$

and

$$\log\left(\frac{p_{i0}}{1-p_{i0}}\right) = \alpha^* + \gamma_i - 0.5\beta_i$$

where $\alpha^* = \alpha - 0.5\theta$ after re-parametrisation, giving

$$\begin{pmatrix} \log\left(\frac{p_{i0}}{1-p_{i0}}\right) \\ \log\left(\frac{p_{i1}}{1-p_{i1}}\right) \end{pmatrix} \sim N\left(\begin{pmatrix} \alpha^* \\ \alpha^*+\theta \end{pmatrix}, \begin{pmatrix} \sigma^2+0.25\tau^2 & \sigma^2-0.25\tau^2 \\ \sigma^2-0.25\tau^2 & \sigma^2+0.25\tau^2 \end{pmatrix}\right)$$

Whereas the standard random-effects meta-analysis approach includes the single heterogeneity parameter τ^2 (as σ^2 are assumed to be known), the above models include two heterogeneity parameters (σ^2, τ^2), although this can increase to three when the additional parameter $\omega = Cov(\gamma_i, \beta_i) \neq 0$. Similarly to the GLMM with fixed intercept discussed in Section 4.2.3, the unknown parameters for this case ($\alpha, \theta, \sigma^2, \tau^2$ and ω) can be estimated iteratively using a range of approaches (Turner et al. (2000)). Such random-intercept logistic regression models have previously been investigated for their performance in the meta-analysis of proportion data (Hamza et al. (2008)), and also briefly in the case of sparse data (Kuss (2015)).

4.3 Poisson mixed regression model

The first GLMM that we are suggesting for application to meta-analyses, and the subsequent estimation of τ^2, with rare-event log-risk ratio data is the Poisson mixed-effects model. One of the key benefits of this model is that it can include double-zero trials without the need to correct for them, which other approaches and models are generally unable to do. Although this method has previously been proposed for the case of rare-event meta-analyses (Böhning et al. (2015)), it has not yet been compared to other heterogeneity variance estimators for a range of realistic scenarios meeting this rare-event definition. As a result, we aim to fill this gap in the research, using a version of the approach that is suitable for use with the log-risk ratio measure. Below we shall outline the theory behind this proposed model approach.

4.3.1 Theory behind approach

Böhning et al. (2015) suggested Poisson mixed regression models (PMRMs) with random study effect and zero-inflation models for the analysis of rare-event data. The benefit of using a Poisson model approach lies in its ability to incorporate additional covariates as fixed and/or random effects, creating an easier approach for dealing with effect heterogeneity.

The idea of Poisson modelling is to consider the count of events X as a Poisson distributed variable with mean $E(X) = \mu P$ (Breslow and Day (1987); Clayton et al. (1993)), where P is the person-time, meaning that $\mu = E(X)/P$ is the incidence rate. We have that, for each trial i and each treatment arm j:

$$E(X_{ij}) = \mu_j P_{ij} \tag{4.3.1}$$

where $j = 1$ refers to the treatment arm and $j = 0$ otherwise, hence giving the risk ratio as $RR = \mu_1/\mu_0$. In terms of the study-specific counts introduced in Table 1.1, $X_{i1} = a_i$ and $X_{i0} = c_i$. Taking logarithms of Equation (4.3.1) yields the basic log-linear model:

$$\begin{aligned} \log E(X_{ij}) &= \log P_{ij} + \log \mu_j \\ &= \log P_{ij} + \alpha + \beta \times j \end{aligned} \tag{4.3.2}$$

where β is the log-risk ratio and α is the baseline risk. Parameter estimates are found by maximising the associated Poisson likelihood. An important feature of the model given in Equation (4.3.2) is that it includes the logarithmic person-times, $\log P_{ij}$, as an offset term. The associated likelihood for this fixed-effects Poisson model is given by:

$$\prod_{i=1}^{k}\prod_{j=0}^{1} Po(x_{ij}|\eta_{ij}) = \prod_{i=1}^{k}[Po(x_{i0}|P_{i0}\exp(\alpha_i)) \times Po(x_{i1}|P_{i1}\exp(\alpha_i+\beta))]$$

where $Po(x|\eta,\sigma_\alpha^2) = \exp(-\eta)\eta^x/x!$ are the associated Poisson probabilities. The model can be modified to include a covariate effect, such as study, to give:

$$\begin{aligned}\log E(X_{ij}) &= \log P_{ij} + \log \mu_{ij} \\ &= \log P_{ij} + \alpha_i + \beta_i \times j\end{aligned}$$

where the case in which $\beta_i = \beta$ corresponds to there being a common slope, reflecting the same effect over all studies. This new model not only allows different study-specific baseline risks α_i, but also study-specific log-risk ratios β_i. As such, this approach can easily be generalised to consider α_i and β_i as random effects, i.e. $\alpha_i \sim N(\alpha,\sigma_\alpha^2)$ and $\beta_i \sim N(\beta,\sigma_\beta^2)$, allowing for the production of mixed models. It is this value of σ_β^2 that we are interested in estimating here, as this is our heterogeneity variance estimate. In this case, the likelihood of the above mixed-effects Poisson model is given as:

$$\prod_{i=1}^{k}\int_{-\infty}^{\infty} Po(x_{i0}|P_{i0}\exp(\alpha_i))\times\left[\int_{-\infty}^{\infty} Po(x_{i1}|P_{i1}\exp(\alpha_i+\beta_i))\phi(\beta_i|\beta,\sigma_\beta^2)d\beta_i\right]\phi(\alpha_i|\alpha,\sigma_\alpha^2)d\alpha_i$$

where $\phi(\beta_i|\beta,\sigma_\beta^2)$ represents the probability density of some normal random variable with mean β and variance σ_β^2, and likewise for $\phi(\alpha_i|\alpha,\sigma_\alpha^2)$.

4.3.2 Zero-inflated Poisson models

The Poisson mixed regression model has several benefits, as it captures variation in the baseline event risks, and may easily be extended to allow for zero-inflation via zero-inflation Poisson (ZIP) models. ZIP models adjust for the over-dispersion that can arise from the occurrence of double-zeros, and are easy to interpret, generally leading to an improved data analysis.

In brief, the idea behind ZIP models involves assuming that some compartment that produces only zero events occurs with probability α, while the standard Poisson model then occurs with probability $1-\alpha$. This gives the ZIP model:

$$Pr[X_{ij}=x] = \begin{cases}\pi_{ij} + (1-\pi+ij)e^{-\eta_{ij}} & ,\ x=0 \\ (1-\pi_{ij})Po(x|\eta_{ij}) & ,\ x=1,2,\ldots\end{cases}$$

where $Po(x|\eta_{ij})$ are the Poisson probabilities, which are modelled as described above. Now, the excess-zero portion of the above model can be modelled via logistic regression, giving:

$$\log \eta_{ij} = \log P_{ij} + \log \mu_{ij} = \log P_{ij} + \alpha + \beta \times j$$
$$\operatorname{logit} \pi_{ij} = \log \pi_{ij} - \log(1 - \pi_{ij}) = \alpha' + \beta' \times j$$

where β' is the log-odds ratio, with β still representing the log-risk ratio. If α and α' are set as random effects, with $\alpha \sim N(\alpha, \sigma_\alpha^2)$ and $\alpha' \sim N(\alpha', \sigma_{\alpha'}^2)$, then the model can be rewritten as:

$$\log \eta_{ij} = \log P_{ij} + \log \mu_{ij} = \log P_{ij} + \alpha_i + \beta \times j$$
$$\operatorname{logit} \pi_{ij} = \log \pi_{ij} - \log(1 - \pi_{ij}) = \alpha_i' + \beta_i' \times j$$

In this case, the associated likelihood is given by:

$$\prod_{i=1}^{k} \int_{-\infty}^{\infty} \left\{ \prod_{j=0}^{1} [\delta_0\{x_{ij}\}\pi + (1-\pi) Po(x_{ij}|P_{i1} \exp(\alpha_i + \beta \times j))] \right\} \phi(\alpha_i|\alpha, \sigma_\alpha^2) d\alpha_i$$

where $\delta_0\{x_{ij}\}$ is a dummy variable that is equal to 1 if $x_{ij} = 0$ and 0 otherwise, and $\phi(\alpha_i|\alpha, \sigma_\alpha^2)$ is defined as previously.

4.3.3 Previous findings

Böhning et al. (2015) compared the various Poisson-based models discussed in this section to the fixed-effect Mantel-Haenszel approach for the case-study dataset regarding Rosiglitazone performance (shown in Table 3.1). They found that the Poisson regression model with random effects resulted in almost identical outcomes to that of the Mantel-Haenszel approach, which has previously been shown to perform well in homogeneous scenarios with few events. As such, it can be deduced that the Poisson model also works well in this scenario. In addition to this, they noted that the intercept variance parameter in the random-effects model, which is equivalent to the variation in baseline risks, appeared to be significant when they looked at both myocardial infarction and cardiovascular-related death outcomes. As such, they were able to conclude that the risk in the control arm varied considerably across the studies included, and more importantly, that this particular model had the power to detect this variation, which other approaches had not.

4.4 Conditional logistic mixed regression model

The second GLMM that we are proposing for use in this field is the conditional logistic mixed regression model (CLMRM), which despite being tested by others recently, has never been applied to the specific scenario of rare-event log-risk ratio meta-analyses. In this section we shall first introduce the hypergeometric-normal model, which led to the original suggestion for this model by Stijnen et al. (2010), and then go on to outline the theory behind the conditional logistic-based approach itself, explaining why we believe the model to be suitable for our problem of interest. We will then summarise previously published results regarding the performance of this model for alternate meta-analysis scenarios, before discussing the options available for its application, specifically for the R package *glmer*. As this model is known to face difficulties with convergence when applied to sparse data, we shall also outline the model options we used to maximise the number of successful applications to real-world rare-event data.

4.4.1 Theory behind approach

Stijnen et al. (2010) proposed a conditional logistic model for use with log-odds ratio meta-analyses, in the form of the two-level exact non-central hypergeometric-normal (NCHGN) model and a Binomial approximation of this. It is from their suggestions that we base our CLMRM approach for log-risk ratio scenarios, and as such we shall first discuss their proposed methods before introducing our own version.

The non-central hypergeometric-normal model

The hypergeometric-normal model was first proposed for use in meta-analysis by Liu and Pierce (1993) and Van Houwelingen et al. (1993), and was then implemented in practice by Stijnen et al. (2010) and Sidik and Jonkman (2008). It has more recently been compared with other models for a selection of meta-analysis scenarios, in the form of both simulation studies and empirical examples (Jackson et al. (2018); Bakbergenuly and Kulinskaya (2018)).

This model, like those discussed previously, can be applied using either a fixed or random intercept, depending on preference and its chosen purpose. When applying to meta-analysis data, the idea behind the use of the NCHGN model stems from conditioning on the total number of events in study i, denoted X_i, so that only the number of events in the treatment arm, X_{i1}, are random. Stijnen et al. (2010) proposed it for use with log-odds ratios, after noting the problems with assuming a normal distribution for the estimated study-specific effects, θ_i (in their case the log-odds ratio), in an attempt to bring about a number of the benefits discussed in Section 4.2.1. In particular, they suggested that given θ_i, which are assumed to be be normally distributed ($\theta_i \sim N(\theta, \tau^2)$), the corresponding event count X_{i1} has an exact non-central hypergeometric distribution.

This non-central hypergeometric distribution depends upon the assumption that the treatment and control groups are binomial distributed, without which use of the former distribution is no longer appropriate. The inference is then be based on the following exact likelihood:

$$\prod_{i=1}^{k} \int_{-\infty}^{\infty} P(E(X_{i1}) = X_{i1}|\theta_i) \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{(\theta_i - \theta)^2}{2\tau^2}\right) d\theta_i$$

with

$$P(E(X_{i1}) = X_{i1}|\theta_i) = \frac{\begin{pmatrix} n_{i1} \\ X_{i1} \end{pmatrix} \begin{pmatrix} n_{i0} \\ X_{i0} \end{pmatrix} \exp(\theta_i X_{i1})}{\sum_{j=max(0,n_{i1}-n_{i0})}^{min(n_{i0},n_{i1})} \begin{pmatrix} n_{i1} \\ j \end{pmatrix} \begin{pmatrix} n_{i0} \\ X_i - j \end{pmatrix} \exp(\theta_i j)} \tag{4.4.1}$$

where n_{i0} and n_{i1} are the sample sizes of the control and treatment arms of study i respectively, and all other parameters are as defined previously in this chapter. The denominator in Equation (4.4.1) represents the normalising constant for this function. For this mixed-effects logistic model, the unknown parameters of interest (θ and τ^2) can be estimated by the Newton-Raphson or EM algorithms, or by maximising the corresponding log-likelihood, in accordance with the methods discussed in Section 4.2.2.

Approximate binomial-normal model for rare events

After proposing the NCHGN model, Stijnen et al. (2010) then went on to suggest an approximation for the case of small numbers of events, in the form of a binomial rather than non-central hypergeometric distribution. They observed that given the total count of events, X_i, the count of events in the treatment group has an approximately binomial distribution:

$$X_{i1} \sim \text{Binomial}\left(X_i, \frac{\exp(\log(n_{i1}/n_{i0}) + \theta_i)}{1 + \exp(\log(n_{i1}/n_{i0}) + \theta_i)}\right)$$

If the count of events is much smaller than the corresponding sample size, i.e. $X << n$, then we have the approximation $\begin{pmatrix} n \\ X \end{pmatrix} \approx n^X/X!$. The combinations in the summations of the normalising constant in Equation (4.4.1) can then be approximated as:

$$\begin{aligned}\binom{n_{i1}}{j}\binom{n_{i0}}{X_i-j} &\approx \frac{n_{i1}^{j}}{j!}\frac{n_{i0}^{X_i-j}}{(X_i-j)!}\\ &= \left(\frac{n_{i1}}{n_{i0}}\right)^{j}\frac{n_{i0}^{X_i}X_i!}{j!(X_i-j)!X_i!}\\ &= \left(\frac{n_{i1}}{n_{i0}}\right)^{j}\binom{n_{i1}}{j}\frac{n_{i0}^{X_i}}{X_i!}\end{aligned}$$

where as before we have $j = max(0, n_{i1} - n_{i0})$. If we set $j = X_{i1}$ then, by the above approximation, the corresponding numerator of Equation (4.4.1) becomes:

$$\begin{aligned}\binom{n_{i1}}{X_{i1}}\binom{n_{i0}}{X_{i0}} &= \binom{n_{i1}}{X_{i1}}\binom{n_{i0}}{X_i-X_{i1}}\\ &\approx \left(\frac{n_{i1}}{n_{i0}}\right)^{X_{i1}}\binom{X_i}{X_{i1}}\frac{n_{i0}^{X_i}}{X_i!}\end{aligned}$$

thus giving the following approximation of Equation (4.4.1):

$$P(E(X_{i1}) = X_{i1}|\theta_i) \approx \frac{\binom{X_i}{X_{i1}}(n_{i1}\exp(\theta_i)/n_{i0})^{X_{i1}}}{\sum_{j=0}^{X_i}\binom{X_i}{j}(n_{i1}\exp(\theta_i)/n_{i0})^{j}}$$

Now, if we divide the numerator and denominator of this by $(1+n_{i1}\exp(\theta_i)/n_{i0})^{X_i}$, and define $\pi_i = (n_{i1}\exp(\theta_i)/n_{i0})/(1+n_{i1}\exp(\theta_i)/n_{i0})$, then we have:

$$P(E(X_{i1}) = X_{i1}|\theta_i) \approx \frac{\binom{X_i}{X_{i1}}\pi_i^{X_{i1}}(1-\pi_i)^{X_i-X_{i1}}}{\sum_{j=0}^{X_i}\binom{X_i}{j}\pi_i^{j}(1-\pi_i)^{X_i-j}}$$

Since the denominator in the above expression is the sum of all probabilities of a binomial distribution with parameters X_i and π_i, it is equal to 1. The numerator, meanwhile, is the probability of a $Binomial(X_{i1}, \pi_i)$ distribution. As a result, when the event of interest is rare, and thus the count of events is very small compared to the sample size, the hypergeometric distribution can be approximated by a simpler binomial distribution, with $X_{i1}|X_i \sim Binomial(X_i, P_{X_{i1}|X_i})$.

By defining $\pi_i = \text{expit}(\log(n_{i1}/n_{i0}) + \theta_i)$ in the case from Stijnen et al. (2010), then by the approximate model, the log-odds becomes equal to $\log(n_{i1}/n_{i0}) + \theta_i$. As the outcome measures are assumed to be normally distributed, i.e. $\theta_i \sim N(\theta, \tau^2)$, this approximate model can be fitted as a mixed-effects conditional logistic regression model containing only the intercept term. In particular, each study constitutes a binomial outcome with X_{i1} events in X_i trials, a random intercept and an offset term of $\log(n_{i1}/n_{i0})$ that is included in the model.

The method that we are proposing is based on this idea from Stijnen et al. (2010), but adapted for the log-risk ratio outcome measure. Although this approach has been tested previously in meta-analyses (the results of which we shall discuss below), it has only been applied for the case of odds ratio, and little has been investigated in terms of rare events, so we shall build on this with our outcome of interest.

4.4.2 Application to the log-risk ratio

As mentioned above, the new approach we are proposing for use with sparse-event data involves conditional logistic mixed regression modelling. It is based on the analysis of 2×2 contingency tables containing zero-event cells, where calculating an effect measure for each stratum, such as a risk ratio, can be problematic. Whereas the standard random-effects meta-analysis approach directly models effect measures as the contrast between two treatment arms (e.g. the risk ratio), this type of mixed-effects logistic regression model will instead focus on the effect measure directly, as trial effects are conditioned out. For our purposes, the idea is to allow random effects on those parameters that provide an estimate of the heterogeneity variance.

Let us recall the classical idea of a logistic regression model:

$$\text{logit}\, p_{ij} = \alpha_i + \beta \times j$$

where p_{ij} is the probability for an event in study i in the j-th treatment arm, and α_i might be a fixed effect $\alpha_i = \alpha$ or a random effect $\alpha_i \sim N(\alpha, \sigma_\alpha^2)$.

Now, this approach occurs in the following specific way, proposed by Stijnen et al. (2010). The basic idea is to consider X_{i1} conditional on $X_i = X_{i1} + X_{i0}$. Since $E(X_{i1}) = \mu_1 P_{i1}$ and $E(X_{i0}) = \mu_0 P_{i0}$, X_{i1} is binomial with size parameter X_i and event parameter:

$$\begin{aligned} q_i &= \frac{\mu_1 P_{i1}}{\mu_1 P_{i1} + \mu_0 P_{i0}} \\ &= \frac{RR_i \frac{P_{i1}}{P_{i0}}}{RR_i \frac{P_{i1}}{P_{i0}} + 1} \end{aligned}$$

With this approach, the event parameter involves only the parameter of interest, RR_i, and its functional form makes it prone to logistic regression:

$$\begin{aligned}\operatorname{logit} q_i &= \log \frac{q_i}{1-q_i} \\ &= \log RR_i + \log \frac{P_{i1}}{P_{i0}} \\ &= \alpha_i + \log \frac{P_{i1}}{P_{i0}}\end{aligned}$$

where the right-hand side of this equation can be used for further modelling such as $\alpha_i = \alpha$ (a common risk ratio across studies) or $\alpha_i \sim N(\alpha, \sigma_\alpha^2)$ (a random effect for the risk ratio). As before, we are interested in estimating the heterogeneity variance, this time represented by the term σ_α^2. The parameters can be estimated by means of (restricted) maximum likelihood approaches with iterative least squares, for example.

The benefit of conditional logistic mixed regression models is that you can focus directly on the treatment heterogeneity distribution, eliminating the intercept or nuisance parameter. Since this conditional inference approach uses the exact distribution, we believe that it is appropriate for the analysis of sparse-event data. A more detailed description of the theory behind our version of the conditional logistic modelling approach is given in Appendix B.1.

4.4.3 Previous findings

To assess the performance of the NCHGN model and its binomial approximation (equivalent to the CLMRM) in comparison to the standard Normal-based random-effects meta-analysis approach, Stijnen et al. (2010) applied them to two example datasets and conducted a simulation study. Their first case study, which is taken from the study by Niel-Weise et al. (2007) that we discussed in Section 3.2.2, included very few events and the estimated log-odds ratio (their outcome measure of interest) was consistently large across all three approaches. Both the NCHGN and CLMLM models were found to perform similarly, however they only focused on scenarios where the summary effect measure was extreme (in both case studies and simulations). The two models were found to outperform the standard normal-based approach in all cases considered, in terms of both $\hat{\theta}$ bias and associated confidence interval coverage. In particular, they had negligible bias and adequate coverage, but this coverage did appear to reduce slightly when between-study variance was present, although never to any significant degree.

Jackson et al. (2018) also recently compared the NCHGN and CLMRM models against others, using empirical and simulated data. Again, they focused only on the log-odds ratio outcome measure, but in addition to measuring bias, precision and coverage, they also

looked at the rate of non-convergence - a key potential problem of applying such models to sparse data. For their empirical analyses, they extracted data from the Cochrane Review relating to use of antibiotics in children with measles. They used a total of 7 meta-analyses, with the number of studies ranging from 1 (a direct singular analysis) to 7, with all meta-analyses containing at least one single-zero trial, and one containing 5 double-zero trials (out of 7). Using this real-life data, they found that they could only apply any of their models to meta-analyses with at least 3 studies containing an event, and as such could only look at 3 of their 7 planned meta-analyses (with one still containing a double-zero study), as the models were too weakly identifiable otherwise. They found that the NCHGN model was robust to inclusion/exclusion of the double-zero study, however with other models this change led to a dramatic difference in results. The reason behind this robustness is that the NCHGN model conditions on the events, so when this is zero it produces a degenerate distribution that is not dependent on the model parameters, and thus does not contribute to the overall likelihood.

In their simulation study, again focusing on the log-odds ratio, Jackson et al. (2018) designed an array of meta-analysis scenarios. They found that, of all the models considered, the NCHGN model was the most computationally demanding, taking up to 4000 times longer to fit than the simpler models, indicative of its increased complexity. In terms of estimation failure, they also found that this model failed the most frequently of those considered. To account for this, they altered the model options to improve convergence, managing to reduce the failure in determining point estimates for θ and τ^2 to 1.4%, and the corresponding standard error rate to 2.8%, of the 30,000 datasets simulated. However, some of the standard errors extracted from this model appeared to be unreliable as they differed significantly from other methods. In particular, they were generally much lower than those produced using other models, and were termed as 'artificially low', however there were also some cases where they were grossly too high, with one simulation generating a standard error of 17.5, far exceeding the next largest estimate of 0.09. This is unsurprising given the number of cases where only point estimates could be determined. As a result of this, they suggested conducting sensitivity analyses when reporting the result of this NCHGN model, as a form of sanity check.

In addition, from the results of their simulation study, Jackson et al. (2018) also found that the NCHGN model did not result in any significant bias in estimates of θ, in contrast to $\hat{\tau}^2$, which tended to have slight negative bias in all cases except for those involving rare events or few studies. As they explained, this makes sense since maximum likelihood estimates of variance tend to be downwardly biased in general, but that this is likely to be overcome by the constraint that the estimate must be positive. For very rare events, they found that the NCHGN model produced very skewed estimates for τ^2 compared to other models, with non-replicated estimates of $\tau^2 > 20$ being produced in some heterogeneous cases. However, this particular estimate was backed up when using the *rma.glmm* command in the R package *metafor*, potentially indicating its accuracy. This

observed bias and skewness lessened when the true value of θ was positive, more events were observed in the treatment group or the probability of events was increased, to the extent that they were no longer of concern. They stated their concern that the NCHGN model may experience a loss of information as a result of conditioning on an almost ancillary statistic, however found this not to be the case as the total count of events has little impact on the odds ratio itself. The quality of the standard errors produced compared to other approaches led them to conclude that the use of GLMMs in general may result in more precise estimates for τ^2, but that this advantage was hindered by the undesirable positive bias in the case of very rare events, and the otherwise negative bias.

Finally, in terms of coverage of 95% confidence intervals, Jackson et al. (2018) pointed out that the NCHGN model, along with all other (non-normal-normal) models they considered, use the maximum likelihood asymptotic theory when computing the intervals, and so some deviation from the nominal 95% is to be expected. They found that the NCHGN model performed similarly to other models except when events were very rare or heterogeneity was considerable, in some cases falling short of the nominal level by 5 percentage points. Similar to other models, it was conservative with homogeneity, and failed to achieve nominal coverage by 1-2 percentage points in general, giving particularly low coverage when $k = 3$. However, all non-normal-normal models performed well in terms of coverage when events were rare, contrasting with the challenge in estimating τ^2. They included the NCHGN model in their final recommendations, however gave caution regarding the fragile nature of, and numerical challenges associated with, this model in several scenarios, potentially making it inappropriate for routine use without statistical expertise. They did however suggest it for use in sensitivity analyses when applying other models, particularly for the estimation of τ^2.

The CLMRM model was compared against others in a simulation setting by Kuss (2015), where they focused on on meta-analyses with varying proportions of double-zero trials. In their study, they looked at a number of outcome measures, including the risk ratio, however only used the odds ratio with this particular model. In the case of non-zero treatment effect, they found the CLMRM model to perform similarly to an alternative Beta-Binomial model in terms of empirical power and convergence time. In fact, CLMRM had the faster convergence of the two, but had slightly less power, with a median of 9.1% power. In terms of coverage, however, the model performed very poorly, with coverage probability regularly below 70%, well below the nominal 95% level

Recently, Bakbergenuly and Kulinskaya (2018) looked at both the NCHGN and CLMRM models in terms of log-odds ratios in a simulation study, and found that for the estimation of τ^2, the NCHGN model produced the highest bias of all models considered when sample sizes were below 100 participants, but had negative bias in all other cases. However, despite the high bias in $\hat{\tau}^2$, the model was found to be one of the best for the scenario of small sample sizes, with the authors noting that it was the magnitude

of sample size inequality that appeared to have a considerable effect on the bias of this parameter. The model was also found to perform well when τ^2 was large, sample size was small and event probability was low, being considered the best for sparse data in general. However, the NCHGN model was found to behave erratically when sample sizes were large, particularly in the estimation of θ, and the CLMRM model gave the worst performance in this scenario as it was asymptotically biased. In particular, the CLMRM model had large negative bias when estimating θ, but produced the smallest estimates for this parameter, with the NCHGN model producing the second-lowest estimates but outperforming standard approaches in the case of sparse data and large heterogeneity. When the event probability was high (> 0.4), both models performed equally well, and their respective biases were smaller in magnitude when event probability in the control arm was not too low, i.e. > 0.1.

In terms of coverage, Bakbergenuly and Kulinskaya (2018) found that the NCHGN model always had lower than nominal coverage, while the CLMRM model had this for all cases except when both τ^2 was small and sample size was large. In general, coverage improved as the size of the treatment effect increased away from zero, with the NCHGN model displaying the poorest coverage at all levels of τ^2. When applying the models to an empirical dataset, they found that the CLMRM model produced a much lower estimate of τ^2 compared to the alternate methods considered, with the estimate of θ also being noticeably different. However, the authors point out that this latter estimate may in fact be close to the true value, as this model consistently produced lower estimates in simulations of data of a similar structure, but had the least bias (almost being unbiased at times) in such a scenario. Both models gave narrower confidence intervals than the other approaches, with the simulations indicating that the CLMRM model tends to have the best coverage when the effect size is zero, but the worst coverage otherwise. However, increased bias in the treatment effect, combined with underestimation of the associated standard error, resulted in both of these models regularly generating a coverage far lower than the nominal level of 95%.

4.4.4 Choice of model options

When applying the NCHGN model, Jackson et al. (2018) chose to implement it in using the *rma.glmm* command in the R package *metafor*, initially using no command options other than setting the model to be *'CM.EL'* - indicating a conditional model with exact likelihood. For the CLMRM model, they applied it using the same command but with option *'CM.AL'*, corresponding to a conditional model with approximate likelihood. The *metafor* R package, along with *lme4*, requires a value to be set for the command option *nAGQ* - the points per axis used to calculate the adaptive Gauss-Hermite approximation of the log-likelihood. For *metafor*, the default value is 7, and Jackson et al. (2018) used this default in order to limit the computational demands of all models they applied with

this package. The *metafor* package removes all studies containing zero events (double-zero studies) by default, however this option can be changed, just as Jackson et al. (2018) did in order to compare the effect of including vs. excluding these studies when applying their models.

When applying the CLMRM model to a meta-analysis of 5 studies (containing one single-zero and one double-zero trial), Jackson et al. (2018) found that the original model options for *metafor* had to be altered in order to ensure convergence. In particular, they increased the maximum number of iterations to 20,000 and the relative tolerance level to 0.0001, and changed the maximisation algorithm to the more robust Nelder-Mead approach. However, they left the option $nAGQ$ as the default (7 points), as this particular model involves only a single random effect. They pointed out that these changes were only made for the CLMRM approach in order to obtain the numerical maximisation initial values, and so did not influence the number of quadrature points applied in the model fitting itself. They also investigated various options relating to the production of the standard errors, via the *hessianCtrl* command options, in an attempt to produce more stable errors, however they determined that the default setting in *metafor* was best and so kept to this.

4.5 Alternative GLMM approaches

So far in this chapter, we have focused on GLMMS that we believe to be appropriate and theoretically beneficial in the meta-analysis of rare-event log-risk ratio data. However, a number of alternate models have also been proposed and tested for use with meta-analyses of different outcome measures and varying scenarios. Below we shall outline some of these alternate GLMMs for completeness, and to provide an idea of models that could be adapted for potential future approaches in our area of interest.

4.5.1 Binomial mixed regression model

In addition to proposing the CLMRM model discussed in Section 4.4, Stijnen et al. (2010) also looked at the Binomial-Normal model, for both proportion data and incidence rate ratios.

Proportion outcome data

For the case of proportion data, Stijnen et al. (2010) used a basic model where the outcome was the incidence of the event of interest, and the parameter of interest in the model was some proportion, e.g. the occurrence of an event with a new treatment. In order to overcome many of the issues with the standard normal-normal model and

the associated use of continuity corrections, Hamza et al. (2008) proposed assuming a Binomial distribution for the study-specific event counts, in the form of:

$$X_i \sim Binomial\left(n_i, \frac{\exp(\theta_i)}{1+\exp(\theta_i)}\right)$$

where θ_i is the proportion of study i in this case. As a result of this assumption, the within-study likelihood is given by the following Binomial likelihood:

$$\exp(\theta_i)^{X_i}/(1+\exp(\theta_i))^{n_i}$$

Now, if the proportions $\theta_1, \ldots, \theta_k$ are assumed to be normally distributed with

$$\theta_i \sim N(\theta, \sigma^2) \tag{4.5.1}$$

as is the case in the normal-normal model, then we have the Binomial-Normal (BN) or random-intercept logistic regression model. By design, this model is a simpler version of the CLMRM we discussed previously, that does not involve any conditioning on parameters.

Some investigation has been conducted to determine the performance of this model in the special case of proportion-based outcomes. For example, Hamza et al. (2008) found the BN model to regularly outperform the normal-normal model in a simulation study, generating unbiased point estimates and adequate confidence interval coverage. As a result, they recommended its use for the case of meta-analyses with few studies, and suggested pairing it with profile likelihood summary-effect confidence intervals.

When Stijnen et al. (2010) proposed this model for use with this data type, they applied it to the rare-event case study from Niel-Weise et al. (2007) that we discussed in Section 3.2.2. In their case, they looked at the incidence of catheter-related blood stream infections in each of the treatment groups. They compared the model outcomes with those produced using the standard normal-normal approach, and found that their results mirrored those found in the simulations of Hamza et al. (2008). In particular, they observed that the BN model produced larger estimates of τ^2, particularly when looking at the active treatment group, while the normal-normal model consistently underestimated this parameter. However, the BN model also produced larger standard errors for the summary effect measure, although this can be explained by the fact that the BN model takes into account the uncertainty of the within-study variances, which the standard approach does not, and also because its corresponding τ^2 estimates are greater in magnitude. They also applied the model to log-odds ratios, using the same dataset, and found that the BN model performed similarly to the NCGHN model, which is to be expected given the data type.

Incidence rate ratio outcome data

Stijnen et al. (2010) also looked at the use of the BN model with incidence rate ratio data, where in this case the aim is to compare the performance of some intervention between two groups. If X_{ij} and P_{ij} are the count of events and total follow-up time of the treatment group j in study i respectively, and θ represents the log-incident rate ratio, then the corresponding meta-analysis model is given by:

$$\hat{\theta}_i = \log\left(\frac{X_{i1}/P_{i1}}{X_{i0}/P_{i0}}\right)$$

and

$$\hat{\sigma}_i^2 = \frac{1}{X_{i0}} + \frac{1}{X_{i1}}$$

The CLMRM-based models that we discussed earlier were based on exact within-study likelihoods that exploit the fact the number of events in the treatment arm given the total number of events follows a binomial distribution, as follows:

$$X_{i1} \sim Binomial\left(X_i, \frac{\exp(\log(P_{i1}/P_{i0}) + \theta_i)}{1 + \exp(\log(P_{i1}/P_{i0}) + \theta_i)}\right) \tag{4.5.2}$$

Combining the model given in Equation (4.5.2) with that from Equation (4.5.1) gives the incidence rate ratio BN model, or logistic regression model with random intercept and offset term $\log(P_{i1}/P_{i0})$, so again a modification of our CLMRM model. Stijnen et al. (2010) applied this model to another dataset from the study by Niel-Weise et al. (2007), comparing its results to that of the traditional normal-normal approach as before. They found that the BN method produced a summary effect of greater magnitude but with a smaller p-value than the normal-normal model, but both gave similar estimates for τ^2. They hypothesised that the difference in results between these models would increase if the true value of θ increased, and so conducted a small simulation study to investigate this theory. From these simulations, they found that whereas point estimate bias increased with the value of θ for the normal-normal approach, it remained negligible in all cases investigated for the BN model, demonstrating the consistent reliability of the BN model over the standard approach. The coverage of the summary-effect confidence interval was also found to be better with the BN model, as it was less sensitive to the magnitude of θ.

4.5.2 Beta-Binomial mixed regression model

An alternative to the BN model discussed above would be to assume that the effect measures represent a random sample from a Beta, rather than Normal, distribution,

generating the Beta-Binomial (BB) model, as used by Zhou et al. (1999). This change in distribution results in a simple expression for the marginal likelihood of the study-specific summary effect. Once again, this model represents another version of logistic regression, particularly for responses that are correlated. For this model, we look at case of proportion data and assume that study-specific proportions are observed from independent Binomial distributions, where the probability parameter (π) of each of these distributions in turn has a Beta distribution with parameters α and β. The mean of the Beta distribution is defined as $E(\pi) = \mu = \alpha/(\alpha + \beta)$, with variance $Var(\pi) = \mu(1-\mu)\theta/(1+\theta)$ when $\theta = 1/(\alpha+\beta)$. This is equivalent to saying that the event counts have a beta-binomial distribution with mean $E(X_i) = n_i\mu$ and variance $Var(X_i) = n_i\mu(1-\mu)\left[1 + (n_i - 1)\theta/(1+\theta)\right]$ when the average is taken with respect to the beta distribution of π. The correlation between individual event counts across study-specific arms can then be defined as $\rho = 1/(\alpha+\beta+1)$. The corresponding log-likelihood of this beta-binomial distribution is as follows:

$$l(\alpha, \beta) = \sum_{i=1}^{k} l_i(\alpha, \beta)$$

with

$$\begin{aligned} l_i(\alpha,\beta) = {} & lgamm(n_i+1) + lgamm(X_i+\alpha) + lgamm(n_i - X_i + \beta) + lgamm(\alpha+\beta) \\ & - lgamm(X_i+1) - lgamm(n_i - X_i + 1) - lgamm(n_i + \alpha + \beta) - lgamm(\alpha) \\ & - lgamm(\beta) \end{aligned}$$

where $\alpha = \mu(1-\rho)/\rho$, $\beta = (1-\mu)(1-\rho)/\rho$ and $lgamm$ represents the natural log of the gamma function. As each study contributes two proportions (one from each treatment arm), the treatment effect, μ, is modelled as follows:

$$g(\mu) = b_0 + b_1 j$$

where g represents some common link function, b_0 and b_1 are the model intercept and slope respectively, and $j = 0, 1$ is a dummy variable for the control and treatment arms as before. The choice of the link function is important as this dictates the outcome measure, e.g. logit link gives the log-odds ratio while log link corresponds to the log-risk ratio. Kuss (2015) subtracted the treatment arm-specific logit model event probability estimates in order to produce the risk difference, and then generated the associated standard error via the delta method. As this model incorporates a closed-form log-likelihood, it allows for parameter estimation despite being a random-effects model. As such, it has the potential to be adaptable to a range of outcome measures.

Kuss (2015) compared the performance of this model against others and found that, for the log-risk ratio, the BB model performed similarly to others in terms of convergence - performing well for large numbers of studies but struggling when few are present. Despite performing well in terms of coverage for 95% confidence intervals, with coverages consistently above 90%, the model was found to generate considerable bias, the magnitude of which resulted in the model being deemed unacceptable for reliable use with this particular outcome measure. For the risk difference, as before they found that the model performed similarly to others in terms of convergence, but this time performed well in terms of bias, while also having the best coverage of all models considered (with only one corresponding coverage below 90%). As mentioned previously, this approach performed similarly to the CLMRM model for log-odds ratio, but had higher power despite producing lower coverage. In general, for the log-risk ratio, the BB model had similar properties to other models, but had better convergence compared to others that struggled with the increasing sparsity of the data. As a result, Kuss (2015) recommended this model for the three outcome measures discussed here for rare-event binary outcome data, as it performed well in all cases for both zero and moderate treatment effect.

4.5.3 Approximate version of hypergeometric-normal model with random-effects variation of Peto's method

As discussed in Section 4.4, the NCHGN model was found to perform well with rare-event data, however this was not the case for much more common events, where the model could become increasingly challenging to fit as a result of small differences between event counts and corresponding sample sizes. Jackson et al. (2018) proposed a variation of the NCHGN model that could account for this, and thus has the potential to be used when events are more common. Their approach involves Peto's method for odds ratio, which naturally incorporates the non-central hypergeometric distribution with a fixed-effect model, thus being an approximation of the NCHGN model with $\tau^2 = 0$. They use this approximation and transform it into a random-effects model, by using the study-specific log-odds estimates within the non-central hypergeometric distributions given in Section 4.4. These estimates are then normally approximated, and an approximate version of the NCHGN model is fit using the estimates as outcome data and the random-effects model. As the NCHGN is fit via maximum likelihood expectation, they suggested estimating τ^2 via the corresponding maximum likelihood approach in order to produce estimates as close to those of the original model as possible.

When applying this variation of the NCHGN model to empirical data containing large numbers of zero counts, Jackson et al. (2018) found that it produced similar point estimates to the original NCHGN model. As such, this approximate version of the approach can act as a sanity check for the original approach's results, or as an alternative when it fails to converge in this scenario. In fact, to reduce the high rate of non-convergence

and investigate any non-replicated extreme estimates produced by the NCHGN model in their simulation study, they replaced point estimates and standard errors that were missing or differed by a certain level with those produced using this approximation. This adjustment in results allowed estimates to now be obtained for all 30,000 simulated studies, altering 3.3% of the standard errors in total.

4.6 Application of our chosen models

We are interested in applying the PMRM and CLMRM models introduced in this chapter to rare-event log-risk ratio meta-analysis data, with the primary aim being to extract an estimate for the heterogeneity variance. We chose to use the *glmer* command in the R package *lme4* to apply these models, because it allowed for the application of both types of model with meta-analysis data and could handle zero-count data - an important criterion in our case. This command includes double-zero studies by default, which the previously mentioned *rma.glmm* command does not. In addition, we observed that the *glmer* command could apply our chosen models to more scenarios than any other package or command, making it the superior choice for our purposes. In terms of the number of axis points used to estimate the adaptive Gauss-Hermite approximation, its default option is 1, which is equivalent to the Laplacian approximation.

When applying our models, we initially used the default options for *glmer*, with the plan to change these model options as required in order to cope with the zero counts present, and to allow for convergence in as many scenarios as possible. Any changes to these options will be discussed with the results of the simulation study, along with the number of non-convergences that may have occurred. Our aim is to produce models that can be applied to the largest selection of rare-event scenarios possible, which may be difficult given previous results regarding non-convergence with sparse data, as discussed above. As a result, there may have to be a pay-off between model accuracy and total number of scenarios it can be applied to. However, we are assuming that any changes in options will not change the outcomes considerably, in the sense that the relative performance of the models compared to the other methods will likely remain the same. As such, we aim to generate models that can be applied to a diversity of rare-event scenarios, and so can be applied in the majority of real-life datasets.

4.7 Scenarios where our chosen models cannot be applied

Despite amending the *glmer* command options to allow for the models to be applied in the maximum number of scenarios possible, there will remain some cases where the models simply cannot be used on certain types of data, and thus no τ^2 estimate can be retrieved. In these cases, it is the result of the structure of the data itself not being

compatible with the type of model being applied. Examples of this include the event of interest being too rare and thus resulting in outcome data that consists of too many zero counts, or the data simply meeting some incompatible form. In the case of extreme rarity of events, there are more model parameters than outcomes, which leads to over-parameterisation of the model and thus no convergence, ensuring the model cannot be applied. The conditions that govern this lack of convergence differ depending on the type of model being fitted and its associated family distribution. Below we shall outline the different scenarios that lead to non-convergence, as a direct result of the structure of the data and the corresponding model, for each of the models that we are considering here. Any additional non-compatible scenarios that we identify during our simulation study will be discussed with the results.

It is worth noting that when applying these models with the chosen *glmer* command, some warnings may appear, with more appearing as the rarity of the outcome increases. These warnings can be safely ignored, as they are merely appearing as the result of convergence difficulties due to the rarity of the data being input, something that cannot be avoided with these types of models. In particular, they are informing us that the models may be close to being non-identifiable (a particular issue for random-effects models), which will result over-parameterisation. However, the results produced are still acceptable and can be taken as the desired solution of the estimates here.

4.7.1 Poisson mixed regression model

The Poisson mixed regression model can be applied to most meta-analysis scenarios, however there are some where the model cannot be applied as a result of the structure of that data. The data is simply incompatible with the model when the ratio of events to sample size in the treatment arm is equal across all studies and the ratio of events to sample size in the control arm is also equal across all studies AND one of the following holds:

- The count of events in the control arm is equal across all studies and the sample size of the control arm is equal across all studies in the meta-analysis
- There are zero events in the control arm for all studies in the meta-analysis.

4.7.2 Conditional logistic mixed regression model

There are a number of meta-analysis scenarios for which the conditional logistic mixed regression model cannot be applied. These include cases where there is either no variation, and so this random-intercept model cannot converge, or the response is constant, meaning that this is simply not the correct type of input data for this particular family of model. These cases are listed below:

- The number of double-zero trials in the meta-analysis is greater or equal to $k-1$, where k is the number of studies in the meta-analysis
- There are zero events in the control arm for every study in the meta-analysis
- There are zero events in the treatment arm for every study in the meta-analysis
- The ratio of the number of events in the treatment arm to the number of events in the control arm is equal across each of the non-double-zero studies in the meta-analysis.

4.8 Comparison with alternative statistical packages

To determine whether we were applying our chosen GLMMs correctly using the *glmer* command in the R package *lme4*, and to check that we were extracting the correct parameters for our estimates, we also applied the models using a different statistical software package. We chose to use STATA as this alternate software as it allowed for the application of both the PMRM and CLMRM models using the *mepoisson* and *melogit* commands respectively (StataCorp (2013)). These STATA commands gave the option to change the integration method being used - something that would also allow us to check whether the command options we had chosen to use with *glmer* in R were appropriate. In particular, the integration methods that *mepoisson* and *melogit* can implement, along with the labels that we shall use to refer to them, are:

- Non-adaptive Gauss-Hermite quadrature (*nonadaptGH*):
  This transforms the multivariate integral in the associated likelihood into a series of manageable nested univariate integrals, which can then be calculated via use of a polynomial that optimally approximates the integrand.
- Mean-variance adaptive Gauss-Hermite quadrature - the default approach with these commands (*Default*):
  Similar to *nonadaptGH*, but the univariate integrals have their intervals divided into subintervals, where the subintervals are further shortened if the integrand behaves poorly. In this case, parameter estimates for the normal random variable in the likelihood are iteratively generated from re-evaluated posterior moments of the Gauss-Hermite quadrature approximation. Starting parameters for this are a mean of $(0, \ldots, 0)$ and identity variance matrix.
- Mode-curvature adaptive Gauss-Hermite quadrature (*modeGH*)
  Similar to *Default*, but the parameter estimates of the normal random variable are generated via optimisation of the integrand with respect to the normal random variable, where the associated mean is taken as the optimal value and the variance is the curvature at this value.

- Laplacian approximation (*Laplace*):
  Expressing the integrand as an exponential function, this method involves a second-order Taylor expansion on the argument of this function, about the value of the normal-based mean that maximises it.

4.8.1 Case studies

In order to make these comparisons between statistical software packages and integration methods, and also to compare the two models themselves with the estimators discussed in Chapter 2, we applied them to the rare-event case studies introduced in Chapter 3. The results of applying the PMRM model to those datasets in R and STATA are given in Table 4.1, while the respective results for the CLMRM model are displayed in Table 4.2.

From looking at Table 4.1, it can be observed that when applying the PMRM approach, the results from our R code most closely correlate to those produced using the default (mean-variance) and mode-curative Gauss-Hermite quadrature integration methods with STATA's *mepoisson* command. This is unsurprising given that the default integration approach that we applied using the *glmer* command in R is more similar to these than the alternate STATA options of non-adaptive Gauss-Hermite quadrature and the Laplacian approach. These differences are only present with the τ^2 estimates however, as the log-risk ratio estimates are actually very similar between all integration methods for the C-section, albumin and transplant datasets, which corresponds to the type of data these case studies represent. There are no results for non-adaptive Gauss-Hermite quadrature with PMRM in STATA for the transplant dataset as this model could not converge, most likely the result of the very small number of studies in this meta-analysis.

In terms of the CLMRM approach, Table 4.2 displays a similar story, although in this case all three Gauss-Hermite quadrature methods provide similar results to that of our R code for $\hat{\tau}^2$. Again, this is indicative of the integration method used in the default option for *glmer* when applying this particular model. In terms of the log-risk ratio, all integration methods considered provide very similar, and in some cases identical, estimates for each of the case studies. From both of these tables, we can conclude that the results obtained using our *R* code are backed up by those obtained using an alternative statistical software package, and as such can be assumed to be reliable. In addition to this, the integration method is observed to have a significant impact on the estimates of τ^2 - particularly between the Gauss-Hermite quadrature methods and the Laplacian approximation. In these tables, we have not included the confidence intervals that would be generated from the STATA parameter estimates, as we were only interested in comparing the direct point estimates extracted from the model parameters.

TABLE 4.1: Summary of results from Poisson mixed regression model according to our written R code and STATA for case studies.

Dataset	**Software**	**Integration Method**	$\hat{\tau}^2$	$\hat{I}^2$	$\widehat{\log RR}$	**Confidence Interval**		
						Z-type CI	***t*-type CI**	**HKSJ CI**
Rosig (MI)	R code	-	0	0	0.34	(0.03, 0.65)	(0.02, 0.66)	(0.13, 0.55)
	STATA	Default	0	0	0.30			
		modeGH	0	0	0.29			
		nonadaptGH	0.99	44.57	-0.07			
		Laplace	0.51	29.29	-0.05			
Rosig (death)	R code	-	0.04	1.86	0.55	(0.11, 0.98)	(0.09, 1.00)	(0.31, 0.78)
	STATA	Default	0	0	0.44			
		modeGH	0	0	0.44			
		nonadaptGH	0.54	20.37	-0.08			
		Laplace	0.52	19.77	0.34			
CRBSI	R code	-	0.58	36.95	-1.17	(-1.78, -0.56)	(-1.83, -0.51)	(-1.67, -0.66)
	STATA	Default	0.54	35.30	-1.35			
		modeGH	0.54	35.30	-1.35			
		nonadaptGH	1.00	50.26	-1.17			
		Laplace	1.00	50.26	-1.17			
C-section	R code	-	0.21	28.00	-1.12	(-1.36, -0.88)	(-1.36, -0.87)	(-1.33, -0.90)
	STATA	Default	0.25	31.64	-1.11			
		modeGH	0.25	31.64	-1.11			
		nonadaptGH	0.91	62.76	-0.97			
		Laplace	0.25	31.64	-1.11			
Albumin	R code	-	0.02	3.74	0.57	(0.28, 0.86)	(0.26, 0.88)	(0.32, 0.82)
	STATA	Default	0	0	0.54			
		modeGH	0	0	0.54			
		nonadaptGH	5.11×10^{-10}	9.92×10^{-8}	0.54			
		Laplace	0.61	54.21	0.54			
Transplant	R code	-	3.40×10^{-13}	1.26×10^{-11}	0.64	(-1.19, 2.46)	(-3.37, 4.64)	(-1.49, 2.76)
	STATA	Default	1.46×10^{-32}	5.41×10^{-31}	0.64			
		modeGH	3.45×10^{-35}	1.28×10^{-33}	0.64			
		nonadaptGH	-	-	-			
		Laplace	0.86	24.15	0.64			

TABLE 4.2: Summary of results from conditional logistic mixed regression model according to our written R code and STATA for case studies.

Dataset	**Software**	**Integration Method**	$\hat{\tau}^2$	$\hat{I}^2$	$\widehat{\log RR}$	**Confidence Interval**		
						Z-type CI	**t-type CI**	**HKSJ CI**
Rosig (MI)	R code	-	0	0	0.35	(0.05, 0.66)	(0.04, 0.67)	(0.15, 0.56)
	STATA	Default	0	0	0.35			
		modeGH	0	0	0.35			
		nonadaptGH	0	0	0.35			
		Laplace	0.70	36.25	0.34			
Rosig (death)	R code	-	0	0	0.51	(0.08, 0.93)	(0.07, 0.94)	(0.29, 0.73)
	STATA	Default	0	0	0.51			
		modeGH	0	0	0.51			
		nonadaptGH	0	0	0.51			
		Laplace	0.79	27.24	0.50			
CRBSI	R code	-	0.56	36.14	-1.30	(-1.91, -0.69)	(-1.95, -0.65)	(-1.82, -0.78)
	STATA	Default	0.60	37.74	-1.30			
		modeGH	0.60	37.74	-1.30			
		nonadaptGH	0.57	36.54	-1.30			
		Laplace	1.00	50.26	-1.18			
C-section	R code	-	0.22	28.95	-1.04	(-1.28, -0.80)	(-1.29, -0.79)	(-1.25, -0.83)
	STATA	Default	0.23	29.87	-1.04			
		modeGH	0.23	29.87	-1.04			
		nonadaptGH	0.23	29.87	-1.05			
		Laplace	0.22	28.95	-1.04			
Albumin	R code	-	0	0	0.54	(0.26, 0.82)	(0.25, 0.84)	(0.30, 0.79)
	STATA	Default	0	0	0.54			
		modeGH	0	0	0.54			
		nonadaptGH	0	0	0.54			
		Laplace	0.81	61.12	0.54			
Transplant	R code	-	0	0	0.69	(-1.13, 2.52)	(-3.31, 4.70)	(-1.53, 2.91)
	STATA	Default	4.00×10^{-34}	1.48×10^{-32}	0.69			
		modeGH	5.16×10^{-34}	1.91×10^{-32}	0.69			
		nonadaptGH	4.16×10^{-34}	1.54×10^{-32}	0.69			
		Laplace	0.83	23.51	0.68			

Now, comparing the results generated from our R code in these two tables directly against each other, we can compare our two proposed GLMM approaches in terms of how they appear to perform in the case of rare-event data. Firstly, we can note that both models could be applied to each of the 6 case studies when using our written R code, which is itself a testament to their consistent suitability in rare-event scenarios, even when there are also very few studies. It can be observed that several zero τ^2 estimates have been generated, particularly with the CLMRM model, with both producing zero (or close to zero) estimates for the rosiglitazone (MI) and transplant datasets - again most likely linked with the type of data these two case studies represent. In terms of the overall τ^2 and log-risk ratio estimates, the results are consistently very similar between the two models, with each giving authenticity to the other, and potentially indicating that the two models may perform rather similarly in general. In terms of the summary-effect confidence intervals, the Z-type and t-distribution methods can be seen to produce similar intervals for all datasets other than transplant, with the HKSJ method producing the narrowest intervals in most scenarios. For the transplant dataset, where there were few studies present, the t-distribution method appears to consistently perform very poorly, generating extremely wide intervals with both models.

Finally, comparing these results to those displayed in Chapter 3, we can compare how these models perform compared to the existing τ^2 estimators discussed in Chapter 2. In terms of both the rosiglitazone datasets, the τ^2 estimates are very similar (as many of the existing approaches also produced zero estimates), however the log-risk ratio estimates are consistently higher in the models. For the CRBSI case study, the model τ^2 estimates are much higher than the existing methods in general (with them being slightly larger than the highest estimates from Sidik-Jonkman), while the corresponding log-risk ratio estimates are far further away from zero. The model log-risk ratio estimates for the C-section case study also have a greater magnitude than the existing approaches, while the τ^2 estimates are similar to those from the Sidik-Jonkman approach. The model τ^2 estimates for the albumin dataset were similar to those given in Chapter 3, as most existing approaches gave zero estimates, while the log-risk ratio estimates again had a greater magnitude and were similar to those from the Mantel-Haenszel approach. Finally, for the transplant meta-analysis with few studies, it can be observed that zero τ^2 estimates were produced in both of the models and many existing methods, with the log-risk ratio estimates being much greater in the models, and identical to those from the fixed-effect Mantel-Haenszel approach in the case of the CLMRM model.

4.9 Conclusions

In this chapter, we have outlined two novel approaches based on the use of GLMMs that we are proposing for the estimation of τ^2 in the case of rare-event log-risk ratio meta-analyses. There are several benefits of using GLMMs over standard approaches to

conduct meta-analyses, particularly in the case of zero count data, as they do not require bias-inducing continuity corrections and also avoid the use of the inappropriate normal approximation. As a result, some recent simulation studies have focused on comparing the performance of a range of GLMMs on varying meta-analysis scenarios, however none have yet focused on the combined case of rare events and the easily interpretable log-risk ratio outcome measure. As such, we aim to focus on this gap in knowledge, by applying such models to rare-event scenarios in order to extract estimates for the heterogeneity variance.

We are proposing the Poisson and conditional logistic mixed regression models, as these both have beneficial qualities that we believe make them appropriate for the estimation of τ^2 with rare-event data. In particular, the CLMRM model conditions on the total number of events, and incorporates the person times of both the control and treatment arms. Both models have previously displayed evidence of performing well in recent simulation studies looking at alternate meta-analysis scenarios. However, the CLMRM model can be very computationally demanding, and as a result can struggle with convergence and take a long time to fit compared with other approaches, especially when the data is sparse. For example, Jackson et al. (2018) found that rare events presented difficulties in model convergence, but noted that events needed to be very rare to be considered a serious problem, and also found that differences in estimation quality were dependent on which trial arm the events were more prevalent in.

As well as the models potentially struggling in the case of rare-event data, there are a number of additional scenarios where the models cannot be applied as a result of incompatibility between the data structure and model distribution being used. As a result, there are some meta-analysis scenarios where our models cannot be used, however these are few in number and so should not heavily impact their potential application to real-life data. We have chosen to apply both our chosen models using the *glmer* command in the R package *lme4*, as this is suitable for both models and allows application to a high number of meta-analysis scenarios. In addition, we modified the associated command options in order to maximise this range of applicable scenarios, in order to make the approaches usable for the majority of real-life datasets.

We applied our models to the rare-event case studies introduced in Chapter 3, using both our chosen R command and the statistical software package STATA, in order to check that our R code was correct and the appropriate parameters were being extracted. We found that the results produced were very similar between the software packages, particularly when the integration method used by STATA was of the Gauss-Hermite quadrature type, providing evidence that our R code and corresponding command-line options are appropriate. When comparing the results between the two models, it could be seen that the two models produced very similar estimates for τ^2 and the log-risk ratio for all case studies. In addition, when comparing these estimates to those produced using the existing estimators discussed in Chapter 2, we found that the model

estimates for τ^2 were generally similar or slightly larger, indicating that they may be detecting heterogeneity that the other methods could not. However, the model log-risk ratio estimates were consistently of greater magnitude, indicating that the models were potentially capable of detecting stronger effects in rare-event scenarios. As our case studies only represent a small number of possible meta-analysis scenarios, we shall also include the models in a simulation study, which will allow us to develop a greater insight into their performance in general, as well as determine any further scenarios where the models are unable to converge.

Chapter 5

Conditional approach to estimate heterogeneity variance

5.1 Introduction

A wealth of methods have been proposed for the estimation of the heterogeneity variance parameter (τ^2) in random-effect meta-analyses. In the previous chapter, we discussed the use of generalised linear mixed models (GLMMs) for the estimation of this value via a one-step meta-analysis approach. In particular, we focused on the use of Poisson mixed regression models (PMRM) and conditional logistic mixed regression models (CLMRM) when dealing with zero-count data, our scenario of interest. However, as noted in previous simulation studies (Bakbergenuly and Kulinskaya (2018); Jackson et al. (2018)), these model types can face problems with convergence when the data are sparse, whether that be in terms of low numbers of studies, sample sizes or event counts. In addition to this, some of these models are computationally complex - the CLMRM in particular can take an inordinately long time to fit, with convergence potentially being difficult to achieve without statistical assistance.

Other estimation methods exist that do not share these computational issues, and are instead used in two-step inverse-variance meta-analysis approaches. We discussed a number of these τ^2 estimators in Chapter 2, looking at both iterative and non-iterative estimators in frequentist and Bayesian frameworks. The majority of these estimators were based on the Normal approximation, and as such could be conflicted with bias when many zero event counts are present and the data subsequently does not follow the Normal distribution. However, not all of the two-step based estimators we discussed made this inappropriate assumption of normality. For example, the approach proposed by Böhning and Sarol (2000), and presented in Section 2.8.2, estimates the τ^2 parameter without making any assumptions regarding its associated distribution, and can be applied to rate data where the event count conditional upon the study follows a Poisson distribution.

We are interested in investigating a novel method for the estimation of τ^2 in the meta-analysis of rare-event data that is based on the estimator from Böhning and Sarol (2000), but which is appropriate for the log-risk ratio outcome measure. Similar to their existing method, the estimating approach that we are proposing is based on a conditional approach, which we believe makes it appropriate for use with low event rates.

In this chapter, we shall outline the motivation and theory behind this approach, and the estimation equations and statistical techniques that are involved in its application. As with the previous methods that we have discussed earlier, we shall then apply this approach, including any variations that we feel are appropriate, to the rare-event case study datasets described in Chapter 3. This will allow us to gain an initial overview of its performance against existing estimators with empirical clinical trial data containing zero event counts.

5.2 Theory behind approach

As mentioned briefly in Section 2.8.2, Böhning and Sarol (2000) proposed a novel approach to estimate the variance of the heterogeneity distribution without having to estimate the associated distribution, for use with both Binomial and Poisson counts. The motivation behind their approach begins by assuming that a given variable of interest, Y, is modelled through a parametric density $p(y, \theta)$ with scalar parameter θ. If θ has variance distribution G and associated density $g(\theta)$, then the marginal, and unconditional, density of Y is given as $f(y) = \int_{-\infty}^{\infty} p(y, \theta)g(\theta)d\theta$. The associated variance of Y, $Var(Y)$, can be separated into two terms, as follows:

$$\begin{aligned} Var(Y) &= \int_{-\infty}^{\infty} Var(Y|\theta)g(\theta)d\theta + \int_{-\infty}^{\infty} (\mu(\theta) - \mu_y)^2 g(\theta)d\theta \\ &= E(\sigma^2(\theta)) + \delta^2 \end{aligned} \tag{5.2.1}$$

where $\mu(\theta) = E(Y|\theta)$ and μ_y is the mean of the variable of interest Y. In some cases, the term $\delta^2 = \int_{-\infty}^{\infty}(\mu(\theta) - \mu_y)^2 g(\theta)d\theta$ instead has the form

$$\delta^2 = K \int_{-\infty}^{\infty} (\theta - \mu_\theta)^2 g(\theta)d\theta = K\tau^2 \tag{5.2.2}$$

where K is some constant, μ_θ is the mean of θ and $\tau^2 = \int_{-\infty}^{\infty}(\theta - \mu_\theta)^2 g(\theta)d\theta$ is the variance of θ, and as such the heterogeneity variance of interest. In scenarios where δ^2 satisfies the alternative definition given in Equation (5.2.2), then Equation (5.2.1) can be thought of as separating the total variance into that of the subpopulation with parameter

θ and that of the heterogeneity distribution G of θ. This latter term of variance, the latent factor, contains the parameter τ^2 via its distribution G.

The aim was to calculate an estimate of the heterogeneity variance τ^2 with no knowledge or assumptions regarding its distribution G. They proposed replacing the values of $Var(Y)$ and $E(\sigma^2(\theta))$ in Equation (5.2.1) with the respective study sample estimates, and rearranging in order to generate this estimate of τ^2. The value would then be truncated to zero if it was negative.

5.3 Standardised mortality ratio and proportion outcome estimators

The approach by Böhning and Sarol (2000) was designed for application to meta-analysis data, in particular the standardised mortality ratio (SMR) and proportion outcome data. However, as mentioned previously, it could potentially be applied to any rate-based data where it is assumed that the observed count conditional upon the study has a Poisson distribution, where the rate may involve some study-specific parameter. In their paper, they introduce an iterative and nonparametric estimator for heterogeneity variance, based on the theory outlined in Section 5.2.

The SMR is defined as the ratio of the number of observed mortality cases, O, and the respective expected (non-random) number of such cases, e. Here, e is used to denote the expected number of cases in order to differentiate from the action of taking expected values, which is denoted by E. In general, O is assumed to follow a Poisson distribution with mean $E(O, e|\theta) = \mu(\theta) = \theta e$. If extra-Poisson variation is allowed for, then the total variance of O can be separated as follows:

$$\begin{aligned} Var(O) &= \int_{-\infty}^{\infty} Var(O|\theta)g(\theta)d\theta + \int_{-\infty}^{\infty} (\theta e - \mu e)^2 g(\theta)d\theta \\ &= E(\sigma^2(\theta)) + e^2\tau^2 \\ &= e\mu + e^2\tau^2 \end{aligned}$$

Rearranging gives

$$\tau^2 = Var(O)/e^2 - \mu/e$$

where $\tau^2 = Var(\theta)$ and θ is the SMR in this case.

As the expected number of mortality cases may differ within a sample subpopulation, let $O_1, \ldots, O_k$ be a random sample of the number of mortality cases, with associated

expected numbers $e_1, \ldots, e_k$, where k can be seen as the number of studies. Then, for the estimation of heterogeneity variance with SMR outcome meta-analysis data, they suggest the following equation:

$$\hat{\tau}^2_\mu = \frac{1}{k}\left[\sum_{i=1}^{k}(O_i - e_i\mu)^2/e_i^2 - \mu\sum_{i=1}^{k}\frac{1}{e_i}\right] \tag{5.3.1}$$

where O_i is the observed number of mortality cases in study i, e_i is the expected (non-random) number of mortality cases in study i, k is the number of studies in the meta-analysis, and μ is an unknown parameter to be estimated.

Sometimes the outcome of interest is based on a proportion of the events of interest. Let proportions $r_1 = X_1/n_1, \ldots, r_k = X_k/n_k$ be a sample from k studies, where X_i is the count of events and n_i is the sample size of study i. Since $X_1, \ldots, X_k$ can be viewed as sample of Binomial counts, it follows by definition that the variance of X is defined as:

$$\begin{aligned} Var(X) &= E(\sigma^2(\theta)) + n^2\tau^2 \\ &= n\mu(1-\mu) + n(n-1)\tau^2 \end{aligned}$$

If μ is small and n is large, then the Poisson approximation for the binomial holds, and $Var(X) \approx n\mu + n^2\tau^2$. However, if this approximation is not valid, as would generally be the case, then they proposed the alternate formula:

$$\hat{\tau}^2 = \frac{1}{k-1}\sum_{i=1}^{k}\frac{(X_i - n_i\hat{\mu})^2}{n_i(n_i-1)} - \frac{1}{k}\hat{\mu}(1-\hat{\mu})\sum_{i=1}^{k}\frac{1}{n_i - 1} \tag{5.3.2}$$

where X_i is the number of events and n_i is the sample size of study i, and $\hat{\mu}$ is an estimate of the unknown parameter μ.

5.3.1 Previous findings

Böhning and Sarol (2000) performed an analysis using both of the estimators described above on empirical data meeting their requirements, in order to determine what effect heterogeneity has on the efficiency of the estimation of the mean effect of interest. They then also investigated how their methods compared to others, including the non-parametric mixture model approach, in terms of their estimation of the heterogeneity variance. From this they were able to report that their proposed estimators are beneficial over parametric models, including Gamma and Beta-distributed models for the data types considered, as they are more flexible due to not assuming a distribution of the heterogeneity.

5.4 Outline of log-risk ratio approach

We shall now outline an alteration of the above described approach, that we are proposing for application to log-risk ratio meta-analyses, our outcome of interest.

5.4.1 Estimation of the event probability

Recall that the Mantel-Haenszel estimate for the risk ratio (RR), as outlined in Section 1.9.5, is defined as:

$$\widehat{RR}_{MH} = \frac{\sum_{i=1}^{k}(X_{i1}n_{i0})/n_i}{\sum_{i=1}^{k}(X_{i0}n_{i1})/n_i} \tag{5.4.1}$$

where X_{i1} and X_{i0} are the count of events in the treatment and control arms, n_{i1} and n_{i0} are the number of subjects in the treatment and control arms, respectively, and $n_i = n_{i1} + n_{i0}$ is the total number of subjects in study i. For simplicity, in our simulation study we shall be setting the within-trial arm sample sizes to be equal, i.e. $n_{i1} = n_{i0}$, so we can simplify Equation (5.4.1) as follows:

$$\widehat{RR}_{MH} = \frac{\sum_{i=1}^{k} X_{i1}}{\sum_{i=1}^{k} X_{i0}}$$

Now, if we define θ as our outcome of interest, the log-risk ratio, i.e. $\theta = \log RR$, we can write

$$e^{\hat{\theta}} = \widehat{RR}_{MH} = \frac{\sum_{i=1}^{k} X_{i1}}{\sum_{i=1}^{k} X_{i0}}$$

The probability of events in the meta-analysis, which we shall denote as p, can then be estimated using

$$\hat{p} = \frac{\sum_{i=1}^{k} X_{i1}}{\sum_{i=1}^{k} X_i} \tag{5.4.2}$$

An approximation for this estimate of $\hat{p}$, for use when the values of X_{i1} and X_i for $i = 1, \ldots, k$ are not available to us, is given as follows:

$$\hat{p} = \frac{e^{\hat{\theta}}}{1 + e^{\hat{\theta}}}$$

since

$$\begin{aligned}\hat{p} &= \frac{e^{\hat{\theta}}}{1+e^{\hat{\theta}}} \\ &= \frac{\sum_{i=1}^{k} X_{i1} / \sum_{i=1}^{k} X_{i0}}{1+\sum_{i=1}^{k} X_{i1} / \sum_{i=1}^{k} X_{i0}} \\ &= \frac{\sum_{i=1}^{k} X_{i1}}{\sum_{i=1}^{k} X_i}\end{aligned}$$

Equation (5.4.2) for $\hat{p}$ is obviously more accurate than its approximation, and as we shall have the event counts available to us in our simulation study, we shall use this equation to estimate the probability of events for use in this approach.

5.4.2 Heterogeneity variance estimating equations

The estimate of the event probability described in the previous section shall be a crucial component in our estimators for τ^2. All of the estimating equations that we shall be discussing below are modified versions of those proposed in the paper by Böhning and Sarol (2000). The τ^2 estimates described in this section are based on an approximation of the log-risk ratio, due to the estimation of p outlined in Section 5.4.1, and as such will later need to be converted for the desired log-risk ratio.

The first estimating equation for τ^2 that we shall consider is based on Equation (5.3.1), as taken from the original paper, and is defined by the following:

$$\hat{\tau}_p^2 = \frac{1}{k}\left[\sum_{i=1}^{k}(X_{i1} - X_i\hat{p})^2/X_i^2 - \hat{p}\sum_{i=1}^{k}\frac{1}{X_i}\right] \tag{5.4.3}$$

where X_{i1} is the number of events in the treatment arm, X_{i0} is the number of events in the control arm, $X_i = X_{i1} + X_{i0}$ is the total number of events in study i, k is the number of studies in the meta-analysis, and $\hat{p}$ is the estimated probability of an event as defined in Equation (5.4.2).

A modified version of the above equation, which is also based partially on Equation (5.3.2) from the same paper, is formed by separating the two terms in parentheses in Equation (5.4.3), and adjusting the components by which each term is multiplied. It is defined as follows:

$$\hat{\tau}_p^2 = \frac{1}{k-1}\sum_{i=1}^{k}\frac{(X_{i1} - X_i\hat{p})^2}{X_i^2} - \frac{\hat{p}(1-\hat{p})}{k}\sum_{i=1}^{k}\frac{1}{X_i} \tag{5.4.4}$$

The third estimating equation for τ^2 that we will consider is again based on Equation (5.3.2), and as such is a slightly modified version of Equation (5.4.4) above. In this case, we modify the X_i elements in the denominators of each term, but only for those meta-analyses that do not contain any studies with only one event across both arms ($X_i = 1$). If the meta-analysis doesn't meet this condition, then Equation (5.4.4) is used. Thus, this equation is given by:

$$\hat{\tau}_p^2 = \begin{cases} \frac{1}{k-1}\sum_{i=1}^{k}\frac{(X_{i1}-X_i\hat{p})^2}{X_i(X_i-1)} - \frac{\hat{p}(1-\hat{p})}{k}\sum_{i=1}^{k}\frac{1}{X_i-1} & , X_i > 1 \text{ for all } i = 1,\ldots,k \\ \frac{1}{k-1}\sum_{i=1}^{k}\frac{(X_{i1}-X_i\hat{p})^2}{X_i^2} - \frac{\hat{p}(1-\hat{p})}{k}\sum_{i=1}^{k}\frac{1}{X_i} & , \text{otherwise} \end{cases} \tag{5.4.5}$$

The fourth and final equation, which is a variation of Equation (5.4.5), involves using the two alternative terms in the summations when the meta-analysis contains both studies with $X_i > 1$ and $X_i \leq 1$, i.e. changing the elements of the summation term-wise, depending on which condition study i meets. As such, it is given by the following formula:

$$\hat{\tau}_p^2 = \begin{cases} \frac{1}{k-1}\sum_{i=1}^{k}\frac{(X_{i1}-X_i\hat{p})^2}{X_i(X_i-1)} - \frac{\hat{p}(1-\hat{p})}{k}\sum_{i=1}^{k}\frac{1}{X_i-1} & , X_i > 1 \text{ for all } i = 1,\ldots,k \\ \frac{1}{k-1}\sum_{i=1}^{k}\frac{(X_{i1}-X_i\hat{p})^2}{X_i^2} - \frac{\hat{p}(1-\hat{p})}{k}\sum_{i=1}^{k}\frac{1}{X_i} & , X_i \leq 1 \text{ for all } i = 1,\ldots,k \\ \frac{1}{k-1}\left(\sum_{i\in k_{(1)}}\frac{(X_{i1}-X_i\hat{p})^2}{X_i(X_i-1)} + \sum_{j\in k_{(2)}}\frac{(X_{j1}-X_j\hat{p})^2}{X_j^2}\right) & \\ \quad -\frac{\hat{p}(1-\hat{p})}{k}\left(\sum_{i\in k_{(1)}}\frac{1}{X_i-1} + \sum_{j\in k_{(2)}}\frac{1}{X_j}\right) & , \text{otherwise} \end{cases} \tag{5.4.6}$$

where in the last case, $k_{(1)}$ is the subset of studies i for which $X_i > 1$ is satisfied, and $k_{(2)}$ is the subset of studies j for which $X_j \leq 1$ are satisfied, with $k_{(1)} + k_{(2)} = k$.

For all of the estimating equations outlined in this section, they are undefined in the case where $X_i = 0$, i.e. for double-zero (DZ) studies. If such a trial exists within the data then it must be omitted from the respective meta-analysis to allow the application of this approach, and the number of non-DZ studies (k^*) should be used in place of k in the estimating equations. As a result of this, if the meta-analysis contains $k-1$ DZ studies, then $k^* = 1$ and Equations (5.4.4), (5.4.5) and (5.4.6) produce infinite estimates as they divide by $k-1$. In these cases, the estimate should be set to undefined and ignored.

5.4.3 Transformation to variance of log-risk ratio

So far when outlining this approach, we have dealt with the variance of a transformation of the log-risk ratio, $\hat{p} = \frac{e^{\hat{\theta}}}{1+e^{\hat{\theta}}}$ to be exact, and not with the variance of the log-risk ratio itself - our outcome measure of interest. To correct for this, we will now transform the

estimate that we proposed to the required value. However, it should be noted that this transformation is based an approximation, the δ-method in this case, and thus may not be as accurate as other τ^2 estimators as a result of this.

To conduct this transformation, we note that $\theta = \log RR$ and $\hat{p} = \frac{e^{\hat{\theta}}}{1+e^{\hat{\theta}}}$. Now from the δ-method, we know that if Y has variance $Var(Y)$, then

$$Var(T(Y)) \approx T'(E(Y))^2 Var(Y)$$

where T is some differentiable transformation. Now, if we input our value of $\hat{p}$ as the transformation of $\hat{\theta}$, we have that

$$Var(\hat{p}) \approx \left[\left(\frac{e^{\hat{\theta}}}{1+e^{\hat{\theta}}}\right)'\right]^2 Var(\hat{\theta}) \tag{5.4.7}$$

Looking at the differentiation term we have:

$$\begin{aligned}\left(\frac{e^{\hat{\theta}}}{1+e^{\hat{\theta}}}\right)' &= \frac{e^{\hat{\theta}}(1+e^{\hat{\theta}}) - e^{2\hat{\theta}}}{(1+e^{\hat{\theta}})^2} \\ &= \frac{e^{\hat{\theta}}}{(1+e^{\hat{\theta}})^2} \\ &= \hat{p}\frac{1}{1+e^{\hat{\theta}}} \\ &= \hat{p}^2\frac{1}{e^{\hat{\theta}}} \\ &= \hat{p}^2 e^{-\hat{\theta}}\end{aligned}$$

Finally, inputting this derivative back into Equation (5.4.7) gives us

$$Var(\hat{p}) \approx \hat{p}^4 \left(e^{-\hat{\theta}}\right)^2 Var(\hat{\theta}) \tag{5.4.8}$$

which can be arranged in terms of $Var(\hat{\theta})$:

$$Var(\hat{\theta}) \approx \left[e^{2\hat{\theta}}/\hat{p}^4\right] Var(\hat{p})$$

Another way to present Equation (5.4.8) is as follows:

$$Var(\hat{p}) \approx \hat{p}^2 \frac{1}{\left(1+e^{\hat{\theta}}\right)^2} Var(\hat{\theta})$$

which again can be rearranged as

$$Var(\hat{\theta}) \approx \left[\left(1+e^{\hat{\theta}}\right)^2 / \hat{p}^2\right] Var(\hat{p})$$

where $\hat{p} = \frac{\sum_{i=1}^{k} X_{i1}}{\sum_{i=1}^{k} X_i}$ and $\hat{\theta} = \log \frac{\hat{p}}{1-\hat{p}}$.

The equations given in Section 5.4.2 estimate $Var(\hat{p})$, as this is the value $\hat{\tau}_p^2$. The estimate of the heterogeneity variance that we are interested in obtaining, $\hat{\tau}^2$, is given by $Var(\hat{\theta})$ as $\theta = \log RR$, our outcome of interest. As such, the estimate for τ^2, using this approach, is calculated via either of the following two equivalent equations:

$$\hat{\tau}^2 = \left[e^{2\hat{\theta}} / \hat{p}^4\right] \hat{\tau}_p^2 \tag{5.4.9}$$

$$\hat{\tau}^2 = \left[\left(1+e^{\hat{\theta}}\right)^2 / \hat{p}^2\right] \hat{\tau}_p^2 \tag{5.4.10}$$

where $\hat{p} = \frac{\sum_{i=1}^{k} X_{i1}}{\sum_{i=1}^{k} X_i}$, $\hat{\theta} = \log \frac{\hat{p}}{1-\hat{p}}$ and $\hat{\tau}_p^2$ has been estimated using one of the four estimating equations given in Equations (5.4.3) to (5.4.6).

In summary, to apply this approach we shall use one of the equations given in Section 5.4.2 to produce an estimate for $Var(\hat{p})$, and then use this value in either Equation (5.4.9) or (5.4.10) to approximate the estimate for the heterogeneity variance $Var(\hat{\theta})$. As with the existing estimators described in Chapter 2, we shall then obtain our estimate of the outcome measure, the log-risk ratio, by inputting our estimate of τ^2 into the inverse-variance approach.

It should be noted that when all of the studies in the meta-analysis have zero events in the treatment arm, i.e. $\sum_{i=1}^{k} X_{i1} = 0$, then the transformation conducted via Equation (5.4.9) or (5.4.10) is undefined, as the denominator $\hat{p}$ $\left(= \frac{\sum_{i=1}^{k} X_{i1}}{\sum_{i=1}^{k} X_i}\right)$ is 0.

5.5 Verification of approach via simulation study

When measuring the performance of this approach, we shall look at its accuracy in estimating the parameter τ_p^2 in addition to the heterogeneity variance τ^2. In order to determine the accuracy of this τ_p^2 estimate, we will first need to determine the corresponding

true value, which can be approximated via a simulation study with data generated using characteristics of the meta-analysis (or scenario) of interest. This approximation of τ_p^2 will in turn allow us to approximate the true value of τ^2 via the transformation described in the previous section. Calculating both of these values will allow us to independently measure the performance of the various aspects of this method. When conducting a general simulation study, the approximation of τ^2 can be compared to its corresponding true value, in order to ensure that the estimates $\hat{\tau}^2$ are being compared to the correct value that has accounted for the approximation used in this approach. In order to determine these approximations of τ^2 and τ_p^2, we shall conduct a separate simulation alongside our main simulation study, for each scenario we investigate, and we shall outline this below.

5.5.1 Calculation of true τ_p^2

We begin by assuming that the true log-risk ratio outcome measures for each study (denoted by θ_j for hypothetical study j) follow a Normal distribution:

$$\theta_j \sim N(\theta, \tau_\theta^2) \tag{5.5.1}$$

where the value of θ is calculated using the trial-arm event probabilities, which in our simulation study will be set and used to sample the corresponding count of events. To observe how this value is calculated, first note that θ_i, the log risk-ratio for study i, is defined as:

$$\hat{\theta}_i = \log \widehat{RR}_i = \log \left(\frac{X_{i1}/n_{i1}}{X_{i0}/n_{i0}} \right)$$

and using the definition for the estimated event probability, we then have

$$\theta_i = \log RR_i = \log \left(\frac{p_{i1}}{p_{i0}} \right)$$

where p_{i1} is the probability of an event in the treatment arm and p_{i0} is the probability of an event in the control arm of study i.

We shall simulate values of θ_j from the distribution in Equation (5.5.1) a large number of times, denoted by s, where say $s = 10,000$, and so $j = 1, \ldots, s$. In order to determine the true value of τ_p^2, we need to use the simulated values of θ_j to determine the associated values of p_j (the probability of an event occurring in study j). This can be done using the equation:

$$p_j = \frac{e^{\theta_j}}{1 + e^{\theta_j}}$$

From this, we can then determine the mean probability of events in these s simulations, denoted by $\bar{p}$:

$$\bar{p} = \frac{1}{s}\sum_{j=1}^{s} p_j$$

Finally, the true value of τ_p^2 can be calculated as follows:

$$\tau_p^2 = \frac{1}{s}\sum_{j=1}^{s}(p_j - \bar{p})^2 \tag{5.5.2}$$

Given this true value of τ_p^2 for each scenario, we can then determine the average value of τ_p^2 resulting from each chosen value of τ_θ^2 that is used for the Normal variance parameter in Equation (5.5.1).

We shall conduct this side-simulation for each scenario we investigate in our main simulation study, as the pairings of θ and τ_θ^2 will remain constant within given scenarios. In addition, since we are determining the true value of τ_p^2 using data that represents $s = 10,000$ studies, this value is unlikely to change significantly if it were calculated separately for meta-analyses from the same scenario, as each of these is calculated using averages from the s simulated hypothetical studies, and so little variation would exist. As such, we believe that it is reasonable to determine this value once for each scenario rather than for each meta-analysis within said scenario.

5.5.2 Approximation of true heterogeneity variance

In order to convert the true value of τ_p^2 into the variance of the log-risk ratio, our heterogeneity variance of interest, we input τ_p^2 into either Equation (5.4.9) or (5.4.10), as we described with the transformation of the estimate in Section 5.4.3. It should be noted again that this is an approximation, based on the δ-method, and so we cannot guarantee it is the correct true value, however we can assume that it has the same level of uncertainty as the respective estimate of τ^2, since they were constructed in the same manner. Thus, we have the approximation of the true value of the heterogeneity variance as:

$$\tau^2 \approx \left[e^{2\hat{\theta}}/\hat{p}^4\right]\tau_p^2$$

where τ_p^2 is calculated using Equation (5.5.2).

To ensure that the approach works correctly, we shall calculate this approximation of τ^2 and compare it to the true value of τ^2, equal to τ_θ^2 in Equation (5.5.1). This will tell

us whether our final estimates for $\hat{\tau}^2$ are being compared against the correct true value, which takes into account the approximation used in this method.

5.6 Summary of variations of estimating equation

In Section 5.4 we outlined the application of this approach to the log-risk ratio, proposing four alternate estimating equations for τ_p^2. In Table 5.1 we summarise these variations of approach in terms of choice of estimating equation, giving each its own notation that we shall use in the remainder of this thesis.

TABLE 5.1: A summary of the conditional-based heterogeneity variance estimating equations along with their respective abbreviations used in this thesis.

Estimator	**Abbreviation**	**Section**	**Estimating equation**
Conditional approach		5.4.2	-
Variation 1	CO1	-	(5.4.3)
Variation 2	CO2	-	(5.4.4)
Variation 3	CO3	-	(5.4.5)
Variation 4	CO4	-	(5.4.6)

5.7 Application to case studies

In order to gain a first impression regarding how the four variations of this approach compare with each other, and with those methods already discussed in Chapters 2 and 4, we applied them to the case studies introduced in Chapter 3. However, when comparing its output to that of existing estimators, we do so tentatively, with the important note that it may not perform as well owing to its approximation. It should also be noted that the estimates for the log-risk ratio were generated using the inverse-variance approach, similar to those methods discussed in Chapter 2. The results of these analyses can be seen in Table 5.2.

Comparing the variations of this conditional approach for each of the case studies, it can be seen that they all produce identical results for 4 of the 6 datasets. The results only vary for the C-section and albumin datasets given in Tables 3.6 and 3.8, respectively, which is probably a result of the change in type of data these case studies represent from the otherwise standard rare-event examples we have used. Even within these two case-studies, the variations CO2 and CO3 still produce identical results, meaning that these are identical for all case studies considered, which is unsurprising given the similarity of their associated estimating equations (Equations (5.4.4) and (5.4.5)). However, it is not known whether this equivalence is only the case for rare-event data, or whether they

would produce very similar results for all data types - something that will need to be investigated in the simulation study.

From the C-section and albumin case study results, it can be seen that CO1 consistently produces zero estimates for τ^2, while the identical estimates from CO2 and CO3 are slightly higher ($\hat{\tau}^2 = 0.04, 0.13$ for the two case studies respectively), and the estimates from CO4 are much higher ($\hat{\tau}^2 = 0.14, 0.27$). These increasing estimates of τ^2 result in associated log-risk ratio estimates of increasing magnitude.

By looking at the different summary effect confidence interval methods that we used, it can be seen that t-distribution intervals are consistently slightly wider than the Normal-based Z-type intervals for the CRBSI and both rosiglitazone datasets. These two methods produce very similar intervals for the C-section and albumin datasets, with the widest interval depending the conditional estimating equation applied. For the transplant meta-analysis that involves only 3 studies, the t-distribution intervals are consistently much larger than those of the Z-type and HKSJ methods. For the albumin, transplant and two rosiglitazone datasets, the HKSJ method produced the narrowest intervals of all methods considered. Meanwhile, for the CRBSI and C-section meta-analyses, the HKSJ method produces intervals similar to those generated by the alternate methods.

When these results are compared to those estimates generated from the pre-existing and GLMM-based approaches, displayed in Chapter 3 and Tables 4.1 and 4.2 respectively, it can be seen that their performance over the different datasets is similar to that of the other estimators. In particular, those datasets that struggled with many zero estimates using the pre-existing methods in Chapter 2 also produce only zero estimates with this approach. For the C-section meta-analysis, the results from CO1 are similar to those produced by the variations of the Der-Simonian Laird estimator, while the results of CO4 mirror those of the maximum likelihood based approaches. For the albumin analysis, the CO4 approach produces similar estimates to the Sidk-Jonkman estimator, however this dataset resulted in largely zero estimates for the Normal-based approaches. In contrast, the results generated by CO1 and CO4 for both datasets vary significantly from those of the Poisson and conditional logistic mixed regression models.

5.8 Conclusions

In this chapter, we proposed a variation of a τ^2 estimator originally proposed by Böhning and Sarol (2000) for use with proportion and SMR data. The benefit of their approach was that it did not assume any distribution for the heterogeneity variance, which is beneficial in the case of rare-event data, where the normal approximation that many other approaches make tends to be inappropriate. Compared to previously discussed GLMM-based methods, this estimator has the benefit that it is based on a non-iterative

TABLE 5.2: Summary of results from 4 alterations of the conditional-based approach for case studies.

Dataset	**Estimator**	$\hat{\tau}^2$	$\hat{I}^2$	$\log \widehat{RR}$	**Confidence Interval**		
					Z-type CI	**t-type CI**	**HKSJ CI**
Rosig (MI)	CO1	0.00	0.00	0.23	(-0.08, 0.54)	(-0.09, 0.54)	(0.03, 0.43)
	CO2	0.00	0.00	0.23	(-0.08, 0.54)	(-0.09, 0.54)	(0.03, 0.43)
	CO3	0.00	0.00	0.23	(-0.08, 0.54)	(-0.09, 0.54)	(0.03, 0.43)
	CO4	0.00	0.00	0.23	(-0.08, 0.54)	(-0.09, 0.54)	(0.03, 0.43)
Rosig (death)	CO1	0.00	0.00	0.13	(-0.30, 0.55)	(-0.31, 0.57)	(-0.06, 0.31)
	CO2	0.00	0.00	0.13	(-0.30, 0.55)	(-0.31, 0.57)	(-0.06, 0.31)
	CO3	0.00	0.00	0.13	(-0.30, 0.55)	(-0.31, 0.57)	(-0.06, 0.31)
	CO4	0.00	0.00	0.13	(-0.30, 0.55)	(-0.31, 0.57)	(-0.06, 0.31)
CRBSI	CO1	0.00	0.00	-0.92	(-1.36, -0.47)	(-1.40, -0.43)	(-1.37, -0.46)
	CO2	0.00	0.00	-0.92	(-1.36, -0.47)	(-1.40, -0.43)	(-1.37, -0.46)
	CO3	0.00	0.00	-0.92	(-1.36, -0.47)	(-1.40, -0.43)	(-1.37, -0.46)
	CO4	0.00	0.00	-0.92	(-1.36, -0.47)	(-1.40, -0.43)	(-1.37, -0.46)
C-section	CO1	0.00	0.00	-0.81	(-0.99, -0.63)	(-1.00, -0.63)	(-1.00, -0.62)
	CO2	0.04	6.79	-0.84	(-1.04, -0.65)	(-1.04, -0.64)	(-1.04, -0.65)
	CO3	0.04	6.79	-0.84	(-1.04, -0.65)	(-1.04, -0.64)	(-1.04, -0.65)
	CO4	0.14	20.21	-0.89	(-1.11, -0.66)	(-1.12, -0.66)	(-1.09, -0.69)
Albumin	CO1	0.00	0.00	0.38	(0.10, 0.66)	(0.09, 0.68)	(0.15, 0.61)
	CO2	0.13	20.17	0.42	(0.08, 0.76)	(0.06, 0.78)	(0.17, 0.68)
	CO3	0.13	20.17	0.42	(0.08, 0.76)	(0.06, 0.78)	(0.17, 0.68)
	CO4	0.27	34.22	0.44	(0.05, 0.82)	(0.03, 0.84)	(0.17, 0.70)
Transplant	CO1	0.00	0.00	0.27	(-1.55, 2.10)	(-3.73, 4.28)	(-1.54, 2.09)
	CO2	0.00	0.00	0.27	(-1.55, 2.10)	(-3.73, 4.28)	(-1.54, 2.09)
	CO3	0.00	0.00	0.27	(-1.55, 2.10)	(-3.73, 4.28)	(-1.54, 2.09)
	CO4	0.00	0.00	0.27	(-1.55, 2.10)	(-3.73, 4.28)	(-1.54, 2.09)

estimating equation, that won't be subject to the computational demands and challenging application of the former. We proposed four different variations of this approach, to determine whether altering the estimating equation slightly might improve its performance.

One negative aspect of this approach compared with the GLMM-based methods is that it would be used within two-step meta-analyses, and so the τ^2 estimates are inserted into the inverse-variance approach that has been heavily criticised for use with sparse data. However, while this may impact the accuracy of the log-risk ratio estimates, we can still determine whether the τ^2 estimates are better than previous mentioned methods, and if this is the case, then these estimates can be used with future updated meta-analysis approaches, acting as a step towards improved analysis.

Another drawback of this approach is that double-zero trials have to be omitted, when we were aiming to propose and investigate methods that could incorporate and include zero event-containing studies. Although this is unfortunate, the approach still allows for the inclusion of single-zero studies without the use of bias-causing continuity corrections, which in itself is an advantage over a number of the methods discussed in Chapter 2.

When we compared our proposed variations of this approach against each other in the form of our rare-event case studies, we found that CO1 consistently produced zero estimates for τ^2, with all four variations producing identical results for four of the six datasets, potentially demonstrating their difficulty in working with certain types of rare-event data. In the two other cases, CO2 and CO3 produced identical estimates, while CO4 generated estimates of the greatest magnitude. When compared with the existing and GLMM-based methods, these non-zero conditional estimates were found to be similar to some method-of-moment and maximum likelihood approaches, but very different from the GLMM approaches. In terms of their performance when combined with varying summary effect confidence intervals, their results were found to be similar to those from the previous results, with the HKSJ method generally producing narrower intervals than the Z-type and t-distribution methods. These empirical analysis results give us an impression of the performance of this approach, however to fully determine which methods perform well in given rare-event scenarios, and also determine the accuracy of the approach-specific parameter τ_p^2, these methods will need to be applied to simulated data.

Chapter 6

Mixture model approach to estimate heterogeneity variance

6.1 Introduction

So far in this thesis we have looked at a range of heterogeneity variance (τ^2) estimators that we believe are appropriate for use in log-risk ratio meta-analyses, with the aim of finding those suitable for rare-event data. In Chapter 2, we began by outlining a selection of pre-existing τ^2 estimators appropriate for two-step meta-analysis approaches, most of which were based on the Normal approximation and required the use of continuity corrections, which has been shown to result in their poor performance with a high number of zero counts. We then proposed two novel approaches for the estimation of this parameter that we believed were more appropriate for application to rare-event scenarios. The first of these was based on the use of generalised linear mixed models (GLMMs), where we focused primarily on Poisson and conditional logistic mixed regression models, which benefit from not making the inappropriate assumption of normality and do not require the use of continuity corrections. Secondly, we proposed a conditional approach based on estimating equations suggested by Böhning and Sarol (2000) for use in meta-analyses with outcome measures other than the log-risk ratio, which benefited from not assuming a distribution for τ^2. Although we believe that our previously proposed approaches should perform better than existing τ^2 estimators in most rare-event scenarios, they still have their own drawbacks. For example, the GLMMs suffer from computational complexity and long fitting times, while the altered conditional approach is based on a δ-method approximation.

We shall now propose a final novel approach for the estimation of τ^2 that involves the use of mixture models, and is based on the theory behind the computer-assisted analysis of mixtures (C.A.MAN) meta-analysis approach proposed by Böhning et al. (1992). This approach models the count of events as a mixture of Binomial models, which we believe

is more appropriate and will provide a better fit than the standard Binomial distribution that this parameter is assumed to follow in other approaches, particularly for the case of rare-event data. Using this mixture model, each meta-analysis is assumed to consist of a chosen number of subpopulations, where this number is decided upon by determining the model of best fit. As such, the model will be sculpted for each individual meta-analysis that it is applied to, incorporating the clustering of subgroups of studies. Such clusters may be grouped in terms of their drug efficiency or respective number of events, as a result of differences in study factors such as gender, ethnicity and previous medical history.

In this chapter, we shall outline the theory behind this novel approach, and describe the methodology required to apply it to meta-analysis data, giving a detailed protocol of its complex application. We shall then discuss the importance of the choice of initial parameter values for this approach, and suggest possible values that we shall use ourselves. We will then outline the process for selecting the model of best fit, and list any types of meta-analyses for which the approach cannot be applied due to the structure of the data being unsuitable. Finally, the approach will be applied to the case studies introduced in Chapter 3, and the resulting estimates compared with the previous methods discussed in this thesis, in order to make conclusions regarding its general performance with empirical rare-event data.

6.2 Theory behind approach

Here we outline a novel approach involving mixture models that we are proposing for the estimation of τ^2 in rare-event meta-analyses. As with the conditional logistic model approach proposed in Section 4.4, we have that the count of events in the treatment arm of study i, denoted X_{i1}, follows a Binomial distribution:

$$X_{i1} \sim Bi(X_i, q_i)$$

where X_i is the total number of events in study i (over both the treatment and control arms) and q_i is defined as:

$$q_i = \frac{\theta'_i \frac{P_{i1}}{P_{i0}}}{\theta'_i \frac{P_{i1}}{P_{i0}} + 1} \quad , \quad q_i \in [0, 1] \tag{6.2.1}$$

where θ'_i is the outcome measure (in this case the risk-ratio), and P_{i1} and P_{i0} are the person times of the treatment and control groups respectively, of study i, where $i = 1, \ldots, k$ with k being the number of studies in the meta-analysis. The Binomial distribution of X_{i1} is a consequence of our assumption that X_i follows the Poisson

distribution. Previously, we have suggested to model the outcome measure ${\theta'}_i$ with the Normal random effect, i.e. ${\theta'}_i \sim N(\theta', {\tau^2}')$, where θ' is the overall effect measure for the meta-analysis and ${\tau^2}'$ is the heterogeneity variance. However, in this approach we shall instead propose modelling ${\theta'}_i$ with a non-parametric random effect.

It should be noted that we are currently working with the risk ratio (${\theta'}_i$), rather than the log-risk ratio, our outcome measure of interest. The rational for using the risk ratio rather than the log-risk ratio for estimating ${\tau^2}'$ in this method is that it is more direct, and as such is beneficial in this case. In order to generate estimates of τ^2 associated with the desired log-risk ratio, we shall later transform our estimates from this approach, in a similar fashion to that used in Chapter 5. We believe that the main benefit of this mixture modelling approach is that we would always obtain a valid estimate for this heterogeneity variance, ${\tau^2}'$.

6.2.1 Case of homogeneity

If homogeneity holds, i.e. there is no heterogeneity present in the meta-analysis (the heterogeneity measure $I^2 = 0\%$), then the value of q_i originally given in Equation (6.2.1) is instead defined as follows:

$$q_i = \frac{\theta' r_i}{\theta' r_i + 1}$$

where θ' is the overall outcome effect measure for the meta-analysis and $r_i = P_{i1}/P_{i0}$. This follows from the assumption that if homogeneity is present in a meta-analysis, then the true outcome effect is equal across all studies of the meta-analysis, i.e. ${\theta'}_i = \theta'$ for all $i = 1, \ldots, k$. Therefore, when homogeneity holds, the probability mass function (pmf) for the Binomial distribution of X_{i1} is defined as follows:

$$Bi\left(X_i, \frac{\theta' r_i}{\theta' r_i + 1}\right) = \begin{pmatrix} X_i \\ X_{i1} \end{pmatrix} \left(\frac{\theta' r_i}{\theta' r_i + 1}\right)^{X_{i1}} \left(\frac{1}{\theta' r_i + 1}\right)^{X_{i0}} \tag{6.2.2}$$

where X_{i0} is the count of events in the control arm of study i, and thus $X_{i0} = X_i - X_{i1}$. Equation (6.2.2) corresponds to a homogeneous Binomial model for the count of events in the treatment arm, X_{i1}.

6.2.2 Mixture of Binomials

We believe that assuming a mixture model for the count of events would be more appropriate than the Binomial model given in Equation (6.2.2), and this is a well-studied problem that we believe to be the next logical step for our meta-analysis case. As such,

we shall now look at modelling the count of events as a mixture of Binomial models, by allowing the population (in this case our meta-analysis) to consist of subpopulations represented by groups of studies with different values of the outcome measure θ'. We shall take J to be the number of subpopulations or subgroups, so $J = 1$ represents one subgroup (equivalent to the original population), $J = 2$ is two subgroups, and so forth.

To view Equation (6.2.2) as a mixing distribution, we replace the overall outcome measure θ' with the subgroup-specific outcome measures θ'_j for $j = 1, \ldots, J$, and sum the resulting j equations multiplied by some subgroup-specific weight. The mixture model that we shall consider is therefore defined as:

$$\sum_{j=1}^{J} Bi\left(X_i, \frac{\theta'_j r_i}{\theta'_j r_i + 1}\right) \pi_j = \sum_{j=1}^{J} \begin{pmatrix} X_i \\ X_{i1} \end{pmatrix} \left(\frac{\theta'_j r_i}{\theta'_j r_i + 1}\right)^{X_{i1}} \left(\frac{1}{\theta'_j r_i + 1}\right)^{X_{i0}} \pi_j \qquad (6.2.3)$$

where π_j are positive weights with $\sum_{j=1}^{J} \pi_j = 1$.

6.2.3 Mixing distribution

In order to generate the above mixed model for a meta-analysis, we need to estimate the parameters θ'_j and π_j for $j = 1, \ldots, J$. In other words, we need to estimate the following matrix of parameters:

$$\begin{pmatrix} \boldsymbol{\theta'} \\ \boldsymbol{\pi} \end{pmatrix} = \begin{pmatrix} \theta'_1 & \ldots & \theta'_J \\ \pi_1 & \ldots & \pi_J \end{pmatrix} \qquad (6.2.4)$$

which can be described as the mixing distribution.

In the classical problem of the estimation of $\tau^{2'}$, only θ' was present as the Binomial parameter and needed to be estimated. In this case however, the Binomial parameter is more complicated, as all of the elements in the array in Equation (6.2.4) need to be estimated. To estimate this parameter array, we shall use the expectation-maximisation (EM) algorithm, which we shall describe in the next section. Once this array has been estimated, the estimates for the heterogeneity variance ($\tau^{2'}$) and overall effect size ($\bar{\theta}'$), both corresponding to the risk ratio, are then calculated as follows:

$$\hat{\tau}^{2'} = \sum_{j=1}^{J} \hat{\pi}_j (\hat{\theta'}_j - \hat{\bar{\theta}'})^2 \qquad (6.2.5)$$

$$\widehat{RR} = \hat{\bar{\theta}'} = \sum_{j=1}^{J} \hat{\pi}_j \hat{\theta'}_j \qquad (6.2.6)$$

6.3 Outline of approach

6.3.1 EM algorithm

As discussed in the previous section, we shall estimate the array in Equation (6.2.4) using the EM algorithm. To conduct this algorithm, we first calculate the observed, or incomplete, likelihood, which for our mixture model is given as:

$$L_O = \prod_{i=1}^{k} \left(\sum_{j=1}^{J} Bi\left(X_i, \frac{\theta'_j}{\theta'_j + 1}\right) \pi_j \right) \tag{6.3.1}$$

This likelihood is the product over the observed data, in our case over the k studies in the meta-analysis. Here we have simplified the mixture density, $Bi\left(X_i, \frac{\theta'_j r_i}{\theta'_j r_i + 1}\right)$, and assumed that the person time is equal across treatment arms in each study i, i.e. $P_{i1} = P_{i0}$, and thus the ratios $r_i = P_{i1}/P_{i0}$ simplify to 1 in Equation (6.3.1).

The observed log-likelihood is therefore given as:

$$l_O = \log L_O = \sum_{i=1}^{k} \log \sum_{j=1}^{J} Bi\left(X_i, \frac{\theta'_j}{\theta'_j + 1}\right) \pi_j$$

The associated complete likelihood will involve a latent, unobserved variable, Z_j, which is an indicator that takes on the value 1 if X_1 is from the j^{th} subpopulation, and 0 otherwise. Let z_{ij} denote the value of Z_j for observation X_{i1}. Then the complete likelihood, also called the complete data likelihood, is given by:

$$L_C = \prod_{i=1}^{k} \prod_{j=1}^{J} Bi\left(X_i, \frac{\theta'_j}{\theta'_j + 1}\right)^{z_{ij}} \pi_j^{z_{ij}} \tag{6.3.2}$$

Here we operate as though the classification or subgroup that each study belongs to is known. The benefit of this is that when we take the log-likelihood, i.e. when we take $\log(L_C)$, then the resulting equation separates into summed elements that can be maximised independently. Thus, when we take the logarithm of Equation (6.3.2), we produce the following complete log-likelihood:

$$l_C = \log L_C = \sum_{i=1}^{k} \sum_{j=1}^{J} z_{ij} \log Bi\left(X_i, \frac{\theta'_j}{\theta'_j + 1}\right) + \sum_{i=1}^{k} \sum_{j=1}^{J} z_{ij} \log \pi_j \tag{6.3.3}$$

where the term $\sum_{i=1}^{k} \sum_{j=1}^{J} z_{ij} \log Bi\left(X_i, \frac{\theta'_j}{\theta'_j + 1}\right)$ is independent of $\boldsymbol{\pi}$. The aim is to determine the values of θ'_j from Equation (6.3.3). This equation can be separated because

when taking the partial derivative with respect to θ'_j, it only depends on a single component. Since this component is solvable in the homogeneous case, then it follows that it can also be solved here in the case where heterogeneity is present.

The maximisation of Equation (6.3.3) leads to $\hat{\pi}_j = \sum_{i=1}^{k} z_{ij}/k$ for $j = 1, \ldots, J$. It remains to obtain values for z_{ij}, since these have not been observed. It is natural to replace them with their expected values given the data $X_{i1}, \ldots, X_{k1}$ and $\boldsymbol{\pi}$: $E(z_{ij}|X_{i1}, \boldsymbol{\pi}) = Pr_{\boldsymbol{\pi}}(Z_{ij} = 1|Y_{i1} = X_{i1})$.

Here we shall prove that $\sum_{j=1}^{J} \hat{\pi}_j = 1$ when the weight estimates satisfy $\hat{\pi}_j = \sum_{i=1}^{k} z_{ij}/k$:

$$\begin{aligned}
\sum_{j=1}^{J} \hat{\pi}_j &= \sum_{j=1}^{J} \sum_{i=1}^{k} z_{ij}/k \\
&= \sum_{i=1}^{k} \sum_{j=1}^{J} z_{ij}/k \\
&= \sum_{i=1}^{k} 1/k \\
&= k/k \\
&= 1
\end{aligned}$$

where the third equality comes from the fact that each study i belongs in only one subgroup j, and since z_{ij} is an indicator variable for whether study i belongs to subgroup j, i.e. $z_{ij} = 0, 1$, then it follows that $\sum_{j=1}^{J} z_{ij} = 1$ for $i = 1, \ldots, k$.

6.3.2 E-step

To conduct the expectation (E)-step of the EM algorithm, we take the expected values of z_{ij} using a base generic formula. By Bayes Theorem, we have that:

$$\begin{aligned}
E(z_{ij}) &= e_{ij} \\
&= \frac{Bi\left(X_i, \frac{\theta'_j}{\theta'_j+1}\right)\pi_j}{\sum_{j'=1}^{J} Bi\left(X_i, \frac{\theta'_{j'}}{\theta'_{j'}+1}\right)\pi_{j'}}
\end{aligned}$$

where $J = 1, 2, \ldots$ is the number of subgroups to be decided upon, and j' refers to the total J subgroups. Replacing z_{ij} by e_{ij} in Equation (6.3.3) leads to the expected complete log-likelihood:

$$E(l_C) = \sum_{i=1}^{k}\sum_{j=1}^{J} e_{ij} \log Bi\left(X_i, \frac{\theta'_j}{\theta'_j + 1}\right) + \sum_{i=1}^{k}\sum_{j=1}^{J} e_{ij} \log \pi_j \tag{6.3.4}$$

6.3.3 M-step

Since the component parameters $\theta'_1, \ldots, \theta'_j$ are unknown, we must take the expression $\sum_{i=1}^{k}\sum_{j=1}^{J} e_{ij} \log \pi_j$ in Equation (6.3.4) into account. As the complete log-likelihood separates into two parts, the first depending on $\boldsymbol{\pi}$ only and the second depending on $\theta'_1, \ldots, \theta'_j$ only, we can incorporate the estimation of $\theta'_1, \ldots, \theta'_j$ by finding the maximum of $\sum_{i=1}^{k}\sum_{j=1}^{J} e_{ij} \log \pi_j$.

Maximising Equation (6.3.4) leads to the maximisation (M)-step and the solution:

$$\begin{aligned}
\pi_j^{(new)} &= \frac{\sum_{i=1}^{k} e_{ij}}{k} \\
&= \frac{\sum_{i=1}^{k} \frac{Bi\left(X_i, \frac{\theta'_j}{\theta'_j+1}\right)\pi_j}{\sum_{j'=1}^{J} Bi\left(X_i, \frac{\theta'_{j'}}{\theta'_{j'}+1}\right)\pi_{j'}}}{k} \\
&= \sum_{i=1}^{k} \frac{Bi\left(X_i, \frac{\theta'_j}{\theta'_j+1}\right)\pi_j}{k\sum_{j'=1}^{J} Bi\left(X_i, \frac{\theta'_{j'}}{\theta'_{j'}+1}\right)\pi_{j'}}
\end{aligned}$$

From the M-step, we also find that:

$$\frac{\theta_j^{'(new)}}{\theta_j^{'(new)} + 1} = \frac{\sum_{i=1}^{k} e_{ij} X_{i1}}{\sum_{i=1}^{k} e_{ij} X_i}$$

Thus, the M-step leads to a form of the weighted mean of the observed counts of events X_{i1}. Rearranging the above formula gives us:

$$\theta_j^{'(new)} = \frac{\sum_{i=1}^{k} e_{ij} X_{i1}}{\sum_{i=1}^{k} e_{ij} X_i - \sum_{i=1}^{k} e_{ij} X_{i1}} \tag{6.3.5}$$

The approach is based on obtaining the original estimate of θ'_j and then generating the updated estimate $\theta_j^{'(new)}$ in the M-step, so that the the expected values can be calculated and used to obtain new values, which in turn shall replace the existing estimates. The value of z_{ij} can also be replaced by e_{ij}, then by $e_{ij}^{(new)}$, and so forth. By doing this for

both of these parameters, we shall determine the values for ${\theta'}_{ij}$ that we are interested in.

Since the M-step is solvable in the homogeneous case, we assume that it is also solvable in our scenario where heterogeneity is present. If it is solvable in the homogeneous case, then the first task is to find the solution for this homogeneous case, i.e. the estimate of θ'. The required estimates for θ'_j will be alternate variations of that value that also involve a weight.

We shall fix the number of components, J, so that we start with $J = 1$ (the homogeneous case), and then look at $J = 2$, $J = 3$, and so forth. In order to assess the performance of these mixture models with varying values of J in estimating our parameters of interest, we shall look at measures of fit such as the Akaike information criterion (AIC) and Bayesian information criterion (BIC). In particular, we will look at the cases where $J = 1, 2, \ldots, s$ with some pre-specified value s that reflects the predicted maximum degree of clustering of studies. While $J = 1$ corresponds to ${\tau^2}' = 0$, increasing the number of subgroups J will in turn increase the value of ${\tau^2}'$:

$$ {\tau^2}'_{(J+1)} > {\tau^2}'_{(J)} > \cdots > {\tau^2}'_{(1)} = 0 \qquad , \; J \geq 2 $$

We shall set the weights, π_j, to be equal, i.e. $\pi_j = 1/J$ for $j = 1, \ldots, J$, and will estimate the k study-specific risk ratios $(\hat{\theta}'_{(1)}, \ldots, \hat{\theta}'_{(k)})$ after each loop of the algorithm.

6.3.4 Algorithm protocol

In summary, the EM algorithm required for this approach will be applied as follows:

1. Let $\boldsymbol{\pi}$ be any initial vector of weights and $\boldsymbol{\theta'}$ any initial vector of component parameters (in this case the risk ratio outcome measure).
2. Compute the E-step according to Section 6.3.2.
3. Compute the M-step according to Section 6.3.3, leading to $\boldsymbol{\pi}^{(new)}$ and the appropriate updated vector of $\boldsymbol{\theta'}^{(new)}$.
4. Set $\boldsymbol{\pi} = \boldsymbol{\pi}^{(new)}$ and $\boldsymbol{\theta'} = \boldsymbol{\theta'}^{(new)}$.
5. Repeat steps 1 to 4 until the condition of some pre-specified stopping rule is met, e.g. the difference in observed log-likelihoods is less than say 1×10^{-7} or some maximum number of iterations is achieved.

6.3.5 Conversion of estimates to log-risk ratio

After applying the mixture model approach described above to estimate the risk ratio and its associated heterogeneity variance, we want to convert these estimates to correspond with our outcome measure of interest - the log-risk ratio. To convert the risk ratio to the log-risk ratio, we derive the following relationship between θ'_j and θ_j:

$$\begin{aligned} q_j &= \frac{e^{\theta_j}}{1+e^{\theta_j}} \\ \Rightarrow q_j + q_j e^{\theta_j} &= e^{\theta_j} \\ \Rightarrow q_j &= (1-q_j)e^{\theta_j} \\ \Rightarrow \theta_j &= \log \frac{q_j}{1-q_j} \\ \Rightarrow \theta_j &= \log \theta'_j \end{aligned}$$

By substituting this expression for θ_j into the estimate of the overall risk ratio given in Equation (6.2.6), we can determine the associated estimate for the overall log-risk ratio, denoted $\bar{\theta}$:

$$\widehat{\log RR} = \hat{\bar{\theta}} = \sum_{j=1}^{J} \pi_j \log \theta'_j \tag{6.3.6}$$

where the hat over the entirety of the $\log RR$ term comes from the fact that the sum of logs is not equal to the log of sums, and so this is an estimate of the function $\log RR$ itself. This estimate is classed as a non-parametric maximum-likelihood estimate (MLE). Similarly, for the conversion of $\hat{\tau}^{2'}$ to $\hat{\tau}^2$, we can identify the estimates that $\hat{\tau}^{2'}$ depends on in Equation (6.2.5):

$$\hat{\tau}^{2'} \leftarrow \begin{pmatrix} \hat{\theta'}_1 & \dots & \hat{\theta'}_j \\ \hat{q}_1 & \dots & \hat{q}_J \\ \hat{\pi}_1 & \dots & \hat{\pi}_J \end{pmatrix}$$

and so can list the transformations that are required for the construction of $\hat{\tau}^2$:

$$\hat{\tau}^2 \leftarrow \begin{pmatrix} \hat{\theta}_1 & \dots & \hat{\theta}_J \\ \log \frac{\hat{q}_1}{1-\hat{q}_1} & \dots & \log \frac{\hat{q}_J}{1-\hat{q}_J} \\ \hat{\pi}_1 & \dots & \hat{\pi}_J \end{pmatrix}$$

By substituting those values seen in the above matrices into Equation (6.2.5), and using the relationship between θ'_j and θ_j described above, we can determine the desired log RR-associated estimate for the heterogeneity variance $\hat{\tau}^2$:

$$\hat{\tau}^2 = \sum_{j=1}^{J} \hat{\pi}_j \left(\hat{\theta}_j - \hat{\bar{\theta}}\right)^2 \tag{6.3.7}$$

It is these converted values of $\widehat{\log RR}$ and $\hat{\tau}^2$, from Equations (6.3.6) and (6.3.7) respectively, that we shall calculate from the extracted model output and represent our final estimates of interest.

6.4 Case if the within-study person times are unequal

Up until this point, we have assumed that the within-trial person times are equal, i.e. $P_{i0} = P_{i1}$, and so $r_i = 1$ in the mixture density given in Equation (6.2.3). However, if we look at the case where $P_{i0} \neq P_{i1}$, and so $r_i \neq 1$, the corresponding observed likelihood in Equation (6.3.1) can be rewritten as:

$$L_O = \prod_{i=1}^{k} \left(\sum_{j=1}^{J} Bi\left(X_i, \frac{\theta'_j r_i}{\theta'_j r_i + 1}\right) \pi_j \right)$$

Updating the corresponding likelihoods and applying the EM algorithm as before, the j iterative estimates of $\boldsymbol{\theta'}$ constructed in the M-step of the algorithm, denoted by $\theta_j^{'(new)}$, now have to be generated using the iterative equation:

$$\theta_j^{'(new)} = \frac{\sum_{i=1}^{k} e_{ij} X_{i1}}{\sum_{i=1}^{k} e_{ij} \frac{X_i r_i}{\theta_j^{'(new)} r_i + 1}}$$

with initial estimate θ'_j. The proof for this is given in Appendix C. The corresponding derived estimates for the log-risk ratio in this case are identical to those given previously, i.e. $\widehat{\log RR} = \sum_{j=1}^{J} \pi_j \log \theta'_j$ and $\hat{\tau}^2 = \sum_{j=1}^{J} \hat{\pi}_j (\hat{\theta}_j - \hat{\bar{\theta}})^2$, but with an alternate definition for the input parameter θ'_j:

$$\begin{aligned} \theta'_j &= \frac{q_{ij}}{r_i(1 - q_{ij})} \\ &= \frac{P_{i0} q_{ij}}{P_{i1}(1 - q_{ij})} \end{aligned}$$

where $q_{ij} = \frac{r_i e^{\theta_j}}{r_i e^{\theta_j}+1}$.

As such, the main difference between this case and the simplified case where $r_i = 1$, i.e. $P_{i1} = P_{i0}$, is the computation of $\theta_j^{'(new)}$ in the M-step, which is now much more complicated due to the need for iteration. Apart from this step, when $r_i = 1$, all other equations simplify to the estimates given in Section 6.3 and reflect the more general case where the person times may or may not be equal.

6.5 Additional aspects of approach

6.5.1 Choice of initial values of $\boldsymbol{\pi}$ and $\boldsymbol{\theta'}$

At the start of the protocol detailed in Section 6.3.4, we have to input initial values for the vectors $\boldsymbol{\pi}$ and $\boldsymbol{\theta'}$. In order to obtain more accurate results from the algorithm, these initial values should reflect realistic values that could be representative of the scenario being investigated. In order for the initial values of $\boldsymbol{\pi}$ to be fair and realistic for each component-based model considered, we set $\pi_1 = \cdots = \pi_J = 1/J$, which gives us $\sum_{j=1}^{J} \pi_j = 1$ as required. We based the initial values of $\boldsymbol{\theta'}$ on the calculated risk ratio for each of the studies in the meta-analysis to which the method is being applied (including those calculated from single-zero trials where a continuity correction of 0.5 has been applied). We outline the process of choosing the initial values of $\boldsymbol{\theta'}$ (and subsequently $\boldsymbol{q} = \frac{\boldsymbol{\theta'}}{\boldsymbol{\theta'}+1}$) below:

1. Apply a continuity correction of 0.5 to any single-zero trials in the meta-analysis (double-zero trials do not need to be considered as these will be omitted in order for this method to be applied).
2. Calculate the risk ratio for each of the studies in the meta-analysis (ignoring double-zero trials), giving us the vector $\boldsymbol{RR} = (RR_1, \ldots, RR_k)$ where k is the number of studies in the meta-analysis.
3. Set $\theta'_1 = min(\boldsymbol{RR})$ and $\theta'_J = max(\boldsymbol{RR})$.
4. Calculate the associated Binomial probabilities q_1 and q_J from these values of θ'_1 and θ'_J using the equation $q_j = \frac{\theta'_j}{\theta'_j+1}$ for $j = 1, J$.
5. If the number of subgroups $J > 2$, set the remaining non-extreme values of $\boldsymbol{q}$ to be equally-spaced apart, i.e. $q_j = q_1 + (J-1)\frac{(q_J - q_1)}{J}$ for $j = 2, \ldots, (J-1)$.

By choosing the initial values of $\boldsymbol{\theta'}$ (and $\boldsymbol{q}$) in this manner, we are guaranteeing that they represent the risk ratios of the studies being used, thus tailoring the algorithm to the meta-analysis that the approach is being applied to, with the intention that the algorithm should then provide us with the most accurate estimates.

6.5.2 Selection of best-fitting model

To find the model of best fit for each meta-analysis, we compare models with varying numbers of components. We decided to base the selection of this best-fitting model on the similarity of the Binomial probabilities, $\boldsymbol{q} = (q_1, \ldots, q_J)$, computed for each component-specific model. If none of the models produce similar probabilities across their individual components, then a secondary model-selection approach must be specified and applied - either the Bayesian Information Criterion (BIC) or likelihood ratio test (LRT). Below is an outline of the steps taken to determine the best-fitting model in our approach and the code we have written for its application:

1. Fit 5 models, with the number of components (J) ranging from 1 to 5.
2. For each model with $J \geq 2$, determine if the absolute difference between any two of the adjacent components' Binomial probabilities (q) is < 0.001, i.e. whether $|q_i - q_{i+1}| < 0.001$ for any $i = 1, \ldots, (J-1)$. If this is the case, take the lowest J that satisfies this requirement, and choose the model of best fit as that with $J-1$ components.
3. If the condition in Step 2 is not satisfied, and no model of best fit can be determined via that selection technique, choose a second model selection criteria out of BIC and LRT.
4. If BIC is selected, the best-fitting model is chosen as that with the lowest BIC score.
5. If LRT is selected, then a likelihood ratio test is conducted to find the best-fitting model.
6. Estimates of τ^2 and $\log RR$ are extracted from the resulting model of best fit using the transformations described in the above section.

Here we shall briefly discuss the choice of model-selection approach, and in particular the LRT, for which mixture models experience a boundary problem. To demonstrate this issue, consider a simple two-component mixture, $(1-\alpha)f_1(x) + \alpha f_2(x)$. We are interested in the hypothesis test H_0: one component is present vs. H_1: two components are present. If $\alpha = 0$ or $\alpha = 1$, then H_0 is true, and so the null hypothesis lies entirely in the boundary of the interval $[0, 1]$, which is the feasible parameter space. Therefore, if the null hypothesis is true, we would expect a normal distribution for the maximum likelihood estimate of α, however this cannot be the case as values smaller than 0 or larger than 1 cannot occur. Hence the likelihood ratio statistic will follow a non-standard distribution. This distribution can be determined in special cases (Böhning et al. (1994)), but is complex to determine in more general settings. A bootstrap approach has been

suggested as a solution (Richardson and Green (1997)), but this is quite computer-intensive. A simpler solution is to use the BIC for model selection, as this works well for mixture models, and so has been set as the default approach in our algorithm here.

Figure 7.1 displays the selection technique for the model of best fit, as applied with our written code for the approach.

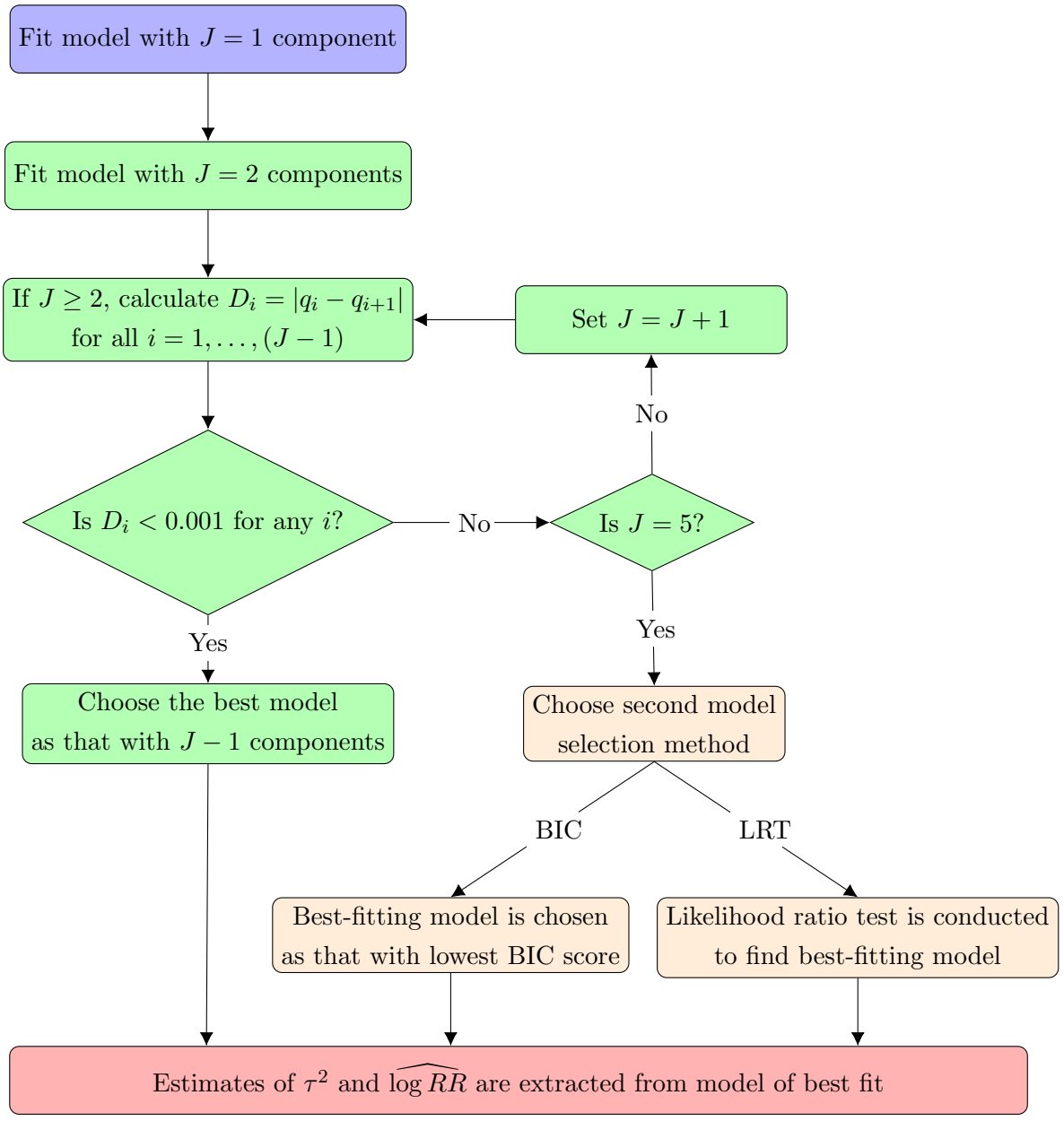


FIGURE 6.1: Outline of steps taken to determine the best-fitting model in mixture model approach.

We decided that a maximum of 5 components was sufficient to be tested, as none of the case studies we investigated reached this maximum of 5 when applying the above

algorithm, and the types of meta-analyses that we are simulating are unlikely to exceed this number of clusters. We set the value for the the cut-off marker for the difference between Binomial probabilities across components within the same model as 0.001, as this allowed us to distinguish those components giving similar values and thus representing excessive components to the model, and agreed with other sources of information. For the stopping rules of the EM algorithm protocol, we set the absolute difference between observed log-likelihoods to be 1×10^{-7} as this was the default value in the comparable C.A.MAN package in R, and it seemed an cut-off value, and the maximum number of iterations to be 5000.

6.5.3 Unsuitable meta-analyses

There are a number of scenarios where this approach cannot be applied. As mentioned previously, all double-zero trials have to be omitted from the meta-analysis in order to apply this approach. In addition to this, the approach can not be applied to meta-analyses where all studies have zero events in the control arm, i.e. $X_{i0} = 0$ for all $i = 1, \ldots, k$, as this is equivalent to $X_i = X_{i1}$ for all $i = 1, \ldots, k$, which leads to $\theta_j^{'(new)}$ being infinite in the M-step, and thus the Binomial probability q is undefined by construction, and the algorithm breaks down.

During the algorithm, we have also had to identify those cases where the value of e_{ij} (from the E-step) is so small (e.g. $< 1 \times 10^{-15}$) that the resulting denominator of $\theta_j^{'(new)}$ is taken to be zero in R due to rounding errors, meaning that $\theta_j^{'(new)}$ itself becomes infinite in the M-step and the algorithm breaks down as described above. In order to prevent this from happening, when constructing the values of e_{ij}, we checked whether any of the values of e_{ij} for a fixed j will cause the resulting value of $\theta_j^{'(new)}$ to be infinite. If this is the case, then we set the smallest e_{ij} for the respective value of j to instead be $1/j$ until $\theta_j^{'(new)}$ becomes finite, and the algorithm can continue. Unfortunately, this alteration tends to result in the number of iterations reaching the set maximum number defined as the cut-off (e.g. 5000), meaning the output estimates are undefined as a result. Meta-analyses that have been identified as causing this type of error include those where the number of studies $k \leq 5$ as a result of omitting double-zero trials, or where the sample size is small, e.g. 10, and the proportion of single-zero studies is significantly high. We could not find any patterns between the meta-analyses that lead to this problem, so assumed that it is not a simple trait to identify, and merely related to the severe rarity of the data identified. However, as we could apply the above mentioned fix to the method without identifying the specific meta-analyses that would cause a problem, it was not necessary to identify them here.

6.6 Comparison with existing mixture model packages

We constructed a function in the statistical software package R to apply this mixture model approach via the EM algorithm described in the previous sections. To ensure that this user-written code worked correctly, and that the appropriate parameters were being extracted after model fitting, we applied it to the case study datasets introduced in Chapter 3 and compared the results with those produced using pre-existing packages capable of fitting this type of model. We decided to use the C.A.MAN package in R as our pre-existing method for comparison, because it was designed to be based on a similar mixture model approach, and met the requirements of our own proposed approach (Schlattmann et al. (2016); Böhning et al. (1998); Schlattmann (2009)).

6.6.1 Case studies

We applied both our user-written code and the C.A.MAN package in R to the rare-event meta-analysis case studies described previously, and the results can be seen below in Table 6.1. Although we are interested in comparing the estimates for the overall log-risk ratio and the heterogeneity variance parameters, we have also recorded characteristics of the models that the two approaches deemed to be best-fitting for each of the datasets, including the number of components, weights and associated Binomial probabilities.

From Table 6.1, we can see that the user-written code and C.A.MAN. package behaved very similarly for each of the rare-event case studies investigated, selecting models of best fit with the same number of components, and producing identical results to several decimal places in all cases. As our user-written code agrees with that of a pre-existing mixture model package, we can say with confidence that our code correctly applies the approach, and extracts the appropriate estimates as required. Focusing on the results produced using our own code, we can see that the HKSJ confidence intervals for the summary effect are consistently narrowest of those considered for all datasets, while the alternate Z-type and t-distribution methods produce very similar intervals. We only present the confidence intervals for our own code here, as we were only interested in comparing them against each other, not against that produced via the C.A.MAN. package, which we have used to confirm the appropriateness of our code using the point estimates alone.

