Received: 25 February 2019 | Revised: 9 October 2019 | Accepted: 13 November 2019

DOI: 10.1002/jrsm.1388

RESEARCH ARTICLE

Research Synthesis Methods WILEY

A simulation study to compare robust tests for linear mixed-effects meta-regression

Thilo Welz | **Markus Pauly**

Faculty of Statistics, Technical University of Dortmund, Dortmund, Germany

Correspondence
Thilo Welz, Technische Universität Dortmund Joseph-von-Fraunhofer-Straße 2-4, A 3.06 44227 Dortmund, Germany.
Email: thilo.welz@tu-dortmund.de

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Number: PA-2409 7-1

The explanation of heterogeneity when synthesizing different studies is an important issue in meta-analysis. Besides including a heterogeneity parameter in the statistical model, it is also important to understand possible causes of between-study heterogeneity. One possibility is to incorporate study-specific covariates in the model that account for between-study variability. This leads to linear mixed-effects meta-regression models. A number of alternative methods have been proposed to estimate the (co)variance of the estimated regression coefficients in these models, which subsequently drives differences in the results of statistical methods. To quantify this, we compare the performance of hypothesis tests for moderator effects based upon different heteroscedasticity consistent covariance matrix estimators and the (untruncated) Knapp-Hartung method in an extensive simulation study. In particular, we investigate type 1 error and power under varying conditions regarding the underlying distributions, heterogeneity, effect sizes, number of independent studies, and their sample sizes. Based upon these results, we give recommendations for suitable inference choices in different scenarios and highlight the danger of using tests regarding the study-specific moderators based on inappropriate covariance estimators.

KEYWORDS
heteroscedasticity, meta-regression, robust covariance estimation, standardized mean difference

1 | INTRODUCTION

Recently, Jackson and White (2018) raised the question "When should meta-analysis avoid making hidden normality assumptions?" In the current paper, we investigate this in the context of meta-regression models while also studying the effect of employing different methods to account for heteroscedasticity. Here, the notion meta-regression refers to a regression, in which the effect sizes from various studies are modeled by means of certain study characteristics. Thus, the effect sizes are the dependent (or outcome) variables and the study characteristics are the independent variables (also called moderators or explanatory variables).

As the effect sizes are usually certain summary statistics within diverse studies (as, eg, Cohen's d or a log-odds ratio), the study-specific moderators can only account for a part of the between-study heterogeneity. Thus, to "fully" account for heterogeneity, the introduction of a random effect is necessary, naturally leading to linear mixed-effects regression models. This was, for example, proposed[1] for the case of a single covariate and later extended.[2-7] In this context, a specific question of interest is to test for an effect of a certain

moderator, that is, to test the null hypothesis whether the corresponding regression coefficient is zero. Here, Viechtbauer et al made a thorough comparison of different existing methods in extensive simulations. In particular, they compared tests based on the Wald-type, Knapp-Hartung (with and without truncation), Permutation, Huber-White, and the likelihood ratio method together with seven different estimators of the so-called between-study heterogeneity. It turned out that the choice of heterogeneity estimator did not affect the results greatly, while the choice of methods mattered: They found a certain preference for the Knapp-Hartung method[3] and also concluded that "Huber-White and likelihood ratio tests (...) cannot be recommended for routine use, at least in their present form." Moreover, they stressed that "additional simulations are needed to assess the performance (...) under more adverse conditions, such as non-normal random errors and/or true effects." In the current paper, we follow this suggestion and continue their work by investigating the effect of non-normal random effects. In addition, we analyze the effect of choosing different versions of the Huber-White heteroscedasticity consistent (HC) covariance estimators. These estimators are typically applied when the assumption of homogeneous variance of the residuals is not plausible, to avoid inconsistent inference. In particular, there exist the six versions HC_0-HC_5 of the Huber-White estimator for regression models, of which Sidik and Jonkman[8] proposed the HC_0 and HC_1-type in the meta-analytic context. For fixed-effects regression models, the estimators HC_3 and HC_4 are often recommended.[9,10] Thus, it is of interest to also investigate the influence of the different choices in the context of meta-regression models. This becomes especially important under adverse conditions, such as non-normally distributed effect sizes and/or unbalanced study sizes or arms. As already shown,[11] such circumstances can lead to poor control of type 1 error and/or poor coverage of confidence intervals when using standard meta-analytic techniques. For this paper, we therefore investigate the performance of the different estimators in different scenarios, utilizing both standardized mean differences and log-odds-ratios as effect measures.

In the following sections, we start with a formal introduction of the mixed-effects meta-regression model and introduce inference procedures for testing moderator effects (Section 2). Next, we focus on a motivational data analysis (Section 3) that illustrates the practical importance of the choice of covariance estimator and we analyze the data example using the previously introduced procedures. The data analysis motivates the need for an extensive simulation study (Section 4). In this section, we explain the various simulation designs and illustrate and discuss our main findings. We end with concluding remarks and an outlook for further research (Section 5).

2 | THE SETUP

The usual mixed-effects meta-regression model is given for independent outcome/effect variables

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im} + u_i + \varepsilon_i, \quad i = 1, \ldots, K \qquad (1)$$

where x_{ij} denotes the jth moderator variable in the ith study, β_j is the corresponding model coefficient, and K the number of independent studies. Furthermore, u_i is a random effect that is typically assumed to be normally distributed[12] with $u_i \sim N(0, \tau^2)$ and ε_i is the *within-study* error with distribution $\varepsilon_i \sim N\left(0, \sigma_i^2\right)$. However, to give answers on the opening question of "When should meta-analysis avoid making hidden normality aumptions?," we also study non-normal situations regarding the random effects u_i: We do not specify a particular distribution and only assume $\mathbb{E}(u_i) = 0$ and $\mathrm{Var}(u_i) = \tau^2$. From a practical point of view, u_i accounts for the variability not explained by the trial-specific moderators, leading to the notion of *between-study heterogeneity* for its variance τ^2. We point out here that the study-level outcome of each individual patient may be binary. In this case, inference is based on normal approximations to discrete (binomial) likelihoods. Caution should be used with such normal approximations, as highlighted by a recent discussion paper on the topic of hidden normality assumptions in meta-analysis.[13] Here, an alternative approach would be exact GLMM approaches, as considered by Stijnen et al and others.[14,15]

Anyhow, model (1) involves several unknown parameters $\left(\sigma_i^2, \boldsymbol{\beta}, \tau^2\right)$, which have to be estimated. Thereof, the *within-study* sampling variance σ_i^2 is estimated from the observations in the study and typically assumed to be known. To provide a simple expression of the weighted least-squares estimate for $\boldsymbol{\beta}$ and the corresponding covariance estimators presented below, we rewrite model (1) in matrix notation as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} + \boldsymbol{\varepsilon}, \qquad (2)$$

where $\mathbf{X} \in \mathbb{R}^{K \times (m+1)}$, $\boldsymbol{\beta} \in \mathbb{R}^{m+1}$, and $\mathbf{u}$, $\boldsymbol{\varepsilon} \in \mathbb{R}^K$. The weighted least-squares estimator for $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\hat{\boldsymbol{W}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\hat{\boldsymbol{W}}\boldsymbol{y}. \qquad (3)$$

The weight matrix is $\hat{\boldsymbol{W}} = \mathrm{diag}\left(\left(\sigma_i^2 + \hat{\tau}^2\right)^{-1}\right)$. In this setup, we are now interested in testing the null hypothesis of no moderator effect

$$H_0 : \left\{\beta_j = 0\right\} \quad \text{for} j \in \{1, ..., m\}$$

against two-sided alternatives H_1: $\{\beta_j \neq 0\}$.

