technische universität
dortmund

fakultät
statistik

Robust Covariance Estimation in Mixed-Effects Meta-Regression Models

Dissertation zur Erlangung des Doktorgrades Dr. rer. nat. der Fakultät Statistik der Technischen Universität Dortmund

Vorgelegt von

Thilo Welz

geboren in Filderstadt

Dortmund, März 2022

Amtierende Dekanin:
Prof. Dr. Katja Ickstadt

Gutachter:
Prof. Dr. Markus Pauly (Technische Universität Dortmund)
Prof. Dr. Guido Knapp (Technische Universität Dortmund)

Tag der Prüfung:
15.08.2022

Abstract

In this PhD thesis we consider robust (sandwich) variance-covariance matrix estimators in the context of univariate and multivariate meta-analysis and meta-regression. The underlying model is the classical mixed-effects meta-regression model. Our goal is to enable valid statistical inference (testing and construction of confidence regions) for the model coefficients. Specifically, we employ heteroscedasticity consistent (HC) and cluster-robust (CR) sandwich estimators in the univariate and multivariate setting, respectively. Such robust estimators are generally applicable in semiparametric linear models. A key aim is to provide better small sample solutions for meta-analytic research and application. Tests based on the original formulations of these estimators are known to produce highly liberal results, especially when the number of studies included in the analysis is small. We therefore transfer results for improved sandwich estimation such as the HC_4 estimator by Cribari-Neto and Zarkos (2004) to the meta-analytic context. We prove the asymptotic equivalence of HC estimators and compare them with commonly suggested techniques such as the Knapp-Hartung (KH) method or standard plugin covariance matrix estimation in extensive simulation studies. The new versions of HC estimators considerably outperform their older counterparts, especially in small samples, achieving comparable results to the KH method.

In a slight excursion, we focus specifically on constructing confidence regions for (Pearson) correlation coefficients as the main effect of interest in a random-effects meta-analysis. We develop a beta-distribution model for generating data in our simulations in addition to the commonly used truncated normal distribution model. We utilize different variance estimation approaches such as HC estimators, the KH method and a wild bootstrap approach in combination with the Fisher-z transformation and an integral z-to-r back-transformation to construct confidence regions. In simulation studies, our novel proposals improve coverage over the Hedges-Olkin-Vevea-z (HOVz) approach and Hunter-Schmidt approaches, enabling reliable inference for a greater range of true correlations.

Finally, we extend our results for the HC estimators to construct CR sandwich estimators for multivariate meta-regression. The aim is to achieve valid inference for the model coefficients, based on Wald-type statistics (WTS), even in small samples. Our simulations

show that previously suggested CR estimators such as the bias reduced linearization approach, can have unsatisfactory small sample performance for bivariate meta-regression. Furthermore, they show that the adjusted Hotelling's T^2-test suggested by Tipton and Pustejovsky (2015) can yield negative estimates for the degrees of freedom when the number of studies K is small ($K \leq 5$). We suggest an adjustment to the classical F-test, truncating the denominator degrees of freedom at two, which corresponds to a well defined expected value for an F-distribution. Our CR extensions of HC_3 and HC_4, using only the diagonal elements of the hat matrix to adjust residuals, improve coverage considerably in small samples over the standard CR_1 and bias reduced linearization approaches. We focus specifically on the bivariate case in our simulations but the discussed approaches can also be applied more generally.

We analyze both small and large sample behavior of all considered tests / confidence regions in extensive simulation studies. Furthermore, we apply the discussed approaches in real life datasets from psychometric and medical research.

Zusammenfassung

In dieser Dissertation betrachten wir robuste (Sandwich-)Varianz-Kovarianz-Matrix Schätzer sowohl im Kontext der univariaten als auch der multivariaten Meta-Analyse und Meta-Regression. Das zugrunde liegende Modell ist das klassische mixed-effects Meta-Regressionsmodell. Ziel ist es, eine valide statistische Inferenz (Testen und Konstruktion von Vertrauensbereichen) für die Modellkoeffizienten zu ermöglichen. Konkret verwenden wir Heteroskedastizitäts-konsistente (HC) und cluster-robuste (CR) Sandwich-Schätzer jeweils im univariaten bzw. multivariaten Fall. Solche robusten Schätzer sind allgemein in semiparametrischen linearen Modellen anwendbar. Ein wichtiges Ziel ist es, bessere Lösungen für Meta-Analysen mit kleinen Stichproben zu entwickeln. Tests, die auf den ursprünglichen Formulierungen dieser Schätzer beruhen, führen bekanntermaßen zu sehr liberalen Ergebnissen, insbesondere wenn die Anzahl der in die Analyse einbezogenen Studien gering ist. Wir übertragen daher Ergebnisse für verbesserte Sandwich-Schätzungen wie den HC_4-Schätzer von Cribari-Neto and Zarkos (2004) auf den meta-analytischen Kontext. Wir beweisen die asymptotische Äquivalenz der HC-Schätzer und vergleichen sie mit alternativen vorgeschlagenen Techniken wie der Knapp-Hartung (KH)-Methode oder der standard Plugin-Kovarianzmatrix-Schätzung in umfangreichen Simulationsstudien. Die neueren Versionen der HC-Schätzer übertreffen ihre älteren Gegenstücke beträchtlich, insbesondere bei kleinen Stichproben, und erzielen annähernd Ergebnisse wie die KH-Methode.

In einem kleinen Exkurs konzentrieren wir uns speziell auf die Konstruktion von Konfidenzbereichen für (Pearson) Korrelationskoeffizienten als interessierenden Haupteffekt in einer Meta-Analyse mit zufälligen Effekten. Wir entwickeln ein Beta-Verteilungsmodell für die Generierung von Daten in unseren Simulationen, zusätzlich zu dem üblicherweise verwendeten Modell der trunkierten Normalverteilung. Wir verwenden HC-Schätzer, die KH-Methode und einen Wild Bootstrap Ansatz in Kombination mit der Fisher-z-Transformation und einer integralen z-zu-r-Transformation, um Vertrauensbereiche zu konstruieren. Unsere neuartigen Vorschläge verbessern in Simulationen die Abdeckung gegenüber dem Hedges-Olkin-Vevea-z (HOVz)-Ansatz und den Hunter-Schmidt-Ansätzen und ermöglichen zuverlässige Schlussfolgerungen für einen größeren Bereich von wahren Korrelationen.

Schließlich erweitern wir unsere Ergebnisse für die HC-Schätzer, um CR-Sandwich-Schätzer für multivariate Meta-Regression zu konstruieren. Ziel ist es, auch bei kleinen Stichproben eine valide Inferenz für die Modellkoeffizienten auf der Grundlage der Wald-Typ-Statistik (WTS) zu erreichen. Unsere Simulationen zeigen, dass zuvor vorgeschlagene CR-Schätzer, wie der Ansatz der bias-reduzierten Linearisierung, eine unbefriedigende Leistung bei kleinen Stichproben für bivariate Meta-Regression aufweisen können. Darüber hinaus zeigen sie, dass der von Tipton and Pustejovsky (2015) vorgeschlagene angepasste Hotelling's T^2-Test negative Schätzungen für die Freiheitsgrade liefern kann, wenn die Anzahl der Studien klein ist ($K \leq 5$). Wir schlagen eine Anpassung des klassischen F-Tests vor, indem wir die Freiheitsgrade des Nenners bei zwei trunkieren, was einem wohl definierten Erwartungswert einer F-Verteilung entspricht. Unsere CR-Erweiterungen von HC_3 und HC_4, die nur die Diagonalelemente der Hutmatrix zur Anpassung der Residuen verwenden, verbessern die Abdeckung bei kleinen Stichproben beträchtlich gegenüber den Standardansätzen CR_1 und bias-reduzierter Linearisierung. Wir konzentrieren uns in unseren Simulationen speziell auf den bivariaten Fall, aber die diskutierten Ansätze können auch allgemeiner angewendet werden.

In umfangreichen Simulationsstudien analysieren wir sowohl das Verhalten bei kleinen als auch bei großen Stichproben für alle betrachteten Tests/Konfidenzbereiche. Darüber hinaus wenden wir die diskutierten Ansätze in realen Datensätzen aus der psychometrischen und medizinischen Forschung an.

Acknowledgments

First of all, I would like to thank my supervisor Markus Pauly for his support throughout my dissertation, for always having an open ear and for the opportunity to work on various interesting projects along the way. I gratefully acknowledge the funding by the DFG and I also thank Guido Knapp for being the second supervisor for my dissertation.

Furthermore, I would like to thank my co-authors Philipp Doebler and Wolfgang Viechtbauer for their helpful comments and discussions, which substantially contributed to this thesis. A special thank you also goes to Eric Knop for his great work in assisting our research and the excellent collaboration.

What really made my time as a PhD candidate special were my colleagues both at the Institute of Statistics in Ulm and in Dortmund. It was a great relief not having to go through such a huge transition – switching Universities and moving from Ulm to Dortmund – alone. Thank you for all the helpful and entertaining discussions, fun times at conferences, for the coffee breaks, Institutsstammtische and everything else!

Finally, I thank my family and Sarah for their continuous love and support, which made all of this possible. Thank you for helping me to get through the tough times and for helping me celebrate the good ones! Last but not least, I would like to thank Lizzy and Nala for being who they are and for continuing to be a constant source of joy in my life.

"Everything is going to be fine in the end. If it's not fine it's not the end."

– Oscar Wilde

Contents

List of Publications 1

I Introduction 5

1 Motivation 7

2 Statistical Methods 11
2.1 The Model 11
2.2 Hypotheses for Model Coefficients 13
2.3 Sandwich Estimators 14
2.4 Meta-Analysis of Correlation Coefficients 18

3 Summary of the Articles 21
3.1 Comparing Robust Tests for Linear Mixed-Effects Meta-Regression . . 21
3.2 Fisher Transformation based Confidence Intervals of Correlations . . . 23
3.3 Cluster-Robust Estimators for Bivariate Meta-Regression 26

4 Discussion 29

Bibliography 31

II Publications 37

List of Publications

This cumulative thesis is based on the following three manuscripts:

Article 1: Welz, T., and Pauly, M. (2020). A simulation study to compare robust tests for linear mixed-effects meta-regression. *Research Synthesis Methods*, **11**(3), 331–342. https://doi.org/10.1002/jrsm.1388.

Contribution of the author:
The author of this thesis implemented the extensive simulation studies and conducted the analysis of the data example, formulation of the manuscript and the mathematical proofs under Prof. Pauly's guidance.

Article 2: Welz, T., Doebler, P. and Pauly, M. (2021). Fisher transformation based confidence intervals of correlations in fixed- and random-effects meta-analysis. *British Journal of Mathematical and Statistical Psychology*, **75**(1), 1–22. https://doi.org/10.1111/bmsp.12242.

Contribution of the author:
The author of this thesis had a leading role in the preparation and structuring of the manuscript. He mainly implemented the simulation studies and conducted the mathematical proofs as well as the analysis of the data example.

Article 3: Welz, T., Viechtbauer, W. and Pauly, M. (2022). Cluster-Robust Estimators for Bivariate Mixed-Effects Meta-Regression. *arXiv preprint arXiv:2203.02234*. https://doi.org/10.48550/arXiv.2203.02234.

Contribution of the author:
The author of this thesis prepared and structured the manuscript mainly on his own. He conducted the methodological development, simulation study and empirical data analysis with helpful comments by the coauthors.

Further publications:

(1) Knop., E., Pauly, M., Friede, T. and Welz, T. (2022). Robust Confidence Intervals for Meta-Regression with Correlated Moderators. *arXiv preprint arXiv:2201.05491*.

(2) Gumpp, A. M., Boeck, C., Behnke, A., Bach, A.M., Ramo-Fernandez, L., Welz., T., Gündel, H., Kolassa, I. and Karabatsiakis, A. (2020). Childhood maltreatment is associated with changes in mitochondrial bioenergetics in maternal, but not in neonatal immune cells. *Proceedings of the National Academy of Sciences*, **117**(40), 24778–24784.

(3) Pauly, M. and Welz, T. (2018). Contribution to the discussion of "When should meta-analysis avoid making hidden normality assumptions?", **60**(6). *Biometrical Journal*, 1075–1076.

(4) Welz, T. (2018). A simulation study of random-effects meta-analysis methods in unbalanced designs. Technical Report, Ulm University. Available at *https://www.uni-ulm.de/fileadmin/website_uni_ulm/mawi2/dokumente/preprint-server/2018/tech_report_tw.pdf*.

Notation

Throughout the thesis, vectors and matrices are denoted by bold symbols, e.g., $\boldsymbol{M}$.

$\mathbb{N}$	Natural numbers
$\mathbb{R}$	Real numbers
$\mathbb{1}\{\cdot\}$	Indicator function
$\boldsymbol{M}'$	The transpose of a matrix or (column) vector $\boldsymbol{M}$
$\boldsymbol{M}^{-1}$	The inverse of a square matrix $\boldsymbol{M}$
$\boldsymbol{M}^{+}$	Moore-Penrose inverse
$\mathbf{I_t}$	$t \times t$ identity matrix, $t \in \mathbb{N}$
$\mathbf{1}_t$	t-dimensional column vector of 1's , $t \in \mathbb{N}$
$\mathrm{diag}(\ldots)$	Diagonal matrix with the values ... on the diagonal
$\oplus$	Direct sum
$\mathrm{tr}()$	The trace of a square matrix
$\mathrm{rank}()$	The rank of a matrix
$\mathbb{E}()$	The expectation of a random variable
$\mathrm{Var}()$	Variance of a random variable
$\xrightarrow{P}$	Convergence in probability
$\xrightarrow{d}$	Convergence in distribution
$\xrightarrow{a.s.}$	Almost sure convergence

Part I

Introduction

1 Motivation

Meta-analysis is a widely used statistical technique for synthesizing the results of multiple trials on the same or closely related research questions. Meta-analyses, like systematic reviews and umbrella reviews belong to secondary research in the hierarchy of evidence (Fusar-Poli and Radua, 2018). Systematic reviews identify published studies for a specific research question, discuss used methods, summarize results, highlight key findings and cite limitations (Garg et al., 2008). When study findings are pooled mathematically, we speak of meta-analysis. An umbrella review is a review of systematic reviews or meta-analyses. This is in contrast to individual studies such as randomized controlled trials or observational studies, which are part of primary research. A visualization of the hierarchy of evidence is given in Figure 1.1.

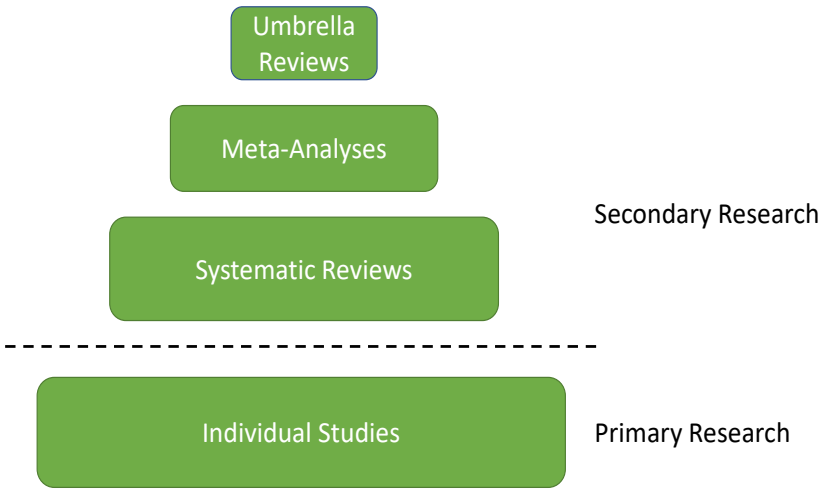

Figure 1.1: Hierarchy of Evidence Visualization.

As shown in Figure 1.2[1], which depicts the number of meta-analysis related publications on PubMed, research volume has increased strongly over the last few decades. This is indicative of the perceived high value of meta-analysis in the research community. Despite meta-analyses being performed routinely, scientists commonly face challenges in their data analyses including a small number of available studies, large variations in study sizes or substantial heterogeneity. However, standard meta-analytic techniques frequently assume normally distributed data, based on asymptotic arguments. Such assumptions are often neither reasonable nor are their violations adequately addressed, as recently pointed out by Jackson and White (2018), see also Pauly and Welz (2018). Not being able to rely on asymptotic arguments is very common, as can be seen in the empirical distribution of sample sizes of published meta-analyses. Davey et al. (2011) showed in a descriptive analysis of 22 453 meta-analyses from the Cochrane Database that the median number of trials per meta-analysis was only three. Evidently, the development of reliable methods for small sample meta-analyses is obligatory.

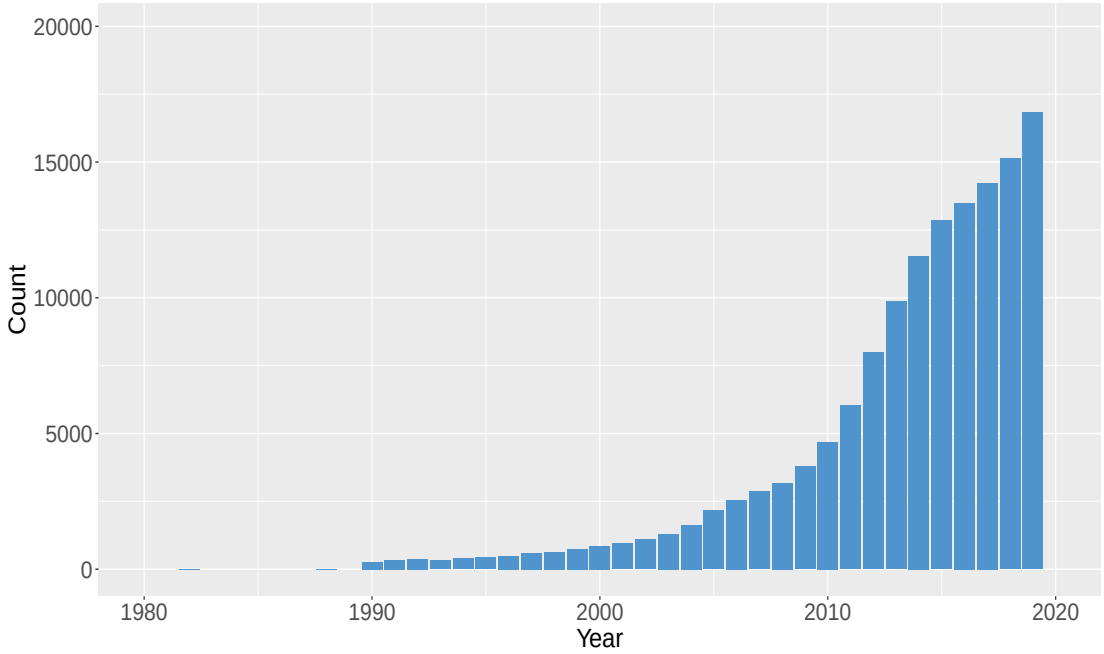

Figure 1.2: Meta-analysis related publications on PubMed between 1980 and 2020.

Such issues become even more pronounced in the multivariate setting, where normality is an even stronger and questionable assumption than in the univariate case. However, studies frequently report multiple effect sizes collected from overlapping patient groups.

[1]This figure was created based on a PubMed search for the term “meta-analysis” in December 2021.

This results in dependencies in the underlying true effects. Such data should generally be analyzed using a multivariate model, as a univariate analysis would lead to inefficient estimates. The multivariate normality assumption is especially difficult to verify for the model's random effects, when many of the studies do not report all parameters of interest, as pointed out by Jackson et al. (2011).

In both univariate or multivariate meta-analyses the trials may include study-level covariates (also called moderators). This is the meta-regression setting. Such moderators, for example a study's publication year, can account for systematic differences between trials and may therefore reduce the between-study heterogeneity in a given model. It is also possible for meta-regression to incorporate multiple moderators per study, categorial or continuous, as well as their interactions, see e.g. Knop et al. (2022). This makes meta-regression very versatile. However, it comes with certain caveats and is sometimes improperly used and interpreted in practice, as pointed out by Higgins and Thompson (2004).

Therefore, it is important to understand where the breakdown points lie in routinely applied methods. One way to check this is via Monte Carlo simulation studies. Furthermore, we need the development of robust methods in order to improve statistical inference when normality assumptions or asymptotic arguments do not hold. An obvious example is meta-analysis based on a small number of studies, which occurs frequently in practice (see above).

This thesis is organized as follows: Chapter 2 describes the underlying statistical models and methods such as the robust sandwich estimators considered in our work. Chapter 3 provides a summary of the three research articles underlying this dissertation, followed by Chapter 4, which contains a discussion of the results as well as an outlook for future research. Finally, Part II contains the three manuscripts summarized in Chapter 3.

In order to further motivate the methods analyzed in this thesis, we give some real data examples from different areas of application. The datasets can be found in the `R` packages `metafor` and `metadat`.

Effectiveness of Azithromycin for treating lower respiratory tract infections

To demonstrate the importance of the choice of heteroscedasticity consistent estimators for statistical inference in practice, we consider six studies with data on the effectiveness of Azithromycin versus Amoxycillin or Amoxycillin/clavulanic acid (Amoxyclav) in the the treatment of acute lower respiratory tract infections. Azithromycin is an antibiotic, which is useful for the treatment of different bacterial infections (Foulds et al., 1990).

The data were previously analyzed in a meta-analysis by Laopaiboon et al. (2015). We evaluate the question, whether a trial having included patients with a diagnosis of pneumonia has a statistically significant effect on the effectiveness of Azithromycin for patients with a diagnosis of acute bacterial bronchitis. Our analysis is based on a mixed-effects meta-regression model, where the study level effects are log odds ratios.

Correlation between conscientiousness and medication adherence

This dataset underlying a meta-analysis by Molloy et al. (2013) contains sixteen studies reporting a correlation between conscientiousness (from the five-factor model of personality) and medication adherence. The authors report that overall a higher level of conscientiousness is associated with better medication adherence. Using the various methods for constructing confidence intervals of correlations in meta-analysis discussed in our manuscript Welz et al. (2021), we construct confidence intervals for the main effect in a random-effects meta-analysis. The data can be divided into two subgroups based on trial design, namely cross-sectional or prospective. We consider these subgroups individually as well as a summary effect for all studies.

Overall and disease-free survival in neuroblastoma patients

These 81 trials from Riley et al. (2003, 2007) examine overall and/or disease-free survival in neuroblastoma patients with amplified (extra copies) versus normal MYC-N genes. Neuroblastoma is an embryonal cancer of the autonomic nervous system generally diagnosed in young children with a median age at diagnosis of 17 months (Maris, 2010). Amplified MYC-N levels are associated with poorer outcomes. The effect measures are log hazard ratios with positive values indicating an increased risk of death or relapse/death for patients with higher MYC-N levels compared to patients with lower levels. We evaluate model coefficients based on a bivariate meta-analysis, applying cluster robust estimators.

2 Statistical Methods

2.1 The Model

The usual (univariate) mixed-effects meta-regression model, also called random-effects model, is given by

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_m x_{im} + u_i + \varepsilon_i,\ i = 1, \ldots, K, \tag{2.1}$$

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. Consider for example the dataset from Laopaiboon et al. (2015) contained in the `R` package `metafor`. In this case the $y_i's$ are log odds ratios, $m = 1$ and x_{i1} is a binary covariate that is equal to one if study i included patients with a diagnosis of pneumonia and zero otherwise. Generally, we assume the number of studies is greater than the number of study-level moderators, i.e. $K > m$, and the model errors ε_i and random effects u_i are assumed to be independent. The within-study error ε_i is usually assumed to have distribution $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$. Furthermore, the within-study sampling variances σ_i^2 are typically assumed to be known, although actually estimated from the data. Additionally, u_i is a random effect that is also typically assumed to be normally distributed with $u_i \sim \mathcal{N}(0, \tau^2)$. Together this yields what is also known as a normal-normal hierarchical model (NNHM) (Friede et al., 2017). However, in our first manuscript we also consider the more general semiparametric setting with the moment assumptions $\mathbb{E}(u_i) = 0$ and $\mathrm{Var}(u_i) = \tau^2$ without other distributional restrictions. Setting $\tau^2 \coloneqq 0$ yields the fixed-effect model (also known as common-effect model) as a special case of the random-effects model. The moderators x_{ij} are study level covariates, which are supposed to reflect systematic differences between studies that are related to the effect size. These moderators are supposed to account for part of the heterogeneity in effect sizes, which is typically greater than what would be expected based on the sampling variability alone (Viechtbauer et al., 2015).

In the absence of moderators ($x_{ij} \equiv 0,\ i = 1, \ldots, K, j = 1, \ldots, m$), we are in the

classical meta-analysis setting, where a synthesized effect can be obtained via

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{K} \hat{w}_i y_i}{\sum_{i=1}^{K} \hat{w}_i},$$

with $\hat{w}_i$ representing study weights. These are usually defined as inverse variance weights with $\hat{w}_i = (\hat{\tau}^2 + \sigma_i^2)^{-1}$, where $\hat{\tau}^2$ is some estimate of the heterogeneity variance τ^2. Random-effects models give larger weights to smaller studies, compared with fixed-effect models. This can easily be verified if we consider two hypothetical studies 1 and 2 with $\sigma_2^2 > \sigma_1^2$, i.e. study 1 is larger than study 2. Let $\hat{w}_{1F}, \hat{w}_{2F}$ define the inverse variance fixed-effect weights and $\hat{w}_{1R}, \hat{w}_{2R}$ the inverse variance random-effects weights respectively. Then it holds

$$\hat{w}_{1F}/\hat{w}_{2F} = \sigma_2^2/\sigma_1^2 \geq (\hat{\tau}^2 + \sigma_2^2)/(\hat{\tau}^2 + \sigma_1^2) = \hat{w}_{1R}/\hat{w}_{2R}.$$

This has led some authors to argue against the use of random-effects models (Greenland, 1994), as smaller studies are more susceptible to bias. Nevertheless, random-effects models are more flexible and offer another distinction versus fixed-effect models. Fixed-effect models only allow for conditional inference (on the studies at hand), whereas random-effects models allow for unconditional inference (Hedges and Vevea, 1998). This means they allow for a generalization beyond the studies being meta-analyzed, assuming that they are a random sample from a greater, more general population of studies. The choice of model will depend on the goals of the analysis. In any case, in this thesis we will focus on the more flexible random-effects models, as study effect estimates are typically more variable than is assumed in fixed-effect models.

There are many suggestions for the estimation of the heterogeneity variance τ^2. Arguably the most well known is the method of moments estimator introduced by DerSimonian-Laird (DL) (DerSimonian and Laird, 1986). Although simple to calculate, the DL estimator is biased due to truncation at 0. Even though it is often used in applications, several authors have argued against the use of this estimator (Veroniki et al., 2016). Maximum likelihood (ML) estimation is also possible, but because ML estimation of variance components is usually negatively biased (Harville, 1977), the restricted maximum likelihood (REML) estimator is a better option (Raudenbush, 2009). Furthermore, there are the closely related (Viechtbauer et al., 2015) moment-based Paule-Mandel (PM) and the empirical Bayes estimator (Morris, 1983). Many comparative simulation studies have been undertaken to determine the best choice. A summary of this research can be found in Veroniki et al. (2016), who generally recommend the PM and for continuous data the REML estimator. In our work we follow their recommendation and use REML

estimation, as we mostly focus on continuous data.

In certain applications it is useful to consider multiple effects per study in a multivariate model. The multivariate mixed-effects model is given by

$$\boldsymbol{y_i} = \boldsymbol{X_i}\boldsymbol{\beta} + \boldsymbol{u_i} + \boldsymbol{\varepsilon_i},\ i = 1, \ldots, K \tag{2.2}$$

where the usual assumptions are $\boldsymbol{u_i} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{T})$ and $\boldsymbol{\varepsilon_i} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{V_i})$ with $\boldsymbol{u_i},\ \boldsymbol{\varepsilon_i}$ independent. Assuming p effects per study and that all studies report all effects of interest, we have $\boldsymbol{y_i}, \boldsymbol{u_i}, \boldsymbol{\varepsilon_i} \in \mathbb{R}^p$, and $\boldsymbol{T}, \boldsymbol{V_i} \in \mathbb{R}^{p \times p}$, $i = 1, \ldots, K$. Furthermore, for design matrix and coefficient vector $\boldsymbol{X_i} \in \mathbb{R}^{p \times q},\ \boldsymbol{\beta} \in \mathbb{R}^q$ for a $q \in \mathbb{N}$. In the bivariate setting, as considered in the third manuscript of this thesis, we have

$$\boldsymbol{T} = \begin{pmatrix} \tau_1^2 & \rho\tau_1\tau_2 \\ \rho\tau_1\tau_2 & \tau_2^2 \end{pmatrix} \text{ and } \boldsymbol{V_i} = \begin{pmatrix} \sigma_{i1}^2 & \sigma_{i12} \\ \sigma_{i12} & \sigma_{i2}^2 \end{pmatrix}.$$

Studies often do not report the sample covariances σ_{i12} between effects. This can be seen in the dataset on disease-free and overall survival in patients with neuroblastoma from Riley et al. (2007) and also in a dataset on the effects of deep-brain stimulation on the motor skills of patients with Parkinson's disease from Ishak et al. (2007). The latter is not analyzed in this thesis. Missing sample covariances frequently make the construction of the within-study variance-covariance matrices $\boldsymbol{V_i}$ difficult in practice.

