Outlier Detection Methods for Meta-analyses of Site-specific Effect Estimates from a Multi-site Network

by

Henry Ratul Halder

A Thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba

In partial fulfilment of the requirements of the degree of

MASTER OF SCIENCE

Department of Community Health Sciences

University of Manitoba

Winnipeg

Date: 21 August 2023

Abstract

Introduction: Data privacy legislation in Canada prohibits patient-level administrative health data from crossing jurisdictional boundaries. Accordingly, multi-site research networks often conduct distributed analyses and pool site-specific effect estimates (EEs) using meta-analysis models. Rare outcomes and heterogeneity in site-specific EEs can produce potential outliers that may bias pooled EEs. Limited research has compared outlier detection methods and the impact of potential outliers on meta-analysis results.

Purpose and Objectives: The research purpose was to examine outlier detection methods for meta-analyses of site-specific EEs from a multi-site network. The objectives were to: 1) compare outlier detection methods for random-effects meta-analysis (REM) models, and 2) apply these methods to site-specific EEs from systematically selected real-world meta-analyses.

Methods: We compared studentized residual estimates (StdR), relative change in pooled EE variance (RCPEV), relative change in estimated between-site variance (RCEBV), and model-based mean-shift method (MMS) using computer simulation. EEs were simulated assuming a normal distribution. Accuracy, misclassification error (ME), and F-1 score were assessed using random-effects analysis of variance models. We systematically selected meta-analyses conducted by investigators from the Canadian Network for Observational Drug Effect Studies (CNODES), applied outlier detection methods, and assessed the impact of potential outliers on REM results.

Results: StdR had the highest accuracy (median: 89.9%) and lowest ME (median: 10.2%). RCPEV was the most consistent in all metrics. For StdR, the number of sites explained 95.1% and 93.0% of the variation in accuracy and ME values. For RCEBV and MMS, between-site variance described the most variation in accuracy and ME values. StdR and

RCPEV were most sensitive to detect potential outliers in re-analyses of 39 published CNODES meta-analyses. Heterogeneity in site-specific EEs was reduced to zero in two-thirds of the meta-analyses when potential outliers were removed, and the precision of pooled EEs increased.

Conclusions: StdR and RCPEV outperformed RCEBV and MMS in outlier detection. The number of sites and between-site variance explained the most variation in performance metrics for all methods. Excluding potential outliers from published meta-analyses, substantially reduced heterogeneity in site-specific EEs and increased the precision of pooled EEs.

Acknowledgements

I would like to thank and acknowledge my advisor, Dr. Lix, for the knowledge, support, and feedback during this thesis research, throughout my graduate program, and for the opportunity to work in the field of pharmacoepidemiology. I would also like to thank my advisory committee members for their key perspectives in this research, in particular, Dr. Rabbani, Dr. Gerstein, and Dr. Schneider-Lindner for their expertise in meta-analysis models, microbiology and statistical methods, and pharmacoepidemiology, respectively. This research was financially supported through the Canadian Institutes of Health Research (CIHR) funding to Dr. Lix, the Visual and Automated Disease Analytics (VADA) Graduate Training Program, and the Graduate Fellowship provided by Dr. Lix. I would also like to thank and acknowledge Dr. Sanusi and Dr. Ayilara for their support and guidance in computer programming and parallel computing.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Abbreviation	Definition
AHD	Administrative health data
EE	Effect estimate
REM	Random-effects meta-analysis
CNODES	Canadian Network for Observational Drug Effect Studies
FEM	Fixed-effects meta-analysis
CI	Confidence interval
StdR	Studentized residual
RCPEV	Relative change in the pooled effect estimate variance
RCEBV	Relative change in the estimated between-site variance
RVSOM	Random-effects variance shift outlier model
FS	Forward search
MMS	Model-based mean-shift
DL	DerSimonian and Laird
LRT	Likelihood-ratio test
ML	Maximum likelihood
ME	Misclassification error
TP	True positive
FP	False positive
FN	False negative
TN	True negative
Interquartile range	IQR
RE-ANOVA	Random-effects analysis of variance
Q-Q	Quantile-quantile
LCI	Lower confidence interval
UCI	Upper confidence interval
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Chapter 1: Introduction

1.1 Background

Administrative health data (AHD) are routinely collected for health system management and monitoring purposes. In Canada, data-sharing restrictions associated with provincial and territorial health privacy legislation do not typically allow patient-level AHD to cross jurisdictional boundaries. Accordingly, distributed analyses are often adopted for multi-site studies; data are analyzed within each site to produce site-specific effect estimates (EEs). These site-specific EEs are then pooled to produce a single overall estimate. Variability in site-specific EEs may be observed. The random-effects meta-analysis (REM) model is usually chosen to pool site-specific EEs because site-specific EEs are assumed to be randomly sampled from their population distribution with an aim to find the mean estimate of that population distribution, known as the pooled EEs [1,2]. The Canadian Network for Observational Drug Effect Studies (CNODES) is an example of a multi-site network that conducts population-based drug safety and effectiveness studies using distributed analyses of AHD and pool site-specific EEs with the REM model [3,4].

If some site-specific EEs do not appear to represent their population distribution in a multi-site network, they can be considered potential outliers. One challenge when performing a meta-analysis is the potential presence of outliers which can influence parameter estimates and lead to biased pooled EEs [5,6]. If some site-specific EEs appear as potential outliers, detecting potential outliers and understanding the subject area in-depth can reveal the reasons for their presence. The Cochrane Collaboration has developed a risk of bias tool, enabling researchers to detect potential outliers based on careful investigation of differences in treatment implementation and patient populations [2]. However, each site in a multi-site network follows a study protocol to mitigate bias arising from inconsistency in site-specific data analyses and cohort construction [3,4,7–9].

The key issues in meta-analyses of site-specific EEs from distributed analyses of a multi-site network are rare outcomes and heterogeneity in site-specific EEs [10]. When an outcome is rare, many site-specific EEs are calculated from small event data. Consequently, calculated standard errors of site-specific EEs are often large and lead to a loss in the precision of pooled EEs [11]. Furthermore, there is the possibility for heterogeneity in site-specific EEs because of differences in population size across sites and data variability due to differences in the source, construction, and contents of site-specific AHD [8]. These factors can result in the potential presence of outliers. A subgroup analysis can be conducted to account for heterogeneity in site-specific EEs [2]. However, subgroup analysis has low statistical power when the number of site-specific EEs is small or when an outcome is rare [12].

The Cochrane Handbook provides guidelines for conducting systematic reviews and meta-analyses and recommends performing sensitivity analyses, in which meta-analyses are conducted with and without site-specific EEs detected as potential outliers [2]. Multi-site networks, including CNODES, follow this recommendation [3]. However, if there are changes in the direction of pooled EEs between REM models with and without detected outliers, researchers may not reach a consensus about which result (i.e., the pooled EEs with or without detected outliers) can be used for decision-making [5]. Site-specific EEs that are potential outliers can be detected graphically with a forest plot [6], but a graphical representation is sensitive to subjective bias (i.e., a potential outlier site-specific EE to one may not be apparent to another) [1,6].

Previous research has proposed several rule-, model-, and algorithm-based methods to detect potential outlier site-specific EEs, in which some methods are proposed for the REM model [13–18]. Due to the availability of rule- and model-based outlier detection methods through “metafor” [19], “altmeta” [20], and “meta” [21] packages in R, researchers widely

apply these methods in real-world meta-analyses. However, limited research has compared the performance of rule- and model-based outlier detection methods under a wide range of study design and analytic conditions [17,22]. In real-world settings, the impact of detected outliers on the REM model parameters has not been previously investigated.

1.2 Purpose and objectives

The research purpose was to examine outlier detection methods for meta-analyses of site-specific EEs from distributed analyses of a multi-site network. The research objectives were to:

1) Compare the performance of outlier detection methods for the REM model, and
2) Apply outlier detection methods to site-specific EEs from real-world distributed analyses of a multi-site network.

1.3 Thesis organization

The thesis is organized as follows: Chapter 2 provides a review of literature; it includes a framework for meta-analysis of site-specific EEs from distributed analyses of a multi-site network, a brief overview of rule-, model- and algorithm-based outlier detection methods in meta-analysis models, and a discussion of gaps in the relevant literature. Chapter 3 describes the methods used to conduct a simulation study to compare the performance of outlier detection methods and is followed by a description of the methods to conduct analyses of real-world data from distributed analyses of a multi-site network using outlier detection methods. Chapter 4 presents the results of the simulation study and findings from the real-world meta-analyses. The final chapter of the thesis discusses the key findings and their interpretations, research strengths and limitations, opportunities for future research, and conclusions.

Chapter 2: Literature Review

In this chapter, we define the framework for meta-analysis of site-specific EEs from distributed analyses of a multi-site network and provide a brief overview of outlier detection methods in meta-analysis models. Application to outlier detection methods in real-world meta-analyses and published simulation studies that evaluated the performance of outlier detection methods in meta-analysis models are also described. The chapter concludes with a summary of the literature review that identifies gaps in existing research about outlier detection methods and the novelty of the proposed research.

2.1 Framework for meta-analysis of site-specific effect estimates from distributed analyses of a multi-site network

Both fixed-effects meta-analysis (FEM) and REM models can be used to pool site-specific EEs from distributed analyses in multi-site networks. Our research framework is based on the REM model for several reasons. Heterogeneity in site-specific EEs may be observed when site-specific EEs do not appear to represent their population distribution. Differences in the source, construction, and contents of site-specific AHD may be another reason for heterogeneity in site-specific EEs [10]. Also, variability in the precision of site-specific EEs may occur because of differences in site-specific sample sizes [10]. A subgroup analysis can be conducted to account for heterogeneity in site-specific EEs [7]. However, a subgroup analysis has low statistical power, and it may falsely detect differences in subgroups when the number of site-specific EEs is small or when an outcome is rare [12].

The REM model has the advantage of capturing both within-site variance (i.e., heterogeneity in site-specific EEs due to sampling error) and between-site variance (i.e., heterogeneity in site-specific EEs not attributable to sampling error) [1]. Sources of variation may be unknown or not measured directly [1]. Some portions of unknown or unmeasurable

variation in site-specific EEs from distributed analyses of a multi-site network can also be explained by between-site variance with the REM model [1].

In the FEM model, site-specific EEs are assumed to be sampled from the same population, indicating a common true EE [1]. However, if all site-specific EEs have a common true EE, then the REM model can reduce to the FEM model [1]. The pooled EEs from a REM model can be generalized to a larger population because site-specific EEs are assumed to be randomly sampled from their population distribution [1]. Therefore, the REM model is often chosen to calculate pooled EEs along with their 95% confidence intervals (CIs).

2.2 Outlier detection methods in meta-analysis models

We identified several rule-, model-, and algorithm-based outlier detection methods for meta-analysis models. Outliers in site-specific EEs are detected using rules of thumb or decision criteria in rule-based methods. Model-based methods rely on assumptions about the distribution of model parameters; outlier site-specific EEs in meta-analysis models are detected using hypothesis testing or parametric bootstrapping. Algorithm-based methods detect outlier site-specific EEs without relying on decision criteria and model assumptions.

2.2.1 Rule-based methods

Hedges and Olkin proposed rule-based outlier detection methods for the FEM model [6]. The authors suggested to visualize and inspect site-specific EEs with a forest plot. If a site-specific EE is visually different in a forest plot (such as non-overlapping confidence intervals of the pooled EEs, site-specific EEs are not aligning close to the pooled EEs) and the remaining site-specific EEs are homogeneous after excluding the visually different site-specific EEs, then the visually different site-specific EEs may be detected as potential outliers. However, this method is sensitive to subjective bias [1,6].

The studentized residual estimates (StdR) method by Hedges and Olkin can also be used to detect potential outliers in FEM models [6]. The distance between site-specific EEs and the pooled EE is calculated. Then, estimated residuals are standardized with the standard error of the distance between site-specific EEs and the pooled EE. The pooled EE in a meta-analysis is calculated from site-specific EEs. To mitigate the dependency between site-specific EEs and the pooled EE in a meta-analysis, it is recommended to exclude one site-specific EE at a time and estimate StdR after each exclusion [6]. As a rule of thumb, a site-specific EE with an estimated StdR larger than three standard deviations is a potential outlier [6]. However, this method is applicable to the FEM model and can detect potential outliers that appeared only due to sampling error [13].

Another method proposed by Hedges and Olkin to detect potential outliers relies on estimating the Q-statistic for both FEM and REM models [6]. The Q-statistic measures the degree of variation or heterogeneity in site-specific EEs, by calculating the weighted sum of squared differences between site-specific EEs and the pooled EE, with the weights estimated using the inverse-variance method [23]. In this rule-based outlier detection method, the Q-statistic is estimated using all site-specific EEs. Then, each site-specific EE is excluded, and the Q-statistic is estimated after each exclusion. The relative change in the estimated Q-statistic is captured by comparing the Q-statistic estimated from all site-specific EEs with the Q-statistics estimated after exclusion of each site-specific EE. An estimated Q-statistic will change substantially when a potential outlier is excluded from a FEM model, indicating homogeneity in site-specific EEs. However, the Q-statistic has low statistical power when the number of site-specific EEs included in meta-analysis models is small, and it may falsely detect heterogeneity in site-specific EEs when many site-specific EEs are pooled [24,25].

Viechtbauer and Cheung proposed to calculate StdR by excluding one site-specific EE at a time to detect potential outliers for REM models. The StdR method estimates the distance between site-specific EEs and the pooled EEs and standardize estimated residuals with the standard error of the distance between site-specific EEs and the pooled EEs [13]. As a rule of thumb, a site-specific EE is detected as a potential outlier if it has a StdR estimate > 1.96 [13]. However, at least one site-specific EE is always detected as a potential outlier by chance in this method due to the exclusion of one site-specific EE at a time, resulting in a high false discovery rate or family-wise type-1 error [13].

Two additional methods to detect potential outliers in REM models are: 1) relative change in the pooled EE variance (RCPEV) and 2) relative change in the estimated between-site variance (RCEBV) [13]. Ratios of the pooled EE variance that excludes one site-specific EE at a time with all site-specific EEs are compared in the RCPEV method. Comparisons among ratios of estimated between-site variance that exclude one site-specific EEs at a time with all site-specific EEs are estimated in the RCEBV method. As a rule of thumb, site-specific EEs with RCPEV and RCEBV < 1 are detected as potential outliers [13]. The RCPEV method cannot be used for the FEM model because RCPEV estimates reflect the information sizes (i.e., proportional sample sizes) [22]. The RCEBV method cannot be applied to the FEM model and in the absence of substantial heterogeneity between site-specific EEs [13,22].

2.2.2 Model-based methods

A model-based method, the random-effects variance shift outlier model (RVSOM), has been proposed to detect potential outliers [14]. Each site-specific EE is tested to determine if it is a potential outlier in the RVSOM, where a potential outlier is assumed to have a distribution with a higher within-site variance than the remaining site-specific EEs. Parametric bootstrapping is used to find the empirical sampling distribution of the test

statistic. Multiple outliers can be detected using second, third, and possibly a higher-order statistic. However, the RVSOM is not recommended when multiple potential outliers are present in a meta-analysis [14].

Beath proposed a finite mixture method to detect potential outliers for the REM model in which a site-specific EE is assumed to belong to one of the two classes [15]. The classes in a finite mixture method are the same, except the variance of site-specific EEs is higher in the class that contains potential outliers. The empirical sampling distribution of the test statistic is derived using parametric bootstrapping. Results are then visualized using the posterior probability of membership in each class (i.e., whether or not a site-specific EE is a potential outlier). There is strong evidence of a site-specific EE detected as a potential outlier if the posterior probability is > 0.9. However, a large sample size is required to distinguish the two classes in a finite mixture method [26].

A model-based mean-shift (MMS) method for meta-analysis was proposed by Negeri and Beyene [17] and also by Noma et al. [18]. The MMS method assumes that the random-effects distribution for a potential outlier has a shifted mean compared to the remaining site-specific EEs. Parametric bootstrapping is used to find the empirical sampling distribution of the test statistic [14,17,18,22]. One of the key strengths of the MMS method is that one can use it in a meta-analysis of site-specific EEs from distributed analyses of a multi-site network [22], diagnostic test accuracy meta-analysis [17], network meta-analysis in frequentist [18] and Bayesian [27] frameworks. The MMS method for meta-analysis is easy to understand, simple to apply, and familiar to researchers because it is extended from an outlier detection method in linear regression [28]. However, when the number of site-specific EEs < 10, the resampled site-specific EEs in the parametric bootstrapping method do not represent their population distribution [29,30].

2.2.3 Algorithm-based methods

An algorithm-based outlier detection method is described by Mavridis and colleagues [16]. They used a forward search (FS) algorithm to detect a potential outlier in meta-analysis models. In the FS algorithm, an initial clean (i.e., likely outlier-free) subset of site-specific EEs is chosen from all site-specific EEs included in a meta-analysis. With an initial subset of site-specific EEs, the FS algorithm searches for potential outliers and constitutes the "basic set", while the "non-basic set" consists of the remaining site-specific EEs. The FS algorithm then adds site-specific EEs from the non-basic set to the basic set until all site-specific EEs are in the basic set. The FS algorithm sorts site-specific EEs by their closeness to the hypothesized model and monitors various test statistics (such as the pooled EEs, between-site variance, and Cook's distance) throughout the search, allowing one to explore if a change occurred due to a potential outlier or random error. Once a potential outlier has been detected from the FS algorithm, researchers can follow the recommendation of the Cochrane Handbook and exclude the potential outlier from a meta-analysis as a part of sensitivity analysis. However, the FS algorithm can only detect a single potential outlier. Furthermore, the order of parameters (i.e., choice of the initial subset, size of the initial subset, and number of samples for site-specific EEs explored) in the FS algorithm can influence site-specific EEs that are detected as potential outliers [16].

2.3 Application of outlier detection methods in real-world meta-analyses

Several methods were proposed in the literature to detect potential outliers in the REM model, namely rule-based (StdR, RCPEV, and RCEBV) [13], model-based (RVSOM method [14], finite mixture method [15], the MMS method [17,31]), and algorithm-based (the FS algorithm [16]) methods. Rule-based outlier detection methods proposed by Viechtbauer and Cheung [13] and the MMS method [17,31] in the REM model have received notable attention

in real-world meta-analyses because these methods are readily available through “metafor” [19], “altmeta” [20], and “meta” [21] packages in R.

The “metafor” and “altmeta” packages include the StdR, RCPEV, RCEBV, and MMS methods [19,20], while only the RCPEV method is available in the “meta” package [21]. In the “metafor” package, estimates from the methods are presented in a table and site-specific EEs detected as potential outliers by any of the methods are marked with asterisks. Additionally, the “metafor” package allows for plotting the index of site-specific EEs on the horizontal axis and the estimates of each method on the vertical axis, enabling visual detection of potential outliers [19].

2.4 Simulation studies about outlier detection methods in meta-analysis models

A small number of simulation studies have compared the performance of outlier detection methods in meta-analysis models [17,22]. A simulation study evaluated the performance of 1) StdR, 2) RCPEV, 3) RCEBV, and 4) MMS methods in the context of meta-analyses for a multi-site clinical trial [22]. Parametric bootstrapping was used in all four methods to find the empirical sampling distributions of the test statistics. As between-site variance increased, the performance of all methods to detect outliers decreased [22]. When the difference between outlier site-specific EEs and the remaining site-specific EEs was small, the performance of all methods decreased [22]. In the case of multiple outliers and small heterogeneity in site-specific EEs, the StdR and MMS methods detected at least one and all three outliers correctly 90.2% and 93.6% of the time, respectively [22]. However, the probability of detecting multiple outliers decreased for the StdR and MMS methods when between-site variance and the difference between outlier site-specific EEs and the remaining site-specific EEs increased [22].

Another simulation study applied 1) StdR, 2) RCPEV, 3) RCEBV, and 4) MMS methods in the context of meta-analyses for diagnostic test accuracy [17]. Parametric bootstrapping was applied to all methods to find the empirical sampling distributions of the test statistics. The simulation results indicated that the performance of StdR and MMS methods improved when the number of site-specific EEs increased and outlier site-specific EEs had substantially larger means than the means for site-specific EEs that were not detected as outliers [22]. On average, more than 80% of the time, the RCPEV and RCEBV methods detected at least two of the three outliers correctly [22]. This study showed that the RCPEV and RCEBV methods could detect multiple outliers more accurately than the StdR and MMS methods [22].

2.5 Summary

In summary, previous research has proposed rule-, model-, and algorithm-based outlier detection methods for meta-analysis models, in which some methods are proposed specifically for the REM model. Only rule and model-based methods are widely accepted among researchers because these methods can be easily applied to detect potential outliers in real-world meta-analyses through statistical packages in R [19–21].

The StdR, RCPEV, RCEBV, and MMS methods were assessed in two simulation studies [17,22]. In the case of multiple outliers, one simulation study concluded that the StdR and MMS methods performed better than the RCPEV and RCEBV methods [22]. However, the second simulation study results concluded the opposite [17]. The conflict in the results of simulation studies occurred likely due to variations in simulation parameters and the choice of simulation conditions [17,22]. Both of the simulation studies were unable to capture variations in the performance of methods due to a narrow exploration of simulation conditions [17,22]. Besides, rule-based methods (StdR, RCPEV, and RCEBV) were used differently in both simulation studies [17,22]. Comparisons among the StdR, RCPEV, and

RCEBV methods were drawn using parametric bootstrapping to find the empirical sampling distributions of the test statistics [17,22]. However, when the number of site-specific EEs < 10, the resampled site-specific EEs in the parametric bootstrapping method may not be representative of their population distribution [29,30].

