AHMAD SOFI-MAHMUDI

IDENTIFYING AN OPTIMAL STRATEGY FOR CONVERTING PAIN AS A CONTINUOUS OUTCOME TO A RESPONDER ANALYSIS

By AHMAD SOFI-MAHMUDI, D.D.S.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfilment of the Requirements for the Degree Master of Science

McMaster University MASTER OF SCIENCE (2024) Hamilton, Ontario (Health Research Methodology)

TITLE: Identifying an optimal strategy for converting pain as a continuous outcome to a responder analysis

AUTHOR: Ahmad Sofi-Mahmudi, D.D.S. (McMaster University)

SUPERVISOR: Dr. Behnam Sadeghirad

NUMBER OF PAGES: xv, 47

Lay Abstract

Pain studies often use scales that are difficult to interpret. To make results more meaningful, researchers sometimes estimate the percentage of patients achieving important pain relief. This study tested four methods for estimating these percentages when combining results from multiple pain studies.

Using computer simulations, we created thousands of hypothetical pain studies and meta-analyses. We then applied the four estimation methods and compared their results to the "true" simulated values, assessing accuracy, precision, and reliability across various scenarios.

Overall, differences between methods were small. A method estimating results for each study individually before combining them performed slightly better in some ways. All methods were less accurate with non-normally distributed data and worked best for small treatment differences.

While these methods can provide useful estimates in many cases, they should be used cautiously, especially for large treatment effects or non-normal data. More pain studies would directly report meaningful patient outcomes.

Abstract

Background: In pain relief research, meta-analyses often combine continuous outcomes from various studies using mean differences. However, this approach can be difficult to interpret clinically. An alternative method involves aggregating the risk difference for patients who achieve a minimally important difference (MID) in pain reduction. The challenge is that many trials do not report responder analyses, necessitating continuous data conversion.

Objective: To conduct a simulation study assessing the performance of four proposed methods for estimating the pooled risk difference (RD) of achieving the MID in meta-analyses of pain measured on a 10cm visual analogue scale (VAS).

Methods: Individual patient data for VAS pain scores were simulated across 4,752 scenarios varying the treatment effect as change score in the intervention (-1.0 to 4.0) and control (-1.0 to 3.0) groups, study sample size (10-1000), number of studies per meta-analysis (3 to 30), shape of distribution (normal or skewed), and MID (1.0 or 1.5). The true pooled RD and 95% confidence interval (CI) were calculated from the simulated individual data. Four methods were evaluated: calculating RD based on pooled 1) median mean differences, 2) unweighted average differences, 3) weighted average differences, and 4) calculating RD for each individual study and then meta-analysing RDs. Bias, mean squared error, confidence interval (CI) coverage of true value, and empirical standard error (SE), and model-based SE were evaluated.

Results: The median method showed the lowest bias (2.048; 95% CI: 1.759-2.338), while the individual method demonstrated the lowest RMSE (4.852; 95% CI: 4.661-5.044), empirical SE (0.148; 95% CI: 0.141-0.154), and model-based SE (2.198; 95% CI: 2.108-2.288), and highest CI coverage (55.717%; 95% CI: 53.185-58.250%).

Differences between methods were minimal and not statistically significant. Performance was optimal when treatment effects were similar between groups and declined with increasing effect size differences. All methods performed poorly with skewed distributions.

Conclusion: While the evaluated methods can provide useful estimates in many scenarios, they should be used cautiously, especially for large treatment effects or non-normal data. Researchers should prioritize conducting and reporting responder analyses in primary studies to reduce reliance on these estimation methods in meta-analyses.

Acknowledgements

I am truly thankful for all the support that I got from my lovely family, Qadir, Mapere, Birayîm, Meryem, Minîre, and Simayîl, during all these years.

I would also like to express my heartfelt gratitude to my friends, Sahar Khademioore and Ali Torabi, for all their unwavering support and kindness during my studies at McMaster University.

Table of Contents

Lists of Figures and Tables

List of all Abbreviations and Symbols

- **ANOVA:** Analysis of Variance
- **CI:** Confidence Interval
- **IMMPACT:** Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials
- **IPD:** Individual Patient Data
- **MD:** Mean Difference
- **MID:** Minimally Important Difference
- **OMERACT:** Outcome Measures in Rheumatology
- **PROM:** Patient-Reported Outcome Measure
- **RD:** Risk Difference
- **RMSE:** Root Mean Squared Error
- **RR:** Risk Ratio
- **SD:** Standard Deviation
- **SE:** Standard Error
- **SMD:** Standardized Mean Difference
- **VAS:** Visual Analogue Scale
- **WMD:** Weighted average Difference

Declaration of Academic Achievement

This thesis presents the work conducted by Ahmad Sofi-Mahmudi, hereafter referred to as the "primary researcher." The project was carried out under the supervision of Dr. Behnam Sadeghirad, with Dr. Jason W. Busse and Dr. Aaron Jones serving as committee members. Dr. Kristian Thorlund acted as the external committee member.

This research constituted a simulation study, encompassing the following key components:

1. **Data generation:** The primary researcher developed the code necessary for generating the datasets used in the study.
2. **Data analysis:** The primary researcher wrote and implemented the code for analyzing the generated datasets.
3. **Interpretation:** The primary researcher was responsible for interpreting the results of the analyses.
4. **Manuscript preparation:** The primary researcher authored the manuscript detailing the study's methodology, findings, and conclusions.

Roles and Contributions

Primary Researcher (Ahmad Sofi-Mahmudi)

- Developed code for data generation and analysis
- Conducted data interpretation
- Authored the manuscript

Supervisor (Dr. Behnam Sadeghirad)

- Oversaw the development of the study concept and design
- Supervised the analysis process

- Verified all outputs for accuracy and completeness

Committee Members (Dr. Jason W. Busse and Dr. Aaron Jones)

- Provided guidance on study design
- Assisted with data analysis strategies
- Contributed to the interpretation of results

This declaration affirms that the primary researcher, Ahmad Sofi-Mahmudi, made substantial and original contributions to all aspects of the study, from conception to completion, under the guidance and support of the supervisor and committee members.

Introduction

Systematic reviews are undertaken to summarize the effect of interventions and often pool data from randomized controlled trials. The most common approach, which allows for pooling of all instruments that measure a common domain, is the standardized mean difference (SMD) (J. P. Higgins et al., 2019). SMD, endorsed by Cochrane, is calculated by dividing weighted average differences (WMDs) between intervention and control in each study by the study's standard deviation (SD) (J. P. Higgins et al., 2019). Therefore, the magnitude of the pooled estimate is presented in SD units. However, this is the least understood effect estimate (Johnston et al., 2016) and is vulnerable to baseline heterogeneity of trial participants (Johnston et al., 2010).

A better approach is to convert all instruments that measure a common domain to a common tool and then pool them in natural units as the WMD. However, this effect estimate is also poorly understood because it needs contextual information on the topic under investigation (e.g., the relation between different tools of measuring an outcome).

The challenges with WMD and SMD are particularly apparent when considering patient-reported outcome measures (PROMs). PROMs offer valuable insights into a patient's health condition by directly capturing information from the patient without any interpretation from a clinician or other individuals (Carrasco-Labra et al., 2021). However, for example, patients and practitioners may have difficulty understanding the implication of a 1-point reduction in pain from baseline on a 0–10 visual analogue scale (VAS) as a result of an intervention (Busse et al., 2015). To address these issues, Johnston et al. proposed two approaches for enhancing the interpretability of pooled effect estimates (Johnston et al., 2013).

The first approach uses minimally important difference (MID) units, which consider the smallest difference in the outcome that patients or clinicians would find meaningful. By dividing the mean difference of each study by the MID associated with the outcome, one can calculate the difference between two groups in MID units. This approach provides a more meaningful interpretation of pooled effect estimate (Jaeschke et al., 1989; Schünemann & Guyatt, 2005) but relies on obtaining accurate MID values for an outcome, which may not always be readily available. However, MID has been defined for some outcomes such as pain which is 1.5 (Wang et al., 2023).

Co-presenting the WMD with the associated MID may improve interpretability, but risks having readers conclude that average effects below the MID mean the intervention is ineffective, whereas in reality there is a distribution of effects around the mean, with some patients doing better than the average and some doing worse. The second approach aims to cover this limitation.

The second approach involves applying a threshold, usually the MID, to determine the proportion of patients experiencing a benefit or harm beyond that threshold (Dworkin et al., 2008; Froud et al., 2011). Reporting proportion of patients who respond to an intervention provides valuable insights into treatment effectiveness. While this method has limitations, such as loss of power due to dichotomization and uncertainty in the definition of a MID (Cappelleri & Chambers, 2021; Collister et al., 2021; Snapinn & Jiang, 2007), it is rarely employed in randomized controlled trials (RCTs), with fewer than one-fifth of low back pain trials incorporating responder analysis (Henschke et al., 2014). Compounding the issue, systematic reviewers often lack access to individual patient data (IPD) (Esmail et al., 2023; Gabelica et al., 2022;

Gabelica et al., 2019; Polanin, 2018; Scutt et al., 2020), making it difficult to calculate the true proportions of responders.

As such, there is a need to convert continuous data to estimate the results of a responder analysis. Thorlund et al. proposed a statistical approach to estimate the proportion of responders that uses normal distribution assumptions to convert individual trial continuous data to probabilities (or risks) using WMD and the corresponding SD and an established MID for that instrument (Thorlund et al., 2011). As Thorlund et al. have indicated in Formula 8 of their paper, to find the probability of achieving a threshold or larger, we can use the cumulative distribution function of the normal distribution (usually symbolized as Φ), as follows: $P(\text{Responders}) = 1 - \Phi(\frac{\text{MID} - \text{mean}}{\text{SD}})$ (Formula 5 below). After obtaining the probabilities for each trial arm, risk ratio (RR) and risk difference (RD) can be calculated for each study and pooled using conventional meta-analysis technique. These outputs are proven to be more intuitive for patients, clinicians, and policymakers to communicate desirable and undesirable outcomes associated with an intervention (Akl et al., 2011).

Busse et al. (Busse et al., 2018) suggested using meta-analytic pooled MD to impute the percentage of patients achieving MID reduction based on methods introduced by Thorlund et al. (Thorlund et al., 2011). There are three options for choosing pooled MD and corresponding SD to be used in Formula 1: (1) median, (2) unweighted average, or (3) weighted average of the MDs and SDs of the included study (using inverse variance as weight). Utilizing each of these methods may yield different estimates for RR and RD of the responder analysis. Despite frequent use of responder analysis and its apparent benefits in interpretations of meta-analytic findings, there has

been no attempt to evaluate the accuracy, validity, or reliability of suggested methods (Busse et al., 2015) using a simulation study.

To address the uncertainties and obtain an optimal approach, we performed a simulation study on the visual analogue scale (VAS) measures of chronic pain to assess bias, root mean squared errors (RMSEs), coverage of the true value in the estimated confidence interval (CI), and empirical standard error (SE) of estimated pooled RD using three methods, namely, median, unweighted average, and weighted average. A fourth method, estimating RRs for each individual study and then meta-analyzing estimated RRs, was also investigated.

Methods

All the codes and data are available in its pertaining OSF and GitHub repositories.

This simulation study followed the aims, data-generating mechanisms, estimands, methods, and performance measures (ADEMP) framework suggested by Morris et al. (Morris et al., 2019).

Aims

To estimate the pooled RD and the associated 95% CI for achieving ≥ to the MID on the 10cm VAS for pain – termed 'responders' - in meta-analyses of pain relief, based on MDs and SDs of the included studies, using four different methods: 1) median, 2) unweighted average, 3) weighted average, and 4) estimating responders individually and then pooling them. These estimates were then compared to the true pooled RD for achieving ≥ to the MID obtained from IPD.

Data-generating mechanisms

We generated IPD of pain relief measured using a 10cm VAS for meta-analyses of continuous data. A VAS is a straight line, typically presented horizontally, with the two ends labelled to represent the extreme limits of the sensation, feeling, or response being assessed (Scott & Huskisson, 1976). For example, a VAS to measure pain could be labelled "no pain" on one end (often left) and "pain as bad as it could possibly be" on the other (often right) (Figure 1).

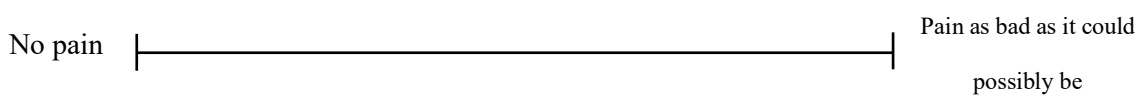


Figure 1. A simple visual analogue scale (horizontal version).

The patient reports their perceived level of pain by placing a mark on the line between these two boundaries. Then, the distance between the "no pain" boundary (0) and the marked point is measured. For this study, we used a 10cm VAS with one decimal (equal to in mm). Thus, our generated data were approximately continuous.

Meta-analyses that utilize VAS scores as their outcome can use the following three options: 1) use outcomes at the end of study, 2) use change from baseline, or 3) use outcomes at the end of study, adjusting for baseline. Whereas the third option is more powerful and makes fewer assumptions, it needs IPD data for analysis which may not resemble the real-world situation in the majority of meta-analyses. Therefore, we used the second option. To do so, we generated random change scores based on a predefined mean and SD of change score for each scenario.