2.2 Hypotheses for Model Coefficients

In matrix notation Models (2.1) and (2.2) can be written as $\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u} + \boldsymbol{\varepsilon}$. The weighted least squares estimate for $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X'}\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X'}\widehat{\boldsymbol{W}}\boldsymbol{y}, \tag{2.3}$$

where the weight matrix $\widehat{\boldsymbol{W}}$ is again defined via inverse variances. For Model (2.1) these are given by $\widehat{\boldsymbol{W}} = \text{diag}\left((\hat{\tau}^2 + \sigma_1^2)^{-1}, \ldots, (\hat{\tau}^2 + \sigma_K^2)^{-1}\right)$. We denote the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ by $\boldsymbol{\Sigma} = \text{Cov}(\hat{\boldsymbol{\beta}})$. Hedges et al. (2010) show that, given regularity conditions, $\hat{\boldsymbol{\beta}} \overset{a.s.}{\to} \boldsymbol{\beta}$ as $K \to \infty$ and $\hat{\boldsymbol{\beta}}$ asymptotically follows a normal distribution.

In this set-up, we are interested in constructing valid confidence regions for $\boldsymbol{\beta}$ or components thereof. Furthermore, we may want to test hypotheses $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta_0}$ vs. $H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta_0}$ or more generally $H_0 : \boldsymbol{H\beta} = \boldsymbol{c}$ vs. $H_1 : \boldsymbol{H\beta} \neq \boldsymbol{c}$ for some hypothesis matrix $\mathbf{H} \in \mathbb{R}^{a \times (m+1)}$, vector $\mathbf{c} \in \mathbb{R}^a$ and $a \in \mathbb{N}$. In the special case $\boldsymbol{c} = \mathbf{0}$ we can

define a unique projection matrix with $\boldsymbol{P} := \boldsymbol{H'}(\boldsymbol{HH'})^{+}\boldsymbol{H}$, where $\boldsymbol{H}^{+}$ denotes the Moore-Penrose inverse of $\boldsymbol{H}$. The matrix $\boldsymbol{P}$ is symmetric and idempotent. It then holds that $\boldsymbol{P\beta} = \boldsymbol{0}$ if and only if $\boldsymbol{H\beta} = \boldsymbol{0}$ (Brunner and Puri, 2001). For tests regarding single model coefficients β_j, we can consider test statistics of t-type $t_j = \hat{\beta}_j / \sqrt{\widehat{\boldsymbol{\Sigma}}_{jj}}$, see Welz and Pauly (2020). For more general hypotheses as listed above, we base inference on Wald-type statistics (WTS)

$$Q_H = (\boldsymbol{H}\hat{\boldsymbol{\beta}} - \boldsymbol{c})'(\boldsymbol{H}\widehat{\boldsymbol{\Sigma}}\boldsymbol{H'})^{-1}(\boldsymbol{H}\hat{\boldsymbol{\beta}} - \boldsymbol{c}), \tag{2.4}$$

where we generally assume that $\widehat{\boldsymbol{\Sigma}}$ is positive definite, cf. the third manuscript. Alternatively one might consider ANOVA-type statistics (ATS) or even modified ATS (MATS) as in Brunner et al. (2017) or Friedrich and Pauly (2018).

In any case we require a reliable, consistent estimator for the variance-covariance matrix $\boldsymbol{\Sigma} = \text{Cov}(\hat{\boldsymbol{\beta}})$, ideally with good small sample properties. This leads us to the next section on sandwich estimators.

2.3 Sandwich Estimators

In order to construct valid confidence regions and tests regarding the vector of model coefficients $\boldsymbol{\beta}$, we require a reliable, consistent estimator for $\boldsymbol{\Sigma}$. Given the true (unknown) weights $\mathbf{W}$, the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ is given by $(\boldsymbol{X'WX})^{-1}$. A standard approach is therefore to calculate the plug-in estimate $(\boldsymbol{X'}\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}$ of $\boldsymbol{\Sigma}$ based on the estimated weights $\widehat{\boldsymbol{W}}$. In the case of univariate random effects meta-analysis this estimator reduces to $\left(\sum_{i=1}^{K} \hat{w}_i\right)^{-1}$, as shown in the technical appendix of the supplement to Welz and Pauly (2020).

A naive test for individual components β_j of $\boldsymbol{\beta}$ for a fixed j can then be obtained via comparison of the test statistic $t_j = \beta_j / \sqrt{\widehat{\boldsymbol{\Sigma}}_{jj}}$ with a standard normal quantile, where $\widehat{\boldsymbol{\Sigma}}_{jj}$ is the jth diagonal element of $\widehat{\boldsymbol{\Sigma}}$. However, this approach ignores the imprecision in estimating τ. So when the estimate of the heterogeneity τ^2 is poor and therefore the weights incorrect, these tests can have undesirable properties, such as an inflated Type I error rate, especially in small samples (Knapp and Hartung, 2003). Knapp and Hartung, as well as Sidik and Jonkman (2002), proposed to use a refined variance estimator for the main effect instead of the standard plug-in estimate. If $\hat{\theta}$ is the main effect estimate of a random effects meta-analysis and $\hat{\theta}_i$ is the effect estimate in study i, Knapp and Hartung (2003) suggest to use

$$\widehat{\text{Var}}_{KH}(\hat{\theta}) = \frac{1}{K-1}\sum_{i=1}^{K}\frac{\hat{w}_i}{w}(\hat{\theta}_i - \hat{\theta})^2 \tag{2.5}$$

with weights $\hat{w}_i = (\sigma_i^2 + \hat{\tau}^2)^{-1}$ and $w = \sum_{i=1}^{K}\hat{w}_i$. Hartung (1999) showed that, given normally distributed study effects $\hat{\theta}_i$, $(\hat{\theta} - \theta)/\sqrt{\widehat{\text{Var}}_{KH}(\hat{\theta})}$ follows a t-distribution with $K-1$ degrees of freedom, K being the number of studies. This approach has been suggested over a standard test, based on a plug-in estimate for $\boldsymbol{\Sigma}$ (see the opening remarks of Section 2.3) i.e. the comparison of the statistic $(\hat{\theta} - \theta)/\sqrt{\left(\sum_{i=1}^{K}\hat{w}_i\right)^{-1}}$ with the standard normal quantile $z_{1-\alpha/2}$, by multiple authors (Viechtbauer et al., 2015; IntHout et al., 2014). The KH approach was mostly successful in controlling the nominal significance level in our own work on univariate meta-regression (Welz and Pauly, 2020).

Our research indicates that it is generally preferable to use t_{K-1} quantiles instead of standard normal quantiles for constructing valid confidence intervals and tests. In the multivariate setting it may also be difficult to construct the sampling variance-covariance matrices $\boldsymbol{V_i}$ because no estimates of covariances between effects are available. Furthermore, there can be various kinds of model misspecification, such as heteroscedastic or auto-correlated errors. Using so called sandwich estimators, also known as Huber-White estimators, in combination with t_{K-1} quantiles, when constructing CIs for individual components of $\boldsymbol{\beta}$, is a possible remedy. They are defined as

$$\widehat{\boldsymbol{\Sigma}} = (\boldsymbol{X'}\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X'}\widehat{\boldsymbol{W}}\widehat{\boldsymbol{\Omega}}\widehat{\boldsymbol{W}}\boldsymbol{X}(\boldsymbol{X'}\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}, \tag{2.6}$$

with $\boldsymbol{\Omega} = \text{Cov}(\boldsymbol{y})$. The idea is to crudely estimate $\boldsymbol{\Omega}$ using the squared residuals. Nevertheless, certain estimators of this type are consistent as $K \to \infty$ under heteroscedasticity of unknown form (White, 1980; Cribari-Neto, 2004), hence they are also referred to as HC estimators. We introduce these estimators in the following. The original formulation sets the "meat" of the sandwich to $\widehat{\boldsymbol{\Omega}} = \text{diag}(\widehat{\boldsymbol{E}})^2$ with $\widehat{\boldsymbol{E}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ (White, 1980). However, this estimator, sometimes called HC_0, is known to yield inflated Type I errors due to a negative bias for variance components (Sidik and Jonkman, 2005a). Therefore, when testing multiple improvements have been put forth, such as multiplication by the constant $\frac{K}{K-m-1}$ (Hedges et al., 2010), called HC_1, or a direct transformation of the residuals by discounting according to the observations' leverages. The leverage of a data point is a measure of its distance from other observations and therefore indicates whether an observation could potentially be highly influential with regard to parameter estimations. The leverage of the ith observation corresponds to the ith diagonal element h_{ii} of the hat matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X'}\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X'}\widehat{\boldsymbol{W}}$. Various suggestions have

been developed for incorporating the observations' leverages (Cribari-Neto and Zarkos, 2004; Cribari-Neto et al., 2007). These estimators were first proposed for the special case of ordinary least squares (OLS) regression, which corresponds to a weight matrix $\boldsymbol{W} = \boldsymbol{I}$. A weighted least squares version was introduced to the meta-analytic context by Sidik and Jonkman (2005b). For meta-analysis and meta-regression, we are in this more general weighted least squares context.

Commonly suggested in the regression literature are HC_3 (Long and Ervin, 2000) and HC_4 (Cribari-Neto and Zarkos, 2004), due to reduced bias and better control of the Type I error compared with alternative HC estimators. HC_3 sets

$$\widehat{\Omega} = \text{diag}\left(\widehat{E}_1^2/(1-h_{11})^2, \ldots, \widehat{E}_K^2/(1-h_{KK})^2\right)$$

and closely approximates the leave-one-out Jackknife estimator for $\boldsymbol{\beta}$, as introduced by Efron (1982). The HC_3 estimator discounts the effect of the leverages more than HC_2, which defines $\widehat{\Omega} = \text{diag}\left(\widehat{E}_1^2/(1-h_{11}), \ldots, \widehat{E}_K^2/(1-h_{KK})\right)$. The latter is rarely recommended because HC_3 typically has better small sample behavior (Cribari-Neto and Zarkos, 2004). HC_4 sets $\widehat{\Omega} = \text{diag}\left(\widehat{E}_1^2/(1-h_{11})^{\delta_1}, \ldots, \widehat{E}_K^2/(1-h_{KK})^{\delta_K}\right)$, where $\delta_i = \min\{4, h_{ii}/\bar{h}\}$, $i = 1, \ldots, K$, and $\bar{h}$ is the sample mean of the diagonal values of $\boldsymbol{H}$. The discounting is truncated at 4, which corresponds to twice as much discounting as for the HC_3 estimator (Cribari-Neto and Zarkos, 2004). Finally, there is also HC_5, which is similar to HC_4 but additionally incorporates a tuning parameter and explicity takes into account the effect of the maximal leverage (Cribari-Neto et al., 2007).

If studies report multiple effects of interest that are based on overlapping patient groups, the study effects will be correlated. This is the setting of multivariate meta-analysis or meta-regression as in Model (2.2). However, information on the correlation structure between these effects is rarely reported and individual patient data (IPD) unlikely to be available. This results in very crude estimates of the within-study covariance structure $\boldsymbol{V_i}$, at best. The estimator $\hat{\boldsymbol{\beta}}$ should still be approximately unbiased, but the standard errors are likely to be incorrect (Hedges et al., 2010), resulting in invalid inference. A proposed solution, which does not require the full availability of the within-study covariance structure, is to use a cluster-robust (CR) estimation approach. These are sandwich estimators, similar to before, except that now we set

$$\widehat{\boldsymbol{\Sigma}} = (\boldsymbol{X'}\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}\left(\sum_{\boldsymbol{j=1}}^{\boldsymbol{K}} \boldsymbol{X_j'}\widehat{\boldsymbol{W}}_{\boldsymbol{j}}\widehat{\boldsymbol{\Omega}}_{\boldsymbol{j}}\widehat{\boldsymbol{W}}_{\boldsymbol{j}}\boldsymbol{X_j}\right)(\boldsymbol{X'}\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}, \tag{2.7}$$

with $\widehat{\boldsymbol{\Omega}}_{\boldsymbol{j}} = \mathbf{A}_j \widehat{\boldsymbol{E}}_{\boldsymbol{j}} \widehat{\boldsymbol{E}}_{\boldsymbol{j}}' \mathbf{A}_j'$, where $\mathbf{A}_j$ are some adjustment matrices and $\boldsymbol{X_j}, \widehat{\boldsymbol{E}}_{\boldsymbol{j}}, \widehat{\boldsymbol{W}}_{\boldsymbol{j}}$ refer to the rows and columns of $\boldsymbol{X}, \widehat{\boldsymbol{E}}, \widehat{\boldsymbol{W}}$ pertaining to study j. For the original CR_0 estimator these adjustment matrices are set equal to identity matrices $\boldsymbol{I}$ and CR_1 can be defined analogously to HC_1 above, although alternative correction factors are possible, as implemented in the `clubSandwich` `R` package (Pustejovsky, 2021). However, it has been shown that tests based on CR_1 estimators still possess inflated Type I errors for small to moderate numbers of studies (Hedges et al., 2010). Tipton and Pustejovsky (2015) proposed a cluster robust estimation approach we call CR_2, extending work by Bell and McCaffrey (2002). Also called bias reduced linearization approach, it is based on a working model for the variance-covariance matrix $\boldsymbol{\Sigma}$. It is designed to be exactly unbiased given that the working model is correct, i.e. $\mathbf{W} = \boldsymbol{\Sigma}^{-1}$. Simulations have been undertaken to assess the decline in performance, when the working model is in fact wrong, as is likely to be the case in practice (Tipton and Pustejovsky, 2015; Tipton, 2015). Results indicate that CR_2 is not very sensitive to mistakes in the working model. However, in our third manuscript we demonstrate that CR_2 can perform poorly in bivariate meta-regression, especially when synthesizing few studies. Finally, Bell and McCaffrey (2002) also introduced CR_3 as a cluster-robust extension to HC_3. In equation (2.7) they set $\mathbf{A}_j = \sqrt{\frac{K-1}{K}}(\boldsymbol{I}_j - \boldsymbol{H}_j)^{-1}$, corresponding to the leave-one-out Jackknife variance estimate of $\hat{\boldsymbol{\beta}}$.

In the third manuscript of this thesis we consider an adjustment to CR_3 for bivariate meta-regression that incorporates only the diagonal elements of $\boldsymbol{H}$. This is due to the natural interpretation of diagonal elements of $\boldsymbol{H}$ as leverage of data points, whereas the off-diagonal elements of $\boldsymbol{H}$ lack an obvious interpretation. We call our estimator $\widehat{\boldsymbol{\Sigma}}_{CR_3^*}$, which has the form (2.7) with

$$\widehat{\boldsymbol{\Omega}}_j = \widehat{\boldsymbol{E}}_{\boldsymbol{j}} \widehat{\boldsymbol{E}}_{\boldsymbol{j}}' - \Delta + \Delta \cdot (\boldsymbol{I}_{p_j} - \mathrm{diag}(\boldsymbol{H}_j))^{-2}, \tag{2.8}$$

where $\boldsymbol{H}_j$ refers to the submatrix of $\boldsymbol{H}$ with entries pertaining to study j, p_j is the number of observed effects in study j and $\Delta = \mathrm{diag}(e_{11}, \ldots, e_{p_j p_j})$ with e_{ii} the diagonal elements of $\widehat{\boldsymbol{E}}_{\boldsymbol{j}} \widehat{\boldsymbol{E}}_{\boldsymbol{j}}'$ for $i = 1, \ldots, p_j$. Additionally, we consider a cluster-robust extension to HC_4 for the bivariate setting, which we denote by $\widehat{\boldsymbol{\Sigma}}_{CR_4^*}$. This time we set $\widehat{\boldsymbol{\Omega}}_j$ equal to (2.8) except the diagonal Δ is multiplied with $(\boldsymbol{I}_{p_j} - \mathrm{diag}(\boldsymbol{H}_j))^{-\delta_j}$. Here $\delta_j = \min\{4, h_{jj}/\bar{h}\}$ with h_{jj} denoting the j-th diagonal element of $\boldsymbol{H}$ and $\bar{h}$ is the average of the diagonal values of the hat matrix $\boldsymbol{H}_j$.

We apply the discussed HC and CR estimators, as well as the standard and KH approaches, in univariate (Welz and Pauly, 2020) and multivariate (Welz et al., 2022) meta-regression models. We perform extensive simulation studies to assess perfor-

mance in both small and large samples as well as evaluations of illustrative datasets. Additionally we consider these methods in random-effects meta-analyses of correlation coefficients (Welz et al., 2021). Some mathematical background for the latter is provided in the next section.

2.4 Meta-Analysis of Correlation Coefficients

The sample correlation coefficient r, based on n observations $(x_i, y_i),\ i = 1, \ldots, n$, corresponding to a pair of random variables (X, Y) is given by

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}. \tag{2.9}$$

Also known as the Pearson correlation coefficient, or product moment correlation (Schulze, 2004), r is a measure of the linear dependence between two variables. Although we focus on Pearson correlations in the following, we note that due to susceptibility to outliers, more robust, rank-based alternatives such as the Spearman correlation coefficient (Myers and Sirois, 2004) and Kendall's tau (Kendall, 1938) have been proposed.

Special care must be taken when correlations are the study effects of interest in a meta-analysis, since it holds $r \in [-1, 1]$ and a normal approximation as in the NNHM is difficult to justify. The exact, involved probability density function of the distribution of r was derived in an article by Hotelling (1953). However, working with the exact distribution of r is unfeasible. Assuming bivariate normality of (X, Y), r is approximately $\mathcal{N}(\varrho, (1-\varrho^2)^2/n)$ distributed for large samples n, where ϱ is the true correlation between X and Y (Lehmann, 1999). However, the corresponding asymptotic confidence interval for ϱ has poor coverage when the underlying data are non-normal, as demonstrated in Welz et al. (2021).

A popular approach is to apply the variance stabilizing Fisher-z transformation (Fisher, 1921) to the correlation coefficients to be synthesized. This transformation is equal to the inverse hyperbolic tangent and is given by $\operatorname{atanh} : (-1, 1) \to \mathbb{R},\ r \mapsto \operatorname{atanh}(r) = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$. Its inverse is given by the hyperbolic tangent with $\tanh : \mathbb{R} \to (-1, 1),\ z \mapsto \tanh(z) = \frac{\exp(2z)-1}{\exp(2z)+1}$. The main advantages of Fisher-z transformed correlations is that they are approximately normally distributed and the transformation is variance stabilizing, since it holds for $z = \operatorname{atanh}(r)$ that $\operatorname{Var}(z) \approx \frac{1}{n-3}$ (Schulze, 2004). The usual methodology for random-effects models can then be used to construct a confidence interval for the main effect on the z-scale. The final step is to then back-transform this

confidence interval using the $\tanh$ function. This is the idea of the Hedges-Olkin-Vevea-z (HOVz) approach (Hafdahl and Williams, 2009; Hedges and Vevea, 1998). Given study effects $r_i,\ i = 1, \ldots, K$, and their respective Fisher-z transforms $z_i,\ i = 1, \ldots, K$, the resulting confidence interval is given by

$$\tanh\left(\bar{z} \pm u_{1-\alpha/2} / \Big(\sum_{i=1}^{K} \hat{w}_i\Big)^{1/2}\right), \tag{2.10}$$

with $\hat{w}_i = (1/(n_i - 3) + \hat{\tau}^2)^{-1}$, $u_{1-\alpha/2}$ referring to the $(1-\alpha/2)$-quantile of the standard normal distribution and $\bar{z} = (\sum_{i=1}^{K} \hat{w}_i z_i)/(\sum_{i=1}^{K} \hat{w}_i)$. This approach, although commonly used, has some drawbacks. We highlight these shortcomings and propose improvements in the second article of this dissertation (Welz et al., 2021). A more in-depth discussion of this approach can be found in Hafdahl and Williams (2009).

An alternative is to aggregate correlations with the Hunter-Schmidt (HS) approach, which utilizes sample size weighting:

$$r_{HS} = \frac{\sum_{i=1}^{K} n_i r_i}{\sum_{i=1}^{K} n_i}.$$

Multiple approaches have been proposed for estimating the sampling variance σ^2_{HS} of r_{HS}, summarized in Schulze (2004). We highlight a suggestion by Osburn and Callender (1992),

$$\hat{\sigma}^2_{HS} = \frac{1}{K}\left(\frac{\sum_{i=1}^{K} n_i (r_i - r_{HS})^2}{\sum_{i=1}^{K} n_i}\right),$$

which is said to perform reasonably well in both heterogeneous and homogeneous settings (Schulze, 2004). In the second manuscript we test the validity of the resulting CI for a main effect ($r_{HS} \pm u_{1-\alpha/2}\hat{\sigma}_{HS}$) as a competing approach to the Fisher-z-based CIs. Use of a standard normal quantile in the Hunter-Schmidt CI is based on recommendations by Hunter and Schmidt (2004).

3 Summary of the Articles

3.1 Article 1: 'A simulation study to compare robust tests for linear mixed-effects meta-regression' (RSM, 2020)

We consider the classical mixed-effects meta-regression model described in Section 2.1. In order to obtain valid tests regarding components of the vector of model coefficients

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

we require consistent and efficient estimators for $\text{Cov}(\widehat{\boldsymbol{\beta}})$. Specifically, we wish to test the null hypothesis of no moderator effect, given by $H_0 : \{\beta_j = 0\}$, for $j = 1, \ldots, m$. We turn to the for regression often fruitful approach of robust estimators. More specifically, we consider heteroscedasticity consistent (HC) covariance estimators, also called Huber-White or sandwich estimators. These sandwich estimators are all of the general form

$$\widehat{\boldsymbol{\Sigma}} = (\boldsymbol{X}'\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X}'\widehat{\boldsymbol{W}}\widehat{\boldsymbol{\Omega}}\widehat{\boldsymbol{W}}\boldsymbol{X}(\boldsymbol{X}'\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}.$$

We introduce and compare a wide range of HC estimators, namely HC_0-HC_5. We study their performance in an extensive simulation study, where we focus on Type I error and power of statistical tests regarding $\widehat{\boldsymbol{\beta}}$, based on these estimators. We compare these approaches with the Knapp-Hartung (KH) approach (Hartung and Knapp, 2001), which was also proposed by Sidik and Jonkman (2002). In our simulation study, we focus on standardized mean differences as effect sizes. We also consider log odds ratios, with similar results. We simulate a wide range of parameter choices, considering 5, 10, 20 and 40 studies, an average of 30, 50 and 100 study participants with unequal study sizes and balanced treatment and control groups as well as varying amounts of heterogeneity τ^2 between 0.1 and 0.9, i.e. little to substantial heterogeneity. The single moderator was drawn from a $\mathcal{N}(0, 1)$ distribution. We performed a total of 1 000 simulation runs, corresponding to a Monte Carlo standard error of approximately 0.689% (Morris et al., 2019).

The results regarding Type I error indicate that changes in the between-study heterogeneity τ^2, the number of subject per study and underlying distributions of the random effects had little effect on the behavior of the procedures under the null hypothesis. In contrast, the number of studies K and the chosen test procedure were the main deciding factors for changes in Type I error control. Tests based on HC_0-HC_2 were quite liberal, especially for a maximum of 10 studies. The other estimators controlled the nominal level α well, except for HC_3 with only five studies, which was quite conservative.

When considering power, we observed that power increased for decreasing amounts of heterogeneity τ^2, increasing number of studies K as well as increasing (average) study size n. On average the KH approach yielded slightly more power than HC_3-HC_5 for up to 10 studies, with approximately equal power for more studies.

We also consider the (empirical) bias $\mathbb{E}[\widehat{\beta}_1] - \beta_1$ and variance $\mathrm{Var}(\widehat{\beta}_1) = \Sigma_{11}$. The simulation results indicate that the estimator $\widehat{\beta}_1$ is approximately unbiased for $\beta_1 = 0$ and becomes increasingly negatively biased for larger effect sizes β_1. Moreover, the variance increases with newer versions of the HC estimator. The KH method has a smaller variance than the newer iterations HC_3-HC_5.

We motivate the methods with a data analysis from medical research that also illustrates the practical importance of the choice of covariance estimator. The data contains six studies, which investigate the effectiveness of Azithromycin versus Amoxycillin or Amoxyclav in the treatment of acute lower respiratory tract infections. The results show that p-values of tests regarding the model coefficients can vary considerably depending on the choice of HC estimator.

In the accompanying supplement we also provide a proof for the asymptotic equivalence (for $K \to \infty$) of all considered HC estimators, given mild regularity conditions on the moderators, as well as the complete simulation results. Additionally, we derive a general formula for HC-type estimators in the case of no moderators (i.e. meta-analysis).

To sum up, we propose the use of updated versions of robust HC-type estimators. We compare these in an extensive simulation study with the older HC_1 estimator considered by Viechtbauer et al. (2015) and the Knapp-Hartung method. In the supplement we prove the asymptotic equivalence of all HC estimators and derive their form analytically for the case of no moderators (meta-analysis instead of meta-regression). Finally, we exemplify the different methods and demonstrate the large influence the choice of estimator can have in practice with a meta-analysis of six studies from medical research, which considers treatments for acute lower respiratory tract infections.

3.2 Article 2: 'Fisher transformation based confidence intervals of correlations in fixed- and random-effects meta-analysis' (BJMSP, 2021)

Pearson correlation coefficients are a common statistical instrument for quantifying strengths of association. They frequently occur in psychometric research. When multiple studies are available for comparable underlying participant groups, meta-analytic techniques enable the pooling of evidence and improve precision and stability of estimates (Hedges and Olkin, 1985). A standard approach for constructing confidence intervals for Pearson correlations as effects of a meta-analysis is the HOVz approach (Hafdahl and Williams, 2009). The idea is to first transform correlations using the Fisher-z transformation, then construct confidence intervals (CIs) on the z-scale and finally to backtransform CIs with the inverse Fisher-z transformation. Results from simulations, however, indicate that in random-effects models the performance of the HOVz confidence interval can be unsatisfactory, regarding control of the nominal level (Hafdahl and Williams, 2009).

We propose multiple improvements to the HOVz approach. Our improvements are based on alternative variance estimates of the pooled effect estimate

$$\bar{z} = \frac{\sum_{i=1}^{K} \left(\frac{1}{n_i - 3} + \hat{\tau}^2 \right)^{-1} z_i}{\sum_{i=1}^{K} \left(\frac{1}{n_i - 3} + \hat{\tau}^2 \right)^{-1}},$$

where $z_i = \operatorname{atanh}(r_i)$, n_i refers to the size of study i and r_i is the estimated correlation coefficient (effect) in stduy i. Specifically, we propose using either the KH approach, a robust estimator of HC-type, as considered in Article 1 of this thesis, or a wild bootstrap (Wu, 1986) approach to estimate the variance of $\bar{z}$. We compare these approaches with the Hunter-Schmidt (HS) method, which proposes sample size weighting.

Additionally, we propose using the integral z-to-r transformation, as suggested by Hafdahl (2009), instead of the inverse Fisher transformation $\tanh$ as in HOVz for the construction of CIs. The motivation behind is that for $\xi \sim \mathcal{N}(\operatorname{atanh}(\varrho), \sigma^2)$ for some $\sigma^2 > 0$ and $\varrho \neq 0$ it holds that $\varrho = \tanh(\mathbb{E}(\xi)) \neq \mathbb{E}(\tanh(\xi))$. Therefore using the inverse Fisher transformation $\tanh$ introduces bias into the analysis. The integral z-to-r transformation, which aims to alleviate this bias, is equal to the expected value of $\tanh(z)$, so if $z \sim \mathcal{N}(\mu, \tau^2)$, it is defined as

$$\psi(\mu \mid \tau^2) = \int_{-\infty}^{\infty} \tanh(t) f(t \mid \mu, \tau^2) dt,$$

where f is the density of z. This transformation is used in practice by applying it to the lower and upper confidence limits on the z-scale, plugging in estimates $\hat{z}$ and $\hat{\tau}^2$.