There already exist several procedures applicable for this purpose and most of them are mainly based on a test statistic of (Welch)-*t*-type. In particular, these basically differ in how both, the between-study heterogeneity τ^2 as well as the within-study variances σ_i^2, are accounted for. To define them, denote by $\hat{\boldsymbol{\beta}}$ the weighted least-squares estimator for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma} = \text{cov}\left(\hat{\boldsymbol{\beta}}\right)$. For all choices of (co)variance estimator $\hat{\boldsymbol{\Sigma}}$ considered in part 2.1, a two-sided test statistic of *t*-type for testing for the presence of the *j*th model coefficient, that is, for inferring H_0: $\{\beta_j = 0\}$, is then calculated via

$$T_j = \frac{\hat{\beta}_j}{\sqrt{\hat{\boldsymbol{\Sigma}}_{jj}}}. \tag{4}$$

Here, $\hat{\boldsymbol{\Sigma}}_{jj}$ is the *j*th diagonal element of the covariance estimator $\hat{\boldsymbol{\Sigma}}$. For large *K*, the statistic T_j approximately follows a *t*-distribution with $K - m - 1$ degrees of freedom under the null hypothesis H_0.[16] Comparing $|T_j|$ against the $1 - \alpha/2$ quantile of the *t*-distribution with $K - m - 1$ degrees of freedom yields the corresponding test and *P* values. Under mild regularity conditions on the moderators, these tests are asymptotically correct. We summarize this in Theorem 1, which is given in the supplement along with a proof.

As has already been pointed out, the testing procedures are not greatly affected by the choice of residual heterogeneity estimator.[17] We therefore solely focus on one estimator for τ^2: the restricted maximum likelihood (REML) estimator, which was recently propagated as a good choice for continuous data.[18,19] Details regarding the REML estimator are presented in the Supplementary Materials (cf. Equation S8). Note that in this context, REML estimates are more suitable than naive ML estimates of variance components as the latter may have a negative bias.[20]

As we have fixed estimators for $\boldsymbol{\beta}$ and τ^2, we now turn to the question of how to estimate the covariance of the estimated model coefficient $\hat{\boldsymbol{\beta}}$, given in Equation (3). Here, the Knapp-Hartung method[3] has been recommended.[17] However, in case of semiparametric linear models, robust Huber-White estimators are often seen as a reasonable solution; especially when the type of heteroscedasticity is not specified.[9,10,21] As Viechtbauer et al[17] only investigated the HC_1 estimator of the six Huber-White estimators HC_0-HC_5, we complement their study by also investigating the other versions with respect to their applicability in meta-regression. To this end they are detailed in the next subsection. These HC-estimators are furthermore compared to the (untruncated) Knapp-Hartung method, which provided adequate control of the type 1 error rate in previous research.[17]

2.1 | Robust (Huber-White) approach

In semiparametric linear models, the assumption of homogeneous variance of the residuals is often not plausible, possibly leading to invalid inference from classical methods based on homoscedasticity. Here, the typical solution is to apply sandwich estimators. These are also known as Huber-White estimators, to recognize the contributions of Peter J. Huber and Halbert White.[22,23] In model (1), it especially makes sense to consider such estimators because the marginal variances $\sigma_i^2 + \tau^2$ of the effect size estimates are heteroscedastic. We are now interested in consistent estimators of the (co)variance matrix $\boldsymbol{\Sigma} = \text{cov}\left(\hat{\boldsymbol{\beta}}\right)$. The classical White-estimator of type HC_0 that was proposed Sidik and Jonkman[8] in the meta-analytic context is given by

$$\hat{\boldsymbol{\Sigma}}_0 = \left(\boldsymbol{X}'\hat{\boldsymbol{W}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\hat{\boldsymbol{W}}\hat{\boldsymbol{E}}^2\hat{\boldsymbol{W}}\boldsymbol{X}\left(\boldsymbol{X}'\hat{\boldsymbol{W}}\boldsymbol{X}\right)^{-1}, \tag{5}$$

where $\hat{\boldsymbol{E}} = \text{diag}\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)$. Multiplying it with $K/(K - m - 1)$ leads to the HC_1-type estimator, which was considered in the above mentioned work by Viechtbauer et al[17] and is given by $\hat{\boldsymbol{\Sigma}}_1 = K\hat{\boldsymbol{\Sigma}}_0/(K - m - 1)$, which is known to be more conservative. However, even in classical regression models Wald- or *t*-tests based on both (co)variance estimators are known to yield inflated type 1 error rates for small to moderate sample sizes.[10,24,25] This was also shown to be the case in meta-regression models.[17] Therefore, improved versions of the original Huber-White estimator have been suggested, namely White estimators of type HC_2, HC_3, HC_4, and HC_5. We introduce these estimators but refer to the papers in which they were originally discussed for further details.[26-28] As their general forms are rather complex (cf. Equations 5 and 6), we have also worked out the analytical form of the HC estimators in the simplest case of no moderators, that is, random-effects meta-analysis. Please refer to the Supplementary Material and the discussion for details. The form of the respective Huber-White covariance estimators in the context of the mixed-effects meta-regression model (2) is described below: we first introduce the HC_2 and HC_3 estimators given by

$$\text{HC}_\ell = \hat{\boldsymbol{\Sigma}}_\ell = \left(\boldsymbol{X}'\hat{\boldsymbol{W}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\hat{\boldsymbol{W}}\hat{\boldsymbol{E}}_\ell^2\hat{\boldsymbol{W}}\boldsymbol{X}\left(\boldsymbol{X}'\hat{\boldsymbol{W}}\boldsymbol{X}\right)^{-1}, \quad \ell = 2, 3. \tag{6}$$

Here, $\hat{\boldsymbol{E}}_2 = \text{diag}\left(\left(1-x_{jj}\right)^{-1/2}\right)\cdot\hat{\boldsymbol{E}}$ and $\hat{\boldsymbol{E}}_3 = \text{diag}\left(\left(1-x_{jj}\right)^{-1}\right)\cdot\hat{\boldsymbol{E}}$, where x_{jj} is the jth diagonal element of the hat matrix $\boldsymbol{X}\left(\boldsymbol{X}'\hat{\boldsymbol{W}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\hat{\boldsymbol{W}}$. Thereof, the HC_3 estimator gives a very close approximation to the computationally more expensive jackknife estimator described in Reference 26 and given by

$$\text{HC}_3^{JK} = \hat{\boldsymbol{\Sigma}}_3^{JK} = \frac{K-1}{K}\sum_{t=1}^{K}\left(\hat{\boldsymbol{\beta}}_{(t)} - \frac{1}{K}\sum_{s=1}^{K}\hat{\boldsymbol{\beta}}_{(s)}\right)\left(\hat{\boldsymbol{\beta}}_{(t)} - \frac{1}{K}\sum_{s=1}^{K}\hat{\boldsymbol{\beta}}_{(s)}\right)'. \tag{7}$$