Our intended data-generating mechanisms differed based on six factors:

A. Treatment effect (change score) in the intervention group;

B. Treatment effect (change score) in the control group (placebo);

C. The number of patients in the studies;

D. The number of studies included in the meta-analysis;

E. Shape of distribution; and,

F. MID.

A) Treatment effect in the intervention group: We considered 11 different scenarios for the mean VAS change score of the treatment group: -1.0 (worsening of pain), -0.5, 0.0 (no effect), 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0 (strong treatment effect).

B) Treatment effect in the control group: The mean change score for the control groups were between -1 to 3 (to compensate for the placebo effect): -1.0 (worsening of the pain), -0.5, 0.0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0 (pain reduction due to placebo effect).

C) The number of patients in the studies: We considered three types of studies: 1) small: n=10–200, 2) medium: n=201–400, and 3) large: n=401–1000.

D) The number of studies included in the meta-analysis: We generated meta-analyses of 3, 5, 10, 20, 30, and 50 studies. The meta-analyses included either all very small/small studies or a composition of studies with different numbers of patients, as follows: 10% large studies, 30% medium-sized studies, and 60% small studies. For example, meta-analyses with 10 studies of mixed size on average included one large study, three medium-size studies, and six small studies.

E) Shape of distribution: The main analyses were based on normally distributed data. To perform a sensitivity analysis, we also generated skewed data.

F) MID: Scenarios with MID=1.5 were the main interest of this simulation study. However, since many previously published meta-analyses have used MID=1, we compared these two thresholds as a sensitivity analysis.

Since the VAS range is 0 to 10 cm, the VAS cannot follow a completely normal distribution, as a normal distribution has no boundaries (Heller et al., 2016). Nevertheless, in real-world scenarios, we might observe VAS scores that show a close

approximation to a normal distribution if the scores are not heavily concentrated at the extreme ends of the scale. Furthermore, we used change scores which exhibit more normally distributed behaviour than end-of-study scores. Thus, we used a normal distribution for generating our simulated data and round numbers to one decimal place for our main scenarios. The range of treatment effects were between −4cm and +4cm, with an associated standard deviation (SD) randomly generated between 1–3.

Overall, we had 11 (intervention effects) × 9 (control effects) × 6 (number of studies) × 2 (study sizes) × 2 (distribution) × 2 (MID) = 4,752 different scenarios for our simulation study.

Estimands

In clinical trials measuring pain with a continuous outcome (e.g., the VAS, numerical rating scale, behavioural pain scale), a responder is someone who has achieved a pre-defined level of improvement on an outcome at a certain time point. A change in the outcome that has been established as the smallest improvement important to most patients is defined as the MID. Once a valid MID is established for a tool, a responder will be someone whose measured outcome achieves or surpasses the MID threshold (Wang et al., 2023) (Formula 1).

$$\text{Responder}_i = \text{Patient}_i - \text{MID} \geq 0$$

or

$$\text{Responder}_i = \frac{\text{Patient}_i}{\text{MID}} \geq 1$$

(1)

Some trials report the proportion of patients in each group who were responders (Formula 2).

$$P(\text{ Responders }) = \frac{\sum(\text{Responder})}{\text{n}}$$

(2)

After obtaining the proportion of responders in each arm, we can calculate RD and/or RR of achieving ≥MID using the following formulae (P_e: proportion in the intervention group, P_c: proportion in the control group):

$$RD = P_e - P_c$$

(3)

$$RR = \frac{Pe}{Pc}$$

(4)

Then, we can pool the RDs/RRs using DerSimonian and Laird method, which produces a random-effects meta-analysis (Higgins & Green, 2022).

However, most trials capturing pain as an outcome do not report the proportion of responders. To derive the probability of achieving a threshold or larger, we can use the cumulative distribution function. Because of the ubiquity of the normal distribution and the continuous nature of proportions, the normal cumulative distribution function

(usually symbolized as Φ) has been traditionally used (Thorlund et al., 2011) to estimate responders' proportion using the following formula:

$$P(\text{ Responders }) = 1 - \Phi(\frac{\text{MID} - \text{MD}}{\text{SD}})$$

(5)

If the MID is a negative number (e.g., improvement in pain on a 10cm VAS), then we use the output of the Φ function and do not subtract 1.

Formula 5 can be used to calculate pooled estimates of proportions as well as for single studies. The pooled mean and pooled SD for this formula can be calculated in three ways: 1) median, 2) unweighted average, or 3) weighted average. For example, for the first option, we can calculate the median of all mean intervention effects that contribute to a meta-analysis and do the same for the associated SDs. Then, using formula 5, we can calculate the estimated pooled proportion of responders in that meta-analysis. The second option follows the same approach as the first, except that instead of the median, we calculate the average mean and associated SD of contributing treatment effects. For the third option, we conduct a meta-analysis for both means and SDs in both the treatment and control groups, typically using the inverse-variance approach (J. Higgins et al., 2019). In this method, studies with more participants (and therefore smaller standard errors) are given higher weights and contribute more to the pooled estimate of effect.

Our fourth method is to estimate the RR for each individual study and then meta-analyze RRs, as the suggested method by Thorlund et al. (Thorlund et al., 2011). After calculating each of the proportions for the intervention and control groups, RR will be

calculated as shown in Formula 4. We use the median of the control groups to inform P_c (baseline risk).

The 95% CI for the RR is calculated as follows:

$$95\%\ \text{CI: Exp}[\ \ln(\ \text{RR}\ ) \pm 1.96 \times \text{SE}(\ \ln(\ \text{RR}\ )\ )\ ] \tag{6}$$

Where:

$$SE\ (\ln(\ \text{RR}\ )\ ) = \sqrt{Var(\ \ln\ (RR\ ))}$$

Var(ln(RR)) is calculated as follows (Ne: number in the intervention group, Nc: number in the control group):

$$Var(\ln(RR)) = \frac{1}{Pe \times \text{Ne}} + \frac{1}{Pc \times \text{Nc}} - \frac{1}{Ne} - \frac{1}{Nc} \tag{7}$$

We calculated RD based on RR and its associated 95% CI since studies have shown that directly calculating the RD directly yields wide CIs (Newcombe & Bender, 2014). To do so, we will use the following formula to convert RR and its 95%CI to an RD and associated 95%CI:

$$\text{RD} = P_c \times (1-\text{RR}) \tag{8}$$

In which P_c is the baseline risk (or risk in the control group).

Hence, our estimands were:

1. The RD between the intervention and control groups' responders; and,

2. The 95% CI for RD.

Simulation methodology

Each simulated dataset were analyzed in five ways:

1. Calculating the true estimands based on individual data of the datasets;
2. Estimating the estimands using the median of means and SDs;
3. Estimating the estimands using the unweighted average of means and SDs;
4. Estimating the estimands using the weighted average of means and SDs;
5. Estimating the estimands using the individual method.

Performance measures

We assessed bias, RMSEs, CI coverage percentage, and empirical SE, and model-based SE for our estimates of the RD for achieving ≥MID. Bias is our key performance measure of interest and was calculated as follows:

$$\text{Bias} = \text{average estimated value } (\hat{\theta}) - \text{true value } (\theta) \tag{9}$$

RMSE was calculated as:

$$\sqrt{\frac{\sum(\hat{\theta} - \theta)^2}{n}} \tag{10}$$

Coverage percentage was the proportion of times where the CI for the estimated RD for achieving ≥MID includes the true RD times 100:

$$P\,(\theta \in 95\%\ \text{CI}\ [\hat{\theta}]) \times 100 \tag{11}$$

The empirical SE of the estimator $\hat{\theta}$ is calculated solely from the observed estimates without knowing the true value of θ. It provides an estimate of the variability of $\hat{\theta}$ across the number of simulation replications by computing the standard deviation of the estimates:

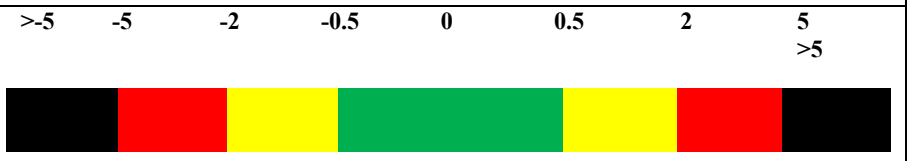
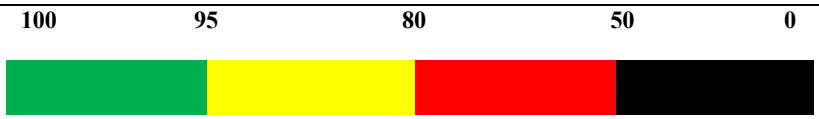
$$\sqrt{\frac{1}{nsim-1}\sum_{i=1}^{nsim}(\hat{\theta}-\bar{\theta})^2} \tag{12}$$

The model-based SE is calculated by taking the square root of the average of the estimated variances from each simulation replication:

$$\sqrt{\frac{1}{nsim}\sum_{i=1}^{nsim}\widehat{Var}\,(\hat{\theta}_i)} \tag{13}$$

To assess the superiority of each method in each performance measure, we defined the following thresholds (Table 1. Thresholds for performance measurement of each method.):

Table 1. Thresholds for performance measurement of each method.

Measure	**Thresholds**
Bias	
RMSE	
CI coverage	

We assumed that SD(θ) ≤ 0.2, meaning that Var(θ) ≤ 0.04, which was a conservative estimate based on an initial small simulation run. We required the Monte Carlo SE of bias to be lower than 0.005. Given that:

$$\text{Monte Carlo SE(Bias)} = \sqrt{\text{Var}(\theta) \div \text{nsim}}$$

This implies that each of our simulations required 1600 repetitions:

$$\text{nsim} = \text{Var}(\theta) \div \text{SE}^2$$

$$= 0.04 \div 0.005^2$$

$$= 0.04 \div 0.000025$$

$$= 1600$$

If coverage of all methods is 95%, the implication of using nsim = 1600 is:

$$\text{Monte Carlo SE(Coverage)} = \sqrt{95 \times 5 \div 1600} = 0.54.$$

With 50% coverage, the Monte Carlo SE is maximized at 1.25. We considered this satisfactory and so proceeded with nsim =1600 (to be revised if, for example, SD(θ) > 0.2).

The summarized workflow of the simulation study is available in Table 2.

Table 2. The workflow for the simulation part of the project.

Step	Calculating true values	Formulae	Calculating estimations	Formulae
1	Calculating the proportion of responders for each arm in each study	$\frac{\sum(\text{Responder})}{n}$	- Calculating the proportion of responders for the treatment arm based on the formula proposed by Thorlund et al. (Thorlund et al., 2011) using median, unweighted average, and weighted average of MDs and SDs - Calculating the same for the control arm but only using the median effect	$\Phi(\frac{\text{MID} - \text{MD}}{\text{SD}})$
2	Calculating RD for each study	$RD = P_e - P_c$	Calculating pooled RR and its 95% CI	$RR = P_e \div P_c$ 95% CI: $\text{Exp}[\ \ln(\ RR\ ) \pm 1.96 \times SE(\ \ln(\ RR\ )\ )\ ]$
3	Pooling RDs in a random-effects meta-analysis	Using DerSimonian and Laird (DL) method	Calculating pooled RD and its 95% CI	$RD = P_c \times (1-RR)$

	Performance measures	Formulae
4	Bias	Average $\hat{\theta} - \theta$
5	Root MSE	$\sqrt{\frac{\sum(\hat{\theta} - \theta)^2}{n}}$
6	CI overage	P ($\theta \in$ 95% CI [$\hat{\theta}$])
7	Empirical SE	$\sqrt{\frac{1}{nsim-1}\sum_{i=1}^{nsim}(\hat{\theta} - \bar{\theta})^2}$
8	Model-based SE	$\sqrt{\frac{1}{nsim}\sum_{i=1}^{nsim}\widehat{Var}(\hat{\theta}_i)}$

MD: mean difference, **MID**: minimally important difference, **SD**: standard deviation, **SE**: standard error, **RD**: risk difference, **Pe**: proportion in intervention group, **Pc**: proportion in control group, **RR**: risk ratio, **MSE**: mean squared error, **CI**: confidence interval.

Analysis

We explored if there were any missing estimates and reported the number of missing values and their location. We used scatterplots to visualize the overall performance measurements, and for each of the six factors data were created based on them. We used Chi-square and ANOVA tests with α=0.05 to test the statistical differences between the four methods. We used R to generate the simulation datasets and analyze the results.

Results

Overall

Missing values

Two missing data points (CIs of true risk difference) were in one of the scenarios with normal distribution and MID=1.5. Normal distribution with MID=1 had eight, skewed distribution with MID=1 had 38, and skewed distribution with MID=1.5 had 56 missing data points. The main reasons for missing were division by zero and non-convergence of estimates in meta-analyses.

Bias

The median method had the least bias (2.048; 1.759-2.338), and the weighted average had the highest (2.144; 1.846-2.442) (Table 3). The median method also had the highest proportion of biases in the range of [-0.5, 0.5] (17.834%; 15.209%-20.790%) and the lowest beyond -5 and 5 (28.402%; 25.241%-31.785%) (Table 4 and Figure 2). The individual method (30.251%; 27.022%-33.684%) had the highest proportion of biases beyond -5 and 5. The ANOVA test showed that the differences between groups were not statistically significant ($F(3, 3020)=0.097$, $P=0.962$). The Chi-square test revealed that the proportions for low ($\chi^2=0.111$, $P=0.991$) and high ($\chi^2=0.824$, $P=0.844$) bias among groups were not statistically significant either.