We compare the mentioned approaches in an extensive Monte Carlo simulation study, focusing on coverage and interval lengths. Notably, we consider two main simulation designs. First, a truncated normal distribution model, where the true study level effects ϱ_i were sampled from normal distributions $\mathcal{N}(\varrho, \tau^2)$ truncated so the samples lie within the interval $[-0.999, 0.999]$, as in Hafdahl and Williams (2009). This truncation results in bias. Therefore we also consider a second model, where we generate (true) study level effects ϱ_i from transformed beta distributions. The idea is to set $Y_i = 2(X_i - 0.5)$ with $X_i \sim Beta(\alpha, \beta)$ with α and β chosen such that $\mathbb{E}(Y_i) = \varrho$ and Var$(Y_i) = \tau^2$.

We simulate a variety of parameter choices such as ϱ between 0 and 0.9, $\tau \in \{0, 0.16, 0.4\}$, $K \in \{5, 10, 20, 40\}$ and (average) trial sizes of 20 and 80 with 10 000 simulation runs respectively, corresponding to a Monte Carlo standard error of approximately 0.218% (Morris et al., 2019). The results and thus recommended methods depend on the assumed model and the amount of heterogeneity present in the data. In general, we believe that the beta distribution model is better suited for random-effects meta-analysis of correlations. For most settings we recommend using either the KH, HC_3 or HC_4 confidence intervals. The wild bootstrap (WBS) CIs have comparable coverage but are wider on average. HS and HOVz mostly have unsatisfactory coverage and therefore cannot be recommended. An exception is for the underlying beta distribution model with $K \geq 40$ studies and $|\varrho| > 0.7$, where we recommend using the HS approach.

Finally, we exemplify the methods in an illustrative meta-analysis of 16 trials from Molloy et al. (2013) on the correlation between medication adherence and conscientiousness (from the five-factor model of personality). The results indicate a small positive (statistically significant) correlation, 0.13 based on a fixed-effect and 0.15 for a random-effects model.

In summary, we proposed new methods for the construction of confidence intervals for Pearson correlation coefficients as the main effect in a random-effects meta-analysis. Specifically, these were based on robust estimators of HC type as well as a wild bootstrap approach. Additionally, we consider Knapp-Hartung based variance estimation in this context. We recommend using these estimators in combination with the integral z-to-r transformation. We compare our novel methods with the HOVz and HS approaches in an extensive simulation study with focus on empirical coverage and interval lengths. Based

on these extensive simulations we were able to derive recommendations for applied researchers. Finally, we exemplify all methods in a real life dataset from psychometric research, considering the correlation between medication adherence and the personality trait conscientiousness.

3.3 Article 3: 'Cluster-robust estimators for bivariate mixed-effects meta-regression' (arXiv, 2022)

Trials frequently report multiple effect sizes that are based on (at least in part) overlapping patient groups. This results in correlated observations and these dependencies should be taken into account. When formulating a multivariate statistical meta-analysis or meta-regression model in practice, it can be difficult to adequately construct the within-study sampling variance-covariance matrix $\boldsymbol{V}$. Even if $\boldsymbol{V}$ is specified incorrectly, estimates of the fixed effects are still approximately unbiased, but with incorrect standard errors. In the setting of the multivariate meta-regression model described in Section 2.1, we follow a fruitful approach to deal with such inefficient estimates by using cluster-robust (CR) covariance estimators.

The general sandwich form of CR estimators is

$$\widehat{\boldsymbol{\Sigma}}_{CR} = (\boldsymbol{X'}\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1} \left(\sum_{\boldsymbol{i=1}}^{\boldsymbol{K}} \boldsymbol{X_i'}\widehat{\boldsymbol{W}}_{\boldsymbol{i}}\widehat{\boldsymbol{\Omega}}_{\boldsymbol{i}}\widehat{\boldsymbol{W}}_{\boldsymbol{i}}\boldsymbol{X_i} \right) (\boldsymbol{X'}\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1},$$

where $\widehat{\boldsymbol{\Omega}}_{\boldsymbol{i}} = \boldsymbol{E_i}\boldsymbol{E_i'}$ with $\boldsymbol{E_i} = \boldsymbol{Y_i} - \boldsymbol{X_i}\hat{\boldsymbol{\beta}}$. Since this estimator is known to be downward biased (Hedges et al., 2010), it needs to be adjusted. Previous suggestions are multiplication with a constant such as $\frac{K}{K-q}$ or transforming the residuals themselves by setting $\widehat{\boldsymbol{\Omega}}_{\boldsymbol{i}} = \boldsymbol{A_i}\boldsymbol{E_i}\boldsymbol{E_i'}\boldsymbol{A_i'}$ for adjustment matrices $\boldsymbol{A}_i$ for $i = 1, \ldots, K$ (Hedges et al., 2010; Tipton, 2015). Tipton and Pustejovsky (2015) developed a bias reduced linearization approach, which is designed to be exactly unbiased under the correct specification of a working model. It is implemented in the `clubSandwich` `R` package (Pustejovsky, 2021). However, this estimator can perform poorly in certain settings, such as bivariate meta-regression with few studies, as we demonstrate in simulations.

In this paper, we present two new CR estimators, which are extensions of the estimators HC_3 and HC_4. We investigated the latter in the univariate meta-regression setting in Welz and Pauly (2020) and applied them to improve confidence regions for meta-analyzed correlation coefficients in Welz et al. (2021). The main idea is to transform the residual variances, i.e. the diagonal elements of $\widehat{\boldsymbol{\Omega}}_i$ using (only!) the diagonal elements of the hat matrix. This is in contrast to the CR_3 estimator as suggested by Bell and McCaffrey (2002), which utilizes the entire hat matrix.

In order to obtain valid tests and confidence regions for $\boldsymbol{\beta} \in \mathbb{R}^q$, Tipton and Pustejovsky (2015) proposed a small sample adjustment. Their idea for testing $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta_0}$ vs

$H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta_0}$ at level α is to consider the WTS $Q = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0})'\widehat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0})$ and based on this the Hotelling's T^2 test $\mathbb{1}\{Q > \frac{\eta q}{\eta - q + 1} F_{q,\eta-q+1,1-\alpha}\}$. In the latter term η denotes a degree of freedom that needs to be estimated. Tipton and Pustejovsky (2015) suggest to use an estimator ($\hat{\eta}_Z$) originally proposed by Zhang (2012) for heteroscedastic one-way MANOVA. However, we found that for a small number of studies ($K \leq 5$) it often happens that $\hat{\eta}_Z - q + 1 < 0$. As the degrees of freedom in an F distribution cannot be negative, we follow an alternative approach. Our proposal is to use an F-test with a degree of freedom adjustment, given by

$$\mathbb{1}\{Q > qF_{q,\max\{2,K-q\},1-\alpha}\}. \tag{3.1}$$

We obtain the corresponding confidence region Λ via test inversion with

$$\Lambda := \left\{\boldsymbol{\beta} \in \mathbb{R}^q : (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\widehat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \leq qF_{q,\max\{2,K-q\},1-\alpha}\right\}.$$

Following Johnson et al. (2014), we obtain a confidence ellipsoid centered around $\hat{\boldsymbol{\beta}}$, whose axes are given by $\hat{\boldsymbol{\beta}} \pm \sqrt{\hat{\lambda}_j qF_{q,\max\{2,K-q\},1-\alpha}}\hat{e}_j,\ j = 1, \ldots, q$ where $\hat{\lambda}_j$ and $\hat{e}_j$ are the eigenvalues and eigenvectors of $\widehat{\boldsymbol{\Sigma}}$ respectively. The volume of the confidence ellipsoid Λ is given by (Wilson, 2010)

$$V_\Lambda = \frac{2\pi^{q/2}}{q\Gamma(q/2)} \prod_{i=1}^{q} \sqrt{\hat{\lambda}_i qF_{q,\max\{2,K-q\},1-\alpha}}. \tag{3.2}$$

We compare the approaches in a simulation study, examining the coverage of confidence regions Λ and power of the test (3.1) for bivariate meta-regression with a single moderator, drawn from a $\mathcal{N}(0,1)$ distribution. We consider $K \in \{5, 10, 20, 40\}$ studies, average study sizes of 40 and 100, coefficient vectors $\boldsymbol{\beta} \in \{(0,0,0,0)', (0.2,0.2,0.1,0.1)', (0.4,0.4,0.2,0.3)'\}$, correlations (between effects) $\varrho \in \{0, 0.3, 0.7\}$ and between 0 and 40% of studies that only report one of the two effects. We performed a total of 5 000 Monte Carlo iterations. This corresponds to a Monte Carlo standard error for the empirical coverage of approximately 0.308% (Morris et al., 2019). Additionally, we apply the methods in a real life dataset containing studies on overall and/or disease-free surival in neuroblastoma patients with amplified versus normal MYC-N genes. The data are contained in the `R` package `metafor` and stem from Riley et al. (2003, 2007).

The results indicate that for bivariate meta-regression and $K \leq 10$ studies, CR_3^* and CR_4^* provide the closest to nominal coverage. CR_3^* performed just very slightly better than

CR_4^*. For $K \geq 20$ studies the standard estimator $(\boldsymbol{X}'\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}$ yielded most accurate coverage. Both CR_1^* and CR_2 were highly liberal, even for larger number of studies. In general the smaller the number of studies, the more pronounced the difference between the estimators become.

To sum up, we introduced two novel CR estimators for multivariate meta-regression, utilizing only the diagonal elements of the hat matrix. We examine the performance of tests and confidence regions based on these estimators for bivariate meta-regression, comparing them with state-of-the-art methods such as the bias reduced linearization approach. In extensive simulations, we show improvements in coverage of the respective confidence regions, especially for few studies. Furthermore, we analyzed a real life dataset on overall and disease-free survival in neuroblastoma patients. In this paper we focus on the bivariate setting. However, the discussed methods are applicable more generally. Further research is necessary to evaluate the viability of our proposed estimators in general multivariate settings.

4 Discussion

In this thesis, we considered various sandwich estimators in order to improve small sample performance of corresponding tests and confidence regions and thus to allow for valid inference in meta-analytic models. Starting with univariate meta-analysis and meta-regression, we also extended the results to the multivariate setting.

We were able to improve the small sample behavior of tests and confidence regions regarding the model coefficients over previously suggested HC estimators in univariate mixed effects meta-regression with a single moderator (Welz and Pauly, 2020). Here we achieved comparable results with the Knapp-Hartung method. We proved the asymptotic equivalence of all HC estimators, given regularity conditions, and derived their analytic form for the special case of no moderators (meta-analysis). Simulations suggest that the results can seemingly be transferred to semi-parametric models, as performance remained stable in settings with non-normal random effects.

For random effects meta-analysis of (Pearson) correlation coefficients (Welz et al., 2021) we applied variance estimation methods discussed in Welz and Pauly (2020) in combination with the Fisher transformation and an integral z-to-r transformation suggested by Hafdahl (2009). We used these methods to construct novel confidence intervals for the main effect ρ, considerably improving coverage over state-of-the-art approaches such as the one by Hedges, Olkin and Vevea (Hedges and Olkin, 1985; Hedges and Vevea, 1998). As part of our simulation study, we develop an arguably more appropriate simulation model than the truncated normal distribution model used in previous work (Field, 2005; Hafdahl and Williams, 2009), using transformed beta distributions to generate study effect sizes. Our proposed approaches reduce bias, allowing for valid inference for larger correlations $|\varrho|$, as compared with state-of-the-art methods. However, there is still room for improvement to even further reduce bias, which remains an issue for large $|\varrho|$, especially when dealing with considerable heterogeneity.

Finally, we extended our research results on heteroscedasticity consistent estimators for univariate meta-regression to cluster robust estimators for multivariate meta-regression. Inference regarding model coefficients is made using Wald-type statistics. We made two novel proposals for CR estimators. Furthermore, our simulations showed that the

approximate Hotelling's T^2 test proposed in Tipton and Pustejovsky (2015) frequently breaks down for $K \leq 5$ studies. Also, coverage based on the state-of-the-art bias reduced linearization approach can be unsatisfactory for few studies. We therefore suggest a small sample adjustment for the classical F-test. Our extensive simulations in the bivariate setting show improved coverage of confidence regions based on our novel proposals. The approaches can also be applied in higher dimensional settings, but further work is necessary to assess the viability of our proposals there. A possible avenue for further improvements in small sample behavior lies in using ANOVA-type (Brunner and Puri, 2001) instead of Wald-type statistics, e.g. in combination with resampling approaches as is done in MANOVA settings (Friedrich and Pauly, 2018). Another potential approach, when studies report incomplete information, is to apply imputation methods such as the MissForest algorithm (Stekhoven and Bühlmann, 2012) or multivariate imputation by chained equations (Van Buuren and Groothuis-Oudshoorn, 2011).

In summary, in this dissertation we investigated the use of sandwich estimators in both univariate und multivariate meta-analytic models. The main goal was to improve small sample behavior of tests and confidence regions regarding the model coefficients. This is crucially important because meta-analyses of few studies occur frequently in practice (Davey et al., 2011) and yet commonly used methods often rely on asymptotic arguments regarding the number of studies. Based on our findings, we were able to derive recommendations for applied researchers that improve the validity of statistical inference regarding model coefficients, especially when synthesizing a small number of studies.

Future work considering alternative statistics such as ATS or MATS statistics, as mentioned in Section 2.2, or the combination of individual participant data with novel resampling techniques may further improve meta-analytic methodology. A starting point for the latter may be extensions to the paper by Van Den Noortgate and Onghena (2005) on parametric and non-parametric bootstrap methods for meta-analysis. A further avenue for future research is to extend the results from Welz et al. (2021) to more robust correlation measures such as Spearman rank correlations, Kendall's tau or other estimators with bounded support like e.g. Wilcoxon-Mann-Whitney effects.

Bibliography

Bell, R. M. and McCaffrey, D. F. (2002). Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology*, 28(2):169–182.

Brunner, E., Konietschke, F., Pauly, M., and Puri, M. L. (2017). Rank-based procedures in factorial designs: Hypotheses about non-parametric treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(5):1463–1485.

Brunner, E. and Puri, M. L. (2001). Nonparametric methods in factorial designs. *Statistical Papers*, 42(1):1–52.

Cribari-Neto, F. (2004). Asymptotic inference under heteroskedasticity of unknown form. *Computational Statistics & Data Analysis*, 45(2):215–233.

Cribari-Neto, F., Souza, T. C., and Vasconcellos, K. L. (2007). Inference under heteroskedasticity and leveraged data. *Communication in Statistics - Theory and Methods*, 36(10):1877–1888.

Cribari-Neto, F. and Zarkos, S. G. (2004). Leverage-adjusted heteroskedastic bootstrap methods. *Journal of Statistical Computation and Simulation*, 74(3):215–232.

Davey, J., Turner, R. M., Clarke, M. J., and Higgins, J. P. (2011). Characteristics of meta-analyses and their component studies in the cochrane database of systematic reviews: a cross-sectional, descriptive analysis. *BMC Medical Research Methodology*, 11(1):1–11.

DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3):177–188.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*. Society for Industrial and Applied Mathematics, Philadelphia.

Field, A. P. (2005). Is the meta-analysis of correlation coefficients accurate when population correlations vary? *Psychological Methods*, 10(4):444–467.

Fisher, R. A. (1921). On the 'probable error' of a coefficient of correlation deduced from a small sample. *Metron*, 1:1–32.

Foulds, G., Shepard, R., and Johnson, R. (1990). The pharmacokinetics of Azithromycin in human serum and tissues. *Journal of Antimicrobial Chemotherapy*, 25(suppl_A):73–82.

Friede, T., Röver, C., Wandel, S., and Neuenschwander, B. (2017). Meta-analysis of few small studies in orphan diseases. *Research Synthesis Methods*, 8(1):79–91.

Friedrich, S. and Pauly, M. (2018). MATS: Inference for potentially singular and heteroscedastic MANOVA. *Journal of Multivariate Analysis*, 165:166–179.

Fusar-Poli, P. and Radua, J. (2018). Ten simple rules for conducting umbrella reviews. *Evidence-Based Mental Health*, 21(3):95–100.

Garg, A. X., Hackam, D., and Tonelli, M. (2008). Systematic review and meta-analysis: when one study is just not enough. *Clinical Journal of the American Society of Nephrology*, 3(1):253–260.

Greenland, S. (1994). Invited commentary: a critical look at some popular meta-analytic methods. *American Journal of Epidemiology*, 140(3):290–296.

Hafdahl, A. R. (2009). Improved Fisher z estimators for univariate random-effects meta-analysis of correlations. *British Journal of Mathematical and Statistical Psychology*, 62(2):233–261.

Hafdahl, A. R. and Williams, M. A. (2009). Meta-analysis of correlations revisited: Attempted replication and extension of Field's (2001) simulation studies. *Psychological Methods*, 14(1):24–42.

Hartung, J. (1999). An alternative method for meta-analysis. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 41(8):901–916.

Hartung, J. and Knapp, G. (2001). A refined method for the meta-analysis of controlled clinical trials with binary outcome. *Statistics in Medicine*, 20(24):3875–3889.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72(358):320–338.

Hedges, L. and Vevea, J. (1998). Fixed-and random-effects models in meta-analysis. *Psychological Methods*, 3(4):486–504.

Hedges, L. V. and Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press. San Diego, CA, USA.

Hedges, L. V., Tipton, E., and Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, 1(1):39–65.

Higgins, J. P. and Thompson, S. G. (2004). Controlling the risk of spurious findings from meta-regression. *Statistics in Medicine*, 23(11):1663–1682.

Hotelling, H. (1953). New light on the correlation coefficient and its transforms. *Journal of the Royal Statistical Society. Series B (Methodological)*, 15(2):193–232.

Hunter, J. E. and Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Sage.

IntHout, J., Ioannidis, J. P., and Borm, G. F. (2014). The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Medical Research Methodology*, 14(1):1–12.

Ishak, K. J., Platt, R. W., Joseph, L., Hanley, J. A., and Caro, J. J. (2007). Meta-analysis of longitudinal studies. *Clinical Trials*, 4(5):525–539.

Jackson, D., Riley, R., and White, I. R. (2011). Multivariate meta-analysis: potential and promise. *Statistics in Medicine*, 30(20):2481–2498.

Jackson, D. and White, I. R. (2018). When should meta-analysis avoid making hidden normality assumptions? *Biometrical Journal*, 60(6):1040–1058.

Johnson, R. A., Wichern, D. W., et al. (2014). *Applied multivariate statistical analysis*, volume 6. Pearson London, UK:.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Knapp, G. and Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine*, 22(17):2693–2710.

Knop, E. S., Pauly, M., Friede, T., and Welz, T. (2022). Robust confidence intervals for meta-regression with correlated moderators. *arXiv preprint arXiv:2201.05491*.

Laopaiboon, M., Panpanich, R., and Mya, K. S. (2015). Azithromycin for acute lower respiratory tract infections. *Cochrane Database of Systematic Reviews*, (3).

Lehmann, E. L. (1999). *Elements of large-sample theory*. Springer.

Long, J. S. and Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, 54(3):217–224.

Maris, J. M. (2010). Recent advances in neuroblastoma. *New England Journal of Medicine*, 362(23):2202–2211.

Molloy, G., O'carroll, R., and Ferguson, E. (2013). Conscientiousness and medication adherence: a meta-analysis. *Annals of Behavioral Medicine*, 47(1):92–101.

Morris, C. N. (1983). Parametric empirical Bayes inference: theory and applications. *Journal of the American Statistical Association*, 78(381):47–55.

Morris, T. P., White, I. R., and Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11):2074–2102.

Myers, L. and Sirois, M. J. (2004). Spearman correlation coefficients, differences between. *Encyclopedia of Statistical Sciences*, 12.

Osburn, H. and Callender, J. (1992). A note on the sampling variance of the mean uncorrected correlation in meta-analysis and validity generalization. *Journal of Applied Psychology*, 77(2):115.

Pauly, M. and Welz, T. (2018). Contribution to the discussion of "when should meta-analysis avoid making hidden normality assumptions?". *Biometrical Journal*, 60(6):1075–1076.

Pustejovsky, J. (2021). *clubSandwich: Cluster-Robust (Sandwich) Variance Estimators with Small-Sample Corrections*. R package version 0.5.3.

Raudenbush, S. W. (2009). Analyzing effect sizes: Random-effects models. *The Handbook of Research Synthesis and Meta-Analysis*, 2:295–316.

Riley, R. D., Abrams, K., Lambert, P., Sutton, A., and Thompson, J. (2007). An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes. *Statistics in Medicine*, 26(1):78–97.

Riley, R. D., Burchill, S., Abrams, K. R., Heney, D., Lambert, P. C., Jones, D. R., Sutton, A. J., Young, B., Wailoo, A. J., and Lewis, I. (2003). A systematic review and evaluation of the use of tumour markers in paediatric oncology: Ewing's sarcoma and neuroblastoma. *Health Technology Assessment*, 7(5).

Schulze, R. (2004). *Meta-analysis: A comparison of approaches*. Hogrefe Publishing.

Sidik, K. and Jonkman, J. N. (2002). A simple confidence interval for meta-analysis. *Statistics in Medicine*, 21(21):3153–3159.

Sidik, K. and Jonkman, J. N. (2005a). A note on variance estimation in random effects meta-regression. *Journal of Biopharmaceutical Statistics*, 15(5):823–838.

Sidik, K. and Jonkman, J. N. (2005b). Simple heterogeneity variance estimation for meta-analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(2):367–384.

Stekhoven, D. J. and Bühlmann, P. (2012). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118.

Tipton, E. (2015). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods*, 20(3):375.

Tipton, E. and Pustejovsky, J. E. (2015). Small-sample adjustments for tests of moderators and model fit using robust variance estimation in meta-regression. *Journal of Educational and Behavioral Statistics*, 40(6):604–634.

Van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45:1–67.

Van Den Noortgate, W. and Onghena, P. (2005). Parametric and nonparametric bootstrap methods for meta-analysis. *Behavior Research Methods*, 37(1):11–22.

Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., Kuss, O., Higgins, J., Langan, D., and Salanti, G. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods*, 7(1):55–79.

Viechtbauer, W., López-López, J. A., Sánchez-Meca, J., and Marín-Martínez, F. (2015). A comparison of procedures to test for moderators in mixed-effects meta-regression models. *Psychological Methods*, 20(3):360–374.

Welz, T., Doebler, P., and Pauly, M. (2021). Fisher transformation based confidence intervals of correlations in fixed-and random-effects meta-analysis. *British Journal of Mathematical and Statistical Psychology*.

Welz, T. and Pauly, M. (2020). A simulation study to compare robust tests for linear mixed-effects meta-regression. *Research Synthesis Methods*, 11(3):331–342.

Welz, T., Viechtbauer, W., and Pauly, M. (2022). Cluster-robust estimators for bivariate mixed-effects meta-regression. *arXiv preprint arXiv:2203.02234*.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838.

Wilson, J. (2010). Volume of n-dimensional ellipsoid. *Sciencia Acta Xaveriana*, 1(1):101–6.

Wu, C.-F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *Annals of Statistics*, 14(4):1261–1295.

Zhang, J.-T. (2012). An approximate Hotelling T2-test for heteroscedastic one-way MANOVA. *Open Journal of Statistics*, 2(1):1–11.

Part II

Publications

Article 1

Welz, T., and Pauly, M. (2020). A simulation study to compare robust tests for linear mixed-effects meta-regression. *Research Synthesis Methods*, **11**(3), 331–342. https://doi.org/10.1002/jrsm.1388.

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

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

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 $\varepsilon_i \sim N(0, \sigma_i^2)$. 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 $(\sigma_i^2, \boldsymbol{\beta}, \tau^2)$, 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

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

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]:

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

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

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

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 $\mathbf{V} = \text{diag}\left(\sigma_i^2 + \hat{\tau}^2\right)$ 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,

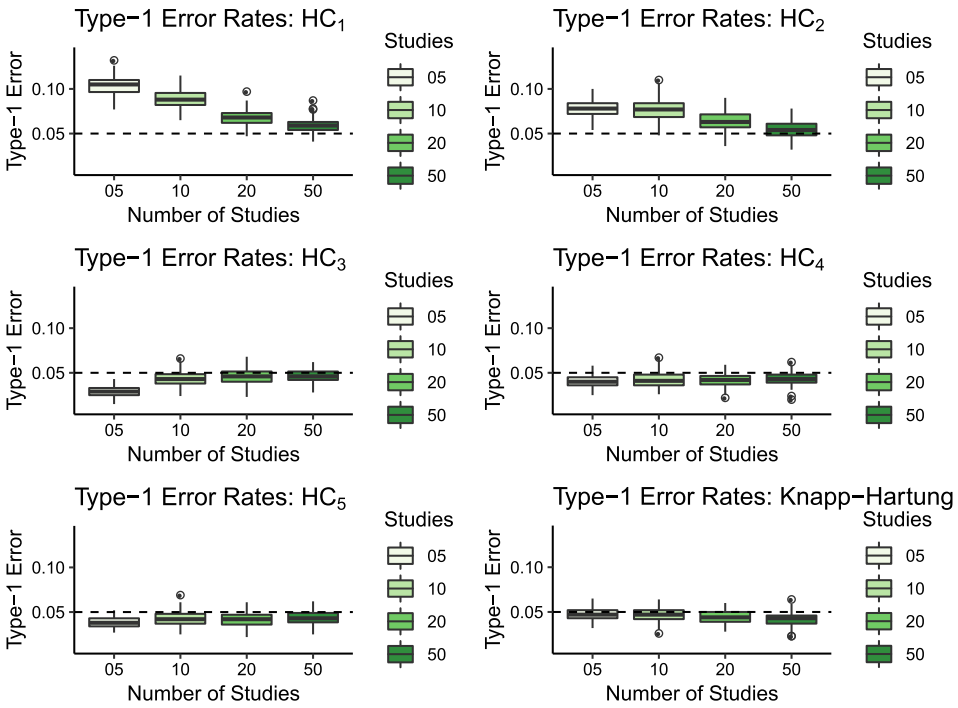

FIGURE 1 Type 1 error of tests based on the White-type estimators HC_1-HC_5 and the Knapp-Hartung correction HC_{KH} for varying number of studies $K \in \{5, 10, 20, 50\}$ and $\tau^2 \in \{0.1, 0.2, ..., 0.9\}$—with unbalanced study sizes and standardized mean difference (SMD) as effect measure. Each 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]

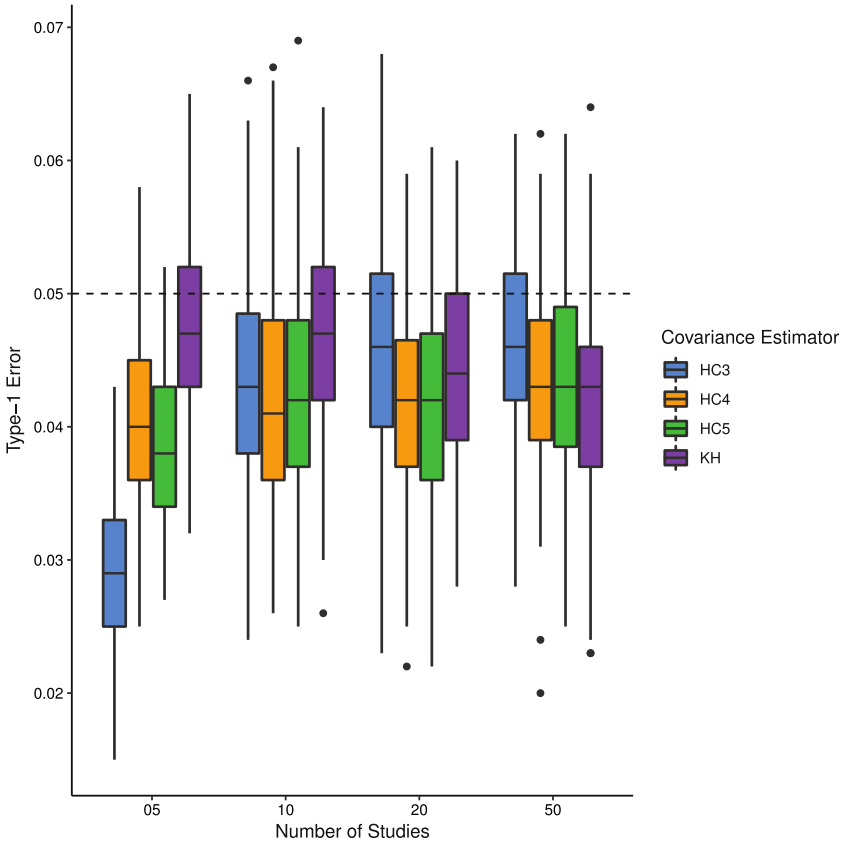

FIGURE 2 Type 1 error based on HC_3-HC_5 and HC_{KH} for $K \in \{5, 10, 20, 50\}$, $\tau^2 \in \{0.1, 0.2, ..., 0.9\}$—with unbalanced study sizes and standardized mean difference (SMD) as effect measure [Colour figure can be viewed at wileyonlinelibrary.com]