Chapter 3: Methods and Materials

In this chapter, we describe the REM model and propose methods to detect potential outliers. To achieve objective 1, a simulation study was conducted to compare the performance of outlier detection methods in the REM model. Our focus was on widely used rule-based outlier detection methods (StdR, RCPEV, RCEBV) and the MMS method in the REM model. To address objective 2, we systematically selected meta-analyses from peer-reviewed publications of distributed analyses from a multi-site network. We then applied outlier detection methods to site-specific EEs and estimated the REM model parameters with and without potential outliers.

We sought an ethics application waiver because data were collected from peer-reviewed publications. This waiver was approved by the Department of Community Health Sciences, Max Rady College of Medicine, University of Manitoba.

3.1 Random-effects meta-analysis model

Assume y_i is the site-specific EE for the i^{th} site and randomly sampled from the population distribution of site-specific EEs, where $i = 1,2, \dots, k$ and k indicates the number of site-specific EEs. The REM model is,

$$y_i = \theta_i + e_i;\ e_i \sim N(0, v_i), \tag{1}$$

where,

$$\theta_i = \mu + u_i;\ u_i \sim N(0, \tau^2). \tag{2}$$

In equation 1, each y_i is assumed to vary by a random amount, determined by the within-site variance (v_i); θ_i denotes the estimated site-specific EEs weighted by within-site error variance for the i^{th} site and is assumed to differ from the pooled EE (μ) by a random amount, determined by the between-site variance (τ^2). Within-site errors (e_i) and between-site errors (u_i) are assumed to be independent.

Then, $\hat{\mu}$ is estimated using a weighted average of y_i, where each site-specific weight is,

$$w_i = \frac{1}{v_i+\tau^2}, \tag{3}$$

such that,

$$\hat{\mu} = \frac{\sum_{i=1}^{k} y_i w_i}{\sum_{i=1}^{k} w_i}. \tag{4}$$

The Q-statistic is used to test for homogeneity in θ_i under the null hypothesis, where all θ_is are assumed to be homogeneous. Homogeneity in θ_i can be tested using $Q = \sum w_i(y_i - \hat{\theta})^2$, where $Q \sim \chi^2_{(k-1)}$. Moreover, $\hat{\tau}^2$ is estimated based on the methods of moments framework and is known as the DerSimonian and Laird (DL) method [32]. The percentage of variation due to heterogeneity in site-specific EEs can be estimated from I^2 such that $I^2 = \frac{(Q-df)}{Q} \times 100\%$ where $df = k - 1$ [33,34]. Note that when $\hat{\tau}^2 = 0$, the FEM and REM models are equivalent.

3.2 Outlier detection methods in random-effects meta-analysis model

3.2.1 Studentized residual estimates

We begin by describing an outlier detection method similar to the dfbeta statistic used in conventional regression analyses [35,36]. Outlier detection methods based on StdR estimates generally measure the distance between observed and predicted values and standardize estimated residuals with the standard error of the distance between observed and predicted values [35,36]. In a meta-analysis, $\hat{\mu}$ is estimated from y_i and $\hat{\mu}$ is dependent on y_i [1]. A conventional StdR estimate is not suitable to detect site-specific EEs that are potential outliers due to the dependency of $\hat{\mu}$ on y_i [13]. This limitation of a conventional StdR estimate can be addressed by excluding one site-specific EE at a time while estimating StdR [6,13].

Let $\hat{\mu}^{(-i)}$ and $\hat{\tau}^{2(-i)}$ be estimated using the DL method based on site-specific EEs of $k-1$ sites. Then, StdR estimates are defined as,

$$t_i = \frac{y_i - \hat{\mu}^{(-i)}}{\sqrt{Var[y_i - \hat{\mu}^{(-i)}]}}, \tag{5}$$

where,

$$Var\left[y_i - \hat{\mu}^{(-i)}\right] = (\widehat{w}_i)^{-1} - \left(\sum_{i=1}^{k} \widehat{w}_i^{(-i)}\right)^{-1}, \tag{6}$$

and,

$$\widehat{w}_i^{(-i)} = \left(\hat{\tau}^{2(-i)} + \hat{\sigma}_i^2\right)^{-1}. \tag{7}$$

We can interpret t_i as a StdR estimate for the i^{th} site estimated from the REM model by the other $k-1$ sites. Under the REM model, t_i is assumed to follow the standard normal distribution [13]. If the REM model assumptions are violated, the parametric bootstrapping method can be used to derive the empirical sampling distribution of t_i [22]. However, when the number of site-specific EEs < 10, the resampled site-specific EEs in the parametric bootstrapping method do not represent their population distribution [29,30]. Therefore, as a rule of thumb, we compare the absolute value of t_i with the threshold of two standard deviations from the mean of standard normal distribution. A site-specific EE from distributed analyses of a multi-site network is detected as a potential outlier if $|t_i| > 1.96$ [13].

3.2.2 Relative change in the pooled effect estimate variance

The RCPEV method estimates ratios of pooled EEs for a dataset that excludes each site-specific EE one at a time and a dataset containing all site-specific EEs. Viechtbauer and Cheung proposed this method for the REM model [13].

$$COVRATIO_i = \frac{Var\left[\hat{\mu}^{(-i)}\right]}{Var[\hat{\mu}]} = \frac{\sum_{i=1}^{k} \widehat{w}_i^{(-i)}}{\sum_{i=1}^{k} \widehat{w}_i}. \tag{8}$$

This method examines the impact of a site-specific EE on the precision of $\hat{\mu}$, where $COVRATIO_i$ ranges between 0 and ∞. The sampling distribution of $COVRATIO_i$ can be derived using the parametric bootstrapping method. However, when the number of site-specific EEs < 10, the parametric bootstrapping method for deriving the sampling distribution of $COVRATIO_i$ is not reliable because the resampled site-specific EEs do not represent their population distribution [29,30]. As a rule of thumb, $COVRATIO_i < 1$ indicates that the exclusion of a site-specific EE increases the precision of $\hat{\mu}$ and the excluded site-specific EE is detected as a potential outlier. When $COVRATIO_i > 1$, the excluded site-specific EE is not a potential outlier [13].

3.2.3 Relative change in the estimated between-site variance

Viechtbauer and Cheung proposed using the ratio of $\hat{\tau}^2$ for a dataset that excludes each site-specific EE one at a time and a dataset that contains all site-specific EEs [13]. This method is effective for outlier detection in the presence of substantial $\hat{\tau}^2$:

$$R_i = \frac{\hat{\tau}^{2(-i)}}{\hat{\tau}^2}. \tag{9}$$

R_i ranges between 0 and ∞. The sampling distribution of R_i can be derived using the parametric bootstrapping method. When the number of site-specific EEs < 10, the parametric bootstrapping method to derive the sampling distribution of R_i is not reliable because the resampled site-specific EEs do not represent their population distribution [29,30]. As a rule of thumb, $R_i < 1$ indicates that the exclusion of a site-specific EE decreases between-site variance, and the excluded site-specific EE is detected as a potential outlier. For $R_i > 1$, the excluded site-specific EE is not a potential outlier [13].

3.2.4 Model-based mean-shift method

Negeri and Beyene [17] and Noma et al. [18] proposed the MMS method, which assumes that 1) the random-effects distribution of a site-specific EE for the corresponding j^{th} site is

shifted as $\theta_j \sim N(\mu + \delta, \tau^2)$, where j indicates a potential outlier and 2) site-specific EEs of $k-1$ sites follow the same distribution as a REM model, that is, $\theta_i \sim N(\mu, \tau^2)$. Then, we can test the hypothesis,

$$H_0: \delta = 0 \text{ vs. } H_1: \delta \neq 0,$$

using a likelihood ratio test (LRT). The log-likelihood function under H_0 corresponds to the REM model and is written as,

$$l_0(\mu, \tau^2) = -\frac{1}{2}\sum_{i=1}^{k}\left\{\log 2\pi\,(v_i + \tau^2) + \frac{(y_i - \mu)^2}{v_i + \tau^2}\right\}. \quad (10)$$

The log-likelihood function under the alternative hypothesis is,

$$l_{1[j]}(\mu, \tau^2, \delta) = -\frac{1}{2}\sum_{i=1}^{k}\left\{\log 2\pi\left(v_j + \tau^2\right) + \frac{\left(y_j - \mu - \delta\right)^2}{v_j + \tau^2}\right\}$$
$$-\frac{1}{2}\sum_{i\neq j}\left\{log2\pi(v_i + \tau^2) + \frac{(y_i - \mu)^2}{v_i + \tau^2}\right\} \quad (11)$$

Then, the LRT statistic is given by

$$LRT_{[j]} = -2\left\{l_0(\hat{\mu}, \hat{\tau}^2) - l_{1[j]}\left(\hat{\mu}_{[j]}, \hat{\tau}^2_{[j]}, \delta_{[j]}\right)\right\}, \quad (12)$$

where $\{\hat{\mu}, \hat{\tau}^2\}$ is the maximum likelihood (ML) estimate of H_0 and $\left\{\hat{\mu}_{[j]}, \hat{\tau}^2_{[j]}, \delta_{[j]}\right\}$ is the ML estimate of the MMS method for j^{th} site-specific EEs. The $LRT_{[j]}$ is assumed to follow a χ^2 distribution with 1 degree of freedom under H_0. However, asymptotic theory, which relates the distribution of $LRT_{[j]}$ to a χ^2 distribution under H_0, does not apply here because H_0 falls on the boundary of the parameter space, and regularity conditions are not met [14,17,18]. Accordingly, we used a parametric bootstrap method to find the empirical sampling distribution of $LRT_{[j]}$ [14,17,18,22].

In summary, the MMS method requires four steps. First, we compute the ML estimates for $\{\hat{\mu}, \hat{\tau}^2\}$ and $\{\hat{\mu}_{[j]}, \hat{\tau}^2_{[j]}, \delta_{[j]}\}$. Then, we resample $y_1^{(b)}, y_2^{(b)}, \ldots, y_k^{(b)}$ from y_i with replacement and apply parametric bootstrap method *B* times, where $b = 1,2, \ldots, B$. We then compute the ML estimates $\left\{\hat{\mu}^{(b)}, \hat{\tau}^{2(b)}\right\}$ and $\left\{\hat{\mu}^{(b)}_{[j]}, \hat{\tau}^{2(b)}_{[j]}, \hat{\delta}^{(b)}_{[j]}\right\}$ for the b^{th} bootstrap sample of $y_1^{(b)}, y_2^{(b)}, \ldots, y_k^{(b)}$ with $\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_k$ and calculate the LRT statistic,

$$LRT^{(b)}_{[j]} = -2\left\{l_0\left(\hat{\mu}^{(b)}, \hat{\tau}^{2(b)}\right) - l_{1[j]}\left(\hat{\mu}^{(b)}_{[j]}, \hat{\tau}^{2(b)}_{[j]}, \delta^{(b)}_{[j]}\right)\right\}. \quad (13)$$

This test statistic is computed for all *B* bootstrap samples. Finally, we obtain the bootstrap estimate of $LRT_{[j]}$ and its *p*-value from the empirical sampling distribution of $LRT^{(1)}_{[j]}, LRT^{(2)}_{[j]}, \ldots, LRT^{(B)}_{[j]}$.

3.3 Simulation design and method

3.3.1 Data generation

For objective 1, computer simulation was used to evaluate the performance of the proposed outlier detection methods. We manipulated two simulation design characteristics: 1) number of site-specific EEs (*k*) and 2) number of outliers (*m*). Meta-analyses in previously published simulation studies manipulated *k* ranging from 4 to 30 and *m* between 1 and 2 [17,22]. Besides, a meta-analysis for a multi-site network typically consists of a small number of site-specific EEs and a small number of potential outliers [15]. Therefore, our simulation design characteristics consisted of *k* = 10, 20, 30 and *m* = 1, 2.

The investigated simulation parameters were considered from previously published simulation studies, including 1) within-site variance (*v*), 2) mean EE for *k*-*m* sites (μ_0), 3) mean EE for *m* sites (μ_1), and 4) between-site variance (τ^2) [17,22]. Site-specific EEs were randomly generated from the mean of site-specific EEs, assuming a normal distribution, where the mean of site-specific EEs differed between μ_1 and μ_0 [22]. Briefly, μ_0 = 0.50, 1.00,

and 1.50 represent decreased, no, and increased risk or odds for an outcome of interest relative to a reference category. In previously published simulation studies, values of μ_1 were selected based on real-world meta-analyses [17,22]. We calculated the values of μ_1 based on the quartiles of μ_0, representing small and large differences between outlier site-specific EEs and the remaining site-specific EEs. Large differences between outlier site-specific EEs and the remaining site-specific EEs were defined as site-specific EEs smaller than the first quartile of μ_0 – 3×(inter-quartile range of μ_0) or greater than the third quartile of μ_0 + 3×(inter-quartile range of μ_0). Small differences between outlier site-specific EEs and the remaining site-specific EEs were defined as site-specific EEs smaller than the first quartile of μ_0 – 1.5×(inter-quartile range of μ_0) or greater than the third quartile of μ_0 + 1.5×(inter-quartile range of μ_0) [37,38]. Accordingly, the calculated μ_1 from corresponding μ_0 were used as parameter values (see **Table 1**). Then, we randomly generated site-specific EEs for *k-m* sites from $N(\mu_0, \tau^2)$ and for *m* sites from $N(\mu_1, \tau^2)$.

Various values of *v* can be used as a proxy for the number of observations at each site to capture heterogeneity within sites. That is, large sample sizes within a site produce smaller *v*, which influences parameter estimation in a meta-analysis [39]. Therefore, we evaluated *v* at 0.05, 0.50, and 1.00. When $\tau^2 > 0.05$, the REM model is likely to be applied in practice [40,41]. We considered $\tau^2 = 0.05$, 0.50, and 1.00 while randomly generating site-specific EEs.

Table 1. Simulation conditions for various combinations of simulation design characteristics with simulation parameters.

Simulation design characteristic		**Simulation parameter**			
Number of sites (k)	Number of outliers (m)	Within-site variance (v)	Mean EE for k-m sites (μ_0)	Mean EE for m sites (μ_1)	Between-site variance (τ^2)
10, 20, 30	1	0.05, 0.50, 1.00	0.50	0.35	0.05
					0.50
					1.00
				0.15	0.05
					0.50
					1.00
			1.00	0.85	0.05
					0.50
					1.00
				0.65	0.05
					0.50
					1.00
			1.50	1.35	0.05
					0.50
					1.00
				1.15	0.05
					0.50
					1.00
10, 20, 30	2	0.05, 0.50, 1.00	0.50	0.35	0.05
					0.50
					1.00
				0.15	0.05
					0.50
					1.00
			1.00	0.85	0.05
					0.50
					1.00
				0.65	0.05
					0.50
					1.00
			1.50	1.35	0.05
					0.50
					1.00
				1.15	0.05
					0.50
					1.00

In summary, we assessed the proposed outlier detection methods under two main scenarios. In the presence of a single site-specific EE that was an outlier, the performance of outlier detection methods was evaluated for various simulation conditions in scenario 1. In scenario 2, the performance of outlier detection methods was assessed for various simulation conditions when two site-specific EEs were outliers. For each combination of *k* and *m*, we fixed v = 0.05, 0.50, and 1.00. For each combination of *k*, *m*, and v, we considered μ_0 = 0.50, 1.00, and 1.50. For every combination of *k*, *m*, v, and μ_0, we used two values of μ_1 representing small and large differences between outlier site-specific EEs and the remaining site-specific EEs based on μ_0 values. Finally, for each combination of *k*, *m*, v, μ_0, and μ_1, we considered τ^2 = 0.05, 0.50, and 1.00. In total, we investigated 54 sets of simulation conditions for each *k* and *m* (**Table 1**).

3.3.2 Validity of simulation results

For objective 1, site-specific EEs without outliers were randomly generated to check the validity of simulation results. Values of *k* and simulation parameters (v, μ_0, and τ^2) were the same as **Table 1** while generating site-specific EEs, when *m*= 0. For each combination of *k* and *m*, we had 27 simulation conditions in randomly generated site-specific EEs (**Appendix A**). Based on previous literature, a total of *n* = 1000 simulation iterations were performed for each set of simulation conditions [22].

StdR estimates for each simulation condition were visualized using density plots for validation purposes because it is known that StdR estimates under the REM model are assumed to follow a standard normal distribution [13]. If the majority of StdR estimates for each simulated condition were concentrated at 0, this suggests that StdR estimates have a mean of 0, and according to that the simulation is valid.

3.3.3 Performance metrics

In our simulation study, observed classes of randomly generated site-specific EEs were known. The expected classes of randomly generated site-specific EEs were then predicted or classified using the proposed outlier detection methods. For simulation iteration, a confusion matrix was constructed, and the performance of each method was evaluated using median accuracy and median misclassification error (ME). Due to the class imbalance between site-specific EEs that are outliers and the remaining site-specific EEs, we additionally estimated the median F-1 score for each method [42].

Consider the following confusion matrix,

	Observed class	
Predicted class	Positive	Negative
Positive	True positive (TP)	False positive (FP)
Negative	False negative (FN)	True negative (TN)

Here, true positive (TP) is the number of site-specific EEs that are observed and predicted as outliers by each of the investigated outlier detection methods. False positive (FP) is the number of site-specific EEs that are incorrectly predicted to be outliers but were observed as non-outliers. False negative (FN) is the number of site-specific EEs that are observed as outliers but predicted as not being outliers. The number of site-specific EEs that are correctly predicted as not being outliers is known as true negative (TN). The median values of the performance metrics were estimated as follows:

$$\text{Median accuracy} = \left(\frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}\right)/1000, \tag{14}$$

$$\text{Median ME} = (1 - \text{Mean accuracy})/1000, \text{ and} \tag{15}$$

$$\text{Median F-1 score} = \left(\frac{2\text{TP}}{2\text{TP}+\text{FP}+\text{FN}}\right)/1000. \tag{16}$$

The higher the values of accuracy and F-1 score, the better a method is performing in detecting outlier site-specific EEs. Low values of ME indicate good performance for a method because low values reflect the ability of an outlier detection method to minimize incorrect predictions. We estimated medians of performance metrics because median estimates are consistent for all distributions of performance metrics whether or not they follow symmetric distributions.

3.3.4 Analyses of performance metrics

Median accuracy, F-1 score, and ME values with their interquartile ranges (IQR) for each outlier detection method were analyzed descriptively and visualized with boxplots to achieve objective 1.

For each method, random-effects analysis of variance (RE-ANOVA) models were used to assess the impact of treatment (i.e., simulation design characteristics and parameters) on the variation of performance metrics. RE-ANOVA models were appropriate for our case as treatments were randomly sampled from a large population of treatment values, where random-effects residuals were assumed to follow a standard normal distribution [43]. Simulation design characteristics and parameters as main effects along with two-way their interactions were considered as treatments in RE-ANOVA models. We reported eta-squared (η^2) and partial eta-squared $\left(\eta_p^2\right)$ as measures of effect size [44].

The coefficient η^2 is defined as:

$$\eta_p^2 = \left(\frac{SS_{treatment}}{SS_{total}}\right) \times 100. \tag{17}$$

The coefficient η_p^2 is defined as:

$$\eta_p^2 = \left(\frac{SS_{treatment}}{SS_{treatment} + SS_{residual}}\right) \times 100. \quad (18)$$

where $SS_{treatment}$ denotes the sum of squares of a treatment (i.e., a simulation design characteristic or parameter), SS_{total} denotes the total sum of squares, and $SS_{residual}$ denotes the sum of squared residuals. We interpret η^2 estimates as the percentage of variance explained by a treatment as main effect or interaction between two treatments. In contrast, η_p^2 estimates can be interpreted as the percentage of variance explained by a treatment as a main effect or interactions between two treatments, after adjusting for other treatments included in a RE-ANOVA model. When there is only one treatment included in a RE-ANOVA model, η^2 and η_p^2 estimates are equivalent.

The assumption of a standard normal distribution for random-effects residuals was validated by quantile-quantile (Q-Q) plots for each method by their performance metrics. In Q-Q plots, standard normal quantiles were plotted on the horizontal axis against the quantiles of random-effects residuals on the vertical axis for each method by their performance metrics. When the random-effects residuals aligned closely with the reference line of the standard normal distribution, it indicates that the random-effects residuals follow a standard normal distribution.

The assumption of homogeneity of variances in a performance metric was assessed by plotting fitted values obtained from the RE-ANOVA model on the horizontal axis against the estimated random-effects residuals on the vertical axis for each method. If the random-effects residuals were equally spread around the reference line without any distinct patterns, it indicates that the variances in performance metrics are homogeneous in RE-ANOVA models.

3.4 Application to real-world data

3.4.1 Data source

For objective 2, we systematically selected meta-analyses from peer-reviewed publications of site-specific EEs conducted by investigators from CNODES. CNODES is a multi-site network of seven Canadian provinces (Alberta, British Columbia, Manitoba, Nova Scotia, Ontario, Québec, and Saskatchewan) and investigates drug safety and effectiveness at the population level [3]. All peer-reviewed publications (between 2011 and 2022) for meta-analyses of site-specific EEs conducted by investigators from CNODES can be found here: https://www.cnodes.ca/about/publications/.