Here, $\hat{\boldsymbol{\beta}}_{(i)}$ is the weighted least-squares estimate of $\boldsymbol{\beta}$ based on all observations except the ith. It is important to note that HC_3, unlike HC_2, is biased under homoscedasticity.[28] To improve HC_3, the following variation was suggested[27]:

$$\text{HC}_4 = \hat{\boldsymbol{\Sigma}}_4 = \left(\boldsymbol{X}'\hat{\boldsymbol{W}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\hat{\boldsymbol{W}}\hat{\boldsymbol{E}}_4^2\hat{\boldsymbol{W}}\boldsymbol{X}\left(\boldsymbol{X}'\hat{\boldsymbol{W}}\boldsymbol{X}\right)^{-1}, \tag{8}$$

where $\hat{\boldsymbol{E}}_4 = \text{diag}\left(\left(1-x_{ii}\right)^{-\frac{\delta_i}{2}}\right)\cdot\hat{\boldsymbol{E}}$ and $\delta_i = \min\left\{4, \frac{x_{ii}}{\bar{x}}\right\}$. Finally, there is

$$\text{HC}_5 = \hat{\boldsymbol{\Sigma}}_5 = \left(\boldsymbol{X}'\hat{\boldsymbol{W}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\hat{\boldsymbol{W}}\hat{\boldsymbol{E}}_5^2\hat{\boldsymbol{W}}\boldsymbol{X}\left(\boldsymbol{X}'\hat{\boldsymbol{W}}\boldsymbol{X}\right)^{-1}, \tag{9}$$

where $\hat{\boldsymbol{E}}_5 = \text{diag}\left(\left(1-x_{ii}\right)^{-\frac{\alpha_i}{2}}\right)\cdot\hat{\boldsymbol{E}}$ and $\alpha_i = \min\left\{\frac{x_{ii}}{\bar{x}}, \max\left\{4, \frac{\gamma x_{\max}}{\bar{x}}\right\}\right\}$ with a predefined constant $0 < \gamma < 1$. Based on findings from simulation studies, the value $\gamma := 0.7$ was recommended.[28] We follow this suggestion below.

The asymptotic behavior (for large K) is the same for all of the considered covariance estimators. However, for small to moderate numbers of studies K, the respective behavior may be vastly different, as asymptotic arguments and limit theorems no longer hold. This is particularly apparent in the illustrative data example presented in the next section.

3 | DATA EXAMPLE

Table 1 contains data on six studies, which investigate the effectiveness of Azithromycin vs Amoxycillin or Amoxycillin/clavulanic acid (Amoxyclav) in the treatment of acute lower respiratory tract infections. An explanation of the different variables can be found in Table 2. Azithromycin is an antibiotic, which is useful for the treatment of various bacterial infections.[29] The data are contained in the **R** package **metafor** and have previously been analyzed.[30] We want to investigate whether the respective trial having included patients

TABLE 1 Data collected to investigate effectiveness of Azithromycin vs Amoxycillin or Amoxyclav in the treatment of acute lower respiratory tract infections

Study	Author	Year	ai	n1i	ci	n2i	Age	diag.ab	diag.cb	diag.pn	ctrl	bi	di	mod	$\hat{\theta}_i$	vi
1	Balmes	1991	4.50	48	7.50	56	Adults	1	0	0	Amoxyclav	44.50	49.50	0	−0.40	0.40
2	Biebuyck	1996	53.50	497	53.50	257	Adults	1	1	0	Amoxyclav	444.50	204.50	0	−0.78	0.04
3	Daniel	1991	5.50	121	10.50	120	Adults	1	0	0	Amoxycillin	116.50	110.50	0	−0.70	0.29
4	Gris	1996	6.50	34	2.50	33	Adults	1	1	1	Amoxyclav	28.50	31.50	1	1.06	0.62
5	Hoepelman	1993	4.50	48	4.50	51	Adults	1	0	0	Amoxyclav	44.50	47.50	0	0.07	0.49
6	Zachariah	1996	8.50	173	7.50	173	Adults	1	1	1	Amoxyclav	165.50	166.50	1	0.13	0.26

TABLE 2 Explanation of variables in Table 1

Variable	Meaning
ai	Number of clinical failures in the group treated with Azithromycin
n1i	Number of patients in the group treated with Azithromycin
ci	Number of clinical failures in the group treated with amoxycillin or amoxyclav
n2i	Number of patients in the group treated with amoxycillin or amoxyclav
age	Whether the trial included adults or children
diag.ab	Trial included patients with a diagnosis of acute bacterial bronchitis
diag.cb	Trial included patients with a diagnosis of chronic bronchitis with acute exacerbation
diag.pn	Trial included patients with a diagnosis of pneumonia
ctrl	Antibiotic in control group (amoxycillin or amoxyclav)
bi	n1i - ai
di	n2i - ci
mod	1 {diag.ab == 1 & diag.pn == 1}
$\hat{\theta}_i$	Estimated effect (here the log-odds-ratio)
vi	Sampling variance

with a diagnosis of pneumonia has a significant effect on the effectiveness of Azithromycin within the subgroup of trials containing patients with a diagnosis of acute bacterial bronchitis. We will attempt to answer this question using a mixed-effects meta-regression model.

Although in the original work on these data[30] the authors used risk ratios as the effect measure, we decided to utilize the log-odds ratio as the effect measure of choice, due to its favorable statistical properties, such as an approximate normal distribution.[31] Moreover, the log-odds ratio behaved similarly to the standardized mean difference in our preliminary simulations. The resulting *P*-values and test statistic values (4) for each choice of estimator HC_i, $i = 0, ..., 5$ and the Knapp-Hartung method are given in Table 3.

The estimators HC_0 and HC_1 lead to a rejection of the null hypothesis at nominal level $\alpha = 0.05$, while the test based on HC_2 still leads to a significant moderator effect at the 10% level. On the contrary, tests based on HC_3-HC_5 do not reject the null hypothesis. If we compare the newer covariance estimators HC_3-HC_5 and HC_{KH}, the Knapp-Hartung method rejects the null hypothesis at the nominal level α, whereas the formerly mentioned methods do not.

These results illustrate that the choice of covariance estimator can have a large influence on results in practice and that the wrong choice of HC-estimator may lead to possibly false-positive or -negative test results. In particular, it is unclear whether the above rejections/non-rejections are due to a potentially liberal/conservative behavior or different power characteristics of the corresponding tests. In any case, researchers should take care when performing inference on study-specific moderators, especially when the number of investigated studies

TABLE 3 Test statistics and *P*-values for the data example in Table 1 based on various HC-type covariance estimators and the Knapp-Hartung method

Estimator	$T_j = \frac{\hat{\beta}_j}{\sqrt{\hat{\Sigma}_{jj}}}$	$\sqrt{\hat{\Sigma}_{jj}}$	*P*-value
HC_0	3.777	0.288	***.019***
HC_1	3.084	0.352	***.037***
HC_2	2.423	0.449	*.073*
HC_3	1.434	0.758	.225
HC_4	1.367	0.795	.244
HC_5	1.367	0.795	.244
HC_{KH}	2.943	0.369	***.042***

K is small. In order to help guide researchers' decision of which covariance estimator to use in their analysis, we perform an extensive simulation study regarding type 1 error and power.