When comparing these estimates to those produced with the existing estimators, our two GLMM-based approaches and our conditional approach, that are presented in Chapter 3, and Tables 4.1, 4.2 and 5.2 respectively, we can see that the similarity in results depends very much on the dataset itself. In terms of the pre-proposed estimators, this approach produces similar τ^2 estimates to the Sidik-Jonkman estimator for the rosiglitazone (MI) and C-section datasets, and zero estimates that mirror the majority of these normal-based methods for the rosiglitazone (death), albumin and transplant case studies. With

TABLE 6.1: Summary of results from the best-fitting mixture models according to our written code and the C.A.MAN package for case studies; **J** is the number of components for the best-fitting model.

Dataset	**R Package**	**J**	**Weights**	**Probabilities**	$\hat{\tau}^2$	$\hat{I}^2$	$\widehat{\log RR}$	**Confidence Interval**		
								Z-type CI	***t*-type CI**	**HKSJ CI**
Rosig (MI)	Own code	2	0.198, 0.802	0.415, 0.679	0.189	13.31	0.535	(0.13, 0.94)	(0.12, 0.95)	(0.26, 0.81)
	C.A.MAN	2	0.198, 0.802	0.415, 0.679	0.189	13.31	0.535			
Rosig (death)	Own code	2	0.468, 0.532	0.629, 0.649	0.000178	0.0084	0.574	(0.15, 1.00)	(0.14, 1.01)	(0.34, 0.81)
	C.A.MAN	2	0.471, 0.529	0.688, 0.641	5.60×10^{-5}	0.0027	0.573			
CRBSI	Own code	2	0.348, 0.654	0.072, 0.339	0.68	40.73	-1.14	(-1.98, -0.66)	(-2.03, -0.61)	(-1.84, -0.80)
	C.A.MAN	2	0.346, 0.654	0.0719, 0.3393	0.68	40.73	-1.14			
C-section	Own code	3	0.3425, 0.6026, 0.0548	0.174, 0.348, 0.666	0.332	38.07	-0.875	(-1.14, -0.61)	(-1.14, -0.61)	(-1.08, -0.66)
	C.A.MAN	3	0.3429, 0.6023, 0.0548	0.174, 0.348, 0.666	0.332	38.07	-0.875			
Albumin	Own code	1	1	0.628	0	0	0.525	(0.24, 0.80)	(0.23, 0.82)	(0.28, 0.77)
	C.A.MAN	1	1	0.628	0	0	0.525			
Transplant	Own code	1	1	0.667	0	0	0.69	(-1.13, 2.52)	(-3.31, 4.70)	(-1.53, 2.91)
	C.A.MAN	1	1	0.667	0	0	0.69			

the CRBI meta-analysis, however, this approach produces a τ^2 estimate higher than any of the existing methods. In terms of the log-risk ratio estimate, this mixture model approach produces estimates of higher magnitude to that of the pre-existing methods for almost all of the datasets considered. Compared to the Poisson and conditional logistic mixed regression models, the estimates produced here were consistently very similar for all datasets, with the conditional logistic model actually having identical results for the 3-study transplant meta-analysis, potentially demonstrating their joint ability to perform well with such small meta-analyses. Finally, comparing these results to those of our conditional approach, we can see that the results in general are not very similar. In particular, the non-zero CRBSI estimate from the conditional approach was not replicated here, and the τ^2 estimates for albumin were different, although the associated log-risk ratio estimates were very similar.

6.7 Conclusions

In this chapter, we proposed and outlined a novel approach for the estimation of the heterogeneity variance in rare-event meta-analyses that is based on the use of a mixture model. In particular, it assumes that the count of events is modelled by a mixture of Binomial models, rather than a singular standard Binomial distribution like other approaches. This use of a mixture model allows for the potential clusters of studies within a meta-analysis to be taken into account. The estimates of interest are generated via an EM algorithm, with the associated model of best fit being found via an intense protocol for model selection. As the model of best fit is determined for each individual meta-analysis, based on its own individual characteristics, the approach is very specific in its application, which we believe will result in more accurate estimates for τ^2.

As with the other methods discussed previously in this thesis, this approach has some drawbacks. In particular, double-zero trials must be omitted in for the approach to be applied to the respective meta-analysis, as these are undefined in the application of the model. Additionally, meta-analyses that met certain requirements could not be used with the approach as a result of the sparsity of their events. For example, meta-analyses that contained only studies with zero events in their control arms were unsuitable, as were those with less than 5 studies and small sample sizes. Although these restrictions for the applicability of the approach are obviously a disadvantage, the characteristics defining these unsuitable meta-analyses are unlikely to be seen very frequently in empirical data.

We have written a function to apply this approach, including options for the user to adjust model selection criteria to those desired. In order to confirm that this function applied the approach correctly, and extracted the appropriate final estimates, we compared it to a pre-existing package that can perform the mixture model section of the

approach. We applied both sets of code to our rare-event case studies, and found that the results, and accompanying characteristics of models of best fit, were identical in all cases, thus backing up the validity of our extended version of code. When these results were compared with those produced via the methods proposed previously in this thesis, it could be seen that this approach generated very similar results to the GLMM-based estimators, particularly for the conditional logistic regression model, where the estimates were identical for one dataset. In terms of the summary effect confidence intervals, this approach mirrored our previous estimators in terms of the HKSJ method consistently generating the narrowest intervals.

Chapter 7

Methods for simulation study

7.1 Introduction

So far in this thesis we have outlined a number of methods to estimate the heterogeneity variance (τ^2) in meta-analyses, some of which we have proposed ourselves, and applied them to a selection of empirical rare-event case studies. In order to gain a comprehensive overview of the performance of these estimators in a wide range of real-life rare-event scenarios, we shall also conduct a simulation study.

The aim of our simulation study is to produce a set of realistic meta-analyses that we can use to compare the novel approaches we proposed in Chapters 4 to 6 to the existing τ^2 estimators discussed in Chapter 2. We are interested only in designing meta-analyses based on binary outcome data, and using the log-risk ratio as our outcome measure since this represents the preferential choice in medical research analysis. Our focus will be on examining different scenarios based on characteristics such as event probability and sample size, in order to determine which τ^2 estimators perform better in given situations. As we are concerned with rare-event data, we shall focus primarily on those situations which would fall under this definition, i.e. meta-analyses containing single- and double-zero studies. However, we shall also examine less restrictive, more common-event scenarios in order to provide a complete overview of the estimators of interest.

In order to assess the performance of our chosen methods under the simulated scenarios, we shall investigate how they perform in terms of estimating τ^2 and the log-risk ratio outcome measure, as well as the proportion of zero estimates they generate and their coverage when applied with certain confidence interval methods. In terms of the two parameter point estimates, we shall measure the estimation performance using a range of measures including the bias and mean squared error. Once we have this information, graphs and tables displaying these results can then be produced, and an analysis of the results will be conducted. Finally, we shall make recommendations based on these

results with respect to the application of τ^2 estimators in the case of a variety of scenarios relating to sparse-event data.

In this chapter we shall outline the methods required to conduct our simulation study, giving details of how we chose to design our simulations and subsequently compare our τ^2 estimators of interest. All simulations and calculations described in this chapter were conducted using the statistical software package R, unless otherwise stated.

7.2 Simulation study design

As we are interested only in binary outcome measures, we needed to generate a pair of sample sizes and event counts for each study, corresponding to those of the study-specific treatment and control arms, along with an individual study effect measure, of each simulated meta-analysis. There are a number of parameters that we controlled the values of during our simulation study by selecting a range of values that reflect diverse realistic situations, with each independent combination of these parameter values forming one of our distinct simulation scenarios. Such parameters include the sample sizes of each study, which for simplicity we set as equal across the treatment and control arms, and the number of studies (k). With these two parameters in particular, we chose to include values that loosely correspond to scenarios of extreme sparsity in meta-analyses, i.e. $k \leq 5$ and/or sample sizes of ≤ 20. We also needed to define the true heterogeneity variance, τ^2, which we chose to correspond with a specified level of heterogeneity, given by the measure I^2.

Given the scenarios governed by our specified range of set parameters, we began simulating values for the probability of events. This is denoted by p_{ij}, which represents the probability of the event of interest occurring in treatment arm j of study i in the meta-analysis, where $j = 1$ represents the active treatment arm and $j = 0$ the control/placebo arm. From this, we were then able to calculate the effect measure of each study i, denoted by θ_i. Since we are interested in the log-risk ratio, $\theta_i = \log RR_i$ in our simulation study. Finally, using all of the previously defined and simulated parameters, we were able to sample the count of events, X_{ij}, in the treatment and control arms of each study of the meta-analysis. Once all of the above steps were complete, we had accumulated all of the essential information to be collected from a meta-analysis, and thus had all of the parameters necessary for use with our heterogeneity variance estimating approaches.

We have based the design of our simulation study on previous studies that have been conducted to assess the performance of heterogeneity variance estimators in meta-analyses (Langan (2015); Langan et al. (2016); Veroniki et al. (2016)). Throughout the remainder of this section we shall describe in detail the processes undertaken to simulate our data, as well as the reasons behind our choice in parameter ranges and sampling techniques. A step-by-step outline of the protocol for our simulation methods is as follows:

1. Set values for number of studies
2. Sample or set values for the study sample sizes, as required
3. Calculate true heterogeneity variance based on chosen values of I^2
4. Sample the probability of events
5. Sample the study-specific effect measure
6. Sample the count of events.

7.2.1 Summary of simulated scenarios

A full list of parameter values and distributions investigated in our simulation study is given in Table 7.1. Each individual combination of these parameter values and distributions shall constitute as one scenario, giving a total of 6720 scenarios for investigation. Although this is a large number of scenarios to simulate, and will therefore be computationally demanding, we believe that it will provide us with the most complete overview of the estimators available in all possible cases. For each defined scenario, we simulated 1000 meta-analyses, as we believed this to be sufficient based on similar simulation studies.

We avoided simulating any scenarios that we found our novel approaches could not be applied to during the course of the simulation study, and will list these with the results. The values for the mean baseline risk (α) and mean log-relative risk (β) in Table 7.1 are paired to correspond to specific event probability scenarios, as discussed later in the chapter. As stated above, all stages of the simulation study were conducted using the statistical software package R, and our corresponding code is given in Appendix D.1. In order to make our results reproducible, we used a seed of 24601 within our script.

TABLE 7.1: Set of parameter values and distributions that shall define simulated meta-analysis scenarios; * denotes parameters paired to correspond to set event probabilities.

Parameter		**Value/distribution**
k	Number of studies in the meta-analysis	2, 3, 5, 10, 20, 30, 50, 100
n_{i0}, n_{i1}	Study sample sizes	(a) Small studies: $n_{i0} = 20$ (b) Small to medium sized studies: $n_{i0} \sim U(20, 200)$ (c) Medium sized studies: $n_{i0} = 200$ (d) Small and large studies: $n_{10}, \ldots, n_{m0} = 20$ and $n_{(m+1)0}, \ldots, n_{k0} \sim U(1000, 2000)$ where m is the integer half way between 1 and k (when k is odd, one study is to be generated from one of the two distributions at random) (e) Large studies: $n_{i0} \sim U(1000, 2000)$ In all scenarios, sample sizes are equal between groups ($n_{i0} = n_{i1}$)
τ^2	True value of heterogeneity variance	0, 0.2, 0.4, 0.6, 0.8, 1
α	Mean baseline risk*	-6.9, -5.3, -4.6, -3.0, -2.3, -0.7
σ_α^2	Variance of study-specific baseline risk	0.1, 3
β	Mean log-relative risk*	-1.6, 0, 1.6
X_{i0}, X_{i1}	Count of events	$X_{ij} \sim Binomial(n_{ij}, p_{ij})$, $X_{ij} \sim Poisson(n_{ij} \times p_{ij})$

7.2.2 Number of studies

As mentioned in Chapter 1, meta-analyses with few numbers of studies ($k \leq 5$) represent a special area of interest with regards to methodology, as most approaches perform very poorly in such scenarios, regardless of whether the events are rare or not. We have chosen to include the scenarios $k = 2, 3, 5$ in our range of values for k, where $k = 2$ represents the extreme minimum case of having only 2 studies in the meta-analysis, a scenario that has been previously covered briefly by Friede et al. (2017b). Although 95% of meta-analyses in the Cochrane Database of Systematic Reviews contain fewer than 16 studies (Handoll et al. (2008)), we extended our range of values for k from 2 to 100,

as this upper boundary value aims to account for meta-analyses with a higher number of studies in fields outside of medicine.

7.2.3 Study sample sizes

As another special area of interest in meta-analysis methodology revolves around small sample sizes, i.e. $\leq$ 20 subjects per trial arm in each study, we also included this scenario in our simulation design, looking at cases where the trial arms contain as little as 20 individuals. To represent this situation, as well as a wide range of other realistic scenarios that occur in empirical datasets, study sample sizes (n_{ij}) were generated from five different distributions (labelled as scenarios (a) to (e)):

(a) Small equally-sized studies

(b) Medium equally-sized studies

(c) Small to medium uniformly sampled sized studies

(d) A mixture of small and large sized studies

(e) Large uniformly sampled sized studies

The exact sample sizes that these scenarios correspond to can be seen in Table 7.1. For simplicity, and to mimic reality, we set sample sizes to be equal across treatment arms, i.e. $n_{i0} = n_{i1}$, as a difference in sample sizes has previously been shown to have no significant impact on heterogeneity variance estimation (Langan et al. (2016)). We used a uniform distribution when sampling n_{ij} from an interval of specified values, however past simulation studies have also used normal and χ^2 distributions, and evidence has been given to suggest that the distribution used may impact on the performance of heterogeneity variance estimators (Langan (2015)). As a result, we also aim to conduct simulations using these two alternate distributions for study sample size, if time permits, and then compare them to the results from the uniform sampling.

7.2.4 Probability of events

As we are interested in rare-event data, we were primarily focused on simulating meta-analyses where the probability, p_{ij}, of the event of interest occurring is very low. We used a regression model to simulate these values of p_{ij}, aiming for them to range from 0.5 to 0.001. This allowed us to explore a range of scenarios, from the probability of an event being equal to that of no event; to those where the probability of an event is 1 in 1000, which could represent a potential rare adverse effect in a large clinical trial.

For our simulation study, we used the random-effects version of our probability generating regression model, as this allowed us to set the true heterogeneity variance (together with a separate method described later). Although we used a random-effects version of the model, we have also outlined the fixed and mixed-effects models below, in order to demonstrate how these models can be built upon and used to produce the final random-effects model of interest.

The basic, fixed-effects version of the regression model is as follows:

$$\log p_{ij} = \alpha + \beta \times j \qquad , \qquad j = 0, 1 \tag{7.2.1}$$

where p_{ij} is the probability of an event in the j-th treatment arm of study i (with $j = 1$ representing the active treatment arm, and $j = 0$ the control/placebo arm), α is the baseline risk in the control/placebo arm, and β represents the log-relative risk. To determine values of p_{ij}, we shall choose a realistic range of values for α and β that will produce the p_{ij} we desire for our rare-event specification, e.g. $p_{ij} = 0.001$.

Next, we consider the case where α is treated as a random effect, and given by $\alpha_i \sim N(\alpha, \sigma_\alpha^2)$. This gives us the mixed-effects model below:

$$\log p_{ij} = \alpha_i + \beta \times j \qquad , \qquad j = 0, 1 \tag{7.2.2}$$

where α_i represents the study-specific baseline risk, given by $\alpha_i \sim N(\alpha, \sigma_\alpha^2)$. To construct this model, we can set the mean of the study-specific baseline risk, α, to equal the fixed-effect baseline risk from the model in Equation (7.2.1), as this will simplify calculations and provide a reasonable estimate for the mean. We will choose σ_α^2 so as to achieve reasonable values of α_i and p_{ij}, and to reflect our chosen variation in the baseline risk between studies, and we shall discuss the selection of this parameter later.

Finally, we shall consider the case where both α and β are treated as random effects, giving us the desired random-effects model:

$$\log p_{ij} = \alpha_i + \beta_i \times j \qquad , \qquad j = 0, 1 \tag{7.2.3}$$

where α_i is the study-specific baseline risk, with $\alpha_i \sim N(\alpha, \sigma_\alpha^2)$, and β_i is the study-specific log-relative risk, given by $\beta_i \sim N(\beta, \sigma_\beta^2)$. Here, α and σ_α^2 can be chosen to be those used in the model in Equation (7.2.2) above, again for reasons relating to simplicity. In a similar manner to the formulation of α_i in Equation (7.2.2), the mean of the study-specific log-relative risk, β, may be chosen to reflect the value of the fixed-effect log-relative risk given in Equations (7.2.1) and (7.2.2). It is the value of σ_β^2, the variance of the study-specific log-relative risk, that is of particular interest to us, however, as this

represents the true heterogeneity variance (τ^2). This value was chosen to correspond with pre-specified levels of heterogeneity (I^2), as will be described later.

Once we had decided upon the fixed-effect values needed to construct the mean and variances of the parameters in this final random-effects model, we sampled values for α_i and β_i using the normal distributions stated above. Finally, this gave us the required probabilities $p_{i0} = \exp(\alpha_i)$ in the control/placebo group, and $p_{i1} = \exp(\alpha_i + \beta_i)$ in the active treatment group.

We were only interested in working with the model in Equation (7.2.3), as this allowed us to set the true heterogeneity variance, which is essential for our simulation study. However, as mentioned above, the previous two models provide a good set-up for this model, and constructing them first allowed us to determine parameter values that were of assistance to us when designing the more complicated random-effects model.

Choice of model event probabilities

For our research, we wanted to review the performance of a number of heterogeneity estimating procedures under the scenario of rare-event data in a meta-analysis. As such, when sampling the count of events for the meta-analyses in our simulation study, we needed to pay special attention to the probability of events occurring in each arm, as we wanted to concentrate on those that represented a sparse event in a clinical trial.

To decide upon the event probabilities that we would use to sample the count of events in each simulated meta-analysis, we studied the rare-event case studies introduced in Chapter 3. These included a meta-analysis investigating the effect of anti-infective-treated central venous catheters versus standard catheters on catheter-related bloodstream infection (CRBSI) events (Niel-Weise et al. (2007)), and another looking at the effect of antibiotic prophylaxis for caesarean section (Hofmeyr and Smaill (2002)). Using the study-level data, we calculated the average probability of an event

$$p_j = \frac{\sum_{i=1}^{k} X_{ij}/n_{ij}}{k}$$

occurring in the treatment and control arm in each of these two case studies. The average event probability in the control arm (p_0) of the CRBSI dataset was 0.03, and the treatment arm probability (p_1) was 0.01, while the C-section study gave us $p_0 = 0.09$ and $p_1 = 0.04$.

Both of the case studies represent the scenario where an event in the control group is more likely than an event in the treatment group, i.e. $p_0 > p_1$. In the case of rare outcomes, such a scenario could be the result of a treatment effectively reducing the occurrence of some rare event in a clinical trial, which may be the primary function of the treatment or the result of some beneficial side-effect. The alternative scenario,

where the probability of an event in the control group is less than that of the treatment group, i.e. $p_0 < p_1$, is the more common case with rare-event data in clinical trials. This scenario could occur when the event of interest is an adverse reaction in a clinical trial, and the treatment inadvertently increases the risk of some very rare outcome occurring in the participant. We shall simulate both of these probability scenarios, as previous results have suggested that this characteristic may affect the performance of τ^2 estimators.

Using the parameters extracted from the two case studies above, we decided to focus on event probabilities in the range of 0.001 to 0.5, representing a 1 in 1000 to a 1 in 2 chance of event occurrence, respectively. This would allow us to compare the τ^2 estimation results for sparse-event data with those for more common events, which in turn could be compared to results from similar studies that have focused on higher probability (and normally distributed) data. In particular, we chose to simulate three alternate event probability pairings for both scenarios of $p_0 > p_1$ and $p_0 < p_1$:

1. Very rare events: $p_j = 0.001$ and $p_{j*} = 0.005$
2. Rare events: $p_j = 0.01$ and $p_{j*} = 0.05$
3. Common events: $p_j = 0.1$ and $p_{j*} = 0.5$

where $j = 0, 1$ and $j* = \{0, 1\} \setminus \{j\}$. We believed that pairings 1 and 2 adequately represented the majority of medical rare-event scenarios, both in the form of clinical trials investigating the side-effects of medications and observational studies looking at the occurrence of rare diseases. For example, there is a 1 in 2500 chance of Caucasians developing the chronic, autosomal recessive disorder cystic fibrosis (Scotet et al. (2012)), while the risk of developing serious complications (e.g. a blood clot) as a result of taking the contraceptive pill is 1 in 10,000 (Mohanna and Chambers (2008)). In this simulation study, we did not investigate risks lower than 1 in 1000, which are defined as minimal to negligible (Mohanna and Chambers (2008)), as the performance of τ^2 estimators is unlikely to differ considerably below this level of event occurrence.

For completeness, we also chose to simulate the scenario where the event probabilities are equal across treatment and control arms, which corresponds to the log-risk ratio being zero. For this case, we looked only at the rare-event scenario, using $p_j = 0.01$ for both arms. In total we simulated seven probability pairings, and these are listed below in Table 7.2.

TABLE 7.2: Pairings of mean baseline risk and mean log-relative risk to produce model event probabilities for treatment and control arms.

Probability scenario	**Model event probabilities**		**Mean baseline risk**	**Mean log-relative risk**
	p_0	p_1	α	β
$p_0 < p_1$	0.001	0.005	-6.9	1.6
	0.01	0.05	-4.6	1.6
	0.1	0.5	-2.3	1.6
$p_0 = p_1$	0.01	0.01	-4.6	0
$p_0 > p_1$	0.005	0.001	-5.3	-1.6
	0.05	0.01	-3.0	-1.6
	0.5	0.1	-0.7	-1.6

Sampling the event probabilities

The arm-specific event probabilities (p_0, p_1) discussed above are only the model or ideal probabilities for our simulation study, as we sampled these values to mirror real-life variation between studies. As mentioned previously, we did this using the random-effects model in Equation (7.2.3), after sampling values for the model parameters - the study-specific baseline risk $\alpha_i \sim N(\alpha, \sigma_\alpha^2)$ and study-specific log-relative risk $\beta_i \sim N(\beta, \sigma_\beta^2)$.

The means of these Normal distributions correspond to the mean baseline risk (α) and mean log-risk ratio (β). These values were chosen to produce our model pairs of probabilities when inserted into Equation (7.2.1) (with some minimal degree of error). For this fixed-effects model, the chosen model probability p_0 solely dictates the necessary value of α. With this knowledge of α, and the required probability p_1 from the respective pairing, the fixed value of β can also then be determined. The values of α and β required to produce each of our chosen event probability pairings can be seen in Table 7.2.

The values of σ_α^2 and σ_β^2 correspond to the variance of the baseline risk and between-study variance (as $\sigma_\beta^2 = \tau^2$), and these also need to be chosen. We shall discuss how we decided on ranges for these parameters in our simulation study later in the chapter. Once we had all of the above parameter values (α, β, σ_α^2 and σ_β^2) for each of our scenarios of interest, we could then sample the required event probabilities for the treatment and control groups in the studies within our simulated meta-analyses.

7.2.5 True heterogeneity variance

As heterogeneity variance (τ^2) estimators have been shown to vary in performance depending on the true value of τ^2, we chose a range of τ^2 that represented the full spectrum of variability present in rare-event meta-analyses, using our case studies from Chapter 3 as a guideline. Having applied the pre-existing estimators and our proposed methods to these datasets, we observed that the estimates of τ^2 ranged from 0 to 4.2. However, this higher value was generated for the case when $k = 3$ and was not backed up by any other estimators, so we decided to ignore this value and look at the next highest estimate for any case study, which was $\hat{\tau}^2 = 1$. As a result, we chose to use values of τ^2 ranging from 0 to 1, increasing in increments of 0.2, as we believed this to represent a range of real-life rare-event scenarios. We also looked at previously published simulation studies and meta-analyses of rare-event data, and found that our chosen range was appropriate, as it was representative of all possible rare-event scenarios, as well as a wide selection of higher-probability (more common) event meta-analyses.

Levels of heterogeneity

We conducted a preliminary simulation study to ensure that our chosen range for τ^2 represented all potential levels of heterogeneity, defined by I^2 in Section 1.7.3, as this is more interpretable in empirical settings. In particular, I^2 can vary between 0%, corresponding to within-study variance accounting for all of the observed variability (homogeneity), and near 100%, where heterogeneity forms the major contributor of variability. In this primary investigation, we calculated the average I^2 for each chosen value of τ^2 for defined clusters of simulation scenarios sharing parameters that affect the computation of I^2. We calculated I^2 for each simulated meta-analysis using the following formula:

$$I^2 = \frac{\tau^2}{\tau^2 + \overline{\sigma^2}} \times 100\%$$

where the true typical study variance $\overline{\sigma^2}$ (a form of average of the study-specific variances, and also a type of harmonic mean) is calculated as follows:

$$\overline{\sigma^2} = \frac{(k-1)\sum_{i=1}^{k} 1/\hat{\sigma}_i^2}{(\sum_{i=1}^{k} 1/\hat{\sigma}_i^2)^2 - \sum_{i=1}^{k}(1/\hat{\sigma}_i^2)^2}$$

where $\hat{\sigma}_i^2 = \frac{1}{X_{i1}} - \frac{1}{n_{i1}} + \frac{1}{X_{i0}} - \frac{1}{n_{i0}}$ is the within-study variance for the log-risk ratio, and X_{ij} and n_{ij} are the event count and sample size for treatment arm j of study i, with $j = 1$ referring to the active treatment arm and $j = 0$ otherwise.

Our aim was to determine whether our set values for τ^2 approximated a wide range of I^2 in the simulated scenario clusters, where clusters were defined by the sample size

and probability of event only. It has previously been shown that I^2 is not sensitive to changes in the number of studies k, so we fixed this and set $k = 3$ when running this preliminary simulation, in order to shorten computation time and still provide well-grounded derivations of the corresponding I^2. As I^2 values are likely to vary to some extent as a result of sampling of the aforementioned parameters and sampling error within the simulated meta-analyses, we defined the resulting value of I^2 as the average produced over 1000 replications.

We found that our chosen selection of τ^2 produced a range of I^2 values from 0% to 95% as required, certifying our choice for the τ^2 parameter. Threshold values can be defined for I^2 to help interpret the results of the simulation study: 15% and 30% represent low inconsistency; 45% and 60% represent moderate inconsistency; and 75%, 90% and 95% represent considerable inconsistency. These threshold values roughly correspond to the guidelines in the Cochrane handbook (Higgins and Green (2011)), but are modified slightly to represent the simulated I^2 values produced using our chosen τ^2.

It should be noted that the value of $\overline{\sigma^2}$, and subsequently I^2, cannot be calculated when $X_{i1} = X_{i0} = n_{i1} = n_{i0}$, i.e. when the number of events is equal to the sample size of each arm, and the sample sizes in turn are equal, for at least one study in the meta-analysis. This is because the calculated $\log RR$ and its associated standard error are both zero in this case, which leads to $\overline{\sigma^2}$ and thus I^2 being undefined. However, this should not affect our simulation study since we are focusing on rare-event data rather than very common outcomes with high probability of event.

7.2.6 Variance of the baseline risk

Another parameter that is essential in the simulation of event counts in meta-analyses is the variance of the baseline risk (the risk in the control or placebo group), which is denoted by σ_α^2. For our simulation study, we based the choice of these parameter values on those characteristic of empirical rare-event meta-analyses, by looking at our case studies in Chapter 3. For each of these real-life datasets, we determined the baseline risk by calculating the probability of an event in the control arm ($= X_{i0}/n_{i0}$) for each study i. We then sampled from a Normal distribution with mean -4.6 (to reflect our median event probability of 0.01) and various values of the variance, and took the exponential of these values as the associated baseline probabilities. Repeating this step 1000 times, with 100 studies (our largest simulated size of meta-analysis), allowed us to determine which value of the variance would result in an average standard deviation of event probabilities similar to that observed in the relevant case study.

By conducting the above simulation, we found that the smallest standard deviation of baseline event probabilities seen in the case studies was 0.003, while the largest was 0.16. To generate values similar to these, using the process described above, we determined

that σ_α^2 values of 0.1 and 3 were required, respectively. Thus, we decided to use these two values for σ_α^2 in our simulation study, as they reflect both small and considerable variability in baseline risk. Although this larger variability could reduce the desired frequency of rare events simulated in the study control arms, it is still a plausible scenario as study populations may differ in terms of baseline characteristics such as gender, ethnicity and medical history.

7.2.7 Count of events

Having completed the previous steps in this chapter, we then had all the parameters necessary to generate the count of events in the treatment and control arms for our simulated meta-analyses. We initially sampled the count of events, X_{ij}, in the j-th treatment arm of study i, using the Binomial distribution:

$$X_{ij} \sim Binomial(n_{ij}, p_{ij})$$

where p_{ij} is the event probability generated according to the model in Equation (7.2.3) in this case.

However, to provide an alternate method for sampling the count of events in our scenarios, we also sampled them using the Poisson distribution. This provides an approximation to the Binomial distribution when n_{ij} is large and p_{ij} is small, and so complements the rare-event property of our simulated data:

$$X_{ij} \sim Poisson(n_{ij} \times p_{ij})$$

Although we initially used the sample sizes, n_{ij}, for simplicity here, we later replaced them with the more accurate trial arm-specific person times, P_{ij}. The methods that we have outlined here to sample the count of events represent only a few of many options. However, we believe that the majority of realistic scenarios, fitting with our rare-event focus, will correspond well with the above sampling approaches.

7.2.8 Study-specific effect measure

To determine the effect measure, θ_i, for study i (in our case the log-relative risk, $\log RR_i$), we re-write the model equations given in Section 7.2.4 as follows. For the model in Equation (7.2.1) we have:

$$\log p_{i1} - \log p_{i0} = \alpha + \beta - \alpha = \beta$$

giving us the fixed-effect measure:

$$\log RR = \log \frac{p_{i1}}{p_{i0}} = \beta = \theta$$

Similarly, for the model in Equation (7.2.2) we obtain:

$$\log p_{i1} - \log p_{i0} = \alpha_i + \beta - \alpha_i = \beta$$

again providing us with the fixed-effect measure:

$$\log RR = \log \frac{p_{i1}}{p_{i0}} = \beta = \theta$$

Finally, for Equation (7.2.3) (our random-effects model of interest), we have:

$$\log p_{i1} - \log p_{i0} = \alpha_i + \beta_i - \alpha_i = \beta_i$$

This provides us with the required study-specific effect measure:

$$\log RR_i = \log \frac{p_{i1}}{p_{i0}} = \beta_i = \theta_i$$

It follows from our previous assumptions that $\theta_i \sim N(\theta, \sigma_\beta^2)$, with the mean $\theta = \beta$ and variance $\sigma_\beta^2 = \tau^2$ (our true heterogeneity variance determined in Section 7.2.5). Since this full random-effects model is the model that we used to sample probability values in our simulation study, as described in Section 7.2.4, we were also able to determine the study-specific log-relative risk in this same manner. At this point in our simulation study, we had generated all of the meta-analysis parameters necessary to apply the τ^2 estimators and novel approaches of interest.

7.2.9 Meta-analyses avoided during simulation

There were a number of types of meta-analysis that, if simulated, could not be used with some of the τ^2 estimators that we were investigating. The most important, and destructive, of these meta-analyses were those for which the methods that we are proposing could not be applied. As it was these novel methods that we are most interested in assessing the performance of, we avoided generating meta-analyses that met such undesirable properties in our simulation study. This ensured that none of our simulated scenarios contained high proportions of meta-analyses that could not be used with our

novel approaches. If meta-analyses meeting these conditions were produced, then the count of events (X_{i0}, X_{i1}) were resampled until the conditions were no longer met, and the meta-analysis was thus deemed as adequate for inclusion in our study.

This resampling of meta-analyses was deemed a fair and reasonable approach to deal with the situation of inapplicability with our proposed methods, and allowed us to maximise the efficiency of our simulations. The conditions that we defined for resampling corresponded to very few scenarios, and consequently few meta-analyses were resampled during our simulation study, thereby having little impact on the study itself. This approach is common practice in other simulation studies in this field (Bakbergenuly and Kulinskaya (2018)), where the authors have chosen to ignore scenarios for which the method(s) of interest cannot be applied.

For each of our generalised linear mixed model (GLMM) approaches proposed in Chapter 4, there are several types of meta-analysis that cannot be used as a result of model structure, and these are listed in Section 4.7. We avoided simulating meta-analyses that could not be used with both of our GLMM approaches - the Poisson mixed regression model (PMRM) and conditional logistic mixed regression model (CLMRM) - but allowed the simulation of those that could be used with one of these methods. This is because we didn't want to overly restrict the types of meta-analyses being simulated, as only producing certain data types may introduce selection bias into our results. There are also a number of scenarios for which our proposed mixture model approach from Chapter 6 cannot be applied, and these are listed in the method guidelines in Section 6.5.3. As before, we avoided simulating these incompatible meta-analysis scenarios.

We also chose to avoid the simulation of meta-analyses consisting entirely of double-zero trials, as these meta-analyses contain very weak information in terms of event counts, and so are unlikely to be published as a result. As such, they do not represent realistic data from publications, and so were deemed unnecessary for inclusion in our simulation study. Such meta-analyses would have been omitted from simulation as a result of incompatibility with both the GLMM methods and the mixture model approach regardless of this choice, but if this had not been the case we still would have avoided this scenario.

If we determined any further scenarios that were incompatible with the τ^2 estimators during the course of the simulation study, particularly those that could not be used with our novel approaches, we also dropped them from our simulation study. Such scenarios will be listed in the results chapter.

Problematic meta-analyses not omitted

It is worth mentioning here a meta-analysis scenario that could not be used with the majority of τ^2 estimators, but which we decided to keep in our simulation study. This scenario relates to meta-analyses containing at least one study where the number of

events in each arm was equal to the sample size, and as a result of our simulation design, were also equal across treatment arms. This represents the scenario where the event of interest is extremely common, and has occurred for all subjects in both the treatment and control arms. If this is the case in a given simulated study, then the log risk-ratio for that study will be zero, as will the associated standard error. If this is the case for at least one study in the meta-analysis, then the following pre-existing τ^2 estimators cannot be calculated and thus are undefined:

- DerSimonian-Laird
- Positive DerSimonian-Laird
- DerSimonian-Laird bootstrap
- Paule-Mandel
- Hartung-Makambi
- Hunter-Schmidt
- Maximum likelihood
- Restricted maximum likelihood
- Approximate restricted maximum likelihood
- Rukhin Bayes with simple prior estimate
- Rukhin Bayes with zero prior estimate ($\hat{\tau}_0^2 = 0$)
- Bayes Modal

As a result, the only methods from Chapter 2 that can be applied to this 'all-events study' meta-analysis scenario are the Hedges-Olkin (HO), Sidik-Jonkman and Sidik-Jonkman with HO initial estimate. In addition to these, both of our GLMM methods could be applied to this scenario. Despite the majority of pre-existing estimators being unsuitable for this scenario, we included it in our simulation study because our novel approaches were appropriate, and these were the methods of primary interest to us.

7.3 Continuity corrections

As discussed in Section 1.9.1, a variety of continuity corrections have been proposed to counteract the issues associated with single-zero and double-zero trials (Jewell and Holford (2005)). We shall apply a selection of these continuity correction approaches in our simulation study in order to determine which performs best in our area of interest (rare-event log-risk ratio meta-analyses). Specifically, we shall apply the following continuity corrections:

- Constant continuity correction using $c = 0.5$ (standard method)
- Reciprocal of the opposing trial arm's sample size

Both of these approaches are discussed in detail in Section 1.9.1. We also discussed an empirical continuity correction in this section, however we shall not include this in our simulation study as is not recommended for use with random-effects models (Jewell and Holford (2005)).

7.3.1 Continuity corrections for all-event studies

Continuity corrections are not only required for the case of single-zero and double-zero studies, but also for studies where every participant in both treatment arms has had the event of interest (the scenario discussed in Section 7.2.9). In these cases, where $X_{i1} = n_{i1} = X_{i0} = n_{i0}$ (as we are setting the sample sizes to be equal in our simulation study, i.e. $n_{i1} = n_{i0}$) for at least one study i in the meta-analysis, we applied the following continuity correction:

- Add a constant c to each event count, and add $2c$ to the sample sizes

To complement our zero-combatting constant continuity correction above, we set $c = 0.5$ in this case. We chose this correction as it corresponded with those proposed previously for similar issues. While this particular 'all-events' scenario is only likely to occur in cases with high event probabilities and small sample sizes, and so not related to our primary focus on rare events, it is still a viable possibility in meta-analyses and as such we did take it into account and adjusted as necessary.

7.3.2 Calculation of log-risk ratio and associated standard error

Finally, with the appropriate continuity corrections applied, we were then able to calculate the log-risk ratio and its associated standard error for each of our simulated studies. It should be noted that the choice of continuity correction will only affect the pre-existing τ^2 estimators in Chapter 2, as the study-specific estimates of $\log RR_i$ and $s.e.(\log RR_i)$ are involved in the calculation of these τ^2 estimates. In contrast, the continuity corrections will have not effect on our GLMM methods, as these do not use the study-specific effect size estimates, requiring only the uncorrected original event counts and sample sizes.

7.4 Heterogeneity variance estimates

7.4.1 Pre-existing estimators

Once we had simulated our meta-analyses as described in Section 7.2, and applied any necessary continuity corrections in order to calculate study-specific log-risk ratio, we had collected all of the information necessary to apply the pre-existing τ^2 estimators from Chapter 2. For each of our simulated meta-analyses, estimates of τ^2 were calculated using the following approaches:

- DerSimonian-Laird
- Positive DerSimonian-Laird
- DerSimonian-Laird bootstrap
- Hedges-Olkin (Cochran's ANOVA)
- Paule-Mandel
- Hartung-Makambi
- Hunter-Schmidt
- Sidik-Jonkman
- Sidik-Jonkman with Hedges-Olkin initial estimate
- Maximum likelihood
- Restricted maximum likelihood
- Approximate restricted maximum likelihood
- Rukhin Bayes with simple prior estimate
- Rukhin Bayes with zero prior estimate ($\hat{\tau}_0^2 = 0$)
- Bayes Modal

Details of how to estimate τ^2 using each of the methods listed above are given in Chapter 2. As can be seen from the above list, not all of the estimators discussed in Chapter 2 and appropriate for our area of interest were included in our simulation study. For example, we decided not to include the two-step DerSimonian-Laird and Hedges-Olkin estimators discussed in Section 2.2.5, as we believed these would have little improvement in performance over their respective one-step alternatives, particularly in the challenging case of rare-event data. In addition, we also chose not to include the fully Bayesian

approach using MCMC methods for τ^2 estimation in our simulation study. This is because we wanted to concentrate on methods that currently dominate the practice of meta-analysis, and so our focus here is not on Bayesian methodology. We believe, however, that Bayesian methods will gain increasing importance in this area, and so we will need to consider them in future methodological developments.

As mentioned previously, these estimators were applied to our simulated meta-analyses using code that we designed in R, which can be seen in Appendix D.1. For these pre-existing estimators, we based our code on that given by Langan (2015). To check the validity of our code, we applied it to empirical datasets that had previously been used with some of the estimators listed above in published meta-analyses. We then compared the τ^2 and effect size estimates that our code produced to those given in the associated publications, in order to confirm that our code was working correctly.

7.4.2 Proposed methods

The main aim of our simulation study was to determine how our generalised linear mixed model (GLMM) approaches (proposed in Chapter 4) performed in terms of estimating τ^2, compared to pre-existing normal-based approaches. The GLMMs that we will apply to our simulated meta-analyses, in order to extract the model parameters corresponding to τ^2 and overall effect size estimates, are:

- Poisson mixed regression model
- Conditional logistic mixed regression model

We applied these listed models using the *glmer* command of the *lme4* package in R (Bates et al. (2015)). We chose to fit the models using command-based options that would maximise the application of these methods, as described in Section 4.6. However, if any further adjustments to these options were found to be beneficial during the course of the simulation study, then these changes were made and will listed with the results.

We also applied our novel conditional-based and mixture modelling approaches, proposed in Chapters 5 and 6 respectively, extracting the relevant estimates as described in the corresponding chapters. As our proposed conditional-based approach estimates the probability-associated variance τ_p^2 rather than τ^2, we cannot compare this directly with the other estimators being considered in the simulation study. However, we shall compare its performance within the four variations of estimating equation that we proposed, in order to determine which scenarios these perform well in. For the mixture modelling approach, there are arguments in our corresponding R code that allow the user to select the number of iterations and cut-off value for the expectation-maximisation algorithm, however we used those options discussed in Chapter 6.

7.5 Summary effect-size estimates

We also generated estimates of the overall effect size (θ), the log-risk ratio, for every estimate of τ^2 in each of our simulated meta-analysis. For the normal-based τ^2 estimators from Chapter 2, we estimated θ for each meta-analysis using the inverse-variance approach described in Section 1.5. We also used this approach with our τ_p^2 estimates from the novel conditional-based method. For our novel GLMM-based methods, the estimates of θ were extracted from the relevant parameters of the associated model output, as described in Chapter 4. Finally, for our mixture model approach, these θ estimates were generated according to the protocol given in Chapter 6.

7.5.1 Confidence intervals

We next calculated the confidence intervals for the overall effect size estimates discussed above. We used the following methods to generate these confidence intervals:

- Wald-type method
- t-distribution method
- Hartung-Knapp-Sidik-Jonkman method

Each of these methods is described in detail in Section 1.6.

7.6 Performance measures

In order to determine the performance of the methods listed in Section 7.4, we used a variety of universally comparable and meaningful performance measures. These measures were chosen to allow us to compare the accuracy of our τ^2 estimators easily and with reliability. In particular, estimators were compared in terms of the following performance measures:

- Median and mean absolute bias in estimate of τ^2
- Median and mean squared error of estimate of τ^2
- Proportion of zero estimates of τ^2
- Mean absolute bias in estimate of the mean treatment effect
- Mean squared error of estimate of the mean treatment effect

- Coverage of confidence interval for treatment effect
- Power, which we defined by the percentage of meta-analyses meeting the requirement $\frac{CI_{upper,\theta}-CI_{lower,\theta}}{2} < c$, where we have chosen the constant $c = 2$ as this appeared appropriate for our needs and was the constant used in similar simulation studies (Langan (2015))
- Mean and variance of the error.

Definitions for each of these measures are given in Appendix D.2. A good estimator should have small bias (if not unbiased) and a low mean squared error. If the estimator can produce zero estimates of τ^2, and so is not strictly positive by construction, then it should produce a high proportion of zero estimates when the true value of $\tau^2 = 0$, and a low proportion of zero estimates in all other scenarios.

7.7 Analysis

In summary, analysis of the results was undertaken after the following steps had been performed:

1. A meta-analysis dataset is generated for a sampled set of specified parameter values
2. Step 1 is repeated 1000 times, to produce 1000 meta-analyses for a given scenario of interest
3. Heterogeneity variance estimators are applied to the 1000 meta-analyses, and the τ^2 estimates are saved
4. Summary effect measures (the log-risk ratio) are generated for each τ^2 estimate in each of the 1000 meta-analyses
5. Performance measures are calculated for the 1000 τ^2 and θ estimates produced from each of the investigated methods
6. Steps 1-5 are repeated for all combinations of parameter values and distributions defining meta-analysis scenarios.

A flow-chart detailing this simulation study protocol is displayed in Figure 7.1. All of the steps outlined in the above protocol were carried out in R, and the estimates and performance measures produced were stored for analysis. All τ^2 estimation methods were applied, and thus also compared, using the same simulated meta-analysis datasets in order to reduce sampling error.

Of the pre-existing estimators listed in Section 7.4, the three maximum likelihood-based methods are iterative in design, and so may fail to converge to a solution when applied to some meta-analyses. Such failure to converge would likely be a result of the chosen iteration algorithm, rather than reflect poor performance of the estimator itself. When applying these iterative estimators in our simulation study, we used the default iteration algorithm of the *metafor* package in R (Fisher's scoring method with Hedges-Olkin estimate as the initial value), as this seemed an appropriate choice (Viechtbauer (2010)). Any simulated meta-analyses that caused such failures in convergence were not replaced, however the instances were recorded so that the characteristics of the associated datasets could be examined for patterns and similarities.

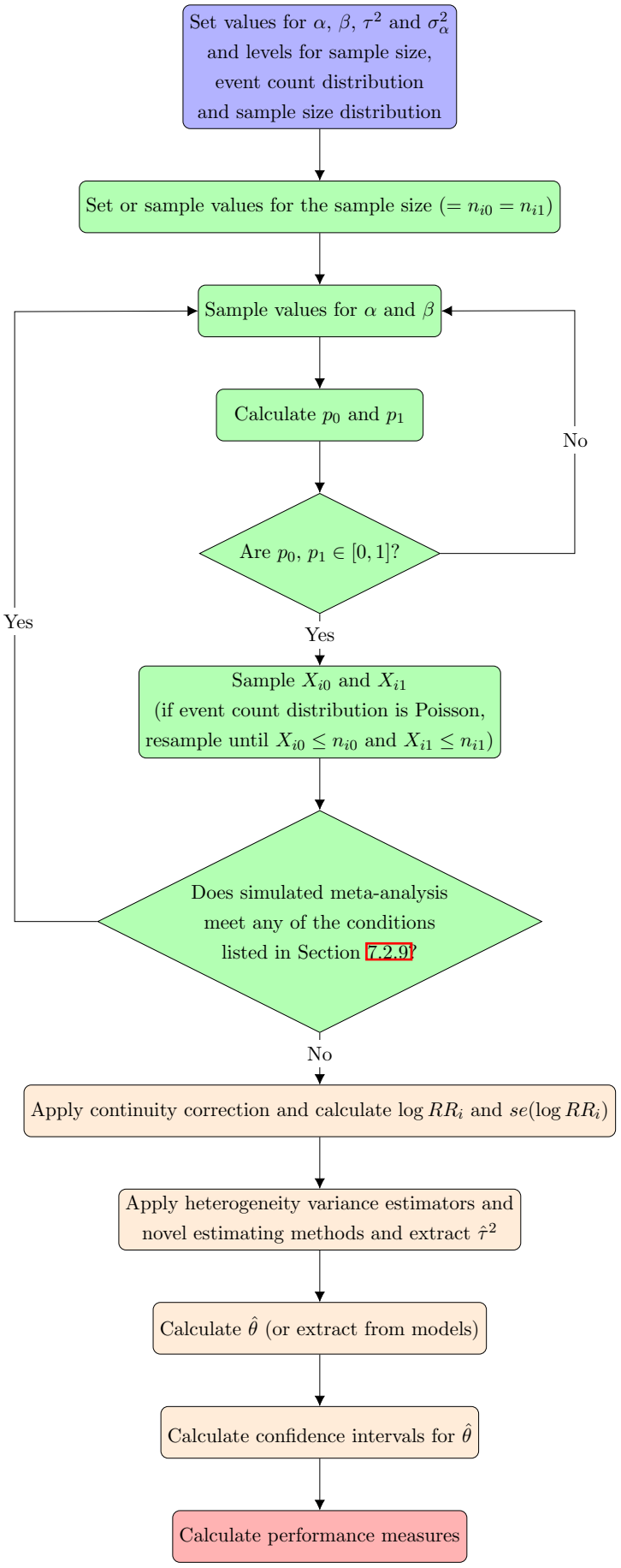


FIGURE 7.1: Outline of simulation study protocol.

Similarly, our proposed GLMMs may fail to converge to a solution in certain scenarios, particularly those with very few or no events in the study arms of the meta-analysis. As a result of this, the models may fail to produce estimates of τ^2 and θ in these cases. As with the iterative maximum likelihood estimators mentioned above, we did not replace the meta-analyses for which the GLMM-based methods could not produce estimates, but instead recorded their instances in order to inspect these meta-analyses for similarities in parameters and other defining features. By recording instances of failures, this also allowed us to compare the efficiency of these estimators.

7.7.1 Primary analysis

The τ^2 estimation methods under investigation were compared in terms of the performance measures listed in Section 7.6, and then were presented graphically to provide a visual representation of the results. For each performance measure and meta-analysis scenario of interest, graphs were produced displaying the value of the given performance measure against the number of studies in the meta-analysis, allowing easy comparisons to be made between the estimators. For each scenario, the average $\widehat{I^2}$ values (that the true τ^2 had approximated, as described in Section 7.2.5) were presented along with the graphs to convey the level of heterogeneity associated with the given scenario. In cases where $\tau^2 \neq 0$, 90% confidence intervals for $\widehat{I^2}$ were also presented, allowing us to show how accurate τ^2 was in approximating $\widehat{I^2}$, which may be of interest when all estimators perform poorly for a given level of heterogeneity.

As mentioned previously, the maximum likelihood and GLMM-based approaches may fail to converge in some cases. To account for this in the presentation of results, the number and percentage of failures were tabulated, and these results were then taken into account when making recommendations regarding the best methods to use in specific sparse-event scenarios. If any of the other τ^2 estimators failed to produce an estimate, then such failures were also recorded.

7.7.2 Secondary analysis

If all of our τ^2 estimators performed poorly in a given scenario of interest, then this setting was investigated further, in order to determine what characteristics of this situation may have been associated with the inability to accurately estimate τ^2. By further investigating the given scenario, it could also be determined which of these poorly performing methods had the most potential, as an alternate estimator of the same type (e.g. method of moments, GLMMs) could be sought for use with this context in the future.

7.8 Recommendations

Once the analysis of results had been conducted, recommendations were made regarding the best choice of τ^2 estimator in a rare-event meta-analysis for each specific scenario under investigation. These recommendations were made based on a compromise between the outcomes of all performance measures of interest and the ease of applying the associated estimating procedure, e.g. whether it is iterative or not. If the difference in performance measures was negligible, then simple estimators were recommended over more complex iterative procedures.

7.9 Overview

Our simulation study was conducted using the protocol detailed in this chapter. Alterations from this protocol may have been made if they were deemed necessary during the course of the study, and if so shall be recorded with the results. We simulated 1000 meta-analyses for each scenario under investigation as this was deemed adequate and corresponded with previous simulation studies in similar fields. We also believed that the number and range of our investigated scenarios was adequate and covered all major classes of potential real-life meta-analyses. As a result, we believe that additional simulations or scenarios would have increased computation time and not significantly added to or altered the overall results.

Chapter 8

Main simulation study results

8.1 Introduction

The principal aim of our simulation study was to determine the performance of our novel methods to estimate the heterogeneity variance (τ^2) compared to existing approaches, in the case of rare-event meta-analyses. We did this by simulating meta-analyses for a wide range of possible real-life meta-analysis scenarios, and then computing performance measures such as bias and mean squared error for the corresponding estimates. All stages of our simulation study were conducted using the statistical software package R, and the code for this was validated before conducting the study to ensure correctness. This ensured that we produced a set of reliable and trustworthy results from which valid recommendations could then be made.

In our simulation study, we looked at over a thousand meta-analysis scenarios governed by characteristics such as the number of studies, sample size and true heterogeneity variance, and looked at both rare and common events. Additional scenarios were generated to investigate the effect of altering aspects such as the type of continuity correction used, and the sampling of sample size and event counts in the meta-analysis simulation. As we generated such a large number of scenarios, we shall group them according to certain characteristics when presenting their results, and report only the most informative of these results.

In this chapter, we shall summarise how the simulation study was conducted - paying special attention to any changes that were made to the original simulation protocol outlined in Chapter 7. We shall detail each of these changes along with explanations as to why they were deemed necessary. We will then summarise the characteristics of the simulated meta-analyses themselves in terms of the proportion of single-zero and double-zero trials present, which are of particular importance in the case of rare-event data. In Chapters 4 to 6 we listed scenarios that, by definition, our novel approaches could

not be applied to. Here we shall list any further scenarios that we found our proposed methods, particularly the generalised liner mixed models (GLMMs), could not be used with during the course of the simulation study, or which failed convergence in a high percentage of cases. Similarly, we shall present the proportion of meta-analyses that did not result in convergence for those estimators that were iterative or model-based, and characterise these failures in terms of scenario-defining parameters.

We will then focus on outlining the main results from our simulation study in terms of the performance of methods in estimating τ^2 and the summary effect size in terms of measures such as bias and mean squared error (MSE). We shall also look at the proportion of zero τ^2 estimates generated for varying τ^2, and the coverage of confidence intervals for the effect size when the estimators were paired with various interval types. We shall summarise these results for each of the major event probability scenarios we are looking at, paying particular attention to rare and very rare events - our area of interest. We shall present the outcomes of these performance measures graphically, plotting only those methods and scenarios that we deem are of interest, in order to prevent over-information. We will also score the estimators in terms of their MSE for grouped scenarios, and provide a general summary of their performance in estimating τ^2. Finally, to ensure credibility of our results, we shall compare our results regarding the pre-existing estimators to those achieved in previous simulation studies.

8.2 Amendments made to simulation study protocol

In Chapter 7, we gave a detailed protocol for our simulation study. Although we tried to follow this protocol as closely as possible, there were a number of aspects that we found needed altering during the course of the study. Such modifications included improving the simulation study itself and omitting certain scenarios that were found to be challenging to apply in terms of convergence, particularly with our novel methods.

8.2.1 Rounding of zero estimates

Some of the estimators included in our simulation study produced very exact τ^2 estimates that never reached zero, but did generate very small values that were close. An example of this is our proposed Poisson mixed regression model (PMRM) approach, which despite not producing zero estimates, resulted in many estimates less than 1×10^{-5}. To account for this and ensure that any methods producing values of $\hat{\tau}^2$ very close to zero had these estimates classed as zero, we added code that rounded any τ^2 estimate $< 1 \times 10^{-5}$ to zero. This ensured that such estimates were correctly classed as being zero, which could have a profound impact on the associated estimator's performance in estimating τ^2, particularly in cases of homogeneity.

8.2.2 Scenarios excluded from simulation study

During the course of the simulation study, it became evident that a number of the iterative and model-based estimators had difficulty converging in certain scenarios. As a result of this, we decided to drop certain parameters from our simulations, especially those that were not compatible with our novel approaches, in order to focus only on cases where the majority of methods could be applied. In particular, from those scenarios originally defined in Table 7.1, we excluded the following parameters:

- Number of studies (k) = 2,3: These scenarios were dropped completely from the simulation study (for all settings) because they were generally inapplicable to our novel GLMM-based approaches, even when sample sizes were large.

We also added to the parameters given in Table 7.1, as we decided to sample any non-constant study sample sizes from uniform, normal and chi-squared distributions rather than only uniform. After making the above described amendments, we simulated a total of 2520 scenarios.

8.2.3 Application of generalised linear mixed models

Two of our novel approaches involved the use of GLMMs - the Poisson mixed regression model (PMRM) and conditional logistic mixed regression model (CLMRM) methods introduced in Chapter 4. To apply these models to our simulated meta-analyses we used the *glmer* command in the R package *lme4* (version 1.1-19). We discussed our decision to use this command in Section 4.6, however the specific package version is also very important here. This is because while there were a number of scenarios for which these models could not be applied (which we shall list later in this chapter), modifications may be made to later versions that make the package appropriate for these cases. Such an increase in applicable scenarios was observed when this particular version became available while conducting our simulation study, as the prior version of the package could be applied to fewer scenarios.

We chose to make some changes to the model application outlined in Section 4.6 as a result of observations made during the simulation study. In particular, when applying the PMRM model we used the following control parameters in the *glmer* command options:

- Nelder-Mead optimising function
- The maximum number of function evaluations the optimising function could make was set to 100,000

- For the residual sum-of-squares step, the tolerance level for convergence was set as 1×10^{-3}
- When determining the adaptive Gauss-Hermite approximation of the log-likelihood, zero points per axis were used - this involves a quicker but less precise version of parameter estimation by optimising the fixed and random effect coefficients in the iteratively re-weighted penalised least-squares step.

When applying the CLMRM method, we used all of the default options given with the *glmer* command, which can be seen in Bates et al. (2015) and includes:

- When determining the adaptive Gauss-Hermite approximation of the log-likelihood, one point per axis was used - this corresponds to the Laplace approximation.

The final R code used to apply these models is given in Appendix D.1. For both models, the command settings were chosen as those resulting in the greatest applicability in terms of range of potential scenarios, after comparison among varying option combinations.

8.3 Summary of simulation study

8.3.1 Characteristics of simulated meta-analyses

In order to gain an overview of our simulated data, we recorded certain characteristics of the meta-analyses generated during our simulation study. In particular, we focused on the number and proportion of single-zero and double-zero studies present in each of the simulated meta-analyses. This allowed us to then summarise this information for each of the study scenarios, so that they can be compared to the observed properties of empirical rare-event studies, in order to measure how well our simulated data mirrors that from real-life cases. Tables 8.1 and 8.2 display the total percentage of single and double-zero trials present in all simulated meta-analyses for scenarios grouped by study sample size and event probability.

TABLE 8.1: Summary of total percentage of single-zero studies produced by scenario groupings.

Probability scenario		**Study sample sizes**				
		Small	**Small-to-medium**	**Medium**	**Small and large**	**Large**
$p_0 < p_1$	Very rare	11.82	37.97	46.72	17.89	25.21
	Rare	39.41	33.68	19.48	21.20	2.97
	Common	39.88	6.91	3.06	20.07	0.25
$p_0 = p_1$	Rare	20.71	33.65	24.45	12.21	4.05
$p_0 > p_1$	Very rare	10.49	36.83	46.21	18.15	26.64
	Rare	38.24	34.09	21.63	20.95	3.76
	Common	40.03	8.33	3.85	20.22	0.39

Table 8.1 shows that for meta-analyses with large study sample sizes, as the rarity of events increases, the percentage of studies containing a single zero event also increases, as expected. However, for small sample sizes, the reverse is true, with the proportion of single-zero studies dropping dramatically as events become rarer. This is because small studies (which in our case have only 20 participants) are more likely to consist of double-zero trials when events are very rare, thereby reducing the proportion of single-zero trials present. The overall results appear to be very similar regardless of whether the event probability is higher in the treatment arm ($p_0 < p_1$) or control arm ($p_0 > p_1$), which is to be expected as this feature would only affect the arm that the zero count occurs in, not the occurrence of the zero count itself.

TABLE 8.2: Summary of total percentage of double-zero studies produced by scenario groupings.

Probability scenario		**Study sample sizes**				
		Small	**Small-to-medium**	**Medium**	**Small and large**	**Large**
$p_0 < p_1$	Very rare	86.91	50.87	33.76	47.00	5.57
	Rare	51.49	9.60	4.10	26.13	0.44
	Common	11.49	1.28	0.39	5.75	0.02
$p_0 = p_1$	Rare	73.51	22.84	10.41	37.68	1.42
$p_0 > p_1$	Very rare	88.40	51.38	32.82	47.08	4.87
	Rare	51.86	8.42	3.57	26.13	0.35
	Common	9.79	1.05	0.29	4.90	0.01

In contrast, Table 8.2 shows that the proportion of studies containing zero events in both arms increases with event rarity in all sample size scenarios. This result is to be expected, and confirms our above theory that for small sample sizes and very rare scenarios, double-zero trials would be far more prominent than single-zero alternatives. In fact, by referring back to the previous table, it can be seen that for small sample sizes and very rare events, 98% of studies in the simulated meta-analyses contained at least one zero count. As before, we can see that the results do not differ significantly between cases where $p_0 < p_1$ and $p_0 > p_1$, backing up the consistency of our simulated data.

8.3.2 Running time of simulation study

We ran our simulation study on the IRIDIS super computer at the University of Southampton in order to reduce running time and increase efficiency. Scenarios were run in parallel in batches, with each scenario taking between 1 and 4 hours to complete, depending on the complexity of the scenario itself and the sampling required in the parameter-based simulations. As such, the average running time was 2.5 hours per scenario, giving a total running time of approximately 6300 hours (262.5 days) for all of our 2520 scenarios.

8.4 Efficiency of heterogeneity variance estimators

Not all of the heterogeneity variance estimating techniques listed in Section 7.4 could be applied to some of our simulated meta-analysis scenarios. The main example of this is our proposed GLMM-based approaches, which failed to converge to a solution for a number of scenarios not originally listed in Section 4.7. If an estimator fails to converge for several scenarios then their efficiency will be lower than that of other approaches that can be applied. Only those estimators that involve iteration in their application or are based on the use of models would be affected by convergence-based efficiency issues, as these methods involve converging to a solution in order to generate an estimate. As such, estimators from our simulation study that meet this description include the Paule-Mandel (PM) method, maximum likelihood-based approaches and our novel GLMM and mixture model-based approaches. The percentage of simulated meta-analyses for which these estimators were unable to converge in our simulation study, grouped by distinct event probabilities and sample sizes, can be seen below in Table 8.3. The results in this table are based on the simulation setting $p_0 < p_1$ with $\sigma_\alpha^2 = 0.1$, binomial event count sampling, uniform sample size sampling and constant continuity corrections (where appropriate).

TABLE 8.3: Summary of percentage of non-convergences of iterative estimators by scenario groupings for event probability scenario $p_0 < p_1$.

Probability	**Sample size**	**Iterative or model-based estimator**						
		PM	**ML**	**REML**	**AREML**	**PMRM**	**CLMRM**	**MM**
Very rare	Small	0	0	0	0	16.67	24.63	23.60
	Small-to-medium	0	1.81	1.97	1.89	16.67	2.27	5.03
	Medium	0	2.43	2.36	2.45	16.67	0.55	3.91
	Small and large	0	0.83	0.72	0.79	33.33	0.71	4.42
	Large	0	0.13	0.16	0.12	16.67	0	3.43
Rare	Small	0	0.10	0.13	0.11	0	2.23	5.21
	Small-to-medium	0	0.31	0.23	0.29	0	0.02	3.54
	Medium	0	0.06	0.10	0.06	0	0.01	3.68
	Small and large	0	0.01	0.20	0.01	0	0.01	2.74
	Large	0	0.01	0.03	0.01	0	0	1.53
Common	Small	0	1.77	1.65	1.77	0	0.07	3.76
	Small-to-medium	0	0.02	0.04	0.02	0	0	2.15
	Medium	0	0.01	0.02	0.01	0	0	1.36
	Small and large	0	0.01	0.09	0.01	0	0	2.64
	Large	0	0	0	0	0	0	5.54

Table 8.3 shows that the PM approach is the most efficient of the iterative estimators included in our simulation study, as it was always able to converge for all scenarios portrayed here. The maximum likelihood approaches (ML, REML and AREML) perform very similarly to one another, only failing to converge for a very small percentage of meta-analyses in general ($\leq 2.4\%$), and even managing to successfully converge for all simulations when very rare events were coupled with small sample sizes. Our novel

methods do not perform as consistently across scenarios. The PMRM approach always converges for rare and common events, however when events are very rare, it fails at least 16.7% of the time, struggling most when sample sizes are unbalanced. In contrast, our CLMRM method only performs very poorly with very rare events and small sample sizes, and is comparable to the maximum likelihood-based results in all other cases, even managing to successfully converge 100% of the time when sample sizes were large and in almost all common-event scenarios. Finally, our mixture model (MM) approach can be seen to always suffer from non-converges, but only to a small extent in general, with the only major problem being with very rare events and small sample sizes again.

8.4.1 Cases where the PMRM method could not be applied

As mentioned previously, in Section 4.7 we outlined scenarios for which the PMRM approach could not be applied as a result of its model structure. During the course of our simulation study, we discovered further scenarios for which the method could not converge in most simulations, as well as one-off meta-analyses that shared no common pattern but which suffered from difficulties with convergence. These meta-analysis scenarios for which our PMRM R code could not be applied are as follows:

- The number of studies $k = 5$ with $k - 1$ zero events in the control arm and some double-zeros studies: We found that the Poisson model sometimes failed to converge when applied to meta-analyses of this structure, however we could not find any obvious pattern linking these problematic meta-analyses. Therefore it is likely to merely be a result of the severe rarity of the data involved. In order to avoid any such non-convergences, we decided not to apply the method to any meta-analyses that were simulated with both $k = 5$ and a very rare events scenario ($\alpha = -6.9$ or $\alpha = -5.3$).
- The sample size is unbalanced (small and large), $k = 10$ and $\alpha = -6.9$ or $\alpha = -5.3$.
- Three unrelated simulated meta-analyses that have the following characteristics:
  1. Unbalanced small and large studies, $k = 5$, $\tau^2 = 0.8$, $\alpha = -4.6$, $\theta = 0$, $\sigma_\alpha^2 = 3$, Poisson event count sampling and Chi-squared sample size sampling - 2 studies were double-zero, the remainder were single-zero (but not in the same arm).
  2. Unbalanced small and large studies, $k = 5$, $\tau^2 = 0$, $\alpha = -3$, $\theta = -1.6$, $\sigma_\alpha^2 = 3$, binomial event count sampling and uniform sample size sampling - 3 studies were double-zero.
  3. Unbalanced small and large studies, $k = 20$, $\tau^2 = 0.8$, $\alpha = -5.3$, $\theta = -1.6$, $\sigma_\alpha^2 = 3$, Poisson event count sampling and Chi-squared sample size sampling - 13 studies were double-zero, 6 were single-zero.

Although the individual problematic meta-analyses shared some characteristics (e.g. unbalanced sample sizes, $\sigma_\alpha^2 = 3$), there were no other discernible links in terms of their structure and/or simulation, and so we prevented the approach from being applied to these specific meta-analyses rather than a class of scenarios.

8.4.2 Cases where the CLMRM method could not be applied

As with the PMRM method above, our CLMRM approach also encountered a number of unforeseen difficulties in terms of convergence with our simulated meta-analyses. In particular, the meta-analyses for which we found the CLMRM to be incompatible were:

- Three unrelated simulated meta-analyses:
  1. Meta-analyses 1 and 2 from the unrelated list above for PMRM
  2. Unbalanced small and large studies, $k = 5$, $\tau^2 = 0$, $\alpha = -5.3$, $\theta = -1.6$, $\sigma_\alpha^2 = 3$, Poisson event count sampling and Chi-squared sample size sampling - 3 studies were double-zero.

As with the PMRM method, we were unable to find any obvious similarities in terms of data structure between these 3 particular meta-analyses, and so simply prevented the CLMRM approach from being applied to them in our study.

8.5 Performance in estimating τ^2

As we have investigated so many scenarios in our simulation study, we shall primarily focus on presenting the results of the main scenarios of interest - rare and very rare event probabilities. Therefore, in this section and the remainder of this chapter, unless stated otherwise, we shall present only the results for these scenarios, combined with the following simulation options: $p_0 < p_1$, $\sigma_\alpha^2 = 0.1$, constant continuity corrections, and event count and sample size sampling from binomial and uniform distributions respectively. However, we shall mention the results associated with our alternative simulation parameters where appropriate and present some of these in Appendix E.

If any of the estimators consistently produce outlying results over a set scenario grouping (e.g. a particular sample size), to the extent that their inclusion in our figures would significantly distort the plot scale, then these will be omitted from the respective figures. However, we shall mention their absence and the direction that the outlying effect was present in. We shall also crop the number of studies (k) if outlying values are consistently present for specific k, e.g. $k = 5$. Finally, for each of the plots we shall display the mean heterogeneity I^2 for that scenario given the chosen value of τ^2, as this can vary considerably over different sample sizes.

8.5.1 Bias of τ^2

Figure 8.1 displays the mean bias of τ^2 estimates in the case of very rare events, for small, unbalanced and large studies, and varying degrees of heterogeneity. For small sample sizes (plots A1-A3), we did not include our CLMRM method as this produced extremely large bias (> 1000). Similarly we excluded the semi-Bayesian RB and RB0 approaches, which consistently produced very unusual results, as can be seen in Appendix E. As a result, the RB and RB0 will not be included for small studies in the majority of plots presented in this chapter. In addition, three versions of our conditional-based approach (CO2, CO3, CO4) and MM method are not included in this figure, as they produced outlying high bias (as is demonstrated in Figure 8.1), and $k = 5$ is not displayed for similar reasons.

By looking at Figure 8.1, it can be seen that PMRM appears to consistently perform best in terms of bias in this scenario when heterogeneity is present (i.e. $\tau^2 > 0$), but only when $k > 20$ if sample sizes are unbalanced. It is not displayed for $k < 20$ in B1-B3 as it could not be applied in this scenario. Our CLMRM method performs very similarly, for unbalanced and large studies. However, both of these GLMM approaches perform poorly in the case of homogeneity, indicating their prevalence to produce non-zero estimates, with the majority of pre-existing estimators performing better. The pre-existing methods all appear to perform similarly except for the semi-Bayesian BM approach, which consistently has a dramatic crop in bias over increasing k, leading it to at least pass through the zero line with small studies. In that particular scenario, the only other estimator to near the desired zero bias was PMRM, which appeared to oscillate around the optimum bias when $\tau^2 > 0$. Whereas the pre-existing estimators always underestimate τ^2, our GLMM approaches tend to overestimate in general.

Figure 8.2 displays the mean bias in the rare events scenario, again with $p_0 < p_1$. The results differ slightly from those of the previous scenario, as PMRM can be computed for $k < 20$ in unbalanced cases, and all included estimators produce reasonable estimates when $k = 5$ (and thus are included here). For small studies, the methods perform as above, with PMRM and BM being closest to zero when $\tau^2 > 0$, with PMRM unusually having maximum bias around $k = 20$. For the remaining sample sizes, when $\tau^2 = 0$, the SJ estimator performs very poorly compared to the alternate estimators. When $\tau^2 > 0$ in large studies our modified conditional methods (CO2-CO4) have very high bias, but when $\tau^2 = 0$ the original CO1 performs best, having consistent near-zero bias. Our two GLMM approaches generally perform similarly to the pre-existing methods for unbalanced and large studies, however they can again be seen to perform best for high k (> 20) when $\tau^2 > 0$.

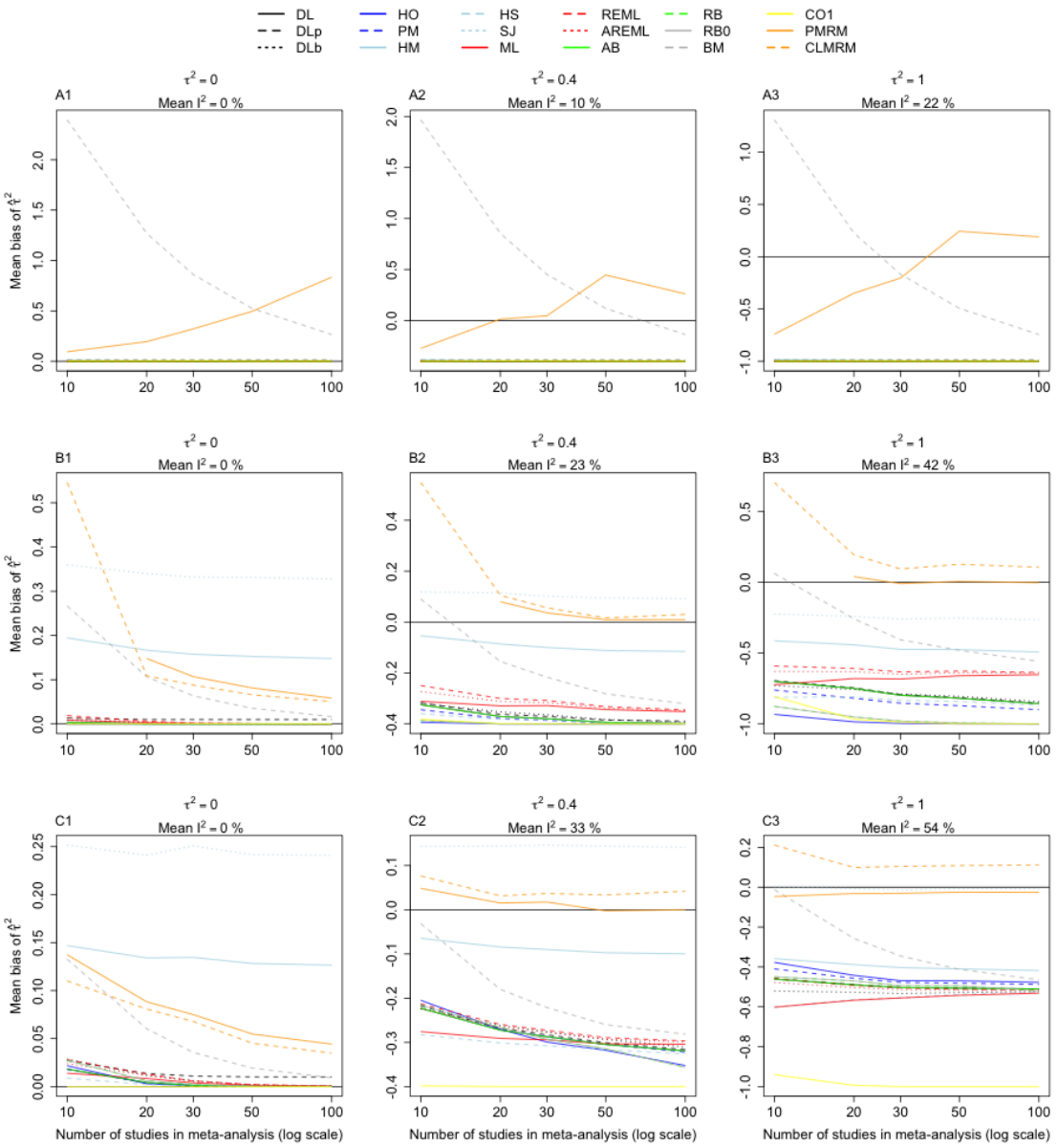


FIGURE 8.1: Mean bias of heterogeneity variance estimates in very rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). RB, RB0 and CLMRM have been omitted from A1-A3; CO2, CO3, CO4 and MM have been omitted from all.

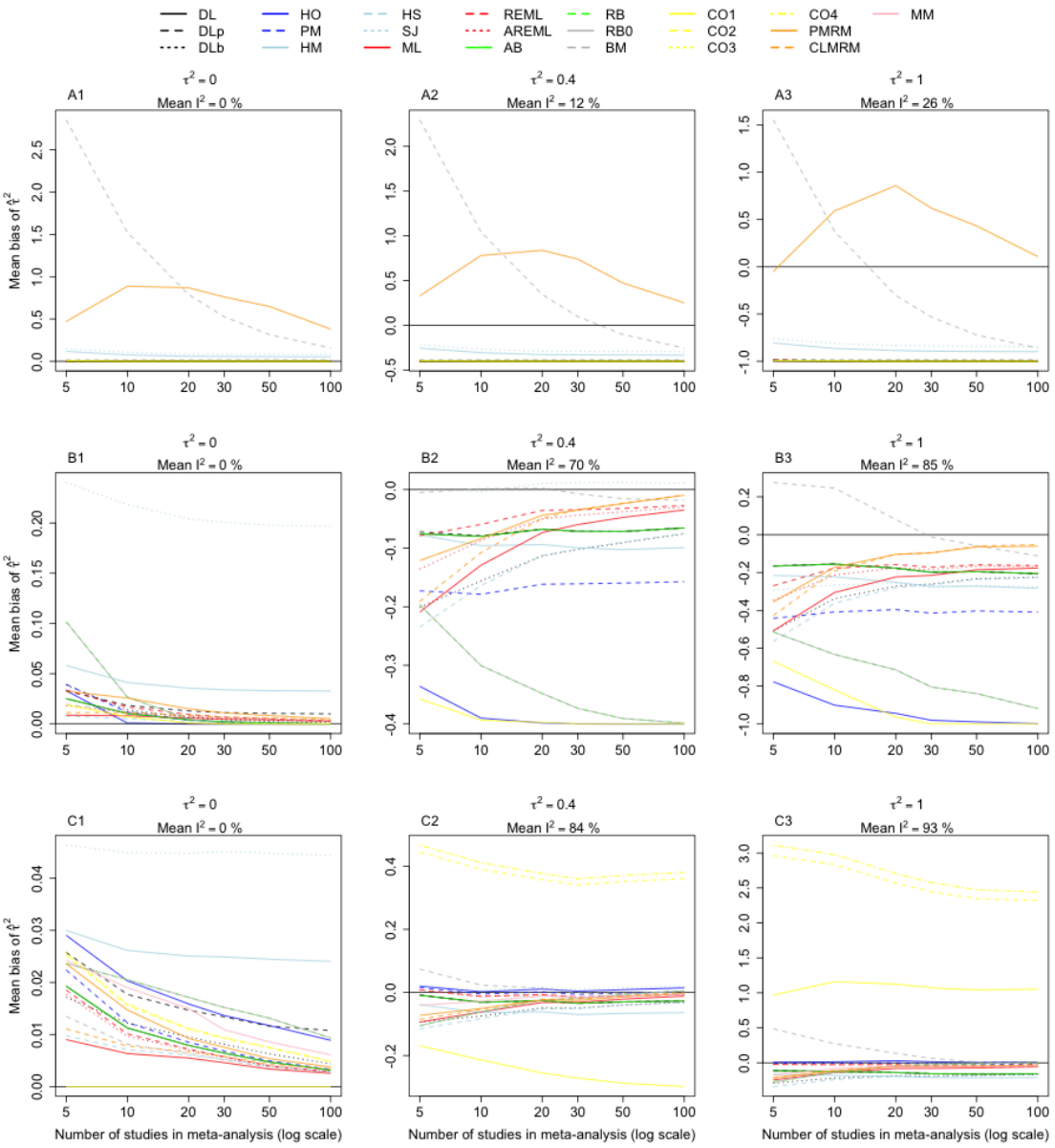


FIGURE 8.2: Mean bias of heterogeneity variance estimates in rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). RB, RB0 and CLMRM have been omitted from A1-A3; CO2, CO3, CO4 and MM have been omitted from A1-B3.

It is difficult to identify a general pattern in the plots given in the previous figures, as the varying estimator types behave differently (and in some cases converge to different values). As a result, we shall also look at other simulation scenarios, to determine whether these changes in behaviour are consistent across different settings and whether any general patterns can be identified.

The results for the alternate and common probability scenarios can be seen in Section E.1. Above we presented the results for $p_0 < p_1$, which we found to be similar to the results for $p_0 = p_1$. In the alternate scenario with $p_0 > p_1$, we find that the results in general are very comparable to these previous cases for the majority of estimators presented, however the conditional-based methods (CO1-CO4) notably change direction of bias. For common events, CLMRM has a much more reasonable level of bias for small samples compared to above. Additionally, when $p_0 < p_1$, the alternative conditional-based methods (CO2-CO4), which consistently performed poorly in the previous cases, appear to perform the best when $\tau^2 > 0$ and $k > 10$, demonstrating that they are applicable for this scenario. However, when $p_0 > p_1$, they do not perform as well, with ML-based methods generally performing best. Additionally, for common events, MM has bias comparable to the other methods when studies are large.

In our simulation study, we also looked at the effect of sampling our parameters from various distributions. The results produced from these alternate sampling techniques are also displayed in Section E.1. We found these results to be very similar to those presented so far in this chapter. The degree of bias differs in some cases when k is small, however the differences between estimators is identical, confirming the results we have presented here. In addition to this, the pre-proposed estimators were also calculated using the reciprocal continuity correction outlined in Section 1.9.1, however very few differences in results were found between this and the original constant correction. We also calculated the median bias of τ^2 for all of these scenarios, and a subset of these results are presented in Section E.4. In summary, the results for median bias do not differ substantially from those presented here for the mean bias, so again back up our findings.

8.5.2 Mean squared error of τ^2

We also looked at the mean squared error (MSE) of the τ^2 estimates, and Figures 8.3 and 8.4 display these results for very rare and rare events, again grouped by sample size and heterogeneity. By looking at Figure 8.3, we can see that in the case of small studies, all included estimators perform very similarly for very rare events, apart from CO1 which displays unusual behaviour when $\tau^2 = 0$. Our CO2-CO4 and MM methods had extremely high MSE in all cases, and so are not included in this figure. Similarly, as before, RB and RB0 were not included for small studies due to their outlying results, and our two GLMM approaches were also omitted in this case. However, when sample sizes

increased, the performance of PMRM and CLMRM improved notably, with their MSE being some of the lowest when $\tau^2 > 0$ and $k > 30$. In this plot, we only presented $k > 20$ as the majority of estimators performed very poorly otherwise. Our CO1 approach appears to perform well for larger studies and homogeneity, with close to zero MSE, however as τ^2 increases above zero so does their MSE. In terms of the pre-existing estimators, they tend to perform rather similarly to each other, apart from HM and SJ, which perform poorly for $\tau^2 = 0$, but significantly outperform the other approaches in heterogeneous cases.

Figure 8.4 shows us that, in the case of rare events, HM and SJ appear to perform best for small studies when $\tau^2 > 0$, but have the largest MSE of the pre-existing estimators with homogeneity (regardless of sample size). As before, our GLMM approaches had too high MSE to include in the small sample size plots, and our MM method, as well as all of the conditional-based methods, was omitted in general from the figure. For unbalanced and large studies, the GLMM-based approaches both performed very similarly to the existing estimators, which all demonstrated a smooth decline in MSE towards zero as k increased. The exceptions to this are HO and HM, which both maintained a constant high MSE in the case of unbalanced sample sizes. It should also be noted that, in comparison to the previous figure, the results for $k = 10$ are displayed in this case, as the MSEs for this scenario were much more reasonable when events are still rare but slightly more common.

As before, we also looked at the probability scenarios where $p_0 > p_1$ and $p_0 = p_1$, and at the case of common events, and these results can be seen in Section E.2. For very rare events with $p_0 > p_1$, the results are very similar to the reverse case, with the exception that the magnitude of MSE is reduced for CLMRM in small k (< 30). In addition, the MSE of HO appears to be consistently slightly elevated when $p_0 > p_1$, while the MSE of CO1 can be seen to do the opposite. When events are rare, MSE results are again very similar regardless of the relationship between p_0 and p_1. However, when $p_0 > p_1$, CO1 can be seen to have a MSE similar to the other estimators, whereas in the alternate case it was omitted for outlying. In addition, when $p_0 = p_1$, CO2-CO4 perform similarly to the other methods when k and studies are large and heterogeneity is at an extreme, compared to the other scenarios where their MSE was very high.

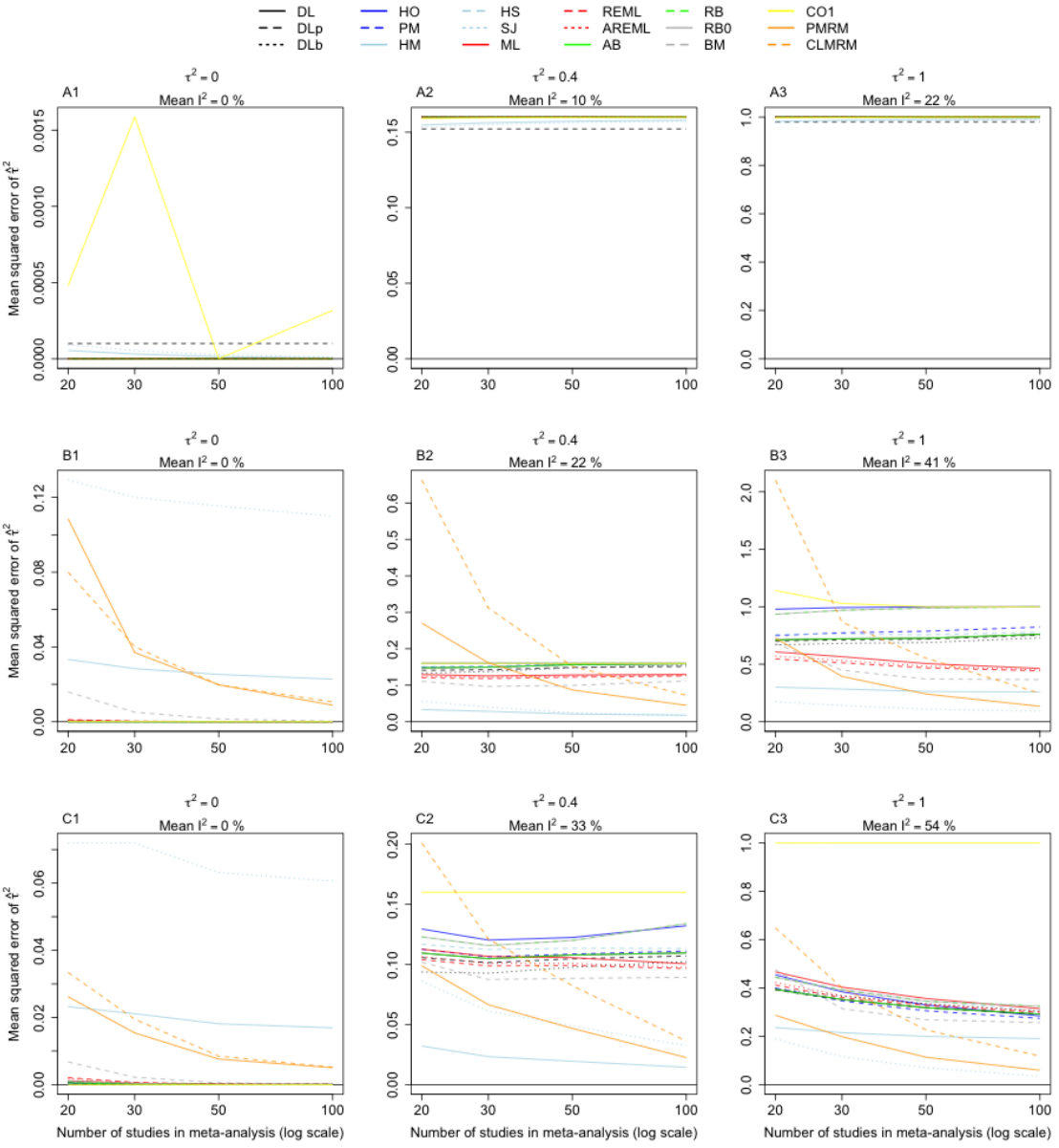


FIGURE 8.3: Mean squared error of heterogeneity variance estimates in very rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). RB, RB0, BM, PMRM and CLMRM have been omitted from A1-A3; CO2, CO3, CO4 and MM have been omitted from all.

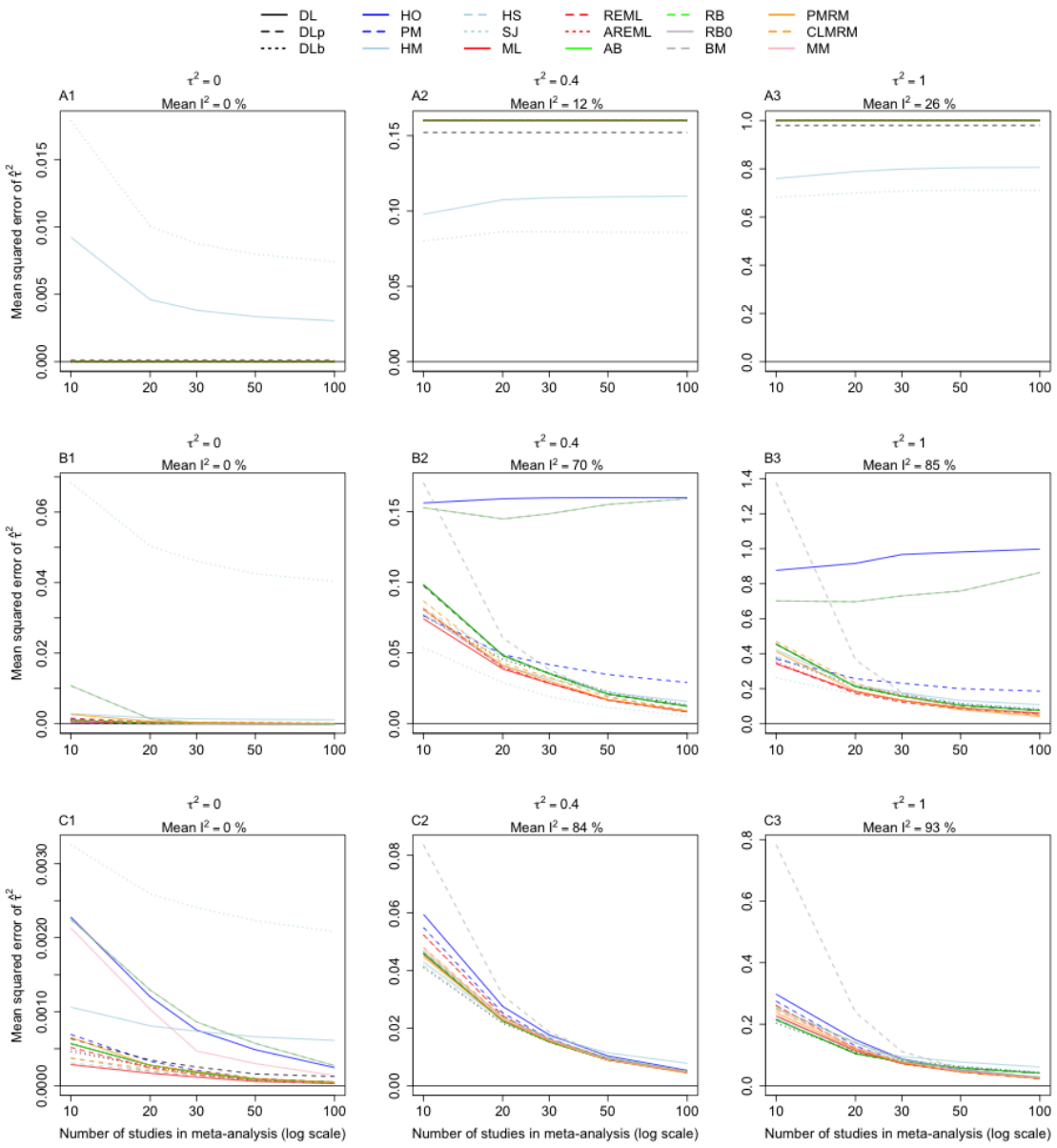


FIGURE 8.4: Mean squared error of heterogeneity variance estimates in rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). RB, RB0, BM, PMRM and CLMRM have been omitted from A1-A3; MM has been omitted from A1-B3; CO1, CO2, CO3 and CO4 have been omitted from all.

In terms of common events, when $p_0 < p_1$, CLMRM has a very low MSE when $k > 20$, particularly in the case of small studies with $\tau^2 > 0$, although HM and SJ are the best performers in this specific scenario. All the conditional-based approaches have similar and the lowest MSE of all methods considered when $k > 20$ and studies are large, however produce very poor results when $k < 10$ (as does CLMRM). In the case of $p_0 > p_1$, the MSE of CLMRM in the case of small studies is again very low, being the optimum estimator when $\tau^2 = 1$ and $k > 20$. CO1-CO4 appear to have very high MSE when homogeneity is present, while our GLMM-based estimators produce results similar to the existing methods when at least some studies are large. In both probability scenarios, MM has a MSE similar to other methods when sample sizes are large, while DLp has the highest MSE of those presented when $\tau^2 = 0$ here.

The results for MSE of τ^2 in alternate sampling scenarios in our simulation study can also be seen in Section E.2. As expected, the results for MSE are very similar in general regardless of the distributions used to sample the event count and non-constant sample sizes. As with the bias, we also calculated the median squared error of τ^2 for all of the scenarios discussed here, and some of these results can be seen in Section E.5. These results do not differ substantially from those presented here for the MSE, and as such provide confirmation to their findings.

In order to provide a summary of the estimators' performance in comparison to each other for the main probability scenarios (very rare, rare and common), we choose to rank the estimators in terms of MSE for their estimation of τ^2. We then tabulated the average ranking of the estimators in terms of MSE for τ^2 for the various sample sizes, in the case of $p_0 < p_1$ only (as event probability relationship did not appear to have a significant effect on the majority of estimators considered in our study). The top 10 ranked estimators in terms of MSE for very rare, rare and common events can be seen in Tables 8.4, 8.5 and 8.6 respectively.

From looking at Table 8.4, we can see that for very rare events, the non-truncated method of moments-based approaches (SJ and HM) appear to consistently perform best regardless of sample size. DLb and the ML-based methods perform next best, with DLp, PM, DL and AB occurring towards the bottom of the table (the latter two performing at the same level consistently). Only one of our proposed methods (PMRM) occurs in the top 10, although it is very low down, and this is for large studies. For rare events, Table 8.5 shows that while the non-truncated moments-based methods perform best while studies are medium-sized or less, ML-based methods outperform them when at least some studies are large. In this case, PMRM is third and CLMRM sixth when studies are large, and are respectively ranked seventh and eighth for unbalanced sample sizes. Finally, PMRM is ranked at the very bottom of the table for medium-sized studies, and none of our proposed methods are in the top 10 for smaller studies.

TABLE 8.4: Average rankings of top 10 estimators for MSE by scenario groupings for very rare event probability scenario $p_0 < p_1$.

Ranking of estimators	**Sample size**				
	Small	Small-to-medium	Medium	Small and large	Large
1	SJ	SJ	SJ	HM	HM
2	DLp	HM	HM	SJ	DLb
3	HM	DLb	DLb	REML	REML
4	DLb	REML	REML	ML	SJ
5	AREML	AREML	AREML	DLb	AREML
6	ML	ML	ML	AREML	ML
7	REML	DLp	DLp	HS	DLp
8	DL AB	DL AB	PM	PM	PMRM
9			DL AB	DLp	DL AB
10	PM	PM		DL	

TABLE 8.5: Average rankings of top 10 estimators for MSE by scenario groupings for rare event probability scenario $p_0 < p_1$.

Ranking of estimators	**Sample size**				
	Small	Small-to-medium	Medium	Small and large	Large
1	SJ	HM	HM	ML	ML
2	HM	DLb	DLb	REML	REML
3	DLp	SJ	DLp	AREML	PMRM
4	REML	REML	REML	SJ	AREML
5	DLb	ML	DL	DLb	DLb
6	AREML	AREML	AB	HS	CLMRM
7	ML	DLp	SJ	PMRM	HS
8	DL AB	DL AB	ML	CLMRM	DL AB
9			AREML	DL AB	
10	PM	BM	PMRM		DLp

TABLE 8.6: Average rankings of top 10 estimators for MSE by scenario groupings for common event probability scenario $p_0 < p_1$.

Ranking of estimators	**Sample size**				
	Small	Small-to-medium	Medium	Small and large	Large
1	HM	SJ	SJ	REML	REML
2	SJ	REML	REML	AREML	AREML
3	DLb	PM	PM	CLMRM	PM
4	REML	AREML	AREML	ML	SJ
5	DLp	ML	ML	SJ	CLMRM
6	BM	PMRM	CLMRM	BM	ML
7	AREML	CLMRM	BM	PMRM	PMRM
8	ML	DL AB	PMRM	DL AB	BM
9	DL AB		HO		HO
10		DLp	DL AB	DLb	MM

In terms of common events, Table 8.6 shows that as before the non-truncated moments-based methods and ML-based methods consume the top of the table in terms of ranking. However, our CLMRM method is ranked third for unbalanced studies and fifth for large studies, and PMRM (and CLMRM where not already mentioned) consistently appear in the table, albeit in the bottom half, in all scenarios other than small studies. In addition, our MM method appears at the very bottom of the table for large sample sizes. It should be noted that these tables give a very crude summary of the estimators' performance in terms of only MSE, and as such do not describe overall performance.

8.5.3 Proportion of zero τ^2 estimates

We also looked at the proportion of zero τ^2 estimates produced by all methods, for both cases when $\tau^2 = 0$ and $\tau^2 > 0$. We chose to look at the case when $\tau^2 > 0$, in addition to the obvious homogeneous case, in order to identify those estimators that have a tendency to produce zero estimates when this is not the case, demonstrating their inability to detect heterogeneity. Ideally an estimator would produce more zero estimates when homogeneity is present, but should not produce any such estimates in heterogeneous cases, and so the best estimator would have a step-function behaviour dropping from 100% to 0% as τ^2 becomes positive.

Figure 8.5 displays the percentage of zero τ^2 estimates in the case of very rare events with $p_0 < p_1$. We can see that the semi-Bayesian RB and RB0 behave erratically for small studies, agreeing with their behaviour with bias and MSE. It is very clear in these plots for which scenarios PMRM cannot be applied (generally $k < 10$ or $k < 20$). When homogeneity is present (plots A1, B1 and C1) and we would expect the percentage of

zero estimates to near 100, the existing estimators all perform very similarly and well in general. However, our CO1 method consistently performs poorly with all sample sizes (with a maximum of 60%), while our PMRM approach generates an undesirable non-zero results for unbalanced and large studies. Our PMRM and MM methods also perform fairly poorly in homogeneous cases. When heterogeneity is present, however, PMRM performs best (when it can be applied) when studies are not small, along with the DLp and HM methods. CO2-CO4 consistently produce near-zero percentages when studies are small, and together with CLMRM and MM produce results better than existing methods in all other scenarios.

In terms of rare events, Figure 8.6 shows a similar pattern, however PMRM can be applied in all cases and is consistently the best estimator (along with DLp and HM) when $\tau^2 > 0$ and at least some studies are large. In the homogeneous case, CO1 is the best performer with a constant percentage of near 100, however it performs very poorly when $\tau^2 > 0$ as it retains that high number of zero estimates. While the majority of estimators tend towards the optimal result as k increases, there are some cases where this is not the case (particularly for unbalanced sample sizes where the opposite trend is observed with some methods). It can also be observed that for large studies, all estimators (with the exception of CO1) have desired near-zero percentages when $k > 10$, displaying the general ability to detect no heterogeneity in these cases.

As before, the results for the alternate scenarios not displayed here are given in Section E.3. For very rare events with $p_0 > p_1$, the results are very similar, however CO1 does appear to perform much better when $\tau^2 > 0$, producing a percentage that tends towards zero at a similar rate to the other estimators. This is also seen for rare events with $p_0 > p_1$, as CO1 again produces a percentage of zero estimates that is comparable to the other methods. However, this percentage has also dropped in the homogeneous case, resulting in the estimator losing its consistent near-100 success rate. When $p_0 = p_1$ with rare events, the drop for CO1 is not as extreme as that for $p_0 > p_1$, but it still appears to perform much better than the case displayed here. For common events, the majority of pre-existing, MM and GLMM methods perform very well when $\tau^2 > 0$ and studies are not small, producing near-zero percentages for $k > 10$. When sample sizes are small and $\tau^2 > 0$, PMRM again performs well alongside DLp and HM for $k > 20$, while CO1 consistently performs best in all cases where $\tau^2 = 0$. The main difference between $p_0 < p_1$ and $p_0 > p_1$ when events are common is that CO1 significantly improves in performance when $p_0 > p_1$, as seen above.

Very similar results can be seen between those presented here and those generated from alternate sampling distributions (as shown in Section E.3), again providing confirmation of the results obtained in our simulation study.

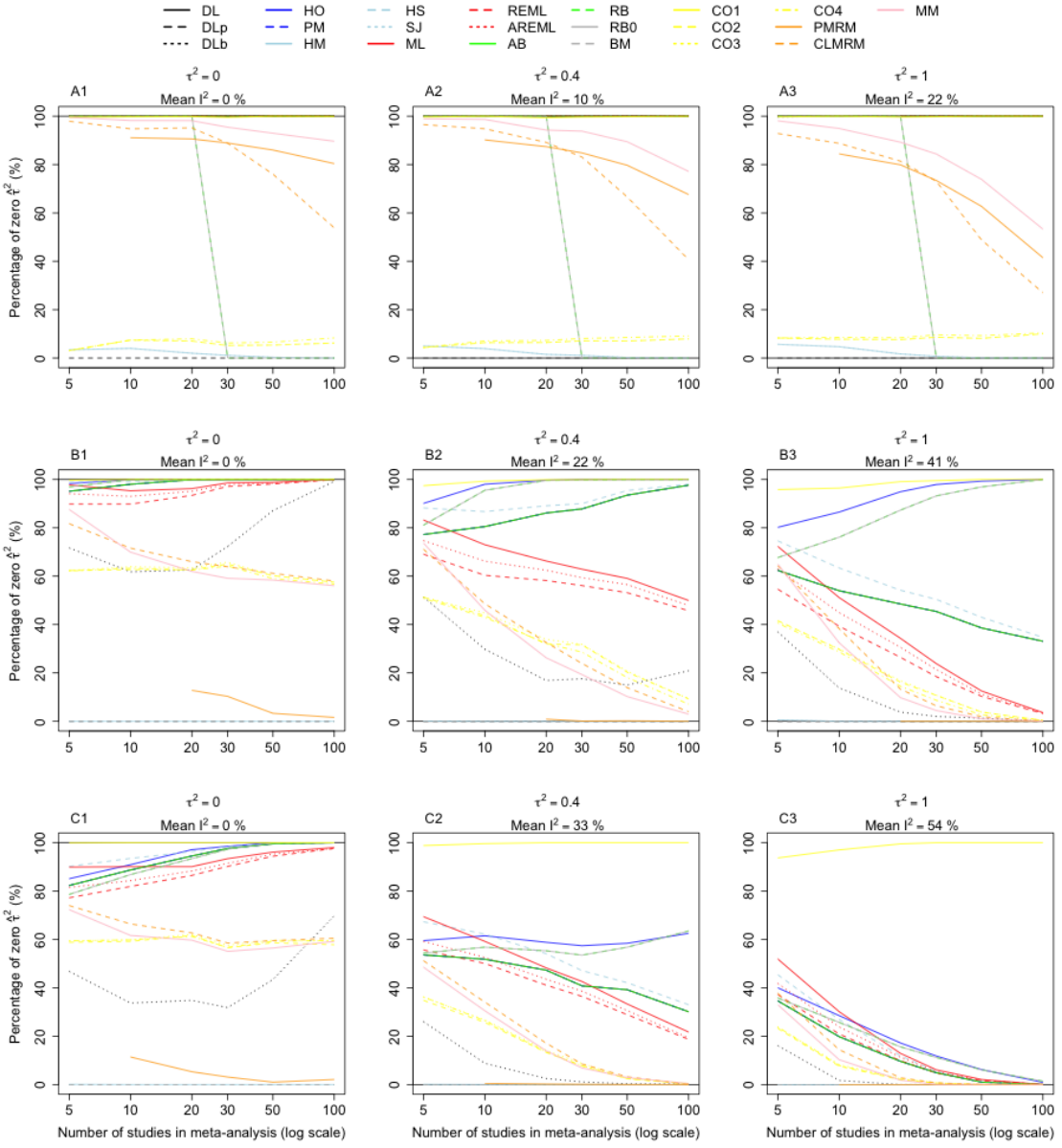


FIGURE 8.5: Proportion of zero heterogeneity variance estimates in very rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

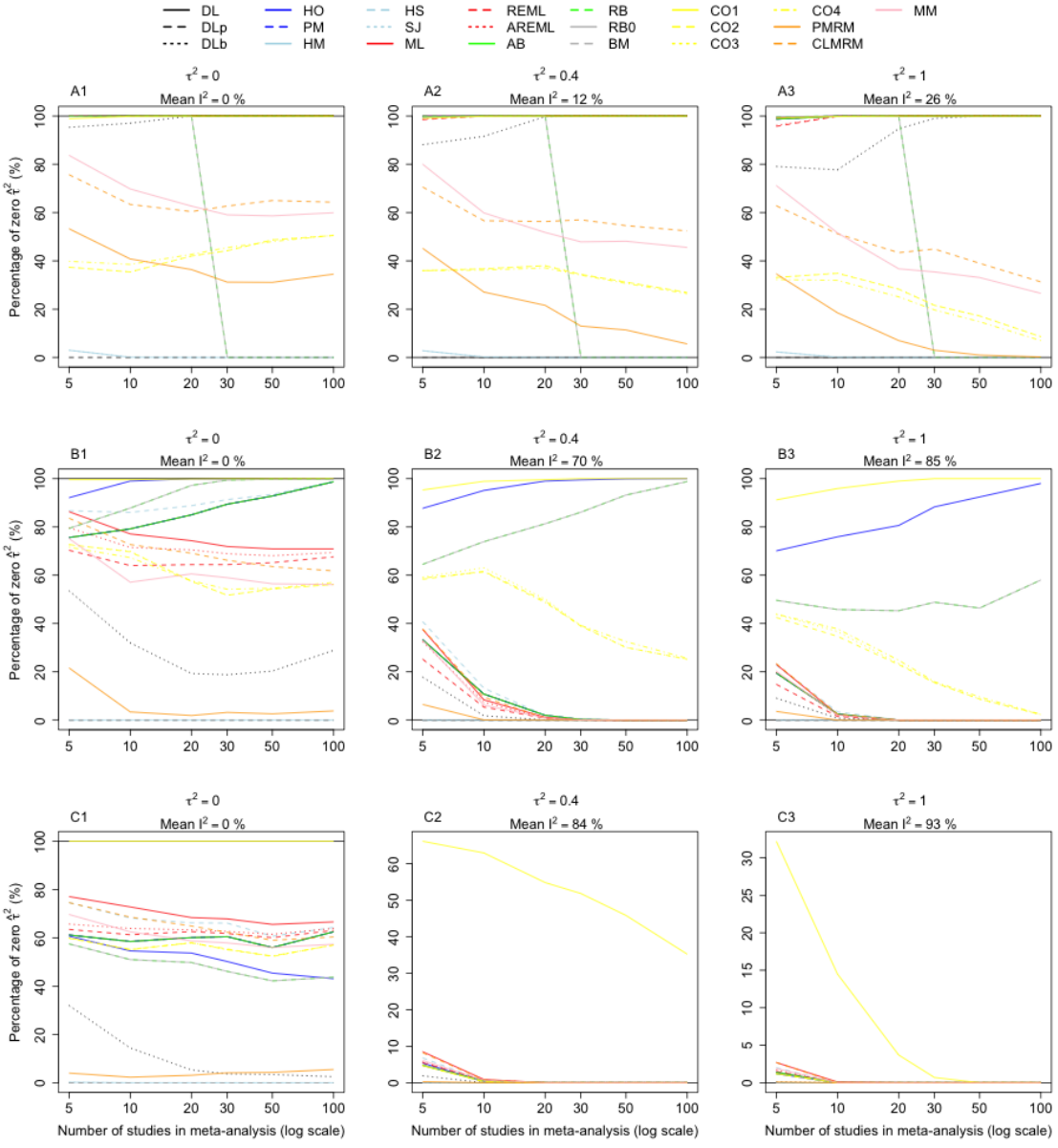


FIGURE 8.6: Proportion of zero heterogeneity variance estimates in rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

8.5.4 Summary of performance

Table 8.7 provides a summary of the performance of τ^2 estimators in terms of the performance measures we have presented thus far, with respect to their estimation of τ^2 itself. We have focused on our novel approaches and their performance with respect to mean bias and MSE.

TABLE 8.7: Summary of performance of estimators in estimating τ^2 by scenario groupings.

<table>
<tr><th rowspan="2">Sample size</th><th colspan="3">Probability scenario</th></tr>
<tr><th>Very rare</th><th>Rare</th><th>Common</th></tr>
<tr><td>Small</td><td rowspan="3">When $\tau^2 > 0$, PMRM has low bias but high MSE, CO1 has low MSE</td><td>PMRM and CO1 both have low bias and MSE</td><td rowspan="3">GLMM methods have low bias and MSE when $k > 30$</td></tr>
<tr><td>Small-to-medium</td><td rowspan="2">GLMM methods have low bias and MSE when $\tau^2 > 0$</td></tr>
<tr><td>Medium</td></tr>
<tr><td>Small and large</td><td colspan="2" rowspan="2">GLMM methods have low bias and MSE when $\tau^2 > 0$ and $k > 30$</td><td>CO2-CO4 have small bias but high MSE, PMRM has low MSE</td></tr>
<tr><td>Large</td><td>CO2-CO4 have small bias and MSE for $k > 10$</td></tr>
</table>

8.5.5 Performance of conditional-based approaches in estimating τ_p^2

As the conditional-based methods we introduced in Chapter 5 generate their τ^2 estimates using an estimate of probability-based τ_p^2 (discussed in detail in the chapter), and we were able to determine the true value of τ_p^2, we decided to also look at the performance of this group of methods in terms of estimating this parameter. This will allow us to determine whether they can estimate this parameter well, as if it transpires that they cannot estimate τ^2 then it is possible that an alternate conversion between the two parameters should be sought. A selection of results produced relating to the performance of these methods in estimating this method-specific value is given in Section E.6. We presented the results for all probabilities (i.e. our cases with $p_0 < p_1$, $p_0 > p_1$ and $p_0 = p_1$) because we observed such dramatic differences in these for the conditional-based approaches above.

In terms of the mean bias of τ_p^2, CO1 can be seen to consistently outperform the alternate methods (CO2-CO4) in homogeneous scenarios, regardless of event probability or study sample size. In general, CO1 can be seen to have negative bias, while the alternate methods overestimate τ_p^2. The alternative approaches (based on variations of the original estimating equation), perform very similarly to each other, and appear to only have less bias than CO1 when heterogeneity is present and samples sizes are either unbalanced or large. The only scenario where this is generally not the case is with large studies and considerable heterogeneity ($\tau^2 = 0$). The bias of CO2-CO4 also drops dramatically

when the probability scenario is based on the relationship $p_0 > p_1$, compared to cases with $p_0 < p_1$ or $p_0 = p_1$, agreeing with the results for τ^2 discussed above.

When looking at the MSE of τ_p^2, CO1 is observed to have lower MSE in the majority of scenarios investigated, however all methods have similar results (particularly for high k). As with the bias, MSE can be seen to be lower in CO2-CO4 in certain scenarios with heterogeneity and at least some large studies, particularly for the cases where $p_0 > p_1$.

8.6 Performance in estimating θ

In addition to determining the performance of methods in estimating τ^2, we also looked at their performance in estimating the summary effect size measure, θ, in this case the log-risk ratio. This will allow us to determine their ability to produce an accurate result for the meta-analysis using their respective τ^2, via either the inverse-variance approach or otherwise (as with our novel methods). In addition to the τ^2 estimators, we have also included the fixed-effect Mantel-Haenszel (MH) approach, as well as a variation including the addition of a constant continuity correction of 0.5 (which will shall denote by MHc), as discussed in Section 1.9.5. The plots given here are based on the scenarios portrayed above, with results for further scenarios again being available in Appendix E.

8.6.1 Bias of θ

The mean bias of the overall log-risk ratio estimates for very rare events can be seen in Figure 8.7. From looking at this we can see that for small studies, while the majority of estimators have a bias near zero that closes in as k increases, CLMRM actually has a near-zero bias for lower k but increases away from this as k increases. MM shows a similar pattern but to a lower extent. In the case of small sample size, all methods have negative bias (for small k at least), and the fixed-effect MH and MHc actually perform best in general here. When studies are unbalanced or large, the pre-existing and conditional methods all have negative bias in general, while our GLMM and MM methods, and the MH-based approaches, consistently overestimate θ. When PMRM can be applied (for either $k > 10$ or $k > 20$), it has the bias closest to zero in all cases where studies are not small except when studies are large and $\tau^2 = 1$. In this latter case, all pre-existing methods and CO1-CO4 (which perform similarly here) also have positive bias, but this is closer to zero than with our approaches.

Figure 8.8 shows the mean bias of θ when events are rare. These results differ quite prominently from those discussed above, as our CLMRM and MM methods have very high bias for small studies, with the bias forming an n-shape over increasing k. For this sample size, PMRM and the MH-approaches perform very similarly and have the best results in terms of bias, while all pre-existing methods (and CO1-CO4) consistently

underestimate θ. When $\tau^2 = 0$ and studies are unbalanced or large, our GLMM methods and the MH approaches consistently have the minimal amount of bias, and MM is also in this group when studies are large in size. However, when $\tau^2 > 0$ the picture varies significantly for cases with at least some large studies, where fixed-effect MH methods always have the highest bias. PMRM appears to consistently have the least bias in these cases, with CO1 and MM also performing well when sample sizes are highly unbalanced.

The results for the alternative probability scenarios relating to this performance measure are in Section E.7. When $p_0 > p_1$ and events are very rare, the methods are always positively biased, with the exception of the GLMM and MH-based approaches. In particular, PMRM significantly underestimates θ when studies are small, as does CLMRM for high k. However, these two methods do appear to perform best in terms of bias for all other scenarios, along with the MH methods when τ^2 is low. In all cases, CO1-CO4 perform very similarly to the existing estimators, while MM consistently produces very high bias when studies are not small. This is the major difference between the case with $p_0 < p_1$, where MM also performed fairly poorly but not to such an extent. When events are rare with $p_0 > p_1$, the results are very different to those described above. In particular, for small studies, both GLMM methods are now very negatively biased, with CLMRM instead producing a U-shaped pattern. In all other scenarios, these two approaches perform best, while MM again has consistently high bias. Finally, when $p_0 = p_1$, CLMRM needs to be slightly positively biased on average with small studies, while PMRM is heavily negatively biased again. In alternate sample size scenarios, MM performs similarly to the other methods when $\tau^2 > 0$, and PMRM consistently performs best regardless of τ^2.

When events are more common and $p_0 < p_1$, CLMRM actually has one of the best results in terms of bias for small sample sizes, along with PMRM and the MH-based approaches. When at least some large studies are present, then all of the estimators are negatively biased, with the exception of the case where $\tau^2 = 0$, where the GLMM and MH methods are the best performing with little positive bias (accompanied by MM for all large studies). When $\tau^2 > 0$, the MH-based approaches consistently have the least bias. When $p_0 > p_1$, the results are similar to those observed above for this probability relation, in that MM is consistently very highly biased in all scenarios, while the bias of the fixed-effect MH approaches can be seen to increase as the degree of heterogeneity increases. Our GLMM-based methods, however, consistently perform best in terms of bias for all scenarios displayed.

As before, we also plotted this performance measure for various sampling distributions in our simulation study, and these can be observed in Section E.7. These results mirror those already discussed here.

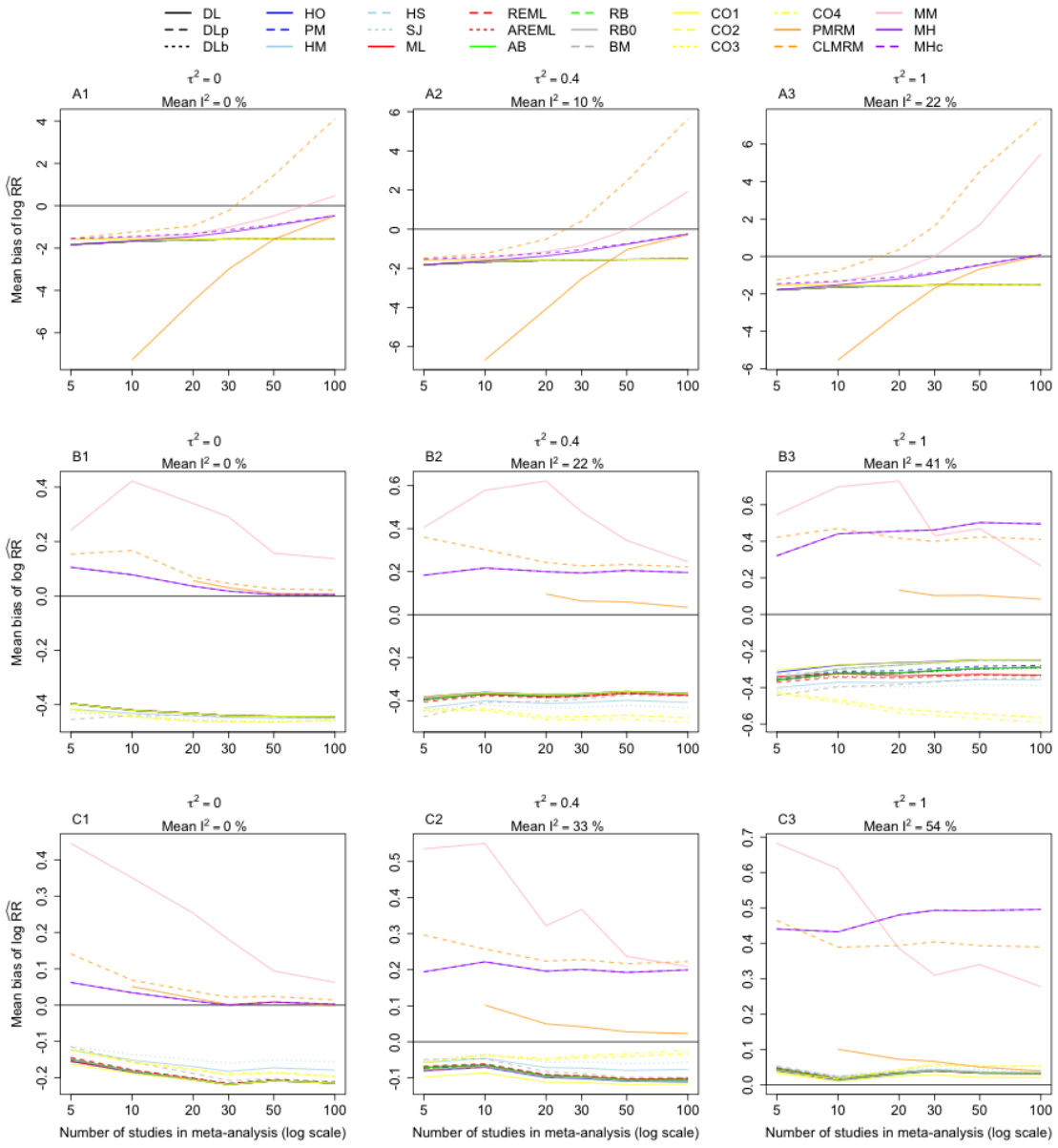


FIGURE 8.7: Mean bias of log-risk ratio estimates in very rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

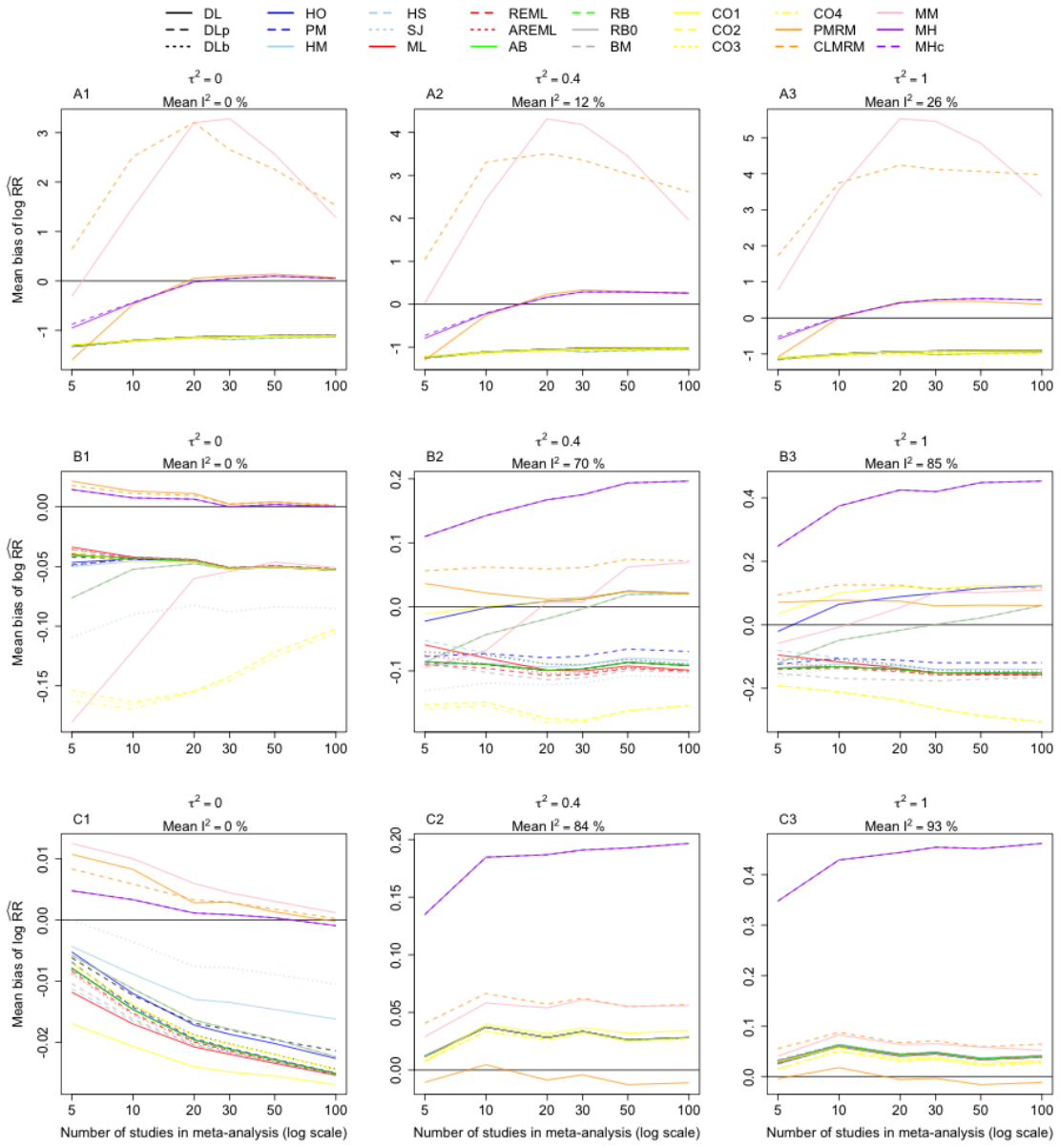


FIGURE 8.8: Mean bias of log-risk ratio estimates in rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

8.6.2 Mean squared error of θ

To measure the performance of the methods in terms of estimating the summary effect θ, we also calculated the mean squared error of these estimates. The results relating to very rare events with $p_0 < p_1$ can be seen in Figure 8.9. By looking at these plots, we can see that all τ^2 estimator-based approaches performed poorly when sample sizes are small, with the fixed-effect MH-based methods outperforming them considerably for moderate-to-high k. When sample sizes are unbalanced or large, however, and $\tau^2 = 0$, CLMRM performs equally well with the MH-methods when $k > 20$, with them having the least MSE of all methods included. It should be noted that the GLMM methods both perform very poorly with small studies, as did MM in all scenarios, and as such are omitted from the respective plots. When sample sizes are unbalanced and τ^2 is moderate, all estimators perform equally well when $k > 5$, with CLMRM performing poorly for $k < 5$ in all cases considered and PMRM only being applicable for $k > 10$ or $k > 20$. Finally, when $\tau^2 > 0$ with large studies, CO1-CO4, PMRM and the pre-existing all perform similarly and generate the least MSE.

Figure 8.10 displays the MSE of the log-risk ratio estimates for rare events. Here we can see that for small studies, despite performing very poorly for $k < 10$, our PMRM approach otherwise has the lowest MSE, along with the fixed-effect MH methods. When sample sizes are unbalanced, all estimators perform very similarly apart from MM and CO2-CO4, which consistently have higher MSEs, and the MH methods whose MSE increases considerably as τ^2 increases. Finally, when studies are large in size, all methods behave in the same manner, with their MSE decreasing as k increases. While all methods have very similar MSEs in this case, the MSE of the fixed-effect MH-based methods again appears to increase with τ^2.

The results for the alternate and common probability scenarios can be seen in Section E.8. For very rare events with $p_0 > p_1$, MM can be seen to perform reasonably for small studies and, although it still has greater MSE than the others, it is not outlying. The other methods perform very similarly to the case described above, although the original MH approach has a lower MSE for small k. When sample sizes are unbalanced or large and τ^2 is high, our GLMM-based methods appear to perform better in respect to the other estimators for this probability relation. When $p_0 > p_1$ but events are classed as rare, MM appears to perform more reasonably in the case of small studies again, while PMRM does the opposite. Meanwhile, when at least some studies are large, the MM method performs more poorly and has an outlying MSE in this case. It can also be seen here that our GLMM-based methods again perform better compared to the other methods when studies are imbalanced and $\tau^2 > 0$. Finally, when $p_0 = p_1$, the MH-based approaches produce an unusual peak-style pattern in terms of their MSE as k increases for small studies. However, MM shows good promise for unbalanced and large studies in this scenario, where it consistently has the lowest bias of those considered.

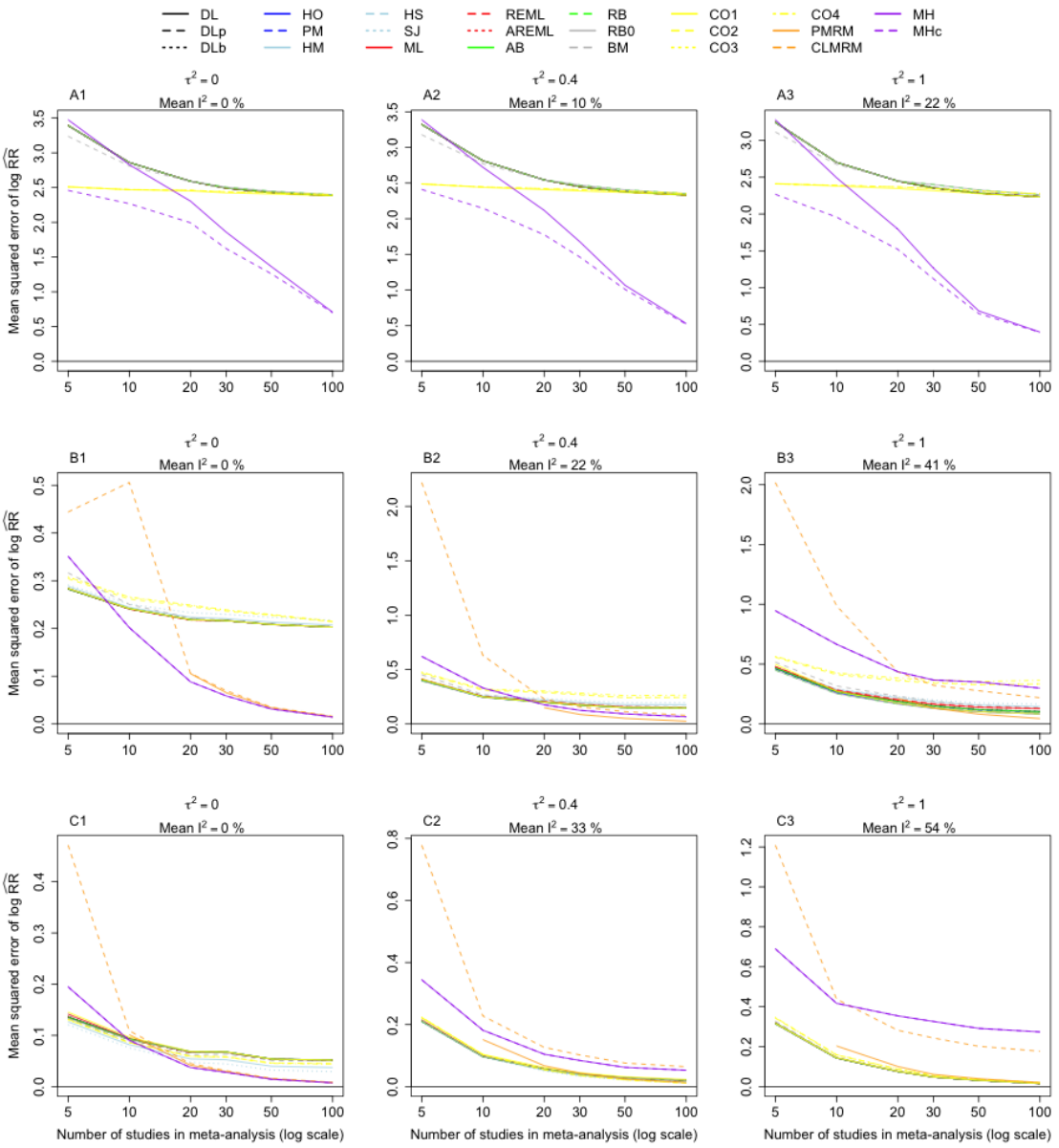


FIGURE 8.9: Mean squared error of log-risk ratio estimates in very rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). PMRM and CLMRM have been omitted from A1-A3; MM has been omitted from all.

