Computational Statistics
https://doi.org/10.1007/s00180-024-01530-0

ORIGINAL PAPER

Robust confidence intervals for meta-regression with interaction effects

Maria Thurow[1,2] · Thilo Welz[1] · Eric Knop[1] · Tim Friede[3] · Markus Pauly[1,2]

Received: 15 February 2023 / Accepted: 4 July 2024

Abstract

Meta-analysis is an important statistical technique for synthesizing the results of multiple studies regarding the same or closely related research question. So-called meta-regression extends meta-analysis models by accounting for study-level covariates. Mixed-effects meta-regression models provide a powerful tool for evidence synthesis, by appropriately accounting for between-study heterogeneity. In fact, modelling the study effect in terms of random effects and moderators not only allows to examine the impact of the moderators, but often leads to more accurate estimates of the involved parameters. Nevertheless, due to the often small number of studies on a specific research topic, interactions are often neglected in meta-regression. In this work we consider the research questions (i) how moderator interactions influence inference in mixed-effects meta-regression models and (ii) whether some inference methods are more reliable than others. Here we review robust methods for confidence intervals in meta-regression models including interaction effects. These methods are based on the application of robust sandwich estimators of Hartung-Knapp-Sidik-Jonkman (**HKSJ**) or heteroscedasticity-consistent (**HC**)-type for estimating the variance-covariance matrix of the vector of model coefficients. Furthermore, we compare different versions of these robust estimators in an extensive simulation study. We thereby investigate coverage and width of seven different confidence intervals under varying conditions. Our simulation study shows that the coverage rates as well as the interval widths of the parameter estimates are only slightly affected by adjustment of the parameters. It also turned out that using the Satterthwaite approximation for the degrees of freedom seems to be advantageous for accurate coverage rates. In addition, different to previous analyses for simpler models, the **HKSJ**-estimator shows a worse performance in this more complex setting compared to some of the **HC**-estimators.

Maria Thurow, Thilo Welz and Eric Knop have contributed equally to this work.

Extended author information available on the last page of the article

Published online: 02 August 2024

Keywords Confidence intervals · Meta-analysis · Random effects · Robust covariance estimation · Regression · Interactions

1 Introduction

Meta-analysis is a statistical technique that combines the results of multiple studies to arrive at a single, more precise estimate of the effect size of a particular intervention or treatment. It aims to provide a comprehensive and quantitative summary of the available evidence on a particular topic, taking into account the heterogeneity of the studies and the sample sizes. By pooling the data from multiple studies, meta-analysis can increase the statistical power and accuracy of the results, and provide a more robust understanding of the effects of an intervention. This statistical technique is routinely applied in different areas of research, such as biology, medicine or psychology. In meta-regression, study-level covariates or moderators, which may influence the observed outcome in the respective study, are accounted for. A meta-regression combines the advantages of a linear regression model and a meta-analysis. On the one hand information from different studies is taken into account. On the other hand one is able to test not only for an overall effect, which is the case for most meta analyses, but also on effects of relevant study characteristics. The characteristics are used as study-level covariates and often called moderators. In contrast to a usual regression model, the mixed-effects model assumes that the estimated treatment effect is influenced by two different types of uncertainty: First, the estimated effect of a single study is assumed to be different from the studies' true effect by a random error. Second, the analyzed studies are assumed to have different true effects caused by differences between the studies, the so called between-study heterogeneity. Therefore, the treatment effects of the studies differ from the treatment effect for the entire population. It is important to account for this additional variation when confidence intervals (CIs) of moderators are calculated (Raudenbush 2009).

A simulation study by Viechtbauer et al. (2015) showed that the type I error rate could be adequately controlled by using the Hartung-Knapp-Sidik-Jonkmann (**HKSJ**) method and permutation tests. Due to their computational extensiveness and the focus on CIs in this study, permutation tests are not considered here. However, Viechtbauer et al. (2015) also showed that when heterogeneity is present, the choice of estimator for the covariance of the vector of model coefficients has a large impact on test results. More specifically, in a model with only one moderator large differences in Type 1 error rates and the power of *t*-type tests were determined. Amongst others, tests based on a heteroscedasticity consistent (**HC**) estimate of the covariance matrix introduced by White (1980) and a modified covariance matrix estimate (**HKSJ**) introduced by Knapp and Hartung (2003) (and also Sidik and Jonkman (2005b)) were considered. The **HC** estimate is an established approach in econometrics, but not commonly applied in meta-analysis, in particular when used in medical research. Because of its structure, it is also known as a sandwich estimator and is used for robust inference. The **HKSJ** estimate is common in meta-analyses applied in medicine. It performed well in a meta-analytic context in previous research (Viechtbauer et al. 2015; Welz and Pauly 2020; Welz et al. 2022; Sidik and Jonkman 2005a, 2006). In Viechtbauer et al. (2015) simulation study, tests based on **HC** estimators ($\mathbf{HC}_0$ and $\mathbf{HC}_1$) turned out to be too liberal. In contrast, the test based on the **HKSJ** estimate performed the best among all considered tests. Since their results are limited to special settings, Viechtbauer et al. (2015) suggested additional future simulation studies that consider, e.g., non-normal random effects, multiple covariates with multicollinearity and coverage probability of coefficients' CIs. Welz and Pauly (2020) extended their research by comparing tests on the significance of the moderator based on six different versions of White's covariance matrix estimator and the Hartung-Knapp-Sidik-Jonkman variance-covariance matrix estimator for several random effect distributions. The six heteroscedasticity consistent covariance estimators are known as $\mathbf{HC}_0, \dots, \mathbf{HC}_5$. The main difference between these different versions is how they transform the model residuals by discounting the observations' leverages (Cribari-Neto et al. 2007; Welz and Pauly 2020). In regression, leverage is a measure for how far away the covariate values of an observation are from those of the other observations. The newer **HC** estimators discount the leverages more strongly than earlier version. In a simulation study Welz and Pauly (2020) also found the **HKSJ** based tests to perform the best compared to the **HC** estimators. Amongst the **HC** estimators the $\mathbf{HC}_3 - \mathbf{HC}_5$ based tests controlled the nominal significance level well and had power close to the **HKSJ** based tests for larger number of studies. The distribution of the random effect turned out to have almost no effect on the results (Welz and Pauly 2020). The **HKSJ** was already compared to the $\mathbf{HC}_2$ in simulation studies, e.g. by Sidik and Jonkman (2005a, 2006). The results were in accordance with the findings of Welz and Pauly (2020).

In a recent meta-analysis, including meta-regression analyses, Kimmoun et al. (2021) analyzed mortality and readmission to hospital after acute heart failure. They found a statistically significant decline of death rates over calendar time. However, the median year of recruitment was correlated with the average age of the patients. This suggests that the observed trend might be explained by a neglected interaction of those variables. In fact, Knop et al. (2023) showed in a re-analysis of the above mentioned data that it is vitally important to account for confounding and interaction effects, when making inference based on meta-regression with multiple moderators. Note that the importance of investigating interactions in meta-regression was already acknowledged by Li et al. (2017).

Motivated by this meta-analysis, the current paper extends the research of Welz and Pauly (2020) in two directions. Firstly, two moderators and, based on the important findings in Knop et al. (2023), their interaction term are modelled. Modelling interactions is required in situations where not only the influence of a moderator itself is of interest but its influence in the presence of other factors. Interactions are also helpful to assess the circumstances under which the influences of certain moderators on the estimated effect size are stronger or weaker (Aiken et al. 1991). Although modelling interaction terms is useful in providing additional insights, they are often neglected in meta-regression. However, neglecting existing interactions may dramatically alter conclusions drawn from quantitative research synthesis, as seen in a recent data analysis from acute heart failure research (Knop et al. 2023). Secondly, CIs are considered instead of hypothesis tests, since CIs can be interpreted more easily, as they also state the involved estimation uncertainty. The equivalence theorem between statistical tests and CIs ensures that the results of this study have direct implications on the behaviour of corresponding two-sided t-tests. In addition, focusing on CIs is also in line with the ICH E9 guidelines that specifically propagate the use of confidence intervals ("*Estimates of treatment effects should be accompanied by confidence intervals, whenever possible [...]*", see Section 5.5. therein).

The methodological aim of this work is to determine the performance of confidence intervals based on the seven covariance estimators $\mathbf{HC}_0 - \mathbf{HC}_5$ and $\mathbf{HKSJ}$ in extensive simulations. On the one hand, it is investigated whether the confidence intervals of a single moderator's coefficient perform different in presence of an interaction. On the other hand, confidence intervals for the interaction coefficient itself are considered. For these more complex models it is of interest whether the estimators have the same properties as in the univariate model. Furthermore, we check how introducing non-normal distributions for the random effects influences results, similar to Welz and Pauly (2020).

In Sect. 2 we introduce the relevant methods, starting with the mixed-effects meta-regression model in Sect. 2.1, followed by weighted least squares (WLS) estimation in Sect. 2.2 and different estimators for the variance-covariance matrix of the estimated vector of coefficients in Sect. 2.3. In Sect. 3 we describe the design and results of our extensive simulation study and provide recommendations for practical applications. Finally, we close with a discussion and an outlook for future research in Sect. 4.