3.1 | Software

Although this data set was analyzed using the open source software `R`, other statistical software packages are available for meta-regression. Two examples are metareg in Stata as well as various procedures in SAS. Metareg in Stata, for example, implements the REML method as the default estimation procedure regarding the between-study variance τ^2. In both Stata and SAS, the covariance matrix estimation approach can be specified: Metareg implements the Knapp-Hartung method as the default covariance estimation approach. In SAS, the PROC

PANEL procedure per default uses the standard sample covariance estimator but allows the option to specify one of the HC covariance estimators HC_0-HC_4 using the HCCME= option in the MODEL statement.[32,33] In the **rma** function in the **metafor** package in **R**, the default covariance matrix is simply $\mathbf{V} = \text{diag}(\sigma_i^2 + \hat{\tau}^2)$ with REML as the default estimation procedure for the between-study heterogeneity τ^2. The Knapp-Hartung method can be specified via the option test = "knha" in the **rma** function.

4 | SIMULATION STUDY

We conducted a Monte Carlo simulation using standardized mean differences and log-odds-ratios as the effect size measures. As we do not want to assume individual patient data, only the study effects $\hat{\theta}_i$ are available (cf. Equation 10). As in previous work,[17] we assumed a single moderator influencing the true study-specific effects resulting in the model

$$\theta_i = \beta_0 + \beta_1 x_i + u_i. \tag{10}$$

The values of the moderator x_i were independently generated from a standard normal distribution and without loss of generality, β_0 was set to 0. Moreover, the random effects u_i were chosen to be either standard normal-, (standardized) exponential-, double exponential-, log-normal-, or t_3-distributed. For a detailed definition of the corresponding data generating processes, we refer to Section S6 in the Supplementary Material.

For the effect size, we considered the standardized mean difference in the ith study. We generated the true parameter θ_i directly, analogously to Viechtbauer et al,[17] according to Equation (10). An unbiased estimator of θ_i is given by Hedges' g[34]

$$g_i = \left(1 - \frac{3}{4(n_i^T + n_i^C) - 9}\right) d_i, \tag{11}$$

where n_i^T and n_i^C denote the size of treatment and control group, respectively, which are specified below. Moreover, d_i denote the effect size estimates (Cohen's d) from study i which were generated via

$$d_i = \phi_i / \sqrt{X_i / n_i},$$

where $\phi_i \tilde{N}\left(\theta_i, 1/n_i^T + 1/n_i^C\right)$, $X_i \tilde{\chi}^2_{n_i}$ with $n_i = n_i^T + n_i^C - 2$ and then applying expression (11).

For the between-study heterogeneity τ^2, we chose the values {0.1, 0.2, ..., 0.9} and for β_1 we considered the choices {0, 0.2, 0.5}, where 0 corresponds to no effect of the moderator variable. The number K of independent studies was chosen from {5, 10, 20, 50}. Finally, a good approximation of the sampling variance of y_i is given by[35]

$$v_i = 1/n_i^T + 1/n_i^C + \frac{g_i^2}{2(n_i^T + n_i^C)}. \tag{12}$$

In order to see if and in what way the results depended on the chosen effect size measure, we also investigated log-odds ratios for binary data. Simulating data in a manner analogous to the one described in foundational work,[8,12] (results not shown) it turned out that the change of effect size did not alter the general conclusion. Therefore, we focus on the standardized mean difference alone.

Regarding study size, we considered balanced experimental and control groups, that is, $n_i^T = n_i^C$. We then considered the case of equal study sizes ($n_i^T \equiv \eta$ for some η) and unbalanced study sizes. In the former case, we simulated the values $\eta \in$ {5, 10, 20, 40, 80} and in the latter we chose the study size vectors (6, 8, 9, 10, 42), (16, 18, 19, 20, 52), and (41, 43, 44, 45, 77) in accordance with previous work.[17] For $K > 5$, these study size vectors were simply repeated accordingly, for example, for $K = 10$ a study size vector might be (6, 8, 9, 10, 42, 6, 8, 9, 10, 42).

In total, we simulated 30, 240 $= 9(\tau^2) \times 3(\beta_1) \times 4$ $(K) \times 8(n_i) \times 5(u_i) \times 7$ (6 HC and Knapp-Hartung) different configurations with $N = 1000$ simulation runs, respectively. The simulation study was conducted in **R**, using the **metafor** package.[36] All tests were performed with a nominal significance level of $\alpha = 0.05$.

In practice, the study-specific moderators are oftentimes binary, as can be seen in our data example. For this reason, we have also (exemplarily) considered binary moderators in the case of balanced study sizes, considering normal and exponential random effects. So, instead of generating the x_{1i} from a $\mathcal{N}(0,1)$ distribution, we generated them from a Bernoulli distribution with parameter $P = .2$. It is necessary to exclude the case where all moderators are equal to 1 or 0. Furthermore, it is sufficient to consider only power for the binary moderators, as the type 1 error will be the same as in the case of standard normally generated moderators because for $\beta_1 = 0$ the choice of x_{1i} does not matter.

5 | RESULTS

In this section, we describe the results of the simulation study. In particular, we present type 1 error and power based on the different covariance estimators under

various simulation configurations. For power, we considered both the case of a (comparatively) smaller effect size $\beta_1 = 0.2$ and a (comparatively) larger effect $\beta_1 = 0.5$ of the study-specific moderator. For ease of presentation, we focus on the most important results and general trends and refer the interested reader to the Supplementary Material for the complete simulation results.

5.1 | Type 1 error rate

Studying the type 1 error results for all configurations given in the Supplementary Material, we can draw the first general conclusion that changes in the between-study heterogeneity τ^2, the number of subjects in each study and the underlying distributions of the random effects had little effect on the behavior of the procedures under the null hypothesis. In comparison, the number of studies K and the chosen test procedure were the driving forces for changes in type 1 error control. We therefore start by presenting a summary of the results of type 1 error simulations for different combinations of these two forces in boxplots given in Figure 1. The results shown in Figure 1 are for the scenario of unequal study sizes. We present results for HC_1-HC_5 and the Knapp-Hartung method, referring HC_0 to the Supplementary Material, due to its known liberal behavior.

Here, each boxplot represents the $9(\tau^2) \times 3(n_i) \times 5(u_i) = 135$ different empirical type 1 error rates for each test in case of $K \in \{5, 10, 20, 50\}$ studies. The White-type test based on the classical HC_0-estimator exhibits highly inflated type 1 error rates, as expected; particularly for a smaller number of studies. The type 1 error rates are even more inflated than for HC_1. For details we refer to the Supplementary Material. A similar, but less pronounced behavior can be observed for the tests based upon HC_1 and HC_2. On the contrary, all other procedures control the nominal level $\alpha = 0.05$ quite well. HC_3-HC_5 are slightly conservative for $K = 5$ studies. HC_3 has a type 1 error around 3% and HC_4 and HC_5 around 4% for $K = 5$. For these three estimators, the type 1 error converges to the nominal level α for increasing number of studies K. The Knapp-Hartung method holds the nominal level exactly for $K = 5$ studies but seems to become (only slightly) conservative for increasing number of studies K. It is interesting to note that there was no significant correlation between type 1 error and different study sizes n (for a fixed number of studies K), see the Supplement for details. Finally, the Knapp and Hartung method controlled the nominal level α very well for a smaller number of studies $K \in \{5, 10\}$, which is in line with previous research.[17] On the contrary, the other HC estimators were either liberal or slightly conservative in the scenario of $K = 5$ studies.