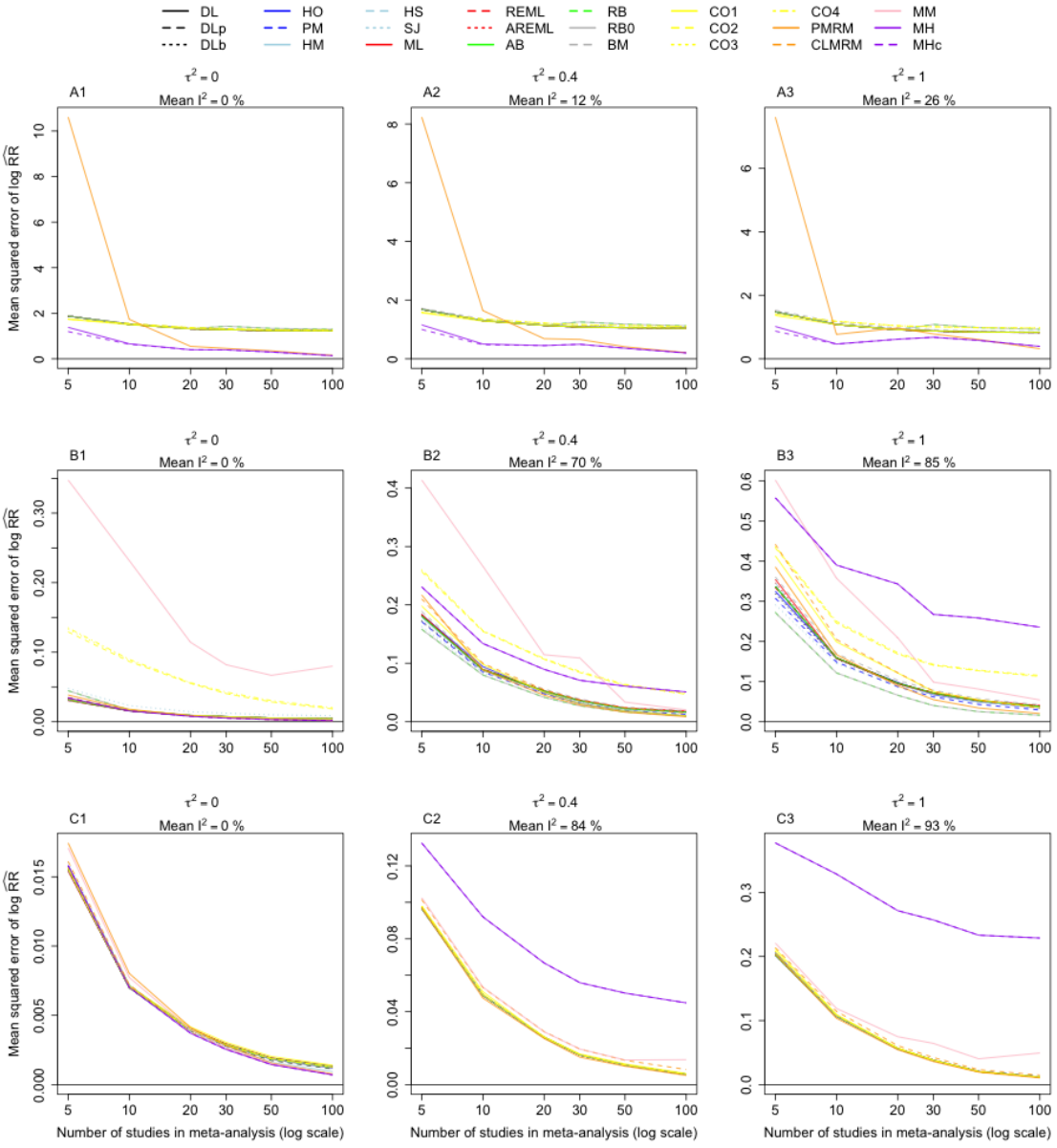


FIGURE 8.10: Mean squared error of log-risk ratio estimates in rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). CLMRM and MM have been omitted from A1-A3.

In the case of common events, when $p_0 < p_1$ our CLMRM method performs more reasonably when sample sizes are small, and is actually one the best for $k > 20$, along with PMRM and the MH-based methods. For unbalanced sample sizes, MM performs much worse and is not included for this reason. However, the fixed-effect MH-based methods consistently have the smallest MSE for all τ^2 with unbalanced and large studies, a very different result to that seen with rare data. When $p_0 > p_1$, the GLMM and MH-based approaches have the lowest MSE for small studies and $k > 10$ ($k > 20$ for CLMRM). MM appears to perform very poorly in all scenarios and is omitted from the plots. The results displayed here for unbalanced and large studies mirror those generated using rare data, however, as the fixed-effect MH-based approaches consistently have a much higher MSE when heterogeneity is present.

The results generated via alternate sampling in the simulation study are also given in Section E.8, and these mirror the MSE-based results discussed here.

8.7 Performance when paired with confidence intervals for θ

As an additional aspect of our simulation study, we investigated how various confidence interval methods for the summary effect θ performed when combined with our τ^2 estimators of interest. The confidence intervals that we considered were Wald-type, t-distribution, Hartung-Knapp-Sidik-Jonkman (HKSJ) and modified Knapp-Hartung (mKH). These are described in detail in Section 1.6. As with the θ estimation in the previous section above, we have included the fixed-effect MH (and continuity corrected MHc) approaches when looking at the performance of the confidence intervals, as these are based on the value of θ itself.

8.7.1 Coverage

We generated 95% confidence intervals for the summary effect θ using the above listed methods, and so an optimum interval would have a corresponding coverage of 95%. Here we shall investigate the coverage of each of our combinations of τ^2 estimator and confidence interval method. Figure 8.11 shows the coverage of our log-risk ratio confidence intervals for very rare events (with $p_0 < p_1$) and medium sample sizes. Each row of plots corresponds to a unique confidence interval method. We can see that for each of the interval methods, the majority of estimators perform fairly well when homogeneity is present, with their coverage remaining fairly constant around 95. The exceptions to this are the semi-Bayesian and conditional-based methods, whose coverage rapidly decreases with increasing k for $\tau^2 = 0$. However, in heterogenous cases, all estimators demonstrate this pattern of decreasing coverage away from the optimal 95 as k increases. In all cases,

however, the best estimators appear to be the PMRM and MH-based approaches, which work well with all interval methods. All estimators be seen to perform more poorly with the HKSJ method though, as coverages never reach 95, not even for small k.

The coverage in the case of medium-sized studies and rare events can be seen in Figure 8.12. Here, we can see that the events having increased in number has resulted in the coverage being far more optimal, near the 95 level, for at least some estimators in all scenarios considered. As before, there are some estimators where the coverage drops as k increases, and these poor performers are the pre-existing methods when $\tau^2 = 0$, and the CLMRM, MM and MH-based methods for $\tau^2 > 0$. As a result, the estimators with the best coverage appear to change depending on whether heterogeneity is present or not. In this case, all confidence interval methods appear to perform very similarly again, with the HKSJ method actually performing best with the appropriate estimators for $\tau^2 > 0$.

We also looked at the coverage for meta-analyses with unbalanced study sample sizes, and these results can be seen for very rare and rare events in Figures 8.13 and 8.14 respectively. For rare events, Figure 8.13 shows that the pre-existing estimators appear to perform poorly in all scenarios, with their respective coverage moving away from the optimal 95 level as k increases. When $\tau^2 = 0$, the GLMM and MH-based approaches consistently have a coverage close to 95, with the MM method also having a coverage just below this level. Meanwhile, in heterogeneous cases, PMRM produces a coverage close to 95 when it can be applied ($k > 20$ here), which is far greater than the coverage of any other estimator for high k. As with the previous very rare event scenario discussed above, all interval methods perform similarly, however the HKSJ appears to perform slightly worse, particularly with the optimum estimators.

Figure 8.14 shows that the coverage of our estimators with rare events and unbalanced samples sizes differs significantly from that for very rare events above. The majority of pre-existing estimators have a constant near-95 coverage when $\tau^2 > 0$, with the exceptions being RB, RB0 and HO, which move away from 95 as k increases for the Wald-type and t-distribution methods. CO2-CO4 and the MH-based approaches also perform very poorly in these cases, and for the HKSJ and mKH methods (although the conditional-based methods perform well with mKH). In the case where $\tau^2 = 0$, all pre-existing estimators appear to perform poorly, again moving away from 95 coverage as k increases. The GLMM and MH-based approaches appear to perform best in this scenario, with MM also performing rather well with near-95 coverage. All estimators tend to have coverage greater than 95 in this particular scenario with small k, for all interval methods except HKSJ. In terms of the interval methods themselves, they all perform very similar again in terms of coverage, although mKH may appear to slightly outperform the others.

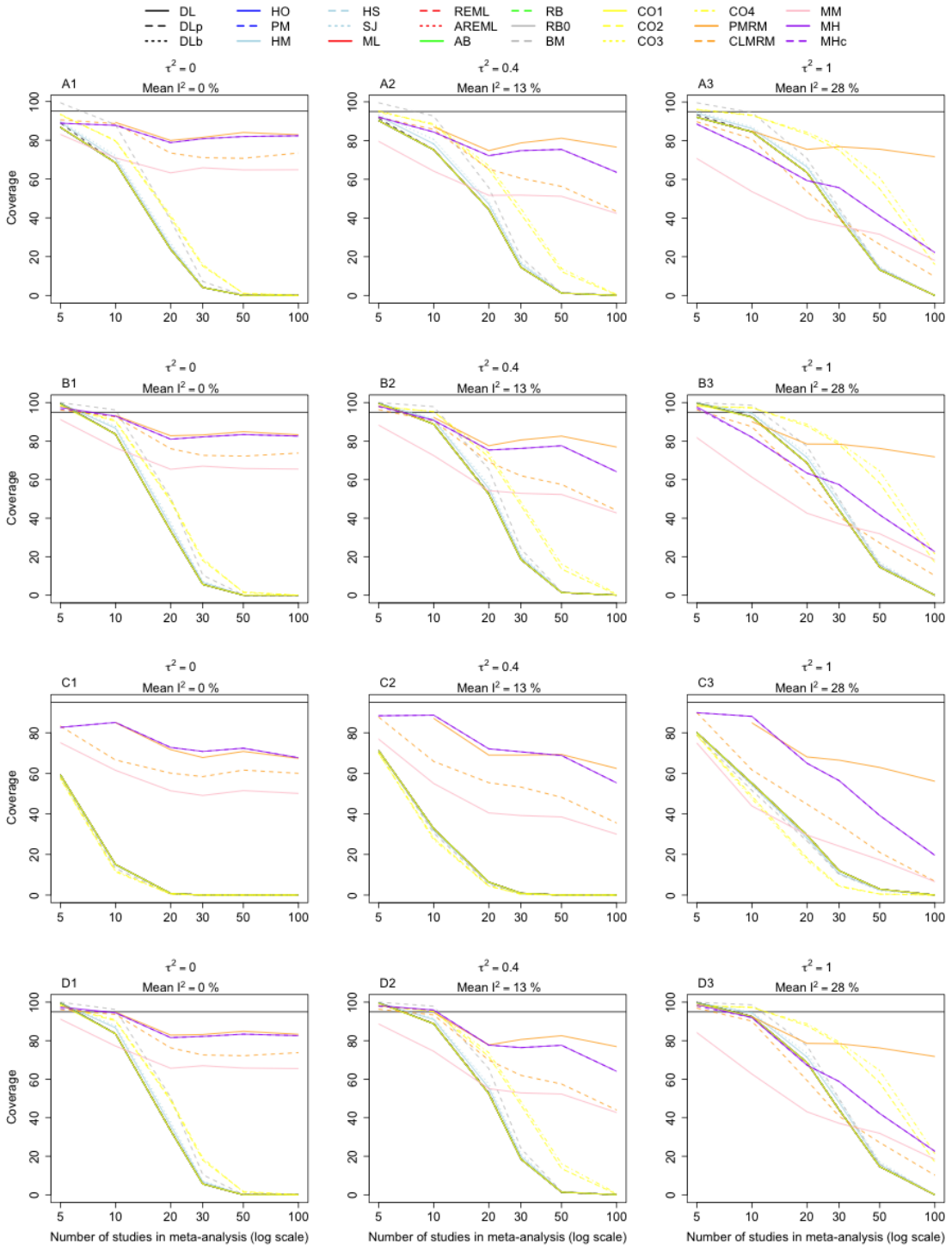


FIGURE 8.11: Coverage of log-risk ratio confidence intervals in very rare events scenario with $p_0 < p_1$ and medium sample sizes; confidence intervals are Wald-type (A1-A3), t-distribution (B1-B3), HKSJ (C1-C3) and mKH (D1-D3).

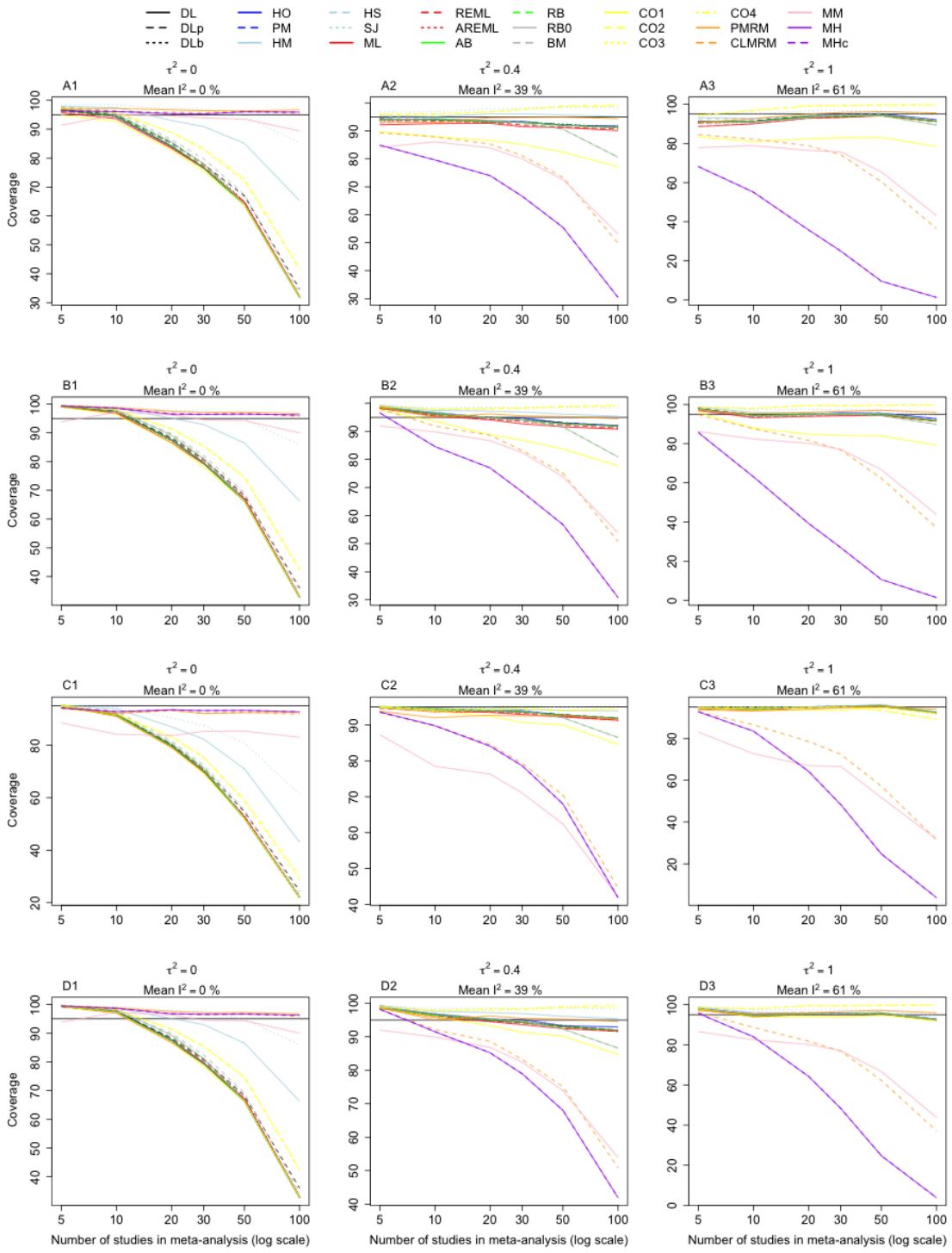


FIGURE 8.12: Coverage of log-risk ratio confidence intervals in rare events scenario with $p_0 < p_1$ and medium sample sizes; confidence intervals are Wald-type (A1-A3), t-distribution (B1-B3), HKSJ (C1-C3) and mKH (D1-D3).

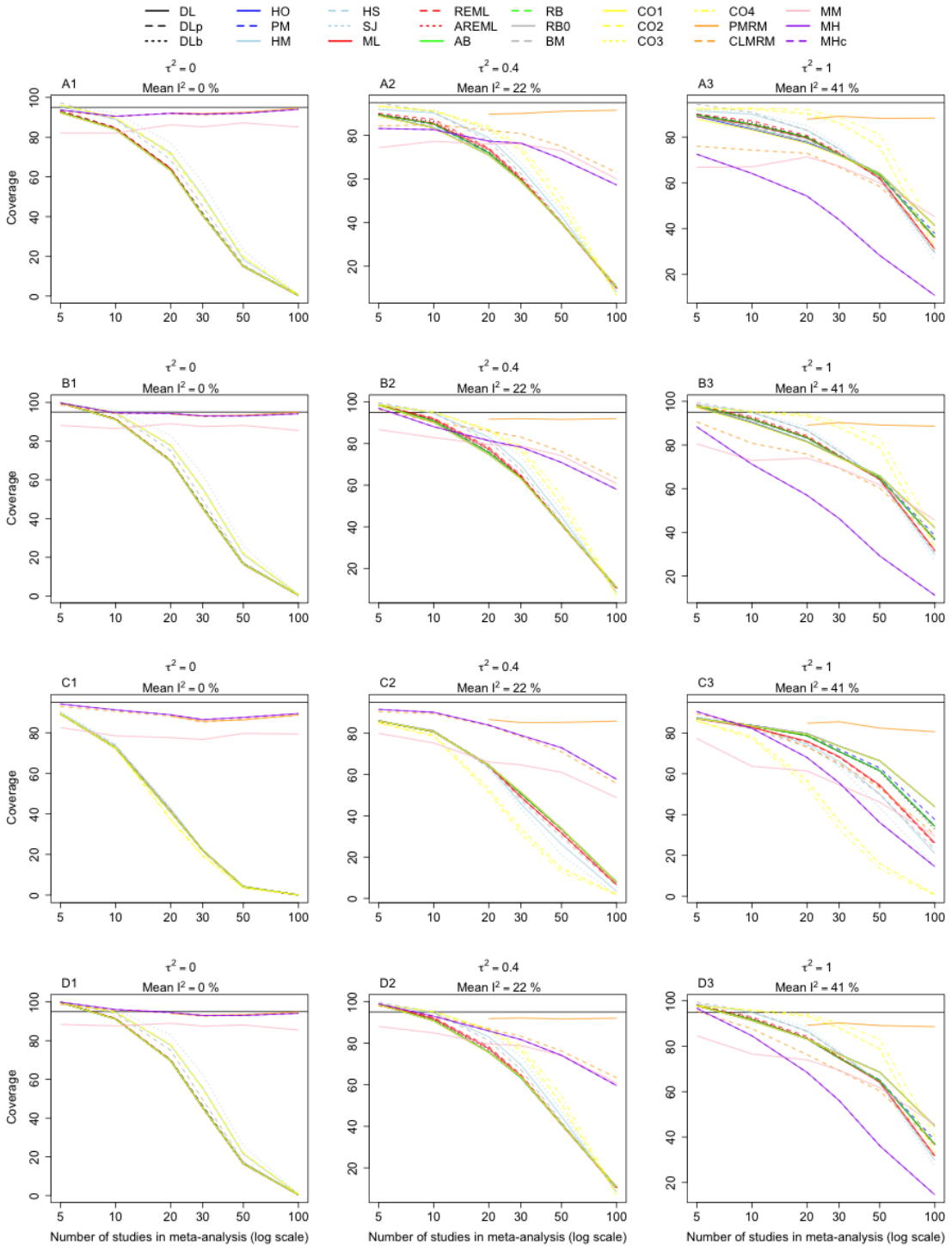


FIGURE 8.13: Coverage of log-risk ratio confidence intervals in very rare events scenario with $p_0 < p_1$ and small and large sample sizes; confidence intervals are Wald-type (A1-A3), t-distribution (B1-B3), HKSJ (C1-C3) and mKH (D1-D3).

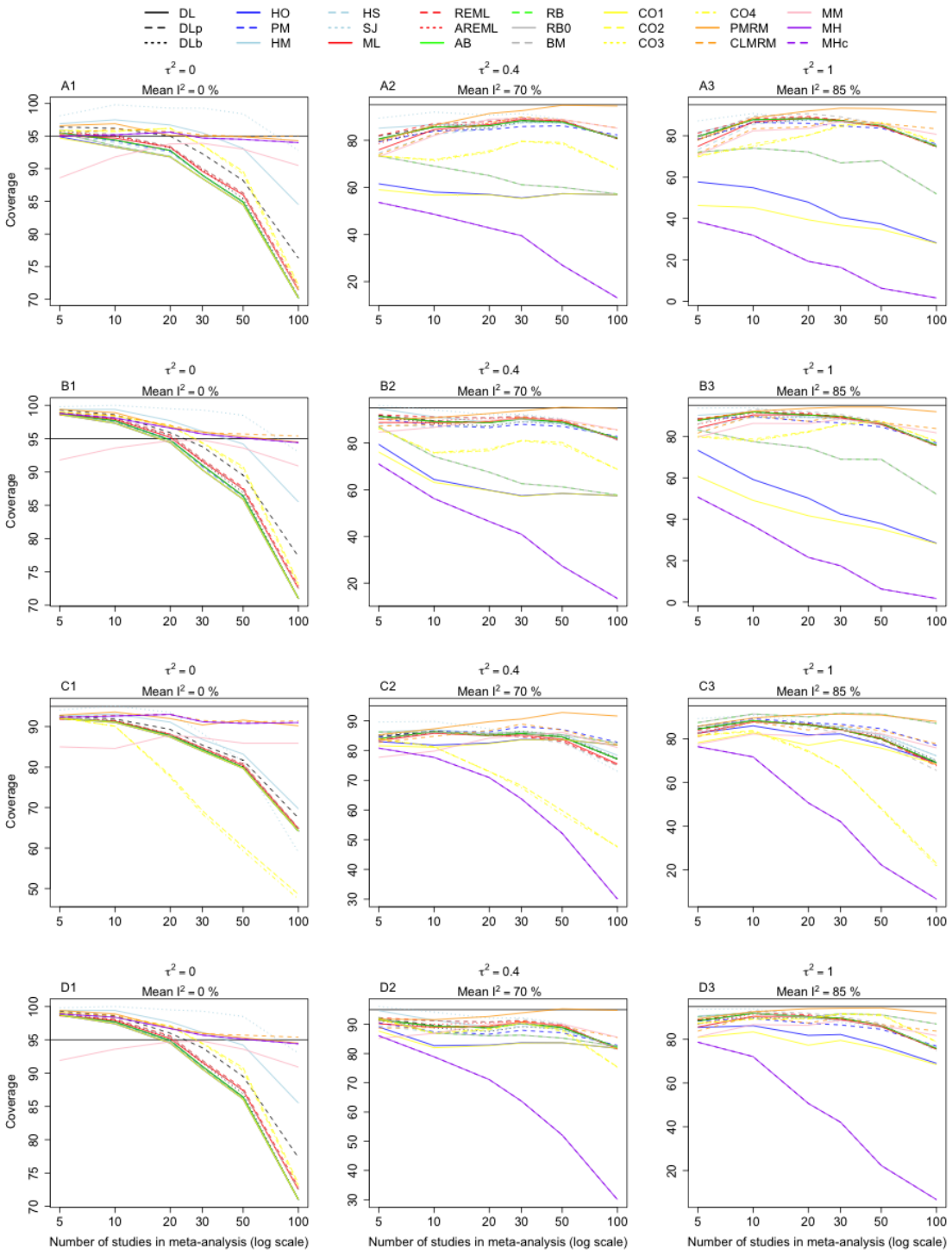


FIGURE 8.14: Coverage of log-risk ratio confidence intervals in rare events scenario with $p_0 < p_1$ and small and large sample sizes; confidence intervals are Wald-type (A1-A3), t-distribution (B1-B3), HKSJ (C1-C3) and mKH (D1-D3).

We also generated results in terms of the coverage for the alternate sample sizes (small, small-to-medium and large) for very are and rare events, and for medium and unbalanced sample sizes with common events. These can be seen in Section E.9. These results fit with the ones that we have discussed here, showing no unexpected outcomes given those already presented.

8.7.2 Power and error

Finally, we assessed the confidence interval methods in terms of their power and error, and the results corresponding to these are given in Sections E.10 and E.11 respectively. In terms of the power, when events are very rare and studies are medium-sized, the majority of estimators have an optimum power of 100 when $k > 10$. However, the CLMRM and MM approaches appear to have much lower power with all confidence interval methods included. CO1-CO4 and BM also have lower power for certain interval methods, with the HKSJ appearing to work best with the majority of estimators. Although if events are very rare, MM consistently has low power for all k in all cases, with CO2-CO4 also having low power for smaller k for all methods other than HKSJ. This result also holds for the scenarios with unbalanced sample sizes, as MM is found to be poorly powered in all scenarios, the majority of estimators have near optimal power for $k > 10$, and HKSJ appears to have the most power in general.

To measure the error of the confidence interval methods, we looked at the mean and variance of the error. The mean error was found to be very high for the CLMRM and MM methods for the case of very rare events and medium-sized studies. Meanwhile, all other estimators had a low mean error (< 2), with the best results again generated from the HKSJ method. When events are rare, the CLMRM method has mean error comparable to the other methods, while MM still has very high error and HKSJ still results in the lowest error of the four interval methods. This pattern is also observed when events are very rare with unbalanced studies. However, when unbalanced studies are coupled with rare events, the MM method is found to produce a low error similar to the other estimators when using the HKSJ method, although it remains an outlier for the alternate interval methods, along with CO1-CO4 in this case.

In terms of the variance of the error of these confidence intervals, it can be seen to be very high for the MM method (and in some cases the CLMRM and conditional-based approaches) for both rare and very rare scenarios, regardless of sample size. Additionally, the HKSJ was found to consistently produce the lowest variance in all these scenarios, agreeing with the results discussed above.

8.8 Conclusions

By following the protocol detailed in Chapter 7, we were able to successfully conduct our simulation study and extract the results. However, during the course of the study, we made some amendments from this protocol that were deemed either beneficial or essential to its completion. For example, we decided to round very small τ^2 estimates in order to account for those methods that have a tendency to produce very exact estimates. We also found it necessary to exclude several scenarios from our simulation study, as some estimators (particularly our novel approaches) were found to be incompatible with certain scenarios, e.g. when $k < 5$, despite modifying the code to apply our GLMM-based methods to make them maximally applicable. While conducting our simulation study, we also noted some of its characteristics in order to generate an overall picture of the meta-analyses simulated so that they could later be compared to empirical cases. We found that the percentage of single and double-zero trials increased as the rarity of event probability also increased, regardless of whether $p_0 < p_1$ or $p_0 > p_1$.

In addition to looking at the performance measures in Section 7.6, we also investigated the efficiency of the estimators, by counting the number of cases where the iterative or model-based methods did not generate a result due to lack of convergence. The pre-existing PM estimator was found to most efficient in this subgroup, successfully converging for all simulated meta-analyses in all scenarios. In terms of our proposed methods, PMRM also successfully converged 100% of the time when events were rare or common, with CLMRM having similar success for most scenarios when events were common. However, our CLMRM approach could be seen to outperform PMRM in terms of convergence when events were very rare as long as study sample sizes were not consistently small. Meanwhile, our MM method had minimal cases of non-convergence in all cases other than when very rare events were coupled with small sample sizes. We also noted a number of specific cases where our GLMM-based methods could not be applied, that had not already been listed elsewhere.

Finally, we looked at the performance of all methods in estimating the parameters of interest for very rare and rare events, focusing on the case with $p_0 < p_1$. In terms of estimating τ^2, our PMRM method was found to have the least bias when $k > 20$ and $\tau^2 > 0$, with CLMRM generating very similar results when sample sizes were not small. While these methods, along with the majority of pre-existing estimators, performed poorly in terms of bias when homogeneity was present, our CO1 approach consistently had near-zero bias here. In terms of MSE, PMRM again performs very well when $k > 30$, with CLMRM also producing one of the lowest MSEs when $k > 50$, however both perform extremely poorly when studies are small (or $\tau^2 = 0$ with very rare events). However, CLMRM was ranked third in terms of MSE for unbalanced sample sizes when events were common. Finally, when looking at the proportion of zero τ^2 estimates produced, we found that CO1 generated nearly 100% zero estimates in homogeneous

cases, while PMRM had constant near-zero percentages when heterogeneity was present and studies were not small (with CLMRM also performing well in these cases for high k).

In terms of estimating the summary log-risk ratio (θ), we found the PMRM to have the least bias in all cases where it could be applied (apart from the combined scenario of small sample sizes and very rare events). CLMRM also performed rather well in these cases, along MM and CO1 when events were rare and $\tau^2 > 0$. The fixed-effect MH-based methods had some of the lowest biases when homogeneity was present. For MSE, PMRM again performed very well in all cases where it could be applied and $k > 10$ with small studies (although not paired with very rare events again). CLMRM and CO1 also produced very low MSEs in the majority of these scenarios, with the fixed-effect approaches again generating some of the best results for homogeneous cases.

Finally, we also looked at how the estimators performed when paired with various confidence interval methods for the summary log-risk ratio. In terms of coverage, PMRM always produced near-optimal results in all cases when it could be applied, regardless of event rarity or sample size balance. Our CLMRM and MM methods performed similarly well when homogeneity was present, and in the case of unbalanced sample sizes. The conditional-based methods, however, only performed well when events were common, sample sizes were fixed and $\tau^2 > 0$. We compared four methods to calculate the confidence intervals themselves, and found that they all behaved very similarly in terms of coverage, although HKSJ may have produced slightly poorer coverage with all estimators. We also discussed the power and error of these methods, and found that the HKSJ method had the greatest power and least error in general. For each of the performance measures investigated, we found no identifiable general patterns in the results obtained. This is because the varying estimator types appeared to perform very differently regardless of the scenario setting that was simulated.

Using the results presented in this chapter and Appendix E, we shall now be able to discuss what these results mean in terms of the estimators themselves and their associated methodology. This will allow us to generate summaries of our novel approaches, determining in what scenarios they outperform existing estimators and should thus be preferred, as well as when they may need to be avoided. We will also be able to compare our results to those generated in previous simulation studies, allowing us to confirm whether they are in agreement and so our results can be deemed as reliable. Finally, these results will be used to generate guidelines as to which estimator to use in specific scenarios, so that the appropriate methods could be determined given a particular meta-analysis dataset. All of the above can also be applied to the pairings of τ^2 estimators with confidence intervals for the summary effect size (an important element of the output for any meta-analysis).

Chapter 9

Discussion and conclusions

9.1 Introduction

In this thesis, we explored the methodology used to estimate the heterogeneity variance (τ^2) in the case of rare-event meta-analyses. This parameter is a key component in the estimation of the summary effect measure in random-effects meta-analyses - the preferred choice when studies are heterogeneous, which is more likely in the case of few events and/or studies. Several methods have previously been proposed to estimate τ^2, including the DerSimonian-Laird estimator (the default choice in many statistical software packages). However, they have been shown to perform very poorly in the case of sparse events, producing inaccurate meta-analysis conclusions, thus calling for more appropriate methodology to take their place.

We have proposed a number of novel approaches to estimate τ^2, and also the summary effect size (in our case the log-risk ratio), which we believe are appropriate for the case of rare-event data. These methods are based on the use of generalised linear mixed models (GLMMs), mixture models (MM) and another approach suggested elsewhere previously (Böhning and Sarol (2000)). In terms of the GLMM-based approaches, we looked at Poisson mixed regression models (PMRM) and conditional logistic mixed regression models (CLMRM), and applied them using an R package with options chosen to maximise convergence success and overall accuracy for our scenarios of interest.

In order to compare our proposed methods to existing τ^2 estimators, we conducted a simulation study, where we varied a range of parameters when simulating our meta-analyses in order to create a diverse range of scenarios. These parameters included the number of studies, number of participants and true heterogeneity variance. We also used differing distributions to sample non-constant parameter values (e.g. unbalanced study sample sizes), in order to determine whether this has any effect on our results as

well as generating the greatest range of realistic simulations. With our simulated meta-analyses, we were then able to apply both our novel and the pre-existing τ^2 estimators, and calculate performance measures such as the bias and mean squared error (MSE).

In Chapter 8, we summarised the results of our simulation study in terms of these performance measures, focusing on the ability of our novel approaches to estimate τ^2 as well as the summary log-risk ratio (θ). We found that our two GLMM-based approaches performed well in terms of bias and MSE when study sample sizes were not small, heterogeneity was present and for moderately high numbers of studies ($k > 30$), for both τ^2 and θ estimates. They also generated appropriately low numbers of zero τ^2 estimates when heterogeneity was present. Our MM approach appeared to perform reasonably well in terms of these measures when studies were large but events were not extremely rare, while our conditional-based methods varied in performance depending on the specific estimating equations used in their application, with them generally having optimum results for large studies with high event frequency.

We also compared four different methods for calculating confidence intervals for the summary effect measure, looking at how these performed in terms of coverage and power when paired with the τ^2 estimators of interest. From looking at the results relating to these confidence intervals, we found that the Hartung-Knapp-Sidik-Jonkman (HKSJ) method had the lowest coverage of the four approaches considered, but performed best in terms of power and error. In terms of τ^2 estimators, our PMRM approach resulted in the best coverage in the majority of scenarios investigated.

In this chapter, we shall discuss our simulation study results presented in Chapter 8 and Appendix E, summarising what these mean in terms of overall performance of each of our novel approaches. As well as identifying scenarios where our novel methods outperform the pre-existing estimators, we shall also note those cases where the existing estimators remain optimal, comparing these cases back to the results from previously published simulation studies. We will then propose how estimator performance may relate to the corresponding methodology, allowing us to identify cases where amendments in methodology (or corresponding code application) may improve performance in other areas. This information will allow us to provide recommendations on which τ^2 estimators (and paired confidence intervals) to use in particular scenarios, and thus generate guidelines for others to follow when conducting rare-event meta-analyses.

In addition to this, we shall identify the advantages and disadvantages of our proposed methods and simulation study design, noting where improvements could have been made to improve the accuracy or range of the results. Finally, we shall discuss what our next stages would be in order to further explore this research topic, as well as discuss further methods that could have the potential to be reliable solutions for our problem of interest.

9.2 Checking of methods and code for correctness

Prior to conducting our simulation study, we conducted a series of investigations to ensure the correctness of our proposed τ^2 estimation methods and the R code we designed to apply both them and the pre-existing estimators of interest. This allowed us to develop trust in our R code and confirm that it performed reliably, so as to ensure the reliability and validity of the associated results produced.

Firstly, we first simulated meta-analyses for more common-event benchmark cases, and applied our estimator code to these simulated datasets. We then calculated performance measures for the pre-existing estimators and compared these to the results found in similar simulation studies, e.g. those by Friede et al. (2017a) and Langan (2015), as the results in these studies represent the asymptotics that should be produced in such scenarios. We also applied our code to empirical data with very large study sample sizes, as there would be little concern of random error in these cases, and so the results of the pre-existing approaches should represent those reported elsewhere for similar scenarios. In both cases, when using our code, we produced appropriate results that were similar to those expected.

As empirical datasets were used in the proposal papers for many of the pre-existing estimators, we also chose to apply our code for the corresponding methods to the example datasets provided in these studies. In all of these cases, we successfully generated identical results to those of the proposed estimator as well as any other methods investigated in the respective study, providing further evidence for the correctness of our application of these methods in the form of our code.

Finally, in order to check the validity of our proposed GLMM and MM approaches, we applied the R code that we developed to utilise these models in the rare-event case studies described in Chapter 3. We then applied the same methods to this data using alternative software or packages - for the GLMMs we applied them in the statistical software package STATA, while with the MM approach we used the C.A.MAN R package developed for the original proposed approach. In both cases, estimates of τ^2 were extracted from the model output and compared with those produced from our written R code. By doing this, we found that our written code produced almost identical results to those from alternate software or packages, confirming the correctness of the application of our proposed methods.

After we had clarified and investigated our code and the associated results, we then formed a collaboration with other researchers to work on and assist with the programming element of this simulation study.

9.3 Discussion of results

In Chapter 8 we presented the main results obtained from our simulation study, with additional results from alternate scenarios given in Appendix E. We found that no general patterns were present for any of the various performance measures, as estimator types behaved so differently regardless of simulated scenario. Here we shall discuss what the individual results mean in terms of overall performance, paying particular attention to our novel approaches.

It is worth noting here that, as mentioned in Chapter 8, 'all-zero events' and 'all events' scenarios were not removed during our simulation study. These scenarios represent meta-analyses where all studies were double-zero trials, and the number of events was equal to the sample size for each study, respectively. Both of these scenarios represent trivial cases, and real-life datasets of this kind are unlikely to be studied and form the entirety of a meta-analysis. Despite this fact, we retained these scenarios in our simulation study as they could be used with some of the estimators considered, and represent a potential (although unlikely) scenario for which recommendations can be made (for methodological knowledge, if not real-life applications). As a result, in our simulated scenarios with very rare events, 'all-zero events' meta-analyses may be present and so incorporated into the results, however 'all events' meta-analyses were unlikely to be produced as we did not simulate any extremely common event scenarios.

9.3.1 Performance of GLMM-based approaches

During our simulation study, we found that both of our novel GLMM-based methods performed extremely poorly in terms of bias and MSE when the number of studies (k) in the meta-analysis was less than 5, and thus we chose not to present these scenarios in our results. However, in alternative scenarios, we found them to have great potential when heterogeneity was present in the meta-analysis and $k > 10$.

In particular, PMRM had one of the lowest biases and MSEs for both τ^2 and θ estimates when these conditions were met and studies were balanced and small-to-medium to large, or extremely unbalanced in size. This performance was irrespective of event probability (it performed well for vary rare, rare and common events with either $p_0 < p_1$ or $p_0 > p_1$), variability of baseline risk (σ_α^2) and degree of heterogeneity present (as long as $\tau^2 > 0$). In addition, it produced the most appropriate numbers of zero τ^2 in these scenarios when heterogeneity was present, consistently generating almost no zero estimates, far better than any other estimator considered. These results were also consistent over the varying sampling distributions for event count and sample size used in our simulation study, confirming the results seen here. The inability of this approach to perform well in homogeneous cases is likely the result of its tendency to generate non-zero parameter

estimates, a result confirmed via the percentage of zero estimates reported in Section 8.5.3.

The major drawback of our PMRM approach, however, is that there exist certain scenarios where this method cannot be applied, either naturally due to model construction or as a result of poor convergence. These cases are listed in Sections 4.7 and 8.4.1 respectively. In particular, PMRM should not be applied when $k < 10$, and this is increased to $k < 20$ when sample sizes are severely unbalanced. As we discovered later, both the GLMMs considered perform poorly when $k = 5$ in general, and we recommend against using them in such situations, and so our decision not to apply the PMRM method in the described scenarios is of little importance or significance to our findings. Despite these drawbacks, our PMRM was consistently found to be the optimal approach for the cases outlined above, in terms of estimating both τ^2 and θ parameters.

The CLMRM approach, our proposed GLMM method, appeared to perform very similarly to the PMRM for the cases described above (non-small studies and $\tau^2 > 0$), producing almost identical values of bias in all given scenarios but resulting in slightly higher MSE when the events were extremely rare. It also generated relatively low numbers of zero τ^2 estimates when heterogeneity was present, as desired, but these were not as small as those from the PMRM approach. However, this method can be applied to a larger range of scenarios than the PMRM, since the number of scenarios for which the model cannot be applied to (as a result of model construction or convergence failure) is fewer, as detailed in Sections 4.7 and 8.4.2 respectively. In addition, we found CLMRM to have higher convergence rates than PMRM for data with very rare events (providing the study sample sizes were not all small).

Our CLMRM model is also able to be applied when $k = 5$, although it has extreme bias and MSE in this case when events are very rare. As a result, this method is still affected by some of the same issues described with PMRM above, in that it is likely to fail to converge when the data is too sparse, in particular when events are very rare and either the sample size is small (e.g. less than 50) or few studies are present. When both of these conditions are met together, the model may to fail to converge or will generate extreme unreliable estimates.

A benefit of both of the GLMM methods discussed here is that they can be applied to the 'all-events' scenario, where the number of events is equal to the sample size across all studies in the meta-analysis. Although this represents a trivial case, similar to the 'all-zero events' scenario mentioned previously, which is unlikely to be seen in practice, the majority of pre-existing normal-based estimators included in our simulation study are not able to work with such meta-analyses, demonstrating the methodological (if not practical) advantage of the GLMMs in this case.

The model constrictions that we described above represent obvious limitations associated with applying a regression model to rare-event data, and can become even more

troublesome when the number of studies in the meta-analysis is low ($k < 5$) or the sample sizes are small (e.g. generally less than 50) and the data is rare-event (so contains a number of zero counts). When either of these conditions are met, then the model will fail to converge as a result of over-parameterisation, as there are more parameters in the models than there are outcomes from the datasets being input. As a result, using another type of model family would not fix this problem, as the data for such examples has so many zero counts in comparison to the number of studies that any estimates produced using such a dataset would not be reliable.

When we ranked all the estimators by MSE and presented the top 10 for each probability scenario, our GLMM methods made very few appearances and were largely towards the bottom of the ranking, even for non-small studies, as shown in Section 8.5.2. The reason behind this is because these rankings were based on all combinations of parameters other than sample size, so this would have incorporated all values of k and τ^2 included in our simulation study. As these methods were found to perform poorly for small k and homogeneous scenarios, this would have dramatically reduced their average ranking.

9.3.2 Performance of conditional-based approaches

We looked at four different variations of our conditional-based approach proposed in Chapter 5, each with their own version of the method-required estimating equation for the parameter τ_p^2. We included all of these variations in our simulation study in order to determine if one outperformed the others, and denoted them by CO1-CO4. From our results, we found that CO2-CO4 behaved very similarly to each other in general, while the original method CO1 behaved differently in the majority of cases. All versions were found to perform poorly in general in terms of bias and MSE when events were rare or very rare and heterogeneity was present, with CO2-CO4 generating extreme outlying estimates. In the case of homogeneity however, CO1 had the lowest bias when estimating τ^2, likely the result of the near-100% of zero τ^2 estimates generated in all levels of heterogeneity considered. This inaccuracy of estimating τ^2 in general lead these methods to have poor performance when estimating θ, or a performance similar to that of the pre-existing estimators. As these conditional-based methods, like the pre-existing estimators, make use of the previously criticised inverse-variance approach when estimating the summary log-risk ratio, this result is not that surprising.

However, the CO2-CO4 methods were found to perform far better in terms of both bias and MSE when events were common or the probability scenario was defined by $p_0 \leq p_1$. This is likely the result of their structure of τ_p^2 estimating equation, which is very dependant on the relationship between the event probability in the control arm (p_0) and treatment arm (p_1), and are likely to perform better in all cases when the event probability is high. In particular, CO2-CO4 had the lowest bias when events were common with $p_0 > p_1$ and heterogeneity was present, regardless of study sample size.

However, this was only the case when $k > 10$ with highly unbalanced sample sizes. As before, the alternative sampling of non-constant parameters in the simulation study agreed with these results, backing up their reliability.

We also investigated the performance of these methods in estimating the study-specific τ_p^2, to determine if they had performed well in terms of this in a broader range of cases than above, and thus an amended transformation to τ^2 could be of interest. From our results, we found that in the case of very rare events, CO1 outperformed CO2-CO4 in terms of both bias and MSE when homogeneity was present or studies were small, but the inverse was true otherwise. However, as the event probability increased, CO2-CO4 outperformed CO1 in all cases when heterogeneity was present, regardless of study sample size. In all cases, the bias and MSE of these methods in estimating τ_p^2 can be seen to be very small, indicating that they perform very well when estimating this parameter, and that their inconsistent performance in estimating τ^2 may be the result of an inappropriate transformation. As a result, it is possible that an alternate form of amendment may increase the overall performance of the estimators and so should be sought.

9.3.3 Performance of mixture model approach

Finally, in terms of our novel MM approach proposed in Chapter 6, the results from our simulation study showed that this estimator performed very poorly in the majority of scenarios investigated. In particular, it had extreme bias in all very rare scenarios, only produced results comparable to the other approaches when studies were large in size and events were rare (with $p_0 < p_1$) or common. Although its results were comparable in these cases, they were not remarkable and as such would not be recommended above the others based on bias and MSE. The proportion of zero τ^2 estimates remained high for small sample sizes or homogenous cases, but dropped as τ^2 increased above zero and the studies increased in size and diversity. However, it only performed well with respect to the other approaches for this measure when heterogeneity was present, outperforming even the previously mentioned CLMRM approach in this case.

When estimating the summary log-risk ratio, our MM approach was found to be one of the better performers in terms of bias and MSE when events were common with $p_0 < p_1$, however in all other scenarios it produced some of the worst results in general. This indicates that while it performed moderately well with estimating τ^2, it was not as successful in generating an estimate for θ (the main purpose of a meta-analysis). As a result, this method appeared to perform fairly poorly overall in this simulation study, particularly in the case of rare-event data, our area of interest. As before, results generated from using alternate sampling distributions in this study agreed with those discussed here.

9.3.4 Performance of pre-existing τ^2 estimators

In addition to looking at the performance of our novel approaches, we shall also briefly discuss the results corresponding to the pre-proposed τ^2 estimators, focusing on cases where they remain the optimal methods, as well as comparing our results with those presented in previously published studies. From the results for performance measures presented in Chapter 8, in general the performance of all estimators tends to improve as the number of studies in the meta-analysis (k) increases, a result that we expected to observe. Our results relating to the mean bias and proportion of zero estimates for τ^2 are comparable with those produced by Friede et al. (2017a), and which are summarised in Figures 2.1 and 2.2. We also observed that the estimators performed better in scenarios with balanced and small-to-medium study sample sizes than those with small and large studies. This result agrees with previous knowledge that the majority of pre-proposed τ^2 estimators perform poorly when the size of treatment arms are highly imbalanced within a meta-analysis, a negative association that is only exaggerated by increasing rarity in events.

We found that many of our proposed estimators performed poorly in the case of small sample sizes as a result of convergence difficulties. In terms of the pre-proposed estimators, the semi-Bayesian Bayes Modal (BM) approach appeared to have the least bias in this case, as long as $\tau^2 > 0$ and k was high, i.e. $k > 30$. When events were rare, the non-truncated method-of-moments approaches Hartung-Makambi (HK) and Sidik-Jonkman (SJ) had the least MSE for this sample size scenario, however SJ had both significant bias and MSE when homogeneity was present. This agrees with a result found by Langan (2015), who showed that this estimator produced a consistently high bias in the case of odds-ratio meta-analyses with event probability 0.1 to 0.5.

The method-of-moments Hedges-Olkin (HO) approach performed poorly in terms of both bias and MSE when events were rare and sample sizes unbalanced, as did both versions of the semi-Bayesian Rukhin-Bayes estimators (RB and RB0). These semi-Bayesian approaches also performed very poorly in the case of small sample sizes, consistently producing extreme estimates of τ^2. For all other estimators, we found the MSE decreased as k increased when sample sizes were not balanced and small, dropping to near-zero for most estimators when $k = 100$, as to be expected. When sample sizes were small, however, all estimators appeared to maintain a consistent level of MSE regardless of k, and as such did not approach zero.

In terms of the proportion of zero τ^2 estimates, RB and RB0 again appeared to produce unusual results when sample sizes were small, jumping from all to no zero estimates depending on the value of k. Some estimators appeared to produce consistent results regardless of whether heterogeneity was present or not, which is not preferable. Examples of these include the method-of-moments HM and positive DerSimonian-Laird estimators, which consistently generated almost no zero τ^2 estimates, regardless of the true value of

τ^2. The reason for this with the latter approach is the correction of 0.01 that is added when a zero estimate is produced, making this an impossible outcome.

All of the pre-proposed estimators appeared to perform very similarly when estimating the summary log-risk ratio, in terms of both bias and MSE, and regardless of changes in sample size or k. However the HO, RB and RB0 did appear to have the least bias of all methods considered when events were rare and sample sizes were unbalanced, contradicting the τ^2 estimation results discussed above. We also included the fixed-effect Mantel-Haenszel (MH) approach when estimating the summary effect size, to determine how this performs compared to methods that incorporate τ^2 (i.e. our random-effects methods). We applied both the original version of this method, as well as one that included a continuity correction in the case of double-zero trials (to prevent them being omitted from the approach). In terms of both bias and MSE, both variations of the MH approach outperformed all other estimators in homogeneous cases. This is to be expected, as there is no heterogeneity present that needs to be accounted for, and this approach is designed to work well with rare-event data. However, as the true value of τ^2 increased, the performance of these approaches dropped rapidly, displaying the importance of accounting for heterogeneity when it is present, and the effect it has on the summary effect measure calculated.

In Chapter 8, we also briefly looked at the efficiency of the iterative estimators by measuring the number of cases where they failed to converge in our simulation study. These iterative pre-existing estimators consisted of the Paule-Mandel (PM), maximum likelihood (ML), restricted maximum likelihood (REML) and approximate restricted maximum likelihood (AREML) methods. The PM appeared to be the most efficient of these, converging to a solution in all cases, regardless of sample size or event probability. Meanwhile, while the maximum likelihood-based approaches did have some convergence issues, these were very few in number (with a maximum of 2.45% of non-applicable meta-analyses in the case of very rare events), and the converge rates were very similar across this group of methods.

In this section we have confirmed that the results discussed here appear to closely follow those seen in previous studies, allowing us to develop further trust in the results we obtained in our simulation study in respect to both pre-existing and novel approaches.

9.3.5 Performance of summary-effect confidence intervals

As an additional aspect of out simulation study, we also looked at the performance of various methods in calculating 95% confidence intervals for the summary log-risk ratio, as this would allow us to determine not only the preferred confidence interval, but also the preferred combination of τ^2 estimator and confidence interval. As such,

this would allow us to make recommendations on preferential τ^2 estimators that incorporated all aspects of a meta-analysis. The methods that we chose to investigate were the Wald-type, t-distribution, Hartung-Knapp-Sidik-Jonkman (HKSJ) and modified Hartung-Knapp (mKH) approaches. Each of these methods is discussed in detail in Section 1.6. In our simulation study, we measured these methods according to their coverage, power, mean error and error variance.

Our results show that the coverage of these four methods is better when the sample sizes are balanced and when the events are less rare, as would be expected. The four methods performed very similarly in terms of coverage over all scenarios considered, however the HKSJ had slightly poorer coverage than the others in general. In terms of the τ^2 estimators, our novel PMRM was found to result in the best coverage in all scenarios where it could be applied, as some of the other τ^2 estimators resulted in coverages that moved away from the optimal 95% level as k increased, an unusual observation, but likely a result of the rare-event nature of the data included. Similar to the results discussed above, the coverage generated from the fixed-effect MH-based approaches appeared to reduce dramatically as τ^2 increased. In terms of the power and error, the HKSJ was actually found to perform the best of all those considered, with high power and minimal mean error (and associated variance) for all scenarios investigated. As a result, when deciding on the confidence interval method of choice, a trade-off needs to be made between coverage and power/error.

9.4 Limitations of simulation study

We were successful in conducting our simulation study and meeting all aims of our project, producing reliable results after checking all methods and code for correctness. However, there were obviously some limitations of our study design and elements that would be changed if we were to conduct it again.

The principal limitation that we faced was the constraint on time to conduct further simulations or scenarios, as well as limited memory to store all of the simulation output and results. These issues were of particular significance since our simulations and estimation techniques were both intensive and time-consuming. We used a super-computer in the form of the University of Southampton's IRIDIS High Performance Computing Facility to combat both of these issues, however we still would have liked to investigate further scenarios in order to provide a broader picture of estimator performance. For example, it would have been preferable to increase the range for parameters such as the true log-risk ratio (θ), for which we only investigated three cases ($\theta = -1.6, 0, 1.6$), as well the variation in baseline risk (σ_α^2). We only conducted 1000 simulations for each investigated scenario as well, which could be increased to 5000 or 10,000 to ensure further reliability of the results generated.

The limitation on time and computing resources also restricted the number of estimators that we could investigate. In particular, it may have been of interest to explore the fully Bayesian approach to estimating τ^2 and conducting meta-analyses, which we discussed briefly in Section 2.6.1, but did not include in our simulation study. This is because we had limited knowledge with regards to prior distributions and parameters that may be suitable for our scenario of interest, and so could be used as starting points, and application of such an approach would involve significant additional computational burden for the simulation study.

Some of our proposed approaches were also limited in their application by the complexity and advancement of statistical software that is currently available. After much investigation into the current methods available to apply GLMMs in statistical software packages, we chose to use the R package *lme4* when applying our GLMM-based approaches (PMRM and CLMRM). We were recommended to use this particular package by others in the field because it currently has the largest scope in terms of applicable scenarios, and can be used with the more complicated CLMRM model. However, there were still certain scenarios where the models failed to converge, likely due to the sparseness of the associated data. At the time of conducting our simulation study, we used the most up-to-date version of this package (version 1.1-19), which was released in 2018. We noted a number of differences in terms of scenarios that could be applied between this and the previous version of the package, leading us to believe that future versions may be able to accommodate some of the currently incompatible scenarios. As a result, it would be suggested that the approach is always applied using the most-up-to-date package.

Finally, a trade-off had to be made between including as many estimators as possible and the ease in presenting an overall message. We chose to include a large number of τ^2 estimators as this would provide the most complete picture, and would allow us to compare our novel approaches to a range of pre-existing estimators with differing methodologies. However, we then found it challenging to extract the overall message from the results of the simulation study without generalising in some manner. We tried to overcome this complexity by presenting summaries of particular performance measures (e.g. MSE ranking tables), however we are aware that these only provide a brief idea of the results and in no way summarise the overall performance. As it was difficult to define clear winners, in this chapter we have also listed those estimators that should certainly be avoided in given scenarios according to the results from our study, in order to provide an alternative take on the recommendations.

9.5 Potential future work

9.5.1 Modifications to simulation study design

As mentioned above, we were restricted by time constraints when conducting our simulation study. If we had additional time and resources available to us, there are a number of potential extensions and additional projects that could be conducted to expand on our current work. For example, an additional aspect that would have been of interest to investigate is the time taken for methods to be applied, which would have been of particularly importance for our novel model-based approaches. This could have considerable impact on the overall performance of these methods, as they make take a very long time to converge to a solution in some cases, particularly those involving very rare-event data.

During the course of our study, we found that our novel approaches performed very poorly when few studies were included in the meta-analysis, particularly when $k < 5$. As a result, we chose to omit these scenarios entirely from our simulation study, and thus not present their results. If we had additional time, we could add to our additional study plan by also focusing on these particular scenarios, paying particular attention to methods that may be more suitable (such as Bayesian approaches). In the previous section, we mentioned briefly how we had not had the opportunity to investigate fully Bayesian approaches due to lack of information on potentially appropriate priors. Incorporating such methodology into our study could allow us to retain cases with $k < 5$, where others have found Bayesian methods to perform well (Günhan et al. (2018)), and investigate the performance of varying prior distributions and prior values.

With additional time and resources, we would re-conduct our simulation study, increasing the number of simulations from 1000 to 5000 or 10,000 replications per scenario, in order to improve the reliability of our current results. We would also present further results, as we only presented the most relevant and impactful results obtained in Chapter 8. It would also be of interest to look at further performance measures, such as the power of the τ^2 estimator itself. In the case of the ML and REML estimators, this would be in the form of the likelihood ratio test (LRT), however the Wald test can also be used to generate the estimate and corresponding standard error. For such tests, our null hypothesis would be that $\tau^2 = 0$, with the alternative hypothesis that $\tau^2 > 0$. We could also inspect the correlation between our log-relative risk estimates and their variance estimates, in order to determine if this has any role in the performance of the corresponding τ^2 estimator. Following this, methods with the intention to control for this correlation could be sought for the estimators of interest.

Finally, it would also be of interest to look at alternative effect size measures, as we focused solely on the log-risk ratio due to its simple interpretation and common use in practice with binary data. We could expand our research by also investigating the

performance of modified τ^2 estimators for the log-odds ratio or risk difference outcome measures, to determine whether outcome measure has any impact in the overall performance of the τ^2 estimate in estimating the overall summary effect. We could also expand on this by also looking at continuous meta-analysis data and corresponding outcome measures.

9.5.2 Modifications to novel τ^2 estimators

In terms of our novel GLMM-based approaches, we only investigated using the PMRM and CLMRM models. However, it would be of interest to look at applying additional GLMMs, such as Binomial mixed regression models, to the problem of estimating τ^2 with sparse data, as discussed briefly in Section 4.5. Such models have potential in our scenario of interest, and could be applied using the packages that are regularly being released and updated. We could also look at making further modifications to our two existing GLMM-based approaches, potentially using alternative packages that will undoubtedly be released in the future.

In Chapter 5, we proposed our conditional-based approach with four alternate estimating equations for the method-specific parameter τ_p^2. Based on the results we have generated for these various approaches, it could be of interest to investigate further variations of the estimating equation, as these may be found to perform better overall. In our results we also found our MM-based method to perform poorly in general, so modified approaches to applying this method could be investigated. In particular, the application of the EM algorithm could be altered and alternative methods for the selection of the model of best fit could be used, e.g. LRT, as we only used the Bayesian information criterion (BIC) in our simulation study.

9.5.3 Future publications

Using the R code presented in Appendix D, we will shortly produce a user-friendly R package that can be used to apply all of the τ^2 estimators considered here to a meta-analysis dataset. This will allow researchers to apply the appropriate estimator of their choice to their own datasets. We shall make this package available on GitHub and promote it in future research papers that we will produce and put forward for publication. Such papers include a literature review of all of the pre-proposed τ^2 estimators currently available, along with the results relating to these from our simulation study, outlining their performance in a wide range of rare-event scenarios. A further potential paper involves a description of our proposed novel methods, with particular attention on the GLMMs, and the results from this simulation study demonstrating their performance compared to a selection of the pre-proposed estimators in the case of rare-event data. We are also collaborating with others on a number of related projects.

9.6 Conclusions

The aim of our simulation study was to compare existing τ^2 estimators with our proposed approaches that we believe should be appropriate for the case of rare-event data. These novel approaches include two based on the use of GLMMs - one based on Poisson regression models, with a random effect on the treatment parameter, the other on conditional logistic regression models. The others consisted of several variations of a previously proposed conditional-based approach and an approach based on the use of mixture models. In order to assess these approaches under sparsity, we designed realistic simulations that satisfied our sparsity requirements, varying parameters such as the sample size and number of studies, while also looking at more common events. We focused on binary endpoints only, and were interested in 2×2 contingency tables with many zero events, and single-zero and double-zero trials. The log-risk ratio outcome measure was our summary effect measure of choice as it is easy to interpret and popular in published studies.

In order to develop trust in the results of our simulation study, we investigated the correctness of our novel approaches and the R code used to apply both these and the pre-existing estimators. To do this, we applied our code to empirical datasets that had been used with some of the pre-existing estimators in previously published studies, and compared our results with those presented there. We also applied these pre-existing methods to simulated common-event cases and empirical data with large study sample sizes, as these cases should have little random error and thus represent those results reported elsewhere. In all of the above checks, we found our code to produce results that mirrored those to be expected. In order to ensure the correctness of our novel GLMM and MM methods, we applied them using alternative statistical software packages or packages previously constructed to apply the modifications of the methods, and in both cases found our written R code to generate identical results to those from alternate packages. As such, we were able to fully ensure the correctness of both our R code and novel approaches, thus developing trust in the results we have presented here.

From the results of our simulation study, we found our novel PMRM approach to generally outperform all other estimators in terms of estimating τ^2 and θ when heterogeneity was present and $k > 10$ in all scenarios where study sample sizes were not small, regardless of event probability. However, there are a number of particular scenarios where this method cannot be applied due to either model design or convergence issues, and so the approach has efficiency issues, particularly when events are very rare. Our other GLMM approach, the CLMRM method, performed very similarly to PMRM in the scenarios described above, however it did have much better convergence rates in the case of very rare events (although this was coupled with slightly higher MSE). While this method was applicable in the case of $k = 5$, it was found to have extreme bias and MSE here. Meanwhile, the variations of our conditional-based approach were found to perform poorly

for rare events, but succeeded in outperforming the other estimators when events were common. These methods appeared to be very sensitive to the relationship between p_0 and p_1, as they performed far better in general when $p_0 \leq p_1$. Finally, our novel MM approach was found to perform very poorly in the majority of scenarios considered here, and only managed to generate comparable results when events were common and studies large.

In terms of the pre-existing τ^2 estimators, we generated results in our simulation study that mirrored those seen in previous studies. For example, we found the SJ estimator to consistently suffer from high bias, particularly when $\tau^2 = 0$, a result previously reported elsewhere. In terms of estimating τ^2, the semi-Bayesian RB and RB0 approaches performed very poorly in all cases when sample sizes were small, as they did with rare events and unbalanced sample sizes, together with the method-of-moments HO estimator. We included the fixed-effect MH approach when investigating the performance in estimating the summary log-risk ratio, and found that while this approach performed best when $\tau^2 = 0$, the performance rapidly decreased as τ^2 increased - a result to be expected from a fixed-effect approach. We also looked at the efficiency of the iterative estimators included in our study, and found that the PM estimator did not suffer from any convergence issues at all, while the ML-based methods did, but only to a very small degree.

We also looked at how the τ^2 estimators performed in terms of generating confidence intervals for the summary log-risk ratio, when combined with four alternative methods for calculating these intervals. The PMRM method was consistently found to result in the best coverage when it could be applied, regardless of the method applied. The methods themselves did not differ dramatically in terms of coverage, however the HKSJ method was observed to have a slightly lower coverage in general. In terms of power and error, however, the HKSJ method outperformed the other approaches, resulting in any decisions regarding method choice to be made via a trade-off between the performance measures discussed here.

Although we successfully conducted our simulation study as planned, and confirmed the correctness of the results produced, there are still some limitations to our study. For example, we were restricted on the time and memory available to us, meaning we could only generate 1000 simulations and were restricted to limited parameter ranges in the 2520 scenarios we investigated here. These limitations also prevented us from including some pre-existing approaches, in particular the fully Bayesian approach, where we were further restricted by lack of published knowledge on appropriate prior values and distributions for our scenario of interest. Our proposed GLMM-based methods were also limited in their applicability, and resulting efficiency, by the performance of the R packages used to apply them.

If we had more time and memory available, we would make some modifications to our simulation study in the form of including the number of simulations and scenarios, and investigating methods specific to the problematic scenario of $k = 5$ (such as Bayesian approaches). We could also look at additional performance measures such as the power of the τ^2 estimators and the correlation between the log-risk ratio estimates and their corresponding variance estimates. It could also be of interest to include additional outcome measures in our simulation study, or focus on the case of continuous data. In terms of our novel methods, we could also look at investigating further GLMMs or make modifications to the application of the ones proposed here. For our conditional-based methods, we could investigate the effect of modifying the associated estimating equation further, as well as look at alternative methods for model selection. We plan to make an R package of our code available to enable easy application of all of the methods included in our study, as well as publish guidelines on which estimators to use.

In summary, in our simulation study we found our proposed GLMM-based methods to perform well in the case of rare-event data as well as more common event scenarios, when heterogeneity was present, sample sizes were not small and $k > 20$. However, they are restricted in terms of efficiency, particularly when events are very rare, although CLMRM can be used in more cases than PMRM. Our novel conditional-based methods only performed well in the case of common events, and were found to be sensitive to the relationship between p_0 and p_1, while our MM approach appeared to perform poorly in the majority of scenarios investigated. In terms of the summary effect confidence interval, PMRM consistently resulted in the best coverage, and while the HKSJ method appeared to perform well in terms of calculating the interval itself, trade-offs may need to be made between power and coverage. As a result, we have proposed some novel approaches, in the form of the GLMM-based methods, that have been shown to outperform existing estimators in the case of rare-event data, and so should act as preferential alternatives.

9.7 Recommendations

Based on the results presented in Chapter 8 and Appendix E, we can now make recommendations on which methods to apply in real-life meta-analyses meeting given scenarios. When deciding on the appropriate method, the first step should be to identify the particular characteristics of the meta-analysis dataset of interest. In particular, it is important to obtain an idea of the following:

- Number of studies - easy to determine
- Sample size of the included studies - again easy to determine. Sample sizes should be grouped according to whether they are relatively well balanced or not, and

whether they can be classed as small ($n \approx 20$), medium-sized ($n \approx 200$) or large ($n \geq 1000$).

- Event probability - easy to determine by comparing the number of events to sample size. It is likely that all studies will have relatively similar event probability as they all relate to the same or very similar outcome, and the event probability should also be in one direction in general, i.e. $p_0 < p_1$ or $p_0 > p_1$. This should be classified according whether they are very rare (with an event rate of around 1 in 1000), rare (1 in 100) or common (1 in 10 or higher). They should also be classed as to whether the event probabilities in general follow $p_0 < p_1$ or $p_0 > p_1$.

- Level of heterogeneity - this has to be determined based on initial impressions of the data, and thus is more complicated to predict. However, it should be easy to detect when absolutely no heterogeneity is present, as these studies will all have been conducted using extremely similar groups of participants and conducted in identical manners. If this is not the case, then heterogeneity can be assumed to be present. In addition, as the differences in these aspects increase, then the level of heterogeneity can also be assumed to increase.

Once these aspects of the meta-analysis have been determined, an appropriate τ^2 estimator can be chosen using the following criteria. If events are rare or very rare, sample sizes are not small, $k > 20$ and heterogeneity is suspected to be present, then the PMRM method should be used. However, this depends on the meta-analysis not meeting one of the patterns incompatible with this approach (defining features of these scenarios are listed in Sections 4.7 and 8.4.1). If the meta-analysis does meet the conditions of one of the incompatible scenarios, or if the events are very rare, then the CLMRM approach should instead be used. However, this approach also has its own set of inapplicable scenarios, and if the meta-analyses meets the conditions of these (listed in Sections 4.7 and 8.4.2), then the SJ estimator is recommended as a final choice.

In the above described scenario, but with $5 < k < 20$, the CLMRM approach is the recommended choice, with the SJ estimator again being the alternative option when the former cannot be applied. In cases where sample sizes are small and heterogeneity is suspected, the SJ estimator is again the preferred choice and would be recommended. If heterogeneity is not believed to be present, and so homogeneity is assumed, then the HO estimator should be used when sample sizes are small, while our CO1 approach is the preferred choice for all other sample size scenarios.

When events are classified as being common, CO1 is recommended when $p_0 < p_1$, sample sizes are not small and heterogeneity is suspected. When $p_0 > p_1$ is the case, however, REML should be the method of choice. In both probability scenarios, when sample sizes are small, SJ is the recommended approach. Finally, when homogeneity is assumed for common-event data, the REML approach is again recommended.

9.7.1 Poorly performing τ^2 estimators to be avoided

As well as producing recommendations on the τ^2 estimators to be used in given scenarios, we can also use our simulation study results to list those estimators that should definitely be avoided in future meta-analyses of rare-event data. For example, the pre-existing HO, RB and RB0 methods should be avoided when sample sizes are rare and sample sizes unbalanced. Meanwhile, if there exists strong evidence that heterogeneity is not present, then the SJ, HM and positive-DL estimators are not recommended and should be avoided, with the latter struggling to produce zero τ^2 estimates. As expected, we found the fixed-effect Mantel-Haenszel approach to perform poorly in terms of estimating the summary log-risk ratio when heterogeneity was present, and so (as is standard protocol) this approach should be avoided if this is suspected to be the case.

9.7.2 Guidelines

We shall now present guidelines for the choice of τ^2 estimators in given rare-event scenarios, which researchers can follow to select the most appropriate method given a particular meta-analysis dataset. These guidelines are presented below in Table 9.1. The definitions of the varying study sample sizes and probability scenarios are listed in Section 9.7 for reference.

TABLE 9.1: Recommended τ^2 estimators to use in various meta-analysis scenarios. Estimators that depend on the applicability of the given meta-analysis are ranked as first/second/third choice, etc.

Sample size	**Probability scenario**		
	Very rare	**Rare**	**Common**
Small	$\tau^2 > 0$: SJ, $\tau^2 = 0$: HO		$\tau^2 > 0$: SJ, $\tau^2 = 0$: REML
Small-to-medium	$\tau^2 > 0$ and $k > 20$: PMRM/CLMRM/SJ, $\tau^2 > 0$ and $5 < k < 20$: CLMRM/SJ, $\tau^2 = 0$: CO1		$\tau^2 > 0$ and $p_0 < p_1$: CO1, $\tau^2 > 0$ and $p_0 > p_1$: REML, $\tau^2 = 0$: REML
Medium			
Small and large			
Large			

Appendix A

Proof of heterogeneity variance estimating approaches

A.1 Derivation of method of moments estimator for the heterogeneity variance

Recall that $Var(\hat{\theta}_i) = \sigma_i^2 + \tau^2$. We begin by taking the expected value of the unweighted squared error for a given study i (Kacker (2004)):

$$\begin{aligned} E\left[(\hat{\theta}_i - \hat{\theta})^2\right] &= Var(\hat{\theta}_i - \hat{\theta}) = Var(\hat{\theta}_i) + Var(\hat{\theta}) - 2\,Cov(\hat{\theta}_i, \hat{\theta}) \\ &= (\sigma_i^2 + \tau^2) + \frac{\sum_{i=1}^k w_i^2(\sigma_i^2 + \tau^2)}{(\sum_{i=1}^k w_i)^2} - \frac{2w_i(\sigma_i^2 + \tau^2)}{\sum_{i=1}^k w_i} \end{aligned}$$

The expected value of the generalised Q-statistic, $Q_{MM} = \sum_{i=1}^k w_i(\hat{\theta}_i - \hat{\theta})^2$, is therefore:

$$\begin{aligned} E\left[\sum_{i=1}^k w_i(\hat{\theta}_i - \hat{\theta})^2\right] &= \sum_{i=1}^k w_i E\left[(\hat{\theta}_i - \hat{\theta})^2\right] \\ &= \sum_{i=1}^k w_i(\sigma_i^2 + \tau^2) + \frac{\sum_{i=1}^k w_i^2(\sigma_i^2 + \tau^2)}{\sum_{i=1}^k w_i} - 2\frac{\sum_{i=1}^k w_i^2(\sigma_i^2 + \tau^2)}{\sum_{i=1}^k w_i} \\ &= \sum_{i=1}^k w_i(\sigma_i^2 + \tau^2) - \frac{\sum_{i=1}^k w_i^2(\sigma_i^2 + \tau^2)}{\sum_{i=1}^k w_i} \end{aligned}$$

Equating the expected value to its observed value gives us:

$$\begin{aligned}\sum_{i=1}^{k} w_i(\hat{\theta}_i - \hat{\theta})^2 &= \sum_{i=1}^{k} w_i\sigma_i^2 + \hat{\tau}^2 \sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2\sigma_i^2}{\sum_{i=1}^{k} w_i} - \hat{\tau}^2 \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i} \\ &= \hat{\tau}^2 \left(\sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i} \right) + \sum_{i=1}^{k} w_i\sigma_i^2 - \frac{\sum_{i=1}^{k} w_i^2\sigma_i^2}{\sum_{i=1}^{k} w_i}\end{aligned}$$

Finally, solving this equation for $\hat{\tau}^2$ gives us the method of moments estimator for τ^2:

$$\Rightarrow \hat{\tau}^2_{MM} = \frac{\sum_{i=1}^{k} w_i(\hat{\theta}_i - \hat{\theta})^2 - \left(\sum_{i=1}^{k} w_i\sigma_i^2 - \frac{\sum_{i=1}^{k} w_i^2\sigma_i^2}{\sum_{i=1}^{k} w_i} \right)}{\sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i}}$$

Appendix B

Logic behind generalised linear mixed model approaches

B.1 Proof of conditional logistic mixed regression model approach

Here we shall discuss the idea beyond our choice of the conditional logistic mixed regression model in the estimation of heterogeneity variance for rare-event meta-analysis data. First, we will look at the idea behind the use of this model, using results taken from Ross (2014), and then we shall apply these results to the case of a meta-analysis, in order to show how the heterogeneity variance can be estimated.

B.1.1 Idea for conditional logistic mixed regression model

From Ross (2014), if we let X and Y be independent Poisson random variables with respective means λ_1 and λ_2, i.e. $X \sim Pois(\lambda_1)$ and $Y \sim Pois(\lambda_2)$, then the conditional probability mass function of X given that $X + Y = n$ can be calculated as follows:

$$\begin{aligned} P\{X = k | X + Y = n\} &= \frac{P\{X = k, X + Y = n\}}{P\{X + Y = n\}} \\ &= \frac{P\{X = k, Y = n - k\}}{P\{X + Y = n\}} \\ &= \frac{P\{X = k\}P\{Y = n - k\}}{P\{X + Y = n\}} \end{aligned} \tag{B.1.1}$$

where the last equality follows from the assumed independence of X and Y. Given the properties of the Poisson distribution, the sum of the random variables $X + Y$ also has

a Poisson distribution, with mean $\lambda_1 + \lambda_2$, i.e. $X + Y \sim Pois(\lambda_1 + \lambda_2)$. Using this information, Equation (B.1.1) can be rewritten as:

$$\begin{aligned} P\{X = k|X + Y = n\} &= \frac{e^{-\lambda_1}\lambda_1^k}{k!}\frac{e^{-\lambda_2}\lambda_2^{n-k}}{(n-k)!}\left[\frac{e^{-(\lambda_1+\lambda_2)}(\lambda_1+\lambda_2)^n}{n!}\right]^{-1} \\ &= \frac{n!}{(n-k)!k!}\frac{\lambda_1^k\lambda_2^{n-k}}{(\lambda_1+\lambda_2)^n} \\ &= \binom{n}{k}\left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^k\left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^{n-k} \end{aligned} \tag{B.1.2}$$

From looking at the last line in Equation (B.1.2), it is clear that the conditional distribution of X given that $X + Y = n$ follows a Binomial distribution with parameters n and $\lambda_1/(\lambda_1 + \lambda_2)$, i.e. $X|(X + Y = n) \sim Bin(n, \lambda_1/(\lambda_1 + \lambda_2))$. It follows from this, and the properties of the Binomial distribution, that the conditional expected value of X given that $X + Y = n$ is given by:

$$E\{X|X + Y = n\} = n\frac{\lambda_1}{\lambda_1 + \lambda_2} \tag{B.1.3}$$

B.1.2 Application of conditional logistic mixed regression model to a meta-analysis scenario

Next we shall apply the results from Section B.1.1 to our meta-analysis situation. If we let X be X_{i1} and Y be X_{i0}, where X_{i1} represents the count of events in the treatment arm of study i and X_{i0} is the count of events in the control arm of study i, then $X_i = X_{i1} + X_{i0}$ is the total count of events in study i.

The relative risk, RR (the outcome measure of interest in our meta-analysis scenario), can be defined in terms of the incidence rate in the treatment arm (μ_1) and the incidence rate in the control arm (μ_0) as follows:

$$RR = \mu_1/\mu_0$$

For the treatment arm of study i, the expected count of events is defined as:

$$E(X_{i1}) = \mu_1 P_{i1}$$

where P_{i1} is the person time for the treatment arm in study i. Similarly, the expected count of events in the control arm is given by:

$$E(X_{i0}) = \mu_0 P_{i0}$$

where P_{i0} is the person time for the control arm in study i.

It follows from the properties of expected values that $E(X_{i1} + X_{i0}) = \mu_1 P_{i1} + \mu_0 P_{i0}$. Inputting the above meta-analysis parameters into Equation (B.1.3), we achieve the following conditional expected value of X_{i1} given $X_i = X_{i1} + X_{i0}$, i.e. the conditional expected value of the event count in the treatment arm given the total event count in study i:

$$\begin{aligned} E(X_{i1}|X_i) &= X_i \frac{\mu_1 P_{i1}}{\mu_1 P_{i1} + \mu_0 P_{i0}} \\ &= X_i \frac{RR_i \frac{P_{i1}}{P_{i0}}}{1 + RR_i \frac{P_{i1}}{P_{i0}}} \end{aligned} \tag{B.1.4}$$

where RR_i represents the relative risk in study i. We can see that Equation (B.1.4) depends only on this value of RR_i, the parameter of interest in our meta-analysis scenario.

From the general results obtained in Section B.1.1, we know that the conditional distribution of X_{i1} given that $X_i = X_{i1} + X_{i0}$ has a Binomial distribution with parameters q_i and X_i, i.e. $X_{i1}|X_i \sim Bin(q_i, X_i)$, where q_i is defined as:

$$\begin{aligned} q_i &= \frac{\mu_1 P_{i1}}{\mu_1 P_{i1} + \mu_0 P_{i0}} \\ &= \frac{RR_i \frac{P_{i1}}{P_{i0}}}{1 + RR_i \frac{P_{i1}}{P_{i0}}} \end{aligned} \tag{B.1.5}$$

Now, if we let $RR_i = \exp(\alpha)$, where α is the common log-relative risk, then Equation (B.1.5) can be rewritten as:

$$q_i = \frac{\exp\left[\alpha + \log\left(\frac{P_{i1}}{P_{i0}}\right)\right]}{1 + \exp\left[\alpha + \log\left(\frac{P_{i1}}{P_{i0}}\right)\right]}$$

which can then be rearranged to give:

$$\frac{q_i}{1 - q_i} = \exp\left[\alpha + \log\left(\frac{P_{i1}}{P_{i0}}\right)\right]$$

Taking logarithms, we obtain the conditional logistic model:

$$\log\left(\frac{q_i}{1-q_i}\right) = \alpha + \log\left(\frac{P_{i1}}{P_{i0}}\right) \tag{B.1.6}$$

The estimate of the relative risk, $\widehat{RR}$, can then be obtained through logistic regression using Model B.1.6, which is composed of only an intercept term and an offset term of $\log(P_{i1}/P_{i0})$. The above fixed conditional logistic model can easily be extended to a random-effects model:

$$\log\left(\frac{q_i}{1-q_i}\right) = \alpha_i + \log\left(\frac{P_{i1}}{P_{i0}}\right) \tag{B.1.7}$$

with $\alpha_i \sim N(\alpha, \sigma_\alpha^2)$, where α is the common relative risk across studies and the value σ_α^2 represents the heterogeneity variance estimate. Thus, by applying the conditional logistic mixed regression model given in Equation (B.1.7) to a meta-analysis dataset, an estimate of the heterogeneity variance for that dataset can be determined and extracted from the associated model output.

Appendix C

Proof for mixture model approach

C.1 Proof for case when within-study person times are unequal

In Chapter 6 we proposed a novel method for the estimation of the heterogeneity variance (τ^2) in rare-event meta-analyses, that is based on the use of mixture models. In our initial outline of the method, we assumed that the within-study person times were equal, i.e. $P_{i1} = P_{i0}$ for all $i = 1, \ldots, k$, where P_{i1} and P_{i0} are the person times in the treatment and control groups of study i respectively and k is the number of studies in the meta-analysis. We made this assumption for simplicity, and because we would be setting this restriction when designing and running our own simulation study. Here we shall outline the proof for the case when the within-study person times are unequal, i.e. $P_{i1} \neq P_{i0}$ for at least one $i = 1, \ldots, k$.

C.1.1 EM algorithm

As in Section 6.3, we shall begin by outlining the input for the EM algorithm to be applied, in particular the associated likelihoods. Recall that the mixture model we are considering is defined as:

$$\sum_{j=1}^{J} Bi\left(X_i, \frac{\theta'_j r_i}{\theta'_j r_i + 1}\right) \pi_j = \sum_{j=1}^{J} \begin{pmatrix} X_i \\ X_{i1} \end{pmatrix} \left(\frac{\theta'_j r_i}{\theta'_j r_i + 1}\right)^{X_{i1}} \left(\frac{1}{\theta'_j r_i + 1}\right)^{X_{i0}} \pi_j$$

where X_{i1} and X_{i0} are the count of events in the treatment and control arms of study i respectively, $X_i = X_{i1} + X_{i0}$, θ'_j is the risk ratio for subgroup j, π_j are positive weights that satisfy $\sum_{j=1}^{J} \pi_j = 1$, and r_i is the ratio of person times for study i, i.e. $r_i = P_{i1}/P_{i0}$, for $i = 1, \ldots, k$ and $j = 1, \ldots, J$ where J is the number of subgroups.

The observed likelihood for our mixture model when the within-study person times are not assumed to be equal is given by:

$$L_O = \prod_{i=1}^{k} \left(\sum_{j=1}^{J} Bi\left(X_i, \frac{\theta'_j r_i}{\theta'_j r_i + 1}\right) \pi_j \right)$$

with the observed log-likelihood therefore being

$$l_O = \log L_O = \sum_{i=1}^{k} \log \sum_{j=1}^{J} Bi\left(X_i, \frac{\theta'_j r_i}{\theta'_j r_i + 1}\right) \pi_j$$

The corresponding complete likelihood is then

$$L_C = \prod_{i=1}^{k} \prod_{j=1}^{J} Bi\left(X_i, \frac{\theta'_j r_i}{\theta'_j r_i + 1}\right)^{z_{ij}} \pi_j^{z_{ij}}$$

where z_{ij} denotes the value of Z_j (an indicator variable that takes on the value 1 if X_1 is from the j^{th} subpopulation, and 0 otherwise) for observation X_{i1}. The complete log-likelihood is then given as

$$l_C = \log L_C = \sum_{i=1}^{k} \sum_{j=1}^{J} z_{ij} \log Bi\left(X_i, \frac{\theta'_j r_i}{\theta'_j r_i + 1}\right) + \sum_{i=1}^{k} \sum_{j=1}^{J} z_{ij} \log \pi_j$$

As in the previous case, maximisation dictates the weight estimates to be $\hat{\pi}_j = \sum_{i=1}^{k} z_{ij}/k$ for subgroups $j = 1, \ldots, J$.

C.1.2 E-step

In the E-step of the EM algorithm, we have by Bayes Theorem that

$$\begin{aligned} E(z_{ij}) &= e_{ij} \\ &= \frac{Bi\left(X_i, \frac{\theta'_j r_i}{\theta'_j r_i + 1}\right) \pi_j}{\sum_{j'=1}^{J} Bi\left(X_i, \frac{\theta'_{j'} r_i}{\theta'_{j'} r_i + 1}\right) \pi_{j'}} \end{aligned}$$

which leads to the updated expected complete log-likelihood:

$$E(l_C) = \sum_{i=1}^{k}\sum_{j=1}^{J} e_{ij} \log Bi\left(X_i, \frac{\theta'_j r_i}{\theta'_j r_i + 1}\right) + \sum_{i=1}^{k}\sum_{j=1}^{J} e_{ij} \log \pi_j \tag{C.1.1}$$

C.1.3 M-step

For the M-step, maximising Equation (C.1.1) leads to the solution

$$\begin{aligned}
\pi_j^{(new)} &= \frac{\sum_{i=1}^{k} e_{ij}}{k} \\
&= \frac{\sum_{i=1}^{k} \frac{Bi\left(X_i, \frac{\theta'_j r_i}{\theta'_j r_i + 1}\right)\pi_j}{\sum_{j'=1}^{J} Bi\left(X_i, \frac{\theta'_{j'} r_i}{\theta'_{j'} r_i + 1}\right)\pi_{j'}}}{k} \\
&= \sum_{i=1}^{k} \frac{Bi\left(X_i, \frac{\theta'_j r_i}{\theta'_j r_i + 1}\right)\pi_j}{k\sum_{j'=1}^{J} Bi\left(X_i, \frac{\theta'_{j'} r_i}{\theta'_{j'} r_i + 1}\right)\pi_{j'}}
\end{aligned}$$

Now, the θ'-relevant component of the expected complete log-likelihood given in Equation (C.1.1) can be rewritten as:

$$\begin{aligned}
&\sum_{i=1}^{k}\sum_{j=1}^{J} e_{ij}[X_{i1} \log \theta'_j - X_{i1} \log(\theta'_j r_i + 1) - X_{i0} \log(\theta'_j r_i + 1)] \\
&= \sum_{i=1}^{k}\sum_{j=1}^{J} e_{ij}[X_{i1} \log \theta'_j - X_i \log(\theta'_j r_i + 1)]
\end{aligned}$$

In order to generate the updated estimate $\theta_j^{'(new)}$, we set $\theta'_j = \theta_j^{'(new)}$ and solve for this by differentiating the expected complete log-likelihood with respect to $\theta_j^{'(new)}$, which is equivalent to differentiating the above formula (written in terms of $\theta_j^{'(new)}$) with respect to $\theta_j^{'(new)}$:

$$\frac{\partial}{\partial \theta_j^{'(new)}} E(l_C) = \sum_{i=1}^{k} e_{ij} \left[\frac{X_{i1}}{\theta_j^{'(new)}} - \frac{X_i r_i}{\theta_j^{'(new)} r_i + 1} \right] = 0$$

$$\Leftrightarrow \sum_{i=1}^{k} e_{ij} \frac{X_{i1}}{\theta_j^{'(new)}} = \sum_{i=1}^{k} e_{ij} \frac{X_i r_i}{\theta_j^{'(new)} r_i + 1}$$

$$\Leftrightarrow \theta_j^{'(new)} = \frac{\sum_{i=1}^{k} e_{ij} X_{i1}}{\sum_{i=1}^{k} e_{ij} \frac{X_i r_i}{\theta_j^{'(new)} r_i + 1}}$$

which is an implicit solution that needs to be iterated for $\theta_j^{'(new)}$ in order to generate a solution. This is the major difference between the case where $r_i = 1$, as the computation of $\theta_j^{'(new)}$ (where $j = 1, \ldots, J$) is more complicated and requires an initial estimate in the form of the original estimate of θ'_j. Once the value of $\theta_j^{'(new)}$ had been determined, the EM algorithm would then be applied in the same manner as described in Section 6.3.4, but with these updated estimates.

C.1.4 Conversion of estimates to log-risk ratio

After applying the EM algorithm for the mixture model approach described above, the estimates need to be converted to those corresponding with the log-risk ratio, our outcome measure of interest. We derived the following relationship between the subgroup-specific risk ratios θ'_j and log-risk ratios θ_j:

$$\begin{aligned} q_{ij} &= \frac{e^{\theta_j} r_i}{1 + e^{\theta_j} r_i} \\ \Rightarrow q_{ij} + q_{ij} e^{\theta_j} r_i &= e^{\theta_j} r_i \\ \Rightarrow q_{ij} &= (1 - q_{ij}) e^{\theta_j} r_i \\ \Rightarrow \theta_j &= \log \frac{q_{ij}}{r_i (1 - q_{ij})} \\ \Rightarrow \theta_j &= \log \theta'_j \end{aligned}$$

as expected.

Inputting this estimate into Equations (6.2.6) and (6.2.5) results in the respective overall estimates for the log-risk ratio and associated heterogeneity variance:

$$\widehat{\log RR} = \hat{\bar{\theta}} = \sum_{j=1}^{J} \pi_j \log \theta'_j$$

$$\hat{\tau}^2 = \sum_{j=1}^{J} \hat{\pi}_j \left(\hat{\theta}_j - \hat{\bar{\theta}}\right)^2$$

It should be noted that these estimates are identical to those for the case of equal person times given in Equations (6.3.6) and (6.3.7) respectively, with the only difference being the definition of the input parameter θ'_j. In addition, all of the results for this alternate case can be simplified to those in Section 6.3 by setting r_i equal to 1 - no further difference is present. As a result, the estimates and results given for this case where the person times are unequal reflect those of the more general case, where the within-study person times may or may not be equal.