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

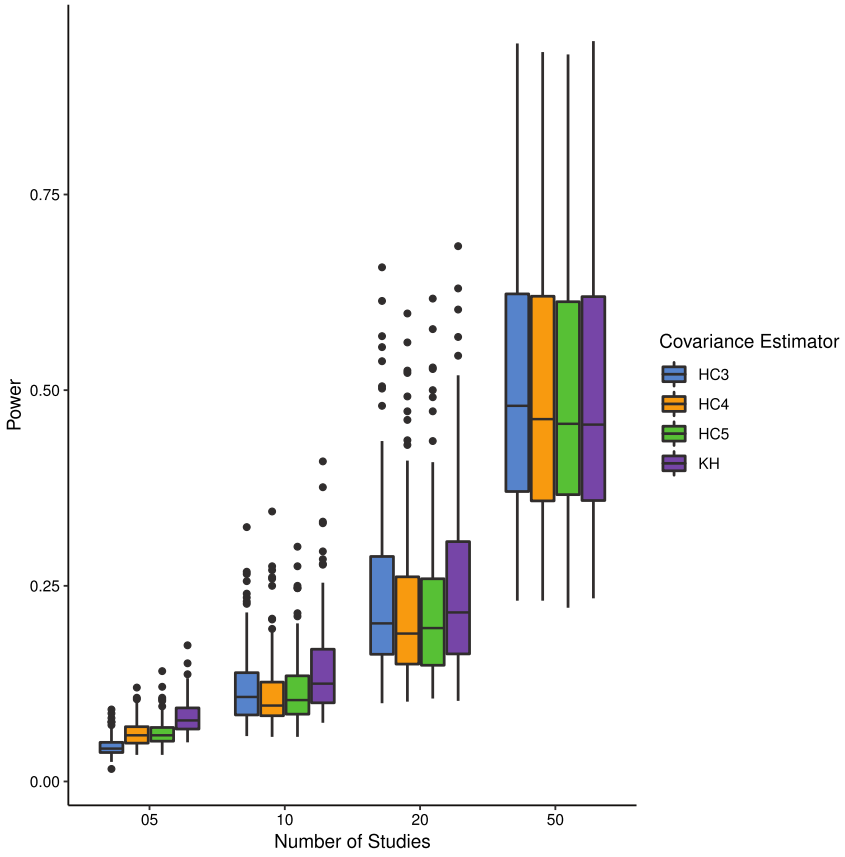

FIGURE 3 Power of tests based on HC_3-HC_5 and HC_{KH} for $K \in \{5, 10, 20, 50\}$, $\tau^2 \in \{0.1, 0.2, ..., 0.9\}$, and $\beta_1 = 0.2$—with unbalanced study sizes and standardized mean difference (SMD) as effect measure [Colour figure can be viewed at wileyonlinelibrary.com]

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 $\hat{\mathbf{\Sigma}}_\ell = \sum_{j=1}^{K} v_{j,\ell} \cdot \hat{\epsilon}_j^2$, $\ell = 0, \ldots, 5$ 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

Article 2

Welz, T., Doebler, P. and Pauly, M. (2021). Fisher transformation based confidence intervals of correlations in fixed- and random-effects meta-analysis. *British Journal of Mathematical and Statistical Psychology*, **75**(1), 1–22. https://doi.org/10.1111/bmsp.12242.

the british psychological society
promoting excellence in psychology

British Journal of Mathematical and Statistical Psychology (2021)

published by John Wiley & Sons Ltd
on behalf of British Psychological Society

www.wileyonlinelibrary.com

Fisher transformation based confidence intervals of correlations in fixed- and random-effects meta-analysis

Thilo Welz*, Philipp Doebler and Markus Pauly

Department of Statistics, Mathematical Statistics and Applications in Industry, TU Dortmund University, Germany

Meta-analyses of correlation coefficients are an important technique to integrate results from many cross-sectional and longitudinal research designs. Uncertainty in pooled estimates is typically assessed with the help of confidence intervals, which can double as hypothesis tests for two-sided hypotheses about the underlying correlation. A standard approach to construct confidence intervals for the main effect is the Hedges-Olkin-Vevea Fisher-z (HOVz) approach, which is based on the Fisher-z transformation. Results from previous studies (Field, 2005, *Psychol. Meth*., 10, 444; Hafdahl and Williams, 2009, *Psychol. Meth*., 14, 24), however, indicate that in random-effects models the performance of the HOVz confidence interval can be unsatisfactory. To this end, we propose improvements of the HOVz approach, which are based on enhanced variance estimators for the main effect estimate. In order to study the coverage of the new confidence intervals in both fixed- and random-effects meta-analysis models, we perform an extensive simulation study, comparing them to established approaches. Data were generated via a truncated normal and beta distribution model. The results show that our newly proposed confidence intervals based on a Knapp-Hartung-type variance estimator or robust heteroscedasticity consistent sandwich estimators in combination with the integral z-to-r transformation (Hafdahl, 2009, *Br. J. Math. Stat. Psychol*., 62, 233) provide more accurate coverage than existing approaches in most scenarios, especially in the more appropriate beta distribution simulation model.

1. Introduction

Quantifying the association of metric variables with the help of the Pearson correlation coefficient is a routine statistical technique for understanding patterns of association. It is a basic ingredient of the data analysis of many cross-sectional and longitudinal designs, and is also indispensable for various psychometric and factor-analytic techniques. When several reports are available for comparable underlying populations, meta-analytic methods allow the available evidence to be pooled (Hedges & Olkin, 1985; Hunter & Schmidt, 2004), resulting in more stable and precise estimates.

*Correspondence should be addressed to Thilo Welz, Department of Statistics, Mathematical Statistics and Applications in Industry, TU Dortmund University, Logistik Campus, Joseph- von-Fraunhofer-Straße 2-4, 44227 Dortmund, Germany (email: thilo.welz@tu-dortmund.de).

DOI:10.1111/bmsp.12242

Systematic reviews based on meta-analyses of correlations are among the most cited in industrial and organizational psychology, clinical psychology and educational psychology (e.g. Aldao, Nolen-Hoeksema, & Schweizer, 2010; Barrick & Mount, 1991; Sirin, 2005 each with several thousand citations), and the methodological monograph on pooling correlations of Hunter and Schmidt (2004) is approaching 10,000 citations on Google Scholar at the time of writing. In addition, pooled correlations are the basis for meta-analytic structural equation modelling (e.g., Cheung, 2015; Jak, 2015, and registered replication efforts pool correlations to reassess findings of others (e.g., Open Science Collaboration, 2015).).

1.1. The importance of confidence intervals for pooled correlations

Schulze (2004) provides a comprehensive summary of fixed- and random-effects meta-analysis of correlations. The best-known approaches are based on Fisher's z transformation (Field, 2001, 2005; Hafdahl & Williams, 2009; Hedges & Olkin, 1985) or on direct synthesis of correlations via the Hunter–Schmidt (HS) method (Hunter & Schmidt, 1994; Schulze, 2004). Regardless of the method and the purpose of the meta-analysis, the point estimate of the correlation is accompanied by an estimate of its uncertainty, in the form of a standard error (SE) or a confidence interval (CI). Since the absolute value of a correlation is bounded by 1, a CI might be asymmetric in this context, that is, not centred around the point estimate. Also, CIs are often more useful than SEs, because a null hypothesis of the form $H_0 : \rho = \rho_0$ can be rejected at level α if a $100\,(1-\alpha)\%$ CI does not include ρ_0 (duality of hypothesis testing and CIs). A CI's coverage is ideally close to the nominal $1-\alpha$ level; for example, a multi-centre registered replication report does want to rely either on an anti-conservative (too narrow) CI that is overly prone to erroneously rejecting previous research, or on a conservative (too wide) CI lacking statistical power to refute overly optimistic point estimates. Despite methodological developments since the late 1970s, the choice of a CI for a pooled correlation should be a careful one: simulation experiments reported in this paper reinforce the finding that CIs are too liberal when heterogeneity is present. The main objective of this paper is a systematic investigation of competing methods, especially when moderate or even substantial amounts of heterogeneity are present, promising refined meta-analytic methods for correlations, especially those based on the Fisher z transformation. The remainder of this introduction reviews results for (z-transformation-based) pooling, and briefly introduces relevant methods for variance estimation.

1.2. Pooling (transformed) correlation coefficients

A line of research summarized in Hunter and Schmidt (1994) pools correlation coefficients on the original scale from -1 to 1. One of the merits of the HS methodology is a clear rationale for artefact corrections, that is, correlations are disattenuated for differences at the primary report level in reliability or variable range. While this part of the HS methodology is beyond the scope of the current paper, CIs originating from Osburn and Callender (1992) are studied here as an HS-based reference method (see also Field, 2005).

Fisher's z-transformation (= *areatangens hyperbolicus*) maps the open interval $(-1, 1)$ to the real number line. Working with z values of correlations avoids problems arising at the bounds and makes normality assumptions of some meta-analytic models more plausible (Hedges & Olkin, 1985). Field (2001) presents a systematic simulation study, and describes scenarios with too liberal behaviour of the HS methodology, but also reports problems with z-transformed pooled values. A simulation strategy is also at the

core of Field (2005), who places a special emphasis on heterogeneous settings. He finds similar point estimates for z-transformation-based and HS pooling, with the CIs from the HS method too narrow in the small-sample case. The simulation study of Hafdahl and Williams (2009) includes a comprehensive account of random-effects modelling and related sources of bias in point estimates. Focusing on point estimation, Hafdahl and Williams (2009) defend z-transformed pooling, but Hafdahl (2009) recommends the integral z-to-r transformation as a further improvement. In the spirit of Hafdahl and Williams (2009), the current paper focuses on variance estimators and resulting CIs, especially in the case of heterogeneity.

1.3. Estimating between-study variance

All CIs studied here are of the form $g\left(\hat{\theta} \pm \hat{\sigma}_{\hat{\theta}}\right)$, for an appropriate back-transformation g (which is not needed in the HS approach), a point estimator $\hat{\theta}$ and its SE estimator $\hat{\sigma}_{\hat{\theta}}$, which depends on the between-study variance estimation. The quality of the CI will depend on an appropriate choice. In other words, especially when primary reports are heterogeneous and the underlying study-specific true correlations vary, good estimators of the between study variance are needed to obtain neither too wide nor too narrow CIs.

The comprehensive study of Veroniki et al., (2016) supports restricted maximum likelihood estimation (REML) as a default estimator of the between-study variance. Since large values of the mean correlation cause REML convergence problems, the robust two-step Sidik and Jonkman (2006) estimator is adopted here. Recently, Welz and Pauly (2020) showed that in the context of meta-regression, the Knapp–Hartung (KH) adjustment (Hartung, 1999; Hartung & Knapp, 2001) aided (co)variance estimation, motivating the inclusion of KH-type CIs in the subsequent comparison.

Less well known in the meta-analysis literature are bootstrap methods for variance estimation, which are not necessarily based on a parametric assumption for the random-effects distribution. The Wu (1986) wild bootstrap intended for heteroscedastic situations is evaluated here. Bootstrapping is complemented by sandwich estimators (heteroscedasticity consistent, HC; White, 1980) which Viechtbauer, López-López, Sánchez-Meca, and Marn-Martnez (2015) introduced in the field of meta-analysis. Recently, a wide range of HC estimators were calculated by Welz and Pauly (2020), whose comparison also includes the more recent HC4 and HC5 estimators (Cribari-Neto, Souza, & Vasconcellos, 2007; Cribari-Neto & Zarkos, 2004). In sum, the following comparison includes a comprehensive collection of established and current variance estimators and resulting CIs.

In Section 2 we introduce the relevant models and procedures for meta-analyses of correlations with more technical detail, as well as our proposed refinements. In Section 3 we perform an extensive simulation study and present the results. In Section 4 we present an illustrative data example on the association of conscientiousness (in the sense of the NEO-PI-R; Costa Jr and McCrae, 1985, 2008) and medication adherence (Molloy, O'Carroll, & Ferguson, 2013). Section 5 concludes the paper with a discussion of our findings and give an outlook for future research.

2. Meta-analyses of Pearson correlation coefficients

For a bivariate metric random vector (X,Y) with existing second moments the correlation coefficient $\rho = \mathrm{Cov}(X,Y)/\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}$ is usually estimated with the (Pearson) correlation coefficient

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}, \tag{1}$$

where $(x_i, y_i),\ i = 1, \cdots, n$, are independent observations of (X, Y).

The Pearson correlation coefficient is asymptotically consistent, that is, for large sample sizes, its value converges to the true ρ. It is also invariant under linear transformations of the data. However, its distribution is difficult to describe analytically and it is not an unbiased estimator of ρ, with an approximate bias of $\mathbb{E}(r - \rho) \approx -\frac{1}{2}\rho(1 - \rho^2)/(n - 1)$ (Hotelling, 1953).

As correlation-based meta-analyses with r as effect measure occur frequently in psychology and the social sciences we briefly recall the two standard models (see Schwarzer, Carpenter, & Rücker, 2015): the fixed- and random-effects models. The *fixed-effect* meta-analysis model is defined as

$$y_i = \mu + \varepsilon_i, i = 1, \ldots, K, \tag{2}$$

where μ denotes the common (true) effect, that is, the (transformed) correlation in our case, K the number of available primary reports, and y_i the observed effect in the i th study. The model errors ϵ_i are typically assumed to be normally distributed with $\varepsilon_i \overset{\text{ind}}{\sim} N\left(0, \sigma_i^2\right)$. In this model the only source of sampling error comes from *within* the studies. The estimate of the main effect μ is then computed as a weighted mean via

$$\hat{\mu} = \sum_{i=1}^{K} \frac{w_i}{w} y_i, \tag{3}$$

where $w := \sum_{i=1}^{K} w_i$ and the study weights $w_i = \hat{\sigma}_i^{-2}$ are the reciprocals of the (estimated) sampling variances $\hat{\sigma}_i^2$. This is known as the *inverse variance method*. The fixed-effect model typically underestimates the observed total variability because it does not account for between-study variability (Schwarzer et al., 2015). However, it has the advantage of being able to pool observations, if individual patient data (IPD) are in fact available, allowing for greater flexibility in methodology in this scenario.

The *random-effects* model extends the fixed-effect model by incorporating a random effect that accounts for between-study variability, such as differences in study population or execution. It is given by

$$\mu_i = \mu + u_i + \varepsilon_i,\ \ i = 1, \cdots, K, \tag{4}$$

where the random effects u_i are typically assumed to be independent and $N(0, \tau^2)$distributed with between-study variance $\tau^2$and $\varepsilon_i \overset{\text{ind}}{\sim} \mathcal{N}\left(0, \sigma_i^2\right)$. Furthermore, the random effects $(u_i)_i$ and the error terms $(\epsilon_i)_i$ are jointly independent. Thus, for $\tau^2 = 0$, the fixed-effect model is a special case of the random-effects model. The main effect is again estimated via the weighted mean $\hat{\mu}$ given in equation (3) with study weights now defined as $w_i = \left(\hat{\sigma}_i^2 + \hat{\tau}^2\right)^{-1}$.

A plethora of approaches exist for estimating the heterogeneity variance τ^2. Which estimator should be used has been discussed for a long time, without reaching a definitive conclusion. However, a consensus has been reached that the popular and easy-to-calculate DerSimonian–Laird estimator is not the best option. Authors such as Veroniki et al., (2016) and Langan et al., (2019) have recommended using iterative estimators for τ^2.

We therefore (initially) followed their suggestion and used the REML estimator. However, in some settings, such as large ρvalues, the REML estimator had trouble converging, even after the usual remedies of utilizing step halving and/or increasing the maximum number of permitted iterations. We therefore opted to use the two-step estimator suggested by Sidik and Jonkman (SJ), which is defined by starting with a rough initial estimate of $\hat{\tau}_0^2 = \frac{1}{K}\sum_{i=1}^{K}(y_i - \bar{y})^2$ and is then updated via the expression

$$\hat{\tau}_{\mathrm{SJ}}^2 = \frac{1}{K-1}\sum_{i=1}^{K} w_i (y_i - \hat{\mu})^2, \tag{5}$$

where $w_i = \left(\hat{\tau}_0^2 / \left(\hat{\sigma}_i^2 + \hat{\tau}_0^2\right)\right)^{-1}$ and $\hat{\mu} = \sum_{i=1}^{K} w_i y_i / \sum_{i=1}^{K} w_i$ (Sidik & Jonkman, 2005). A comprehensive comparison of heterogeneity estimators for τ^2in the context of random-effects meta-analyses for correlations would be interesting but is beyond the scope of this paper.

Before discussing different CIs for the common correlation μ within model (4), we take a short excursion on asymptotics for r in the one-group case.

2.1. Background: Asymptotic confidence intervals

Assuming bivariate normality of (X, Y), r is approximately distributed as $\mathcal{N}(\rho, (1-\rho^2)^2/n)$ for large sample sizes n (Lehmann, 2004). Here, bivariate normality is a necessary assumption to obtain $(1-\rho^2)^2$ in the asymptotic variance (Omelka & Pauly, 2012). Plugging in r, we obtain an approximate $100(1-\alpha)\%$ CI of the form $r \pm u_{1-\alpha/2}(1-r^2)/\sqrt{n}$, where $u_{1-\alpha/2}$ denotes the $(1-\alpha/2)$ quantile of the standard normal distribution.

In fixed-effect meta-analyses, when IPD are available, this result can be used to construct a CI based on pooled data: calculating $\hat{\rho}_{\mathrm{pool}}$, the pooled sample correlation coefficient, we obtain an approximate CI for ρ as

$$\hat{\rho}_{\mathrm{pool}} \pm u_{1-\alpha/2}\frac{\left(1-\hat{\rho}_{\mathrm{pool}}^2\right)}{\sqrt{N}}, \tag{6}$$

where $N := \sum_{i=1}^{K} n_i$ is the pooled sample size. As this pooling of observations only makes sense if we assume that each study has the same underlying effect, this approach is not feasible for a random-effects model, even if IPD were available. In any case, even under IPD and a fixed-effects model, this CI is sensitive to the normality assumption and the underlying sample size, as we demonstrate in Table 1 for the case $K = 1$. We simulated bivariate data from standard normal and standardized lognormal distributions[1] with correlation $\rho \in \{.3, .7\}$and study size $n \in \{20, 50, 100\}$. In each setting we performed N = 10,000 simulation runs. For the lognormal data coverage is extremely poor in all cases, ranging from 53–80%. For the normally distributed case coverage was somewhat low at 90% for n = 20 but improved for larger sample sizes. This case study clearly illustrates that alternatives are needed when the data cannot be assumed to stem from a normal distribution or sample sizes are small.

[1] Further details regarding the data generation can be found in the online supplementary materials.

Table 1. Empirical coverage of the asymptotic confidence interval for $K = 1$, study size $n \in \{20, 50, 100\}$ and correlation $\rho \in \{0.3, 0.7\}$

	N			
Distribution	ρ	20	50	100
Normal	.3	.90	.93	.94
	.7	.90	.92	.94
Lognormal	.3	.79	.80	.79
	.7	.63	.57	.53

After this short excursion we return to model (4) and CIs for ρ.

2.2. The Hunter–Schmidt approach

The aggregation of correlations in the Hunter–Schmidt approach is done by sample size weighting:

$$r_{\mathrm{HS}} = \frac{\sum_{i=1}^{K} n_i r_i}{\sum_{i=1}^{K} n_i}. \tag{7}$$

Several formulae have been recommended for estimating the sampling variance of this mean effect size estimate. We opted for a suggestion by Osburn and Callender (1992),

$$\hat{\sigma}^2_{\mathrm{HS}} = \frac{1}{K}\left(\frac{\sum_{i=1}^{K} n_i (r_i - r_{\mathrm{HS}})^2}{\sum_{i=1}^{K} n_i}\right), \tag{8}$$

which is supposed to perform reasonably well in both heterogeneous and homogeneous settings (Schulze, 2004). In the simulation study we will investigate whether this is in fact the case for the resulting CI, $r_{\mathrm{HS}} \pm u_{1-\alpha/2}\hat{\sigma}_{\mathrm{HS}}$.

2.3. Confidence intervals based on the Fisher z transformation

A disadvantage of the asymptotic confidence interval (6) is that the variance of the limit distribution depends on the unknown correlation ρ. This motivates a variance-stabilizing transformation. A popular choice for correlation coefficients is the *Fisherz transformation* (Fisher, 1915),

$$\rho \mapsto z = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right) = \operatorname{atanh}(\rho). \tag{9}$$

The corresponding inverse Fisher transformation is $z \mapsto \tanh(z) = (\exp(2z) - 1)/(\exp(2z) + 1)$.

The variance-stabilizing property of the Fisher transformation follows from the δ-method (Lehmann, 2004); that is, if $\sqrt{n}(r - \rho) \to^d \mathcal{N}(0, (1-\rho^2)^2)$ then $\sqrt{n}(\hat{z} - z) = \sqrt{n}(\operatorname{atanh}(r) - \operatorname{atanh}(\rho)) \to^d \mathcal{N}(0, 1)$. Following, it is reasonable to substitute $\sqrt{n}$by, $\sqrt{n-3}$that is, to approximate the distribution of $\hat{z}$by, $\mathcal{N}(\operatorname{atanh}(r), 1/(n-3))$ still assuming bivariate normality. Thus, a single-group approximate $100(1-\alpha)\%$ CI can be constructed via $\tanh(\hat{z} \pm u_{1-\alpha/2}/\sqrt{N-3})$.

In the random-effects model (4), the z transformation may also be used to construct a CI for the common correlation ρ. Here, the idea is again to use inverse variance weights to define

$$\overline{z} = \frac{\sum_{i=1}^{K} \left(\frac{1}{n_i - 3} + \hat{\tau}^2 \right)^{-1} z_i}{\sum_{i=1}^{K} \left(\frac{1}{n_i - 3} + \hat{\tau}^2 \right)^{-1}}, \tag{10}$$

where $z_i = \operatorname{atanh}(r_i)$. A rough estimate of the variance of $\overline{z}$ is given by $\left(\sum_{i=1}^{K} w_i\right)^{-1}$. In the fixed-effect case with $\tau^2 = 0$ this yields the variance estimate $\left(\sum_{i=1}^{K} (n_i - 3)\right)^{-1} = (N - 3K)^{-1}$. Then $\overline{z}\sqrt{N - 3K}$ approximately follows a standard normal distribution and an approximate $100(1-\alpha)\%$ CI is given by $\tanh\left(\overline{z} \pm u_{1-\alpha/2}/\sqrt{N - 3K}\right)$. Proceeding similarly in the random-effects model (4), one obtains the *Hedges–Olkin–Vevea Fisher-z*(HOV z) CI

$$\tanh\left(\overline{z} \pm u_{1-\alpha/2} / \left(\sum_{i=1}^{K} w_i\right)^{1/2}\right), \tag{11}$$

with $w_i = \left(1/(n_i - 3) + \hat{\tau}^2\right)^{-1}$ (Hafdahl & Williams, 2009; Hedges & Olkin, 1985; Hedges & Vevea, 1998).

2.3.1. Knapp–Hartung-type CI

The above approximation of the variance of $\overline{z}$ via $\left(\sum_{i=1}^{K} w_i\right)^{-1}$ can be rather inaccurate, especially in random-effects models. Although this is the exact variance of $\overline{z}$ when the weights are chosen perfectly as $w_i = \left(\sigma_i^2 + \tau^2\right)^{-1}$, this variance estimate does not protect against (potentially substantial) errors in estimating $\hat{\sigma}_i^2$ and $\hat{\tau}^2$ (Sidik & Jonkman, 2006). Therefore, we propose an improved CI based on the KH method (Hartung & Knapp, 2001). Knapp and Hartung proposed the following variance estimator for the estimate $\hat{\mu}$ of the main effect μ in a random-effects meta-analysis (REMA):

$$\hat{\sigma}_{\mathrm{KH}}^2 = \hat{\mathrm{Var}}_{\mathrm{KH}}(\hat{\mu}) = \frac{1}{K-1} \sum_{i=1}^{K} \frac{w_i}{w} (\hat{\mu}_i - \hat{\mu})^2, \tag{12}$$

where again $w = \sum_{i=1}^{K} w_i$. showed that if $\hat{\mu}$ is normally distributed, then $(\hat{\mu} - \mu)/\hat{\sigma}_{\mathrm{KH}}$ follows a t distribution with $K - 1$ degrees of freedom. Therefore an approximate $100(1-\alpha)\%$ CI for μ is given by

$$\tanh\left(\overline{z} \pm t_{K-1,1-\alpha/2} \cdot \hat{\sigma}_{\mathrm{KH}}\right), \tag{13}$$

where $t_{K-1,1-\alpha/2}$ is the $1 - \alpha/2$ quantile of the t distribution with $K - 1$ degrees of freedom. Because of the approximately normal distribution of z-transformed correlations, the CI ((13)) seems justified. Various authors have highlighted the favourable performance of the KH approach compared to alternative meta-analytic methods (IntHout, Ioannidis, & Borm, 2014; Viechtbauer et al., 2015; Welz & Pauly, 2020). Analogously to (13), we can construct further CIs by using other variance estimation procedures for $\mathrm{Var}(\hat{\mu})$.

2.3.2. Wild bootstrap approach

Another possibility for estimating the variance of $\overline{z}$ is through bootstrapping. Bootstrapping belongs to the class of resampling methods. It allows the estimation of the sampling distribution of most statistics using random sampling methods. The wild bootstrap is a subtype of bootstrapping that is applicable in models which exhibit heteroscedasticity. Roughly speaking, the idea of the wild bootstrap approach is to resample the response variables based on the residuals. The idea was originally proposed by Wu (1986) for regression analysis.