Table 3. The mean of bias, root mean squared error, confidence interval (CI) coverage percentage, empirical standard error (SE), and model-based SE for each imputation method with their 95% CI in parentheses.

Indicator	Median	Unweighted average	Weighted average	Individual	*P*-value
Bias	2.048 (1.759-2.338)	2.064 (1.774-2.355)	2.144 (1.846-2.442)	2.126 (1.831-2.420)	0.962

RMSE	5.065 (4.884-5.246)	5.114 (4.934-5.294)	5.044 (4.856-5.231)	4.852 (4.661-5.044)	0.221
CI coverage %	55.162 (52.799-57.525)	54.649 (52.305-56.993)	54.693 (52.237-57.150)	55.717 (53.185-58.250)	0.922
Empirical SE	0.154 (0.147-0.16)	0.154 (0.148-0.161)	0.151 (0.145-0.158)	0.148 (0.141-0.154)	0.463
Model-based SE	2.212 (2.122-2.303)	2.212 (2.121-2.302)	2.208 (2.118-2.299)	2.198 (2.108-2.288)	0.996

CI: confidence interval; **RMSE:** root mean squared error; **SE:** standard error.

P-values from one-way analysis of variance (ANOVA) tests.

Table 4. The mean percentage of bias, root mean squared error, and confidence interval (CI) coverage percentage in the optimal threshold for each imputation method.

Indicator	**Threshold**	**Median**	**Unweighted average**	**Weighted average**	**Individual**	***P*-value**
Bias	[-0.5, 0.5]	17.857 (15.230-20.817)	17.725 (15.106-20.678)	17.460 (14.859-20.399)	17.196 (14.612-20.120)	0.991
	<-5 or >5	28.439 (25.275-31.826)	28.439 (25.275-31.826)	30.159 (26.930-33.592)	30.291 (27.058-33.727)	0.844
RMSE	[0, 1]	2.381 (1.460-3.812)	2.249 (1.358-3.652)	3.704 (2.520-5.378)	6.217 (4.650-8.244)	**<0.001**
	>5	45.767 (42.181-49.397)	46.296 (42.704-49.927)	46.032 (42.442-49.662)	44.048 (40.482-47.674)	0.916
CI coverage %	[100, 95]	14.55 (12.155-17.314)	10.979 (8.885-13.479)	17.593 (14.982-20.538)	20.635 (17.84-23.734)	**<0.001**
	(50, 0]	42.063 (38.528-45.680)	42.725 (39.179-46.346)	42.857 (39.309-46.479)	41.931 (38.398-45.547)	0.990

CI: confidence interval; **RMSE:** root mean squared error.

P-values from Chi-squared tests.

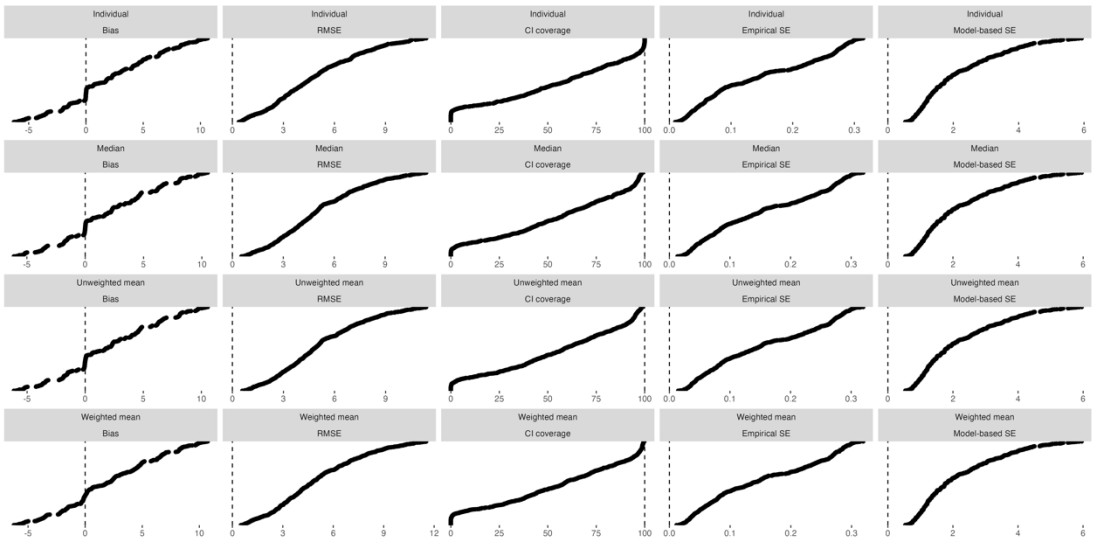


Figure 2. Bias, RMSE, CI coverage percentage, empirical SE, and model-based SE for each imputation method compared to the optimal value for each indicator.

RMSE

The individual method (4.852; 4.661-5.044) outperformed all the others in terms of low RMSE (Table 3). However, this was not statistically significant (F(3, 3020)=1.469, *P*=0.221). Also, the individual method had the highest percentage of RMSE between [0, 1] (6.209%; 4.644%-8.233%, *P*<0.001)) and the lowest proportion of RMSE>5 (44.048%; 40.482%-47.674%, *P*=0.916) (Table 4 and Figure 2).

CI coverage percentage

55.717% (53.185%-58.250%) of times the CI of the individual method included the true value. The lowest percentage belonged to the unweighted average with 54.649% (52.305%-56.993%) (Table 3). However, the differences between the groups were not statistically significant (*P*=0.922). The individual method was the only one with 100% coverage in scenarios where treatment effects were the same between the intervention and the control groups (Figure 2). The individual method also had the highest proportion

of scenarios with CI coverage of more than 95% (20.635%; 17.840%-23.730%). This difference was statistically significant (P<0.001). The individual method also had the proportions lowest with CI coverage of less than 50% (42.008%; 38.476%-45.622%), but the differences between the groups were not statistically significant (P=0.990) (Table 4).

Empirical SE

The individual method (0.148; 0.141-0.154) had the lowest mean empirical SE (Table 3). However, this was not statistically significant (F(3, 3020)= 0.857, P= 0.463). The distribution of empirical SEs was similar for all methods (Figure 2).

Model-based SE

Similar to empirical SE, the mean model-based SE of the individual method (2.198; 2.108-2.288) was lower compared to other methods (Table 3). This difference was not statistically significant (F(3, 3020)= 0.021, P= 0.996). The distribution of model-based SEs was similar for all methods (Figure 2).

Subgroups

Treatment effect in the intervention group

The lowest bias, RMSE, empirical SE, and model-based SE was seen when the treatment effect was -1. Then it gradually increased. There was a drop when the treatment effect was 1.5 or 2.0, and then it increased again. The highest bias was when the treatment effect was 4.0. The same higher performance in the treatment effect of 1.5 and 2.0 was also seen in CI coverage. Table 5 and Figure 3 illustrate the detailed information for each treatment effect.

Table 5. Bias, RMSE, CI coverage percentage, empirical SE, and model-based SE percentage for each intervention group effect size category.

Indicator	Treatment effect	Median	Unweighted average	Weighted average	Individual
Bias	-1.0	-0.005 (-0.023-0.012)	-0.038 (-0.057--0.019)	-0.207 (-0.257--0.157)	-0.022 (-0.041--0.003)
	-0.5	-0.91 (-1.308--0.511)	-0.949 (-1.349--0.550)	-1.154 (-1.556--0.752)	-0.994 (-1.426--0.562)
	0.0	-1.86 (-2.378--1.342)	-1.9 (-2.418--1.382)	-2.11 (-2.628--1.592)	-1.997 (-2.550--1.444)
	0.5	-2.521 (-3.094--1.947)	-2.556 (-3.132--1.981)	-2.73 (-3.306--2.155)	-2.641 (-3.245--2.037)
	1.0	-2.524 (-3.081--1.966)	-2.545 (-3.101--1.989)	-2.642 (-3.197--2.086)	-2.577 (-3.150--2.004)
	1.5	-1.604 (-2.097--1.111)	-1.603 (-2.096--1.110)	-1.601 (-2.093--1.108)	-1.552 (-2.049--1.055)
	2.0	0.065 (-0.358-0.487)	0.082 (-0.339-0.504)	0.189 (-0.233-0.611)	0.2 (-0.221-0.620)
	2.5	2.036 (1.635-2.438)	2.073 (1.671-2.474)	2.248 (1.844-2.651)	2.221 (1.818-2.625)
	3.0	3.855 (3.414-4.296)	3.894 (3.453-4.334)	4.102 (3.661-4.543)	4.029 (3.583-4.475)
	3.5	5.783 (5.340-6.226)	5.822 (5.379-6.266)	6.024 (5.582-6.467)	5.918 (5.466-6.369)
	4.0	7.236 (6.795-7.677)	7.273 (6.832-7.714)	7.444 (7.002-7.885)	7.322 (6.866-7.777)
RMSE	-1.0	1.471 (0.933-2.010)	1.52 (0.974-2.066)	1.367 (0.866-1.868)	1.18 (0.733-1.628)
	-0.5	2.207 (1.809-2.604)	2.255 (1.863-2.646)	2.157 (1.769-2.546)	1.969 (1.589-2.350)
	0.0	3.326 (2.841-3.811)	3.376 (2.897-3.854)	3.292 (2.790-3.794)	3.139 (2.603-3.675)
	0.5	4.398 (3.809-4.986)	4.451 (3.866-5.035)	4.347 (3.737-4.957)	4.202 (3.550-4.853)
	1.0	4.865 (4.241-5.490)	4.916 (4.295-5.537)	4.734 (4.093-5.375)	4.564 (3.887-5.242)
	1.5	4.399 (3.852-4.946)	4.455 (3.913-4.997)	4.175 (3.620-4.730)	3.98 (3.400-4.560)
	2.0	3.815 (3.470-4.159)	3.861 (3.520-4.203)	3.595 (3.265-3.925)	3.405 (3.064-3.745)
	2.5	4.069 (3.818-4.320)	4.122 (3.873-4.370)	3.998 (3.741-4.254)	3.799 (3.537-4.061)
	3.0	5.059 (4.677-5.442)	5.117 (4.736-5.497)	5.119 (4.720-5.518)	4.912 (4.506-5.317)
	3.5	6.589 (6.142-7.035)	6.635 (6.190-7.079)	6.737 (6.286-7.188)	6.535 (6.074-6.997)
	4.0	7.875 (7.415-8.336)	7.914 (7.454-8.373)	8.028 (7.566-8.490)	7.835 (7.359-8.311)
	-1.0	93.073 (92.618-93.528)	92.281 (91.463-93.100)	93.974 (93.607-94.340)	95.375 (94.997-95.753)
	-0.5	79.026 (70.696-87.356)	77.935 (69.456-86.414)	78.32 (68.957-87.683)	80.036 (70.814-89.259)

CI coverage %	0.0	71.01 (63.173-78.848)	70.144 (62.302-77.986)	70.365 (61.948-78.781)	72.111 (63.684-80.538)
	0.5	69.301 (62.695-75.907)	68.582 (62.012-75.152)	69.146 (62.230-76.062)	70.548 (63.534-77.563)
	1.0	71.275 (65.577-76.973)	70.626 (64.967-76.285)	72.245 (66.396-78.093)	73.414 (67.458-79.369)
	1.5	75.413 (70.197-80.630)	74.77 (69.598-79.942)	77.108 (71.778-82.437)	78.195 (72.759-83.632)
	2.0	74.37 (69.670-79.070)	73.754 (69.076-78.433)	75.654 (70.816-80.492)	76.823 (71.788-81.858)
	2.5	63.57 (58.054-69.086)	63.112 (57.669-68.555)	63.051 (57.153-68.949)	63.879 (57.647-70.111)
	3.0	50.135 (43.923-56.346)	49.602 (43.435-55.769)	48.714 (42.220-55.208)	49.573 (42.806-56.341)
	3.5	32.056 (26.481-37.630)	31.784 (26.250-37.317)	29.929 (24.346-35.511)	30.933 (24.960-36.907)
	4.0	17.856 (13.604-22.109)	17.679 (13.468-21.889)	16.257 (12.101-20.413)	16.788 (12.282-21.294)
Empirical SE	-1.0	0.036 (0.023-0.049)	0.037 (0.024-0.050)	0.033 (0.021-0.045)	0.028 (0.017-0.038)
	-0.5	0.047 (0.038-0.056)	0.048 (0.039-0.058)	0.043 (0.035-0.052)	0.039 (0.031-0.046)
	0.0	0.056 (0.047-0.064)	0.057 (0.048-0.065)	0.051 (0.043-0.059)	0.047 (0.040-0.054)
	0.5	0.062 (0.054-0.071)	0.063 (0.055-0.072)	0.056 (0.048-0.064)	0.051 (0.045-0.058)
	1.0	0.085 (0.077-0.092)	0.086 (0.079-0.094)	0.079 (0.072-0.086)	0.074 (0.068-0.081)
	1.5	0.13 (0.121-0.140)	0.131 (0.122-0.141)	0.127 (0.117-0.136)	0.123 (0.113-0.132)
	2.0	0.176 (0.162-0.190)	0.177 (0.163-0.190)	0.174 (0.160-0.188)	0.17 (0.156-0.185)
	2.5	0.201 (0.183-0.218)	0.201 (0.184-0.219)	0.2 (0.182-0.218)	0.196 (0.178-0.214)
	3.0	0.199 (0.180-0.218)	0.199 (0.180-0.218)	0.197 (0.178-0.217)	0.193 (0.173-0.213)
	3.5	0.192 (0.173-0.211)	0.192 (0.174-0.211)	0.191 (0.172-0.209)	0.187 (0.168-0.207)
	4.0	0.176 (0.159-0.193)	0.176 (0.159-0.193)	0.175 (0.158-0.192)	0.172 (0.155-0.190)
Model-based SE	-1.0	1.535 (0.964-2.105)	1.533 (0.963-2.103)	1.525 (0.959-2.091)	1.516 (0.953-2.079)
	-0.5	1.766 (1.344-2.188)	1.765 (1.343-2.187)	1.758 (1.338-2.177)	1.747 (1.330-2.164)
	0.0	2.007 (1.624-2.390)	2.007 (1.624-2.390)	2.002 (1.620-2.383)	1.99 (1.610-2.370)
	0.5	2.23 (1.866-2.595)	2.23 (1.865-2.594)	2.227 (1.863-2.591)	2.215 (1.853-2.577)
	1.0	2.403 (2.054-2.753)	2.403 (2.053-2.752)	2.403 (2.053-2.753)	2.391 (2.043-2.738)
	1.5	2.493 (2.163-2.822)	2.492 (2.163-2.822)	2.493 (2.164-2.823)	2.481 (2.153-2.809)
	2.0	2.477 (2.174-2.779)	2.476 (2.174-2.779)	2.476 (2.174-2.779)	2.464 (2.163-2.765)
	2.5	2.376 (2.105-2.646)	2.375 (2.104-2.646)	2.373 (2.102-2.643)	2.361 (2.092-2.630)