3.4.2 Selection of real-world meta-analyses

FEM models were initially used by investigators from CNODES to pool site-specific EEs [45,46], as multi-site networks conduct research with an identical site-specific study protocol to minimize heterogeneity in site-specific EEs and assume site-specific EEs are not randomly sampled from a potentially larger set of sites [3]. Since 2016, investigators from CNODES adopted REM models to pool site-specific EEs because REM models can reduce to FEM models when estimated between-site variances are small and provide more unbiased pooled EEs than FEM models in the presence of potential outliers [9]. Therefore, peer-reviewed publications before 2016 were excluded from the application of outlier detection methods.

We selected meta-analyses from peer-reviewed publications of site-specific EEs conducted by investigators from CNODES based on estimated $\hat{\tau}^2$ and the number of site-specific EEs. When $\hat{\tau}^2$ estimates were small, REM models can reduce to FEM models, leading to identical pooled EEs with their 95% CIs [1]. Meta-analyses conducted by investigators from CNODES also corroborate this observation [47,48]. Hence, meta-analyses were excluded from our investigation of potential outliers when the estimates derived from both FEM and REM

models were equivalent. Previous research recommended investigating potential sources of variability in site-specific EEs when the number of site-specific EEs < 4, rather than applying methods to detect potential outliers that could produce high heterogeneity in site-specific EEs [49,50]. We thus included meta-analyses, where the number of site-specific EEs > 4 for the application of outlier detection methods.

In summary, we systematically selected peer-reviewed publications of site-specific EEs conducted by investigators from CNODES. Following that, we applied outlier detection methods to the selected meta-analyses and performed REM models with and without potential outliers to capture the impact of potential outliers on the estimated pooled EEs, $\hat{\tau}^2$, and $\hat{I}^2$.

3.4.3 Required estimates of the random-effects model parameters for selected meta-analyses

Estimates of REM model parameters were not reported in several peer-reviewed publications conducted by investigators from CNODES, including: 1) standard errors of site-specific EEs, 2) $\hat{\tau}^2$ estimates, and 3) site-specific weights. However, site-specific EEs with their 95% lower confidence intervals (LCIs) and upper confidence intervals (UCIs) were reported in every selected meta-analysis. Therefore, standard errors of site-specific EEs were calculated as follows: 1) convert 95% LCIs and UCIs of site-specific EEs on the log scale, and 2) standard error estimates $= \frac{\ln(95\%\ UCIs)-\ln\ (95\%\ LCIs)}{2*1.96}$ [51].

We performed REM models using site-specific EEs with their standard error estimates to calculate $\hat{\tau}^2$. If standard error estimates were not reported, we first estimated standard errors and then fit REM models to estimate $\hat{\tau}^2$. Site-specific weights were estimated from equation (3).

3.4.4 Data analysis for selected meta-analyses

We compared outlier detection methods for selected meta-analyses to achieve objective 2. Questions addressed when comparing outlier detection methods were: 1) did the method detect any outlier site-specific EEs? 2) how many outlier site-specific EEs were detected by the method? and 2) did the method detect the same site-specific EEs as outliers with the other methods?

We gained insight into the methods' sensitivity to detect outliers from question 1. Variation in the number of potential outliers detected by each method was captured in question 2. Question 3 described agreement among methods for the site-specific EEs that were detected as potential outliers. For questions 1 and 3, if the method detected potential outliers and the detected outliers agreed with other methods, we marked the response as "✓"; otherwise it was "✕". Numeric responses were recorded for question 2.

The impact of potential outliers was captured by performing REM models with and without site-specific EEs that were detected as potential outliers. $\hat{\tau}^2$, $\hat{I}^2$, and pooled EEs from REM models with and without potential outliers were estimated. We then examined whether the estimated $\hat{\tau}^2$ and $\hat{I}^2$ were present or absent (i.e. $\hat{\tau}^2$ and $\hat{I}^2$ estimates reduced to 0) once potential outliers were excluded from REM models. Changes in the magnitude (decreased, same, increased), and direction (same, opposite) of pooled EEs were monitored from REM models with and without potential outliers. The change in the precision of pooled EEs was measured by taking the difference between 95% UCIs and LCIs. If the difference in UCIs and LCIs for REM models without potential outliers was narrower than REM models with potential outliers, it indicates an increase in the precision of pooled EEs. On the contrary, a larger difference suggests a decrease in the precision of pooled EEs.

Chapter 4: Results

4.1 Simulation study

For objective 1, a computer simulation was conducted using the following steps: 1) randomly generated data from the prespecified simulation conditions, 2) checked the validity of simulation results, 3) calculated performance metrics for each outlier detection method, and 4) analyzed performance metrics for each method using boxplots and RE-ANOVA models.

4.1.1 Overall performance of outlier detection methods

StdR estimates for each simulation condition were visualized using density plots for validation purposes. The StdR method was selected for validation because it is known that StdR estimates under the REM model are assumed to follow a standard normal distribution [13]. StdR estimates for each simulated condition were concentrated at 0, which suggested that StdR estimates have a mean of 0 and indicated valid simulation findings (**Appendix B**).

4.1.1.1 Overall accuracy values

Figure 1 and **Table 2** provide information about the overall accuracy values for each outlier detection method. The highest median accuracy value was observed for the StdR (89.9%, IQR: 86.3%-92.5%), followed by RCPEV (87.9%, IQR: 83.6%-92.3%). The lowest median accuracy was for the RCEBV (68.6%, IQR: 38.2%-94.9%).

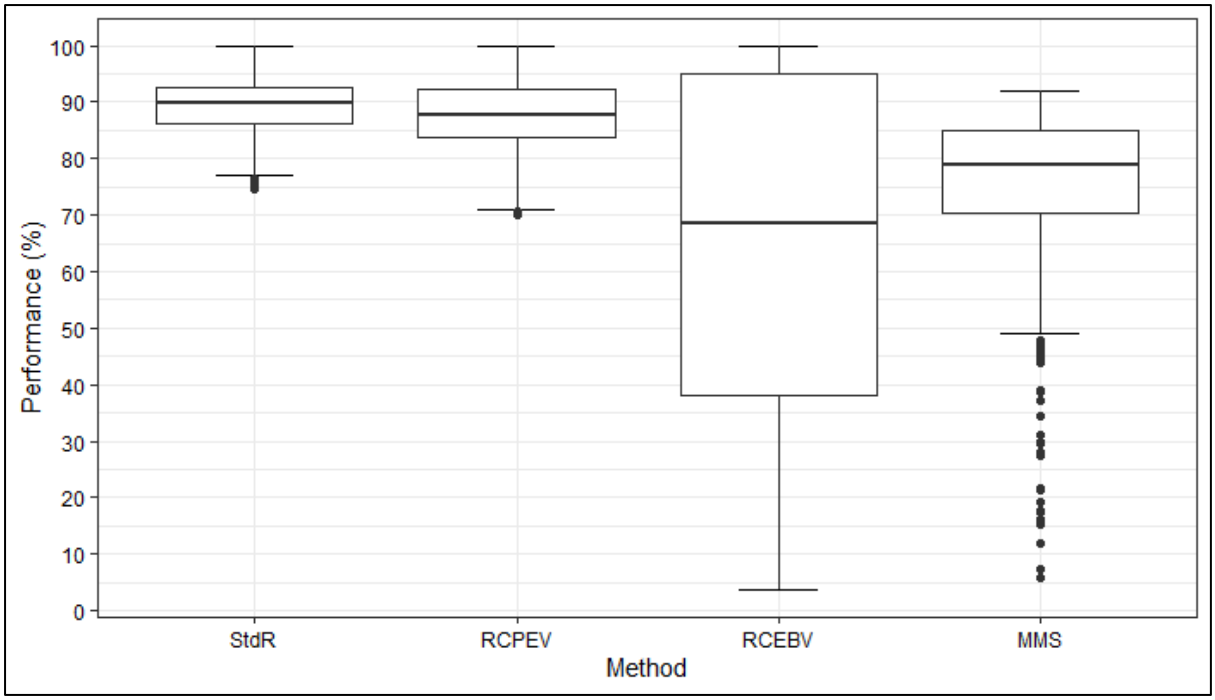

Figure 1. Overall accuracy of outlier detection methods.

We observed symmetric distributions in accuracy values for StdR and RCPEV methods. The RCEBV had a larger variation in accuracy values when compared to other outlier detection methods.

Table 2. Estimated quartiles for the overall accuracy values of outlier detection methods.

	Quartiles		
Method	25%	50%	75%
StdR	86.3	89.9	92.5
RCPEV	83.6	87.9	92.3
RCEBV	38.2	68.6	94.9
MMS	70.3	79.0	84.8

StdR= Studentized residual estimates, RCPEV= Relative change in the pooled effect estimate variance, RCEBV= Relative change in the estimated between-site variance, MMS= Model-based mean-shift method.

4.1.1.2 Overall F-1 score values

Figure 2 and **Table 3** summarize overall F-1 score values for each outlier detection method. The highest median F-1 score value was achieved by RCEBV (38.8%, IQR: 20.0%-66.9%).

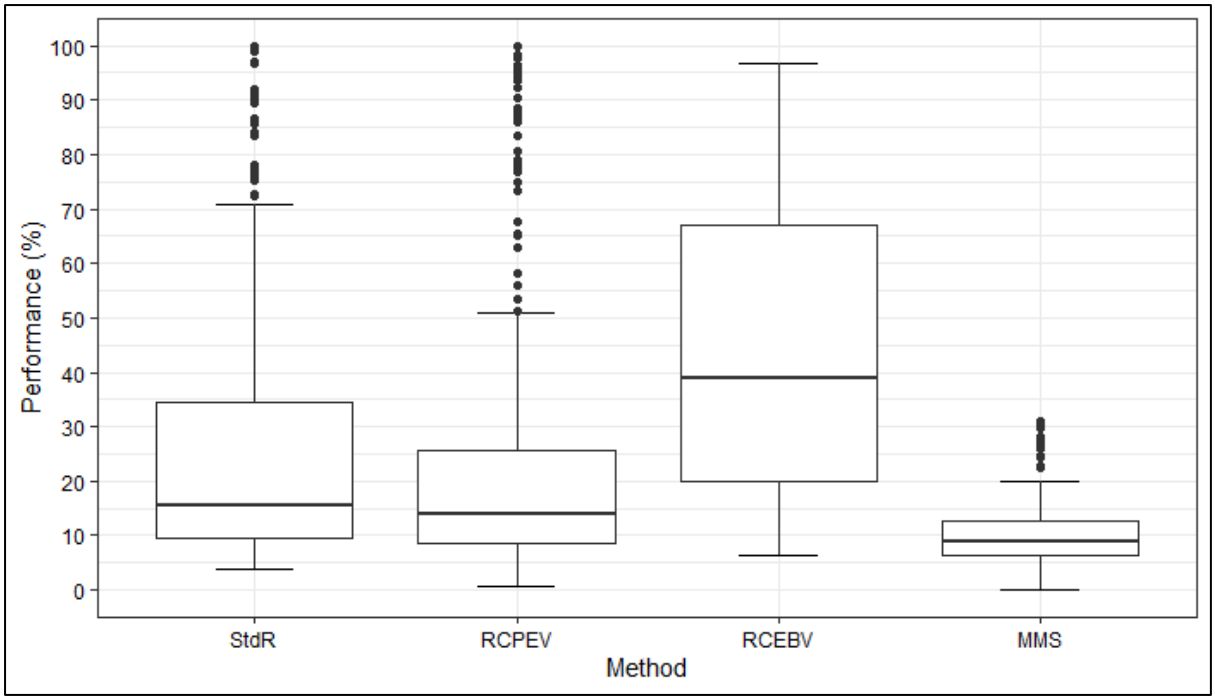

Figure 2. Overall F-1 score of outlier detection methods.

MMS had the lowest median F-1 score value of 9.0% (IQR: 6.4%-12.7%). However, the overall F-1 score values never exceeded 50%. Except for the MMS method, skewed distributions were observed in F-1 score values for the other methods. A larger variation in F-1 score values was observed for RCEBV when compared to other outlier detection methods.

Table 3. Estimated quartiles for the overall F-1 score of outlier detection methods.

	Quartiles		
Method	25%	50%	75%
StdR	9.4	15.6	34.5
RCPEV	8.5	13.8	25.5
RCEBV	20.0	38.8	66.9
MMS	6.4	9.0	12.7

StdR= Studentized residual estimates, RCPEV= Relative change in the pooled effect estimate variance, RCEBV= Relative change in the estimated between-site variance, MMS= Model-based mean-shift method.

4.1.1.3 Overall misclassification error values

A summary of the overall ME values for each outlier detection method is present in **Table 4** and **Figure 3**. StdR had the lowest (i.e., best) median ME value (10.2%, IQR: 7.5%-13.7%), followed by RCPEV (12.1%, IQR: 7.7%-16.4%). RCEBV had the highest median ME value (31.4%, IQR: 5.1%-61.8%).

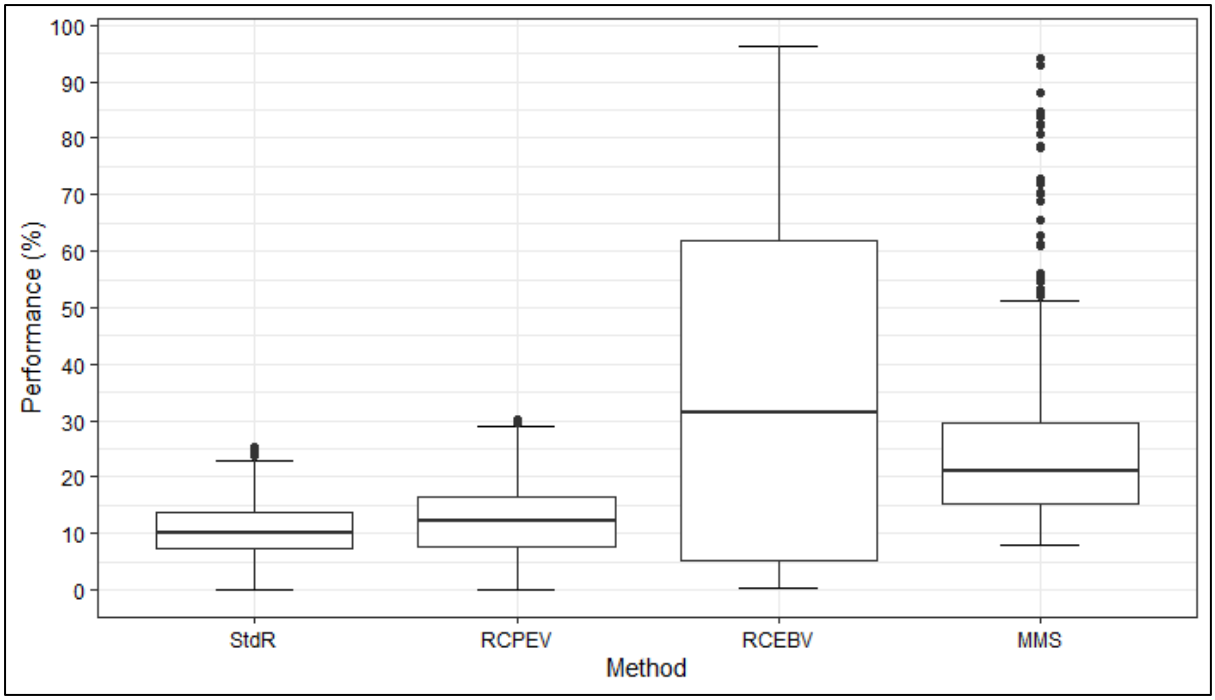

Figure 3. Overall misclassification error of outlier detection methods.

We observed symmetric distributions in ME values for StdR and RCPEV methods. Similar to accuracy and F-1 score values, a larger variation in ME values was observed for RCEBV when compared to other outlier detection methods.

Table 4. Estimated quartiles for the overall performance of outlier detection methods.

	Quartiles		
Method	25%	50%	75%
StdR	7.5	10.2	13.7
RCPEV	7.7	12.1	16.4
RCEBV	5.1	31.4	61.8
MMS	15.2	21.0	29.7

StdR= Studentized residual estimates, RCPEV= Relative change in the pooled effect estimate variance, RCEBV= Relative change in the estimated between-site variance, MMS= Model-based mean-shift method.

4.1.1.4 Summary

For accuracy and ME values, StdR outperformed other outlier detection methods, whereas RCEBV outperformed other methods for F-1 score values. We observed a larger variation in all performance metrics for RCEBV when compared to other outlier detection methods. The performance of RCPEV was consistent for all performance metrics compared to other methods.

A summary of performance metrics by simulation design characteristics and simulation parameters for each method is included in **Appendix C**. We observed that the performance of outlier detection methods was associated with simulation parameters. Specifically, the performance of the StdR and RCEBV methods to detect outliers improved when between-site variance and mean EE for outlier sites increased, whereas the performance of the RCPEV and MMS methods declined under these conditions.

4.1.2 Random-effects analysis of variance for performance metrics

For objective 1, the assumption of random-effects residuals in RE-ANOVA models was assessed using Q-Q plots for each method and each performance metric [43]. In Q-Q plots, all of the performance metrics for each method fall on the reference line of the standard normal distribution, indicating random-effects residuals follow a standard normal distribution (**Appendix D**).

The assumption of homogeneity of variances in performance metrics was assessed by plotting fitted values obtained from RE-ANOVA models on the horizontal axis and estimated random-effects residuals on the vertical axis for each method. In all of the plots, random-effects residuals were equally spread around the reference line and there was no evidence of distinct patterns. These results indicate the variances in performance metrics are homogeneous in RE-ANOVA models (**Appendix E**).

4.1.2.1 Accuracy values

Table 5 displays estimated η^2 (percentages) for accuracy values of outlier detection methods by individual simulation design characteristic/parameter and their 2-way interactions. The main effects and two-way interactions that were fitted in RE-ANOVA models for the accuracy values explained 90.9%, 85.5%, 85.1%, and 78.8% of the variation in RCEBV, StdR, RCPEV, and MMS methods. k and τ^2 explained the majority of variation

in accuracy values for the StdR, RCPEV, and MMS methods. While 64.3% and 15.4% of the variation in accuracy values were explained by σ^2 and τ^2 for the RCEBV method.

Table 5. Estimated η^2 (percentages) for accuracy values of outlier detection methods by individual simulation design characteristic/parameter and their 2-way interactions.

Simulation design characteristic/parameter	**StdR**	**RCPEV**	**RCEBV**	**MMS**
k	35.0	27.8	0.2	14.9
m	16.6	10.9	0.8	0.1
σ^2	0.1	3.8	64.3	8.0
μ_0	0.3	0.0	0.0	0.0
μ_1	0.3	0.2	0.4	3.4
τ^2	25.8	28.1	15.4	36.9
k:m	4.4	2.4	0.0	0.0
k:σ^2	0.0	0.2	0.3	0.1
k:μ_0	0.0	0.0	0.0	0.0
k:μ_1	0.0	0.1	0.0	0.1
k:τ^2	0.6	0.6	2.2	1.2
m:σ^2	0.2	0.6	0.2	0.3
m:μ_0	0.0	0.0	0.0	0.0
m:μ_1	0.0	0.0	0.0	0.1
m:τ^2	0.1	0.1	0.7	1.2
σ^2:μ_0	0.0	5.0	0.0	0.0
σ^2:μ_1	0.1	0.0	0.1	0.2
σ^2:τ^2	1.5	5.0	5.7	7.4
μ_0:μ_1	0.0	0.0	0.0	0.0
μ_0:τ^2	0.5	0.3	0.6	4.9
μ_1: τ^2	0.0	0.0	0.0	0.0
Total	**85.5**	**85.1**	**90.9**	**78.8**

: refers to an interaction between two simulation design characteristics/parameters.
StdR= Studentized residual estimates, RCPEV= Relative change in the pooled effect estimate variance, RCEBV = Relative change in the estimated between-site variance, MMS = Model-based mean-shift method, *k*= number of sites, *m*= number of outliers, σ^2= within-site variance, μ_0= mean effect estimate excluding outliers, μ_1= mean effect estimate for outliers, τ^2= between-site variance.

From **Table 6**, *k*, τ^2, and *m* accounted for 95.1%, 93.5%, and 90.2% of the variation in accuracy values for StdR among other simulation design characteristics and parameters included in RE-ANOVE models as main effects. For RCPEV, *k* and τ^2 explained 85.2% and 85.3% of variation in accuracy values. Overall, τ^2 contributed to 82.0% and 89.1% of variation in accuracy values for the RCEBV and MMS methods, respectively.

Amongst the two-way interactions of simulation design characteristics and parameters included in RE-ANOVA models, *k*:*m* explained 71.0% of variation in accuracy values for the StdR method. 50.7% of variation in accuracy values was described by σ^2:τ^2 in the RCPEV method, while σ^2:τ^2 accounted for 62.6% of variation in accuracy values for the RCEBV method. For the MMS method, σ^2:τ^2 and μ_0:τ^2 explained 62.0% and 52.0% of the variation in accuracy values.

Table 6. Estimated η_p^2 (percentages) for accuracy values of outlier detection methods by individual simulation design characteristic/parameter and their 2-way interactions.