We now propose a confidence interval for ρ based on a (data-dependent) *wild bootstrap* (WBS) approach combined with the $z$transformation. The idea works as follows. We assume an REMA model with Pearson's correlation coefficient as the effect estimate (and $K>3$ studies). Given the estimated study-level correlation coefficients r_i, $i=1,\cdots,K$, we transform these using $z$transformation to $\hat{z}_i$, $i=1,\cdots,K$, and estimate $z=\text{atanh}(\rho)$via $\hat{z}=\sum_i(w_i/w)\hat{z}_i$, where again $w_i=(\hat{\sigma}_i+\hat{\tau}^2)^{-1}$ with $\hat{\sigma}_i^2=\frac{1}{n_i-3}$ and $w=\sum_i w_i$. Here, $\hat{\tau}^2$ may be any consistent estimator of the between-study heterogeneity τ^2, where we have chosen the SJ estimator. We then calculate the estimated residuals $\hat{\varepsilon}_i=\hat{z}-\hat{z}_i$ and use these to generate B new sets of study-level effects $\hat{z}^*_{1b},\ldots,\hat{z}^*_{Kb}, b=1,\ldots,B$. Typical choices for $B$are 1,000 or 5,000. The new study-level effects are generated via

$$\hat{z}^*_{ib}:=\hat{z}_i+\hat{\varepsilon}_i\cdot v_i, \tag{14}$$

where $v_i\sim\mathcal{N}(0,\gamma)$. The usual choice of variance in a WBS is $\gamma=1$. However, we propose a data-dependent choice of either $\gamma_K=(K-1)/(K-3)$ or $\gamma_K=(K-2)/(K-3)$. These choices are based on simulation results, which will be discussed in detail in Section 3. We will later refer to these approaches as WBS1, WBS2 and WBS3, respectively. The corresponding values for γ are 1, $(K-1)/(K-3)$ and $(K-2)/(K-3)$. This allows us to generate B new estimates of the main effect z by calculating

$$\hat{z}^*_b=\frac{\sum_{i=1}^K w^*_{ib}\hat{z}^*_{ib}}{\sum_{i=1}^K w^*_{ib}}, \tag{15}$$

with $w^*_{ib}\equiv w_i$. We then estimate the variance of $\hat{z}$ via the empirical variance of $\hat{z}^*_1,\cdots,\hat{z}^*_B$,

$$\sigma^{*2}_z:=\frac{1}{B-1}\sum_{i=1}^B(\hat{z}^*_i-\overline{z}^*)^2,\quad\text{with}\,\overline{z}^*=\frac{1}{B}\sum_{i=1}^B\hat{z}^*_i$$

It is now possible to construct a CI for z as in equation (13) but with this new variance estimate of $-z$. The CI is back-transformed via the inverse Fisher transformation to obtain a CI for the common correlation ρ, given by

$$\tanh\left(\hat{z}\pm\hat{\sigma}^*_z\cdot t_{K-1,1-\alpha/2}\right). \tag{16}$$

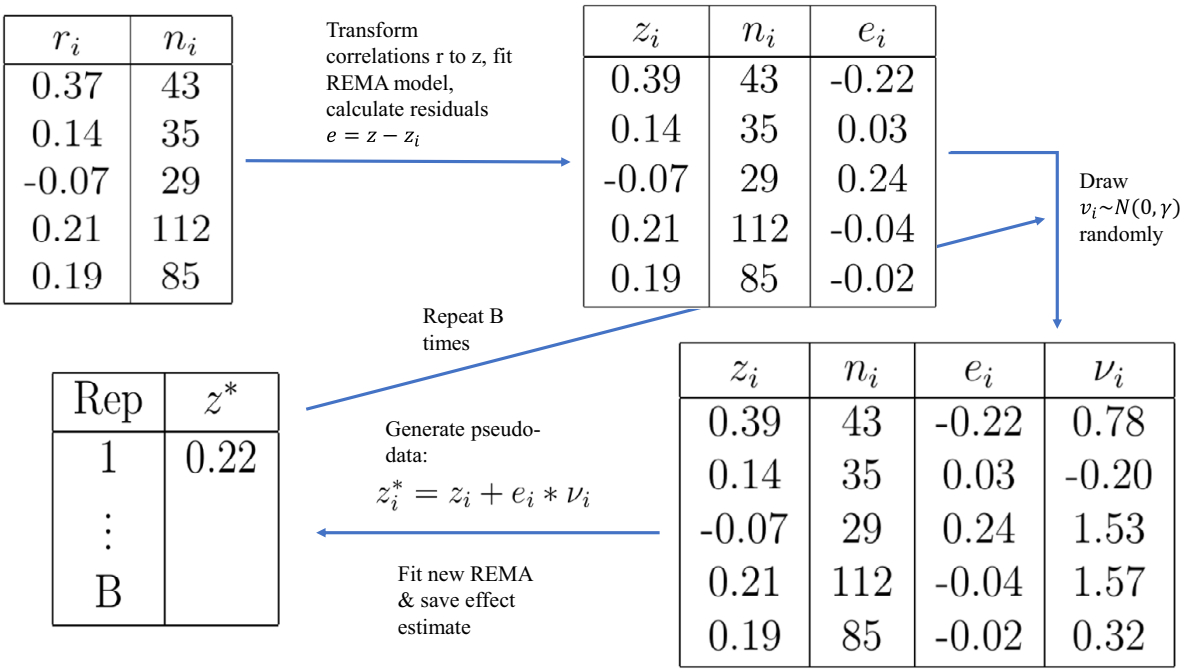

Figure 1. Visual illustration of the wild bootstrap procedure for generating B bootstrap samples of the main effect estimate on the z scale. REMA, random-effects meta-analysis.

Figure 1 provides a visual illustration of the WBS procedure discussed above.

2.3.3. HC-type variance estimators

Last but not least, we employ *heteroscedasticity consistent* variance estimators [sandwich estimators; White, 1980). Different forms (HC0,...,HC5) are in use for linear models (Rosopa, Schaffer, & Schroeder, 2013). The motivation for the robust HC variance estimators is that in a linear regression setting the usual variance estimate is unbiased when unit-level errors are independent and identically distributed. However, when the unit-level variances are unequal, this approach can be biased. If we apply this to the meta-analysis context, the study-level variances are almost always unequal due to varying sample sizes. Therefore, it makes sense to consider variance estimators that are unbiased even when the variances of the unit (study) level variances are different.

The extension of HC estimators to the meta-analysis context can be found in Viechtbauer et al., (2015) for HC_0 and HC_1 and in Welz and Pauly (2020) for the remaining $HC_2, \cdots, HC_5$. Statistical tests based on these robust estimators have been shown to perform well, especially those of types HC_3 and HC_4. In the special case of an REMA they are defined as

$$\hat{\sigma}^2_{HC_3} = \frac{1}{\left(\sum_{i=1}^{K} w_i\right)^2} \sum_{j=1}^{K} w_j^2 \hat{\varepsilon}_j^2 (1 - x_{jj})^{-2}$$

$$\hat{\sigma}^2_{HC_4} = \frac{1}{\left(\sum_{i=1}^{K} w_i\right)^2} \sum_{j=1}^{K} w_j^2 \hat{\varepsilon}_j^2 (1 - x_{jj})^{-\delta_j}, \quad \delta_j = \min\left\{4, \frac{x_{jj}}{\bar{x}}\right\},$$

with $\hat{\varepsilon}_j = \hat{z}_j - \hat{z}, x_{ij} = w_j / \sum_{i=1}^{K} w_i$ and $\overline{x} = K^{-1} \sum_{i=1}^{K} x_{ij}$ [see the Appendix S1 of Welz and Pauly, 2020 for details). Plugging them into equation (13) leads to the confidence intervals

$$\tanh\left(\hat{z} \pm \hat{\sigma}_{\mathrm{HC}_j} \cdot t_{K-1,1-\alpha/2}\right),\ j = 3, 4. \tag{17}$$

2.3.4. Integral z-to-r transformation

There is a fundamental problem with back-transforming CIs on the z scale using the inverse Fisher transformation tanh. Consider a random variable $\xi : \mathcal{N}(\operatorname{artanh}(\rho), \sigma^2)$ with some variance $\sigma^2 > 0$ and $\rho \neq 0$. Then $\rho = \tanh(\mathbb{E}(\xi)) \neq \mathbb{E}(\tanh(\xi))$ by Jensen's inequality. This means the back-transformation introduces an additional bias. A remedy was proposed by Hafdahl (2009), who suggested back-transforming from the z scale using an integral z-to-r transformation. This transformation is the expected value of $\tanh(z)$ where $z : \mathcal{N}(\mu_z, \tau_z^2)$ that is,

$$\psi(\mu_z | \tau_z^2) = \int_{-\infty}^{\infty} \tanh(t) f(t | \mu_z, \tau_z^2) \mathrm{d}t, \tag{18}$$

where f is the density of z. In practice we apply this transformation to the lower and upper confidence limits on the z scale, plugging in the estimates $\hat{z}$ and $\hat{\tau}_z^2$. For example, for the KH-based CI (13) with z scale confidence bounds $\ell = \overline{z} - t_{K-1,1-\alpha/2} \cdot \hat{\sigma}_{\mathrm{KH}}$ and $u = \overline{z} + t_{K-1,1-\alpha/2} \cdot \hat{\sigma}_{\mathrm{KH}}$, with an estimated heterogeneity $\hat{\tau}_z^2$ (on the z scale), the CI is given by

$$\left(\psi(\ell | \hat{\tau}_z^2), \psi(\mathrm{u} | \hat{\tau}_z^2)\right).$$

If the true distribution of $\hat{z}$ is well approximated by a normal distribution and $\hat{\tau}_z^2$ is a good estimate of the heterogeneity variance (on the z scale), ψ should improve the CIs as compared to simply back-transformation with tanh (Hafdahl, 2009). Following this argument, we also suggest using ψ instead of tanh. We calculate the integral with Simpson's rule (Süli & Mayers, 2003), which is a method for the numerical approximation of definite integrals. Following Hafdahl (2009), 150 subintervals over $\hat{z} \pm 5 \cdot \hat{\tau}_{\mathrm{SJ}}$ were used. Note that the HOVz CI is implemented in its original formulation, using tanh.

3. Simulation study

We have suggested several new CIs for the mean correlation ρ, all based on the z transformation, applicable in both, fixed- and random-effects models. In order to investigate their properties (especially coverage of ρ), we perform extensive Monte Carlo simulations. We focus on comparing the coverage of our newly suggested CIs with existing methods.

3.1. Simulation study design

The Pearson correlation coefficient is constrained to lie in the interval $[-1, 1]$. The typical random-effects model $\mu_i = \mu + u_i + \varepsilon_i$, assuming a normal distribution for the random

effect $u_i \sim \mathcal{N}(0, \tau^2)$ and error term $\varepsilon_i \sim \mathcal{N}\left(0, \sigma_i^2\right)$, needs to be adjusted, since values outside of $[-1, 1]$ could result when sampling without any modification.

3.1.1. Model 1

As a first option for generating the (true) study-level correlations, we consider a truncated normal distribution $\rho_i \sim \mathcal{N}(\rho, \tau^2)$: Sampling of ρ_i is repeated until a sample lies within the interval $[-0.999, 0.999]$. This type of truncated normal distribution model was also used in Hafdahl and Williams (2009) and Field (2005). A problem with this modelling approach is that the expected value of the resulting truncated normal distribution is in general not equal to ρ. For a random variable X stemming from a truncated normal distribution with mean μ, variance σ^2, lower bound a and upper bound b,

$$\mathbb{E}(X) = \mu + \sigma \frac{\phi(\Delta_1) - \phi(\Delta_2)}{\delta},$$

where $\Delta_1 = (a - \mu)/\sigma$, $\Delta_2 = (b - \mu)/\sigma$ and $\delta = \Phi(\Delta_2) - \Phi(\Delta_1)$ (Johnson, Kotz, & Balakrishnan, 1994). Here $\phi(\cdot)$ is the probability density function of the standard normal distribution and $\Phi(\cdot)$ its cumulative distribution function. Figure S15 shows the bias in our setting with $a = -0.999$ and $b = 0.999$. The bias is equal to $\sigma(\phi(\Delta_1) - \phi(\Delta_2))/\delta$. In addition to generating a biased effect, the truncation also leads to a reduction of the overall variance, which is smaller than τ^2.

3.1.2. Model 2

We therefore studied a second model, in which we generate the (true) study-level effects ρ_i from transformed beta distributions: $Y_i = 2(X_i - 0.5)$ with $X_i \sim \text{Beta}(\alpha, \beta)$ for studies $i = 1, \cdots, K$. The idea is to choose the respective shape parameters α, β such that

$$E(Y_i) = 2 \cdot \left(\frac{\alpha}{\alpha + \beta} - 0.5 \right) = \rho,$$

$$\text{Var}(Y_i) = \frac{4\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} = \tau^2.$$

The solution to the system of equations above is

$$\alpha = \frac{(1 - \rho)(1 + \rho) - \tau^2}{\tau^2} \cdot \left(\frac{1 + \rho}{2} \right),$$

$$\beta = \left(\frac{1 - \rho}{1 + \rho} \right) \alpha.$$

In this second simulation scenario we also truncate the sampling distribution of the correlation coefficients to $[-0.999, 0.999]$, but values outside of this interval are considerably rarer. The second model has the advantages that the expected value and variance are approximately correct, unlike in the first (truncated) model. A disadvantage is

that for extreme τ^2 values, the above solution for α (and thus β) may become negative, which is undefined for parameters of a beta distribution. However, this was not a concern for the parameters considered in our simulation study and only occurs in more extreme scenarios.

3.1.3. Parameter choices

In order to get a broad overview of the performance of all methods, we simulated various configurations of population correlation coefficient, heterogeneity, sample size and number of studies. Here we chose the correlations $\rho \in \{0, .1, .3, .5, .6, .7, .8, .9\}$and heterogeneity $\tau \in \{0, 0.16, 0.4\}$. We used the same values for τ as Hafdahl and Williams (2009), to enable comparability of our simulation studies. Moreover, we considered small to large numbers $K \in \{5, 10, 20, 40\}$ of studies with different study sizes. For $K = 5$, we considered $\vec{n} = (15, 16, 19, 23, 27)$ as vector of 'small' study sizes and $4 \cdot \vec{n}$ for larger study sizes, corresponding to an average study size $(\overline{n})$ of 20 and 80 subjects, respectively. For all other choices of K we proceeded similarly, stacking copies $\vec{n}$ behind each other, for example, the sample size vectors $\left(\vec{n}, \vec{n}\right)$ and $4 \cdot \left(\vec{n}, \vec{n}\right)$ for $K = 10$. By way of comparison, Hafdahl and Williams (2009) considered $5 \leq K \leq 30$. As we wanted to capture the methods' behaviour when many studies are present, we also included the setting $K = 40$ in our simulation study. Additionally, we accounted for variability in study sizes, which will be present in virtually any meta-analysis in practice. Additionally, we considered two special scenarios: the case of few and heterogeneous studies, with study size vector $(23, 19, 250, 330, 29)$ and the case of many large studies, with study size vector $\left(\vec{n}^*, \vec{n}^*\right)$ with $\vec{n}^* = (210, 240, 350, 220, 290, 280, 340, 400, 380, 290)$. The latter case corresponds to $K = 20$ studies with an average of 300 study subjects.

Thus, in total we simulated $8(\rho) \times 3(\tau^2) \times 10(K, \text{studysizevector}) \times (\text{models}) = 480$ different scenarios for each type of confidence interval discussed in this paper. For each scenario we performed $N = 10,000$ simulation runs, where for the WBS CI each run was based upon $B = 1,000$ bootstrap replications. The primary focus was on comparing empirical coverage, with nominal coverage being $1 - \alpha = .95$. For 10,000 iterations, the Monte Carlo standard error of the simulated coverage will be approximately $\sqrt{.95 \times .05/10000} \approx 0.218\%$, using the formula provided in the recent work on simulation studies by Morris, White, and Crowther (2019).

All simulations were performed using the open-source software R. The R scripts written by the first author especially make use of the *metafor* package for meta-analysis (Viechtbauer, 2010).

3.2. Results

For ease of presentation, we aggregated the multiple simulation settings with regard to number and size of studies. The graphics therefore display the mean observed coverage for each confidence interval type and true main effect ρ. Results are separated by heterogeneity τ^2 and simulation design. The latter refers to the truncated normal distribution approach and the transformed beta distribution approach, respectively. More detailed simulation results for all settings considered are given in the Appendix S1.

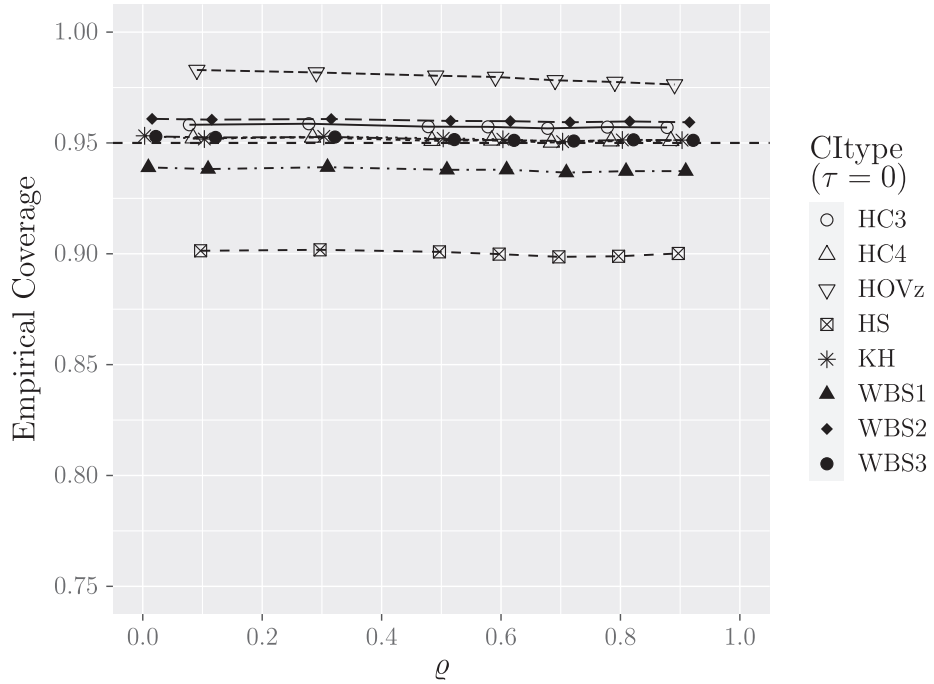

Figure 2. Mean Coverage for truncated normal distribution model with $\tau = 0$, aggregated across all number of studies and study size settings. HC, heteroscedasticity-consistent; HOVz, Hedges–-Olkin–Vevea Fisher z; HS, Hunter–Schmidt; KH, Knapp–Hartung; WBS, wild bootstrap

3.2.1. Coverage

We first discuss the results based on the truncated normal distribution (model 1). In the case of no heterogeneity (fixed-effect model), Figure 2 shows that the new methods control the nominal coverage of 95% well. Only the first wild bootstrap (WBS1) CI exhibits liberal behaviour, yielding empirical coverage of approximately 93.5% . The HS approach only provides 90% coverage, and HOVz was slightly conservative with (mean) coverage of around 97–98% . Moreover, in the fixed-effect model the value of ρ did not affect any of the methods.

In the truncated normal set-up with moderate heterogeneity of $\tau = 0.16$ in Figure 3, several things change. First, there is a strong drop-off in coverage for higher correlations $\rho \geq .8$. For HS this drop-off occurs earlier for $\rho \geq .7$. Second, for $\rho \leq .7$, HS is even more liberal than for $\tau = 0$, with coverage around 87.5%. Additionally, HOVz is no longer conservative but becomes more liberal than WBS1 with estimated coverage probabilities around 90–94% for $\rho \leq .7$. For all new methods a slight decrease in coverage can be observed for increasing values of ρ from 0 to .7. Moreover, there is a slight uptick at $\rho = .8$ for HOVz, followed by a substantial drop-off. Overall the WBS3, HC_3, HC_4 and KH CIs show the best control of nominal coverage in this setting.

We now consider model 2 with a transformed beta distribution model. In the fixed-effects case ($\tau^2 = 0$) the two models are equivalent so we obtain the same coverage as in Figure 2. For moderate heterogeneity ($\tau = 0.16$; see Figure 4), our newly proposed methods clearly outperform HOVz and HS, with a good control of nominal coverage. Only for $\rho = .9$ is their coverage slightly liberal. WBS1 performs just slightly worse than the other new CIs. The observed coverage for HS is around 86–88% for $\rho \leq .7$ and drops to just below 80% for $\rho = .9$. For $\rho > .6$ the HOVz CI is even worse, with values dropping (substantially) below 75%.

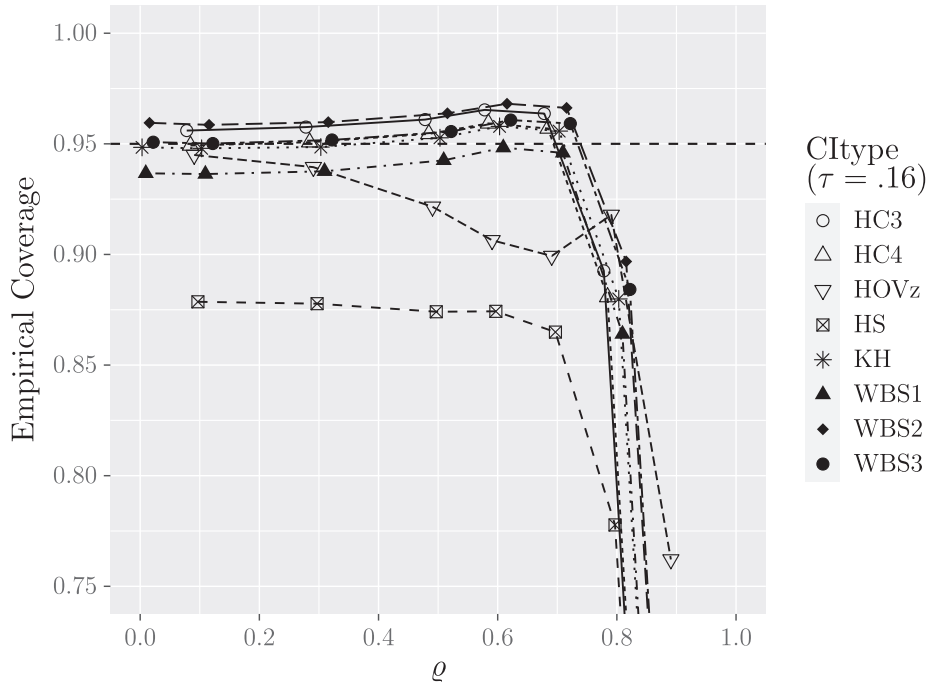

Figure 3. Mean coverage for truncated normal distribution model with $\tau = 0.16$, aggregated across all number of studies and study size settings. HC, heteroscedasticity-consistent; HOVz, Hedges–-Olkin–Vevea Fisher z; HS, Hunter–Schmidt; KH, Knapp–Hartung; WBS, wild bootstrap

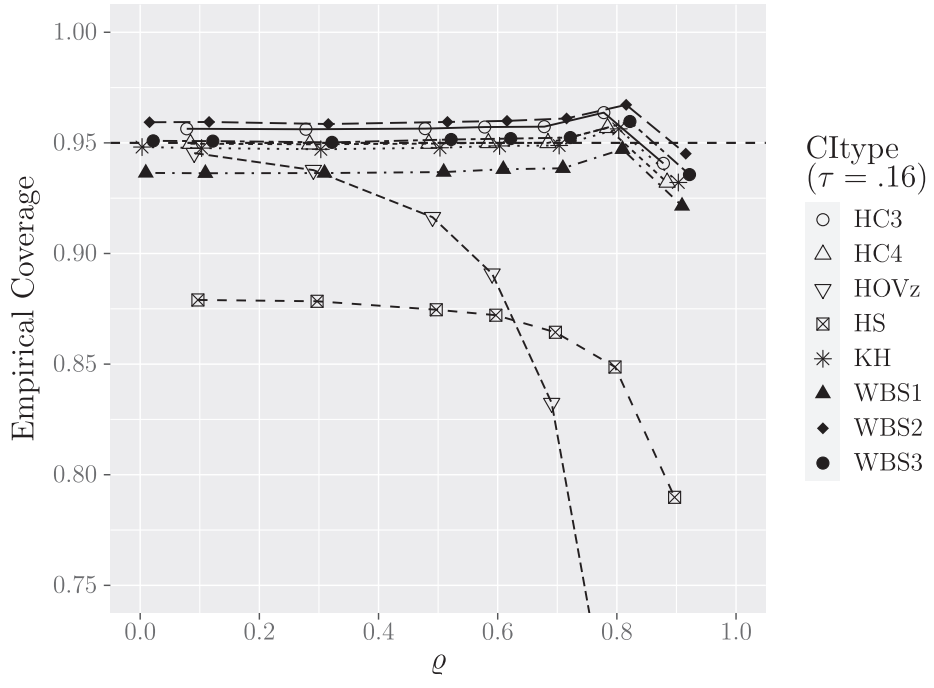

Figure 4. Mean coverage for transformed beta distribution model with $\tau = 0.16$, aggregated across all number of studies and study size settings. HC, heteroscedasticity-consistent; HOVz, Hedges–-Olkin–Vevea Fisher z; HS, Hunter–Schmidt; KH, Knapp–Hartung; WBS, wild bootstrap

For ease of presentation, the results for the case of extreme heterogeneity with $\tau = 0.4$ are given in the Appendix S1. Here, we only summarize important points from Figures S13–S14. In the truncated normal distribution model we observe that HS again has

unsatisfactory coverage, compared with the other approaches. For our new CIs based on the Fisher transformation, for small K, coverage is approximately correct for $\rho \leq .6$ and then drops off considerably. HOVz is slightly liberal with coverage around 90% for $\rho \leq .6$ and then drops off strongly. This holds for both smaller and larger studies with $\overline{n} \in \{20, 80\}$, respectively. For an increasing number of studies K, HOVz remains largely unchanged, whereas coverage of the new methods gets progressively worse (i.e., the drop-off in coverage occurs earlier for an increasing number of studies). For $K = 40$ the new CIs only have correct coverage for $\rho \leq .3$. In the case of the beta distribution model with $\tau = 0.4$ the new CIs provide correct coverage for $\rho \leq .7$ in all scenarios, dropping off after this threshold. HOVz is highly inadequate, with coverage growing progressively worse for increasing K. HOVz only has correct coverage for simultaneously $\rho \leq .1$ and large K. For $K = 5$, HS has coverage up to 82%, decreasing for increasing values of ρ. However, for increasing number of studies (whether large or small), HS appears to converge towards nominal coverage. In particular, for $K = 40$ and $\rho > .7$, HS provides the most accurate coverage under the beta distribution model.

3.2.2. Interval lengths

We simulated the expected confidence interval lengths for all methods discussed in this paper. The detailed results are provided in Figures S7–S12. The results again depend on both the assumed model and the amount of heterogeneity τ.

Generally we observe that the confidence intervals become increasingly narrow for increasing values of ρ and increasingly wide for larger values of τ. For the truncated normal distribution model and $\tau = 0$, HS (on average) yields the shortest confidence intervals and HOVz the widest, with the other CIs lying in between with quite similar lengths. Only for $K = 5$ are the CIs based on the wild bootstrap quite wide, indicating that potentially more studies are required to reliably use WBS-based approaches. For $\tau = 0.16$, HS again yields the shortest CIs in all scenarios. For small K, the WBS approaches yield the widest CIs, and for more studies, HOVz is the widest, when ρ is small, but becoming nearly as narrow as HS when ρ is close to 1. The lengths of the other CIs are nearly identical for $K = 40$, whereas for fewer studies there are considerable differences. This relative evaluation also holds for $\tau = 0.4$.

When the underlying model is the beta distribution model and $\tau = 0$, the results are equivalent to the truncated normal distribution model. For $\tau = 0.16$ and $K = 5$ the widths of the new CIs decrease with increasing ρ until $\rho = .7$. Interestingly, the widths of these CIs then increase again for $\rho > .7$, which was not observed in the truncated normal model. This effect becomes much less pronounced for increasing number of studies K. HS is always narrower than the new CIs, and, for $K \geq 20$, HOVz is the widest at $\rho = 0$ but even narrower than HS for $\rho \geq .8$. For $\tau = 0.4$ the results are similar, except that the widths of the CIs now decrease monotonously for increasing ρ and HOVz is narrowest for $\rho > .5$.

3.2.3. Recommendations

We summarize our findings by providing recommendations to practitioners wishing to choose between the methods considered. The recommendations will depend on the assumed model and how much heterogeneity is present in the data. We believe the beta distribution model is better suited for random-effects meta-analyses of correlations. Recall that HOVz employs the inverse Fisher transformation, whereas our newly proposed confidence intervals employ the integral z-to-r transformation suggested by

- $\tau = 0$ (*fixed-effect model*). HS and HOVz are not recommended. We recommend using KH, HC3 or HC4.
- $\tau = 0.16$. For the *truncated normal model*, HS and HOVz are not recommended and we recommend using KH, HC3 or HC4. For $|\rho| > .7$, all methods are unsatisfactory and only in the case of $K = 40$ may HOVz be preferable. For the *beta distribution model*, HS and HOVz are not recommended. All new confidence intervals exhibit satisfactory coverage. For small K, WBS approaches yield wider confidence intervals, therefore preferably use KH, HC3 or HC4.
- $\tau = 0.4$. For the *truncated normal model*, HS is not recommended. For $K = 5$ and $|\rho| \leq .7$ we again recommend KH, HC3 or HC4. For $K \geq 10$ and $|\rho| \leq .7$ we recommend HOVz. For $|\rho| > 0.7$ none of the methods is satisfactory. For the *beta distribution model*, HOVz is not recommended. For $|\rho| \leq .7$ we recommend KH, HC3 or HC4. For $K \geq 40$ and $|\rho| > .7$ we recommend using HS. For $K \leq 20$ and $|\rho| > .7$ none of the methods is satisfactory.