2 Statistical methods

2.1 The mixed-effects meta-regression model

The study characteristics which are used as covariates in the meta-regression model are called moderators and are denoted with $\boldsymbol{x}_j = (x_{j1}, \dots, x_{jk})'$, where k is the number of studies and $j \in \{0, 1, \dots, m\}$, with m as the number of moderators. Functions of other moderators such as interactions of the form $x_{3i} = x_{1i}x_{2i}$ could be moderators themselves. The true outcome of an individual study $i \in \{1, \dots, k\}$ is denoted with θ_i. The model equation for the true outcome of study i is

$$\theta_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_m x_{mi} + u_i. \tag{1}$$

The parameters $\beta_1, \dots, \beta_m$ are the regression coefficients of the associated moderators. We generally assume that the number of studies is greater than the number of study-level moderators, i.e. $k > m$. The deviation of the i^{th} studies' true outcome θ_i is modelled by the random effect u_i. The random effect u_i is usually assumed to be normally distributed with $u_i \sim \mathcal{N}(0, \tau^2)$. Furthermore, the observed outcome for study i is modelled as

$$y_i = \theta_i + \varepsilon_i, \tag{2}$$

with model errors $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$. The model errors ε_i and random effects u_i are assumed to be independent. Together this yields what is also known as a normal-normal hierarchical model (NNHM) (Friede et al. 2017). It is also possible to consider a more general semiparametric setting with the moment assumptions $\mathbb{E}(u_i) = 0$ and $\mathrm{Var}\,(u_i) = \tau^2$ without other distributional restrictions on the random effects, as in Welz and Pauly (2020). In matrix notation the model can be rewritten as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u} + \boldsymbol{\varepsilon}, \tag{3}$$

where

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_k \end{pmatrix} \in \mathbb{R}^k,\ \boldsymbol{X} = \begin{pmatrix} 1 & \dots & x_{1m} \\ \vdots & & \vdots \\ 1 & \dots & x_{km} \end{pmatrix} \in \mathbb{R}^{k \times (m+1)}, \tag{4}$$

$$\boldsymbol{u} = \begin{pmatrix} u_1 \\ \vdots \\ u_k \end{pmatrix} \in \mathbb{R}^k \text{ and } \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_k \end{pmatrix} \in \mathbb{R}^k. \tag{5}$$

The design matrix $\boldsymbol{X}$ is assumed to have full rank. Under the assumption that $\boldsymbol{u}$ and $\boldsymbol{\varepsilon}$ are independent, the variance-covariance matrix of $\boldsymbol{y}$ is $\mathrm{Var}\,(\boldsymbol{y}) = \boldsymbol{V} = \mathrm{diag}\,(\sigma_1^2 + \tau^2, \dots, \sigma_k^2 + \tau^2)$.

2.2 Weighted-least-squares estimation

The weighted least squares estimate for the model coefficients $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X}'\widehat{\boldsymbol{W}}\boldsymbol{y}, \tag{6}$$

with the weight matrix $\widehat{\boldsymbol{W}}$ typically (but not always) defined as the inverse variance matrix. For Model (3) it is given by $\widehat{\boldsymbol{W}} = \mathrm{diag}\left((\sigma_1^2 + \hat{\tau}^2)^{-1}, \dots, (\sigma_k^2 + \hat{\tau}^2)^{-1}\right)$. It should be noted that the sampling variances σ_i, $i = 1, \dots, k$ are assumed as known, although they are in fact estimated from the data. This is done for mathematical convenience and is common practice in meta-analysis (DerSimonian and Laird 1986). Various estimators are available for the between-study variance τ^2 (Veroniki et al. 2016). The recommendation for meta-analysis is to use either the restricted maximum likelihood (REML) or the Paule-Mandel estimator, both of which are iterative (Veroniki et al. 2016). We denote the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ by $\boldsymbol{\Sigma} = \mathrm{Cov}\,(\hat{\boldsymbol{\beta}})$. It was shown that, given certain regularity conditions, $\hat{\boldsymbol{\beta}} \xrightarrow{a.s.} \boldsymbol{\beta}$ as $k \longrightarrow \infty$ and $\hat{\boldsymbol{\beta}}$ asymptotically follows a normal distribution (Hedges et al. 2010).

Given a consistent estimator $\widehat{\boldsymbol{\Sigma}}$ for the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$, an approximate $(1-\alpha)$, $\alpha \in (0, 1)$ confidence interval (CI) for a coefficient β_j, $j \in \{0, 1, \dots, m\}$, is given by

$$\left[\hat{\beta}_j \pm t_{\mathrm{df},1-\alpha/2}\sqrt{\widehat{\boldsymbol{\Sigma}}_{jj}}\right], \tag{7}$$

where $t_{df,1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of the t-distribution with df degrees of freedom and $\widehat{\boldsymbol{\Sigma}}_{jj}$ is the jth diagonal element of $\widehat{\boldsymbol{\Sigma}}$ (Sterchi and Wolf 2017). In the following we discuss various possibilities for estimating $\boldsymbol{\Sigma}$. A typical choice for the degrees of freedom is $\mathrm{df} = k - m - 1$. However, Tipton (2015), Tipton and Pustejovsky (2015) showed that using the Satterthwaite approximation for the degrees of freedom improves the properties of the confidence interval in meta-regression. Therefore, this approach is used here to approximate the degrees of freedom.

2.3 Estimators for the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$

There are several ways to estimate the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$. In the following section we introduce $\mathbf{HC}_0$, $\mathbf{HC}_1$, $\mathbf{HC}_2$ according to MacKinnon and White (1985), $\mathbf{HC}_3$, $\mathbf{HC}_4$ according to Cribari-Neto (2004) and $\mathbf{HC}_5$ according to Cribari-Neto et al. (2007) if not stated otherwise.

The **HC** estimators are all based on $\mathbf{HC}_0$ which was originally introduced by White (1980) for an ordinary least squares (OLS) estimator. For the meta-regression model in (3) and the estimator $\hat{\boldsymbol{\beta}}$ given in (6) the estimator $\mathbf{HC}_0$ can be written as

$$\mathbf{HC}_0 = (\boldsymbol{X}^\top \widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \widehat{\boldsymbol{W}}\widehat{\boldsymbol{\Omega}}_0\widehat{\boldsymbol{W}}\boldsymbol{X}(\boldsymbol{X}^\top \widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}, \tag{8}$$

where $\widehat{\boldsymbol{\Omega}}_0 = \mathrm{diag}\,(\hat{e}_1^2, \dots, \hat{e}_n^2)$ is a matrix containing the squared residuals $\hat{e}_i = y_i - \boldsymbol{x}_i\hat{\boldsymbol{\beta}}$ on its diagonal (Welz and Pauly 2020).

How the formula for $\mathbf{HC}_0$ in (8) can be derived from the representation in MacKinnon and White (1985) is shown in Section A of the Supplement. The formulas for $\mathbf{HC}_1 - \mathbf{HC}_5$ can be derived analogously. Because the usual residuals tend to be too small (MacKinnon 2013), $\mathbf{HC}_0$ tends to underestimate the variance of the components of $\hat{\boldsymbol{\beta}}$. A simple adjustment of this estimator is given by $\mathbf{HC_1} = k(k-m-1)^{-1}\mathbf{HC_0}$, which takes the models' degrees of freedom $(k-m-1)$ into account.

Another approach to fix this problem of $\mathbf{HC}_0$ is to modify the residuals themselves. One possible modification is to take the leverage scores h_{ii} into account. The h_{ii} denotes the i^{th} diagonal element of the hat matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\widehat{\boldsymbol{W}}$. By using $\tilde{e}_i = \hat{e}_i/\sqrt{1-h_{ii}}$ instead of $\hat{e}_i$ there is more weight on residuals with higher leverage scores. A representation of $\mathbf{HC}_2$ is given by (8) using $\widehat{\boldsymbol{\Omega}}_2 = \mathrm{diag}\,((1-h_{ii})^{-1} \cdot \hat{e}_i^2)$ instead of $\widehat{\boldsymbol{\Omega}}_0$.

An estimator of similar form is $\mathbf{HC}_3$. It can be written by using $\widehat{\boldsymbol{\Omega}}_3 = \mathrm{diag}\,((1-h_{ii})^{-2} \cdot \hat{e}_i^2)$ in place of $\widehat{\boldsymbol{\Omega}}_0$ in (8). The estimator $\mathbf{HC}_3$ introduced here is a close approximation of Efrons' jackknife estimator (Efron 1982). A property of this estimator is that it takes the leverage scores stronger into account than $\mathbf{HC}_2$.

The following estimator, $\mathbf{HC}_4$, also differs from the former estimator in the way that it incorporates the leverage scores. The idea is to weight the residuals stronger, when the leverage score h_{ii} of a residual is relatively high compared to the average leverage score $\bar{h} = k^{-1} \sum_{i=1}^{k} h_{ii}$. This is done by using some δ_i as exponent for $(1 - h_{ii})$, where $\delta_i = \min\left\{4, h_{ii}/\bar{h}\right\}$. In this way the exponent $h_{ii}/\bar{h}$ is truncated at $\delta_i = 4$. The resulting estimator $\mathbf{HC}_4$ is given by (8) with $\widehat{\mathbf{\Omega}}_4 = \operatorname{diag}\left((1 - h_{ii})^{-\delta_i} \cdot \hat{e}_i^2\right)$ instead of $\widehat{\mathbf{\Omega}}_0$, see Zimmermann et al. (2020) for a similar estimator for multivariate analysis of covariance (MANCOVA).