Simulation design characteristic/parameter	**StdR**	**RCPEV**	**RCEBV**	**MMS**
k	95.1	85.2	6.4	24.0
m	90.2	69.2	20.1	2.9
σ^2	3.4	44.2	95.0	64.0
μ_0	16.2	0.0	0.0	0.0
μ_1	16.2	4.4	9.9	43.3
τ^2	93.5	85.3	82.0	89.1
k:m	71.0	33.6	0.8	0.8
k:σ^2	2.0	3.5	9.3	2.4
k:μ_0	0.0	0.0	0.0	0.0
k:μ_1	0.0	1.0	0.4	2.3
k:τ^2	24.5	10.9	39.7	20.7
m:σ^2	7.9	11.4	4.3	7.1
m:μ_0	0.0	0.0	0.0	0.0
m:μ_1	0.7	0.4	0.1	1.8
m:τ^2	3.9	2.1	16.4	21.2
σ^2:μ_0	0.0	50.7	0.0	0.0
σ^2:μ_1	2.9	0.0	1.7	5.0
σ^2:τ^2	45.4	50.7	62.6	62.0
μ_0:μ_1	0.0	0.0	0.0	0.0
μ_0:τ^2	20.5	5.6	15.3	52.0
μ_1: τ^2	0.0	0.0	0.0	0.0

: refers to an interaction between two simulation design characteristics/parameters.
StdR= Studentized residual estimates, RCPEV= Relative change in the pooled effect estimate variance, RCEBV = Relative change in the estimated between-site variance, MMS = Model-based mean-shift method, *k*= number of sites, *m*= number of outliers, σ^2= within-site variance, μ_0= mean effect estimate excluding outliers, μ_1= mean effect estimate for outliers, τ^2= between-site variance.

4.1.2.2 F-1 score values

Estimated η^2 for F-1 score values of outlier detection methods by individual simulation design characteristic/parameter and their 2-way interactions are present in **Table 7**. The main effects and two-way interactions that were included in RE-ANOVA models for the F-1 score values explained 86.1% of the variation in RCEBV, followed by 77.3%, 70.7%, and 70.2% of the variation in RCPEV, StdR, and MMS methods. Overall, *k* and τ^2 explained the majority of the variation in F-1 score values for all methods.

Table 7. Estimated η^2 (percentages) for F-1 score values of outlier detection methods by individual simulation design characteristic/parameter and their 2-way interactions.

Simulation design characteristic/parameter	StdR	RCPEV	RCEBV	MMS
k	8.2	18.7	69.4	13.4
m	0.0	2.7	2.6	12.8
σ^2	0.6	1.8	3.8	1.6
μ_0	0.0	0.0	0.0	0.0
μ_1	2.9	1.6	0.2	0.4
τ^2	48.5	28.9	5.3	30.8
k:m	0.1	0.8	0.1	0.2
k:σ^2	0.1	0.6	0.5	2.4
k:μ_0	0.0	0.0	0.0	0.0
k:μ_1	0.0	0.0	0.0	0.3
k:τ^2	0.5	0.5	1.1	2.9
m:σ^2	0.0	0.2	0.3	0.6
m:μ_0	0.0	0.0	0.0	0.0
m:μ_1	0.0	0.1	0.6	0.0
m:τ^2	0.3	0.9	0.0	0.6
σ^2:μ_0	0.0	0.0	0.1	0.0
σ^2:μ_1	0.3	0.2	0.0	0.3
σ^2:τ^2	5.5	18.2	1.7	2.5
μ_0:μ_1	0.0	0.0	0.4	0.0
μ_0:τ^2	3.7	2.1	0.0	1.4
μ_1: τ^2	0.0	0.0	0.0	0.0
Total	**70.7**	**77.3**	**86.1**	**70.2**

: refers to an interaction between two simulation design characteristics/parameters.
StdR= Studentized residual estimates, RCPEV= Relative change in the pooled effect estimate variance, RCEBV = Relative change in the estimated between-site variance, MMS = Model-based mean-shift method, *k*= number of sites, *m*= number of outliers, σ^2= within-site variance, μ_0= mean effect estimate excluding outliers, μ_1= mean effect estimate for outliers, τ^2= between-site variance.

Overall, τ^2 explained 94.2% and 70.5% of the variation in F-1 score values for StdR and RCPEV methods among other simulation design characteristics and parameters that were fitted in RE-ANOVA models as main effects (**Table 8**). While *k* described 86.5% and 67.5% of the variation in F-1 score values for RCEBV and MMS methods.

Table 8. Estimated η_p^2 (percentages) for F-1 score values of outlier detection methods by individual simulation design characteristic/parameter and their 2-way interactions.

Simulation design characteristic/parameter	StdR	RCPEV	RCEBV	MMS
k	63.4	65.6	86.5	67.5
m	0.3	21.7	19.3	47.4
σ^2	17.2	15.7	25.9	9.8
μ_0	0.0	0.0	0.0	0.0
μ_1	39.2	14.4	1.9	2.5
τ^2	94.2	70.5	32.7	46.4
k:m	3.1	8.0	1.1	1.2
k:σ^2	4.5	5.9	4.8	14.1
k:μ_0	0.0	0.0	0.0	0.0
k:μ_1	0.5	0.3	0.1	2.0
k:τ^2	14.5	4.6	9.0	16.5
m:σ^2	0.1	2.1	2.7	3.8
m:μ_0	0.0	0.0	0.0	0.0
m:μ_1	0.2	0.5	5.1	0.1
m:τ^2	8.2	8.9	0.0	4.2
σ^2:μ_0	0.0	0.0	1.3	0.0
σ^2:μ_1	9.8	2.3	0.0	2.2
σ^2:τ^2	60.1	62.1	13.4	14.4
μ_0:μ_1	0.0	0.0	3.9	0.0
μ_0:τ^2	47.4	18.3	0.0	8.5
μ_1: τ^2	0.0	0.0	4.0	0.0

: refers to an interaction between two simulation design characteristics/parameters.
StdR= Studentized residual estimates, RCPEV= Relative change in the pooled effect estimate variance, RCEBV = Relative change in the estimated between-site variance, MMS = Model-based mean-shift method, *k*= number of sites, *m*= number of outliers, σ^2= within-site variance, μ_0= mean effect estimate excluding outliers, μ_1= mean effect estimate for outliers, τ^2= between-site variance.

Amongst the two-way interactions of simulation design characteristics and parameters included in RE-ANOVA models, σ^2:τ^2 accounted for 60.1%, 62.1%, 13.4%, and 14.4% of variation in F-1 score values for StdR, RCPEV, RCEBV, and MMS methods. While μ_0:τ^2

explained 47.4%, 18.3% and 8.5% of variation in F-1 score values for StdR, RCPEV, and MMS methods.

4.1.2.3 Misclassification error values

From the estimated η^2 for ME values, the main effects and two-way interactions that were fitted in RE-ANOVA models explained 94.7% of the variation in RCEBV, followed by 85.8%, 84.3%, and 76.0% of the variation in StdR, RCPEV, and MMS methods (**Table 9**).

Table 9. Estimated η^2(percentages) for misclassification error values of outlier detection methods by individual simulation design characteristic/parameter and their 2-way interactions.

Simulation design characteristic/parameter	StdR	RCPEV	RCEBV	MMS
k	31.0	25.8	0.2	1.6
m	15.3	12.9	0.9	0.1
σ^2	0.1	4.8	64.3	10.4
μ_0	0.0	0.0	0.0	0.0
μ_1	0.5	0.5	0.4	3.4
τ^2	27.8	25.7	18.4	36.9
k:m	5.3	4.8	0.0	0.0
k:σ^2	0.0	0.7	0.3	0.2
k:μ_0	0.0	0.0	0.0	0.0
k:μ_1	0.0	0.1	0.0	0.1
k:τ^2	1.2	1.2	2.2	1.8
m:σ^2	0.5	0.8	0.4	0.3
m:μ_0	0.0	0.0	0.0	0.0
m:μ_1	0.0	0.0	0.0	0.1
m:τ^2	0.6	0.1	1.0	1.2
σ^2:μ_0	0.0	0.0	0.0	0.0
σ^2:μ_1	0.2	0.1	0.1	0.6
σ^2:τ^2	2.5	6.0	5.9	9.7
μ_0:μ_1	0.0	0.0	0.0	0.0
μ_0:τ^2	0.8	0.8	0.6	9.6
μ_1: τ^2	0.0	0.0	0.0	0.0
Total	**85.8**	**84.3**	**94.7**	**76.0**

: refers to an interaction between two simulation design characteristics/parameters.
StdR= Studentized residual estimates, RCPEV= Relative change in the pooled effect estimate variance, RCEBV = Relative change in the estimated between-site variance, MMS = Model-based mean-shift method, *k*= number of sites, *m*= number of outliers, σ^2= within-site variance, μ_0= mean effect estimate excluding outliers, μ_1= mean effect estimate for outliers, τ^2= between-site variance.

From **Table 9**, *k* and τ^2 explained the majority of the variation in ME values for StdR and RCPEV methods, whereas σ^2 and τ^2 described most of the variation in ME values for RCEBV and MMS methods.

Table 10. Estimated η_p^2 (percentages) for misclassification error values of outlier detection methods by individual simulation design characteristic/parameter and their 2-way interactions.

Simulation design characteristic/parameter	**StdR**	**RCPEV**	**RCEBV**	**MMS**
k	93.0	80.0	5.0	27.0
m	90.0	72.0	23.0	2.9
σ^2	5.4	48.2	85.0	67.0
μ_0	0.0	0.0	0.0	0.0
μ_1	18.2	6.0	13.0	47.3
τ^2	82.0	82.3	91.0	89.1
k:m	75.0	36.6	0.0	0.6
k:σ^2	5.0	5.0	6.0	2.7
k:μ_0	0.0	0.0	0.0	0.0
k:μ_1	2.0	1.0	0.4	2.3
k:τ^2	30.0	15.0	39.7	29.7
m:σ^2	10.0	18.0	7.3	7.0
m:μ_0	0.0	0.0	0.0	0.0
m:μ_1	0.7	0.4	0.1	2.0
m:τ^2	6.0	3.0	22.0	22.2
σ^2:μ_0	0.0	0.0	0.0	0.0
σ^2:μ_1	2.7	2.1	1.7	9.0
σ^2:τ^2	55.4	55.7	65.6	65.0
μ_0:μ_1	0.0	0.0	0.0	0.0
μ_0:τ^2	28.5	6.6	15.3	64.0
μ_1: τ^2	0.0	0.0	0.0	0.0

: refers to an interaction between two simulation design characteristics/parameters.
StdR= Studentized residual estimates, RCPEV= Relative change in the pooled effect estimate variance, RCEBV = Relative change in the estimated between-site variance, MMS = Model-based mean-shift method, *k*= number of sites, *m*= number of outliers, σ^2= within-site variance, μ_0= mean effect estimate excluding outliers, μ_1= mean effect estimate for outliers, τ^2= between-site variance.

Overall, *k* described 93.0% and 80.0% of the variation in ME values for StdR and RCPEV methods among other simulation design characteristics and parameters that were fitted in RE-ANOVA models as main effects (**Table 10**). As well, τ^2 described 91.0% and 89.1% of the

variation in ME values for the RCEBV and MMS methods. Amongst the two-way interactions of simulation design characteristics and parameters included in RE-ANOVA models, *k*:*m* accounted for 75.0% of variation in ME values for StdR. For RCPEV, RCEBV, and MMS methods, σ^2:τ^2 explained 55.7%, 65.6%, and 65.0% of variation in ME values, respectively.

4.1.2.4 Summary

Among simulation design characteristics and parameters included in RE-ANOVA models as main effects, *k* and τ^2 explained the majority of the variation in all performance metrics. The two-way interactions of σ^2:τ^2 and *k*:*m* captured most of the variation in accuracy and ME values compared to other two-way interactions fitted in RE-ANOVA models. In terms of F-1 score values, σ^2:τ^2 and μ_0:τ^2 explained the majority of variation among other two-way interactions of simulation design characteristics and parameters included in the models.

4.2 Application to real-world meta-analyses

In objective 2, we illustrated the application of outlier detection methods to real-world meta-analyses using the following steps: 1) systematically selected meta-analyses from peer-reviewed publications of site-specific EEs conducted by investigators from CNODES, 2) applied outlier detection methods to the selected meta-analyses, 3) compared outlier detection methods based on the selected meta-analyses, and 4) performed REM models with and without potential outliers to capture the impact of site-specific EEs that are detected as potential outliers.

4.2.1 Systematic selection of real-world meta-analyses

As **Figure 4** reveals, we identified a total of 249 meta-analyses of site-specific EEs that were published between 2011 and 2022. Subsequently, we excluded 55 meta-analyses published before 2016 in which FEM models were applied by investigators from CNODES to pool site-specific EEs.

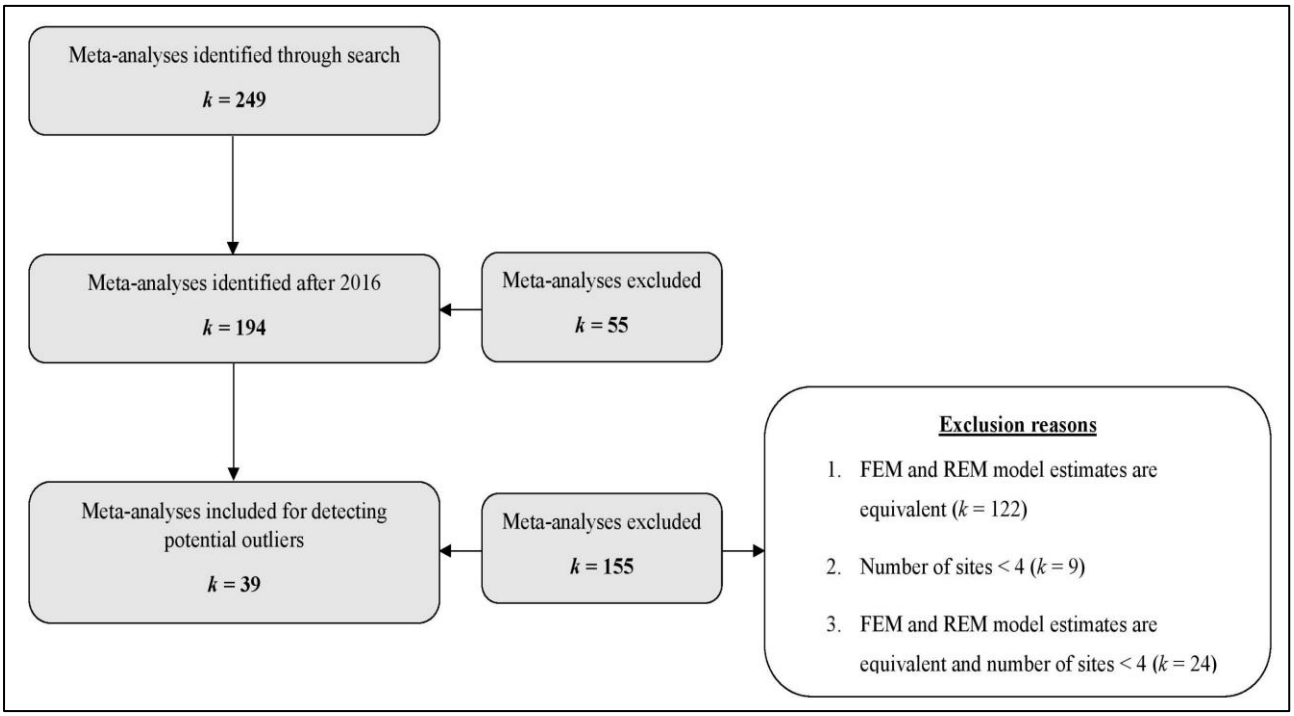

Figure 4. Flow diagram for selected meta-analyses from the Canadian Network for Observational Drug Effect Studies.

155 meta-analyses were further excluded: 122 of these meta-analyses had equivalent parameter estimates for FEM and REM models, another 9 meta-analyses consisted of less than 4 sites, and 24 meta-analyses were excluded for both reasons. Accordingly, we included 39 meta-analyses of site-specific EEs for the application of outlier detection methods. The selected meta-analyses can be found in **Appendix F**.

4.2.3 Comparison of outlier detection methods

A comparison of outlier detection methods for the selected meta-analyses is illustrated in **Figure 5**. In **Panel A**, we observed that the StdR and RCPEV methods detected potential outliers in all selected meta-analyses.

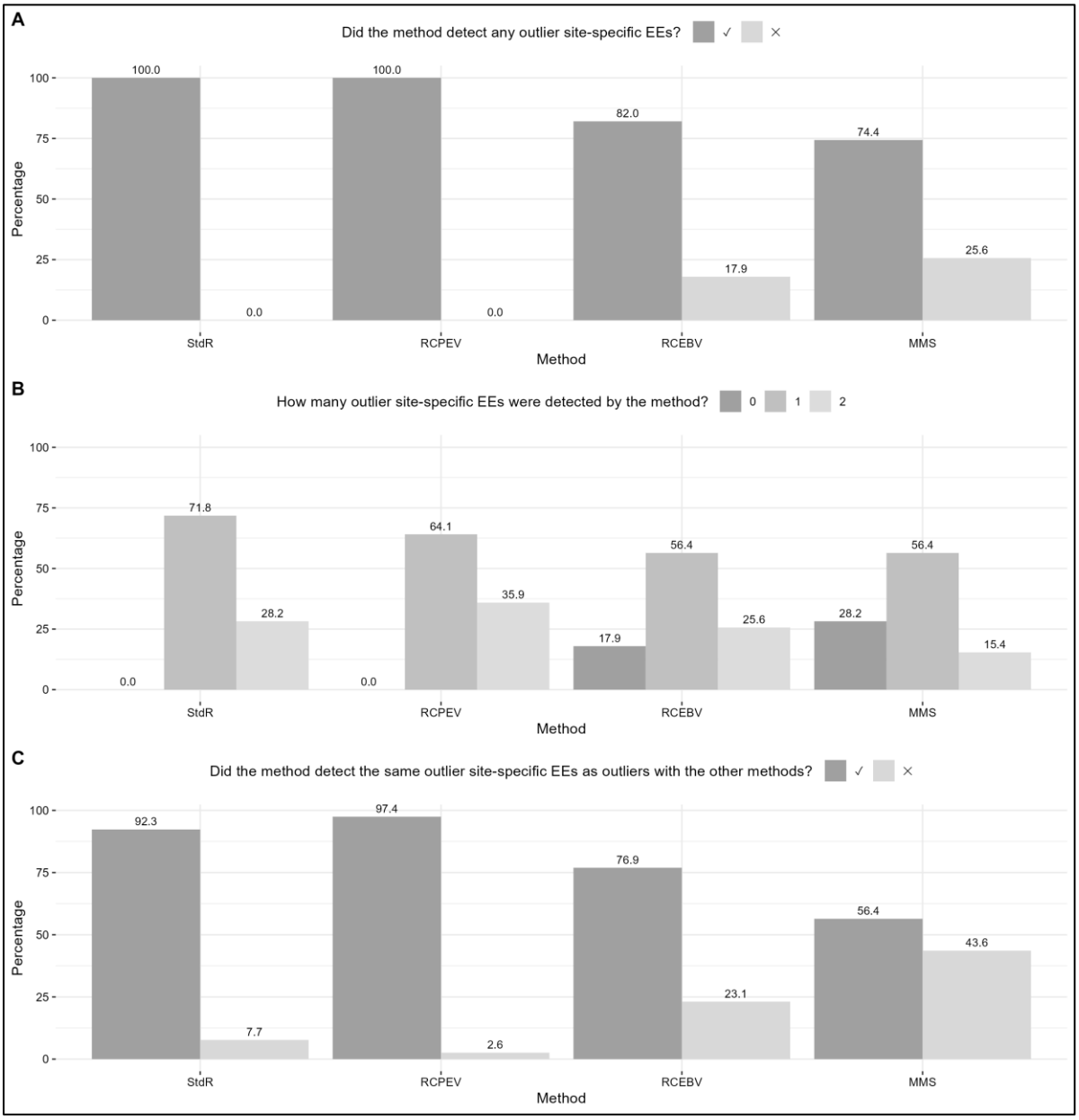

Figure 5. Comparison of outlier detection methods from selected meta-analyses conducted by investigators from the Canadian Network for Observational Drug Effect Studies (k = 39).

However, the RCEBV and MMS methods did not detect any site-specific EEs as potential outliers in 17.9% and 25.6% of the selected meta-analyses, respectively. Overall, the StdR and RCPEV methods demonstrated higher sensitivity in detecting potential outliers compared

to the RCEBV and MMS methods. From **Panel B**, the StdR, RCPEV, RCEBV, and MMS methods detected a single potential outlier in 71.8%, 64.1%, 56.4%, and 56.4% of the selected meta-analyses, respectively. Across the selected meta-analyses, the StdR, RCPEV, RCEBV, and MMS methods consistently detected potential outliers in site-specific EEs, with percentages of 92.3%, 97.4%, 76.9%, and 56.4% (**Panel C**).

4.2.4 Impact of potential outliers

The impact of potential outliers was captured by performing REM models with and without site-specific EEs that were detected as potential outliers (**Figure 7**). We estimated $\hat{\tau}^2$, $\hat{I}^2$, and pooled EEs from REM models with and without potential outliers.

