# Robust Variance Estimation to Model Complex Meta-Analytic Structures in the Immediate Antihypertensive Effects of Exercise

Magdalene Mlynek

University of Connecticut, Statistics

June 2020

*Elizabeth D. Schifano, PhD., Department of Statistics*

*Blair T. Johnson, Ph.D., Department of Psychological Sciences*

*Linda S. Pescatello, PhD., Department of Kinesiology*

## Table of Contents

## ACKNOWLEDGMENTS

I would like to briefly thank the many people who have supported my ambitions and helped me throughout the entirety of this project. Without their guidance and expertise, this would not have been possible. I would like to express my great appreciation to Dr. Schifano for her unwavering guidance and patience throughout the project. She is an excellent advisor, and has been a tremendous mentor in beginning my career in statistics. I'd also like to extend my sincerest thanks to Dr. Pescatello for her practical suggestions in applying the methods and results to the clinician's perspective. I'm also extremely grateful to Dr. Johnson and Dr. Hennessey for the extensive knowledge of meta-analysis that they brought to the project, as well as their invaluable insight they had in identifying ways to apply Robust Variance Estimation meta-analysis to real-life data. In addition, I'd like to extend my appreciation to Dr. Wu, for his guidance and support in understanding previous meta-analysis methods used in Post-Exercise Hypotension research. Lastly, I'd like to thank Dr. James Pustejovsky at the University of Texas Austin for his detailed recommendations on our simulation.

## ABSTRACT

**Objective:** To determine and compare performance of the classical random effects and Robust Variance Estimation meta-analysis methods among studies with dependencies and moderators. We will apply these findings through the lens of the immediate blood pressure response to exercise termed, *postexercise hypotension* (PEH) and provide practical solutions for modeling common dependencies. **Methods:** Through the use of a Monte Carlo simulation, we created several simulation scenarios modeling a variety of common settings in PEH studies, including repeated measurements within the same subject. We compared the estimation and inference results of the Robust Variance Estimation (RVE) method to three variations of classical random effects meta-analysis and meta-regression with a time moderator: average effect across all time points ("Average"), one randomly selected observation ("Random") across time points, and all observations across all time points ("Naïve). **Results:** Overall, RVE method performs best in terms of: 1) reducing bias of the overall and moderator effects; 2) correctly estimating coverage probability containing the true effect; 3) having a high power in detecting a significant non-zero overall or moderator effect; and 4) most accurately estimating the true $\tau^2$ value. Under a few specific simulation settings, the classical methods performed just as well as, or better than the RVE method in some performance metrics when there was no time moderator effect. **Conclusions:** The RVE method performs best overall across the simulation scenarios considered and is recommended for PEH studies with time moderators and various levels of dependencies of effect sizes, along with other types of studies.

## SECTION 1: INTRODUCTION

As an alternative to antihypertensive medication, many physicians are prescribing exercise to treat hypertension[1,2]. Indeed, two recent network meta-analyses have concluded that exercise is as effective as antihypertensive medication in lowering blood pressure (BP) and frequent exercise lowers BP on average by 5-8 mmHg[34]. The 2018 Physical Activity Guidelines state that "about 90 minutes a week of the equivalent amount of vigorous intensity activity helps to substantially lower the risk of heart disease"[4]. While the Guidelines and supporting research shows that frequent exercise is an effective approach to prevent, treat, and control hypertension, to date there are few meta-analyses of randomized controlled trials on the immediate BP effects of a single session of resistance and aerobic exercise combined or concurrent exercise, termed *postexercise hypotension*.

Typically, meta-analyses on the more long-term or chronic BP benefits of exercise training utilize traditional methods relying on statistical independence to determine the overall effect size. However, many of the studies included within these meta-analyses have study designs that contain multiple arms that cause an increased dependency between the BP observations, as well as multiple dependent effect sizes measured over time. To overcome these dependency issues, authors commonly aggregate to a single effect size or select one and drop others, which reduces the ability to understand the heterogeneity of effects. An alternative option is to use a method that appropriately accounts for the dependencies, such as the Robust Variance Estimate (RVE)[5] method. While the RVE method is favored over the traditional method, few meta-analyses have utilized this method when examining the antihypertensive effects of exercise. Through the lens of improving meta-analysis methods to study the effects of a single concurrent exercise session on PEH, we will compare the RVE and traditional meta-analytic methods using a simulation to evaluate the performance of the methods in a variety of settings.

The remainder of this thesis is organized as follows. In section 2, we explain the motive for comparing the meta-analysis methods used in this project and introduce a PEH meta-analysis dataset, which we will use to illustrate the differences between the methods. In section 3, we introduce the statistical models used in classical and RVE meta-regression and explain the estimated variance calculations for each. Section 4 describes the Monte Carlo simulation used to compare the performance of the estimators across three variations of the classical random effects and RVE meta-analysis methods, as well as provides an example of a simulated dataset. In section 5, we provide the simulation results of the estimation performance of the bias, standard deviation, coverage probability, and power for the four methods. In addition, section 6 applies the findings of the Monte Carlo simulation to PEH research, and using a real PEH meta-analytic dataset, compares the results of classical random effects meta-analysis using the average BP reading across all time points to the results of the RVE method across all time points. Lastly, in section 7, we propose which methods are appropriate for different study design settings according to the method performance found in the simulation.

## SECTION 2: APPLICATION TO POSTEXERCISE HYPOTENSION

The purpose of this analysis is to compare the performance of meta-analytic procedures across complex study designs, specifically in PEH research. As defined by the American College of Cardiology and American Heart Association, hypertension is the presence of elevated blood pressure in which systolic blood pressure (SBP) $\geq$ 130 mmHg and diastolic blood pressure (DBP) $\geq$ 80 mmHg [2,4]. Elevated blood pressure is a common condition; it affects nearly half of all American adults[6] and is the most important risk factor of cardiovascular disease[1,2,7]. Previous studies have indicated that there is a significant and sustained reduction in arterial blood pressure after a single exercise session[2,8]. This phenomenon, known as *postexercise hypotension* (PEH), is more specifically defined as an immediate reduction in blood pressure following a single bout of exercise that can be sustained for up to 24 hours[3]. The effects of exercise on BP vary due to a variety of factors that include the duration, intensity and type of exercise.

Previous meta-analyses show that aerobic exercise and dynamic resistance exercise lowers SBP and DBP 5-8 mmHg among adults with hypertension[3]. However, little is known about the acute PEH effects of concurrent exercise, defined as aerobic and dynamic resistance training performed in close proximity[3,9]. Only two previous meta-analyses have demonstrated the effects of concurrent exercise on PEH: Carpio-Rivera *et al.* found a significant reduction in SBP/DBP of 7/3 mmHg[10] and similarly, Baffoni found a significant reduction in SBP/DBP of 13.8/4.9 mmHg[11] following a single bout of exercise. Unlike the random-effects analysis of variance (ANOVA) method utilized by Carpio-Rivera *et al.*[10], Baffoni was able to account for the dependency occurring in a variety of study design settings, using the Robust Variance Estimation (RVE) method. While both meta-analyses had limitations due to the small number of included studies, the difference in effect size estimates across the two studies is likely due to the differences in statistical methods and design that were used.

Due to the differences in variance estimation techniques used in random effects and RVE methods, we expect that in PEH meta-analyses, the RVE method will perform better in detecting the true effect size. Because of the complex design of PEH studies with both within subject and between group effects over time, we expect that the RVE method will be better able to account for the variance due to dependent intervention groups and multiple BP collection time points. Table 1 provides a summary of all included studies within the Baffoni meta-analysis[11]:

**Table 1** Summary of studies included in Baffoni meta-analysis

| Study ID | Author | Year | Design type | Number of Exercise groups | Subjects (n) | Repeated Measurements (k) |
|---|---|---|---|---|---|---|
| **1** | Del Pozo Cruz | 2012 | Randomized to group | 1 | 22 | 1 |
| **2** | Favaji | 2012 | Crossover | 2 | 10 | 4 |
| **3** | Keese | 2012 | Crossover | 3 | 21 | 12 |
| **4** | Mendes | 2014 | Crossover | 1 | 23 | 4 |
| **5** | Tibana | 2015 | Crossover | 2 | 16 | 6 |
| **6** | Teixeira | 2011 | Crossover | 1 | 20 | 4 |
| **7** | Meneses | 2015 | Crossover | 2 | 19 | 1 |
| **8** | Dos Santos | 2014 | Randomized to group | 2 | 20 | 1 |
| **9** | Azevedo | 2017 | Crossover | 3 | 11 | 4 |
| **10** | Ferrari | 2017 | Crossover | 1 | 20 | 7 |
| **11** | Abrahin | 2016 | Crossover | 1 | 15 | 4 |

Table 1: Summary of studies included in Baffoni meta-analysis[11]

Using the Baffoni meta-analysis[11] as our dataset, we modeled many aspects of the simulation after the designs of included studies. In an effort to improve the applicability of the results using the RVE method, our simulation will compare the estimation and testing performance of the RVE method to three different variations of the classical random effects method.

## SECTION 3: METHODS

To determine the overall effect of a research question across many similar studies, meta-analysis is a statistical technique that allows us to: 1) summarize effect sizes across all studies; and 2) observe the moderator level effects to determine how study design settings are related to the effects[5,12,13]. For example, in many PEH studies, common moderators include resting BP, measures of body composition, and features of the exercise intervention such as intensity, and the time BP was recorded after exercise. In cases where: 1) multiple measurements of the same outcome variable are collected across each subject and compared to the same baseline measurement; 2) several outcome variables are measured across each subject, or 3) within the study there are multiple treatment groups that are being compared to the same control group, standard meta-analysis techniques will incorrectly assume independence of the effect sizes[5]. If effect sizes are actually dependent, the grouped effect sizes could be summarized or collapsed into independent synthetic effect sizes so that the traditional methods could be used. Alternatively, methods that explicitly account for the dependence in effect sizes are needed. The Robust Variance Estimation (RVE) meta-analytic method is able to account for dependence between effect sizes, without making assumptions about the specific form of the sampling distributions of the effect sizes or requiring knowledge of the covariance structure of the dependent estimates[5].

For comparison, the traditional random effects meta-analytic model with $p$ moderators is written as:

$$Y_j = \delta + \beta_1 x_{1j} + \cdots + \beta_p x_{pj} + \zeta_j + \varepsilon_j \tag{1}$$

where $Y_j$ is the effect size for study $j$ ($j$=1,…$m$); $\delta$ is the overall effect size (also known as the intercept in meta-regression); $\beta_l$ is the unknown regression coefficient for covariate $l$ ($l$ = 1…$p$); $\zeta_j$ is the study-level random effect for study $j$ where $\zeta_j \sim N(0, \tau^2)$; and $\varepsilon_j$ is the random error of study $j$ = 1,… $m$ where $\varepsilon_j \sim \mathrm{N}(0, v_j)$[5,14].

The weighted least squares estimate of $\beta = (\delta, \beta_1, \dots, \beta_p)'$ is given by

$$b = (X'WX)^{-1}(X'WY),$$

where $X$ is the m × ($p$+1) design matrix, $Y$=$(Y_1, \dots, Y_m)'$, and $W$ is an $m \times m$ diagonal weight matrix. The variance of $b$, in general, can be estimated as

$$var(b) = (X'WX)^{-1}(X'W\Sigma WX)(X'WX)^{-1},$$

where Σ is the $m \times m$ covariance matrix of $\varepsilon_j$. In a classical meta-analysis model where the effect sizes from each study are assumed to be independent, the diagonal elements of Σ are the variances $v_j$ and the off-diagonal covariance estimates are assumed to be zero. Since the weights are typically defined to be inverse variances, with the diagonals of $W$ equal to $(1/v_j)$ for each effect size, the estimator for the variance of $b$ reduces to:

$$V_{classical}(b) = \ (X'WX)^{-1}.$$

The estimator $V_{classical}(b)$ is appropriate for use when the effect sizes are independent, but inappropriate for use when effect sizes are statistically dependent, however, because it results in standard error estimates that are too small[9].

With dependent effect sizes, and assuming a correlated effects model where multiple effect size estimates are nested within studies, the meta-analytic model with $p$ covariates can now be written as:

$$Y_{ij} = \delta + \beta_1 x_{1ij} + \cdots + \beta_p x_{pij} + \zeta_j + \varepsilon_{ij}, \tag{2}$$

where for $i = 1 \ldots k_j$, ( j = 1…$m$), $Y_{ij}$ is the $i$th effect size in the $j$th study; $\delta$ is the intercept in the meta-regression model; $\beta_l$ is the unknown regression coefficient for covariate $l$ ($l =$ 1…$p$); $\zeta_j$ is the study-level random effect such that $\mathrm{Var}(\zeta_j) = \tau^2$ is the between-study variance component; and $\varepsilon_{ij}$ is the random error for the $i$th effect size in the $j$th study. For $\varepsilon_{ij}$, the errors are assumed independent across the $m$ studies, but they are not assumed independent within the same study. We can write the covariance matrix of $\varepsilon_{ij}$ as a block diagonal matrix of m $k_j \times k_j$ matrices where $\Sigma = diag(\Sigma_1, \ldots \Sigma_m)$[5,14].

To calculate the estimates for the regression coefficients $\beta = (\delta, \beta_1, \ldots, \beta_p)$' for the RVE method, we use the same weighted least squares procedure:

$$b = \left(\sum_{j=1}^{m} X'_j W_j X_j\right)^{-1} \left(\sum_{j=1}^{m} X'_j W_j Y_j\right), \qquad (3)$$

where $W_j$ is the diagonal weight matrix of the jth study; $X_j$ is the design submatrix for the $j$th study; and here $Y_j$ is the $k_j \times 1$ vector containing the effect sizes for study $j$[5,14]. However, the classical random effects and Robust Variance Estimation methods differ in the variance of the coefficients[14]. The variance using the RVE method is estimated as:

$$V_{RVE}(b) = \left(\sum_{j=1}^{m} X'_j W_j X_j\right)^{-1} \left(\sum_{j=1}^{m} X'_j W_j A_j e_j e'_j A_j' W_j X_j\right) \left(\sum_{j=1}^{m} X'_j W_j X_j\right)^{-1}, \qquad (4)$$

where $A_j$ is an adjustment matrix to adjust for small sample bias when m is small, $e_j$ is the $k_j \times 1$ vector containing the estimated residual errors of the $j$th study, and $e_j e'_j$ is an estimate of $\Sigma_j$[5,14,15]. The choice of weights and other details may be found in Hedges *et al.*[5,15]

## SECTION 4: MONTE CARLO SIMULATION

In order to better understand the complexities of real-world PEH meta-analyses and find results that could be applied to a variety of study designs, we designed a Monte Carlo simulation study to compare the performance of the estimators from three variations of the classical random effects meta-analysis methods to the preferred RVE method. The simulated replicates were created and analyzed using R software and the simulation design was inspired from the design in Hedges, Tipton, and Johnson (2010)[8].

### *Parameters*

Within this simulation, we only consider two-groups studies with an equal number of independent subjects in the control and treatment groups within study. Each simulation scenario examined a fixed combination of three parameter values, *m* [number of studies], *slope* [of a single time moderator variable], and $\tau^2$ [between-study variance] over 1000 replicated data sets. Across all simulation scenarios, we set the overall effect size, δ, to 0.5. Within our meta-regression model, this value indicates the intercept, or the effect when the *slope* of the moderator variable is zero.

Indicating the number of studies included in the meta-analysis, we set $m$=10, 50, 100. Based on the included studies within the previous Baffoni meta-analysis[11] where $m$=11, we set these values where the findings could be applied to small and large meta-analyses.

As time is added as a moderator variable in our meta-regression model, we chose four possible slope values for this moderator; *slope* = 0, 0.05, 0.10, 0.15, indicating the increase in overall effect size as time increases.

For random effects meta-analysis, we must also specify the between-study variance of the effect size in our meta regression model. This $\tau^2$ value indicates the between study variance and reflects the amount of true heterogeneity. In this simulation, we set $\tau^2$ to be one of three values; $\tau^2$= 0, 0.15, 0.30.

From the combinations of these three parameters, we have (3×4×3) 36 simulated scenarios. In each of these combinations/scenarios, we generated 1000 simulated data sets, in which we varied the following parameters to account for differences within individual study designs, such as those included in the Baffoni meta-analysis[11]:

Sample size within study: Within a single simulated dataset, the m studies were allowed to have different sample sizes. For a given study, the sample size of the control group, *n*, was one of 10, 20, 200, and the number of total study subjects in the study across the control and experimental groups was 2**n*, as we are assuming independent groups. These specific values were chosen based on previous exercise intervention meta-analyses where a smaller sample sizes are common. We chose a larger third sample size to allow our results to be applicable to fields in which studies can be completed with large sample sizes. Within each replicated dataset, 30% of the m studies had $n$=10 (20 subjects), 60% of the $m$ studies had $n$=20 (40 subjects) and the remaining 10% of the m studies had have $n$=200 (400 subjects).

Number of time points: Using the designs of the studies included in Baffoni meta-analysis[11] as an example, we introduced parameter $k$ to indicate the number of time points BP was recorded after exercise. We chose values $k$=2, 4, 8 as they are similar to those seen in the included studies, where BP was measured 5.2±3.3 times after exercise[11]. Within each simulated dataset, 30% of the m studies had 2 time points, 60% of studies had 4 time points and 10% of studies had 8 time points.