For a better comparison of the procedures with the overall best type 1 error control (HC_3-HC_5 and HC_{KH}), we present the boxplots of their simulated type 1 error rates together in one figure (see Figure 2). The results shown are from the simulation configuration of unbalanced study sizes and the standardized mean difference as effect measure.

Figure 2 summarizes the observed type 1 error rates. These are fairly close to the nominal level $\alpha = 5\%$, albeit being slightly conservative at the median with median type 1 error rates between 4% and 5%. The exception is the HC_3 estimator in the case of five studies, which is much more conservative with a median type 1 error rate just below 3% and the entire boxplot has whiskers lying below the nominal level α. For HC_3-HC_5, the type 1 error rates increase monotonically toward the nominal level for an increasing number of studies K, and for the Knapp-Hartung method the type 1 error rates start close to nominal for the case of $K = 5$ studies and decrease (slowly) away from the nominal level for increasing numbers of studies K.

Based on these results, we conclude that for ≤10 studies the Knapp-Hartung method is to be recommended (in terms of type 1 error control) and for the case of ≥20 studies, especially when the number of studies is very large, for example, 50 studies as in Figure 2, HC_3 is the preferred estimator with regards to type 1 error control. For the case of $10 < K < 20$ studies, further simulations need to be done in order to give a clear recommendation for the choice of covariance estimator. For more comprehensive recommendations, we compare the procedures' power behavior in the next section.

5.2 | Power

In addition to the type 1 error rate, we investigated the power of the respective tests to reject the null hypothesis of no effect of the moderator variable, when it is in fact false. To this end, we consider alternatives with (comparatively) small and (comparatively) larger effects by setting $\beta_1 = 0.2$ and $\beta_1 = 0.5$, respectively.

For all methods, the observed general trend was that power increased monotonically for decreasing amounts of heterogeneity τ^2, increasing number of studies K as well as increasing study size n. In the following, we again concentrate on power for the procedures based on HC_3-HC_5 and Knapp-Hartung, as these were the only tests with a satisfactory type 1 error control. The detailed power simulation results, for each separate simulation scenario, for these, and all other methods are given in Section S6.2 of the Supplement. As the results for heterogeneous and homogeneous study sizes were very similar,

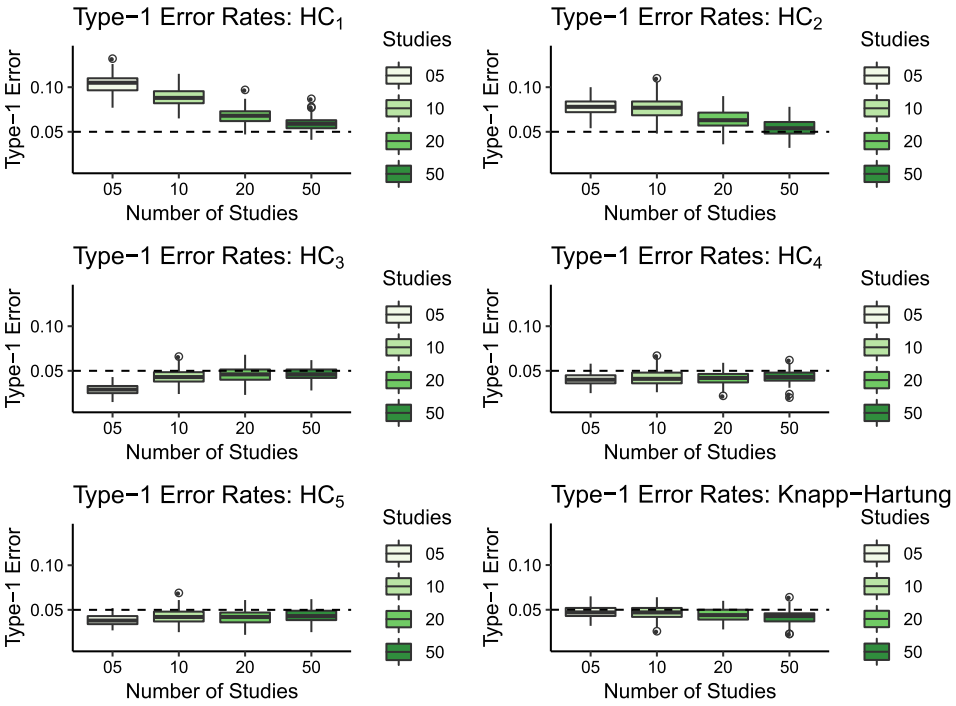

FIGURE 1 Type 1 error of tests based on the White-type estimators HC_1-HC_5 and the Knapp-Hartung correction HC_{KH} for varying number of studies $K \in \{5, 10, 20, 50\}$ and $\tau^2 \in \{0.1, 0.2, ..., 0.9\}$—with unbalanced study sizes and standardized mean difference (SMD) as effect measure. Each boxplot represents 135 type 1 error rates. For detailed individual simulation results, please refer to the Supplement [Colour figure can be viewed at wileyonlinelibrary.com]

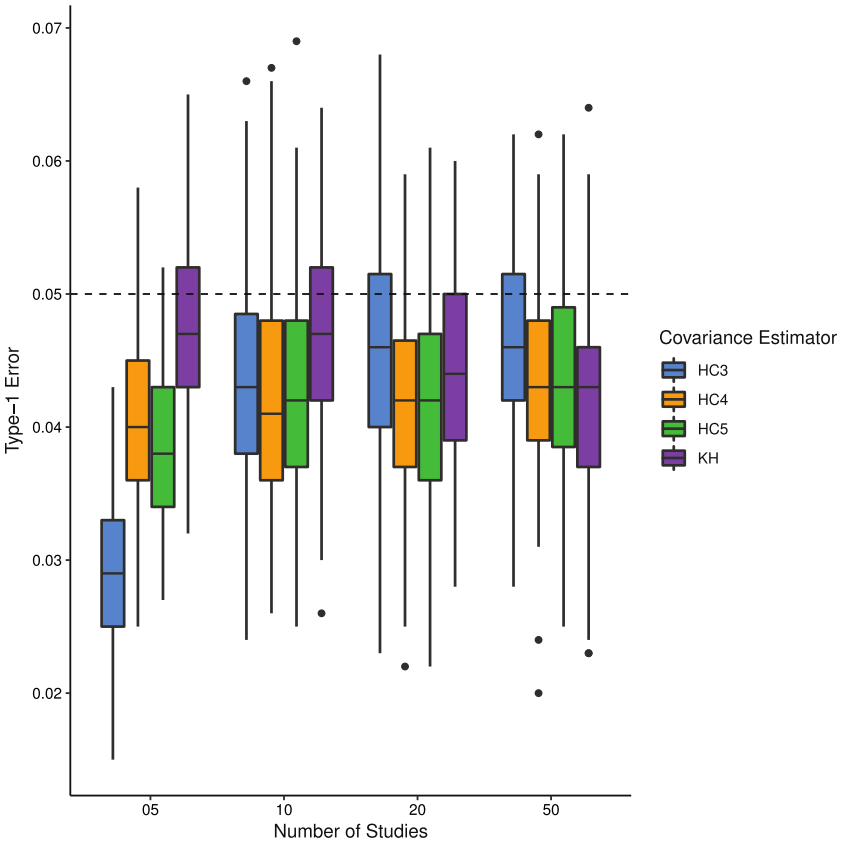

FIGURE 2 Type 1 error based on HC_3-HC_5 and HC_{KH} for $K \in \{5, 10, 20, 50\}$, $\tau^2 \in \{0.1, 0.2, ..., 0.9\}$—with unbalanced study sizes and standardized mean difference (SMD) as effect measure [Colour figure can be viewed at wileyonlinelibrary.com]

we restrict ourselves to the former case and again refer to the Supplement (Section S6.2) for the complete results.