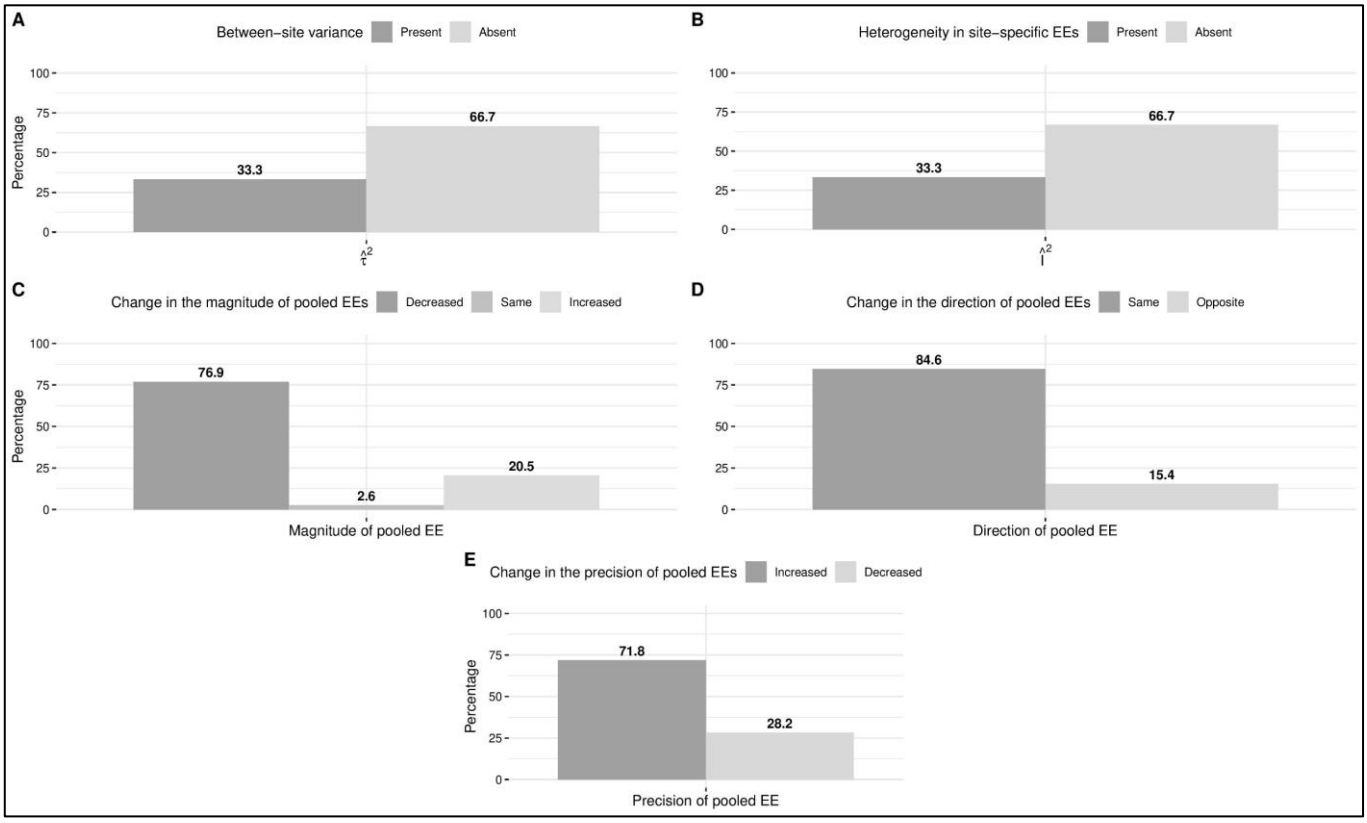

Figure 6. Impact of potential outliers on $\hat{\tau}^2$ (Panel A), $\hat{I}^2$ (Panel B), change in the magnitude (Panel C), direction (Panel D), and precision (Panel E) of pooled EEs (k = 39).

From **Figure 6**, estimated $\hat{\tau}^2$ (**Panel A**) and $\hat{I}^2$ (**Panel B**) were reduced to 0.0 in 66.7% of the selected meta-analyses when potential outliers were excluded from REM models. In the majority of selected meta-analyses, a decrease in the magnitude of pooled EEs was observed

in REM models without potential outliers compared to REM models with potential outliers (**Panel C**). The direction of change in the pooled EEs remained the same between REM models with and without potential outliers in 84.6% of the selected meta-analyses (**Panel D**). In 71.8% of selected meta-analyses, we observed an increase in the precision of pooled EEs for REM models without potential outliers when compared to the precision of pooled EEs for REM models with potential outliers (**Panel E**).

Chapter 5: Discussion and Conclusions

5.1 Summary and discussion

The purpose of this research was to examine outlier detection methods for meta-analyses of site-specific EEs from distributed analyses of a multi-site network. For objective 1, a simulation study was conducted to compare the performance of outlier detection methods for the REM model. The steps in the simulation study were as follows: 1) randomly generated data from the prespecified simulation conditions, 2) checked the validity of simulation results, 3) calculated performance metrics for each outlier detection method, and 4) analyzed performance metrics for each method using boxplots and RE-ANOVA models.

For objective 2, we illustrated the application of outlier detection methods to real-world meta-analyses using the following steps: 1) systematically selected meta-analyses from peer-reviewed publications of site-specific EEs conducted by investigators from CNODES, 2) applied outlier detection methods to the selected meta-analyses, 3) compared outlier detection methods based on the selected meta-analyses, and 4) performed REM models with and without potential outliers to capture the impact of site-specific EEs that are detected as potential outliers.

The overall performance of outlier detection methods from the simulation study was consistent with previous research, which also indicated that the StdR and RCEBV methods outperformed other outlier detection methods [17,22]. However, a large variation in the performance of the RCEBV method was observed because this method is sensitive to small between-site variance [13]. Our simulation study also captured differences in the performance of outlier detection methods associated with simulation parameters. Specifically, the performance of the StdR and RCEBV methods to detect outliers improved when between-site variance and mean EE for outlier sites increased, whereas the performance of the RCPEV and

MMS methods declined. The selection of a wide range of simulation conditions enabled us to capture differences in the performance of outlier detection methods, addressing a limitation in previous research [17,22]. Briefly, previous research considered arbitrarily selected values of outlier site-specific EEs and small between-site variance [17,22]. Whereas we randomly generated site-specific EEs from the mean of site-specific EEs, assuming a normal distribution, where the mean of site-specific EEs differed between outlier site-specific EEs and the remaining site-specific EEs. Values of between-site variance comprised a range representing small to large heterogeneity in site-specific EEs. While previous research explored only 6 and 24 simulation conditions [17,22], we explored 54 simulation conditions.

The current simulation study represents a novel contribution to the literature on outlier detection methods in meta-analysis models. Using RE-ANOVA models, we identified the extent to which simulation design characteristics and parameters explained variations in the performance of each method. Among simulation design characteristics and parameters included in RE-ANOVA models as main effects, k and τ^2 explained the majority of the variation in all performance metrics. The two-way interactions of σ^2:τ^2 and k:m captured most of the variation in accuracy and ME values compared to other two-way interactions fitted in RE-ANOVA models. In terms of F-1 score values, σ^2:τ^2 and μ_0:τ^2 explained the majority of variations among two-way interactions of simulation design characteristics and parameters included in the models.

From systematically selected meta-analyses of site-specific EEs conducted by investigators from CNODES, a single site-specific EE was detected as a potential outlier more frequently than multiple site-specific EEs as potential outliers. This finding was consistent with the characteristics of meta-analyses for multi-site networks, which typically involve a small number of site-specific EEs [3] and a limited number of potential outliers

[15]. Suissa et al. recommended using the FEM model when pooling site-specific EEs, as multi-site networks conduct research with an identical site-specific study protocol to minimize heterogeneity in site-specific EEs and assume site-specific EEs are not randomly sampled from a potentially larger set of sites [3]. However, we recommend using REM models to pool site-specific EEs because REM models can reduce to FEM models, in cases of limited heterogeneity in site-specific EEs [1,49]. REM models also provide more unbiased pooled EEs than FEM models when dealing with potential outliers [49].

The StdR and RCPEV methods detected the same site-specific EEs as potential outliers in the bulk of real-world meta-analyses, suggesting limited variation in these methods. Aoki et al. and Viechtbauer et al. applied the StdR, RCPEV, RCEBV, and MMS methods in real-world meta-analyses and found similar results, further supporting the limited variation between the StdR and RCPEV methods for detecting potential outliers [13,22]. It appears that the StdR and RCPEV methods can detect potential outliers consistently in real-world meta-analyses when compared to RCEBV and MMS methods.

In the real-world meta-analyses, we performed REM models with and without potential outliers to assess their impact on heterogeneity parameters (i.e., $\hat{\tau}^2$and $\hat{I}^2$) and pooled EEs. When we excluded potential outliers from REM models, we observed a substantial reduction in estimated $\hat{\tau}^2$and $\hat{I}^2$ for the REM models when compared to REM models that contained potential outliers. Demir et al. observed a similar result, demonstrating that heterogeneity in site-specific EEs appeared due to potential outliers [49]. Moreover, detecting and excluding outliers from real-world meta-analyses can provide precise pooled EEs, thereby improving the ability to make informed decisions about the effectiveness of treatments. However, researchers should report the results of pooled EEs with and without potential outliers when a change is observed in the direction of pooled EEs from REM models [13,49].

In the meta-analyses, we identified discrepancies in the reporting of REM model parameter estimates. Specifically, there was inconsistent reporting of the standard errors of site-specific EEs, heterogeneity of site-specific EEs (i.e., $\hat{\tau}^2$and $\hat{I}^2$), and site-specific weights. We estimated the REM model parameters that were not reported, but it is important to note that these estimates may vary depending on the statistical package and programming language used. The Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) should be used to guide the reporting of meta-analyses conducted by investigators from CNODES [52]. Particularly, PRISMA item number 20 refers to reporting standard errors of site-specific EEs, along with a measure of heterogeneity in site-specific EEs (i.e., $\hat{\tau}^2$and $\hat{I}^2$) and site-specific weights for any meta-analyses [52]. Therefore, we recommend that investigators from CNODES adhere to PRISMA guidelines when presenting and reporting the results of meta-analyses.

Overall findings from the simulation study and real-world meta-analyses suggest that RCEBV and MMS methods did not perform better than StdR and RCPEV methods. The RCEBV method compares the ratio of $\hat{\tau}^2$ that excludes site-specific EE one at a time with estimated $\hat{\tau}^2$ that contains all site-specific EEs [13]. This ratio of estimated $\hat{\tau}^2$ makes the RCEBV method sensitive to instances where $\hat{\tau}^2$ estimates are small [13,22], limiting its ability to detect potential outliers. For the MMS method, a parametric bootstrap method was used to find the empirical sampling distribution of the test statistic [22]. However, the parametric bootstrap requires the number of site-specific EEs > 4 to ensure that the resampled EEs are representative of the population distribution, which may not be feasible for some meta-analyses [3,4]. Based on these findings, we thus recommend using StdR and RCPEV methods to detect potential outliers.

Although CNODES uses a structured process to investigate potential outlier site-specific EEs, this process may fail to determine any apparent reason for their potential presence [4]. Therefore, we recommend that investigators from CNODES implement StdR and RCPEV methods to detect potential outliers that are not detected by the current process [4]. Once outliers are detected, researchers should investigate those outlier sites. For example, differences in the site-specific AHD sources can contribute to data variability [9,10]. Therefore, it is important to examine differences in data sources and determine if they might contribute to the presence of potential outliers. By taking these steps, researchers can ensure that their meta-analyses are based on reliable and accurate site-specific EEs, leading to unbiased pooled EEs.

5.2 Strengths and limitations

This research has many strengths. For objective 1, one major strength is that we compared the performance of outlier detection methods under a wide range of simulation conditions, enabling us to capture variations in the performance of methods. We considered medians of performance metrics because median estimates are consistent for all distributions of performance metrics whether or not they are symmetric. As well, the median is less sensitive than the mean to potential outliers. Additionally, we addressed the class imbalance between site-specific EEs that are outliers and the remaining site-specific EEs with F-1 scores for each method [42]. Previous studies have used parametric bootstrapping to determine the empirical sampling distributions of the test statistics for the StdR, RCEBV, and RCPEV methods [17,22]. However, parametric bootstrapping requires a large number of site-specific EEs to ensure that the resampled dataset represents the population distribution, which may not be feasible when the number of site-specific EEs < 10 [28,29]. Therefore, we used the StdR, RCEBV, and RCPEV methods as rules of thumb in this research [13]. For objective 2, we systematically selected published studies and applied outlier detection methods to selected

meta-analyses, which helped us to compare outlier detection methods and to evaluate the impact of potential outliers on the REM model parameters in real-world scenarios.

Despite these strengths, this research has some limitations. First, in the presence of multiple potential outliers, excluding site-specific EEs one at a time from meta-analysis models and calculating estimates using the StdR, RCPEV, and RCEBV methods may lead to biased results due to masking and swamping effects [16,53,54]. The masking effect refers to unidentified potential outliers because outlier site-specific EEs may be clustered together, and detecting a single potential outlier is affected by the presence of other potential outliers present in meta-analysis models [16,53,54]. The swamping effect refers to site-specific EEs misclassified as potential outliers because true outliers may shift estimates toward them and away from site-specific EEs that are not potential outliers. In other words, the masking effect is equivalent to false negatives, while the swamping effect is false positives [16,53,54]. Second, the RCEBV method is sensitive to detect potential outliers in cases of limited between-site variance and the StdR method has high false discovery rates [13]. Third, although we applied the DL method to estimate between-site variance, many other methods have been proposed in the literature for this purpose [5,49–51,55]. Our research is primarily focused on outlier detection methods rather than exploring various methods to estimate between-site variance; nonetheless, we discuss opportunities for future research in this area.

5.3 Opportunities for future research

Several opportunities for future research exist. First, several methods have been proposed to estimate between-site variance and these methods can directly impact the REM model parameters, including the estimation of site-specific weights [49,50]. Previous research compared numerous methods to estimate between-site variance; however, there is a lack of agreement on which method is more suitable than others under different conditions [5,49–

51,55]. For example, Langan et al. have proposed alternative methods for estimating between-site variance than the DL method [56,57], whereas Thorlund et al. suggest that the restricted maximum likelihood estimator produces more unbiased estimates for between-site variance compared to other methods [58]. Petropoulou and Mavridi found that the Sidik-Jonkman estimator for between-site variance performs poorly when compared to other estimators [50]. There is also a lack of information about the impact of various methods to estimate between-site variance in the presence of potential outliers. Thus, a comparison of methods to estimate between-site variance in the presence of potential outliers would be an important step as these methods can directly impact the REM model parameters and site-specific EEs [49]. In real-world settings, researchers would benefit from understanding the variability and sensitivity of methods for estimating between-site variance in the presence of potential outliers [5].

Lastly, a meta-analysis is typically comprised of a small number of site-specific EEs, and it is generally not recommended to exclude potential outliers from meta-analysis models [13,59]. Potential outliers may represent important sources of heterogeneity in site-specific EEs and excluding potential outliers from meta-analysis models may lead to the loss of important information [13,59]. Instead, researchers have proposed using methods that can accommodate potential outliers by downweighting them in meta-analysis models assuming heavy-tailed distributions (such as a t-distribution, arcsinh distribution, and beta distribution) for site-specific EEs [5,15,60]. Further research is required to investigate the impact of various model assumptions and specifications on the pooled EEs that can account for potential outliers in meta-analysis models.

5.4 Conclusions

Overall, the StdR and RCPEV methods outperformed the RCEBV and MMS methods for detecting outliers. The RCEBV method showed larger variation in accuracy, F-1 score, and misclassification error values compared to other methods. The majority of variance in the performance metrics was attributed to the number of sites, between-site variance, and the interaction between within- and between-site variance across all methods. A substantial reduction in the heterogeneity of site-specific EEs was observed between REM models with and without potential outliers. Excluding potential outliers from the REM models resulted in narrower 95% CIs for the pooled EEs compared to REM models that included potential outliers.

Practical implications for researchers and organizations conducting multi-site research include the following: 1) REM models should be preferred for pooling site-specific EEs because REM models can reduce to FEM models when estimated between-site variance is small and they provide more unbiased pooled EEs than FEM models in the presence of potential outliers, 2) reporting of results from meta-analyses should adhere to PRISMA guidelines, 3) StdR and RCPEV methods should be used to detect potential outliers, and 4) sites that are detected as potential outliers should be further investigated to examine differences in data sources that may contribute to the estimation of very large or very small site-specific EEs.

References

1. Borenstein M, Hedges L V., Higgins JPT, Rothstein HR. A Basic Introduction to Fixed-effect and Random-effects Models for Meta-analysis. Res Synth Methods [Internet]. 2010 Apr;1(2):97–111.

2. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. Cochrane Handbook for Systematic Reviews of Interventions. John Wiley & Sons; 2019. 1–694 p.

3. Suissa S, Henry D, Caetano P, Dormuth CR, Ernst P, Hemmelgarn B, et al. CNODES: The Canadian Network for Observational Drug Effect Studies. Open Med [Internet]. 2012;6(4):134–40.

4. Platt RW, Platt R, Brown JS, Henry DA, Klungel OH, Suissa S. How Pharmacoepidemiology Networks Can Manage Distributed Analyses to Improve Replicability and Transparency and Minimize Bias. Pharmacoepidemiol Drug Saf [Internet]. 2020 Jan 1;29(S1):3–7.

5. Lin L, Chu H, Hodges JS. Alternative Measures of Between-study Heterogeneity in Meta-analysis: Reducing the Impact of Outlying Studies. Biometrics [Internet]. 2017 Mar 1;73(1):156–66.

6. Hedges L V., Olkin I. Statistical Methods for Meta-Analysis. Academic Press. Orlando, FL: Academic Press; 1985. 369 p.

7. Reynolds RF, Kurz X, de Groot MCH, Schlienger RG, Grimaldi-Bensouda L, Tcherny-Lessenot S, et al. The IMI PROTECT Project: Purpose, Organizational Structure, and Procedures. Pharmacoepidemiol Drug Saf [Internet]. 2016 Mar 1;25:5–10.

8. Behrman RE, Benner JS, Brown JS, McClellan M, Woodcock J, Platt R. Developing the Sentinel System — A National Resource for Evidence Development. N Engl J Med [Internet]. 2011 Jan 12;364(6):498–9.

9. Platt RW, Dormuth CR, Chateau D, Filion K. Observational Studies of Drug Safety in Multi-database Studies: Methodological Challenges and Opportunities. eGEMs [Internet]. 2016;4(1):9.

10. Doyle CM, Lix LM, Hemmelgarn BR, Paterson JM, Renoux C. Data Variability Across Canadian Administrative Health Databases: Differences in Content, Coding, and Completeness. Pharmacoepidemiol Drug Saf [Internet]. 2020 Jan 1;29(S1):68–77.

11. Turner RM, Bird SM, Higgins JPT. The Impact of Study Size on Meta-analyses: Examination of Underpowered Studies in Cochrane Reviews. PLoS One [Internet]. 2013 Mar 27;8(3).

12. Cuijpers P, Griffin JW, Furukawa TA. The Lack of Statistical Power of Subgroup Analyses in Meta-analyses: A Cautionary Note. Epidemiol Psychiatr Sci [Internet]. 2021;30.

13. Viechtbauer W, Cheung MW-L. Outlier and Influence Diagnostics for Meta-analysis. Res Synth Methods [Internet]. 2010 Apr;1(2):112–25.

14. Gumedze FN, Jackson D. A Random Effects Variance Shift Model for Detecting and Accommodating Outliers in Meta-analysis. BMC Med Res Methodol [Internet]. 2011;11.

15. Beath KJ. A Finite Mixture Method for Outlier Detection and Robustness in Meta-analysis. Res Synth Methods [Internet]. 2014 Dec 1;5(4):285–93.

16. Mavridis D, Moustaki I, Wall M, Salanti G. Detecting Outlying Studies in Meta-regression Models Using a Forward Search Algorithm. Res Synth Methods [Internet]. 2017 Jun 1;8(2):199–211.

17. Negeri ZF, Beyene J. Statistical Methods for Detecting Outlying and Influential Studies in Meta-analysis of Diagnostic Test Accuracy Studies. Stat Methods Med Res [Internet]. 2020 Apr 1;29(4):1227–42.

18. Noma H, Gosho M, Ishii R, Oba K, Furukawa TA. Outlier Detection and Influence Diagnostics in Network Meta-analysis. Res Synth Methods [Internet]. 2020 Nov 1;11(6):891–902.

19. Viechtbauer W. Conducting Meta-analyses in R with the metafor. J Stat Softw [Internet]. 2010;36(3):1–48.

20. Lin L, Rosenberger KJ, Shi L, Wang Y, Chu H. Package “altmeta.” CranRstudioOrg [Internet]. 2022;

21. Schwarzer G, Mair P, Hatzinger R. meta : An R Package for Meta-Analysis. CranRstudioOrg [Internet]. 2016;7(January):40–5.

22. Aoki M, Noma H, Gosho M. Methods for Detecting Outlying Regions and Influence Diagnosis in Multi-regional Clinical Trials. Biostat Epidemiol [Internet]. 2021;5(1):30–48.