Finally, $\mathbf{HC}_5$ is defined similar to $\mathbf{HC}_4$ but uses the exponents $\alpha_i = \min\left\{h_{ii}/\bar{h}, \max\left\{4, \eta \cdot h_{max}/\bar{h}\right\}\right\}$ instead of δ_i. Here, $h_{max} = \max\{h_{11}, \dots, h_{kk}\}$ and $\eta \in (0, 1)$ is a predefined constant used as a tuning parameter. The simulation study of Cribari-Neto et al. (2007) suggests $\eta = 0.7$ as a reliable choice for finite samples; we follow this recommendation here. Notably α_i is only different from δ_i when $(\eta \cdot h_{max})/\bar{h} > 4$. In this situation α_i is not truncated at $\alpha_i = 4$ but at $\alpha_i = (\eta \cdot h_{max})/\bar{h}$. A representation of $\mathbf{HC}_5$ is given by (8) plugging in $\widehat{\mathbf{\Omega}}_5 = \operatorname{diag}\left((1 - h_{ii})^{-\alpha_i} \cdot \hat{e}_i^2\right)$ for $\widehat{\mathbf{\Omega}}_0$. The Hartung-Knapp-Sidik-Jonkman estimator for the mixed-effects meta-regression model was independently introduced by Knapp and Hartung (2003) and Sidik and Jonkman (2005b). It can be derived as follows. Let $\boldsymbol{P} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top \widehat{\boldsymbol{W}} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \widehat{\boldsymbol{W}}$ and $s^2 = (k - m - 1)^{-1}(\boldsymbol{y}^\top \boldsymbol{P}^\top \widehat{\boldsymbol{W}} \boldsymbol{P} \boldsymbol{y}) = (k - m - 1)^{-1}(\boldsymbol{y}^\top \widehat{\boldsymbol{W}} \boldsymbol{P} \boldsymbol{y})$. Then the **HKSJ** estimator for $\operatorname{Cov}(\hat{\boldsymbol{\beta}})$ is given as

$$\mathbf{HKSJ} = s^2(\boldsymbol{X}^\top \widehat{\boldsymbol{W}} \boldsymbol{X})^{-1}.$$

3 Simulation study

3.1 Simulation design

The simulation was conducted using the open source software package `R`. Relevant packages that were used for the analyses are `metafor`, `MASS` and `mvtnorm`. Visualizations, such as boxplots, were created using the `ggplot2`, `reshape2`, `grid` and `gridExtra` packages. The simulation setup expands upon the one by Welz and Pauly (2020). The Satterthwaite approximation for the degrees of freedom is available in the `robust` function of the `metafor` package (version 3.4 or later) by specifying the `clubSandwich` argument as `TRUE`.

We start with a description of relevant effect measures for the simulation study. We consider the standardized mean difference (SMD), estimates of which are therefore the dependent variable in our meta-regression models. In many applications, θ_i is considered as the true SMD between the means of an experimental and a control group in the ith study. An unbiased estimator y_i for θ_i can be derived via a modification of Hedges' g. We describe the effect measure in the following, according to Hedges (1981). An unbiased estimator for the SMD is given by (Lin and Aloe 2021)

$$g := \frac{\Gamma(n/2)}{\sqrt{n/2}\,\Gamma((n-1)/2)} d \tag{9}$$

with $n = n_T + n_C - 2$, where n_T and n_C refer to the treatment and control group sizes. The regular Hedges' g is defined as $d = (\bar{x}_T - \bar{x}_C)/s$, where s is the pooled standard deviation with $s = \sqrt{\frac{(n_T-1)s_T^2+(n_C-1)s_C^2}{n_T+n_C-2}}$ and s_T^2, s_C^2 refer to the variances in the treatment and control groups respectively. The sampling variance of g can be approximated by (Hedges and Olkin 1985)

$$v = \frac{1}{n_T} + \frac{1}{n_C} + \frac{g^2}{2(n_T + n_C)}. \tag{10}$$

A mixed-effects meta-regression model with two covariates and their interaction is considered. The y_i are assumed to be influenced by two covariates and their interaction. The interaction is modelled as $x_{i12} := x_{i1}x_{i2}$. Thus the model equation is given as

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12} x_{i12} + u_i + \varepsilon_i. \tag{11}$$

The dependent variable y_i is assumed to be the estimated SMD between an experimental and a control group in the i^{th} study for $i = 1, \dots, k$. There are four choices for the number of studies, $k \in \{6, 10, 20, 50\}$. We note that test runs with $k = 5$ frequently resulted in either a rank-deficient design matrix $\boldsymbol{X}$ or extremely wide confidence intervals. Therefore it cannot be recommended to use only $k = 5$ studies for a model with two covariates and interaction. We assume balanced study designs, i.e. $n_{T,i} = n_{C,i} =: n_i$ for each study. For each choice of $k \in \{6, 10, 20, 50\}$ three different vectors of group sizes are considered. In the situation $k = 6$, five studies contain the group sizes according to the following three vectors: $n_{15} = (6, 8, 9, 10, 42)', n_{25} = (16, 18, 19, 20, 52)'$ or $n_{50} = (41, 43, 44, 45, 77)'$. The size of the sixth study is set to the mean $\bar{n}$ of the corresponding vector, either 15, 25 or 50. For $k \in \{10, 20, 50\}$ the vectors are repeated $k/5$ times and the resulting vector is used as the vector of study sizes. With this choice for the number of participants the study size vectors all have the same variance for a fixed k.

The covariates x_{i1} and x_{i2} are sampled from a joint normal distribution

$$\begin{pmatrix} x_{i1} \\ x_{i2} \end{pmatrix} \sim \mathcal{N}\begin{pmatrix} 1 & \varrho \\ \varrho & 1 \end{pmatrix},$$

where ϱ is the correlation between x_{i1} and x_{i2}. We examined the settings of no correlation ($\varrho = 0$), small correlation ($\varrho = 0.2$), large correlation ($\varrho = 0.5$) and large negative correlation ($\varrho = -0.5$). Possible adjustments for β_1, β_2 and β_{12} are 0, 0.2 and 0.5. Additionally, the situation $\beta_{12} = -0.5$ is considered in order to check whether the estimates differ for a negative coefficient.

The random effects u_i are chosen as $u_i = \tau q_i$, where $\tau^2 \in \{0.10, 0.15, \dots, 0.40\}$ (see, e.g., Linden and Hönekopp (2021) for a motivation of the τ^2 range) and the q_i's are independently sampled from either a standard normal- or a standardized exponential-, Laplace-, log-normal- or t_3-distribution. Here t_3 denotes the t distribution with three degrees of freedom. If q_i is drawn from a standardized exponential distribution, then $q_i := a_i - 1$ where $a_i \sim \exp(1)$. The q_i's following a standardized Laplace distribution, are generated via $q_i = (a_i - b_i)/\sqrt{2}$ where $a_i, b_i \sim \exp(1)$ are sampled independently. For the q_i's following a standardized log-normal distribution, q_i is set to

$$q_i = \frac{\exp(z_i) - \exp(1/2)}{\sqrt{\exp(1)(\exp(1) - 1)}},$$

where $z_i \sim \mathcal{N}(0, 1)$. Finally, q_i's following a standardized t_3 distribution are set as $q_i = t_i/\sqrt{3}$ with $t_i \sim t_3$. The standardization of the q_i's ensures that the corresponding u_i's all have expectation $\mathbb{E}(u_i) = 0$ and variance $\mathrm{Var}\,(u_i) = \tau^2$. Note, that if u_i is not normally distributed, the y_i are not normally distributed and the t-quantile used in (7) is not correct. However, results by Kontopantelis and Reeves (2012) suggest that the distribution of the study outcomes has almost no impact on the resulting confidence intervals. Therefore the quantile of the t-distribution is used for this simulation as well.

The estimated effects (Hedges' g) y_i are generated according to

$$g_i = \frac{\phi_i}{\sqrt{X_i/(2n_i - 2)}}, \tag{12}$$

where $\phi_i \sim \mathcal{N}(\theta_i, 2/n_i)$ and $X_i \sim \chi^2_{(2n_i-2)}$ are sampled. The sampling variance σ_i^2 of y_i is estimated using (10). In total there are $77,760 = 3(\bar{n}) \times 4(k) \times 9(\tau^2) \times 3(\beta_1) \times 3(\beta_2) \times 4(\beta_{12}) \times 5(u_i) \times 4(\varrho)$ different combinations of simulation parameters. For each combination the model is generated $N = 10,000$ times. The confidence level is chosen as $1 - \alpha = 0.95$. For this choice of N and α the expected Monte Carlo standard error of empirical coverage is approximately equal to $\sqrt{\alpha(1-\alpha)/N} = 0.22\%$ (Morris et al. 2019). For each model the estimators $\mathbf{HC}_0$–$\mathbf{HC}_5$ and **HKSJ** are calculated and τ^2 is estimated using the REML estimator, with a maximum of 5, 000 iterations and a default step length of 0.5. Based on each estimator a $(1 - \alpha)$ confidence interval is estimated for the coefficient β_1 of a single moderator and for the coefficient β_{12} of the interaction term. Since x_1 and x_2 have the same distribution, intervals for β_2 are not considered. The proportion of estimated confidence intervals that cover the true coefficient is used as an estimate of the coverage probability. As an estimate of the interval width the 10 %-trimmed mean of the widths of the estimated intervals is calculated in order to robustify the results. For comparability reasons, the same parameter ranges as in Welz and Pauly (2020) were used, except for the values for τ^2.

3.2 Simulation results

In confidence interval estimation two properties are relevant, namely coverage and interval width. The actual coverage of the interval should be at least equal to the nominal confidence level $(1-\alpha)$. Second, we want to determine the interval, where the true parameter is included in with probability $(1-\alpha)\cdot 100\%$, as precisely as possible. This means of the intervals that have sufficient coverage, we choose the narrowest one. Therefore, the coverage and widths of the simulated intervals for β_1 (and β_2) as well as β_{12} are compared in respect of the covariance estimators they are based on. Due to the high number of parameter adjustments not every adjustment is considered separately. Hence, the coverage and interval widths of different settings are summarized by boxplots. That is, e.g., the boxplots in Sect. 3.2.1 based upon the results for every adjustment of $\beta_1, \beta_2, \beta_{12}, \rho, \tau^2, u_i$ and $\bar{n}$ and thus consider the overall performance of the estimators. The aim of this section is to investigate, whether one estimator has a better overall performance compared to all other estimators. It is also of interest, whether there are any estimators that are outperformed by at least one other estimator in each situation. Because the intervals for β_1 and β_{12} performed similarly for the most estimators and parameter adjustments, only the results for the confidence intervals of β_1 are shown in detail. The differences to the intervals for β_{12} are highlighted in Sect. 3.2.1, the full results for the intervals for β_{12} are shown in Section B the Supplement.