Selection of time points: To account for the differences in time points BP was collected across the studies included in the previous Baffoni meta-analysis[11] in which BP was collected over 75.3±36.2 minutes, we introduced a parameter, *kgrid*, which listed all of the possible time points from which *k* are chosen for each study within a simulated dataset. We set *kgrid* = 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2, 2.25, 2.5, 2.75, 3; where 0.25 represents 15 minutes, 0.5 represents 30 minutes, and so on up to three hours after the exercise intervention. For each of the m studies in a simulated dataset, *k* (as determined above) elements from *kgrid* were randomly selected, each with equal probability.

Correlation over time: After introducing a time moderator, we must also specify the correlation across the *k* time points. In real-world meta-analysis this value may be different for each included study. For our simulation we set $\rho = 0.2$ or 0.8 for the correlation between any two different time points within a study. More specifically 50% of the *m* studies for a given simulated dataset had $\rho = 0.2$ and the remaining 50% had $\rho = 0.8$.

### *Data Generation*

To generate a single simulated dataset under a specific simulation scenario (fixed *m*, *slope*, and $\tau^2$ combination), we first randomly assigned *n* (sample size with 2**n* subjects between the treatment and control groups), *k* (time points selected out of *kgrid*) and $\rho$ (correlation coefficient across time points) values to each of the *m* studies, where 30% of studies had *n*=10, 60% of studies had *n*=20 and 10% of studies had *n*=200. Additionally, 30% of studies were selected to have 2 time points, 60% to have 4 time points, and 10% to have 8 time points. We created a vector, *x*, indicating which time points are collected for that particular study. To construct this vector, we randomly selected *k* time points out of all possible time points in *kgrid*. Lastly, 50% of m studies were selected to have low correlation between time points (ρ =0.2) and the other 50% of studies had high correlation between time points (ρ = 0.8). Following Hedges *et al.*[5], for each study within a single simulated dataset, we generated treatment group data and control group data as follows.

To simulate the response, $\mathrm{Y}_{\mathrm{treatment,l,i,j}}$, for the *i*th treatment subject (*i*=1…*n*) at the *j*th time point (*j*=1…*k*) within *l*th study (*l*=1…*m*), we used the model:

$$Y_{treatment,l,i,j} = \delta + slope * x_j + \eta_l + \xi_i + \zeta_{(i-1)k+j}, \tag{5}$$

where a) $\delta = 0.5$ throughout, b) *slope* is set to one of 0, 0.05, 0.10, or 0.15, c) $x_j$ is the time the response measurement was taken in hours, d) $\eta_l$ is the study-specific random effect, where each $\eta_l$ was generated from a $N(0,\tau^2)$ distribution, e) $\xi_i$ is the within-subject random effect, taken from the vector $[\xi_1, \xi_2, \ldots \xi_{2*n}]$ where each element was generated independently from a $\mathrm{N}(0, \rho)$ distribution, f) and the random error $\zeta_{(i-1)k+j}$ is taken from vector $[\zeta_1, \zeta_2, \ldots \zeta_{2*n*k}]$, where each element was generated independently from a $\mathrm{N}(0, 1-\rho)$ distribution. Note that new within-subject random effects, random errors, and time points are generated for each of the *m* studies within a simulated data, although the study index for these variables have been suppressed for ease of notation.

As we wish to compare the treatment group to a control group, we generated control responses, $\mathrm{Y}_{\mathrm{control,l,i,j}}$, for *i*th control subject (*i*=1…*n*) at the *j*th time point (*j*=1…*k*) within the *l*th study (*l*=1…*m*) using the model:

$$Y_{control,l,i,j} = \xi_{n+i} + \zeta_{nk+(i-1)k+j}, \tag{6}$$

where the random effects were taken as elements from the same vectors described above.

Ultimately, we used standardized mean difference as the effect size of choice. To determine the differences, we calculate the mean difference and pooled standard deviation between the treatment and control responses at each of the *k* time points. The effect size estimate for the *j*th time point within the *l*th study is defined as

$$Y_{lj} = \frac{\bar{Y}_{treatment,l,\cdot,j} - \bar{Y}_{control,l,\cdot,j}}{S_{l,\cdot,j}},$$

where $\bar{Y}_{treatment,l,\cdot,j}$ and $\bar{Y}_{control,l,\cdot,j}$ are the sample means for the treatment and control groups, respectively, at the *j*th time point, and $S_{l,\cdot,j}$ is the corresponding pooled within groups standard deviation estimate. Lastly, we store the effect sizes in a dataframe where each row corresponds to one time point within study *l*, *l*=1,…,*m*, and with columns:

- studyid=study *l* (in 1…*m*)
- effectsize
- var. effectsize = (2/*n*) + (effectsize)$^2$/(2*(2*$n$-2))
- mod= time point used (in hours)
- samp.size= sample size in one group, *n*, (total sample size is 2**n*)
- mean.e = mean of the effect size of the treatment group within the corresponding time point across all subjects

- mean.c= mean of the effect size of the control group within the corresponding time point across all subjects
- sd.e = standard deviation of the effect size of the treatment group within the corresponding time point across all subjects
- sd.c = standard deviation of the effect size of the control group within the corresponding time point across all subjects

Using this process to generate datasets, we conducted the RVE and classical random effects meta-analyses across all 1000 replicates for each simulation scenario. Table 2 displays an example replicate dataset, where $m = 10$, $\tau^2 = 0.00$, $slope = 0.00$. The R code for generating and analyzing the data is provided in the Appendix.

**Table 2** Example Simulated Dataset

| studyid | effectsize | var.effsize | mod | samp.size | mean.e | mean.c | sd.e | sd.c |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.05262347 | 0.20007692 | 1.5 | 10 | 0.12097329 | 0.06586558 | 0.8821156 | 1.1896056 |
| 1 | 0.37057597 | 0.20381463 | 3 | 10 | 0.7460434 | 0.39419455 | 0.982089 | 0.9156791 |
| 1 | 0.68069548 | 0.21287073 | 1.75 | 10 | 0.54206341 | -0.1550536 | 1.2738355 | 0.6892067 |
| 1 | 0.49031994 | 0.20667816 | 1.25 | 10 | 0.47127448 | -0.1294723 | 1.1582231 | 1.2887268 |
| 2 | 0.15729836 | 0.2006873 | 2.5 | 10 | 0.17578814 | 0.02228416 | 0.9745036 | 0.9772501 |
| 2 | 0.7291087 | 0.21476665 | 0.25 | 10 | 0.26209475 | -0.4396974 | 0.9899361 | 0.9343292 |
| 3 | 0.0030823 | 0.20000026 | 0.5 | 10 | 0.06020441 | 0.05703929 | 1.0507913 | 1.0023813 |
| 3 | 0.31922746 | 0.20283073 | 2.75 | 10 | 0.25727509 | -0.1368294 | 1.2731455 | 1.1947231 |
| 3 | 0.31037551 | 0.20267592 | 0.75 | 10 | 0.31187442 | -0.0652071 | 1.2829409 | 1.1428583 |
| 3 | 0.02806727 | 0.20002188 | 1.75 | 10 | 0.09854139 | 0.07068331 | 1.0069662 | 0.9779148 |
| 4 | 0.71841147 | 0.10679099 | 2.25 | 20 | 0.48969765 | -0.1569916 | 0.9469913 | 0.8507664 |
| 4 | 0.59808637 | 0.10470668 | 1.25 | 20 | 0.56064186 | -0.0757792 | 1.2229622 | 0.876905 |
| 4 | 0.72892701 | 0.10699124 | 2.75 | 20 | 0.59828972 | -0.1425935 | 1.0781746 | 0.9506249 |
| 4 | 0.68965117 | 0.10625814 | 0.25 | 20 | 0.58959254 | -0.1147072 | 1.1334852 | 0.8950284 |
| 5 | 0.6068152 | 0.10484506 | 2.25 | 20 | 0.69564667 | 0.00922512 | 1.0417144 | 1.2140839 |
| 5 | 1.07636937 | 0.11524436 | 0.25 | 20 | 1.04634369 | -0.0476519 | 0.845907 | 1.1621016 |
| 5 | 0.69027738 | 0.10626951 | 1.25 | 20 | 0.73654784 | -0.0045838 | 1.0134943 | 1.1306517 |
| 5 | 0.86508919 | 0.1098471 | 3 | 20 | 0.77552243 | -0.0601494 | 0.8764858 | 1.0478861 |
| 5 | 0.8535626 | 0.10958644 | 0.75 | 20 | 0.9227241 | 0.03342366 | 0.9365583 | 1.1374712 |
| 5 | 0.93405759 | 0.11147978 | 1.5 | 20 | 0.8404184 | -0.0115605 | 0.91545 | 0.9087913 |
| 5 | 0.85171333 | 0.10954494 | 2.75 | 20 | 0.94237061 | 0.07852826 | 1.0601695 | 0.966131 |
| 5 | 0.93170573 | 0.11142205 | 1 | 20 | 1.10397948 | 0.09498799 | 1.007485 | 1.1534899 |
| 6 | 0.69924822 | 0.10643353 | 2.5 | 20 | 0.5027329 | -0.1331752 | 0.9234956 | 0.8951166 |
| 6 | -0.2646856 | 0.10092182 | 1.75 | 20 | 0.05849797 | 0.33521246 | 1.1508829 | 0.928107 |
| 6 | 0.47169961 | 0.10292764 | 3 | 20 | 0.43145124 | -0.0113392 | 0.8967702 | 0.9788598 |
| 6 | 0.20899499 | 0.10057472 | 2.75 | 20 | 0.20670258 | 0.03729902 | 1.0097353 | 0.5426406 |
| 7 | 0.18095768 | 0.10043086 | 1.75 | 20 | 0.47883198 | 0.28101224 | 0.9135775 | 1.2471853 |
| 7 | 0.26478809 | 0.10092254 | 2.5 | 20 | 0.30852882 | 0.01264215 | 1.0998684 | 1.1347534 |
| 7 | 0.22133523 | 0.1006446 | 1.25 | 20 | 0.48424348 | 0.24496934 | 0.9196824 | 1.2212766 |
| 7 | 0.37574469 | 0.10185769 | 2.25 | 20 | 0.45036761 | 0.01780371 | 1.0323269 | 1.2589301 |
| 8 | 0.6176485 | 0.1050196 | 2.5 | 20 | 0.56321795 | -0.108157 | 1.1248458 | 1.047758 |
| 8 | 0.50548052 | 0.10336198 | 1.75 | 20 | 0.44457614 | -0.0645036 | 0.9784001 | 1.035044 |
| 9 | 0.72687984 | 0.10695203 | 0.5 | 20 | 0.50350354 | -0.1770101 | 1.1146119 | 0.7145812 |
| 9 | 0.13218383 | 0.1002299 | 0.25 | 20 | 0.52256762 | 0.39103825 | 0.9200676 | 1.0647629 |
| 10 | 0.5269465 | 0.01034883 | 0.75 | 200 | 0.45793195 | -0.0374194 | 0.9424758 | 0.9376 |
| 10 | 0.60681037 | 0.01046259 | 1.25 | 200 | 0.55348577 | -0.0143078 | 0.8808955 | 0.9874711 |
| 10 | 0.45832745 | 0.0102639 | 2.5 | 200 | 0.46915359 | 0.00349202 | 1.029104 | 1.0027286 |
| 10 | 0.56346999 | 0.01039887 | 1 | 200 | 0.48820456 | -0.0671104 | 0.9209362 | 1.0461375 |

Table 2: Example of simulated dataset for one $m$ /$\tau^2$/slope combination

### *Robust Variance Estimate (RVE) Meta-Analysis*

We fit model (2) in Section 3 with $p$=1 using robu() from the robumeta package, with the following functional inputs: the meta-regression *formula* (effectsize~mod), *studynum* as the studyid, *var.eff.size* is the var.effsize that we calculated in the generated dataset, *modelweights*="CORR" to indicate the weighting based on the correlation coefficient (default), *rho*=0.8 (default)as the user-provided within-study effect size correlation used to determine the weights, and *small*=TRUE to invoke small sample corrections for the residuals and degrees of freedom[14].

For each replicate dataset in each simulation scenario, we extracted the estimated $\tau^2$, $I^2$ values, as well as the coefficient estimates, standard errors, t-statistics, p-values, and 95% confidence intervals for both the intercept and the moderator effects. We also calculated the proportion of simulations in each scenario in which the RVE method was sensitive to the choice of user-specified "*rho* (=0.8)" in the robu() function. Specifically, an RVE analysis for a given dataset was declared "sensitive" to the rho specification if any of the standard deviations of intercept, slope, or tau-squared estimates across rho=(0,0.2,0.4,0.6,0.8,1), computed on the same dataset, was more than 0.05.

### *Classical Random Effects Meta-Analysis*

For the classical random effects meta-analysis, we fit model (1) from Section 3, again with p=1. As we have induced dependence among effect sizes over time, we examined three different variations of the classical random effects meta-analysis method, all of which assume independence of the effect sizes: Naïve, Average and Random, described in more detail below.

We conducted classical random effects meta-analysis using the metacont() function from the R *meta* package, with functional inputs including mean effect size and standard deviation of treatment and control groups, sample size of treatment and control groups, study id, and specifying use of  the pooled variance, DerSimonian-Laird estimation method for $\tau^2$, and Cohen's standardized mean difference. The only difference between the following three methods is the input data over which the meta-analyses are conducted. While Robust Variance Estimation in R can directly incorporate the time moderator within the robu() function, for the classical random effects meta-analysis, we can use the metacont() function in conjunction with the metareg() function to introduce time as a moderator for the Naïve and Random approaches.

### *Classical Random Effects- Naïve*

The first classical random effects method was the naïve method, which naively treats the all effect sizes across all time points as independent.  The independence assumption is clearly violated with this approach.

### *Classical Random Effects- Average*

We also chose to use the average method, in which a typical random effects meta-analysis was conducted on the average response (across all time points), eliminating the possibility for including time of measurement as a moderator in the model. With this method, the effect size is the average effect size across all time points within each study; the sample size of experimental and control groups are the same within one study; the mean of the experimental group is the mean effect size across all experimental time points within one study; the mean of the control group is the mean effect size across all control time points within one study; the SD of the experimental group is the SD of mean effect size across all experimental time points within one study; and the SD of the control group is the SD of mean effect size across all experimental time points within one study. Thus, the independence assumption is be satisfied with this approach, as the study-specific averages are all independent.

### *Classical Random Effects- Random*

The third classical random effects analysis that we chose to include was a random selection of a single effect size from each study. More specifically, we randomly selected one time point within each study to enter into the meta-analysis. The independence assumption is also satisfied with this approach.

For all classical approaches, as with RVE, we extracted the estimated $\tau^2$, $I^2$ values, as well as the coefficient estimates, standard errors, t-statistics, p-values, and 95% confidence levels for both the intercept and the moderator (note that for average, however, no moderator-level information is available).

We compared the performance of the methods by compiling the results of the 1000 replicates from each simulation scenario and calculated coverage probabilities of the 95% confidence intervals of the intercept and moderator coefficients; bias for the estimation of the intercept and slope; power of non-zero intercept and slope tests; standard deviations of the intercept and moderator estimates using both the empirical method and the average of the calculated standard errors. Details of the performance measures and the results of the simulation study are provided in Section 5.

## SECTION 5: RESULTS

Using the *N*=1000 simulated replicates of each combination of *m*, $\tau^2$ and *slope*, we observe the differences in performance between the four meta-analysis methods. Let λ denote is the true value of the parameter (either the intercept or slope parameter), which is known from the simulation design, and let $\hat{\lambda}_i$ denote the estimate of λ from the *i*th replicate dataset. To compare the estimation and testing performance, for both the intercept/overall and moderator effects, we calculate:

1) the estimated bias of the estimator

$$Bias\left(\widehat{\lambda_\imath}\right) = \frac{1}{N}\sum_{i=1}^{N}\left(\widehat{\lambda_\imath} - \lambda\right), \tag{7}$$

2) the standard deviation of the estimator using

a) the empirical standard deviation of N replicate estimates (gold standard)

$$SD_{empirical} = \sqrt{\sum_{i=1}^{N}\left(\bar{\lambda} - \widehat{\lambda_\imath}\right)^2 \Big/ N}, \tag{8}$$

where $\bar{\lambda} = \sum_{i=1}^{N} \widehat{\lambda_\imath}/\mathrm{N}$,

b) and the estimated standard deviation of the estimator using the average of the N calculated standard error (SE) estimates

$$SD_{calculated} = \sqrt{\sum_{i=1}^{N} SE_i^2 \Big/ N}, \tag{9}$$

3) the empirical coverage probability of the 95% confidence interval

$$CP = \frac{1}{N}\sum_{i=1}^{N} I\left(\lambda \in \left[95\%\ CI\ of\ \lambda\ based\ on\ \widehat{\lambda_\imath}\right]\right), \tag{10}$$

where I(a) takes the value 1 when condition a is true, and 0 otherwise,

4) and the empirical power for detecting a non-zero effect,

$$Power = \frac{\sum_{i=1}^{N} I(\ p_i < 0.05)}{N} \tag{11}$$

where $p_i$ is the two-sided p-value from the $i$th replicate dataset for testing the null hypothesis $H_0: \lambda = \ 0$ (either intercept or slope parameter). These values are displayed in a series of plots where Figures 1, 2 and 3 each correspond to a fixed m level and each panel within the Figure refers to a fixed $\tau^2$ level. It should be noted that within a $\tau^2$ panel, the plots show the estimation performance for the intercept (top) and slope (bottom) parameters in the meta-regression model. Within each plot, values are plotted as a function of the true, data-generating slope of the moderator, and are calculated across the 1000 replicates. The four methods are differentiated by line type.