4. Illustrative data analyses

Between 25% and 50% of patients fail to take their medication as prescribed by their caregiver (Molloy et al., 2013). Some studies have shown that medication adherence tends to be better in patients who score higher on conscientiousness (from the five-factor model of personality). Table 2 contains data on 16 studies, which investigated the correlation between conscientiousness and medication adherence. These studies were first analysed in the form of a meta-analysis in Molloy et al. (2013). The columns of Table 2 contain information on the authors of the respective study, the year of publication, the sample size of study i (n_i), the observed correlation in study i, the number of variables controlled for (controls), study design, the type of adherence measure (a_measure), the type of conscientiousness measure (c_measure), the mean age of study participants (mean_age) and the methodological quality (as scored by the authors on a scale from 1 to 4, with higher scores indicating higher quality).

Regarding the measurement of conscientiousness, where NEO (*Neuroticism-Extraversion-Openness*) is indicated as c_measure, the personality trait of conscientiousness was measured by one of the various types of NEO personality inventories (PIs; Costa Jr and McCrae, 1985, 2008).

We performed both a fixed- and random-effects meta-analysis, using all methods considered. For the random-effects model we used the SJ estimator to estimate the between-study heterogeneity variance τ^2. Combining all available studies yielded $r_{\text{FE}} = .130$, $r_{\text{RE}} = .154$ and $\hat{\tau}^2_{\text{SJ}} = 0.012$. In addition to a complete-case study, we also examined the cross-sectional and prospective studies separately. In total there were five cross-sectional and 11 prospective studies in the data set. For the cross-sectional studies $r_{\text{FE}} = .168$ and $r_{\text{RE}} = .170$ resulted and slightly lower values for the prospective studies ($r_{\text{FE}} = .108$, $r_{\text{RE}} = .147$). Heterogeneity estimates were $\hat{\tau}^2_{\text{SJ}} = 0.007$ (cross-sectional) and $\hat{\tau}^2_{\text{SJ}} = 0.016$ (prospective), respectively. In Table 3 we provide values of all CIs discussed in this paper.

In the case of all studies ($K = 16$), all methods yield quite similar CIs except for HS. Additional simulations for this situation ($K = 16$, $\tau^2 = 0.012$, n_i as in Table 3) are given in the Appendix S1 and show a coverage of around 80% for HS, while all other methods exhibit a fairly accurate coverage of around 95% and HOVz with around 94%. Thus, the

Table 2. Data from 16 studies investigating the correlation between conscientiousness and medication adherence

Study *i*	Authors	Year	n_i	r_i	Controls	Design	a_measure	c_measure	mean_age	Quality
1	Axelsson *et al.*	2009	109	.19	None	cross-sectional	Self-report	other	22.00	1
2	Axelsson *et al.*	2011	749	.16	None	Cross-sectional	Self-report	NEO	53.59	1
3	Bruce *et al.*	2010	55	.34	None	Prospective	Other	NEO	43.36	2
4	Christensen *et al.*	1999	107	.32	None	Cross-sectional	Self-report	other	41.70	1
5	Christensen and Smith	1995	72	.27	None	Prospective	Other	NEO	46.39	2
6	Cohen *et al.*	2004	65	.00	None	Prospective	Other	NEO	41.20	2
7	Dobbels *et al.*	2005	174	.17	None	Cross-sectional	Self-report	NEO	52.30	1
8	Ediger *et al.*	2007	326	.05	Multiple	Prospective	Self-report	NEO	41.00	3
9	Insel *et al.*	2006	58	.26	None	Prospective	Other	other	77.00	2
10	Jerant *et al.*	2011	771	.01	Multiple	prospective	Other	NEO	78.60	3
11	Moran *et al.*	1997	56	−.09	Multiple	Prospective	Other	NEO	57.20	2
12	O'Cleirigh *et al.*	2007	91	.37	None	Prospective	Self-report	NEO	37.90	2
13	Penedo *et al.*	2003	116	.00	None	cross-Sectional	Self-report	NEO	39.20	1
14	Quine *et al.*	2012	537	.15	None	Prospective	Self-report	other	69.00	2
15	Stilley *et al.*	2004	158	.24	None	Prospective	Other	NEO	46.20	3
16	Wiebe and Christensen	1997	65	.04	None	Prospective	Other	NEO	56.00	1

n_i, study size; r_i, empirical correlation; controls, number of variables controlled for;design, studydesign; a_measure, type of adherence measure (self-report or other); c_measure, type of conscientiousness measure (NEO or other); mean_age, mean age of study participants; quality, methodological quality.

Table 3. Random-effects model confidence intervals for all studies and subgroups separated by study design, original data from Molloy et al. (2013)

	Study design		
Approach	All designs	Cross-sectional	Prospective
HOVz	[.081, .221]	[.067, .266]	[.050, .240]
HS	[.073, .174]	[.100, .220]	[.035, .166]
KH	[.080, .218]	[.037, .291]	[.043, .239]
WBS1	[.086, .213]	[.063, .267]	[.051, .232]
WBS2	[.079, .219]	[.053, .276]	[.043, .239]
WBS3	[.084, .215]	[.058, .272]	[.048, .234]
HC3	[.081, .218]	[.041, .288]	[.041, .241]
HC4	[.083, .216]	[.054, .276]	[.045, .237]

HC, heteroscedasticity-consistent; HOVz, Hedges–Olkin–Vevea Fisher z; HS, Hunter–Schmidt; KH, Knapp–Hartung; WBS, wild bootstrap.

price paid for the narrow HS CIs is poor coverage. Additional analyses of other data sets are given in the Appendix S1.

5. Discussion

We introduced several new methods to construct confidence intervals for the main effect in random-effects meta-analyses of correlations, based on the Fisher z transformation. We compared these to the standard HOVz and Hunter–Schmidt confidence intervals and, following the suggestion by Hafdahl (2009), utilized an integral z-to-r transformation instead of the inverse Fisher transformation. We performed an extensive Monte Carlo simulation study in order to assess the coverage and mean interval length of all CIs. In addition to the truncated normal distribution model considered by Hafdahl and Williams (2009) and Field (2005), we investigated a transformed beta distribution model which exhibits less bias in the generation of the study-level effects.

The results of our simulations show that for low and moderate heterogeneity and correlations of $|\rho| \leq .7$, our newly proposed confidence intervals improved coverage considerably over the classical HOVz and Hunter-Schmidt approaches. However, for extreme heterogeneity and $|\rho| > .7$ all confidence intervals performed poorly. Therefore, further methodological research is necessary in order to fill this gap. Also, the choice of data-generating model (truncated normal or transformed beta distribution) has substantial influence on results. For various reasons, which we discussed when introducing the two models, the beta distribution model is arguably more appropriate. Based on our findings, we provide recommendations to practitioners looking for guidance in choosing a method for data analysis. These are listed in Section 3.2.3.

5.1. Limitations and further research

In the present paper we focused on the Pearson correlation coefficient, as it is the most commonly used dependence measure. However, a limitation of the Pearson correlation coefficient is that it only considers the linear relationship between variables. If variables are related via some nonlinear function or significant outliers are present, other

correlation coefficients such as Spearman's rank correlation may be more appropriate. The Spearman correlation coefficient is the Pearson correlation coefficient of the rank values of the variables considered. Moreover, it shares similar properties with Pearson's correlation such as taking values in $[-1,1]$ and even being asymptotically normal under relatively weak assumptions (Schmid & Schmidt, 2007). The confidence intervals we discussed in this paper can be calculated analogously for Spearman correlation coefficients, for example when dealing with ordinal data. Evaluating their performance, as we did in our simulation study, in conjunction with Spearman correlations is a topic for future research. A detailed analysis of Spearman's and more general correlations as in Schober, Boer, and Schwarte (2018),, however, is outside the scope of this paper.

When dealing with different underlying data than we considered in our paper, it should be kept in mind that although the underlying normal-normal model (4) is often very useful, it has some limitations. For example, when dealing with binomial variables with extreme observations, normal approximations may perform poorly (Agresti & Coull, 1998).. A context where this might occur are ceiling or floor effects on questionnaires or ability tests; that is, when many participants obtain a near maximal (or minimal) score on some questionnaire, a normal approximation may be invalid. Count data may also be problematic, due to their ordinal nature and especially when zeros frequently occur. Therefore researchers should carefully consider the data being analysed when choosing a fitting model in practical applications.

In real-life data sets model (4) may be improved by including meaningful moderator variables, leading to meta-regression as considered in Viechtbauer et al., (2015) and Welz and Pauly (2020).. This can considerably reduce the heterogeneity present in the model.

We attempted to further improve the proposed confidence intervals with the help of a bias correction for the Pearson correlation coefficient r, given by $r^* = r(1-r^2)/(2(n-1))$, as the (negative) bias of r is usually approximated by $\mathscr{B}_r = -\rho(1-\rho^2)/(2(n-1))$ (Hotelling, 1953; Schulze, 2004). However, this bias correction actually made coverage worse in the settings studied.

Acknowledgments

The authors gratefully acknowledge the computing time provided on the Linux HPC cluster at Technical University Dortmund (LiDO3), partially funded in the course of the Large-Scale Equipment Initiative by the German Research Foundation (DFG) as project 271512359. Furthermore, we thank Marléne Baumeister and Lena Schmid for many helpful discussions, Wolfgang Trutschnig and Sebastian Fuchs for valuable comments on Spearman's Rho and Philip Buczak for finding interesting data sets. This work was supported by the German Research Foundation project (Grant no. PA-2409 7-1).

Conflict of interest

All authors declare no conflict of interest.

Author contributions

Thilo Welz, M.Sc. Mathematical Biometry (Conceptualization; Formal analysis; Methodology; Software; Visualization; Writing – original draft) Philipp Doebler (Methodology;

Writing – review & editing) Markus Pauly (Funding acquisition; Methodology; Project administration; Supervision; Writing – review & editing).

Data availability statement

The R-scripts used for our simulations and data analyses will be made publicly available on osf.io under https://osf.io/t83b7/. The dataset from Molloy et al., (2013) can be found in the metafor package in R and the datasets considered for re-analysis are from Chalkidou et al. (2012) and Santos et al. (2016) respectively.

References

Agresti, A., & Coull, B. A. (1998). Approximate is better than "Exact" for interval estimation of binomial proportions. *The American Statistician*, *52*, 119–126. https://doi.org/10.1080/00031305.1998.10480550

Aldao, A., Nolen-Hoeksema, S., & Schweizer, S. (2010). Emotion-regulation strategies across psychopathology: A meta-analytic review. *Clinical Psychology Review*, *30*, 217–237. https://doi.org/10.1016/j.cpr.2009.11.004

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: a meta-analysis. *Personnel Psychology*, *44*, 1–26. https://doi.org/10.1111/j.1744-6570.1991.tb00688.x

Chalkidou, A., Landau, D. B., Odell, W., Cornelius, V. R., O'Doherty, M. J., & Marsden, P. K. (2012). Correlation between Ki-67 immunohistochemistry and 18F-fluorothymidine uptake in patients with cancer: a systematic review and meta-analysis. *European journal of cancer*, *48*(18), 3499–3513. https://doi.org/10.1016/j.ejca.2012.05.001

Cheung, M.W.L. (2015). *Meta-analysis: A structural equation modeling approach*. Hoboken, NJ: John Wiley & Sons.

Costa, Jr, P. T., & McCrae, R. R. (1985). *The NEO personality inventory*. Lutz, FL: Psychological Assessment Resources Odessa. https://www.parinc.com/PAR_Support

Costa, Jr, P. T., & McCrae, R. R. (2008). *The revised NEO personality inventory (NEO-PI-R)*. Thousand Oaks, CA: Sage Publications. https://us.sagepub.com/en-us/nam/contact-us

Cribari-Neto, F., Souza, T. C., & Vasconcellos, K. L. (2007). Inference under heteroskedasticity and leveraged data. *Communication in Statistics – Theory and Methods*, *36*, 1877–1888. https://doi.org/10.1080/03610920601126589

Cribari-Neto, F., & Zarkos, S. G. (2004). Leverage-adjusted heteroskedastic bootstrap methods. *Journal of Statistical Computation and Simulation*, *74*, 215–232. https://doi.org/10.1080/0094965031000115411

Field, A. P. (2001). Meta-analysis of correlation coefficients: A Monte Carlo comparison of fixed-and random-effects methods. *Psychological Methods*, *6*, 161–180. https://doi.org/10.1037/1082-989X.6.2.161

Field, A. P. (2005). Is the meta-analysis of correlation coefficients accurate when population correlations vary? *Psychological Methods*, *10*, 444–467. https://doi.org/10.1037/1082-989X.10.4.444

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, *10*, 507–521. https://doi.org/10.2307/2331838

Hafdahl, A. R. (2009). Improved Fisher z estimators for univariate random-effects meta-analysis of correlations. *British Journal of Mathematical and Statistical Psychology*, *62*, 233–261. https://doi.org/10.1348/000711008X281633

Hafdahl, A. R., & Williams, M. A. (2009). Meta-analysis of correlations revisited: Attempted replication and extension of Field's (2001) simulation studies. *Psychological Methods*, *14*, 24–42. https://doi.org/10.1037/a0014697

Hartung, J. (1999). An alternative method for meta-analysis. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, *41*, 901–916. https://doi.org/10.1002/(SICI)1521-4036(199912)41:8<901::AID-BIMJ901>3.0.CO;2-W

Hartung, J., & Knapp, G. (2001). A refined method for the meta-analysis of controlled clinical trials with binary outcome. *Statistics in Medicine*, *20*, 3875–3889. https://doi.org/10.1002/sim.1009

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.

Hedges, L., & Vevea, J. (1998). Fixed-and random-effects models in meta-analysis. *Psychological Methods*, *3*, 486–504.

Hotelling, H. (1953). New light on the correlation coefficient and its transforms. *Journal of the Royal Statistical Society, Series B*, *15*, 193–232. https://doi.org/10.1111/j.2517-6161.1953.tb00135.x

Hunter, J. E., & Schmidt, F. L. (1994). Estimation of sampling error variance in the meta-analysis of correlations: Use of average correlation in the homogeneous case. *Journal of Applied Psychology*, *79*, 171.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Thousand Oaks, CA: Sage Publishing.

IntHout, J., Ioannidis, J. P., & Borm, G. F. (2014). The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Medical Research Methodology*, *14*, 1–12. https://doi.org/10.1186/1471-2288-14-25

Jak, S. (2015). *Meta-analytic structural equation modelling*. New York: Springer.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1994). *Continuous univariate distributions*. New York: John Wiley & Sons.

Langan, D., Higgins, J. P., Jackson, D., Bowden, J., Veroniki, A. A., Kontopantelis, E., . . . Simmonds, M. (2019). A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. *Research Synthesis Methods*, *10*, 83–98. https://doi.org/10.1002/jrsm.1316

Lehmann, E. L. (2004). *Elements of large-sample theory*. Springer Science and Business media, Luxemburg.

Molloy, G. J., O'Carroll, R. E., & Ferguson, E. (2013). Conscientiousness and medication adherence: A meta-analysis. *Annals of Behavioral Medicine*, *47*, 92–101. https://doi.org/10.1007/s12160-013-9524-4

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

Omelka, M., & Pauly, M. (2012). Testing equality of correlation coefficients in two populations via permutation methods. *Journal of Statistical Planning and Inference*, *142*, 1396–1406. https://doi.org/10.1016/j.jspi.2011.12.018

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, *349*. https://doi.org/10.1126/science.aac4716

Osburn, H., & Callender, J. (1992). A note on the sampling variance of the mean uncorrected correlation in meta-analysis and validity generalization. *Journal of Applied Psychology*, *77*, 115–122. https://doi.org/10.1037/0021-9010.77.2.115

Rosopa, P. J., Schaffer, M. M., & Schroeder, A. N. (2013). Managing heteroscedasticity in general linear models. *Psychological Methods*, *18*, 335–351. https://doi.org/10.1037/a0032553

Santos, S., Almeida, I., Oliveiros, B., & Castelo-Branco, M. (2016). The role of the amygdala in facial trustworthiness processing: A systematic review and meta-analyses of fMRI studies. *PloS one*, *11*(11).e0167276. https://doi.org/10.1371/journal.pone.0167276

Schmid, F., & Schmidt, R. (2007). Multivariate extensions of Spearman's rho and related statistics. *Statistics & Probability Letters*, *77*, 407–416. https://doi.org/10.1016/j.spl.2006.08.007

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: Appropriate use and interpretation. *Anesthesia & Analgesia*, *126*, 1763–1768. https://doi.org/10.1213/ANE.0000000000002864

Schulze, R. (2004). *Meta-analysis: A comparison of approaches*. Boston, MA: Hogrefe Publishing. https://www.hogrefe.com/us/contact

Schwarzer, G., Carpenter, J. R., & Rücker, G. (2015). *Meta-analysis with R*. Cham: Springer.

Sidik, K., & Jonkman, J. N. (2005). Simple heterogeneity variance estimation for meta-analysis. *Applied Statistics*, *54*, 367–384. https://doi.org/10.1111/j.1467-9876.2005.00489.x

Sidik, K., & Jonkman, J. N. (2006). Robust variance estimation for random effects meta-analysis. *Computational Statistics & Data Analysis*, *50*, 3681–3701. https://doi.org/10.1016/j.csda.2005.07.019

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research*, *75*, 417–453. https://doi.org/10.3102/00346543075003417

Süli, E., & Mayers, D. F. (2003). *An introduction to numerical analysis*. Cambridge, UK: Cambridge University Press.

Veroniki, A. A., 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. *Research Synthesis Methods*, *7*, 55–79. https://doi.org/10.1002/jrsm.1164

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*, 1–48. https://doi.org/10.18637/jss.v036.i03

Viechtbauer, W., López-López, J. A., Sánchez-Meca, J., & Marn-Martnez, F. (2015). A comparison of procedures to test for moderators in mixed-effects meta-regression models. *Psychological Methods*, *20*, 360–374. https://doi.org/10.1037/met0000023

Welz, T., & Pauly, M. (2020). A simulation study to compare robust tests for linear mixed-effects meta-regression. *Research Synthesis Methods*, *11*, 331–342. https://doi.org/10.1002/jrsm.1388

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, *48*, 817–838. https://doi.org/10.2307/1912934

Wu, C.F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *Annals of Statistics*, *14*, 1261–1295. https://doi.org/10.1214/aos/1176350142

Received 10 September 2020; revised version received 15 February 2021

Supporting Information

The following supporting information may be found in the online edition of the article:

Appendix S1 Complete results of simulation study.

Article 3

Welz, T., Viechtbauer, W. and Pauly, M. (2022). Cluster-Robust Estimators for Bivariate Mixed-Effects Meta-Regression. *arXiv preprint arXiv:2203.02234*. https://doi.org/10.48550/arXiv.2203.02234.

Cluster-Robust Estimators for Bivariate Mixed-Effects Meta-Regression

Thilo Welz*, Wolfgang Viechtbauer, Markus Pauly

March 7, 2022

Abstract

Meta-analyses frequently include trials that report multiple effect sizes based on a common set of study participants. These effect sizes will generally be correlated. Cluster-robust variance-covariance estimators are a fruitful approach for synthesizing dependent effects. However, when the number of studies is small, state-of-the-art robust estimators can yield inflated Type 1 errors. We present two new cluster-robust estimators, in order to improve small sample performance. For both new estimators the idea is to transform the estimated variances of the residuals using only the diagonal entries of the hat matrix. Our proposals are asymptotically equivalent to previously suggested cluster-robust estimators such as the bias reduced linearization approach. We apply the methods to real world data and compare and contrast their performance in an extensive simulation study. We focus on bivariate meta-regression, although the approaches can be applied more generally.

Keywords: Meta-regression, multivariate analysis, cluster-robust estimators, Monte-Carlo-simulation

*Correspondence: thilo.welz@tu-dortmund.de

1 Introduction

In psychometric and medical research, studies frequently report multiple dependent outcomes. These effects can be synthesized across studies, while incorporating study level moderators, via multivariate meta-regression (Berkey et al., 1998). This is a more sophisticated approach than averaging the effects within studies to create aggregate effects, which are then synthesized. A fruitful approach to achieve reliable inference in the case of a multivariate meta-regression is to use a cluster-robust (CR) variance-covariance estimator (Hedges et al., 2010). Robust estimators are designed to account for potential model misspecification. They can handle dependent effect size estimates and heteroscedastic model errors. A frequent problem in multivariate meta-analysis models is that it is difficult to impossible to compute the variance-covariance matrix of the vector of effect estimates. This is because trials frequently report neither the sampling covariances between study effects nor individual patient data (IPD). This is where CR estimators come into play: They have multiple advantages, such as providing consistent standard errors and asymptotically valid tests without requiring restrictive assumptions regarding the (correlation) structure of the model errors.

Cluster-robust estimators are an extension of heteroscedasticity consistent (HC) estimators. HC estimators, proposed by White (1980) and later extended in Cribari-Neto (2004) and Cribari-Neto et al. (2007), were first proposed in the meta-analytic literature by Sidik and Jonkman (2005). They have been examined and applied for use in ANCOVA (Zimmermann et al., 2019), ordinary least squares regression (Hayes and Cai, 2007) and mixed-effect meta-regression (Hedges et al., 2010; Viechtbauer et al., 2015; Welz and Pauly, 2020). When trials report multiple effects stemming from the same study participants, their clustered, i.e. correlated nature should be accounted for. This is where CR estimators come in. The original formulations of both HC and CR estimators have been shown to possess a downward bias for variance components, as well as yielding highly inflated Type 1 errors of respective test procedures in case of a small number of studies/clusters (Viechtbauer et al., 2015; Tipton and Pustejovsky, 2015; Welz and Pauly, 2020). Therefore it is recommended to instead use one of various improvements that have been suggested. We discuss some of these, such as the bias reduced linearization approach and CR_3 as introduced in Bell and McCaffrey (2002), as well as two new proposals in the chapter on cluster-robust estimators. These can be applied generally for multivariate meta-regression, but we focus specifically on the bivariate case.

First, we present the statistical model, as well as tests and confidence regions for the model coefficients in Section 2. In Section 3, we describe

multiple CR estimators, including two new suggestions. In Section 4, we conduct a real world data analysis. Section 5 describes the design and results of our simulation study. We close with a discussion of the results and an outlook for future research (Section 6).

2 The Set-up

The usual multivariate mixed-effects meta-regression model (Jackson et al., 2011) is given by

$$\boldsymbol{Y_i} = \boldsymbol{X_i}\boldsymbol{\beta} + \boldsymbol{u_i} + \boldsymbol{\varepsilon_i},\ i = 1, \ldots, l, \tag{1}$$

where k is the number of independent studies, $\boldsymbol{\beta} \in \mathbb{R}^q$ is a vector of coefficients and $\boldsymbol{X_i}$ a $p_i \times q$ design matrix of study-level covariates. In the following we will assume that there are p effects of interest per study, but only $p_i \leq p$ effects are observed (reported) in study i, i.e. $\boldsymbol{Y_i} \in \mathbb{R}^{p_i}$. Furthermore, $\boldsymbol{u_i}$ is a random effect that is typically assumed to be multivariate normally distributed with $\boldsymbol{u_i} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{T_i})$ and $\boldsymbol{\varepsilon_i}$ is the *within-study* error with $\boldsymbol{\varepsilon_i} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{V_i})$. With $\boldsymbol{T_i}$ we refer to the $p_i \times p_i$ submatrix of the matrix $\boldsymbol{T} = \begin{pmatrix} \tau_1^2 & \tau_{12} \\ \tau_{12} & \tau_2^2 \end{pmatrix}$, denoting the $p \times p$ *between*-study variance-covariance matrix (under complete data). $\boldsymbol{V_i}$ refers to the corresponding $p_i \times p_i$ *within*-study variance-covariance matrix. A typical example would be a compound symmetry structure for $\boldsymbol{T_i}$, see Section 5 below. We rewrite model (1) in matrix notation as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u} + \boldsymbol{\varepsilon}, \tag{2}$$

with $\boldsymbol{\beta} \in \mathbb{R}^q$, $\boldsymbol{Y} = (\boldsymbol{Y_1'}, \ldots, \boldsymbol{Y_K'})'$, and design matrix $\mathbf{X}$. Assuming that we have a block diagonal matrix of weights $\widehat{\boldsymbol{W}} = \text{diag}(\widehat{\boldsymbol{W}}_1, \ldots, \widehat{\boldsymbol{W}}_K)$, usually corresponding to the inverse variance weights with $\widehat{\boldsymbol{W}}_{\boldsymbol{i}} = \left(\widehat{\boldsymbol{T}}_{\boldsymbol{i}} + \boldsymbol{V_i}\right)^{-1}$, then the weighted least squares estimator for $\boldsymbol{\beta}$ is given by (Hedges et al., 2010)

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X'}\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X'}\widehat{\boldsymbol{W}}\boldsymbol{Y}. \tag{3}$$

We will focus on constructing (multivariate) confidence regions for $\boldsymbol{\beta}$ and confidence intervals for the individual coefficients $\beta_j,\ j = 1, \ldots, q$ based on testing the hypotheses $H_0 : \{\boldsymbol{\beta} = \boldsymbol{\beta}_0\}$ vs. $H_1 : \{\boldsymbol{\beta} \neq \boldsymbol{\beta}_0\}$. We set $\boldsymbol{\Sigma} = \text{Cov}(\hat{\boldsymbol{\beta}})$ and denote estimates thereof by $\widehat{\boldsymbol{\Sigma}}$. We discuss specific choices for estimating $\boldsymbol{\Sigma}$ in Section 3.

Neglecting multiplicity, we note that a commonly used confidence interval for $\beta_j,\ j = 1, \ldots, q$ is given by

$$\hat{\beta}_j \pm \sqrt{\widehat{\boldsymbol{\Sigma}}_{jj}} z_{1-\alpha/2}. \tag{4}$$

Here $z_{1-\alpha/2}$ denotes the $1-\alpha/2$ quantile of the standard normal distribution and $\widehat{\boldsymbol{\Sigma}}_{jj}$ denotes the j^{th} diagonal element of $\widehat{\boldsymbol{\Sigma}}$. A confidence interval with better small sample performance that is asymptotically equivalent for $k \to \infty$ is given by using the $t_{p(k)-q,1-\alpha/2}$ quantile instead, which refers to the $1-\alpha/2$ quantile of the t-distribution with $p(k)-q$ degrees of freedom. Here $p(k) \coloneqq \sum_{i=1}^{k} p_i$ is the total number of observed effects, which is equal to the number of studies k in the univariate setting (Viechtbauer et al., 2015). Alternatively the degrees of freedom of the t distribution can be estimated via a Satterthwaite approximation, as suggested by Bell and McCaffrey (2002).

In order to construct a $(1-\alpha)$ confidence region for $\boldsymbol{\beta}$ we consider the usual Wald-type test-statistic (Tipton and Pustejovsky, 2015)

$$Q = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)' \widehat{\boldsymbol{\Sigma}}^{-1} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0), \tag{5}$$

Alternatively, if one were interested in testing more general hypotheses of the form $H_0 : \{\boldsymbol{H}\boldsymbol{\beta} = \boldsymbol{c}\}$ vs. $H_1 : \{\boldsymbol{H}\boldsymbol{\beta} \neq \boldsymbol{c}\}$ for some hypothesis matrix $\boldsymbol{H} \in \mathbb{R}^{s \times q}$ (which we assume to be of full rank) and vector $\mathbf{c} \in \mathbb{R}^s$, then the test statistic becomes

$$Q_{\boldsymbol{H}} = (\boldsymbol{H}\hat{\boldsymbol{\beta}} - \boldsymbol{c})' (\boldsymbol{H}\widehat{\boldsymbol{\Sigma}}\boldsymbol{H}')^{-1} (\boldsymbol{H}\hat{\boldsymbol{\beta}} - \boldsymbol{c}),$$

For example, the special case of a test regarding a single regression coefficient β_a would be given by $\boldsymbol{H}$ equal to a vector of length q with a 1 at entry a and 0 otherwise.

Under the null hypothesis Q is approximately χ^2_q-distributed (and $Q_{\boldsymbol{H}}$ approximately χ^2_f-distributed with $f = \text{rank}(\boldsymbol{H})$), assuming $\boldsymbol{\Sigma}$ is positive definite. However, it is known that tests based on this approximation can perform poorly for small to moderate values of k (Tipton and Pustejovsky, 2015). An arguably better alternative is the F-test

$$\mathbb{1}\left\{Q > qF_{q,k-q,1-\alpha}\right\}, \tag{6}$$

where $F_{q,k-q,1-\alpha}$ denotes the $1-\alpha$ quantile of an F-distribution with q and $k-q$ degrees of freedom. This is analogous to the t-tests for univariate coefficients and is superior to the test based on the asymptotic χ^2-approximation (Tipton and Pustejovsky, 2015). However, the F-test has been criticized for only performing well in certain scenarios (Tipton, 2015). As a remedy for smaller k, Tipton and Pustejovsky (2015) proposed to approximate Q by a Hotelling's T^2 distribution with parameters q and (degrees of freedom) η, such that

$$\frac{\eta - q + 1}{\eta q} Q \sim F(q, \eta - q + 1). \tag{7}$$

They discuss different approaches for estimating the degrees of freedom η. Based on their research, they recommend an estimation approach, which they call "HTZ". We briefly summarize this estimator, originally proposed by Zhang (2012) for heteroscedastic one-way MANOVA, and refer to their paper for details.