Appendix D

Simulation study design

D.1 R code for simulation study

```
############################################################
# R code for producing combinations of study parameters #
############################################################

library(plyr)

## Set values for pre-specified parameters and generate dataframe ##
## of all possible combinations of these parameters                ##

## Less rare events and smaller sample sizes

# Set sample size classes (number of participants in each study arm)
Study.sizes <- c("small", "small-to-medium", "medium", "small and large"
    ↪ , "large")

# Set range of values for true heterogeneity variance tau^2
tau2 <- seq(0, 1, by = 0.2)

## Set pairs of probabilities of events for treatment and control groups

# List of variables used and their meanings:
# p0 - probability of event in control arm
# p1 - probability of event in treatment arm

# Pairings are given in the order (first entry = p0, second entry = p1)
```

```
# For the scenario when p0 > p1 (such as that seen in empirical data):
# (0.03, 0.01) - reflecting probabilities seen in the CRBSI dataset

# For the scenario when p0 < p1 (e.g. for adverse reaction in clinical
    ↪ trial):
# (0.04, 0.08) - twice as likely in the treatment group

# Enter paired values of alpha and beta to correspond to above
    ↪ probabilities

# Set fixed effect value for alpha (baseline risk in control group)
alpha <- c(-6.9, -4.6, -2.3, -4.6, -5.3, -3.0, -0.7)
# Set fixed effect value for beta (the mean log-risk ratio)
beta <- c(1.6, 1.6, 1.6, 0, -1.6, -1.6, -1.6)

# Construct the overall paired probability dataset
prob <- cbind(alpha, beta)
# Give each probability scenario an identifying variable
probscenario <- seq(1, nrow(prob), 1)
probc <- data.frame(prob, probscenario) # combine probabilities and
    ↪ identifying variable

# Choose distributions to sample the counts of events from
eventdist <- c("binomial", "poisson")

# Choose distributions to sample the study sample sizes from
sampdist <- c("uniform", "normal", "chisq")

# Set values for the variance of alpha (sigma_alpha^2)
varalpha <- c(0.1, 3)

# Create dataset of all possible combinations of the above parameters
para <- expand.grid(probscenario = probscenario, tau2 = tau2, Study.
    ↪ sizes =
Study.sizes, eventdist = eventdist, sampdist = sampdist, varalpha =
    ↪ varalpha)
combs <- join(probc, para, by = "probscenario")

combs$theta <- combs$beta
```

```
# Save all combinations
write.table(combs, file = "rarecombs")

########################################
# R code for simulating meta-analyses #
########################################

# Read in dataframe of scenarios
allcombs <- read.table("rarecombs")

# Choose number of scenarios to run per array (max. for time is ~10)
n <- 1

# Extract the necessary rows of scenarios
iscen <- as.numeric(Sys.getenv("PBS_ARRAYID"))
imin <- ((iscen-1)*n)+1
imax <- iscen*n
combs <- allcombs[imin:imax, ]

## Set parameters specific to the simulation ##

kvalues <- c(5, 10, 20, 30, 50, 100) # range of k-values (number of
    ↪ studies in meta-analysis)
sims <- 1000 # number of simulations of each scenario to be run e.g.
    ↪ 1000

set.seed(24601) # set seed to save the random numbers generated

# Make variable of the continuity corrections to be applied
# (will be applied in Section 2)
#contc <- c("constant", "reciprocal")
# At least one continuity correction must be given - if contcorr is not
    ↪ defined
# then a constant continuity correction is given as the only correction
    ↪ by default
if (!exists("contc")) {
  contc <- "constant"
}

# Set scenario counter to adjust for number of studies (k) and
    ↪ continuity corrections
```

```
scenmin <- ((iscen-1)*length(kvalues)*length(contc)*n)+1
scenmax <- iscen*length(kvalues)*length(contc)*n
scen <- c(scenmin:scenmax)

# Determine scenarios without cont. corr.'s - for DZ/SZ counter
slength <- length(scen)/length(contc)
scen2 <- c(scenmin:scen[slength])

## List of arguments and their meanings ##

# combs - data frame where each row is a different combination of
# the variables alpha, beta, alpha, tau2, Study.sizes and
# varalpha, and which also contains the simulated values of alphai,
# betai, Event.prob1, Event.prob0 and Thetai

# kvalues - vector of number of studies in meta-analyses

# sims - number of simulations to be conducted for each scenario

# sims2 - number of simulations to be conducted to determine the true
    ↪ value
# of taup2 (for untransformed conditional-based estimators)

## Simulation algorithm code ##

simmeta <- function(combs = combs, kvalues = c(5, 10, 20, 30, 50, 100),
    ↪ sims = 1000, sims2 = 10000) {
  results <- data.frame() # create empty dataframe for results to be
    ↪ stored in
  scenario <- scenmin - 1 # start counter for scenario ID label
  meta <- 0 # start counter for meta-analysis ID label
  for (m in 1:nrow(combs)) {
    varalpha <- combs$varalpha[m] # variance of alpha = sigma_alpha^2
    tau2 <- combs$tau2[m] # true heterogeneity variance (tau^2) value
    alpha <- combs$alpha[m]
    beta <- combs$beta[m]
    theta <- combs$theta[m]
    SampleSize1 <- as.character(combs$Study.sizes[m]) # number of
    ↪ individuals in each study's trial arms (assumed to be equal across
    ↪  treatment and control groups)
    eventdist <- as.character(combs$eventdist[m])
```

```
sampdist <- as.character(combs$sampdist[m])
## Side step - Find true tau^2_p ##
# Simulate theta_i values
thetai <- rnorm(n = sims2, mean = theta, sd = sqrt(tau2))
# Approximate p_i for each simulation
Pi <- exp(thetai)/(1+exp(thetai))
# Mean p
pbar <- (1/sims2)*sum(Pi)
# True tau_p^2
taup2 <- (1/sims2)*sum((Pi-pbar)^2)
for (k in kvalues) {
  scenario <- scenario + 1 # scenario ID label
  for (i in 1:sims) {
    meta <-  meta + 1 # meta-analysis ID label
    SampleSize <- numeric(k) # empty vector for sample size of each
↪ of the k studies
    # Loop for sample size generation (sample sizes assumed to be
↪ equal across treatment and control arms of the same study)
    if (SampleSize1 == "small") {
      SampleSize <- rep(10, times = k)
    }
    if (SampleSize1 == "small-to-medium") {
      if (sampdist == "uniform") {
        SampleSize <- sample(10:200, k, replace = TRUE) # sample
↪ size integers sampled from Uniform (10,200) with replacement
      }
      if (sampdist == "normal") {
        SampleSize <- rnorm(n = k, mean = 105, sd = 105/3)
        while (any(SampleSize < 10)) {
          SampleSize <- rnorm(n = k, mean = 105, sd = 105/3)
        }
      }
      if (sampdist == "chisq") {
        SampleSize <- rchisq(n = k, df = 105)
        while (any(SampleSize < 10)) {
          SampleSize <- rchisq(n = k, df = 105)
        }
      }
    }
    if (SampleSize1 == "medium") {
      SampleSize <- rep(200, times = k)
```

```
    }
    if (SampleSize1 == "small and large") {
      if (k %% 2 == 0) {
        SampleSize[1:(k/2)] <- 10 # half of studies given sample
↪ size of 10
        if (sampdist == "uniform") {
          SampleSize[((k/2)+1):k] <- sample(1000:2000, k/2, replace
↪ = TRUE) # half of studies given sample size from Uniform
↪ (1000,2000)
        }
        if (sampdist == "normal") {
          SampleSize[((k/2)+1):k] <- rnorm(n = k/2, mean = 1500, sd
↪ = 1500/3)
          while (any(SampleSize[((k/2)+1):k] < 10)) {
            SampleSize[((k/2)+1):k] <- rnorm(n = k/2, mean = 1500,
↪ sd = 1500/3)
          }
        }
        if (sampdist == "chisq") {
          SampleSize[((k/2)+1):k] <- rchisq(n = k/2, df = 1500)
          while (any(SampleSize[((k/2)+1):k] < 10)) {
            SampleSize[((k/2)+1):k] <- rchisq(n = k/2, df = 1500)
          }
        }
      } else {
        SampleSize[1:((k-1)/2)] <- 10
        if (sampdist == "uniform") {
          SampleSize[(((k-1)/2)+1):(k-1)] <- sample(1000:2000, (k-1)
↪ /2, replace = TRUE)
        }
        if (sampdist == "normal") {
          SampleSize[(((k-1)/2)+1):(k-1)] <- rnorm(n = (k-1)/2, mean
↪  = 1500, sd = 1500/3)
          while (any(SampleSize[(((k-1)/2)+1):(k-1)] < 10)) {
            SampleSize[(((k-1)/2)+1):(k-1)] <- rnorm(n = (k-1)/2,
↪ mean = 1500, sd = 1500/3)
          }
        }
        if (sampdist == "chisq") {
          SampleSize[(((k-1)/2)+1):(k-1)] <- rchisq(n = (k-1)/2, df
↪ = 1500)
```

```
          while (any(SampleSize[(((k-1)/2)+1):(k-1)] < 10)) {
            SampleSize[(((k-1)/2)+1):(k-1)] <- rchisq(n = (k-1)/2,
↪ df = 1500)
          }
        }
        # If k is odd, one study is given 20 or Uniform (1000,2000)
↪ at random
        if (sample(1:2, 1) == 1) {
          SampleSize[k] <- 10
        } else {
          if (sampdist == "uniform") {
            SampleSize[k] <- sample(1000:2000, 1, replace = TRUE)
          }
          if (sampdist == "normal") {
            SampleSize[k] <- rnorm(n = 1, mean = 1500, sd = 1500/3)
            while (SampleSize[k] < 10) {
              SampleSize[k] <- rnorm(n = 1, mean = 1500, sd = 1500/
↪ 3)
            }
          }
          if (sampdist == "chisq") {
            SampleSize[k] <- rchisq(n = 1, df = 1500)
            while (SampleSize[k] < 10) {
              SampleSize[k] <- rchisq(n = 1, df = 1500)
            }
          }
        }
      }
    }
    if (SampleSize1 == "large") {
      if (sampdist == "uniform") {
        SampleSize <- sample(1000:2000, k, replace = TRUE) # sample
↪ size integers sampled from Uniform (1000,2000) with replacement
      }
      if (sampdist == "normal") {
        SampleSize <- rnorm(n = k, mean = 1500, sd = 1500/3)
        while (any(SampleSize < 10)) {
          SampleSize <- rnorm(n = k, mean = 1500, sd = 1500/3)
        }
      }
      if (sampdist == "chisq") {
```

```
        SampleSize <- rchisq(n = k, df = 1500)
        while (any(SampleSize < 10)) {
          SampleSize <- rchisq(n = k, df = 1500)
        }
      }
    }
    nc <- nt <- round(SampleSize, digits = 0) # save sample size of
↪ treatment/control group
    p0 <- p1 <- rep(2, times = k)
    g <- h <- y <- 1
    while ((g == 1 & h == 1) | y == 1) {
      for (j in 1:k) {
        while (p0[j] < 0 | p0[j] > 1 | p1[j] < 0 | p1[j] > 1) { # in
↪  case the event probability simulated is > 1 (not allowed)
          alphai <- rnorm(n = 1, mean = alpha, sd = sqrt(varalpha))
↪ # re-sample alpha
          betai <- rnorm(n = 1, mean = beta, sd = sqrt(tau2)) # re-
↪ sample beta
          p0[j] <- exp(alphai) # final probability of events in the
↪ control arm
          p1[j] <- exp(alphai + betai) # probability of events in
↪ treatment group
        }
      }
      et <- ec <- nt + 1
      if (eventdist == "binomial") {
        for (j in 1:k) {
          ec[j] <- rbinom(n = 1, size = nc[j], prob = p0[j]) #
↪ produce random vector of number of events in control group
          et[j] <- rbinom(n = 1, size = nt[j], prob = p1[j]) #
↪ produce random vector of number of events in treatment group
        }
      }
      if (eventdist == "poisson") {
        for (j in 1:k) {
          while (ec[j] > nc[j] | et[j] > nt[j]) {
            ec[j] <- rpois(n = 1, lambda = nc[j]*p0[j]) # produce
↪ random vector of number of events in control group
            et[j] <- rpois(n = 1, lambda = nt[j]*p1[j]) # produce
↪ random vector of number of events in treatment group
          }
```

```
            }
          }
          g <- h <- y <- 0
          if (((all(et/nt == et[1]/nt[1])) & (all(ec/nc == ec[1]/nc[1]))
    ↪ ) & (all(ec == 0) | (all(ec == ec[1]) & all(nc == nc[1])))) {
            g <- 1
          }
          if ((sum((ec+et) == 0) >= (k-1)) | all(ec == 0) | all(et == 0)
    ↪  | (all(et[which(!(ec+et) %in% 0)]/ec[which(!(ec+et) %in% 0)] ==
    ↪ et[which(!(ec+et) %in% 0)][1]/ec[which(!(ec+et) %in% 0)][1])) {
            h <- 1
          }
          if (all(ec == 0)) {
            y <- 1
          }
        }
        study <- 1:k # vector used to indentify/label the individual
    ↪ studies
        simulation <- i # variable to identify the simulation number in
    ↪ data
        resultsi <- cbind(scenario, meta, tau2, taup2, p0, p1, alpha,
    ↪ SampleSize, SampleSize1, k, simulation, study, theta, et, ec, nt,
    ↪ nc, varalpha, eventdist, sampdist) # combine results for
    ↪ simulation study i
        results <- rbind(results, resultsi) # combine all results from
    ↪ previous simulations into one dataframe
      }
    }
  }
  return(results) # returns the results of the simulation
  write.table(results, file = "simulation_results") # save the results
    ↪ when the simulation is complete
}

## To conduct simulation ##

# Apply simulation code:
simresults <- simmeta(combs = combs, kvalues = kvalues, sims = sims,
    ↪ sims2 = 10000)

# Change the format of certain variables to be numeric or integers
```

```
simresults[, c("scenario", "meta", "SampleSize", "k", "simulation", "
    ↪ study", "et", "ec", "nt", "nc", "p0", "p1", "alpha", "theta", "
    ↪ varalpha", "tau2", "taup2")] <- sapply(simresults[, c("scenario",
    ↪ "meta", "SampleSize", "k", "simulation", "study", "et", "ec", "nt"
    ↪ , "nc", "p0", "p1", "alpha", "theta", "varalpha", "tau2", "taup2")
    ↪ ], as.character)
simresults[, c("scenario", "meta", "SampleSize", "k", "simulation", "
    ↪ study", "et", "ec", "nt", "nc")] <- sapply(simresults[, c("
    ↪ scenario", "meta", "SampleSize", "k", "simulation", "study", "et",
    ↪  "ec", "nt", "nc")], as.integer)
simresults[, c("p0", "p1", "alpha", "theta", "varalpha", "tau2", "taup2"
    ↪ )] <- sapply(simresults[, c("p0", "p1", "alpha", "theta", "
    ↪ varalpha", "tau2", "taup2")], as.numeric)

# Save all simulation results
write.table(simresults, file = paste("simresults_", iscen, sep = ""))

# Code to count the number of single-zero (SZ) and double-zero trials
# simulated in our meta-analyses by SCENARIO

# Read in dataframe of simulated meta-analyses
allsims <- matrix(NA, nrow = length(scen2), ncol = 6)
allsims <- data.frame(allsims)
colnames(allsims) <- c("scenario", "SZcount", "DZcount", "total", "
    ↪ SZprop", "DZprop")

# Count the number of SZ and DZ trials by scenario
z <- 0
for (i in scen2) {
  z <- z + 1
  simdata <- simresults[simresults$scenario == i, ]
  allsims$scenario[z] <- i
  allsims$DZcount[z] <- sum((simdata$ec+simdata$et) == 0)
  allsims$SZcount[z] <- sum(simdata$ec == 0 | simdata$et == 0) - allsims
    ↪ $DZcount[z]
  allsims$total[z] <- nrow(simdata)
  allsims$SZprop[z] <- allsims$SZcount[z]/allsims$total[z]
  allsims$DZprop[z] <- allsims$DZcount[z]/allsims$total[z]
}

# Combine results and save to output file
```

```
myvars <- names(simresults) %in% c("scenario", "tau2", "alpha", "
    ↪ SampleSize1", "k", "theta", "varalpha", "eventdist", "sampdist")
myvars <- simresults[myvars]
allsims <- merge(unique(myvars), allsims)
write.table(allsims, file = paste("SZDZ_results_", iscen, sep = ""))

################################################
# R code for applying continuity corrections  #
################################################

## List of arguments and their meanings ##

# simresults - results from the simulation study containing event counts
# and all values representative of meta-analysis scenario

# contc - names of contunity corrections to be applied

## PARAMETERS SPECIFIC TO constant
# c - constant to be added for continuity correction and all-event meta-
    ↪ analyses

## PARAMETERS SPECIFIC TO reciprocal
# k - value to be used as the numerator of the continuity corrections

## PARAMETERS SPECIFIC TO empirical
# s - value to be used as an estimate of theta (logRR) in the case of
    ↪ all-
# single-zero studies

## Function for applying different types of continuity corrections ##

ccorr <- function(simresults = simresults, contc = c("constant", "
    ↪ reciprocal", "empirical"), c = 0.5, k = 1, s = 1) {
  if (is.null(contc)) contc <- c("constant", "reciprocal", "empirical")
  # Ensure there is a separate dataset for each correction used
  mydatacorr <- simresults[rep(seq_len(nrow(simresults)), times = length
    ↪ (contc)), ]
  contcorr <- rep(contc, each = nrow(simresults))
  mydatacorr <- cbind(mydatacorr, contcorr)
  # METHOD 1 - Constant continuity correction
  if ("constant" %in% contc) {
```

```
  mydata <- mydatacorr[mydatacorr$contcorr == "constant", ]
  mydata$eccor <- mydata$ec
  mydata$etcor <- mydata$et
  mydata$nccor <- mydata$nc
  mydata$ntcor <- mydata$nt
  for (i in 1:nrow(simresults)) {
    if (mydata$eccor[i] == 0 | mydata$etcor[i] == 0) {
      mydata$eccor[i] <- mydata$eccor[i] + c
      mydata$etcor[i] <- mydata$etcor[i] + c
      mydata$nccor[i] <- mydata$nccor[i] + c
      mydata$ntcor[i] <- mydata$ntcor[i] + c
    }
  }
  mydatacon <- mydata
}
# METHOD 2 - Reciprocal continuity correction
if ("reciprocal" %in% contc) {
  mydata <- mydatacorr[mydatacorr$contcorr == "reciprocal", ]
  mydata$eccor <- mydata$ec
  mydata$etcor <- mydata$et
  mydata$nccor <- mydata$nc
  mydata$ntcor <- mydata$nt
  for (i in 1:nrow(simresults)) {
    if (mydata$eccor[i] == 0 | mydata$etcor[i] == 0) {
      mydata$eccor[i] <- mydata$eccor[i] + (k/mydata$nt[i])
      mydata$etcor[i] <- mydata$etcor[i] + (k/mydata$nc[i])
      mydata$nccor[i] <- mydata$nccor[i] + (k/mydata$nt[i])
      mydata$ntcor[i] <- mydata$ntcor[i] + (k/mydata$nc[i])
    }
  }
  mydatarec <- mydata
}
# METHOD 3 - Empirical continuity correction
if ("empirical" %in% contc) {
  thetaest <- rep(NA, times = nrow(simresults))
  # Set estimates for all studies
  for (i in 1:nrow(simresults)) {
    if (simresults$ec[i] == 0 | simresults$et[i] == 0) {
      metadata <- simresults[simresults$meta == simresults$meta[i], ]
      if (all(metadata$ec == 0 | metadata$et == 0)) {
        thetaest[i] <- s
```

```
      } else {
        for (j in 1:nrow(metadata)) {
          if (metadata$ec[j] != 0 & metadata$et[j] != 0) {
            break
          }
        }
        thetaest[i] <- log((metadata$et[j]/metadata$nt[j])/(metadata$
  ↪ ec[j]/metadata$nc[j]))
      }
    }
  }
  # Calculate values needed for continuity correction
  R <- simresults$nc/simresults$nt # the group ratio imbalance
  kc <- R/(R+thetaest)
  kt <- thetaest/(R+thetaest)
  # Apply continuity corrections
  mydata <- mydatacorr[mydatacorr$contcorr == "empirical", ]
  mydata$eccor <- mydata$ec
  mydata$etcor <- mydata$et
  mydata$nccor <- mydata$nc
  mydata$ntcor <- mydata$nt
  for (i in 1:nrow(simresults)) {
    if (mydata$eccor[i] == 0 | mydata$etcor[i] == 0) {
      mydata$eccor[i] <- mydata$eccor[i] + kc[i]
      mydata$etcor[i] <- mydata$etcor[i] + kt[i]
      mydata$nccor[i] <- mydata$nccor[i] + kc[i]
      mydata$ntcor[i] <- mydata$ntcor[i] + kt[i]
    }
  }
  mydataemp <- mydata
}
# Combine all continuity correction datasets at end
# Need to only combine those DFs that have been created
if ("constant" %in% contc & "reciprocal" %in% contc & !"empirical" %in
  ↪ % contc) {
  mydata <- rbind(mydatacon, mydatarec)
} else if ("constant" %in% contc & "empirical" %in% contc & !"
  ↪ reciprocal" %in% contc) {
  mydata <- rbind(mydatacon, mydataemp)
} else if ("reciprocal" %in% contc & "empirical" %in% contc & !"
  ↪ constant" %in% contc) {
```

```
    mydata <- rbind(mydatarec, mydataemp)
  } else if ("constant" %in% contc & "reciprocal" %in% contc & "
    ↪ empirical" %in% contc) {
    mydata <- rbind(mydatacon, mydatarec)
    mydata <- rbind(mydata, mydataemp)
  }
  # Amend the 'meta' and 'scenario' labels to change over different
    ↪ continuity corrections
  if (length(contc) > 1) {
    for (i in (nrow(simresults)+1):nrow(mydata)) {
      mydata$meta[i] <- mydata$meta[i]+simresults$meta[nrow(simresults)]
      mydata$scenario[i] <- mydata$scenario[i]+(max(simresults$scenario)
    ↪ -min(simresults$scenario)+1)
    }
  }
  if (length(contc) == 3) {
    for (i in (2*(nrow(simresults))+1):nrow(mydata)) {
      mydata$meta[i] <- mydata$meta[i]+simresults$meta[nrow(simresults)]
      mydata$scenario[i] <- mydata$scenario[i]+(max(simresults$scenario)
    ↪ -min(simresults$scenario)+1)
    }
  }
  # Apply (compulsory) continuity correction for all-event meta-analyses
  for (i in 1:nrow(simresults)) {
    if (mydata$eccor[i] == mydata$nccor[i] & mydata$etcor[i] == mydata$
    ↪ ntcor[i]) {
      mydata$eccor[i] <- mydata$eccor[i] + c
      mydata$etcor[i] <- mydata$etcor[i] + c
      mydata$nccor[i] <- mydata$nccor[i] + (2*c)
      mydata$ntcor[i] <- mydata$ntcor[i] + (2*c)
    }
  }
  # Calculate log-risk ratio (logRR) and its standard error for all
    ↪ studies (using corrected counts)
  mydata$logRR <- log((mydata$etcor/mydata$ntcor)/(mydata$eccor/mydata$
    ↪ nccor))
  mydata$selogRR <- sqrt((1/mydata$etcor)-(1/mydata$ntcor)+(1/mydata$
    ↪ eccor)-(1/mydata$nccor))
  return(mydata)
}
```

```
# Apply continuity correction function
mydata <- ccorr(simresults = simresults, contc = c("constant"), c = 0.5,
    ↪ k = 1, s = 1)

# Save the up-to-date data frame of 'mydata'
write.table(mydata, file = paste("ccresults_", iscen, sep = ""))

################################################
# R code for heterogeneity variance estimators #
################################################

# Read in meta-analysis data produced in simulation study
study <- 1:nrow(mydata) # variable to identify each study

## List of arguments and their meanings ##

# xi - vector of effect estimates for each study. If the outcome is
# risk ratio (for example), we assume that xi is already converted to
# log-risk ratios. log argument can be used to convert output back onto
# the original scale after all heterogeneity estimates have been
# calculated.

# sei - vector of standard errors for each study

# hetest - vector of heterogeneity estimators that you would like to be
# calculated. The default is NULL, which means all estimates are
# calculated.

# signiftau2 - number of significant figures to round tau2 estimates

# maxit - maximum number of iterations allowed where the process of
# estimating tau2 involves iteration

# trunc - TRUE if estimators should be truncated to zero, FALSE
    ↪ otherwise

# output - TRUE if output is displayed, FALSE otherwise (stops too much
# output when we are running the program iteratively)

# tau2prior - starting value of iterative estimators
```

```
## PARAMETERS SPECIFIC TO AB - note that 2 out of 3 are required to
# calculate the estimate:
# eta - shape parameter of the prior distribution
# lambda - spread parameter of the prior distribution
# tau2prior - prior estimate of heterogeneity

## PARAMETERS SPECIFIC TO IPM
# nc - sample size of the control group
# nt - sample size of the treatment group
# ec - number of events in the control group

## PARAMETERS SPECIFIC TO DLp
# DLpos - truncation value as an alternative to zero with the original
# DL estimator

## PARAMETERS SPECIFIC TO DLb
# bsamp - number of bootstrap samples

## List of estimators and their acronyms ##

## Method of moment approach
# HO - Hedges-Olkin
# DL - DerSimonian-Laird
# PM - Paule-Mandel
# HO2 - Two step PM with HO initial estimate
# DL2 - Two step PM with DL initial estimate
# IPM - Improved Paule-Mandel - uses arguments ec, nc and nt
# DLp - Positive DerSimonian-Laird estimate, with truncation at 0.01
# DLb - Bootstrap version of DerSimonian-Laird

# Other approaches
# HM - Hartung-Makambi
# HS - Hunter-Schmidt (original estimator using FE weightings)
# SJ - Sidik-Jonkman
# SJ2 - Alternate Sidik-Jonkman

## Maximum Likelihood approach
# ML - Maximum Likelihood
# REML - Restricted Maximum Likelihood
# AREML - Approximate Restricted Maximum Likelihood
```

```
## Bayesian approaches
# AB - Approximate Bayesian
# RB - Rukhin Bayes with simple prior
# RB0 - Rukhin Bayes with zero prior (with correction for sum(n))
# BM - Bayes Modal

## The heterogeneity variance estimation function ##

hetest <- function(xi = logRR, sei = selogRR, Ntot = NULL, nc = NULL, nt
    ↪  = NULL, ec = NULL, eta = NULL, lambda = NULL, tau2prior = NULL,
    ↪ DLpos = 0.01, bsamp = 5000, hetests = NULL, signiftau2 = 6, maxit
    ↪ = 100, trunc = TRUE, output = TRUE) {
  # if no specific set of estimates is required , calculate them all ...
  if (is.null(hetests)) hetests <- c("HO", "DL", "PM", "IPM", "HO2", "
    ↪ DL2", "DLp", "DLb", "HM", "HS", "SJ", "SJ2", "ML", "REML", "AREML"
    ↪ , "AB", "RB", "RB0", "BM")
  # clear the variables that may have been defined previously when this
  # function was run so that we can start again fresh
  HO_est <- DL_est <- PM_est <- IPM_est <- HO2_est <- DL2_est <- DLp_est
    ↪  <- DLb_est <- HM_est <- HS_est <- SJ_est <- SJ2_est <- ML_est <-
    ↪ REML_est <- AREML_est <- AB_est <- RB_est <- RB0_est <- BM_est <-
    ↪ as.numeric(NA)
  # assume equal sample sizes in arms
  if (!is.null(Ntot) & is.null(nc) & is.null(nt)) {
    Ntot <- nc + nt
  }
  if (!is.null(Ntot) & is.null(nc) & is.null(nt)) {
    nc <- nt <- round(Ntot/2, digits = 0)
  }
  Kest <- length (hetests) # number of estimates to be calculated
  esti <- 1 # a counter so that we can create a dataset with a separate
    ↪ estimate on each row - the first specified estimate will be in row
    ↪  1, ..., etc
  ## Specifying all output vectors before replacing the values with
    ↪ actual estimates
  name <- tau2 <- theta <- rep(NA, times = Kest)
  # theta not needed for output, just for the process of calculating
    ↪ some of the tau2 estimates
  K <- length(xi) # K = number of studies in the meta-analysis
  vi <- sei^2 # variance of each study
  wFEi <- 1/vi # fixed-effects weights
```

```
FEtheta <- sum(xi*wFEi)/sum(wFEi)
# DerSimonian-Laird
if ("DL" %in% hetests | "AB" %in% hetests) {
  name[esti] <- "DL"
  DLw <- 1/vi
  theta[esti] <- sum(xi*DLw)/sum(DLw)
  DLtausq1 <- sum(DLw*((xi-theta[esti])^2)) - (sum(DLw*vi)) + (sum((
 ↪ DLw^2)*vi)/sum(DLw))
  DLtausq2 <- sum(DLw)-(sum(DLw^2)/sum(DLw))
  if (trunc) {
    DL_est <- tau2[esti] <- max(0, DLtausq1/DLtausq2)
  } else {
    DL_est <- tau2[esti] <- DLtausq1/DLtausq2
  }
  esti <- esti + 1
}
# positive DerSimonian-Laird
if ("DLp" %in% hetests) {
  name[esti] <- "DLp"
  DLw <- 1/vi
  theta[esti] <- sum(xi*DLw)/sum(DLw)
  DLtausq1 <- sum(DLw*((xi-theta[esti])^2)) - (sum(DLw*vi)) + (sum((
 ↪ DLw^2)*vi)/sum(DLw))
  DLtausq2 <- sum(DLw) - (sum(DLw^2)/sum(DLw))
  if (trunc) {
    DLp_est <- tau2[esti] <- max(DLpos, DLtausq1/DLtausq2)
  } else {
    DLp_est <- tau2[esti] <- DLtausq1/DLtausq2
  }
  esti <- esti + 1
}
# DerSimonian-Laird bootstrap
if ("DLb" %in% hetests) {
  name[esti] <- "DLb"
  DLw <- 1/vi
  theta[esti] <- sum(xi*DLw)/sum(DLw)
  comb_DLb <- t(replicate(bsamp, sample(1:K, K, replace = TRUE)))
  no_samples <- bsamp
  DLb_est2 <- rep(NA, times = no_samples)
  for (i in 1:no_samples) {
    studycomb <- comb_DLb[i,]
```

```
    theta_b <- sum(xi[studycomb] * (DLw[studycomb])) / sum((DLw[
  ↪ studycomb]))
    DLtausq1_b <- sum(DLw[studycomb] * ((xi[studycomb]-theta_b)^2)) -
  ↪ (sum(DLw[studycomb]*vi[studycomb])) + (sum((DLw[studycomb]^2)*vi[
  ↪ studycomb])/sum(DLw[studycomb]))
    DLtausq2_b <- sum(DLw[studycomb]) - (sum(DLw[studycomb]^2) / sum(
  ↪ DLw[studycomb]))
    if (trunc) {
      DLb_est2[i] <- max(0, DLtausq1_b/DLtausq2_b)
    } else {
      DLb_est2[i] <- DLtausq1_b/DLtausq2_b
    }
  }
  DLb_est <- tau2[esti] <- mean(DLb_est2)
  esti <- esti + 1
}
# Hedges-Olkin
if ("HO" %in% hetests | "ML" %in% hetests | "REML" %in% hetests | "BM"
  ↪  %in% hetests) {
  # To calculate REML, we need a starting value of tau2_ML, or else
  ↪ there may be more than 1 solution.
  HOw <- rep(1/K, times = K)
  theta[esti] <- sum(xi*HOw) / sum(HOw)
  HOtausq1 <- sum(HOw*((xi-theta[esti])^2)) - (sum(HOw*vi)) + (sum((
  ↪ HOw^2)*vi)/sum(HOw))
  HOtausq2 <- sum(HOw) - (sum(HOw^2)/sum(HOw))
  HO_est <- max(0, HOtausq1/HOtausq2)
  if ("HO" %in% hetests) {
    if (trunc) {
      tau2[esti] <- max(0, HOtausq1/HOtausq2)
    } else {
      tau2[esti] <- HOtausq1/HOtausq2
    }
    name[esti] <- "HO"
    esti <- esti + 1
  }
}
# Paule Mandel
if ("PM" %in% hetests) {
  quant <- df <- K - 1 # degrees of freedom and expected mean under
  ↪ the fixed effects assumption
```

```
PMtau2out <- 1 # just set an initial value for PM estimate for
↪ output
if (is.null(tau2prior)) {
  PMtausq <- 0 # initial estimate of tau2
} else {
  PMtausq <- tau2prior
}
PMit <- 1 # iteration number
PM_F <- 1 # just to get the iteration started, PM_F=0 implies
↪ convergence
# If the number of events = number of people in trial arm,
# then the estimator cannot be calculated and NA is produced.
# So one must avoid looping through any NA estimates
while (!is.na(PM_F) & PM_F != 0) {
  # First calculate the pooled effect based on present estimate of
↪ tausq
  PMw <- 1/(sei^2+PMtausq)
  PMyW <- sum(xi*PMw)/sum(PMw)
  # Equation comes from DerSimonian and Kacker 2007
  Q1 <- sum(PMw*(xi-PMyW)^2) # generalised Q-statistic
  Q2 <- sum((PMw^2)*(xi-PMyW)^2) # denominator from delta
  # quant = statistic coming from the chi-squared distribution
↪ regardless of data
  if (trunc) {
    PM_F <- max(Q1 - quant, 0)
  } else {
    PM_F <- Q1 - quant
  }
  delta <- PM_F/Q2 # what to add onto the next tausq estimate
  if (is.na(PM_F)) {
    PMtau2out <- PMtausq <- NA
  } else {
    if (PM_F != 0) {
      PMtausq <- PMtausq + delta
    }
    PMit <- PMit + 1
    if (PM_F == 0) {
      PMtau2out <- PMtausq
    }
    if (PMit == maxit) {
      PM_F <- 0
```

```
        if (output == TRUE)
          cat ("PM estimator: Maximum number of iterations reached
  ↪ without convergence \n")
      }
    }
  }
  name[esti] <- "PM"
  if (PMit == maxit) {
    PM_est <- tau2[esti] <- NA
  } else {
    PM_est <- tau2[esti] <- PMtau2out
    PMw <- 1/(vi+tau2[esti])
  }
  theta[esti] <- sum(xi*PMw)/sum(PMw)
  esti <- esti + 1
}
# Improved Paule Mandel (with improved standard errors)
if("IPM" %in% hetests) {
  quant <- df <- K-1 # degrees of freedom and expected mean under the
  # Fixed-effects assumption
  if (is.null(tau2prior)) {
    IPMtausq <- 0 # initial estimate of tau2
  } else {
    IPMtausq <- tau2prior
  }
  IPMdiff <- 1
  IPMit <- 1 # iteration number
  negcount <- 0 # counter for number of negative estimates
  # Calculations needed to calculate standard errors, but that don't
  ↪ change for each iteration
  oddsc <- log(eccor/(nccor-eccor)) # odds in control group
  thetaH0 <- sum(xi)/K # un-weighted average
  # If the number of events = number of people in trial arm,
  # then the estimator cannot be calculated and NA is produced.
  # So one must avoid looping through any NA estimates
  while (!is.na(IPMdiff) & IPMdiff != 0) {
    IPMtausq_prev <- IPMtausq
    # First calculate the standard errors according to the alternative
  ↪  formula proposed by Bhaumik
    # (depends on tau2 estimate so needs to be calculated for each
  ↪ iteration)
```

```
  sei_IPM <- ((exp(-oddsc-thetaHO+(IPMtausq/2)) + 2 + exp(oddsc+
↪ thetaHO+(IPMtausq/2)))/(nt+1)) + ((exp(-oddsc) + 2 + exp(oddsc))/(
↪ nc+1))
  # Calculate the pooled effect based on present estimate of tausq
  IPMw <- 1/(sei_IPM^2+ IPMtausq)
  IPMyw <- sum(xi*IPMw)/sum(IPMw)
  # Equation comes from DerSimonian and Kacker (2007)
  Q1 <- sum(IPMw*(xi-IPMyw)^2) # generalised Q-statistic
  Q2 <- sum(IPMw*(sei_IPM^2)) - (sum((IPMw^2)*(sei_IPM^2)) / sum(
↪ IPMw))
  Q3 <- sum(IPMw)-(sum(IPMw^2) / sum(IPMw))
  IPMtausq <- (Q1-Q2)/Q3
  if (!is.na(IPMtausq)) {
    if (trunc) {
      if (IPMtausq >= 0) {
        IPMdiff <- round (abs(IPMtausq-IPMtausq_prev), digits =
↪ signiftau2)
      } else {
        negcount <- negcount + 1
        # If iteration is negative more than once then final
↪ estimate IPM = 0
        if (negcount >= 2) {
          IPMdiff <-0
        }
        IPMtausq <-0
      }
    } else {
      IPMdiff <- round (abs(IPMtausq-IPMtausq_prev), digits =
↪ signiftau2)
    }
  }
  IPMit <- IPMit + 1
  if (IPMit == maxit) {
    IPMdiff <- 0
    if (output == TRUE) {
      cat ("IPM estimator: Maximum number of iterations reached
↪ without convergence \n")
    }
  }
}
name[esti] <- "IPM"
```

```
  if (IPMit == maxit) {
    IPM_est <- tau2[esti] <- NA
  } else {
    IPM_est <- tau2[esti] <- IPMtausq
    IPMw <- 1/(vi + tau2[esti])
  }
  theta[esti] <- sum(xi*IPMw) / sum(IPMw)
  esti <- esti + 1
}
# Hedges-Olkin initial estimate with PM weightings
if ("HO2" %in% hetests) {
  name[esti] <- "HO2"
  if (trunc) {
    HOtau2 <- max(0, (1/(K-1)) * sum((xi-(sum(xi)/K))^2) - (1/K)*sum(
  ↪ vi))
  } else {
    HOtau2 <- (1/(K-1))*sum((xi-(sum(xi)/K))^2)-(1/K)*sum(vi)
  }
  HO2w <- 1/(HOtau2+vi)
  theta[esti] <- sum(xi*HO2w)/sum(HO2w)
  HO2wtausq1 <- sum(HO2w*((xi-theta[esti])^2)) - (sum(HO2w*vi)) + (sum
  ↪ ((HO2w^2)*vi)/sum(HO2w))
  HO2wtausq2 <- sum(HO2w) - (sum(HO2w^2)/sum(HO2w))
  if (trunc) {
    HO2_est <- tau2[esti] <- max (0, HO2wtausq1/HO2wtausq2)
  } else {
    HO2_est <- tau2[esti] <- HO2wtausq1/HO2wtausq2
  }
  esti <- esti + 1
}
# DerSimonian-Laird initial estimate with PM weightings
if ("DL2" %in% hetests) {
  name[esti] <- "DL2"
  DLw <- 1/(vi)
  DLtheta <- sum(xi*DLw)/sum(DLw)
  if (trunc) {
    DLtau2 <- max(0, (sum(DLw *((xi-DLtheta)^2)) - K+1) / (sum(DLw) -
  ↪ (sum(DLw^2)/sum(DLw))))
  } else {
    DLtau2 <- (sum(DLw*((xi-DLtheta)^2)) - K+1) / (sum(DLw) - (sum(DLw
  ↪ ^2)/sum(DLw)))
```

```
  }
  DL2w <- 1 / (DLtau2+vi)
  theta[esti] <- sum(xi*DL2w) / sum(DL2w)
  DL2tausq1 <- sum(DL2w*((xi-theta[esti])^2)) - (sum(DL2w*vi)) + (sum
  ↪ ((DL2w^2)*vi)/sum(DL2w))
  DL2tausq2 <- sum(DL2w) - (sum(DL2w^2)/sum(DL2w))
  if (trunc) {
    DL2_est <- tau2[esti] <- max(0, DL2tausq1/DL2tausq2)
  } else {
    DL2_est <- tau2[esti] <- DL2tausq1/DL2tausq2
  }
  esti <- esti + 1
}
# Hartung-Makambi
if ("HM" %in% hetests) {
  name [esti] <- "HM"
  HMQ <- sum((1/vi)*((xi-FEtheta)^2))
  HM_est <- tau2[esti] <- (HMQ^2) / ((2*(K-1)+HMQ) * (sum(1/vi) - (sum
  ↪ ((1/vi)^2)/sum(1/vi))))
  esti <- esti + 1
}
# Hunter-Schmidt (original estimator using FE weightings)
if ("HS" %in% hetests) {
  name[esti] <- "HS"
  if (trunc) {
    HS_est <- tau2[esti] <- max(0, (sum(wFEi*(xi-FEtheta)^2)-K) / (sum
  ↪ (wFEi)))
  } else {
    HS_est <- tau2[esti] <- (sum(wFEi*(xi-FEtheta)^2)-K) / (sum(wFEi))
  }
  esti <- esti + 1
}
# Sidik-Jonkman
if ("SJ" %in% hetests) {
  name [esti] <- "SJ"
  ## Estimate of tau2
  # Calculate the pooled estimate
  SJtheta_0 <- sum(xi)/K
  # Cochran's equally weighted estimate of the pooled result
  SJtau2_0 <- (1/K) * sum((xi-SJtheta_0)^2)
```

```
  # If all estimates are identical then we cannot go any further in
 ↪ the calculation and our estimate is zero
  if (SJtau2_0 > 0) {
    # SJ weightings based on initial estimate of tau2 (SJtau2_0)
    SJw <- 1/((vi/SJtau2_0)+1)
    # Random-effects pooled estimate based on the above weightings
    SJtheta_1 <- sum(xi*SJw) / sum(SJw)
    SJ_est <- tau2[esti] <- (1/(K-1)) * sum(SJw*(xi-SJtheta_1)^2)
  } else {
    SJ_est <- tau2[esti] <- 0
  }
  ## Pooled effect estimate
  SJw2 <- 1/((vi/tau2[esti])+1)
  theta[esti] <- sum(SJw2*xi) / sum(SJw2)
  esti <- esti + 1
}
# Alternate Sidik-Jonkman
if ("SJ2" %in% hetests) {
  name[esti] <- "SJ2"
  ## Esitmate of tau2
  # Calculate the pooled estimate
  SJ2theta_0 <- sum(xi)/K
  # If all estimates are identical then tau2 is zero
  if (sum((xi-SJ2theta_0)^2) > 0) {
    # Variance components method (general form of Hedges-Olkin)
    SJ2tau2_0 <- max(0.01, ((1/(K-1))*(sum((xi-SJ2theta_0)^2)))-((1/K)
 ↪ *(sum(vi))))
    # SJ2 weightings based on initial estimate of tau2 (SJ2tau2_0)
    SJ2w <- 1/((vi/SJ2tau2_0)+1)
    # Random-effects pooled estimate based on the above weightings
    SJtheta_1 <- sum(xi*SJ2w)/sum(SJ2w)
    SJ2_est <- tau2[esti] <- (1/(K-1))*sum(SJ2w*(xi-SJtheta_1)^2)
  } else {
    SJ2_est <- tau2[esti] <- 0
  }
  ## Pooled effect estimate
  SJ2w2 <- 1/((vi/tau2[esti])+1)
  theta[esti] <- sum(SJ2w2*xi)/sum(SJ2w2)
  esti <- esti + 1
}
# Maximum Likelihood
```

```
if ("ML" %in% hetests | "BM" %in% hetests) {
  name[esti] <- "ML"
  # Difference between this iteration and previous to assess when we
 ↪ have convergence
  # set MLdiff != 0 initially to get the process of iteration going
  MLdiff <- 1
  MLit <- 0 # counter for number of iterations
  negcount <- 0 # counter for number of negative estimates
  # First set initial estimate of tau2 and theta (fixed-effect
 ↪ estimates)
  if (is.null(tau2prior)) {
    MLtau2 <- H0_est
  } else {
    MLtau2 <- tau2prior
  }
  MLtheta <- FEtheta
  # If the number of events = number of people in trial arm,
  # then the estimator cannot be calculated and NA is produced.
  # So one must avoid looping through any NA estimates
  while (MLdiff != 0 & !is.na(MLtau2)) {
    # Estimate of between-study heterogeneity
    MLtau2_prev <- MLtau2 # record of previous step
    if (-min(vi) >= MLtau2) {
      MLtau2 <- -min(vi) + (10^-signiftau2)
    }
    MLtau2 <- sum(((xi-MLtheta)^2-vi) / (vi+MLtau2)^2) / sum(1/(vi+
 ↪ MLtau2)^2)
    # Estimate for pooled effect
    MLtheta_prev <- MLtheta # record of previous step
    MLtheta <- sum(xi/(vi+MLtau2)) / sum(1/(vi+MLtau2))
    if (!is.na(MLtau2)) {
      if (trunc) {
        if (MLtau2 >= 0) {
          MLdiff <- round(abs(MLtau2 - MLtau2_prev), digits =
 ↪ signiftau2)
        } else {
          negcount <- negcount + 1
          # If iteration is negative more than once then final
 ↪ estimate ML = 0
          if (negcount >= 2) {
            MLdiff <- 0
```

```
          }
          MLtau2 <- 0
          MLtheta <- sum(xi*wFEi)/sum(wFEi)
        }
      } else {
        MLdiff <- round(abs(MLtau2-MLtau2_prev), digits = signiftau2)
      }
    }
    MLit <- MLit + 1
    if (MLit == maxit) {
      MLdiff <- 0
      if (output == TRUE)
        cat ("ML estimator: Maximum number of iterations reached
  ↪ without convergence \n")
    }
  }
  if (MLit == maxit) {
    ML_est <- tau2[esti] <- NA
  } else {
    ML_est <- tau2[esti] <- MLtau2
  }
  # Pooled effect estimate
  theta[esti] <- sum(xi/(vi+MLtau2)) / sum(1/(vi+MLtau2))
  esti <- esti + 1
}
# Restricted Maximum Likelihood
if ("REML" %in% hetests) {
  name[esti] <- "REML"
  # First set initial estimate of tau2 and theta (fixed-effect
  ↪ estimates)
  if (is.null(tau2prior)) {
    REMLtau2 <- HO_est
  } else {
    REMLtau2 <- tau2prior
  }
  REMLtheta <- FEtheta
  # Difference between this iteration and previous to assess when we
  ↪ have convergence
  # set diff != 0 initially to get the process of iteration going
  REMLdiff <- 1
  REMLit <- 0 # counter for number of iterations
```

```
negcount <- 0 # counter for number of negative estimates
# Process of iteration, stop when there is no difference between the
↪  last two steps.
# If the number of events = number of people in trial arm,
# then the estimator cannot be calculated and NA is produced.
# So one must avoid looping through any NA estimates
while (REMLdiff != 0 & !is.na(REMLtau2)) {
  # Estimate of between study heterogeneity
  REMLtau2_prev <- REMLtau2 # record of previous step
  tau2_p1 <- sum((1/((vi+REMLtau2_prev)^2)) * (((xi-REMLtheta)^2) -
↪ vi))
  tau2_p2 <- sum(1/((vi+REMLtau2_prev)^2))
  tau2_p3 <- sum(1/(vi+REMLtau2_prev))
  REMLtau2 <- (tau2_p1/tau2_p2) + (1/tau2_p3)
  if (!is.na(REMLtau2)) {
    if (trunc) {
      if (REMLtau2 >= 0) {
        REMLdiff <- round (abs(REMLtau2 - REMLtau2_prev), digits =
↪ signiftau2)
      } else {
        negcount <- negcount + 1
        # If iteration is negative more than once then final
↪ estimate REML = 0
        if (negcount >= 2) {
          REMLdiff <- 0
        }
        REMLtau2 <- 0
        REMLtheta <- FEtheta
      }
    } else {
      REMLdiff <- round(abs(REMLtau2 - REMLtau2_prev), digits =
↪ signiftau2)
    }
  }
  REMLit <- REMLit + 1
  if (REMLit == maxit) {
    REMLdiff <- 0
    if (output == TRUE)
      cat ("REML estimator: Maximum number of iterations reached
↪ without convergence \n")
  }
```

```
    # This is just to update theta, rather than because this has
    # anything to do with convergence of this outcome.
    # Estimate for pooled effect
    REMLtheta_prev <- REMLtheta # record of previous step
    REMLtheta <- sum(xi/(vi+REMLtau2)) / sum(1/(vi+REMLtau2))
  }
  if (REMLit == maxit) {
    REML_est <- tau2[esti] <- NA
  } else {
    REML_est <- tau2[esti] <- REMLtau2
  }
  # Pooled effect estimate
  theta[esti] <- REMLtheta
  esti <- esti + 1
}
# Approximate Restricted Maximum Likelihood
if ("AREML" %in% hetests) {
  name[esti] <- "AREML"
  # First set initial estimate of tau2 and theta (fixed-effect
  ↪ estimates)
  AREMLtau2 <- 0
  AREMLtheta <- FEtheta
  # Difference between this iteration and previous to assess when we
  ↪ have convergence
  # set diff != 0 initially to get the process of iteration going
  AREMLdiff <- 1
  AREMLit <- 0  # counter for number of iterations
  # Process of iteration, stop when there is no difference between the
  ↪  last two steps.
  # If the number of events = number of people in trial arm,
  # then the estimator cannot be calculated and NA is produced.
  # So one must avoid looping through any NA estimates
  while (AREMLdiff != 0 & !is.na(AREMLtau2)) {
    # Estimate of between study heterogeneity
    AREMLtau2_prev <- AREMLtau2 # record of previous step
    tau2_p1 <- sum((1/((vi+AREMLtau2_prev)^2)) * (((K/(K-1)) * (xi-
  ↪ AREMLtheta)^2) - vi))
    tau2_p2 <- sum(1/((vi+AREMLtau2_prev)^2))
    AREMLtau2 <- tau2_p1/tau2_p2
    if (!is.na(AREMLtau2)) {
      if (trunc) {
```

```
        if (AREMLtau2 >= 0) {
          AREMLdiff <- round(abs(AREMLtau2 - AREMLtau2_prev), digits =
  ↪  signiftau2)
        } else {
          AREMLdiff <- 0
          AREMLtau2 <- 0
          AREMLtheta <- FEtheta
        }
      } else {
        AREMLdiff <- round (abs(AREMLtau2 - AREMLtau2_prev), digits =
  ↪ signiftau2)
      }
    }
    AREMLit <- AREMLit + 1
    if (AREMLit == maxit) {
      AREMLdiff <- 0
      cat ("AREML estimator: Maximum number of iterations reached
  ↪ without convergence \n")
    }
    # This is just to update theta, rather than because this has
    # anything to do with convergence of this outcome.
    # Estimate for pooled effect
    AREMLtheta_prev <- AREMLtheta # record of previous step
    AREMLtheta <- sum(xi/(vi+AREMLtau2)) / sum(1/(vi+AREMLtau2))
  }
  if (AREMLit == maxit) {
    AREML_est <- tau2[esti] <- NA
  } else {
    AREML_est <- tau2[esti] <- AREMLtau2
  }
  # Pooled effect estimate
  theta[esti] <- AREMLtheta
  esti <- esti + 1
}
# Approximate Bayesian
if ("AB" %in% hetests) {
  # Check that exactly 2 of the prior parameters are specified,
  ↪ otherwise return an error
  countarg <- 0
  if (!is.numeric(lambda)) {
    countarg <- countarg + 1
```

```
  }
  if (!is.numeric(eta)) {
    countarg <- countarg + 1
  }
  if (!is.numeric(tau2prior)) {
    countarg <- countarg + 1
  }
  if (countarg == 1) {
    name[esti] <- "AB"
    # Calculate both eta and lambda parameters if they are not both
  ↪ specified
    if (is.null(lambda)) {
      lambda <- tau2prior*(eta-1)
    } else if (is.null(eta)) {
      eta <- (lambda/tau2prior) + 1
    }
    # Compute approximate Bayes estimate of heterogeneity variance
    # from DerSimonian-Laird estimate and prior distribution
    AB_est <- tau2[esti] <- max(0, (2*lambda+K*DL_est) / (2*eta+K-2))
    esti <- esti + 1
  }
  else {cat ("ERROR: AB estimate cannot be calculated as prior
  ↪ parameters have been specified incorrectly \n")}
}
# Rukhin Bayes (simple prior)
if ("RB" %in% hetests) {
  name [esti] <- "RB"
  # Just assume fixed-effects mean for now, paper does not specify
  RBtheta <- theta[esti] <- FEtheta
  if (trunc) {
    RB_est <- tau2[esti] <- max(0, (sum((xi-RBtheta)^2)/(K+1)) + ((sum
  ↪ (nc+nt)-K)*((2*K*tau2prior)-(K-1)*sum(vi)) / (K*(K+1)*sum(nc+nt-K
  ↪ +2))))
  } else {
    RB_est <- tau2[esti] <- (sum((xi-RBtheta)^2)/(K+1)) + ((sum(nc+nt)
  ↪ -K)*((2*K*tau2prior)-(K-1)*sum(vi)) / (K*(K+1)*sum(nc+nt-K+2)))
  }
  esti <- esti + 1
}
# Rukhin Bayes (zero prior)
if ("RB0" %in% hetests) {
```

```
  name [esti] <- "RB0"
  # Just assume fixed-effects mean, this is what Kontopantelis used.
  # Also they do not specify what n_i is - the estimator is not
  ↪ proposed
  # in the context where there are 2 treatment groups per study,
  # so we assume N = nc + nt as used by Kontopantelis.
  # Not exactly the same formula as in Rukhin (2012), because there is
  ↪  a mistake,
  # this is the corrected formula similar to that used by Konto (2012)
  RB0theta <- theta[esti] <- FEtheta
  if (trunc) {
    RB0_est <- tau2[esti] <- max(0, (sum((xi-RB0theta)^2)/(K+1)) - (((
  ↪ sum(nc+nt)-K)*(K-1)*sum(vi)) / (K*(K+1)*sum(nc+nt-K+2))))
  } else {
    RB0_est <- tau2[esti] <- (sum((xi-RB0theta)^2)/(K+1)) - (((sum(nc+
  ↪ nt)-K)*(K-1)*sum(vi)) / (K*(K+1)*sum(nc+nt-K+2)))
  }
  if (is.infinite(RB0_est)) {
    RB0_est <- tau2[esti] <- NA # this is possible of the denominator
  ↪ is zero (rare)
  }
  esti <- esti + 1
}
# Bayes Modal
if ("BM" %in% hetests) {
  name[esti] <- "BM"
  # Similar to the equation from Chung and Veroniki papers, but
  ↪ replacing var(tauML) with var(tauML^2)
  # as the original equation does not work when tauML=0.
  # Still unsure whether the tauML should also then be replaced with
  ↪ tauML^2 to match (not done here).
  varML2 <- 2/(sum(1/(vi+MLtau2)^2))
  # If the number of events = number of people in trial arm,
  # then the estimator cannot be calculated and NA is produced.
  # So one must avoid looping through any NA estimates
  if (!is.na(MLtau2) & MLtau2 == 0) {
    BM_est <- tau2[esti] <- varML2
  } else {
    BM_est <- tau2[esti] <- ((sqrt(MLtau2)/2) + (sqrt(MLtau2)/2)*sqrt
  ↪ (1+(4*varML2/MLtau2)))^2
  }
```

```
  }
  ## Data frame for reporting all output
  # First round off the estimate to specified number of decimal places
    ↪ by signiftau2 argument
  tau2 <- signif(tau2, digits = signiftau2)
  out <- data.frame(name, tau2)
  if (output == TRUE) print(out)
  ## Output that can be used after function has been run
  # Create an output frame that can be used when iterating through this
    ↪ function multiple times.
  # The above dataframe is better when only calculating estimates for
    ↪ one meta-analysis
  res <- as.list(paste(hetests, "_est", sep = ""))
  names(res) <- hetests
  for (j in 1:length(hetests)) {
    res[[j]] <- get(res[[j]])
  }
  return (res)
  # We can refer to the estimates outside of this function by <funct
    ↪ name>$<est name>
}

## Run the heterogeneity variance estimators function ##
# (in loop for each simulated meta-analysis)
estresults <- data.frame()
for (x in 1:mydata$meta[nrow(mydata)]) {
  newdata <- mydata[which(mydata$meta == x), ]
  logRR <- newdata$logRR
  selogRR <- newdata$selogRR
  nt <- newdata$nt
  nc <- newdata$nc
  ec <- newdata$ec
  metaests <- hetest(xi = logRR, sei = selogRR, Ntot = sum(nc) + sum(nt)
    ↪ , nc = nc, nt = nt, ec = ec, eta = 1, lambda = NULL, tau2prior =
    ↪ 0, DLpos = 0.01, bsamp = 5000, hetests = c("DL", "DLp", "DLb", "HO
    ↪ ", "PM", "HM", "HS", "SJ", "ML", "REML", "AREML", "AB", "RB", "RBO
    ↪ ", "BM"), signiftau2 = 6, maxit = 100, trunc = TRUE, output =
    ↪ FALSE)
  myvars <- names(mydata) %in% c("p0", "p1", "et", "ec", "nt", "nc", "
    ↪ etcor", "eccor", "ntcor", "nccor", "logRR", "selogRR", "SampleSize
    ↪ ", "study")
```

```
  keydata <- newdata[!myvars]
  keydata <- keydata[1, ]
  keydata <- cbind(keydata, metaests)
  estresults <- rbind(estresults, keydata)
}

# Create dataframe of counts of meta-analyses for which iterative
    ↪ estimators
# did not converge for each scenario
noncon <- matrix(NA, nrow = (estresults$scenario[nrow(estresults)]-
    ↪ estresults$scenario[1]+1), ncol = 12)
noncon <- data.frame(noncon)
colnames(noncon) <- c("scenario", "PMcount", "PMprop", "IPMcount", "
    ↪ IPMprop", "MLcount", "MLprop", "REMLcount", "REMLprop", "
    ↪ AREMLcount", "AREMLprop", "total")

z <- 0
for (i in estresults$scenario[1]:estresults$scenario[nrow(estresults)])
    ↪ {
  z <- z + 1
  smallests <- estresults[which(estresults$scenario == i), ]
  noncon$scenario[z] <- i
  noncon$total[z] <- nrow(smallests)
  if ("PM" %in% colnames(estresults)) {
    noncon$PMcount[z] <- sum(is.na(smallests$PM))
    noncon$PMprop[z] <- noncon$PMcount[z]/noncon$total[z]
  }
  if ("IPM" %in% colnames(estresults)) {
    noncon$IPMcount[z] <- sum(is.na(smallests$IPM))
    noncon$IPMprop[z] <- noncon$IPMcount[z]/noncon$total[z]
  }
  if ("ML" %in% colnames(estresults)) {
    noncon$MLcount[z] <- sum(is.na(smallests$ML))
    noncon$MLprop[z] <- noncon$MLcount[z]/noncon$total[z]
  }
  if ("REML" %in% colnames(estresults)) {
    noncon$REMLcount[z] <- sum(is.na(smallests$REML))
    noncon$REMLprop[z] <- noncon$REMLcount[z]/noncon$total[z]
  }
  if ("AREML" %in% colnames(estresults)) {
    noncon$AREMLcount[z] <- sum(is.na(smallests$AREML))
```

```
    noncon$AREMLprop[z] <- noncon$AREMLcount[z]/noncon$total[z]
  }
}

# Save this dataframe of non-converging counts per scenario
myvars <- names(estresults) %in% c("scenario", "tau2", "alpha", "
    ↪ SampleSize1", "k", "theta", "varalpha", "eventdist", "sampdist", "
    ↪ contcorr")
myvars <- estresults[myvars]
noncon <- merge(unique(myvars), noncon)
write.table(noncon, file = paste("noncon_results_", iscen, sep = ""))

# Set those estimates < 1x10^-5 to zero (set a class for zero)
for (i in 1:nrow(estresults)) {
  for (j in 14:ncol(estresults)) {
    if (!is.na(estresults[i,j]) & estresults[i,j] < 0.00001) {
      estresults[i,j] <- 0
    }
  }
}

# Save heterogeneity variance estimates data frame to a file
write.table(estresults, file = paste("estresults_", iscen, sep = ""))

##############################
# R code for applying GLMMs #
##############################

# Re-designing datafile to make it the correct format for Poisson model
    ↪ input
meta <- rep(mydata$meta, times = 2)
scenario <- rep(iscen, times = length(meta))
trial <- rep(study, times = 2)
treat <- c(rep(0, times = length(study)), rep(1, times = length(study)))
event <- c(mydata$ec, mydata$et)
number <- ptime <- c(mydata$nc, mydata$nt)
PMRMdata <- data.frame(scenario, meta, trial, treat, event, number,
    ↪ ptime, mydata$ec, mydata$et, mydata$k, mydata$alpha, mydata$
    ↪ SampleSize1)
colnames(PMRMdata)[8:12] <- c("ec", "et", "k", "alpha", "SampleSize")
```

```
# Re-designing datafile to make it the correct format for conditional
    ↪ logistic model input
meta <- mydata$meta
scenario <- rep(iscen, times = length(meta))
pratio <- mydata$nt/mydata$nc
CLMRMdata <- data.frame(scenario, meta, study, pratio, mydata$ec, mydata
    ↪ $et, mydata$k)
colnames(CLMRMdata)[5:7] <- c("ec", "et", "k")

## List of arguments and their meanings ##

# PMRMdata - data frame in format needed to apply the Poisson
# model, and containing the variables meta, trial, treat, event,
# number, ptime, ec, et and k for each study

# CLMRMdata - data frame in the format needed to apply the
# conditional logistic model, and containing the variables meta,
# study, pratio, ec, et, k, event and number for each study

# modests - vector of GLMMs that you would like to be applied to
# estimate the heterogeneity variance. The default is NULL, which
# means all GLMMs are applied.

# signiftau2 - number of significant figures to round tau2 estimates

# signiflogRR - number of significant figures to round logRR estimates

# output - TRUE if output is displayed, FALSE otherwise (stops too much
# output from the GLMMs)

## Function for applying Poisson and conditional logistic models ##
## and extracting the tau^2 estimates ##

modest <- function(PMRMdata = PMRMdata, CLMRMdata = CLMRMdata, modests =
    ↪  NULL, signiftau2 = 6, signiflogRR = 6, output = TRUE) {
  if (is.null(modests)) modests <- c("PMRM", "CLMRM")
  PMRM_est <- CLMRM_est <- PMRM_logRR <- CLMRM_logRR <- as.numeric(NA)
  Kest <- length(modests) # number of heterogeneity variance estimates
    ↪ to be calculated
  esti <- 1 # start counter for number of estimators
  name <- rep(NA, times = Kest)
```

```
tau2 <- rep(NA, times = Kest)
logRR <- rep(NA, times = Kest)
# Poisson regression model
if ("PMRM" %in% modests) {
  name[esti] <- "PMRM"
  k <- PMRMdata$k[1]
  et <- PMRMdata$event[PMRMdata$treat == 1]
  ec <- PMRMdata$event[PMRMdata$treat == 0]
  nt <- PMRMdata$number[PMRMdata$treat == 1]
  nc <- PMRMdata$number[PMRMdata$treat == 0]
  alpha <- PMRMdata$alpha[1]
  # Do not attempt to apply the Poisson model to any meta-analyses
 ↪ that contain problematic studies for which
  # the Poisson model cannot be applied, and instead give their
 ↪ estimates as NA
  if ((((all(et/nt == et[1]/nt[1])) & (all(ec/nc == ec[1]/nc[1]))) & (
 ↪ all(ec == 0) | (all(ec == ec[1]) & all(nc == nc[1])))) | (k == 5 &
 ↪  (alpha == -6.9 | alpha == -5.3)) | (PMRMdata$SampleSize[1] == "
 ↪ small and large" & k == 10 & (alpha == -6.9 | alpha == -5.3)) | (
 ↪ PMRMdata$scenario[1] == 1433 & PMRMdata$meta[1] == 829) | (
 ↪ PMRMdata$scenario[1] == 1999 & PMRMdata$meta[1] == 944) | (
 ↪ PMRMdata$scenario[1] == 1793 & PMRMdata$meta[1] == 2654)) {
    PMRM_est <- tau2[esti] <- NA
    PMRM_logRR <- logRR[esti] <- NA
  } else {
    # Apply the Poisson model to the meta-analyses that will work,
    # and extract the heterogeneity variance and log-risk ratio
    # estimates from the model output parameters
    pois.glmer <- glmer(event ~ 1+treat+(1+treat|trial), offset = log(
 ↪ ptime), data = PMRMdata, family = poisson, control = glmerControl(
 ↪ optimizer = "Nelder_Mead", tolPwrss = 1e-3, optCtrl = list(maxfun
 ↪ = 100000)), nAGQ = 0)
    temp <- VarCorr(pois.glmer)
    PMRM_est <- tau2[esti] <- temp$trial[2, 2]
    PMRM_logRR <- logRR[esti] <- coef(summary(pois.glmer))[2 , "
 ↪ Estimate"]
  }
  esti <- esti + 1
}
# Conditional logistic regression model
if ("CLMRM" %in% modests) {
```

```
  name[esti] <- "CLMRM"
  ec <- CLMRMdata$ec
  et <- CLMRMdata$et
  k <- CLMRMdata$k[1]
  # Do not attempt to apply the conditional logistic model to any meta
 ↪ -analyses that contain problematic studies for which
  # the conditional logistic model cannot be applied, and instead give
 ↪  their estimates as NA
  if (((sum((ec+et) == 0) >= (k-1)) | all(ec == 0) | all(et == 0) | (
 ↪ all(et[which(!(ec+et) %in% 0)]/ec[which(!(ec+et) %in% 0)] == et[
 ↪ which(!(ec+et) %in% 0)][1]/ec[which(!(ec+et) %in% 0)][1])) | (
 ↪ CLMRMdata$scenario[1] == 1433 & CLMRMdata$meta[1] == 829) | (
 ↪ CLMRMdata$scenario[1] == 1999 & CLMRMdata$meta[1] == 944) | (
 ↪ CLMRMdata$scenario[1] == 1789 & CLMRMdata$meta[1] == 540)) {
    CLMRM_est <- tau2[esti] <- NA
    CLMRM_logRR <- logRR[esti] <- NA
  } else {
    # Apply the conditional logistic model to the meta-analyses
    # that will work, and extract the heterogeneity variance and
    # log-risk ratio estimates from the model output parameters
    cond.glmer <- glmer(cbind(et,ec) ~ 1+(1|study), offset = log(
 ↪ pratio), data = CLMRMdata, family = binomial)
    temp <- VarCorr(cond.glmer)
    CLMRM_est <- tau2[esti] <- temp$study[1]
    CLMRM_logRR <- logRR[esti] <- coef(summary(cond.glmer))[ , "
 ↪ Estimate"]
  }
}
# Round off the tau2 estimates to specified number of decimal places
 ↪ by signiftau2 argument
tau2 <- signif(tau2, digits = signiftau2)
# Round off the logRR estimates to specified number of decimal places
 ↪ by signiflogRR argument
logRR <- signif(logRR, digits = signiflogRR)
out <- data.frame(name, tau2, logRR)
if (output == TRUE) print(out)
## Output that can be used after function has been run
# Create an output frame that can be used when iterating through this
 ↪ function multiple times.
# The above dataframe is better when only calculating estimates for
 ↪ one meta-analysis
```

```
  res <- c(as.list(paste(modests, "_est", sep = "")), as.list(paste(
    ↪ modests, "_logRR", sep = "")))
  names(res) <- c(modests, paste(modests, "_logRR", sep = ""))
  for (j in 1:length(res)) {
    res[[j]] <- get(res[[j]])
  }
  return (res)
  # We can refer to the estimates outside of this function by <funct
    ↪ name>$<est name>
}

## Apply Poisson and conditional logistic models and save estimates ##
# (in loop for each simulated meta-analysis)
modresults <- data.frame()
for (x in 1:mydata$meta[nrow(mydata)]) {
  newdata <- PMRMdata[which(PMRMdata$meta == x), ]
  newdata2 <- CLMRMdata[which(CLMRMdata$meta == x), ]
  modelests <- modest(PMRMdata = newdata, CLMRMdata = newdata2, modests
    ↪ = c("PMRM", "CLMRM"), signiftau2 = 6, signiflogRR = 6, output =
    ↪ FALSE)
  keydata <- newdata["meta"]
  keydata <- as.data.frame(keydata[1, ])
  keydata <- cbind(keydata, modelests)
  colnames(keydata)[1] <- "meta"
  modresults <- rbind(modresults, keydata)
}

myvars <- c("scenario", "meta", "tau2", "alpha", "SampleSize1", "k", "
    ↪ simulation", "theta", "varalpha", "eventdist", "sampdist", "
    ↪ contcorr")
modresults <- merge(unique(mydata[myvars]), modresults, by = "meta")

# Create dataframe of counts of meta-analyses which had to be excluded
    ↪ from
# the models for each scenario
excl <- matrix(NA, nrow = (modresults$scenario[nrow(modresults)] -
    ↪ modresults$scenario[1]+1), ncol = 6)
excl <- data.frame(excl)
colnames(excl) <- c("scenario", "PMRMcount", "PMRMprop", "CLMRMcount", "
    ↪ CLMRMprop", "total")
```

```
z <- 0
for (i in modresults$scenario[1]:modresults$scenario[nrow(modresults)])
    ↪ {
  z <- z + 1
  smallests <- modresults[which(modresults$scenario == i), ]
  excl$scenario[z] <- i
  excl$total[z] <- nrow(smallests)
  if ("PMRM" %in% colnames(modresults)) {
    excl$PMRMcount[z] <- sum(is.na(smallests$PMRM))
    excl$PMRMprop[z] <- excl$PMRMcount[z]/excl$total[z]
  }
  if ("CLMRM" %in% colnames(modresults)) {
    excl$CLMRMcount[z] <- sum(is.na(smallests$CLMRM))
    excl$CLMRMprop[z] <- excl$CLMRMcount[z]/excl$total[z]
  }
}

# Save thise data frame of exclusions
myvars <- names(modresults) %in% c("scenario", "tau2", "alpha", "
    ↪ SampleSize1", "k", "theta", "varalpha", "eventdist", "sampdist", "
    ↪ contcorr")
myvars <- modresults[myvars]
excl2 <- merge(unique(myvars), excl)
write.table(excl2, file = paste("modexclude_results_", iscen, sep = ""))

# Set those estimates < 1x10^-5 to zero (set a class for zero)
for (i in 1:nrow(modresults)) {
  for (j in 13:14) {
    if (!is.na(modresults[i,j]) & modresults[i,j] < 0.00001) {
      modresults[i,j] <- 0
    }
  }
}

# Save GLMM estimates data frame to a file
write.table(modresults, file = paste("modresults_", iscen, sep = ""))

############################################################
# R code for conditional heterogeneity variance estimators #
############################################################
```

```
## List of arguments and their meanings ##

# ec - number of events in the control group
# et - number of events in the treatment group

# condests - vector of conditional heterogeneity estimators that you
    ↪ would
# like to be calculated. The default is NULL, which means all
    ↪ conditional
# estimates are calculated.

# logRR - TRUE if the estimates are to be transformed (via an
    ↪ approximation)
# to the variance of the logRR, FALSE if the estimates are to be left as
    ↪  the
# variance of the RR.

# signiftau2 - number of significant figures to round tau2 estimates

# trunc - TRUE if estimators should be truncated to zero, FALSE
    ↪ otherwise

# output - TRUE if output is displayed, FALSE otherwise (stops too much
# output when we are running the program iteratively)

## List of estimators and their acronyms ##

# CO1 - conditional estimating equation (1)
# CO2 - conditional estimating equation (2)
# CO3 - conditional estimating equation (3)
# CO4 - conditional estimating equation (4)

## Function to apply conditional tau2 estimating equations ##

condest <- function(ec = ec, et = et, condests = NULL, logRR = TRUE,
    ↪ signiftau2 = 6, trunc = TRUE, output = TRUE) {
  if (is.null(condests)) condests <- c("CO1", "CO2", "CO3", "CO4")
  CO1_est <- CO2_est <- CO3_est <- CO4_est <- as.numeric(NA)
  yi <- ec + et
  Kest <- length(condests) # number of estimates to be calculated
```

```
esti <- 1 # a counter so that we can create a dataset with a separate
  ↪ estimate on each row - the first specified estimate will be in row
  ↪  1, ..., etc
## Specifying all output vectors before replacing the values with
  ↪ actual estimates
name <- rep(NA, times = Kest)
taup2 <- rep(NA, times = Kest)
# Omit DZ trials
ec <- ec[!yi == 0]
et <- et[!yi == 0]
yi <- yi[!yi == 0]
k <- length(yi)
yia <- yi[yi > 1]
eca <- ec[yi > 1]
eta <- et[yi > 1]
yib <- yi[yi == 1]
ecb <- ec[yi == 1]
etb <- et[yi == 1]
# Find estimate of probability of event (hat(p))
phat <- sum(et)/sum(yi)
# Transform estimates to variance of log-risk ratio (optional)
if (logRR == TRUE) {
  theta <- log(phat/(1-phat))
  dpneg2 <- ((exp(theta))^2)/(phat^4)
  tau2 <- rep(NA, times = Kest)
}
# Find estimate of hat(tau_p)^2 - heterogeneity variance estimate
if ("CO1" %in% condests) {
  name[esti] <- "CO1"
  if (trunc) {
    CO1_est <- taup2[esti] <- max(0, (1/k)*(sum(((et-(yi*phat))^2)/(yi
  ↪ ^2))-(phat*sum(1/yi))))
  } else {
    CO1_est <- taup2[esti] <- (1/k)*(sum(((et-(yi*phat))^2)/(yi^2))-(
  ↪ phat*sum(1/yi)))
  }
  if (logRR == TRUE) {
    CO1_est <- tau2[esti] <- dpneg2*CO1_est
  }
  esti <- esti + 1
}
```

```
if ("CO2" %in% condests) {
  name[esti] <- "CO2"
  if (k == 1) {
    CO2_est <- taup2[esti] <- NA
  }  else {
    if (trunc) {
      CO2_est <- taup2[esti] <- max(0, ((1/(k-1))*sum(((et-(yi*phat))
↪ ^2)/(yi^2)))-((1/k)*phat*(1-phat)*sum(1/yi)))
    } else {
      CO2_est <- taup2[esti] <- ((1/(k-1))*sum(((et-(yi*phat))^2)/(yi
↪ ^2)))-((1/k)*phat*(1-phat)*sum(1/yi))
    }
    if (logRR == TRUE) {
      CO2_est <- tau2[esti] <- dpneg2*CO2_est
    }
  }
  esti <- esti + 1
}
if ("CO3" %in% condests) {
  name[esti] <- "CO3"
  if (k == 1) {
    CO3_est <- taup2[esti] <- NA
  } else {
    if (all(yi > 1)) {
      if (trunc) {
        CO3_est <- taup2[esti] <- max(0, ((1/(k-1))*sum(((et-(yi*phat)
↪ )^2)/(yi*(yi-1))))-((1/k)*phat*(1-phat)*sum(1/(yi-1))))
      } else {
        CO3_est <- taup2[esti] <- ((1/(k-1))*sum(((et-(yi*phat))^2)/(
↪ yi*(yi-1))))-((1/k)*phat*(1-phat)*sum(1/(yi-1)))
      }
    } else {
      if (trunc) {
        CO3_est <- taup2[esti] <- max(0, ((1/(k-1))*sum(((et-(yi*phat)
↪ )^2)/(yi^2)))-((1/k)*phat*(1-phat)*sum(1/yi)))
      } else {
        CO3_est <- taup2[esti] <- ((1/(k-1))*sum(((et-(yi*phat))^2)/(
↪ yi^2)))-((1/k)*phat*(1-phat)*sum(1/yi))
      }
    }
    if (logRR == TRUE) {
```

```
      CO3_est <- tau2[esti] <- dpneg2*CO3_est
    }
  }
  esti <- esti + 1
}
if ("CO4" %in% condests) {
  name[esti] <- "CO4"
  if (k == 1) {
    CO4_est <- taup2[esti] <- NA
  } else {
    if (all(yi > 1)) {
      if (trunc) {
        CO4_est <- taup2[esti] <- max(0, ((1/(k-1))*sum(((et-(yi*phat)
↪ )^2)/(yi*(yi-1))))-((1/k)*phat*(1-phat)*sum(1/(yi-1))))
      } else {
        CO4_est <- taup2[esti] <- ((1/(k-1))*sum(((et-(yi*phat))^2)/(
↪ yi*(yi-1))))-((1/k)*phat*(1-phat)*sum(1/(yi-1)))
      }
    } else if (all(yi <= 1)) {
      if (trunc) {
        CO4_est <- taup2[esti] <- max(0, ((1/(k-1))*sum(((et-(yi*phat)
↪ )^2)/(yi^2)))-((1/k)*phat*(1-phat)*sum(1/yi)))
      } else {
        CO4_est <- taup2[esti] <- ((1/(k-1))*sum(((et-(yi*phat))^2)/(
↪ yi^2)))-((1/k)*phat*(1-phat)*sum(1/yi))
      }
    } else {
      if (trunc) {
        CO4_est <- taup2[esti] <- max(0, ((1/(k-1))*(sum(((eta-(yia*
↪ phat))^2)/(yia*(yia-1)))+sum(((etb-(yib*phat))^2)/(yib^2))))-((1/k
↪ )*phat*(1-phat)*(sum(1/(yia-1))+sum(1/yib))))
      } else {
        CO4_est <- taup2[esti] <- ((1/(k-1))*(sum(((eta-(yia*phat))^2)
↪ /(yia*(yia-1)))+sum(((etb-(yib*phat))^2)/(yib^2))))-((1/k)*phat*
↪ (1-phat)*(sum(1/(yia-1))+sum(1/yib)))
      }
    }
    if (logRR == TRUE) {
      CO4_est <- tau2[esti] <- dpneg2*CO4_est
    }
  }
```

```
  }
  if (logRR == TRUE) {
    tau2 <- signif(tau2, digits = signiftau2)
    out <- data.frame(name, tau2)
  } else {
    taup2 <- signif(taup2, digits = signiftau2)
    out <- data.frame(name, taup2)
  }
  if (output == TRUE) print(out)
  ## Output that can be used after function has been run
  # Create an output frame that can be used when iterating through this
    ↪ function multiple times.
  # The above dataframe is better when only calculating estimates for
    ↪ one meta-analysis
  res <- as.list(paste(condests, "_est", sep = ""))
  names(res) <- condests
  for (j in 1:length(condests)) {
    res[[j]] <- get(res[[j]])
  }
  return (res)
}

condresults <- data.frame()
for (x in 1:mydata$meta[nrow(mydata)]) {
  newdata <- mydata[which(mydata$meta == x), ]
  ec <- newdata$ec
  et <- newdata$et
  metaests <- condest(ec = ec, et = et, condests = c("CO1", "CO2", "CO3"
    ↪ , "CO4"), logRR = TRUE, signiftau2 = 6, trunc = TRUE, output =
    ↪ FALSE)
  myvars <- names(mydata) %in% c("p0", "p1", "ec", "et", "nc", "nt", "
    ↪ eccor", "etcor", "nccor", "ntcor", "logRR", "selogRR", "SampleSize
    ↪ ", "study")
  keydata <- newdata[!myvars]
  keydata <- keydata[1, ]
  keydata <- cbind(keydata, metaests)
  condresults <- rbind(condresults, keydata)
}

# Set those estimates < 1x10^-5 to zero (set a class for zero)
for (i in 1:nrow(condresults)) {
```

```
  for (j in 14:ncol(condresults)) {
    if (!is.na(condresults[i,j]) & condresults[i,j] < 0.00001) {
      condresults[i,j] <- 0
    }
  }
}

write.table(condresults, file = paste("condresults_", iscen, sep = ""))

################################################################
# R code for mixture model heterogeneity variance estimator #
################################################################

## List of arguments and their meanings ##

# xi - vector of effect estimates for each study. If the outcome is
# risk ratio (for example), we assume that xi is already converted to
# log-risk ratios. log argument can be used to convert output back onto
# the original scale after all heterogeneity estimates have been
# calculated.

# ec - number of events in the control group
# et - number of events in the treatment group

# maxJ - the maximum number of model components to consider

# select - the second model selection method (either "BIC" or "LRT",
    ↪ where
# LRT is the likelihood-ratio test) to be used if none of the models
# produced very similar values of q.

# maxit - maximum number of iterations for EM algorithm

# itdiff - stopping value for EM algorithm (the algorithm will stop if
    ↪ the
# absolute difference between observed log-likelihoods is less than
    ↪ itdiff).

# probdiff - value used to find the best-fitting model via the initial
# selection method looking at differences in q (the values of q are
    ↪ deemed
```

```
# to be identical if the difference between them is less than probdiff).

# signiftau2 - number of significant figures to round tau2 estimates
# signiflogRR - number of significant figures to round logRR estimates

# output - TRUE if output is displayed, FALSE otherwise (stops too much
# output when we are running the program iteratively)

## Function to apply EM algorithm and produce tau2 estimates ##

mixest <- function(xi = logRR, ec = ec, et = et, maxJ = 5, select = "BIC
    ↪ ", maxit = 5000, itdiff = 0.0000001, probdiff = 0.001, signiftau2
    ↪ = 6, signiflogRR = 6, output = TRUE) {
  # Method cannot be applied to MAs where all ec = 0, so produce NA
    ↪ estimates for this case
  if (all(ec == 0)) {
    logthetabar <- NA
    logtau2est <- NA
  } else {
    # Calculate total number of events per study over both trial arms
    yi <- ec + et
    # Identify any double-zero (DZ) trials - they cannot be included in
    ↪ this approach
    ec <- ec[!yi == 0]
    et <- et[!yi == 0]
    xi <- xi[!yi == 0]
    yi <- yi[!yi == 0]
    # Number of studies in meta-analysis (excluding any DZ trials
    ↪ omitted)
    k <- length(yi)
    # Apply mixture model for range of components (J)
    bic <- logL <- rep(NA, times = maxJ) # make empty vectors for model
    ↪ comparison measures (BIC and LR test)
    bestcount <- 0 # count to apply method for all J considered, then
    ↪ the best J chosen, and then stop
    while (bestcount < 2) { # to stop applying the method after it has
    ↪ been applied to all J considered and the best J (to obtain results
    ↪ )
      # apply method to each J considered
      J <- 0
```

```
  w <- NULL # count for number of prob's from Binomial dist. that
↪ are essentially identical (diff. is less than probdiff)
  while (is.null(w) & J < maxJ) {
    if (bestcount == 1) { # if the best J/model has been chosen by
↪ model comparison/choice method
      J <- bestJ # set J to be this best J
    } else {
      J <- J + 1 # set a counter for J
    }
    # Weights Pi are equal and sum to 1, i.e. Pi = 1/J
    Pi <- rep(1/J, times = J)
    # Vector of the probabilities (q) for the Binomial distribtuion
↪ (must be of length J)
    # (derived from the initial theta's, where theta' is the RR in
↪ this case)
    qprob <- theta <- rep(NA, times = J) # create empty vectors for
↪ q (the Binomial prob) and theta'
    qprob[1] <- exp(min(xi[is.finite(xi)]))/(exp(min(xi[is.finite(xi
↪ )]))+1) # first element in q vector is based on the minimum (log)
↪ RR value from the data
    qprob[J] <- exp(max(xi[is.finite(xi)]))/(exp(max(xi[is.finite(xi
↪ )]))+1) # last element in q vector is based on the maximum (log)RR
↪  value from the data
    # if J > 2, fill in the remainder of elements in q by equally
↪ spacing between the first and the last elements
    if (J > 2) {
      y <- 0
      for (i in 2:(J-1)) {
        y <- y + 1
        qprob[i] <- qprob[1] + (y*((qprob[J] - qprob[1])/J))
      }
    }
    MMit <- 1 # iteration number
    MMdiff <- 1 # to get iteration started, MMdiff < 0.0001 implies
↪ convergence
    while (!is.na(MMdiff) & MMdiff >= itdiff) {
      # Calculate observed log-likelihood for MMit = 1, set as value
↪  from previous iteration otherwise
      if (MMit == 1) {
```

```
        bi <- e <- matrix(0, nrow = k, ncol = J) # create matrices
↪ of zeros (see while below) for Binomial density function and
↪ expected value
        biPi <- rep(NA, times = k) # create empty vector for
↪ Binomial density function * weight Pi
        for (i in 1:k) {
          for (j in 1:J) {
            bi[i,j] <- dbinom(x = et[i], size = yi[i], prob = qprob[
↪ j]) # Binomial density function
          }
          biPi[i] <- sum(bi[i, ]*Pi)
        }
        obll <- log(prod(biPi)) # observed log-likelihood for MMit =
↪  1
      } else {
        obll <- obllnew # observed log-likelihood for MMit > 1
      }
      # E-step
      # Calculate expected value for z_ij
      for (j in 1:J) {
        for (i in 1:k) {
          e[i,j] <- (bi[i,j]*Pi[j])/(biPi[i]) # expected value
          if (is.na(e[i,j])) { # if e is NA, set it equal to 1/J in
↪ order for algorithm to move forward
            e[i,j] <- 1/J
          }
        }
        while (!is.finite((sum(e[ ,j]*et))/(sum(e[ ,j]*yi) - sum(e[
↪ ,j]*et)))) {
          e[which.min(e[,j]),j] <- 1/J
        }
      }
      # M-step
      # Updated estimates resulting from maximisation of complete
↪ log-likelihood
      for (j in 1:J) {
        Pi[j] <- sum(e[ ,j])/k # updated estimate of Pi
        theta[j] <- (sum(e[ ,j]*et))/(sum(e[ ,j]*yi) - sum(e[ ,j]*et
↪ )) # updated estimate of theta
        qprob[j] <- theta[j]/(1+theta[j]) # updated value of q using
↪  updated estimates of theta
```

```
      }
      # Use updated estimates of Pi and theta to re-calculate
↪ observed log-likelihood
      bi <- matrix(0, nrow = k, ncol = J) # create matrices of zeros
↪  (see while below) for Binomial density function and expected
↪ value
      biPi <- rep(NA, times = k) # create empty vector for Binomial
↪ density function * weight Pi
      for (i in 1:k) {
        for (j in 1:J) {
          bi[i,j] <- dbinom(x = et[i], size = yi[i], prob = qprob[j
↪ ]) # Binomial density function
        }
        biPi[i] <- sum(bi[i, ]*Pi)
      }
      obllnew <- log(prod(biPi)) # updated observed log-likelihood
      # Calculate difference between updated and previous observed
↪ log-likelihood
      MMdiff <- abs(obllnew - obll)
      if (is.na(MMdiff)) {
        Piout <- thetaout <- NA
      } else {
        if (MMdiff >= itdiff) # repeat algorithm
          MMit <- MMit + 1
        if (MMdiff < itdiff) { # condition is met - stop algorithm
↪ and extract estimates
          Piout <- Pi
          thetaout <- theta
        }
        if (MMit == maxit) { # maximum number of iterations has been
↪  reached without convergence
          MMdiff <- 0
          if (output == TRUE)
            cat ("MM estimator: Maximum number of iterations reached
↪  without convergence \n")
        }
      }
    }
    if (MMit == maxit) { # if convergence has not been achieved, set
↪  output estimates to NA
      Piout <- thetaout <- NA
```

```
    }
    if (bestcount == 1) { # if the best J/model has been chosen
↪ already, stop here
      break
    } else { # calculate the model comparison measures
      bic[J] <- (2*J*log(k)) - (2*obllnew) # BIC
      logL[J] <- obllnew # observed log-likelihood for use in
↪ likelihood-ratio (LR) test
      # compile a vector (w) of those J that result in q's with an
↪ absolute difference less than probdiff (i.e. very similar q's)
      if (J >= 2) {
        m <- 0
        while(is.null(w) & m < (J-1)) {
          m <- m + 1
          if (abs(qprob[m] - qprob[m+1]) < probdiff) {
            w <- J # vector of J's where some of the q's are
↪ practically identical/very similar
          }
        }
      }
    }
  }
  bestcount <- bestcount + 1
  if (bestcount == 1) { # if best J/model has yet to be chosen,
↪ choose it via model comparison methods
    # if a number of the q's were deemed to be identical (i.e.
↪ length(w) > 0), then choose best model as min(w)-1 if min(w) > 1,
↪ else min(w)
    if (!is.null(w)) {
      if (w > 1) {
        bestJ <- w - 1
      } else {
        bestJ <- w
      }
    } else { # if none of the models gave q's with differences less
↪ than probdiff, then use the second model selection method of
↪ choice
      if (select == "BIC") { # if second selection method was BIC,
↪ choose the model with lowest value of BIC
        bestJ <- which(bic == min(bic))[[1]]
```

```
      } else if (select == "LRT") { # if second selection method was
↪  LRT, apply the likelihood ratio test to find best model
        chisq <- qchisq(0.95, df = 2) # Chi-square value with df = 2
↪  (difference in number of parameters between each J-based model)
        LR <- chisq + 1 # set likelihood ratio to be an arbitrary
↪ value to begin loop
        i <- 0 # set a counter for J
        # look for J(=i) where LR <= chisq (as the model with less
↪ parameters between this comparison of 2 models is the best)
        while (LR > chisq & i < (maxJ-1)) {
          i <- i + 1 # counter for J
          LR <- -2 * (logL[i] - logL[i+1]) # likelihood-ratio
↪ statistic
        }
        # if counter i has reached maxJ-1, then best fitting model
↪ must be the model with most parameters (maxJ)
        if (LR > chisq & i == (maxJ-1)) {
          bestJ <- maxJ
        } else { # if counter i did not reach maxJ-1, then the above
↪  while loop was satisfied and best J = i
          bestJ <- i
        }
      } else { # if no second selection method was specified, but is
↪  necessary, then give warning
        stop("Second-choice model selection technique (select) is
↪ necessary and must be either BIC or LRT")
      }
    }
  }
 }
 Pi <- Piout # final estimate of Pi
 theta <- thetaout # final estimate of theta
 logtheta <- log(theta) # estimate of the converted logRR
 thetabar <- sum(Pi * theta) # estimate of the overall risk ratio
 tau2est <- sum(Pi * ((theta - thetabar)^2)) # estimate of tau_p^2
 # Conversions to estimates of interest
 logthetabar <- sum(Pi * logtheta) # estimate of overall logRR
 logtau2est <- sum(Pi * ((logtheta - logthetabar)^2)) # estimate of
↪ tau^2
}
# Output estimates
```

```
  tau2 <- signif(logtau2est, digits = signiftau2)
  logRRest <- signif(logthetabar, digits = signiflogRR)
  out <- data.frame("MM", tau2, logRRest)
  if (output == TRUE) print(out)
  ## Output that can be used after function has been run
  # Create an output frame that can be used when iterating through this
    ↪ function multiple times.
  # The above dataframe is better when only calculating estimates for
    ↪ one meta-analysis
  res <- list(tau2, logRRest)
  names(res) <- c("MM", "MM_logRR")
  return(res)
}

# Apply mixture model approach
## Run the heterogeneity variance estimators function ##
# (in loop for each simulated meta-analysis)
mixresults <- data.frame()
for (x in 1:mydata$meta[nrow(mydata)]) {
  newdata <- mydata[which(mydata$meta == x), ]
  logRR <- newdata$logRR
  ec <- newdata$ec
  et <- newdata$et
  metaests <- mixest(xi = logRR, ec = ec, et = et, maxJ = 5, select = "
    ↪ BIC", maxit = 5000, itdiff = 0.0000001, probdiff = 0.001,
    ↪ signiftau2 = 6, signiflogRR = 6, output = FALSE)
  myvars <- names(mydata) %in% c("p0", "p1", "et", "ec", "nt", "nc", "
    ↪ etcor", "eccor", "ntcor", "nccor", "logRR", "selogRR", "SampleSize
    ↪ ", "study")
  keydata <- newdata[!myvars]
  keydata <- keydata[1, ]
  keydata <- cbind(keydata, metaests)
  mixresults <- rbind(mixresults, keydata)
}

# Create dataframe of counts of meta-analyses for which iterative
    ↪ estimators
# did not converge for each scenario
mixnoncon <- matrix(NA, nrow = (mixresults$scenario[nrow(mixresults)]-
    ↪ mixresults$scenario[1]+1), ncol = 4)
mixnoncon <- data.frame(mixnoncon)
```

```
colnames(mixnoncon) <- c("scenario", "MMcount", "MMprop", "total")

z <- 0
for (i in mixresults$scenario[1]:mixresults$scenario[nrow(mixresults)])
    ↪ {
  z <- z + 1
  smallests <- mixresults[which(mixresults$scenario == i), ]
  mixnoncon$scenario[z] <- i
  mixnoncon$total[z] <- nrow(smallests)
  mixnoncon$MMcount[z] <- sum(is.na(smallests$MM))
  mixnoncon$MMprop[z] <- mixnoncon$MMcount[z]/mixnoncon$total[z]
}

# Save this dataframe of non-converging counts per scenario
myvars <- names(mixresults) %in% c("scenario", "tau2", "alpha", "
    ↪ SampleSize1", "k", "theta", "varalpha", "eventdist", "sampdist", "
    ↪ contcorr")
myvars <- mixresults[myvars]
mixnoncon <- merge(unique(myvars), mixnoncon)
write.table(mixnoncon, file = paste("mixnoncon_results_", iscen, sep = "
    ↪ "))

# Set those estimates < 1x10^-5 to zero (set a class for zero)
for (i in 1:nrow(mixresults)) {
  for (j in 14:ncol(mixresults)) {
    if (!is.na(mixresults[i,j]) & mixresults[i,j] < 0.00001) {
      mixresults[i,j] <- 0
    }
  }
}

# Save heterogeneity variance estimates data frame to a file
write.table(mixresults, file = paste("mixresults_", iscen, sep = ""))

############################################################
# R code for calculating confidence interval estimates #
############################################################

## Incorporate theta estimates into simulation results dataframe ##
# (needed for theta performance measures and construction of CIs)
```

```
# Remove the duplicate columns (except for meta to merge)
condnames <- names(condresults) %in% c("scenario", "tau2", "taup2", "
    ↪ alpha", "SampleSize1", "k", "simulation", "theta", "varalpha", "
    ↪ eventdist", "sampdist", "contcorr")
condresults <- condresults[!condnames]
# Combine these 2 data frames together by the meta-analysis ID variable
results <- merge(estresults, condresults, by = "meta")

# Remove the duplicate columns (except for meta to merge)
modnames <- names(modresults) %in% c("scenario", "tau2", "alpha", "
    ↪ SampleSize1", "k", "simulation", "theta", "varalpha", "eventdist",
    ↪  "sampdist", "contcorr")
modresults <- modresults[!modnames]
# Combine these 2 data frames together by the meta-analysis ID variable
results <- merge(results, modresults, by = "meta")

# Remove the duplicate columns (except for meta to merge)
mixnames <- names(mixresults) %in% c("scenario", "tau2", "taup2", "alpha
    ↪ ", "SampleSize1", "k", "simulation", "theta", "varalpha", "
    ↪ eventdist", "sampdist", "contcorr")
mixresults <- mixresults[!mixnames]
# Combine these 2 data frames together by the meta-analysis ID variable
results <- merge(results, mixresults, by = "meta")

# Trim down estresults variables to avoid duplicates when merging
myvars <- names(results) %in% c("scenario", "tau2", "taup2", "alpha", "
    ↪ SampleSize1", "k", "simulation", "theta", "varalpha", "eventdist",
    ↪  "sampdist", "contcorr")
cutresults <- results[!myvars]
myvars <- names(mydata) %in% c("etcor", "eccor", "ntcor", "nccor")
cutdata <- mydata[!myvars]

# Combine theta values with heterogeneity variance estimates
results <- merge(cutdata, cutresults, by.x = "meta", by.y = "meta")

# Save this complete dataframe
write.table(results, file = paste("theta_results_", iscen, sep = ""))

## Construction of confidence intervals (CIs) ##

# Create vector/list of heterogeneity variance estimating methods
```

```
myvars <- names(results) %in% c("SampleSize", "study", "et", "ec", "nt",
    ↪  "nc", "logRR", "selogRR", "meta", "I2", "tau2", "taup2", "p0", "
    ↪ p1", "alpha", "SampleSize1", "k", "simulation", "theta", "varalpha
    ↪ ", "scenario", "eventdist", "sampdist", "contcorr", "PMRM_logRR",
    ↪ "CLMRM_logRR", "MM_logRR")
hetdata <- results[!myvars]
hetnames <- names(hetdata)

# Extract the logRR estimates produced via the poisson and condlog
    ↪ models
models <- c("PMRM_logRR", "CLMRM_logRR", "MM_logRR")
modelthetas <- results[models]

## List of arguments and their meanings ##

# xi - effect estimates of the studies in the meta-analysis
# sei - standard errors of the effect estimates
# ec - number of events in control arm for each study
# et - number of events in treatment arm for each study
# nc - sample size of control arm for each study
# nt - sample size of treatment arm for each study
# hetests - vector of heterogeneity estimates
# modelthetas - theta (logRR) estimates produced by GLMM and mixture
    ↪ model (MM) methods
# CIests - names of the confidence intervals to be calculated
# hetnames - names of the heterogeneity estimators (corresponding to the
    ↪  heterogeneity estimates in hetests argument)
# signif - significance level set for the confidence intervals, the
    ↪ default is 0.05 (i.e. 95% CI)
# signifCI - number of significant figues to round off the CI bounds
# MH - if TRUE then Mantel-Haenszel estimate for logRR is calculated (
    ↪ with NAs for those that cannot be calculated without a cont. corr
    ↪ .)
# MHc - if TRUE then Mantel-Haenszel estimate for logRR is calculated
    ↪ using a continuity correction when required

## PARAMETERS SPECIFIC TO MHc
# c - constant continuity correction to be used in the case of all-zero
    ↪ trial arms

# output - TRUE if output to be displayed, FALSE otherwise
```

```
## List of confidence intervals and their acronyms ##

# Z - Z-type confidence interval
# T - t-distribution confidence interval
# HKSJ - Hartung-Knapp-Sidik-Jonkman confidence interval
# mKH - modified Hartung-Knapp confidence interval

## Confidence interval (CI) estimators code ##

CIest <- function(xi = logRR, sei = selogRR, ec = ec, et = et, nc = nc,
    ↪ nt = nt, hetests = hetests, modelthetas = modelthetas, CIests =
    ↪ NULL, hetnames = NULL, signif = 0.05, signifCI = 4, MH = TRUE, MHc
    ↪  = TRUE, c = 0.5, output = TRUE) {
  if (is.null(CIests)) CIests <- c("Z", "T", "HKSJ", "mKH")
  if (is.null(hetnames)) hetnames <- c("HO", "DL", "PM", "IPM", "HO2", "
    ↪ DL2", "DLp", "DLb", "HM", "HS", "SJ", "SJ2", "ML", "REML", "AREML"
    ↪ , "AB", "RB", "RBO", "BM", "CO1", "CO2", "CO3", "CO4", "PMRM", "
    ↪ CLMRM", "MM")
  if (length(hetests) != length(hetnames)) {
    stop("Number of tau2 estimator names doesn't match the number of
    ↪ estimates given")
  }
  thetahetests <- hetests[!names(hetests) %in% c("PMRM", "CLMRM", "MM")]
  thetahetnames <- names(thetahetests)
  hetests <- as.numeric(hetests[1, ])
  thetahetests <- as.numeric((thetahetests[1, ]))
  if (MH == TRUE & MHc == TRUE) {
    CImat <- matrix(NA, nrow = (2*length(CIests))+1, ncol = length(
    ↪ hetnames)+2)
    colnames(CImat) <- c(hetnames, "MH", "MHc")
  } else if (MH == TRUE) {
    CImat <- matrix(NA, nrow = (2*length(CIests))+1, ncol = length(
    ↪ hetnames)+1)
    colnames(CImat) <- c(hetnames, "MH")
  } else if (MHc == TRUE) {
    CImat <- matrix(NA, nrow = (2*length(CIests))+1, ncol = length(
    ↪ hetnames)+1)
    colnames(CImat) <- c(hetnames, "MHc")
  } else {
```

```
  CImat <- matrix(NA, nrow = (2*length(CIests))+1, ncol = length(
 ↪ hetnames))
  colnames(CImat) <- hetnames
}
rowmat <- rep(NA, times = nrow(CImat))
for (i in 2:nrow(CImat)) {
  if (i %% 2 == 0) {
    rowmat[i] <- paste(CIests[i/2], "_lb", sep = "")
  } else {
    rowmat[i] <- paste(CIests[(i/2)-0.5], "_ub", sep = "")
  }
}
rowmat[1] <- "theta"
rownames(CImat) <- rowmat
k <- length(xi) # number of studies
# Calculate mean effects
thetaests <- rep(NA, times = length(thetahetests))
for (i in 1:length(thetahetnames)) {
  CImat["theta", thetahetnames[i]] <- thetaests[i] <- sum(xi*(1/((sei
 ↪ ^2)+thetahetests[i])))/sum(1/((sei^2)+thetahetests[i]))
}
modelthetas <- as.numeric(modelthetas[1, ])
modnames <- hetnames[hetnames %in% c("PMRM", "CLMRM", "MM")]
for (i in 1:length(modelthetas)) {
  CImat["theta", modnames[i]] <- thetaests[i+length(thetahetnames)] <-
 ↪  modelthetas[i]
}
# Mantel-Haenszel estimate of log risk-ratio
if (MH == TRUE) {
  if (all(ec == 0) | all(et == 0)) {
    CImat["theta", "MH"] <- thetaMH <- NA
  } else {
    CImat["theta", "MH"] <- thetaMH <- log((sum((et*nc)/(nc+nt)))/(sum
 ↪ ((ec*nt)/(nc+nt))))
  }
}
if (MHc == TRUE) {
  if (all(ec == 0) | all(et == 0)) {
    ec <- ec + c
    et <- et + c
    nc <- nc + c
```

```
    nt <- nt + c
  }
  CImat["theta", "MHc"] <- thetaMHc <- log((sum((et*nc)/(nc+nt)))/(sum
 ↪ ((ec*nt)/(nc+nt))))
}
# Z-type CI
if ("Z" %in% CIests) {
  for (i in 1:length(hetnames)) {
    CImat["Z_lb", hetnames[i]] <- thetaests[i]-qnorm(1-(signif/2))*
 ↪ sqrt(1/sum(1/(hetests[i]+(sei^2))))
    CImat["Z_ub", hetnames[i]] <- thetaests[i]+qnorm(1-(signif/2))*
 ↪ sqrt(1/sum(1/(hetests[i]+(sei^2))))
  }
  if (MH == TRUE) {
    CImat["Z_lb", "MH"] <- thetaMH-qnorm(1-(signif/2))*sqrt(1/sum(1/(
 ↪ sei^2)))
    CImat["Z_ub", "MH"] <- thetaMH+qnorm(1-(signif/2))*sqrt(1/sum(1/(
 ↪ sei^2)))
  }
  if (MHc == TRUE) {
    CImat["Z_lb", "MHc"] <- thetaMHc-qnorm(1-(signif/2))*sqrt(1/sum(1/
 ↪ (sei^2)))
    CImat["Z_ub", "MHc"] <- thetaMHc+qnorm(1-(signif/2))*sqrt(1/sum(1/
 ↪ (sei^2)))
  }
}
# t-type CI
if ("T" %in% CIests) {
  for (i in 1:length(hetnames)) {
    CImat["T_lb", hetnames[i]] <- thetaests[i]-qt(1-(signif/2), df = k
 ↪ -1)*sqrt(1/sum(1/(hetests[i]+(sei^2))))
    CImat["T_ub", hetnames[i]] <- thetaests[i]+qt(1-(signif/2), df = k
 ↪ -1)*sqrt(1/sum(1/(hetests[i]+(sei^2))))
  }
  if (MH == TRUE) {
    CImat["T_lb", "MH"] <- thetaMH-qt(1-(signif/2), df = k-1)*sqrt(1/
 ↪ sum(1/(sei^2)))
    CImat["T_ub", "MH"] <- thetaMH+qt(1-(signif/2), df = k-1)*sqrt(1/
 ↪ sum(1/(sei^2)))
  }
  if (MHc == TRUE) {
```

```
    CImat["T_lb", "MHc"] <- thetaMHc-qt(1-(signif/2), df = k-1)*sqrt(1
  ↪ /sum(1/(sei^2)))
    CImat["T_ub", "MHc"] <- thetaMHc+qt(1-(signif/2), df = k-1)*sqrt(1
  ↪ /sum(1/(sei^2)))
  }
}
# Hartung-Knapp-Sidik-Jonkman CI
if ("HKSJ" %in% CIests) {
  for (i in 1:length(hetnames)) {
    varHK <- sum((1/(hetests[i]+(sei^2)))*((xi-thetaests[i])^2))/((k
  ↪ -1)*sum(1/(hetests[i]+(sei^2))))
    CImat["HKSJ_lb", hetnames[i]] <- thetaests[i]-qt(1-(signif/2), df
  ↪ = k-1)*sqrt(varHK)
    CImat["HKSJ_ub", hetnames[i]] <- thetaests[i]+qt(1-(signif/2), df
  ↪ = k-1)*sqrt(varHK)
  }
  if (MH == TRUE) {
    varHK <- sum((1/(sei^2))*((xi-thetaMH)^2))/((k-1)*sum(1/(sei^2)))
    CImat["HKSJ_lb", "MH"] <- thetaMH-qt(1-(signif/2), df = k-1)*sqrt(
  ↪ varHK)
    CImat["HKSJ_ub", "MH"] <- thetaMH+qt(1-(signif/2), df = k-1)*sqrt(
  ↪ varHK)
  }
  if (MHc == TRUE) {
    varHK <- sum((1/(sei^2))*((xi-thetaMHc)^2))/((k-1)*sum(1/(sei^2)))
    CImat["HKSJ_lb", "MHc"] <- thetaMHc-qt(1-(signif/2), df = k-1)*
  ↪ sqrt(varHK)
    CImat["HKSJ_ub", "MHc"] <- thetaMHc+qt(1-(signif/2), df = k-1)*
  ↪ sqrt(varHK)
  }
}
# Modified Knapp-Hartung CI
if ("mKH" %in% CIests) {
  for (i in 1:length(hetnames)) {
    q <- sum((1/(hetests[i]+(sei^2)))*((xi-thetaests[i])^2))/(k-1)
    maxq <- max(1, q, na.rm = TRUE)
    varHK <- maxq/(sum(1/(hetests[i]+(sei^2))))
    CImat["mKH_lb", hetnames[i]] <- thetaests[i]-qt(1-(signif/2), df =
  ↪  k-1)*sqrt(varHK)
    CImat["mKH_ub", hetnames[i]] <- thetaests[i]+qt(1-(signif/2), df =
  ↪  k-1)*sqrt(varHK)
```

```
    }
    if (MH == TRUE) {
      q <- sum((1/(sei^2))*((xi-thetaMH)^2))/(k-1)
      maxq <- max(1, q, na.rm = TRUE)
      varHK <- maxq/(sum(1/(sei^2)))
      CImat["mKH_lb", "MH"] <- thetaMH-qt(1-(signif/2), df = k-1)*sqrt(
    ↪ varHK)
      CImat["mKH_ub", "MH"] <- thetaMH+qt(1-(signif/2), df = k-1)*sqrt(
    ↪ varHK)
    }
    if (MHc == TRUE) {
      q <- sum((1/(sei^2))*((xi-thetaMHc)^2))/(k-1)
      maxq <- max(1, q, na.rm = TRUE)
      varHK <- maxq/(sum(1/(sei^2)))
      CImat["mKH_lb", "MHc"] <- thetaMHc-qt(1-(signif/2), df = k-1)*sqrt
    ↪ (varHK)
      CImat["mKH_ub", "MHc"] <- thetaMHc+qt(1-(signif/2), df = k-1)*sqrt
    ↪ (varHK)
    }
  }
  # Round off the CI estimates to specified number of decimal places by
    ↪ signifCI argument
  CImat <- round(CImat, digits = signifCI)
  if (output == TRUE) print(CImat)
  ## Output that can be used after function has been run
  # Create an output frame that can be used when iterating through this
    ↪ function multiple times.
  # The above dataframe is better when only calculating estimates for
    ↪ one meta-analysis
  return(CImat)
}

## Run the confidence interval estimators function ##
# (in loop for each simulated meta-analysis)
ciresults <- data.frame()
for (x in 1:results$meta[nrow(results)]) {
  newresults <- results[which(results$meta == x), ]
  hetests <- hetdata[which(results$meta == x), ]
  logRR <- newresults$logRR
  selogRR <- newresults$selogRR
  ec <- newresults$ec
```

```
  et <- newresults$et
  nc <- newresults$nc
  nt <- newresults$nt
  newmodelthetas <- modelthetas[which(results$meta == x), ]
  CIresults <- CIest(xi = logRR, sei = selogRR, ec = ec, et = et, nc =
    ↪ nc, nt = nt, hetests = hetests, modelthetas = newmodelthetas,
    ↪ CIests = c("Z", "T", "HKSJ", "mKH"), hetnames = hetnames, signif =
    ↪  0.05, signifCI = 4, MH = TRUE, MHc = TRUE, c = 0.5, output =
    ↪ FALSE)
  CIresults <- data.frame(CIresults)
  CIresults$estimate <- row.names(CIresults)
  row.names(CIresults) <- 1:nrow(CIresults)
  keydata <- newresults["meta"]
  keydata <- keydata[1, ]
  keydata <- cbind(keydata, CIresults)
  colnames(keydata)[1] <- "meta"
  ciresults <- rbind(ciresults, keydata)
}

scenario <- names(mydata) %in% c("scenario", "meta", "tau2", "taup2", "
    ↪ alpha", "SampleSize1", "k", "simulation", "theta", "varalpha", "
    ↪ eventdist", "sampdist", "contcorr")
scenario <- mydata[scenario]
ciresults <- merge(unique(scenario), ciresults, by = "meta")

# Save CI estimates data frame to a file
write.table(ciresults, file = paste("CIresults_", iscen, sep = ""))

################################################
# R code for calculating performance measures #
################################################

# Combine these 3 data frames together by the meta-analysis ID variable
results <- merge(estresults, condresults, by = "meta")
results <- merge(results, modresults, by = "meta")
results <- merge(results, mixresults, by = "meta")
resnames <- names(results) %in% c("PMRM_logRR", "CLMRM_logRR", "MM_logRR
    ↪ ")
results <- results[!resnames]
results$MH <- results$MHc <- NA
results$estimate <- "tau2"
```

```
results <- rbind(results, ciresults)

# Create vector/list of heterogeneity variance estimating methods
hetnames <- colnames(results)[colnames(results) %in% c("HO", "DL", "PM",
    ↪ "IPM", "HO2", "DL2", "DLp", "DLb", "HM", "HS", "SJ", "SJ2", "ML",
    ↪ "REML", "AREML", "AB", "RB", "RBO", "BM", "PMRM", "CLMRM", "CO1",
    ↪ "CO2", "CO3", "CO4", "MM")]
thetanames <- colnames(results)[colnames(results) %in% c(hetnames, "MH",
    ↪ "MHc")]

## List of arguments and their meanings ##

# hetdata - data frame containing the results of the tau2
# estimators for each meta-analysis (including GLMMs)

# thetadata - data frame containing the values of theta (logRR)
# calculated using each of the tau2 estimates for each meta-analysis

# cidata - data frame containing the values of the CIs for each
# meta-analysis (split into lower and upper bounds)

# sedata - data frame containing the standard error of the logRR
# for each study

# hetnames - vector of the names of the tau2 estimators applied
# (including GLMMs)

# thetanames - vector of the names of the theta estimators applied
# (hetnames plus MH, if MH was included)

# cinames - vector of the names of the CIs applied (split into
# lower and upper bounds)

# measure - vector of performance measures that you would like to be
# calculated for each of the tau2 and CI estimates. The default is
# NULL, which means all performance measures are calculated.

# logRR - TRUE if the conditional-based estimates were transformed (via
    ↪ an
# approximation) to the variance of the logRR, FALSE if the conditional-
    ↪ based
```

```
# estimates were left as the variance of the RR.

# signifmes - number of significant figures to round performance
# measure results

# output - TRUE if output is to be displayed, FALSE otherwise

## Function for calculating performance measures for the tau2 estimators
    ↪  ##

perform <- function(hetdata = hetdata, thetadata = thetadata, cidata =
    ↪ cidata, sedata = sedata, hetnames = hetnames, thetanames =
    ↪ thetanames, measure = NULL, logRR = TRUE, signifmes = 6, output =
    ↪ TRUE) {
  if (is.null(measure)) measure <- c("bias", "medbias", "mse", "medmse",
    ↪  "propzero", "biastheta", "msetheta", "coverage", "power", "
    ↪ meanerror", "varerror")
  ciests <- NULL
  for (i in c("Z", "T", "HKSJ", "mKH")) {
    if (paste(i, "_lb", sep = "") %in% cidata$estimate) {
      ciests <- c(ciests, i)
    }
  }
  for (i in c("coverage", "power", "meanerror", "varerror")) {
    if (i %in% measure) {
      measure <- measure[!measure %in% i]
      for (j in ciests) {
        measure <- c(measure, paste(i, "_", j, sep = ""))
      }
    }
  }
  if ("MH" %in% thetanames & "MHc" %in% thetanames) {
    permat <- matrix(NA, nrow = length(measure), ncol = length(hetnames)
    ↪ +2)
    colnames(permat) <- c(hetnames, "MH", "MHc")
  } else if ("MH" %in% thetanames) {
    permat <- matrix(NA, nrow = length(measure), ncol = length(hetnames)
    ↪ +1)
    colnames(permat) <- c(hetnames, "MH")
  } else if ("MHc" %in% thetanames) {
```

```
  permat <- matrix(NA, nrow = length(measure), ncol = length(hetnames)
 ↪ +1)
  colnames(permat) <- c(hetnames, "MHc")
} else {
  permat <- matrix(NA, nrow = length(measure), ncol = length(hetnames)
 ↪ )
  colnames(permat) <- c(hetnames)
}
rownames(permat) <- measure
tau2 <- sedata$tau2[1]
taup2 <- sedata$taup2[1]
theta <- sedata$theta[1]
sei <- sedata$selogRR
k <- sedata$k[1]
# Bias
if ("bias" %in% measure) {
  for (i in 1:length(hetnames)) {
    if (logRR == FALSE & (hetnames[i] == "CO1" | hetnames[i] == "CO2"
 ↪ | hetnames[i] == "CO3" | hetnames[i] == "CO4")) {
      permat["bias", hetnames[i]] <- mean(hetdata[ ,i], na.rm = TRUE)
 ↪ - taup2
    } else {
      permat["bias", hetnames[i]] <- mean(hetdata[ ,i], na.rm = TRUE)
 ↪ - tau2
    }
  }
}
# Median bias
if ("medbias" %in% measure) {
  for (i in 1:length(hetnames)) {
    if (logRR == FALSE & (hetnames[i] == "CO1" | hetnames[i] == "CO2"
 ↪ | hetnames[i] == "CO3" | hetnames[i] == "CO4")) {
      permat["medbias", hetnames[i]] <- median(hetdata[ ,i], na.rm =
 ↪ TRUE) - taup2
    } else {
      permat["medbias", hetnames[i]] <- median(hetdata[ ,i], na.rm =
 ↪ TRUE) - tau2
    }
  }
}
# Mean squared error
```

```
if ("mse" %in% measure) {
  for (i in 1:length(hetnames)) {
    if (logRR == FALSE & (hetnames[i] == "CO1" | hetnames[i] == "CO2"
 ↪ | hetnames[i] == "CO3" | hetnames[i] == "CO4")) {
      permat["mse", hetnames[i]] <- mean((hetdata[ ,i]-taup2)^2, na.rm
 ↪  = TRUE)
    } else {
      permat["mse", hetnames[i]] <- mean((hetdata[ ,i]-tau2)^2, na.rm
 ↪ = TRUE)
    }
  }
}
# Median mean squared error
if ("medmse" %in% measure) {
  for (i in 1:length(hetnames)) {
    if (logRR == FALSE & (hetnames[i] == "CO1" | hetnames[i] == "CO2"
 ↪ | hetnames[i] == "CO3" | hetnames[i] == "CO4")) {
      permat["medmse", hetnames[i]] <- median((hetdata[ ,i]-taup2)^2,
 ↪ na.rm = TRUE)
    } else {
      permat["medmse", hetnames[i]] <- median((hetdata[ ,i]-tau2)^2,
 ↪ na.rm = TRUE)
    }
  }
}
# Proportion of zero estimates
if ("propzero" %in% measure) {
  for (i in 1:length(hetnames)) {
    if (all(is.na(hetdata[ ,i]))) {
      permat["propzero", hetnames[i]] <- NA
    } else {
      permat["propzero", hetnames[i]] <- sum(hetdata[ ,i] == 0 & !is.
 ↪ na(hetdata[ ,i]))/(nrow(hetdata) - sum(is.na(hetdata[ ,i])))
    }
  }
}
# Bias of theta
if ("biastheta" %in% measure) {
  for (i in 1:length(thetanames)) {
    permat["biastheta", thetanames[i]] <- mean(thetadata[ ,i], na.rm =
 ↪  TRUE) - theta
```

```
  }
}
# Mean squared error of theta
if ("msetheta" %in% measure) {
  for (i in 1:length(thetanames)) {
    permat["msetheta", thetanames[i]] <- mean((thetadata[ ,i] - theta)
 ↪ ^2, na.rm = TRUE)
  }
}
# Coverage
if (any(grepl("coverage", measure))) {
  for (z in ciests) {
    lb <- cidata[cidata$estimate == paste(z, "_lb", sep = ""), ]
    ub <- cidata[cidata$estimate == paste(z, "_ub", sep = ""), ]
    for (i in 1:length(thetanames)) {
      if (all(is.na(lb[ ,i])) | all(is.na(ub[ ,i]))) {
        permat[paste("coverage_", z, sep = ""), thetanames[i]] <- NA
      } else {
        permat[paste("coverage_", z, sep = ""), thetanames[i]] <- (sum
 ↪ (!is.na(lb[ ,i]) & !is.na(ub[ ,i]) & lb[ ,i] <= theta & ub[ ,i] >=
 ↪  theta)/nrow(lb))*100
      }
    }
  }
}
# Power
if (any(grepl("power", measure))) {
  for (z in ciests) {
    lb <- cidata[cidata$estimate == paste(z, "_lb", sep = ""), ]
    ub <- cidata[cidata$estimate == paste(z, "_ub", sep = ""), ]
    for (i in 1:length(thetanames)) {
      if (all(is.na(lb[ ,i])) | all(is.na(ub[ ,i]))) {
        permat[paste("power_", z, sep = ""), thetanames[i]] <- NA
      } else {
        permat[paste("power_", z, sep = ""), thetanames[i]] <- (sum(!
 ↪ is.na(lb[ ,i]) & !is.na(ub[ ,i]) & ((ub[ ,i]-lb[ ,i])/2) < 2)/nrow
 ↪ (lb))*100
      }
    }
  }
}
```

```
# Mean error
if (any(grepl("meanerror", measure))) {
  for (z in ciests) {
    lb <- cidata[cidata$estimate == paste(z, "_lb", sep = ""), ]
    ub <- cidata[cidata$estimate == paste(z, "_ub", sep = ""), ]
    error <- matrix(NA, nrow = nrow(lb), ncol = length(thetanames))
    for (i in 1:length(thetanames)) {
      if (all(is.na(lb[ ,i])) | all(is.na(ub[ ,i]))) {
        permat[paste("meanerror_", z, sep = ""), thetanames[i]] <- NA
      } else {
        for (j in 1:nrow(lb)) {
          seii <- sei[(((j-1)*k)+1):(j*k)]
          error[j,i] <- (ub[j,i] - lb[j,i])/(3.92*sqrt(1/sum(1/(tau2+
↪ seii))))
        }
        permat[paste("meanerror_", z, sep = ""), thetanames[i]] <-
↪ mean(error[ ,i], na.rm = TRUE)
      }
    }
  }
}
# Variance error
if (any(grepl("varerror", measure))) {
  for (z in ciests) {
    lb <- cidata[cidata$estimate == paste(z, "_lb", sep = ""), ]
    ub <- cidata[cidata$estimate == paste(z, "_ub", sep = ""), ]
    error <- matrix(NA, nrow = nrow(lb), ncol = length(thetanames))
    for (i in 1:length(thetanames)) {
      if (all(is.na(lb[ ,i])) | all(is.na(ub[ ,i]))) {
        permat[paste("varerror_", z, sep = ""), thetanames[i]] <- NA
      } else {
        for (j in 1:nrow(lb)) {
          seii <- sei[(((j-1)*k)+1):(j*k)]
          error[j,i] <- (ub[j,i] - lb[j,i])/(3.92*sqrt(1/sum(1/(tau2+
↪ seii))))
        }
        permat[paste("varerror_", z, sep = ""), thetanames[i]] <- var(
↪ error[ ,i], na.rm = TRUE)
      }
    }
  }
```

```
  }
  # Round off the performance measures to specified number of decimal
    ↪ places by signifmes argument
  permat <- round(permat, digits = signifmes)
  if (output == TRUE) print(permat)
  ## Output that can be used after function has been run
  # Create an output frame that can be used when iterating through this
    ↪ function multipe times.
  # The above dataframe is better when only calculating estimates for
    ↪ one meta-analysis
  return(permat)
  # We can refer to the estimates outside of this function by <funct
    ↪ name>$<est name>
}

## Apply the performance measures function ##
# (in loop for each scenario)
permes <- data.frame()
for (x in results$scenario[1]:results$scenario[nrow(results)]) {
  newresults <- results[which(results$scenario == x), ]
  hetdata <- newresults[newresults$estimate == "tau2", ]
  hetdata <- hetdata[hetnames]
  thetadata <- newresults[newresults$estimate == "theta", ]
  thetadata <- thetadata[c(hetnames, "MH", "MHc")]
  cidata <- newresults[!newresults$estimate %in% c("tau2","theta"), ]
  cidata <- cidata[c(hetnames, "MH", "MHc", "estimate")]
  sedata <- mydata[which(mydata$scenario == x), ]
  perms <- perform(hetdata = hetdata, thetadata = thetadata, cidata =
    ↪ cidata, sedata = sedata, hetnames = hetnames, thetanames =
    ↪ thetanames, measure = c("bias", "medbias", "mse", "medmse", "
    ↪ propzero", "biastheta", "msetheta", "coverage", "power", "
    ↪ meanerror", "varerror"), logRR = TRUE, signifmes = 6, output =
    ↪ FALSE)
  perms <- data.frame(perms)
  perms$measure <- row.names(perms)
  row.names(perms) <- 1:nrow(perms)
  keydata <- newresults["scenario"]
  keydata <- keydata[1, ]
  keydata <- cbind(keydata, perms)
  colnames(keydata)[1] <- "scenario"
  if (x >= results$scenario[1] + 1) {
```

```
    names(keydata) <- names(permes) # to ensure the names of the
    ↪ dataframes are the same
  }
  permes <- rbind(permes, keydata)
}

# Save performance measures data frame to a file
write.table(permes, file = paste("permes_results_", iscen, sep = ""))

# Graphs will be divided by a number of characteristic parameters
# including the level of heterogeneity (I^2) given the true tau^2.
# Calculate I^2 for each simulated meta-analysis
I2 <- I2p <- NULL
for (i in 1:mydata$meta[nrow(mydata)]) {
  newdata <- mydata[which(mydata$meta == i), ]
  k <- newdata$k[1]
  sei <- newdata$selogRR
  vari <- sei^2
  tau2 <- newdata$tau2[1]
  taup2 <- newdata$taup2[1]
  sigma2 <- ((k-1)*sum(1/vari))/(((sum(1/vari))^2)-(sum((1/vari)^2)))
  I2[i] <- (tau2/(tau2+sigma2))*100
  I2p[i] <- (taup2/(taup2+sigma2))*100
}

# Calculate the mean I^2 value for each scenario
scenario <- seq(1, length(I2), by = max(mydata$simulation))
j <- 0
I2mean <- I2pmean <- NULL
for (i in scenario) {
  newI2 <- I2[i:(i+(max(mydata$simulation)-1))]
  newI2p <- I2p[i:(i+(max(mydata$simulation)-1))]
  j <- j + 1
  I2mean[j] <- mean(newI2, na.rm = TRUE)
  I2pmean[j] <- mean(newI2p, na.rm = TRUE)
}
# Round the mean I^2 to a whole number to make classification easier
I2mean <- round(I2mean, digits = 0)
I2pmean <- round(I2pmean, digits = 0)

# Remove unnecessary variables that vary within scenarios
```

```
myvars <- names(mydata) %in% c("meta", "SampleSize", "simulation", "
    ↪ study", "p0", "p1", "et", "ec", "nt", "nc", "etcor", "eccor", "
    ↪ ntcor", "nccor", "logRR", "selogRR")
cutdata <- mydata[!myvars]

# Produce new data frame with one row for each scenario
results <- data.frame()
for (i in cutdata$scenario[1]:cutdata$scenario[nrow(cutdata)]) {
  newdata <- cutdata[which(cutdata$scenario == i), ]
  newdata <- newdata[1, ]
  results <- rbind(results, newdata)
}

# Combine this data frame of scenario parameter characteristics
# with the performance measures data frame
results <- merge(results, permes, by = "scenario")

# Add mean I^2 to results data frame
results$I2 <- I2mean
results$I2p <- I2pmean

# Save results
write.table(results, file = paste("results_", iscen, sep = ""))
```

D.2 Definition of performance measures

TABLE D.1: Equations for performance measures used in simulation study; $\hat{\boldsymbol{\tau}}^{\mathbf{2}} = (\hat{\tau}_1^2, \ldots, \hat{\tau}_N^2)$, $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_N)$, N is the number of simulations, and $CI_{upper,\theta}$ and $CI_{lower,\theta}$ are the upper and lower bounds of the confidence interval for θ respectively.

Performance measure	**Equation/Definition**
Mean absolute bias of $\hat{\tau}^2$	$mean(\hat{\boldsymbol{\tau}}^{\mathbf{2}}) - \tau^2$
Median absolute bias of $\hat{\tau}^2$	$median(\hat{\boldsymbol{\tau}}^{\mathbf{2}}) - \tau^2$
Mean squared error of $\hat{\tau}^2$	$mean\left[(\hat{\boldsymbol{\tau}}^{\mathbf{2}} - \tau^2)^2\right]$
Median squared error of $\hat{\tau}^2$	$median\left[(\hat{\boldsymbol{\tau}}^{\mathbf{2}} - \tau^2)^2\right]$
Proportion of zero estimates of τ^2	Proportion of meta-analyses meeting $\hat{\tau}^2 = 0$
Mean absolute bias of $\hat{\theta}$	$mean(\hat{\boldsymbol{\theta}}) - \theta$
Median absolute bias of $\hat{\theta}$	$median(\hat{\boldsymbol{\theta}}) - \theta$
Mean squared error of $\hat{\theta}$	$mean\left[(\hat{\boldsymbol{\theta}} - \theta)^2\right]$
Median squared error of $\hat{\theta}$	$median\left[(\hat{\boldsymbol{\theta}} - \theta)^2\right]$
Coverage	Percentage of meta-analyses meeting $CI_{lower,\theta} \leq \theta$ and $CI_{upper,\theta} \geq \theta$
Power	Percentage of meta-analyses meeting $\frac{CI_{upper,\theta} - CI_{lower,\theta}}{2} < c$ with $c = 2$
Error	$\frac{CI_{upper,\theta} - CI_{lower,\theta}}{3.92\sqrt{1/\sum_{i=1}^{k}(1/(\tau^2 + \sigma_i^2))}}$

Appendix E

Further simulation study results

E.1 Bias of τ^2

E.1.1 Examples without omitting outlying estimators

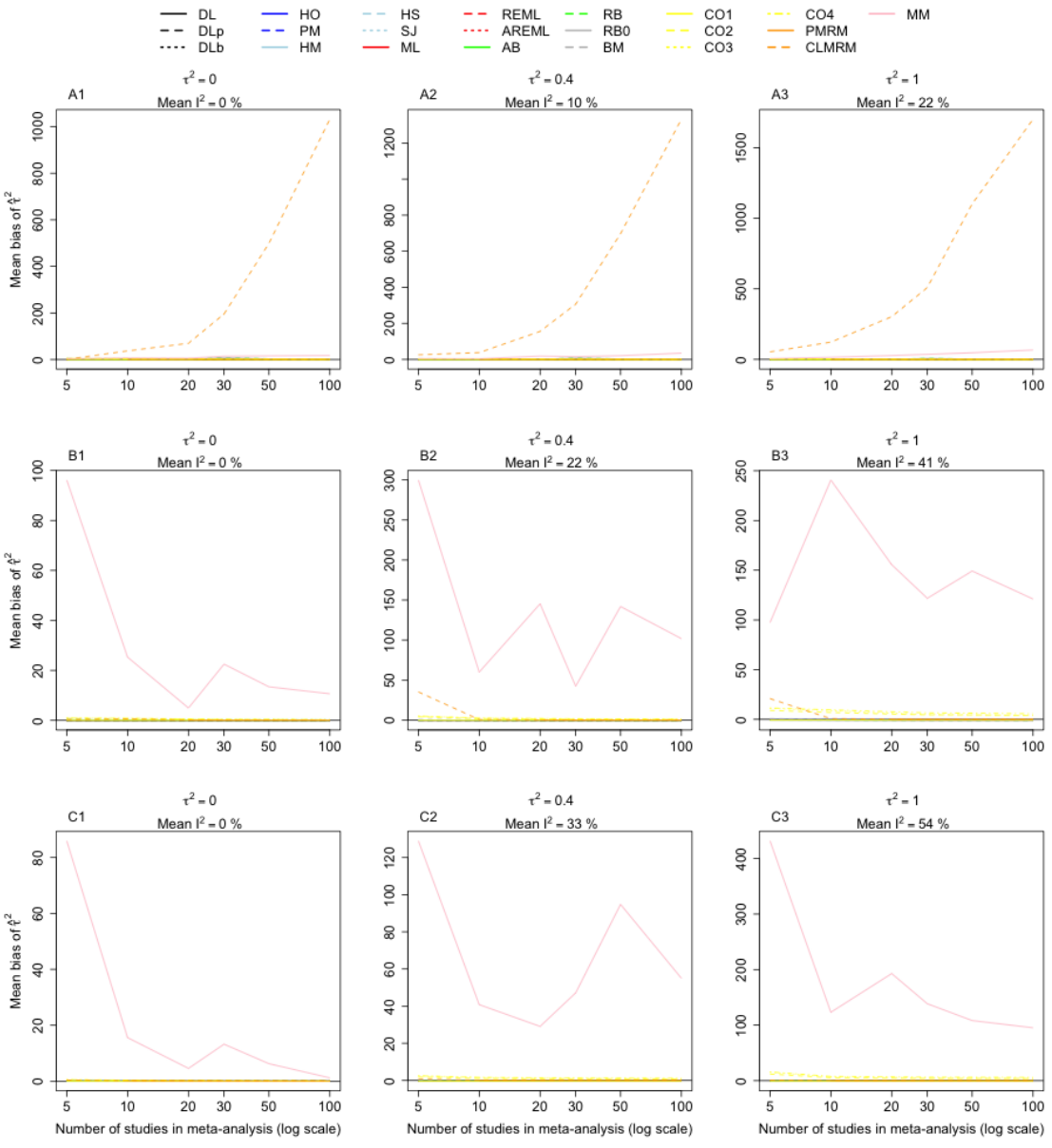


FIGURE E.1: Mean bias of heterogeneity variance estimates in very rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

E.1.2 Alternate values of heterogeneity variance

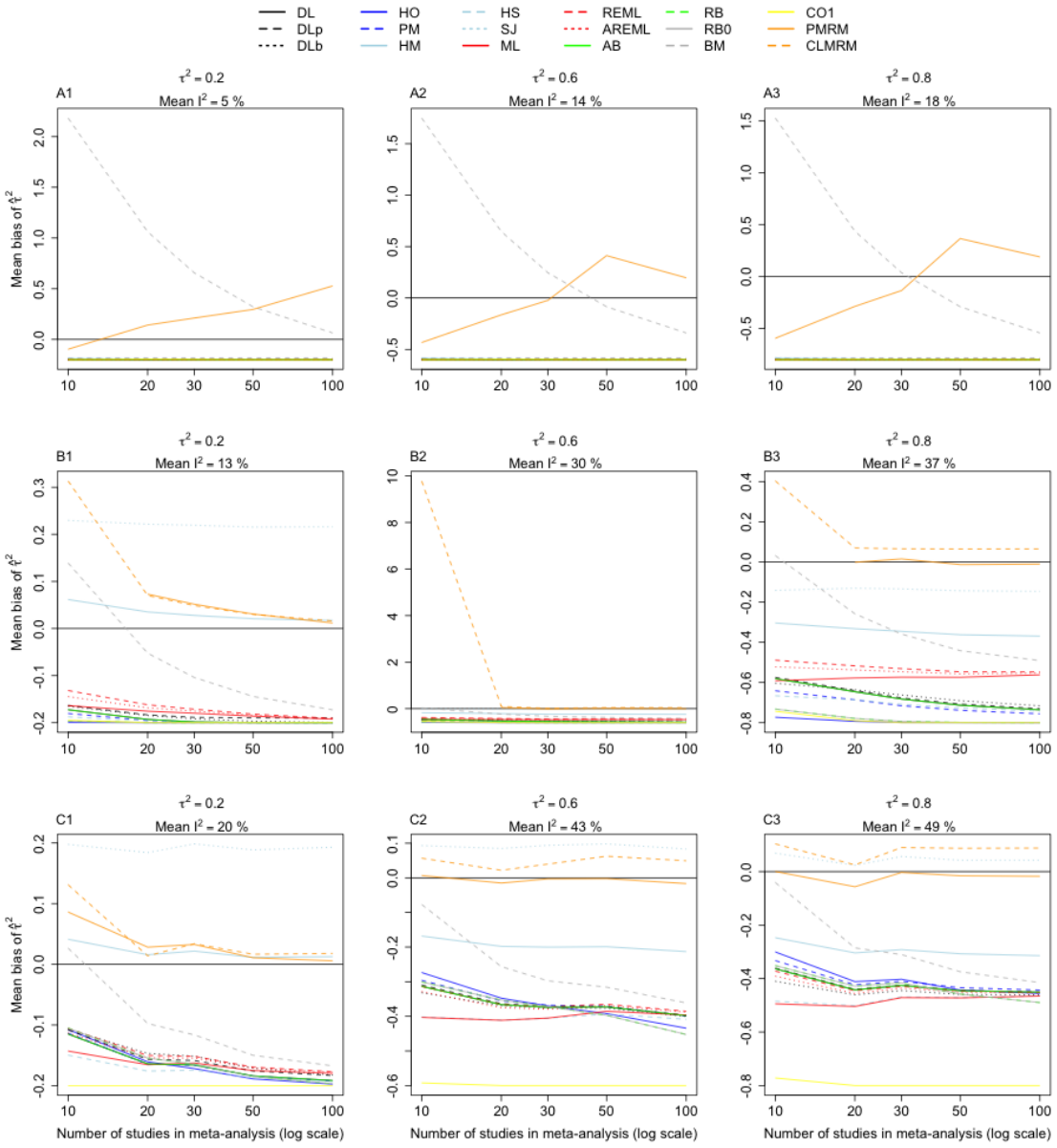


FIGURE E.2: Mean bias of heterogeneity variance estimates in very rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). RB, RB0 and CLMRM have been omitted from A1-A3; CO2, CO3, CO4 and MM have been omitted from all.

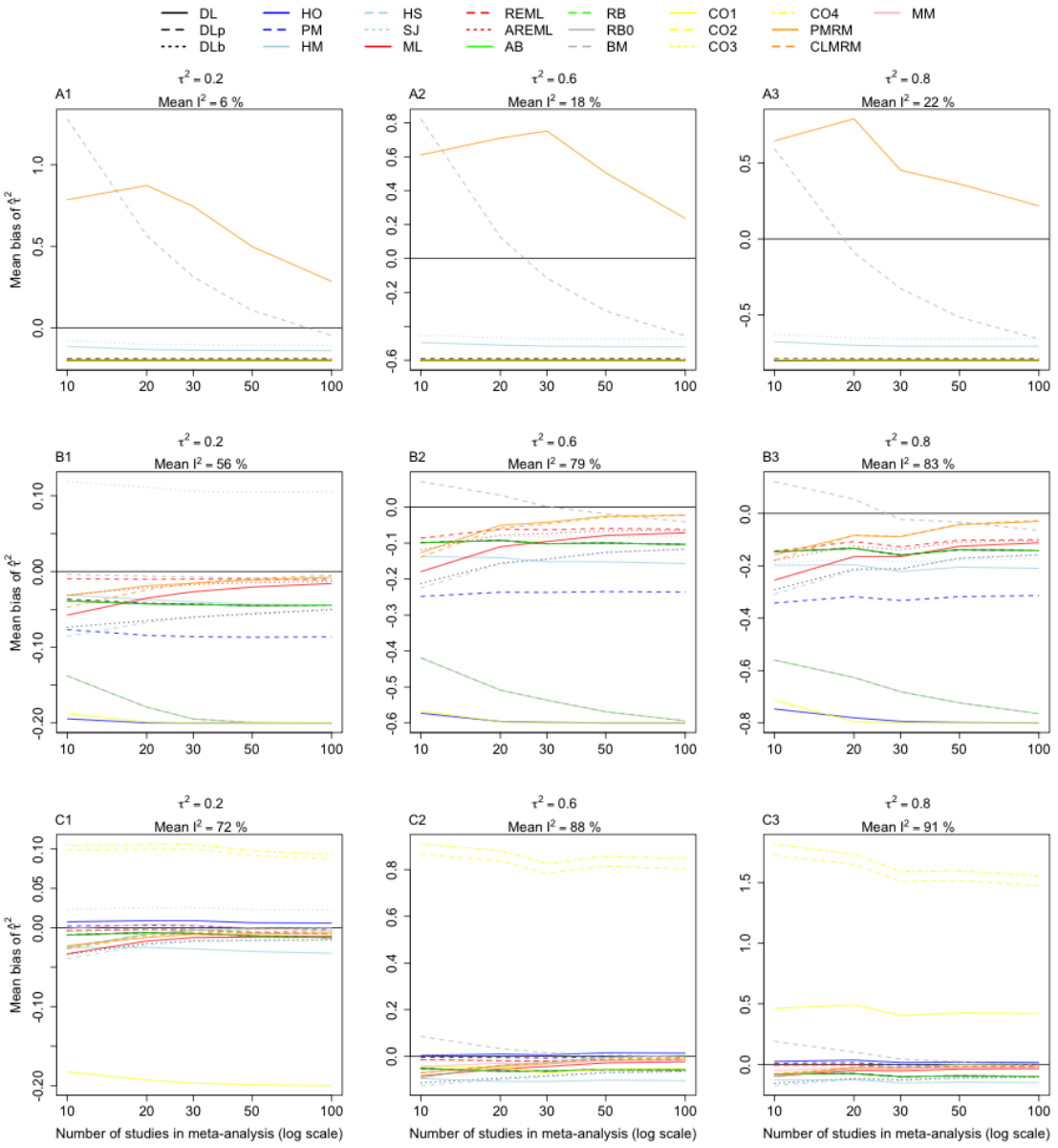


FIGURE E.3: Mean bias of heterogeneity variance estimates in rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). RB, RB0 and CLMRM have been omitted from A1-A3; CO2, CO3, CO4 and MM have been omitted from A1-B3.

E.1.3 Alternate study sample sizes

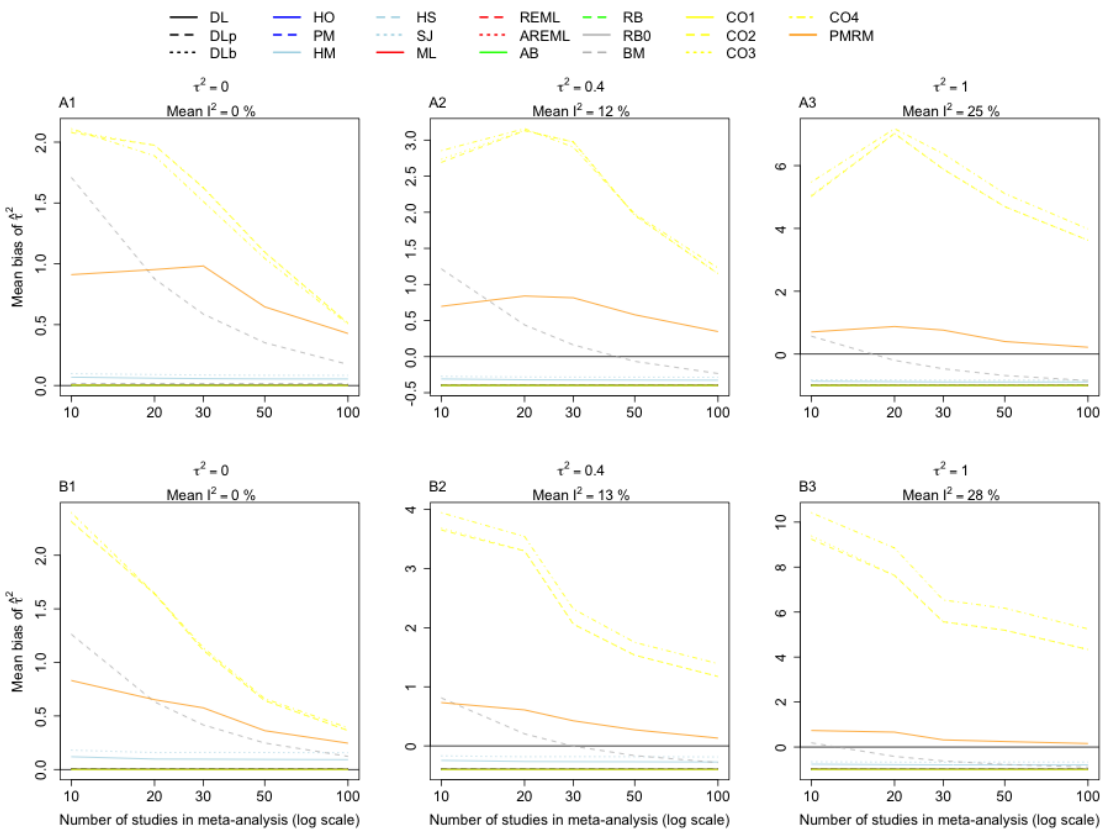


FIGURE E.4: Mean bias of heterogeneity variance estimates in very rare events scenario with $p_0 < p_1$; sample sizes are small-to-medium (A1-A3) and medium (B1-B3). CLMRM and MM have been omitted from all.

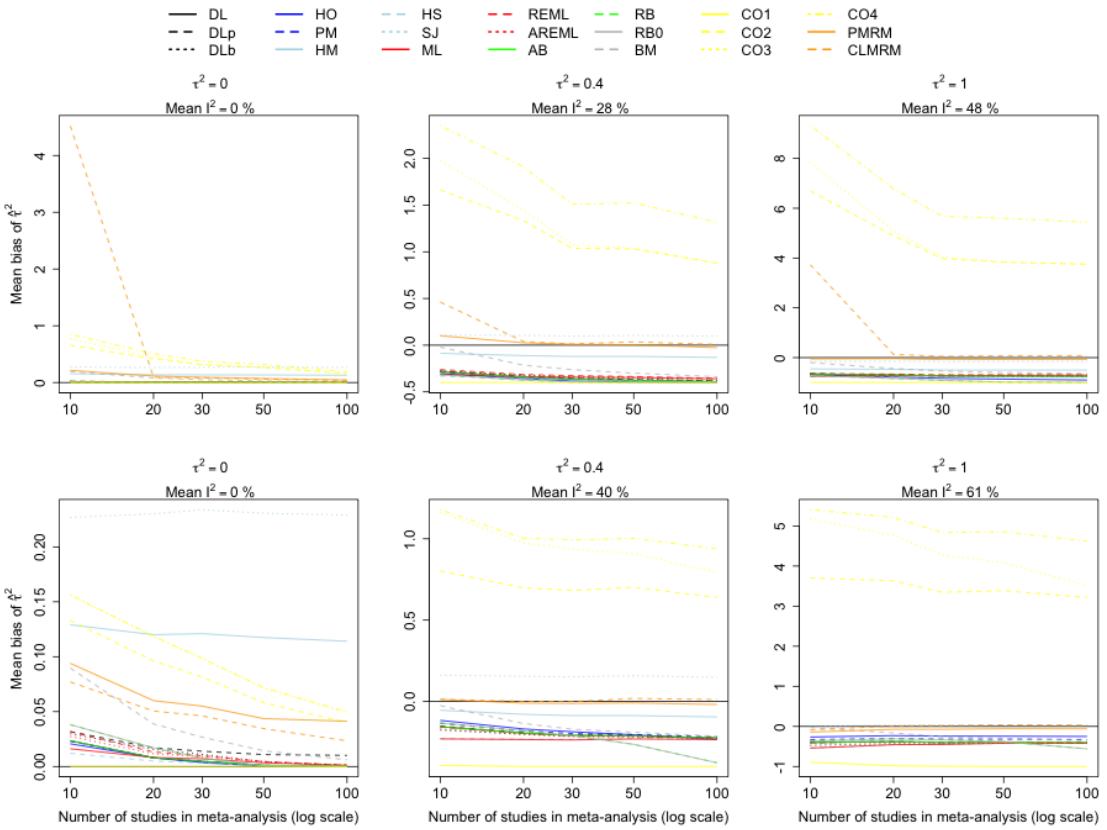


FIGURE E.5: Mean bias of heterogeneity variance estimates in rare events scenario with $p_0 < p_1$; sample sizes are small-to-medium (A1-A3) and medium (B1-B3). MM is omitted from all.

E.1.4 Alternate values of σ_α^2

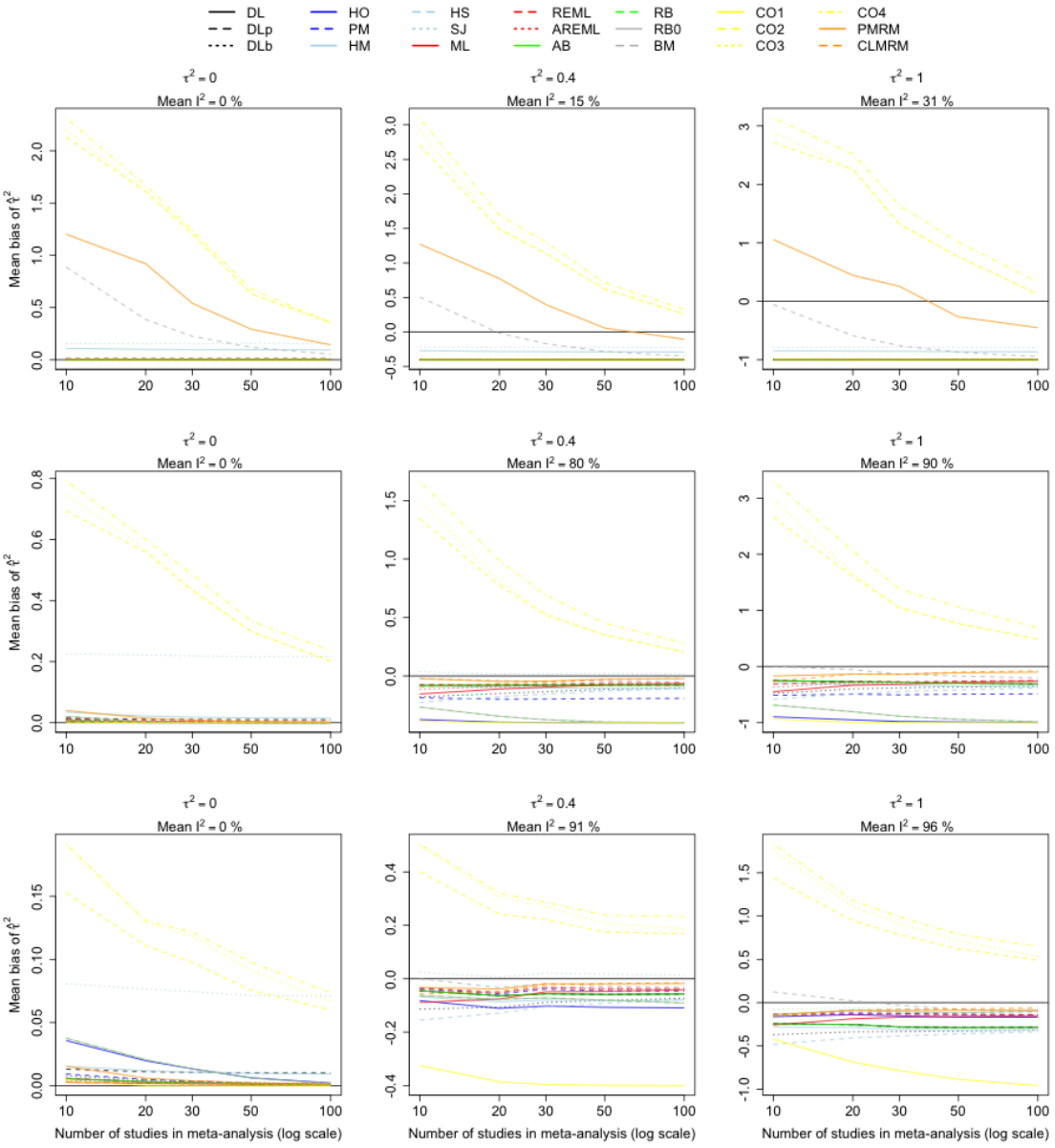


FIGURE E.6: Mean bias of heterogeneity variance estimates in very rare events scenario with $p_0 < p_1$ and $\sigma_\alpha^2 = 3$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). RB, RB0 and CLMRM are omitted from A1-A3; MM is omitted from all.

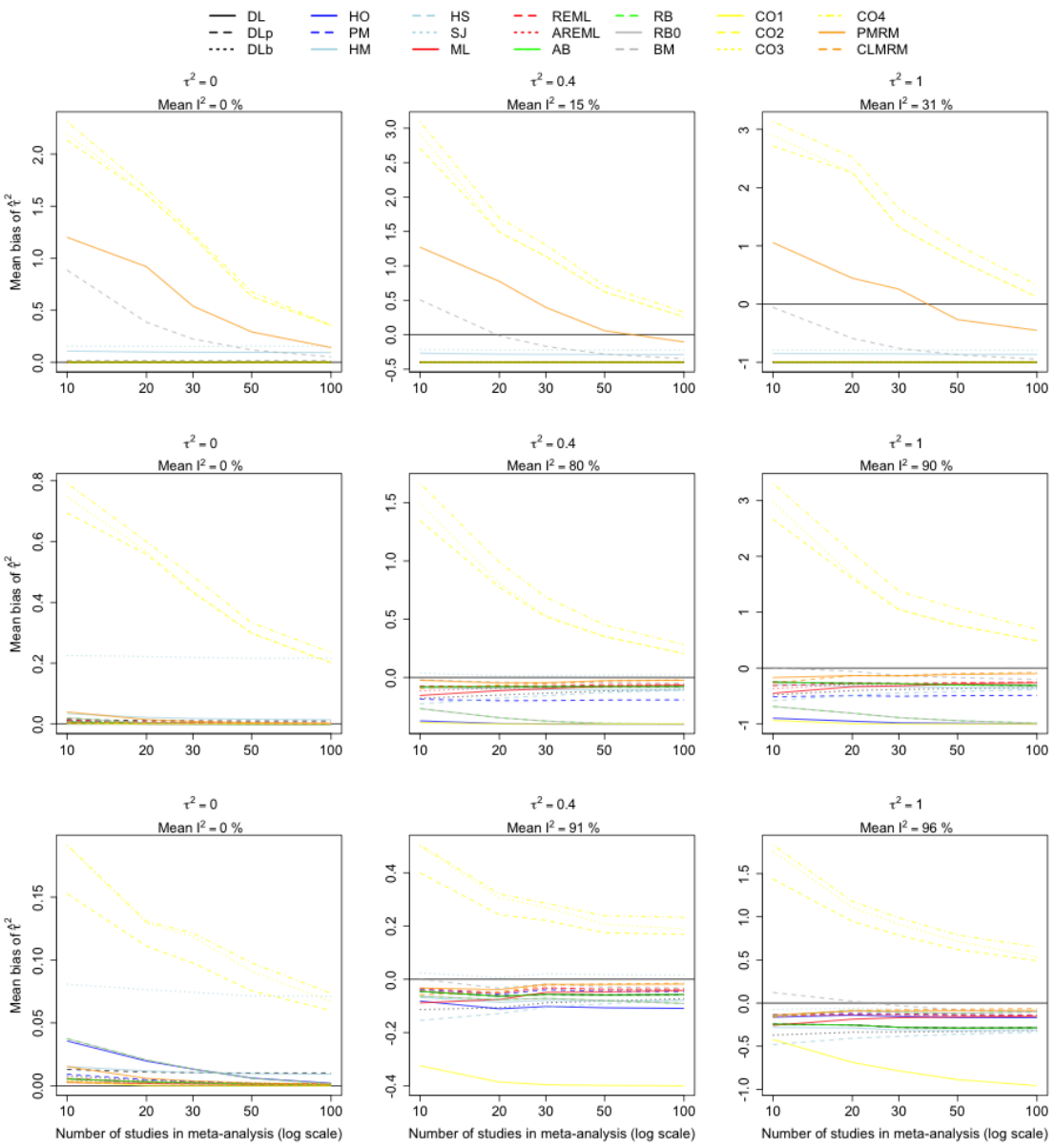


FIGURE E.7: Mean bias of heterogeneity variance estimates in rare events scenario with $p_0 < p_1$ and $\sigma_\alpha^2 = 3$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). RB, RB0 and CLMRM are omitted from A1-A3; MM is omitted from all.

E.1.5 Alternate probability scenarios

Alternate rare events scenarios

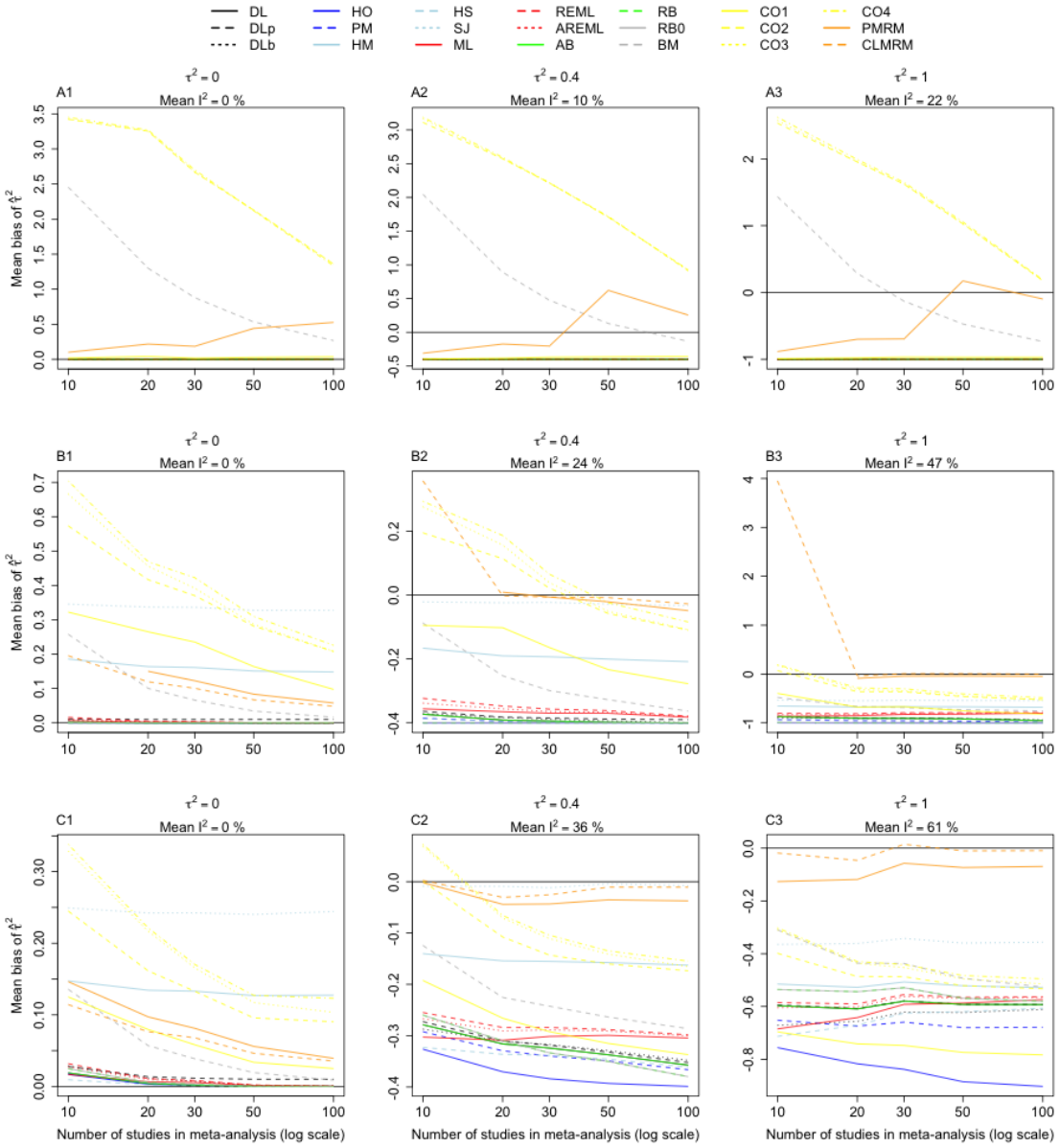


FIGURE E.8: Mean bias of heterogeneity variance estimates in very rare events scenario with $p_0 > p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). RB, RB0 and CLMRM are omitted from A1-A3; MM is omitted from all.

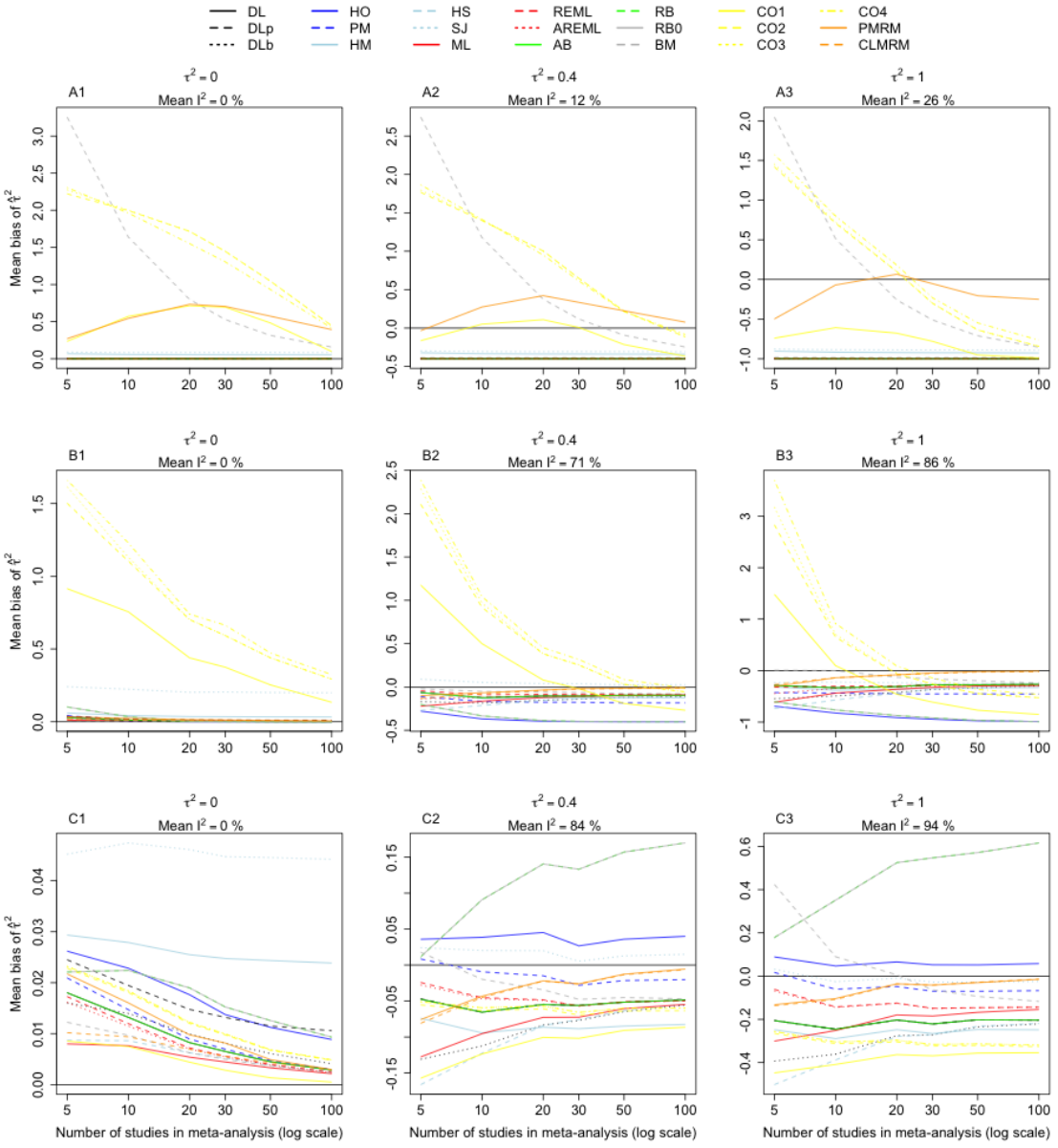


FIGURE E.9: Mean bias of heterogeneity variance estimates in rare events scenario with $p_0 > p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). RB, RB0 and CLMRM are omitted from A1-A3; MM is omitted from all.

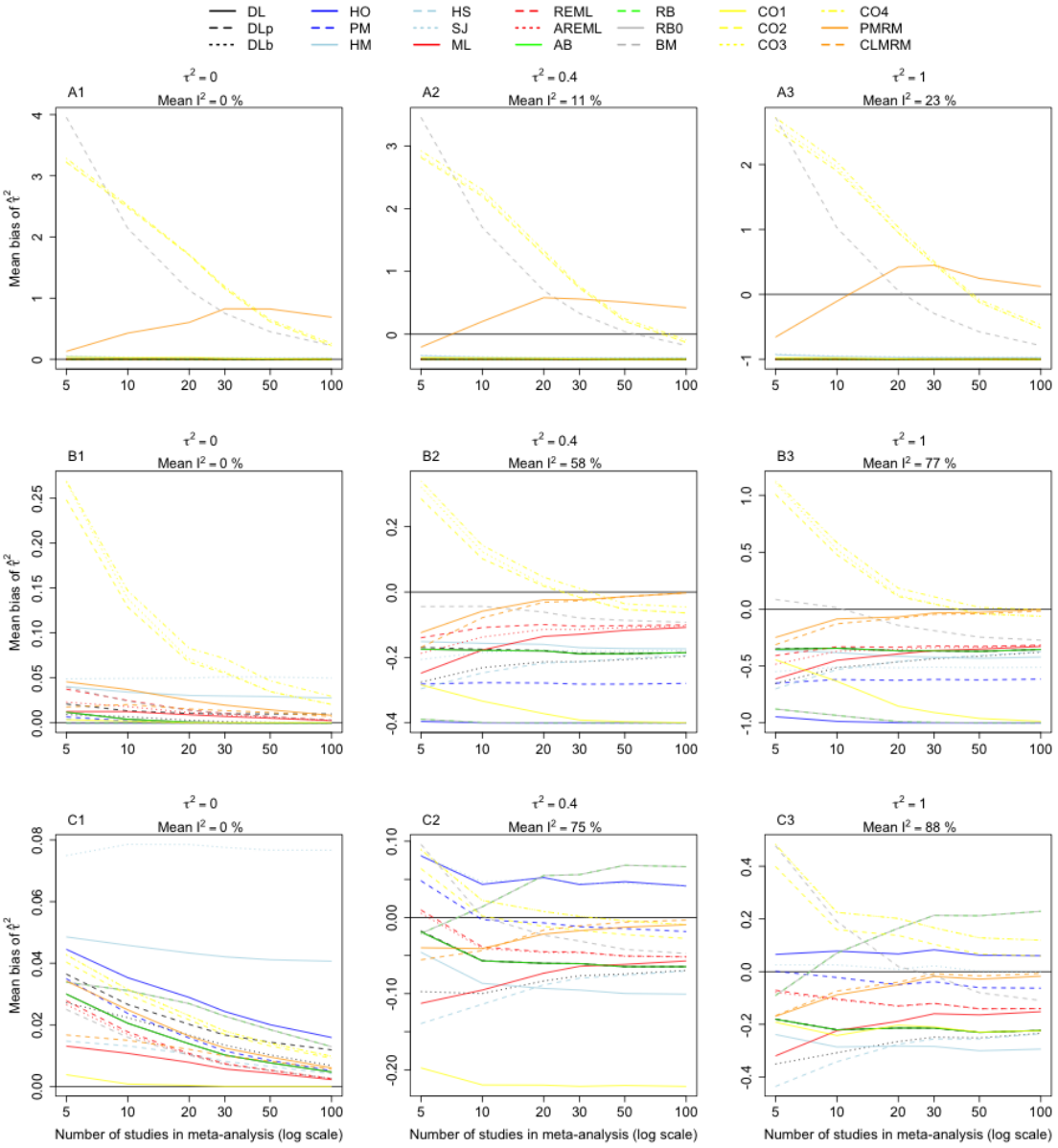


FIGURE E.10: Mean bias of heterogeneity variance estimates in rare events scenario with $p_0 = p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). RB, RB0 and CLMRM are omitted from A1-A3; MM is omitted from all.