We describe the results in detail below, and provide a summary at the end of this section. We also note that while the simulation design allows for scenarios that include a time moderator when $\tau^2$ =0, these arguably do not represent practical scenarios because when studies are indeed homogeneous, an intercept-only model would be a sufficient model. Thus, we do not discuss estimation and inference performance for the moderator when $\tau^2$ =0, and limit discussion of estimation and inference performance for the intercept when $\tau^2$ =0 to only scenarios where the *slope* is also equal to 0.

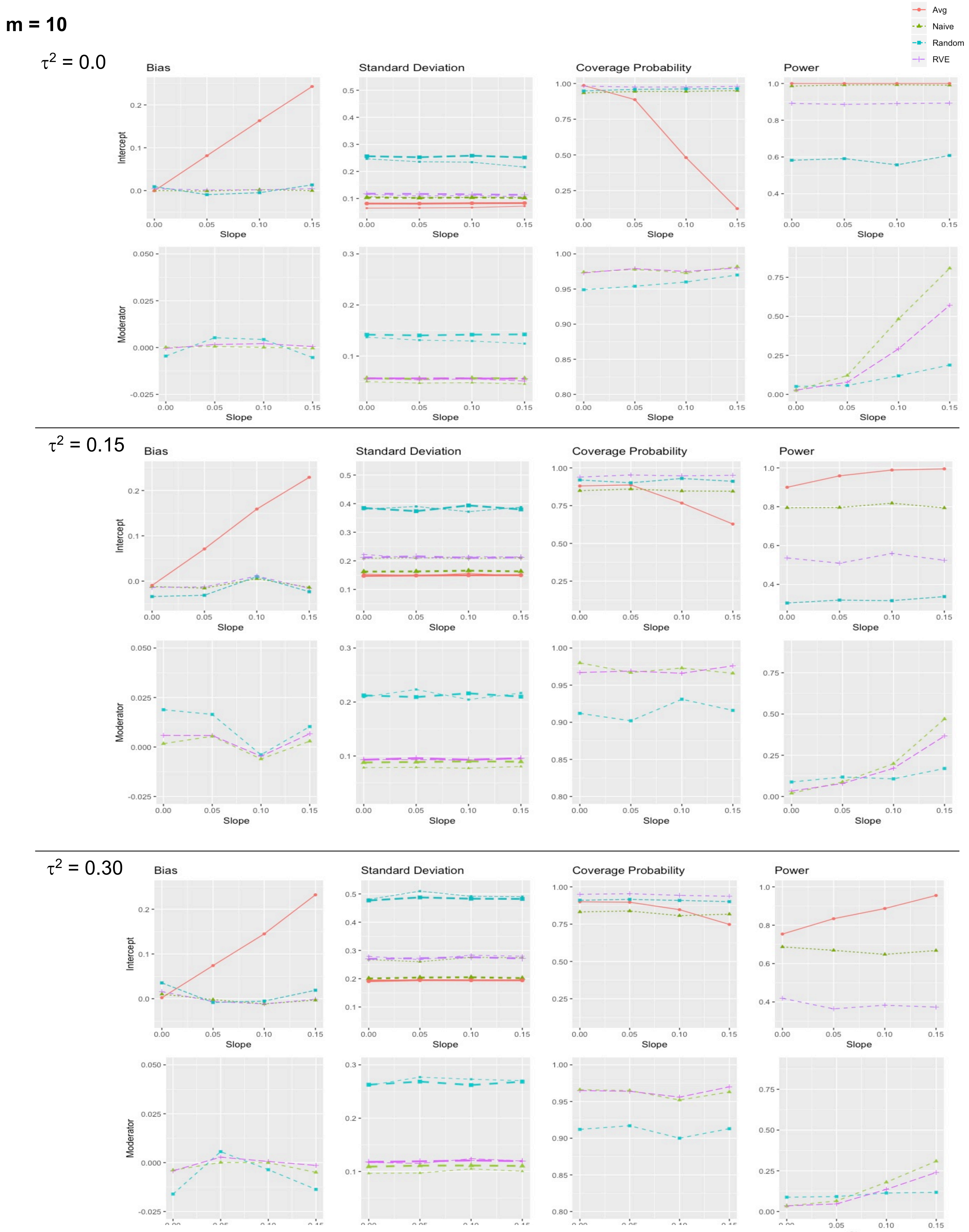


Figure 1: Estimator performance of four methods, where m=10

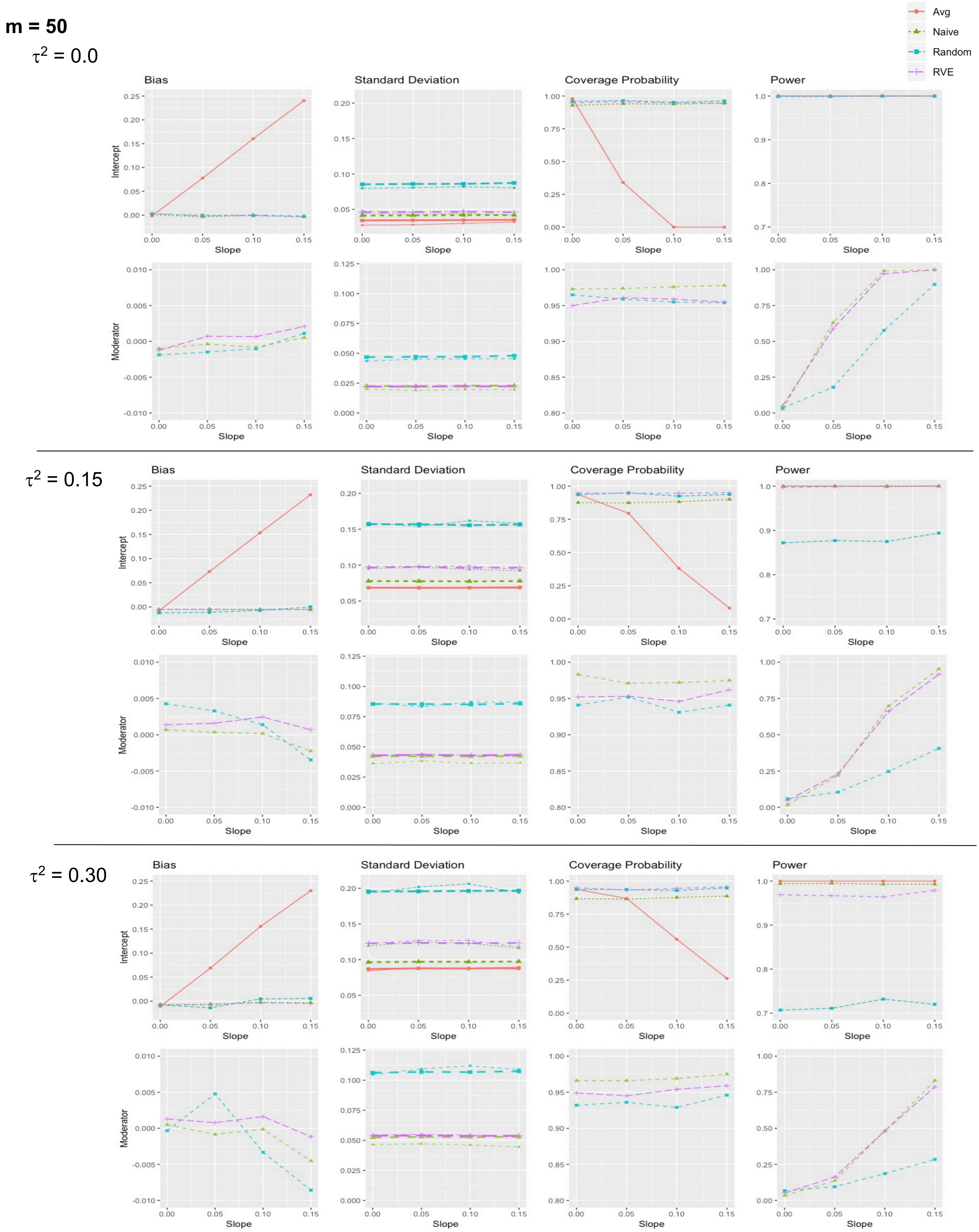


Figure 2: Estimator performance of four methods, where m=50

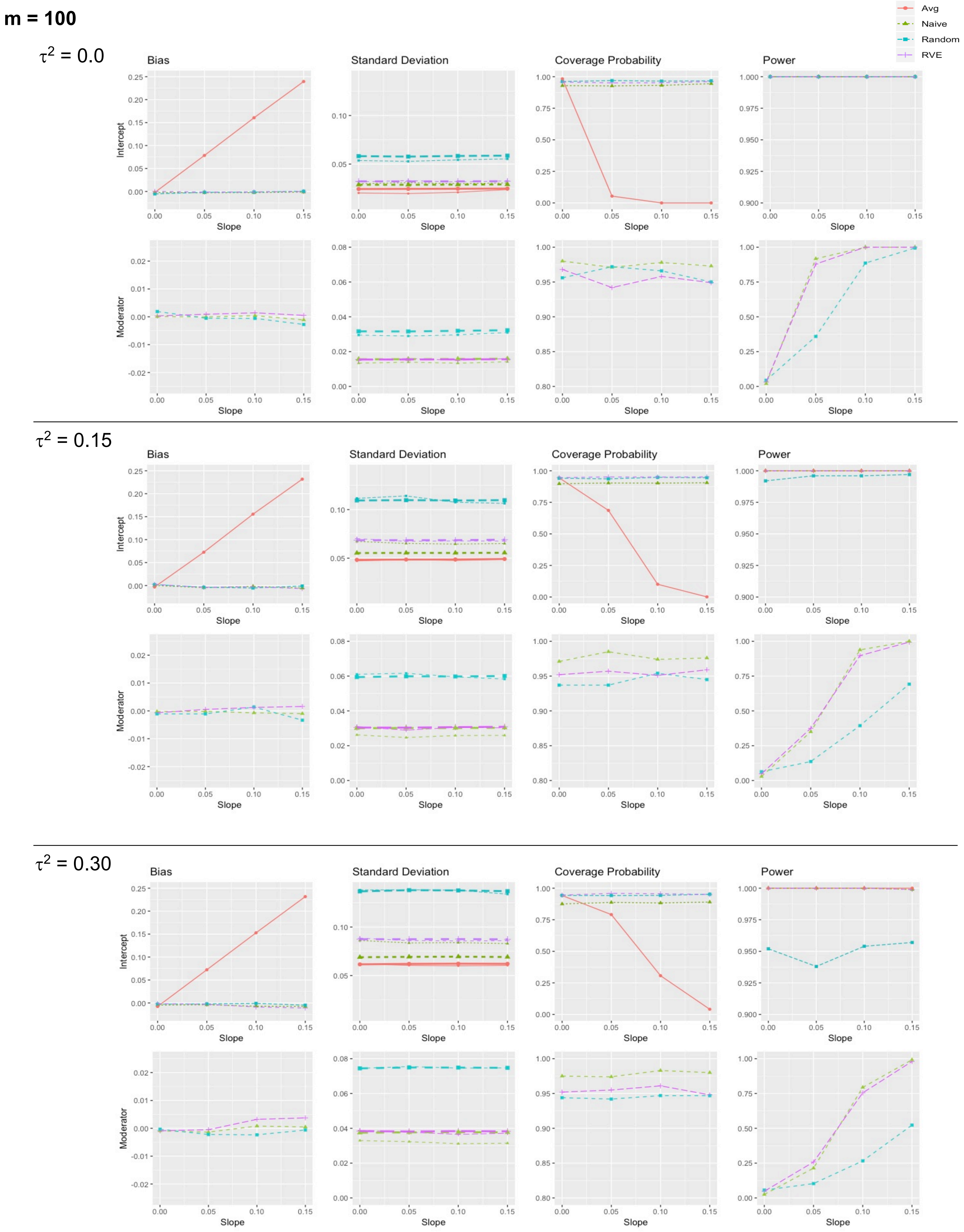


Figure 3: Estimator performance of four methods, where m=100

### *Bias Estimation*

We found that across all *m*, $\tau^2$, and *slope* levels that classical Naïve and Random, and RVE do about the same in terms of providing low bias estimators for both the intercept and moderator effects. However, the estimates using the Naïve and RVE methods are slightly closer to the ideal value of 0. As expected, the bias of the intercept estimator using the Average method increased as slope increased because the average method does not account for the time moderator effect. It instead accounts for this effect within the estimate of the intercept, causing the bias to increase. The four methods were compared on their ability to reduce bias of the intercept/overall and moderator effect sizes in more detail below. We also compared the performance of the bias estimator for both effects while holding two variables constant and changing the other, among *m*, $\tau^2$, and *slope* variables.

#### *Intercept*

Looking at the intercept and overall effect across all practical *m*, $\tau^2$and *slope* values, the Naïve and RVE methods are the most effective in reducing the bias of the intercept estimate. Meanwhile the Random and Average methods did not perform as well.

More specifically, when the number of studies, *m*, is 100, the Random, RVE, and Naïve methods perform roughly equivalently, while Average method performs worse. When the number of studies is equal to 50, the order of the performance of the methods slightly change: Naïve and RVE are approximately the same, followed by Random and Average. Lastly, when m is 10, the order stays the same: Naïve and RVE are roughly equivalent, while Random, and Average perform worse. Therefore, as $\tau^2$ and *slope* are held constant and *m* increases, the RVE and Naive methods are better able to reduce the bias for estimating the intercept/effect size, while the Random and Average methods perform worse.

When the number of studies and the true *slope* of the moderator are held constant and the $\tau^2$ is equal to 0.30, the best methods for reducing bias of the intercept are the Naïve and RVE methods, which are roughly equivalent for bias when estimating the intercept, followed by the Random method and Average methods. When $\tau^2$ is 0.15, the Naïve method still performs the best, followed by RVE, Random and Average. In addition, when $\tau^2 = 0.00$ and *slope* is 0, all four methods preform similarly. Moreover, the RVE method performs well in reducing the bias of the intercept/effect size across all $\tau^2$ levels, however as *m* and *slope* are held constant and $\tau^2$ increases, the Average and RVE methods become less effective in reducing bias of the intercept/overall effect, and the Naïve and Random methods improve.

Holding the *m* and $\tau^2$ levels constant, and the true *slope* is 0.15, the RVE method is best able to reduce the bias of the intercept/overall effect size, followed by the Naïve, Random and Average methods. When the *slope* is held at 0.10, the Naïve method performs the best, closely followed by the RVE method. The worst methods were the Random and the Average methods. At the third *slope* level of 0.05, the method ranking is as follows: RVE, Naïve, Random and Average. Lastly, when the *slope* is equal to 0, the best method is Naïve, then RVE, Average and Random.  Moreover, as slope increases, RVE and Naïve neither improves nor worsens, Random improves and Average worsens in reducing bias of the intercept/overall effect.

### *Moderator*

The best method in reducing the bias when estimating the moderator effect is the Naïve method across all *m* levels. The order of the methods for the performance of the bias of the moderator is as follows: 1) Naïve, 2) RVE and 3) Random. As previously stated, the average method is not used in the moderator analysis.

Within the individual *m* levels, the ranking of the methods is the same; Naïve is the best method in minimizing the bias of the moderator effect, RVE is the second-best method, and Random is the worst method. We can conclude that as $\tau^2$ and *slope* remain constant and *m* increases, the performance of the RVE and Random methods improve in reducing the bias for estimating the moderator effect.

As the *m* and *slope* levels remain constant, and the $\tau^2$ level is non-zero, the Naïve method is the best method for reducing the bias for estimating the moderator effect followed by the RVE and Random methods. As $\tau^2$ increases and is greater than 0, Naïve worsens but remains the best of the three methods, Random also worsens and RVE improves in minimizing the estimation of the bias of the moderator effect.

In addition, when the *m* and $\tau^2$ levels are held constant and the *slope* of the moderator is 0.15, the order of the highest performing methods is Naïve, RVE and Random. When the *slope* is equal to 0.10, Naïve remains the best method, RVE is not far behind and the worst method is still the Random method. If the *slope* is equal to 0.05, the order of the best methods remains the same: Naïve, RVE and Random. Lastly, when the *slope* is equal to zero, the order remains the same yet again: Naïve, RVE and Random. In addition, we found that the bias of the intercept and moderator effects tend to be inversely related, in that as the bias of the moderator increases, the bias of the intercept decreases, and vice versa.