First note that the statistic in (5) can also be written as $Q = \boldsymbol{z}'\boldsymbol{S}^{-1}\boldsymbol{z}$ with $\boldsymbol{z} = \boldsymbol{\Sigma}^{-1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ and $\boldsymbol{S} = \boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1/2}$. Under H_0, $\boldsymbol{z}$ is normally distributed with mean $\boldsymbol{0}$ and covariance $\boldsymbol{I}$ (Tipton and Pustejovsky, 2015). Moreover, if $\boldsymbol{S}$ is a random $q \times q$ matrix such that $\eta\boldsymbol{S}$ follows a Wishart distribution with η degrees of freedom and scale matrix $\boldsymbol{I}_q$, the estimator is given by

$$\hat{\eta}_Z = \frac{q(q+1)}{\sum_{a=1}^{q}\sum_{b=1}^{q}\mathrm{Var}(s_{ab})}.$$

Here s_{ab} denotes the entry (a, b) of $\boldsymbol{S}$. This approach corresponds to setting the total variation in $\boldsymbol{S}$ equal to the total variation in a Wishart distribution (Tipton and Pustejovsky, 2015).

However, our own simulations showed that there are situations when $\hat{\eta}_Z < q - 1$ and therefore $\hat{\eta}_Z - q + 1 < 0$. Specifically this frequently happened in cases with a small number of studies ($k \leq 5$). As the degrees of freedom in an F distribution cannot be negative the HTZ approach is not applicable here. Therefore we will stick to the classical F-test (6), although we propose a small sample adjustment. In our simulations the F-test (6) leads to very liberal or conservative results, depending on the variance-covariance estimator used, in settings with $k = 5$ studies. We therefore propose to truncate the denominator degrees of freedom at the value two, i.e. we consider the F-test

$$\mathbb{1}\left\{Q > qF_{q,\max(2,k-q),1-\alpha}\right\}. \tag{8}$$

The simple motivation behind this adjustment is that for an $F_{m,n}$ distribution with degrees of freedom m and n the expected value $\frac{n}{n-2}$ only exists when $n > 2$. We also tested a truncation of the denominator degrees of freedom at three. However, simulations indicate superior coverage of respective confidence intervals for a truncation at two.

Confidence regions for $\boldsymbol{\beta}$ can be derived via test inversion. For example, if (8) is a test for $H_0 : \{\boldsymbol{\beta} = \boldsymbol{\beta_0}\}$ vs. $H_1 : \{\boldsymbol{\beta} \neq \boldsymbol{\beta_0}\}$, then the set

$$\Lambda := \left\{\boldsymbol{\beta} \in \mathbb{R}^q : (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\widehat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \leq qF_{q,\max(2,k-q),1-\alpha}\right\} \tag{9}$$

is a corresponding confidence region for $\boldsymbol{\beta}$.

A confidence ellipsoid can be obtained following Johnson et al. (2014), based on the eigenvalues $\hat{\lambda}_j$ and eigenvectors $\hat{\boldsymbol{e}}_j$ of $\widehat{\boldsymbol{\Sigma}}$. This means Λ is an ellipsoid centered around $\hat{\boldsymbol{\beta}}$, whose axes are given by

$$\hat{\boldsymbol{\beta}} \pm \sqrt{\hat{\lambda}_j q F_{q,\max(2,k-q),1-\alpha}}\hat{\boldsymbol{e}}_j, \quad j = 1, \ldots, q.$$

This means Λ extends for $\sqrt{\hat{\lambda}_j q F_{q,\max(2,k-q),1-\alpha}}$ units along the estimated eigenvector $\hat{\boldsymbol{e}}_j$ for $j = 1, \ldots, q$. Since the volume of an n-dimensional ellipsoid with axis lengths $a_1, \ldots, a_n$ is given by (Wilson, 2010)

$$V = \frac{2\pi^{n/2}}{n\Gamma(n/2)} \prod_{i=1}^{n} a_i,$$

the volume of the confidence ellipsoid Λ is equal to

$$V_\Lambda = \frac{2\pi^{q/2}}{q\Gamma(q/2)} \prod_{i=1}^{q} \sqrt{\hat{\lambda}_i q F_{q,\max(2,k-q),1-\alpha}}.$$

3 Cluster-Robust Covariance Estimators

Robust variance-covariance estimators, also known as sandwich estimators or Huber-White estimators, have been recommended as a promising alternative in the context of meta-regression (Hedges et al., 2010; Tipton, 2015; Welz and Pauly, 2020). Robust estimators are designed to account for potential model misspecification. They have many desirable properties, such as consistency under heteroscedasticity or asymptotic normality (Hedges et al., 2010) without making restrictive assumptions about the specific form of the effect sizes' sampling distributions.

The reliability of confidence regions based on the statistic (5) depends on the quality of the estimator $\widehat{\boldsymbol{\Sigma}}$ for $\boldsymbol{\Sigma} = \mathrm{Cov}(\hat{\boldsymbol{\beta}})$. The standard (Wald-type) estimator, which we will refer to as ST, is given by $(\boldsymbol{X'}\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}$. The motivation behind this estimator is that the true covariance matrix of $\widehat{\boldsymbol{\beta}}$ (given correct weights) is equal to $\boldsymbol{\Sigma} = (\boldsymbol{X'WX})^{-1}$ with $\boldsymbol{W} = \mathrm{diag}\,(\boldsymbol{W_1}, \ldots, \boldsymbol{W_K})$ and $\boldsymbol{W_i} = \boldsymbol{T_i} + \boldsymbol{V_i}$. However, this ignores the imprecision in the estimation of $\boldsymbol{T}, \boldsymbol{V}$ and therefore in the estimation of $\mathbf{W}$. In fact, if $\boldsymbol{T}$ is estimated poorly, this may lead to deviations from nominal Type 1 error and coverage of corresponding confidence regions (Sidik and Jonkman, 2005).

In the case of univariate meta-analysis and meta-regression heteroscedasticity-consistent (HC) estimators can be applied (Sidik

and Jonkman, 2005; Viechtbauer et al., 2015; Welz and Pauly, 2020). For multivariate meta-regression however, the correlated nature of the study effects needs to be taken into account. We therefore consider cluster-robust (CR) estimators. A selection of CR estimators is, e.g., implemented in the `R` package `clubSandwich` (Pustejovsky, 2021). The package recommendation is the "bias reduced linearization" approach CR_2, which is discussed in detail in Tipton and Pustejovsky (2015); Pustejovsky and Tipton (2018). Sandwich estimators (of HC- as well as CR-type) are all of the general form

$$\widehat{\boldsymbol{\Sigma}} = (\boldsymbol{X'}\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X'}\widehat{\boldsymbol{W}}\widehat{\boldsymbol{\Omega}}\widehat{\boldsymbol{W}}\boldsymbol{X}(\boldsymbol{X'}\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}, \tag{10}$$

with the differences lying in the central "meat" matrix $\widehat{\boldsymbol{\Omega}}$, surrounded by the "bread". This form motivates the name "sandwich" estimator. HC_1[†] is arguably the best known sandwich estimator in the context of univariate meta-regression (Hedges et al., 2010; Viechtbauer et al., 2015; Tipton and Pustejovsky, 2015). However, the extensions HC_3 and HC_4 are frequently recommended as superior alternatives in the non meta-analytic literature, see Cribari-Neto et al. (2007) for details, and have been shown to be superior to HC_1 (Long and Ervin, 2000; Hayes and Cai, 2007; Zimmermann et al., 2019). A natural extension of HC_1 for the multivariate setting and what we will refer to as CR_1^* is defined as

$$\widehat{\boldsymbol{\Sigma}}_{CR_1^*} = \tfrac{k}{k-q}(\boldsymbol{X'}\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}\left(\sum_{\boldsymbol{i=1}}^{\boldsymbol{K}}\boldsymbol{X_i'}\widehat{\boldsymbol{W}}_{\boldsymbol{i}}\widehat{\boldsymbol{\Omega}}_{\boldsymbol{i}}\widehat{\boldsymbol{W}}_{\boldsymbol{i}}\boldsymbol{X_i}\right)(\boldsymbol{X'}\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}, \tag{11}$$

where $\widehat{\boldsymbol{\Omega}}_i = \boldsymbol{E_i}\boldsymbol{E_i'}$ with $\boldsymbol{E_i} = \boldsymbol{Y_i} - \boldsymbol{X_i}\hat{\boldsymbol{\beta}}$ and $\frac{k}{k-q}$ is a correction factor that converges to 1 as k goes to infinity. The motivation for this factor is to correct for a liberal behavior in case of few studies/clusters k; see the `clubSandwich` package for similar choices.

However, as our simulation study below will show, tests based on CR_1^* are still quite liberal when k is small. An alternative is to instead use a bias reduced linearization approach, which was originally proposed by Bell and McCaffrey (2002) and further developed by Pustejovsky and Tipton (2018). This estimator, called CR_2, is designed to be exactly unbiased under the correct specification of a working model. This is achieved via a clever choice of adjustment matrices in the formulation of the estimator, see Tipton and Pustejovsky (2015); Pustejovsky and Tipton (2018) for details. This is the recommended approach in the `clubSandwich` package (Pustejovsky, 2021).

[†] $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{HC_1}} = \tfrac{k}{k-q}(\boldsymbol{X'}\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}\left(\sum_{\boldsymbol{i=1}}^{\boldsymbol{K}}\boldsymbol{X_i'}\widehat{\boldsymbol{W}}_{\boldsymbol{i}}\hat{\boldsymbol{\varepsilon}}_{\boldsymbol{i}}^{\boldsymbol{2}}\widehat{\boldsymbol{W}}_{\boldsymbol{i}}\boldsymbol{X_i}\right)(\boldsymbol{X'}\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}$

Another alternative is the CR_3 estimator, which is a close approximation of the leave-one-(cluster)-out Jackknife variance-covariance estimator. CR_3 is also implemented in the `clubSandwich` package.

However, all of the estimators above can be unsatisfactory for small k, as our simulations will show. Therefore, in addition to these CR-estimators, we propose two others, which are extensions of the HC_3 and HC_4 estimators. Since HC_3 and HC_4 often outperform both HC_1 and HC_2 in the univariate regression setting (Long and Ervin, 2000; Cribari-Neto, 2004; Welz and Pauly, 2020), one would suspect their respective cluster-robust extensions to outperform in the case of multivariate regression. We therefore define CR_3^* and CR_4^* via

$$\widehat{\boldsymbol{\Sigma}}_{CR_3^*} = (\boldsymbol{X'}\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}\left(\sum_{\boldsymbol{i=1}}^{\boldsymbol{K}} \boldsymbol{X_i'}\widehat{\boldsymbol{W}}_{\boldsymbol{i}}\widehat{\boldsymbol{\Omega}}_{\boldsymbol{3i}}\widehat{\boldsymbol{W}}_{\boldsymbol{i}}\boldsymbol{X_i}\right)(\boldsymbol{X'}\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}, \tag{12}$$

$$\widehat{\boldsymbol{\Sigma}}_{CR_4^*} = (\boldsymbol{X'}\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}\left(\sum_{\boldsymbol{i=1}}^{\boldsymbol{K}} \boldsymbol{X_i'}\widehat{\boldsymbol{W}}_{\boldsymbol{i}}\widehat{\boldsymbol{\Omega}}_{\boldsymbol{4i}}\widehat{\boldsymbol{W}}_{\boldsymbol{i}}\boldsymbol{X_i}\right)(\boldsymbol{X'}\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}. \tag{13}$$

Here $\widehat{\boldsymbol{\Omega}}_{3i}$ is defined as

$$\widehat{\boldsymbol{\Omega}}_{3i} = \widehat{\boldsymbol{\Omega}}_i - \Delta + \Delta \cdot (\boldsymbol{I}_{p_i} - \text{diag}(\boldsymbol{H}_i))^{-2}, \tag{14}$$

where $\boldsymbol{H}_i$ refers to the submatrix of $\boldsymbol{H}$ with entries pertaining to study i, p_i is the number of observed effects in study i and $\Delta = \text{diag}(\boldsymbol{E_i}\boldsymbol{E_i'})$. $\boldsymbol{H}$ refers to the hat matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X'}\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X'}\widehat{\boldsymbol{W}}$. Furthermore, $\widehat{\boldsymbol{\Omega}}_{4i}$ is equal to (14) except Δ is multiplied with $(\boldsymbol{I}_{p_i} - \text{diag}(\boldsymbol{H}_i))^{-\delta_i}$, where $\delta_i = \min\left\{4, h_{ii}/\bar{h}\right\}$ with h_{ii} denoting the *i-th* diagonal element of $\boldsymbol{H}$ and $\bar{h}$ is the average of the values in the diagonal of the hat matrix. This data-dependent exponent stems from the HC_4 suggestion by Cribari-Neto (2004). HC_4 performs well in univariate meta-regression (Welz and Pauly, 2020) and therefore motivates an extension to the cluster-robust context.

We highlight that our proposed estimator CR_3^* is different from the estimator CR_3 implemented in the `R` package `clubSandwich` as proposed by Bell and McCaffrey (2002). Whereas the latter uses the entire hat matrix for each cluster, we propose to use just the diagonal elements. In contrast, the "meat" matrix for CR_3 is given by $\sum_{i=1}^{K} \boldsymbol{X_i'}\widehat{\boldsymbol{W}}_{\boldsymbol{i}}(\boldsymbol{I} - \boldsymbol{H_i})^{-1}\widehat{\boldsymbol{\Omega}}_{\boldsymbol{i}}(\boldsymbol{I} - \boldsymbol{H_i})^{-1}\widehat{\boldsymbol{W}}_{\boldsymbol{i}}\boldsymbol{X_i}$. Furthermore note that CR_3^* is not even equal to the estimator with meat matrix given by

$$\sum_{i=1}^{K} \boldsymbol{X_i'}\widehat{\boldsymbol{W}}_{\boldsymbol{i}}(\boldsymbol{I} - \text{diag}(\boldsymbol{H_i}))^{-1}\widehat{\boldsymbol{\Omega}}_{\boldsymbol{i}}(\boldsymbol{I} - \text{diag}(\boldsymbol{H_i}))^{-1}\widehat{\boldsymbol{W}}_{\boldsymbol{i}}\boldsymbol{X_i}$$

because $\widehat{\boldsymbol{\Omega}}_{\boldsymbol{i}}$ is in general not a diagonal matrix (only block-diagonal), due to the clustered nature of the data.

For univariate regression we were able to prove the asymptotic equivalence of all HC estimators, which is formulated in the supplement of Welz and Pauly (2020). Under some some weak regularity conditions it follows that the leverages asymptotically converge to zero, as the number of studies k goes to infinity. Therefore, we expected similar results to hold for CR estimators with analogous arguments. A theorem regarding the asymptotic equivalence of CR estimators under regularity conditions is given in the supplement of this paper, along with a proof.

4 Data Analysis

We exemplify the methods presented in this manuscript with the analysis of a dataset containing 81 trials examining overall (OS) and/or disease-free survival (DFS) in neuroblastoma patients with amplified (extra copies) versus normal MYC-N genes. The data are contained in the `R` package `metafor` and were previously analyzed by Riley et al. (2003, 2007). Amplified MYC-N levels are associated with poorer outcomes. The effect measures are log hazard ratios with positive values indicating an increased risk of death or relapse/death for patients with higher MYC-N levels as compared to patients with lower levels. 17 studies reported both outcomes, 25 studies only reported DFS and 39 studies only reported OS.

The dataset contains the log hazard ratios and the corresponding sampling variances. However, since no information is available on the sampling covariances between OS and DFS we must make some assumptions with regard to our working model. In the spirit of a sensitivity analysis we will first assume a weaker correlation of $\varrho_1 = 0.5$ and subsequently a stronger correlation of $\varrho_2 = 0.8$ and then compare the results. This means for a hypothetical study i that reports log hazard ratios for OS and DFS, $y_{i,OS}$ and $y_{i,DFS}$, with an assumed correlation of 0.5 along with respective sampling variances $\sigma^2_{i,OS}$ and $\sigma^2_{i,DFS}$, we have the sampling variance-covariance matrix $V_i = \begin{pmatrix} \sigma^2_{i,OS} & 0.5 \cdot \sigma_{i,OS}\sigma_{i,DFS} \\ 0.5 \cdot \sigma_{i,OS}\sigma_{i,DFS} & \sigma^2_{i,DFS} \end{pmatrix}$.

We assume a multivariate meta-regression model that includes a random effect as in Section 2 as well as an unstructured (but positive definite) variance-covariance matrix. In the following we are interested in testing whether both pooled effects are different from zero. When the full dataset is analyzed, the Wald-test for $H_0 : \{\boldsymbol{\beta} = \mathbf{0}\}$ vs. $H_1 : \{\boldsymbol{\beta} \neq \mathbf{0}\}$ returns a p-value < 0.001 for all CR estimators and for both ϱ_1 and ϱ_2. However, let us assume we only

had the data from studies 1-5, which all contain results for both OS and DFS. Such a situation is not unrealistic, considering the median number of studies per meta-analyses in a sample of 22,453 published meta-analyses from the Cochrane Database was three (Davey et al., 2011). This reduced dataset is shown in Table 1. The p-values for the estimators CR_1*, CR_3*, CR_4*, CR_2 and ST for assumed correlations $\varrho_1,\ \varrho_2$ are displayed in Table 2.

study	y_i	v_i	outcome
1	-0.11	0.45	DFS
1	-0.14	0.66	OS
2	0.30	0.07	DFS
2	0.67	0.08	OS
3	0.41	0.77	DFS
3	0.43	0.66	OS
4	0.47	0.29	DFS
4	2.08	0.45	OS
5	0.76	0.24	DFS
5	0.70	0.31	OS

Table 1: Sample of five studies containing log hazard ratios (y_i) for disease-free and overall survival and their respective sampling variances (v_i).

The results show that when the number of studies is small the p-values can vary substantially, depending on the choice of estimator. Furthermore, the results based on CR estimators appear to be more stable and depend much less on the underlying $\boldsymbol{V}$ matrix i.e. the assumed correlation between OS and DFS than the standard estimator $\boldsymbol{(X'\widehat{W}X)^{-1}}$. This motivates the use of a CR approach over the standard variance-covariance estimator.

Estimators	p-values	
	ϱ_1	ϱ_2
CR1*	0.073	0.075
CR3*	0.069	0.077
CR4*	0.076	0.090
CR2	0.054	0.055
ST	0.138	0.206

Table 2: p-values of Wald-tests based on CR estimators and the standard variance-covariance estimator $\boldsymbol{(X'\widehat{W}X)^{-1}}$ for assumed correlations of $\varrho_1 = 0.5$ and $\varrho_2 = 0.8$.

5 Simulation Study

Simulation Design In order to assess the performance of the previously discussed methods, we conducted a Monte Carlo simulation. We considered $k \in \{5, 10, 20, 40\}$ studies, average study sizes $N \in \{40, 100\}$ with balanced treatment and control groups, coefficient vectors $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)' \in \{(0, 0, 0, 0)', (0.2, 0.2, 0.1, 0.1)', (0.4, 0.4, 0.2, 0.3)'\}$, correlations $\varrho \in \{0, 0.3, 0.7\}$ and missing data ratios from $\{0, 0.1, 0.2, 0.3, 0.4\}$. The latter refers to the number of studies that only report one of the two effects of interest and ϱ refers to the IPD correlations between the two observed outcomes. In the coefficient vector $\boldsymbol{\beta}$ the first two entries refer to the population means of the two effects of interest and the other two represent the effect of the study-level moderator on each effect respectively. Study sizes were varied, such that for an average study size N, 20% of studies had size $0.8N, 0.9N, \ldots, 1.2N$ respectively. Datasets with missing data were generated by first simulating complete data and then removing entries completely at random.

The simulated study-level effects are (correlated) standardized mean differences (SMD). We estimated these SMDs via the adjusted Hedges' g (Hedges, 1981)

$$g \coloneqq \frac{\Gamma(m/2)}{\sqrt{(m/2)}\Gamma((m-1)/2)} d$$

with $m = n_T + n_C - 2$ and where n_T and n_C refer to the treatment and control group sizes. Hedges' g is defined as $d = (\bar{x}_T - \bar{x}_C)/s^*$, with a pooled standard deviation $s^* = \sqrt{\frac{(n_T-1)s_T^2 + (n_C-1)s_C^2}{m}}$, where s_T^2, s_C^2 refer to the variances in the treatment and control groups respectively (Hedges, 1981). This adjustment to Hedges' g yields an unbiased effect estimator (Lin and Aloe, 2021). We generated the SMDs by first simulating individual participant data (IPD). The treatment and control group IPD observations Y_{ij}^T and Y_{ij}^C were drawn from bivariate normal distributions respectively. More precisely, for study $i = 1, \ldots, k$ and participant $j = 1, \ldots, N_i/2$ the observations are drawn from $Y_{ij}^T \sim \mathcal{N}(\theta_i, P)$ and $Y_{ij}^C \sim \mathcal{N}(\mathbf{0}, P)$ with $\theta_i = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u_i}$ and $P = \begin{pmatrix} 1 & \varrho \\ \varrho & 1 \end{pmatrix}$ is the population correlation matrix of the outcomes in study i. $\boldsymbol{X}$ is a $2 \times q$ design matrix of covariates. In our specific simulation design of a single study-level covariate x with potentially different influence on the two study effects we have $\boldsymbol{X} = \begin{pmatrix} 1 & 0 & x & 0 \\ 0 & 1 & 0 & x \end{pmatrix}$.

For the heterogeneity matrix $\boldsymbol{T}$ we consider the two settings

$$\begin{pmatrix} \tau^2 & 0.2\tau^2 \\ 0.2\tau^2 & \tau^2 \end{pmatrix} \text{ and } \begin{pmatrix} \tau^2 & 0.4\tau^2 \\ 0.4\tau^2 & 2\tau^2 \end{pmatrix}.$$

For $M = N/2$ (average size of the treatment and control groups), we set $\tau^2 := \frac{2}{M} + \frac{\beta_0^2}{4M} = \frac{4}{N} + \frac{\beta_0^2}{2N}$, which is approximately equal to the sampling variance of the standardized mean difference (Borenstein et al., 2021). This corresponds to an I^2 value of 0.5. Here, I^2 refers to the percentage of the total variation across studies that is due to heterogeneity rather than sampling variation (Higgins and Thompson, 2002).

We briefly discuss the covariance between two SMDs in the setting where we have a single treatment and control group but with different outcome measures. The resulting effect sizes will be correlated because the outcomes are collected from the same study participants. Olkin and Gleser (2009) showed that a large sample estimate for the covariance between two SMDs d_1 and d_2 with estimated (raw data) correlation $\hat{\varrho}$ is given by

$$\widehat{\text{Cov}}(d_1, d_2) = \hat{\varrho}\left(\tfrac{1}{n_T} + \tfrac{1}{n_C}\right) + \frac{\hat{\varrho}^2 d_1 d_2}{m}. \tag{15}$$

Thus we obtain

$$\widehat{\text{Cov}}(g_1, g_2) = \left(\frac{\Gamma(m/2)}{\sqrt{(m/2)}\Gamma((m-1)/2)}\right)^2 \left(\hat{\varrho}\left(\frac{1}{n_T} + \frac{1}{n_C}\right) + \frac{\hat{\varrho}^2 d_1 d_2}{m}\right). \tag{16}$$

All results are based on a nominal significance level $\alpha = 0.05$. For each scenario we performed $N = 5000$ simulation runs. The primary focus was on comparing empirical coverage of the confidence regions (9) with nominal coverage being $1 - \alpha = 0.95$. For 5000 iterations, the Monte Carlo standard error of the simulated coverage will be approximately $\sqrt{\frac{0.95 \times 0.05}{5000}} \approx 0.31\%$ and assuming a power of 80% the Monte Carlo standard error of the simulated power will be approximately $\sqrt{\frac{0.8 \times 0.2}{5000}} \approx 0.57\%$ (Morris et al., 2019).

All simulations were performed using the open-source software `R`. The `R` scripts written by the first author especially make use of the `metafor` package for meta-analysis (Viechtbauer, 2010) as well as James Pustejovsky's `clubSandwich` package.

Results

Figures 1–4 display the empirical coverage based on the adjusted F-test (8) and estimators CR_1^*, CR_3^*, CR_4^*, CR_2 and ST. CR_1^* and CR_2 yield much

less than nominal coverage 95% in all settings, but especially for $k < 40$. CR_2 gives around 50% coverage for five studies, between 70-80% for ten, 82-87% for twenty and 88-91% coverage for forty studies. The CR_1^* estimator yields between 25-50% coverage for five studies, 65-75% for ten, 80-86% for twenty and 87-91% for forty studies. It is interesting to observe a clustering of coverage results for the estimator CR_1* and $k = 5$ (depending on the inter-study correlation of effects) that cannot be observed for any other setting or estimator. The standard estimator ST gives approximately correct coverage for $k \geq 20$ but is highly conservative for $k \leq 10$ studies, especially for five. CR_3^* very consistently yields slightly more coverage than CR_4^* in all settings except for $k = 40$ where the difference between the two is negligible. For $k = 5$ coverage based on CR_4^* is approximately nominal and when based on CR_3^* slightly conservative. For $k = 10$ and $k = 20$ CR_4^* gives coverage around 91-92% and CR_3^* around 93-94%. For $k = 40$ both yield coverage around 92-94%.

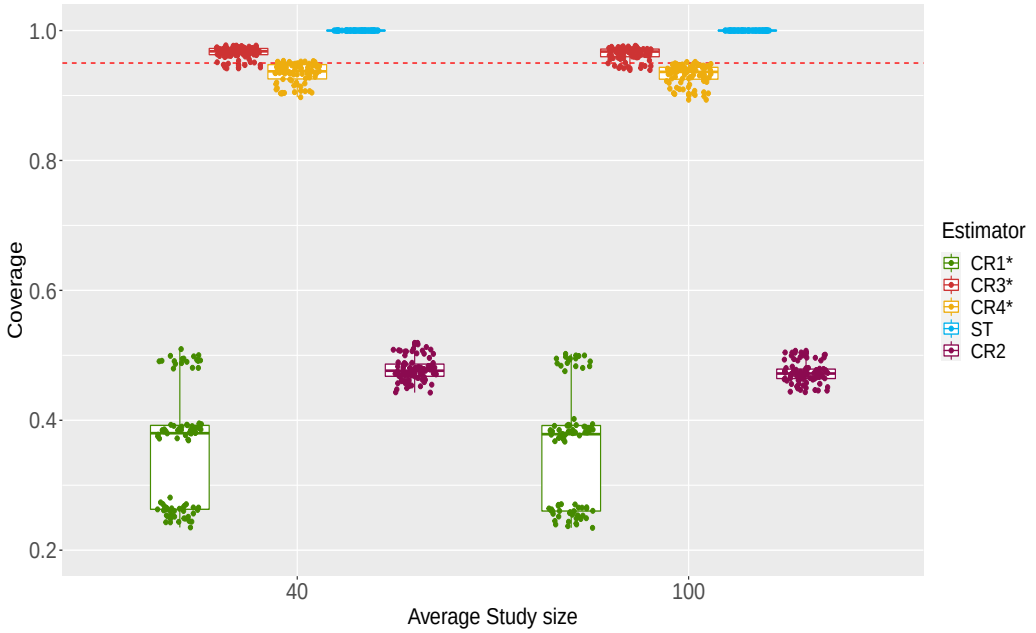

Figure 1: Coverage of the confidence set (9) based on an inversion of the adjusted F-test for $k = 5$ studies.