Since the number of studies k strongly affects the coverage and interval widths (Sect. 3.2.2), the results are compared separately for each k. How the adjustments of other simulation parameters affect the coverage and interval width is discussed in Sect. 3.2.2. There it is of interest, whether the performance of a certain estimator differs from its overall performance for a special adjustment. For example, it is analyzed whether there is an estimator whose intervals have the best performance but only for large correlations. For ease of presentation “confidence interval” is abbreviated with CI in this section. The CIs based on $\mathbf{HC}_0$ are abbreviated with $\mathbf{HC}_0$-CI, the CIs based on other estimators in an analogous manner.

3.2.1 Overall performance of the estimators

Confidence intervals for $\boldsymbol{\beta_1}$ – Coverage Probability. In Fig. 1 the coverages of the CIs for β_1 are summarized using boxplots. Each plot reflects the results for a certain number of studies $k \in \{6, 10, 20, 50\}$. The individual boxplots contain the coverage of all intervals based on the respective estimator and k. In addition, each of the plots contains one boxplot with a reference distribution for the coverage. The values for the reference plot were calculated by sampling 15, 120 values from a binomial distribution ($\mathcal{B}(10000, 0.95)$) and dividing the values by 10, 000. These values show the distribution of coverage rates that can be expected within a simulation and can be used to better interpret the results for the different estimators.

The coverage of the $\mathbf{HC}_0$-CIs ranges from 0.8640 to 0.8904 in case of $k = 6$. For $k = 50$ the coverage ranges from 0.9142 to 0.9569. But only for 2.37% of the adjustments with $k = 50$ the coverage is above the nominal confidence level. Thus, $\mathbf{HC}_0$ is too liberal and an inappropriate choice of estimators regarding their CI coverage.

The $\mathbf{HC}_1$-CIs have higher median coverages than the $\mathbf{HC}_0$-CIs for all k. For $\mathbf{HC}_1$, the median coverage is above the nominal level for $k = 6$, $k = 10$ and $k = 20$. For $k = 50$ it is slightly lower (0.9487). But still for 38.16% of the adjustments, the

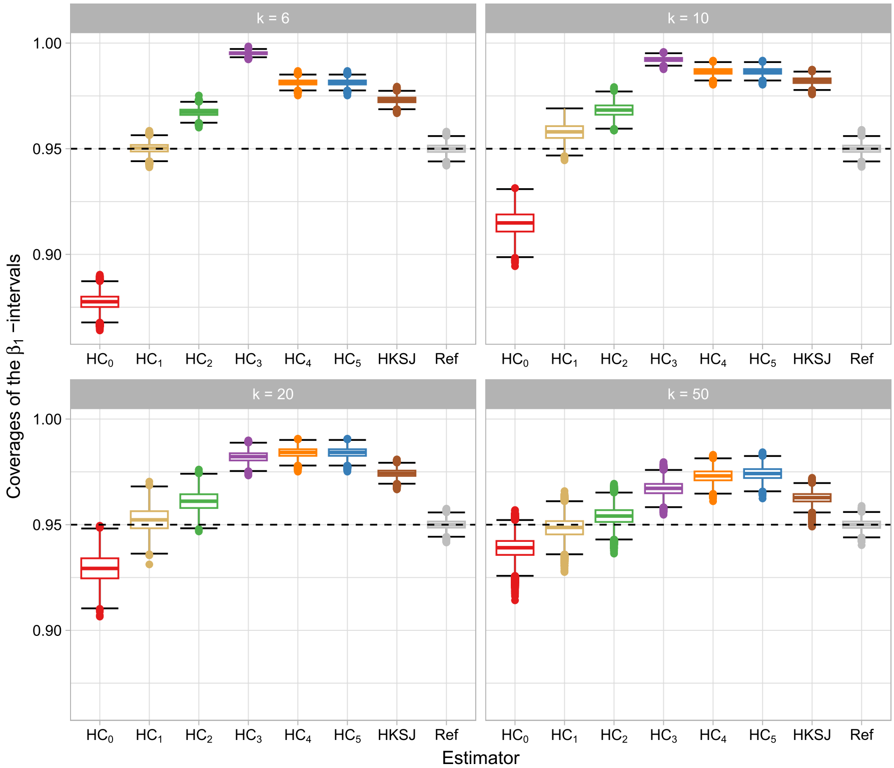

Fig. 1 Coverage probabilities of the confidence intervals for the regression parameter β_1 based on the estimators $\mathbf{HC}_0 - \mathbf{HC}_5$ and **HKSJ** for different numbers of studies k. In addition, a boxplot with a reference distribution for the coverage is included in the plots

coverages are above the nominal confidence level. In summary, the coverage of the $\mathbf{HC}_1$-CI for β_1 is quite accurate.

$\mathbf{HC}_2$ based CIs are more conservative compared to $\mathbf{HC}_1$ based CIs. Their median coverage is above the nominal level for all values of k. The coverage decreases with increasing k. For $k = 50$, the coverage is below 0.95 for 16.12% of the adjustments.

$\mathbf{HC}_3 - \mathbf{HC}_5$ based CIs also exhibit a conservative behaviour with a median coverage larger than 0.95 for all number of studies k. For $k = 6$ the $\mathbf{HC}_3$-CIs are the most conservative with coverages ranging from 0.9923 to 0.9984. However, the coverage of the $\mathbf{HC}_3$-CI is decreasing in k, but is above the nominal level for all of the adjustments. The coverages of the $\mathbf{HC}_4$-CIs and $\mathbf{HC}_5$-CIs range from 0.9752 to 0.9868, and differ only slightly in respect of the number of studies.

The CIs based on the **HKSJ** have median coverages that are slightly lower than for $\mathbf{HC}_4$ and $\mathbf{HC}_5$ for almost all considered values for k. For $k = 50$, the coverages are below the nominal level for only 0.01% of the adjustments.

Among all estimators $\mathbf{HC}_1$-CIs show the closest coverages compared to the nominal confidence level 0.95. The coverage tends to be slightly lower for larger

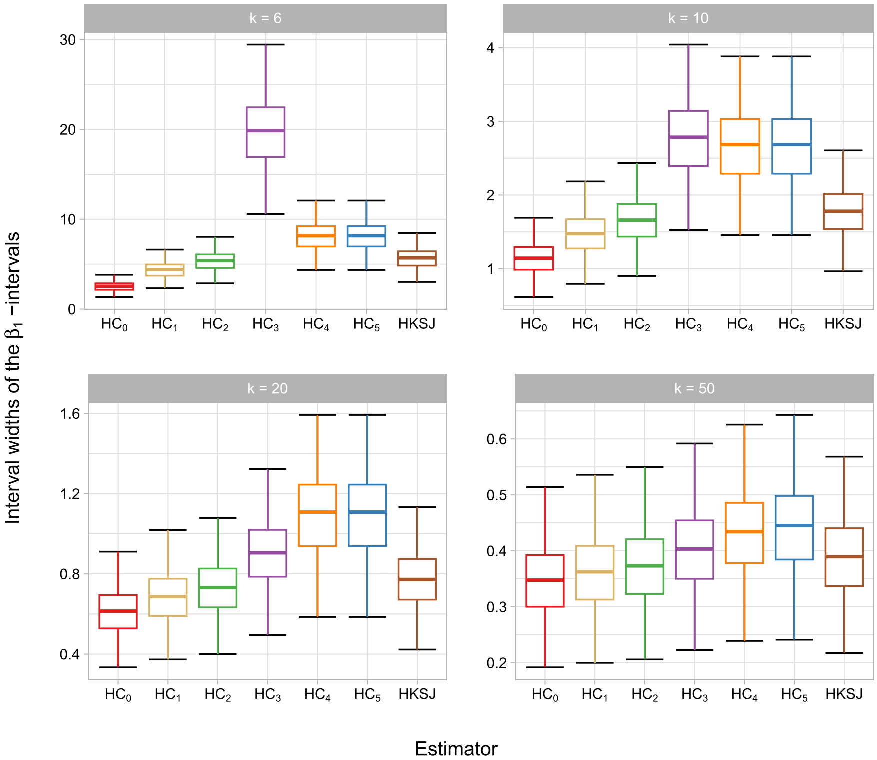

Fig. 2 Widths of the confidence intervals for the regression parameter β_1 based on the estimators $\mathbf{HC}_0 - \mathbf{HC}_5$ and **HKSJ** for different numbers of studies k without outliers

number of studies k. For $k = 6$ the median coverage of the $\mathbf{HC}_1$-CI is almost equivalent to the nominal confidence level and the distribution seems to be similar to the reference distribution. The $\mathbf{HC}_2$-based CIs are slightly more conservative than the intervals based on $\mathbf{HC}_1$. The intervals based on $\mathbf{HC}_3$-$\mathbf{HC}_5$ are most conservative, having coverages above 0.95 for all adjustments. For the **HKSJ**, the coverages are above the nominal level for almost all k. For $k = 50$, only for 0.01% of the adjustments, the coverage is below 0.95.

Confidence intervals for $\boldsymbol{\beta_1}$ – Width. Boxplots of the corresponding interval widths are shown in Fig. 2.

The interval widths of all estimators are monotonically decreasing in the number of studies k.

Widths of the **HKSJ**-CIs range from 3.02 to 8.47 for $k = 6$ and from 0.22 to 0.57 for $k = 50$. Thereby, they are narrower compared to the $\mathbf{HC}_4$- and $\mathbf{HC}_5$-CIs for all considered number of studies k. Except for $k = 50$, where the widths of the $\mathbf{HC}_5$-CIs tend to be slightly larger, the widths of the $\mathbf{HC}_4$- and $\mathbf{HC}_5$-CIs behave almost identically.