### *Standard Deviations in Relation to Coverage and Probability and Power*

In the standard deviation (SD) plots, for each method across the *m*, $\tau^2$ and *slope* levels, the bold line represents the estimated calculated standard deviation in equation (9) based on the standard error estimates of the replicates computed within the corresponding meta-analysis function. Within each replicate of each simulation scenario, these individual SE estimates were used to calculate the 95% confidence intervals and test statistics in the software. Thus, coverage probability and power estimates will be influenced based on the performance of this SD method. The thin line represents the empirical standard deviation of the 1000 replicate estimates of the effect, shown in equation (8), which we consider to be the gold standard. By comparing the calculated standard deviation to the empirical standard deviation, we can judge the propriety of the resulting inferences and relate them to the coverage probability and power estimates.  If the two standard deviation methods match across a single *m*, $\tau^2$ and *slope* combination, then we can conclude that the method is performing well and that the corresponding inferences (e.g., based on confidence interval, p-values) are trustworthy provided the estimators are also unbiased. We expect the results to be trustworthy in this sense when the underlying assumptions of the meta-analytic methods are satisfied.  When underlying assumptions (e.g., independence, large *m*) are violated, performance may be affected. The coverage probability is expected to be 0.95, as we have extracted 95% confidence intervals from each replicate dataset.  For the intercept, the estimated power is expected to be high (close to 1), as the intercept was set to 0.5 across all simulation scenarios, and therefore the non-zero effect should be easily detected in many cases. As expected, the power of the moderator test increases as the true *slope* increases across all methods. In addition, as $\tau^2$ increases, the power for detecting a non-zero intercept and *slope* decreases, as expected. When the slope is equal to zero, the power of the moderator test corresponds to the type I error rate, which should be equal to 0.05. A summary of the estimation performance of the coverage probability and power estimate of the intercept/overall effect size based on standard deviation comparison is shown in Tables 3 through 5 Similarly, a summary of the estimation performance of the coverage probability and power estimate of the moderator effect are shown in Tables 6 through 8.

### *Intercept*

Across all m and $\tau^2$ levels, the calculated and empirical standard deviations using the average classical method were often equal or oscillated between being greater or less than the other across the four slope levels, showing that this method is appropriate in terms of providing accurate standard error estimates. It should be noted that this conclusion was expected since the independence assumption was not violated.  In addition, when *m* is low and there is no significant moderator effect, the Average method performed better than RVE in the simulations in terms of power for non-zero $\tau^2$ values.

According to the standard deviation when $m = 10$, the estimation of the coverage probability and power metrics perform well under the Average method when *slope* = 0, and we can infer that the method performs well under these conditions. When $m$= 50 or $m = 100$, the method is conservative. As previously stated, the Average method produces highly biased estimates for the overall effect size in the presence of a moderator effect (non-zero *slope*), so inferences for the overall effect size in these cases should not be trusted.

Following the Naïve method, within all *m* and $\tau^2$ levels, the calculated standard deviation estimate was less than the empirical standard deviation. This causes the margin of error of the effect size estimate and corresponding confidence interval width to be too small and the test statistic to be too big, which in turn causes the coverage probability to be less than 0.95 and the power to be overstated. Therefore, inference (based on confidence intervals, p-values, etc.) for the intercept using the Naïve method cannot be trusted.

When the Random method is used, the calculated standard deviation is roughly equal to the empirical standard deviation across most *m*, $\tau^2$ and *slope* combinations. Only when $\tau^2 = 0.00$, the calculated standard deviation is greater than the empirical standard deviation, therefore the power is lower and the coverage probability is larger than it should be. Moreover, as $\tau^2 = 0.00$ is not commonly observed in practice and with the exception of the when m=10 and $\tau^2 = 0.0$, we can infer that the method at a low $\tau^2$ level is conservative but otherwise leads to accurate inference for the intercept/effect size.

Using the RVE method, the calculated standard deviation estimate is equal to the empirical standard deviation across all *m* and $\tau^2$ level combinations, except when *m*=10 and $\tau^2 = 0.00$, where the calculated estimate is slightly greater than the empirical estimate. Therefore, the RVE method also leads to accurate inference for the intercept/effect size.

**TABLE 3** Estimation Performance of CP and Power of Intercept/Effect Size Summary Table (*m*=10)

| *m* | $\tau^2$ | *Slope* | **Method** | Calculated SD Compared to Empirical SD | Estimation Performance of Coverage Probability | Estimation of Performance of Power | **Inference** |
|---|---|---|---|---|---|---|---|
| 10 | 0.00 | 0.00 | **RVE** | Equal | Good | Good | **Good** |
| | | | **Naïve** | Equal | Good | Good | **Good** |
| | | | **Average** | Greater | Too wide | Too small | **Conservative** |
| | | | **Random** | Greater | Too wide | Too small | **Conservative** |
| 10 | 0.15 | 0.00 | **RVE** | Less | Too small | Too big | **Not trusted** |
| | | | **Naïve** | Less | Too small | Too big | **Not trusted** |
| | | | **Average** | Equal | Good | Good | **Good** |
| | | | **Random** | Equal | Good | Good | **Good** |
| | | 0.05 | **RVE** | Equal | Good | Good | **Good** |
| | | | **Naïve** | Less | Too small | Too big | **Not trusted** |
| | | | **Average** | - | - | - | **-** |
| | | | **Random** | Less | Too small | Too big | **Not trusted** |
| | | 0.10 | **RVE** | Equal | Good | Good | **Good** |
| | | | **Naïve** | Less | Too small | Too big | **Not trusted** |
| | | | **Average** | - | - | - | **-** |
| | | | **Random** | Greater | Too wide | Too small | **Conservative** |
| | | 0.15 | **RVE** | Equal | Good | Good | **Good** |
| | | | **Naïve** | Less | Too small | Too big | **Not trusted** |
| | | | **Average** | - | - | - | **-** |
| | | | **Random** | Less | Too small | Too big | **Not trusted** |
| 10 | 0.30 | 0.00 | **RVE** | Equal | Good | Good | **Good** |
| | | | **Naïve** | Less | Too small | Too big | **Not trusted** |
| | | | **Average** | Equal | Good | Good | **Good** |
| | | | **Random** | Equal | Good | Good | **Good** |
| | | 0.05 | **RVE** | Equal | Good | Good | **Good** |
| | | | **Naïve** | Less | Too small | Too big | **Not trusted** |
| | | | **Average** | - | - | - | **-** |
| | | | **Random** | Less | Too small | Too big | **Not trusted** |
| | | 0.10 | **RVE** | Equal | Good | Good | **Good** |
| | | | **Naïve** | Less | Too small | Too big | **Not trusted** |
| | | | **Average** | - | - | - | **-** |
| | | | **Random** | Less | Too small | Too big | **Not trusted** |
| | | 0.15 | **RVE** | Equal | Good | Good | **Good** |
| | | | **Naïve** | Less | Too small | Too big | **Not trusted** |
| | | | **Average** | - | - | - | **-** |
| | | | **Random** | Less | Too small | Too big | **Not trusted** |

Table 3: Summary Table of the performance of the coverage probability and power estimates for the intercept/effect size in the meta regression model, according to the differences in empirical and calculated standard deviations of the replicates; final inferences based on these comparisons across each method, *m* and $\tau^2$ level is shown; performance not assessed for simulations where *slope* greater than 0 and $\tau^2$ = 0.
Note: Average method not shown for *slope* greater than 0, since the estimates in these cases are biased.

**TABLE 4** Estimation Performance of CP and Power of Intercept/Effect Size Summary Table (*m*=50)

| *m* | $\tau^2$ | *Slope* | **Method** | Calculated SD Compared to Empirical SD | Estimation Performance of Coverage Probability | Estimation of Performance of Power | **Inference** |
|---|---|---|---|---|---|---|---|
| 50 | 0.00 | 0.00 | **RVE** | Equal | Good | Good | **Good** |
| | | | **Naïve** | Equal | Good | Good | **Good** |
| | | | **Average** | Greater | Too wide | Too small | **Conservative** |
| | | | **Random** | Less | Too small | Too big | **Not trusted** |
| 50 | 0.15 | 0.00 | **RVE** | Equal | Good | Good | **Good** |
| | | | **Naïve** | Less | Too small | Too big | **Not trusted** |
| | | | **Average** | Equal | Good | Good | **Good** |
| | | | **Random** | Equal | Good | Good | **Good** |
| | | 0.05 | **RVE** | Equal | Good | Good | **Good** |
| | | | **Naïve** | Less | Too small | Too big | **Not trusted** |
| | | | **Average** | - | - | - | - |
| | | | **Random** | Greater | Too wide | Too small | **Conservative** |
| | | 0.10 | **RVE** | Equal | Good | Good | **Good** |
| | | | **Naïve** | Less | Too small | Too big | **Not trusted** |
| | | | **Average** | - | - | - | - |
| | | | **Random** | Less | Too small | Too big | **Not trusted** |
| | | 0.15 | **RVE** | Equal | Good | Good | **Good** |
| | | | **Naïve** | Less | Too small | Too big | **Not trusted** |
| | | | **Average** | - | - | - | - |
| | | | **Random** | Equal | Good | Good | **Good** |
| 50 | 0.30 | 0.00 | **RVE** | Equal | Good | Good | **Good** |
| | | | **Naïve** | Less | Too small | Too big | **Not trusted** |
| | | | **Average** | Equal | Good | Good | **Good** |
| | | | **Random** | Equal | Good | Good | **Good** |
| | | 0.05 | **RVE** | Equal | Good | Good | **Good** |
| | | | **Naïve** | Less | Too small | Too big | **Not trusted** |
| | | | **Average** | - | - | - | - |
| | | | **Random** | Less | Too small | Too big | **Not trusted** |
| | | 0.10 | **RVE** | Equal | Good | Good | **Good** |
| | | | **Naïve** | Less | Too small | Too big | **Not trusted** |
| | | | **Average** | - | - | - | - |
| | | | **Random** | Less | Too small | Too big | **Not trusted** |
| | | 0.15 | **RVE** | Greater | Too wide | Too small | **Conservative** |
| | | | **Naïve** | Less | Too small | Too big | **Not trusted** |
| | | | **Average** | - | - | - | - |
| | | | **Random** | Equal | Good | Good | **Good** |

Table 4: Summary Table of the performance of the coverage probability and power estimates for the intercept/effect size in the meta regression model, according to the differences in empirical and calculated standard deviations of the replicates; final inferences based on these comparisons across each method, *m* and $\tau^2$ level is shown; performance not assessed for simulations where *slope* greater than 0 and $\tau^2$ = 0.
Note: Average method not shown for *slope* greater than 0, since the estimates in these cases are biased.

**TABLE 5** Estimation Performance of CP and Power of Intercept/Effect Size Summary Table (*m*=100)

| *m* | $\tau^2$ | *Slope* | **Method** | Calculated SD Compared to Empirical SD | Estimation Performance of Coverage Probability | Estimation of Performance of Power | **Inference** |
|---|---|---|---|---|---|---|---|
| 100 | 0.00 | 0.00 | **RVE** | Equal | Good | Good | **Good** |
| | | | **Naïve** | Equal | Good | Good | **Good** |
| | | | **Average** | Greater | Too wide | Too small | **Conservative** |
| | | | **Random** | Greater | Too wide | Too small | **Conservative** |
| 100 | 0.15 | 0.00 | **RVE** | Equal | Good | Good | **Good** |
| | | | **Naïve** | Less | Too small | Too big | **Not trusted** |
| | | | **Average** | Equal | Good | Good | **Good** |
| | | | **Random** | Equal | Good | Good | **Good** |
| | | 0.05 | **RVE** | Equal | Good | Good | **Good** |
| | | | **Naïve** | Less | Too small | Too big | **Not trusted** |
| | | | **Average** | - | - | - | - |
| | | | **Random** | Greater | Too wide | Too small | **Conservative** |
| | | 0.10 | **RVE** | Equal | Good | Good | **Good** |
| | | | **Naïve** | Less | Too small | Too big | **Not trusted** |
| | | | **Average** | - | - | - | - |
| | | | **Random** | Equal | Good | Good | **Good** |
| | | 0.15 | **RVE** | Equal | Good | Good | **Good** |
| | | | **Naïve** | Less | Too small | Too big | **Not trusted** |
| | | | **Average** | - | - | - | - |
| | | | **Random** | Greater | Too wide | Too small | **Conservative** |
| 100 | 0.30 | 0.00 | **RVE** | Equal | Good | Good | **Good** |
| | | | **Naïve** | Less | Too small | Too big | **Not trusted** |
| | | | **Average** | Equal | Good | Good | **Good** |
| | | | **Random** | Equal | Good | Good | **Good** |
| | | 0.05 | **RVE** | Equal | Good | Good | **Good** |
| | | | **Naïve** | Less | Too small | Too big | **Not trusted** |
| | | | **Average** | - | - | - | - |
| | | | **Random** | Equal | Good | Good | **Good** |
| | | 0.10 | **RVE** | Equal | Good | Good | **Good** |
| | | | **Naïve** | Less | Too small | Too big | **Not trusted** |
| | | | **Average** | - | - | - | - |
| | | | **Random** | Equal | Good | Good | **Good** |
| | | 0.15 | **RVE** | Equal | Good | Good | **Good** |
| | | | **Naïve** | Less | Too small | Too big | **Not trusted** |
| | | | **Average** | - | - | - | - |
| | | | **Random** | Greater | Too wide | Too small | **Conservative** |

Table 5: Summary Table of the performance of the coverage probability and power estimates for the intercept/effect size in the meta regression model, according to the differences in empirical and calculated standard deviations of the replicates; final inferences based on these comparisons across each method, *m* and $\tau^2$ level is shown; performance not assessed for simulations where *slope* greater than 0 and $\tau^2 = 0$.
Note: Average method not shown for *slope* greater than 0, since the estimates in these cases are biased.

### *Moderator*

When the naïve method is used across all $m$ and $\tau^2$ levels, the calculated standard deviation estimate is greater than the empirical standard deviation, therefore the coverage probability is greater than 0.95 and the power estimate is less than expected. Therefore, we can infer that the Naïve method is conservative across all $m$ and $\tau^2$ levels for inference regarding the moderator effect.

The calculated standard deviation estimates for the Random method are less than the empirical standard deviation estimates when the $\tau^2 = 0.30$. However, if $m = 100$ when the $\tau^2 = 0.30$, the SD estimates become roughly equal. Therefore, we can infer that the Random method is not appropriate when the $\tau^2 = 0.30$, except when the $m = 100$, but is otherwise good or conservative.

Using the RVE method, we find that across all $m$ and $\tau^2$ levels, the calculated standard deviation estimate is roughly equal to the empirical standard deviation estimate. Thus, the estimation performance of the coverage probability and power are appropriate, and we can infer that the RVE method leads to accurate inference for the moderator effect across all $m$ and $\tau^2$ levels.

Moreover, the power for detecting a non-zero intercept can only be trusted with the RVE and Random methods, but RVE has higher power to detect the non-zero effect. Power for detecting non-zero moderator effects is roughly equivalent for the RVE and Naïve methods, and is higher than the power from the Random method.

**TABLE 6** Estimation Performance of CP and Power of Moderator Effect Summary Table (*m*=10)

| *m* | $\tau^2$ | *Slope* | **Method** | Calculated SD Compared to Empirical SD | Estimation Performance of Coverage Probability | Estimation of Performance of Power | **Inference** |
|---|---|---|---|---|---|---|---|
| 10 | 0.15 | 0.00 | **RVE** | Equal | Good | Good | **Good** |
| | | | **Naïve** | Greater | Too wide | Too small | **Conservative** |
| | | | **Average** | - | - | - | **-** |
| | | | **Random** | Equal | Good | Good | **Good** |
| | | 0.05 | **RVE** | Equal | Good | Good | **Good** |
| | | | **Naïve** | Greater | Too wide | Too small | **Conservative** |
| | | | **Average** | - | - | - | **-** |
| | | | **Random** | Less | Too small | Too big | **Not trusted** |
| | | 0.10 | **RVE** | Equal | Good | Good | **Good** |
| | | | **Naïve** | Greater | Too wide | Too small | **Conservative** |
| | | | **Average** | - | - | - | **-** |
| | | | **Random** | Greater | Too wide | Too small | **Conservative** |
| | | 0.15 | **RVE** | Equal | Good | Good | **Good** |
| | | | **Naïve** | Greater | Too wide | Too small | **Conservative** |
| | | | **Average** | - | - | - | **-** |
| | | | **Random** | Less | Too small | Too big | **Not trusted** |
| 10 | 0.30 | 0.00 | **RVE** | Equal | Good | Good | **Good** |
| | | | **Naïve** | Greater | Too wide | Too small | **Conservative** |
| | | | **Average** | - | - | - | **-** |
| | | | **Random** | Equal | Good | Good | **Good** |
| | | 0.05 | **RVE** | Equal | Good | Good | **Good** |
| | | | **Naïve** | Greater | Too wide | Too small | **Conservative** |
| | | | **Average** | - | - | - | **-** |
| | | | **Random** | Less | Too small | Too big | **Not trusted** |
| | | 0.10 | **RVE** | Equal | Good | Good | **Good** |
| | | | **Naïve** | Greater | Too wide | Too small | **Conservative** |
| | | | **Average** | - | - | - | **-** |
| | | | **Random** | Less | Too small | Too big | **Not trusted** |
| | | 0.15 | **RVE** | Equal | Good | Good | **Good** |
| | | | **Naïve** | Greater | Too wide | Too small | **Conservative** |
| | | | **Average** | - | - | - | **-** |
| | | | **Random** | Equal | Good | Good | **Good** |

Table 6: Summary table of the performance of the coverage probability and power estimates for the moderator effect in the meta- regression model, according to the differences in empirical and calculated standard deviations of the replicates; final inferences based on these comparisons across each method, *m, slope* and $\tau^2$ level are shown; performance not assessed for simulations where $\tau^2 = 0$.
NOTE: Average method is not shown here as it does not have a time moderator in the model.