Common probability scenarios

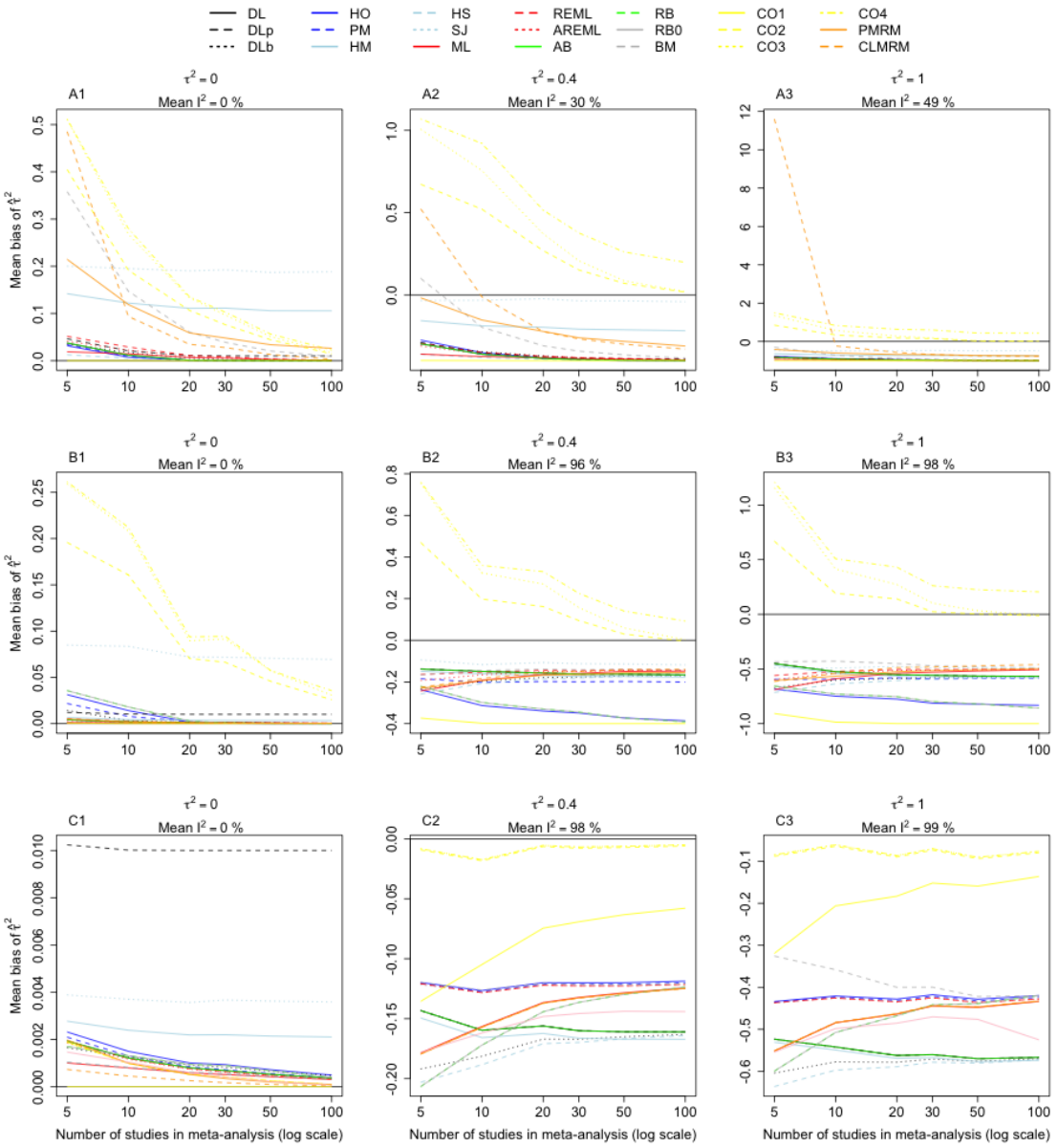


FIGURE E.11: Mean bias of heterogeneity variance estimates in common probability scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). RB and RB0 are omitted from A1-A3; MM is omitted from A1-B3.

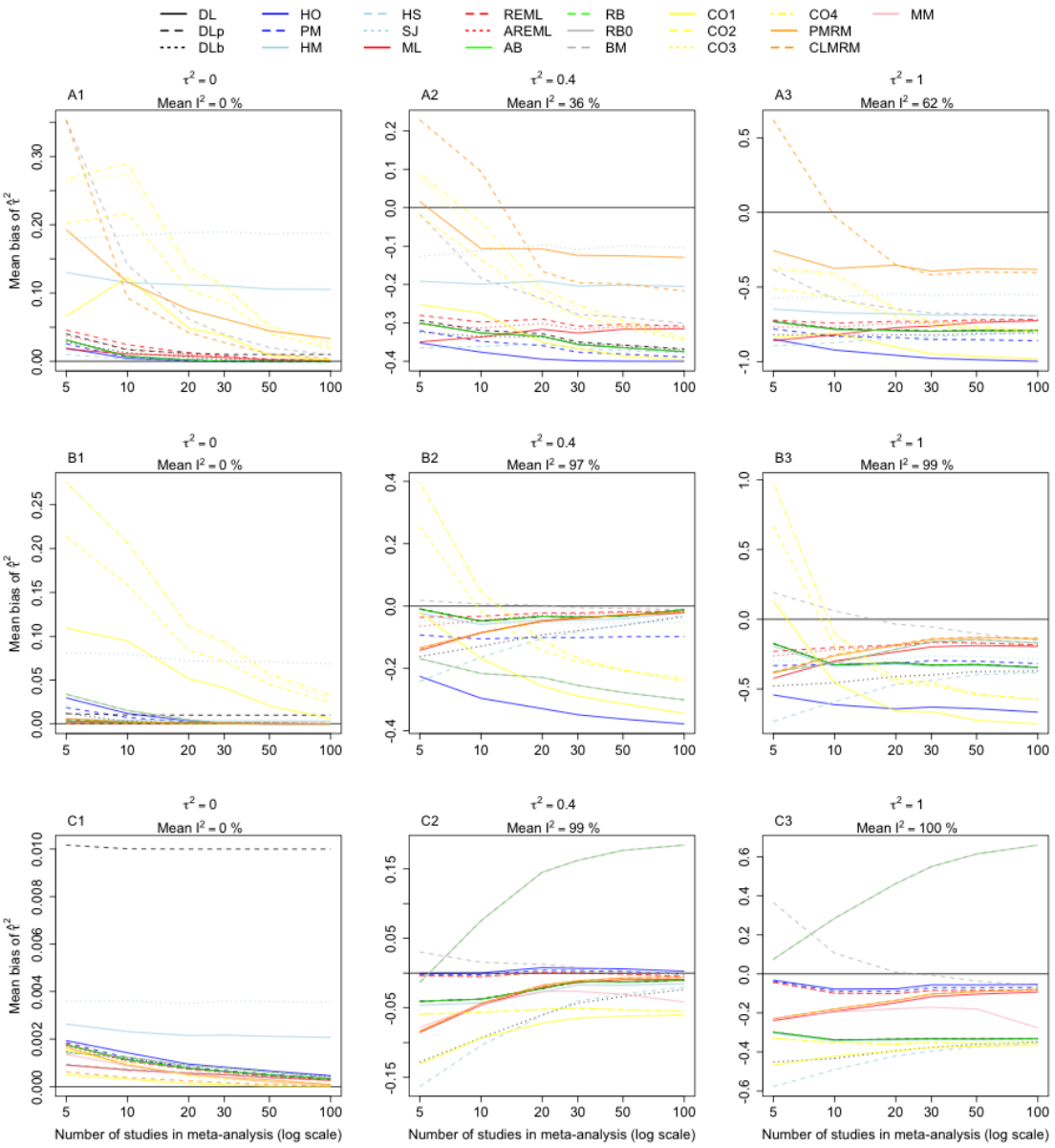


FIGURE E.12: Mean bias of heterogeneity variance estimates in common probability scenario with $p_0 > p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). RB and RB0 are omitted from A1-A3; MM is omitted from A1-B3.

E.1.6 Alternate sampling in simulation study

Alternate event count sampling

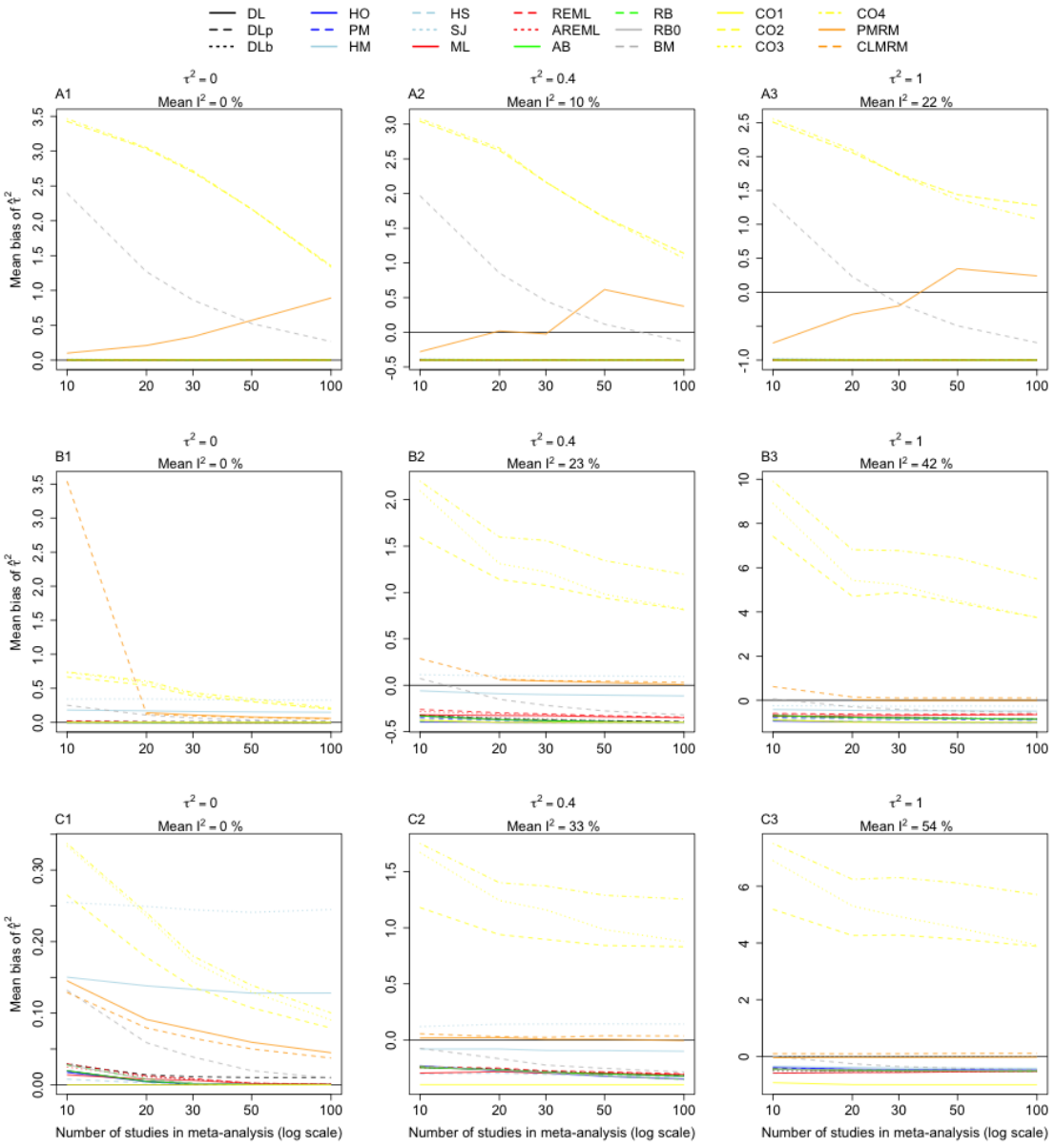


FIGURE E.13: Mean bias of heterogeneity variance estimates in very rare events scenario with $p_0 < p_1$ and poisson event sampling; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). RB, RB0 and CLMRM are omitted from A1-A3; MM is omitted from all.

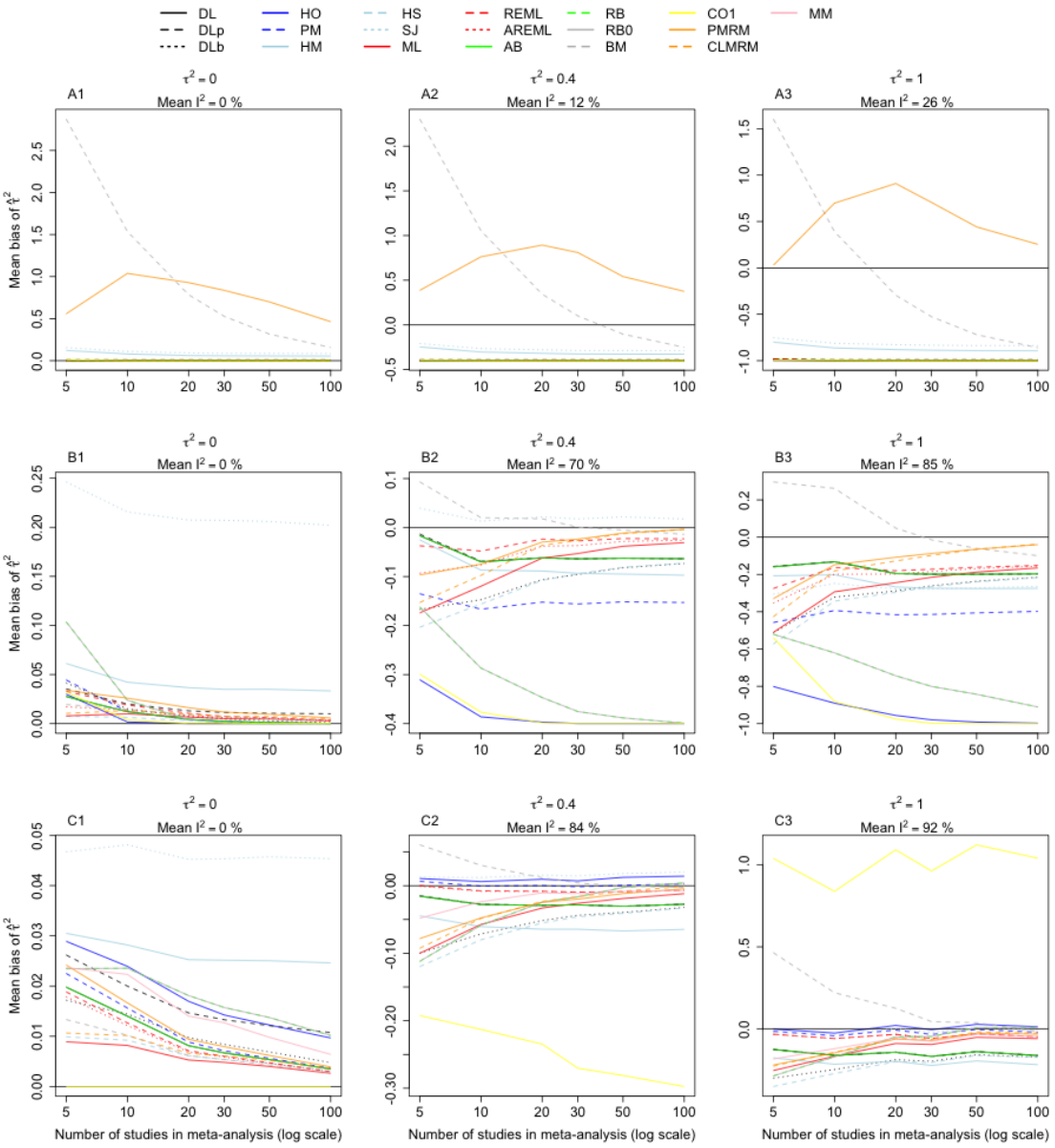


FIGURE E.14: Mean bias of heterogeneity variance estimates in rare events scenario with $p_0 < p_1$ and poisson event sampling; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). RB, RB0 and CLMRM are omitted from A1-A3; MM is omitted from A1-B3; CO2, CO3 and CO4 are omitted from all.

Alternate sample size sampling

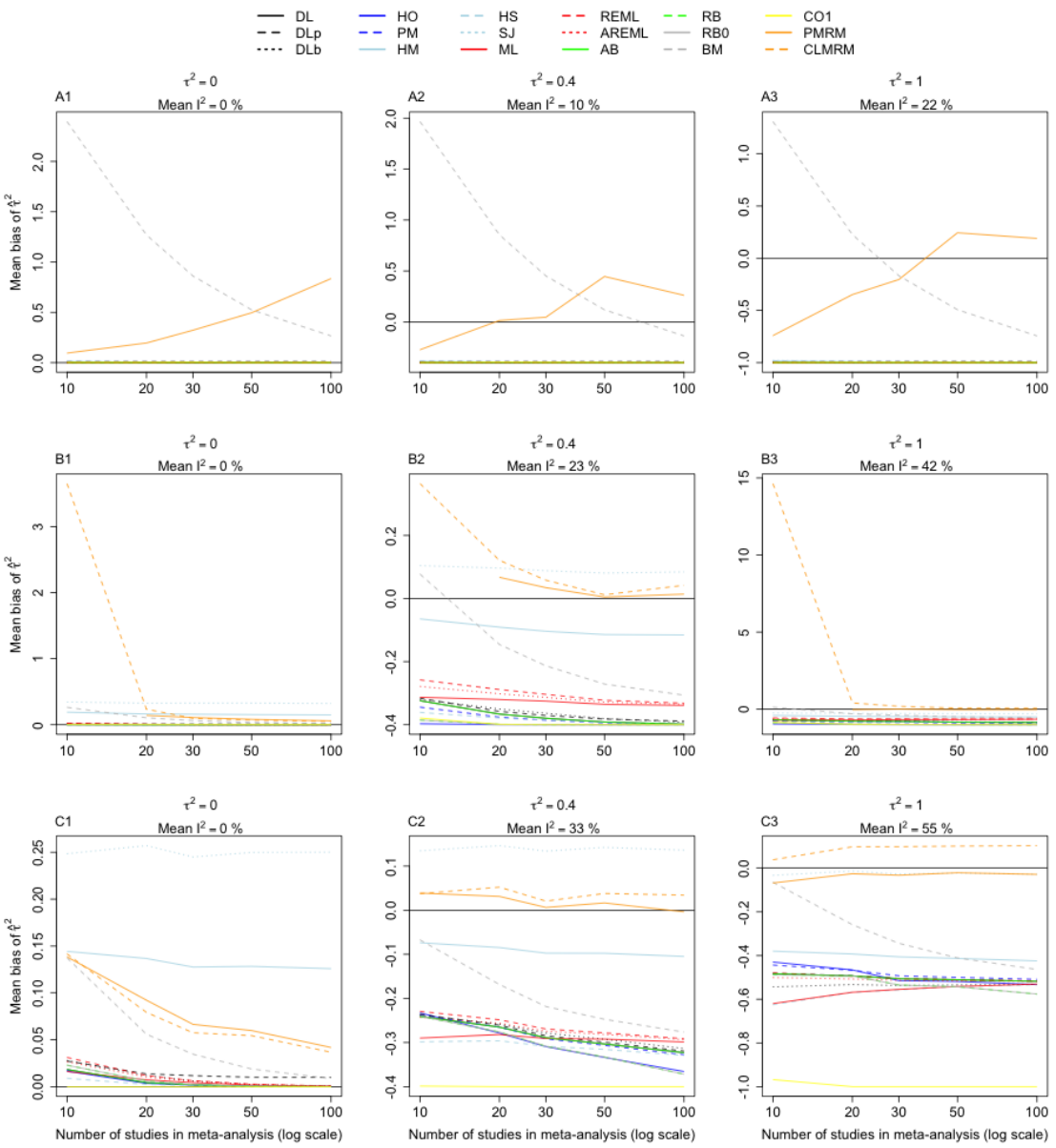


FIGURE E.15: Mean bias of heterogeneity variance estimates in very rare events scenario with $p_0 < p_1$ and normal sample size sampling; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). RB, RB0 and CLMRM are omitted from A1-A3; MM, CO2, CO3 and CO4 are omitted from all.

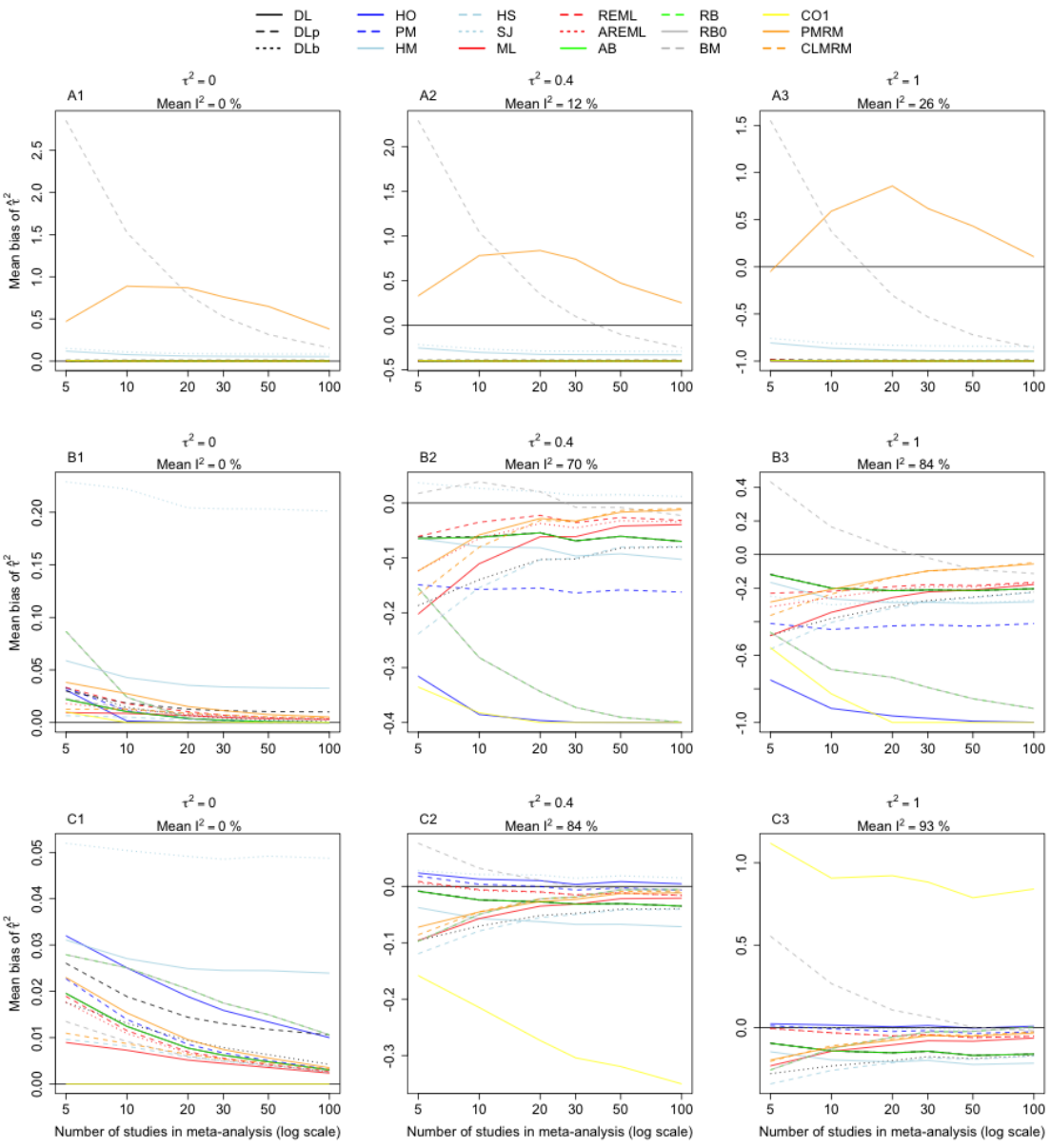


FIGURE E.16: Mean bias of heterogeneity variance estimates in rare events scenario with $p_0 < p_1$ and normal sample size sampling; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). RB, RB0 and CLMRM are omitted from A1-A3; MM, CO2, CO3 and CO4 are omitted from all.

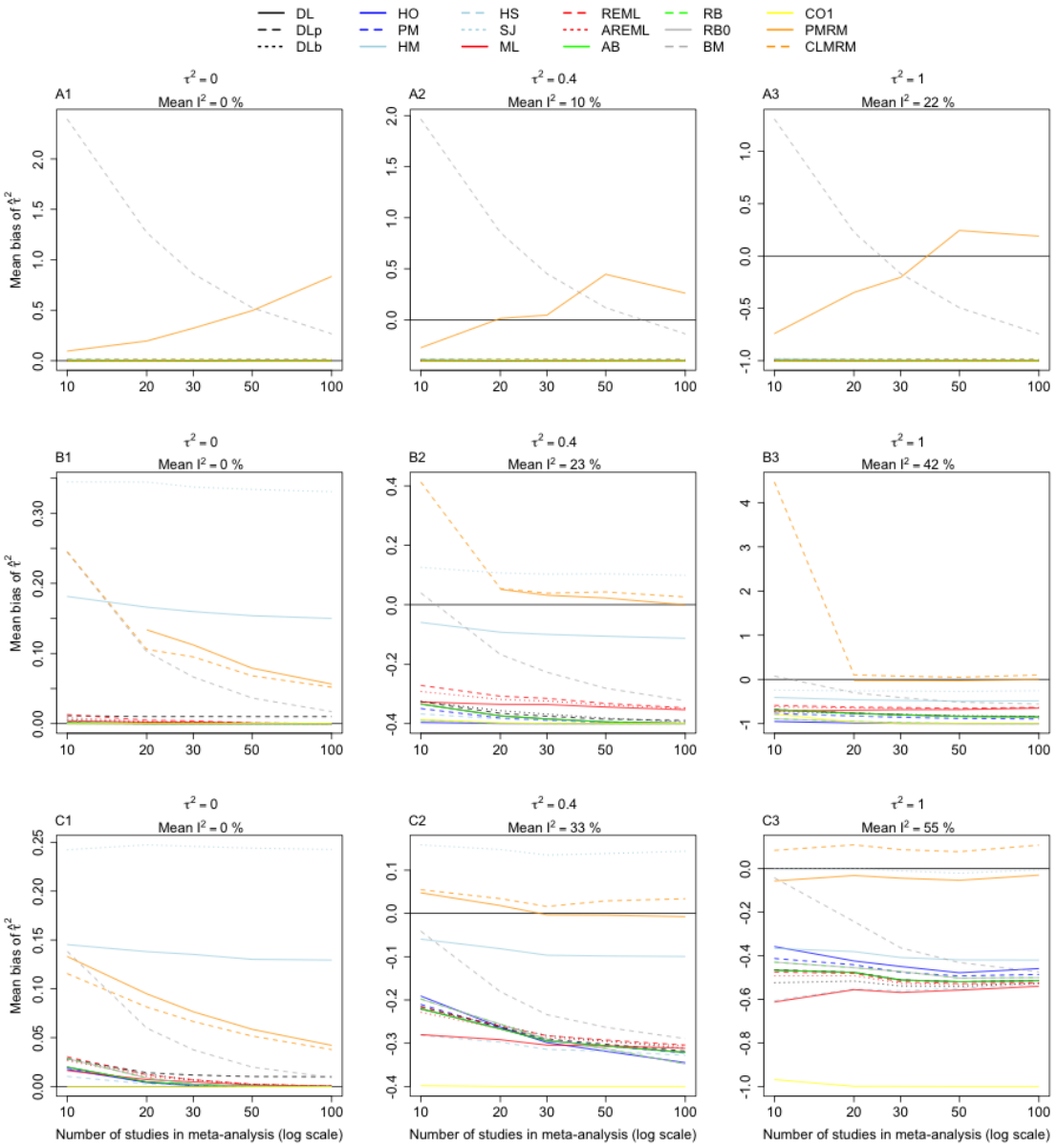


FIGURE E.17: Mean bias of heterogeneity variance estimates in very rare events scenario with $p_0 < p_1$ and Chi-squared sample size sampling; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). RB, RB0 and CLMRM are omitted from A1-A3; MM, CO2, CO3 and CO4 are omitted from all.

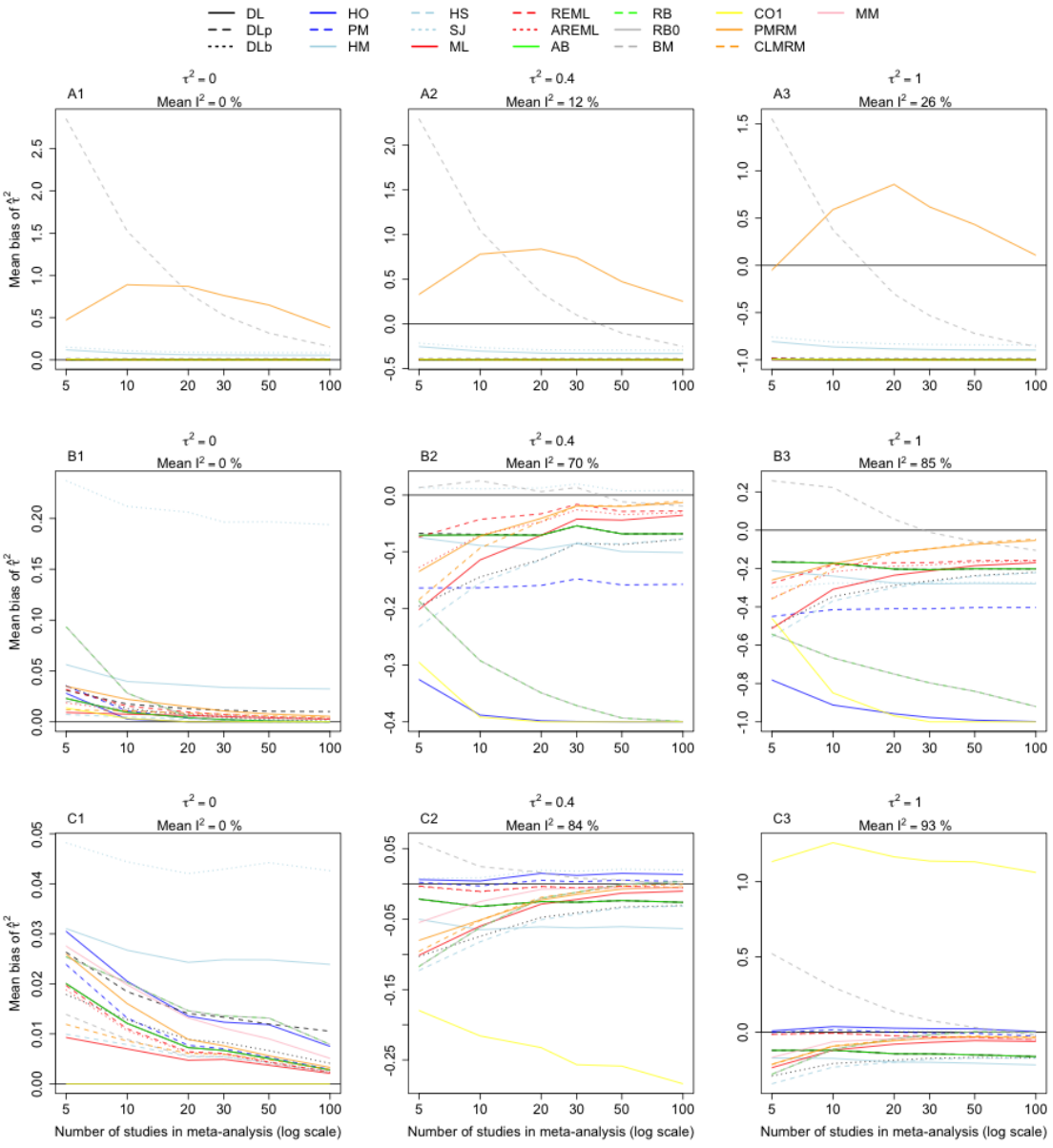


FIGURE E.18: Mean bias of heterogeneity variance estimates in rare events scenario with $p_0 < p_1$ and Chi-squared sample size sampling; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). RB, RB0 and CLMRM are omitted from A1-A3; MM is omitted from A1-B3; CO2, CO3 and CO4 are omitted from all.

E.1.7 Alternate continuity corrections

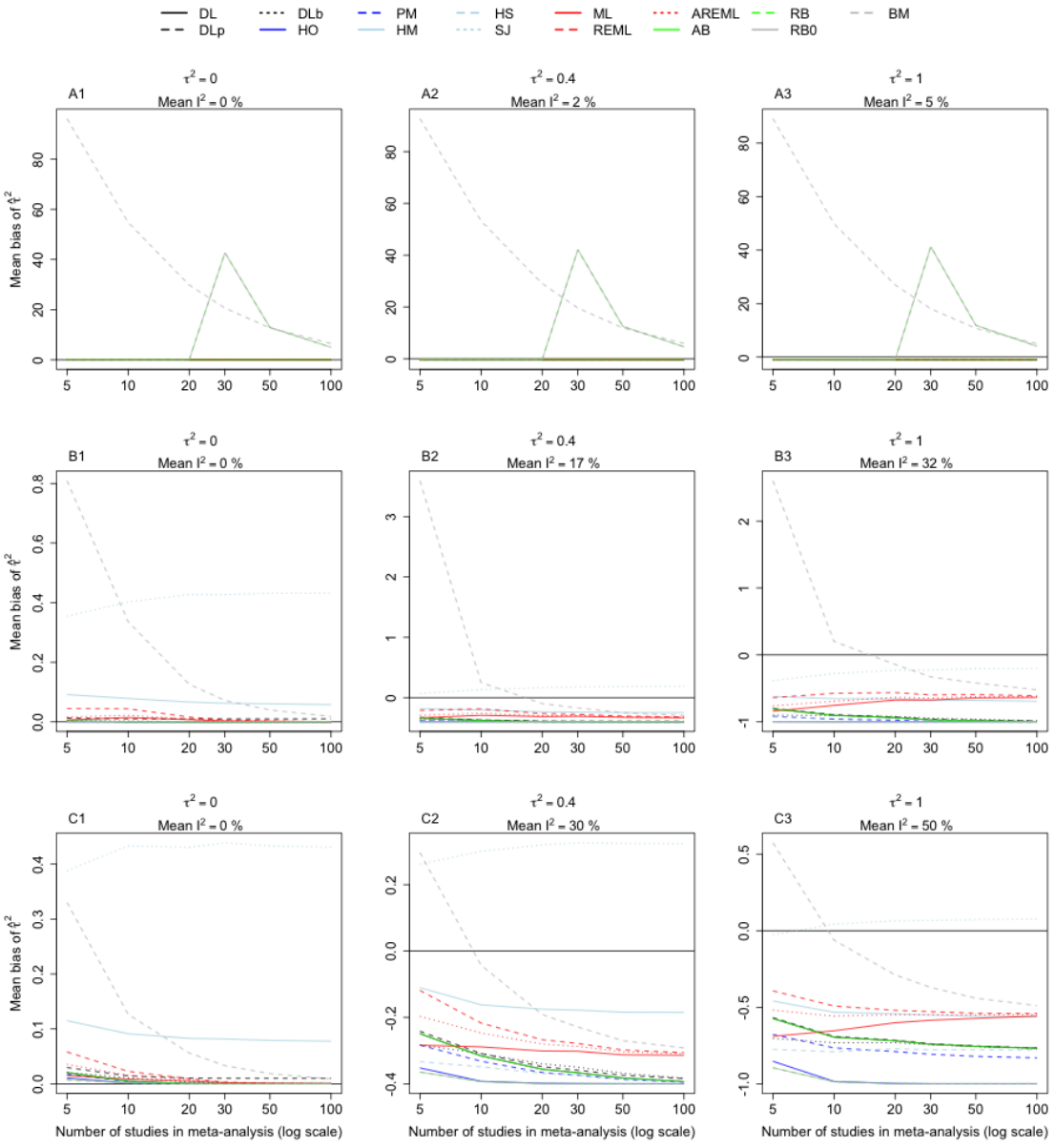


FIGURE E.19: Mean bias of heterogeneity variance estimates in very rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

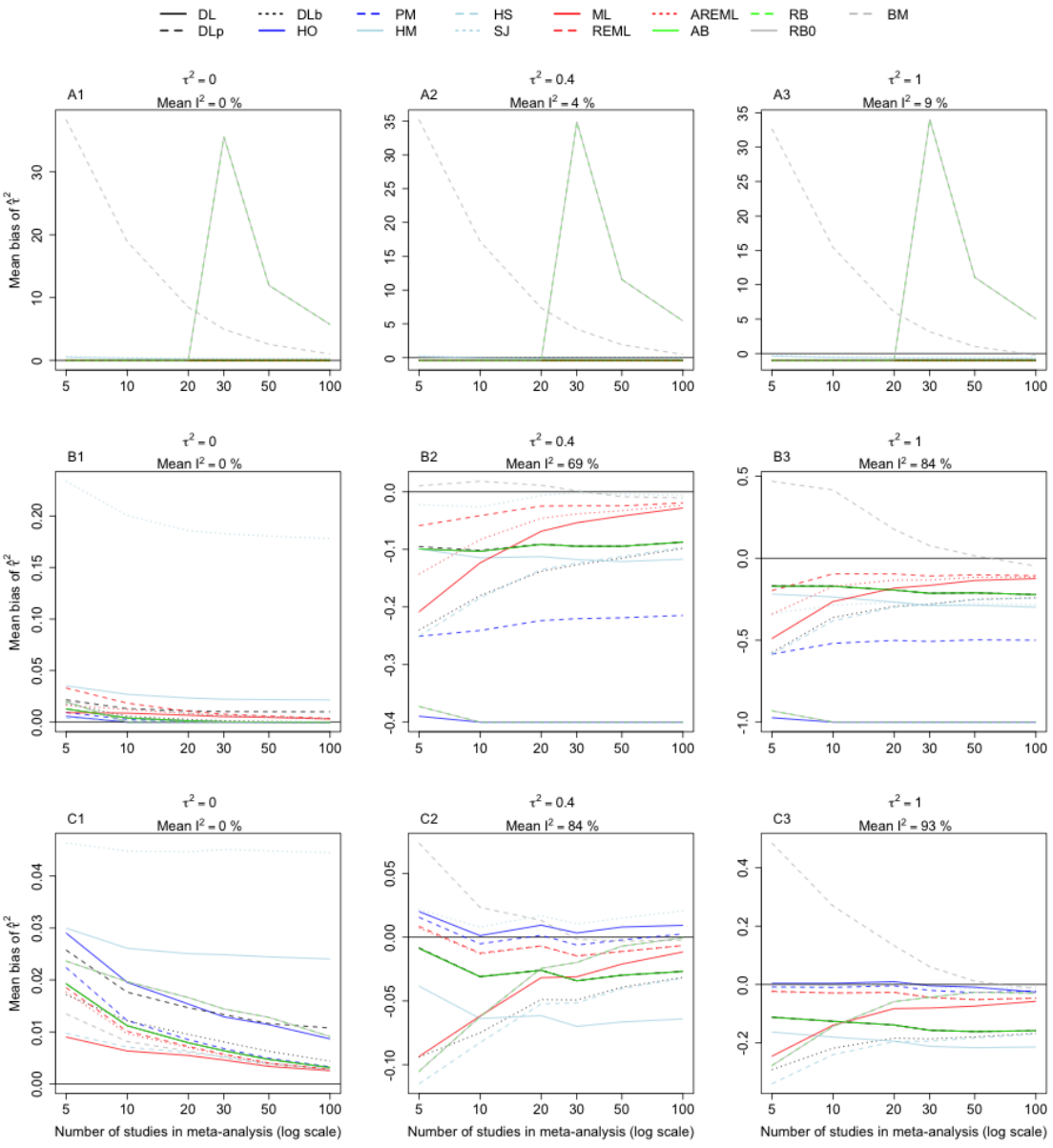


FIGURE E.20: Mean bias of heterogeneity variance estimates in rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

E.2 Mean squared error of τ^2

E.2.1 Examples without omitting outlying estimators

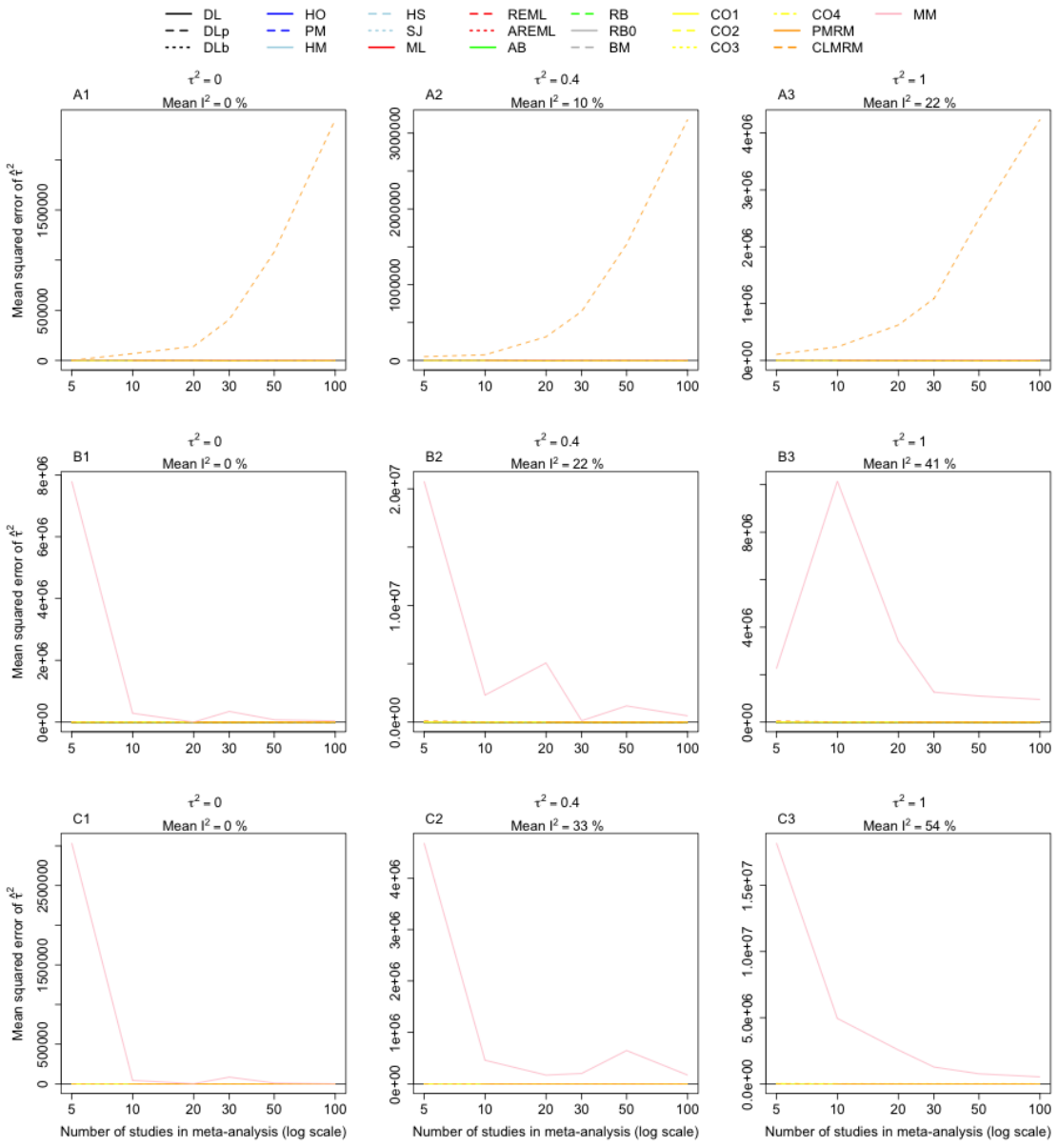


FIGURE E.21: Mean squared error of heterogeneity variance estimates in very rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

E.2.2 Alternate values of heterogeneity variance

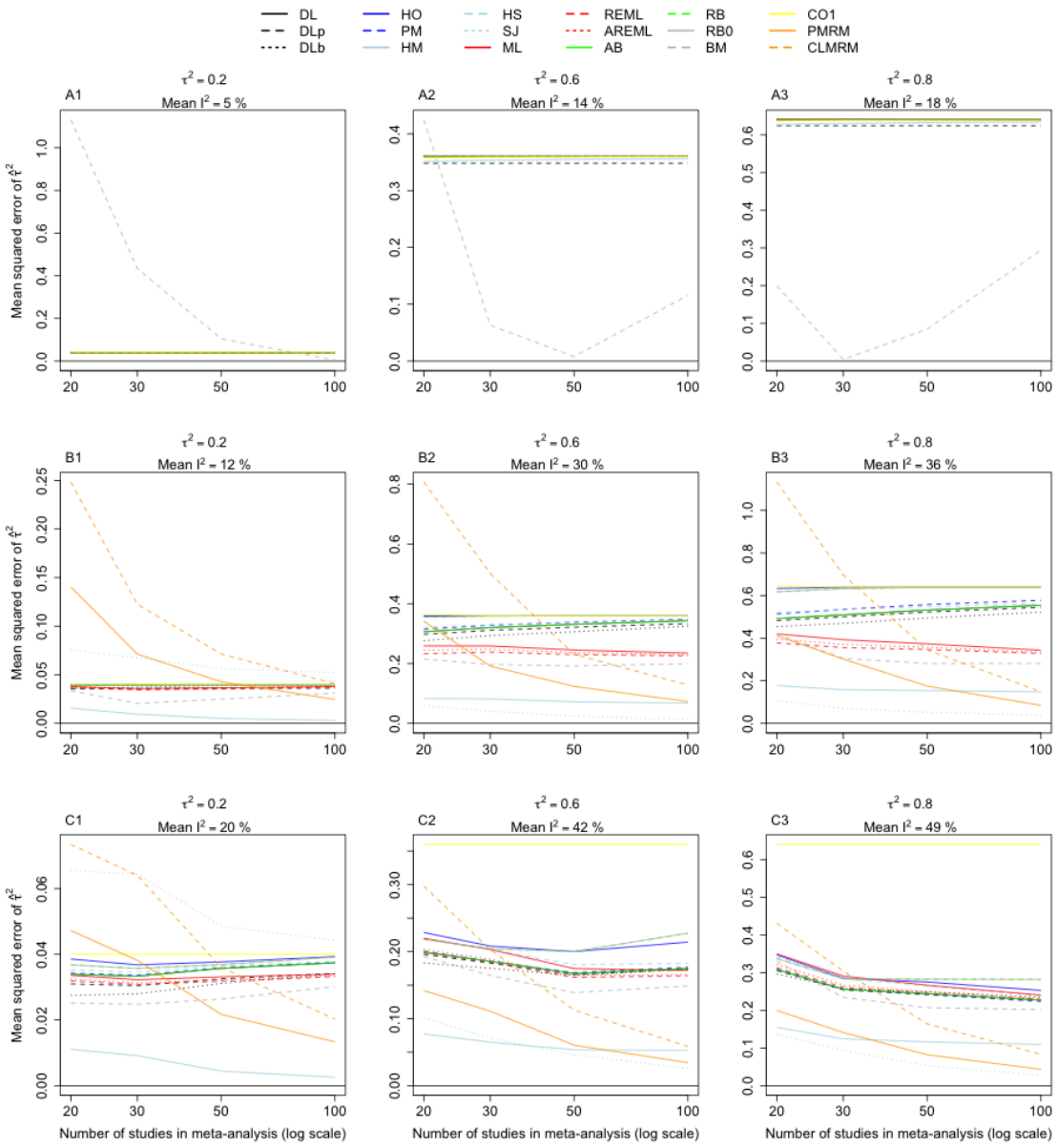


FIGURE E.22: Mean squared error of heterogeneity variance estimates in very rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). RB, RB0, PMRM and CLMRM were omitted from A1-A3; CO2, CO3, CO4 and MM were omitted from all.

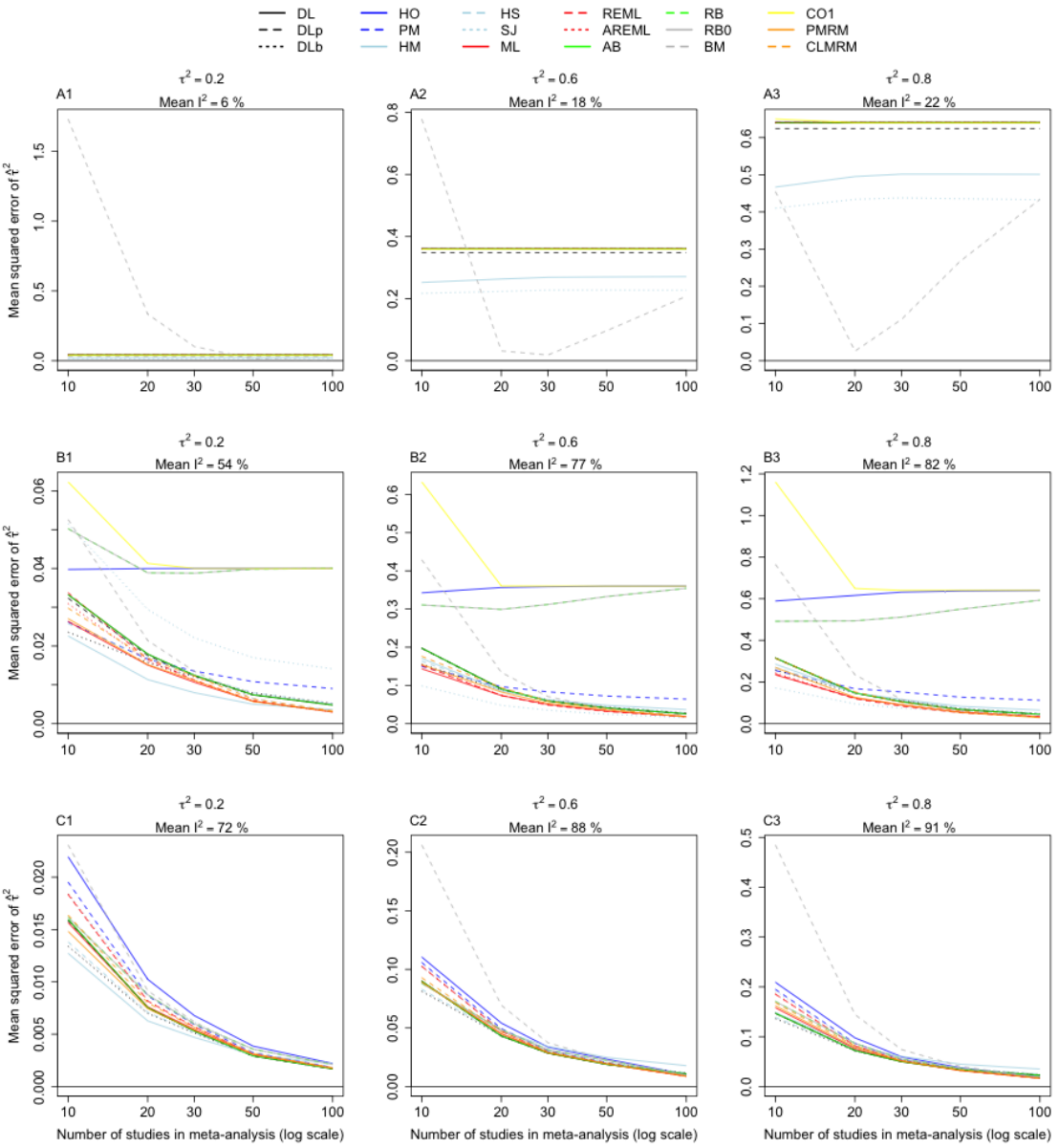


FIGURE E.23: Mean squared error of heterogeneity variance estimates in rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). RB, RB0, PMRM and CLMRM were omitted from A1-A3; CO1 was omitted for C1-C3; CO2, CO3, CO4 and MM were omitted from all.

E.2.3 Alternate study sample sizes

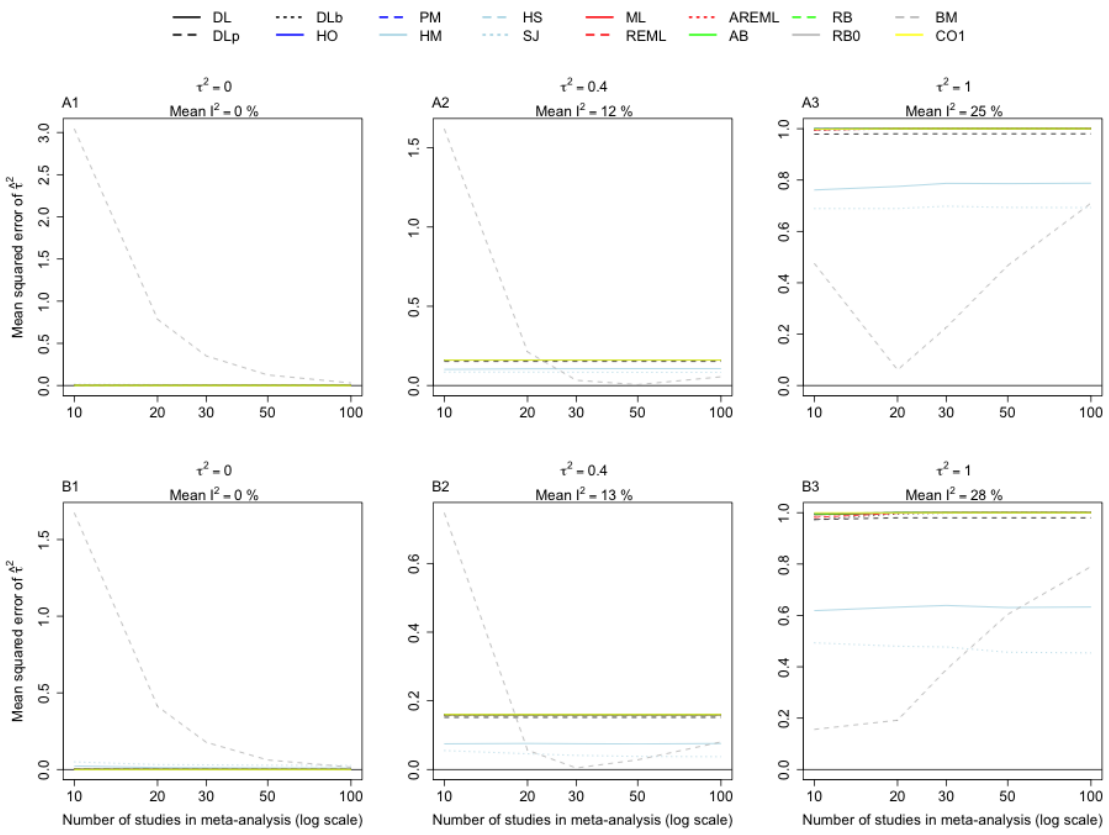


FIGURE E.24: Mean squared error of heterogeneity variance estimates in very rare events scenario with $p_0 < p_1$; sample sizes are small-to-medium (A1-A3) and medium (B1-B3). CO2, CO3, CO4, PMRM, CLMRM and MM were omitted from all.

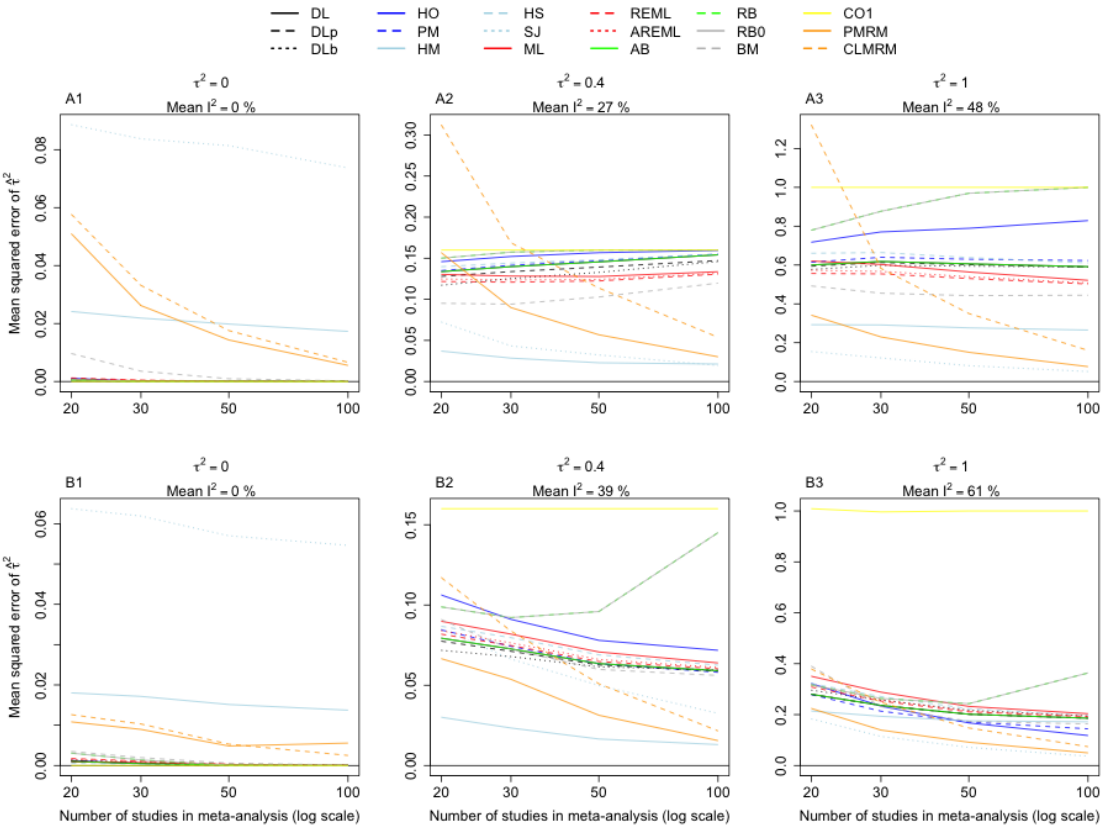


FIGURE E.25: Mean squared error of heterogeneity variance estimates in rare events scenario with $p_0 < p_1$; sample sizes are small-to-medium (A1-A3) and medium (B1-B3). CO2, CO3, CO4 and MM are omitted from all.

E.2.4 Alternate values of σ_α^2

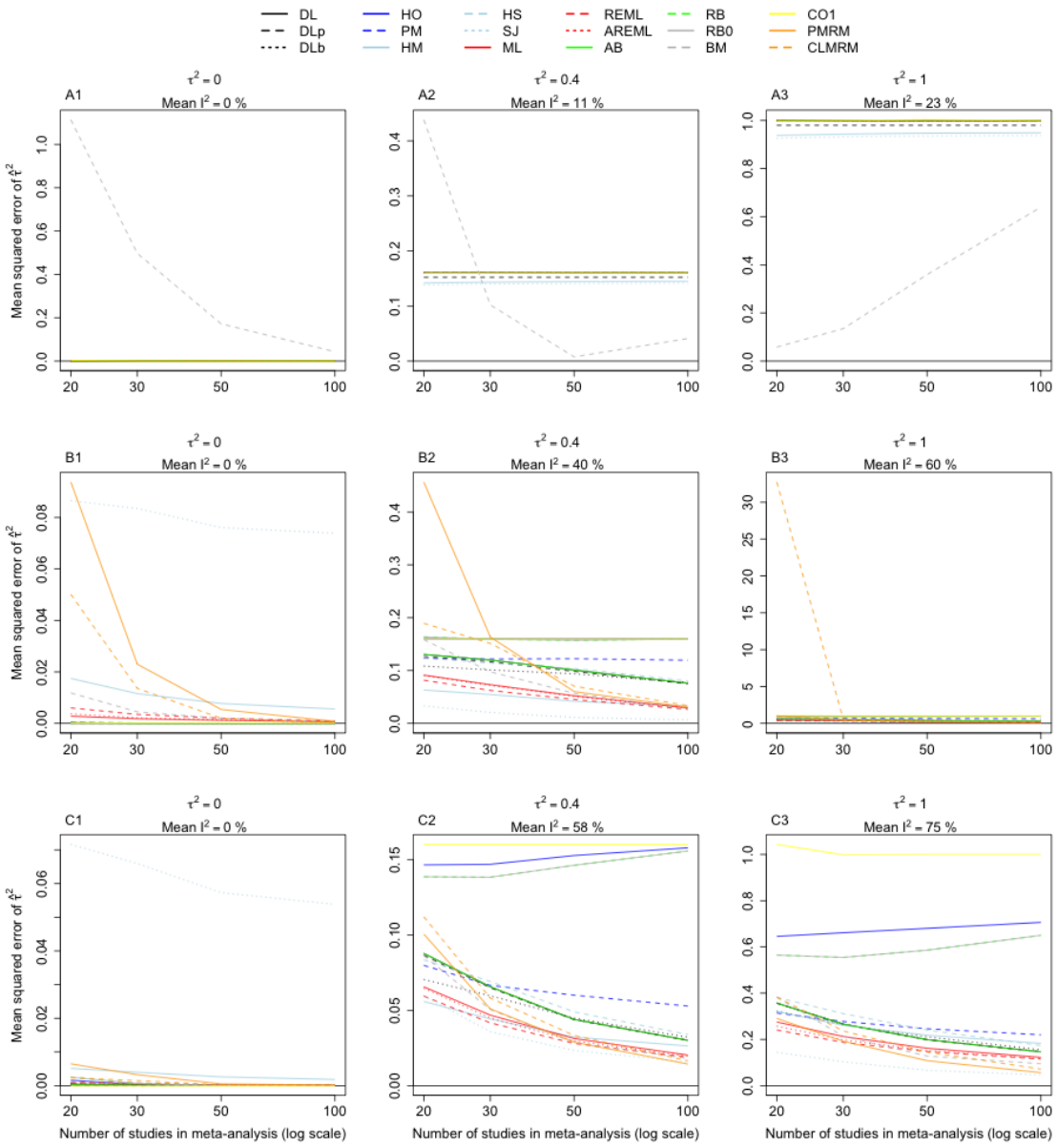


FIGURE E.26: Mean squared error of heterogeneity variance estimates in very rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). RB, RB0, PMRM and CLMRM are omitted from A1-A3; CO2, CO3, CO4 and MM are omitted from all.

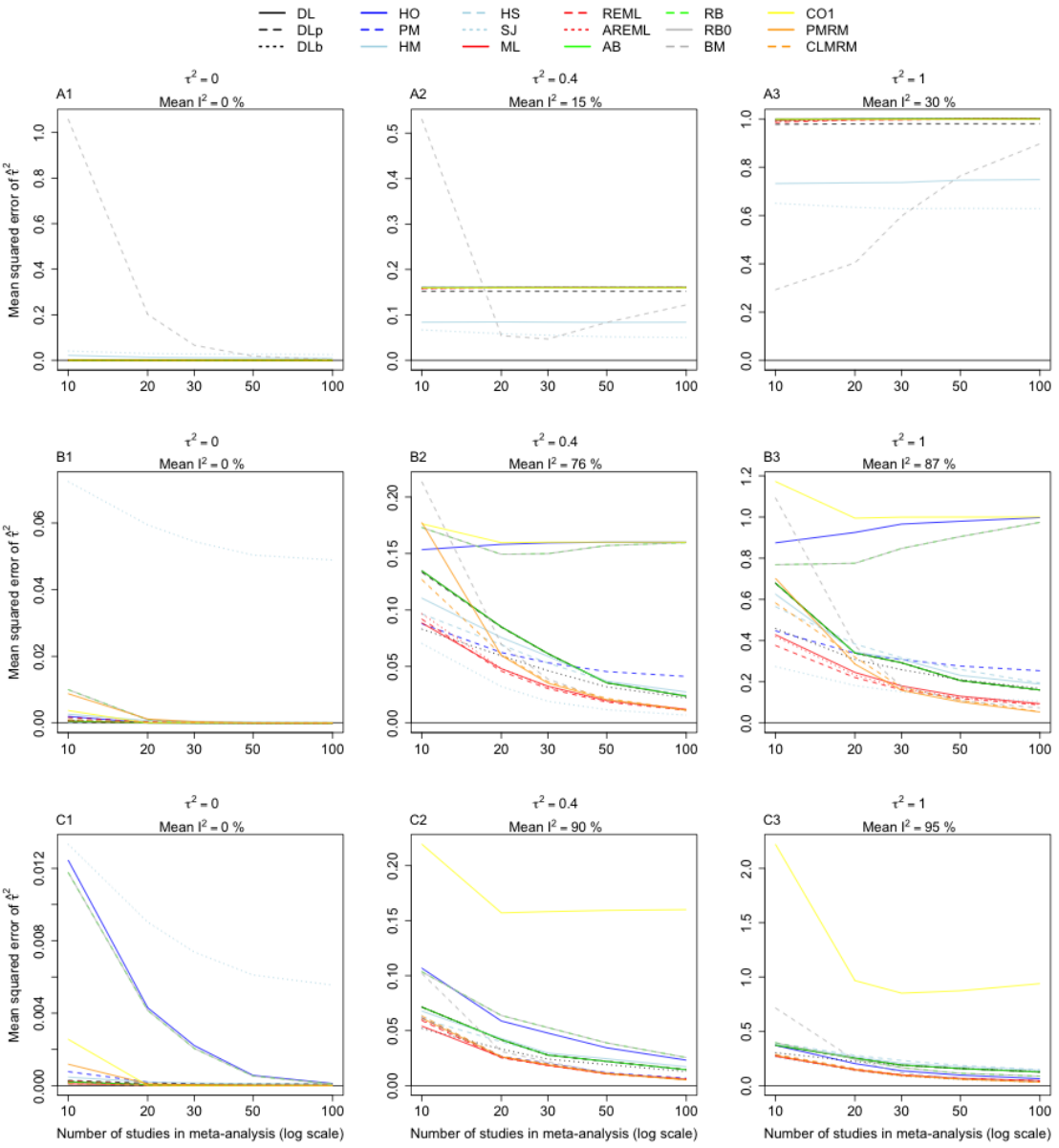


FIGURE E.27: Mean squared error of heterogeneity variance estimates in rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). RB, RB0, PMRM and CLMRM are omitted from A1-A3; CO2, CO3, CO4 and MM are omitted from all.

E.2.5 Alternate probability scenarios

Alternate rare events scenarios

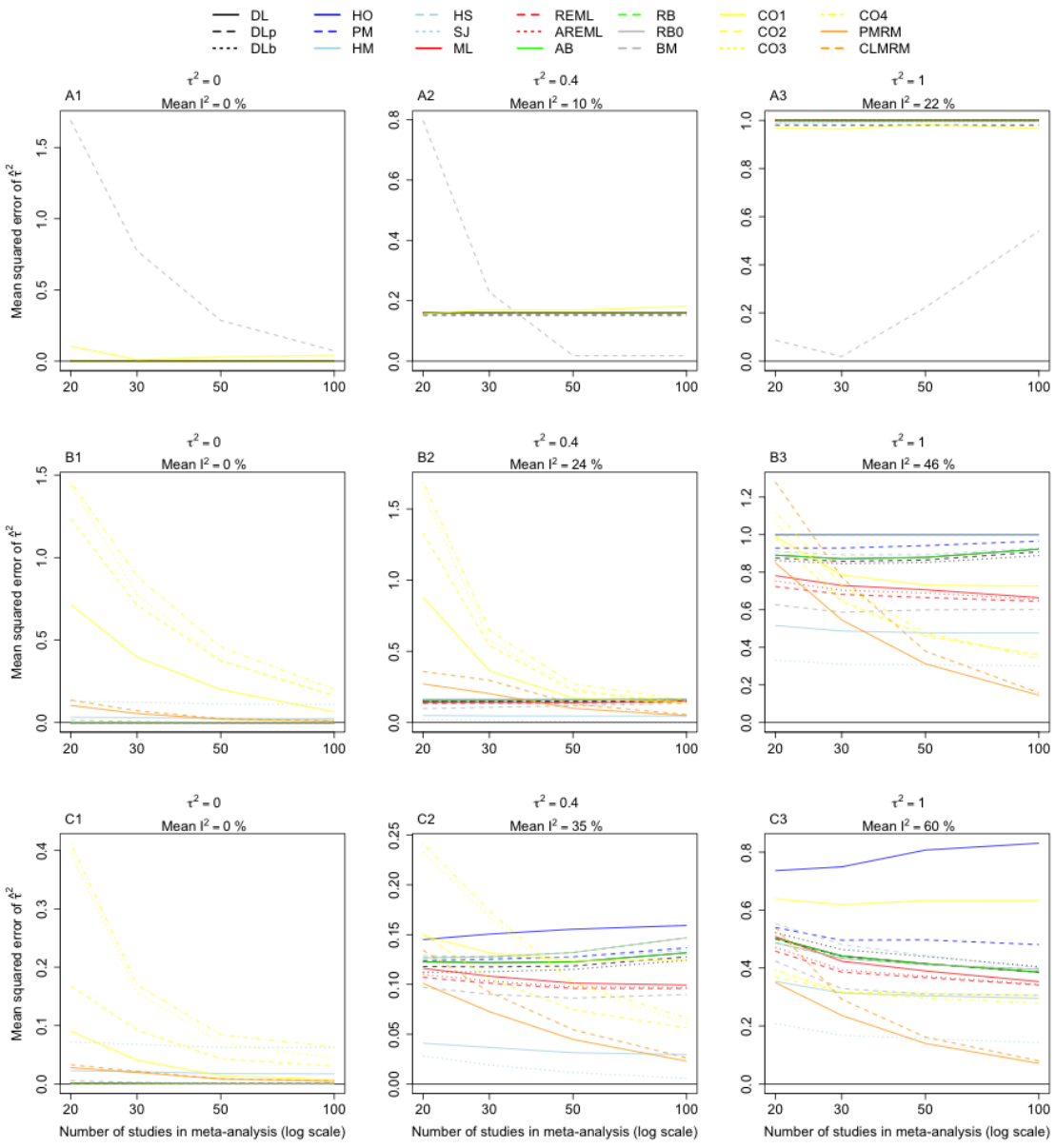


FIGURE E.28: Mean squared error of heterogeneity variance estimates in very rare events scenario with $p_0 > p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). RB, RB0, PMRM, CLMRM, CO2, CO3 and CO4 are omitted from A1-A3; MM is omitted from all.

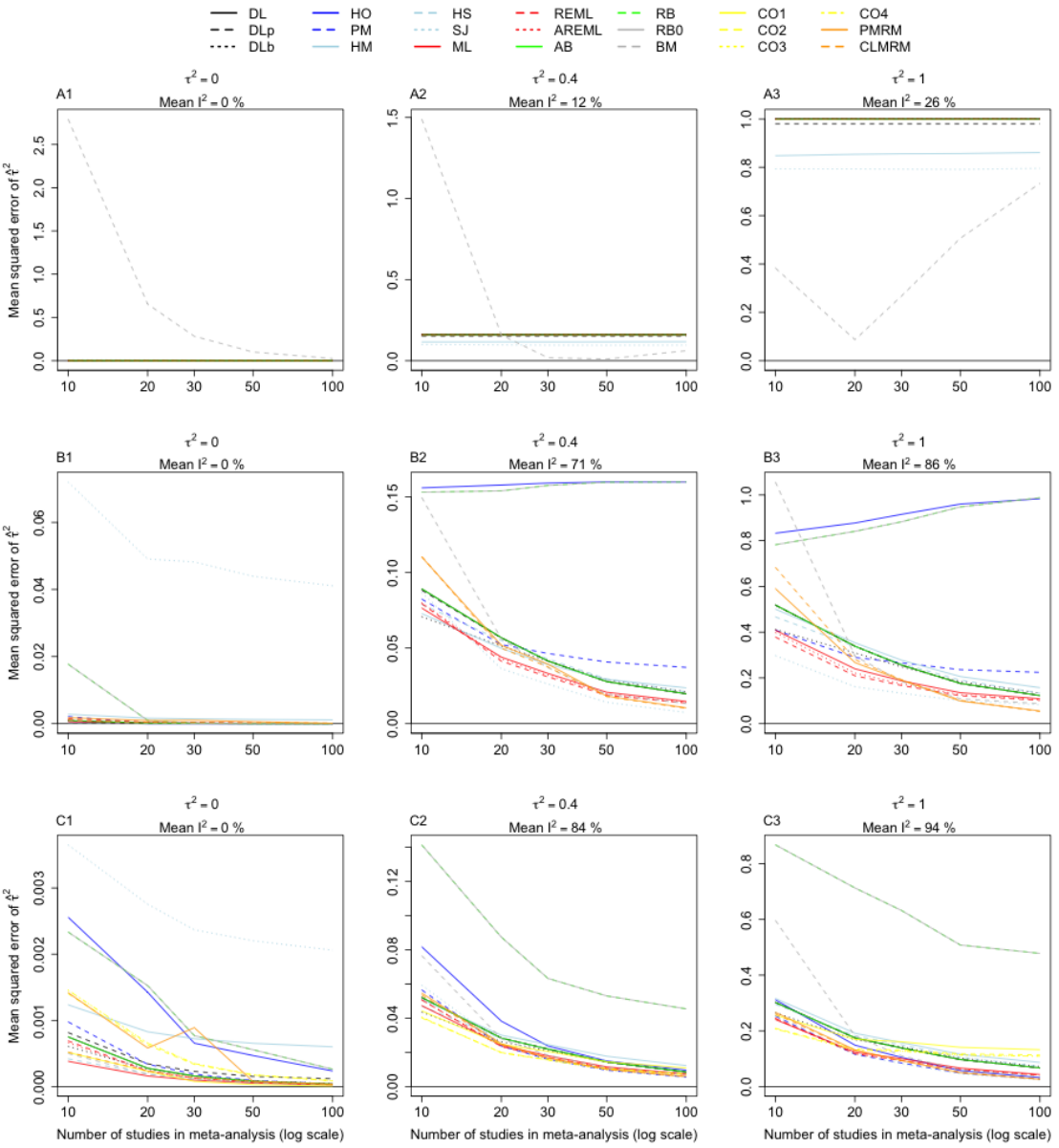


FIGURE E.29: Mean squared error of heterogeneity variance estimates in rare events scenario with $p_0 > p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). RB, RB0, PMRM and CLMRM are omitted from A1-A3; CO2, CO3 and CO4 are omitted from A1-B3; MM is omitted from all.

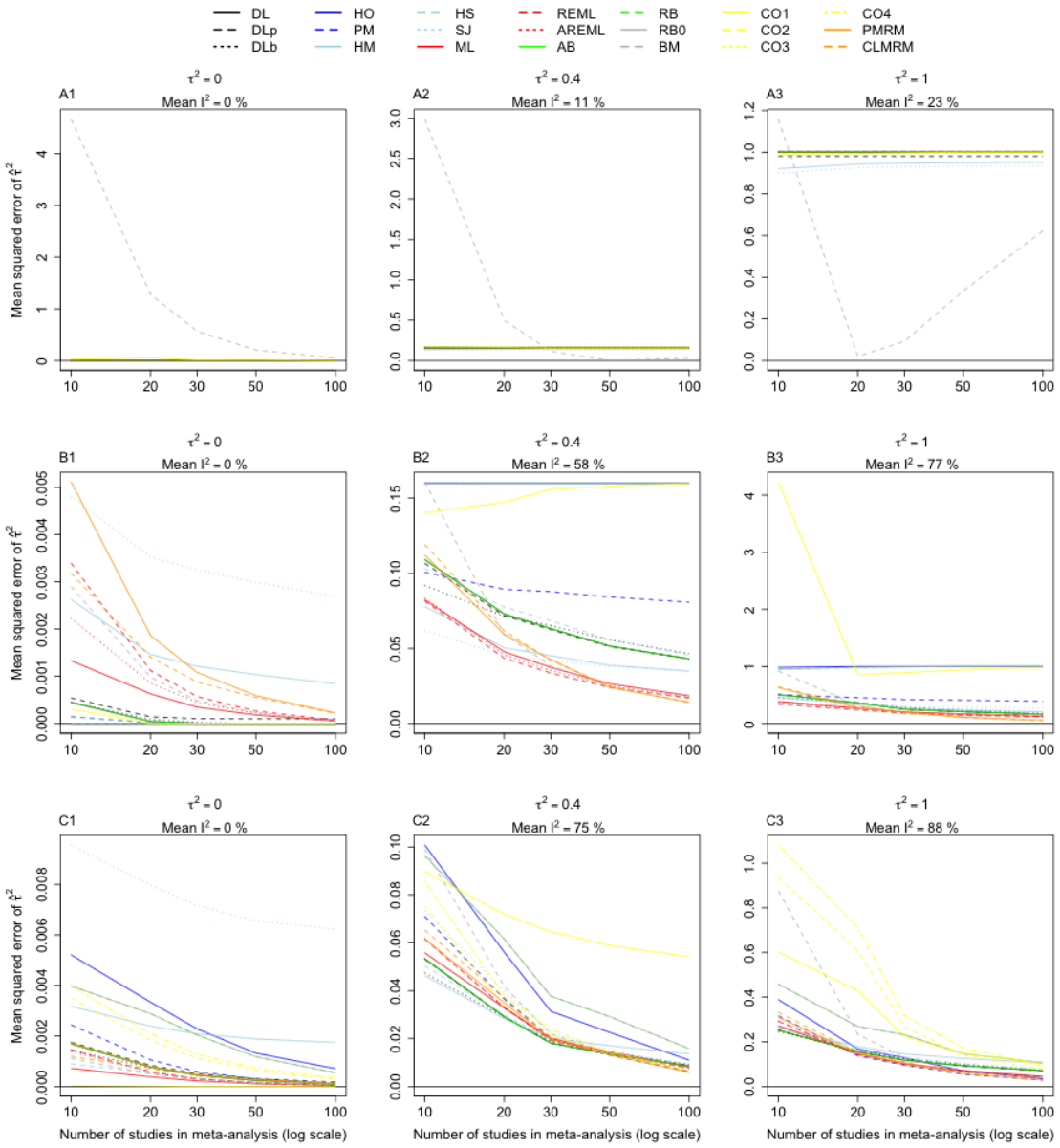


FIGURE E.30: Mean squared error of heterogeneity variance estimates in rare events scenario with $p_0 = p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). RB, RB0, PMRM and CLMRM are omitted from A1-A3; CO2, CO3 and CO4 are omitted from A1-B3; MM is omitted from all.

Common probability scenarios

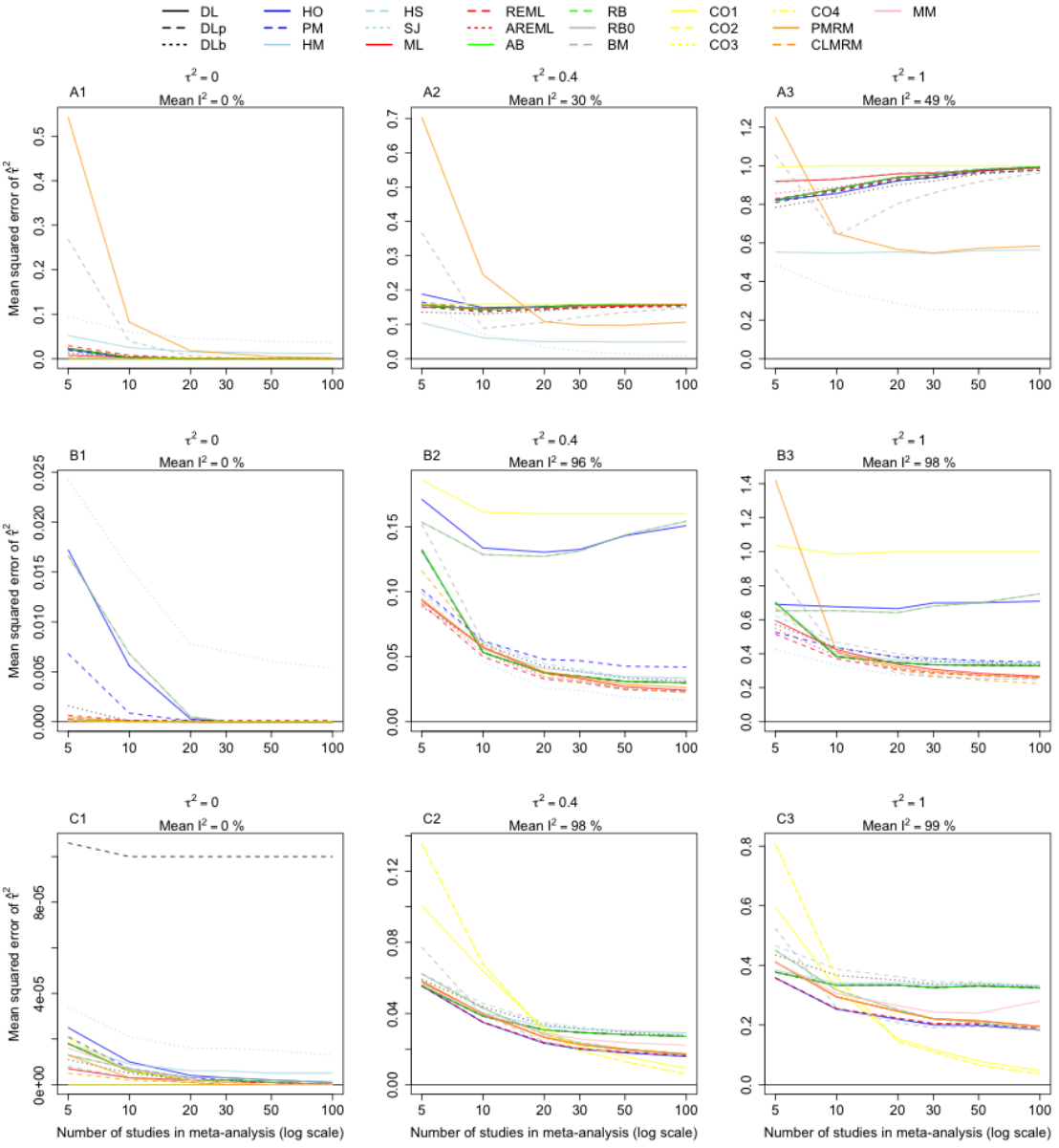


FIGURE E.31: Mean squared error of heterogeneity variance estimates in common probability scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). RB, RB0 and CLMRM are omitted from A1-A3; CO2, CO3, CO4 and MM are omitted from A1-B3.

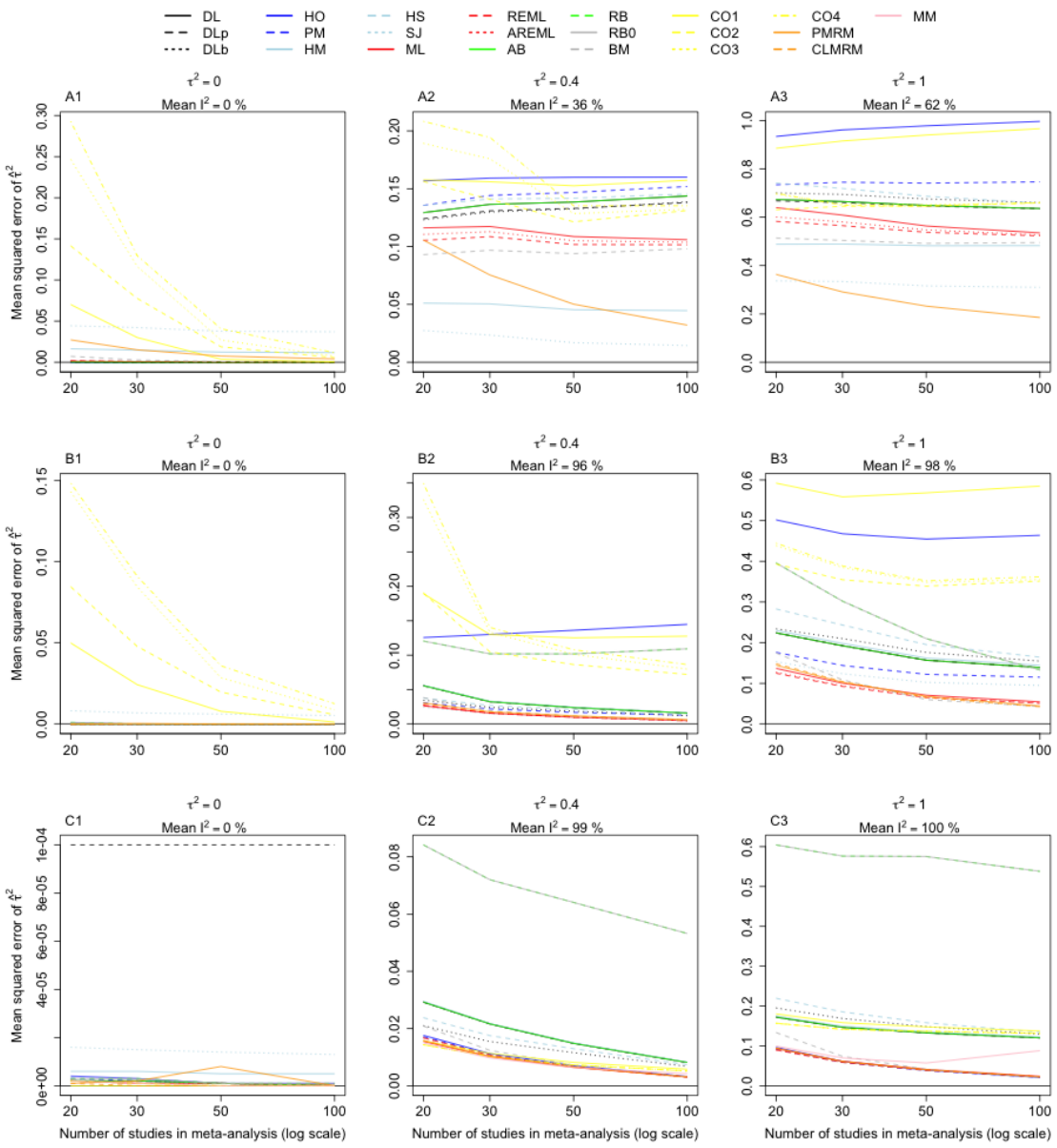


FIGURE E.32: Mean squared error of heterogeneity variance estimates in common probability scenario with $p_0 > p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). RB, RB0 and CLMRM are omitted from A1-A3; MM is omitted from A1-B3.

E.2.6 Alternate sampling in simulation study

Alternate event count sampling

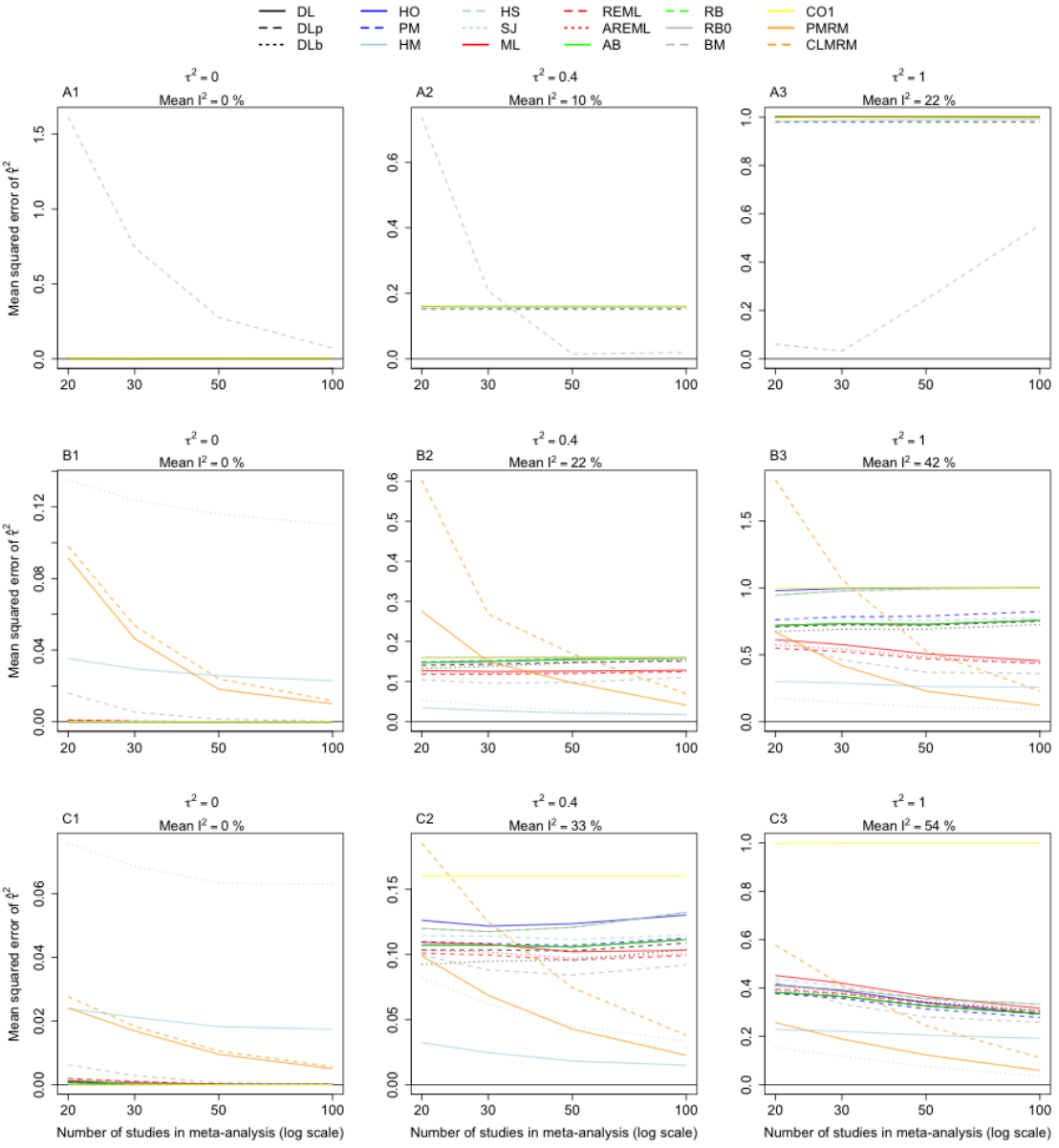


FIGURE E.33: Mean squared error of heterogeneity variance estimates in very rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). RB, RB0, PMRM and CLMRM are omitted from A1-A3; MM, CO2, CO3 and CO4 are omitted from all.

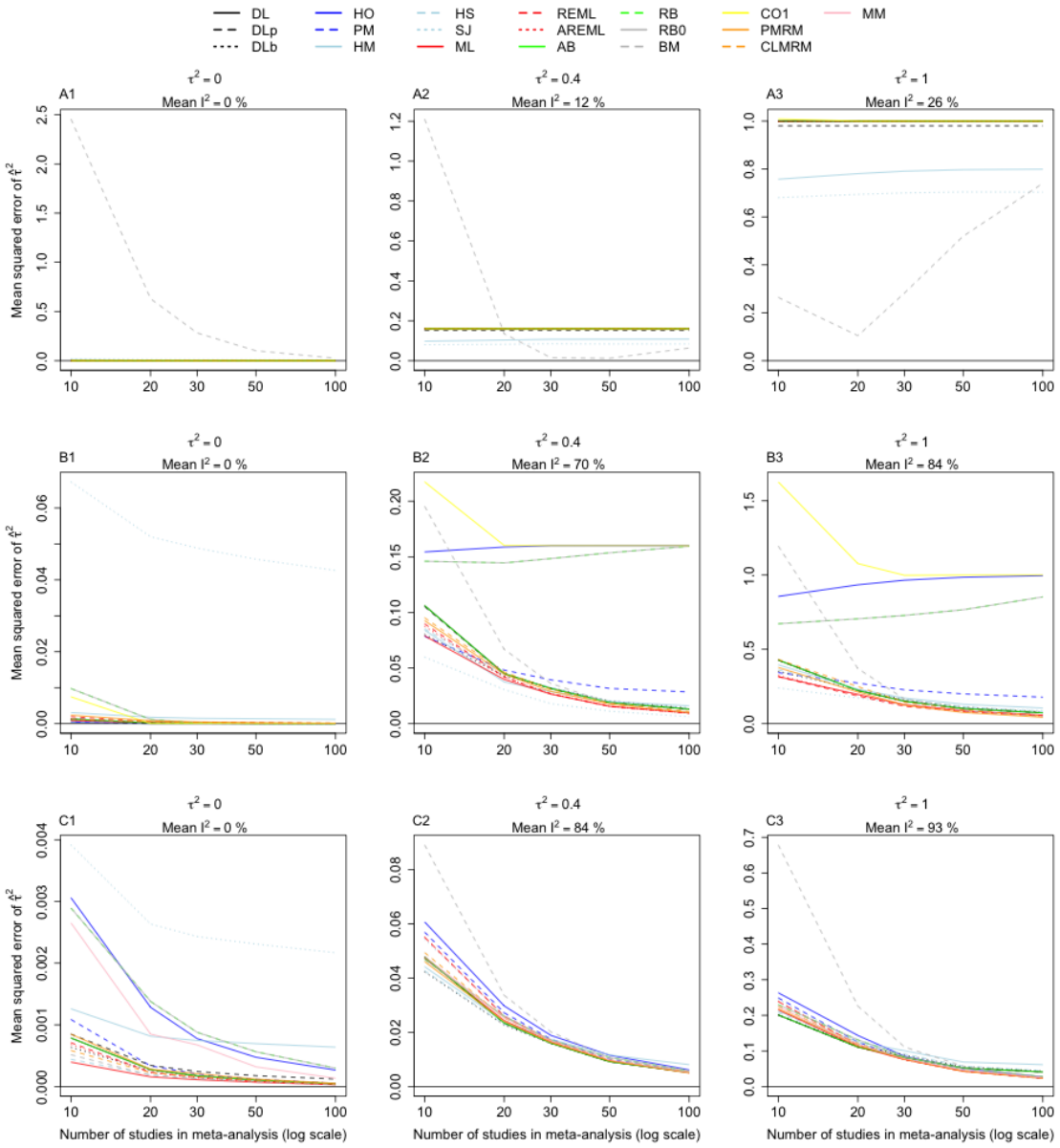


FIGURE E.34: Mean squared error of heterogeneity variance estimates in rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). RB, RB0, PMRM and CLMRM are omitted from A1-A3; MM is omitted from A1-B3; CO1 is omitted form C1-C3; CO2, CO3 and CO4 are omitted from all.

Alternate sample size sampling

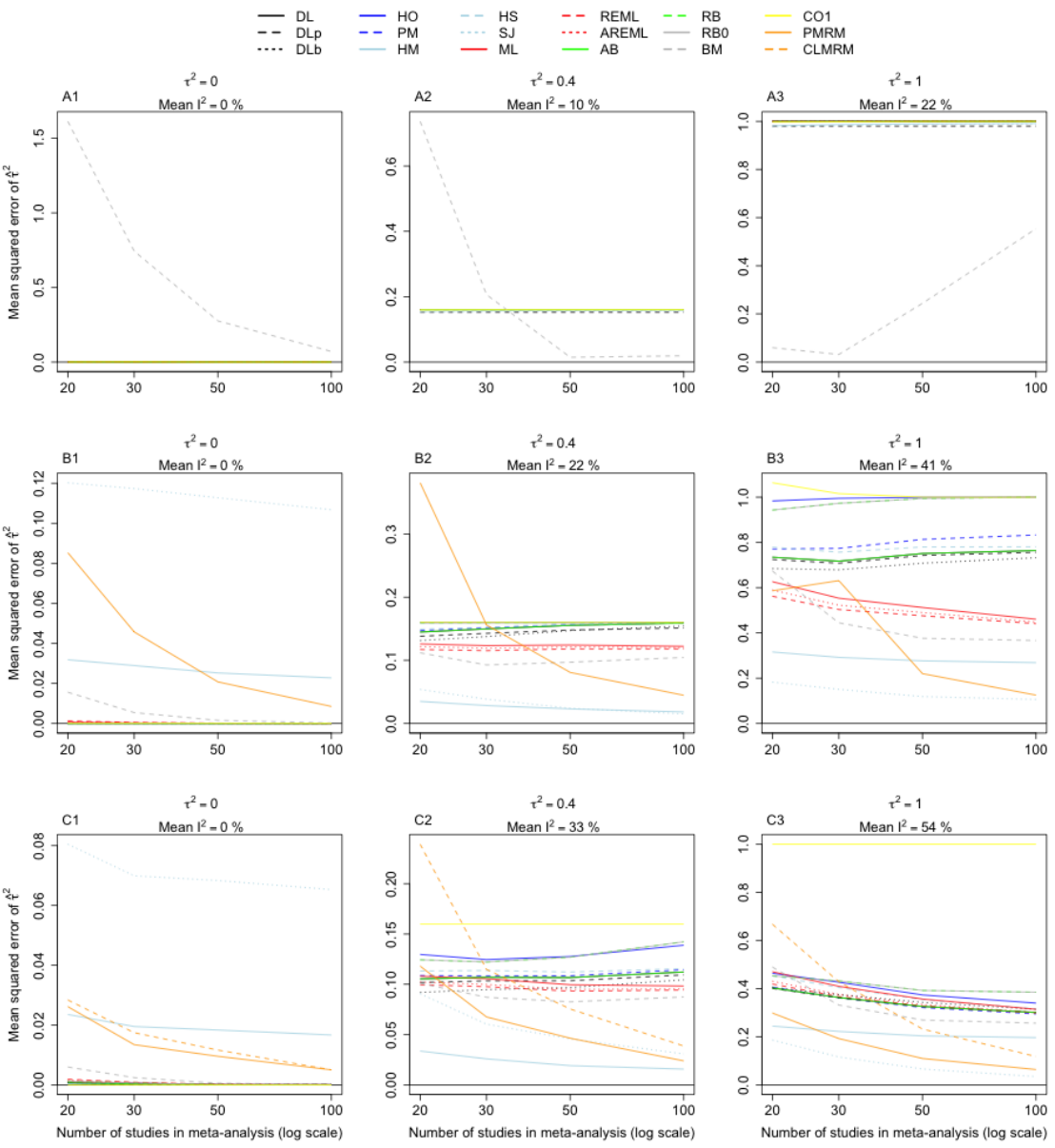


FIGURE E.35: Mean squared error of heterogeneity variance estimates in very rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). RB, RB0 and PMRM are omitted from A1-A3; CLMRM is omitted from A1-B3; MM, CO2, CO3 and CO4 are omitted from all.

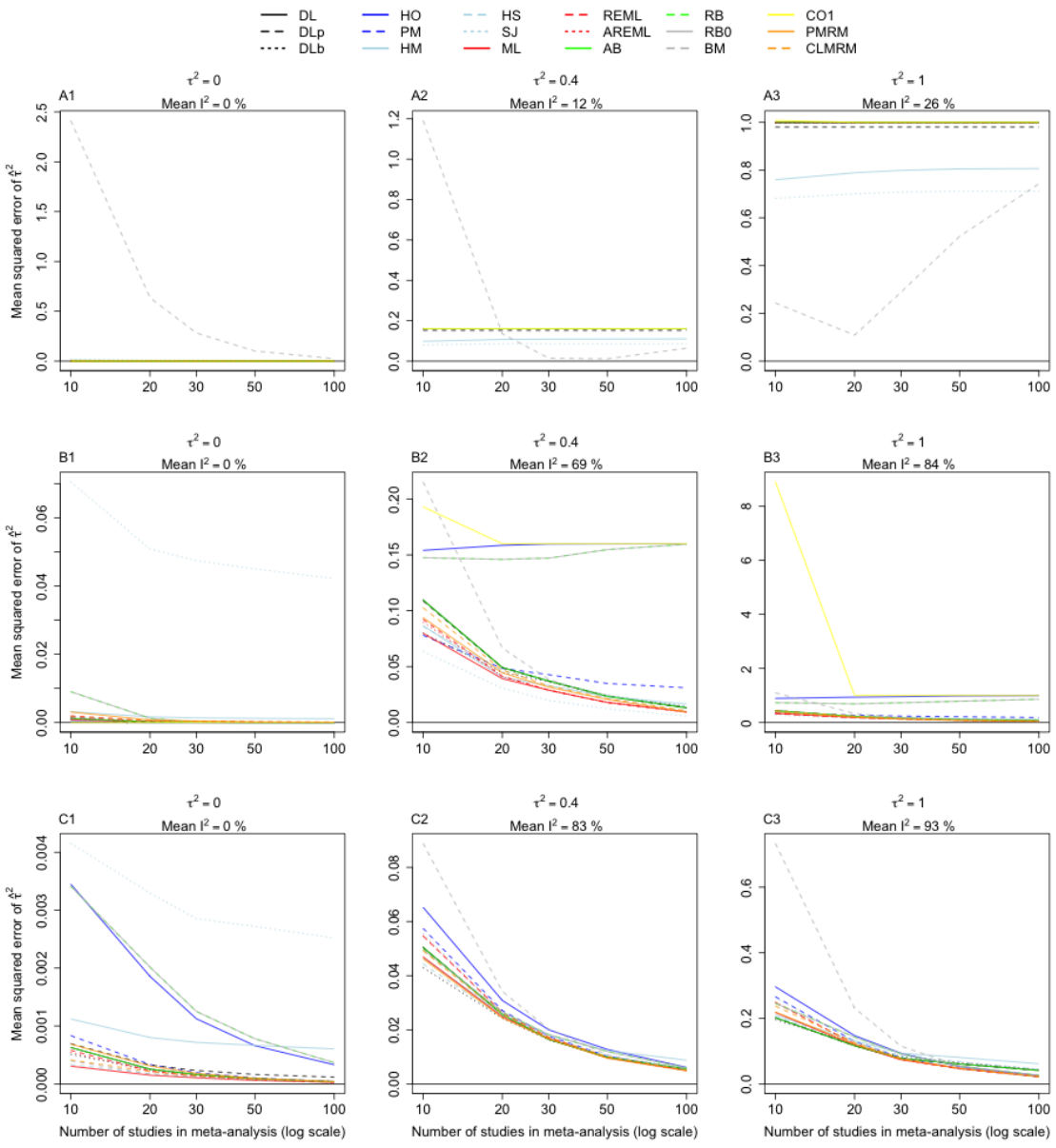


FIGURE E.36: Mean squared error of heterogeneity variance estimates in rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). RB, RB0, PMRM and CLMRM are omitted from A1-A3; CO1 is omitted from C1-C3; MM, CO2, CO3 and CO4 are omitted from all.

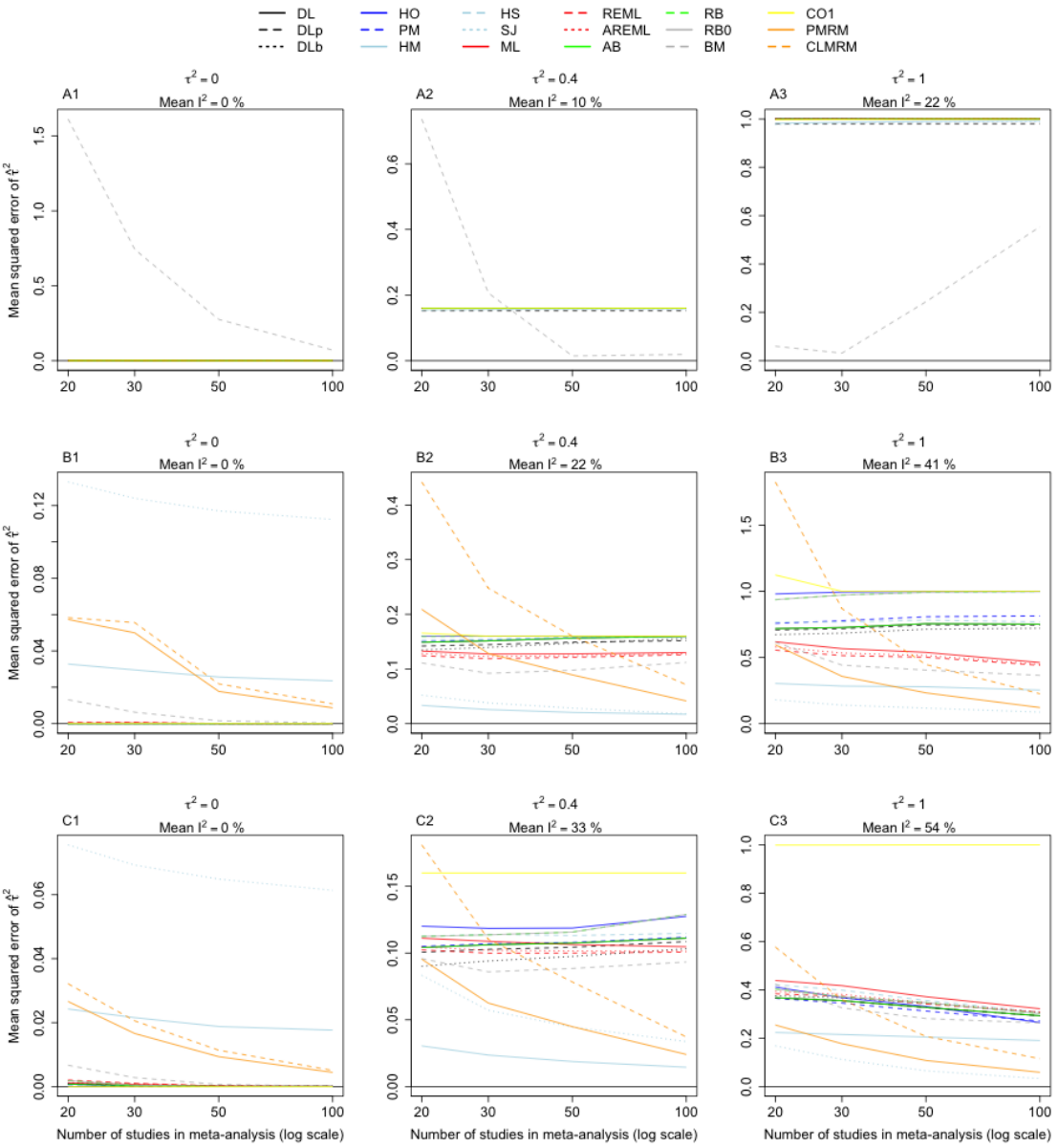


FIGURE E.37: Mean squared error of heterogeneity variance estimates in very rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). RB, RB0, PMRM and CLMRM are omitted from A1-A3; MM, CO2, CO3 and CO4 are omitted from all.

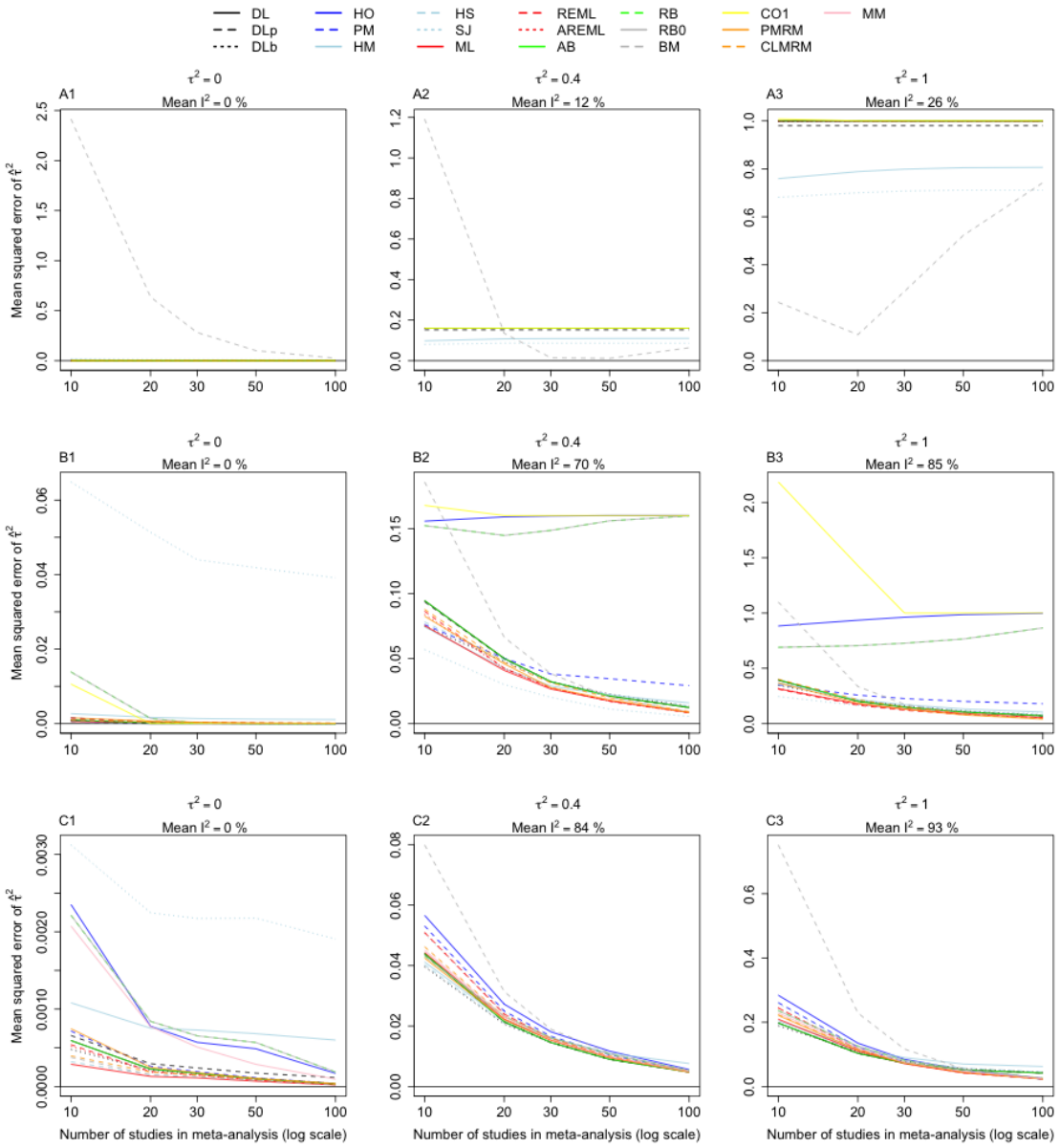


FIGURE E.38: Mean squared error of heterogeneity variance estimates in rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). RB, RB0, PMRM and CLMRM are omitted from A1-A3; MM is omitted from A1-B3; CO1 is omitted from C1-C3; CO2, CO3 and CO4 are omitted from all.