For $k = 6$ the median width of the **HKSJ**-CIs is equal to 5.6, whereas it is equal to 8.12 for the $\mathbf{HC}_4$- and $\mathbf{HC}_5$-CIs. In the situation of $k = 50$ the median interval width of **HKSJ**-CIs is 0.39, which is smaller than the $\mathbf{HC}_4$-CIs with 0.43 and the $\mathbf{HC}_5$-CIs with 0.46.

The $\mathbf{HC}_1$-CIs are in median narrower than the other CIs that control the nominal confidence level. The $\mathbf{HC}_2$-CIs show a tendency of being slightly larger than the $\mathbf{HC}_1$-CIs but are still narrower than the **HKSJ**-CIs.

Widths of the $\mathbf{HC}_3$-CIs are highly inflated for $k = 6$. The lower quartile is equal to 16.93 and the upper quartile's value is 22.45. For $k = 6$ and $k = 10$ the $\mathbf{HC}_4$- and $\mathbf{HC}_5$-CIs are narrower in the median than the $\mathbf{HC}_3$-CIs, for the other values of k they are larger. The $\mathbf{HC}_0$ based CIs tend to be narrower than the other CIs for all k but were the most liberal regarding the coverage.

In comparison of all estimators whose intervals have a suitable coverage, the $\mathbf{HC}_1$-CIs are the narrowest and therefore preferable. Since their CIs are just slightly larger for all values of k, $\mathbf{HC}_2$ has the second best performance. The interval width of **HKSJ**-CIs are just minimally larger.

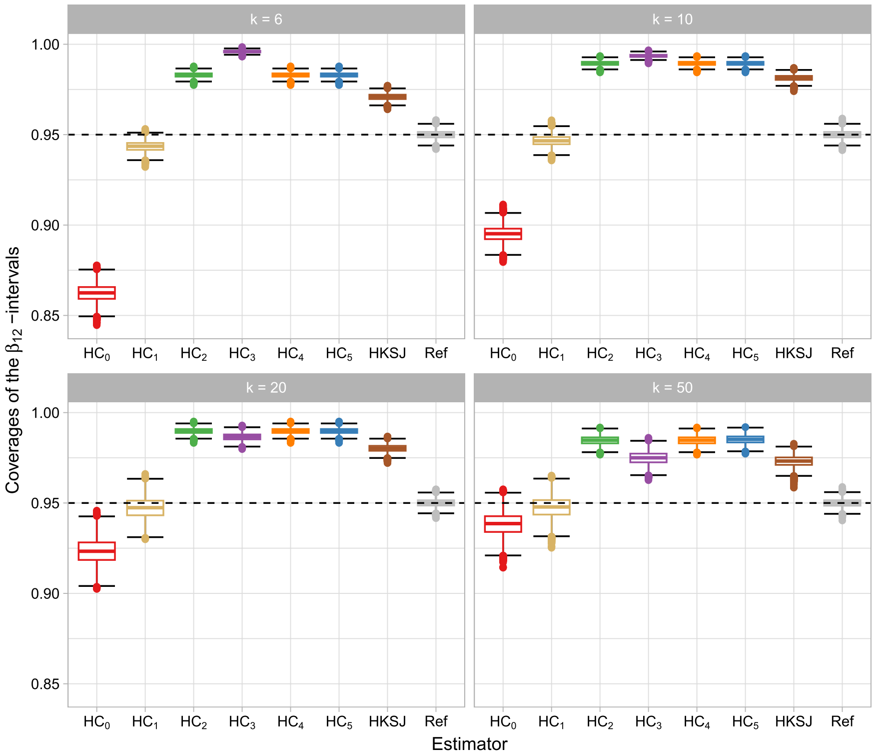

Fig. 3 Coverage probabilities of the confidence intervals for β_{12} based on the estimators $\mathbf{HC}_0 - \mathbf{HC}_5$ and **HKSJ** for different numbers of studies k. In addition, a boxplot with a reference distribution for the coverage is included in the plots

Performance of the intervals for β_{12}. Compared to the intervals for β_1 the intervals for β_{12} tend to be wider for most estimators and adjustments of k.

Regarding the coverage rate, the $\mathbf{HC}_0$-CIs perform better compared to their counterpart for β_1, see Fig. 3. For $k = 50$ in 3.44% of the adjustments, the coverage is above the nominal level but still too liberal for all k. For the $\mathbf{HC}_1$ estimator, the coverage of the CIs increases with an increasing k. For $k = 6$, the median coverage is 0.9436 and only for 0.62% of the adjustments, the coverage is above the nominal level of 0.95. Whereas for $k = 50$, the median coverage is 0.9478 and for 34.71% of the adjustments, the coverage is higher than 0.95. Moreover the coverage is larger than 0.93 in almost all of the cases (99.81%).

The estimators $\mathbf{HC}_2$-$\mathbf{HC}_5$ lead to the most conservative CIs. For coverages for none of the corresponding CIs show coverages below 0.95. For $\mathbf{HC}_2$, $\mathbf{HC}_4$ and $\mathbf{HC}_5$ the coverages of the CIs first increase until $k = 20$ and then slightly decrease for $k = 50$, while for $\mathbf{HC}_3$ the coverages decrease for increasing k. The **HKSJ**-CIs show lower median coverages than the CIs based on $\mathbf{HC}_2$-$\mathbf{HC}_5$ but still are quite conservative. The median coverage of the **HKSJ**-based CIs and $k = 6$ is 0.971.

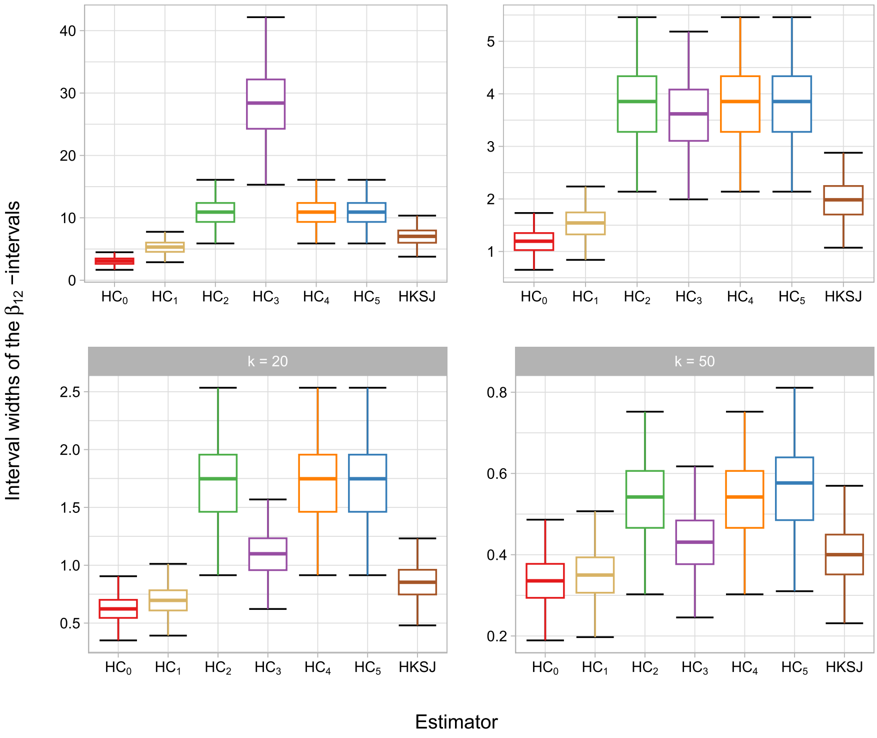

Fig. 4 Widths of the confidence intervals for β_{12} based on the estimators $\mathbf{HC}_0 - \mathbf{HC}_5$ and **HKSJ** for different numbers of studies k without outliers

All estimators perform worse for intervals for β_{12} compared to intervals for β_1 in terms of their CI widths, see Fig. 4. Again, $\mathbf{HC}_0$ led to the smallest widths but also was very liberal.

As for β_1, $\mathbf{HC}_1$ should prefarably be used, especially for $k = 50$. The **HKSJ** shows the next best performance, which is different to the results for β_1, where $\mathbf{HC}_2$ performed slightly better than **HKSJ**. $\mathbf{HC}_2$ now shows the worst performance for $k = 50$ and the CIs based upon this estimator are very wide compared to the other estimators.

Summary of the overall performance. Summing up, the results are different compared to the ones observed for a model with one covariate (Viechtbauer et al. 2015; Welz and Pauly 2020). Among all considered estimators the $\mathbf{HC}_1$ estimator is the most appropriate for a model with an interaction term since it performed the best for the coefficient of the single moderator and was only slightly liberal for the interaction term in case of $k = 6$. Focusing on the other estimators, $\mathbf{HC}_2$ was found to be the second best choice for constructing CIs for β_1 and **HKSJ** for β_{12}, respectively. If you aim for a rather secure conservative coverage, they can even be viewed as best choice. Overall, $\mathbf{HC}_0$ shows the worst performance with very liberal CIs. $\mathbf{HC}_4$ and $\mathbf{HC}_5$ perform bad in terms of interval widths. $\mathbf{HC}_3$ performs even worse for small numbers of studies k but is better than $\mathbf{HC}_4$ and $\mathbf{HC}_5$ (and $\mathbf{HC}_2$ for β_{12} intervals) for $k \geq 20$. Therefore, $\mathbf{HC}_0$ and $\mathbf{HC}_3$-$\mathbf{HC}_5$ are not recommendable for estimating β_1 and β_{12}.

Importance of the Satterthwaite approximation. We also run comparative simulations for the 'classical' t-type intervals that use $df = k - 4$ degrees of freedom in (7). As the performance was worse compared to the Satterthwaite approximation, we decided to only present the detailed results in Appendix D. There, you can see that using $k - 4$ degrees of freedom leads to more liberal CIs (for both, β_1 and β_{12}) with similar interval widths. For that reason we focus on the Satterthwaite approximation for the degrees of freedom in the paper.