**TABLE 7** Estimation Performance of CP and Power of Moderator Effect Summary Table (*m*=50)

| *m* | $\tau^2$ | *Slope* | **Method** | Calculated SD Compared to Empirical SD | Estimation Performance of Coverage Probability | Estimation of Performance of Power | **Inference** |
|---|---|---|---|---|---|---|---|
| 50 | 0.15 | 0.00 | **RVE** | Equal | Good | Good | **Good** |
| | | | **Naïve** | Greater | Too wide | Too small | **Conservative** |
| | | | **Average** | - | - | - | - |
| | | | **Random** | Equal | Good | Good | **Good** |
| | | 0.05 | **RVE** | Equal | Good | Good | **Good** |
| | | | **Naïve** | Greater | Too wide | Too small | **Conservative** |
| | | | **Average** | - | - | - | - |
| | | | **Random** | Greater | Too wide | Too small | **Conservative** |
| | | 0.10 | **RVE** | Equal | Good | Good | **Good** |
| | | | **Naïve** | Greater | Too wide | Too small | **Conservative** |
| | | | **Average** | - | - | - | - |
| | | | **Random** | Less | Too small | Too big | **Not trusted** |
| | | 0.15 | **RVE** | Equal | Good | Good | **Good** |
| | | | **Naïve** | Greater | Too wide | Too small | **Conservative** |
| | | | **Average** | - | - | - | - |
| | | | **Random** | Equal | Good | Good | **Good** |
| 50 | 0.30 | 0.00 | **RVE** | Equal | Good | Good | **Good** |
| | | | **Naïve** | Greater | Too wide | Too small | **Conservative** |
| | | | **Average** | - | - | - | - |
| | | | **Random** | Equal | Good | Good | **Good** |
| | | 0.05 | **RVE** | Equal | Good | Good | **Good** |
| | | | **Naïve** | Greater | Too wide | Too small | **Conservative** |
| | | | **Average** | - | - | - | - |
| | | | **Random** | Less | Too small | Too big | **Not trusted** |
| | | 0.10 | **RVE** | Equal | Good | Good | **Good** |
| | | | **Naïve** | Greater | Too wide | Too small | **Conservative** |
| | | | **Average** | - | - | - | - |
| | | | **Random** | Less | Too small | Too big | **Not trusted** |
| | | 0.15 | **RVE** | Equal | Good | Good | **Good** |
| | | | **Naïve** | Greater | Too wide | Too small | **Conservative** |
| | | | **Average** | - | - | - | - |
| | | | **Random** | Equal | Good | Good | **Good** |

Table 7: Summary table of the performance of the coverage probability and power estimates for the moderator effect in the meta- regression model, according to the differences in empirical and calculated standard deviations of the replicates; final inferences based on these comparisons across each method, *m*, *slope* and $\tau^2$ level are shown; performance not assessed for simulations where $\tau^2 = 0$.
NOTE: Average method is not shown here as it does not have a time moderator in the model.

**TABLE 8** Estimation Performance of CP and Power of Moderator Effect Summary Table ($m$=100)

| $m$ | $\tau^2$ | *Slope* | **Method** | Calculated SD Compared to Empirical SD | Estimation Performance of Coverage Probability | Estimation of Performance of Power | **Inference** |
|---|---|---|---|---|---|---|---|
| 100 | 0.15 | 0.00 | **RVE** | Equal | Good | Good | **Good** |
| | | | **Naïve** | Greater | Too wide | Too small | **Conservative** |
| | | | **Average** | - | - | - | - |
| | | | **Random** | Less | Too small | Too big | **Not trusted** |
| | | 0.05 | **RVE** | Equal | Good | Good | **Good** |
| | | | **Naïve** | Greater | Too wide | Too small | **Conservative** |
| | | | **Average** | - | - | - | - |
| | | | **Random** | Less | Too small | Too big | **Not trusted** |
| | | 0.10 | **RVE** | Equal | Good | Good | **Good** |
| | | | **Naïve** | Greater | Too wide | Too small | **Conservative** |
| | | | **Average** | - | - | - | - |
| | | | **Random** | Equal | Good | Good | **Good** |
| | | 0.15 | **RVE** | Equal | Good | Good | **Good** |
| | | | **Naïve** | Greater | Too wide | Too small | **Conservative** |
| | | | **Average** | - | - | - | - |
| | | | **Random** | Greater | Too wide | Too small | **Conservative** |
| 100 | 0.30 | 0.00 | **RVE** | Equal | Good | Good | **Good** |
| | | | **Naïve** | Greater | Too wide | Too small | **Conservative** |
| | | | **Average** | - | - | - | - |
| | | | **Random** | Equal | Good | Good | **Good** |
| | | 0.05 | **RVE** | Equal | Good | Good | **Good** |
| | | | **Naïve** | Greater | Too wide | Too small | **Conservative** |
| | | | **Average** | - | - | - | - |
| | | | **Random** | Equal | Good | Good | **Good** |
| | | 0.10 | **RVE** | Greater | Too wide | Too small | **Conservative** |
| | | | **Naïve** | Greater | Too wide | Too small | **Conservative** |
| | | | **Average** | - | - | - | - |
| | | | **Random** | Equal | Good | Good | **Good** |
| | | 0.15 | **RVE** | Greater | Too wide | Too small | **Conservative** |
| | | | **Naïve** | Greater | Too wide | Too small | **Conservative** |
| | | | **Average** | - | - | - | - |
| | | | **Random** | Equal | Good | Good | **Good** |

Table 8: Summary table of the performance of the coverage probability and power estimates for the moderator effect in the meta- regression model, according to the differences in empirical and calculated standard deviations of the replicates; final inferences based on these comparisons across each method, *m, slope* and $\tau^2$ level are shown; performance not assessed for simulations where $\tau^2$ = 0.
NOTE: Average method is not shown here as it does not have a time moderator in the model.

### *Estimated $\tau^2$ and $I^2$*

We also assess the performance of these meta-analysis methods in terms of their estimated $\tau^2$ and $I^2$ values. If a method performed well in terms of estimating the heterogeneity, we would expect the $\tau^2$ estimates to be approximately equal to the values set in during the generation of the replicates, which is represented with a black line in Figures A1, A2, and A3 in the Appendix. While $I^2$ does not have a theoretical counterpart with the current simulation design, the estimates can be used to help calibrate inferences in real datasets.

By comparing the estimated $\tau^2$ to the set (simulation) values, we can observe that out of all four methods, the median estimated $\tau^2$ value across all replicates using RVE method is closest to the true $\tau^2$ value we set in the simulation. The only exception to this finding is when the true $\tau^2 = 0.00$, in which we can see that the median estimated $\tau^2$ value using the RVE method is greater than 0. We found that the Average and Naïve methods tended to underestimate heterogeneity, particularly when m is small. We also observed that as m or slope increases, all methods improve, while RVE still performs the best. In addition, as $\tau^2$ increases, $I^2$ estimates become more similar across the methods.

### *Sensitivity Analysis*

We also assessed the performance of the RVE method by conducting a sensitivity analysis for each of the simulation scenarios. However, we found that none of the RVE analyses across all 1000 replicates in each of the 36 scenarios were marked as sensitive. Therefore, we can conclude that the RVE method was not sensitive to the choice of correlation parameter (*rho*) specified in the robu() function.

### *A Summary of the Simulation Results*

*BIAS (AND RELATED)*

- Since the Average method does not include a moderator, we observe the bias and CP performance degrades for the intercept estimation in the Average method as the true value of the slope (used for data generation) increases.
- Naive, Random, RVE performed about the same in terms of bias in both the intercept and moderator effect estimation, with Naïve and RVE performing slightly better.

*SDs (AND RELATION TO CP AND POWER)*

- Naïve method:
    - For the intercept effect: the observed power is higher than it should be, and the CP is smaller than .95 (test statistics are too big and confidence intervals are too small).

      *Therefore, inference (tests, confidence intervals, p-values, power, CI endpoints) for the intercept based on the Naïve method cannot be trusted.*
    - For the moderator effect: the observed power is be lower than it should be (conservative) and the CP is higher than .95 (test statistics are too small and confidence intervals are too wide).
- Random method: when $\tau^2$ is non-zero, there is little difference between the SD methods for both intercept and moderator, so inferences can be trusted. However, when $\tau^2$=0 (viewed as uncommon in practice), the power of the intercept is lower than it should be (conservative) and the CP is higher than it should be. The SDs (both methods) for intercept and moderator estimates tend to be larger using the

Random method than in the other methods, which contributes to its relatively low power.

- Average method: behaves similarly to Random with respect to SDs for intercept: when $\tau^2$ is non-zero, there is little difference between the SD methods for intercept. However when $\tau^2=0$, the empirical SD tends to be lower than the calculated SD, *Since the Average method produces biased estimates when the true data-generating slope is not equal to 0, inference for Average is acceptable (trustworthy or conservative) only when there is no moderator effect.*
- RVE: SDs from both methods for both intercept and moderator estimators are very similar, thus CP is about .95 (and improves as m gets bigger) and inferences can be trusted.

*MORE ON POWER*

- The true data-generating intercept is constant (and non-zero at 0.5) across all simulation scenarios, and we observe high power for detecting this non-zero effect.
- Across all methods, power of moderator test increases as true (data-generating) slope increases.
- Power for detecting non-zero intercept and non-zero slope decreases as $\tau^2$ increases.
- Power results for detecting non-zero intercept can only be trusted with the RVE and Random methods, and RVE has higher power.
- Power for detecting non-zero moderator effects is the highest (and similar) for the RVE and Naïve methods; power for the Random method is the lowest.

- *In meta-regression models containing both intercept and time moderator effects, RVE is the most appropriate and generally the most powerful method to use.*

*ESTIMATING HETEROGENEITY*

- All methods improve in terms of $\tau^2$ estimation as m increases, but the RVE method estimates $\tau^2$ most accurately.
- The Average method median estimates of $I^2$ and $\tau^2$ are much smaller than those from the other methods, suggesting that the Average method underestimates the heterogeneity of the between-study effects.
- All methods improve in terms of $\tau^2$ estimation as the true data-generating slope increases, but the RVE method estimates $\tau^2$ most accurately.

## SECTION 6: APPLICATION OF RVE AND CLASSICAL AVERAGE META-ANALYSIS

In order to show the effect that different meta-analysis methods have on effect sizes, we compare the RVE findings of the Baffoni meta-analysis[11], to the results of the classical Average method across the same dataset. From the simulation, we found that the RVE method performed better than the classical Average method in the presence of a time moderator. If there is, in fact, no significant moderator effect, we should expect to see similar effect size results when using the classical Average and RVE methods based on the same dataset, although we expect that the heterogeneity will be underestimated for the classical Average method as compared to the true heterogeneity.

### *RVE*

Following the RVE procedures used in the Baffoni meta-analysis[11], we reproduced the results across the dataset both with and without a time moderator in the meta-regression model, with the results shown in Table 9. Based on the analysis with the time moderator that included the 92 time points from the studies summarized in Table 1, there were significant reductions in the intercept effect of SBP ($\hat{d}$= -1.12; -12.36 mmHg, 95%*CI:* [-23.51, -1.21]; p=0.034) and trending reductions in the intercept effect of DBP ($\hat{d}$= -1.10; -5.56 mmHg, 95%*CI:* [-11.67, 0.51]; p=0.067) following acute concurrent exercise as compared to the control. In the same meta-regression model, there are no significant reductions in the effect of SBP due to the time moderator (slope estimate=-0.003; -0.03 mmHg, 95%*CI:* [-0.19, 0.12]; p=0.636) and no significant reductions in the effect of DBP due to the time moderator (slope estimate = 0.003; 0.015 mmHg, 95%*CI:* [-0.71, 0.10]; p=0.69), following acute concurrent exercise as compared to the control. Heterogeneity was large across the meta-regression models for both SBP ($I^2$=89.17%, $\tau^2$=1.25) and DBP ($I^2$=85.59%, $\tau^2$=0.83).

As the time moderator effects were not significant for SBP or DBP, we proceed with the model without the time moderator. The observed reductions in the overall effect, without the time moderator, in the Baffoni meta-analysis[11], were significant in SBP ($\hat{d}$= -1.25; -13.8 mmHg, 95%*CI:* [-19.8, -7.6]; p=0.0005) and DBP ($\hat{d}$= -0.97; -4.9 mmHg, 95%*CI:* [-8.1, -1.7]; P=.007) following acute concurrent exercise as compared to the control. Heterogeneity remained large across the meta-analysis models for both SBP ($I^2$=88.79%, $\tau^2$=1.18) and DBP ($I^2$=85.07%, $\tau^2$=0.78)[11].

### *Classical Average*

Similarly, using the classical average method with no moderator effect, and assuming independence of the 19 effect sizes from the 11 studies, there was a significant reduction both in SBP ($\hat{d}$= -1.09; -12.03 mmHg, 95%*CI:* [-15.12, -8.94], p<0.001) and DBP ($\hat{d}$= -1.03; -5.20 mmHg, 95%*CI:* [-7.71, -2.65], p<0.001). It is important to note that in reality, the independence assumption is violated here because multiple treatment groups are compared to the same control group within several of the studies. Estimated heterogeneity was smaller than estimated from the RVE method for SBP ($I^2$=63.7%, $\tau^2$=0.24), but larger for DBP ($I^2$=88.23%, $\tau^2$=1.06).

We also ran the classical average random effects meta-analysis method while satisfying the independence assumption by randomly choosing one treatment group per study ($m$=11). We found that there was a greater reduction in both mean SBP ($\hat{d}$= -1.14; -12.59 mmHg, 95%*CI:* [-17.69, -7.58], p<0.001) and mean DBP ($\hat{d}$= -1.47; -7.42 mmHg, 95%*CI:* [-11.47, 3.36], p<0.001) than when the independence assumption was violated.

**TABLE 9** Applied Comparison of Classical Average and RVE Meta-Analysis

| Description | *m* | Number of Outcomes | *Est.* | SE | t value (df) | p (\|t\| >) | 95% CI | $I^2$(%) | $\tau^2$ | $\hat{\Delta}$ (mm Hg) |
|---|---|---|---|---|---|---|---|---|---|---|
| **Systolic Blood pressure** | | | | | | | | | | |
| **RVE** | | | | | | | | | | |
| Overall (no moderator) | 11 | 92 | -1.25 | 0.25 | -5.01 (9.98) | .0005 | [-1.80, -0.69] | 88.79 | 1.18 | -13.8 |
| Intercept | 11 | 92 | -1.12 | 0.42 | -2.64 (6.79) | 0.034 | [-2.13, -0.11] | 89.17 | 1.25 | -12.36 |
| Moderator | | | -0.003 | 0.006 | -0.50 (5.97) | 0.636 | [-0.017, 0.011] | | | -0.03 |
| **All studies Classical Average** | | | | | | | | | | |
| Overall | 11 | 19 | -1.09 | 0.14 | -7.63 (18) | <0.001 | [-1.37, -0.81] | 63.7 | 0.24 | -12.03 |
| **Independent Treatments Classical Average** | | | | | | | | | | |
| Overall | 11 | 11 | -1.14 | 0.233 | -4.9 (10) | <0.001 | [-1.6024; -0.6867] | 77.1 | 0.4570 | -12.59 |
| **Diastolic Blood Pressure** | | | | | | | | | | |
| **RVE** | | | | | | | | | | |
| Overall (no moderator) | 11 | 92 | -0.97 | 0.29 | -3.4 (9.97) | 0.007 | [-1.61, -0.34] | 85.07 | 0.78 | -4.9 |
| Intercept | 11 | 92 | -1.10 | 0.51 | -2.18 (6.71) | 0.067 | [-2.31, 0.10] | 85.59 | 0.83 | -5.56 |
| Moderator | | | 0.003 | 0.007 | 0.42 (6.01) | 0.69 | [-0.014, 0.019] | | | 0.015 |
| **All studies CLASSICAL AVERAGE** | | | | | | | | | | |
| Overall | 11 | 19 | -1.03 | 0.26 | -4.02 (18) | <0.001 | [-1.5267, -0.5255] | 88.23 | 1.06 | -5.20 |
| **Independent Treatments Classical Average** | | | | | | | | | | |
| Overall | 11 | 11 | -1.47 | 0.41 | -3.58 | <0.001 | [-2.2713; -0.6654] | 91.7 | 1.6374 | -7.42 |

Table 9: Provides a comparison of the results from RVE and classical average methods across real studies, taken from Baffoni RVE results

*Note*. *m* = number of studies included in meta-regression. Est.= estimated effects, corrected for baseline difference. CI = confidence interval. $\hat{\Delta}$ = predicted change. Standardized mean difference effect size was transformed to $\hat{\Delta}$ (mm Hg) using the standard deviation corresponding to the sample mean baseline SBP and DBP for each dataset.