E.2.7 Alternate continuity corrections

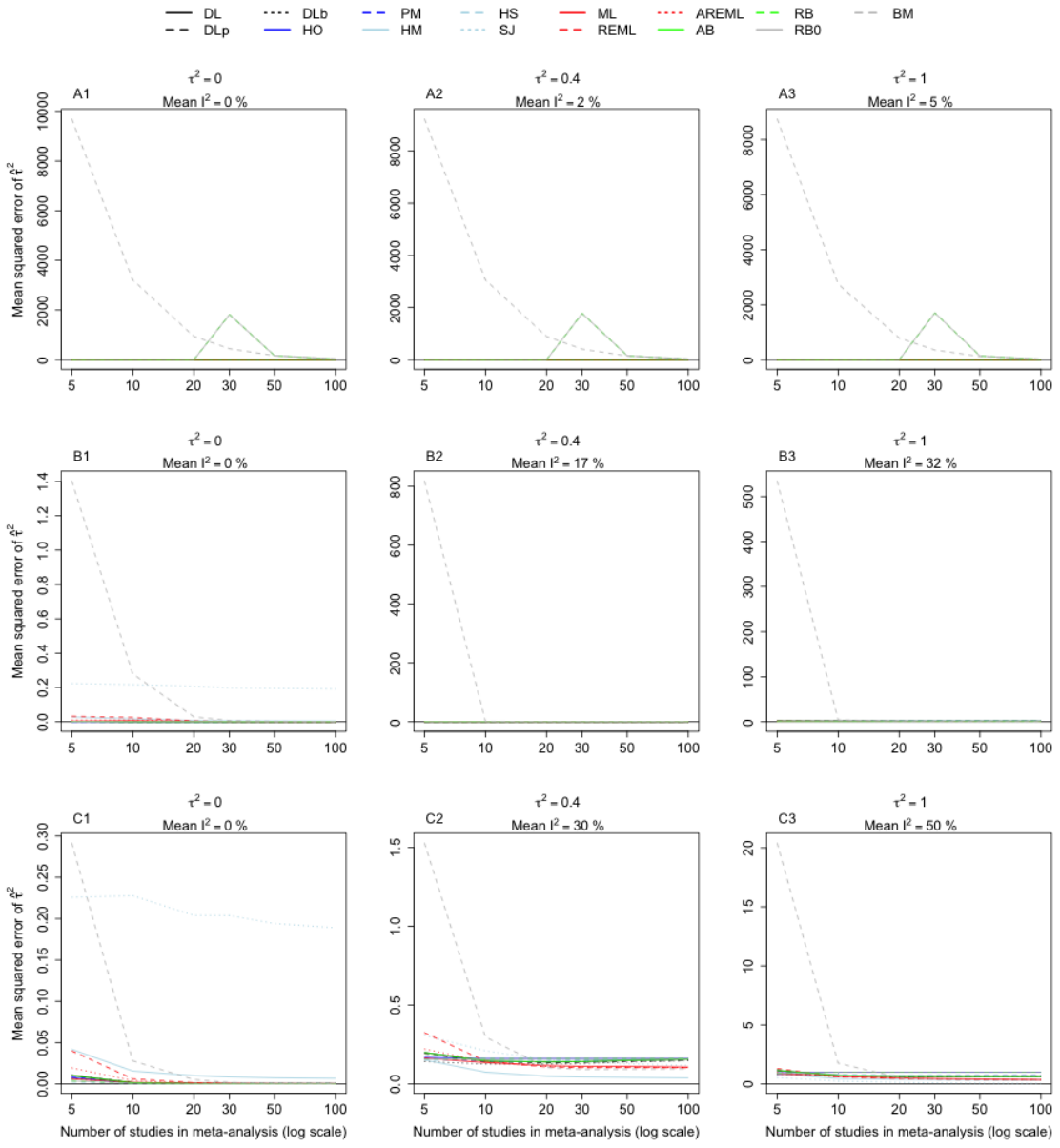


FIGURE E.39: Mean squared error of heterogeneity variance estimates in very rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

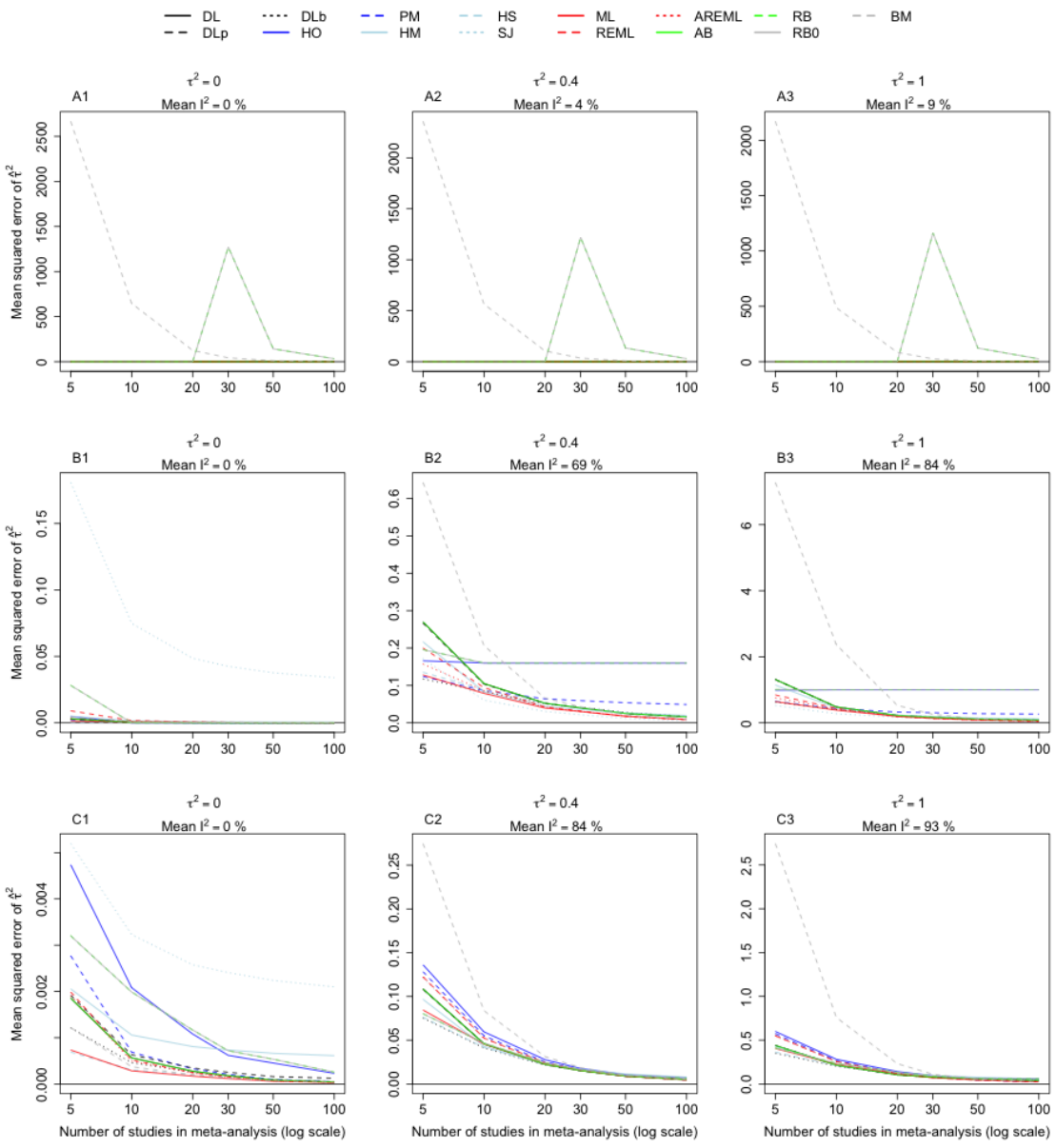


FIGURE E.40: Mean squared error of heterogeneity variance estimates in rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

E.3 Proportion of zero τ^2 estimates

E.3.1 Alternate values of heterogeneity variance

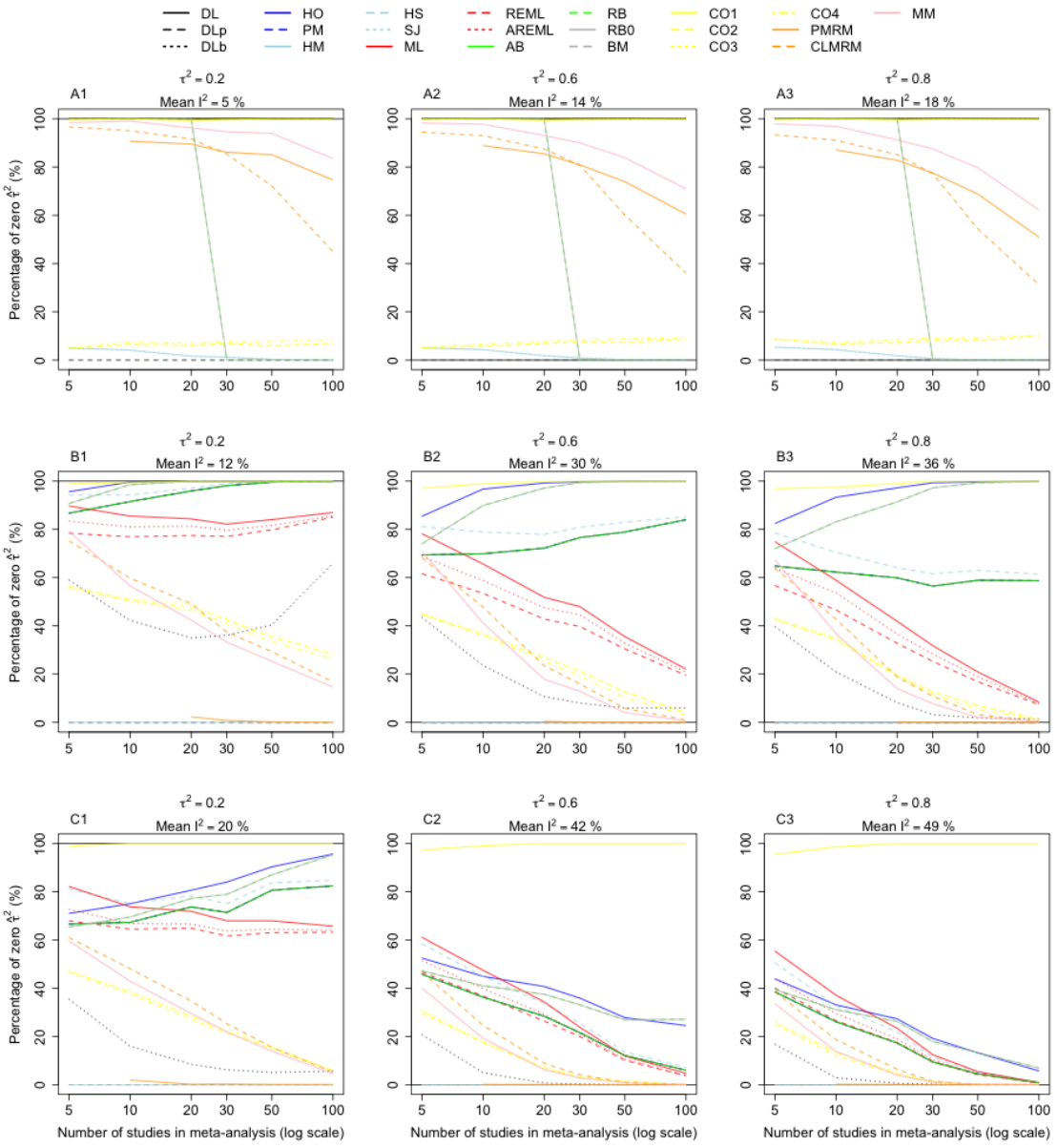


FIGURE E.41: Proportion of zero heterogeneity variance estimates in very rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

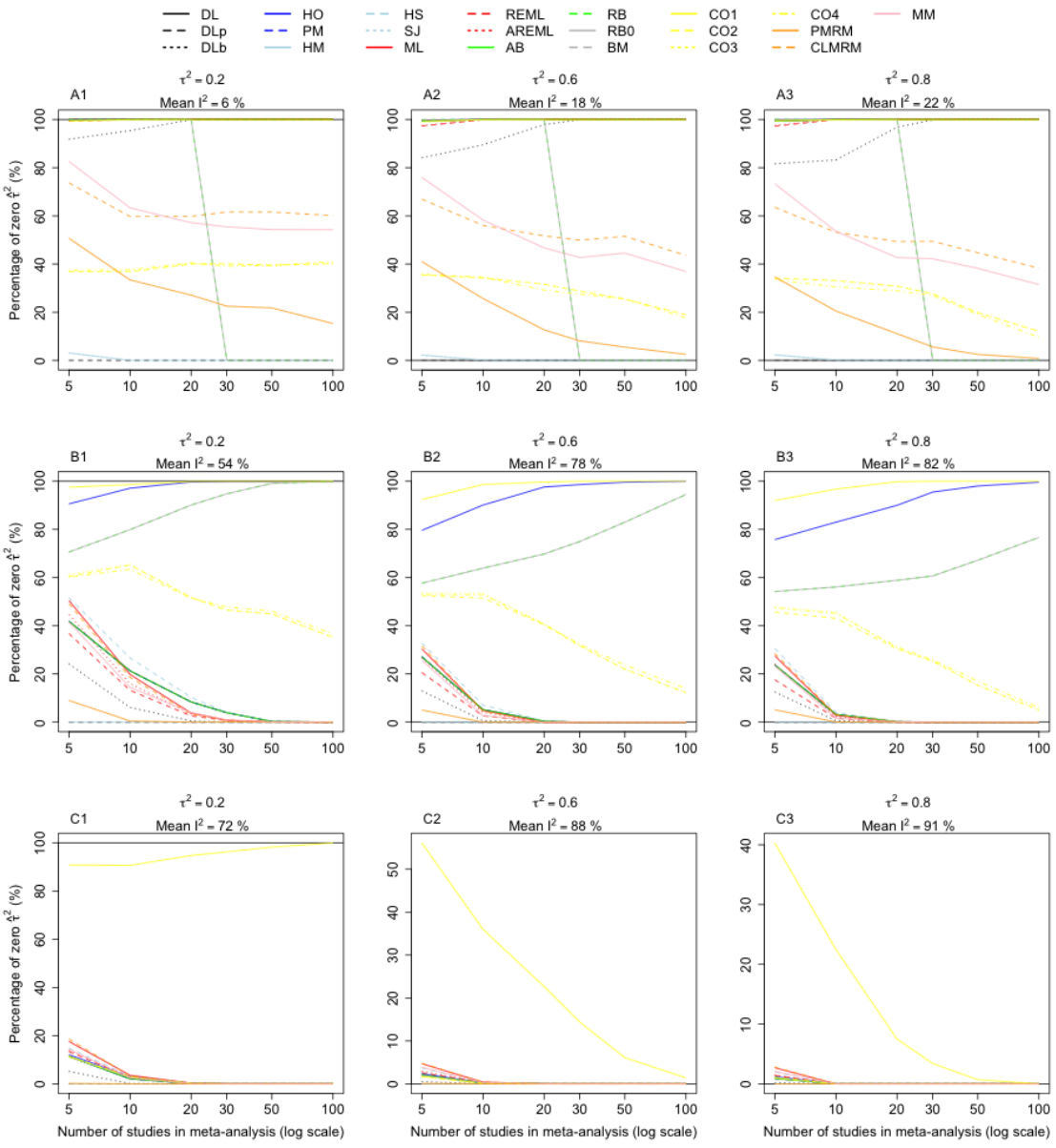


FIGURE E.42: Proportion of zero heterogeneity variance estimates in rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

E.3.2 Alternate study sample sizes

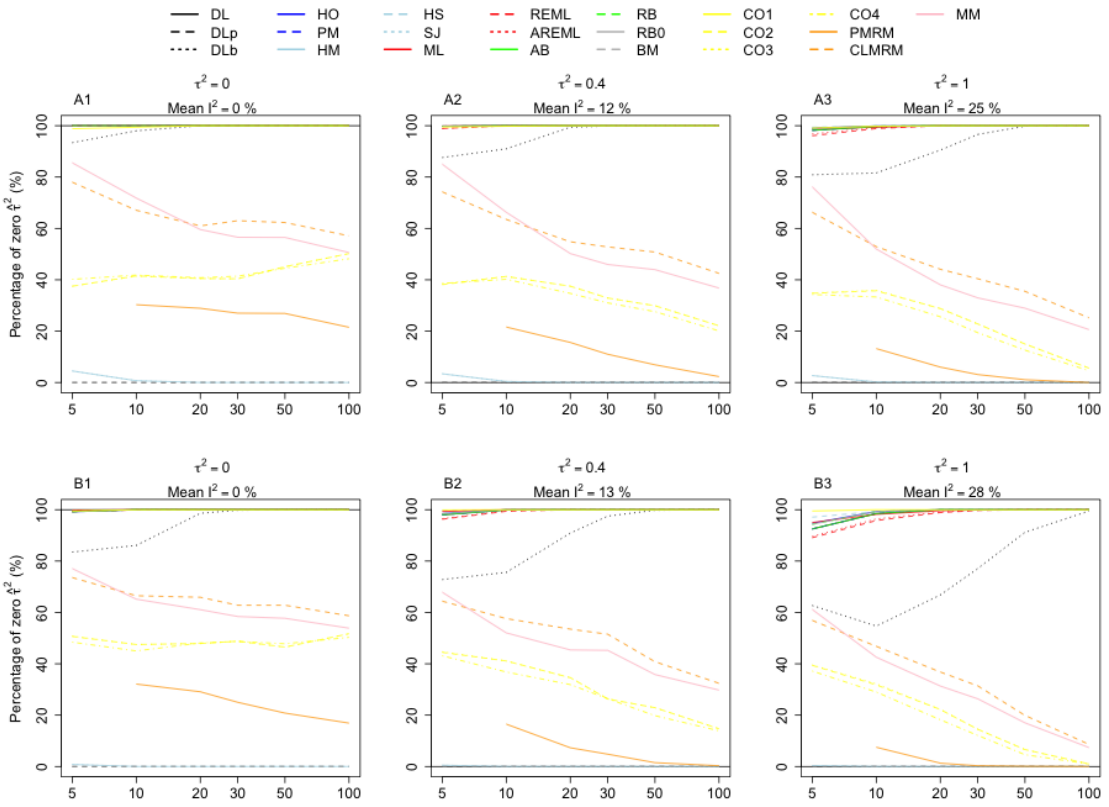


FIGURE E.43: Proportion of zero heterogeneity variance estimates in very rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

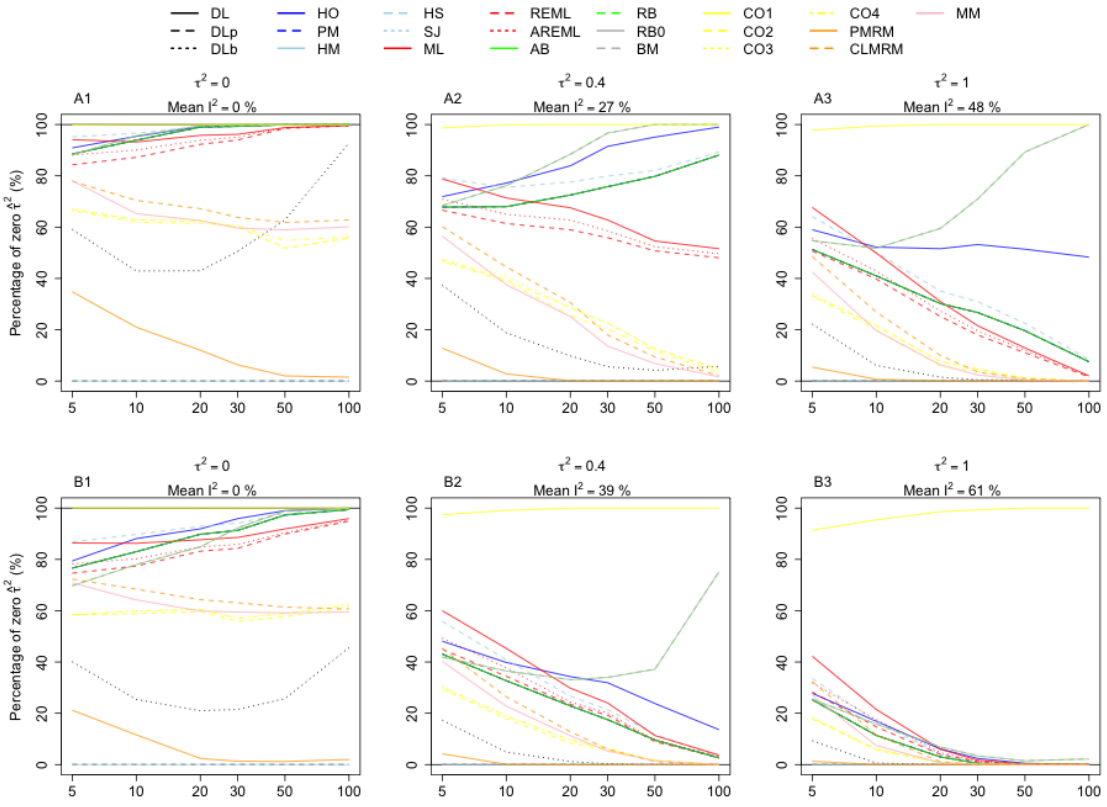


FIGURE E.44: Proportion of zero heterogeneity variance estimates in rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

E.3.3 Alternate values of σ_α^2

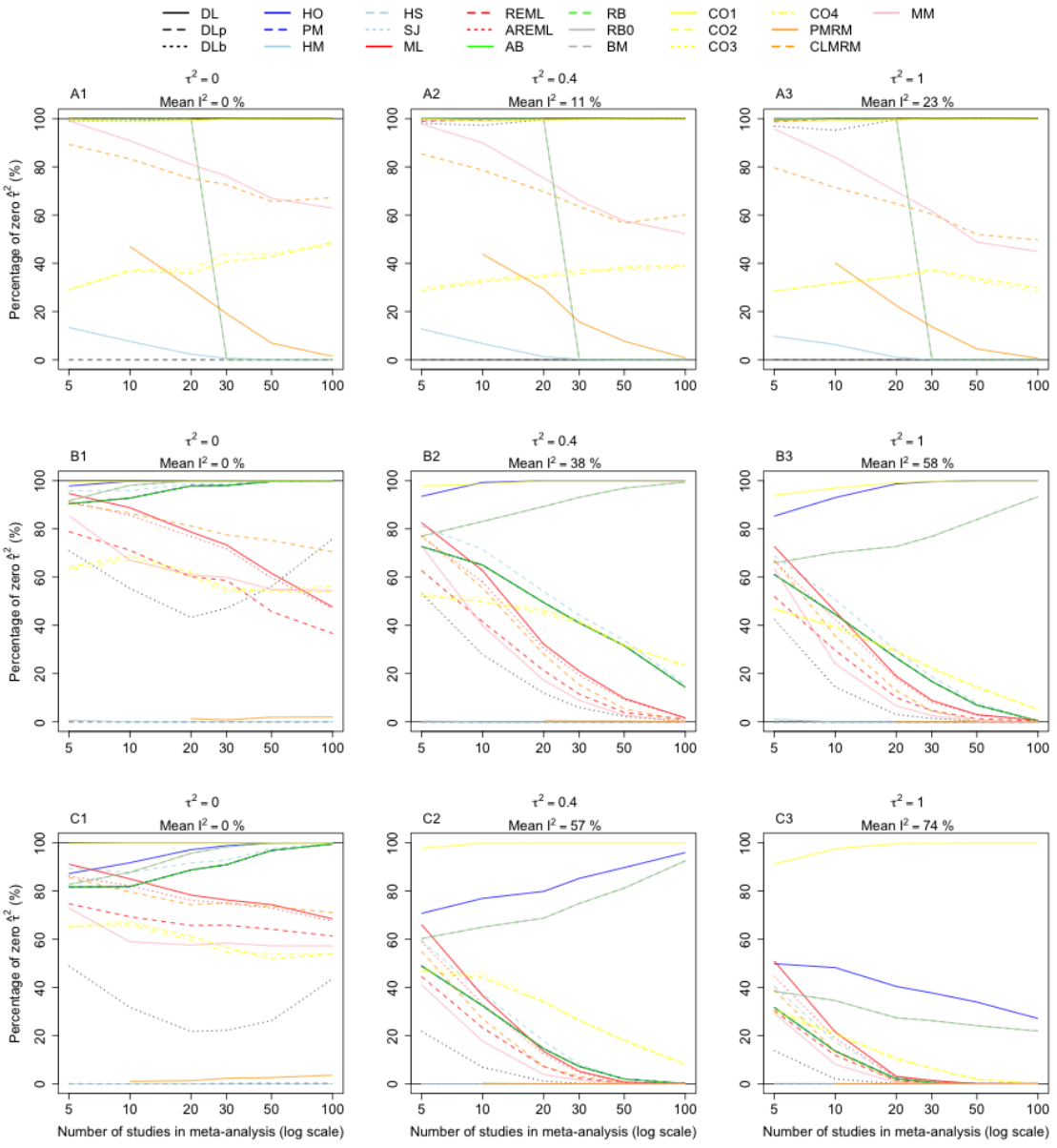


FIGURE E.45: Proportion of zero heterogeneity variance estimates in very rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

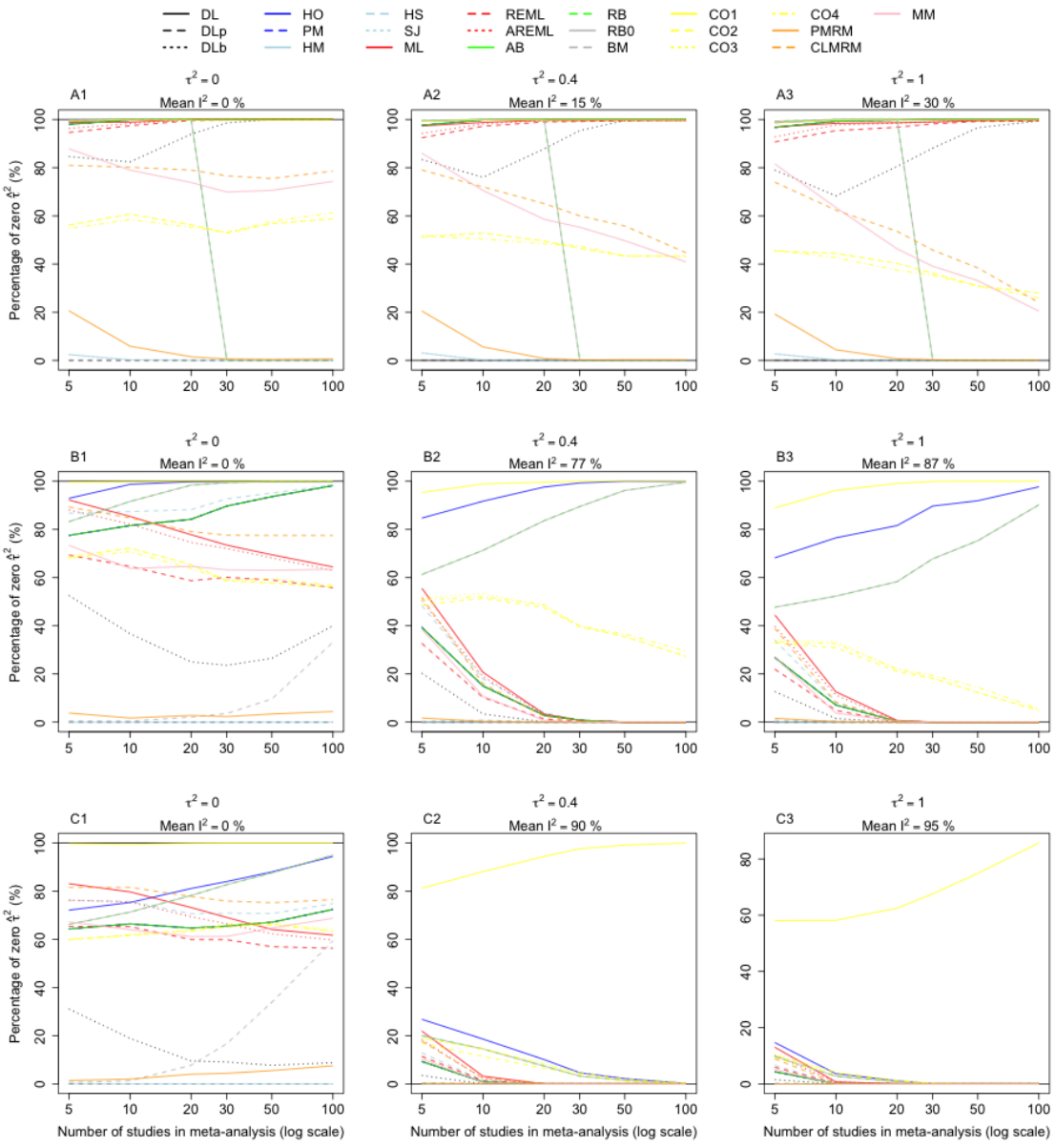


FIGURE E.46: Proportion of zero heterogeneity variance estimates in rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

E.3.4 Alternate probability scenarios

Alternate rare events scenarios

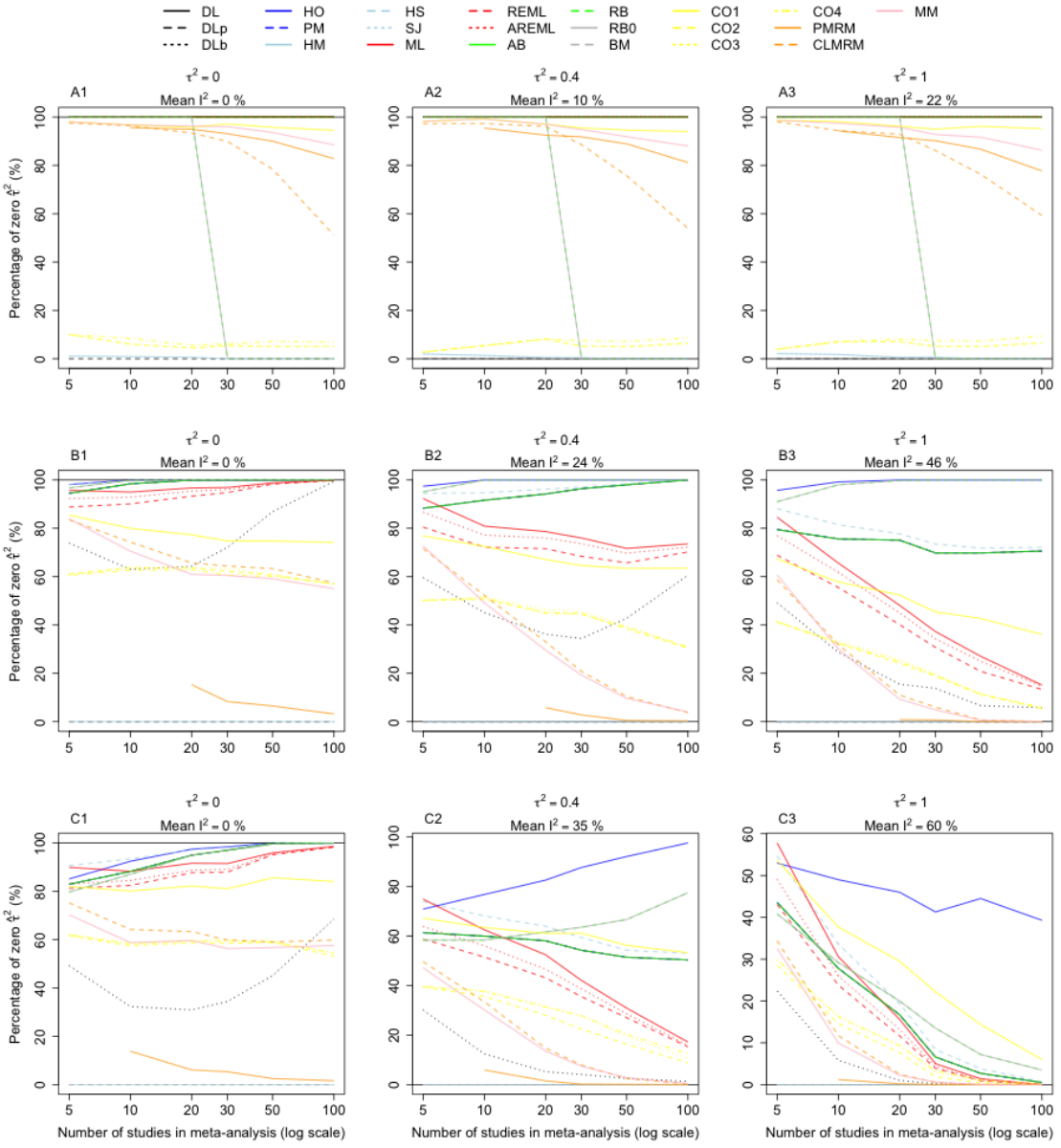


FIGURE E.47: Proportion of zero heterogeneity variance estimates in very rare events scenario with $p_0 > p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

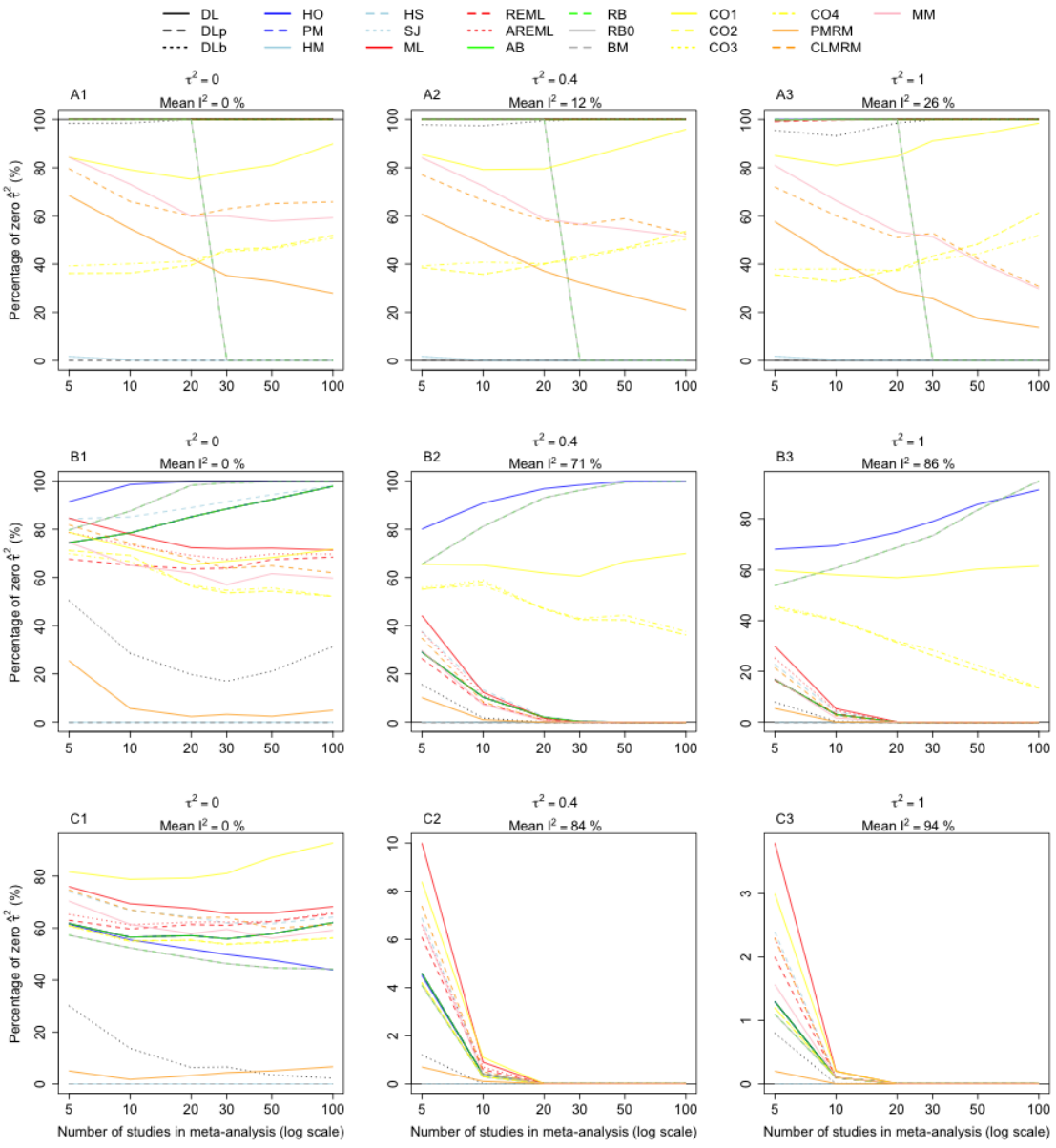


FIGURE E.48: Proportion of zero heterogeneity variance estimates in rare events scenario with $p_0 > p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

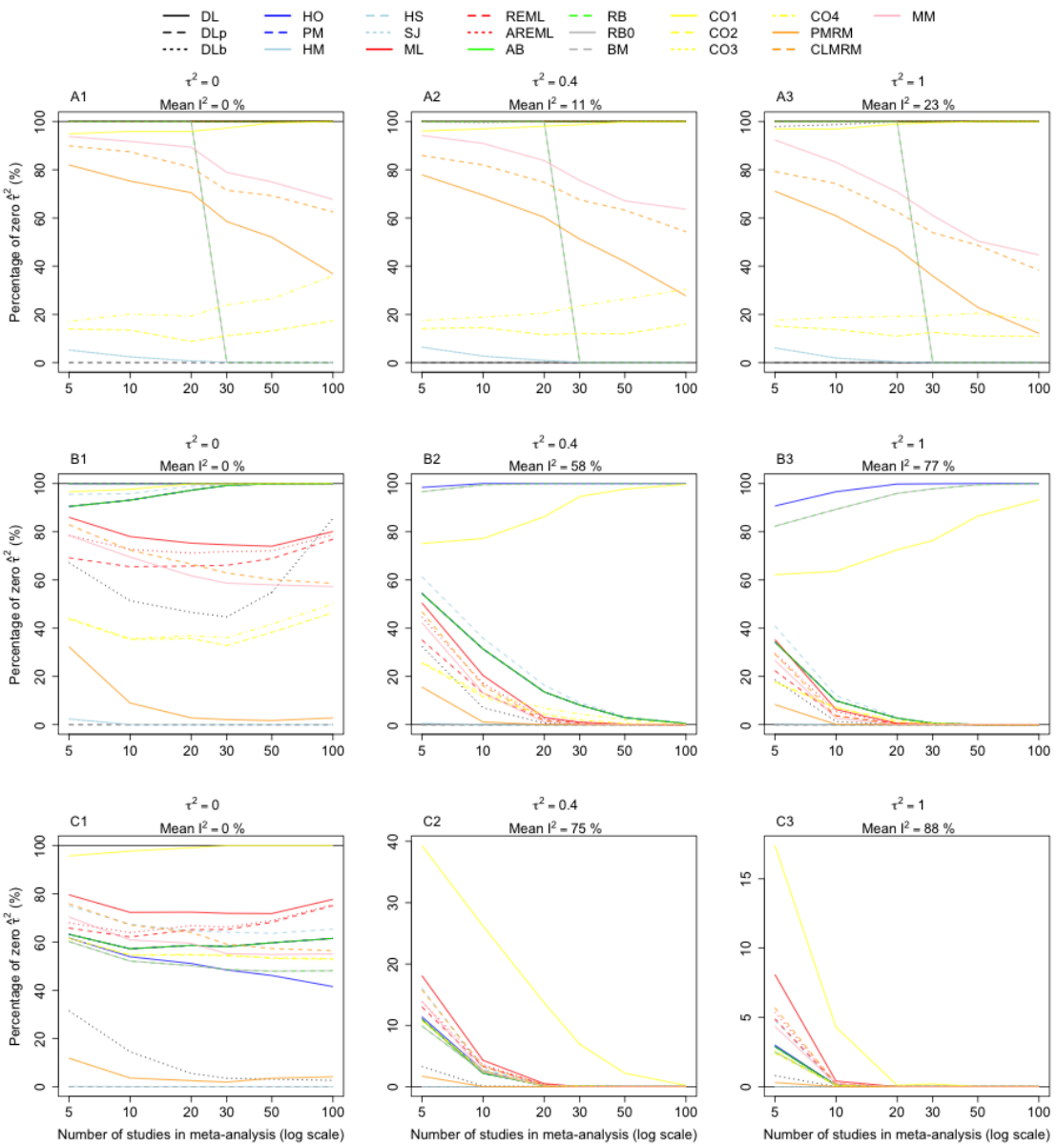


FIGURE E.49: Proportion of zero heterogeneity variance estimates in rare events scenario with $p_0 = p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

Common probability scenarios

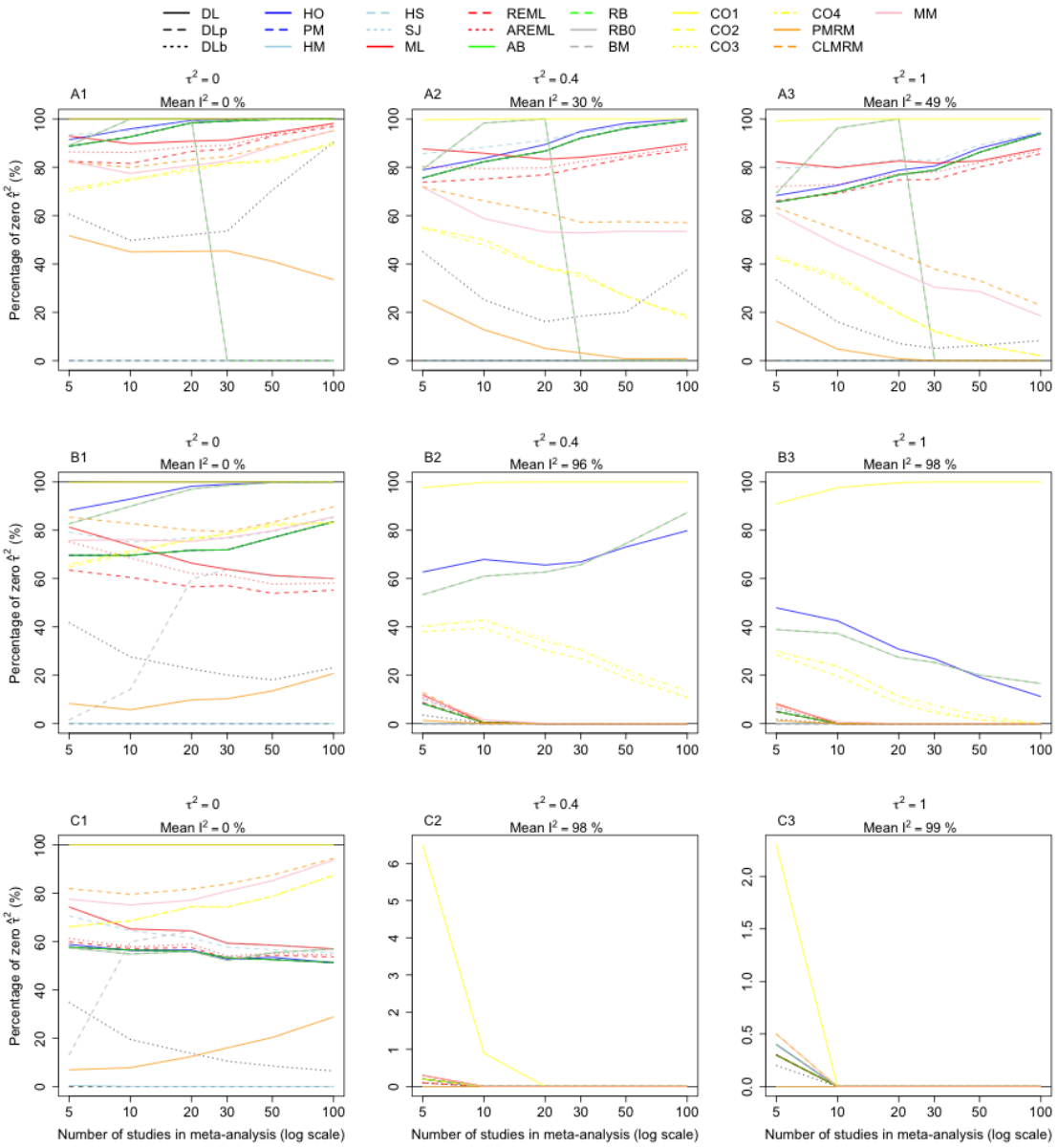


FIGURE E.50: Proportion of zero heterogeneity variance estimates in common probability scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

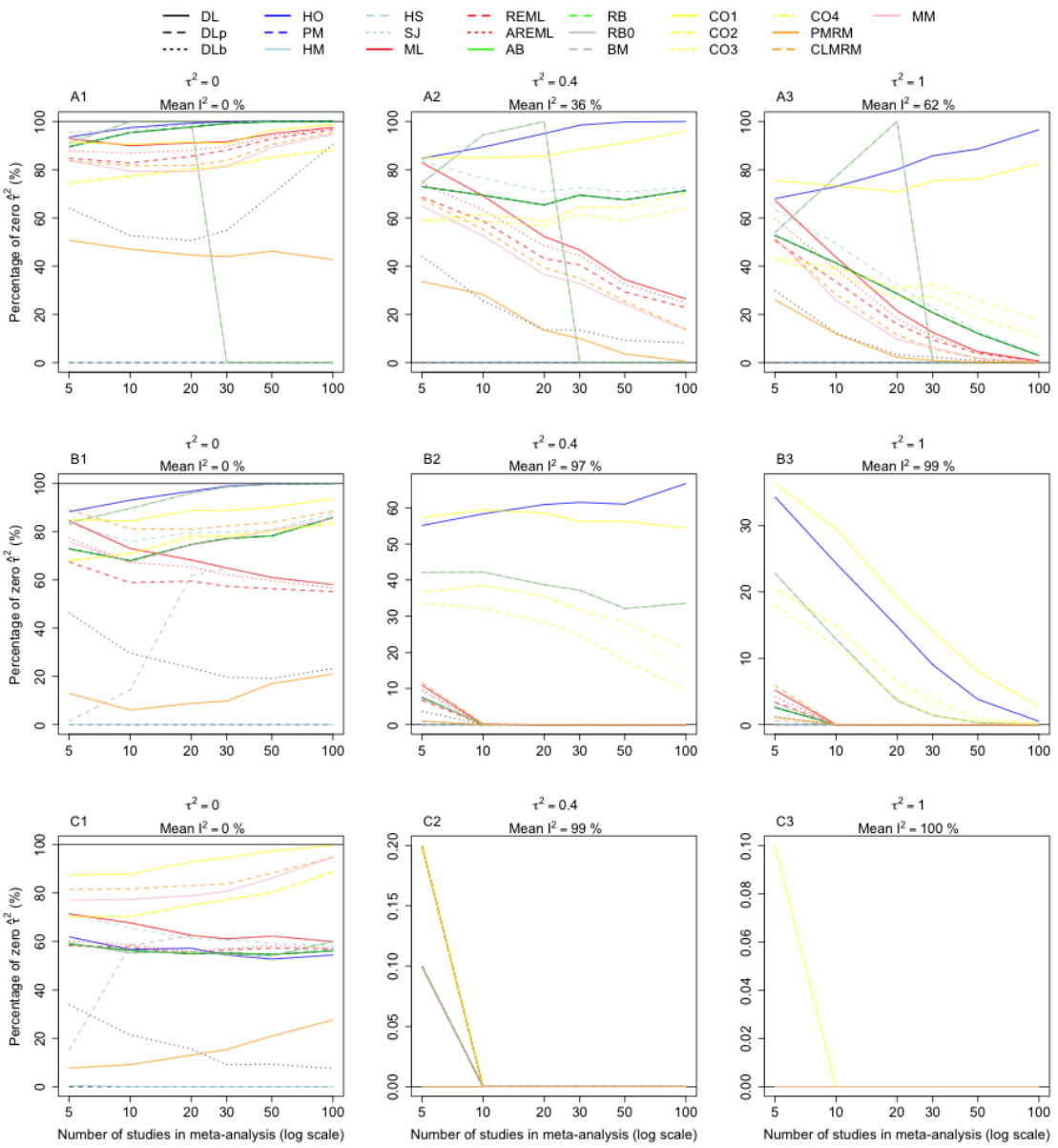


FIGURE E.51: Proportion of zero heterogeneity variance estimates in common probability scenario with $p_0 > p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

E.3.5 Alternate sampling in simulation study

Alternate event count sampling

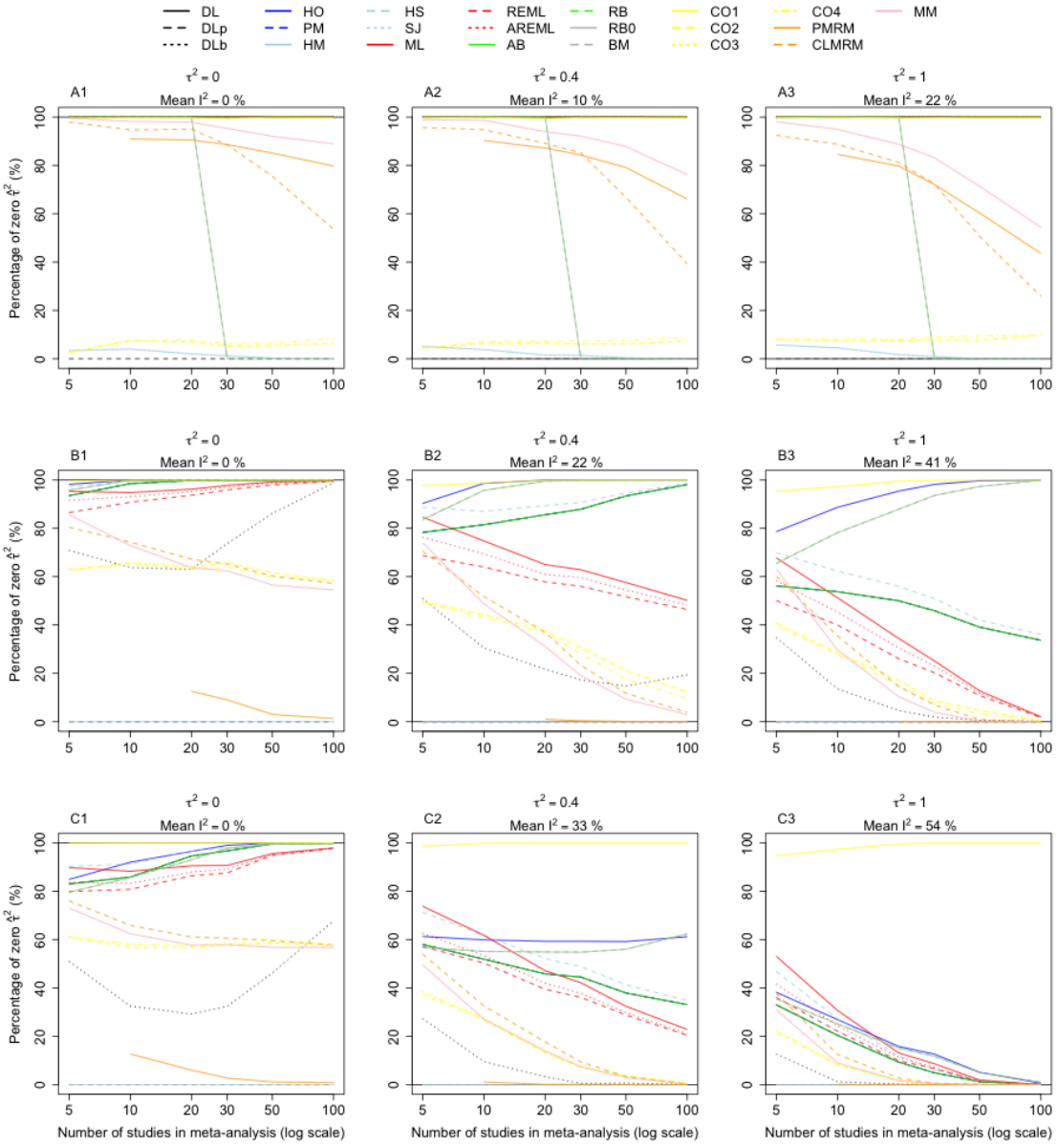


FIGURE E.52: Proportion of zero heterogeneity variance estimates in very rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

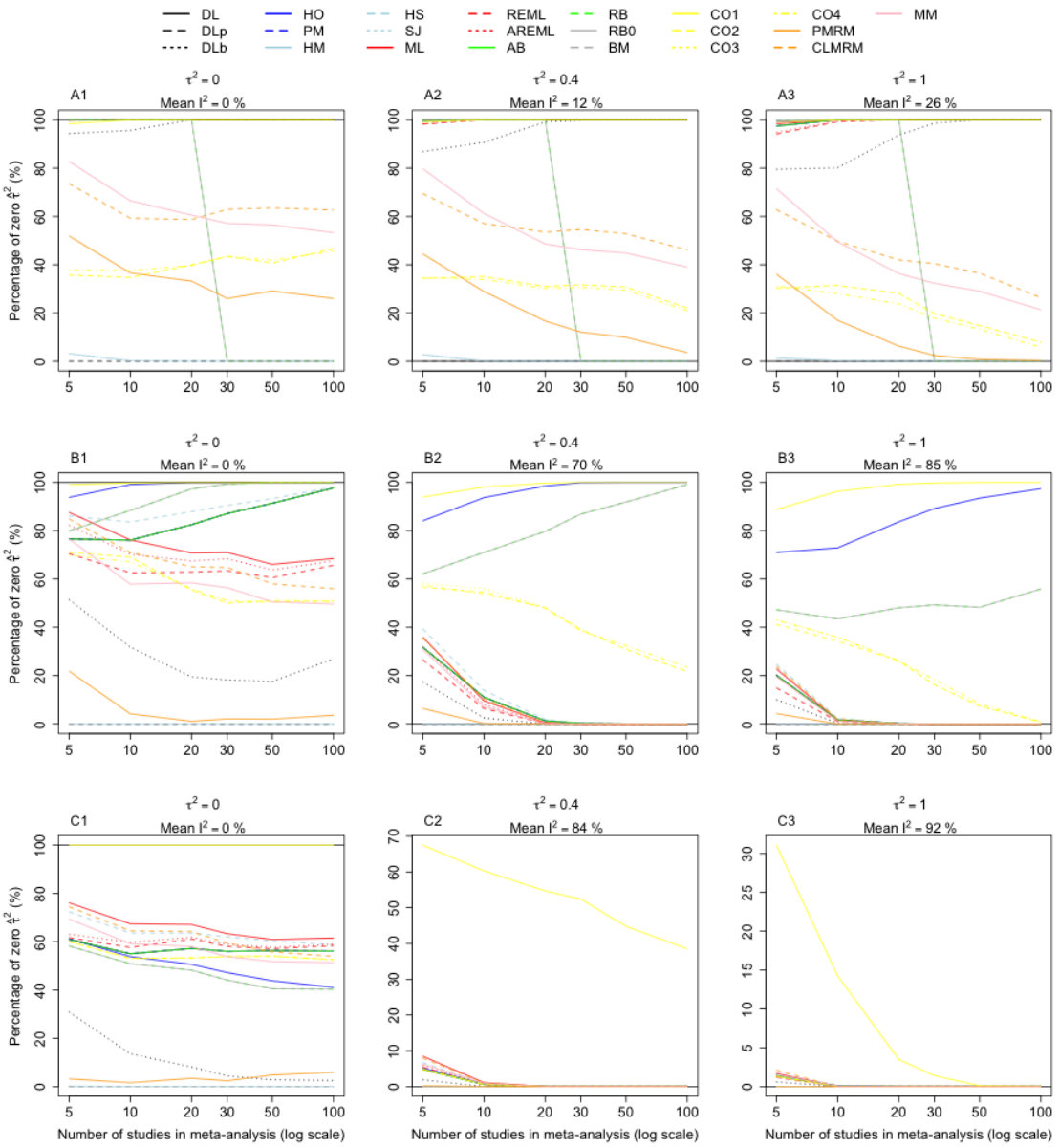


FIGURE E.53: Proportion of zero heterogeneity variance estimates in rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

Alternate sample size sampling

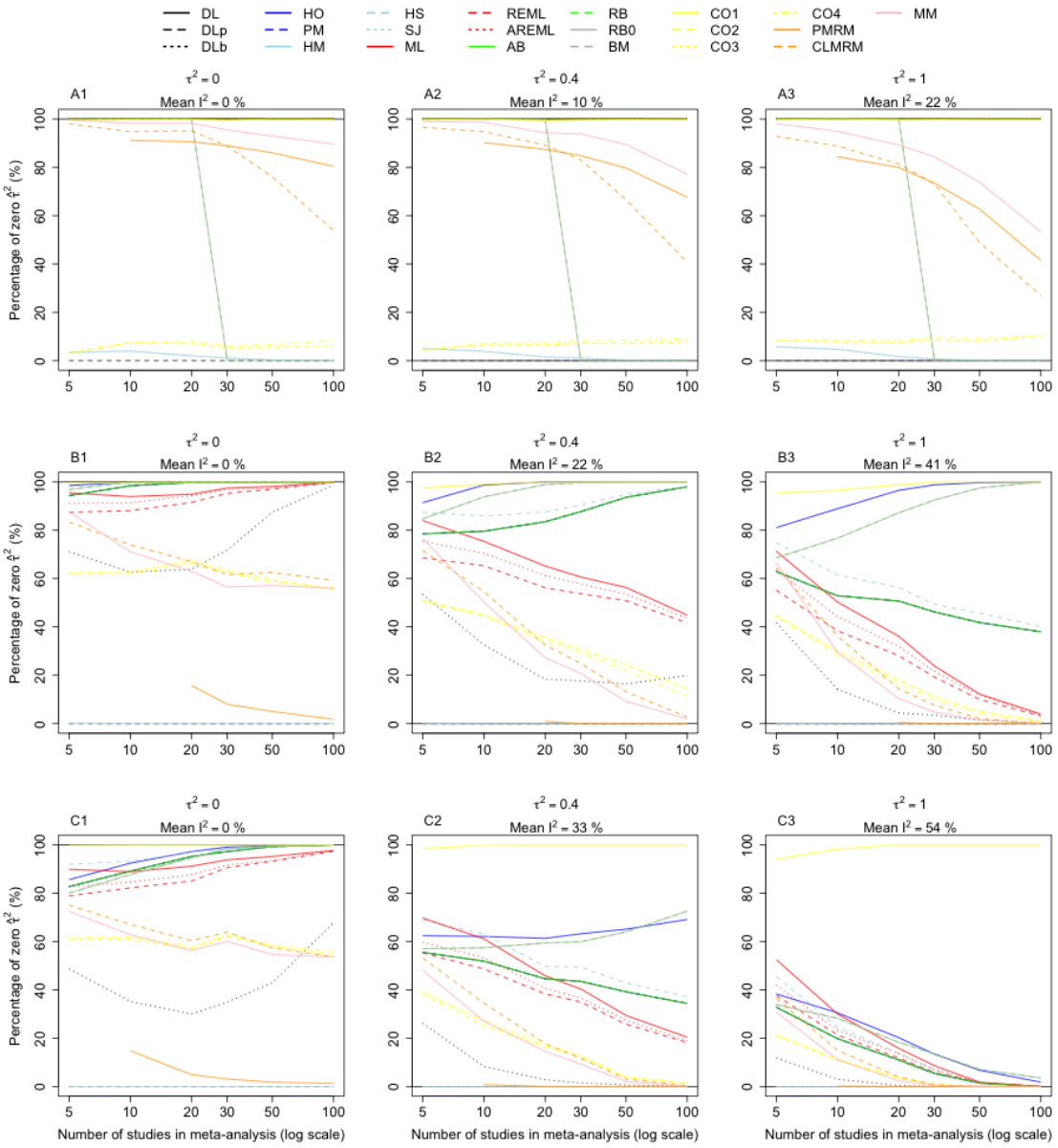


FIGURE E.54: Proportion of zero heterogeneity variance estimates in very rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

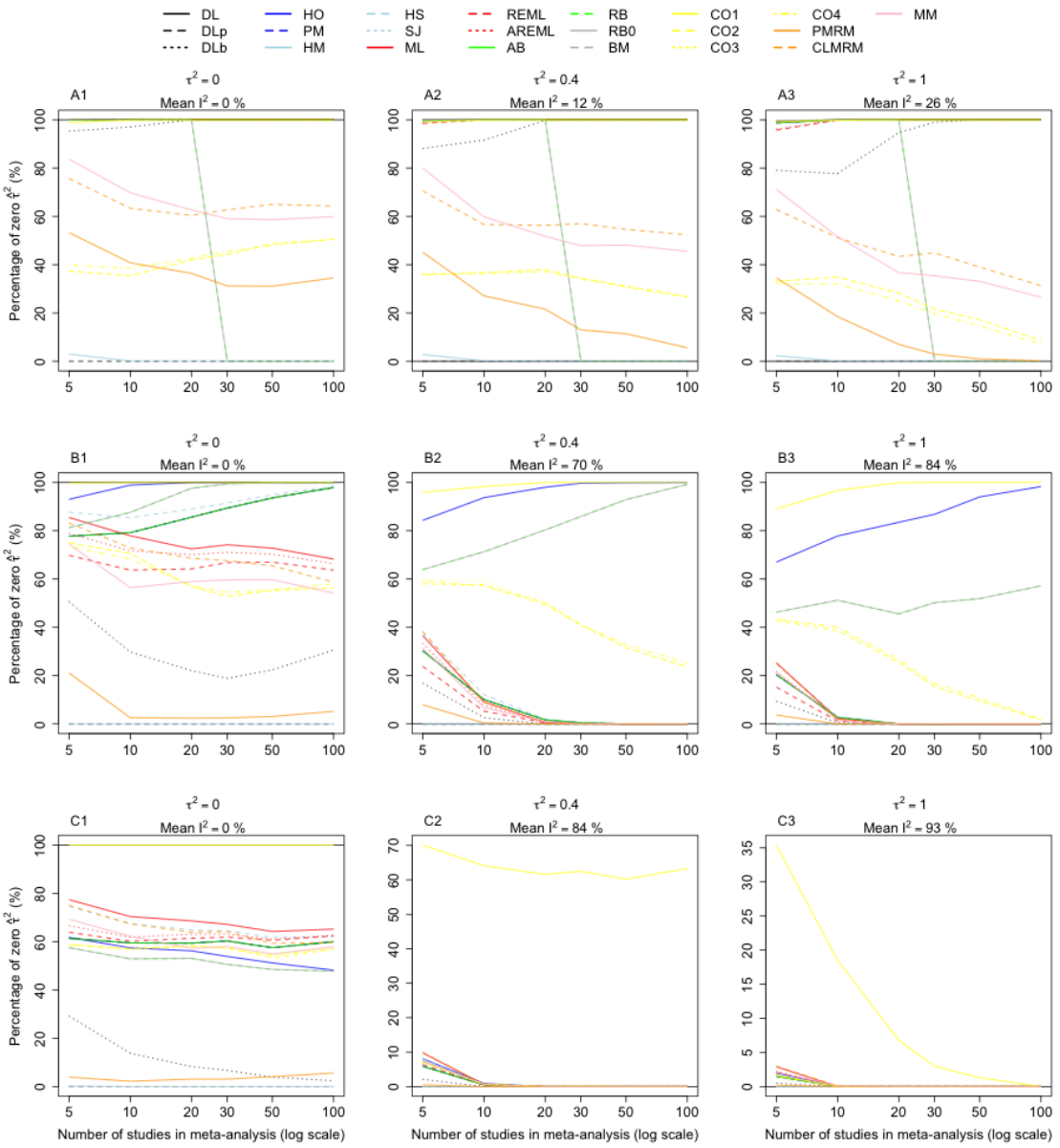


FIGURE E.55: Proportion of zero heterogeneity variance estimates in rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

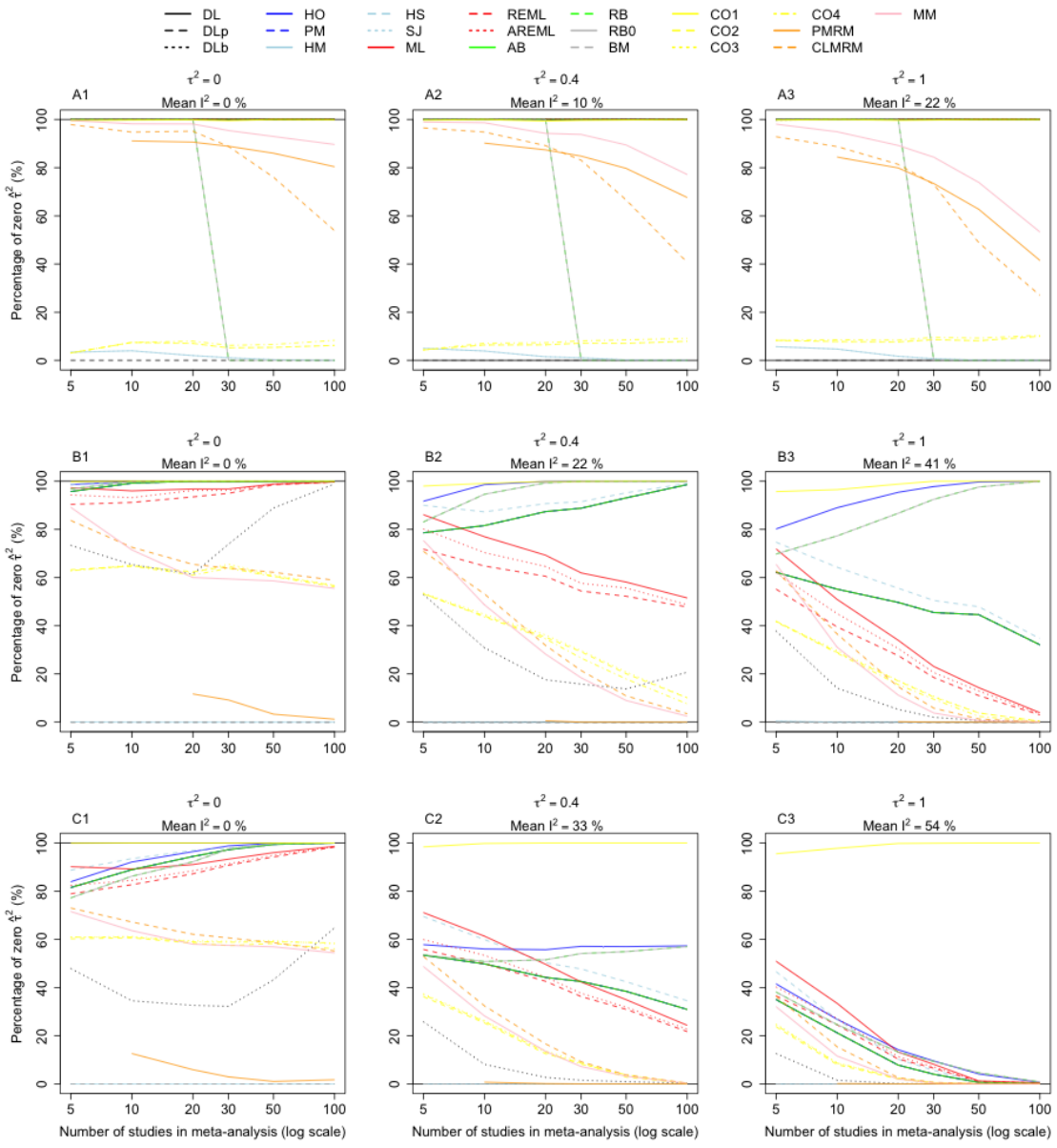


FIGURE E.56: Proportion of zero heterogeneity variance estimates in very rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

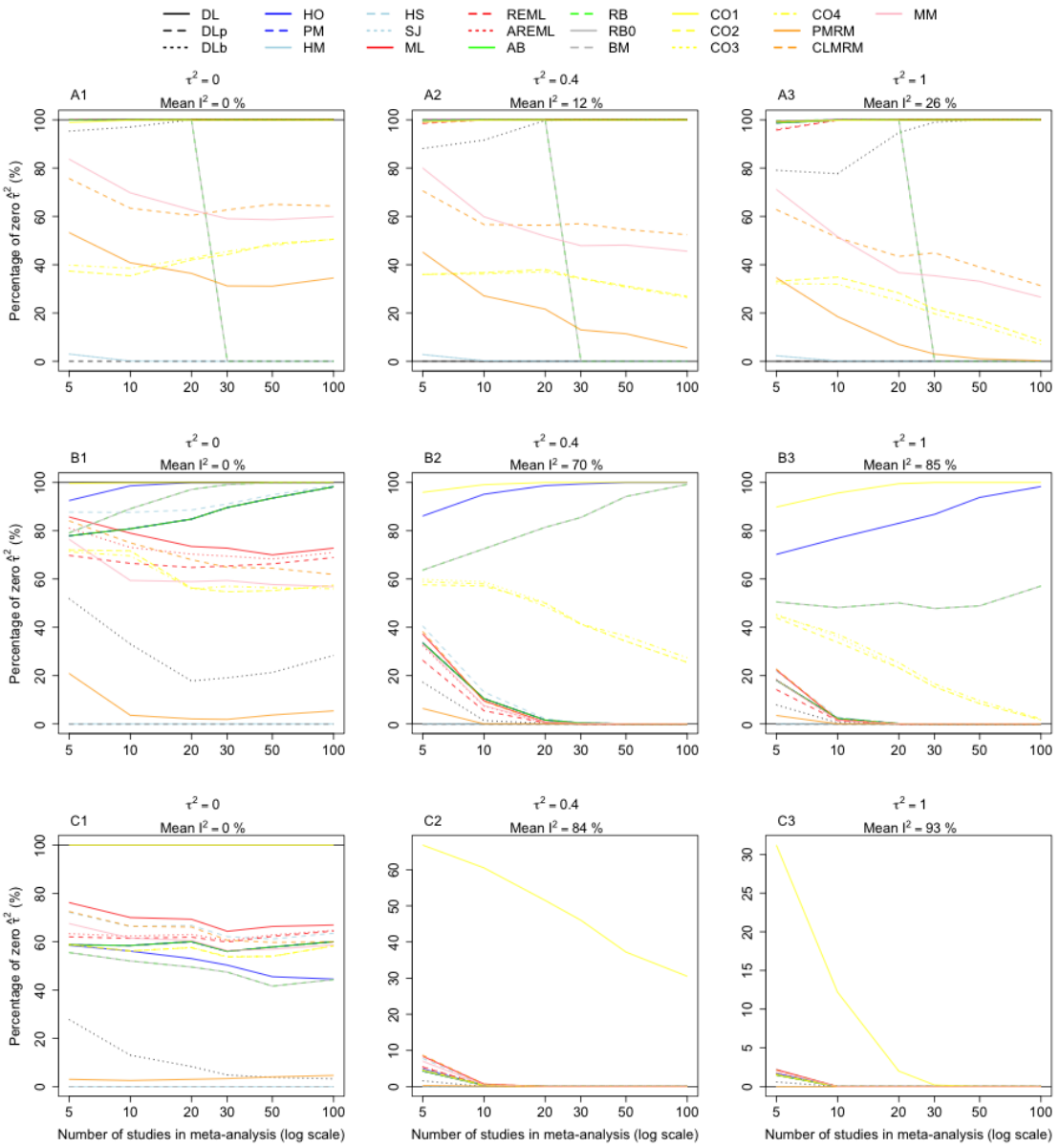


FIGURE E.57: Proportion of zero heterogeneity variance estimates in rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

E.3.6 Alternate continuity corrections

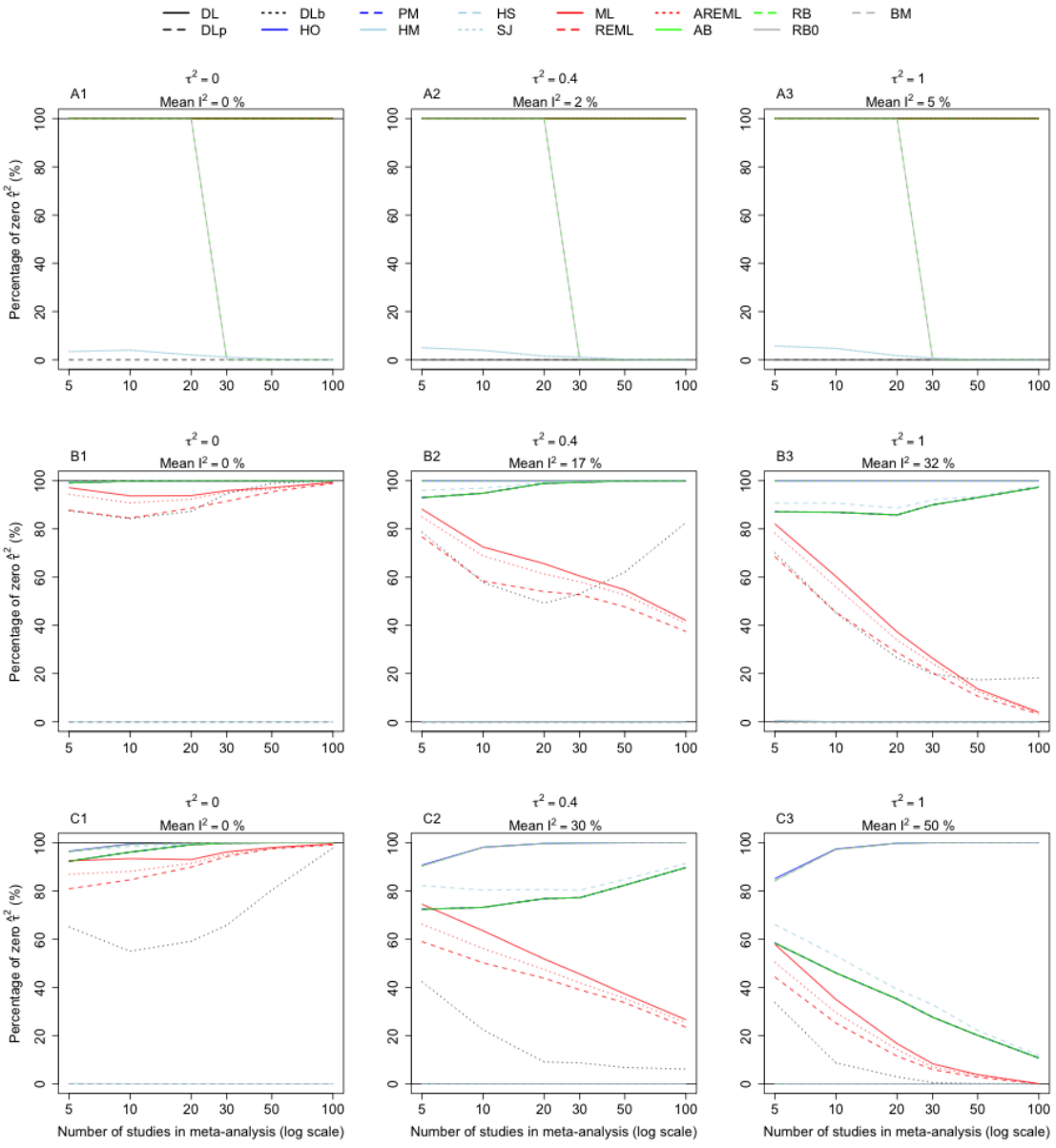


FIGURE E.58: Proportion of zero heterogeneity variance estimates in very rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

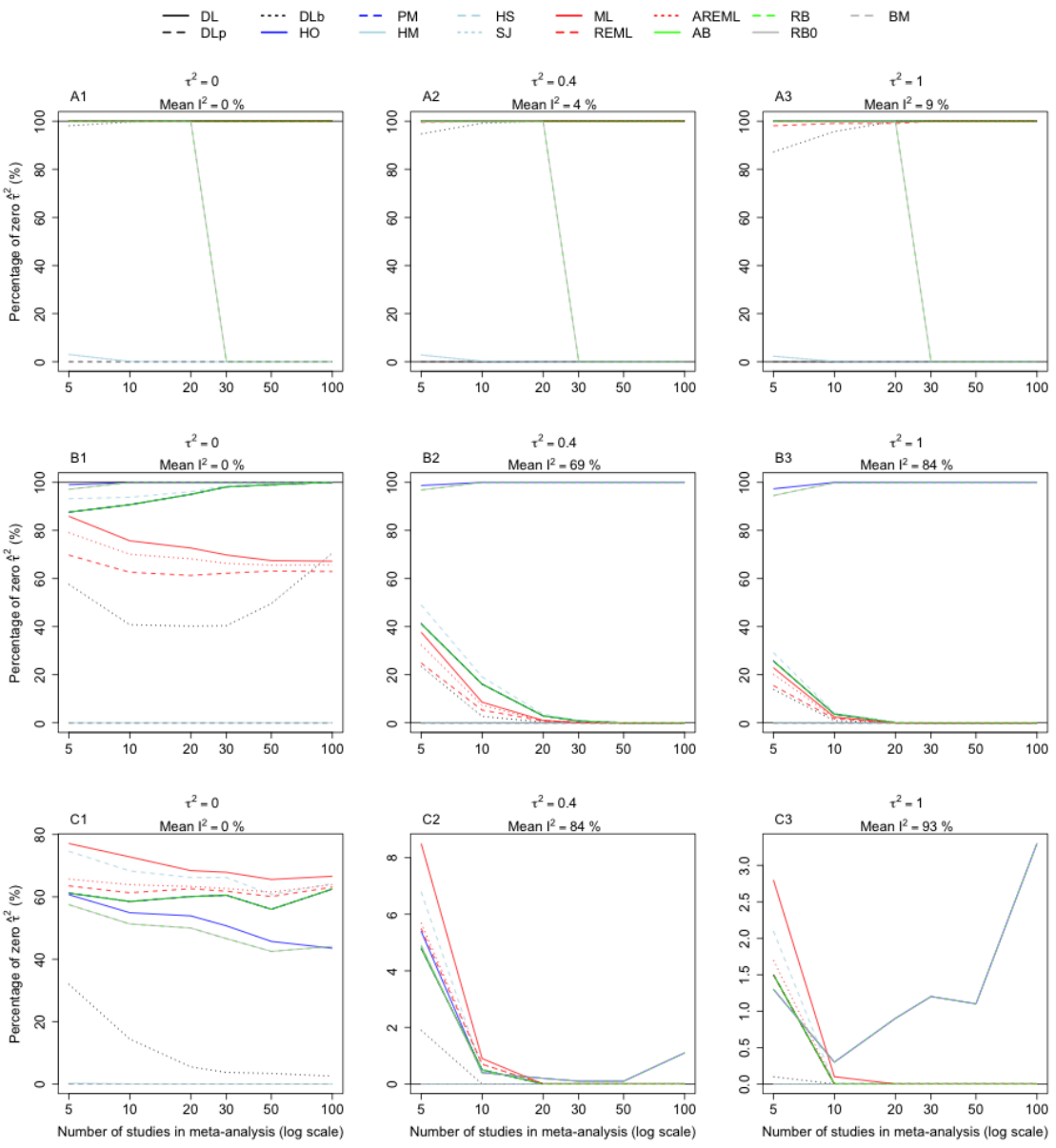


FIGURE E.59: Proportion of zero heterogeneity variance estimates in rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

E.4 Median bias of τ^2

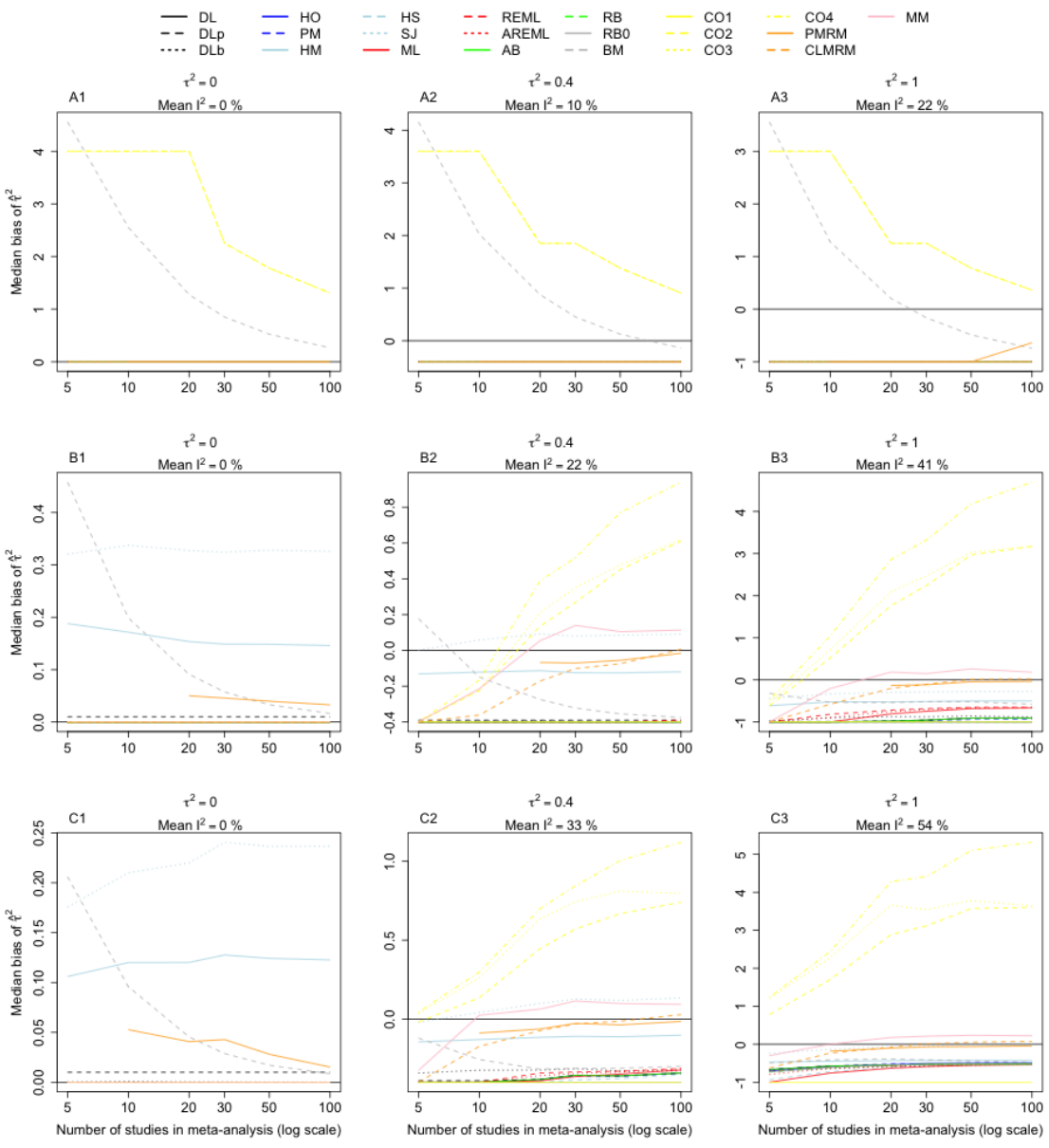


FIGURE E.60: Median bias of heterogeneity variance estimates in very rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). RB, RB0 and CLMRM are omitted from A1-A3.

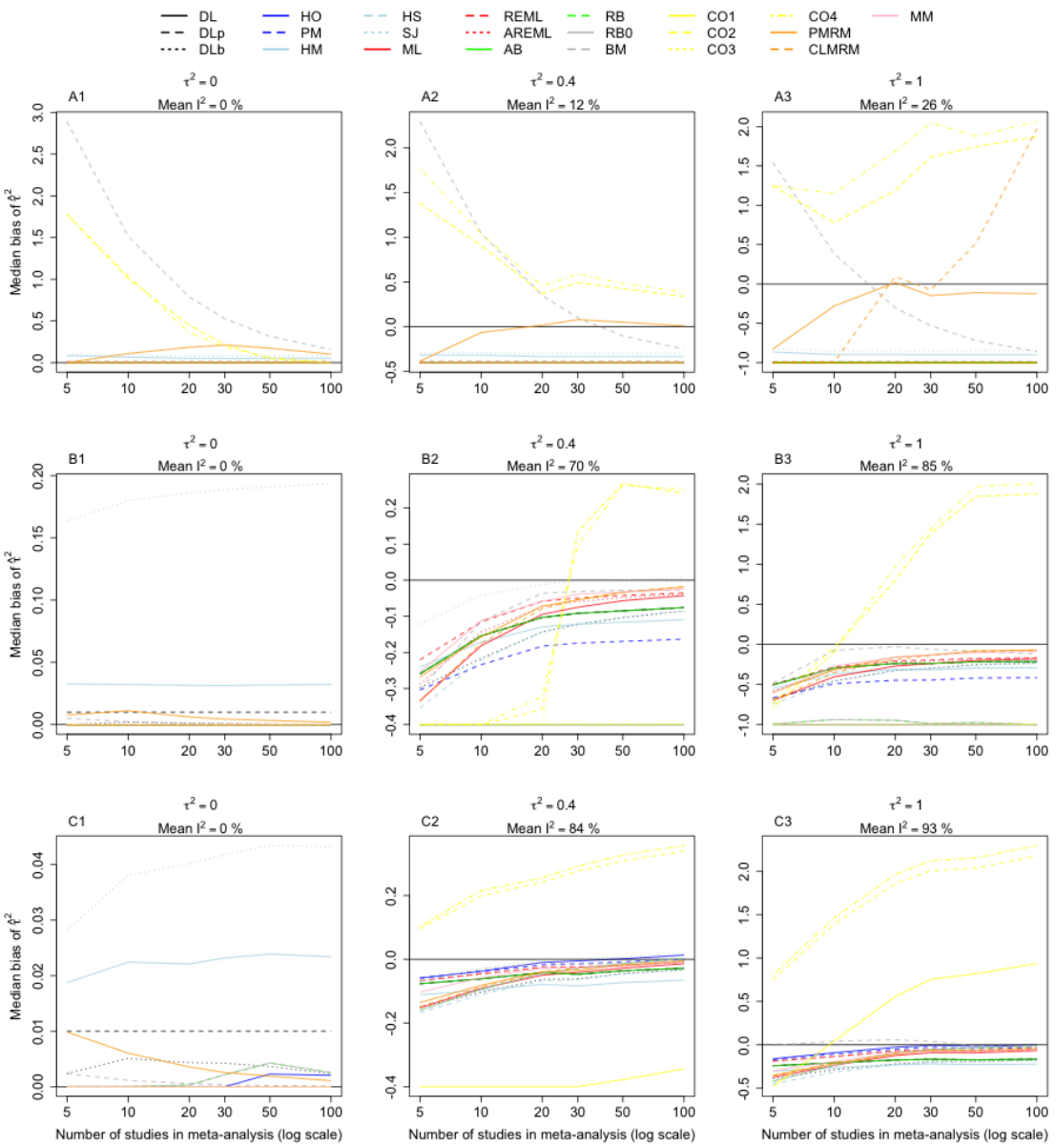


FIGURE E.61: Median bias of heterogeneity variance estimates in rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). RB and RB0 are omitted from A1-A3; MM is omitted from A3.

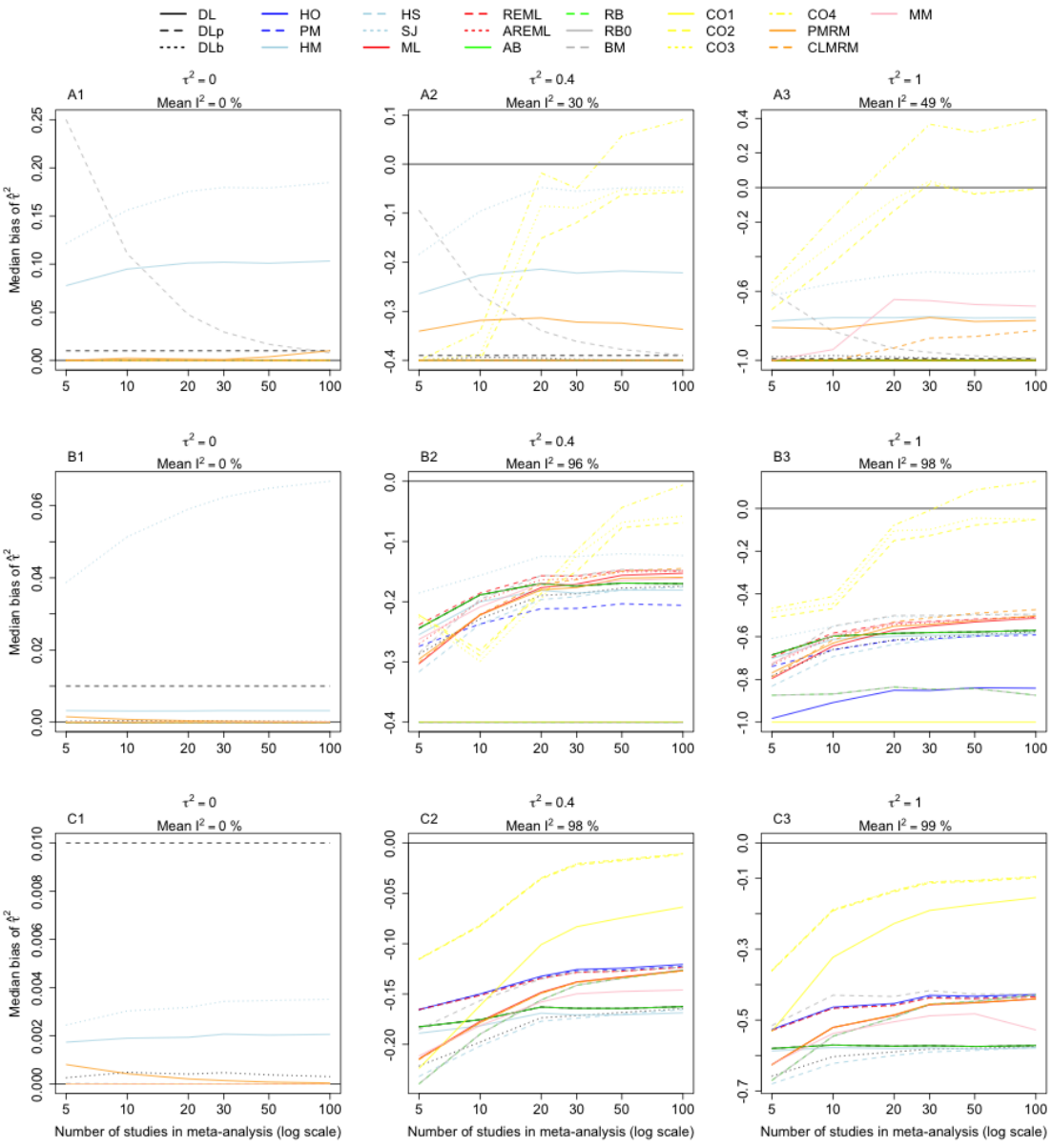


FIGURE E.62: Median bias of heterogeneity variance estimates in common probability scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). RB and RB0 are omitted from A1-A3.

E.5 Median squared error of τ^2

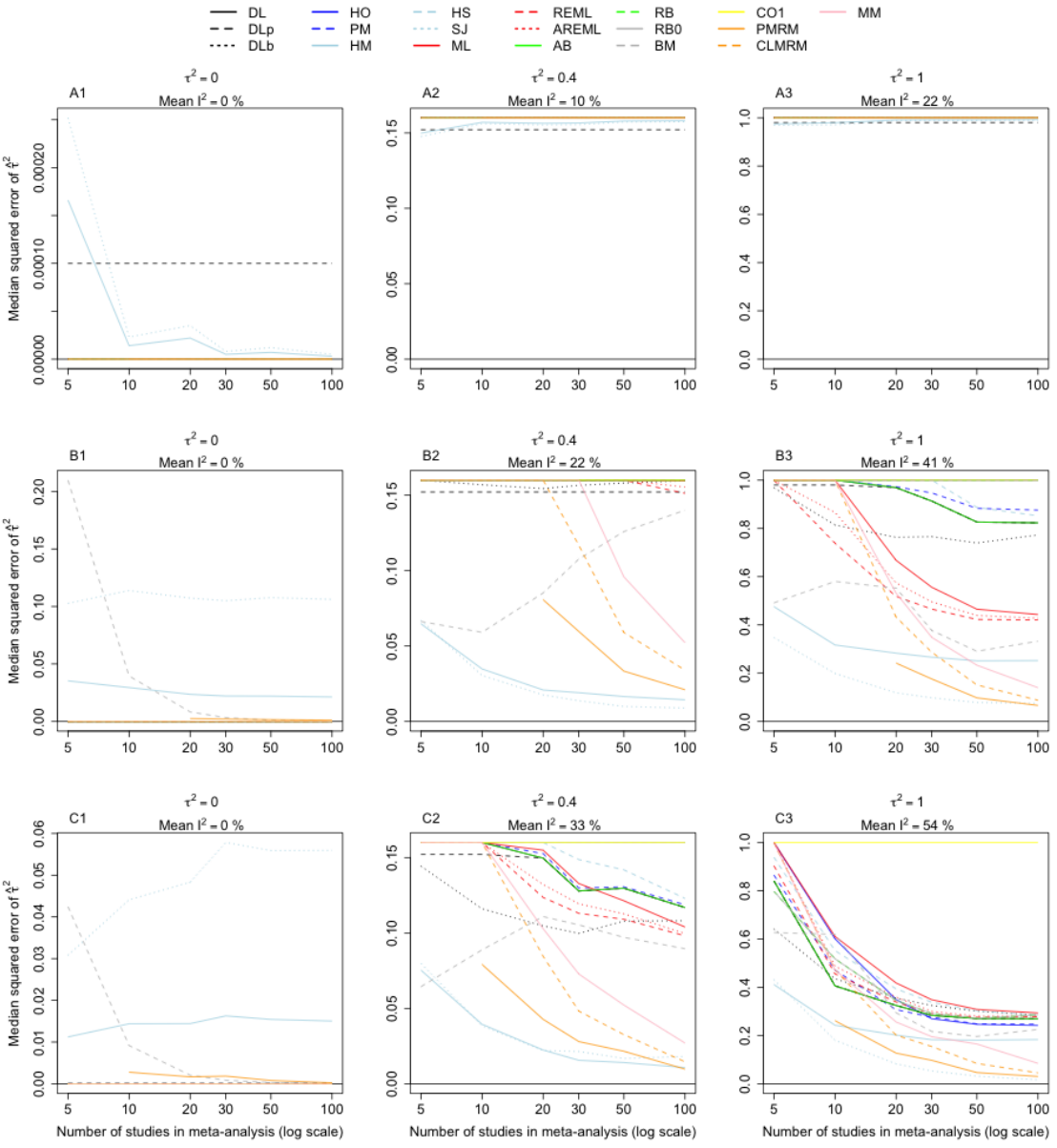


FIGURE E.63: Median squared error of heterogeneity variance estimates in very rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). RB, RB0, BM and CLMRM are omitted from A1-A3; CO2, CO3 and CO4 are omitted from all.

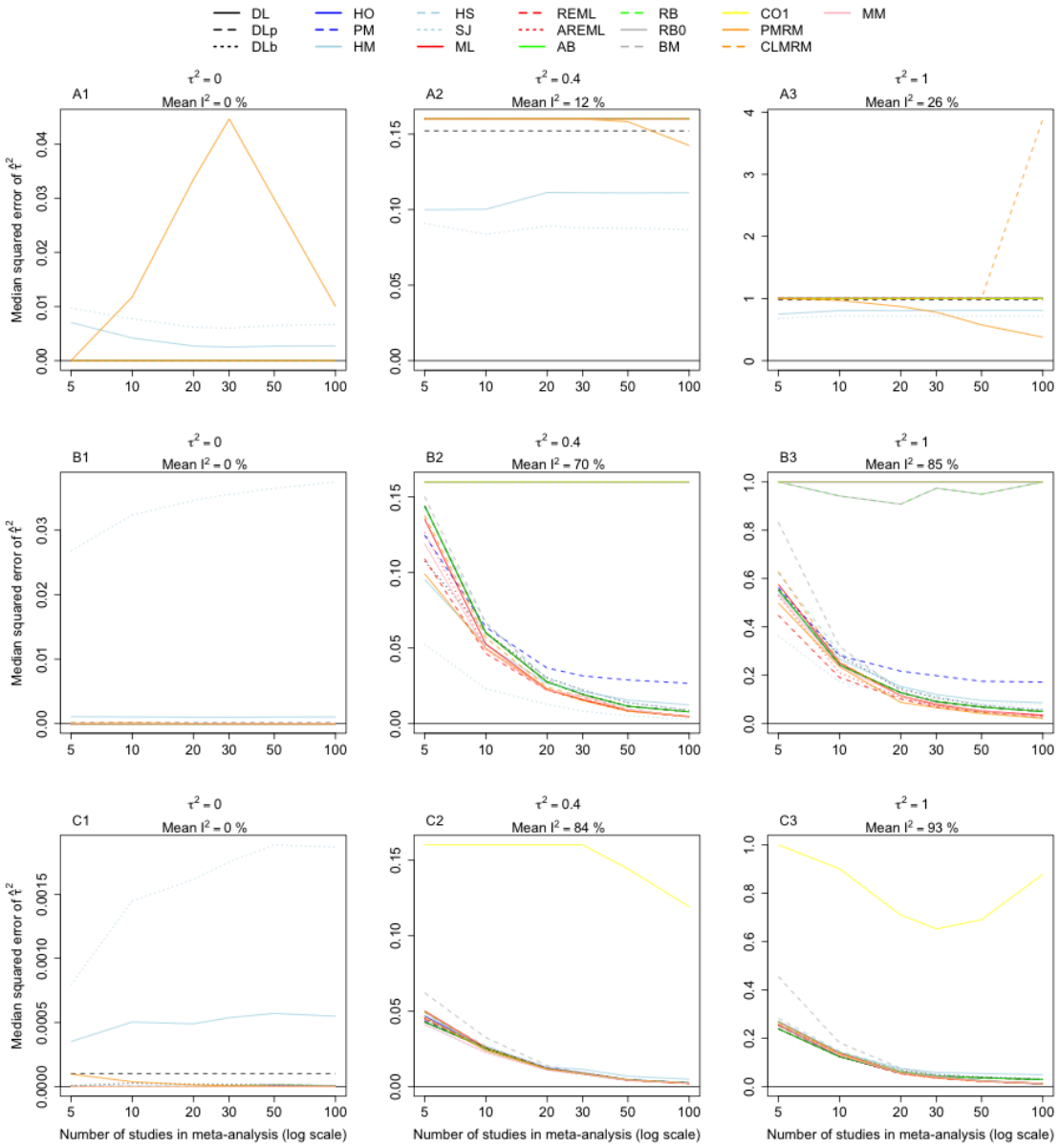


FIGURE E.64: Median squared error of heterogeneity variance estimates in rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). RB, RB0, BM and CLMRM are omitted from A1-A3; MM is omitted from A3; CO2, CO3 and CO4 are omitted from all.

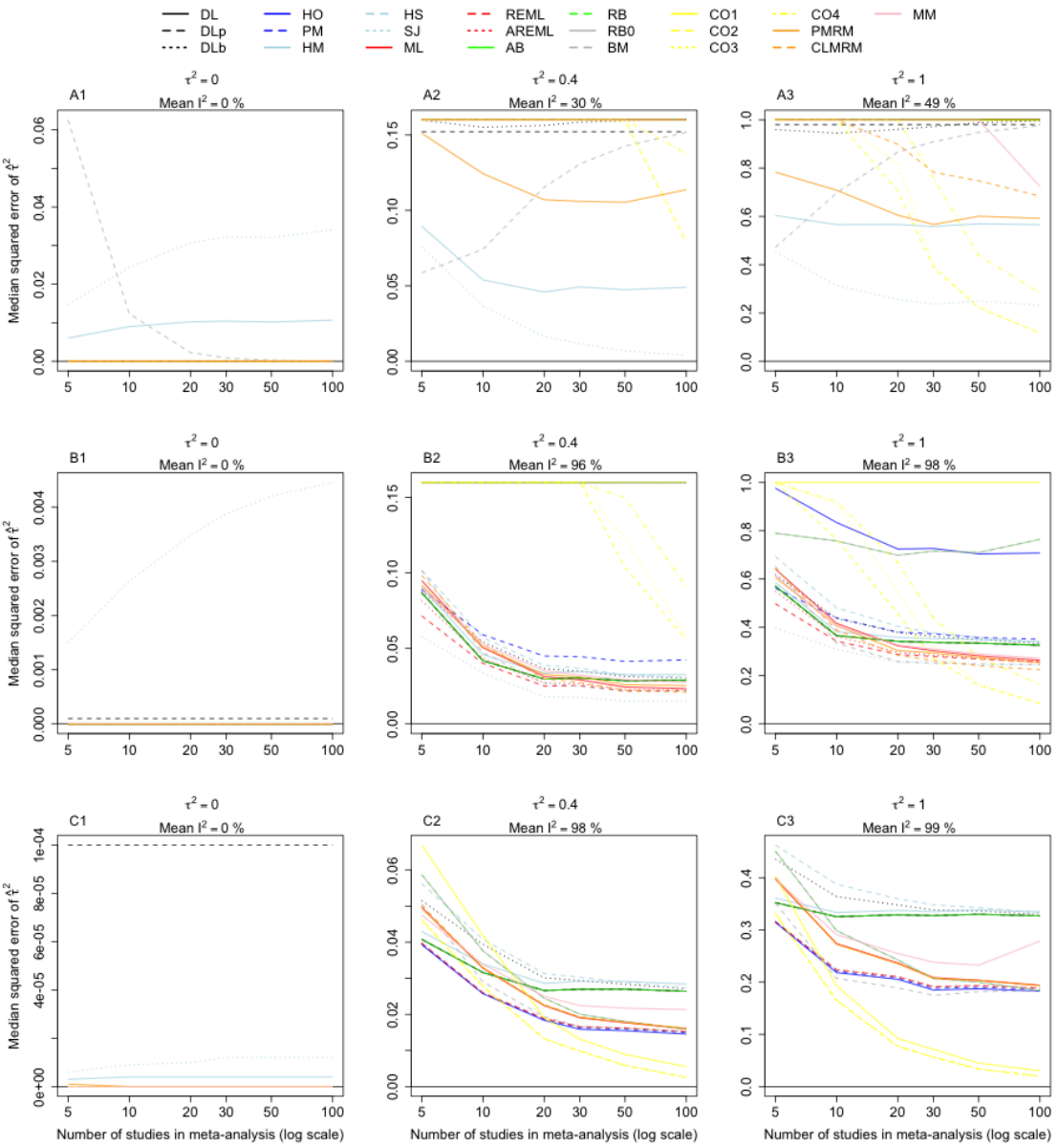


FIGURE E.65: Median squared error of heterogeneity variance estimates in common probability scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). RB and RB0 are omitted from A1-A3.

E.6 Performance of conditional-based methods in estimating τ_p^2

E.6.1 Mean bias of τ_p^2

Very rare events scenarios

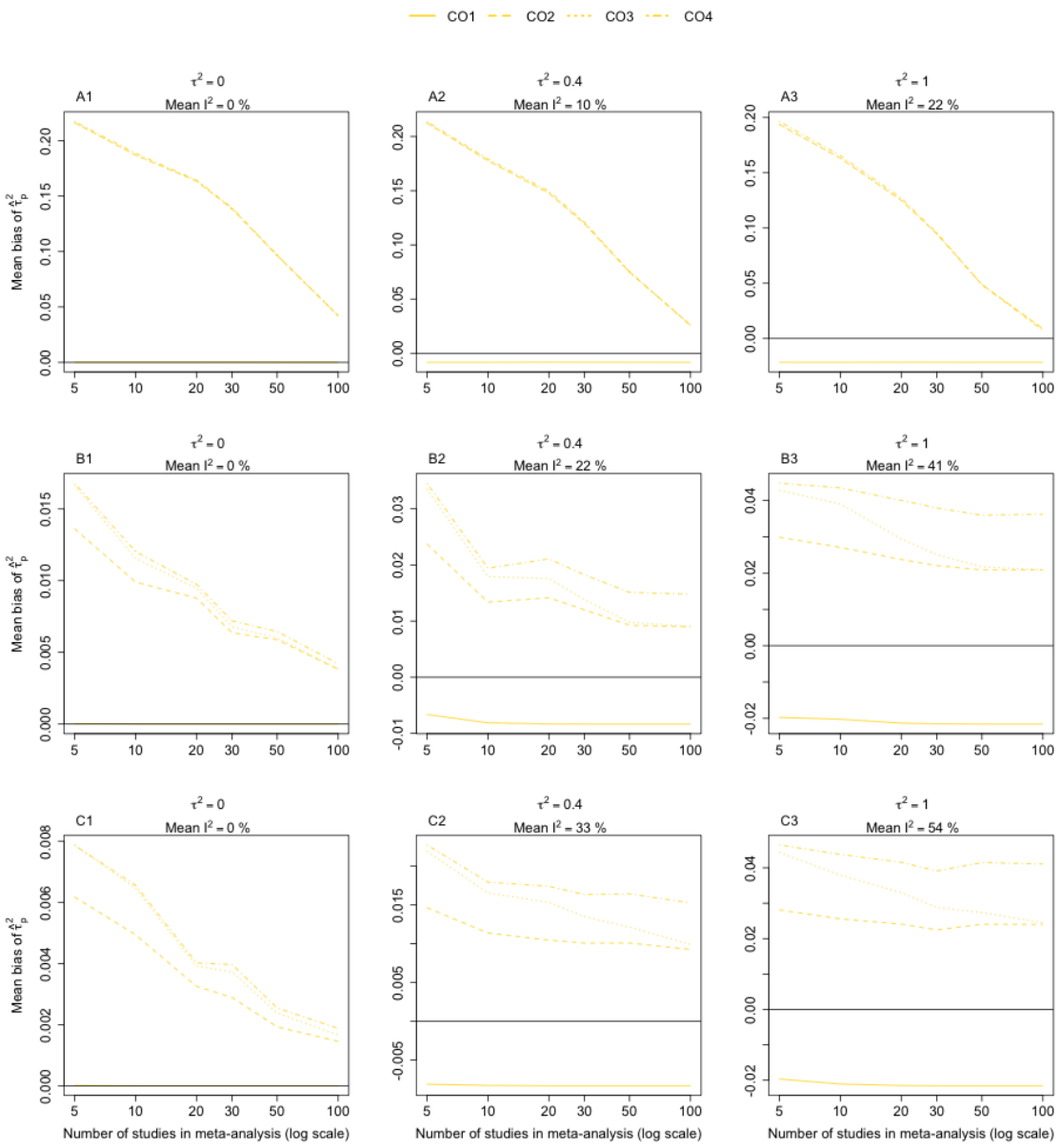


FIGURE E.66: Mean bias of τ_p^2 estimates in very rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

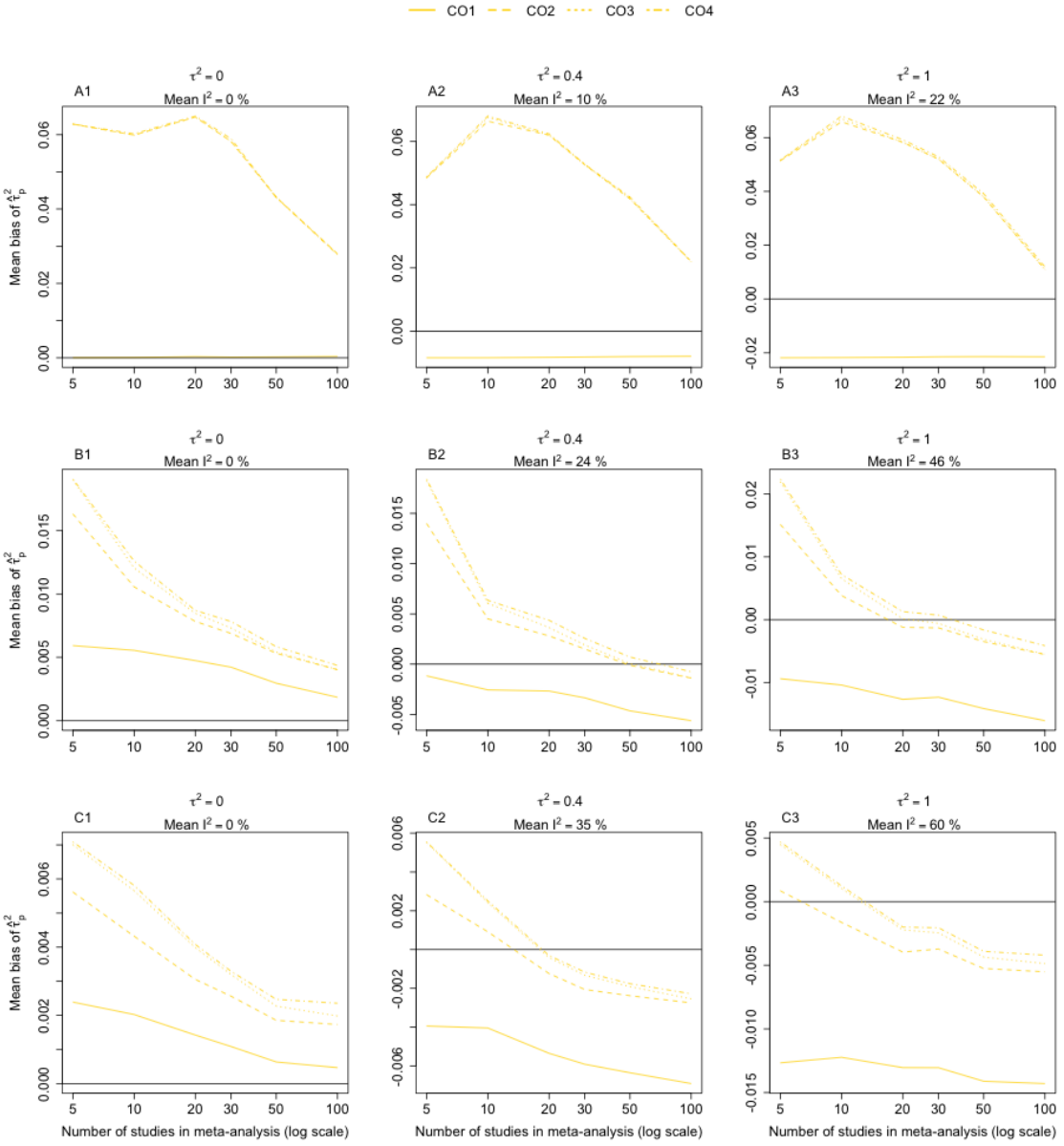


FIGURE E.67: Mean bias of τ_p^2 estimates in very rare events scenario with $p_0 > p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

Rare events scenarios

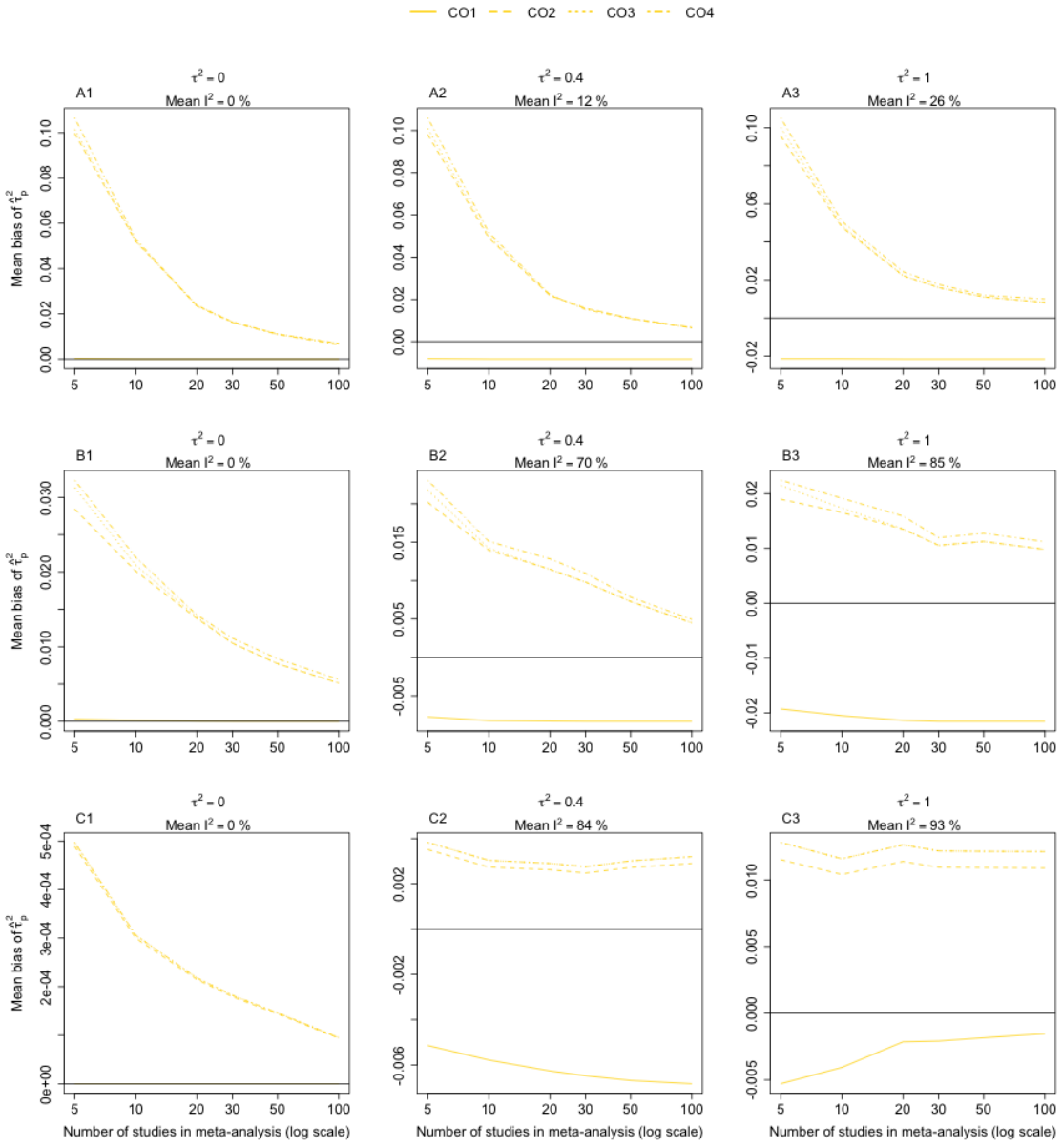


FIGURE E.68: Mean bias of τ_p^2 estimates in rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

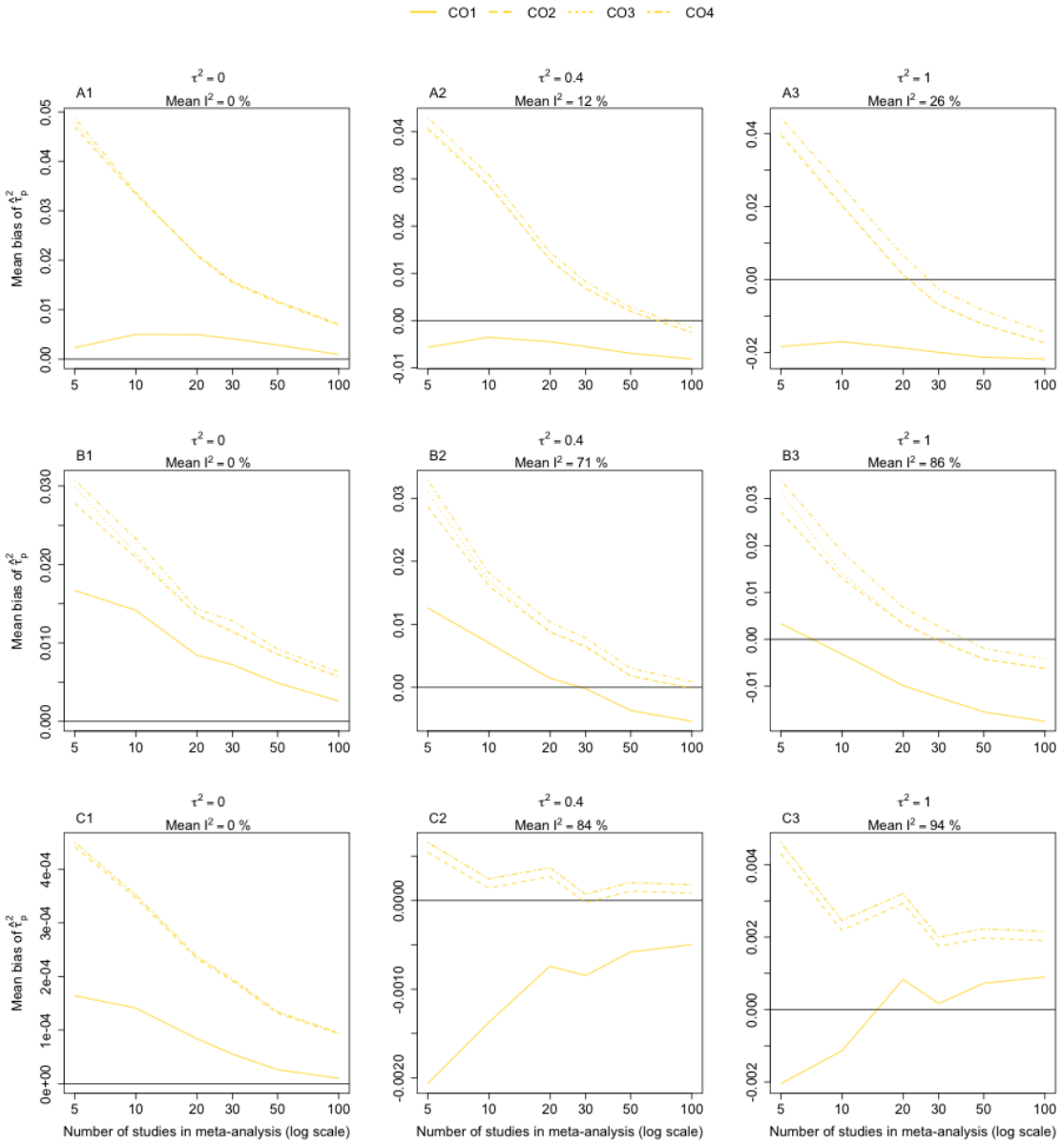


FIGURE E.69: Mean bias of τ_p^2 estimates in rare events scenario with $p_0 > p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

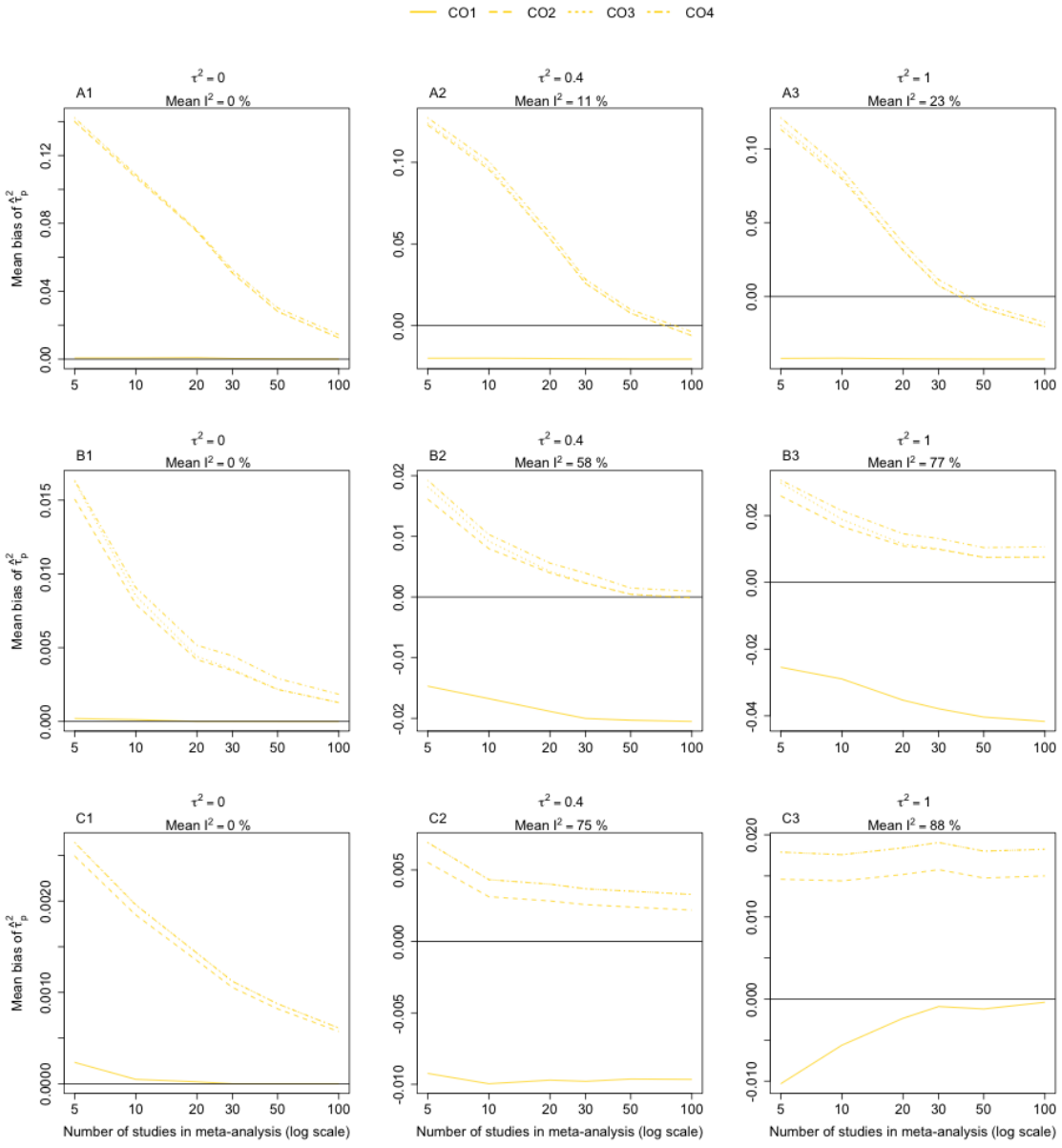


FIGURE E.70: Mean bias of τ_p^2 estimates in rare events scenario with $p_0 = p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

Common probability scenarios

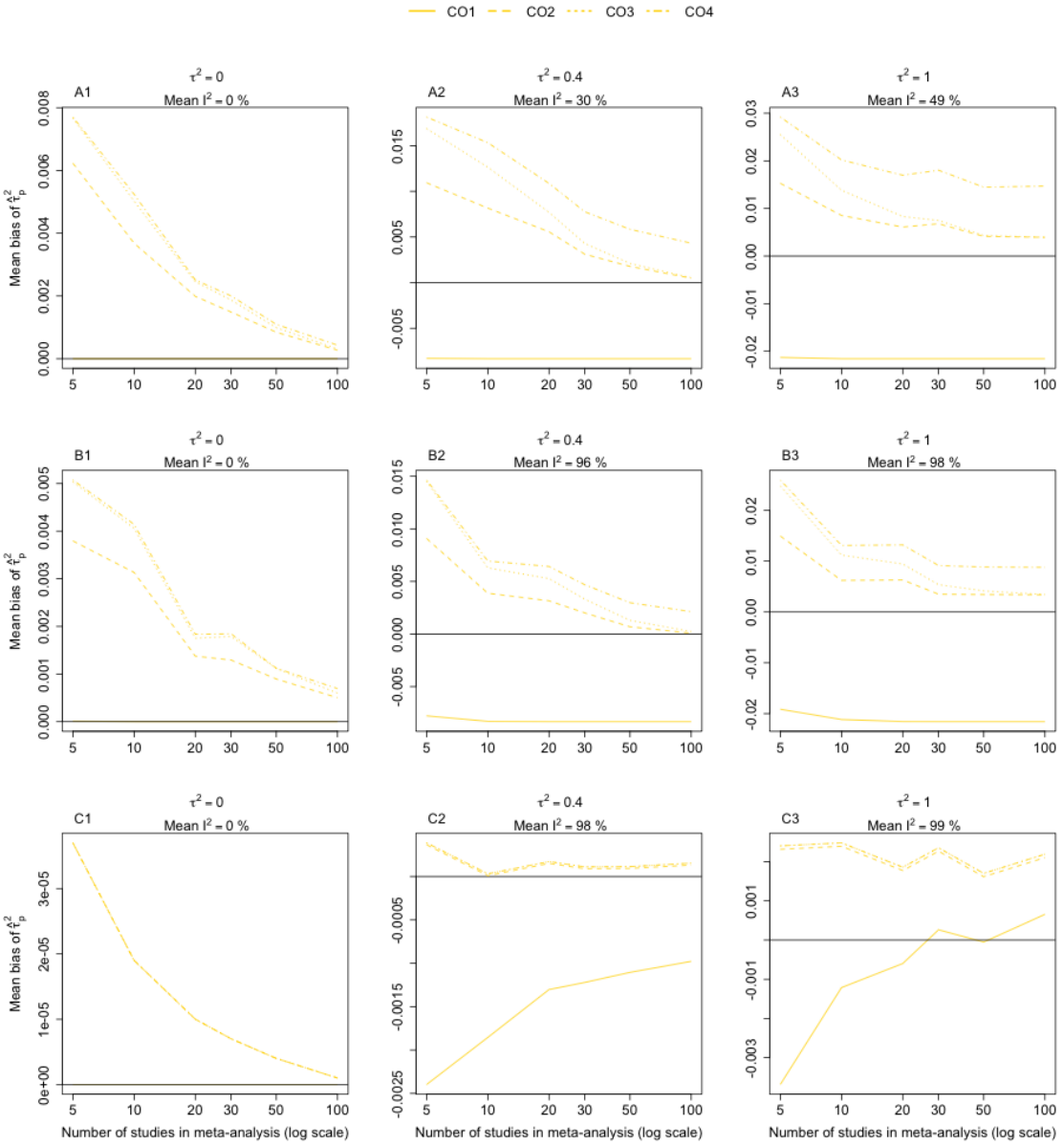


FIGURE E.71: Mean bias of τ_p^2 estimates in common probability scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

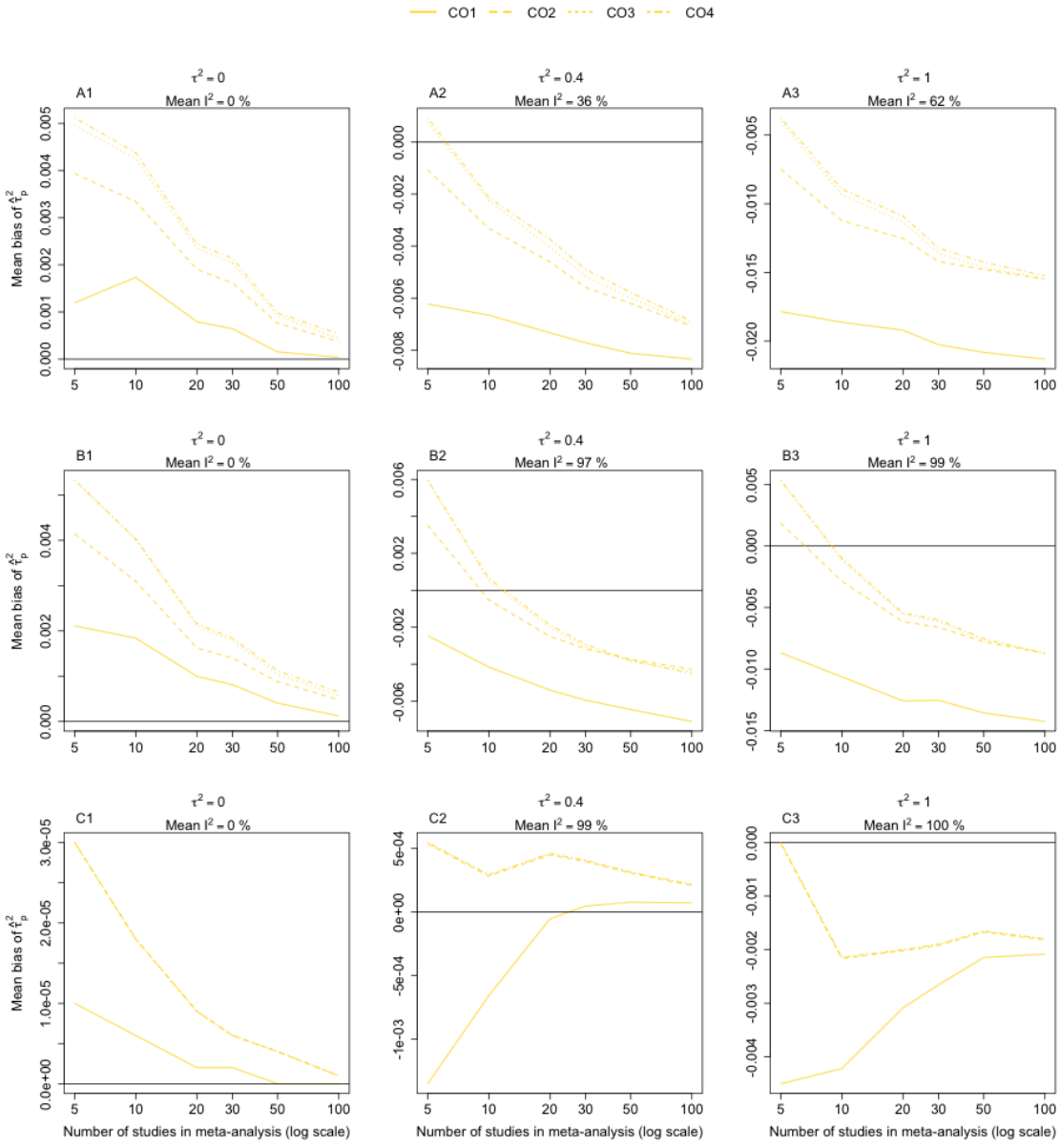


FIGURE E.72: Mean bias of τ_p^2 estimates in common probability scenario with $p_0 > p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

E.6.2 Mean squared error of τ_p^2

Very rare events scenarios

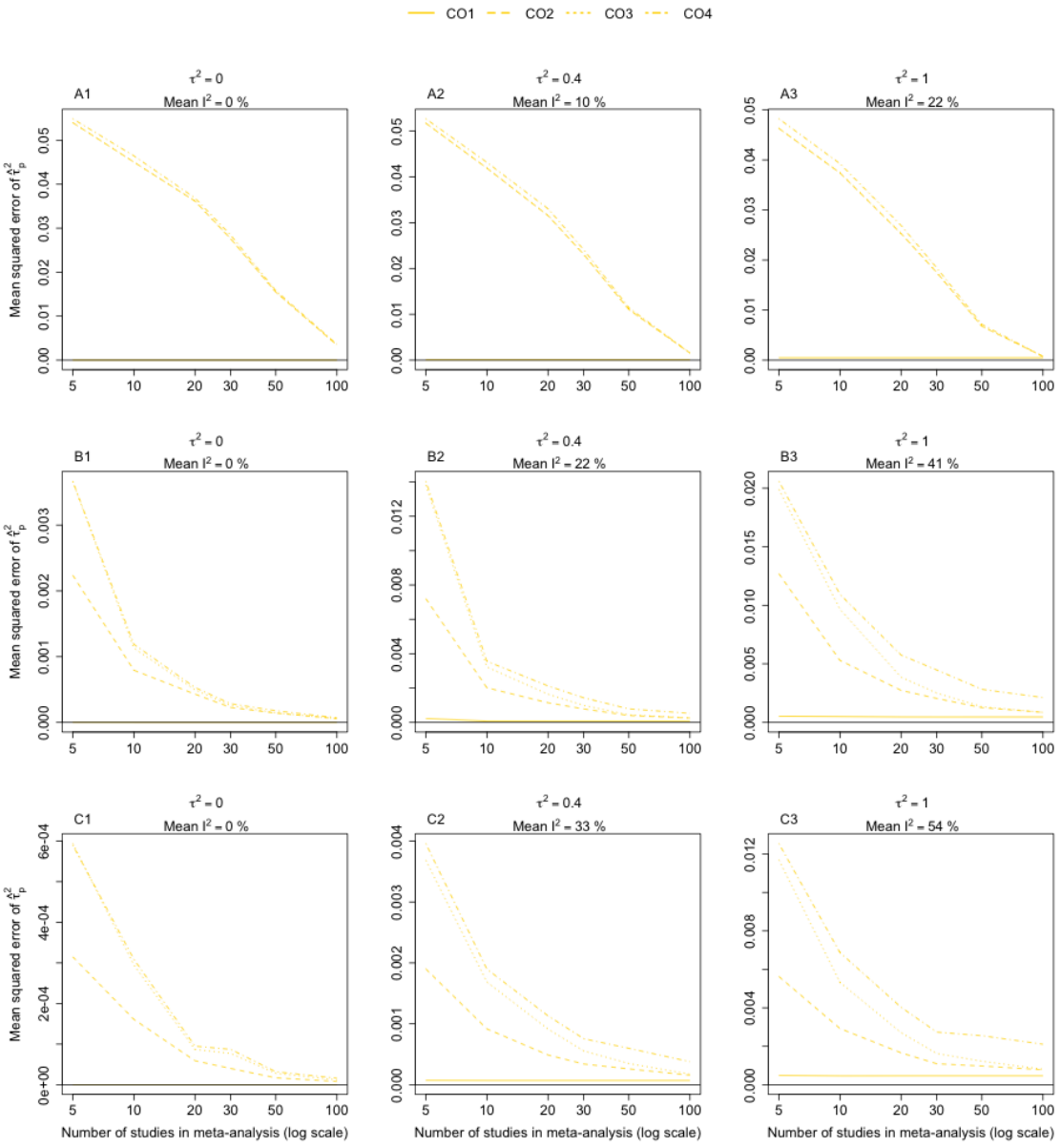


FIGURE E.73: Mean squared error of τ_p^2 estimates in very rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

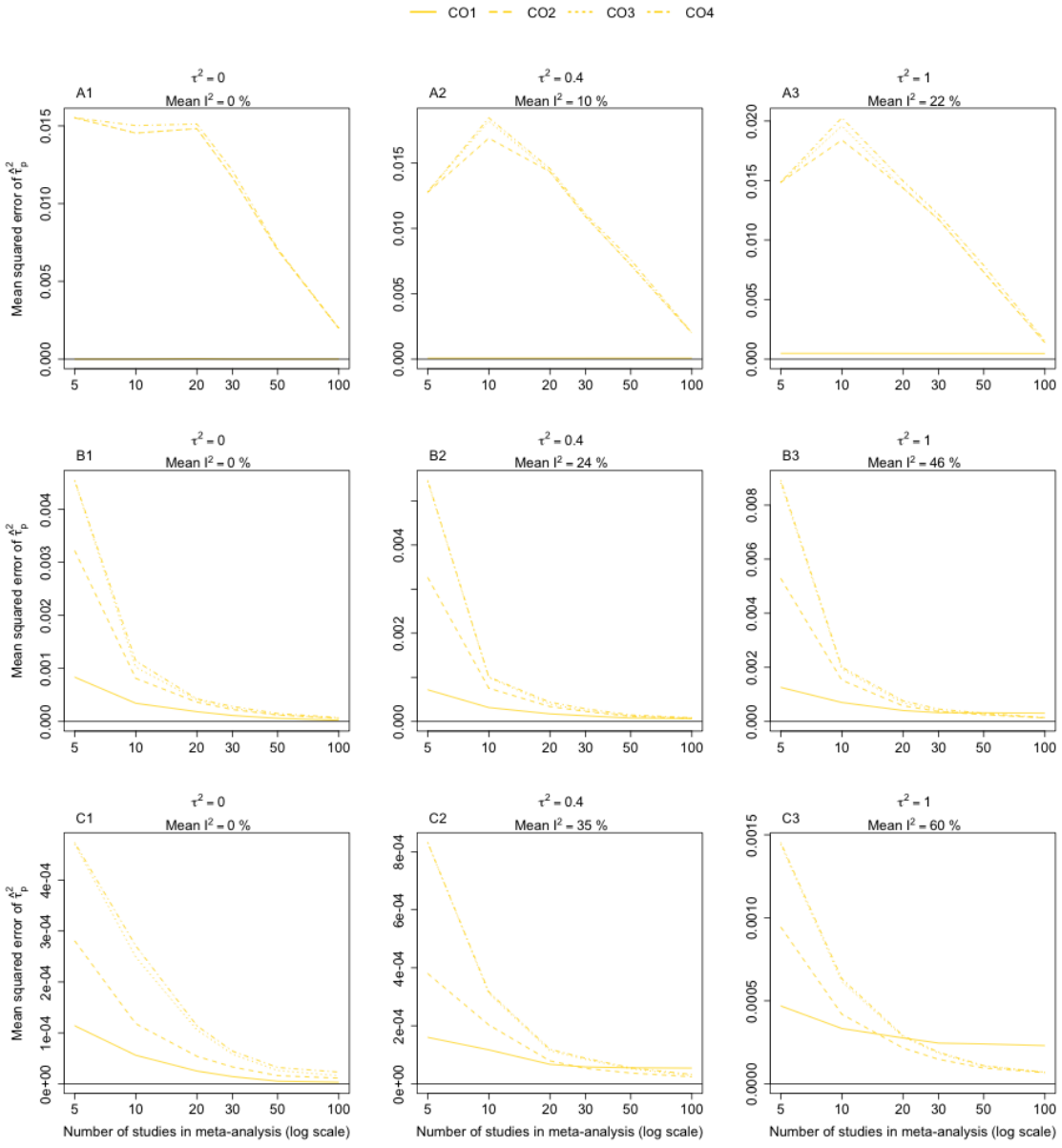


FIGURE E.74: Mean squared error of τ_p^2 estimates in very rare events scenario with $p_0 > p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

Rare events scenarios

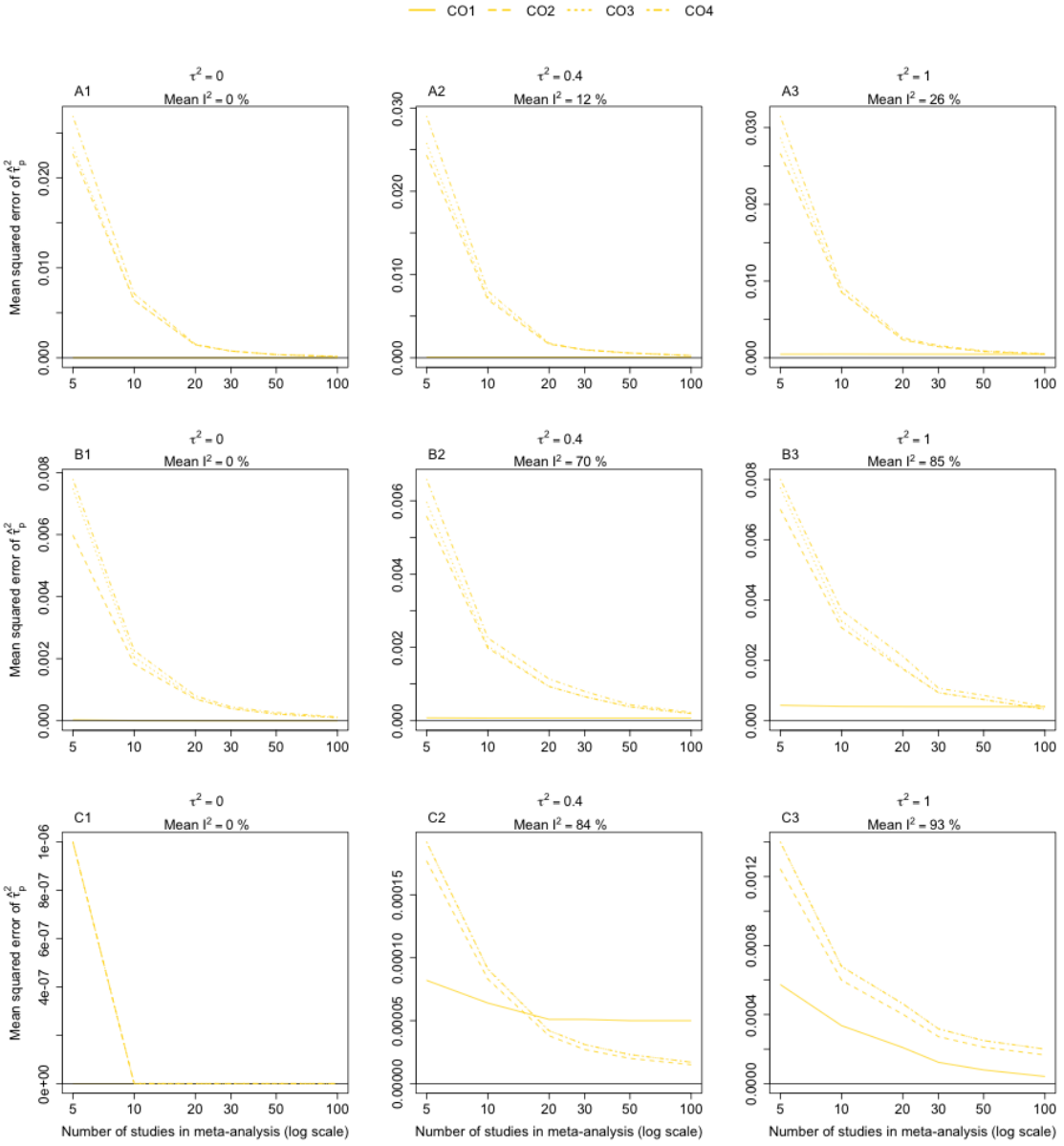


FIGURE E.75: Mean squared error of τ_p^2 estimates in rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

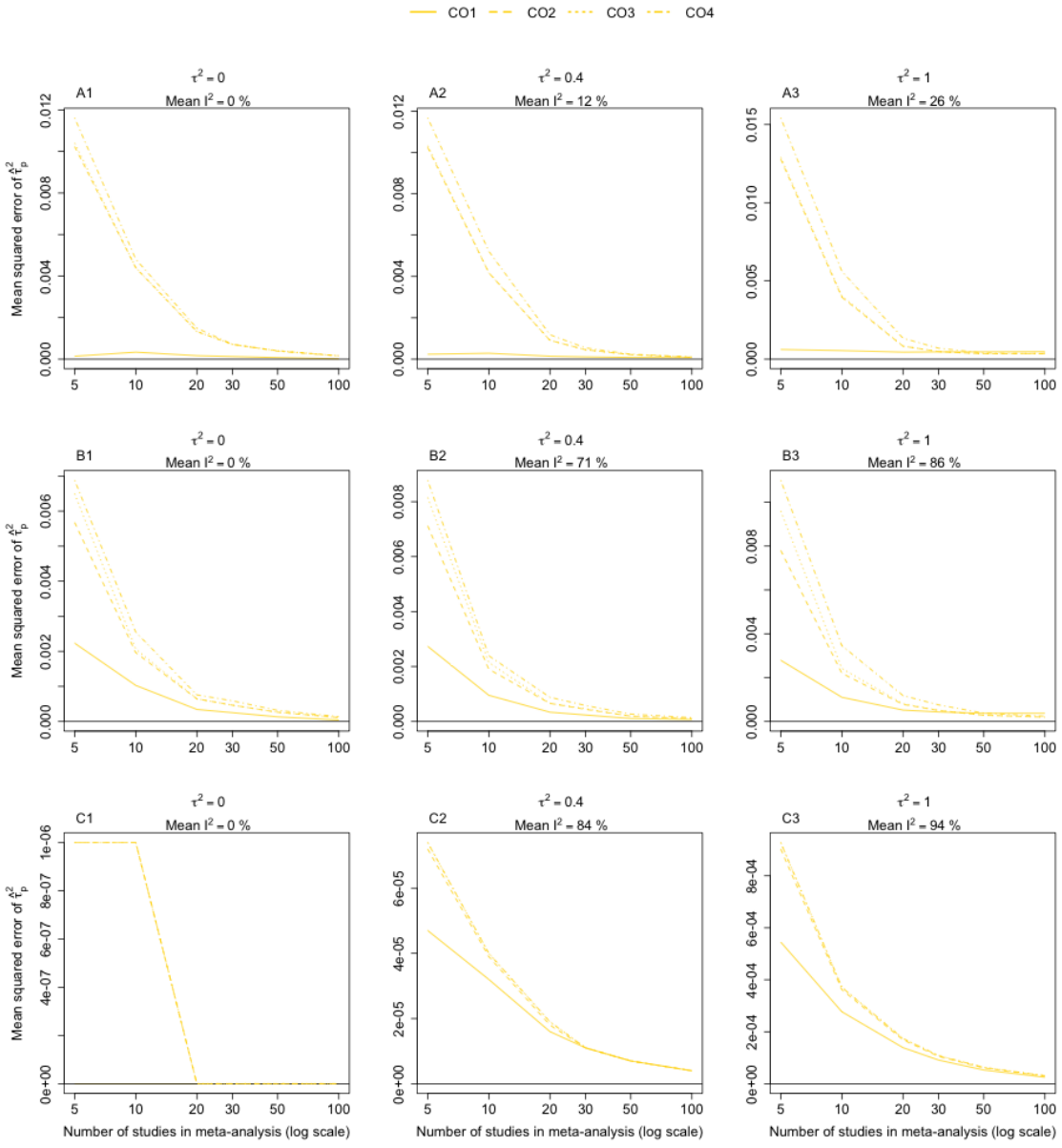


FIGURE E.76: Mean squared error of τ_p^2 estimates in rare events scenario with $p_0 > p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