3.2.2 Effects of parameter adjustments

This section will summarize how the coverages and interval widths are affected by the adjustments of the flexible parameters. Since both coefficients are effected similar by most parameters they are considered together. We highlight the most important results and refer to Section C of the Supplement for complete results.

Adjustments of the number of studies *k*. Considered numbers of studies are 6, 10, 20 and 50. The widths of both coefficients intervals are monotonically decreasing in the number of studies k. The effect of the number of studies on coverage is not constant and depends on the considered covariance estimator. In general coverage tends towards the nominal level $1 - \alpha$ for increasing k. Therefore, for all estimators a large number of studies is preferable. For the $\mathbf{HC}_1$-based CIs also smaller number of studies lead to good coverage rates.

Adjustments of study size. Small ($\bar{n} = 15$), medium ($\bar{n} = 25$) and large ($\bar{n} = 50$) group sizes are compared. For most covariance estimators the median coverage for β_1 is slightly increasing in the study size.

For β_{12}, most of the coverage rates are decreasing (towards the nominal level) with increasing study sizes.

The corresponding interval widths are decreasing as the study sizes increase for all k and estimators. This trend may be caused by the impact of n_i on v_i in Equation (10), which leads to decreasing standard errors in equation (7). Thus, overall larger studies lead to better confidence intervals, since both coverages and interval widths are improved for larger study sizes.

Adjustments of $\boldsymbol{\tau^2}$. Coverages of both coefficients intervals are increasing slightly in the heterogeneity parameter τ^2 for almost all estimators and $k = 20$ and $k = 50$. For $k = 6$ and $k = 10$, almost no changes in the medium coverage rates can be observed for all of the estimators. For a larger number of studies, the effect is stronger. The increasing coverages in τ^2 show that the model used in the simulation is adequate to model a study effect. On the other hand the interval widths are increasing in τ^2 strongly. This result is explicable by the direct impact the value of τ^2 has on the variances of the coefficients and thus on the interval bounds.

Adjustments of $\boldsymbol{\beta_1}$. Examined adjustments of β_1 are 0, 0.2 and 0.5. The CIs for β_{12} were not affected by these adjustments of β_1, whereas the coverage of the CIs for β_1 slightly decrease for an increasing β_1 for $k = 50$ studies and all estimators. Adjustments of β_1 had no influence on the interval widths of the CIs for β_1 and β_{12}.

Adjustments of $\boldsymbol{\beta_2}$. For β_2 the adjustments 0, 0.2 and 0.5 were considered as well. None of the intervals was affected by the adjustment of β_2 regarding the coverage or width.

Adjustments of $\boldsymbol{\beta_{12}}$. Besides the adjustments 0, 0.2 and 0.5 for β_{12} the adjustment −0.5 was simulated as well, to check whether it differs from the 0.5 adjustment. This is neither the case for the interval widths nor for the coverages of the CIs for β_1 and the CIs for β_{12}. However, the $\mathbf{HC}_4$- and $\mathbf{HC}_5$-CIs for β_1 have slightly lower coverage for a high absolute value of β_{12}.

For $k = 50$, the coverage rates for the CIs for β_{12} for all estimators tend to be slightly lower for an absolute value of $|\beta| = 0.5$ than for lower absolute values. This is also the case for the interval widths for the CIs for β_{12} which are slightly lower for higher absolute values of β_{12} for $k = 20$ and $k = 50$ and the estimators $\mathbf{HC}_2, \mathbf{HC}_4$ and $\mathbf{HC}_5$ and slightly higher for the same values of k for the other estimators.

Altogether the true values of the considered parameters do not have a strong impact on the intervals of any estimator. Therefore, there is no coefficient for which an estimator performs better or worse compared to the other estimators than in the overall results.

Adjustments of the correlation $\boldsymbol{\rho}$. Examined adjustments of ρ are 0, 0.2, 0.5 and −0.5.

The sign of the correlation affects neither the coverages nor the interval widths. Intervals for β_1 that are based on $\mathbf{HC}_3 - \mathbf{HC}_5$ tend to have a lower coverage for higher absolute correlations, whereas CIs based on $\mathbf{HC}_0 - \mathbf{HC}_2$ tend to have higher coverages for $|\rho| = 0.5$. For $\mathbf{HC}_2$ and $\mathbf{HC}_3$ the respective effect is only marginal. Large correlations induce wider CIs for β_1 for all number of studies and estimators. There is no consistent impact of the correlation on the CIs for β_{12}. The changes depend on both the estimator and number of studies k. However, these changes are only minor.

$\mathbf{HC}_0$, $\mathbf{HC}_1$, $\mathbf{HC}_3$ and **HKSJ** based CIs have narrower widths for larger values of $|\rho|$ and all k. Intervals based on $\mathbf{HC}_2$, $\mathbf{HC}_4$ and $\mathbf{HC}_5$ have marginally decreasing widths in $|\rho|$ for $k = 6$, slightly increasing widths for $k \in \{10, 20\}$ and again marginally decreasing widths for $k = 50$. It is also interesting to note, that most of the extreme outliers of $\mathbf{HC}_5$ occur for high correlations.

Adjustments of the random effect distribution. Simulated random effect distributions are the standard normal distribution and standardized Laplace-, exponential, t_3- and log-normal-distributions. The results for the β_{12} CIs in case of $k = 10$ are presented in Figs. 5 (coverages) and 6 (widths), respectively, while all other results including the ones for the β_1 CIs are given in the appendix and are summarized here.

In comparison with the other simulated distributions, the coverages of CIs for β_1 based on $\mathbf{HC}_0 - \mathbf{HC}_5$ are on average the lowest with normal distributed u_i and highest with log-normal distributed random effects. The coverages do not differ much in respect of the other random effect distributions. The **HKSJ**-CIs for $k \in \{6, 10\}$ have the highest coverage with normal distributed random effects and the lowest with log-normal random effects. Especially for $k = 10$ the coverages of the **HKSJ**-CIs with non-normal random effects tend to be lower. But in none of the adjustments with non-normal random effects the coverages of the **HKSJ**-CIs are below 0.95. For $k = 20$ the **HKSJ**-CIs show no observable differences between the random effect distributions, whereas for $k = 50$ the order of the median coverages is the same as for the other estimators. Thus, in this situation the coverages of the **HKSJ**-CIs for β_{12} are even less adequate than for the intervals for β_1.

The coverage of the $\mathbf{HC}_0 - \mathbf{HC}_5$ CIs for β_{12} are all slightly better for non-normal random effect distribution compared to the normal case for all $k \in \{10, 20, 50\}$.

For $k \in \{6, 10, 20\}$ the coverage of the $\mathbf{HC}_1$-CIs for β_{12} are on median between 0.943 and 0.949 in the non-normal random effects setting.

The median widths of both coefficients CIs depends on the underlying distribution and can be ordered in the following way for all k and estimators: normal > Laplace > exponential > t_3 > log-normal. Thus, for all approaches the confidence intervals have better properties, when the random effect distribution is different from a normal distribution. Therefore, the quantile used as critical value is suitable, even if the distribution of the u_i is not normal. Thus, if a precise control of the nominal confidence level is required $\mathbf{HC}_1$ (for all k) may be most suitable.

In summary, the estimators are affected by most parameter adjustments in the same way or a similar manner. Only the number of studies k shows a strong varying effect on the coverage of some estimators. Besides the number of studies, the group size and the heterogeneity parameter τ^2 have impact on the interval widths. However, the trend is the same for all estimators and reducible to the direct impact of these parameters on components of the confidence interval in equation (7). The results of the different random effect distributions indicate that all estimators are robust against deviations from the normal distribution. In particular, the CIs were even slightly better compared to the normal setting. There is no situation where any estimator performs superior compared to its overall performance.