In the Baffoni RVE meta-analysis[11] without a moderator, *m*=11 and $I^2 = 88.79\%$ and 85.07% for SBP and DBP, respectively. Within the *m*=10 level and using the RVE method, the $I^2$ levels 88.79% and 85.07% are most similar to the $\tau^2 = 0.30$ simulation setting. Therefore, we will draw interpretations of the methods based on our highest level: $\tau^2$

=0.30. Because Baffoni did not detect a significant time moderator in the meta-regression model, which is likely due to the small sample size, we will use the simulation setting *slope* = 0.00. When $m = 10$, $\tau^2$=0.30 and the *slope* =0.00, we observed that both the RVE and average classical methods performed well in terms of estimation and inference, which explains why the results are similar in Table 9.

In particular, when $m = 10$, $\tau^2 = 0.30$ and there is no significant time moderator, the Average method is appropriate. However, if a significant time moderator were to exist in the meta-regression model, the performance of the Average method worsens and RVE method improves. In addition, when $m = 10$ and there is no significant moderator effect, the Average method also performed better than RVE in the simulations in terms of power for non-zero $\tau^2$ values.

As expected, for SBP, we can see that the estimated heterogeneity using the RVE method is greater than that of the classical Average method, both where the all treatment groups are included, and where only independent treatment groups are included. However, this finding does not hold in the case of DBP. When $m = 10$, $\tau^2$=0.30 and the *slope* =0.00, we observed from the simulations that the variability in the $\tau^2$ and $I^2$ estimates is very high using the classical Average method with many high outlying values. It is possible that the results for DBP correspond to one of these “outlying'' cases in term of its heterogeneity estimate, but this warrants further investigation.

We suspect that the Baffoni meta-analysis[11] did not detect a significant time moderator in the meta-regression model because of the choice of the time ranges used. Baffoni classified time as three ranges: studies that measured BP outcomes from 0 to 40 minutes, 40 to 80 minutes and 80 to 120 minutes after control or exercise intervention (3

40-minute increments over 2 hours). Perhaps with larger time ranges, a significant moderator effect could have been detected.

As Baffoni[11] addresses, this insignificant moderator is also likely due to the small sample number of studies. If more studies were included in the meta-analysis and $m$ increased while $\tau^2$ remains high and no significant time moderator is introduced into the model, the Average method worsens (in terms of bias) and the RVE method improves.

A second possible explanation for an insignificant moderator in this analysis is that time may have a non-linear effect. Here, we assume that the relationship between the BP effect size difference between control and experimental groups, and time BP was collected after intervention, was a linear. However, it is possible that this assumption is incorrect and therefore warrants further analysis.

## SECTION 7: DISCUSSION

The purpose of this study was to compare the performance of RVE and classical random meta-analytic methods under realistic simulation scenarios in the context of PEH and beyond. In comparing the performance of these methods, we applied these findings to a previous meta-analysis "estimating the BP response to a single bout of concurrent exercise"[11] and made recommendations for future meta-analyses.

Meta-analysis offers further insight into the implications of PEH and assesses the findings across multiple studies to determine an overall BP effect. RVE meta-analysis addresses the dependency in effect sizes without making assumptions about the specific form of the sampling distributions of the effect sizes or requiring knowledge of the covariance structure of the dependent estimates[5,14]. To show the reliability of the RVE method, we compared its performance in simulation to that of commonly used variations of classical random effects meta-analysis: Average, Naïve, and Random. In addition, the time BP was collected after exercise may influence the overall BP effect, and was therefore included as a potential moderator in our simulation design.

To assess the performance of these methods, we calculated the bias, standard deviation, coverage probability, and power of the overall effect and the moderator effects. Comparing the software-calculated standard deviation estimated by the four methods to the empirical standard deviation of the replicates, we determined how the coverage probability and power were affected across each method.

As the Average method did not include time of measurement as a moderator, we found that as the true moderator effect, or slope, increased, the bias and coverage probability performance degraded for the estimation of the overall effect size. Meanwhile, the Naïve, Random and RVE methods were similarly able to reduce the bias of the intercept and moderator effects, with RVE and Naïve doing slightly better. We found that based on the standard deviation estimates, that the Naïve method cannot be trusted for inference regarding the overall effect, but is conservative for inference regarding the moderator effect. Based on the standard deviation estimates from the RVE method, the coverage probability and power estimates from the RVE and random methods can be consistently trusted, while RVE has higher power.

To better understand the clinical implications of these simulation results, we conducted the classical average random effects model over the same dataset used to complete the previous Baffoni[11] RVE meta-analysis. We found that when using the classical Average random effects method, there was a significant difference between the intervention and control groups (independent) both mean SBP ($\hat{d}$= -1.14; -12.59 mmHg, 95%*CI:* [-17.69, -7.58], p<0.001) and mean DBP ($\hat{d}$= -1.47; -7.42 mmHg, 95%*CI:* [-11.47, 3.36], p<0.001). In the RVE meta-analysis without a time moderator, Baffoni[11] found that there were significant reductions in SBP ($\hat{d}$= -1.25; -13.8 mmHg, 95%*CI:* [-19.8, -7.6; p<.001) and DBP ($\hat{d}$= -0.97; -4.9 mmHg, 95%*CI:* [-8.1, -1.7]; p=.007) following acute concurrent exercise as compared to the control[11].

We suspect that the similarity between effect sizes is because the small number of repeated observations over a small time range is not large enough to detect a significant moderator[16]. According to the simulation results, when the slope or significance of the moderator effect increases and all other variables are held constant, the RVE method becomes more appropriate than the Average method. In addition, if the number of studies and moderator effect were held constant and the between study variance increased, the

Average method worsens and the RVE method improves. It is also likely that with a larger number of studies, the time moderator could become significant, in which case the RVE method would also be preferred over the Average method.

We can conclude that when the number of studies is small ($m$=11) and there is not a significant moderator, the commonly used classical Average random effects analysis performs as well as the RVE method. However, as the number of studies increases, between study variance increases, or a significant moderator is introduced, using the RVE method is preferred over the three variations of the classical random effects meta-analysis methods considered. Lastly, it is possible that the performance of the Average method could improve with the introduction of an "average time" moderator (e.g., the average time over which the measurements were taken), however we expect that this would underestimate the effect of the moderator and that the RVE method would still perform better. As we did not include a moderator in the classical Average meta-analysis simulations, we leave this topic for further investigation.

# REFERENCES

1. Varini S, Slonso A. Heart disease and stroke Statistics—2020 update: A report from the american heart association. *Circulation*. 2020;141(9):e139-e596. https://www.ahajournals.org/doi/10.1161/CIR.0000000000000757. Accessed Jun 2, 2020. doi: 10.1161/CIR.0000000000000757.

2. Whelton PK, Carey RM, Aronow WS, et al. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: Executive summary: A report of the american college of cardiology/american heart association task force on clinical practice guidelines. *Hypertension*. 2018;71(6):1269-1324. Accessed Jun 2, 2020. doi: 10.1161/HYP.0000000000000066.

3. Pescatello L, MacDonald H, Lamberti L, Johnson B. Exercise for hypertension: A prescription update integrating existing recommendations with emerging research. *Curr Hypertens Rep*. 2015;17(11):1-10. https://www.ncbi.nlm.nih.gov/pubmed/26423529. doi: 10.1007/s11906-015-0600-y.

4. New ACC/AHA high blood pressure guidelines lower definition of hypertension. American College of Cardiology Web site. http%3a%2f%2fwww.acc.org%2flatest-in-cardiology%2farticles%2f2017%2f11%2f08%2f11%2f47%2fmon-5pm-bp-guideline-aha-2017. Accessed May 9, 2020.

5. Hedges LV, Tipton E, Johnson MC. Robust variance estimation in meta-regression with dependent effect size estimates. *Res Synth Methods*. 2010;1(1):39-65. Accessed May 10, 2020. doi: 10.1002/jrsm.5.

6. Hypertension prevalence in the U.S. | million hearts®. Centers for Disease Control and Prevention Web site. https://millionhearts.hhs.gov/data-reports/hypertension-prevalence.html. Updated 2020. Accessed May 9, 2020.

7. Essential hypertension. *Circulation*. 2000;101(3):329-335. https://www.ahajournals.org/doi/full/10.1161/01.CIR.101.3.329. Accessed May 9, 2020. doi: 10.1161/01.CIR.101.3.329.

8. Postexercise hypotension. key features, mechanisms, and clinical significance. *Hypertension*. 1993;22(5):653-664. https://www.ahajournals.org/doi/abs/10.1161/01.HYP.22.5.653. Accessed May 9, 2020. doi: 10.1161/01.HYP.22.5.653.

9. Corso LML, Macdonald HV, Johnson BT, et al. Is concurrent training efficacious antihypertensive therapy? A meta-analysis. *Med Sci Sports Exerc*. 2016;48(12):2398-2406. Accessed Jun 2, 2020. doi: 10.1249/MSS.0000000000001056.

10. Carpio-Rivera E, Moncada-Jiménez J, Salazar-Rojas W, Solera-Herrera A. Acute effects of exercise on blood pressure: A meta-analytic investigation. *Arq Bras Cardiol*. 2016;106(5):422-433. Accessed May 9, 2020. doi: 10.5935/abc.20160064.

11. Baffoni G. *The Immediate Blood Pressure Response to Acute Concurrent Exercise: A Meta-Analysis Update.* Spring 2019.

12. Johnson BT, Huedo-Medina TB. *Meta-analytic statistical inferences for continuous measure outcomes as a function of effect size metric and other assumptions.* Rockville (MD): Agency for Healthcare Research and Quality (US); 2013. http://www.ncbi.nlm.nih.gov/books/NBK140575/. Accessed Jun 2, 2020.

13. Hennessy EA, Johnson BT, Keenan C. Best practice guidelines and essential methodological steps to conduct rigorous and systematic meta-reviews. *Applied Psychology: Health and Well-Being*. 2019;11(3):353-381. https://iaap-journals.onlinelibrary.wiley.com/doi/abs/10.1111/aphw.12169. Accessed Jun 2, 2020. doi: 10.1111/aphw.12169.

14. Tanner-Smith E, Tipton E, Polanin J. Handling complex meta-analytic data structures using robust variance estimates: A tutorial in R. *Journal of Developmental and Life-Course Criminology*. 2016;2:85-112. Accessed May 10, 2020. doi: 10.1007/s40865-016-0026-5.

15. Tipton E. Small sample adjustments for robust variance estimation with meta-regression. *Psychol Methods*. 2015;20(3):375-393. Accessed Jun 4, 2020. doi: 10.1037/met0000011.

16. Pescatello LS, Guidry MA, Blanchard BE, et al. Exercise intensity alters postexercise hypotension. *J Hypertens*. 2004;22(10):1881-1888. Accessed Jun 2, 2020. doi: 10.1097/00004872-200410000-00009.