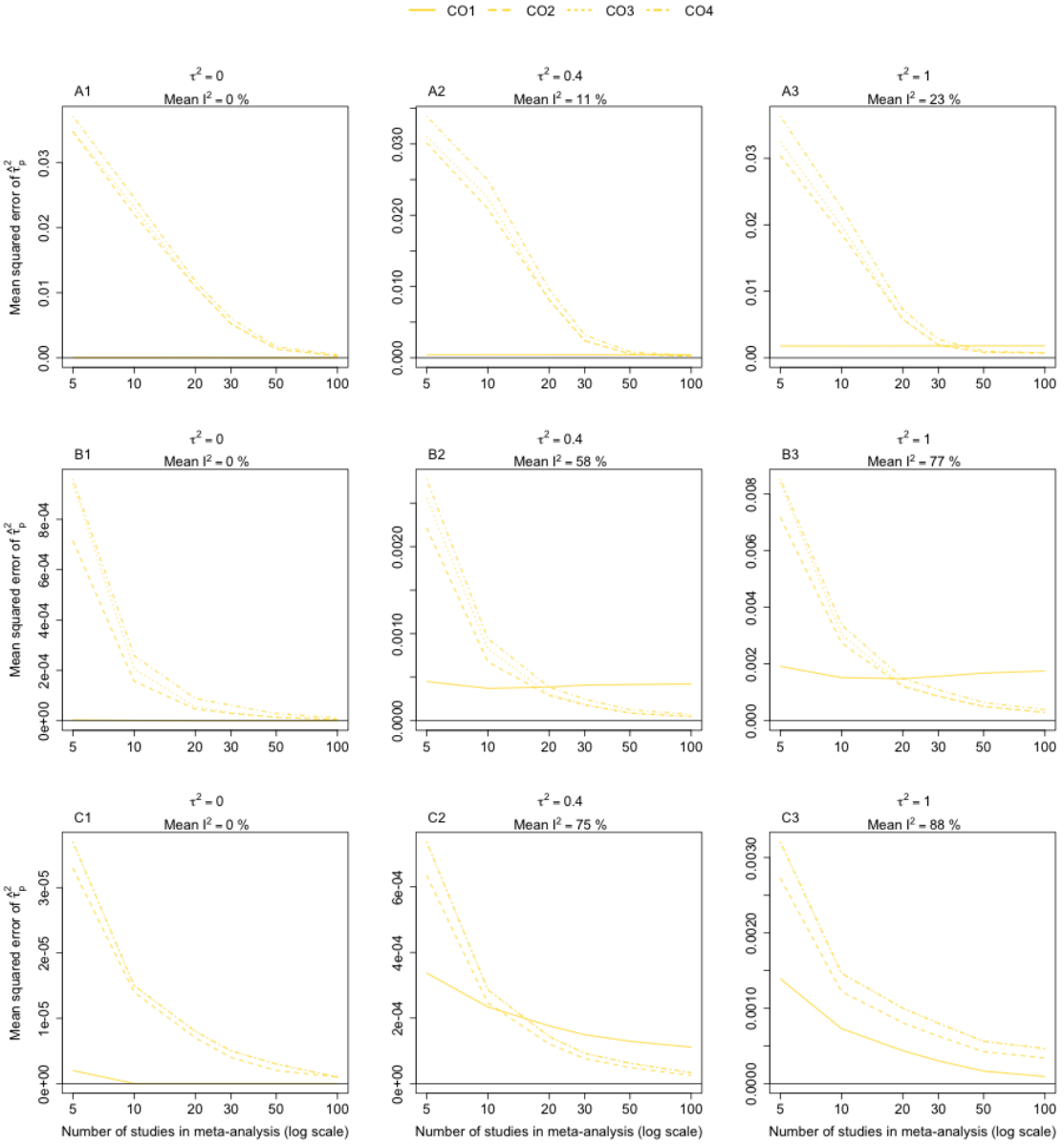


FIGURE E.77: Mean squared error of τ_p^2 estimates in rare events scenario with $p_0 = p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

Common probability scenarios

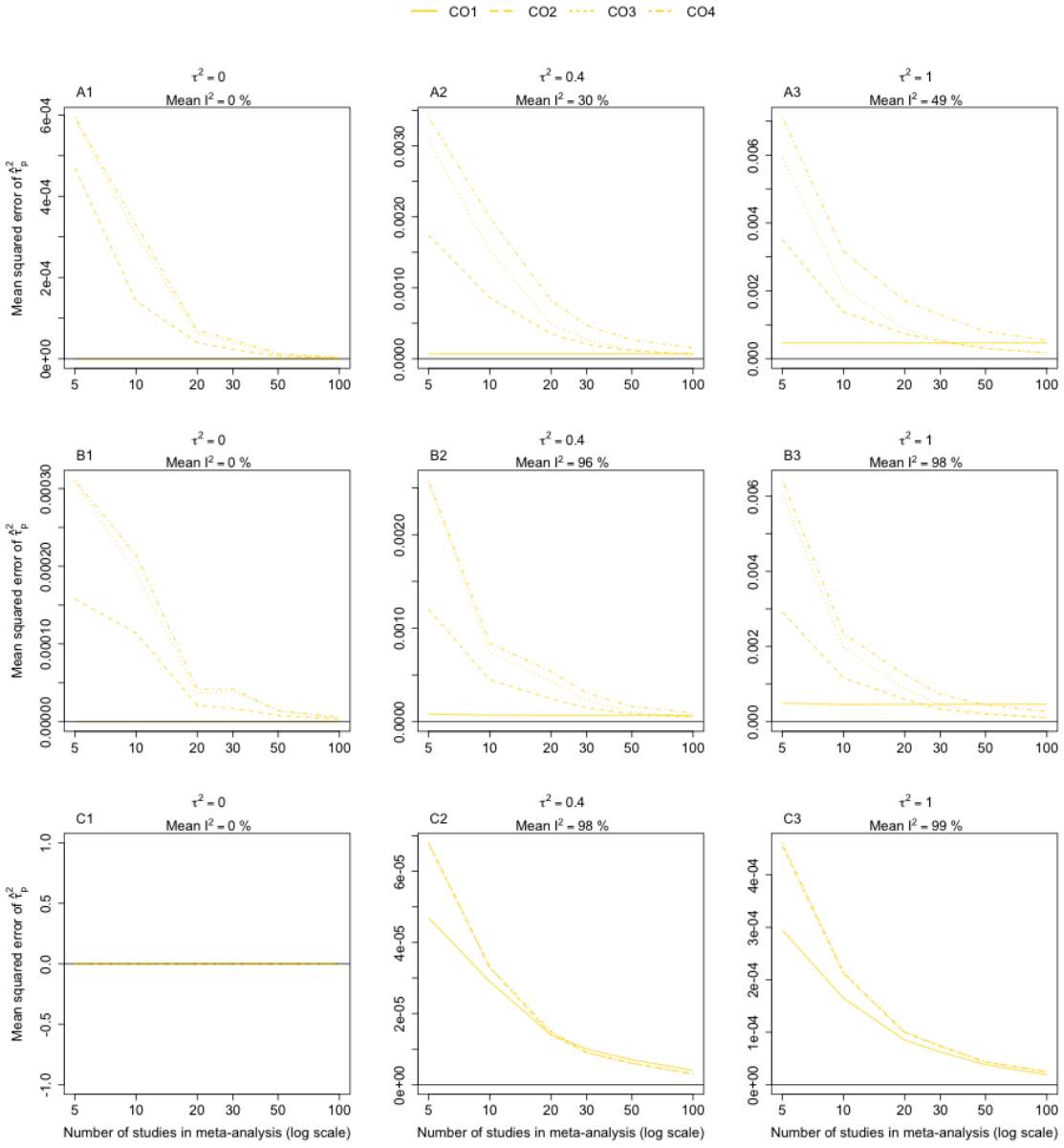


FIGURE E.78: Mean squared error of τ_p^2 estimates in common probability scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

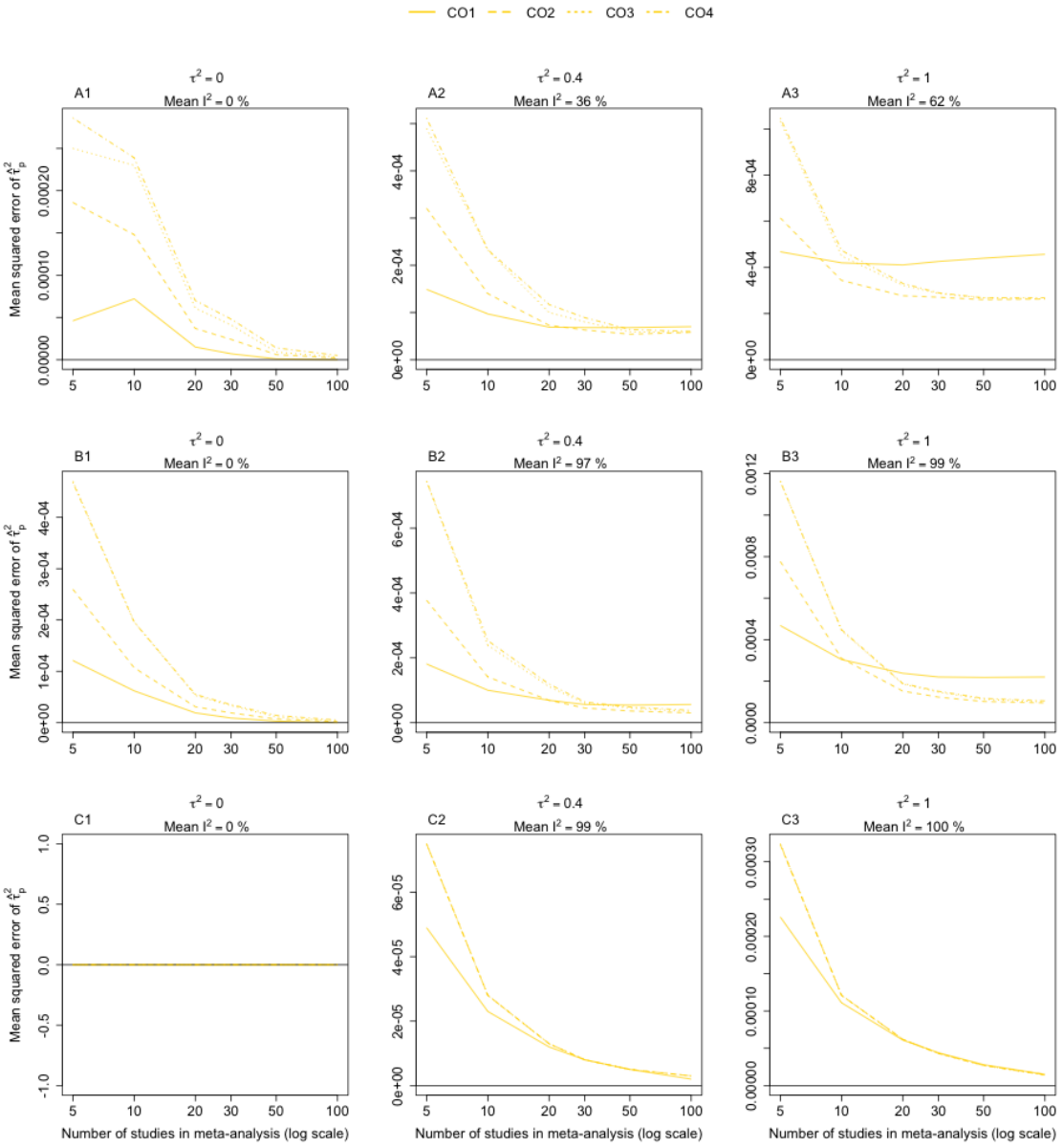


FIGURE E.79: Mean squared error of τ_p^2 estimates in common probability scenario with $p_0 > p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

E.7 Bias of θ

E.7.1 Alternate values of heterogeneity variance

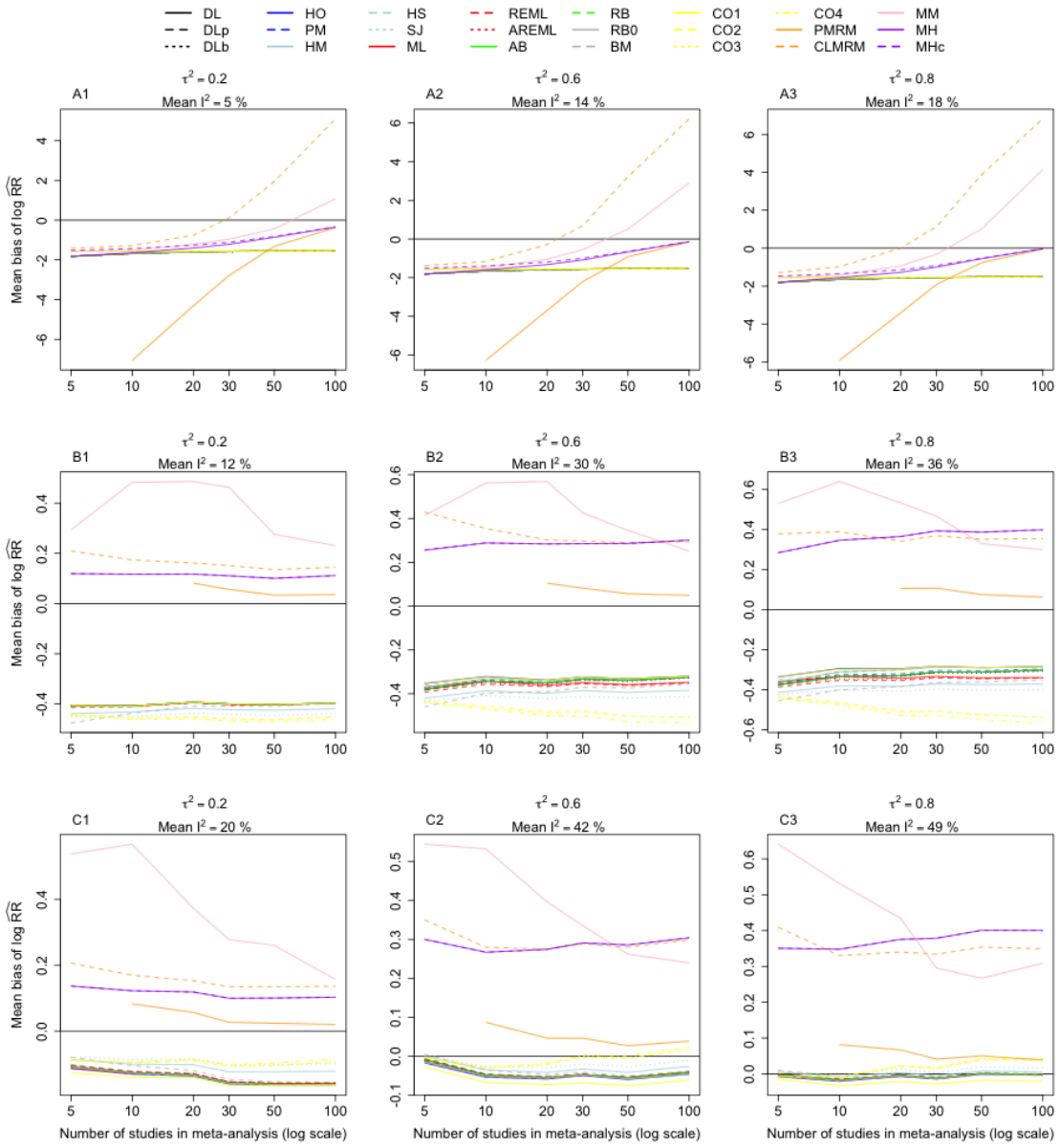


FIGURE E.80: Mean bias of log-risk ratio estimates in very rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

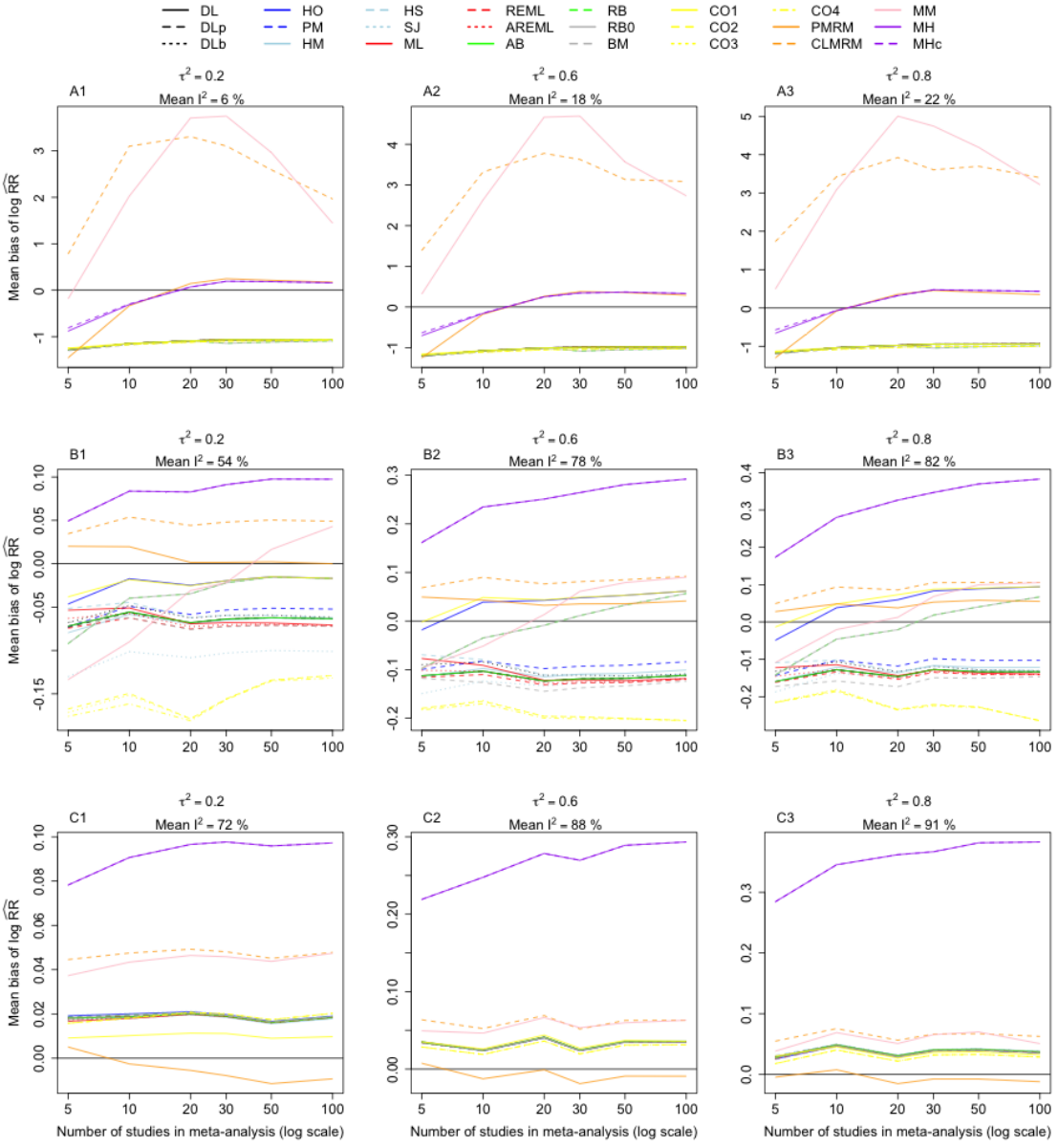


FIGURE E.81: Mean bias of log-risk ratio estimates in rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

E.7.2 Alternate study sample sizes

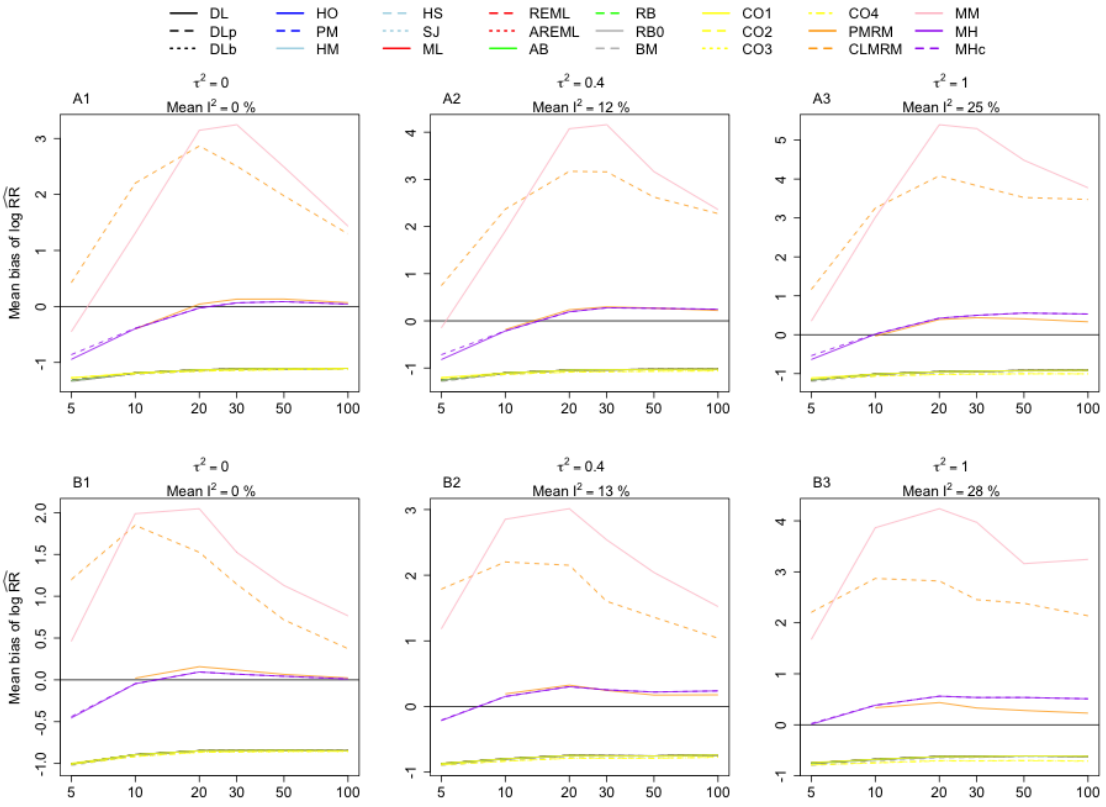


FIGURE E.82: Mean bias of log-risk ratio estimates in very rare events scenario with $p_0 < p_1$; sample sizes are small-to-medium (A1-A3) and medium (B1-B3).

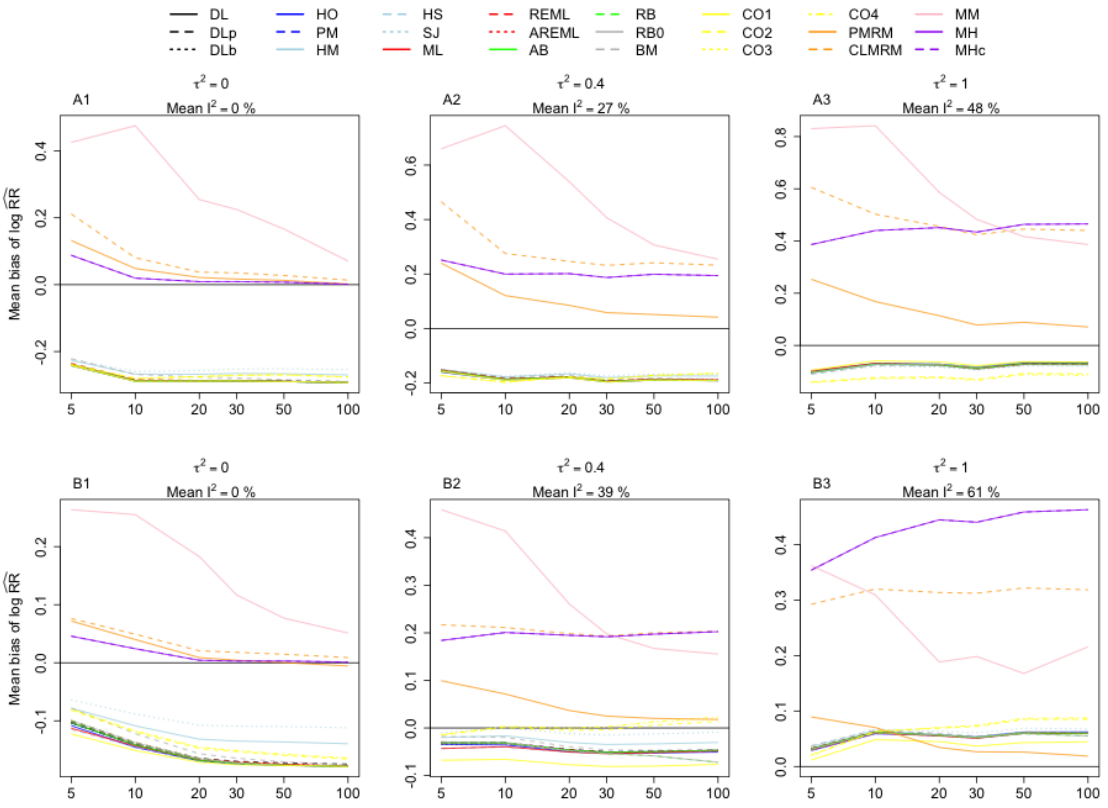


FIGURE E.83: Mean bias of log-risk ratio estimates in rare events scenario with $p_0 < p_1$; sample sizes are small-to-medium (A1-A3) and medium (B1-B3). MM is omitted from all.

E.7.3 Alternate values of σ_α^2

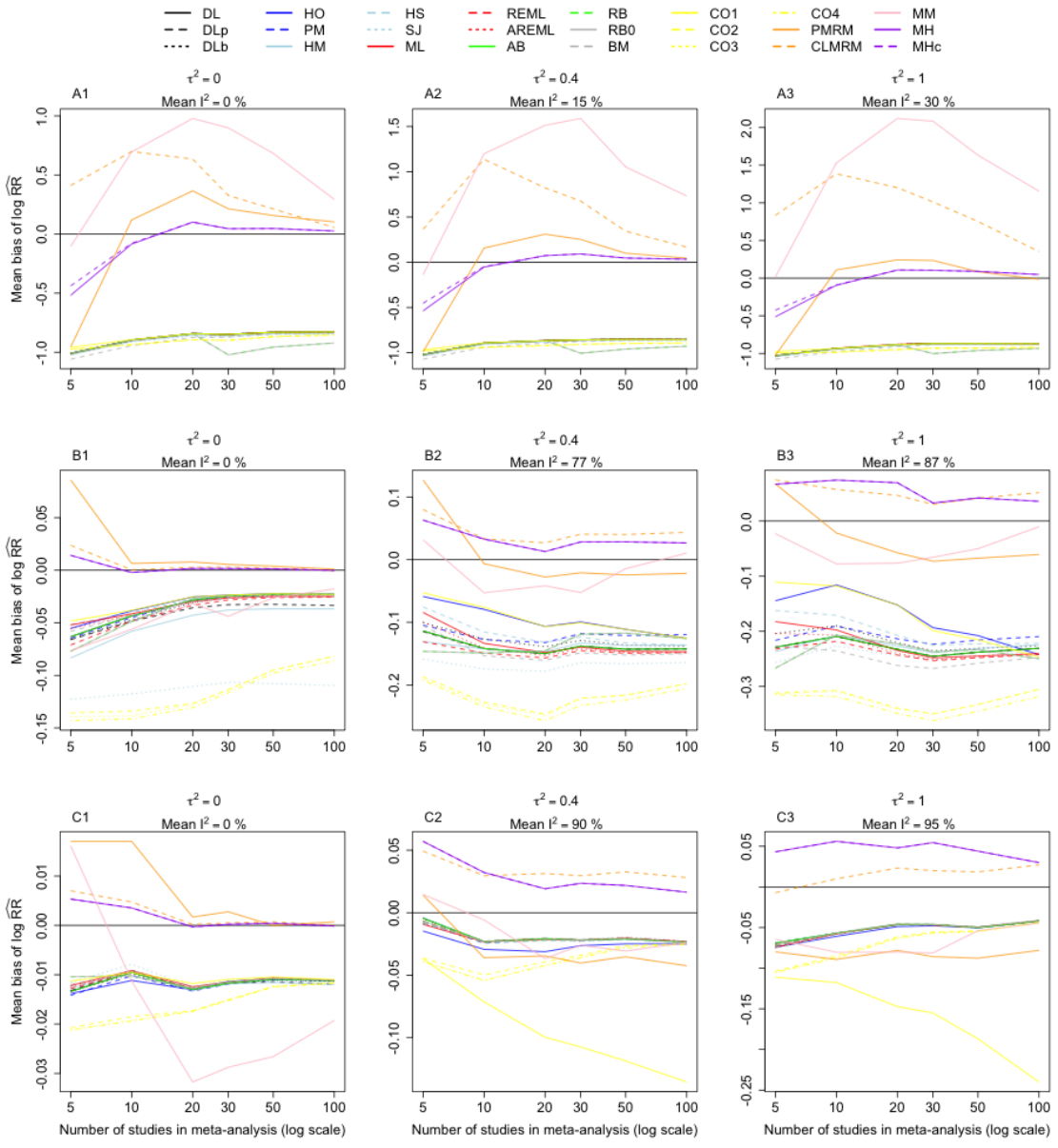


FIGURE E.84: Mean bias of log-risk ratio estimates in very rare events scenario with $p_0 < p_1$ and $\sigma_\alpha^2 = 3$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

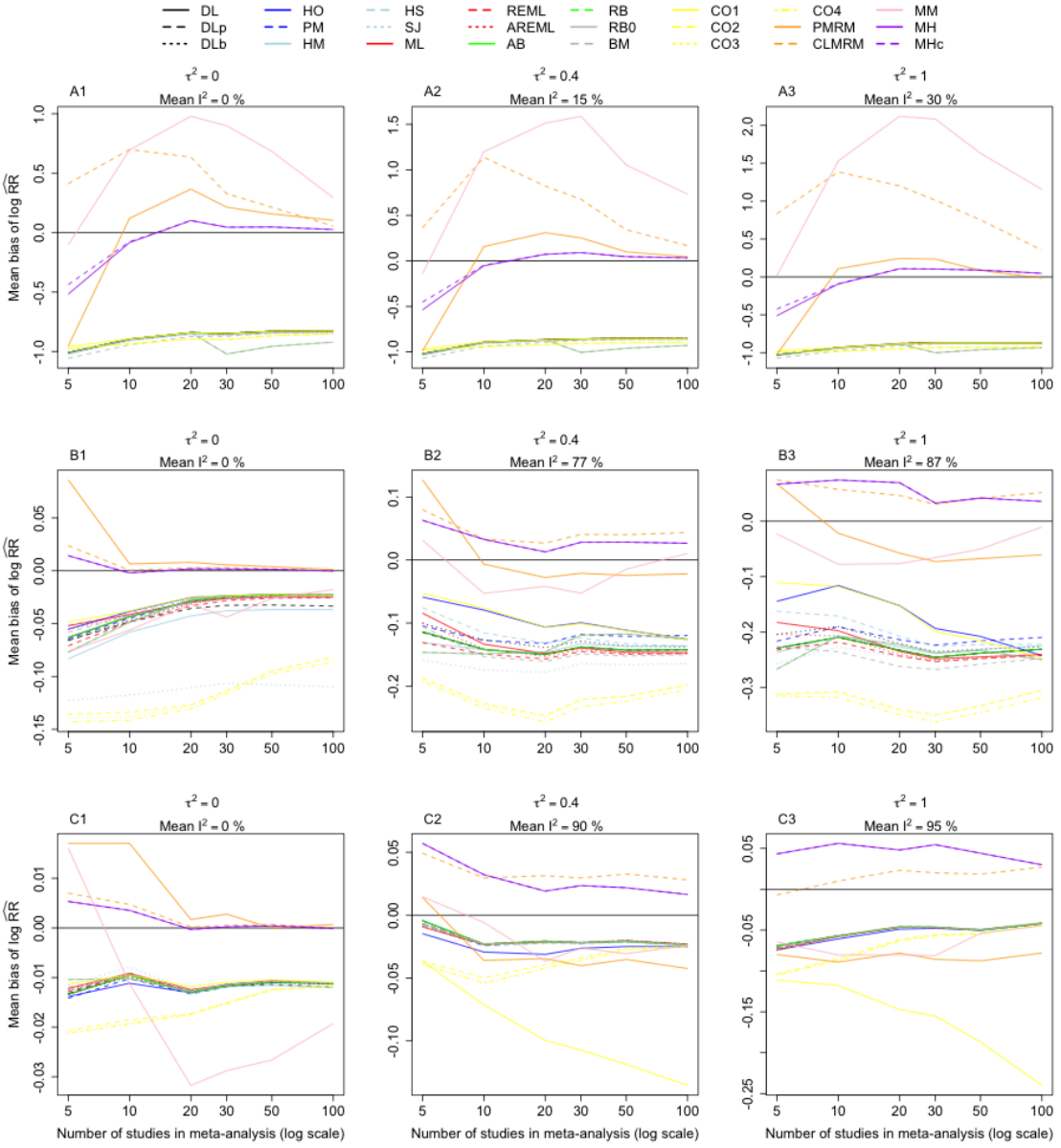


FIGURE E.85: Mean bias of log-risk ratio estimates in rare events scenario with $p_0 < p_1$ and $\sigma_\alpha^2 = 3$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

E.7.4 Alternate probability scenarios

Alternate rare events scenarios

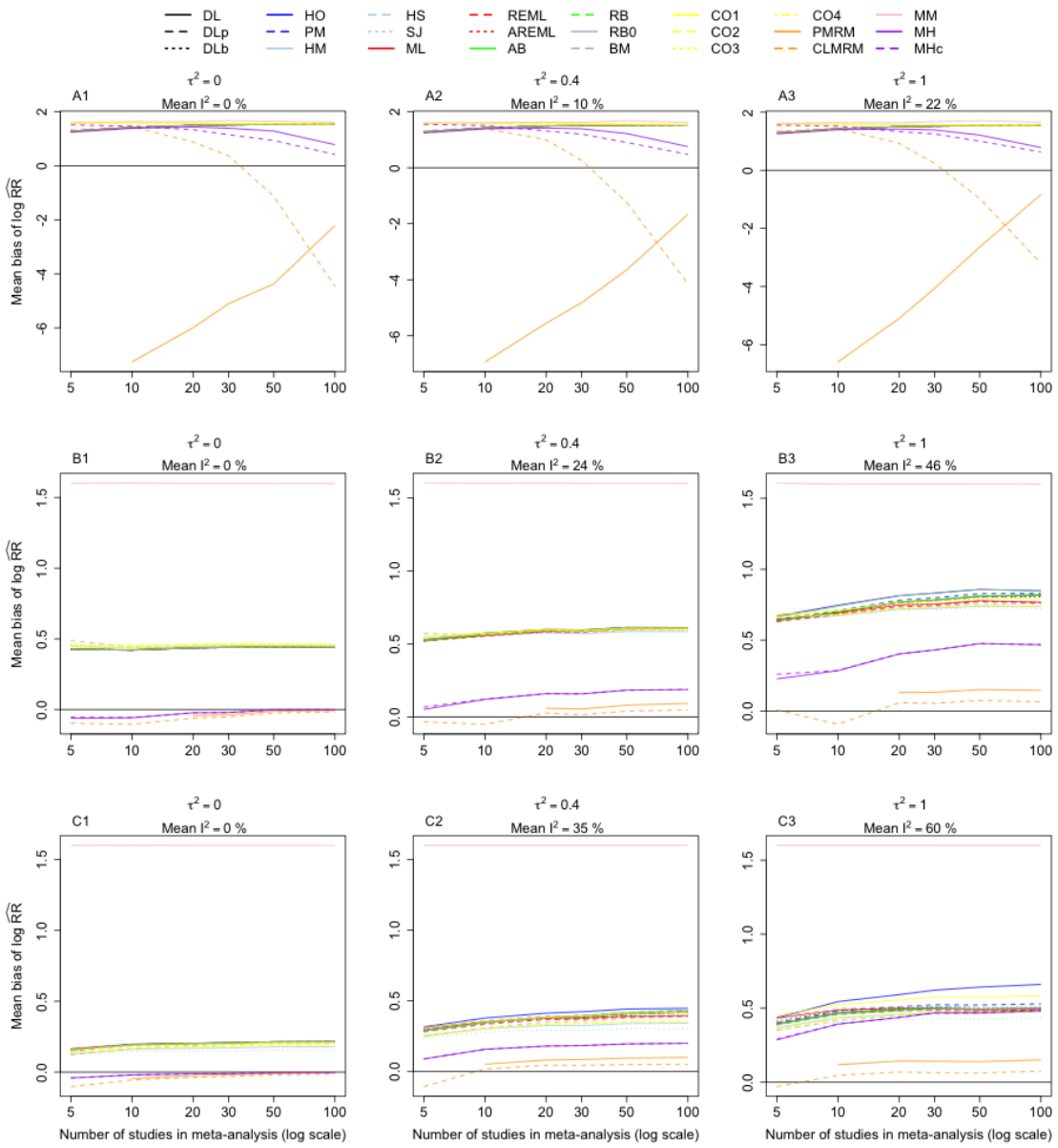


FIGURE E.86: Mean bias of log-risk ratio estimates in very rare events scenario with $p_0 > p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

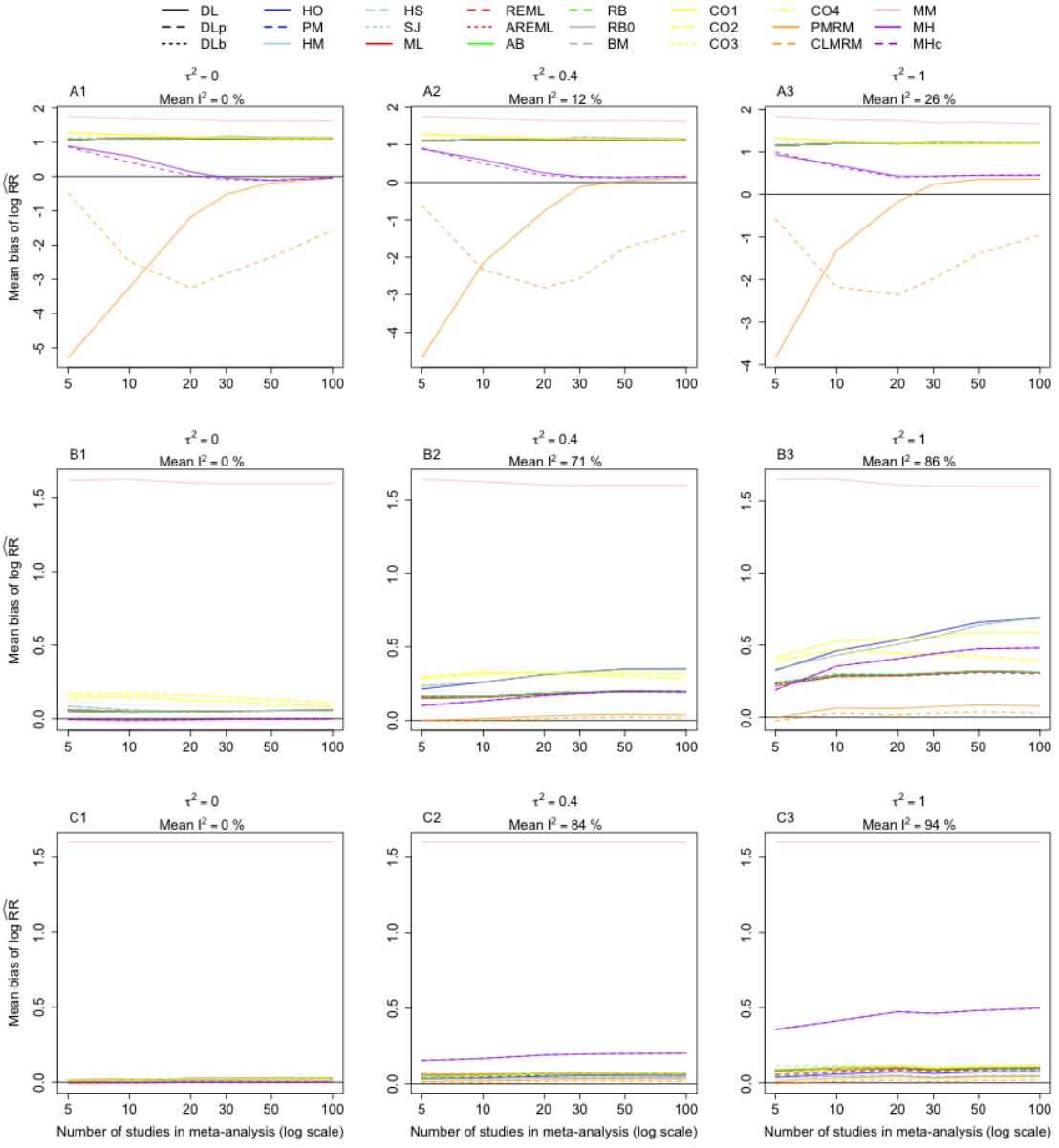


FIGURE E.87: Mean bias of log-risk ratio estimates in rare events scenario with $p_0 > p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

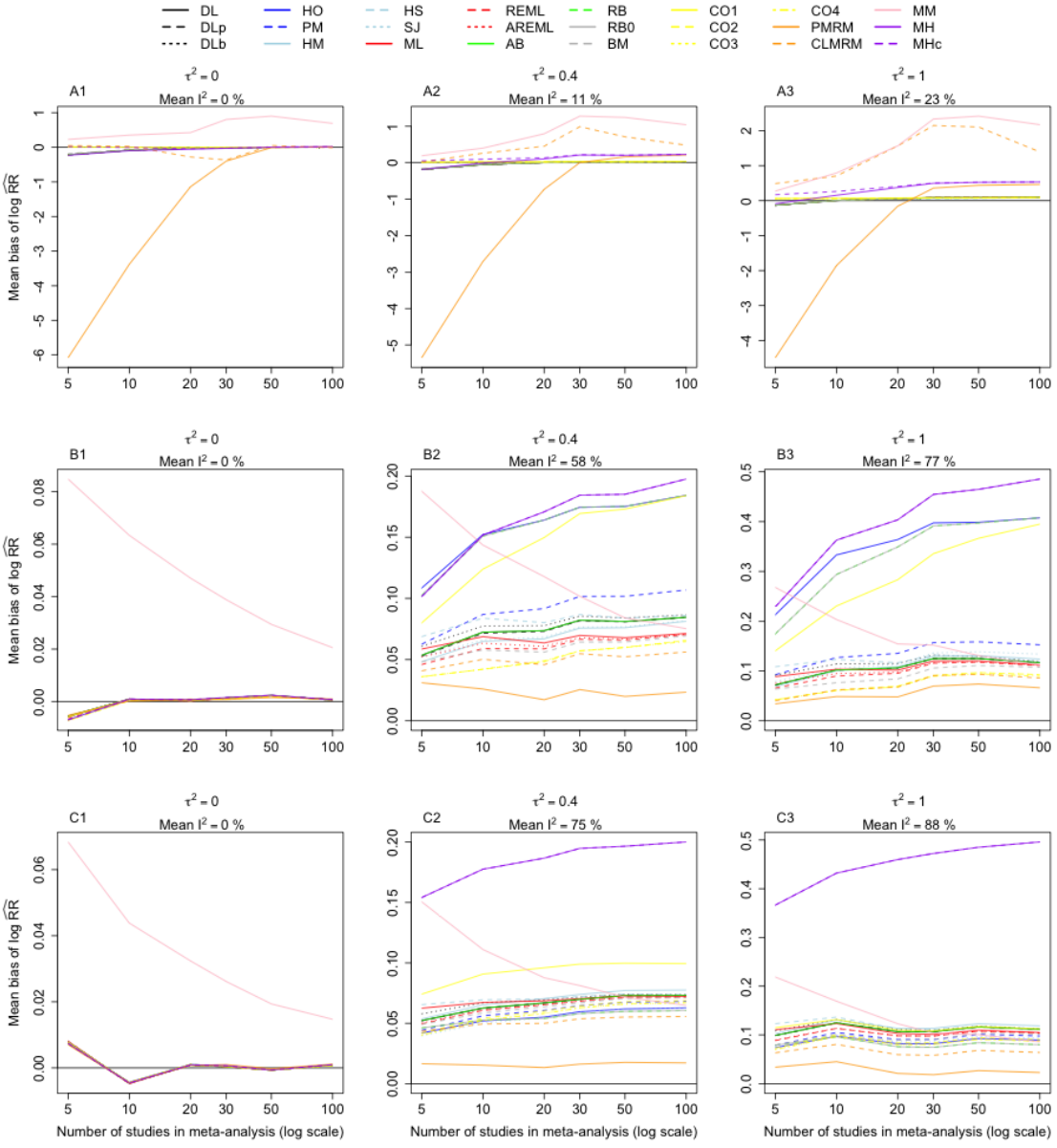


FIGURE E.88: Mean bias of log-risk ratio estimates in rare events scenario with $p_0 = p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

Common probability scenarios

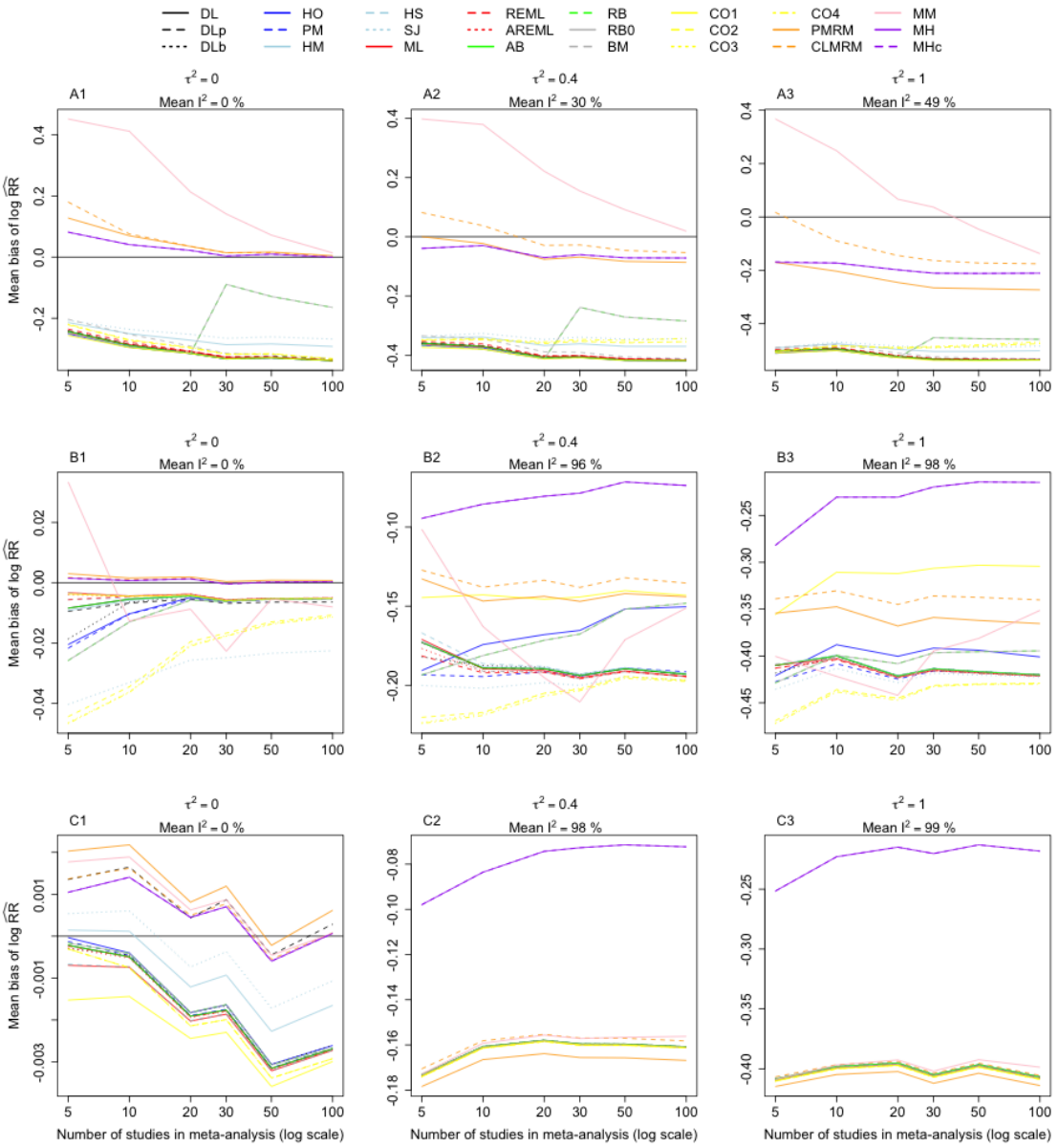


FIGURE E.89: Mean bias of log-risk ratio estimates in common probability scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

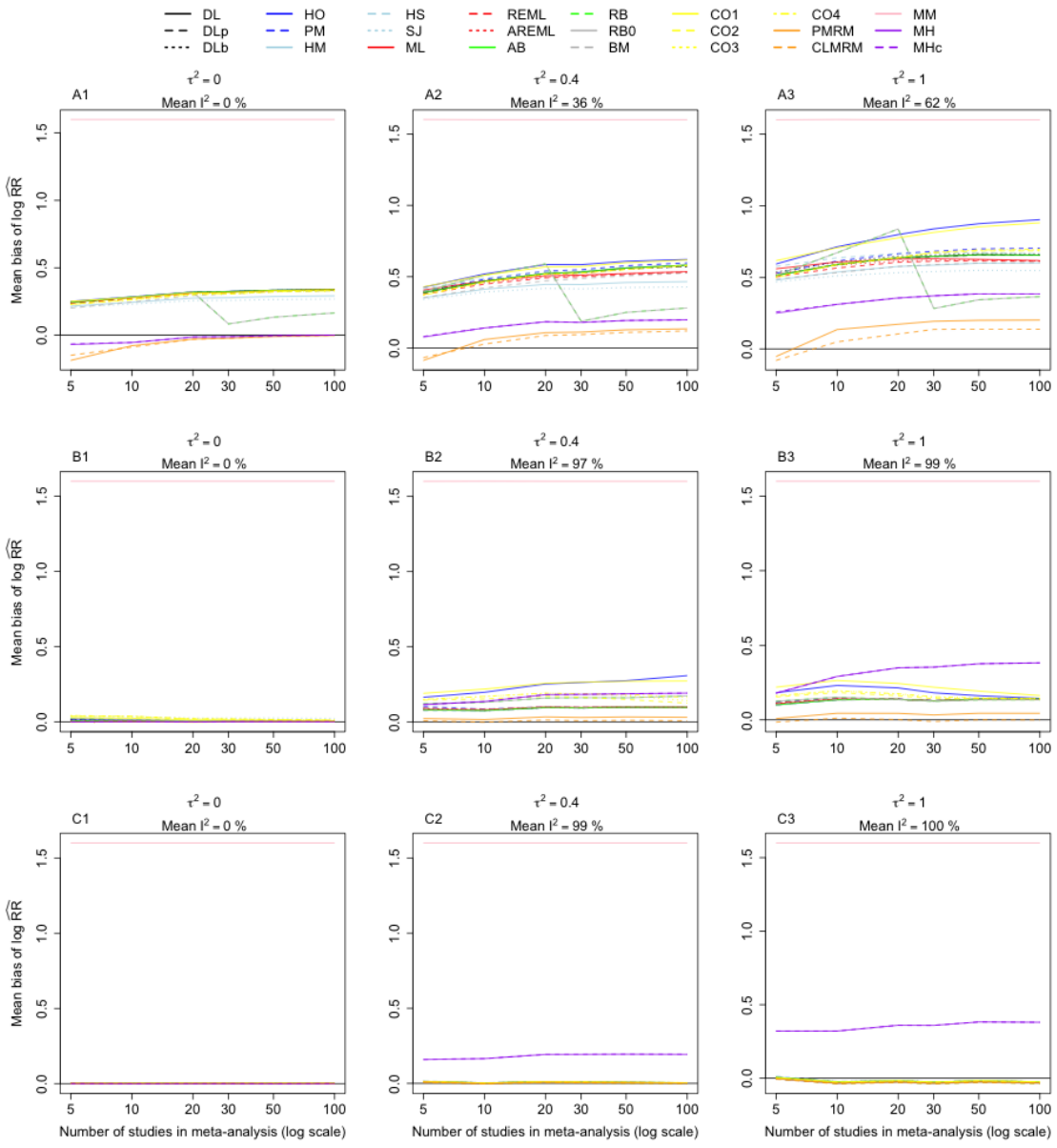


FIGURE E.90: Mean bias of log-risk ratio estimates in common probability scenario with $p_0 > p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

E.7.5 Alternate sampling in simulation study

Alternate event count sampling

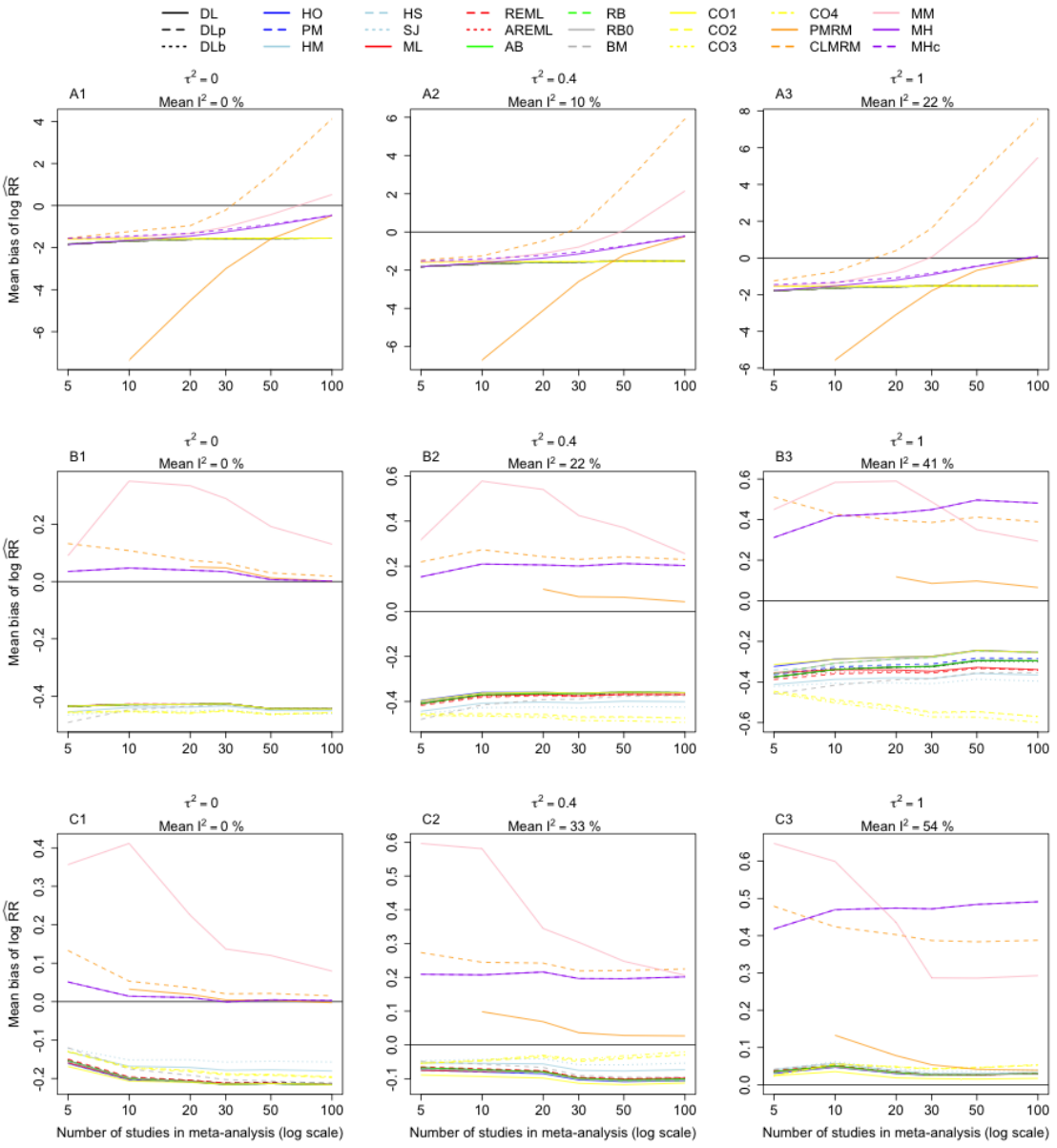


FIGURE E.91: Mean bias of log-risk ratio estimates in very rare events scenario with $p_0 < p_1$ and poisson event sampling; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

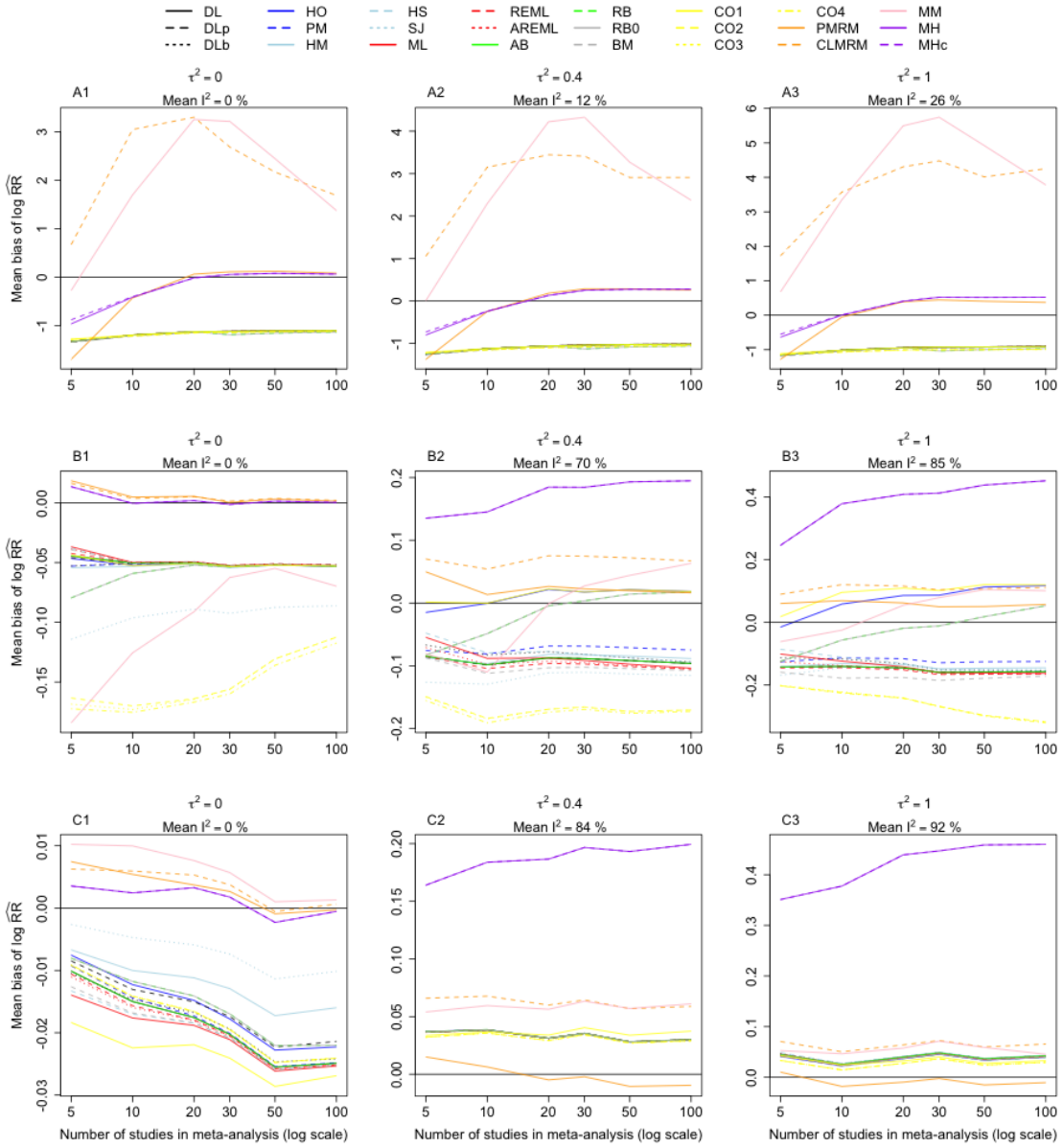


FIGURE E.92: Mean bias of log-risk ratio estimates in rare events scenario with $p_0 < p_1$ and poisson event sampling; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

Alternate sample size sampling

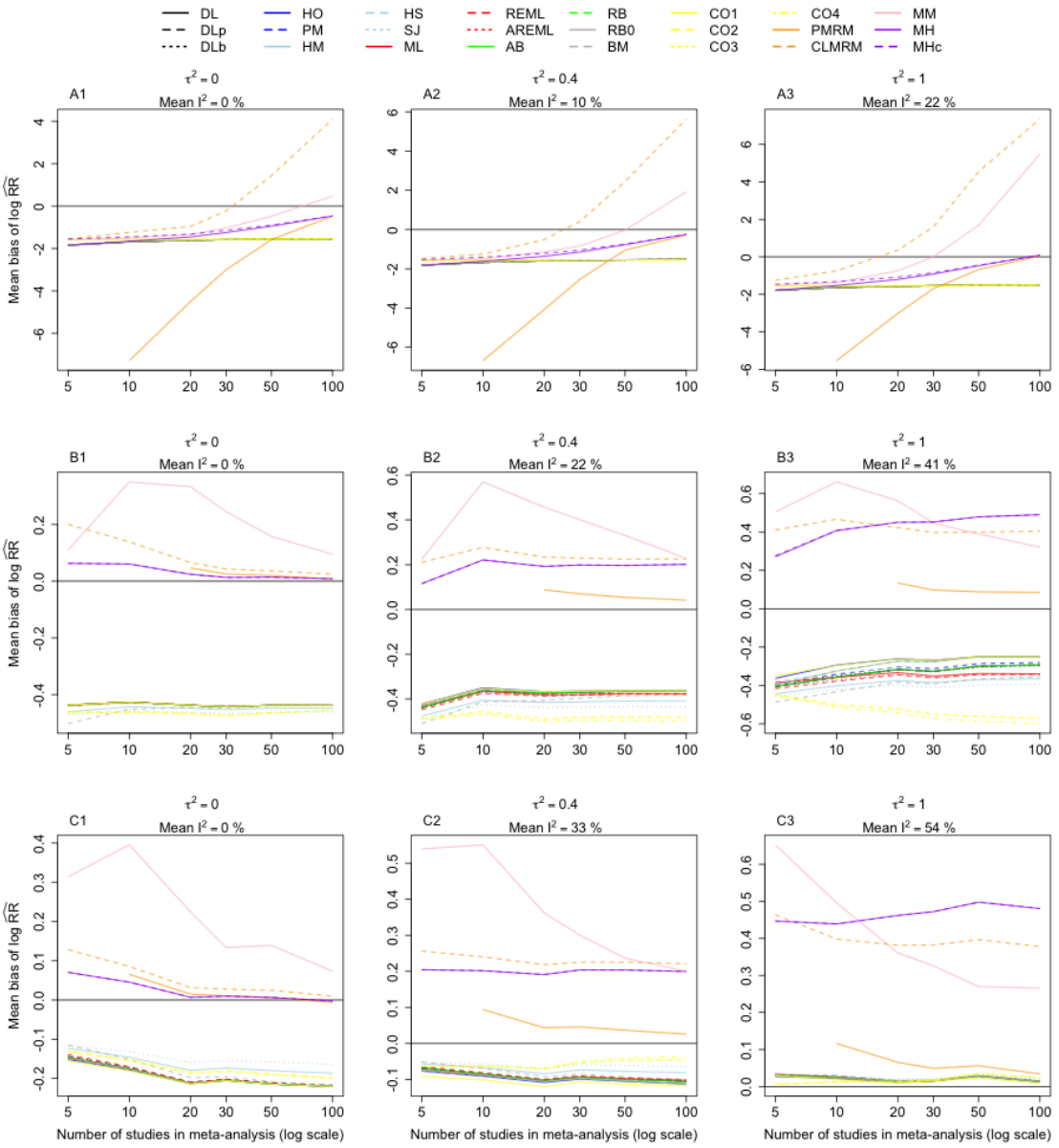


FIGURE E.93: Mean bias of log-risk ratio estimates in very rare events scenario with $p_0 < p_1$ and normal sample size sampling; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

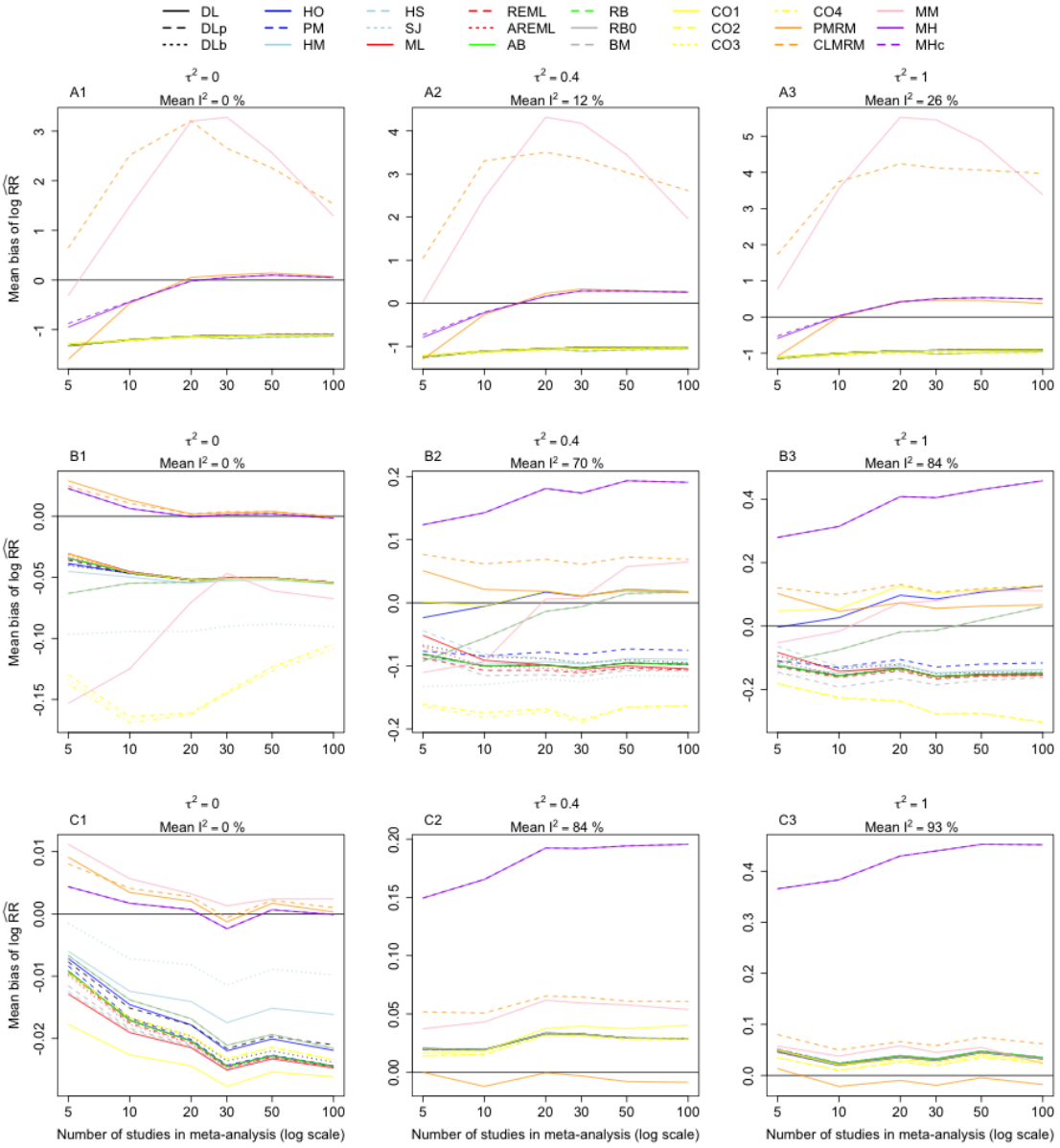


FIGURE E.94: Mean bias of log-risk ratio estimates in rare events scenario with $p_0 < p_1$ and normal sample size sampling; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

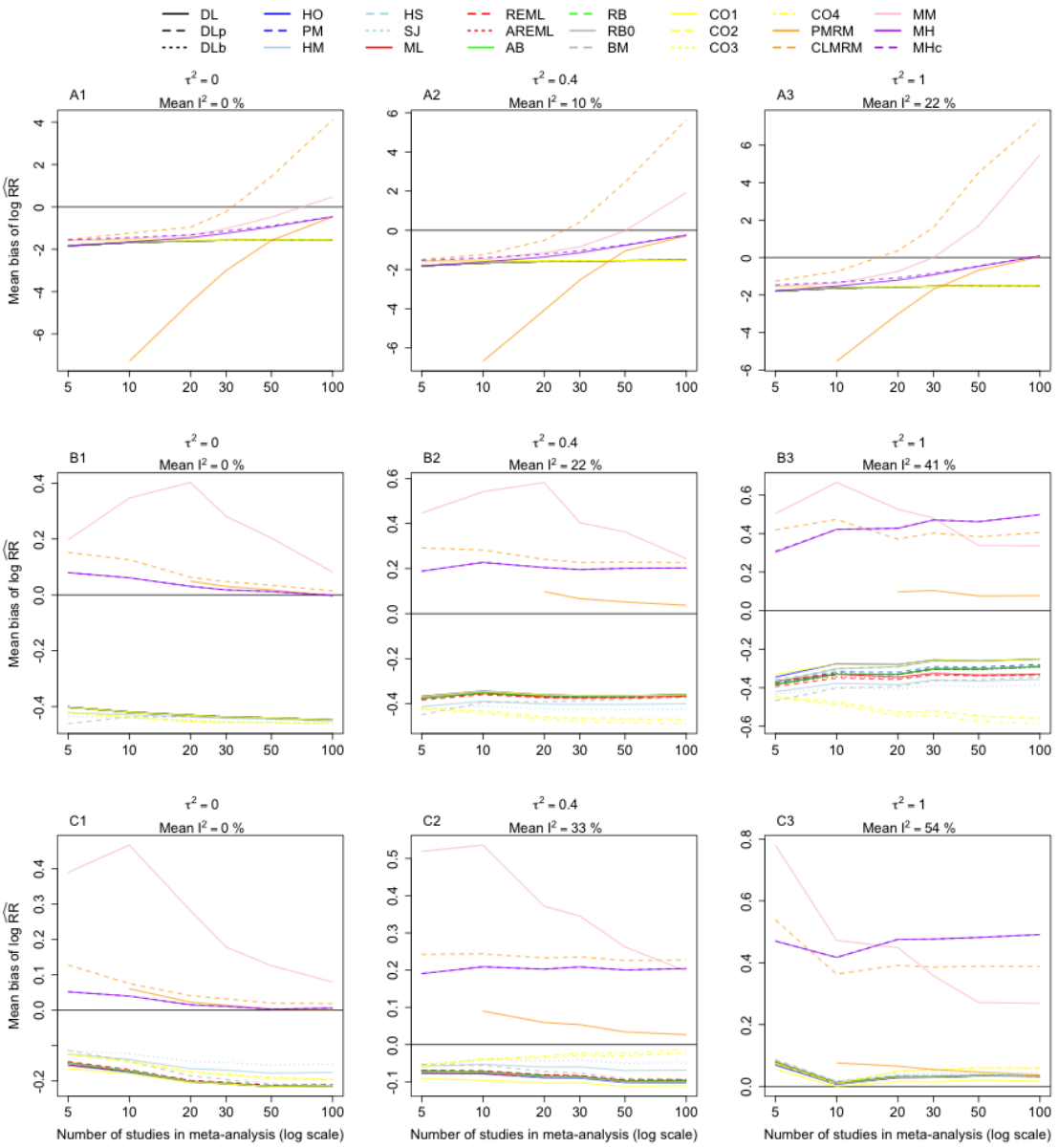


FIGURE E.95: Mean bias of log-risk ratio estimates in very rare events scenario with $p_0 < p_1$ and Chi-squared sample size sampling; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

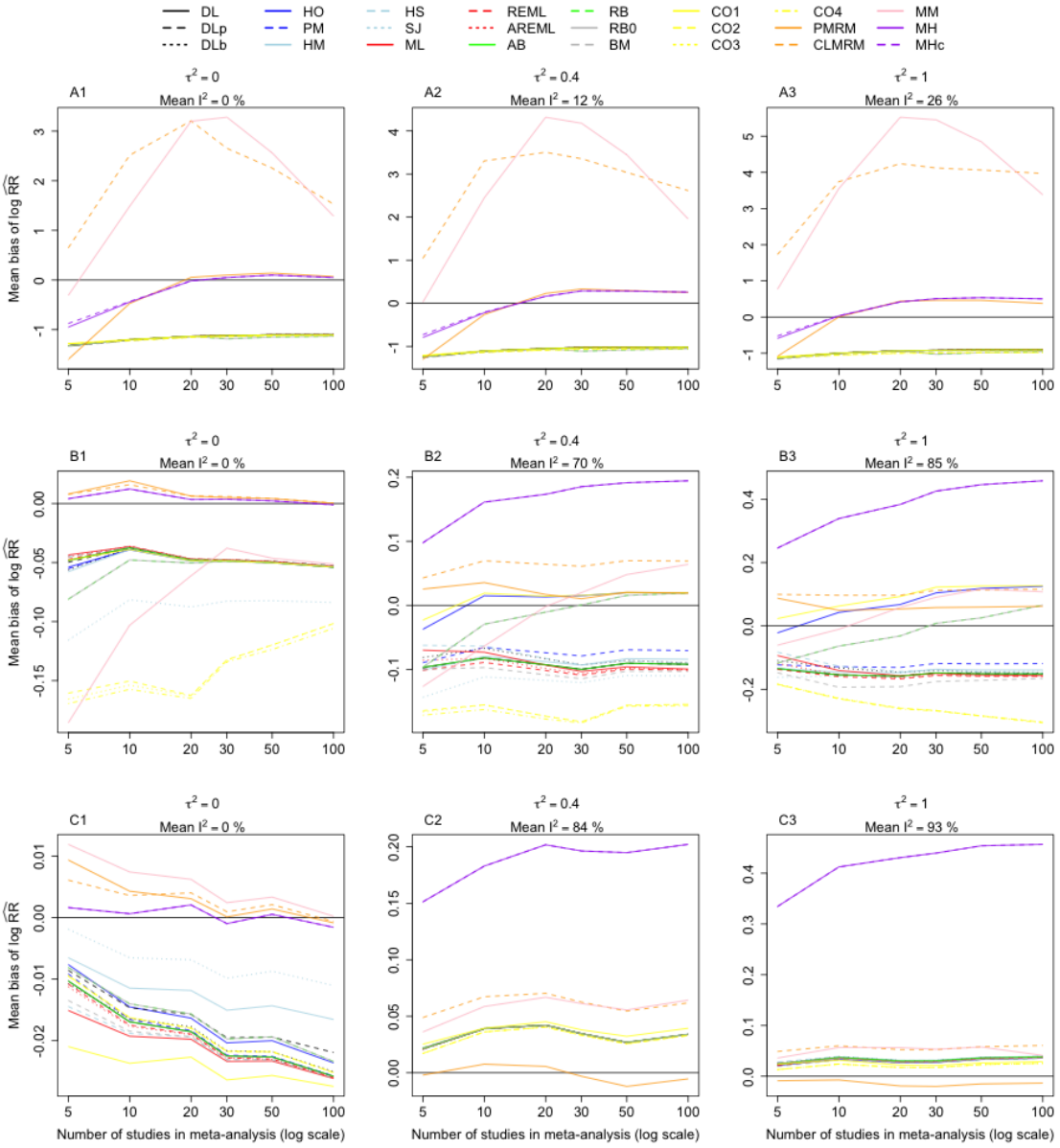


FIGURE E.96: Mean bias of log-risk ratio estimates in rare events scenario with $p_0 < p_1$ and Chi-squared sample size sampling; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

E.7.6 Alternate continuity corrections

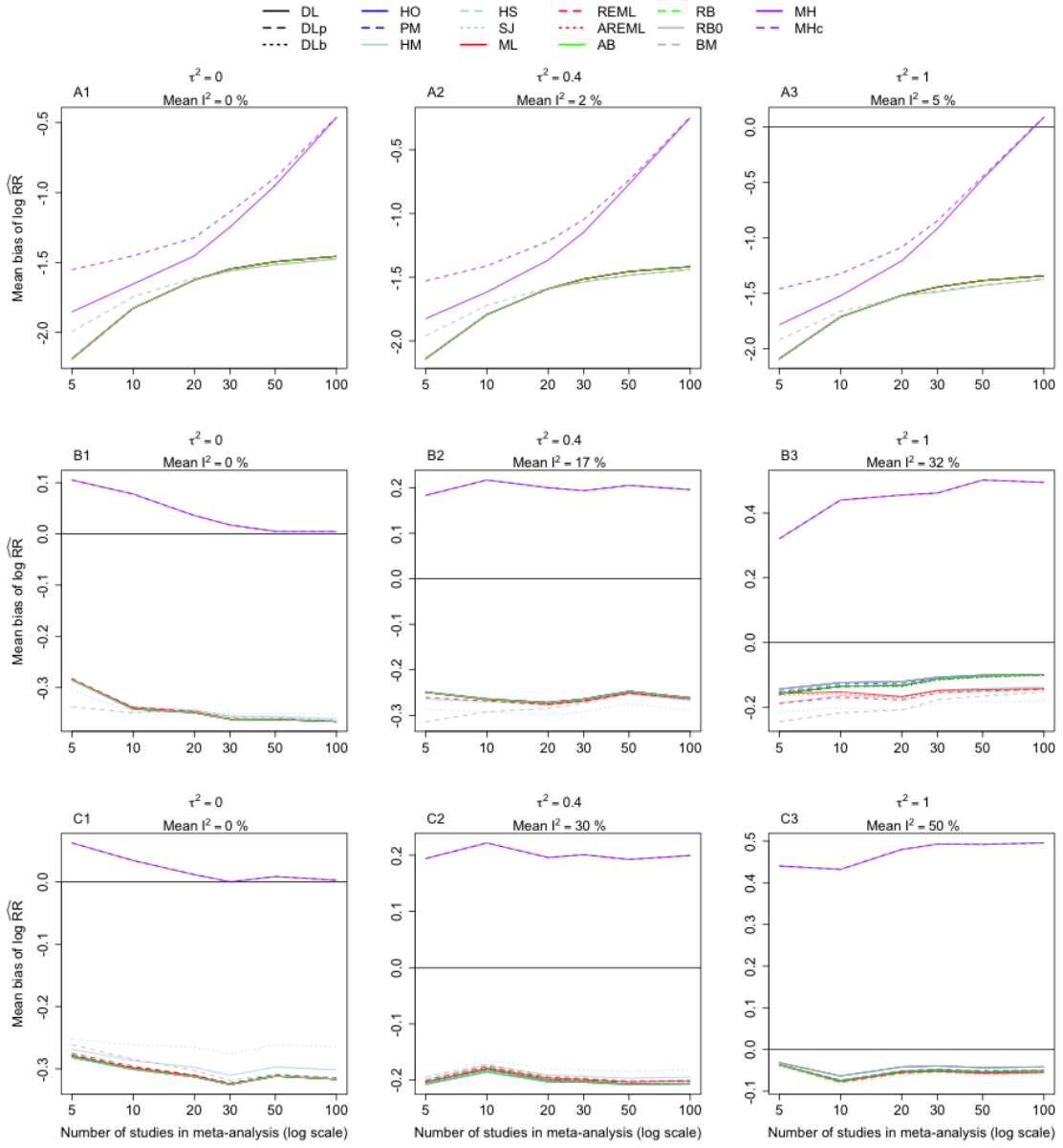


FIGURE E.97: Mean bias of log-risk ratio estimates in very rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

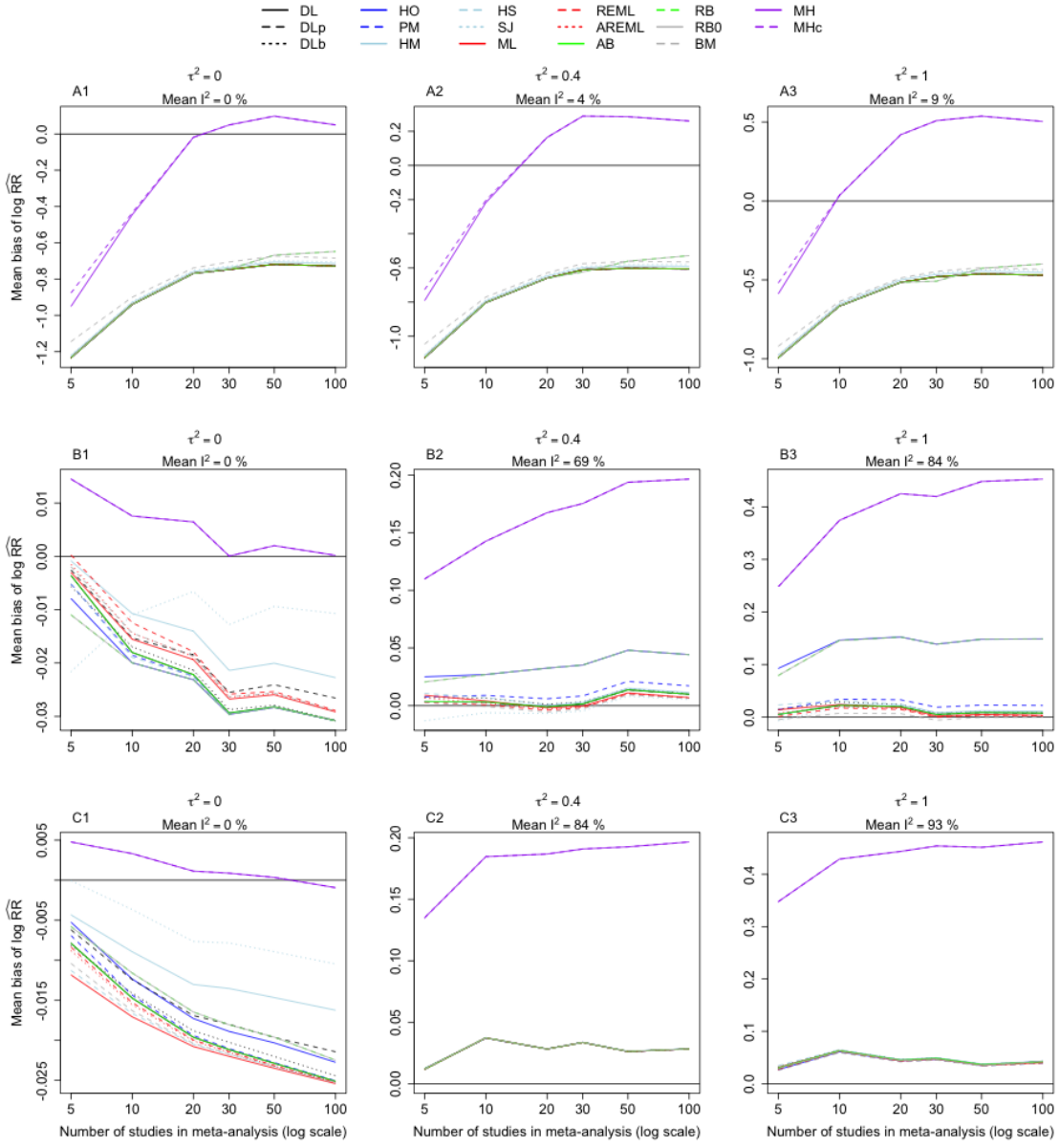


FIGURE E.98: Mean bias of log-risk ratio estimates in rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

E.8 Mean squared error of θ

E.8.1 Examples without omitting outlying estimators

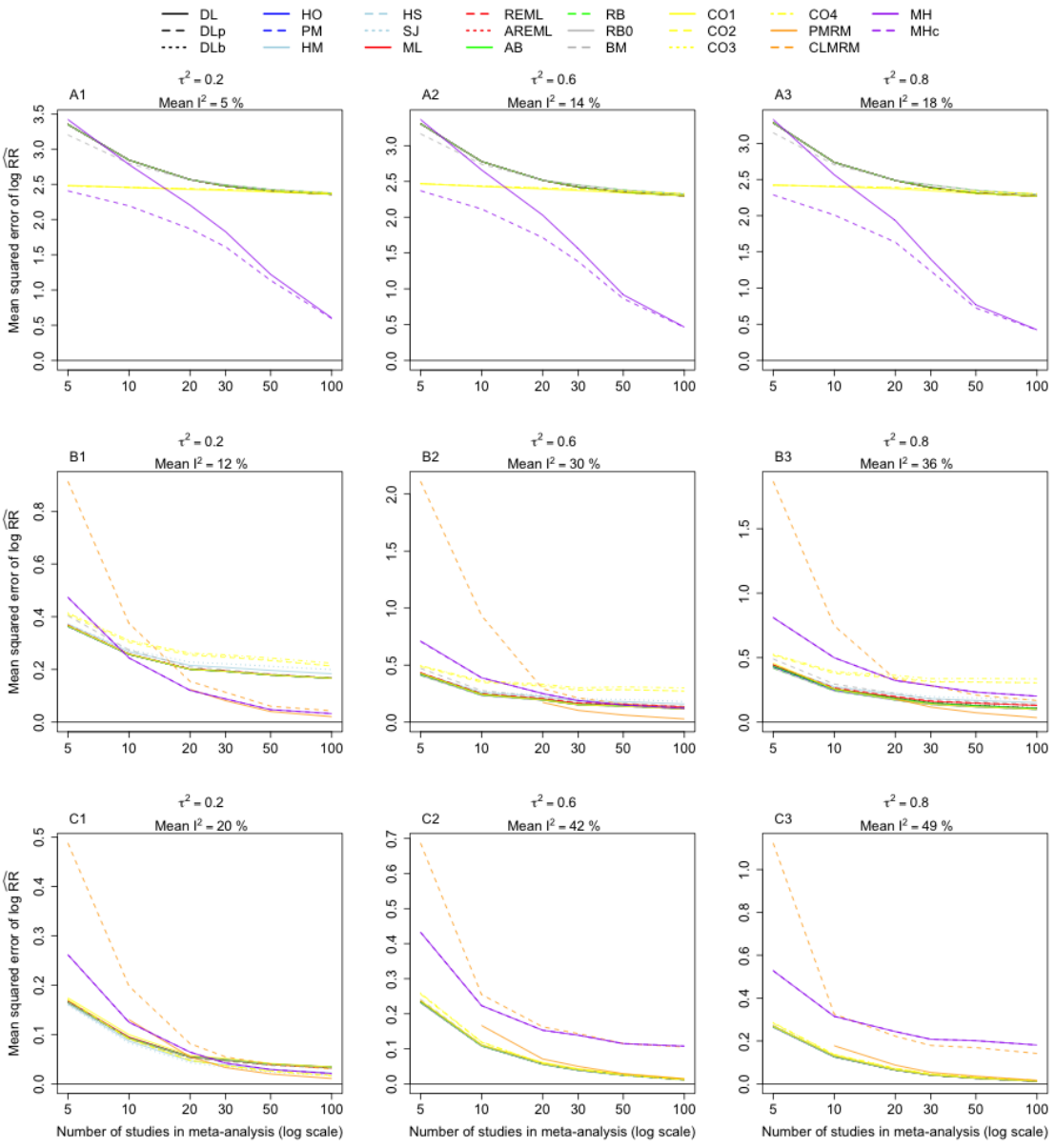


FIGURE E.99: Mean squared error of log-risk ratio estimates in very rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

E.8.2 Alternate values of heterogeneity variance

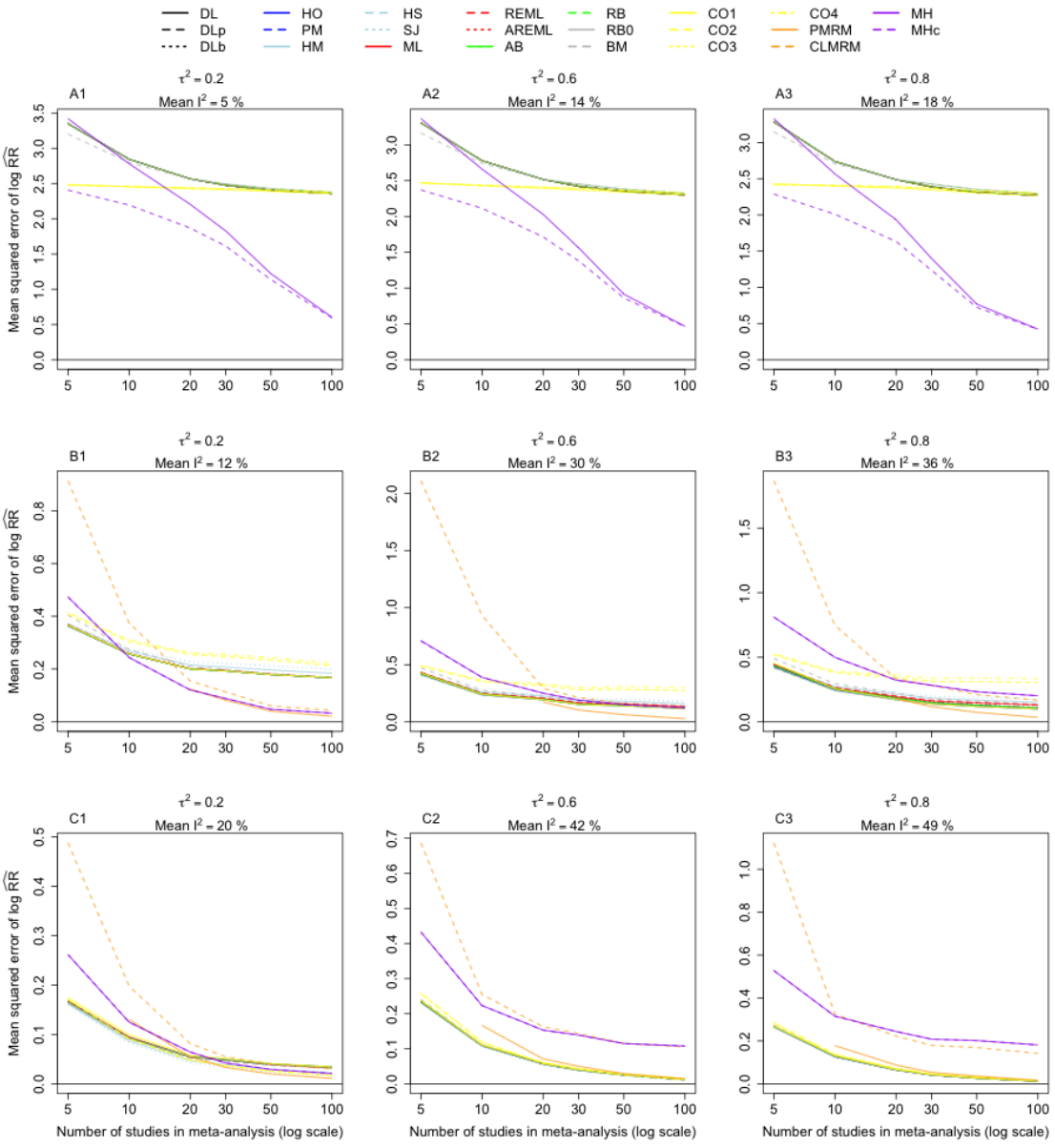


FIGURE E.100: Mean squared error of log-risk ratio estimates in very rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). PMRM and CLMRM are omitted from A1-A3; MM is omitted from all.

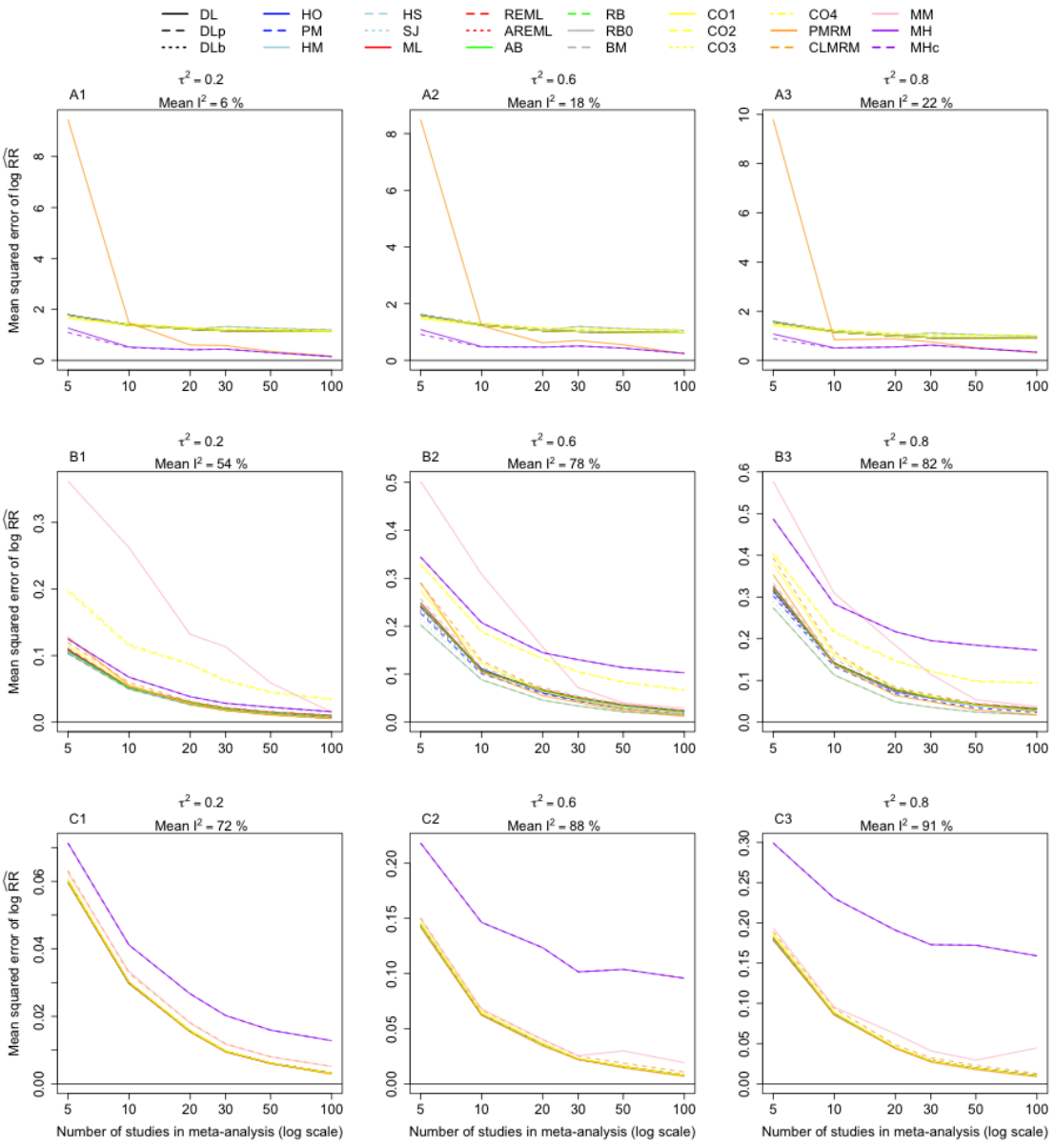


FIGURE E.101: Mean squared error of log-risk ratio estimates in rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). CLMRM and MM are omitted from A1-A3.

E.8.3 Alternate study sample sizes

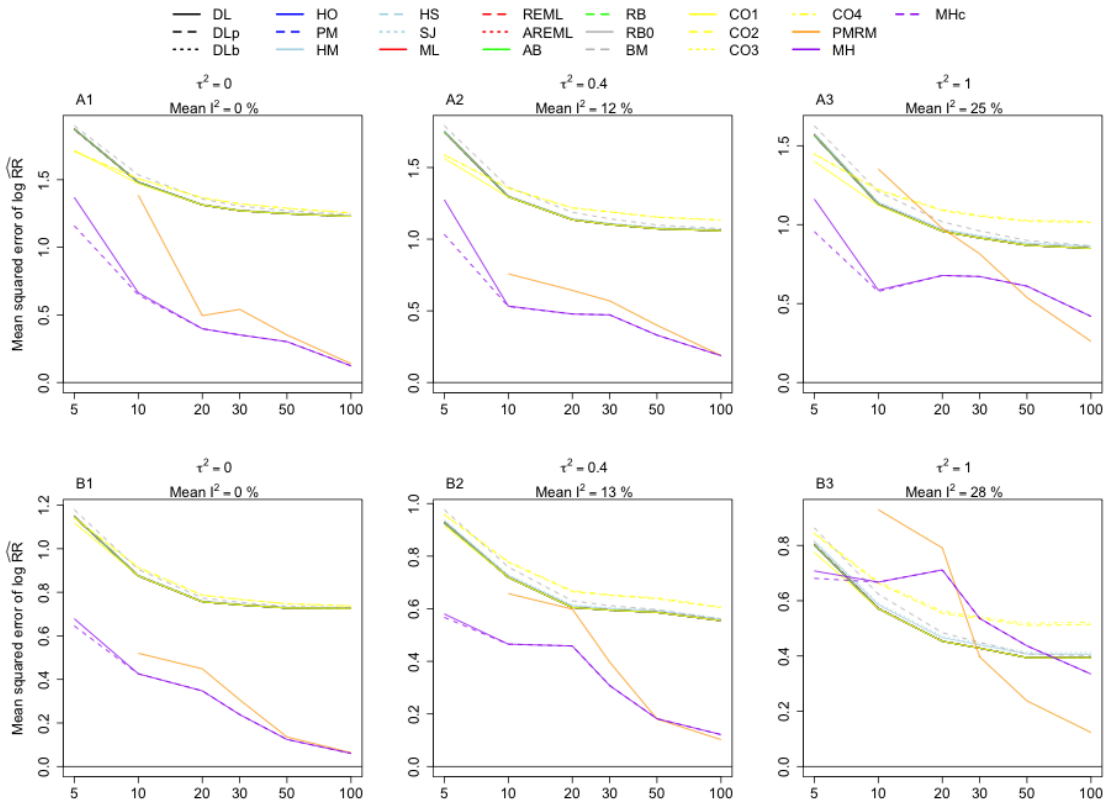


FIGURE E.102: Mean squared error of log-risk ratio estimates in very rare events scenario with $p_0 < p_1$; sample sizes are small-to-medium (A1-A3) and medium (B1-B3). CLMRM and MM are omitted from all.

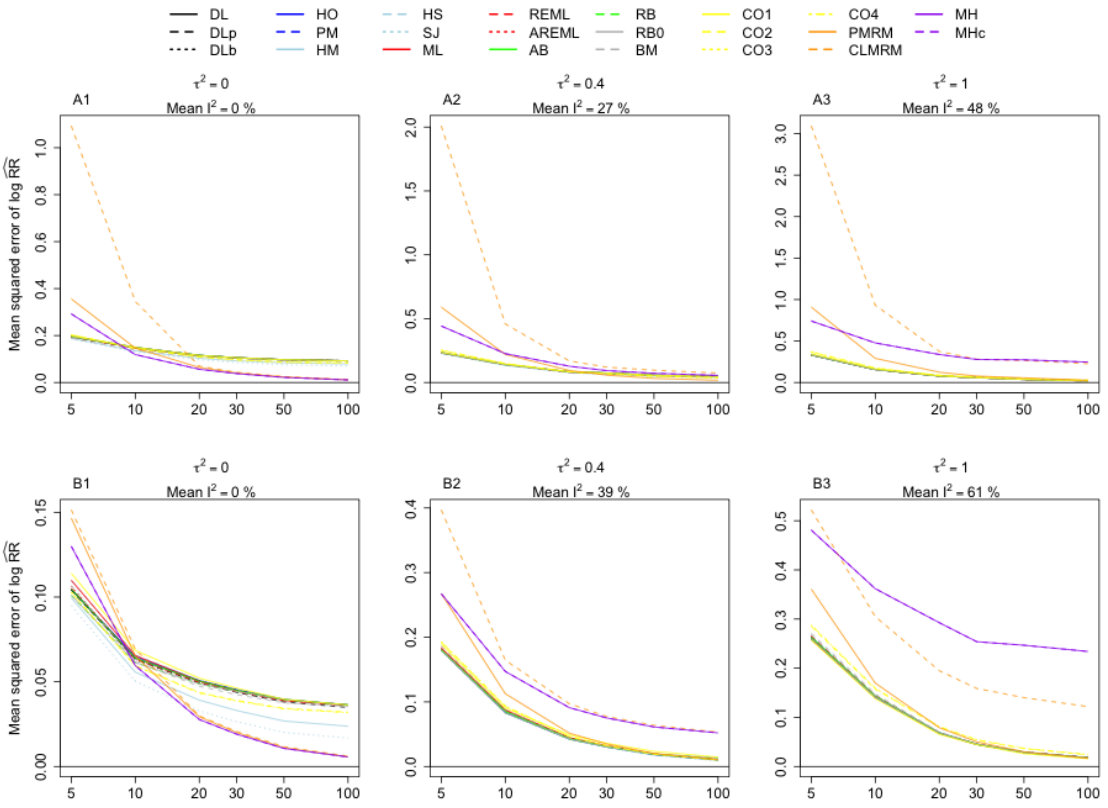


FIGURE E.103: Mean squared error of log-risk ratio estimates in rare events scenario with $p_0 < p_1$; sample sizes are small-to-medium (A1-A3) and medium (B1-B3). MM is omitted from all.

E.8.4 Alternate values of σ_α^2

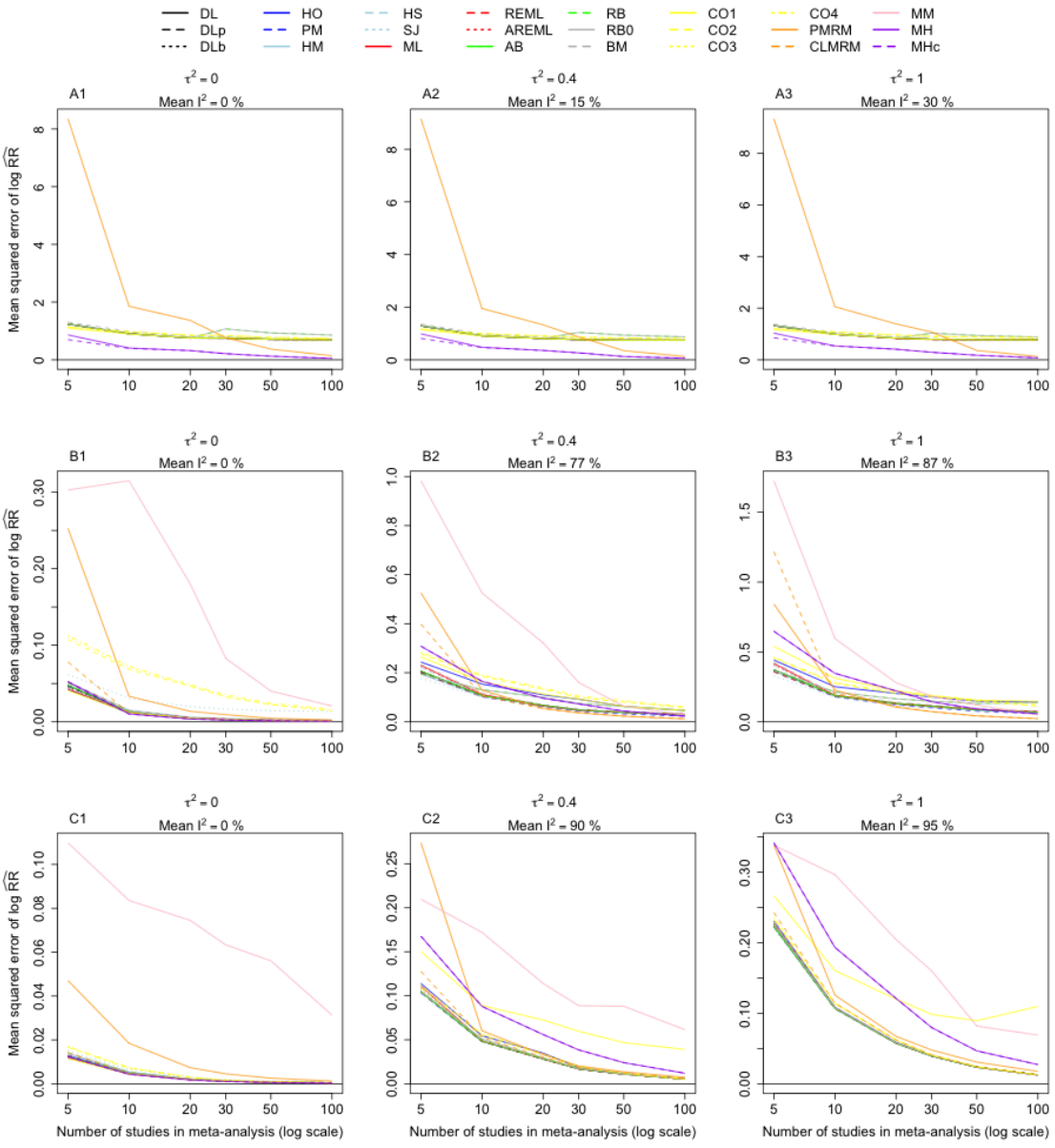


FIGURE E.104: Mean squared error of log-risk ratio estimates in very rare events scenario with $p_0 < p_1$ and $\sigma_\alpha^2 = 3$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

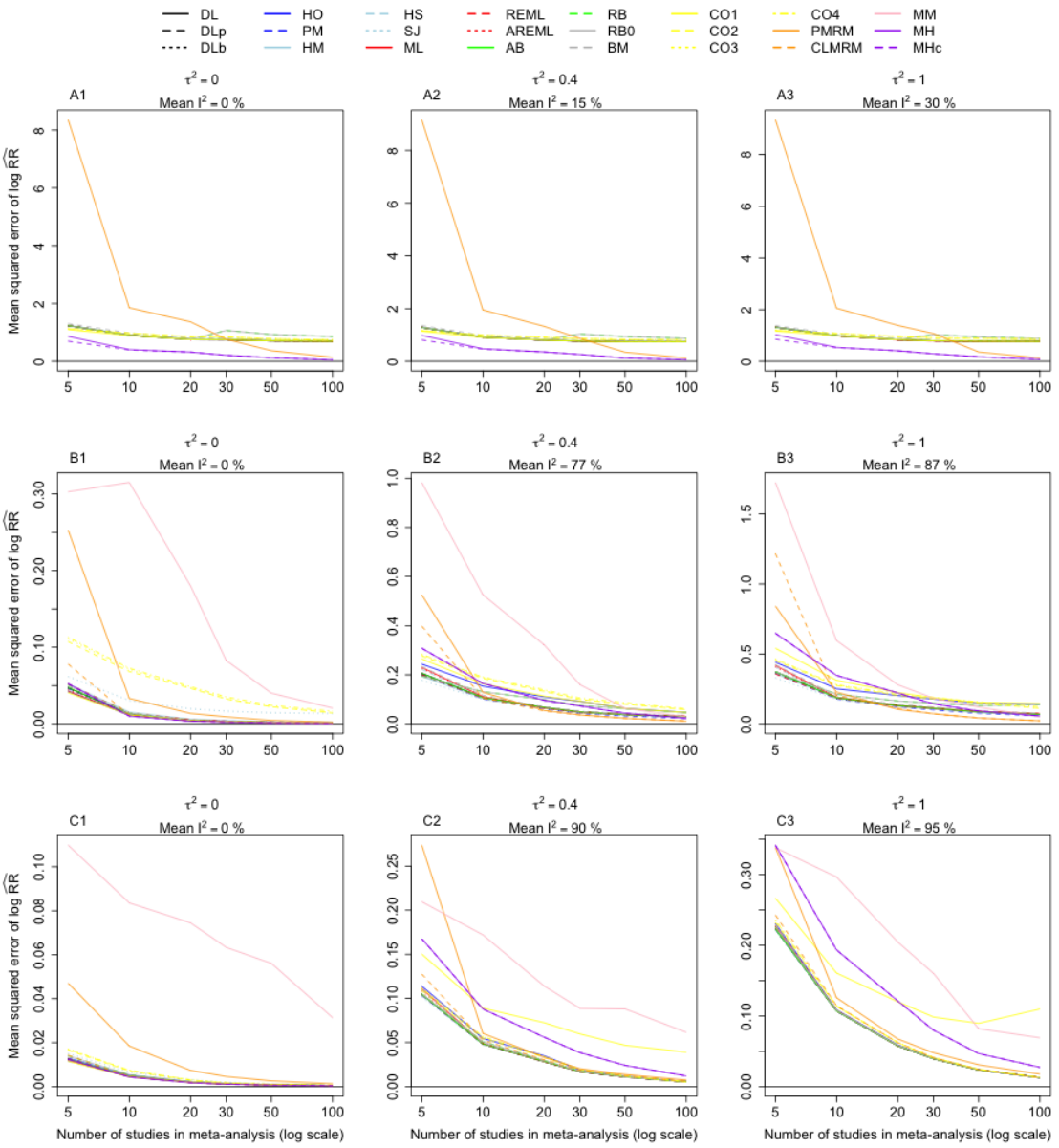


FIGURE E.105: Mean squared error of log-risk ratio estimates in rare events scenario with $p_0 < p_1$ and $\sigma_\alpha^2 = 3$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). CLMRM and MM are omitted from A1-A3.

E.8.5 Alternate probability scenarios

Alternate rare events scenarios

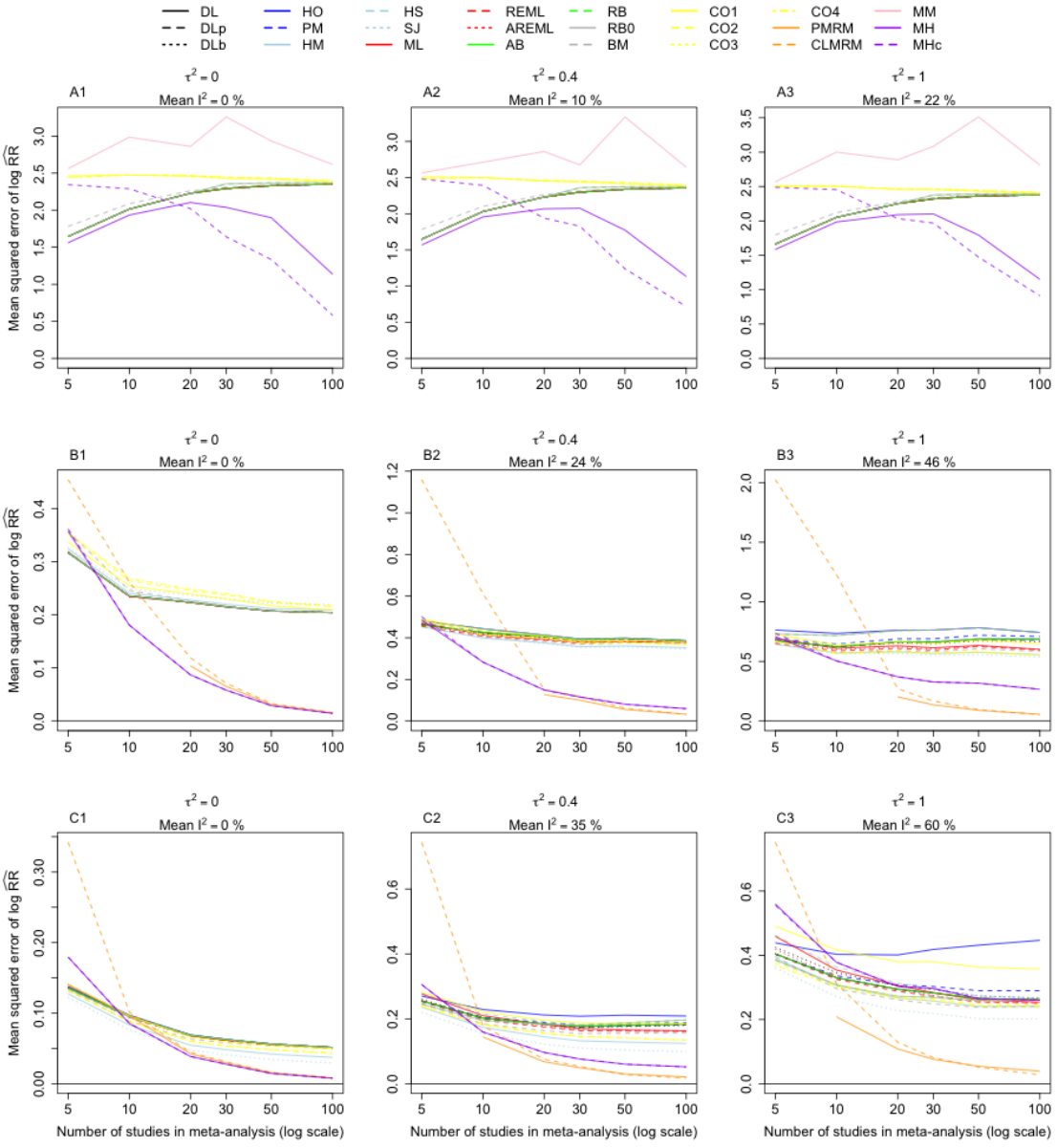


FIGURE E.106: Mean squared error of log-risk ratio estimates in very rare events scenario with $p_0 > p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). PMRM and CLMRM are omitted from A1-A3; MM is omitted from B1-C3.

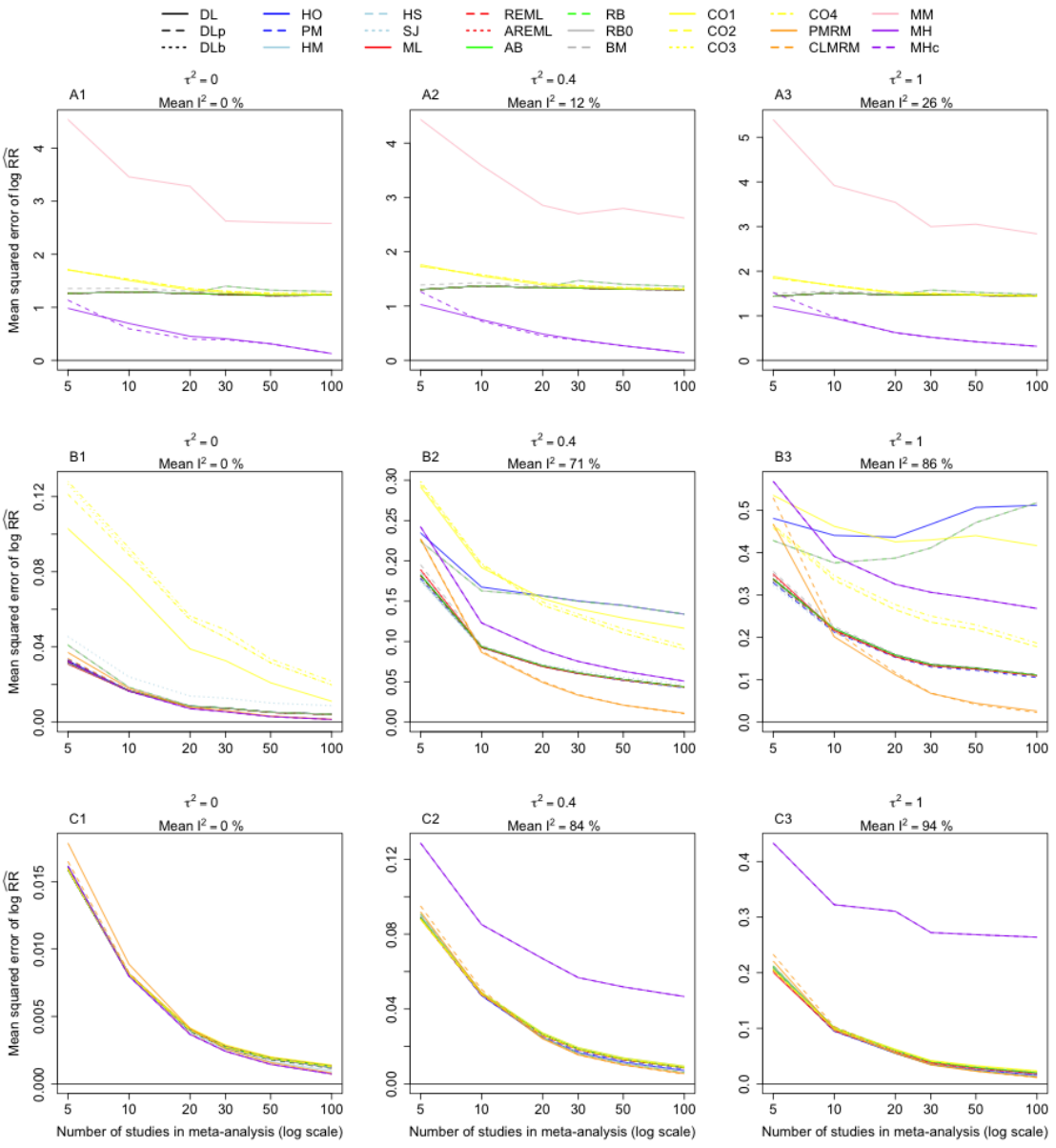


FIGURE E.107: Mean squared error of log-risk ratio estimates in rare events scenario with $p_0 > p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). PMRM and CLMRM are omitted from A1-A3; MM is omitted from B1-C3.

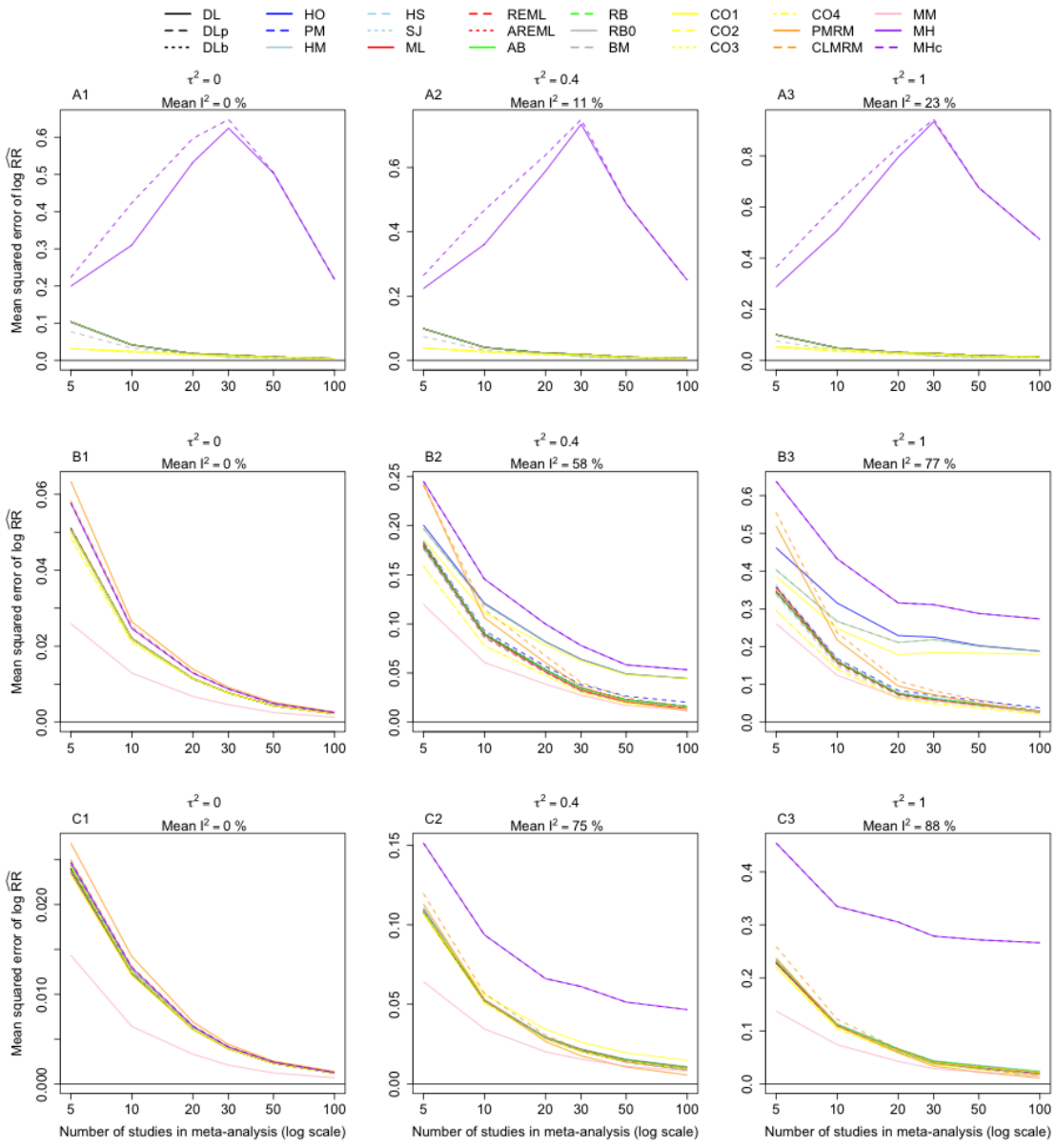


FIGURE E.108: Mean squared error of log-risk ratio estimates in rare events scenario with $p_0 = p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). PMRM, CLMRM and MM are omitted from A1-A3.

Common probability scenarios

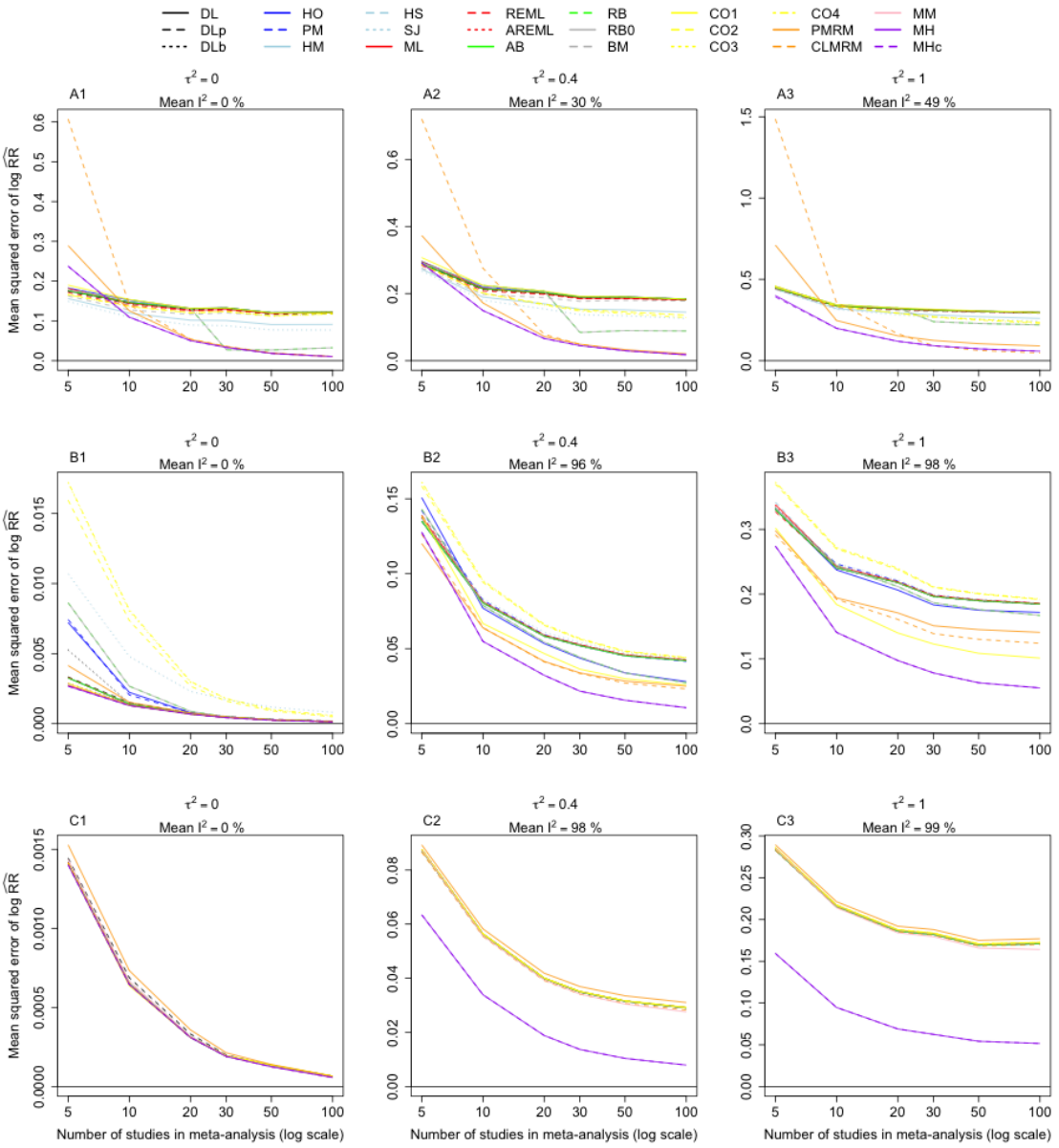


FIGURE E.109: Mean squared error of log-risk ratio estimates in common probability scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). MM is omitted from A1-B3.

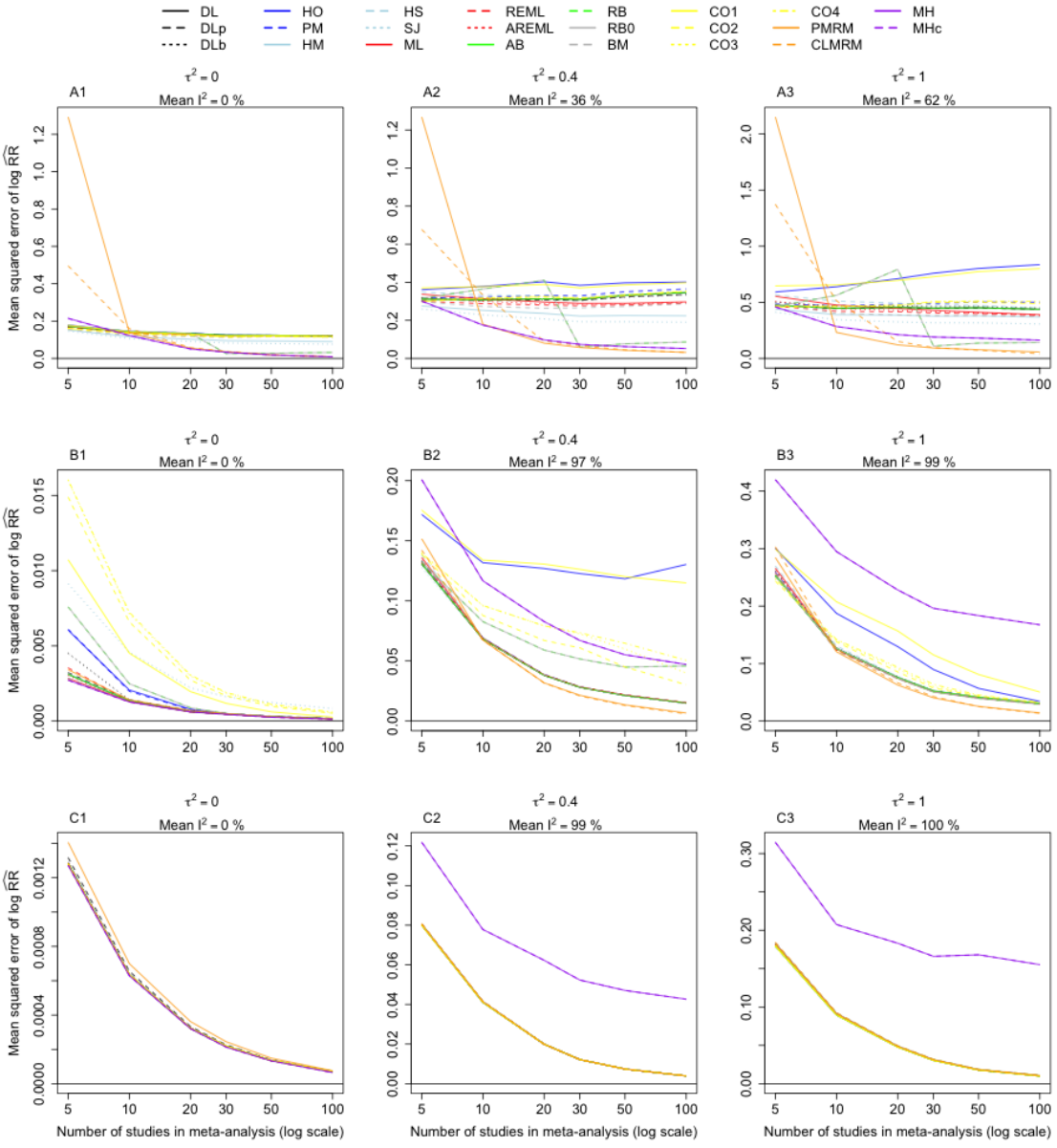


FIGURE E.110: Mean squared error of log-risk ratio estimates in common probability scenario with $p_0 > p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). MM is omitted from all.