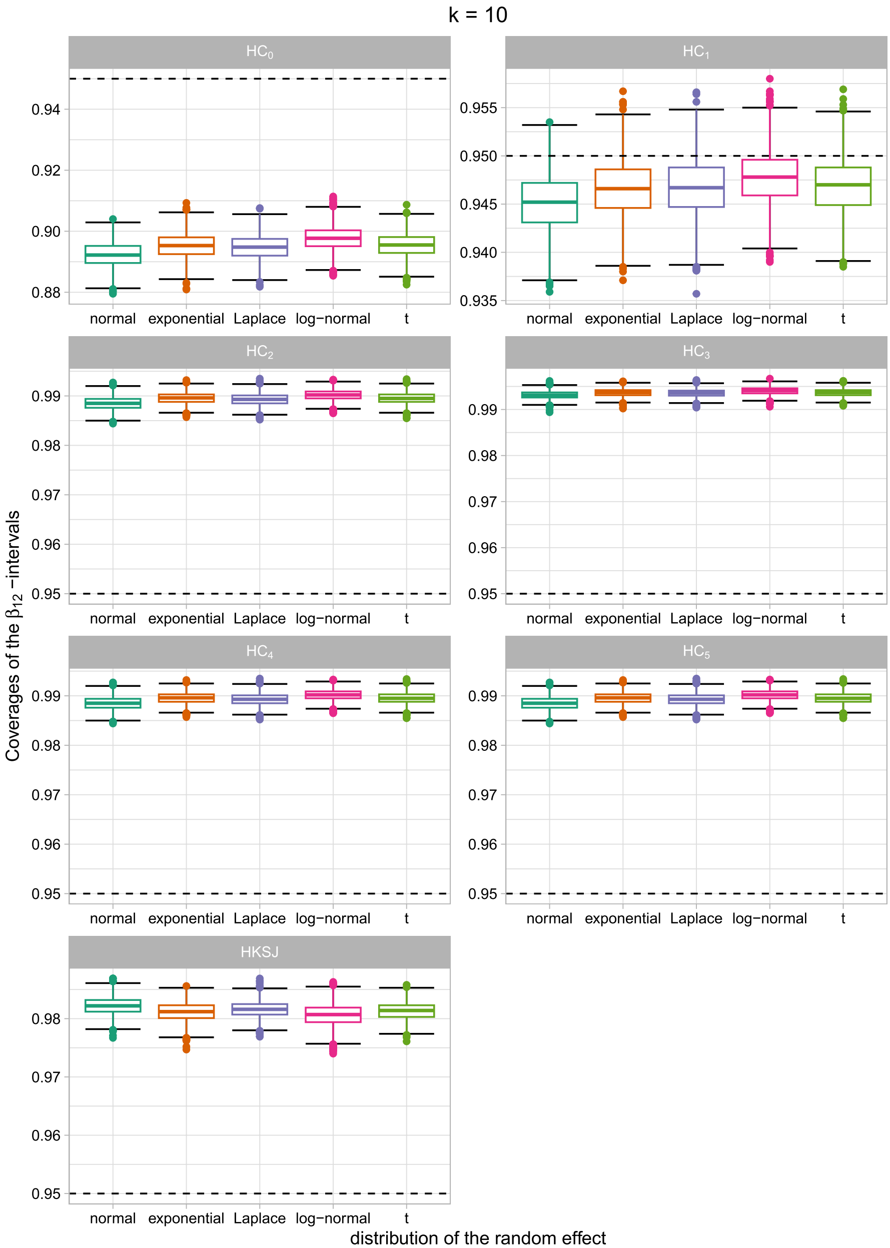

Fig. 5 Coverages of the β_{12}-intervals compared regarding the adjustments of u_i for the estimators $\mathbf{HC_0} - \mathbf{HC_5}$ and **HKSJ** with $k = 10$

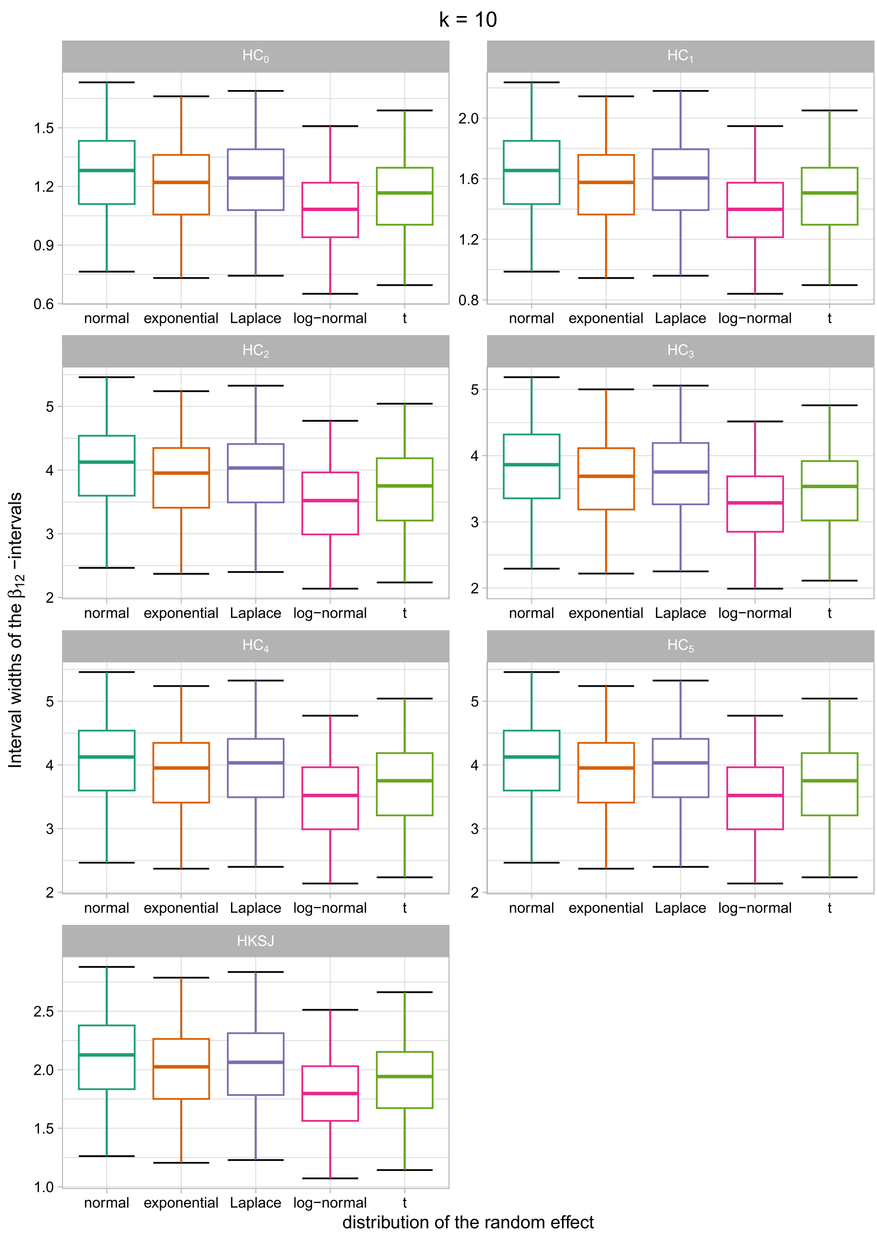

Fig. 6 Widths of the β_{12}-intervals compared regarding the adjustments of u_i for the estimators $\mathbf{HC_0} - \mathbf{HC_5}$ and **HKSJ** with $k = 10$

4 Discussion

Here we compared different confidence intervals for a mixed-effects meta-regression model with two moderators and an interaction term. The confidence intervals were based on one of the six different heteroscedasticity consistent covariance estimators $\mathbf{HC}_0, ..., \mathbf{HC}_5$ or the Hartung-Knapp-Sidik-Jonkman covariance estimator **HKSJ**. In a simulation study the confidence intervals based on these estimators were compared regarding their coverage and widths for numerous combinations of simulation parameters. The simulation settings varied in the number of studies, the study sizes, a heterogeneity parameter, the coefficients of the moderators, the correlation between the covariates and the distribution of the random effect. A total of 60, 480 combinations was simulated 10, 000 times.

The coverage of the confidence intervals based on $\mathbf{HC}_0$ turned out to be below the nominal confidence level 0.95 for almost every setting and are therefore not adequate. Contrary, the $\mathbf{HC}_1$-CIs were on median the most accurate wrt nominal confidence level. Although the coverage of the confidence intervals based on $\mathbf{HC}_2$ ($\mathbf{HC}_2$ -CIs) for β_1 was suitable, $\mathbf{HC}_2$-CIs were the most conservative for β_{12}. The CIs based on the estimators $\mathbf{HC}_3 - \mathbf{HC}_5$ and **HKSJ** showed a conservative coverage for both parameters. Concerning the interval widths the $\mathbf{HC}_1$-CIs performed the best for all settings among all estimators with adequate coverage followed by $\mathbf{HC}_2$ (for β_1) and **HKSJ** (for β_{12}), respectively. For a small number of studies ($k = 6$) the widths of the $\mathbf{HC}_3$-CIs were highly inflated. For larger numbers of studies ($k \geq 10$) the widths of the $\mathbf{HC}_3$-CIs are narrower compared to the $\mathbf{HC}_4$ and $\mathbf{HC}_5$ intervals. Thus, for $k \geq 10$ $\mathbf{HC}_3$-CIs are preferable compared to $\mathbf{HC}_4$- and $\mathbf{HC}_5$-CIs. However, **HKSJ** and $\mathbf{HC}_1$-$\mathbf{HC}_2$ showed a better performance compared to $\mathbf{HC}_0$, $\mathbf{HC}_3$-$\mathbf{HC}_5$.

The results for single parameter adjustments differ only slightly from the overall results. The interval widths were shown to be increasing in the amount of heterogeneity τ^2, whereas they were decreasing in the number of studies k and the mean study sizes $\bar{n}$ for all estimators. Coverages were mostly increasing in $\bar{n}$ and τ^2. The confidence intervals were only slightly affected by the values of the true coefficients. Only high values of β_1 and strong interactions ($|\beta_{12}| = 0.5$) slightly reduced the coverage of some intervals.

For all different estimators for both CIs higher correlations ρ only slightly affected the coverages but the widths of the intervals for β_1 were increasing in ρ. Concerning coverage and widths of the CIs for β_{12} no such trend was observable. The widths of the CIs even were smaller for increasing ρ. Surprisingly, all estimators performed slightly better for non-normal distributed random effects regarding their coverage and widths.

For small numbers of studies $k \in \{6, 10\}$ the coverage of the $\mathbf{HC}_1$-CIs tend to be slightly below the nominal confidence level $(1 - \alpha) = 0.95$ for β_{12}, but the coverage of the $\mathbf{HC}_1$-CIs are still close to 0.95. However, the **HKSJ**-CIs also show suitable coverage rates in these situations, though they are more conservative.

Altogether, the results of this work differ from the results of previous studies for a single moderator Viechtbauer et al. (2015); Welz and Pauly (2020). The superior performance of the **HKSJ** estimator and the behavior of the **HC** estimators observed in the model with one moderator is different when studying the model with two covariates and an interaction. While the **HKSJ**-estimator performed best in their simulations, the $\mathbf{HC}_1$-estimator showed a better performance than the **HKSJ**-estimator for estimating the parameter β_1 for most of the adjustments in the present simulation study for the more complex interaction model. Thus, we recommend $\mathbf{HC}_1$ in this setting. However, since the coverage rates are above the nominal level of 0.95 in all of the adjustments when using the **HKSJ**-estimator, it still is a suitable choice here. For CIs for β_1 or other not-interaction parameters (e.g., β_2), $\mathbf{HC}_2$ is a suitable choice as well, since its intervals held the nominal confidence level for every k and were even narrower compared to the **HKSJ** CIs. In further research it may also be of interest to analyze the situations where highly inflated interval widths of the $\mathbf{HC}_3$-, $\mathbf{HC}_4$- and $\mathbf{HC}_5$-CIs occurred in detail, because they cannot be explained by the results of this work. Furthermore, using the Satterthwaite approximation to calculate the degrees of freedom appears to be crucial for achieving appropriate coverage rates of the confidence intervals.