# APPENDIX 1

**m = 10**

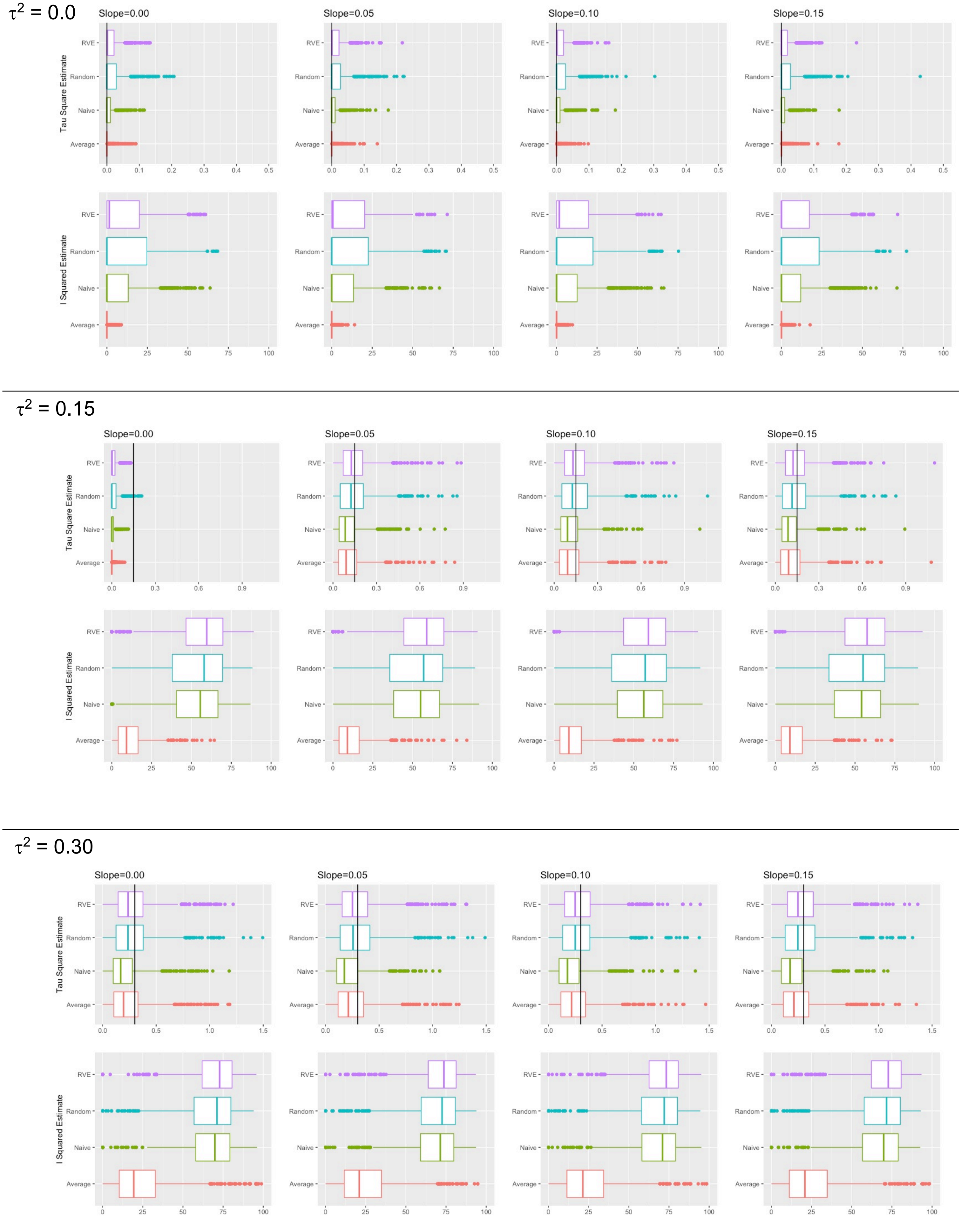


**Figure A1** Heterogeneity Estimation performance of four methods, where m=10.

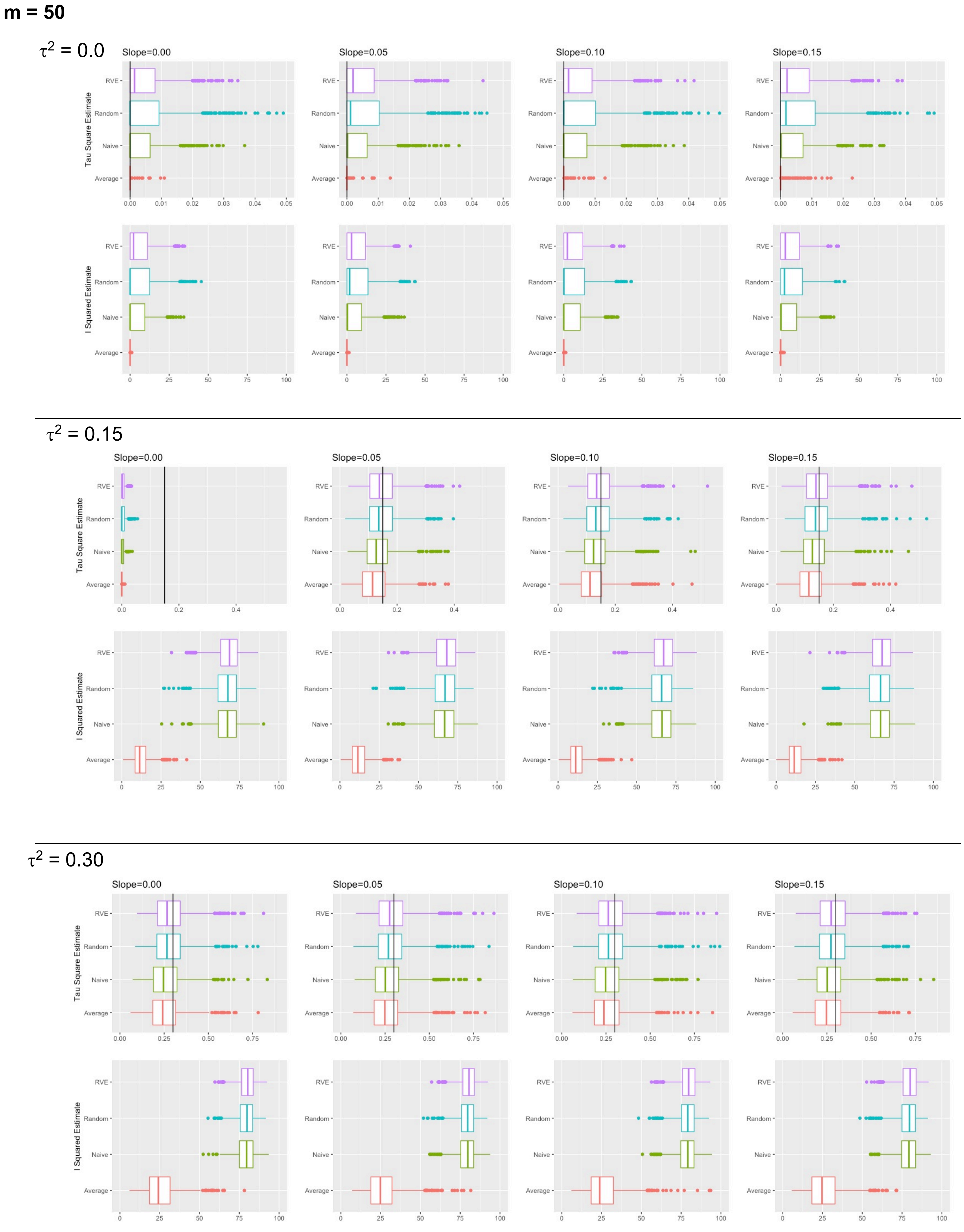
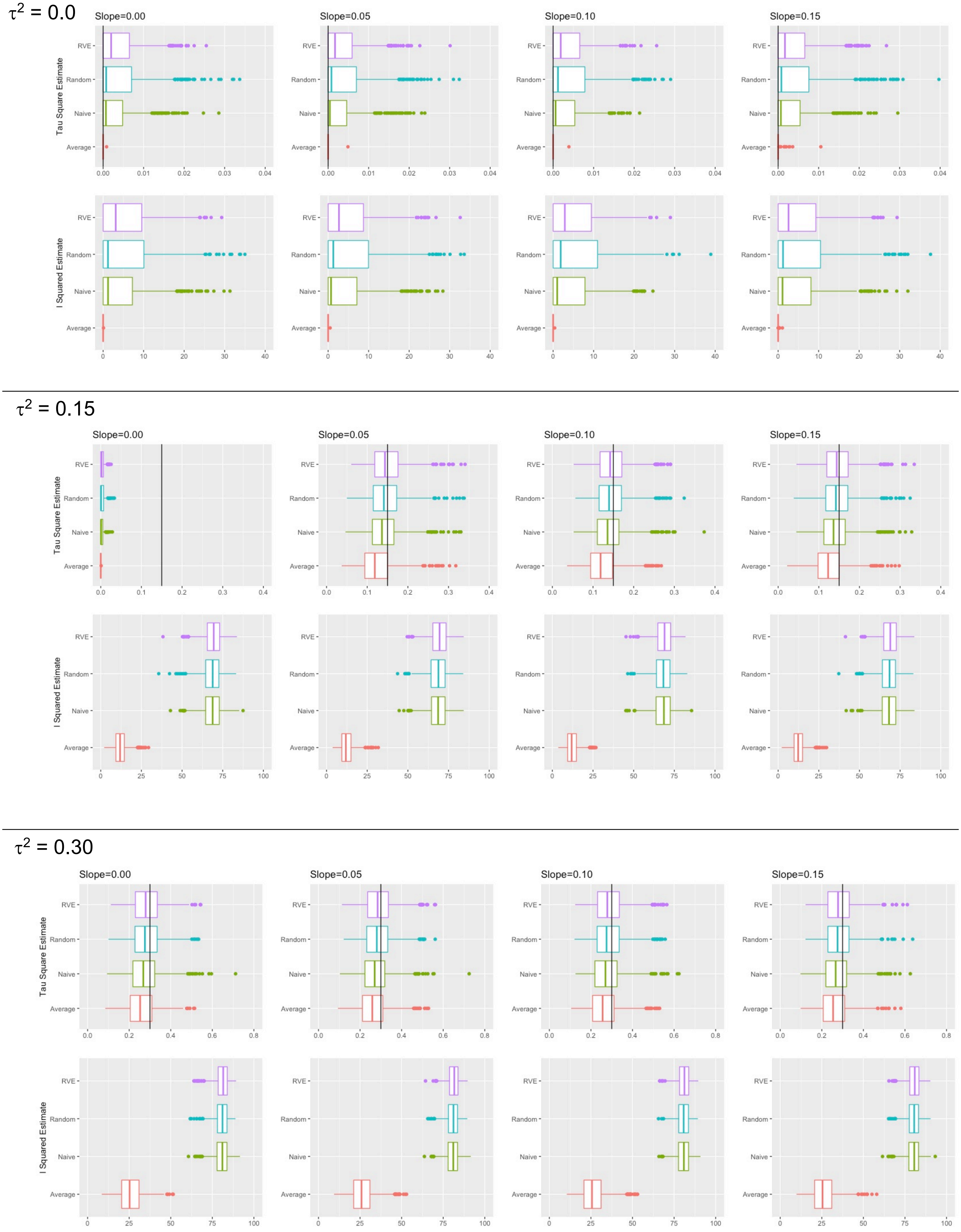