Figure 3 summarizes the power results for the tests based on HC_3-HC_5 and HC_{KH} for a (comparatively) small effect size of $\beta_1 = 0.2$, in the scenario of unbalanced study sizes. Median power ranges from around 5% to 8% for $K = 5$ to around 45% to 47% for $K = 50$. For larger amounts of studies, the power of all shown tests is close together. However, HC_3 does seem to have slightly more median power than the other estimators for $K = 50$. In

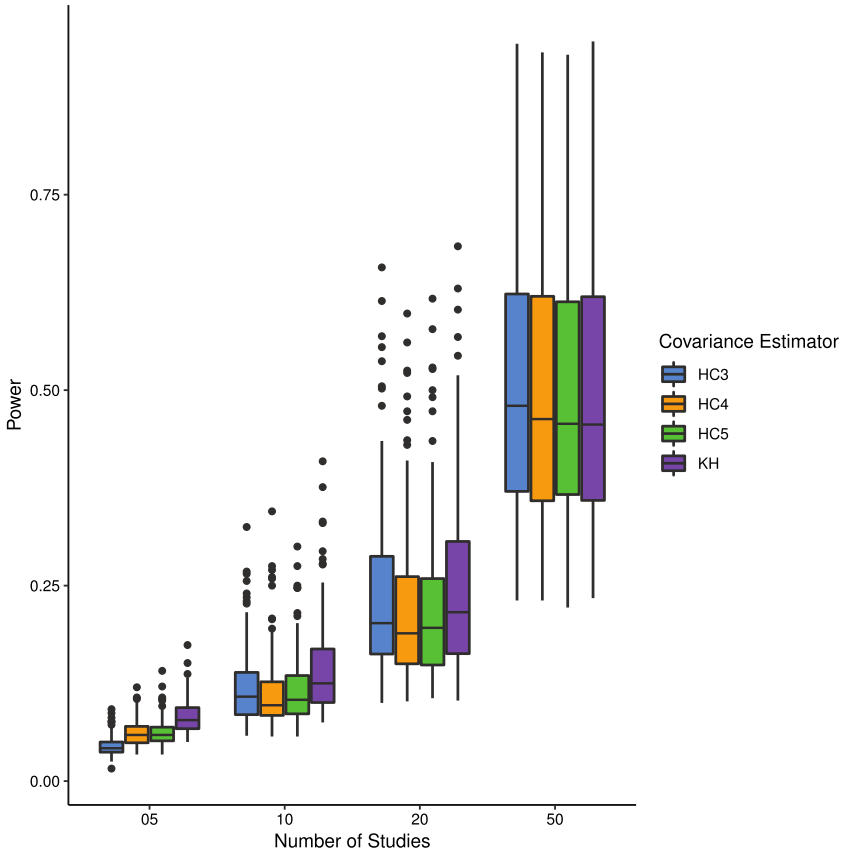

FIGURE 3 Power of tests based on HC_3-HC_5 and HC_{KH} for $K \in \{5, 10, 20, 50\}$, $\tau^2 \in \{0.1, 0.2, ..., 0.9\}$, and $\beta_1 = 0.2$—with unbalanced study sizes and standardized mean difference (SMD) as effect measure [Colour figure can be viewed at wileyonlinelibrary.com]

the scenario of $K = 5$ studies, the Knapp-Hartung method yields much greater power than HC_3 and slightly more power than HC_4 and HC_5. For $K = 10$ and $K = 20$ studies, HC_{KH} has slightly more median power than HC_3-HC_5, as well as having a longer "upper whisker" in the latter case, in comparison to the other methods.

The results for a (comparatively) larger effect of $\beta_1 = 0.5$ can be found in the Supplementary Materials. We again concentrate on the estimators HC_3-HC_5 and HC_{KH}. In the Supplement, we also give the power results for the scenario of balanced study sizes. For $\beta_1 = 0.5$, the difference between methods is more pronounced; especially for a smaller number of studies $K \in \{5, 10\}$. In fact, HC_{KH} has considerably more power than HC_3-HC_5 for $K \in \{5, 10\}$. At the median this difference amounts to 7%-8% more power than HC_3 and around 4% more than HC_4 and HC_5 for $K = 5$ and around 7%-8% more power than HC_3-HC_5 for $K = 5$. For larger study sizes, this effect diminishes and the results are quite close together. Results were very similar for balanced study sizes.

5.3 | Bias and variance estimation

In addition to type 1 error and power, we also study the bias $\mathbb{E}[\hat{\beta}_1] - \beta_1$ and the variance $\text{var}(\hat{\beta}_1) = \Sigma_{11}$ of the effect estimator of $\hat{\beta}_1$ in the Supplement. Clearly $\hat{\beta}_1$ is identical across all variations of variance estimator. Because these values cannot be expressed analytically, we resorted to simulations, which we performed in the scenario of normally distributed random effects and balanced study sizes with moderator variables drawn from a normal distribution. Our findings can be summarized as follows: The estimator $\hat{\beta}_1$ is approximately unbiased for $\beta_1 = 0$ and becomes increasingly negatively biased for larger effect sizes β_1. Moreover, the variance seems to increase with each new version of the HC estimator, that is, from HC_0 to HC_5. The Knapp-Hartung method, however, has a smaller variance than the newer iterations of the HC estimators HC_3-HC_5. The details can be found in the Supplement.

5.4 | Binary moderators

Finally, since moderators can also be binary in practice, we extended the simulations to consider this scenario. The results of the power simulations with binary moderators indicate that use of binary covariates instead of continuous ones reduces power considerably. Furthermore, power did increase for larger numbers of studies K but much more slowly than in the case of continuous moderators. When comparing the power results of the different covariances estimators, it became apparent that the HC

estimators displayed vastly superior power over the Knapp-Hartung method when the number of studies was small ($K \leq 10$). This is interesting, as with continuous moderators Knapp-Hartung often had more power. For large numbers of studies ($K = 50$), Knapp-Hartung had slightly more power than the HC estimators. It therefore seems prudent to use one of the newer HC estimators (HC_3-HC_5) instead of the Knapp-Hartung method when dealing with binary moderators and a small number of studies K. However, if dealing with binary moderators and a large number of studies ($K > 20$), it is probably best to stick with the Knapp-Hartung method. Detailed results can be found in the Supplementary Material.