	3.0	2.22 (1.982-2.458)	2.219 (1.981-2.457)	2.214 (1.977-2.452)	2.204 (1.968-2.441)
	3.5	2.072 (1.848-2.296)	2.071 (1.847-2.294)	2.065 (1.842-2.288)	2.057 (1.835-2.279)
	4.0	1.938 (1.727-2.149)	1.937 (1.726-2.147)	1.931 (1.721-2.140)	1.924 (1.714-2.133)

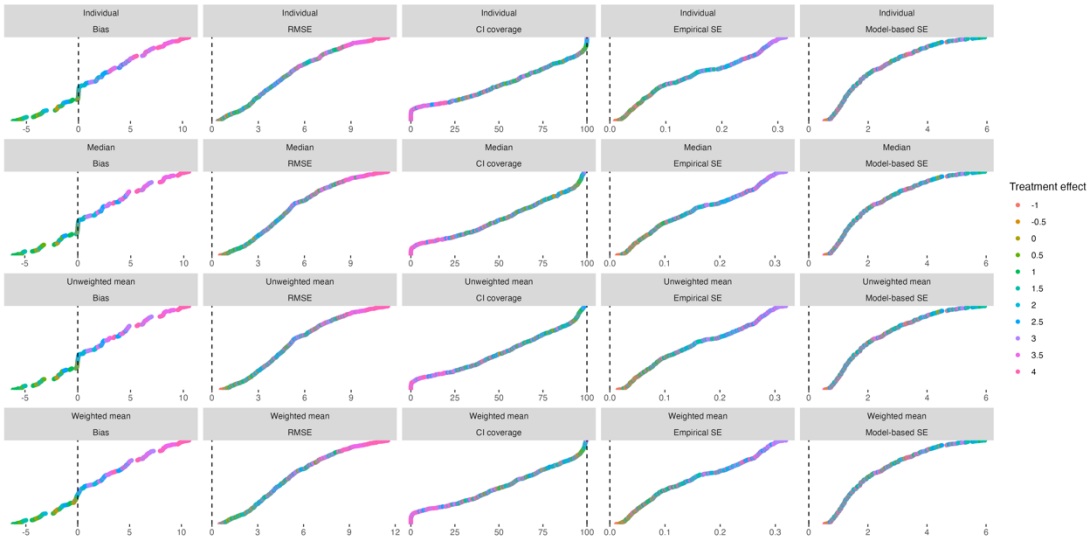


Figure 3. Bias, RMSE, CI coverage percentage, empirical SE, and model-based SE for each intervention group effect size category.

Treatment effect in the control group

The lowest bias was seen in the treatment effect of -0.5, which gradually increased. Then, there was a drop in the treatment effect of 3.0. For RMSE and CI coverage, they moved toward better performance when going from a treatment effect of -1.0 to 0.0, but they started being non-optimal afterward. Empirical SE had a decreasing trend whereas model-based SE increased and then decreased. The detailed information for each treatment effect is shown in Table 6 and Figure 4.

Table 6. Bias, RMSE, CI coverage percentage, empirical SE, and model-based SE for each control group effect size category.

Indicator	Treatment effect	Median	Unweighted average	Weighted average	Individual
Bias	-1.0	-1.656 (-2.221--1.091)	-1.656 (-2.224--1.087)	-1.652 (-2.239--1.066)	-1.601 (-2.198--1.004)
	-0.5	-0.01 (-0.627-0.607)	-0.005 (-0.627-0.616)	0.01 (-0.633-0.653)	0.101 (-0.546-0.749)
	0.0	1.864 (1.178-2.550)	1.874 (1.183-2.565)	1.917 (1.203-2.632)	2.029 (1.316-2.743)
	0.5	3.592 (2.830-4.353)	3.606 (2.840-4.373)	3.682 (2.892-4.471)	3.772 (2.992-4.553)
	1.0	4.738 (3.927-5.549)	4.76 (3.945-5.575)	4.861 (4.026-5.697)	4.868 (4.046-5.690)
	1.5	4.954 (4.145-5.764)	4.976 (4.163-5.789)	5.127 (4.302-5.953)	5.004 (4.196-5.812)
	2.0	4.246 (3.503-4.988)	4.285 (3.544-5.027)	4.455 (3.705-5.205)	4.213 (3.485-4.941)
	2.5	3.054 (2.433-3.675)	3.089 (2.467-3.711)	3.279 (2.655-3.903)	2.951 (2.354-3.549)
	3.0	1.779 (1.301-2.258)	1.817 (1.336-2.297)	2.013 (1.538-2.488)	1.667 (1.225-2.110)
RMSE	-1.0	5.133 (4.770-5.496)	5.172 (4.810-5.534)	5.107 (4.746-5.468)	5.104 (4.733-5.474)
	-0.5	4.751 (4.459-5.043)	4.796 (4.504-5.088)	4.721 (4.426-5.017)	4.669 (4.362-4.976)
	0.0	4.62 (4.240-5.001)	4.674 (4.295-5.054)	4.581 (4.183-4.979)	4.479 (4.065-4.893)
	0.5	5.151 (4.552-5.751)	5.2 (4.602-5.798)	5.096 (4.466-5.726)	4.935 (4.285-5.584)
	1.0	6.04 (5.302-6.778)	6.098 (5.364-6.831)	6.012 (5.240-6.784)	5.794 (5.007-6.580)
	1.5	6.299 (5.553-7.045)	6.346 (5.604-7.088)	6.312 (5.533-7.091)	5.982 (5.198-6.765)
	2.0	5.485 (4.828-6.143)	5.54 (4.888-6.192)	5.505 (4.813-6.197)	5.074 (4.391-5.758)
	2.5	4.13 (3.618-4.643)	4.181 (3.673-4.689)	4.143 (3.596-4.690)	3.639 (3.113-4.164)
	3.0	2.786 (2.406-3.165)	2.849 (2.474-3.223)	2.777 (2.375-3.178)	2.275 (1.916-2.633)
CI coverage %	-1.0	51.977 (47.947-56.007)	51.499 (47.488-55.509)	51.573 (47.381-55.765)	51.733 (47.446-56.020)
	-0.5	55.454 (50.474-60.434)	54.992 (50.050-59.933)	55.08 (49.917-60.244)	55.417 (50.099-60.736)
	0.0	57.281 (50.955-63.607)	56.794 (50.512-63.076)	57.158 (50.585-63.731)	57.576 (50.756-64.396)
	0.5	56.747 (48.937-64.556)	56.301 (48.563-64.040)	56.834 (48.702-64.966)	57.233 (48.807-65.660)
	1.0	52.693 (44.144-61.243)	52.187 (43.736-60.638)	52.44 (43.550-61.331)	52.935 (43.740-62.129)
	1.5	50.34 (41.510-59.170)	49.935 (41.198-58.672)	49.682 (40.548-58.816)	51.045 (41.588-60.503)
	2.0	53.046 (43.725-62.367)	52.357 (43.103-61.611)	52.004 (42.309-61.700)	54.14 (44.165-64.114)

	2.5	59.329 (49.395-69.264)	58.753 (48.870-68.636)	58.013 (47.599-68.428)	61.526 (51.030-72.023)
	3.0	70.182 (60.120-80.244)	69.248 (59.179-79.317)	68.599 (57.778-79.420)	73.922 (63.631-84.213)
Empirical SE	-1.0	0.151 (0.135-0.168)	0.152 (0.136-0.168)	0.149 (0.132-0.165)	0.147 (0.130-0.164)
	-0.5	0.177 (0.159-0.196)	0.178 (0.159-0.197)	0.175 (0.156-0.194)	0.172 (0.152-0.191)
	0.0	0.199 (0.180-0.218)	0.199 (0.180-0.218)	0.196 (0.177-0.216)	0.194 (0.174-0.213)
	0.5	0.201 (0.184-0.219)	0.201 (0.184-0.219)	0.199 (0.181-0.217)	0.196 (0.178-0.214)
	1.0	0.176 (0.162-0.190)	0.177 (0.163-0.190)	0.174 (0.160-0.188)	0.17 (0.156-0.185)
	1.5	0.131 (0.121-0.140)	0.131 (0.122-0.140)	0.128 (0.119-0.138)	0.123 (0.113-0.132)
	2.0	0.085 (0.077-0.092)	0.085 (0.078-0.093)	0.082 (0.075-0.089)	0.074 (0.068-0.081)
	2.5	0.062 (0.054-0.070)	0.063 (0.055-0.071)	0.059 (0.051-0.067)	0.051 (0.045-0.058)
	3.0	0.056 (0.047-0.065)	0.057 (0.049-0.066)	0.053 (0.045-0.061)	0.047 (0.040-0.054)
Model-based SE	-1.0	1.894 (1.706-2.081)	1.893 (1.705-2.080)	1.889 (1.701-2.076)	1.878 (1.692-2.064)
	-0.5	2.055 (1.844-2.266)	2.054 (1.843-2.265)	2.051 (1.840-2.261)	2.039 (1.830-2.248)
	0.0	2.223 (1.984-2.463)	2.223 (1.983-2.462)	2.22 (1.981-2.459)	2.208 (1.971-2.446)
	0.5	2.372 (2.101-2.643)	2.371 (2.101-2.642)	2.369 (2.098-2.639)	2.358 (2.089-2.627)
	1.0	2.468 (2.167-2.769)	2.468 (2.166-2.769)	2.465 (2.164-2.766)	2.456 (2.156-2.756)
	1.5	2.483 (2.154-2.811)	2.482 (2.154-2.811)	2.479 (2.151-2.808)	2.471 (2.144-2.798)
	2.0	2.392 (2.045-2.739)	2.391 (2.044-2.738)	2.388 (2.041-2.734)	2.38 (2.035-2.725)
	2.5	2.222 (1.859-2.586)	2.222 (1.858-2.585)	2.217 (1.854-2.580)	2.208 (1.847-2.569)
	3.0	2 (1.618-2.382)	1.999 (1.617-2.381)	1.993 (1.613-2.374)	1.983 (1.604-2.362)

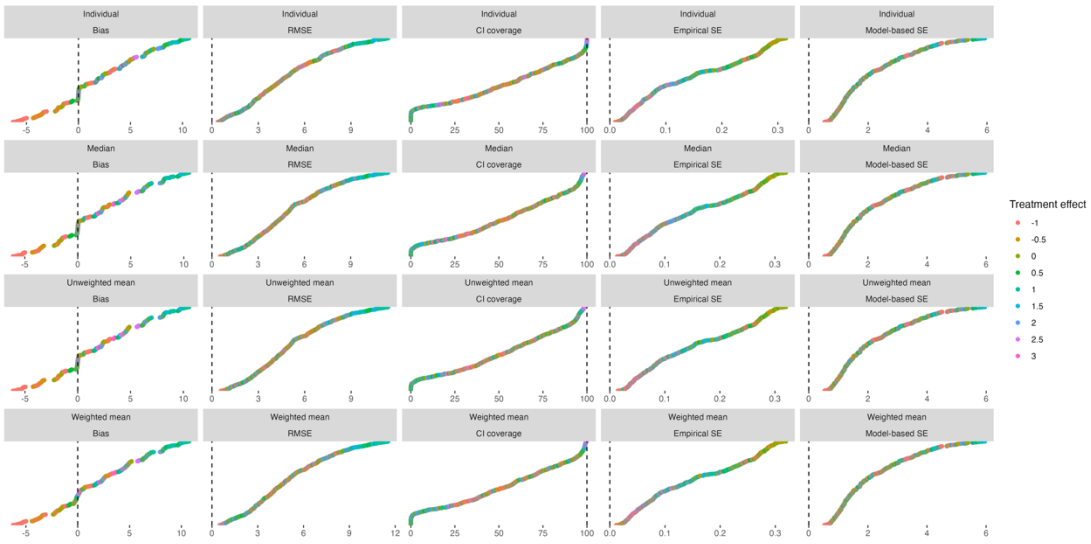


Figure 4. Bias, RMSE, CI coverage percentage, empirical SE, and model-based SE for each control group effect size category compared to the optimal value for each indicator.