23. Cochran WG. The Combination of Estimates from Different Experiments. Biometrics [Internet]. 1954 Mar;10(1):101.

24. Jackson D. The Power of the Standard Test for the Presence of Heterogeneity in Meta-analysis. Stat Med [Internet]. 2006 Aug 15;25(15):2688–99.

25. Hardy RJ, Thompson SG. Detecting and Describing Heterogeneity in Meta-analysis. Stat Med [Internet]. 1998;17(8):841–56.

26. Jaki T, Kim M, Lamont A, George M, Chang C, Feaster D, et al. The Effects of Sample Size on the Estimation of Regression Mixture Models. Educ Psychol Meas [Internet]. 2019;79(2):358–84.

27. Zhang J, Fu H, Carlin BP. Detecting Outlying Trials in Network Meta-analysis. Stat Med [Internet]. 2015 Aug 30;34(19):2695–707.

28. Zhang M, Li Y, Lu J, Shi L. Outlier Detection and Accommodation in Meta-regression Models. Commun Stat - Theory Methods [Internet]. 2021;50(8):1728–44.

29. Hall P. The Bootstrap and Edgeworth Expansion. Vol. 45, Springer Science & Business Media. 1996. 532 p.

30. Chernick M. Bootstrap Methods: A Guide for Practitioners and Researchers. John Wiley & Sons. 2011.

31. Noma H, Gosho M, Ishii R, Oba K, Furukawa TA. Outlier Detection and Influence Diagnostics in Network Meta-analysis. Res Synth Methods [Internet]. 2020;11(6):891–902.

32. DerSimonian R, Laird N. Meta-analysis in Clinical Trials. Control Clin Trials [Internet]. 1986 Sep 1;7(3):177–88.

33. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring Inconsistency in Meta-analyses. Br Med J [Internet]. 2003 Sep 4;327(7414):557–60.

34. Higgins JPT, Thompson SG. Quantifying Heterogeneity in a Meta-analysis. Stat Med

[Internet]. 2002 Jun 15;21(11):1539–58.

35. Belsley DA, Kuh E, Welsch RE. Regression Diagnostics: Influential Data and Sources of Collinearity. Wiley Online Library. Wiley Series in Probability and Statistics; 1980. 292 p.

36. Cook RD, Weisberg S. Residuals and Influence in Regression. J R Stat Soc [Internet]. 1984;147(1):108.

37. Shiffler RE. Maximum Z Scores and Outliers. Am Stat [Internet]. 1988;42(1):79–80.

38. Oyeyemi GM, Bukoye A, Akeyede I. Comparison of Outlier Detection Procedures in Multiple Linear Regressions. Am J Math Stat [Internet]. 2015;5(1):37–41.

39. Andiloro NR. Hierarchical Meta-analysis: A Simulation Study Comparing Classical Random Effects and Fully Bayesian Methods. Vol. 79, Dissertation Abstracts International: Section B: The Sciences and Engineering. 2018.

40. Jackson D, Bowden J. Confidence Intervals for the Between-study Variance in Random-effects Meta-analysis Using Generalised Heterogeneity Statistics: Should We Use Unequal Tails? BMC Med Res Methodol [Internet]. 2016 Sep 7;16(1).

41. Huedo-Medina TB, Sánchez-Meca J, Marín-Martínez F, Botella J. Assessing Heterogeneity in Meta-analysis: Q statistic or I 2 Index? Psychol Methods [Internet]. 2006;11(2):193–206.

42. Van Rijsbergen CJ. Information Retrieval. Butterworth-Heinemann; 1979.

43. Fisher RA. Statistical Methods for Research Workers. Edinburgh: Oliver and Boyd; 1925.

44. Cohen J. Eta-squared and Partial Eta-squared in Fixed Factor Anova Designs. Educ Psychol Meas [Internet]. 1973;33(1):107–12.

45. Jun M, Lix LM, Durand M, Dahl M, Paterson JM, Dormuth CR, et al. Comparative Safety of Direct Oral Anticoagulants and Warfarin in Venous Thromboembolism: Multicentre, Population based, Observational study. BMJ [Internet]. 2017;359.

46. Renoux C, Lix LM, Patenaude V, Bresee LC, Paterson JM, Lafrance JP, et al. Serotonin-norepinephrine Reuptake Inhibitors and the Risk of AKI: A Cohort Study of Eight Administrative Databases and Meta-analysis. Clin J Am Soc Nephrol [Internet]. 2015;10(10):1716–22.

47. van den Ham HA, Souverein PC, Klungel OH, Platt RW, Ernst P, Dell'Aniello S, et al. Major Bleeding in Users of Direct Oral Anticoagulants in Atrial Fibrillation: A Pooled Analysis of Results from Multiple Population-based Cohort Studies. Pharmacoepidemiol Drug Saf [Internet]. 2021 Oct 1;30(10):1339–52.

48. Renoux C, Dell'Aniello S, Khairy P, Marras C, Bugden S, Turin TC, et al. Ventricular Tachyarrhythmia and Sudden Cardiac Death with Domperidone Use in Parkinson's Disease. Br J Clin Pharmacol [Internet]. 2016 Aug 1;82(2).

49. Demir S, Doguyurt MF. A Comparison of Fixed and Random Effect Models by the Number of Research in the Meta-Analysis Studies with and without an Outlier. African Educ Res J [Internet]. 2022;10(3):277–90.

50. Petropoulou M, Mavridis D. A Comparison of 20 Heterogeneity Variance Estimators in Statistical Synthesis of Results from Studies: A Simulation Study. Stat Med [Internet]. 2017 Nov 30;36(27):4266–80.

51. Morton S, Murad M, O'Connor E, Lee C, Booth M, Vandermeer B, et al. Quantitative Synthesis—An Update. Methods Guide for Effectiveness and Comparative Effectiveness Reviews. 2018.

52. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews. Syst Rev [Internet]. 2021 Dec 1;10(1):1–11.

53. Seheult AH, Green PJ, Rousseeuw PJ, Leroy AM. Robust Regression and Outlier Detection. J R Stat Soc [Internet]. 1989;152(1):133.

54. Chatterjee S, Hadi AS. Influential Observations, High Leverage Points, and Outliers in Linear Regression. Stat Sci [Internet]. 1986 Apr 6;1(3):379–93.

55. Veroniki AA, Jackson D, Bender R, Kuss O, Langan D, Higgins JP, et al. Methods to Calculate Uncertainty in the Estimated Overall Effect Size from a Random-effects Meta-analysis. Res Synth Methods [Internet]. 2019 Mar 1;10(1):23–43.

56. Langan D, Higgins JPT, Jackson D, Bowden J, Veroniki AA, Kontopantelis E, et al. A Comparison of Heterogeneity Variance Estimators in Simulated Random-effects Meta-analyses. Res Synth Methods [Internet]. 2019 Mar 1;10(1):83–98.

57. Langan D, Higgins JPT, Simmonds M. An Empirical Comparison of Heterogeneity Variance Estimators in 12894 Meta-analyses. Res Synth Methods [Internet]. 2015 Jun 1;6(2):195–205.

58. Thorlund K, Wetterslev J, Awad T, Thabane L, Gluud C. Comparison of Statistical Inferences from the DerSimonian–Laird and Alternative Random-effects Model Meta-analyses – An Empirical Assessment of 920 Cochrane Primary Outcome Meta-

analyses. Res Synth Methods [Internet]. 2011 Dec;2(4):238–53.

59. Schmidt FL, Hunter JE. Methods of Meta-Analysis: Correcting Error and Bias in Research Findings. Methods of Meta-Analysis: Correcting Error and Bias in Research Findings. 2016.

60. Baker R, Jackson D. A New Approach to Outliers in Meta-Analysis. Health Care Manag Sci [Internet]. 2008 Jun;11(2):121–31.

Appendix A – Simulation conditions in the absence of outliers

Table A1. List of simulation conditions in the absence of outliers when $k = 10$.

Simulation condition	Simulation design characteristic	Simulation parameter		
	Number of sites (k)	Within-site variance (σ^2)	Mean EE for k sites (μ_0)	Between-site variance (τ^2)
1	10	0.05	0.5	0.05
2	10	0.5	0.5	0.05
3	10	1	0.5	0.05
4	10	0.05	1	0.05
5	10	0.5	1	0.05
6	10	1	1	0.05
7	10	0.05	1.5	0.05
8	10	0.5	1.5	0.05
9	10	1	1.5	0.05
10	10	0.05	0.5	0.5
11	10	0.5	0.5	0.5
12	10	1	0.5	0.5
13	10	0.05	1	0.5
14	10	0.5	1	0.5
15	10	1	1	0.5
16	10	0.05	1.5	0.5
17	10	0.5	1.5	0.5
18	10	1	1.5	0.5
19	10	0.05	0.5	1
20	10	0.5	0.5	1
21	10	1	0.5	1
22	10	0.05	1	1
23	10	0.5	1	1
24	10	1	1	1
25	10	0.05	1.5	1
26	10	0.5	1.5	1
27	10	1	1.5	1

Table A2. List of simulation conditions in the absence of outliers when k = 20.

Simulation condition	Simulation design characteristic	Simulation parameter		
	Number of sites (k)	Within-site variance (σ^2)	Mean EE for k sites (μ_0)	Between-site variance (τ^2)
1	20	0.05	0.5	0.05
2	20	0.5	0.5	0.05
3	20	1	0.5	0.05
4	20	0.05	1	0.05
5	20	0.5	1	0.05
6	20	1	1	0.05
7	20	0.05	1.5	0.05
8	20	0.5	1.5	0.05
9	20	1	1.5	0.05
10	20	0.05	0.5	0.5
11	20	0.5	0.5	0.5
12	20	1	0.5	0.5
13	20	0.05	1	0.5
14	20	0.5	1	0.5
15	20	1	1	0.5
16	20	0.05	1.5	0.5
17	20	0.5	1.5	0.5
18	20	1	1.5	0.5
19	20	0.05	0.5	1
20	20	0.5	0.5	1
21	20	1	0.5	1
22	20	0.05	1	1
23	20	0.5	1	1
24	20	1	1	1
25	20	0.05	1.5	1
26	20	0.5	1.5	1
27	20	1	1.5	1

Table A3. List of simulation conditions in the absence of outliers when k = 30.

Simulation condition	Simulation design characteristic	Simulation parameter		
	Number of sites (k)	Within-site variance (σ^2)	Mean EE for k sites (μ_0)	Between-site variance (τ^2)
1	30	0.05	0.5	0.05
2	30	0.5	0.5	0.05
3	30	1	0.5	0.05
4	30	0.05	1	0.05
5	30	0.5	1	0.05
6	30	1	1	0.05
7	30	0.05	1.5	0.05
8	30	0.5	1.5	0.05
9	30	1	1.5	0.05
10	30	0.05	0.5	0.5
11	30	0.5	0.5	0.5
12	30	1	0.5	0.5
13	30	0.05	1	0.5
14	30	0.5	1	0.5
15	30	1	1	0.5
16	30	0.05	1.5	0.5
17	30	0.5	1.5	0.5
18	30	1	1.5	0.5
19	30	0.05	0.5	1
20	30	0.5	0.5	1
21	30	1	0.5	1
22	30	0.05	1	1
23	30	0.5	1	1
24	30	1	1	1
25	30	0.05	1.5	1
26	30	0.5	1.5	1
27	30	1	1.5	1

Appendix B – Distribution of studentized residual estimates in the absence of outliers

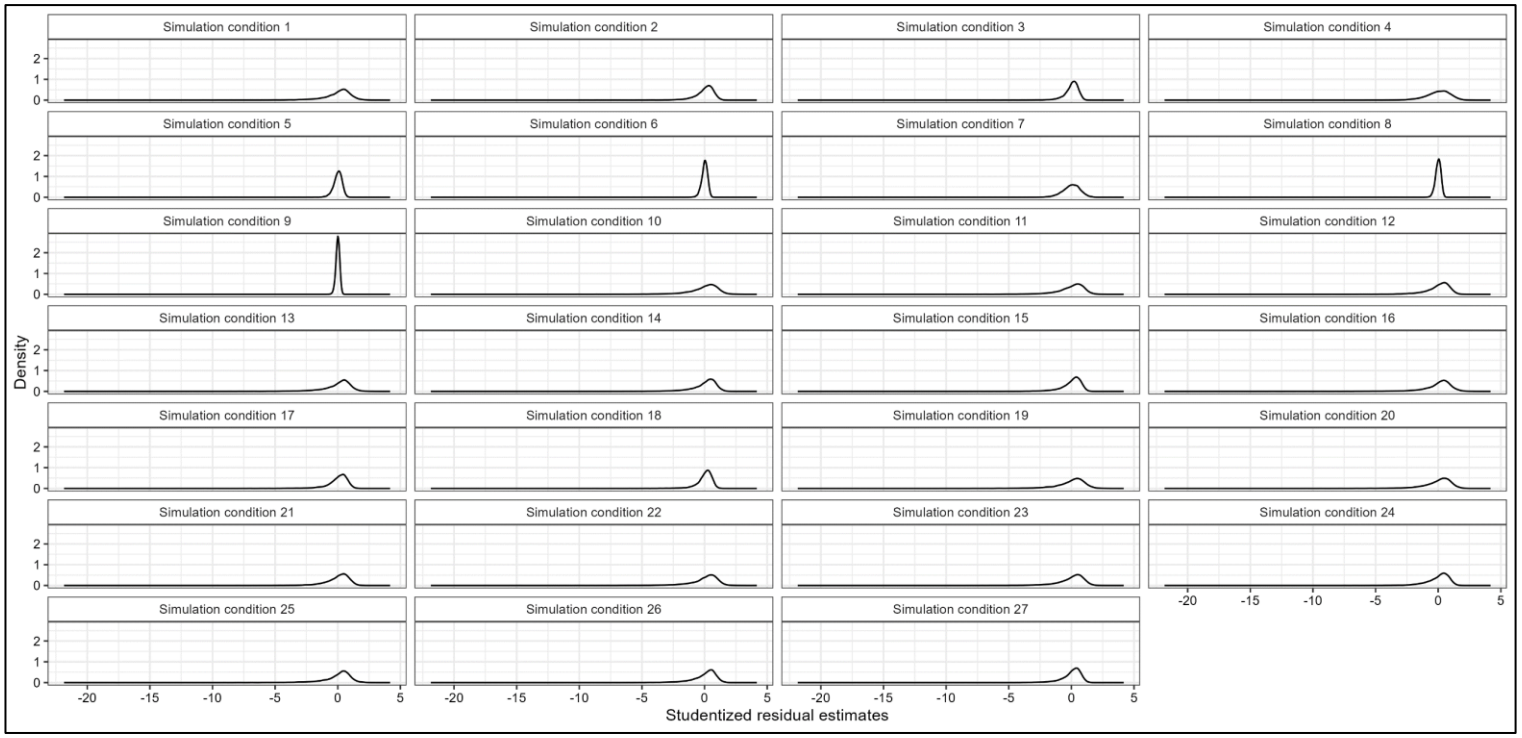

Figure B1. Distribution of studentized residual estimates in the absence of outliers for k = 10.

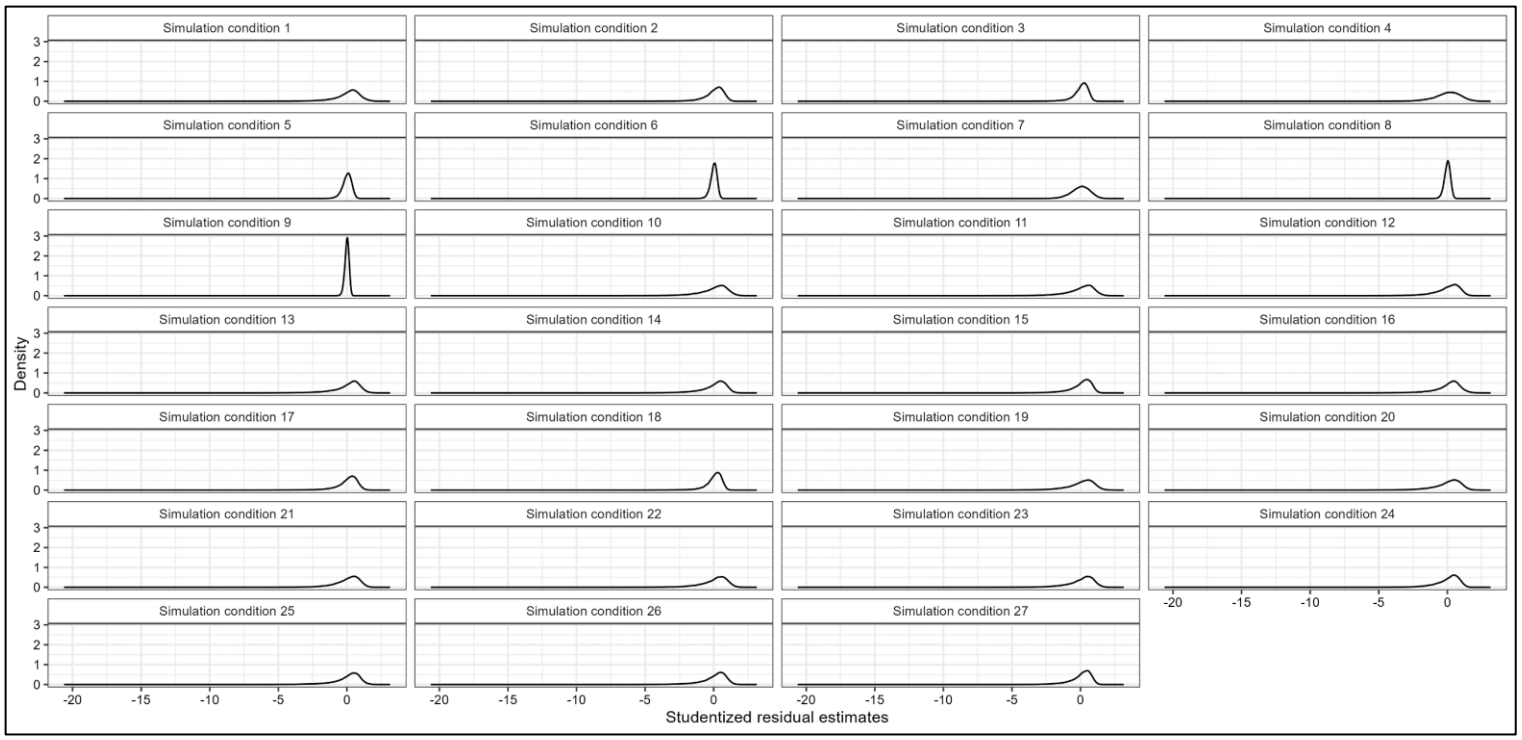

Figure B2. Distribution of studentized residual estimates in the absence of outliers for k = 20.

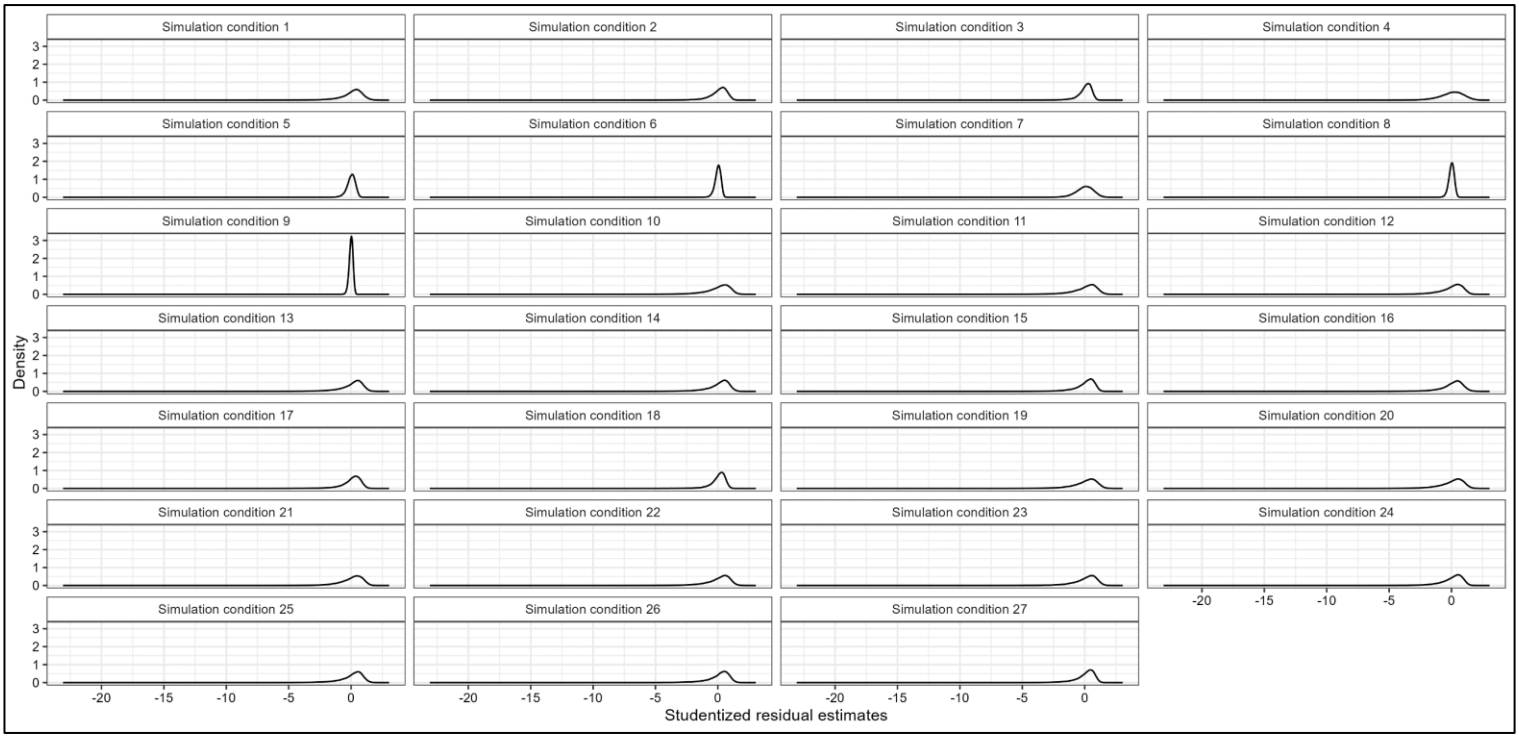

Figure B3. Distribution of studentized residual estimates in the absence of outliers for $k = 30$.