6 | DISCUSSION AND FURTHER RESEARCH

Mixed-effects meta-regression models offer a good possibility to describe and model moderator (covariate) effects from various studies in a meta-analysis. In this context, it is of interest to determine which moderators significantly help to explain heterogeneity. This naturally leads to t-tests for the null hypotheses of no moderator effects. Here, Viechtbauer et al[17] compared several procedures in extensive simulations and recommended the (untruncated) Knapp-Hartung method[3] as the procedure of choice. We complement their investigations by additionally considering all six robust covariance estimators of Huber-White (HC) type suggested in the literature, while also extending their simulation scenarios. In fact, following recent discussions on *hidden normality assumptions* in meta-analyses,[13] we also study situations with non-normal random effects. Although we focus on hypothesis tests for moderator effects, confidence intervals for the unknown regression coefficients based on t-quantiles can easily be constructed via test inversion.[37] The coverage probabilities of these confidence intervals would be given by 1 minus the respective type 1 error.

For a total of 30 240 different simulation configurations we compared the t-tests based on the six different HC-type estimators (HC_0-HC_5) and the (untruncated) Knapp-Hartung method[3] with respect to their type 1 error control and power. As observed in other regression contexts,[9,17,25,27,28] the tests based on the classical Huber-White estimators HC_0, HC_1 as well as HC_2 generally had a highly inflated type 1 error, except for the simulation scenario of $K = 50$ studies. Of the other existing modifications HC_3-HC_5, all managed a satisfactory control of the nominal level α. HC_4 and HC_5 controlled the nominal level more exactly, whereas the HC_3 estimator was conservative in the case of very few studies ($K = 5$), with an observed type 1 error of around 3%. The (untruncated) Knapp-Hartung method also controlled the nominal level α well, albeit being more exact for smaller numbers of studies and slightly conservative for a larger number of studies K.

Regarding the behavior under different alternatives, all tests' power tended to increase monotonically with increasing study numbers K, increasing average study size and decreasing amounts of heterogeneity τ^2—a marked difference when comparing to type 1 error behavior, where τ^2 and study size had little influence.

Somewhat surprisingly the choice of distribution of the random effects in the simulation study had hardly any effect on the type 1 error and power of t-tests based on the considered covariance estimators. This leads us to conclude that the typical normality assumption $u_i \sim N(0, \tau^2)$ for the mixed-model random effects is unproblematic, at least in the scenarios we considered in our simulation study.

Comparing HC_3-HC_5 and the Knapp-Hartung-method, we observed a higher power of the latter; especially in case of larger moderator effects or few studies. Only in case of small moderator effects and a larger number of studies ($K = 50$) a slight power advantage of the HC_3-method was observed. Nevertheless, our findings lead to similar conclusions as drawn in previous research[17] that in most cases the (untruncated) Knapp-Hartung method seems to be the procedure of choice.

In addition to meta-regression, we have considered the special case of no moderators (random-effects meta-analysis) and worked out the formulas for the individual HC-type variance estimators of the main effect $\hat{\theta}$ in this case. These results are presented in Proposition 1 of the technical Appendix in the Supplementary Material, along with a proof. Additionally, the individual formulas of the six HC estimators $\hat{\mathbf{\Sigma}}_0, \ldots, \hat{\mathbf{\Sigma}}_5$ of the form $\hat{\mathbf{\Sigma}}_\ell = \sum_{j=1}^{K} v_{j,\ell} \cdot \hat{\varepsilon}_j^2$, $\ell = 0, \ldots, 5$ for specific weights $v_{j,\ell}$ are presented in Equations (S2)-(S7) of the Supplement along with a numerical example. $\hat{\mathbf{\Sigma}}_0$ and $\hat{\mathbf{\Sigma}}_1$ only differ by a constant, whereas $\hat{\mathbf{\Sigma}}_2$-$\hat{\mathbf{\Sigma}}_5$ differ through the exponent of a weighting factor included in $v_{j,\ell}$. Please refer to the technical Appendix of the Supplementary Material for their explicit form.

In applications, one of the most problematic cases is when only a small number of studies are available. Our data example in Section 3 shows how large the influence of the choice of HC estimator can be in such a scenario. One possible reason may be that all considered estimators make direct use of the residuals. In case of few studies, this may not be too reliable, leading to less stable estimation of the between-study heterogeneity τ^2 and more variable SE. Here, alternative approaches exist, such as higher order likelihood based methods, which aim to improve on inference based on first order

likelihoods. In this context, some authors have, for example, recommended inference based on Skovgaard's second-order statistic.[38,39] Moreover, we additionally conjecture that for such a case of few studies the underlying error distribution plays an important role as well.[13] We leave an exhaustive evaluation of these "residual concerns" to future research.

We conclude this paper with an outlook on ongoing and future research. In most clinical trials, two or more endpoints of interest are measured. Therefore, the current investigations will be extended to the case of multivariate mixed-effects meta-regression models. As the assumption of normality is usually more problematic than in the univariate case,[40-43] an adequate treatment may require the extension and/or improvement of existing methods. In this context, the additional study of modern imputation techniques[44,45] will be mandatory. Moreover, different to the present setting one might explore the methodology under the presence of individual patient data, allowing the application of a multitude of different permutation or resampling procedures.[25,46,47]

ACKNOWLEDGMENTS

This work was supported by the German Research Foundation (Grant no. PA-2409 7-1). We would also like to thank the editor, the reviewers, and the associate editor for their constructive comments, which have helped to improve the quality of our work.

CONFLICT OF INTEREST

The authors reported no conflict of interest.

RECOMMENDATIONS

Based on the results of our simulation study, we give the following recommendations:

In general, we recommend the use of the Knapp-Hartung method. However, there are a few special cases, in which an HC-estimator may be superior. In particular, in the scenario of many studies ($K \geq 50$) and an effect size that is suspected to be "not too large", that is, $\beta_1 \leq 0.2$, the HC_3 estimator seems to yield slightly more power than the Knapp-Hartung method, with both controlling the nominal type 1 error level α well. Furthermore, when dealing with binary moderators and a small number of studies ($K \leq 10$), it seems that the modern HC estimators HC_3-HC_5 have more power than the Knapp-Hartung method, while controlling type 1 error and should therefore be preferred in this scenario.

If a researcher does decide to use one of the HC estimators HC_0-HC_5 then the estimators HC_0-HC_2 should *not* be used, mainly due to their inflated type 1 error behavior. The other three HC-estimators control the nominal type 1 error α well. When deciding between the HC-estimators HC_3-HC_5, the choice can be made based on the number of studies available. For $K \leq 10$ studies (especially for $K = 5$), HC_4 and HC_5 have more power than HC_3. However, for $K \geq 20$ studies, HC_3 yields slightly more power than the other two.

DATA AVAILABILITY STATEMENT

The (simulated) data that support the findings of this study can be generated using our openly available R-scripts. These files are made public on [figshare] at DOI: [10.6084/m9.figshare.10327274]. The data used in Section 3 are freely available in the metafor package of the open source software package R.