Difference between treatment effects

Except for model-based SE, the highest performance was seen in a difference of 0.0. Bias, RMSE, and empirical SE gradually increased, and the highest bias was in the difference of 4.0. Then, it decreased for the differences of 4.5 and 5.0. The highest performance was seen in a difference of 0.0 for the other three performance measurements. The performance decreased until there was a difference of 3.0, and then it increased again. The highest performance of model-based SE was seen in a difference of 5.0. For detailed information, see Table 7 and Figure 5.

Table 7. Bias, RMSE, CI coverage percentage, empirical SE, and model-based SE for each difference between effect sizes categories.

Indicator	Difference	Median	Unweighted average	Weighted average	Individual
Bias	0.0	-0.003 (-0.011-0.006)	-0.012 (-0.024--0.001)	-0.053 (-0.093--0.012)	-0.004 (-0.011-0.004)

	0.5	0.333 (0.007-0.658)	0.333 (0.002-0.665)	0.332 (-0.032-0.696)	0.296 (-0.038-0.630)
	1.0	1.013 (0.403-1.624)	1.024 (0.407-1.641)	1.067 (0.420-1.714)	0.961 (0.341-1.582)
	1.5	1.718 (0.811-2.625)	1.728 (0.815-2.640)	1.808 (0.869-2.748)	1.711 (0.793-2.629)
	2.0	2.532 (1.389-3.676)	2.553 (1.406-3.700)	2.659 (1.490-3.828)	2.594 (1.443-3.745)
	2.5	3.355 (2.073-4.637)	3.383 (2.097-4.669)	3.527 (2.227-4.827)	3.512 (2.225-4.800)
	3.0	4.007 (2.702-5.312)	4.044 (2.738-5.350)	4.214 (2.902-5.527)	4.245 (2.939-5.550)
	3.5	4.44 (3.240-5.640)	4.48 (3.279-5.681)	4.667 (3.466-5.869)	4.745 (3.544-5.946)
	4.0	4.623 (3.634-5.612)	4.662 (3.673-5.652)	4.857 (3.872-5.843)	4.963 (3.974-5.952)
	4.5	4.555 (3.844-5.267)	4.598 (3.886-5.309)	4.786 (4.081-5.491)	4.901 (4.192-5.611)
	5.0	4.404 (4.351-4.457)	4.438 (4.380-4.496)	4.61 (4.514-4.705)	4.727 (4.603-4.850)
RMSE	0.0	2.003 (1.790-2.216)	2.07 (1.858-2.282)	1.752 (1.566-1.937)	1.39 (1.236-1.543)
	0.5	2.973 (2.805-3.141)	3.035 (2.868-3.203)	2.912 (2.764-3.060)	2.579 (2.455-2.703)
	1.0	4.467 (4.264-4.670)	4.516 (4.314-4.719)	4.48 (4.270-4.691)	4.202 (3.998-4.406)
	1.5	5.938 (5.622-6.253)	5.978 (5.663-6.292)	5.966 (5.632-6.299)	5.76 (5.437-6.083)
	2.0	6.944 (6.490-7.397)	6.981 (6.529-7.433)	6.961 (6.486-7.436)	6.813 (6.357-7.269)
	2.5	7.299 (6.701-7.896)	7.34 (6.745-7.936)	7.317 (6.698-7.935)	7.22 (6.617-7.824)
	3.0	7.054 (6.370-7.738)	7.095 (6.413-7.778)	7.067 (6.356-7.778)	7.019 (6.316-7.723)
	3.5	6.442 (5.734-7.151)	6.481 (5.772-7.190)	6.476 (5.736-7.217)	6.475 (5.731-7.220)
	4.0	5.789 (5.115-6.463)	5.839 (5.169-6.510)	5.872 (5.169-6.576)	5.904 (5.191-6.616)
	4.5	5.352 (4.817-5.886)	5.407 (4.875-5.939)	5.483 (4.937-6.029)	5.548 (4.999-6.097)
	5.0	5.146 (5.011-5.282)	5.194 (5.064-5.324)	5.296 (5.158-5.434)	5.38 (5.228-5.531)
CI coverage %	0.0	96.358 (96.038-96.677)	95.444 (95.048-95.840)	97.805 (97.474-98.135)	99.04 (98.756-99.324)
	0.5	77.515 (74.328-80.702)	76.508 (73.261-79.754)	76.804 (73.189-80.419)	80.615 (77.253-83.978)
	1.0	58.822 (53.997-63.647)	58.061 (53.240-62.882)	57.724 (52.497-62.951)	60.304 (54.975-65.632)
	1.5	48.11 (42.371-53.849)	47.806 (42.102-53.510)	47.464 (41.411-53.516)	48.428 (42.167-54.690)
	2.0	40.662 (34.495-46.829)	40.504 (34.354-46.654)	40.263 (33.803-46.723)	40.472 (33.778-47.167)
	2.5	38.321 (31.468-45.175)	37.987 (31.202-44.772)	37.829 (30.800-44.858)	37.681 (30.467-44.894)
	3.0	33.527 (26.014-41.040)	33.433 (25.956-40.911)	33.158 (25.346-40.971)	32.597 (24.614-40.580)

	3.5	34.081 (25.588-42.574)	34.038 (25.575-42.501)	33.172 (24.487-41.857)	32.384 (23.562-41.207)
	4.0	30.238 (20.252-40.223)	30.024 (20.105-39.944)	29.556 (19.217-39.894)	28.908 (18.366-39.450)
	4.5	34.495 (22.126-46.863)	33.901 (21.687-46.115)	31.917 (19.676-44.157)	30.349 (18.027-42.671)
	5.0	22.432 (6.152-38.712)	22.214 (6.079-38.348)	19.745 (3.787-35.703)	18.271 (2.847-33.695)
Empirical SE	0.0	0.058 (0.051-0.064)	0.059 (0.053-0.065)	0.053 (0.047-0.058)	0.045 (0.040-0.051)
	0.5	0.074 (0.068-0.081)	0.075 (0.069-0.082)	0.07 (0.064-0.076)	0.064 (0.059-0.070)
	1.0	0.101 (0.092-0.110)	0.102 (0.093-0.111)	0.098 (0.089-0.107)	0.094 (0.085-0.103)
	1.5	0.134 (0.121-0.148)	0.135 (0.122-0.148)	0.132 (0.119-0.145)	0.128 (0.115-0.142)
	2.0	0.173 (0.157-0.189)	0.173 (0.157-0.190)	0.171 (0.155-0.188)	0.167 (0.151-0.184)
	2.5	0.217 (0.202-0.232)	0.217 (0.202-0.232)	0.215 (0.200-0.230)	0.213 (0.198-0.228)
	3.0	0.254 (0.243-0.266)	0.254 (0.243-0.266)	0.254 (0.242-0.265)	0.253 (0.241-0.264)
	3.5	0.277 (0.271-0.284)	0.277 (0.270-0.284)	0.277 (0.270-0.283)	0.276 (0.270-0.283)
	4.0	0.28 (0.277-0.284)	0.28 (0.277-0.283)	0.28 (0.277-0.283)	0.28 (0.277-0.283)
	4.5	0.268 (0.267-0.270)	0.268 (0.266-0.270)	0.267 (0.266-0.269)	0.267 (0.266-0.268)
	5.0	0.245 (0.242-0.248)	0.245 (0.242-0.248)	0.244 (0.241-0.247)	0.243 (0.241-0.246)
Model-based SE	0.0	2.278 (2.023-2.532)	2.277 (2.023-2.532)	2.274 (2.020-2.528)	2.262 (2.009-2.515)
	0.5	2.3 (2.046-2.553)	2.299 (2.046-2.552)	2.296 (2.043-2.549)	2.284 (2.032-2.536)
	1.0	2.291 (2.041-2.541)	2.29 (2.040-2.540)	2.287 (2.038-2.537)	2.276 (2.028-2.525)
	1.5	2.321 (2.056-2.586)	2.321 (2.056-2.586)	2.318 (2.053-2.583)	2.308 (2.045-2.572)
	2.0	2.312 (2.032-2.593)	2.312 (2.032-2.592)	2.31 (2.029-2.590)	2.3 (2.022-2.579)
	2.5	2.256 (1.960-2.552)	2.255 (1.960-2.551)	2.253 (1.957-2.548)	2.244 (1.950-2.538)
	3.0	2.15 (1.840-2.460)	2.149 (1.839-2.459)	2.145 (1.836-2.454)	2.136 (1.829-2.444)
	3.5	2.013 (1.686-2.340)	2.012 (1.685-2.339)	2.007 (1.681-2.333)	1.997 (1.673-2.322)
	4.0	1.855 (1.504-2.207)	1.854 (1.503-2.206)	1.848 (1.498-2.198)	1.838 (1.490-2.186)
	4.5	1.689 (1.283-2.094)	1.687 (1.282-2.092)	1.68 (1.277-2.082)	1.67 (1.269-2.070)
	5.0	1.537 (0.966-2.108)	1.536 (0.965-2.106)	1.527 (0.961-2.094)	1.517 (0.955-2.078)

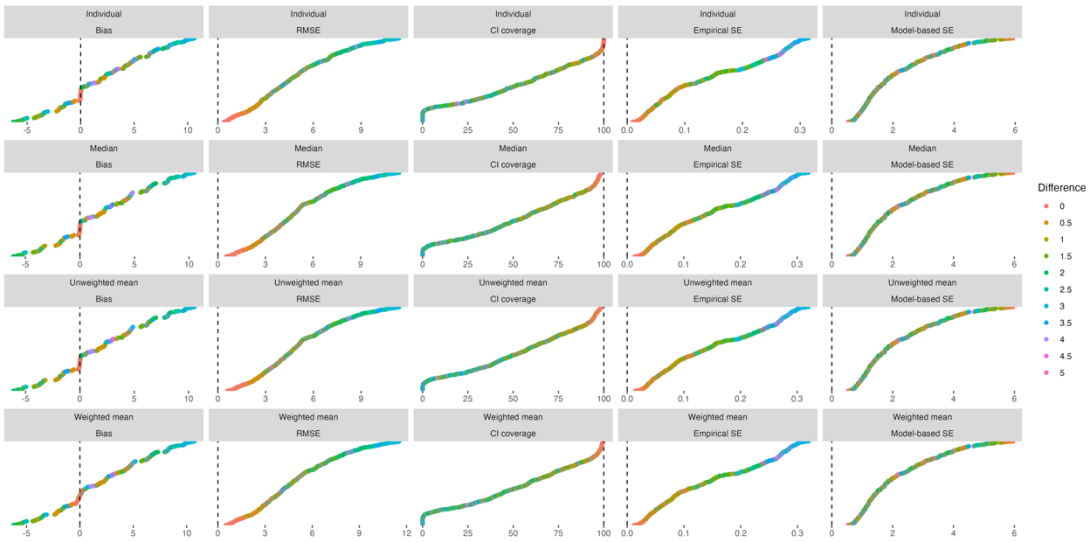


Figure 5. Bias, RMSE, CI coverage percentage, empirical SE, and model-based SE for each difference between effect sizes category compared to the optimal value for each indicator.

Number of studies

Bias and CI coverage percentage had the highest performance when the number of studies was 3. The performance decreased for meta-analyses of higher studies; however, the performance was more evident for CI coverage percentage. On the other hand, RMSE, empirical SE, and model-based SE had the lowest performance when the number of studies was 3, and their performance gradually increased to the highest when the number of studies was 50. Table 8 and Figure 6 illustrate the detailed information for each treatment effect.

Table 8. Bias, RMSE, CI coverage percentage, empirical SE, and model-based SE for each number of study categories.