Limitations and future work. The presented simulation study, though very extensive, had several limitations. The mixed model examined in this work still has a simple structure with just two covariates and one interaction effect and focused on confidence intervals. Thus, our results do not generalize to other settings. In further research it might be of interest to consider the performance of these and other inference methods (confidence ellipsoids, multivariate (Welz et al. 2023) and multiple tests etc.) for more complex models. Interesting settings are mixed regression interaction terms of higher order, other random effect and covariate distributions, more complex dependencies, other nominal levels (Johnson 2013; Benjamin et al. 2018; Noguchi et al. 2021) than $\alpha = 0.05$, more extreme coefficients, different parameter values or even more complex meta analysis models (e.g. network analyses (White 2015)). Moreover, a detailed analysis of the methods' behaviour under model mis-specification would be of its own interest. For example, as pointed out by an expert referee, HC estimators are usually more robust to mis-specification of the random effects distribution compared to the **HKSJ** method. For instance, this might be the case in a random effects location scale model as considered in Viechtbauer and López-López (2022).

Another limitation of our research regarding the estimator $\mathbf{HC}_5$ is that we did not optimize the tuning parameter η, relying on the recommendation of $\eta = 0.7$ by Cribari-Neto et al. (2007). The question whether and how the optimal choice of η depends on a given context remains an open question for further research.

Concluding, meta-regression remains an important field of statistical research. CIs derived from the **HKSJ** estimator, which performed best in earlier studies with simpler models, performed worse than the CIs derived from $\mathbf{HC}_1$ estimator together with the Satterthwaite approximation in this situation and the $\mathbf{HC}_2$ for β_1. However, the CIs based upon the **HKSJ** estimator still perform better than most of the other **HC** based approaches in most of the settings.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00180-024-01530-0.

Acknowledgements The authors gratefully acknowledge the computing time provided on the Linux HPC cluster at TU Dortmund University (LiDO3), partially funded in the course of the Large-Scale Equipment Initiative by the German Research Foundation (DFG) as project 271512359. The code for the simulation is provided in a repository at https://osf.io/q7hz9/.

Funding Open Access funding enabled and organized by Projekt DEAL. This work was supported by the German Research Foundation: project Grant no. PA-2409 7-1 (Markus Pauly) and FR 3070/3-1 (Tim Friede).

Declarations

Conflict of interest The authors have declared no conflict of interest.

References

Aiken LS, West SG, Reno RR (1991) Multiple regression: testing and interpreting interactions. Sage Publications, California

Benjamin DJ, Berger J, Johannesson M, Nosek B, Wagenmakers E, Berk R, Johnson V (2018) Redefine statistical significance. Nat Hum Behav 2:6–10

Cribari-Neto F (2004) Asymptotic inference under heteroskedasticity of unknown form. Comput Stat Data Anal 452:215–233

Cribari-Neto F, Souza TC, Vasconcellos KLP (2007) Inference under heteroskedasticity and leveraged data. Commun Stat- Theo Methods 3610:1877–1888. https://doi.org/10.1080/03610920601126589

DerSimonian R, Laird N (1986) Meta-analysis in clinical trials. Control Clin Trials 73:177–188

Efron B (1982) The jackknife, the bootstrap and other resampling plans. SIAM, Philadelphia

Friede T, Röver C, Wandel S, Neuenschwander B (2017) Meta-analysis of few small studies in orphan diseases. Res Synth Methods 81:79–91

Hayashi F (2000) Econometrics. Princeton University Press, Princeton, NJ

Hedges LV (1981) Distribution theory for glass's estimator of effect size and related estimators. J Educ Stat 62:107–128

Hedges LV, Olkin I (1985) Statistical methods for meta-analysis. Elsevier, Amsterdam. https://doi.org/10.1016/c2009-0-03396-0

Hedges LV, Tipton E, Johnson MC (2010) Robust variance estimation in meta-regression with dependent effect size estimates. Res Synth Methods 11:39–65

Johnson VE (2013) Revised standards for statistical evidence. In: Proceedings of the national academy of sciences 11048:19313–19317

Kimmoun A, Takagi K, Gall E, Ishihara S, Hammoum P, El Bèze N (2021) Temporal trends in mortality and readmission after acute heart failure: a systematic review and meta-regression in the past four decades. Euro J Heart Fail 233:420–431

Knapp G, Hartung J (2003) Improved tests for a random effects meta-regression with a single covariate. Stat Med 2217:2693–2710

Knop ES, Pauly M, Friede T, Welz T (2023) The consequences of neglected confounding and interactions in mixed-effects meta-regression: an illustrative example. Res Synth Methods 144:647–651. https://doi.org/10.1002/jrsm.1643

Kontopantelis E, Reeves D (2012) Performance of statistical methods for meta-analysis when true study effects are non-normally distributed: a simulation study. Stat Methods Med Res 214:409–426
Li X, Dusseldorp E, Meulman JJ (2017) Meta-cart: a tool to identify interactions between moderators in meta-analysis. Br J Math Stat Psychol 701:118–136. https://doi.org/10.1111/bmsp.12088
Lin L, Aloe AM (2021) Evaluation of various estimators for standardized mean difference in meta-analysis. Stat Med 402:403–426
Linden AH, Hönekopp J (2021) Heterogeneity of research results: a new perspective from which to assess and promote progress in psychological science. Perspect Psychol Sci 162:358–376. https://doi.org/10.1177/1745691620964193
MacKinnon JG (2013) Thirty years of heteroskedasticity-robust inference: essays in honor of Halbert L. White Jr. In: Chen X, Swanson N (eds) Recent advances and future directions in causality, prediction, and specification analysis. Springer, New York, pp 437–461
MacKinnon JG, White H (1985) Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. J Econ 293:305–325
Morris TP, White IR, Crowther MJ (2019) Using simulation studies to evaluate statistical methods. Stat Med 3811:2074–2102
Noguchi K, Konietschke F, Marmolejo-Ramos F, Pauly M (2021) Permutation tests are robust and powerful at 0.5% and 5% significance levels. Behav Res Methods 536:2712–2724
Raudenbush SW (2009) Analyzing effect sizes: random-effects models. Handb Res Synth Meta-Anal 2:295–316
Sidik K, Jonkman JN (2005a) A note on variance estimation in random effects meta-regression. J Biopharm Stat 155:823–838. https://doi.org/10.1081/BIP-200067915
Sidik K, Jonkman JN (2005b) Simple heterogeneity variance estimation for meta-analysis. J R Stat Soc: Ser C (Appl Stat) 542:367–384
Sidik K, Jonkman JN (2006) Robust variance estimation for random effects meta-analysis. Comput Stat & Data Anal 5012:3681–3701. https://doi.org/10.1016/j.csda.2005.07.019
Sterchi M, Wolf M (2017) Weighted least squares and adaptive least squares: further empirical evidence. In: Vladik K, Songsak S, Van-Nam H (eds) Robustness in econometrics. Springer, Cham, pp 135–167
Tipton E (2015) Small sample adjustments for robust variance estimation with meta-regression. Psychol Methods 203:375–393. https://doi.org/10.1037/met0000011
Tipton E, Pustejovsky JE (2015) Small-sample adjustments for tests of moderators and model fit using robust variance estimation in meta-regression. J Educ Behav Stat 406:604–634. https://doi.org/10.3102/1076998615606099
Veroniki AA, Jackson D, Viechtbauer W, Bender R, Bowden J, Knapp G, Salanti G (2016) Methods to estimate the between-study variance and its uncertainty in meta-analysis. Res Synth Methods 71:55–79. https://doi.org/10.1002/jrsm.1164
Viechtbauer W, López-López JA (2022) Location-scale models for meta-analysis. Res Synth Methods 136:697–715. https://doi.org/10.1002/jrsm.1562
Viechtbauer W, López-López J, Sanchez-Meca J, Marín-Martínez F (2015) A comparison of procedures to test for moderators in mixed-effects meta-regression models. Psychol Methods 20:360
Welz T, Doebler P, Pauly M (2022) Fisher transformation based confidence intervals of correlations in fixed-and random-effects meta-analysis. Br J Math Stat Psychol 751:1–22
Welz T, Pauly M (2020) A simulation study to compare robust tests for linear mixed-effects meta-regression. Res Synth Methods 113:331–342. https://doi.org/10.1002/jrsm.1388
Welz T, Viechtbauer W, Pauly M (2023) Cluster-robust estimators for multivariate mixed-effects meta-regression. Comput Stat & Data Anal 179:107631
White H (1980) A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica 484:817–838
White IR (2015) Network meta-analysis. Stata J 154:951–985
Zimmermann G, Pauly M, Bathke AC (2020) Multivariate analysis of covariance with potentially singular covariance matrices and non-normal responses. J Multivar Anal 177:104594

Authors and Affiliations

Maria Thurow[1,2] · **Thilo Welz**[1] · **Eric Knop**[1] · **Tim Friede**[3] · **Markus Pauly**[1,2]

✉ Thilo Welz
welz@statistik.tu-dortmund.de

[1] Department of Statistics, TU Dortmund University, Vogelpothsweg 87, 44227 Dortmund, Germany

[2] Research Center Trustworthy Data Science and Security, UA Ruhr, Otto-Hahn-Straße 14, 44227 Dortmund, Germany

[3] Medizinische Statistik, Universitätsmedizin Göttingen, Humboldtallee 32, 37073 Göttingen, Germany