ORCID

Thilo Welz https://orcid.org/0000-0001-6223-5698

REFERENCES

1. Berkey CS, Hoaglin DC, Mosteller F, Colditz GA. A random-effects regression model for meta-analysis. *Stat Med*. 1995;14:395-411.
2. Berkey C, Hoaglin D, Antczak-Bouckoms A, Mosteller F, Colditz G. Meta-analysis of multiple outcomes by regression with random effects. *Stat Med*. 1998;17:2537-2550.
3. Knapp G, Hartung J. Improved tests for a random effects meta-regression with a single covariate. *Stat Med*. 2003;22:2693-2710.
4. Higgins JP, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *J R Stat Soc Ser A Stat Soc*. 2009;172:137-159.
5. Rota M, Bellocco R, Scotti L, et al. Random-effects meta-regression models for studying nonlinear dose–response relationship, with an application to alcohol and esophageal squamous cell carcinoma. *Stat Med*. 2010;29:2679-2687.
6. Huizenga HM, Visser I, Dolan CV. Testing overall and moderator effects in random effects meta-regression. *Br J Math Stat Psychol*. 2011;64:1-19.
7. Jackson D, Turner R, Rhodes K, Viechtbauer W. Methods for calculating confidence and credible intervals for the residual between-study variance in random effects meta-regression models. *BMC Med Res Methodol*. 2014;14:103.
8. Sidik K, Jonkman JN. Simple heterogeneity variance estimation for meta-analysis. *J R Stat Soc Ser C Appl Stat*. 2005;54:367-384.
9. Long JS, Ervin LH. Using heteroscedasticity consistent standard errors in the linear regression model. *Am Stat*. 2000;54:217-224.
10. Rosopa PJ, Schaffer MM, Schroeder AN. Managing heteroscedasticity in general linear models. *Psychol Methods*. 2013;18:335-351.
11. Pauly M, Welz T. Contribution to the discussion of "when should meta-analysis avoid making hidden normality assumptions?". *Biom J*. 2018;60:1075-1076.
12. Viechtbauer W. Confidence intervals for the amount of heterogeneity in meta-analysis. *Stat Med*. 2007;26:37-52.
13. Jackson D, White IR. When should meta-analysis avoid making hidden normality assumptions? *Biom J*. 2018;60:1040-1058.

14. Stijnen T, Hamza TH, Özdemir P. Random effects meta-analysis of event outcome in the framework of the generalized linear mixed model with applications in sparse data. *Stat Med*. 2010;29:3046-3067.
15. Bagos PG, Nikolopoulos GK. Mixed-effects Poisson regression models for meta-analysis of follow-up studies with constant or varying durations. *Int J Biostat*. 2009;5:article 21.
16. Sterchi M, Wolf M. Weighted least squares and adaptive least squares: further empirical evidence. *Robustness in Econometrics*. Cham: Springer; 2017:135-167.
17. Viechtbauer W, López-López JA, Sánchez-Meca J, Marín-Martínez F. A comparison of procedures to test for moderators in mixed-effects meta-regression models. *Psychol Methods*. 2015;20:360-374.
18. Veroniki AA, Jackson D, Viechtbauer W, et al. Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Res Synth Methods*. 2016;7:55-79.
19. Novianti PW, Roes KC, Tweel I. Estimation of between-trial variance in sequential meta-analyses: a simulation study. *Contemp Clin Trials*. 2014;37:129-138.
20. Harville DA. Maximum likelihood approaches to variance component estimation and to related problems. *J Am Stat Assoc*. 1977;72:320-338.
21. Ng M, Wilcox RR. A comparison of two-stage procedures for testing least-squares coefficients under heteroscedasticity. *Br J Math Stat Psychol*. 2011;64:244-258.
22. Huber PJ. The behavior of maximum likelihood estimates under nonstandard conditions. Paper presented at: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability; University of California Press; 1967;1:221–233.
23. White H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*. 1980;48:817-838.
24. Pauly M, Brunner E, Konietschke F. Asymptotic permutation tests in general factorial designs. *J R Stat Soc Series B Stat Methodol*. 2015;77:461-473.
25. Zimmermann G, Pauly M, Bathke AC. Small-sample performance and underlying assumptions of a bootstrap-based inference method for a general analysis of covariance model with possibly heteroskedastic and nonnormal errors. *Stat Methods Med Res*. 2019;28:3808-3821.
26. MacKinnon JG, White H. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *J Econom*. 1985;29:305-325.
27. Cribari-Neto F. Asymptotic inference under heteroskedasticity of unknown form. *Comput Stat Data Anal*. 2004;45:215-233.
28. Cribari-Neto F, Souza TC, Vasconcellos KL. Inference under heteroskedasticity and leveraged data. *Commun Stat Theory Methods*. 2007;36:1877-1888.
29. Foulds G, Shepard R, Johnson R. The pharmacokinetics of azithromycin in human serum and tissues. *J Antimicrob Chemother*. 1990;25:73-82.
30. Laopaiboon M, Panpanich R, Mya KS. Azithromycin for acute lower respiratory tract infections. *Cochrane Database Syst Rev*. 2015;3:CD001954.
31. Bland JM, Altman DG. The odds ratio. *BMJ*. 2000;320:1468.
32. Harbord RM, Higgins JP. Meta-regression in Stata. *Stata J*. 2008;8:493-519.
33. SAS/ETS(R) 9.3 User's Guide 2011.
34. Hedges LV. Distribution theory for Glass's estimator of effect size and related estimators. *J Educ Behav Stat*. 1981;6:107-128.
35. Borenstein M, Hedges LV, Higgins JP, Rothstein HR. *Introduction to Meta-Analysis*. Hoboken, NJ: John Wiley & Sons; 2011.
36. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *J Stat Softw*. 2010;36:1-48.
37. Kelley K. Confidence intervals for standardized effect sizes: theory, application, and implementation. *J Stat Softw*. 2007;20:1-24.
38. Guolo A. Higher-order likelihood inference in meta-analysis and meta-regression. *Stat Med*. 2012;31:313-327.
39. Skovgaard IM. An explicit large-deviation approximation to one-parameter tests. *Ther Ber*. 1996;2:145-165.
40. Xu J, Cui X. Robustified MANOVA with applications in detecting differentially expressed genes from oligonucleotide arrays. *Bioinformatics*. 2008;24:1056-1062.
41. Vallejo G, Ato M. Robust tests for multivariate factorial designs under heteroscedasticity. *Behav Res Methods*. 2012;44:471-489.
42. Konietschke F, Bathke AC, Harrar SW, Pauly M. Parametric and nonparametric bootstrap methods for general MANOVA. *J Multivar Anal*. 2015;140:291-301.
43. Bathke AC, Friedrich S, Pauly M, et al. Testing mean differences among groups: multivariate and repeated measures analysis with minimal assumptions. *Multivariate Behav Res*. 2018;53:348-359.
44. van Buuren S, Groothuis-Oudshoorn K. MICE: multivariate imputation by chained equations in R. *J Stat Softw*. 2010;1-68.
45. Stekhoven DJ, Bühlmann P. MissForest – non-parametric missing value imputation for mixed-type data. *Bioinformatics*. 2011;28:112-118.
46. Flachaire E. Bootstrapping heteroskedastic regression models: wild bootstrap vs. pairs bootstrap. *Comput Stat Data Anal*. 2005;49:361-376.
47. Davidson R, Flachaire E. The wild bootstrap, tamed at last. *J Econom*. 2008;146:162-169.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Welz T, Pauly M. A simulation study to compare robust tests for linear mixed-effects meta-regression. *Res Syn Meth*. 2020;11:331–342. https://doi.org/10.1002/jrsm.1388