Indicator	# Studies	Median	Unweighted average	Weighted average	Individual
Bias	3	2.01 (1.292-2.728)	2.016 (1.296-2.736)	2.092 (1.355-2.830)	2.083 (1.354-2.812)
	5	2.03 (1.314-2.747)	2.048 (1.327-2.769)	2.128 (1.389-2.867)	2.115 (1.385-2.845)
	10	2.051 (1.335-2.768)	2.067 (1.347-2.787)	2.147 (1.408-2.886)	2.128 (1.399-2.858)
	20	2.059 (1.341-2.776)	2.075 (1.353-2.796)	2.156 (1.415-2.896)	2.136 (1.404-2.868)
	30	2.054 (1.337-2.772)	2.074 (1.353-2.796)	2.153 (1.412-2.893)	2.134 (1.401-2.866)
	50	2.069 (1.354-2.785)	2.088 (1.368-2.808)	2.167 (1.428-2.906)	2.149 (1.418-2.880)
RMSE	3	6.029 (5.660-6.398)	6.07 (5.705-6.435)	5.806 (5.410-6.202)	5.425 (5.011-5.839)
	5	5.488 (5.090-5.887)	5.549 (5.153-5.944)	5.371 (4.945-5.796)	5.091 (4.649-5.534)
	10	4.978 (4.541-5.415)	5.051 (4.618-5.484)	4.975 (4.517-5.434)	4.803 (4.333-5.273)
	20	4.729 (4.267-5.192)	4.778 (4.318-5.239)	4.779 (4.299-5.260)	4.655 (4.165-5.145)
	30	4.63 (4.157-5.103)	4.67 (4.198-5.141)	4.702 (4.214-5.191)	4.602 (4.105-5.100)
	50	4.54 (4.057-5.022)	4.574 (4.091-5.056)	4.631 (4.133-5.128)	4.548 (4.043-5.052)
CI coverage %	3	77.203 (73.798-80.609)	76.869 (73.474-80.264)	77.998 (74.418-81.578)	79.482 (75.789-83.176)
	5	69.167 (64.855-73.479)	68.669 (64.362-72.976)	69.4 (64.784-74.016)	70.658 (65.810-75.506)
	10	58.22 (52.797-63.643)	57.513 (52.180-62.846)	57.52 (51.821-63.219)	58.257 (52.335-64.179)
	20	47.649 (41.722-53.576)	47.076 (41.228-52.924)	46.675 (40.546-52.803)	47.63 (41.301-53.959)
	30	42.419 (36.430-48.408)	41.873 (35.947-47.798)	41.348 (35.203-47.494)	42.146 (35.812-48.480)
	50	36.75 (30.772-42.727)	36.326 (30.411-42.242)	35.654 (29.552-41.755)	36.556 (30.266-42.846)
Empirical SE	3	0.184 (0.170-0.197)	0.184 (0.171-0.198)	0.178 (0.164-0.192)	0.17 (0.155-0.185)
	5	0.167 (0.152-0.182)	0.168 (0.153-0.182)	0.162 (0.147-0.178)	0.157 (0.141-0.172)
	10	0.152 (0.136-0.168)	0.153 (0.137-0.169)	0.149 (0.133-0.166)	0.146 (0.129-0.163)
	20	0.144 (0.127-0.161)	0.144 (0.127-0.161)	0.142 (0.125-0.159)	0.14 (0.123-0.157)
	30	0.14 (0.123-0.157)	0.14 (0.123-0.157)	0.139 (0.122-0.156)	0.137 (0.120-0.155)
	50	0.137 (0.120-0.155)	0.137 (0.120-0.155)	0.137 (0.119-0.154)	0.135 (0.118-0.153)
Model-based SE	3	4.238 (4.090-4.386)	4.237 (4.089-4.385)	4.231 (4.082-4.379)	4.213 (4.066-4.360)
	5	3.155 (3.038-3.272)	3.154 (3.037-3.271)	3.149 (3.032-3.267)	3.135 (3.019-3.251)

	10	2.174 (2.090-2.258)	2.173 (2.089-2.257)	2.17 (2.086-2.254)	2.159 (2.076-2.243)
	20	1.519 (1.459-1.579)	1.519 (1.459-1.579)	1.516 (1.456-1.576)	1.509 (1.449-1.568)
	30	1.235 (1.186-1.284)	1.235 (1.185-1.284)	1.233 (1.183-1.282)	1.226 (1.178-1.275)
	50	0.954 (0.915-0.992)	0.953 (0.915-0.991)	0.952 (0.913-0.990)	0.947 (0.909-0.984)

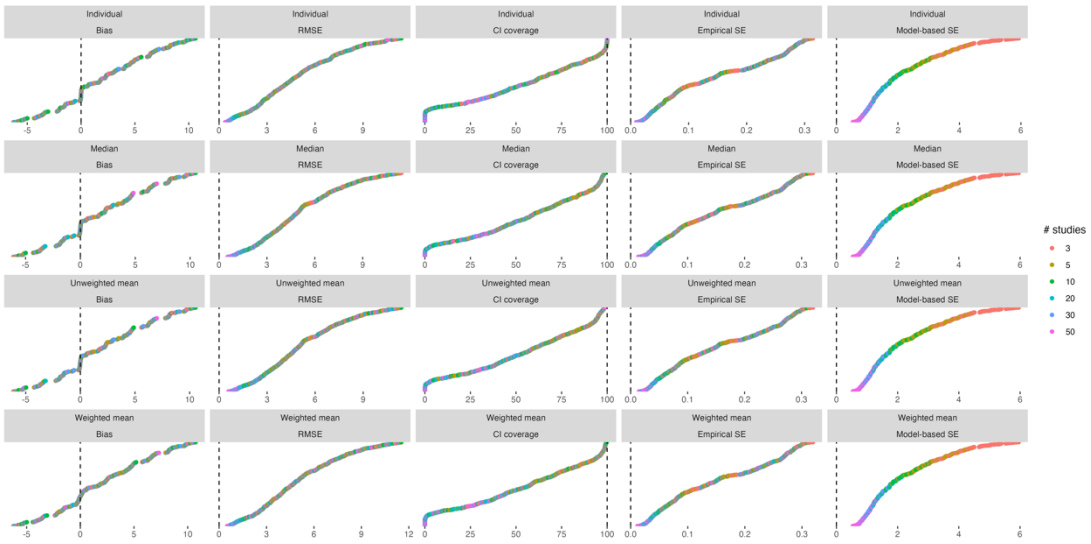


Figure 6. Bias, RMSE, CI coverage percentage, empirical SE, and model-based SE for each number of study categories compared to the optimal value for each indicator.

Study size

All performance measures, except CI coverage percentage were lower for meta-analyses of studies with mixed study sizes. Table 9 and Figure 7 illustrate the detailed information for each treatment effect.

Table 9. Bias, RMSE, CI coverage percentage, empirical SE, and model-based SE for each study size category.

Indicator	Study sizes	Median	Unweighted average	Weighted average	Individual
Bias	mixed	2.032 (1.622-2.441)	2.046 (1.634-2.457)	2.092 (1.674-2.509)	2.08 (1.665-2.494)
	all-small	2.06 (1.649-2.471)	2.077 (1.664-2.490)	2.189 (1.761-2.618)	2.169 (1.747-2.590)
RMSE	mixed	4.912 (4.653-5.171)	4.989 (4.732-5.247)	4.863 (4.595-5.130)	4.708 (4.435-4.981)
	all-small	5.219 (4.966-5.473)	5.241 (4.987-5.494)	5.225 (4.962-5.488)	5 (4.732-5.269)
CI coverage %	mixed	50.361 (46.976-53.747)	49.575 (46.259-52.891)	50.353 (46.834-53.873)	51.164 (47.529-54.798)
	all-small	60.108 (56.873-63.344)	59.867 (56.624-63.110)	59.178 (55.795-62.562)	60.413 (56.932-63.894)
Empirical SE	mixed	0.149 (0.139-0.158)	0.15 (0.141-0.160)	0.146 (0.137-0.156)	0.143 (0.133-0.153)
	all-small	0.159 (0.150-0.168)	0.159 (0.149-0.168)	0.156 (0.147-0.165)	0.152 (0.142-0.161)
Model-based SE	mixed	1.847 (1.742-1.951)	1.846 (1.742-1.951)	1.845 (1.740-1.949)	1.839 (1.735-1.943)
	all-small	2.578 (2.440-2.717)	2.577 (2.439-2.716)	2.572 (2.434-2.711)	2.557 (2.420-2.695)

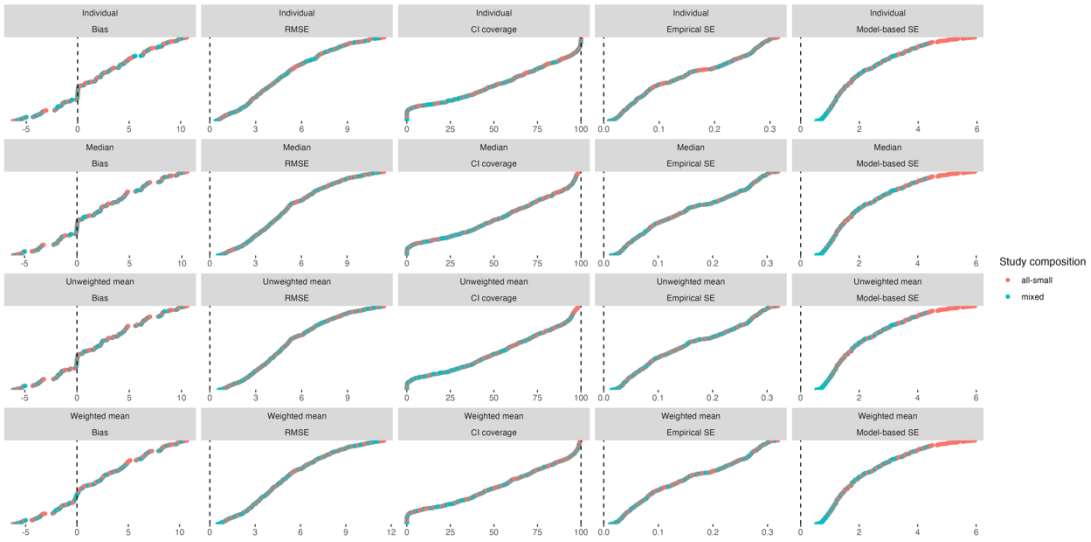


Figure 7. Bias, RMSE, CI coverage percentage, empirical SE, and model-based SE for the study size category compared to the optimal value for each indicator.

Most influential factor

The regression analysis showed that the most influential factors were the difference between treatment effects (R^2 between 0.169 and 0.731) and the treatment effect in the intervention group (R^2 between 0.032 and 0.740) (Table 10). The least influential factor for all the performance measurements was the imputation method.

Table 10. R^2 for regression models for each indicator involving different variables.

Variable	Bias	RMSE	CI coverage %	Empirical SE	Model-based SE
Imputation method	<0.001	0.001	<0.001	0.001	<0.001
Number of studies	<0.001	0.030	0.191	0.026	0.833
Study size	<0.001	0.003	0.020	0.002	0.083
Treatment effect in the intervention group	0.740	0.387	0.414	0.345	0.032
Treatment effect in the control group	0.327	0.093	0.016	0.251	0.028
Difference in treatment effects	0.169	0.548	0.452	0.731	0.019
Overall	0.999	0.880	0.861	0.731	0.968

Sensitivity analyses

Skewed distribution

All the imputation methods performed significantly less optimally when the distribution was skewed. The bias and RMSE were doubled, and the CI coverage percentage dropped by almost 20% (Table 11). The differences between empirical SEs were minimal. The difference was mainly because of the higher proportion of estimates in the extremely suboptimal performance thresholds, whereas the proportions were the same in optimal thresholds (Table 12 and Figure 8).

Table 11. The mean of bias, root mean squared error, confidence interval (CI) coverage percentage, empirical SE, and model-based SE for each imputation method with their 95% CI in parentheses when data is skewed.

Indicator	Median	Unweighted average	Weighted average	Individual	*P*-value
Bias	4.409 (3.829-4.988)	4.394 (3.812-4.977)	4.754 (4.167-5.341)	4.507 (3.926-5.087)	0.815
RMSE	9.132 (8.763-9.502)	9.188 (8.819-9.557)	9.27 (8.891-9.648)	8.955 (8.573-9.338)	0.692
CI coverage %	35.622 (33.085-38.158)	35.454 (32.939-37.968)	34.774 (32.19-37.358)	36.015 (33.348-38.682)	0.926
Empirical SE	0.156 (0.15-0.163)	0.156 (0.15-0.162)	0.154 (0.148-0.161)	0.15 (0.143-0.156)	0.568
Model-based SE	2.537 (2.437-2.638)	2.536 (2.436-2.637)	2.53 (2.429-2.63)	2.523 (2.423-2.623)	0.998

CI: confidence interval; **RMSE:** root mean squared error; **SE:** standard error.

P-values from one-way analysis of variance (ANOVA) tests.

Table 12. The mean percentage of bias, root mean squared error, confidence interval (CI) coverage percentage, empirical SE, and model-based SE in the optimal threshold for each imputation method when data is skewed.

Indicator	Threshold	Median	Unweighted average	Weighted average	Individual	*P*-value
Bias	[-0.5, 0.5]	16.931 (14.365-19.84)	17.063 (14.488-19.98)	13.889 (11.545-16.608)	17.725 (15.106-20.678)	0.256
	<-5 or >5	56.878 (53.256-60.43)	57.804 (54.186-61.342)	59.788 (56.185-63.29)	58.862 (55.252-62.382)	0.892
RMSE	[0, 1]	2.513 (1.563-3.971)	2.116 (1.256-3.491)	3.439 (2.304-5.069)	5.952 (4.421-7.946)	**<0.001**
	>5	72.884 (69.537-75.994)	73.28 (69.946-76.375)	74.339 (71.039-77.387)	71.296 (67.902-74.471)	0.920
CI coverage %	[100, 95]	16.534 (13.995-19.42)	16.402 (13.872-19.28)	16.931 (14.365-19.84)	18.122 (15.477-21.096)	0.845
	(50, 0]	68.386 (64.918-71.665)	68.519 (65.054-71.792)	69.312 (65.866-72.559)	67.989 (64.513-71.281)	0.992

CI: confidence interval; **RMSE:** root mean squared error.

P-values from Chi-squared tests.