E.8.6 Alternate sampling in simulation study

Alternate event count sampling

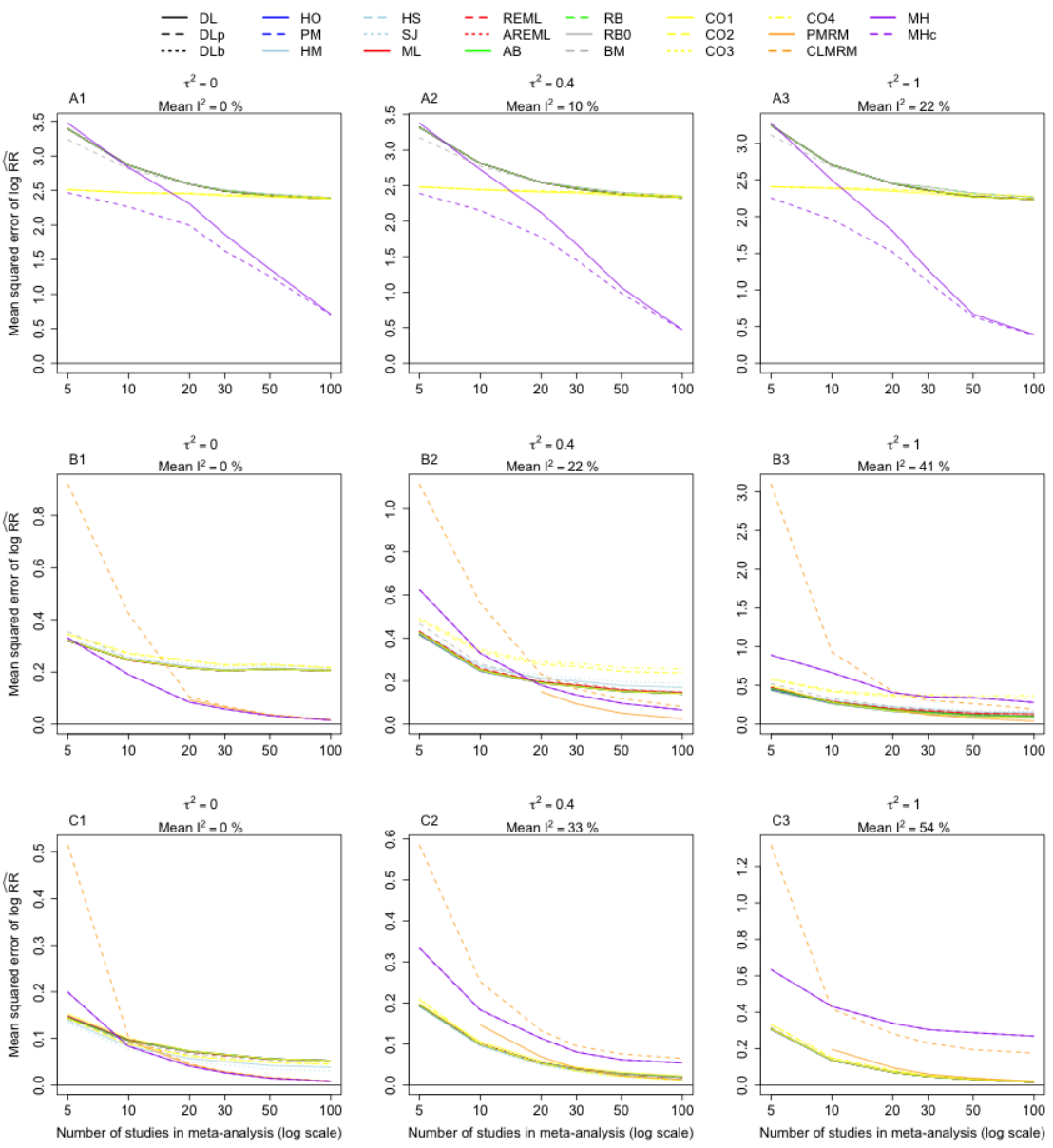


FIGURE E.111: Mean squared error of log-risk ratio estimates in very rare events scenario with $p_0 < p_1$ and poisson event sampling; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). PMRM and CLMRM are omitted from A1-A3; MM is omitted from all.

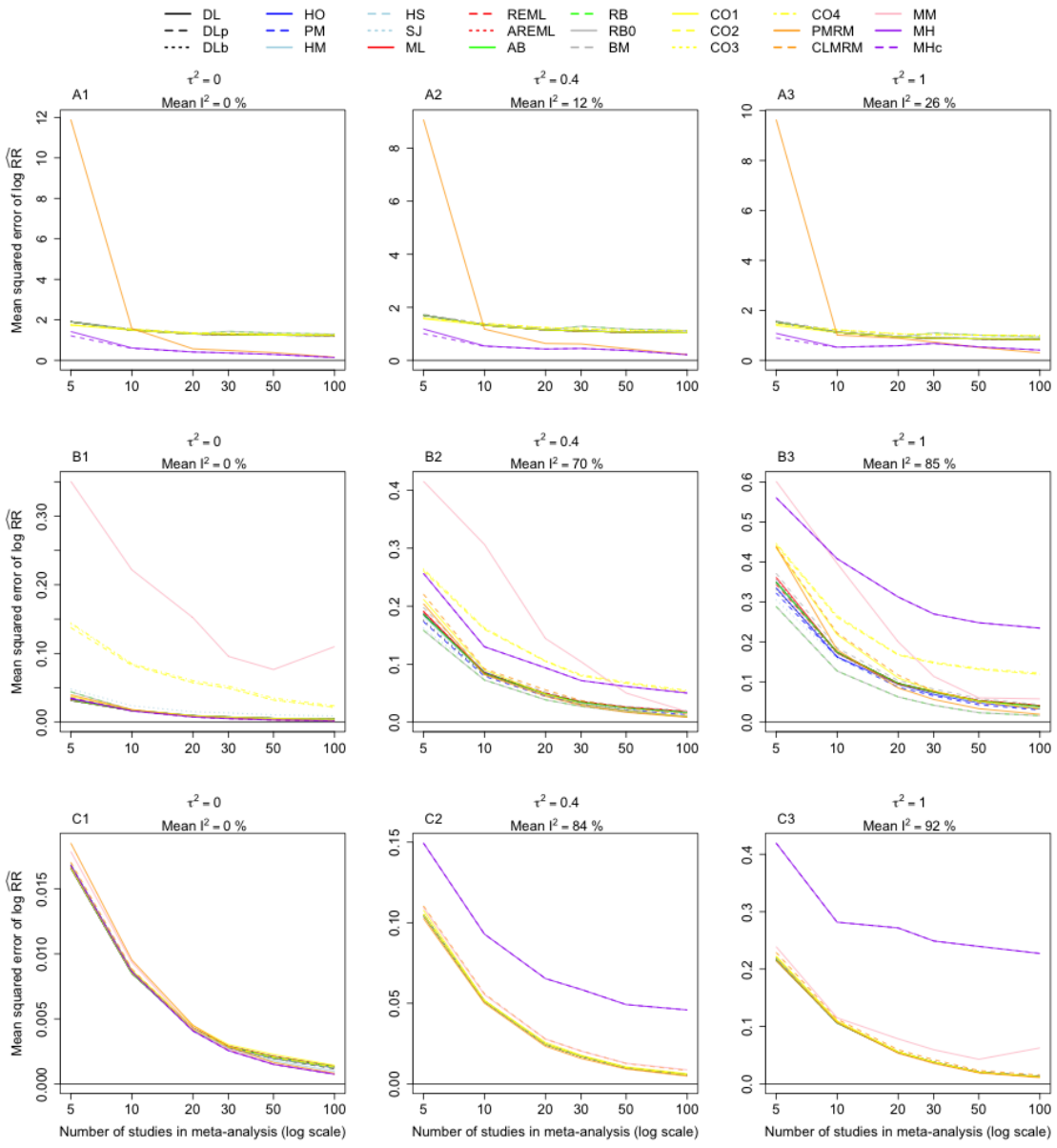


FIGURE E.112: Mean squared error of log-risk ratio estimates in rare events scenario with $p_0 < p_1$ and poisson event sampling; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). CLMRM and MM are omitted from A1-A3.

Alternate sample size sampling

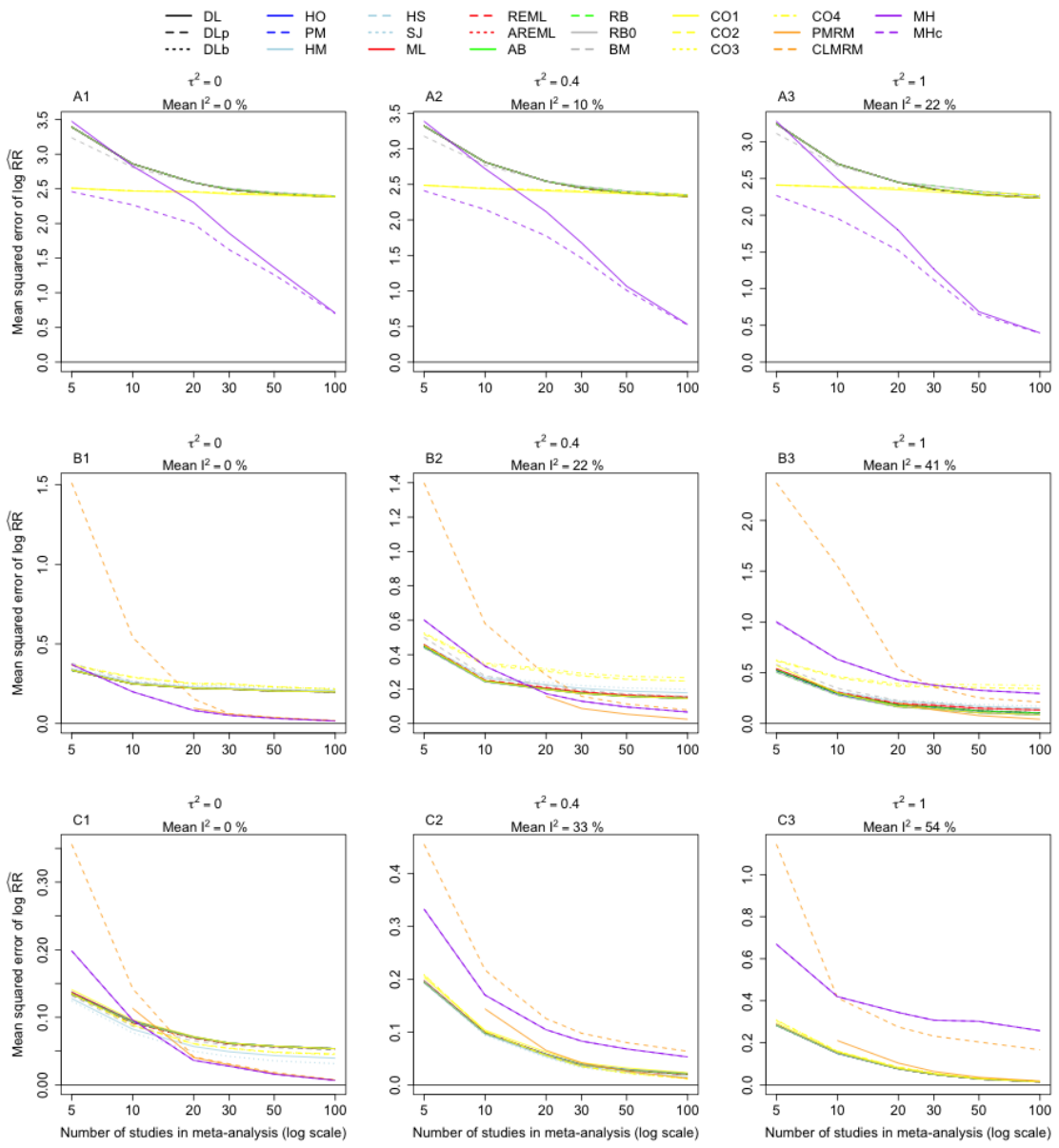


FIGURE E.113: Mean squared error of log-risk ratio estimates in very rare events scenario with $p_0 < p_1$ and normal sample size sampling; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). PMRM and CLMRM are omitted from A1-A3; MM is omitted from all.

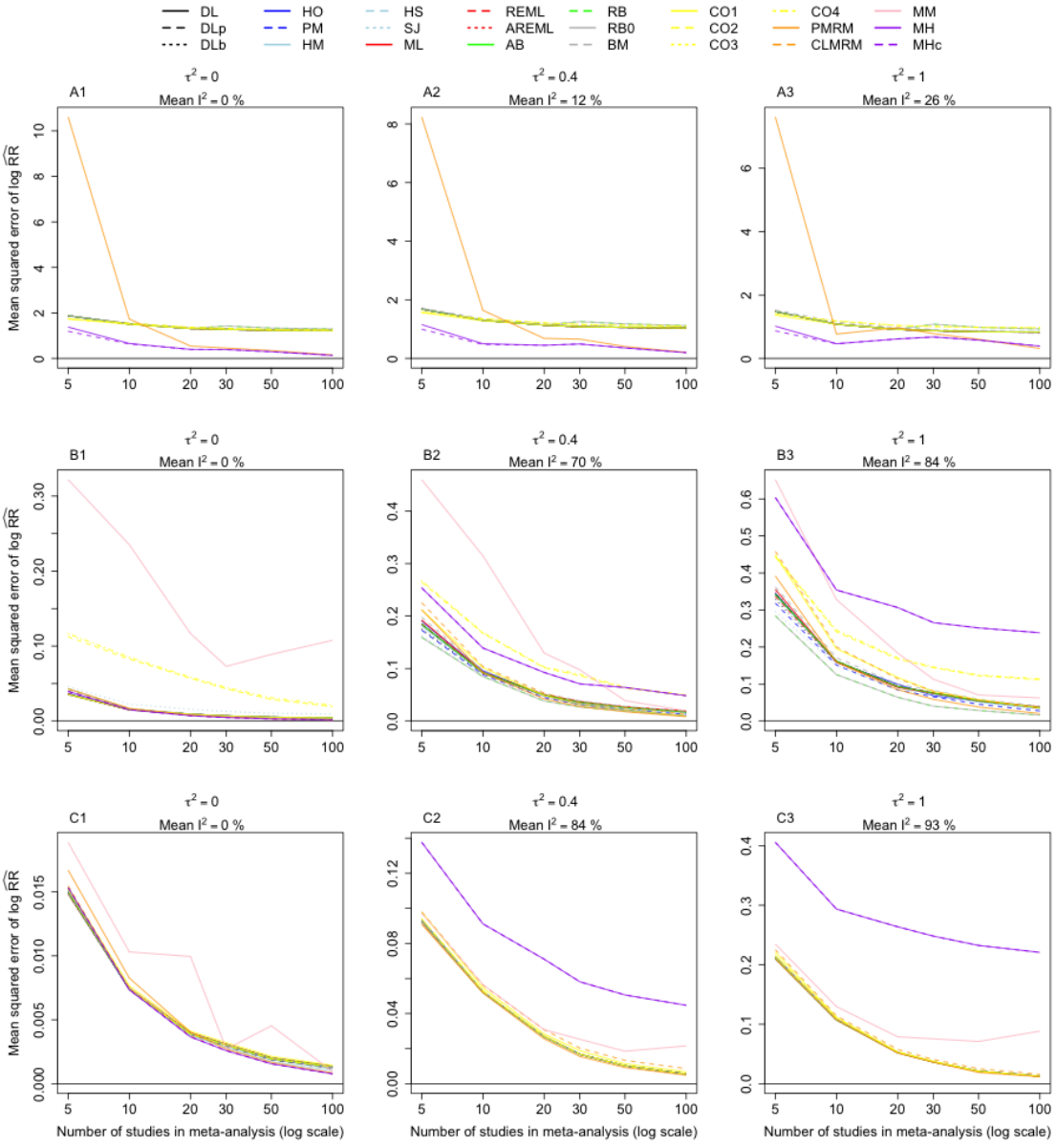


FIGURE E.114: Mean squared error of log-risk ratio estimates in rare events scenario with $p_0 < p_1$ and normal sample size sampling; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). CLMRM and MM are omitted from A1-A3.

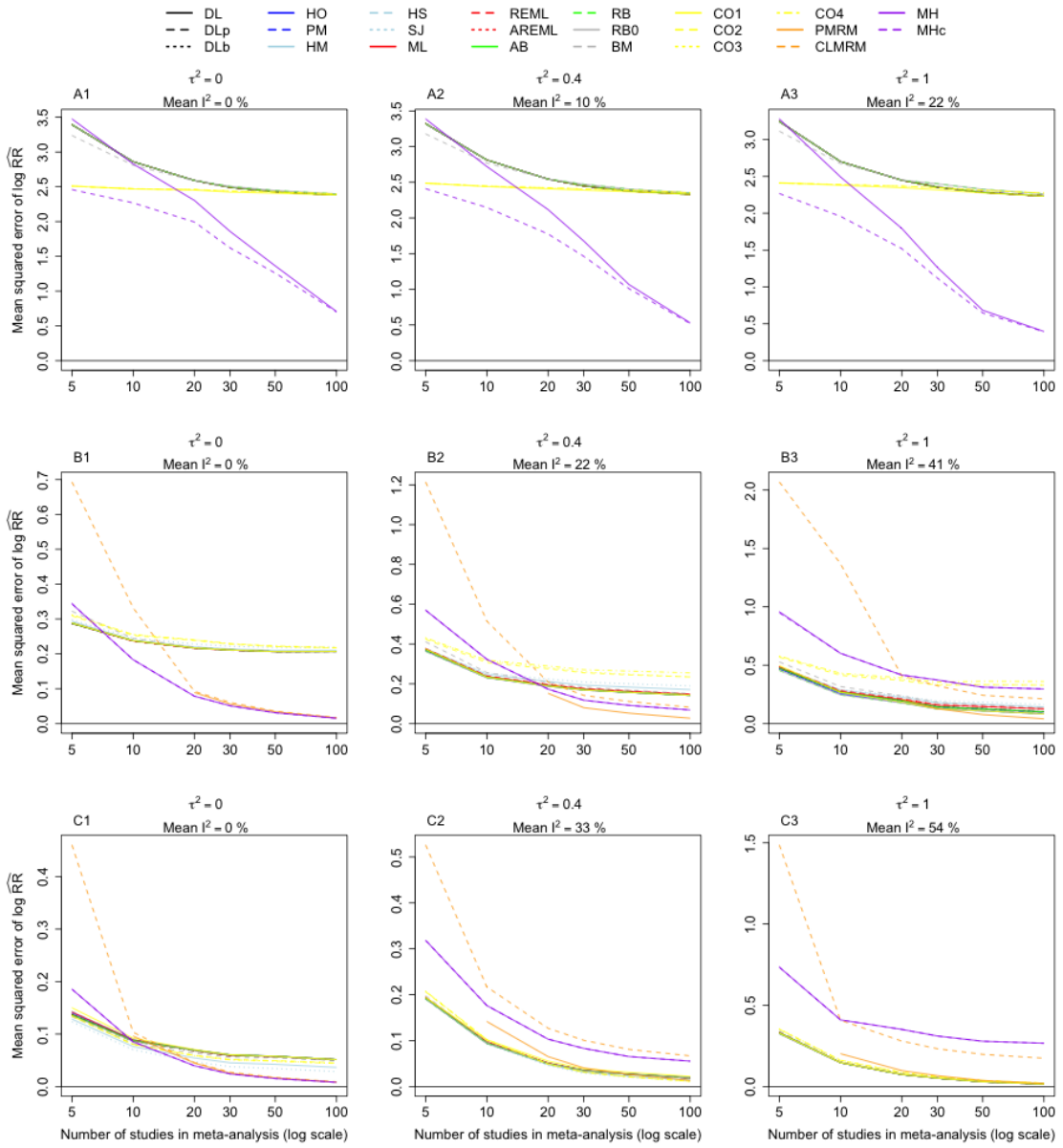


FIGURE E.115: Mean squared error of log-risk ratio estimates in very rare events scenario with $p_0 < p_1$ and Chi-squared sample size sampling; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). PMRM and CLMRM are omitted from A1-A3; MM is omitted from all.

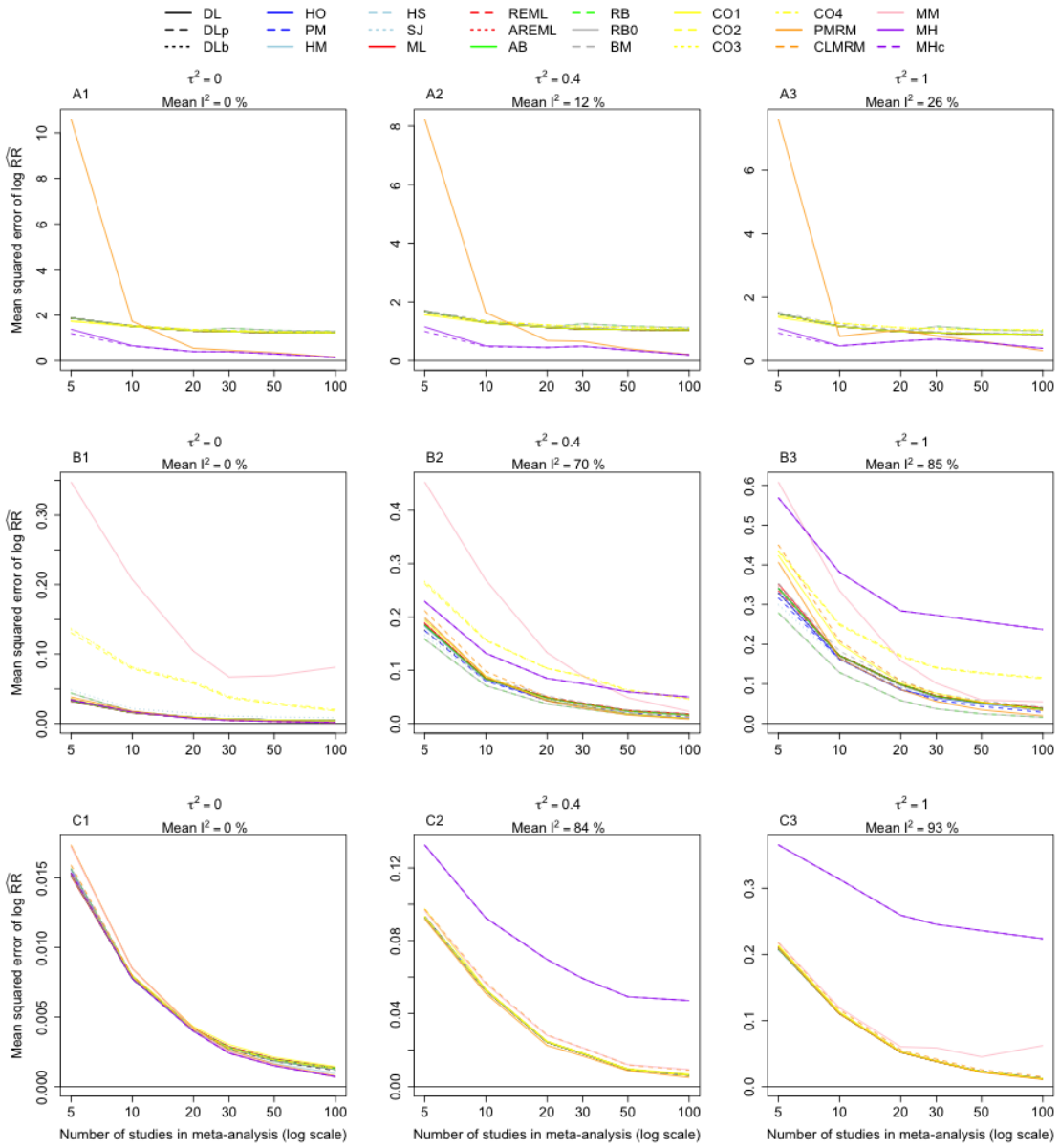


FIGURE E.116: Mean squared error of log-risk ratio estimates in rare events scenario with $p_0 < p_1$ and Chi-squared sample size sampling; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3). PMRM and CLMRM are omitted from A1-A3; MM is omitted from all.

E.8.7 Alternate continuity corrections

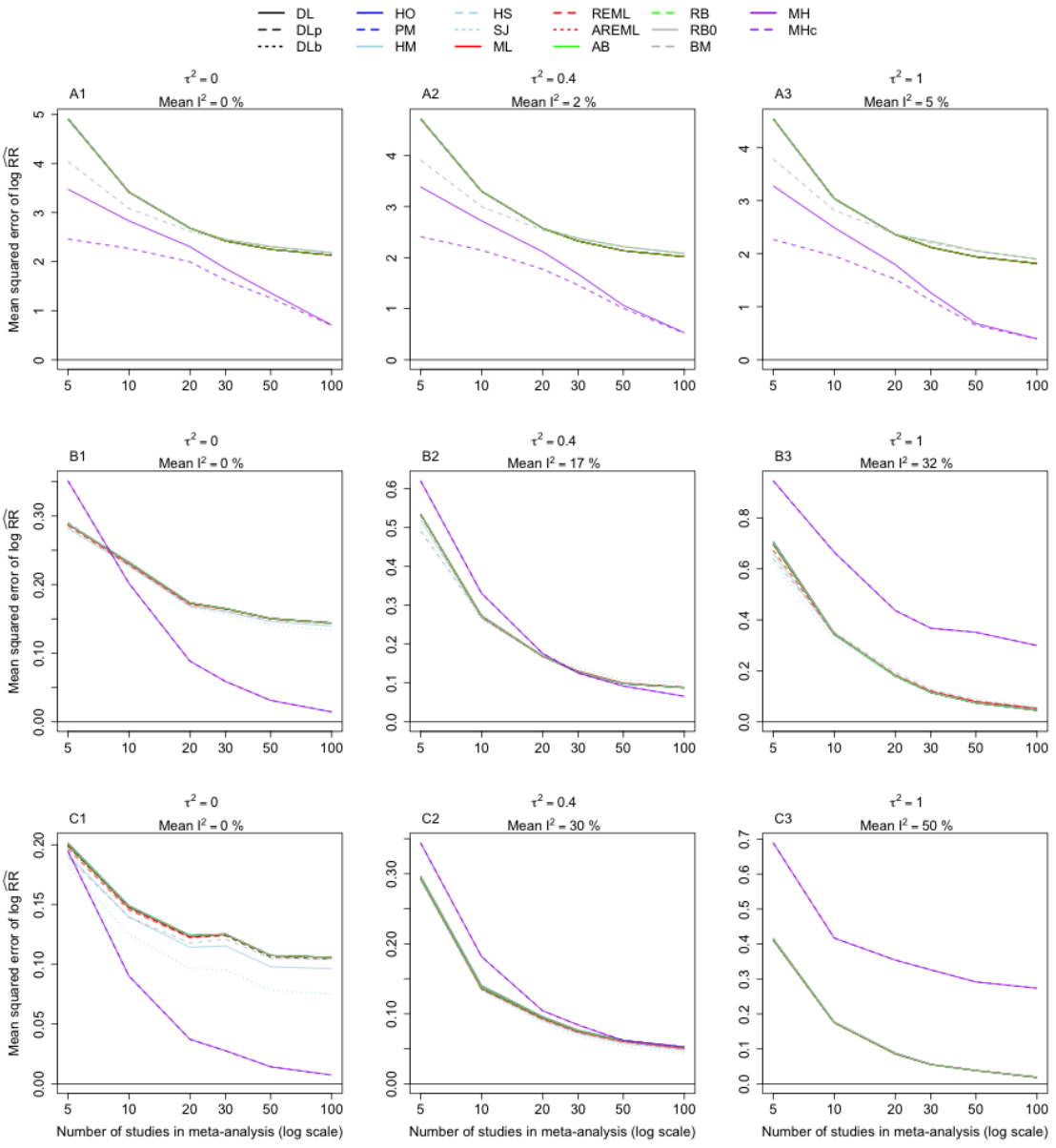


FIGURE E.117: Mean squared error of log-risk ratio estimates in very rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

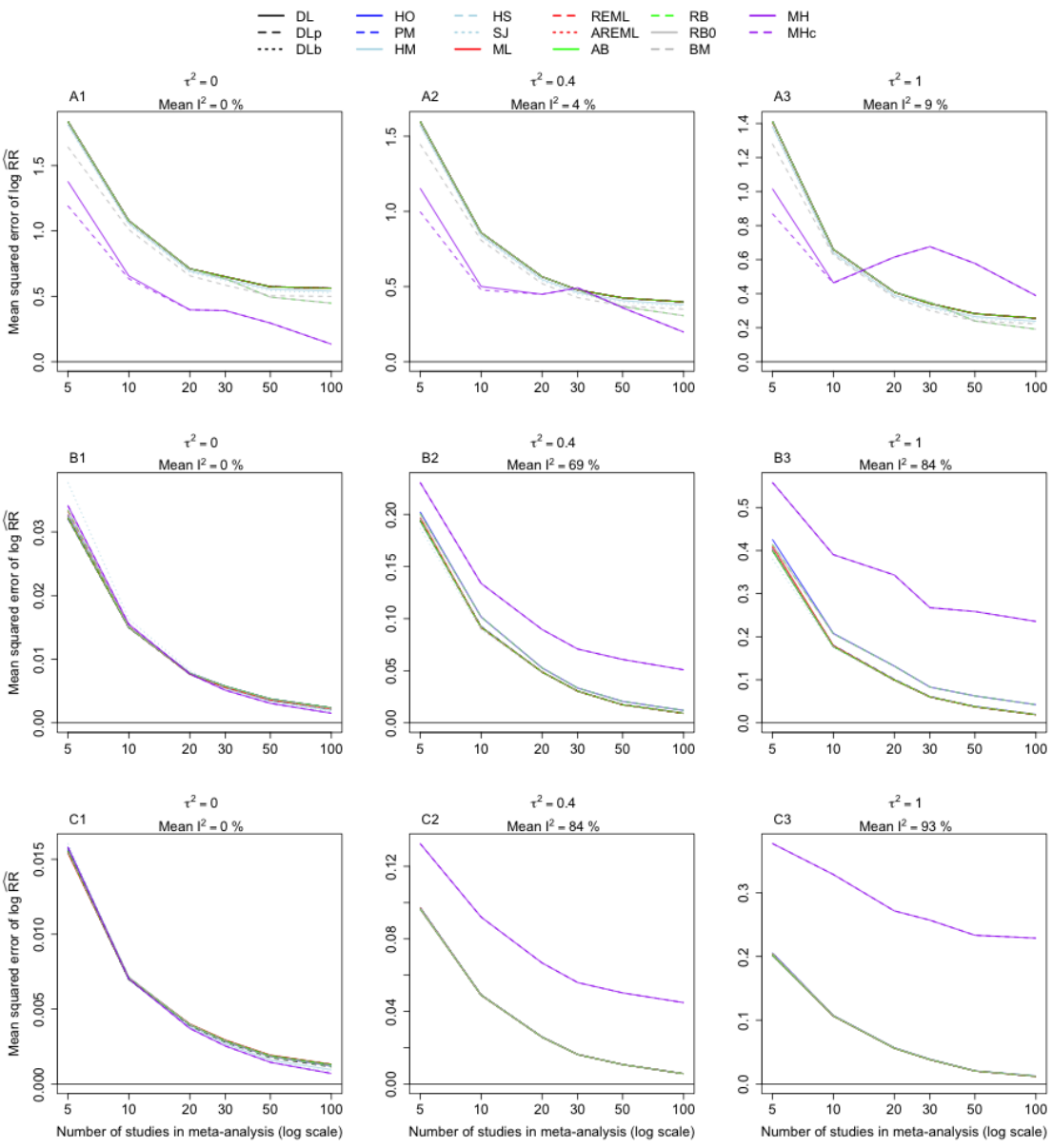


FIGURE E.118: Mean squared error of log-risk ratio estimates in rare events scenario with $p_0 < p_1$; sample sizes are small (A1-A3), small and large (B1-B3) and large (C1-C3).

E.9 Coverage

E.9.1 Rare events scenario

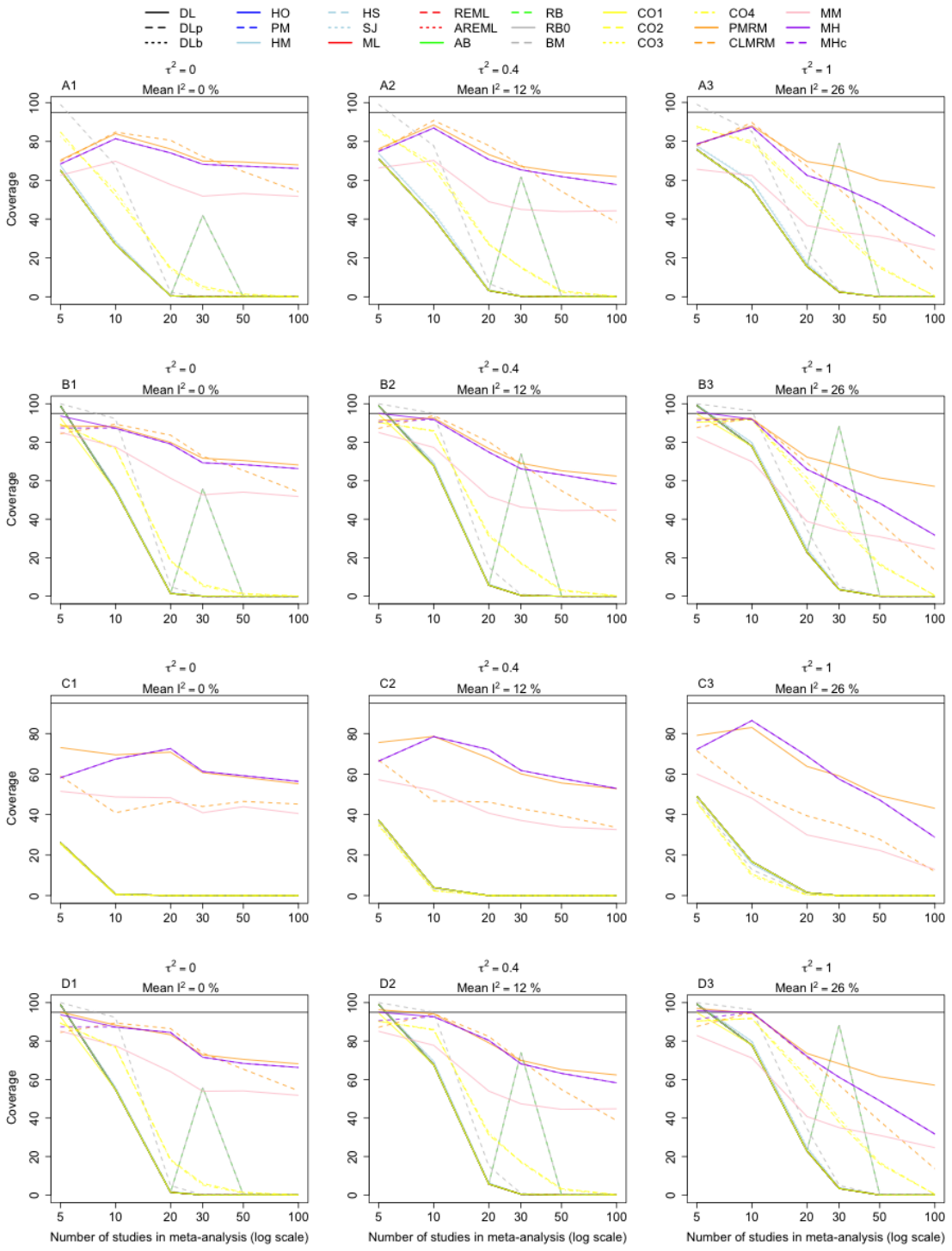


FIGURE E.119: Coverage of log-risk ratio confidence intervals in rare events scenario with $p_0 < p_1$ and small sample sizes; confidence intervals are Wald-type (A1-A3), t-distribution (B1-B3), HKSJ (C1-C3) and mKH (D1-D3).

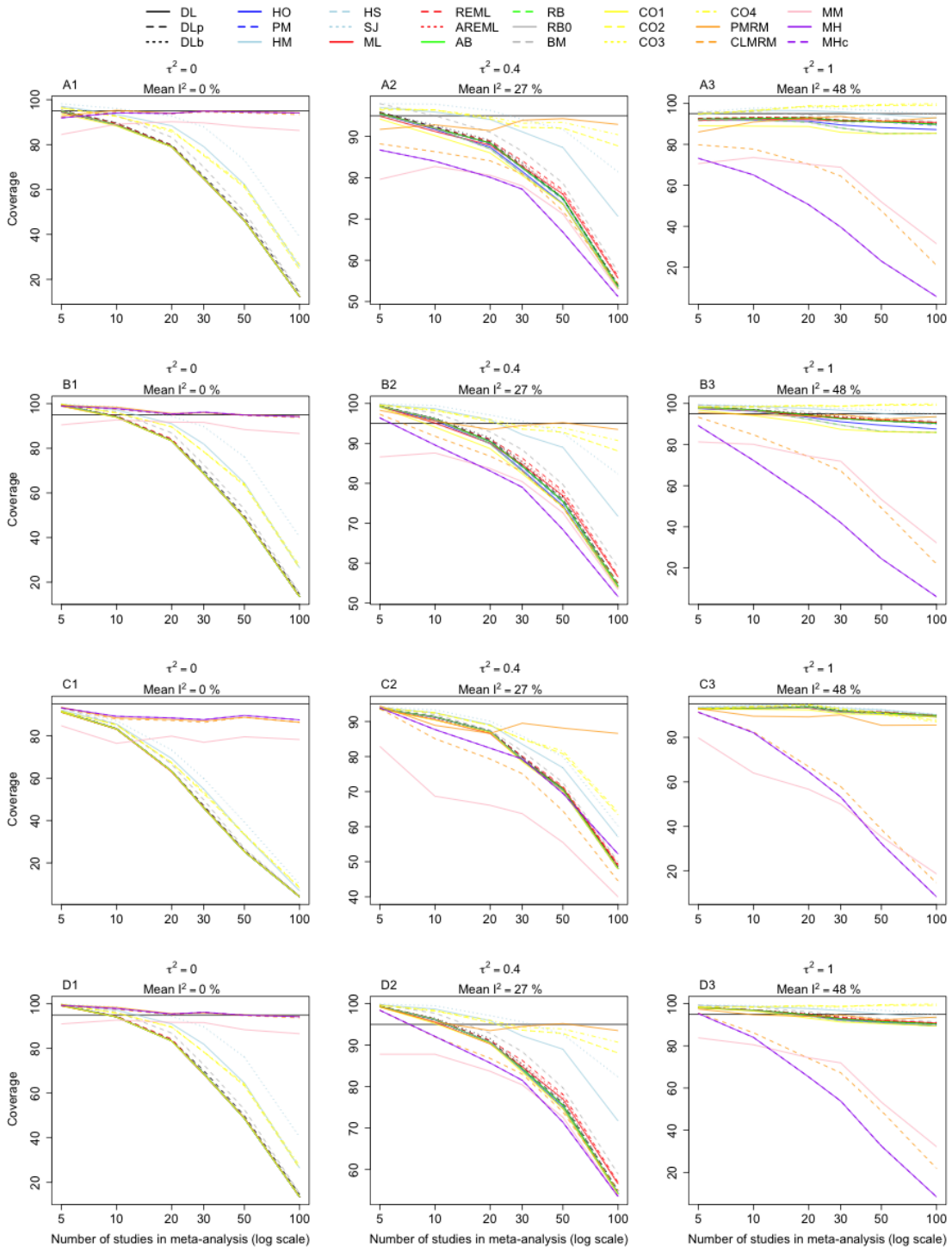


FIGURE E.120: Coverage of log-risk ratio confidence intervals in rare events scenario with $p_0 < p_1$ and small-to-medium sample sizes; confidence intervals are Wald-type (A1-A3), t-distribution (B1-B3), HKSJ (C1-C3) and mKH (D1-D3).

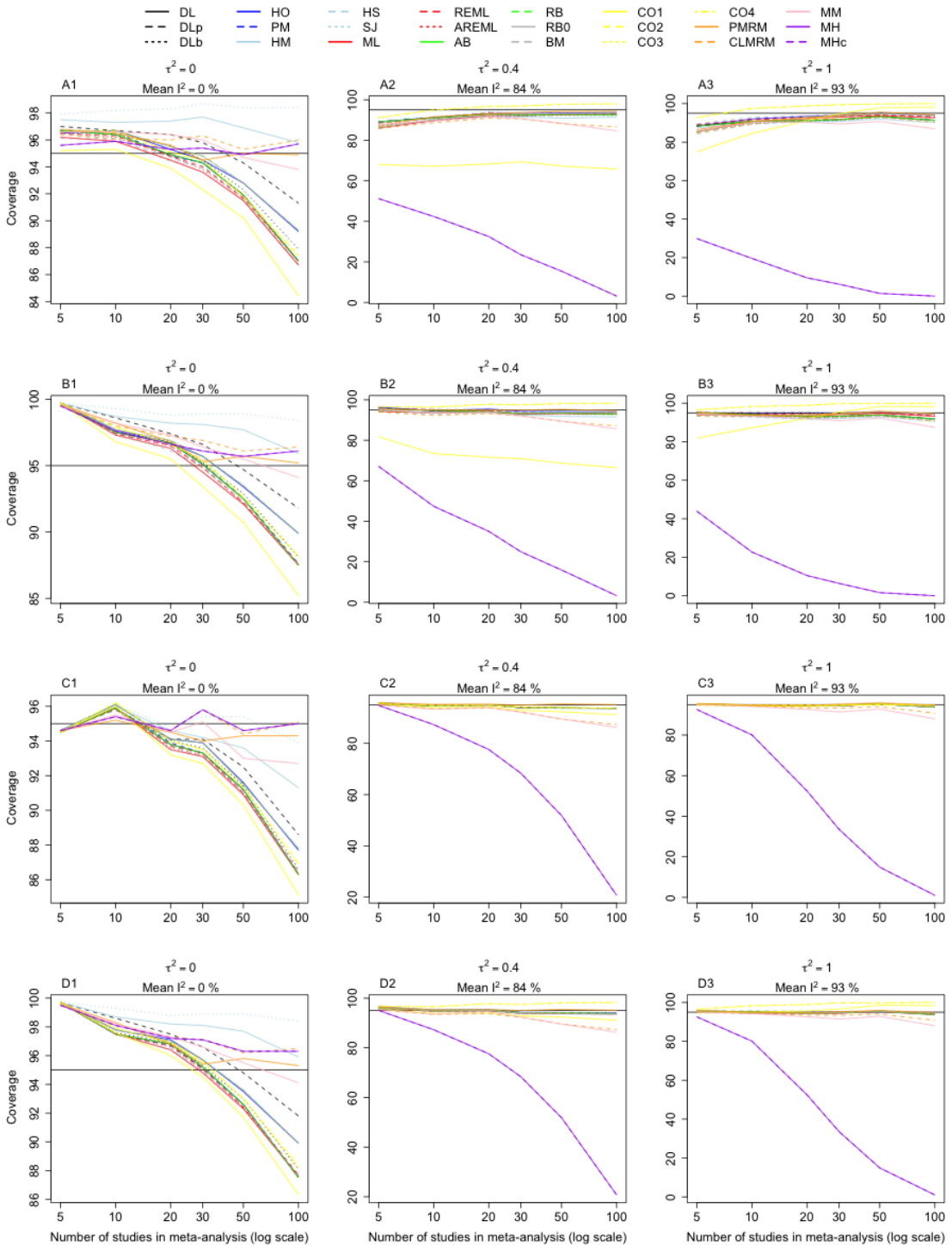


FIGURE E.121: Coverage of log-risk ratio confidence intervals in rare events scenario with $p_0 < p_1$ and large sample sizes; confidence intervals are Wald-type (A1-A3), t-distribution (B1-B3), HKSJ (C1-C3) and mKH (D1-D3).

E.9.2 Very rare events scenario

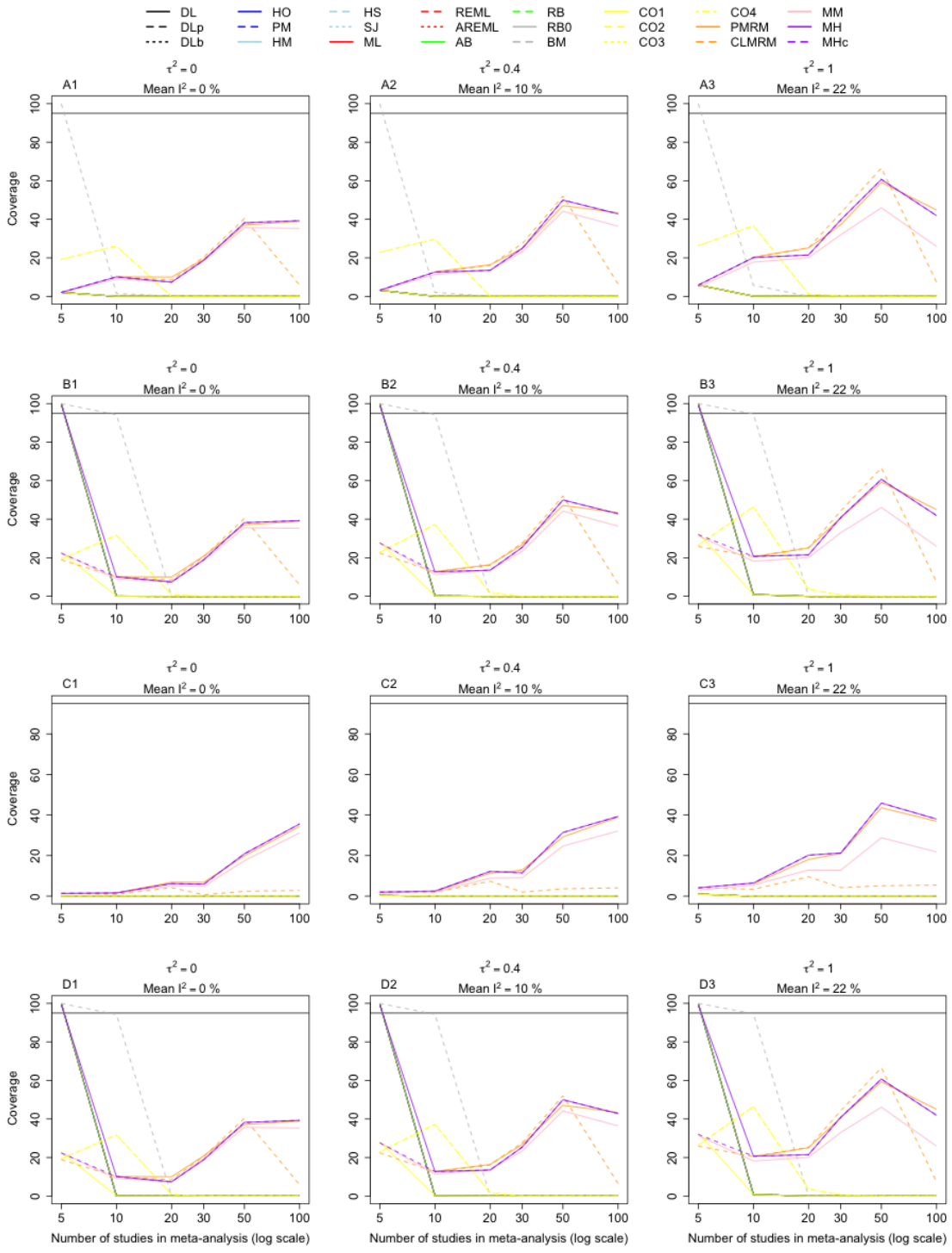


FIGURE E.122: Coverage of log-risk ratio confidence intervals in very rare events scenario with $p_0 < p_1$ and small sample sizes; confidence intervals are Wald-type (A1-A3), t-distribution (B1-B3), HKSJ (C1-C3) and mKH (D1-D3).

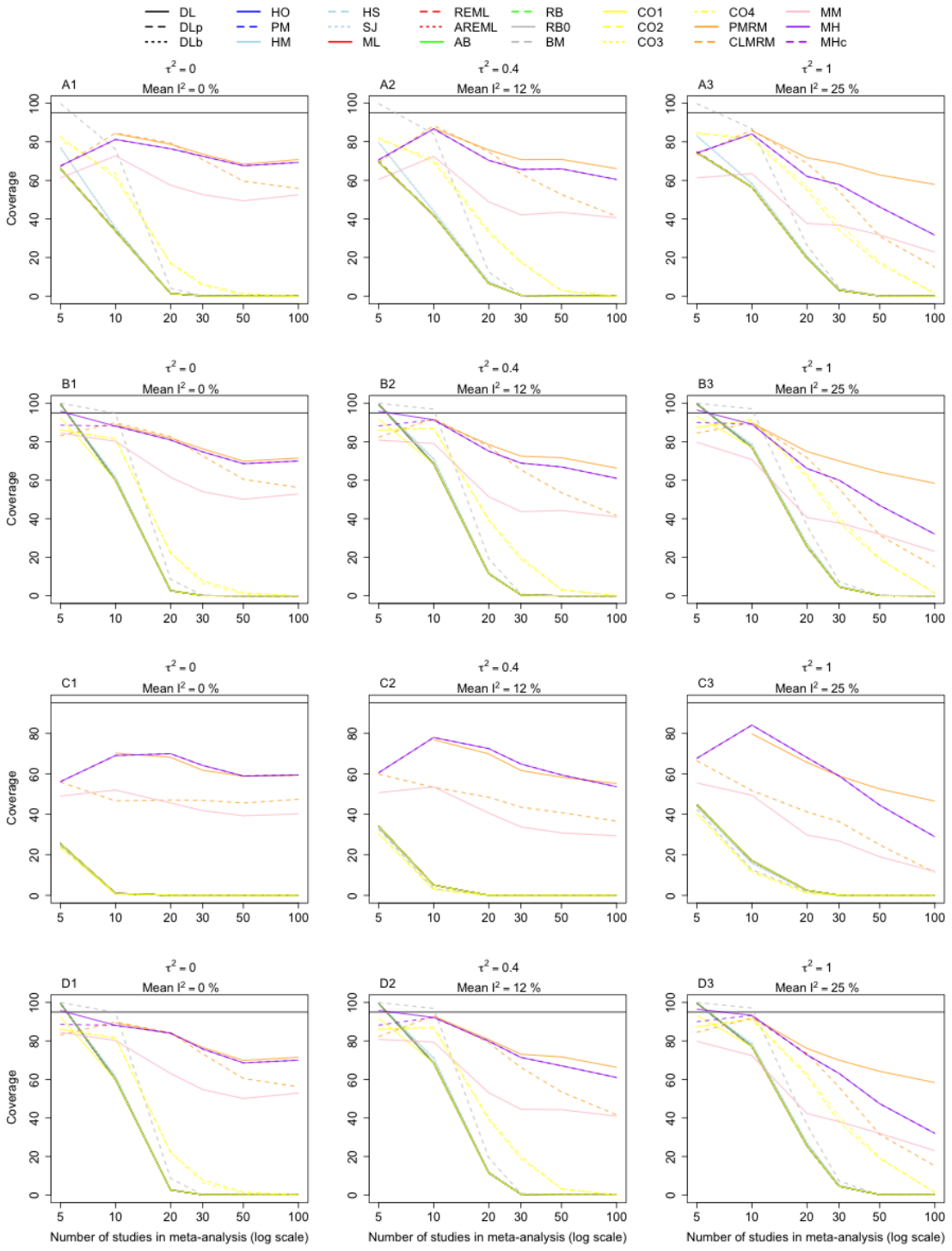


FIGURE E.123: Coverage of log-risk ratio confidence intervals in very rare events scenario with $p_0 < p_1$ and small-to-medium sample sizes; confidence intervals are Wald-type (A1-A3), t-distribution (B1-B3), HKSJ (C1-C3) and mKH (D1-D3).

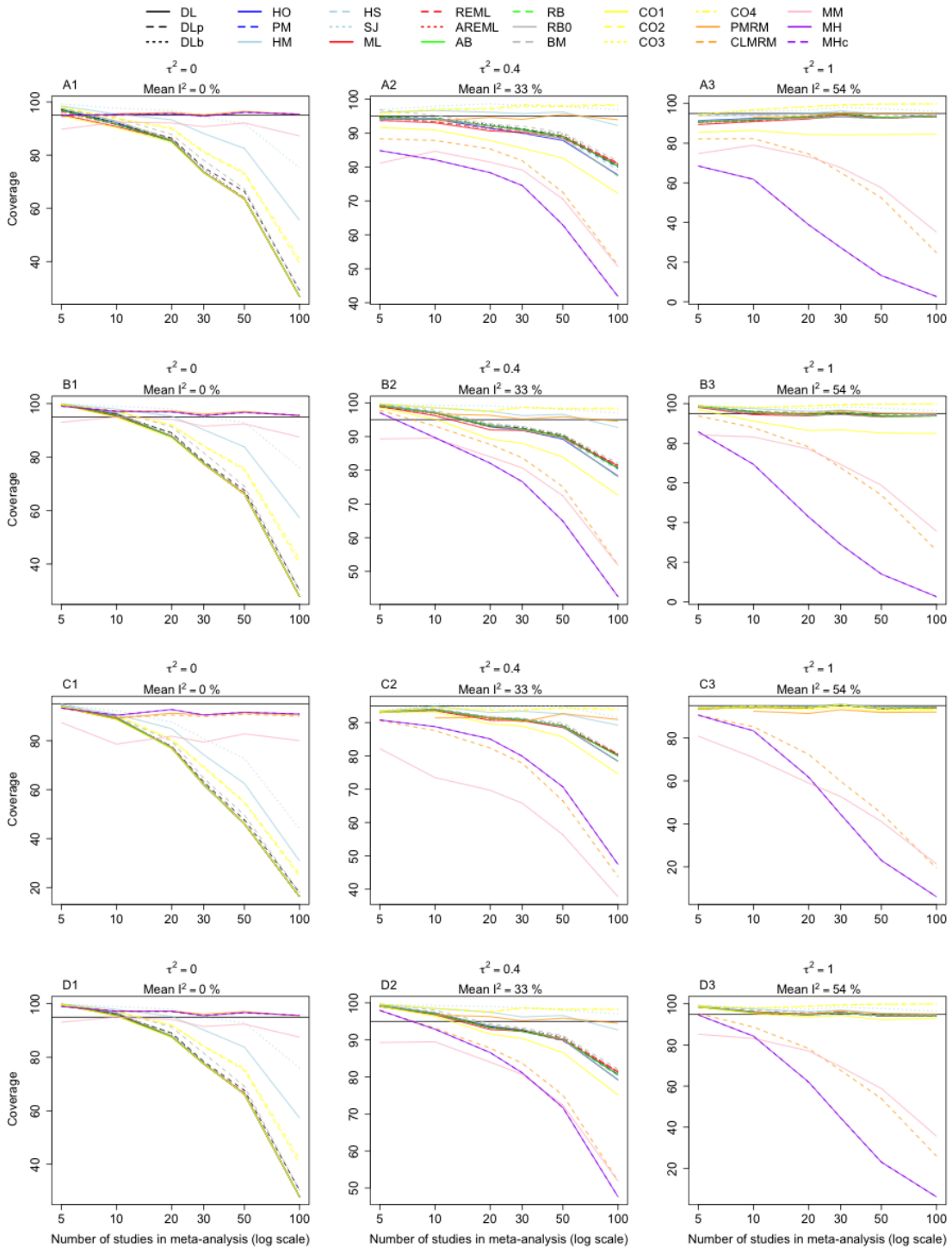


FIGURE E.124: Coverage of log-risk ratio confidence intervals in very rare events scenario with $p_0 < p_1$ and large sample sizes; confidence intervals are Wald-type (A1-A3), t-distribution (B1-B3), HKSJ (C1-C3) and mKH (D1-D3).

E.9.3 Common probability scenario

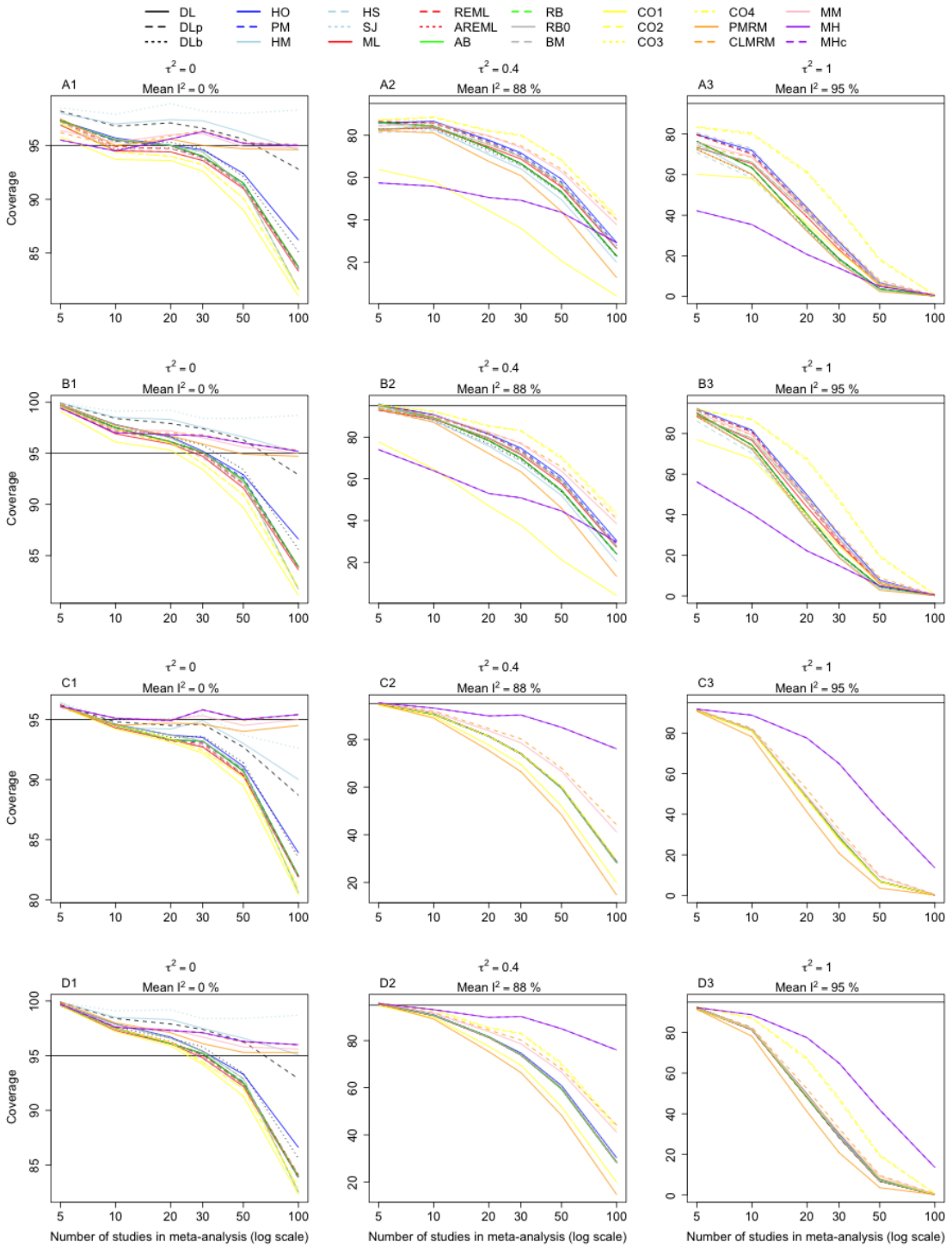


FIGURE E.125: Coverage of log-risk ratio confidence intervals in common probability scenario with $p_0 < p_1$ and medium sample sizes; confidence intervals are Wald-type (A1-A3), t-distribution (B1-B3), HKSJ (C1-C3) and mKH (D1-D3).

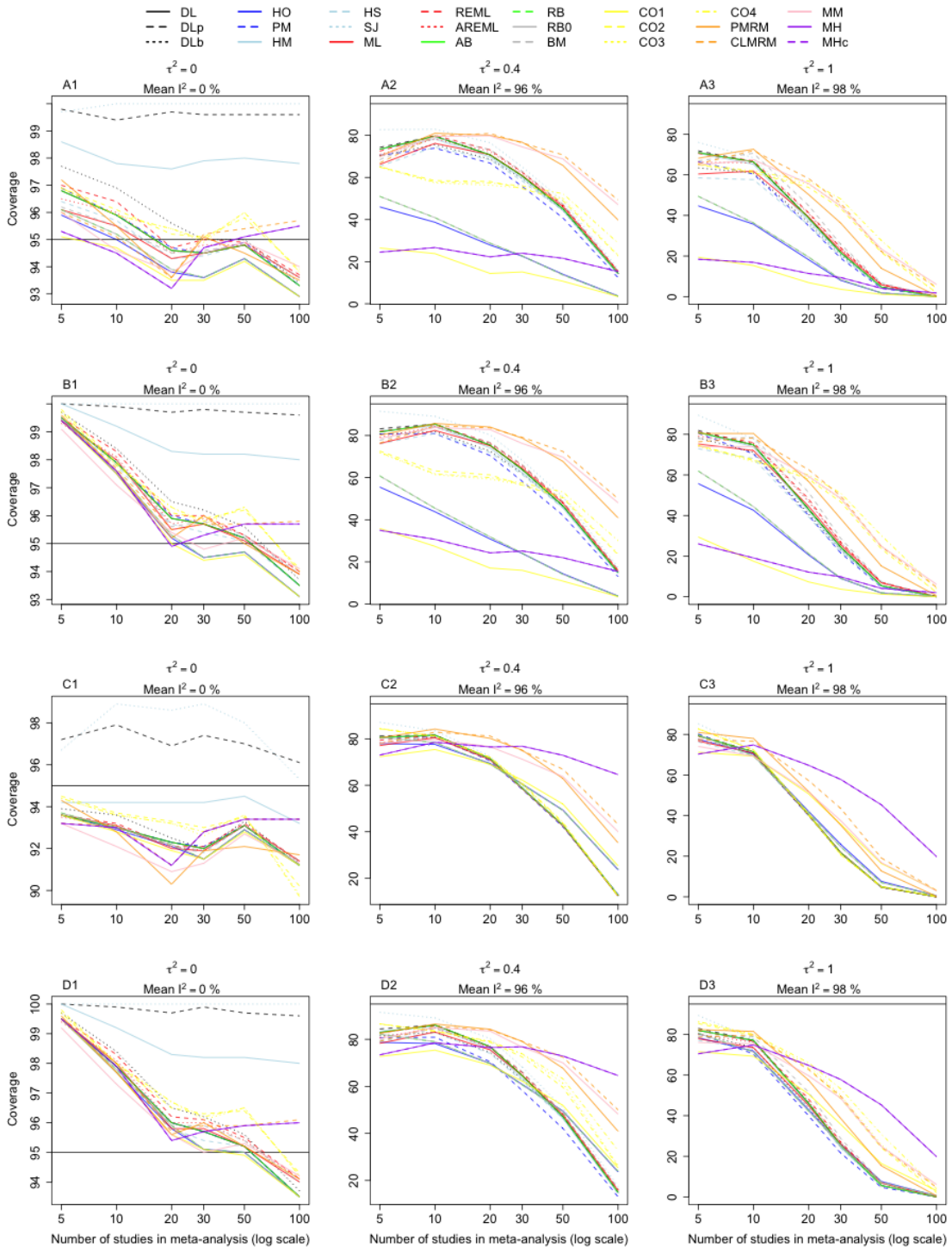


FIGURE E.126: Coverage of log-risk ratio confidence intervals in common probability scenario with $p_0 < p_1$ and small and large sample sizes; confidence intervals are Wald-type (A1-A3), t-distribution (B1-B3), HKSJ (C1-C3) and mKH (D1-D3).

E.10 Power

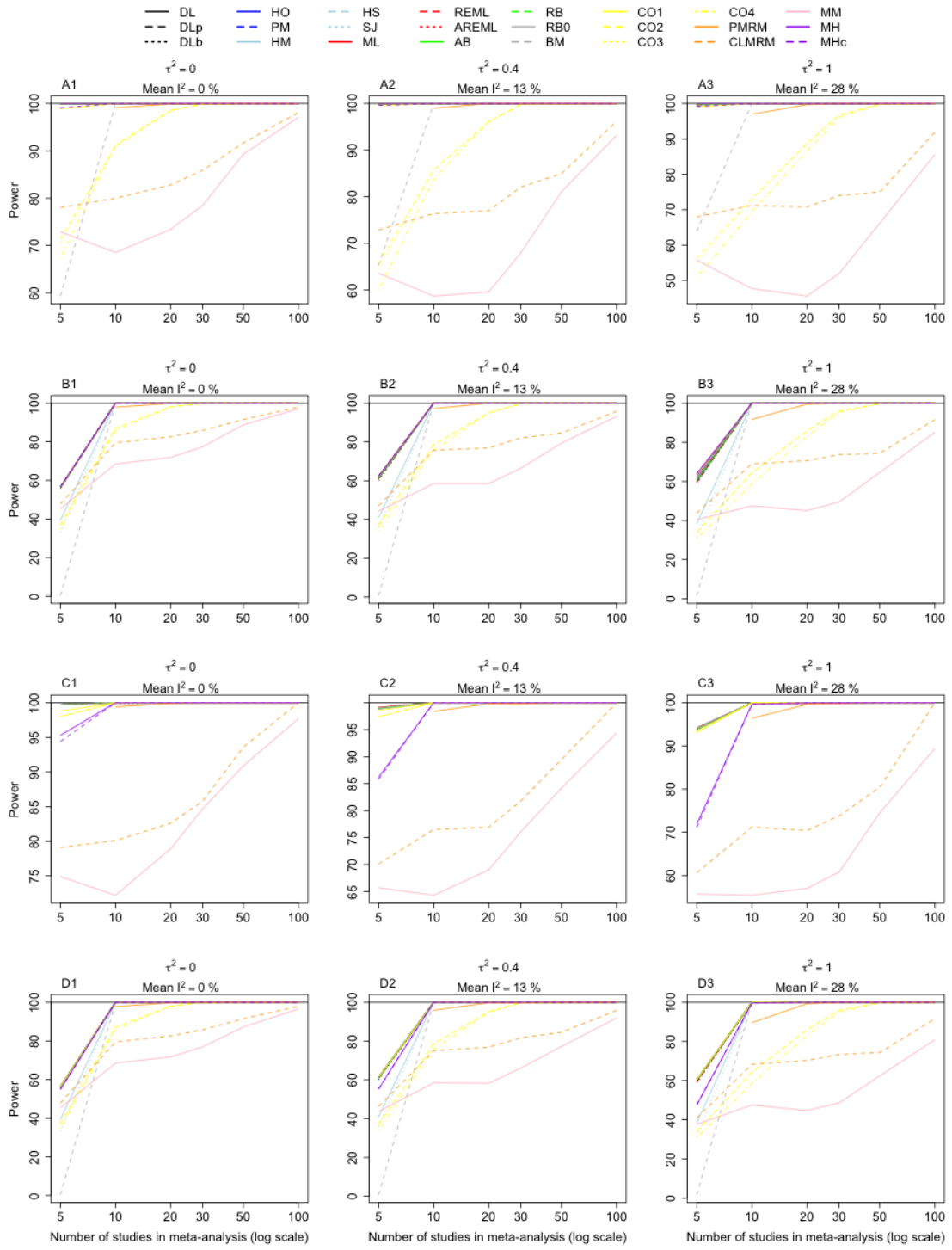


FIGURE E.127: Power of log-risk ratio confidence intervals in very rare events scenario with $p_0 < p_1$ and medium sample sizes; confidence intervals are Wald-type (A1-A3), t-distribution (B1-B3), HKSJ (C1-C3) and mKH (D1-D3).

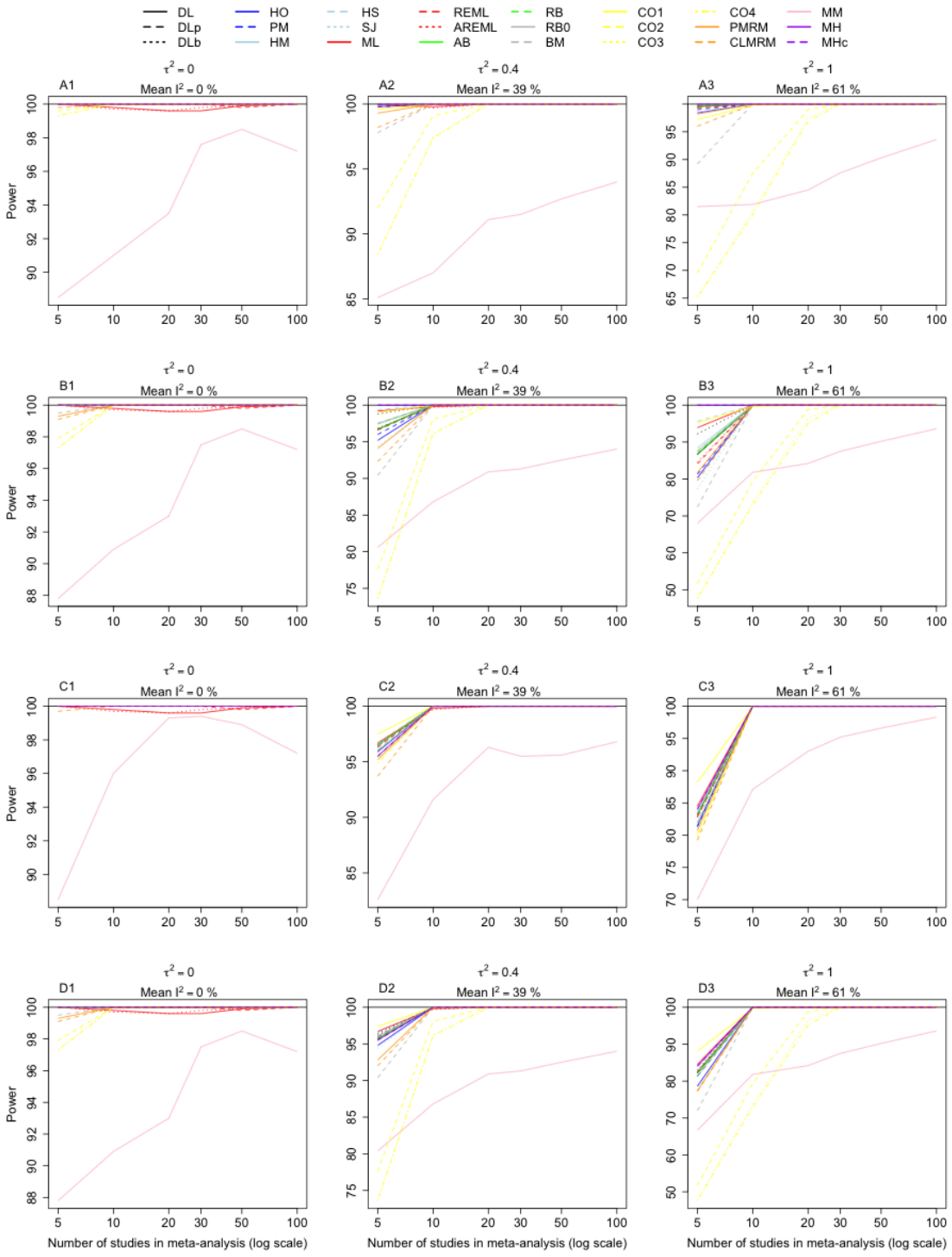


FIGURE E.128: Power of log-risk ratio confidence intervals in rare events scenario with $p_0 < p_1$ and medium sample sizes; confidence intervals are Wald-type (A1-A3), t-distribution (B1-B3), HKSJ (C1-C3) and mKH (D1-D3).

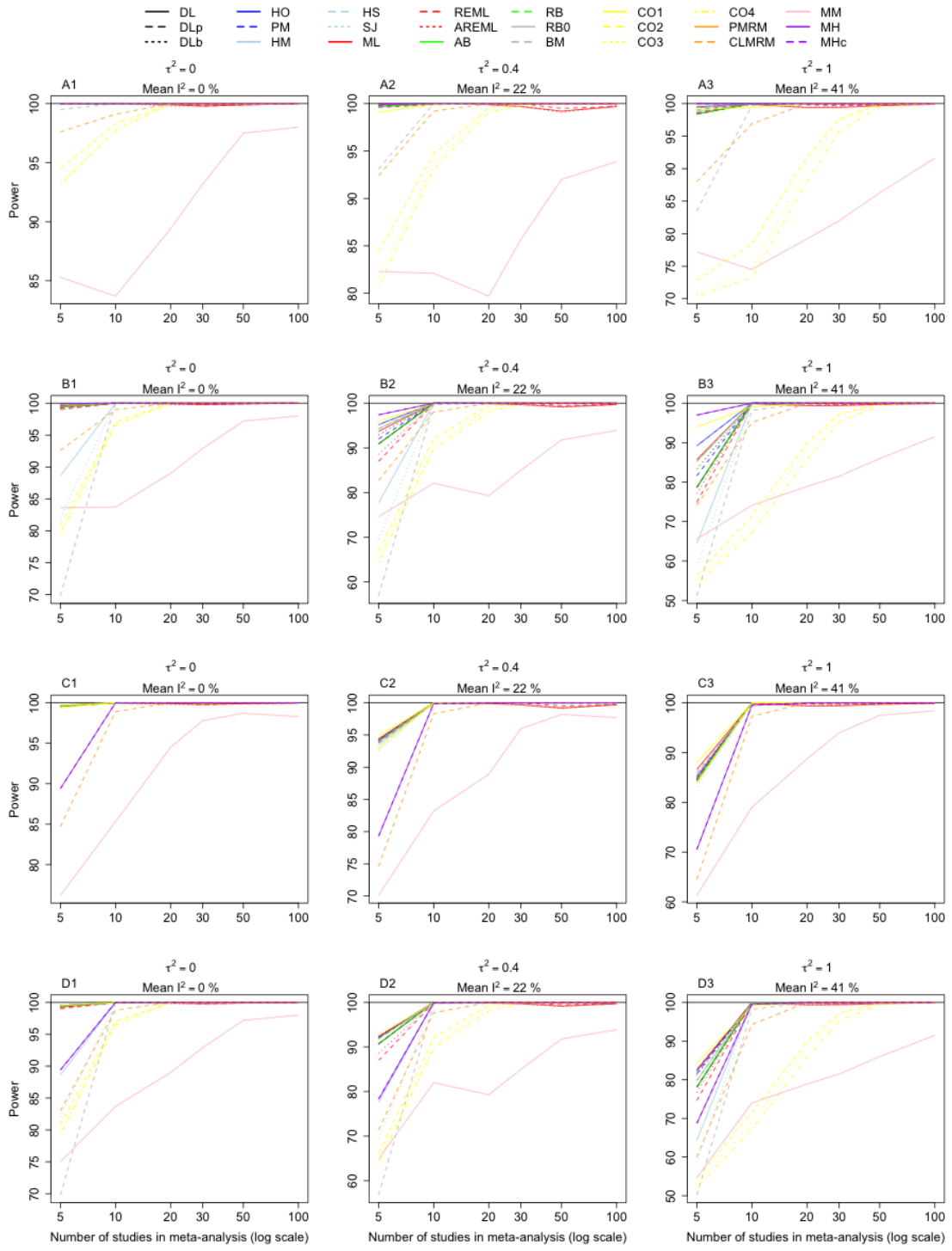


FIGURE E.129: Power of log-risk ratio confidence intervals in very rare events scenario with $p_0 < p_1$ and small and large sample sizes; confidence intervals are Wald-type (A1-A3), t-distribution (B1-B3), HKSJ (C1-C3) and mKH (D1-D3).

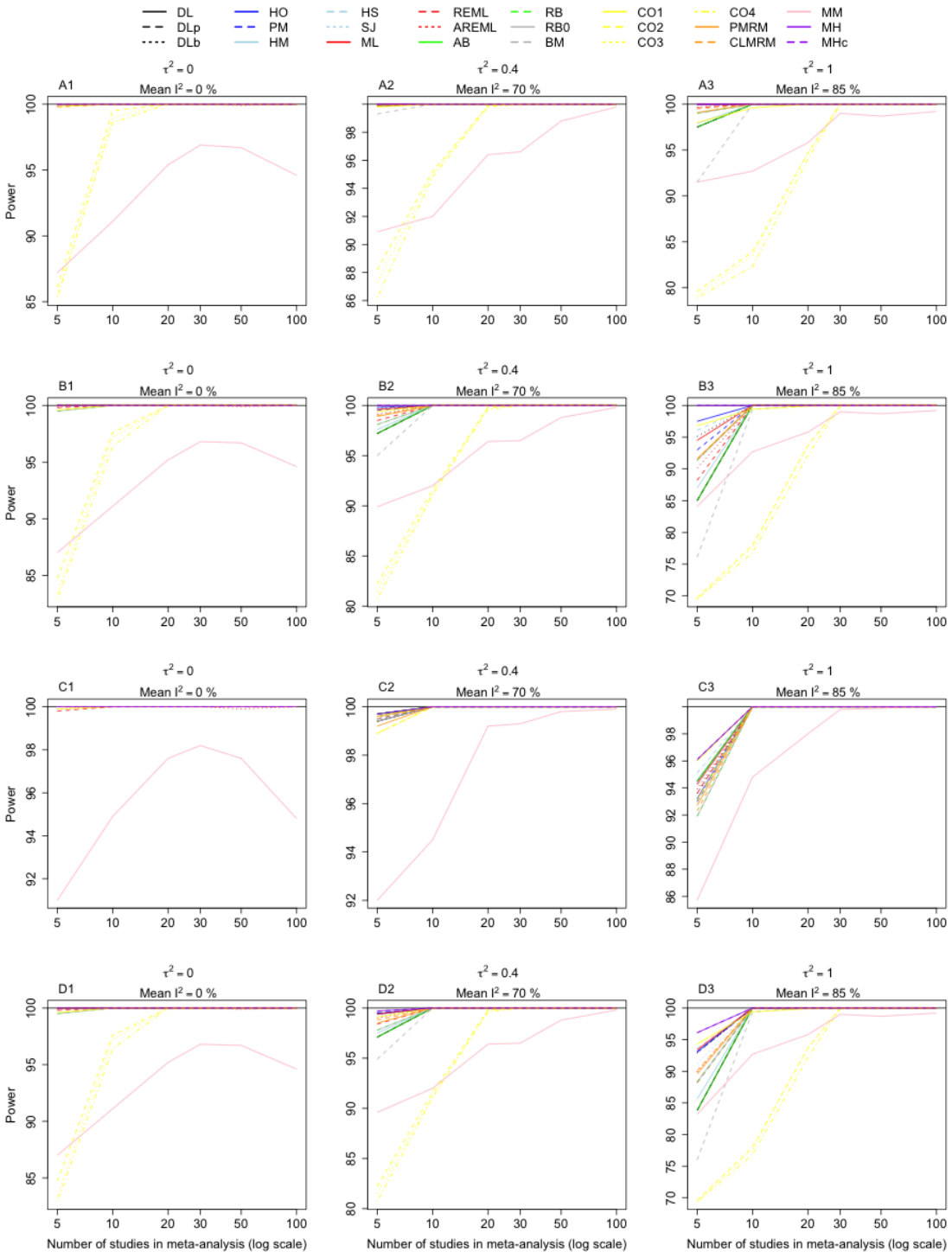


FIGURE E.130: Power of log-risk ratio confidence intervals in rare events scenario with $p_0 < p_1$ and small and large sample sizes; confidence intervals are Wald-type (A1-A3), t-distribution (B1-B3), HKSJ (C1-C3) and mKH (D1-D3).

E.11 Error

E.11.1 Mean error

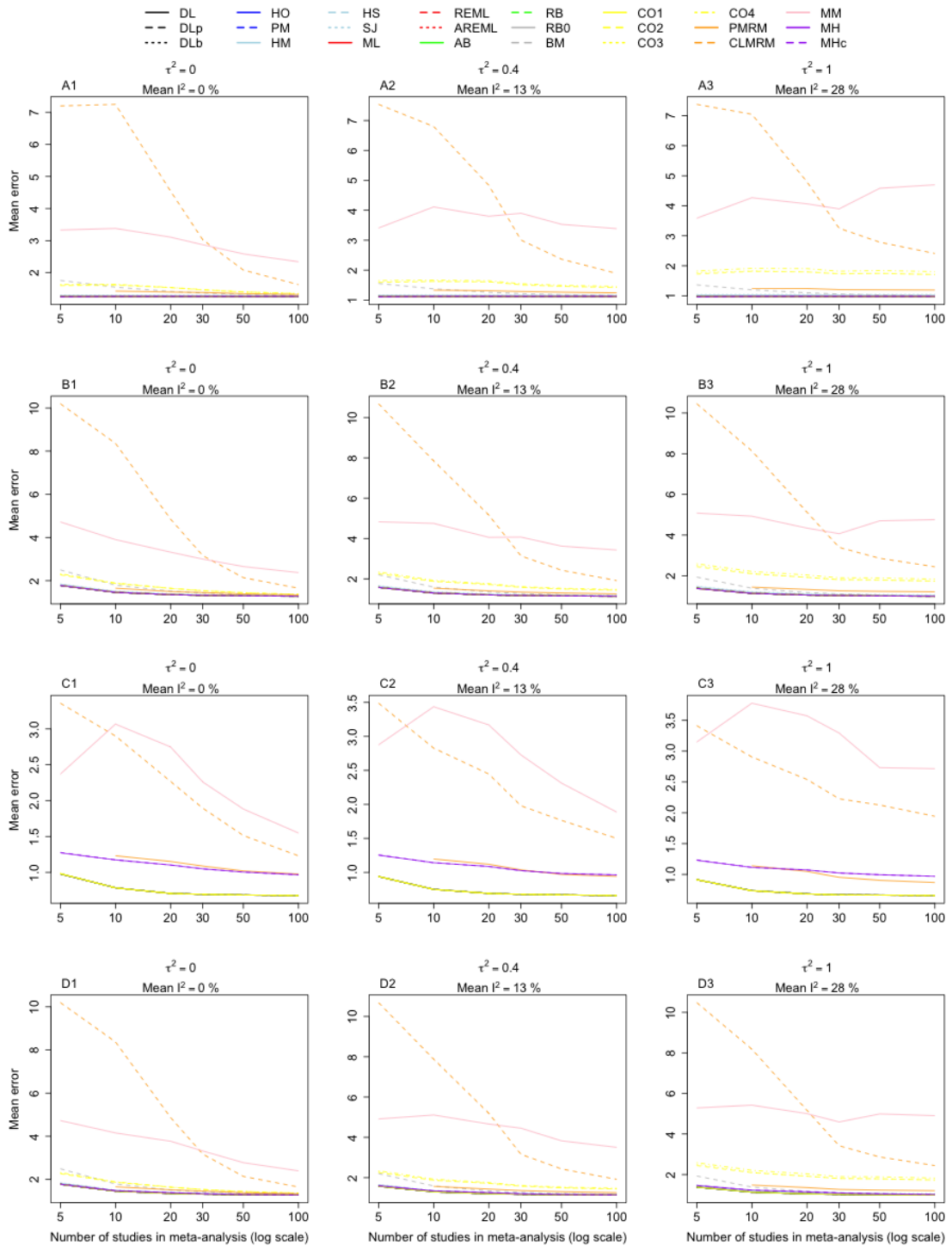


FIGURE E.131: Mean error of log-risk ratio confidence intervals in very rare events scenario with $p_0 < p_1$ and medium sample sizes; confidence intervals are Wald-type (A1-A3), t-distribution (B1-B3), HKSJ (C1-C3) and mKH (D1-D3).

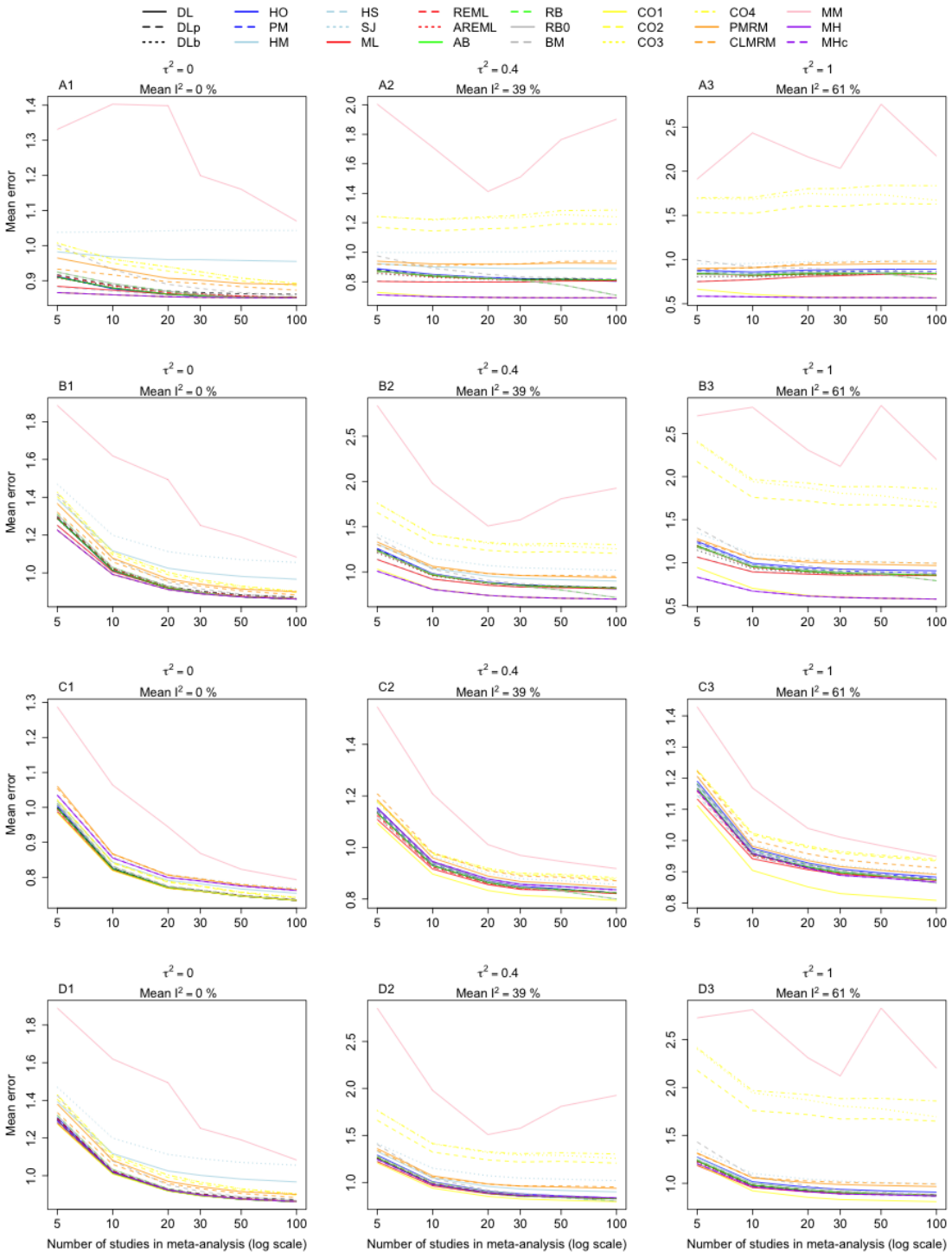


FIGURE E.132: Mean error of log-risk ratio confidence intervals in rare events scenario with $p_0 < p_1$ and medium sample sizes; confidence intervals are Wald-type (A1-A3), t-distribution (B1-B3), HKSJ (C1-C3) and mKH (D1-D3).

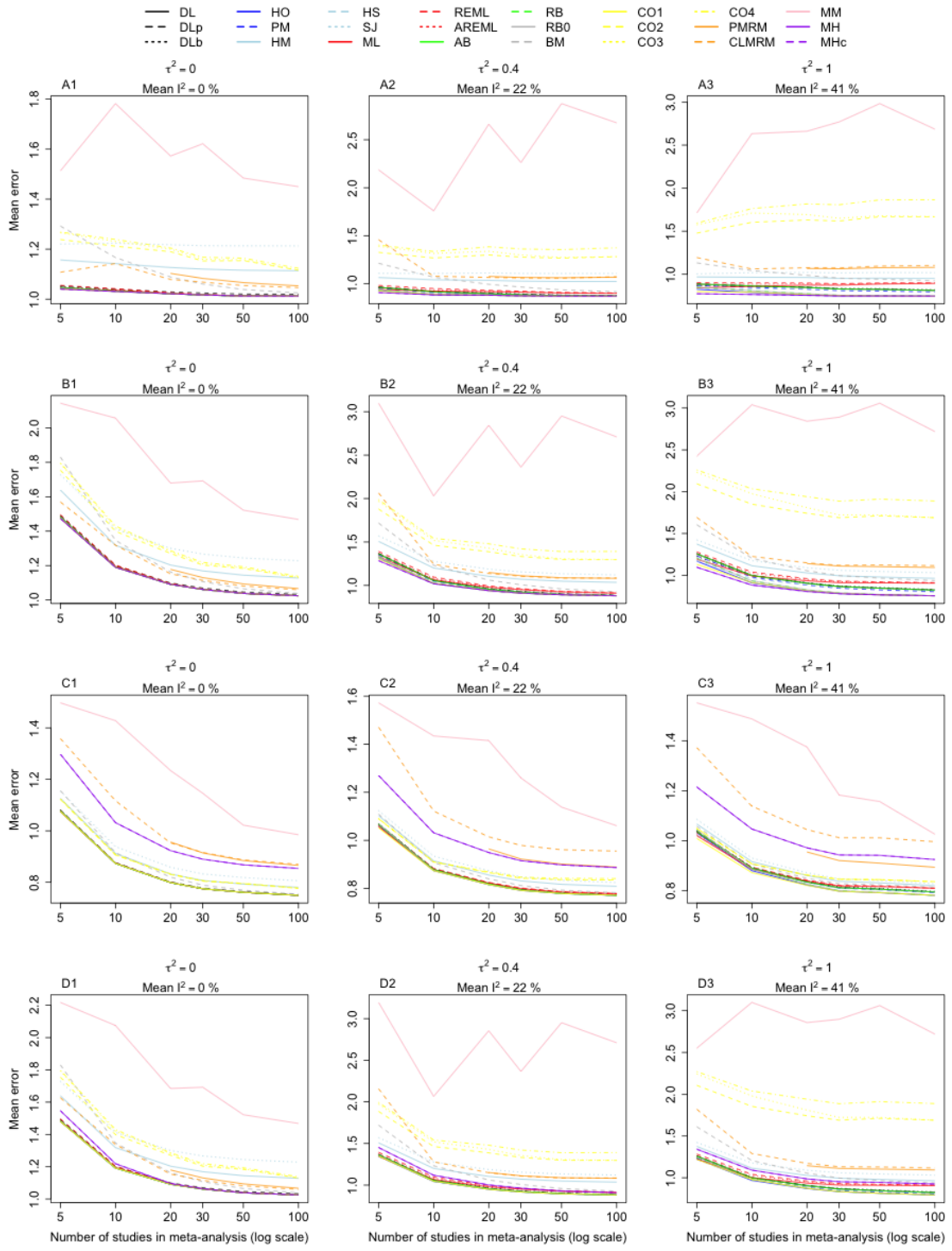


FIGURE E.133: Mean error of log-risk ratio confidence intervals in very rare events scenario with $p_0 < p_1$ and small and large sample sizes; confidence intervals are Wald-type (A1-A3), t-distribution (B1-B3), HKSJ (C1-C3) and mKH (D1-D3).

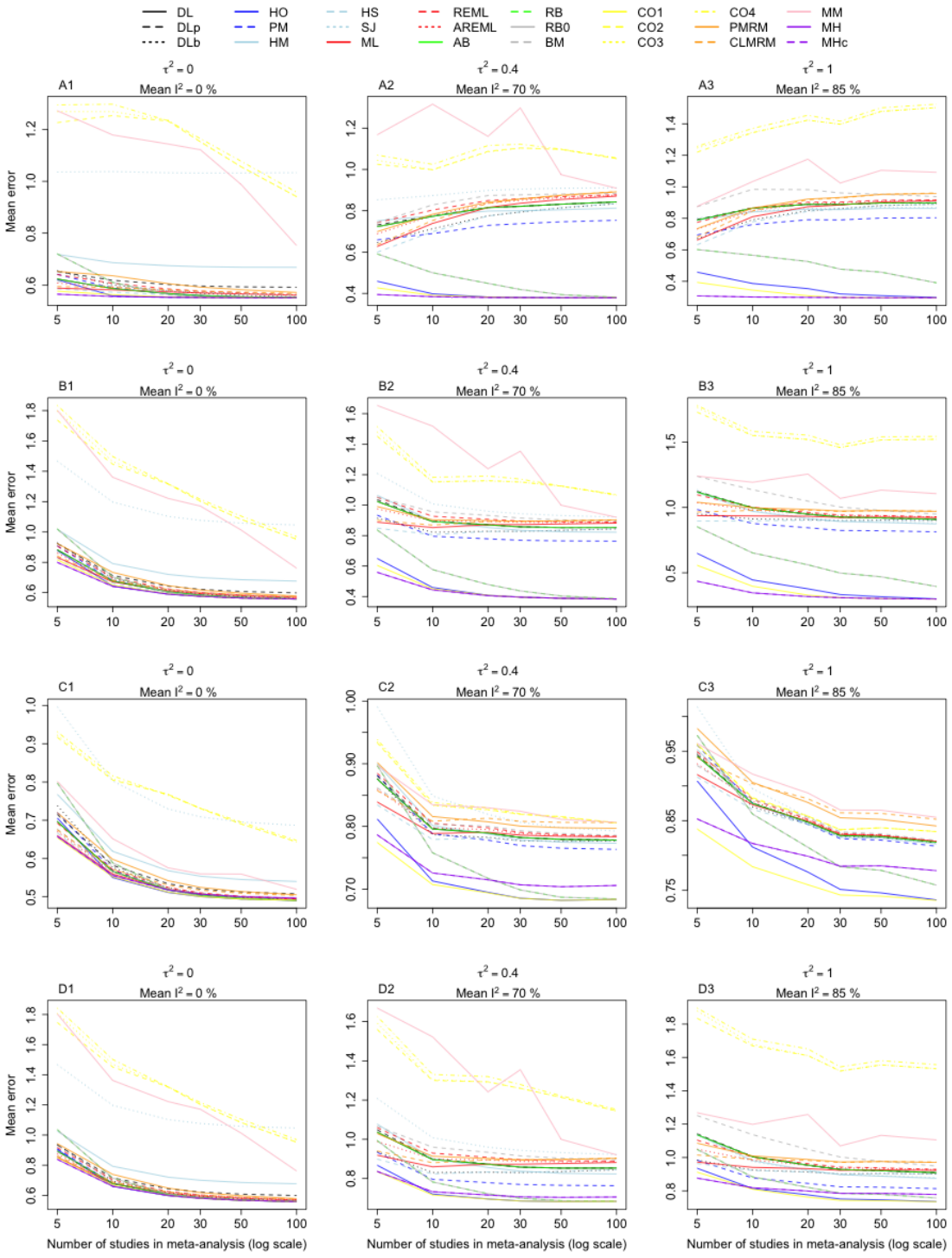


FIGURE E.134: Mean error of log-risk ratio confidence intervals in rare events scenario with $p_0 < p_1$ and small and large sample sizes; confidence intervals are Wald-type (A1-A3), t-distribution (B1-B3), HKSJ (C1-C3) and mKH (D1-D3).

E.11.2 Error variance

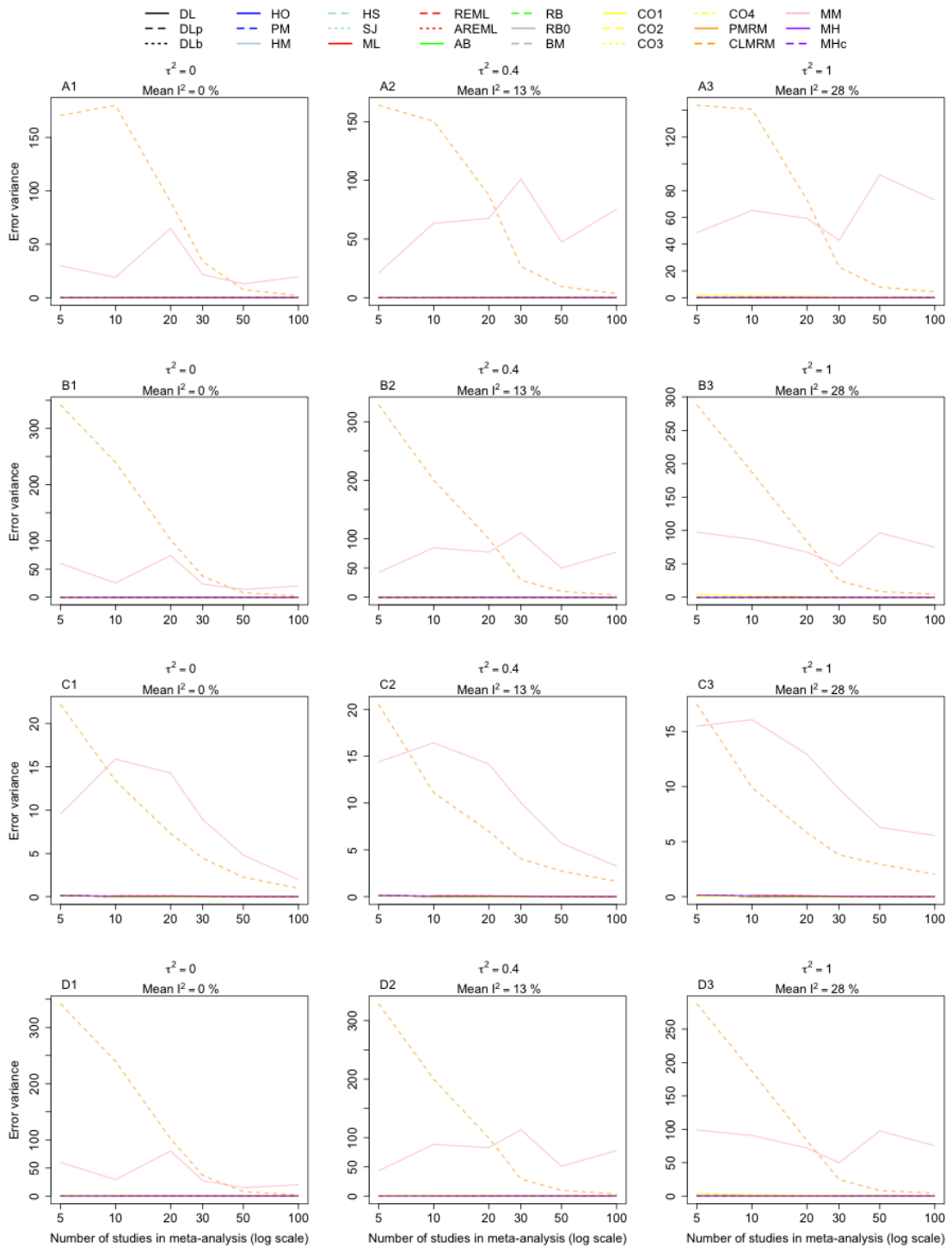


FIGURE E.135: Error variance of log-risk ratio confidence intervals in very rare events scenario with $p_0 < p_1$ and medium sample sizes; confidence intervals are Wald-type (A1-A3), t-distribution (B1-B3), HKSJ (C1-C3) and mKH (D1-D3).

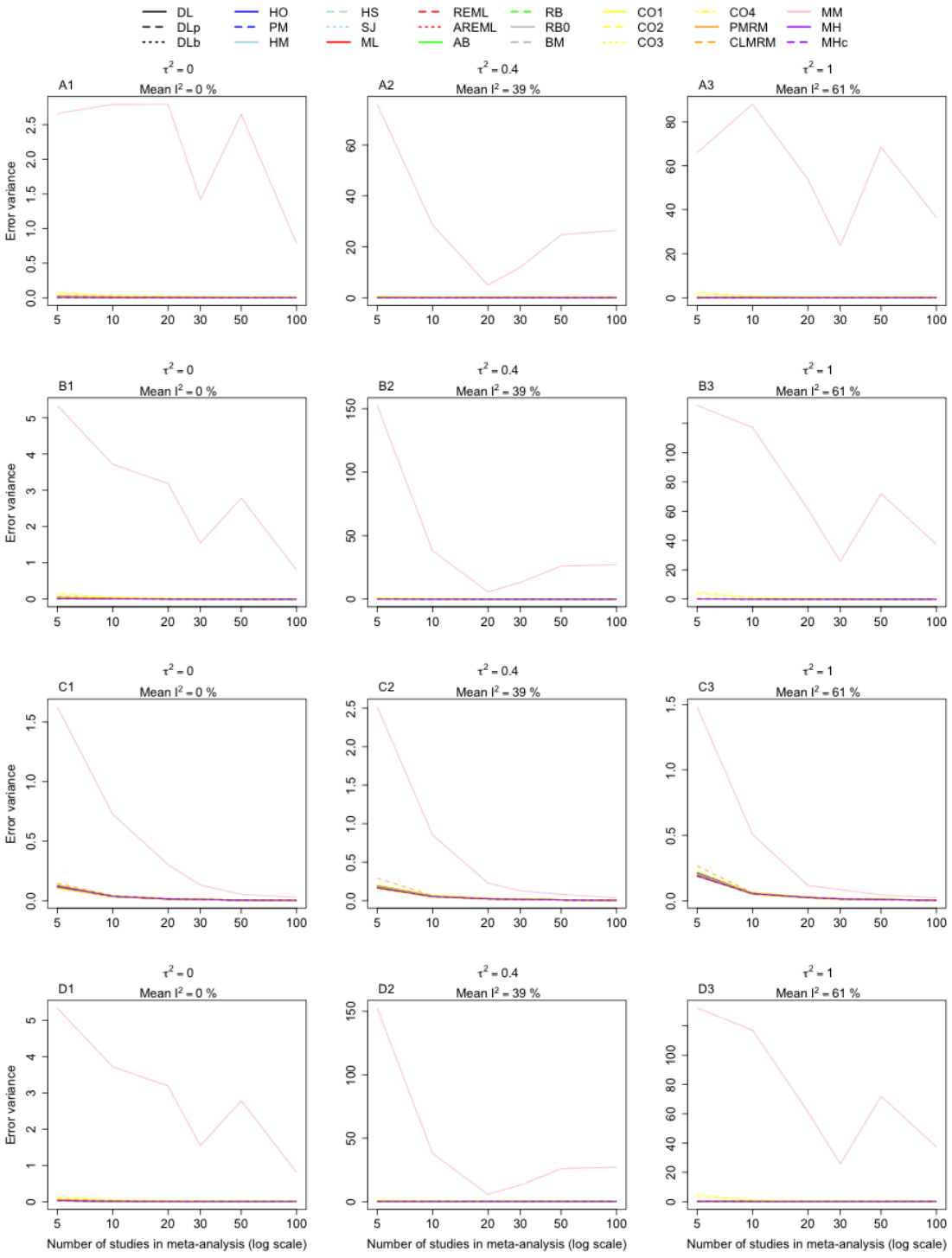


FIGURE E.136: Error variance of log-risk ratio confidence intervals in rare events scenario with $p_0 < p_1$ and medium sample sizes; confidence intervals are Wald-type (A1-A3), t-distribution (B1-B3), HKSJ (C1-C3) and mKH (D1-D3).

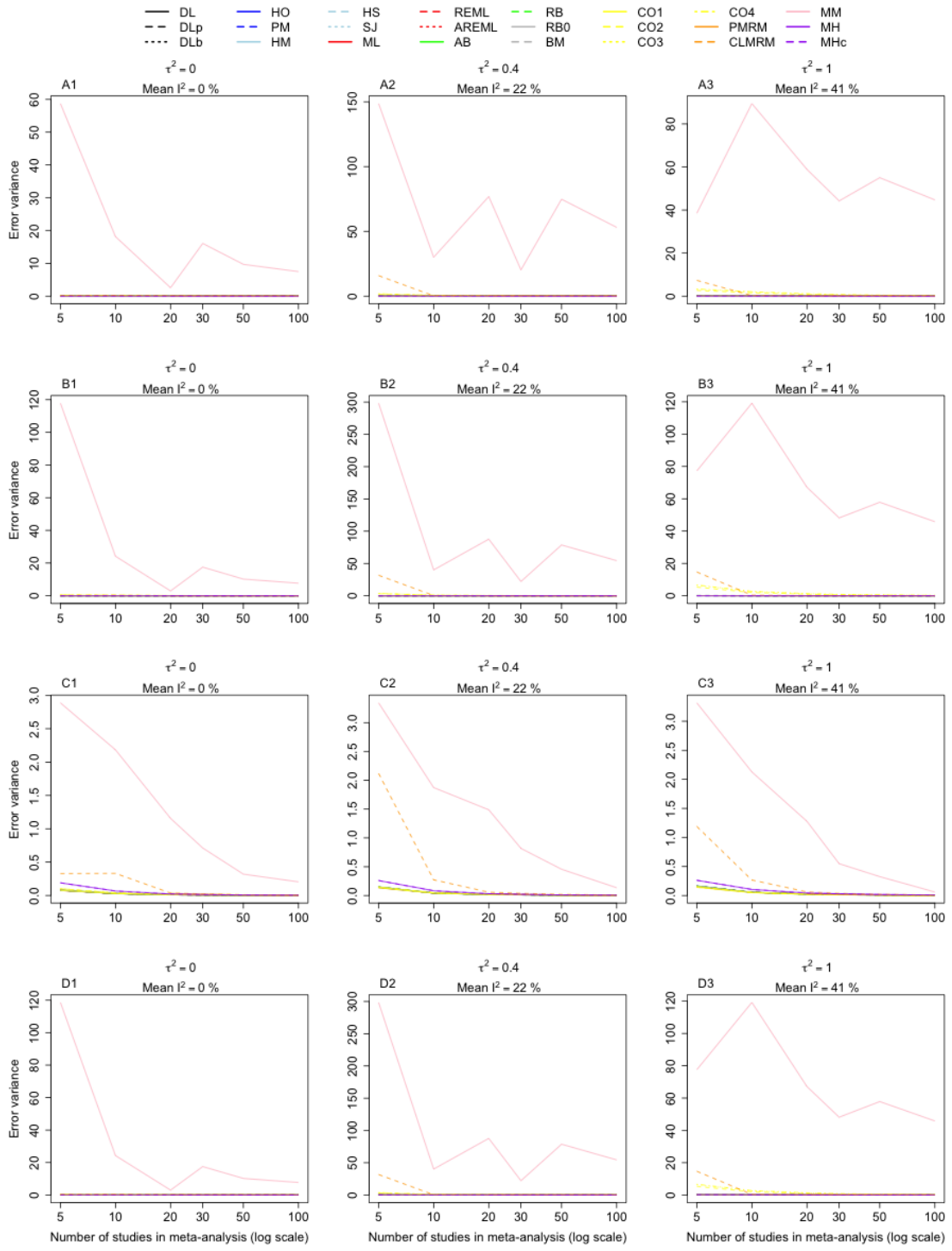


FIGURE E.137: Error variance of log-risk ratio confidence intervals in very rare events scenario with $p_0 < p_1$ and small and large sample sizes; confidence intervals are Wald-type (A1-A3), t-distribution (B1-B3), HKSJ (C1-C3) and mKH (D1-D3).

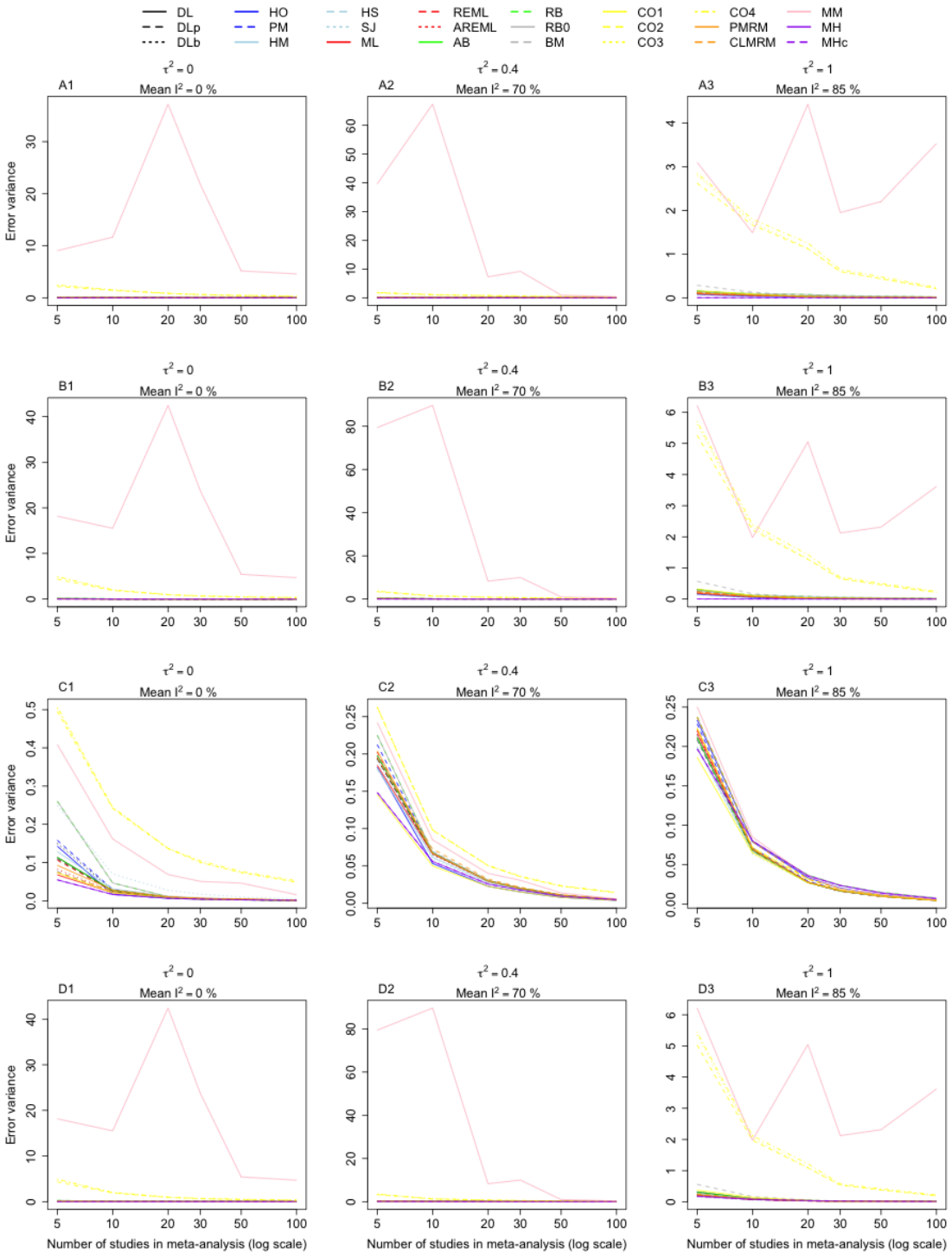


FIGURE E.138: Error variance of log-risk ratio confidence intervals in rare events scenario with $p_0 < p_1$ and small and large sample sizes; confidence intervals are Wald-type (A1-A3), t-distribution (B1-B3), HKSJ (C1-C3) and mKH (D1-D3).

Bibliography

Alderson, P., Bunn, F., Lefebvre, C., Li, W., Li, L., Roberts, I., and Schierhout, G. (2002). Human albumin solution for resuscitation and volume expansion in critically ill patients. *Cochrane Database of Systematic Reviews*, (1):CD001208.

Altman, D. G. and Deeks, J. J. (2002). Meta-analysis, Simpson's paradox, and the number needed to treat. *BMC Medical Research Methodology*, 2(1):3.

Andrade, C. (2015). Understanding relative risk, odds ratio, and related terms: as simple as it can get. *The Journal of Clinical Psychiatry*, 76(7):857–861.

Bakbergenuly, I. and Kulinskaya, E. (2018). Meta-analysis of binary outcomes via generalized linear mixed models: a simulation study. *BMC Medical Research Methodology*, 18(1):70.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Berkey, C. S., Hoaglin, D. C., Mosteller, F., and Colditz, G. A. (1995). A random-effects regression model for meta-analysis. *Statistics in Medicine*, 14(4):395–411.

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.

Böhning, D., Dietz, E., Schaub, R., Schlattmann, P., and Lindsay, B. G. (1994). The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Annals of the Institute of Statistical Mathematics*, 46(2):373–388.

Böhning, D., Dietz, E., and Schlattmann, P. (1998). Recent developments in C.A.MAN (computer assisted analysis of mixtures). *Biometrics*, 54:525–536.

Böhning, D., Mylona, K., and Kimber, A. (2015). Meta-analysis of clinical trials with rare events. *Biometrical Journal*, 57(4):633–648.

Böhning, D. and Sarol, J. (2000). A nonparametric estimator of heterogeneity variance with applications to SMR- and proportion-data. *Biometrical Journal*, 42(3):321–334.

Böhning, D., Schlattmann, P., and Lindsay, B. (1992). C.A.MAN (computer assisted analysis of mixtures): Statistical algorithms. *Biometrics*, 48:283–303.

Borenstein, M., Hedges, L. V., Higgins, J., and Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, 1(2):97–111.

Bowden, J., Tierney, J. F., Copas, A. J., and Burdett, S. (2011). Quantifying, displaying and accounting for heterogeneity in the meta-analysis of RCTs using standard and generalised Q statistics. *BMC Medical Research Methodology*, 11(1):41.

Bradburn, M. J., Deeks, J. J., Berlin, J. A., and Russell Localio, A. (2007). Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. *Statistics in Medicine*, 26(1):53–77.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25.

Breslow, N. E. and Day, N. (1987). *Statistical Methods in Cancer Research: The Design and Analysis of Cohort Studies*, volume 11. International Agency for Research on Cancer, Lyon.

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

Capanu, M., Gönen, M., and Begg, C. B. (2013). An assessment of estimation methods for generalized linear mixed models with binary outcomes. *Statistics in Medicine*, 32(26):4550–4566.

Chung, Y., Rabe-Hesketh, and Choi (2014). Avoiding zero between-study variance estimates in random-effects meta-analysis. *Statistics in Medicine*, 33(4):720–720.

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

Clayton, D., Hills, M., and Pickles, A. (1993). *Statistical Models in Epidemiology*, volume 161. Oxford University Press, Oxford.

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

Cohn, L. D. and Becker, B. J. (2003). How meta-analysis increases statistical power. *Psychological Methods*, 8(3):243.

Colditz, G. A., Brewer, T. F., Berkey, C. S., Wilson, M. E., Burdick, E., Fineberg, H. V., and Mosteller, F. (1994). Efficacy of BCG vaccine in the prevention of tuberculosis: meta-analysis of the published literature. *JAMA*, 271(9):698–702.

Cox, D. R. (2018). *Analysis of Binary Data*. Routledge.

Crins, N. D., Röver, C., Goralczyk, A. D., and Friede, T. (2014). Interleukin-2 receptor antagonists for pediatric liver transplant recipients: a systematic review and meta-analysis of controlled studies. *Pediatric Transplantation*, 18(8):839–850.

Deeks, J., Bradburn, M. J., Localio, R., and Berlin, J. (1999). Much ado about nothing: statistical methods for meta-analysis with rare events. Presented at 2nd Symposium on Systematic reviews: Beyond the basics, Oxford. Abstract available at `http://www.ihs.ox.ac.uk/csm/talks.html#p23`.

Deeks, J. J., Altman, D. G., Bradburn, M. J., et al. (2001). Statistical methods for examining heterogeneity and combining results from several studies in meta-analysis. *Systematic Reviews in Health Care: Meta-Analysis in Context*, 2:285–312.

Demidenko, E. (2013). *Mixed Models: Theory and Applications with R*. John Wiley & Sons.

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

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

Deshpande, A. D., Harris-Hayes, M., and Schootman, M. (2008). Epidemiology of diabetes and diabetes-related complications. *Physical Therapy*, 88(11):1254.

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.

Egger, M., Smith, G. D., and Phillips, A. N. (1997). Meta-analysis: principles and procedures. *BMJ*, 315(7121):1533–1537.

Faraone, S. V. (2008). Interpreting estimates of treatment effects: implications for managed care. *Pharmacy and Therapeutics*, 33(12):700.

Follmann, D. A. and Proschan, M. A. (1999). Valid inference in random effects meta-analysis. *Biometrics*, 55(3):732–737.

Friede, T., Röver, C., Wandel, S., and Neuenschwander, B. (2017a). Meta-analysis of few small studies in orphan diseases. *Research Synthesis Methods*, 8(1):79–91.

Friede, T., Röver, C., Wandel, S., and Neuenschwander, B. (2017b). Meta-analysis of two studies in the presence of heterogeneity with applications in rare diseases. *Biometrical Journal*, 59(4):658–671.

Gelman, A. et al. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1(3):515–534.

Goldstein, H. and Rasbash, J. (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, pages 505–513.

Goldwasser, P. and Feldman, J. (1997). Association of serum albumin and mortality risk. *Journal of Clinical Epidemiology*, 50(6):693–703.

Greenland, S. and Salvan, A. (1990). Bias in the one-step method for pooling study results. *Statistics in Medicine*, 9(3):247–252.

Günhan, B. K., Röver, C., and Friede, T. (2018). Meta-analysis of few studies involving rare events. *arXiv preprint arXiv:1809.04407*.

Hamza, T. H., van Houwelingen, H. C., and Stijnen, T. (2008). The binomial distribution of meta-analysis was preferred to model within-study variability. *Journal of Clinical Epidemiology*, 61(1):41–51.

Handoll, H. H., Gillespie, W. J., Gillespie, L. D., Madhok, R., et al. (2008). The Cochrane Collaboration: a leading role in producing reliable evidence to inform healthcare decisions in musculoskeletal trauma and disorders. *Indian Journal of Orthopaedics*, 42(3):247.

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

Hartung, J. (1998). An alternative method for meta-analysis. Technical Report, Universitätsbibliothek Dortmund.

Hartung, J. and Knapp, G. (2001). A refined method for the meta-analysis of controlled clinical trials with binary outcome. *Statistics in Medicine*, 20(24):3875–3889.

Hartung, J. and Makambi, K. H. (2003). Reducing the number of unjustified significant results in meta-analysis. *Communications in Statistics-Simulation and Computation*, 32(4):1179–1190.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, 6(2):107–128.

Hedges, L. V. and Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. Academic Press.

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

Higgins, J. P. and Green, S. (2011). *Cochrane Handbook for Systematic Reviews of Interventions*, volume 4. John Wiley & Sons.

Hofmeyr, G. J. and Smaill, F. M. (2002). Antibiotic prophylaxis for cesarean section. *The Cochrane Library*.

IntHout, J., Ioannidis, J. P., and Borm, G. F. (2014). The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Medical Research Methodology*, 14(1):1.

Jackson, D., Law, M., Stijnen, T., Viechtbauer, W., and White, I. R. (2018). A comparison of seven random-effects models for meta-analyses that estimate the summary odds ratio. *Statistics in Medicine*, 37(7):1059–1085.

Jewell, N. P. and Holford, T. R. (2005). *Statistics for Epidemiology*. JSTOR.

Kacker, R. N. (2004). Combining information from interlaboratory evaluations using a random effects model. *Metrologia*, 41(3):132.

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.

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.

Kuss, O. (2015). Statistical methods for meta-analyses including information from studies without any events - add nothing to nothing and succeed nevertheless. *Statistics in Medicine*, 34(7):1097–1116.

Lambert, P. C., Sutton, A. J., Burton, P. R., Abrams, K. R., and Jones, D. R. (2005). How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Statistics in Medicine*, 24(15):2401–2428.

Langan, D. (2015). *Estimating the Heterogeneity Variance in a Random-Effects Meta-Analysis*. PhD thesis, University of York.

Langan, D., Higgins, J., and Simmonds, M. (2016). Comparative performance of heterogeneity variance estimators in meta-analysis: a review of simulation studies. *Research Synthesis Methods*.

Langan, D., Higgins, J. P., Jackson, D., Bowden, J., Veroniki, A. A., Kontopantelis, E., Viechtbauer, W., and Simmonds, M. (2018). A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. *Research Synthesis Methods*, pages 1–16.

Lewallen, S. and Courtright, P. (1998). Epidemiology in practice: case-control studies. *Community Eye Health*, 11(28):57.

Liu, Q. and Pierce, D. A. (1993). Heterogeneity in Mantel-Haenszel-type models. *Biometrika*, 80(3):543–556.

Mallett, S. and Clarke, M. (2002). The typical Cochrane review: How many trials? How many participants? *International Journal of Technology Assessment in Health Care*, 18(4):820–823.

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.

Mantel, N. (1963). Chi-square tests with one degree of freedom; extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association*, 58(303):690–700.

Mohanna, K. and Chambers, R. (2008). Risk matters in healthcare: communicating, explaining and managing risk.

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

National Institute for Clinical Excellence et al. (2005). Guideline development methods: information for national collaborating centres and guideline developers. *London: NICE*.

Niel-Weise, B., Stijnen, T., and Van den Broek, P. (2007). Anti-infective-treated central venous catheters: a systematic review of randomized controlled trials. *Intensive Care Medicine*, 33(12):2058–2068.

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.

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.

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.

Panityakul, T., Bumrungsup, C., and Knapp, G. (2013). On estimating residual heterogeneity in random-effects meta-regression: a comparative study. *Journal of Statistical Theory and Applications*, 12(3):253.

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.

Peto, R. and Peto, J. (1972). Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society. Series A (General)*, pages 185–207.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raudenbush, S. W. (2009). Analyzing effect sizes: Random-effects models. *The Handbook of Research Synthesis and Meta-analysis*, 2:295–316.

Reviewers, C. I. G. A. (1998). Human albumin administration in critically ill patients: systematic review of randomised controlled trials. *BMJ: British Medical Journal*, pages 235–240.

Richardson, S. and Green, P. J. (1997). On bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: series B (statistical methodology)*, 59(4):731–792.

Ross, S. M. (2014). *Introduction to Probability Models*. Academic Press.

Röver, C., Knapp, G., and Friede, T. (2015). Hartung-Knapp-Sidik-Jonkman approach and its modification for random-effects meta-analysis with few studies. *BMC Medical Research Methodology*, 15(1):99.

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

Schlattmann, P. (2009). *Medical Applications of Finite Mixture Models*. Springer.

Schlattmann, P., Hoehne, J., and Verba, M. (2016). *CAMAN: Finite Mixture Models and Meta-Analysis Tools - Based on C.A.MAN*. R package version 0.74.

Schmidt, F. L. and Hunter, J. E. (2014). *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. Sage Publications.

Scotet, V., Duguépéroux, I., Saliou, P., Rault, G., Roussey, M., Audrézet, M.-P., and Férec, C. (2012). Evidence for decline in the incidence of cystic fibrosis: a 35-year observational study in Brittany, France. *Orphanet Journal of Rare Diseases*, 7(1):14.

Setia, M. S. (2016). Methodology series module 3: Cross-sectional studies. *Indian Journal of Dermatology*, 61(3):261.

Shen, L. Q., Child, A., Weber, G. M., Folkman, J., and Aiello, L. P. (2008). Rosiglitazone and delayed onset of proliferative diabetic retinopathy. *Archives of Ophthalmology*, 126(6):793–799.

Sibbald, B. and Roland, M. (1998). Understanding controlled trials. why are randomised controlled trials important? *BMJ: British Medical Journal*, 316(7126):201.

Sidik, K. and Jonkman, J. N. (2002). A simple confidence interval for meta-analysis. *Statistics in Medicine*, 21(21):3153–3159.

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.

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.

Sidik, K. and Jonkman, J. N. (2008). Estimation using non-central hypergeometric distributions in combining 2× 2 tables. *Journal of Statistical Planning and Inference*, 138(12):3993–4005.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 238–241.

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.

Song, J. W. and Chung, K. C. (2010). Observational studies: cohort and case-control studies. *Plastic and Reconstructive Surgery*, 126(6):2234.

StataCorp, L. (2013). Stata multilevel mixed-effects reference manual. *College Station, TX: StataCorp LP*.

Stijnen, T., Hamza, T. H., and Özdemir, P. (2010). Random effects meta-analysis of event outcome in the framework of the generalized linear mixed model with applications in sparse data. *Statistics in Medicine*, 29(29):3046–3067.

Sutton, A. J. and Abrams, K. R. (2001). Bayesian methods in meta-analysis and evidence synthesis. *Statistical Methods in Medical Research*, 10(4):277–303.

Sweeting, M. J., Sutton, A. J., and Lambert, P. C. (2004). What to add to nothing? use and avoidance of continuity corrections in meta-analysis of sparse data. *Statistics in Medicine*, 23(9):1351–1375.

Szumilas, M. (2010). Explaining odds ratios. *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, 19(3):227.

Takeshima, N., Sozu, T., Tajika, A., Ogawa, Y., Hayasaka, Y., and Furukawa, T. A. (2014). Which is more generalizable, powerful and interpretable in meta-analyses, mean difference or standardized mean difference? *BMC Medical Research Methodology*, 14(1):30.

Taylor, L. E., Swerdfeger, A. L., and Eslick, G. D. (2014). Vaccines are not associated with autism: an evidence-based meta-analysis of case-control and cohort studies. *Vaccine*, 32(29):3623–3629.

Thompson, S. G. (1994). Why sources of heterogeneity in meta-analysis should be investigated. *BMJ: British Medical Journal*, 309(6965):1351.

Thompson, S. G. and Higgins, J. P. (2002). How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine*, 21(11):1559–1573.

Thompson, S. G. and Sharp, S. J. (1999). Explaining heterogeneity in meta-analysis: a comparison of methods. *Statistics in Medicine*, 18(20):2693–2708.

Thorlund, K., Wetterslev, J., Awad, T., Thabane, L., and Gluud, C. (2011). Comparison of statistical inferences from the DerSimonian–Laird and alternative random-effects model meta-analyses–an empirical assessment of 920 Cochrane primary outcome meta-analyses. *Research Synthesis Methods*, 2(4):238–253.

Turner, R. M., Omar, R. Z., Yang, M., Goldstein, H., and Thompson, S. G. (2000). A multilevel model framework for meta-analysis of clinical trials with binary outcomes. *Statistics in Medicine*, 19(24):3417–3432.

Uman, L. S. (2011). Systematic reviews and meta-analyses. *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, 20(1):57.

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

Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., Kuss, O., Higgins, J. P., 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.

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.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3):1–48.

Walwyn, R. and Roberts, C. (2017). Meta-analysis of standardised mean differences from randomised trials with treatment-related clustering associated with care providers. *Statistics in medicine*, 36(7):1043–1067.

Whitehead, A. and Whitehead, J. (1991). A general parametric approach to the meta-analysis of randomized clinical trials. *Statistics in Medicine*, 10(11):1665–1677.

Wiksten, A., Rücker, G., and Schwarzer, G. (2016). Hartung–Knapp method is not always conservative compared with fixed-effect meta-analysis. *Statistics in Medicine*, 35(15):2503–2515.

Zhou, X.-H., Brizendine, E., and Pritz, M. (1999). Methods for combining rates from several studies. *Statistics in Medicine*, 18(5):557–566.