**Figure A2** Heterogeneity Estimation performance of four methods, where m=50

**m = 100**



$\tau^2$ = 0.0

$\tau^2$ = 0.15

$\tau^2$ = 0.30

**Figure A3** Heterogeneity Estimation performance of four methods, where m=100

# APPENDIX 2

R Code

```
library(robumeta)
library(meta)
library(dplyr)
library(multiApply)
source("helper.r")

# Consider only independent subjects studies (equal sample size in
control and treatment groups);
# as discussed at last meeting, if we go with the Hedges recommendation,
it ignores the correlation
# in matched/paired anaylysis when computing the estimate of delta, so
this seems inconsequential for
# our purposes.
# The following parameters would be fixed at one value within a complete
simulation study (consisting
# of 1000 replicate/simulated datasets):
#  m= 10,50,250
#  delta= .5 (fixed; do we need to vary this?)
#  slope= 0, 0.05, .1, .15  # may need to recalibrate?
#  tau2 = .04, .08, .16, .32  # (do we need to include 0?)
#  kgrid = seq(.25,3,by=.25)
#  (3 * 4* 4 = 48 simulated scenarios)
# vary within each replicate/simulated dataset
#    n= 10,20,200
#        with proportions nprop= .3,.6,.1 of m
#    k= 2, 4, 8 with time points selected randomly and with equal
probability from kgrid
#        with proportions kprop= .3,.6,.1 of m
#    rho = .2, .8
#        with proportions rhoprop=.5,.5 of m

# comments:
# 1. previously discussed m=10,50,500; 250 would be take substantially
less time to run
# than 500
# 2. previously discussed k = 2, 4, 8, 16, 32 but with 2 hr time window
(like in BP study)
# this becomes challenging (unless we make the time points not equally
spaced?);
# ultimately I went with time window of 3 hours, and selecting either 2,
4, or 8,
# but both could be adjusted (e.g., kgrid = seq(.25,5,by=.25) with
k=2,4,8,16), with kprop
# and possible also slope adjusted accordingly

# NOTE: not considering non-linear effects of moderator at this time

#=====================================================================
pooled.sd <- function(vec,nt,nc){
  t <- vec[1:nt]; c<-vec[(nt+1):(nt+nc)]
  sd <- sqrt( ((nt-1)*var(t) + (nc-1)*var(c))/(nt + nc -2) )
}


get.cover <- function(x,p){ifelse(x[5]<= p & x[6]>=p, 1, 0) }


# genData function will build the parameters required for the replicates
genData <- function(m, nvec, kvec, delta, slope, tau2, rhovec,
                    kprop, rhoprop, nprop, kgrid ) {
  #m= number of studies
  #nvec= vector of # of observations in control(=# obs in treatment
group)
  #kvec= vector of # repeated obs for each subject (constant within
study)
  #delta=effect size (intercept in metaregression)
  #slope = slope for moderator in metaregression
  #tau2 = variance of effect size
  #rhovec = vector of correlation of repeated observations
  #kprop = vector indicating proportion of studies with different k
values
  #rhoprop = vector indicating proportion of studies with different rho
values
  #nprop = vector indicating proportion of studies with different rho
values
  #kgrid = timepoints to select from

  # determine number of effect sizes based on m,k,and kprop
  N = sum(m*kprop*kvec)
  Ts <- mods <- varTs<- rep(0,N)
  mn.e <- mn.c <- sd.e <- sd.c <- rep(0,N)
  eta <- rnorm(m,0,sqrt(tau2))

  # create vectors of length m indicating the relevant n, k, rho values
for each study
  ns <- as.numeric(unlist( apply(cbind(nvec,m*nprop), 1, function(x)
rep(x[1],x[2])) ))
  ks <- as.numeric(unlist( apply(cbind(kvec,m*kprop), 1, function(x)
rep(x[1],x[2])) ))
  rhos <- as.numeric(unlist( apply(cbind(rhovec,m*rhoprop), 1,
function(x) rep(x[1],x[2])) ))
  # shuffle ks and rhos so that low ns do not always get low ks and low
rhos
  ks <- sample(ks,m)
  rhos <- sample(rhos,m)
  # to be used later for indexing
  inds <- c(0,cumsum(ks))

  for (l in 1:m){
   n<- ns[l]; k<-ks[l]; rho<-rhos[l]
   xi <- rnorm(2*n, 0, sqrt(rho))
   zeta <- rnorm(2*n*k, 0, sqrt(1-rho))
   x <- sample(kgrid,k)
   Yt <- matrix(0,n,k)
   Yc <- matrix(0,n,k)
   for (i in 1:n){
    for (j in 1:k){
      Yt[i,j] <- delta + slope*x[j] + eta[l] + xi[i] + zeta[(i-1)*k+j]
      Yc[i,j] <- xi[n+i] + zeta[n*k+(i-1)*k+j]
    }
   }
   meandiffs <- colMeans(Yt) - colMeans(Yc)
   psds <- apply(rbind(Yt,Yc),2,pooled.sd, n,n)
   effests<-meandiffs/psds
   Ts[(inds[l]+1):inds[l+1]] <-  effests
   mods[(inds[l]+1):inds[l+1]] <- x
   varTs[(inds[l]+1):inds[l+1]] <- 2/n +  effests^2/(2*(2*n-2))
   mn.e[(inds[l]+1):inds[l+1]] <- colMeans(Yt)
   mn.c[(inds[l]+1):inds[l+1]] <-colMeans(Yc)
   sd.e[(inds[l]+1):inds[l+1]] <- apply(Yt,2,sd)
   sd.c[(inds[l]+1):inds[l+1]] <- apply(Yc,2,sd)
  }
  id = as.numeric(unlist( apply(cbind(1:m,ks), 1, function(x)
rep(x[1],x[2])) ))
  samps = as.numeric(unlist( apply(cbind(ns,ks), 1, function(x)
rep(x[1],x[2])) ))
  data.frame(studyid = id, effectsize=Ts, var.effsize=varTs, mod=mods,
             samp.size =samps, mean.e = mn.e, mean.c=mn.c, sd.e = sd.e,
sd.c=sd.c)
}
# ====

# myFit function will generate simulations based on the parameters
chosen here (using the previous genData function),
# run the simulations through the four methods and present the results
in a list
myFit <- function(m, nvec, kvec, delta, slope, tau2, rhovec,
                  kprop, rhoprop, nprop, kgrid, replicate) {
  tau.sq <- I.sq<-matrix(0,4,replicate)
  eff <- mod <- array(0,dim=c(4,replicate,6))
  mean.sens<- sd.sens <- matrix(0,replicate,5)
   for (i in 1:replicate){
    mydata <- genData(m, nvec, kvec, delta, slope, tau2, rhovec,
                      kprop, rhoprop, nprop, kgrid)

    # RVE
    robufit <-  robu(formula = effectsize ~ mod, data = mydata, studynum
= studyid,
                     var.eff.size = var.effsize, modelweights = "CORR",
rho=0.8, small = TRUE)
    tau.sq[1,i]<-robufit$mod_info$tau.sq
    I.sq[1,i]<-robufit$mod_info$I.2
    #the following provides Estimate StdErr t-value  P(|t|>) 95% CI.L
95% CI.U for int and slope, resp.
    eff[1,i,]<-as.numeric(robufit$reg_table[1,c(2:4,6:8)])
    mod[1,i,]<-as.numeric(robufit$reg_table[2,c(2:4,6:8)])
    sens <- sensit(robufit)
    mat<-matrix(as.numeric(as.matrix(sens[,3:8])),5,6)
    # the following reports means and sds of eff.est, eff.est.se,
mod.est, mod.est.se, tau.sq, resp
    # for different values of rho (seq(0,1,by=.2))
    mean.sens[i,] <- rowMeans(mat)
    sd.sens[i,] <- apply(mat,1,sd)


    # classical naive
    m1 <- metacont(mean.e=mean.e, mean.c=mean.c, sd.e=sd.e, sd.c=sd.c,
                   n.e=samp.size, n.c=samp.size, studlab=studyid,
data=mydata,
                   pooledvar=TRUE, sm="SMD", method.smd="Cohen",
comb.random=TRUE)
    naiv <- metareg(m1,~ mod)
    tau.sq[2,i]<-naiv$tau2
    I.sq[2,i]<-naiv$I2
    est <-
cbind(naiv$beta,naiv$se,naiv$zval,naiv$pval,naiv$ci.lb,naiv$ci.ub)
    eff[2,i,]<-est[1,]
    mod[2,i,]<-est[2,]

    # classical average
    mydata.avg<- mydata %>% group_by(studyid) %>%
      dplyr::summarise(effectsize=mean(effectsize),
                       samp.size=mean(samp.size),
                       mean.e=mean(mean.e),
                       mean.c=mean(mean.c),
                       sd.e=sqrt(mean(sd.e^2)),
                       sd.c=sqrt(mean(sd.c^2)))
    avg <- metacont(mean.e=mean.e, mean.c=mean.c, sd.e=sd.e, sd.c=sd.c,
                   n.e=samp.size, n.c=samp.size, studlab=studyid,
data=mydata.avg,
                   pooledvar=TRUE, sm="SMD", method.smd="Cohen",
comb.random=TRUE)
    tau.sq[3,i]<-avg$tau2
    I.sq[3,i]<-avg$I2
    eff[3,i,]<-c(avg$TE.random,avg$seTE.random,avg$zval.random,
avg$pval.random, avg$lower.random, avg$upper.random)
    # mod[3,i,] remains at 0, since we do not have a moderator test
here.
    # ignore all moderator results for "average" method when compiling
results in figures!


    # classical random
    mydata.rand<-  mydata %>% group_by(studyid) %>% sample_n(1)
    m2 <- metacont(mean.e=mean.e, mean.c=mean.c, sd.e=sd.e, sd.c=sd.c,
                     n.e=samp.size, n.c=samp.size, studlab=studyid,
data=mydata.rand,
                     pooledvar=TRUE, sm="SMD", method.smd="Cohen",
comb.random=TRUE)
    rand <- metareg(m2,~ mod)
    tau.sq[4,i]<-rand$tau2
    I.sq[4,i]<-rand$I2
    est <-
cbind(rand$beta,rand$se,rand$zval,rand$pval,rand$ci.lb,rand$ci.ub)
    eff[4,i,]<-est[1,]
    mod[4,i,]<-est[2,]

  }
  list(tau.sq=tau.sq, I.sq=I.sq,
       eff.est=eff, mod.est=mod,
       rve.mn.sens =mean.sens, rve.sd.sens=sd.sens)
}

# The doit function will calculate performance estimates based off the
output of myFit and display in
# results a dataframe and list
doit <- function(m, nvec, kvec, delta, slope, tau2, rhovec,
                 kprop, rhoprop, nprop, kgrid, replicate, alpha) {
  fit <- myFit(m, nvec, kvec, delta, slope, tau2, rhovec,
               kprop, rhoprop, nprop, kgrid, replicate)


  # results combiled below are reported in this order: rve, naive,
average, random
  # (ignore all mod results for average!)

  # CI coverage of true parameter (intercept, slope)
  ci.cov.eff <- rowMeans(Apply(data=fit$eff, margins=c(1,2),
fun=get.cover, p=delta)$output1)
  ci.cov.mod <- rowMeans(Apply(data=fit$mod, margins=c(1,2),
fun=get.cover, p=slope)$output1)

  # Bias for estimation of tau, intercept, slope
  bias.tau.sq <- rowMeans(fit$tau.sq)-tau2  # rve, naive, average,
random
  bias.eff    <- rowMeans(fit$eff[,,1]) - delta
  bias.mod    <- rowMeans(fit$mod[,,1]) - slope

  # Power of non-zero intercept and slope tests
  pwr.eff         <- rowMeans(fit$eff[,,4]<alpha)
  pwr.mod         <- rowMeans(fit$mod[,,4]<alpha)

  # SD (method 1 = "empirical"= sd of replicate estimates of tau,
intercept and slope)
```

```
  sd.tau.sq <- apply(fit$tau.sq,1,sd) # rve, naive, average, random   #
may want
  sd.eff    <- apply(fit$eff[,,1],1,sd)
  sd.mod    <- apply(fit$mod[,,1],1,sd)

  # SD (method 2 = average of calculated se )
  asd.eff <- sqrt(rowMeans(fit$eff[,,2]^2))
  asd.mod <- sqrt(rowMeans(fit$mod[,,2]^2))

  # sensitivity flag for rve (value of 0 = good; 1=bad)
  rve.sens.flag <- mean(apply(fit$rve.sd.sens,1, function(x,p)
ifelse(any(x>p),1,0), p=.05 ))

  # I2
  mn.I2 <- rowMeans(fit$I.sq)
  sd.I2 <- apply(fit$I.sq,1,sd)

  output <- data.frame( ci.cov.eff = ci.cov.eff, ci.cov.mod =
ci.cov.mod,
                        bias.eff = bias.eff, bias.mod = bias.mod,
                        sd.eff = sd.eff, sd.mod = sd.mod,
                        asd.eff = asd.eff, asd.mod=asd.mod,
                        pwr.eff = pwr.eff, pwr.mod=pwr.mod)

  list( output=output, tau.sq = fit$tau.sq, bias.tau.sq = bias.tau.sq,
sd.tau.sq = sd.tau.sq,
        mn.I2=mn.I2, sd.I2=sd.I2, I.sq = fit$I.sq,
rve.sens.flag=rve.sens.flag)
}

# =======================================================
# set parameters for this simulation
set.seed(0)
m_1<-10;  m_2<-50; m_3 <- 100
delta<-.5;
tau2_1<-0; tau2_2 <- 0.15; tau2_3 <- 0.30;
slope=0; slope2=.05; slope3=0.1; slope4=0.15;
nvec<-c(10,20,200); kvec<-c(2, 4, 8);
rhovec<-c(.2,.8);
kprop<-nprop<-c(.3,.6,.1); kgrid<-seq(.25,3,by=.25); rhoprop=c(.5,.5);
alpha=0.05
replicate <- 1000
# =======================================================

# =======================================================
# create performance estimates using generated replicates for all
combinations
# of m, tau^2 and slope levels

##m=10, tau2=0.0
## slope=0
```

```
outm1tau1slope1<-doit(m_1, nvec, kvec, delta, slope, tau2_1,
           rhovec, kprop, rhoprop, nprop, kgrid,replicate, alpha)
#slope=0.05
outm1tau1slope2<-doit(m_1, nvec, kvec, delta, slope2, tau2_1,
           rhovec, kprop, rhoprop, nprop, kgrid,replicate, alpha)
#slope=0.10
outm1tau1slope3<-doit(m_1, nvec, kvec, delta, slope3, tau2_1,
           rhovec, kprop, rhoprop, nprop, kgrid,replicate, alpha)
#slope=0.15
outm1tau1slope4<-doit(m_1, nvec, kvec, delta, slope4, tau2_1,
           rhovec, kprop, rhoprop, nprop, kgrid,replicate, alpha)

# build new dataframe for m=10, tausq=0.0 containing results for 4
methods, 4 slope levels
method<-rep(c("RVE","Naive","Average","Random"), times=4)
slopes<-rep(c(0,0.05, 0.10,0.15), each=4)
newdf1 <- rbind(outm1tau1slope1$output, outm1tau1slope2$output,
outm1tau1slope3$output, outm1tau1slope4$output)
dfm1tau1<-data.frame(method, slopes, newdf1)
# =====================
# m=10, tau2=0.15
# slope=0
outm1tau2slope1<-doit(m_1, nvec, kvec, delta, slope, tau2_2,
                      rhovec, kprop, rhoprop, nprop, kgrid,replicate,
alpha)
#slope=0.05
outm1tau2slope2<-doit(m_1, nvec, kvec, delta, slope2, tau2_2,
                      rhovec, kprop, rhoprop, nprop, kgrid,replicate,
alpha)
#slope=0.10
outm1tau2slope3<-doit(m_1, nvec, kvec, delta, slope3, tau2_2,
                      rhovec, kprop, rhoprop, nprop, kgrid,replicate,
alpha)
#slope=0.15
outm1tau2slope4<-doit(m_1, nvec, kvec, delta, slope4, tau2_2,
                      rhovec, kprop, rhoprop, nprop, kgrid,replicate,
alpha)

# build new dataframe for m=10, tausq=0.15 containing results for 4
methods, 4 slope levels
newdf2 <- rbind(outm1tau2slope1$output, outm1tau2slope2$output,
outm1tau2slope3$output, outm1tau2slope4$output)
dfm1tau2<-data.frame(method, slopes, newdf2)

# ====================
# m=10, tau2=0.30
# slope=0
outm1tau3slope1<-doit(m_1, nvec, kvec, delta, slope, tau2_3,
                      rhovec, kprop, rhoprop, nprop, kgrid,replicate,
alpha)
#slope=0.05
```

```
outm1tau3slope2<-doit(m_1, nvec, kvec, delta, slope2, tau2_3,
                      rhovec, kprop, rhoprop, nprop, kgrid,replicate,
alpha)
#slope=0.10
outm1tau3slope3<-doit(m_1, nvec, kvec, delta, slope3, tau2_3,
                      rhovec, kprop, rhoprop, nprop, kgrid,replicate,
alpha)
#slope=0.15
outm1tau3slope4<-doit(m_1, nvec, kvec, delta, slope4, tau2_3,
                      rhovec, kprop, rhoprop, nprop, kgrid,replicate,
alpha)

# build new dataframe for m=10, tausq=0.30 containing results for 4
methods, 4 slope levels
newdf3 <- rbind(outm1tau3slope1$output, outm1tau3slope2$output,
outm1tau3slope3$output, outm1tau3slope4$output)
dfm1tau3<-data.frame(method, slopes, newdf3)

#======================
# m=50, tau2=0.0
# slope=0
outm2tau1slope1<-doit(m_2, nvec, kvec, delta, slope, tau2_1,
                      rhovec, kprop, rhoprop, nprop, kgrid,replicate,
alpha)
#slope=0.05
outm2tau1slope2<-doit(m_2, nvec, kvec, delta, slope2, tau2_1,
                      rhovec, kprop, rhoprop, nprop, kgrid,replicate,
alpha)
#slope=0.10
outm2tau1slope3<-doit(m_2, nvec, kvec, delta, slope3, tau2_1,
                      rhovec, kprop, rhoprop, nprop, kgrid,replicate,
alpha)
#slope=0.15
outm2tau1slope4<-doit(m_2, nvec, kvec, delta, slope4, tau2_1,
                      rhovec, kprop, rhoprop, nprop, kgrid,replicate,
alpha)

# build new dataframe for m=50, tausq=0.0 containing results for 4
methods, 4 slope levels
newdf4 <- rbind(outm2tau1slope1$output, outm2tau1slope2$output,
outm2tau1slope3$output, outm2tau1slope4$output)
dfm2tau1<-data.frame(method, slopes, newdf4)

# =====================
# m=50, tau2=0.15
# slope=0
outm2tau2slope1<-doit(m_2, nvec, kvec, delta, slope, tau2_2,
                      rhovec, kprop, rhoprop, nprop, kgrid,replicate,
alpha)
#slope=0.05
outm2tau2slope2<-doit(m_2, nvec, kvec, delta, slope2, tau2_2,
                      rhovec, kprop, rhoprop, nprop, kgrid,replicate,
alpha)
#slope=0.10
outm2tau2slope3<-doit(m_2, nvec, kvec, delta, slope3, tau2_2,
                      rhovec, kprop, rhoprop, nprop, kgrid,replicate,
alpha)
#slope=0.15
outm2tau2slope4<-doit(m_2, nvec, kvec, delta, slope4, tau2_2,
                      rhovec, kprop, rhoprop, nprop, kgrid,replicate,
alpha)

# build new dataframe for m=50, tausq=0.15 containing results for 4
methods, 4 slope levels
newdf5 <- rbind(outm2tau2slope1$output, outm2tau2slope2$output,
outm2tau2slope3$output, outm2tau2slope4$output)
dfm2tau2<-data.frame(method, slopes, newdf5)

# ====================
# m=50, tau2=0.30
# slope=0
outm2tau3slope1<-doit(m_2, nvec, kvec, delta, slope, tau2_3,
                      rhovec, kprop, rhoprop, nprop, kgrid,replicate,
alpha)
#slope=0.05
outm2tau3slope2<-doit(m_2, nvec, kvec, delta, slope2, tau2_3,
                      rhovec, kprop, rhoprop, nprop, kgrid,replicate,
alpha)
#slope=0.10
outm2tau3slope3<-doit(m_2, nvec, kvec, delta, slope3, tau2_3,
                      rhovec, kprop, rhoprop, nprop, kgrid,replicate,
alpha)
#slope=0.15
outm2tau3slope4<-doit(m_2, nvec, kvec, delta, slope4, tau2_3,
                      rhovec, kprop, rhoprop, nprop, kgrid,replicate,
alpha)

# build new dataframe for m=50, tausq=0.30 containing results for 4
methods, 4 slope levels
newdf6 <- rbind(outm2tau3slope1$output, outm2tau3slope2$output,
outm2tau3slope3$output, outm2tau3slope4$output)
dfm2tau3<-data.frame(method, slopes, newdf6)
# ====================

# ====================
# m=100, tau2=0.0
# slope=0
outm3tau1slope1<-doit(m_3, nvec, kvec, delta, slope, tau2_1,
                      rhovec, kprop, rhoprop, nprop, kgrid,replicate,
alpha)
#slope=0.05
outm3tau1slope2<-doit(m_3, nvec, kvec, delta, slope2, tau2_1,
                      rhovec, kprop, rhoprop, nprop, kgrid,replicate,
alpha)
#slope=0.10
outm3tau1slope3<-doit(m_3, nvec, kvec, delta, slope3, tau2_1,
                      rhovec, kprop, rhoprop, nprop, kgrid,replicate,
alpha)
#slope=0.15
outm3tau1slope4<-doit(m_3, nvec, kvec, delta, slope4, tau2_1,
                      rhovec, kprop, rhoprop, nprop, kgrid,replicate,
alpha)

# build new dataframe for m=100, tausq=0.0 containing results for 4
methods, 4 slope levels
newdf7 <- rbind(outm3tau1slope1$output, outm3tau1slope2$output,
outm3tau1slope3$output, outm3tau1slope4$output)
dfm3tau1<-data.frame(method, slopes, newdf7)
# ====================
# m=100, tau2=0.15
# slope=0
outm3tau2slope1<-doit(m_3, nvec, kvec, delta, slope, tau2_2,
                      rhovec, kprop, rhoprop, nprop, kgrid,replicate,
alpha)
#slope=0.05
outm3tau2slope2<-doit(m_3, nvec, kvec, delta, slope2, tau2_2,
                      rhovec, kprop, rhoprop, nprop, kgrid,replicate,
alpha)
#slope=0.10
outm3tau2slope3<-doit(m_3, nvec, kvec, delta, slope3, tau2_2,
                      rhovec, kprop, rhoprop, nprop, kgrid,replicate,
alpha)
#slope=0.15
outm3tau2slope4<-doit(m_3, nvec, kvec, delta, slope4, tau2_2,
                      rhovec, kprop, rhoprop, nprop, kgrid,replicate,
alpha)

# build new dataframe for m=100, tausq=0.15 containing results for 4
methods, 4 slope levels
newdf8 <- rbind(outm3tau2slope1$output, outm3tau2slope2$output,
outm3tau2slope3$output, outm3tau2slope4$output)
dfm3tau2<-data.frame(method, slopes, newdf8)

# =====================
# m=100, tau2=0.30
# slope=0
outm3tau3slope1<-doit(m_3, nvec, kvec, delta, slope, tau2_3,
                      rhovec, kprop, rhoprop, nprop, kgrid,replicate,
alpha)
#slope=0.05
outm3tau3slope2<-doit(m_3, nvec, kvec, delta, slope2, tau2_3,
                      rhovec, kprop, rhoprop, nprop, kgrid,replicate,
alpha)
```

```
#slope=0.10
outm3tau3slope3<-doit(m_3, nvec, kvec, delta, slope3, tau2_3,
                      rhovec, kprop, rhoprop, nprop, kgrid,replicate,
alpha)
#slope=0.15
outm3tau3slope4<-doit(m_3, nvec, kvec, delta, slope4, tau2_3,
                      rhovec, kprop, rhoprop, nprop, kgrid,replicate,
alpha)

# build new dataframe for m=100, tausq=0.30 containing results for 4
methods, 4 slope levels
newdf9 <- rbind(outm3tau3slope1$output, outm3tau3slope2$output,
outm3tau3slope3$output, outm3tau3slope4$output)
dfm3tau3<-data.frame(method, slopes, newdf9)


#=====================================


# Application to Baffoni meta-analysis- RVE and classical average random
effects meta-analysis comparison
# CLASSICAL AVERAGE RANDOM EFFECTS

# All Interventions- independence assumption violated
    # SBP
    avg_sbp_dbp_try3 <-
read.csv("~/Desktop/HonorsThesis/classicalavg_SBP_DBP.csv")
    avg_sbp3 <- metacont(mean.e=mean.e.sbp, mean.c=mean.c.sbp,
sd.e=mean.e.sbp.SD, sd.c=mean.c.sbp.SD,
                     n.e=samp.size, n.c=samp.size,
data=avg_sbp_dbp_try3,
                     pooledvar=TRUE, sm="SMD", method.smd="Cohen",
comb.random=TRUE)
    summary(avg_sbp3)
    avg_sbp3$tau2
    avg_sbp3$I2
    avg_sbp3$TE.random
    avg_sbp3$seTE.random
    avg_sbp3$zval.random
    avg_sbp3$pval.random
    avg_sbp3$lower.random
    avg_sbp3$upper.random
    # DBP
    avg_dbp3 <- metacont(mean.e=mean.e.dbp, mean.c=mean.c.dbp,
sd.e=mean.e.dbp.SD, sd.c=mean.c.dbp.SD,
                     n.e=samp.size, n.c=samp.size,
data=avg_sbp_dbp_try3,
                     pooledvar=TRUE, sm="SMD", method.smd="Cohen",
comb.random=TRUE)
    summary(avg_dbp3)
    avg_dbp3$tau2
    avg_dbp3$I2
    avg_dbp3$TE.random
    avg_dbp3$seTE.random
    avg_dbp3$zval.random
    avg_dbp3$pval.random
    avg_dbp3$lower.random
    avg_dbp3$upper.random

## Independent groups- randomly choose 1 intervention arm per study,
repeat classical analysis
    library(dplyr)
    avg_sbp_dbp_try3_indep <- avg_sbp_dbp_try3 %>% group_by(Studyid) %>%
sample_n(1)
    # SBP
    avg_sbp3_indep <- metacont(mean.e=mean.e.sbp, mean.c=mean.c.sbp,
sd.e=mean.e.sbp.SD, sd.c=mean.c.sbp.SD,
                           n.e=samp.size, n.c=samp.size,
data=avg_sbp_dbp_try3_indep,
                           pooledvar=TRUE, sm="SMD", method.smd="Cohen",
comb.random=TRUE)
    summary(avg_sbp3_indep)
    avg_sbp3_indep$tau2
    avg_sbp3_indep$I2
    avg_sbp3_indep$TE.random
    avg_sbp3_indep$seTE.random
    avg_sbp3_indep$zval.random
    avg_sbp3_indep$pval.random
    avg_sbp3_indep$lower.random
    avg_sbp3_indep$upper.random
    # DBP
    avg_dbp3_indep <- metacont(mean.e=mean.e.dbp, mean.c=mean.c.dbp,
sd.e=mean.e.dbp.SD, sd.c=mean.c.dbp.SD,
                           n.e=samp.size, n.c=samp.size,
data=avg_sbp_dbp_try3_indep,
                           pooledvar=TRUE, sm="SMD", method.smd="Cohen",
comb.random=TRUE)
    summary(avg_dbp3_indep)
    avg_dbp3_indep$tau2
    avg_dbp3_indep$I2
    avg_dbp3_indep$TE.random
    avg_dbp3_indep$seTE.random
    avg_dbp3_indep$zval.random
    avg_dbp3_indep$pval.random
    avg_dbp3_indep$lower.random
    avg_dbp3_indep$upper.random


# RVE
    RVE_Baffoni_data <-
read.csv("~/Desktop/HonorsThesis/RVE_Baffoni_data.csv")
    # SBP
    rve_sbp <- robu(formula = diffsys ~ time -1 , var.eff.size =
diffsysv, data = RVE_Baffoni_data, studynum = Studyid,
                modelweights = "CORR")
    summary(rve_sbp)
    rve_sbp$mod_info$tau.sq
    rve_sbp$mod_info$I.2
    rve_sbp$reg_table
    sensit(rve_sbp)
    # DBP
    rve_dbp <- robu(formula = diffdia ~ time, var.eff.size = diffdiav,
data = RVE_Baffoni_data, studynum = Studyid,
                modelweights = "CORR")
    rve_dbp$mod_info$tau.sq
    rve_dbp$mod_info$I.2
    rve_dbp$reg_table
    sensit(rve_dbp)
### END ###
```