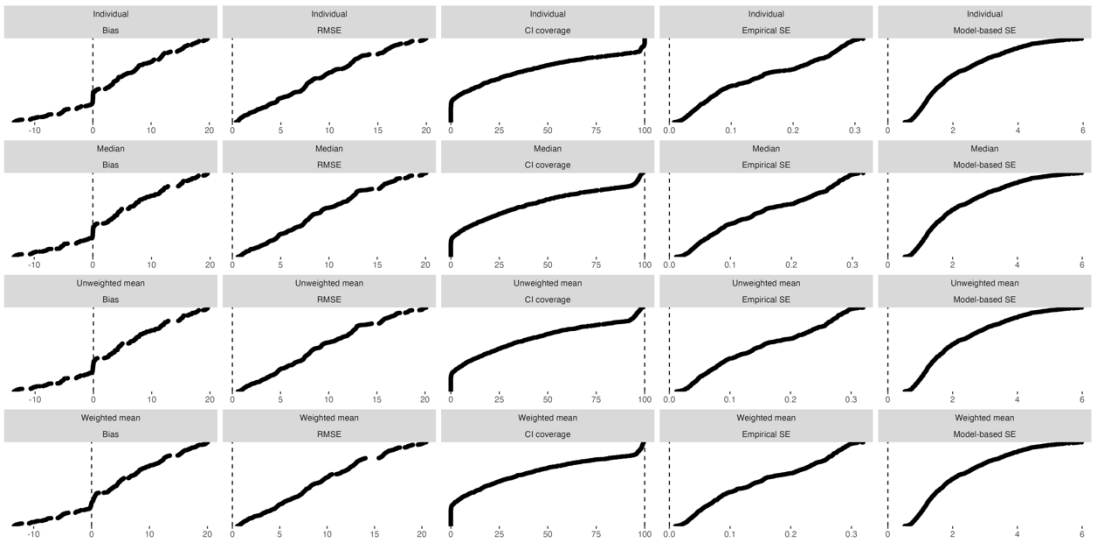


Figure 8. Bias, RMSE, CI coverage percentage, empirical SE, and model-based SE for each imputation method compared to the optimal value for each indicator when data is skewed.

MID=1.0

Choosing an MID=1.0 did not make any difference in the estimates from any of the imputation methods (Table 13, Table 14, and Figure 9).

Table 13. The mean of bias, root mean squared error, confidence interval (CI) coverage percentage, and empirical SE, and model-based SE for each imputation method with their 95% CI in parentheses when MID=1.

Indicator	Median	Unweighted average	Weighted average	Individual	*P*-value
Bias	2.042 (1.753-2.33)	2.058 (1.768-2.348)	2.137 (1.839-2.435)	2.121 (1.827-2.416)	0.961
RMSE	5.055 (4.874-5.235)	5.103 (4.923-5.283)	5.034 (4.846-5.221)	4.845 (4.654-5.036)	0.226
CI coverage %	55.338 (52.974-57.702)	54.796 (52.451-57.141)	54.879 (52.418-57.34)	55.874 (53.339-58.408)	0.923
Empirical SE	0.154 (0.148-0.161)	0.155 (0.148-0.161)	0.152 (0.145-0.159)	0.147 (0.141-0.154)	0.998

Model-based SE	2.289 (2.199-2.379)	2.288 (2.198-2.378)	2.284 (2.194-2.374)	2.276 (2.186-2.365)	0.998

CI: confidence interval; **RMSE:** root mean squared error; **SE:** standard error.

P-values from one-way analysis of variance (ANOVA) tests.

Table 14. The mean percentage of bias, root mean squared error, confidence interval (CI) coverage percentage, empirical SE, and model-based SE in the optimal threshold for each imputation method when MID=1.

Indicator	**Threshold**	**Median**	**Unweighted average**	**Weighted average**	**Individual**	***P*-value**
Bias	[-0.5, 0.5]	17.989 (15.354-20.957)	17.989 (15.354-20.957)	17.46 (14.859-20.399)	17.196 (14.612-20.12)	0.977
	<-5 or >5	28.307 (25.148-31.69)	28.307 (25.148-31.69)	17.196 (14.612-20.12)	30.026 (26.803-33.456)	0.871
RMSE	[0, 1]	2.646 (1.667-4.13)	2.249 (1.358-3.652)	3.968 (2.739-5.685)	6.217 (4.65-8.244)	**<0.001**
	>5	45.899 (42.312-49.53)	46.825 (43.228-50.456)	45.767 (42.181-49.397)	44.048 (40.482-47.674)	0.881
CI coverage %	[100, 95]	22.354 (19.466-25.529)	21.693 (18.84-24.839)	23.545 (20.596-26.767)	25.529 (22.487-28.824)	0.430
	(50, 0]	41.534 (38.008-45.148)	42.196 (38.658-45.814)	42.989 (39.44-46.612)	41.931 (38.398-45.547)	0.977

CI: confidence interval; **RMSE:** root mean squared error.

P-values from Chi-squared tests.

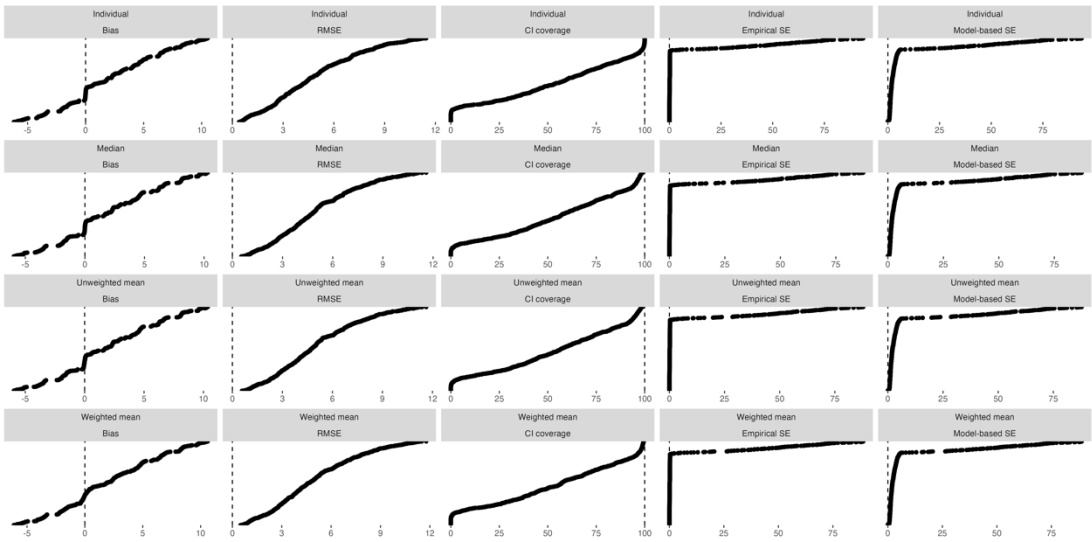


Figure 9. Bias, RMSE, CI coverage percentage, empirical SE, and model-based SE for each imputation method compared to the optimal value for each indicator when MID=1.

Negative treatment effect difference

When the treatment effect in the intervention group was lower than that in the control group, bias, RMSE, and CI coverage percentage showed higher performance for all the imputation methods. The bias was negative, and RMSE dropped by almost 0.5. The CI coverage percentage was also higher by almost 5%. On the other hand, empirical SE and model-based SE showed increase (Table 15). The proportion of performance measurements in lower thresholds was lower and higher in extreme thresholds (Table 16 and Figure 10).

Table 15. The mean of bias, root mean squared error, confidence interval (CI) coverage percentage, empirical SE, and model-based SE for each imputation method with their 95% CI in parentheses the difference in treatment effects is negative.

Indicator	Median	Unweighted average	Weighted average	Individual	*P*-value
Bias	-0.328 (-0.652--0.005)	-0.355 (-0.678--0.032)	-0.493 (-0.812--0.174)	-0.413 (-0.746--0.08)	0.900
RMSE	4.74 (4.573-4.907)	4.77 (4.604-4.936)	4.596 (4.428-4.764)	4.539 (4.366-4.713)	0.168

CI coverage %	60.023 (57.618-62.429)	59.748 (57.361-62.135)	61.117 (58.591-63.642)	61.089 (58.437-63.741)	0.815
Empirical SE	0.163 (0.155-0.171)	0.163 (0.155-0.171)	0.161 (0.153-0.169)	0.157 (0.149-0.165)	0.711
Model-based SE	2.294 (2.172-2.417)	2.294 (2.171-2.416)	2.29 (2.167-2.413)	2.281 (2.159-2.403)	0.999

CI: confidence interval; **RMSE:** root mean squared error; **SE:** standard error.

P-values from one-way analysis of variance (ANOVA) tests.

Table 16. The mean percentage of bias, root mean squared error, confidence interval (CI) coverage percentage, empirical SE, and model-based SE in the optimal threshold for each imputation method when MID=1.

Indicator	**Threshold**	**Median**	**Unweighted average**	**Weighted average**	**Individual**	***P*-value**
Bias	[-0.5, 0.5]	5.729 (4.702-6.958)	4.803 (3.865-5.947)	5.729 (4.702-6.958)	5.382 (4.387-6.58)	0.849
	<-5 or >5	37.211 (34.934-39.545)	36.343 (34.079-38.668)	35.417 (33.168-37.731)	38.889 (36.589-41.239)	0.728
RMSE	[0, 1]	0	0	0	0	NA
	>5	61.979 (59.637-64.268)	62.269 (59.929-64.553)	59.144 (56.779-61.467)	58.333 (55.964-60.665)	0.677
CI coverage %	[100, 90]	8.102 (6.88-9.514)	7.755 (6.558-9.142)	10.706 (9.307-12.283)	11.921 (10.45-13.565)	**0.023**
	(50, 0]	52.778 (50.391-55.152)	53.472 (51.086-55.842)	51.389 (49.003-53.769)	52.836 (50.449-55.209)	0.990

CI: confidence interval; **RMSE:** root mean squared error.

P-values from Chi-squared tests.

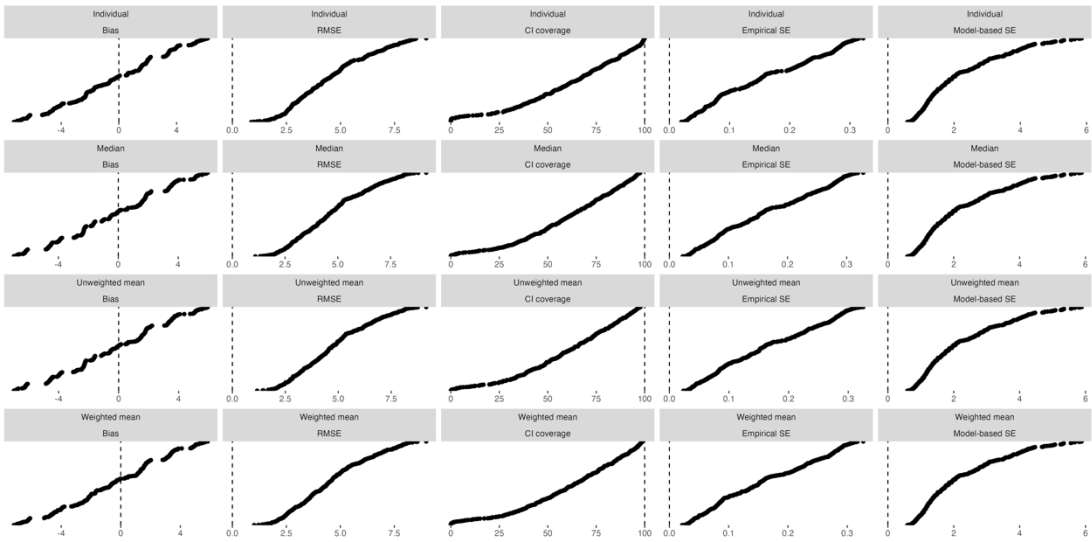


Figure 10. Bias, RMSE, CI coverage percentage, empirical SE, and model-based SE for each imputation method compared to the optimal value for each indicator when the treatment effect in the control group is higher than the intervention group.

Discussion

Summary of findings

This study presents a simulation-based assessment of four methods for converting continuous pain outcomes into responder analyses within the framework of meta-analysis. The primary focus was on estimating the pooled RD of achieving the MID using pain scores from a 10cm VAS. Whereas the median method showed the lowest bias, the individual method, which calculates RRs for each study and then estimates the pooled RD, demonstrated the lowest RMSE, empirical SE, and model-based SE and the highest CI coverage. However, the differences between the groups were minimal and not statistically significant. The only statistically significant differences between the groups were observed when comparing the percentage of RMSE between [0, 1], in which the individual method outperformed other methods. Generally, the weighted average method performed less optimally compared to other methods.

The highest performance was seen when the treatment effect was -1 in the intervention group and -0.5 in the control group. When we explored the performance for the difference between treatment effects in the intervention and control groups, we found that difference=0.0 had the highest performance. In other words, when the treatment does not work, the estimated RD will be less biased. This can be because the normal cumulative distribution function was used to estimate proportions. When there is no difference between the arms, the highest proportion of estimated RDs is close to 0, which is the mean of the standard normal distribution (Blázquez-Rincón et al., 2023).

With an increase in the number of studies in the meta-analysis, the performance in bias and CI coverage became poorer. However, RMSE, empirical SE, and model-based SE showed an opposite trend, with the highest performance being when the number of studies was 50. While the differences in bias were not large, the differences for the others were considerable, such that they performed almost twice as poorly. The lower performance of CI coverage may be because of estimating RDs based on RRs and their associated 95% CI. As mentioned before, this will narrow the 95% CI (Newcombe & Bender, 2014), which in this case yielded poor performance in the meta-analyses with a higher number of studies. About RMSE, empirical SE, and model-based SE, higher number of studies is related to a better estimate of the true value based on the "law of large numbers" (Dekking et al., 2006).