Appendix C – Performance metrics by simulation design characteristics and parameters for each method

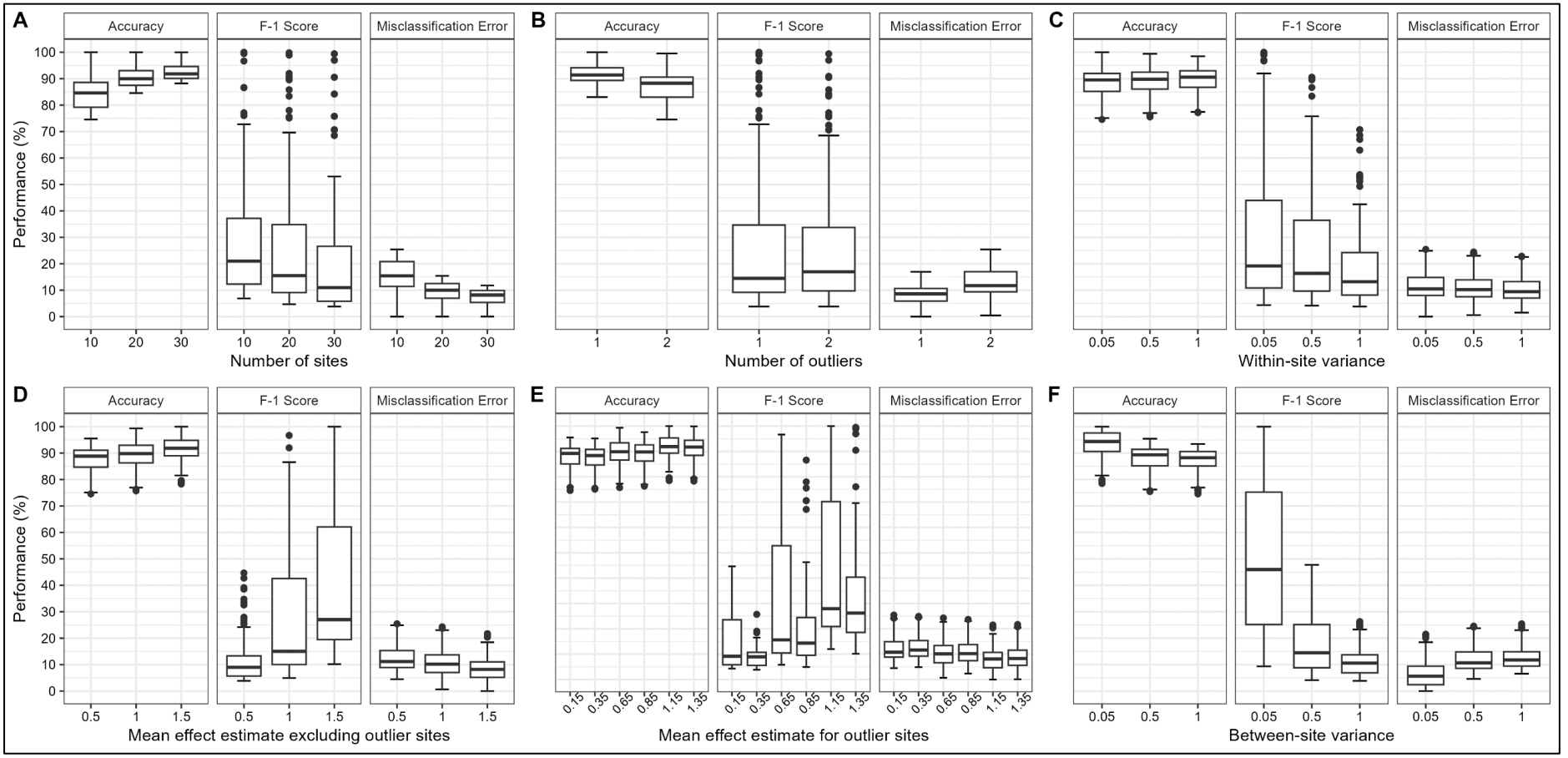

Figure C1. Summary of performance metrics for studentized residual estimates.

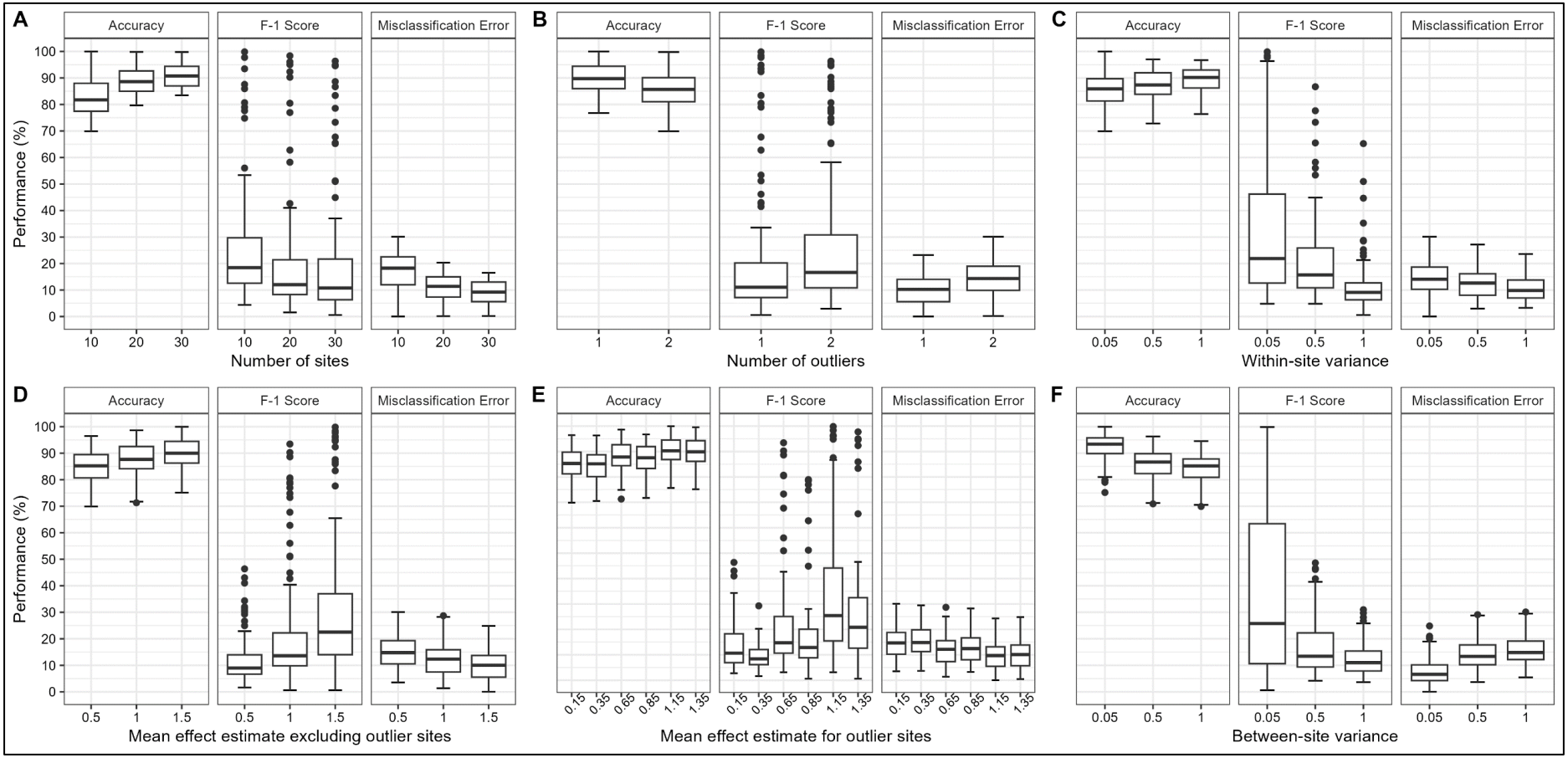

Figure C2. Summary of performance metrics for RCPEV.

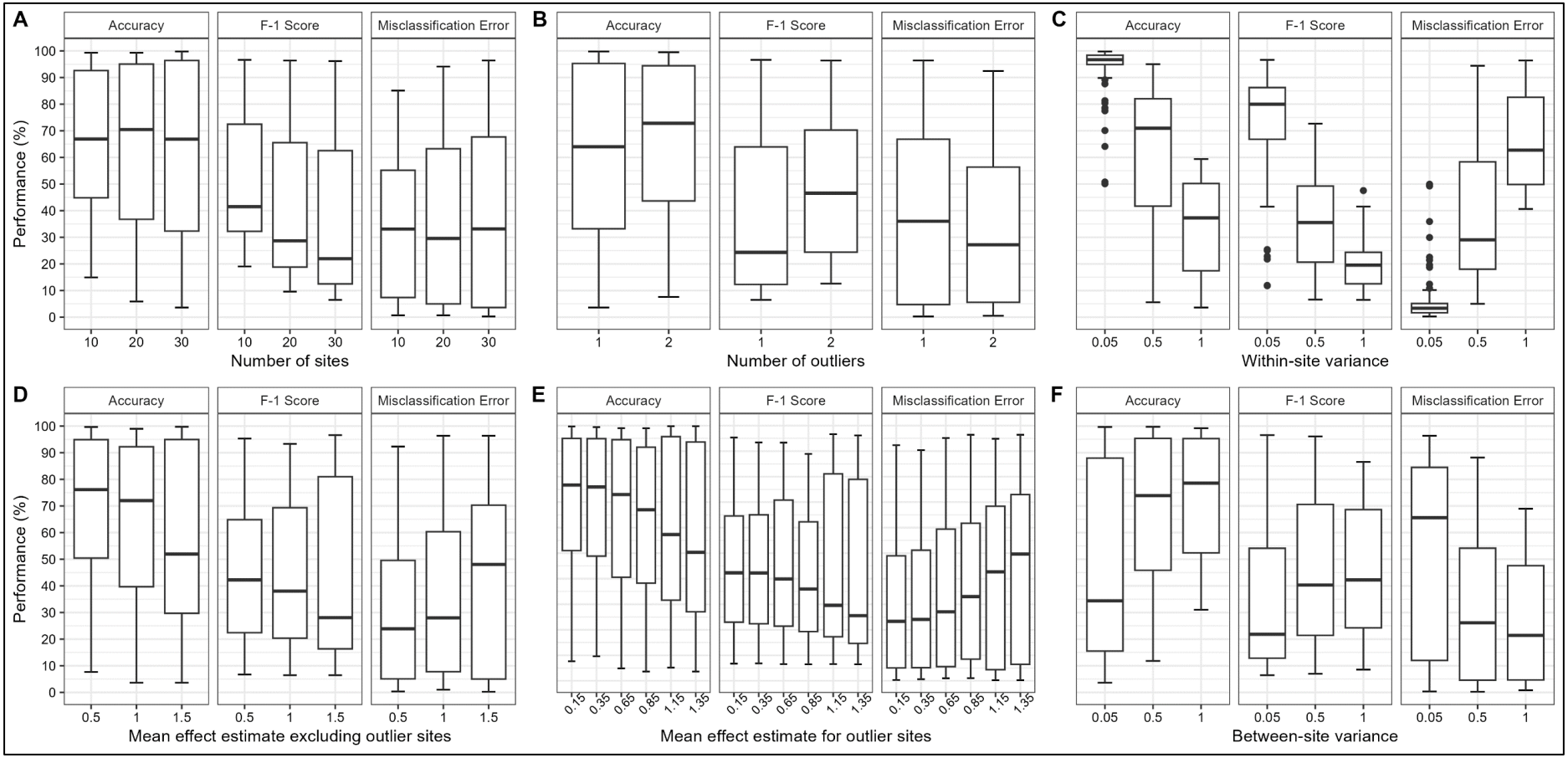

Figure C3. Summary of performance metrics for RCEBV.

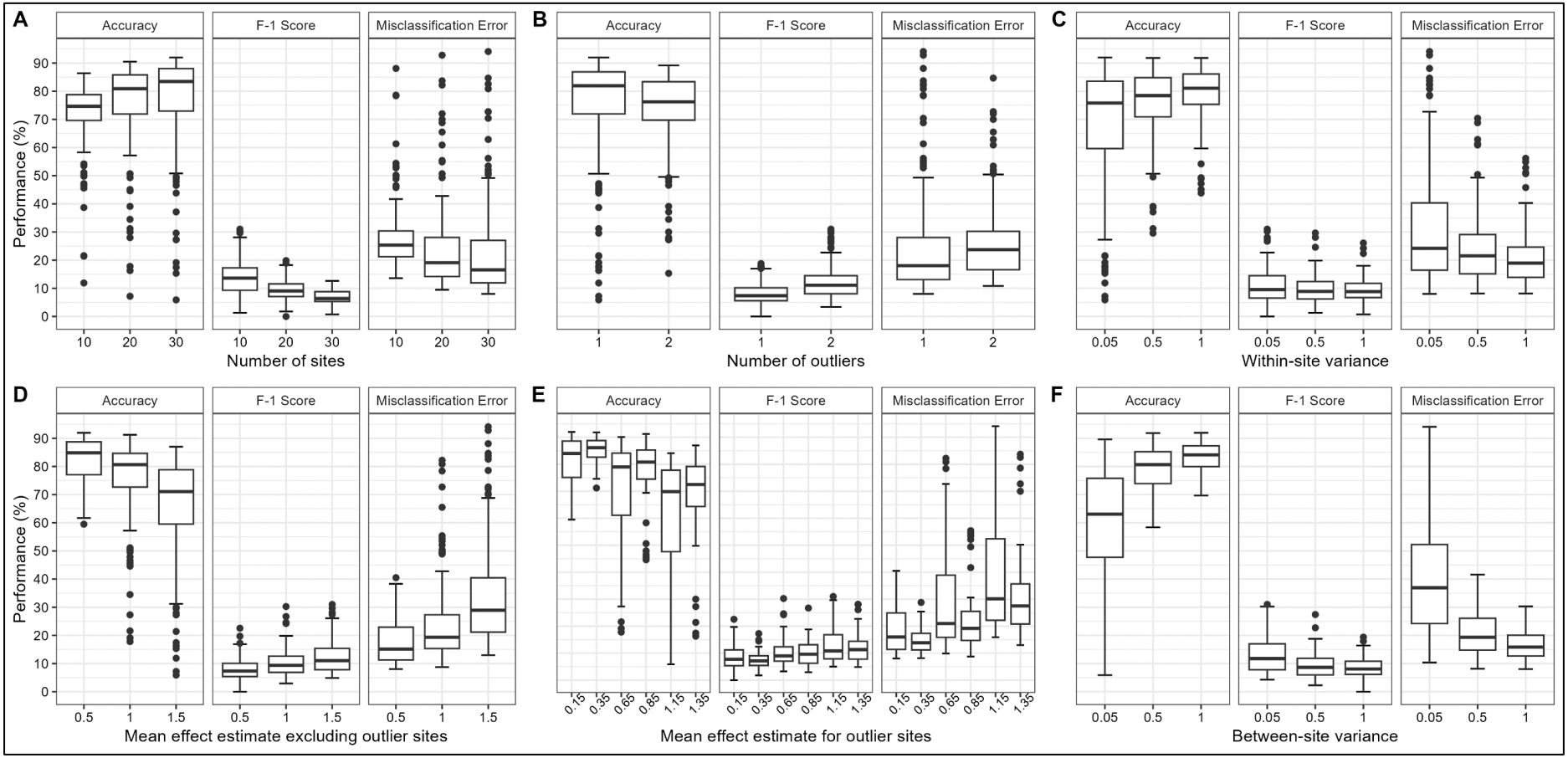

Figure C4. Summary of performance metrics for MMS method.

Appendix D – Distribution of random-effects residuals for random-effects analysis of variance models

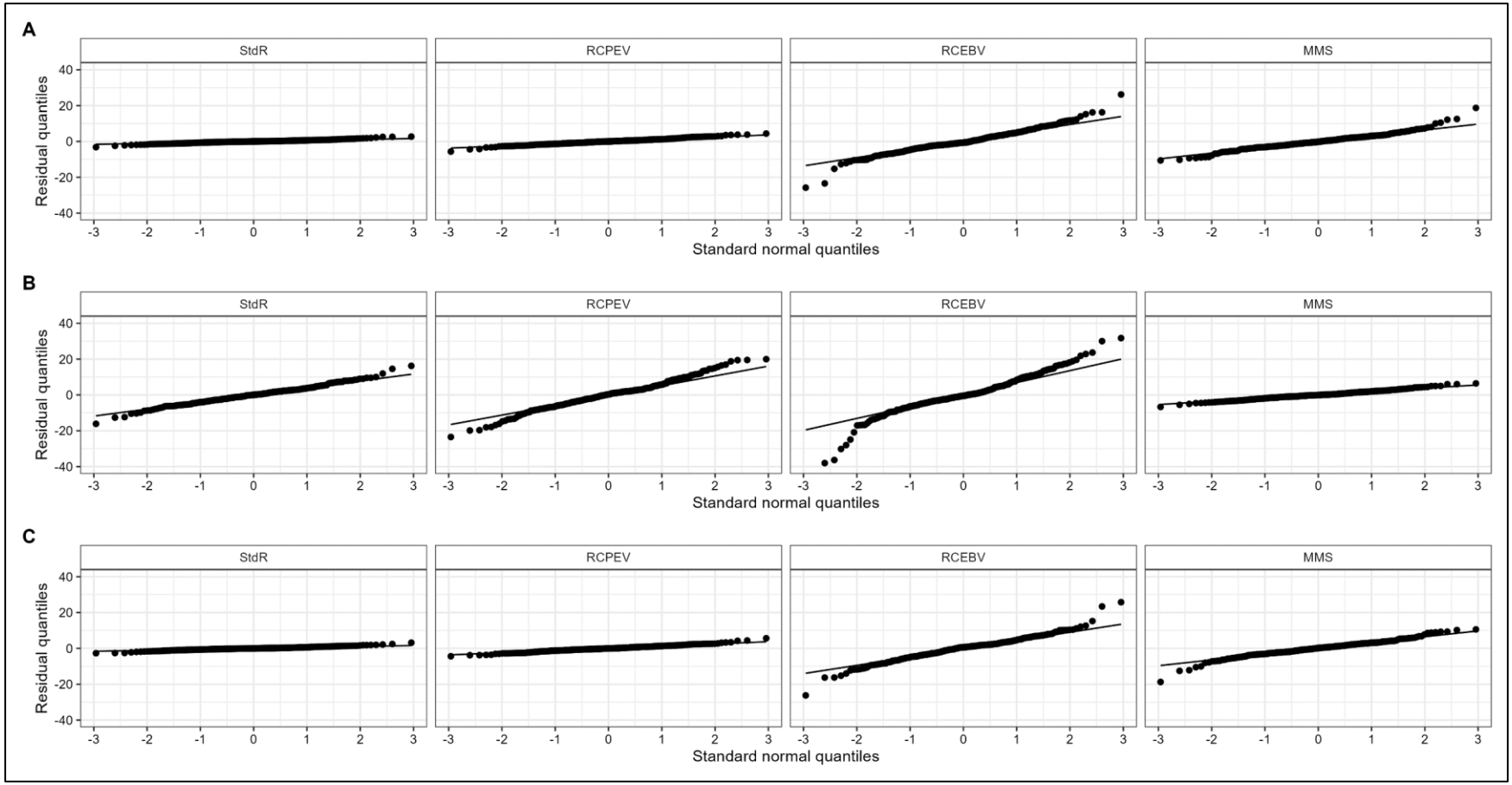

Figure D1. Distribution of random-effects residuals for random-effects analysis of variance models (Panel A: Accuracy, Panel B: F-1 score, Panel C: Misclassification error).