The most influential factor in the methods' performance was the difference between treatment effects. Since the difference includes both treatment effects in the intervention and control groups, they also had a high influence on performance. Other factors had minimal effect on the performance. This shows the effect of the normal distribution assumption in estimating proportions.

We performed three sensitivity analyses. When the distribution of the change scores was skewed, the performance of the methods decreased largely. This is not surprising due to the normal distribution assumption in our estimations. When MID=1.0 was used, the performance did not differ, meaning that other than its effect on the number of responders, it did not affect the performance. Interestingly, when the treatment effect in the intervention group was lower than the control group (negative difference), a higher performance was observed. This means that negative RDs are

better estimated. However, the dataset for negative differences was not as comprehensive as the main dataset.

Implications

The implications of these findings can be significant for the practice of meta-analysis, particularly in the context of pain research, where continuous outcomes are commonly reported. Converting continuous data into responder rates not only enhances the interpretability of pooled estimates but also aligns with the principles of patient-centered care, focusing on outcomes that are meaningful to patients and clinicians. This approach facilitates clearer communication of the clinical significance of treatment effects, aiding in informed decision-making by healthcare professionals and policymakers.

Based on our results, the individual method can be a better option compared to the other three methods for estimating RDs from continuous VAS scores in meta-analyses. However, it needs more computations, which may not add a greater value compared to the median method in most of the cases. Plus, the estimates were poorly performed in some extreme cases and therefore the results of the imputations should be used with caution.

Therefore, we believe the best option would be to encourage researchers to publish the responder analysis of their studies alongside continuous VAS. This can be done through a standardized guideline for reporting trials of pain relief, like the recommendations from OMERACT (Busse et al., 2015).

Strengths and limitations

The methodological strengths of this study lie in its comprehensive simulation framework, which varied key parameters such as treatment effect size, control group effect, study sample size, and the number of studies per meta-analysis. This allowed for a thorough evaluation of each method under diverse scenarios. Multiple performance metrics, including bias, RMSE, CI coverage, empirical SE, and model-based SE, provided a holistic assessment of the strengths and limitations of each method.

However, there are limitations to consider. The assumption of normality for VAS scores, while common in many trials, may not always reflect real-world data distributions. Future research should explore the impact of alternative distributions, such as the beta distribution, on the performance of these methods. Additionally, the individual method, despite its relative advantages, requires significant computational resources, which may pose challenges for researchers with limited access to high-performance computing facilities.

Future research should focus on empirical validation of these methods using real-world data from clinical trials. Extending the applicability of these methods to other continuous outcomes, such as quality of life or depression scales, could further enhance their utility. Integrating these methods with advanced statistical techniques, such as Bayesian meta-analysis, may also provide additional insights and improve accuracy.

In conclusion, this study provides robust evidence supporting the individual method as the most reliable and accurate approach for converting continuous pain outcomes to responder analyses in meta-analyses. By adopting this method, researchers

can improve the interpretability and clinical relevance of meta-analytic findings, ultimately enhancing patient care and clinical decision-making. The insights gained from this study are pivotal for advancing the methodological rigor of meta-analyses and optimizing the use of patient-reported outcomes in evidence synthesis.

Bibliography

Akl, E. A., Oxman, A. D., Herrin, J., Vist, G. E., Terrenato, I., Sperati, F., Costiniuk, C., Blank, D., & Schünemann, H. (2011). Using alternative statistical formats for presenting risks and risk reductions. *Cochrane Database Syst Rev*, *2011*(3), Cd006776. https://doi.org/10.1002/14651858.CD006776.pub2

Blázquez-Rincón, D., Sánchez-Meca, J., Botella, J., & Suero, M. (2023). Heterogeneity estimation in meta-analysis of standardized mean differences when the distribution of random effects departs from normal: A Monte Carlo simulation study. *BMC Med Res Methodol*, *23*(1), 19. https://doi.org/10.1186/s12874-022-01809-0

Busse, J. W., Bartlett, S. J., Dougados, M., Johnston, B. C., Guyatt, G. H., Kirwan, J. R., Kwoh, K., Maxwell, L. J., Moore, A., Singh, J. A., Stevens, R., Strand, V., Suarez-Almazor, M. E., Tugwell, P., & Wells, G. A. (2015). Optimal Strategies for Reporting Pain in Clinical Trials and Systematic Reviews: Recommendations from an OMERACT 12 Workshop. *J Rheumatol*, *42*(10), 1962-1970. https://doi.org/10.3899/jrheum.141440

Busse, J. W., Wang, L., Kamaleldin, M., Craigie, S., Riva, J. J., Montoya, L., Mulla, S. M., Lopes, L. C., Vogel, N., Chen, E., Kirmayr, K., De Oliveira, K., Olivieri, L., Kaushal, A., Chaparro, L. E., Oyberman, I., Agarwal, A., Couban, R., Tsoi, L., . . . Guyatt, G. H. (2018). Opioids for Chronic Noncancer Pain: A Systematic Review and Meta-analysis. *Jama*, *320*(23), 2448-2460. https://doi.org/10.1001/jama.2018.18472

Cappelleri, J. C., & Chambers, R. (2021). Addressing Bias in Responder Analysis of Patient-Reported Outcomes. *Ther Innov Regul Sci*, *55*(5), 989-1000. https://doi.org/10.1007/s43441-021-00298-5

Carrasco-Labra, A., Devji, T., Qasim, A., Phillips, M. R., Wang, Y., Johnston, B. C., Devasenapathy, N., Zeraatkar, D., Bhatt, M., Jin, X., Brignardello-Petersen, R., Urquhart, O., Foroutan, F., Schandelmaier, S., Pardo-Hernandez, H., Hao, Q., Wong, V., Ye, Z., Yao, L., . . . Guyatt, G. H. (2021). Minimal important difference estimates for patient-reported outcomes: A systematic survey. *J Clin Epidemiol*, *133*, 61-71. https://doi.org/10.1016/j.jclinepi.2020.11.024

Collister, D., Bangdiwala, S., Walsh, M., Mian, R., Lee, S. F., Furukawa, T. A., & Guyatt, G. (2021). Patient reported outcome measures in clinical trials should be initially analyzed as continuous outcomes for statistical significance and responder analyses should be reserved as secondary analyses. *J Clin Epidemiol*, *134*, 95-102. https://doi.org/10.1016/j.jclinepi.2021.01.026

Dekking, F. M., Kraaikamp, C., Lopuhaä, H. P., & Meester, L. E. (2006). *A Modern Introduction to Probability and Statistics: Understanding why and how*. Springer Science & Business Media.

Dworkin, R. H., Turk, D. C., Wyrwich, K. W., Beaton, D., Cleeland, C. S., Farrar, J. T., Haythornthwaite, J. A., Jensen, M. P., Kerns, R. D., Ader, D. N., Brandenburg, N., Burke, L. B., Cella, D., Chandler, J., Cowan, P., Dimitrova, R., Dionne, R., Hertz, S., Jadad, A. R., . . . Zavisic, S. (2008). Interpreting the clinical importance of treatment outcomes in chronic pain clinical trials: IMMPACT recommendations. *J Pain*, *9*(2), 105-121. https://doi.org/10.1016/j.jpain.2007.09.005

Esmail, L. C., Kapp, P., Assi, R., Wood, J., Regan, G., Ravaud, P., & Boutron, I. (2023). Sharing of Individual Patient-Level Data by Trialists of Randomized Clinical Trials of Pharmacological Treatments for COVID-19. *Jama*, *329*(19), 1695-1697. https://doi.org/10.1001/jama.2023.4590

Froud, R., Eldridge, S., Kovacs, F., Breen, A., Bolton, J., Dunn, K., Fritz, J., Keller, A., Kent, P., Lauridsen, H. H., Ostelo, R., Pincus, T., van Tulder, M., Vogel, S., & Underwood, M. (2011). Reporting outcomes of back pain trials: a modified Delphi study. *Eur J Pain*, *15*(10), 1068-1074. https://doi.org/10.1016/j.ejpain.2011.04.015

Gabelica, M., Bojčić, R., & Puljak, L. (2022). Many researchers were not compliant with their published data sharing statement: a mixed-methods study. *J Clin Epidemiol*, *150*, 33-41. https://doi.org/10.1016/j.jclinepi.2022.05.019

Gabelica, M., Cavar, J., & Puljak, L. (2019). Authors of trials from high-ranking anesthesiology journals were not willing to share raw data. *J Clin Epidemiol*, *109*, 111-116. https://doi.org/10.1016/j.jclinepi.2019.01.012

Heller, G. Z., Manuguerra, M., & Chow, R. (2016). How to analyze the Visual Analogue Scale: Myths, truths and clinical relevance. *Scand J Pain*, *13*, 67-75. https://doi.org/10.1016/j.sjpain.2016.06.012

Henschke, N., van Enst, A., Froud, R., & Ostelo, R. W. (2014). Responder analyses in randomised controlled trials for chronic low back pain: an overview of currently used methods. *Eur Spine J*, *23*(4), 772-778. https://doi.org/10.1007/s00586-013-3155-0

Higgins, J., & Green, S. (2022). Random-effects (DerSimonian and Laird) method for meta-analysis. *Cochrane handbook for systematic reviews of interventions. Version*, *5*(0).

Higgins, J., Thomas, J., Chandler, J., Cumpston, M., Li, T., & Page, M. (2019). A generic inverse-variance approach to meta-analysis. *Cochrane handbook for systematic reviews of interventions: John Wiley & Sons*, 245.

Higgins, J. P., Li, T., & Deeks, J. J. (2019). Choosing effect measures and computing estimates of effect. *Cochrane handbook for systematic reviews of interventions*, 143-176.

Jaeschke, R., Singer, J., & Guyatt, G. H. (1989). Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials*, *10*(4), 407-415. https://doi.org/10.1016/0197-2456(89)90005-6

Johnston, B. C., Alonso-Coello, P., Friedrich, J. O., Mustafa, R. A., Tikkinen, K. A. O., Neumann, I., Vandvik, P. O., Akl, E. A., da Costa, B. R., Adhikari, N. K., Dalmau, G. M., Kosunen, E., Mustonen, J., Crawford, M. W., Thabane, L., & Guyatt, G. H. (2016). Do clinicians understand the size of treatment effects? A randomized survey across 8 countries. *Cmaj*, *188*(1), 25-32. https://doi.org/10.1503/cmaj.150430

Johnston, B. C., Patrick, D. L., Thorlund, K., Busse, J. W., da Costa, B. R., Schünemann, H. J., & Guyatt, G. H. (2013). Patient-reported outcomes in meta-analyses-part 2: methods for improving interpretability for decision-makers. *Health Qual Life Outcomes*, *11*, 211. https://doi.org/10.1186/1477-7525-11-211

Johnston, B. C., Thorlund, K., Schünemann, H. J., Xie, F., Murad, M. H., Montori, V. M., & Guyatt, G. H. (2010). Improving the interpretation of quality of life evidence in meta-analyses: the application of minimal important difference

units. *Health Qual Life Outcomes*, *8*, 116. https://doi.org/10.1186/1477-7525-8-116

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Stat Med*, *38*(11), 2074-2102. https://doi.org/10.1002/sim.8086

Newcombe, R. G., & Bender, R. (2014). Implementing GRADE: calculating the risk difference from the baseline risk and the relative risk. *Evid Based Med*, *19*(1), 6-8. https://doi.org/10.1136/eb-2013-101340

Polanin, J. R. (2018). Efforts to retrieve individual participant data sets for use in a meta-analysis result in moderate data sharing but many data sets remain missing. *J Clin Epidemiol*, *98*, 157-159. https://doi.org/10.1016/j.jclinepi.2017.12.014

Schünemann, H. J., & Guyatt, G. H. (2005). Commentary--goodbye M(C)ID! Hello MID, where do you come from? *Health Serv Res*, *40*(2), 593-597. https://doi.org/10.1111/j.1475-6773.2005.00374.x

Scott, J., & Huskisson, E. C. (1976). Graphic representation of pain. *Pain*, *2*(2), 175-184.

Scutt, P., Woodhouse, L. J., Montgomery, A. A., & Bath, P. M. (2020). Data sharing: experience of accessing individual patient data from completed randomised controlled trials in vascular and cognitive medicine. *BMJ Open*, *10*(9), e038765. https://doi.org/10.1136/bmjopen-2020-038765

Snapinn, S. M., & Jiang, Q. (2007). Responder analyses and the assessment of a clinically relevant treatment effect. *Trials*, *8*, 31. https://doi.org/10.1186/1745-6215-8-31

Thorlund, K., Walter, S. D., Johnston, B. C., Furukawa, T. A., & Guyatt, G. H. (2011). Pooling health-related quality of life outcomes in meta-analysis-a tutorial and review of methods for enhancing interpretability. *Res Synth Methods*, *2*(3), 188-203. https://doi.org/10.1002/jrsm.46

Wang, Y., Devji, T., Carrasco-Labra, A., King, M. T., Terluin, B., Terwee, C. B., Walsh, M., Furukawa, T. A., & Guyatt, G. H. (2023). A step-by-step approach for selecting an optimal minimal important difference. *Bmj*, *381*, e073822. https://doi.org/10.1136/bmj-2022-073822