Appendix E – Homogeneity of variance in performance metrics for random-effects analysis of variance models

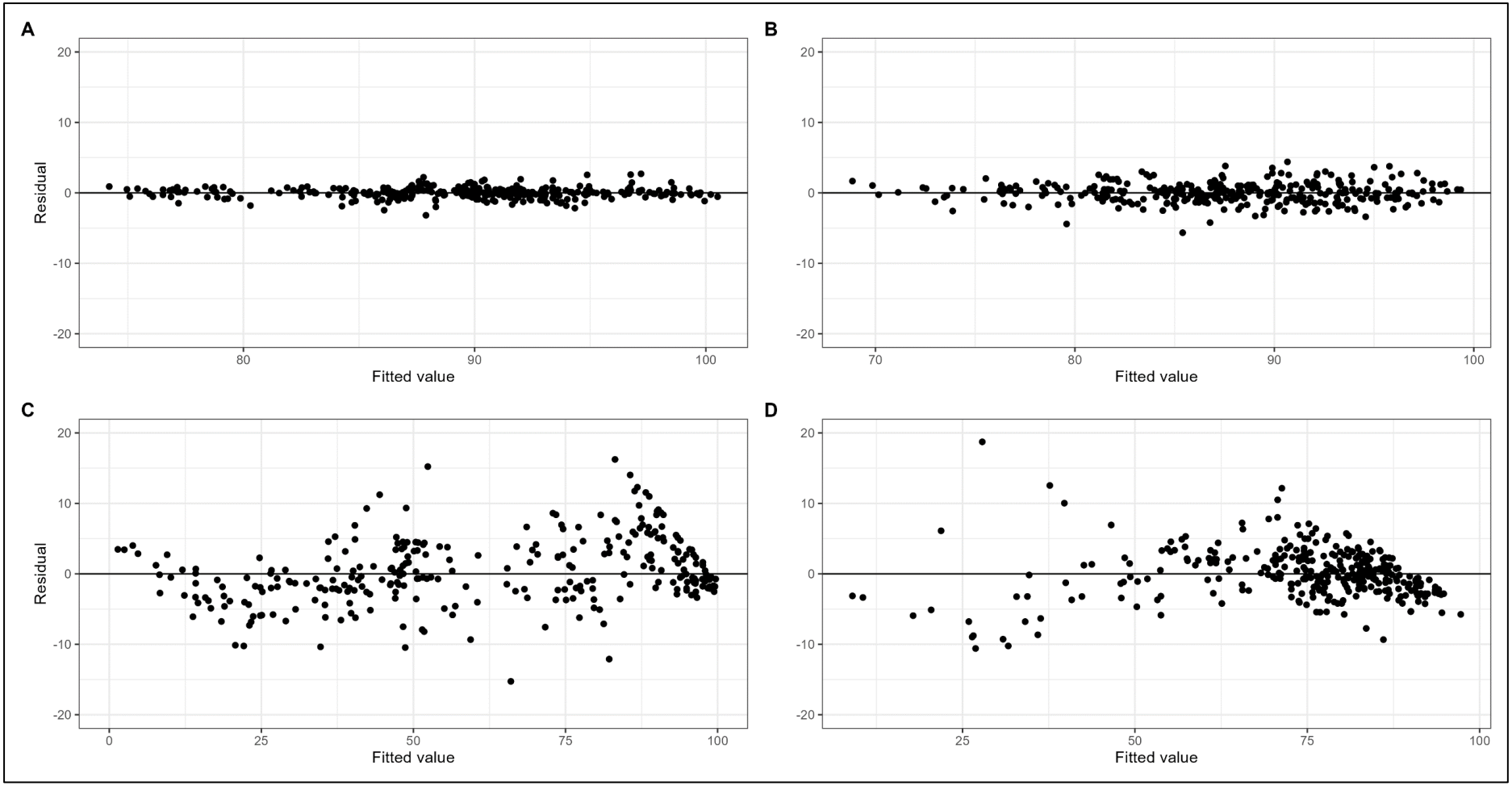

Figure E1. Residual versus fitted value plots to assess homogeneity of variance in accuracy values for random-effects analysis of variance models (Panel A: StdR, Panel B: RCPEV, Panel C: RCEBV, Panel D: MMS).

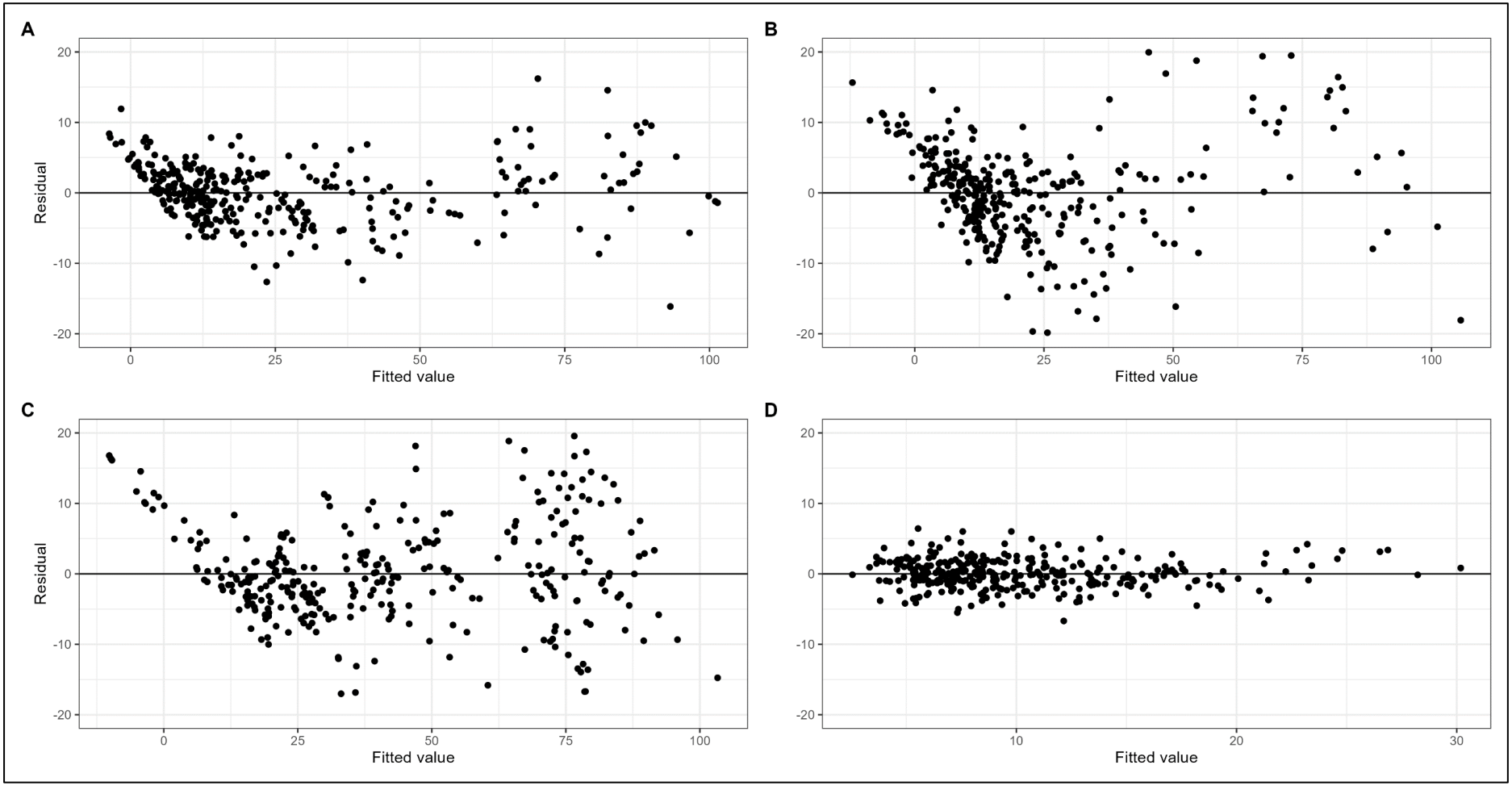

Figure E2. Residual versus fitted value plots to assess homogeneity of variance in F-1 score values for random-effects analysis of variance models (Panel A: StdR, Panel B: RCPEV, Panel C: RCEBV, Panel D: MMS).

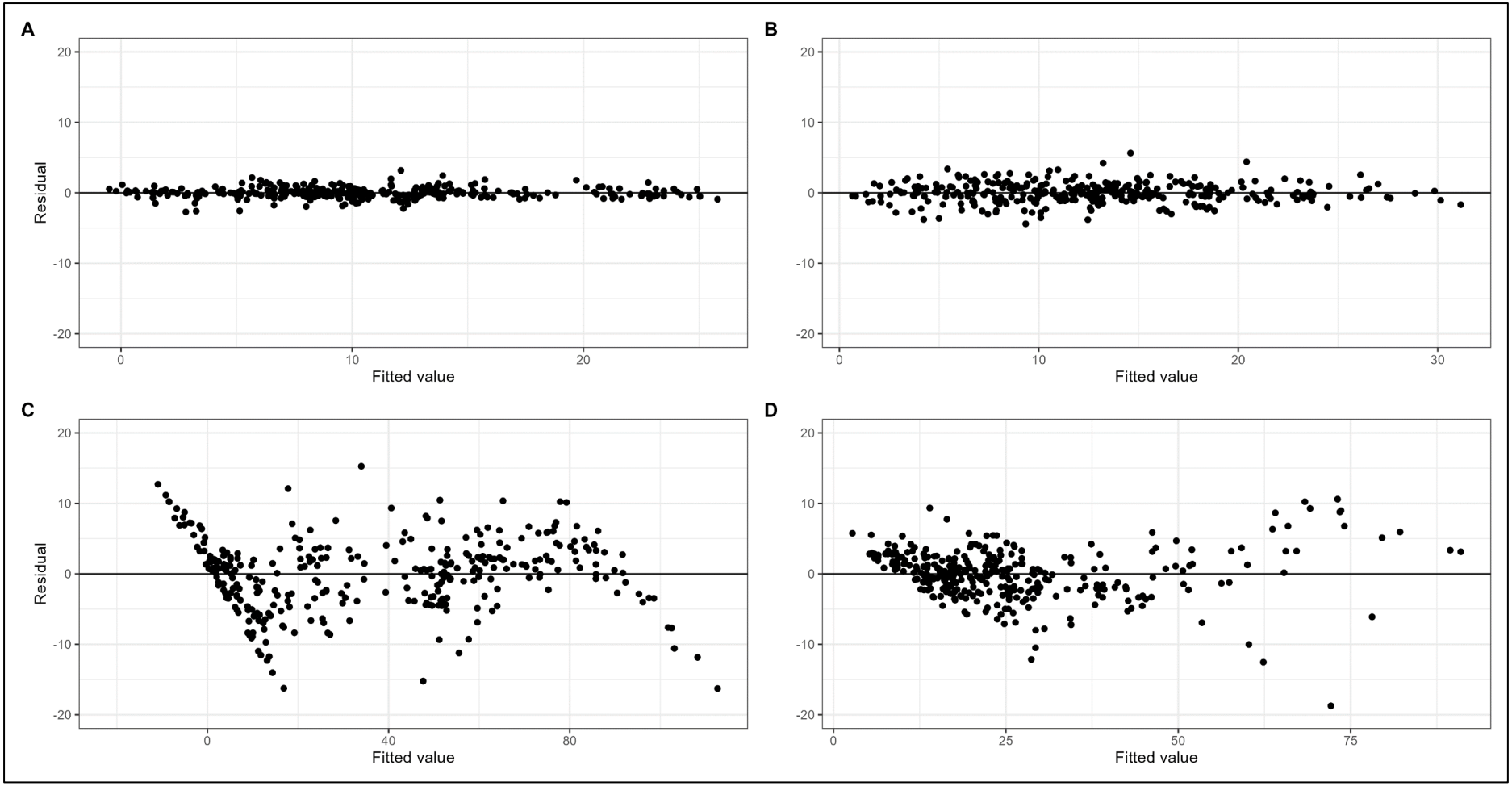

Figure E3. Residual versus fitted value plots to assess homogeneity of variance in misclassification error values for random-effects analysis of variance models (Panel A: StdR, Panel B: RCPEV, Panel C: RCEBV, Panel D: MMS).

Appendix F – Selected meta-analyses conducted by CNODES investigators for application of outlier detection methods

Table F1. Data sources for selected meta-analyses conducted by CNODES investigators for application of outlier detection methods (k = 39).

Peer-reviewed publication	Publication year	Meta-analysis ID	Data source
Azoulay et al. (a)	2016	1	Supplementary Figure 3
		2	Supplementary Figure 5
		3	Supplementary Figure 7
		4	Supplementary Figure 8
		5	Supplementary Figure 12
		6	Supplementary Figure 19
Azoulay et al. (b)	2016	7	Figure 2
Filion et al.	2016	8	Figure 2 (Subgroup: History of heart failure)
		9	Supplementary Figure 6
		10	Supplementary Figure 10
		11	Supplementary Figure 12
		12	Supplementary Figure 15
		13	Supplementary Figure 23
		14	Supplementary Figure 24
		15	Supplementary Figure 29
		16	Supplementary Figure 30
		17	Supplementary Figure 31
		18	Supplementary Figure 33
		19	Supplementary Figure 35
Renoux et al.	2016	20	Figure 3 (Subgroup: Use of 30 days or less)
		21	Figure 6 (Subgroup: Without QT drugs)
Durand et al.	2020	22	Figure 2 (Outcome: Major bleeding)
		23	Figure 2 (Outcome: Composite)
Fisher et al.	2020	24	Figure 2
Filion et al.	2020	25	Figure 3
		26	Figure 5
		27	Supplementary Figure 1
		28	Supplementary Figure 4
Dormuth et al.	2021	29	Figure A (Outcome: Fetal death)
		30	Figure A (Outcome: Spontaneous abortions)
		31	Figure B
Durand et al.	2021	32	Figure 2 (Outcome: Major bleeding)
		33	Figure 2 (Outcome: Composite)
		34	Figure 3 (Outcome: Composite)
van den Ham et al.	2021	35	Figure 3.1
		36	Figure 3.2
		37	Figure 4.1
		38	Figure 4.2
		39	Figure 4.4

Appendix G – Computer program for simulation study

The following is the R program code for the following simulation conditions: k = 10, σ^2 = 0.05, μ_0 = 0.50, μ_1 = 0.35, τ^2 = 0.05 for 10 simulation iterations. The source code for the MMS method can be found here.

```
####################LOADING ALL REQUIRED LIBARIES#####################
library(metafor)
library(doParallel)
library(doSNOW)
library(meta)

####################SETTING WORKING DIRECTORY#####################
setwd("") #Please set the working directory accordingly.

#########MAKING CLUSTERS OF CORES FOR PARALLEL COMPUTING############
cores <- detectCores() - 1; #cores   #Detecting cores in the computer
cluster <- makeCluster(cores)       #Making cluster of cores
registerDoSNOW(cluster)             #Registering the cores

###################5 SIMULATION ITERATIONS#####################
#Setting the seed for reproducibility
random_seed <- sample.int(n=100000)
seed <- random_seed[1]
set.seed(seed)

#Creating a DATAFRAME containing all combination of simulation conditions.
combo_list <- expand.grid(k = 10,
                          s_all = 0.05,
                          m_remain = 0.50,
                          m_out = 0.35,
                          t = sqrt(0.05),
                          stringsAsFactors = FALSE); #combo_list

sim_n <- 10                    #Setting number of simulation iterations.
site <- rep(1:combo_list$k[1])    #Creating site for each simulation iteration.
rep_id <- rep(1:sim_n, each = length(site)) #Creating replication id for simulation iterations.

time0 <- Sys.time()        #To record the run time.

source("sourcecode_LRT.r")  #Running source code for model-based LRT statistic using a
mean-shift method.

#####################################################################
###################INDEX VARIABLES TO STORE VALUES###################
#####################################################################
```

```
st_res <- pooled_ee <- bet_var <- LRT <- true_class <- NULL                #Estimated values
from outlier detection methods.
st_res.est <- pooled_ee.est <- bet_var.est <- LRT_stat <- LRT_stat.est <- LRT_boot_95.est <-
LRT_boot_p.est <- NULL #Index variables to store estimated values from outlier detection
methods.
y <- log_y <- NULL                                                          #Index variables to store EEs (y)
and log scale EEs (log_y).
v <- k <- NULL                                                              #Index variables to store within-
site variance (v) and total number of sites (k).
k_value <- m_remain_value <- m_out_value <- s_all_value <- t_value <- NULL              #Index
variables to store simulation parameter values in EXCEL.



for (j in 1:nrow(combo_list)){  #"for" loop will run through by the row of "combo_list".

  for (sim in 1:sim_n){       #"for" loop will run through "sim_n".

    for (i in 1:combo_list$k[j]){  #"for" loop will run through total number of sites (k).

      #Storing k, m_remain, m_out, s_all, t values by "combo_list" over "sim_n".
      k_value <- c(k_value, combo_list$k[j])                    #Total number of sites.
      m_remain_value <- c(m_remain_value, combo_list$m_remain[j])  #Mean EE for k-m sites.
      m_out_value <- c(m_out_value, combo_list$m_out[j])          #Mean EE for m sites.
      s_all_value <- c(s_all_value, combo_list$s_all[j])          #Within-site variance.
      t_value <- c(t_value, ((combo_list$t)^2)[j])                #Between-site variance (NOTE:
STORING THE SQUARED VALUE, INDICATING TAU-SQUARED).

      #Randomly sampling EEs from NORMAL DISTRIBUTION.
      y[i] <- abs(rnorm(1, mean = combo_list$m_remain[j], sd = combo_list$t[j]))  #NOTE:
TOOK THE ABSOLUTE VALUE OF y[i] BECAUSE EEs CANNOT BE NEGATIVE.
      log_y[i] <- log(y[i])                                 #Converting y[i] into "log scale" because
"metafor" uses log scale of y[i] as input.
      v[i] <- combo_list$s_all[j]                                #Within-site variance.
    }

    #Randomly sampling EEs that are OUTLIERS from NORMAL DISTRIBUTION
    #and replace that value in the FIRST POSITION of y[i].
    y[1] <- abs(rnorm(1, mean = combo_list$m_out[j], sd = combo_list$t[j]))   #NOTE: TOOK
THE ABSOLUTE VALUE OF y[i] BECAUSE EEs CANNOT BE NEGATIVE.
    log_y[1] <- log(y[1])                                   #Converting y[1] into log scale because
"metafor" uses log scale of y as input.
    v[1] <- combo_list$s_all[j]                                #Within-site variance.

    #Outlier detection method using Viechtbauer and Cheung paper
    #using random-effects meta-analysis model (metafor package)
    sim_metafor = rma(yi = log_y, vi = v, method = "DL")   #Performing random-effects meta-
analysis model in "metafor".
```

```
   #Applying outlier detection methods using "metafor" package
   inf_metafor <- influence(sim_metafor)

   ###METHOD 1: STUDENTIZED RESIDUAL ESTIMATES
   st_res <- round(inf_metafor$inf$rstudent,2)   #Estimated values.
   st_res.est <- c(st_res.est, st_res)          #Storing the estimated values.

   ###METHOD 2: RELATIVE CHANGE IN THE POOLED EE VARIANCE
   pooled_ee <- round(inf_metafor$inf$cov.r,2)       #Estimated values.
   pooled_ee.est <- c(pooled_ee.est, pooled_ee)      #Storing the estimated values.

   ###METHOD 3: RELATIVE CHANGE IN THE ESTIMATED BETWEEN-SITE
VARIANCE
   bet_var <- round(inf_metafor$inf$tau2.del,2)      #Estimated values.
   bet_var.est <- c(bet_var.est, bet_var)               #Storing the estimated values.

   ###METHOD 4: MODEL-BASED LRT LRT STATISTIC USING A MEAN-SHIFT
METHOD ("meta" package)
   sim_meta <- metagen(TE = log_y, seTE = sqrt(v), backtransf = TRUE, method.tau = "DL")
   LRT <- PBS_LR(y = log_y, v = sqrt(v), B = 10)      #B = Number of bootstrap resamples.

   LRT_stat <- rbind(LRT_stat, LRT[LRT$X1.n == 1,])    #LR Estimates from the model-based
method.
   LRT_stat.est <- round(LRT_stat$LR, 2)             #Storing estimated LRT statistic values.
   LRT_boot_95.est <- round(LRT_stat$Q95, 2)         #Storing estimated bootstrap 95th
percentile of LRT statistic values.
   LRT_boot_p.est <- round(LRT_stat$P, 2)            #Storing estimated bootstrap p-value.
  }

  #Defining outliers based on threshold from each method
  st_res.count <- ifelse(st_res.est > 1.96 | st_res.est < -1.96, 1, 2)  #1 = Outlier, 2 = Not an outlier
  pooled_ee.count <- ifelse(pooled_ee.est < 1.00, 1, 2)              #1 = Outlier, 2 = Not an outlier
  beta_var.count <- ifelse(bet_var.est < 1.00, 1, 2)                #1 = Outlier, 2 = Not an outlier
  LRT.count <- ifelse(LRT_boot_p.est < 0.05, 1, 2)                  #1 = Outlier, 2 = Not an outlier
  true_class <- ifelse(y == y[1], 1, 2)                             #1 = True outlier, 2 = Not a true outlier

  test = data.frame(rep_id, site, k_value, m_remain_value, m_out_value, s_all_value, t_value,
true_class, st_res.est, st_res.count, pooled_ee.est, pooled_ee.count, bet_var.est, beta_var.count,
LRT_stat.est, LRT_boot_95.est, LRT_boot_p.est, LRT.count)
  write.csv(test, file="1000_10_1_test.csv", row.names = F, quote = T)
}

stopCluster(cluster)
time1 <- Sys.time() - time0; time1
```