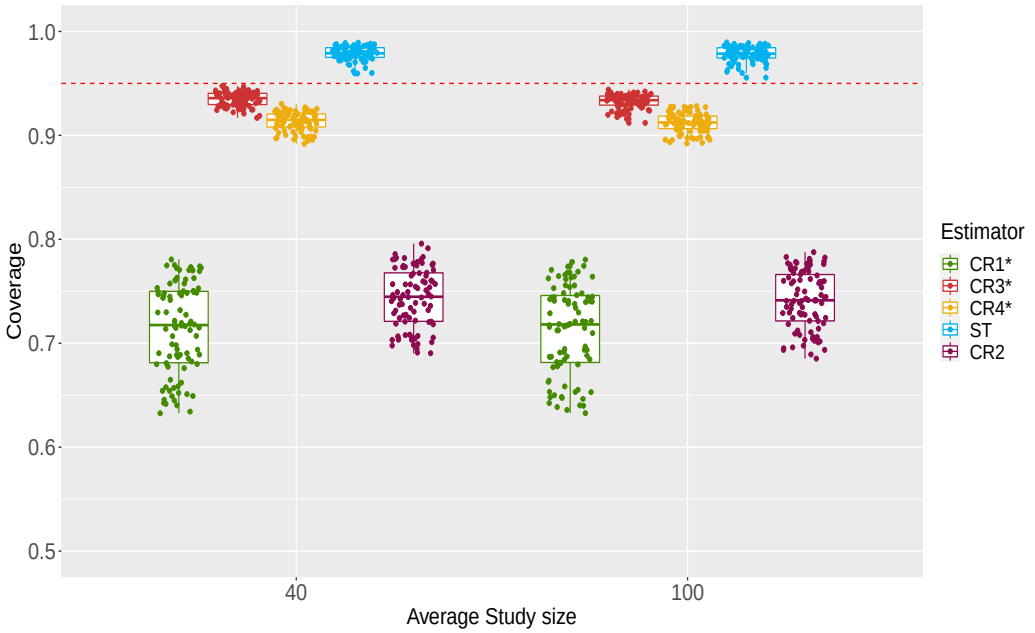

Figure 2: Coverage of the confidence set (9) based on an inversion of the adjusted F-test for $k = 10$ studies.

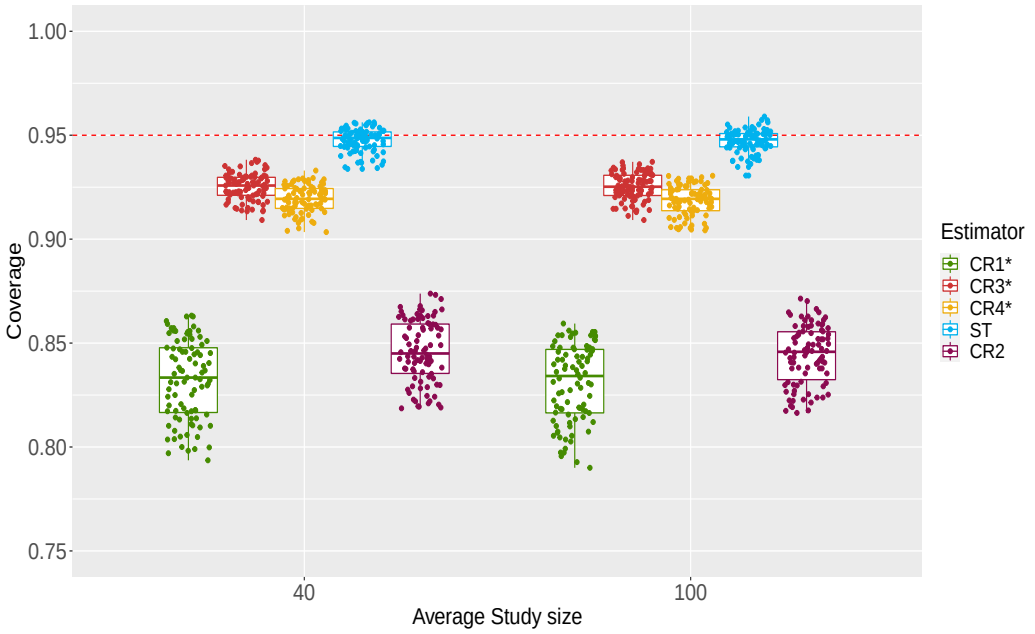

Figure 3: Coverage of the confidence set (9) based on an inversion of the adjusted F-test for $k = 20$ studies.

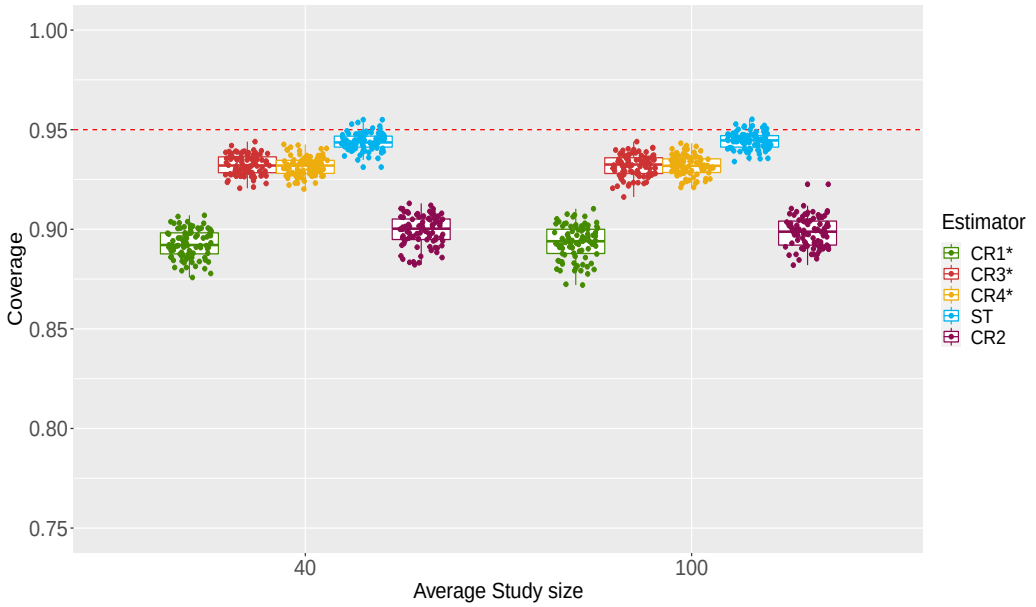

Figure 4: Coverage of the confidence set (9) based on an inversion of the adjusted F-test for $k = 40$ studies.

In addition to these empirical coverage results, we also consider the power related to the respective tests and confidence regions. The power plots are provided in Figures 5 and 6 for $\boldsymbol{\beta} = (0.2, 0.2, 0.1, 0.1)'$ and $\boldsymbol{\beta} = (0.4, 0.4, 0.2, 0.3)'$ respectively. We show box plots to summarize the various simulation settings. For $\boldsymbol{\beta} = (0.4, 0.4, 0.2, 0.3)'$ power is monotone increasing in the number of studies k for all estimators. For $\boldsymbol{\beta} = (0.2, 0.2, 0.1, 0.1)'$ power is monotone increasing in k for CR_3^*, CR_4^* and ST, whereas for CR_1^* and CR_2 power decreases from a median of approximately 70% and 60% to 55% and 52% respectively, when going from five to ten studies and then increases in k beyond this point.

The differences in power between the considered estimators are small for a large number of studies and become more pronounced as the number of studies decreases. For forty studies the power based on all estimators is nearly identical for both choices of $\boldsymbol{\beta}$. For twenty studies power based on CR_1^* and CR_2 is slightly higher than for the other estimators. CR_3^*, CR_4^* and ST yield approximately the same power for both choices of $\boldsymbol{\beta}$ and twenty studies. For $k = 10$ and $\boldsymbol{\beta} = (0.2, 0.2, 0.1, 0.1)'$ the median power for CR_1^* and CR_2 is around 55% and 52% respectively, whereas for CR_3^*, CR_4^* and ST it is around 25%, 31% and 20% respectively. For $k = 10$ and $\boldsymbol{\beta} = (0.4, 0.4, 0.2, 0.3)'$ the median power for CR_1^* and CR_2 is around 87%, whereas for CR_3^*, CR_4^* and ST it is around 70%, 74% and 73% respectively.

For $k = 5$ and $\boldsymbol{\beta} = (0.2, 0.2, 0.1, 0.1)'$ the median power for CR_1^* and CR_2 is around 70% and 60% respectively, whereas for CR_3^*, CR_4^* and ST it is only around 8%, 12% and 0% respectively. For $k = 5$ and $\boldsymbol{\beta} = (0.4, 0.4, 0.2, 0.3)'$ the median power for CR_1^* and CR_2 is around 83% and 70% respectively, whereas for CR_3^*, CR_4^* and ST it is around 13%, 24% and 1% respectively.

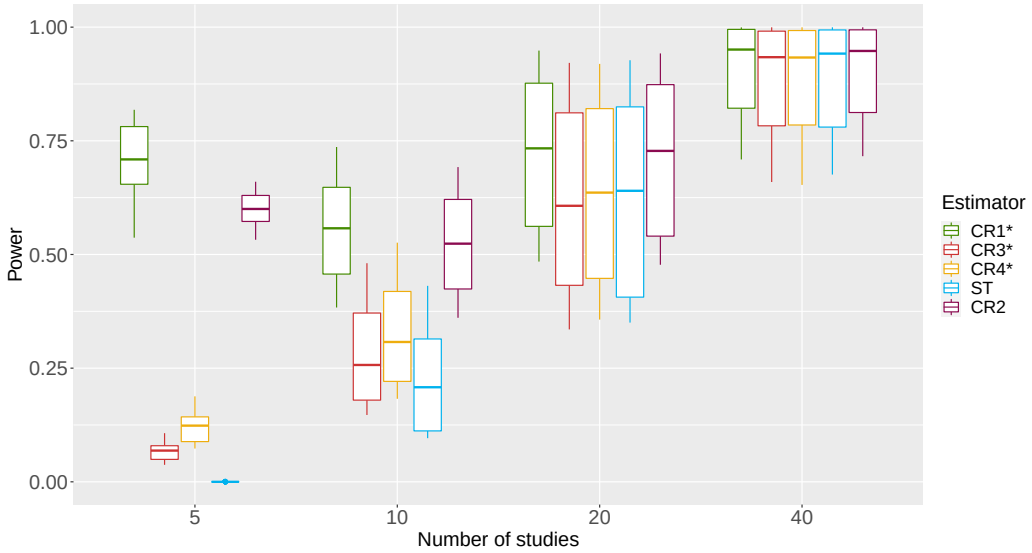

Figure 5: Box plots of power based on adjusted F-test for all settings with $\boldsymbol{\beta} = (0.2, 0.2, 0.1, 0.1)'$.

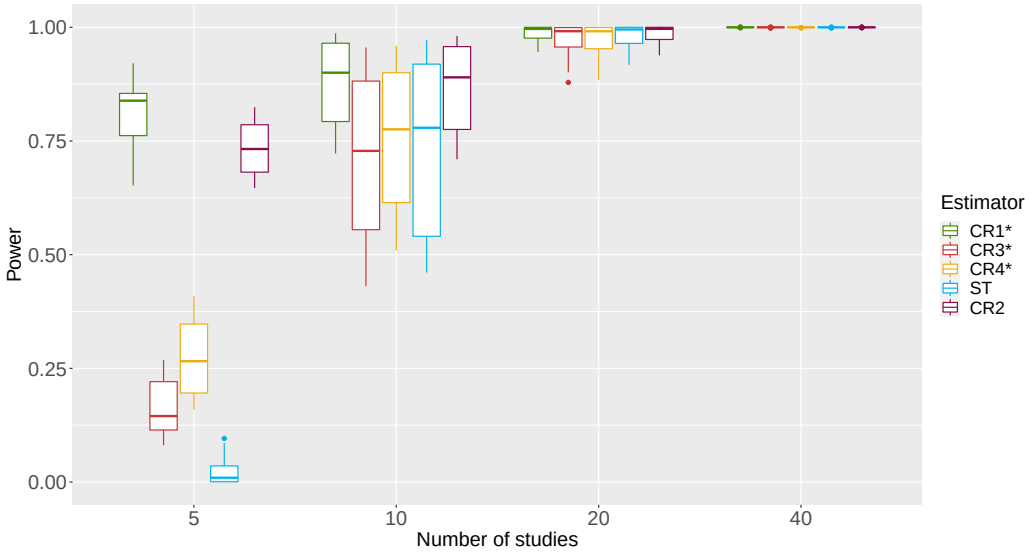

Figure 6: Box plots of power based on adjusted F-test for all settings with $\boldsymbol{\beta} = (0.4, 0.4, 0.2, 0.3)'$.

6 Discussion

Multivariate Meta-Regression is an important tool for synthesizing and interpreting results from trials reporting multiple, correlated effects. However, information on these correlations is rarely available to analysts, making it difficult to construct the variance-covariance $\boldsymbol{V}$ matrix of the studies' sampling errors. Cluster-robust estimators allow for a correction of the standard errors, therefore enabling more reliable inference. In this paper we introduced two new proposals of CR estimators for use in multivariate meta-regression. We performed a simulation study, comparing these estimators with results based on two alternative CR estimators and the standard variance-covariance estimator with a focus on coverage and power of confidence sets and tests, as well as an illustrative real life data analysis. In our manuscript we only investigated the bivariate meta-regression setting, although all methods discussed are also applicable in higher dimensions. Further work is necessary to assess the viability of our suggestions in other settings, such as when the number of effects per study is greater than two.

Our main findings can be summarized as follows: The Zhang estimator, discussed in Tipton and Pustejovsky (2015), can lead to a negative estimate of the denominator degrees of freedom in the F-distribution. This can occur when the number of studies is very small. The AHZ approach is therefore not recommendable for bivariate meta-regression if the number of studies is small ($k \leq 5$). Furthermore, when using the classical F-test in the bivariate setting, we recommend truncating the denominator degrees of freedom at two. The CR_1^* and CR_2 estimators yield an empirical coverage that lies far below the nominal level $1-\alpha$ and the coverage based on the other estimators, especially for smaller numbers of studies. On the flip side the tests based on these two CR-estimators unsurprisingly have superior power. The ST estimator has approximately correct coverage for $k \geq 20$ studies but is highly conservative for $k \leq 10$ studies. CR_3^* and CR_4^* yield approximately correct coverage for five studies. CR_3^* also gives nearly correct coverage for ten studies whereas CR_4^* becomes slightly liberal in this case.

Based on our results we recommend using either the CR_3^* or CR_4^* estimator for bivariate meta-regression if $k \leq 10$ with a very slight preference for CR_3^*. For an analysis with $k \geq 20$ studies the ST estimator seems to work best.

A limitation of our simulation study is that the sampling covariances between study-level effects were available for the construction of weight matrices. As mentioned in the introduction, this is often not feasible in practice, requiring analysts to calculate weights using a specified working model for the covariance structure. Hedges et al. (2010) provide possible working models likely to be found in meta-analyses. They propose the use of

approximately inverse variance weights, based on these working models.

An open question that requires further research is what the best testing procedure is when the number of studies k is no greater than around five. Neither the adjusted Hotelling's T^2 approach in combination with Zhang's estimator for the degrees of freedom, which was recommended by Tipton and Pustejovsky (2015), nor the naive or adjusted F-tests used in our simulations seem to be the ideal approach. This requires more intensive work that is outside the scope of this manuscript. For a discussion of alternative estimation approaches for the degrees of freedom in the adjusted Hotelling approach, we refer to Tipton and Pustejovsky (2015). Another question for future research is whether other statistics or resampling approaches that have shown promising small sample approximations for heterogeneous MAN(C)OVA settings (Friedrich et al., 2017; Friedrich and Pauly, 2018; Zimmermann et al., 2020) can also help in multivariate meta-regression models.

References

Bell, R. M. and McCaffrey, D. F. (2002). Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology*, 28(2):169–182.

Berkey, C., Hoaglin, D., Antczak-Bouckoms, A., Mosteller, F., and Colditz, G. (1998). Meta-analysis of multiple outcomes by regression with random effects. *Statistics in Medicine*, 17(22):2537–2550.

Borenstein, M., Hedges, L. V., Higgins, J. P., and Rothstein, H. R. (2021). *Introduction to meta-analysis*. John Wiley & Sons.

Cribari-Neto, F. (2004). Asymptotic inference under heteroskedasticity of unknown form. *Computational Statistics & Data Analysis*, 45(2):215–233.

Cribari-Neto, F., Souza, T. C., and Vasconcellos, K. L. (2007). Inference under heteroskedasticity and leveraged data. *Communication in Statistics - Theory and Methods*, 36(10):1877–1888.

Davey, J., Turner, R. M., Clarke, M. J., and Higgins, J. P. (2011). Characteristics of meta-analyses and their component studies in the Cochrane database of systematic reviews: a cross-sectional, descriptive analysis. *BMC Medical Research Methodology*, 11(1):1–11.

Friedrich, S., Brunner, E., and Pauly, M. (2017). Permuting longitudinal data in spite of the dependencies. *Journal of Multivariate Analysis*, 153:255–265.

Friedrich, S. and Pauly, M. (2018). MATS: Inference for potentially singular and heteroscedastic MANOVA. *Journal of Multivariate Analysis*, 165:166–179.

Hayes, A. F. and Cai, L. (2007). Using heteroskedasticity-consistent standard error estimators in ols regression: An introduction and software implementation. *Behavior Research Methods*, 39(4):709–722.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, 6(2):107–128.

Hedges, L. V., Tipton, E., and Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, 1(1):39–65.

Higgins, J. P. and Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, 21(11):1539–1558.

Jackson, D., Riley, R., and White, I. R. (2011). Multivariate meta-analysis: potential and promise. *Statistics in Medicine*, 30(20):2481–2498.

Johnson, R. A., Wichern, D. W., et al. (2014). *Applied multivariate statistical analysis*, volume 6. Pearson London, UK:.

Lin, L. and Aloe, A. M. (2021). Evaluation of various estimators for standardized mean difference in meta-analysis. *Statistics in Medicine*, 40(2):403–426.

Long, J. S. and Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, 54(3):217–224.

Morris, T. P., White, I. R., and Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11):2074–2102.

Olkin, I. and Gleser, L. (2009). Stochastically dependent effect sizes. *The Handbook of Research Synthesis and Meta-Analysis*, pages 357–376.

Pustejovsky, J. (2021). *clubSandwich: Cluster-Robust (Sandwich) Variance Estimators with Small-Sample Corrections*. R package version 0.5.3.

Pustejovsky, J. E. and Tipton, E. (2018). Small-sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *Journal of Business & Economic Statistics*, 36(4):672–683.

Riley, R. D., Abrams, K., Lambert, P., Sutton, A., and Thompson, J. (2007). An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes. *Statistics in Medicine*, 26(1):78–97.

Riley, R. D., Burchill, S., Abrams, K. R., Heney, D., Lambert, P. C., Jones, D. R., Sutton, A. J., Young, B., Wailoo, A. J., and Lewis, I. (2003). A systematic review and evaluation of the use of tumor markers in paediatric oncology: Ewing's sarcoma and neuroblastoma. *Health Technology Assessment*.

Sidik, K. and Jonkman, J. N. (2005). A note on variance estimation in random effects meta-regression. *Journal of Biopharmaceutical Statistics*, 15(5):823–838.

Tipton, E. (2015). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods*, 20(3):375.

Tipton, E. and Pustejovsky, J. E. (2015). Small-sample adjustments for tests of moderators and model fit using robust variance estimation in meta-regression. *Journal of Educational and Behavioral Statistics*, 40(6):604–634.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3):1–48.

Viechtbauer, W., López-López, J. A., Sánchez-Meca, J., and Marín-Martínez, F. (2015). A comparison of procedures to test for moderators in mixed-effects meta-regression models. *Psychological Methods*, 20(3):360–374.

Welz, T. and Pauly, M. (2020). A simulation study to compare robust tests for linear mixed-effects meta-regression. *Research Synthesis Methods*, 11(3):331–342.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838.

Wilson, J. (2010). Volume of n-dimensional ellipsoid. *Sciencia Acta Xaveriana*, 1(1):101–6.

Zhang, J.-T. (2012). An approximate Hotelling T2-test for heteroscedastic one-way MANOVA. *Open Journal of Statistics*, 2(1):1–11.

Zimmermann, G., Pauly, M., and Bathke, A. C. (2019). 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. *Statistical Methods in Medical Research*, 28(12):3808–3821.

Zimmermann, G., Pauly, M., and Bathke, A. C. (2020). Multivariate analysis of covariance with potentially singular covariance matrices and non-normal responses. *Journal of Multivariate Analysis*, 177:104594.

Acknowledgements

This work was supported by the German Research Foundation (DFG) (Grant no. PA-2409 7-1). 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 as project 271512359.

We would also like to thank James Pustejowsky for his helpful comments during the research phase for this manuscript.

Data Availability Statement

The neuroblastoma dataset is contained in the `R` package `metafor`. All `R` scripts will be made publicly available, pending publication.

Supplement to "Cluster-Robust Estimators for Bivariate Mixed-Effects Meta-Regression"

Theorem 1. *Suppose there is a* $k_0 \in \mathbb{N}$ *such that* $k(\boldsymbol{X'}\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}$ *exists and is uniformly bounded element-wise for all* $k \geq k_0$*. Furthermore, let* $\widehat{\boldsymbol{T}}$ *be a consistent estimator for* $\boldsymbol{T}$ *and* Λ *the confidence region defined in the main paper. Then the CR estimators* $CR_0, CR_1, CR_2, CR_3, CR_3^*$, CR_4^* *are asymptotically equivalent and we have* $P(\Lambda \ni \boldsymbol{\beta}) \longrightarrow 1-\alpha$ *as* $k \to \infty$*.*

Proof. Let $k \in \mathbb{N}$ be the number of studies, $\boldsymbol{\beta} \in \mathbb{R}^q$ and $\mathbf{X} \in \mathbb{R}^{k \times q}$. Furthermore, $\hat{\boldsymbol{\beta}} = \boldsymbol{(X'\widehat{W}X)^{-1}X'\widehat{W}y}$ and $\mathbf{H} = \boldsymbol{X(X'\widehat{W}X)^{-1}X'\widehat{W}}$. Then

$$\mathbf{H}\boldsymbol{X} = \boldsymbol{X(X'\widehat{W}X)^{-1}X'\widehat{W}X} = \mathbf{X},$$

and since $\mathbf{X}$ is a design matrix with the first column equal to $\mathbf{1}_{kp}$, all row sums in $\mathbf{H}$ are equal to 1. Due to the regularity condition that there exists a $k_0 \in \mathbb{N}$ such that $\forall\ k \geq k_0 :\ k(\boldsymbol{X'}\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}$ exists and is uniformly bounded element-wise, we have that for every $i, j \in \{1, \ldots, kp\} : h_{ij} \overset{a.s.}{\to} 0$ as $k \to \infty$.

So for $i \in \{1, \ldots, k\}$ it holds that $\mathbf{H}_i \to \mathbf{0}$ as $k \to \infty$. Here $\mathbf{H}_i$ refers to the submatrix of $\mathbf{H}$ with entries pertaining to study i. Thus $(\mathbf{I}_{p_i} - \mathbf{H}_i)^\eta \longrightarrow \mathbf{I}_{p_i}$ and also $(\mathbf{I}_{p_i} - \text{diag}(\mathbf{H}_i))^\eta \longrightarrow \mathbf{I}_{p_i}$ as $k \to \infty$ for any $\eta \in \mathbb{R}$. It follows that $\widehat{\boldsymbol{\Sigma}}_a - \widehat{\boldsymbol{\Sigma}}_b \to \mathbf{0}_{q \times q}$ as $k \to \infty$ for any choice of $a, b \in \{CR_0, CR_1^*, CR_3, CR_3^*, CR_4^*\}$, i.e. they are asymptotically equivalent.

Consider the test statistic

$$Q = \boldsymbol{(\hat{\beta} - \beta_0)'\widehat{\Sigma}_{CR}^{-1}(\hat{\beta} - \beta_0)},$$

where $\widehat{\boldsymbol{\Sigma}}_{CR}$ is one of the considered CR variance-covariance estimators. Then for any choice of CR estimator (as they are all consistent) we have $\widehat{\boldsymbol{\Sigma}}_{CR} \to \boldsymbol{\Sigma} = \text{Cov}(\widehat{\boldsymbol{\beta}})$ as $k \to \infty$. It follows with Slutzky's Lemma that $Q \xrightarrow{d} \chi_q^2$ as $k \to \infty$ because with Lemma 2 in White (1980), it holds that $\boldsymbol{\widehat{\Sigma}_{CR}^{-1/2}\widehat{\beta}} \xrightarrow{d} \mathcal{N}(\boldsymbol{\beta}, \boldsymbol{I}_q)$. Furthermore it holds that $qF_{q,k-q,1-\alpha} \xrightarrow{d} \chi^2_{q,1-\alpha}$ as $k \to \infty$ because $F_{q,k-q} \overset{d}{=} \frac{\chi_q^2/q}{\chi^2_{k-q}/(k-q)}$, where χ_q^2, χ^2_{k-q} are independent chi-squared distributed random variables with $q, k-q$ degrees of freedom and $\xi := \chi^2_{k-q}/(k-q) \overset{a.s.}{\to} 1$ as $k \to \infty$ since $\mathbb{E}(\xi) \equiv 1$ and $\text{Var}(\xi) = \frac{2}{k-q} \to 0$ for $k \to \infty$.

Therefore the confidence region Λ from the main paper is an asymptotic $1-\alpha$ confidence region for $\boldsymbol{\beta}$.

□

Technical Appendix

We include a technical appendix, which is an excerpt (pp. 236–242) of the supplement from the first manuscript (Welz and Pauly, 2020). The rest of the supplement, which includes complete simulation results, is omitted for the sake of brevity. The technical appendix includes mathematical details that were not included in the main publication of the first article.

7 Technical Appendix

Proposition 1. *Let $K \geq 2$ be the number of studies. Then the general form of the HC-type estimator in the case of no moderators is*

$$\hat{\boldsymbol{\Sigma}}_\ell = \frac{c_\ell}{(\sum_{i=1}^K w_i)^2} \sum_{j=1}^K w_j^2 \hat{\varepsilon}_j^2 (1 - x_{jj})^{\gamma_\ell},\ \ell = 0, \ldots, 5 \tag{S1}$$

where $w_j = (\sigma_j^2 + \hat{\tau}^2)^{-1}$, $x_{jj} = \frac{w_j}{\sum_{i=1}^K w_i}$ is the j^{th} diagonal element of the hat matrix $\boldsymbol{X}(\boldsymbol{X}'\hat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X}'\hat{\boldsymbol{W}}$ and γ_ℓ takes different values, depending on the type of HC estimator. Moreover, for HC_1, $c_1 = \frac{K}{K-1}$, otherwise $c_\ell = 1$. The residuals are defined as $\hat{\boldsymbol{\varepsilon}} = (\hat{\varepsilon}_j)_{j=1}^K = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$.

Proof. **Proposition 1**
We first note that in the case of no moderators the design matrix takes the simple form $\mathbf{X} = (1, \ldots, 1)' \in \mathbb{R}^K$. The HC estimators from equations (3) - (7) are then given by

$$\hat{\boldsymbol{\Sigma}}_\ell = (\boldsymbol{X}'\hat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X}'\hat{\boldsymbol{W}}\hat{\boldsymbol{\Omega}}_\ell\hat{\boldsymbol{W}}\boldsymbol{X}(\boldsymbol{X}'\hat{\boldsymbol{W}}\boldsymbol{X})^{-1}, \quad \ell = 0, 2, \ldots, 5$$

with $\hat{\boldsymbol{\Sigma}}_1 = \frac{K}{K-1}\hat{\boldsymbol{\Sigma}}_0$, where $\hat{\mathbf{W}} = \text{diag}(\sigma_i^2 + \hat{\tau}^2)^{-1}$ and

$$\hat{\boldsymbol{\Omega}}_\ell = \text{diag}\Big(\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\Big)^2 \cdot \text{diag}\Big((1 - x_{jj})^{\gamma_\ell}\Big).$$

Then

$$\boldsymbol{X}'\hat{\boldsymbol{W}}\boldsymbol{X} = (1, \ldots, 1) \cdot \hat{\boldsymbol{W}} \cdot (1, \ldots, 1)' = \sum_{i=1}^K (\sigma_i^2 + \hat{\tau}^2)^{-1} =: \sum_{i=1}^K w_i,$$

where the w_i are the classical inverse variance weights of the random effects model.

Furthermore the j^{th} diagonal element of the hat matrix is

$$\begin{aligned}
x_{jj} &= [\boldsymbol{X}(\boldsymbol{X'}\hat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X'}\hat{\boldsymbol{W}}]_{jj} \\
&= [(1,\ldots,1)' \cdot \Big(\frac{1}{\sum_{i=1}^{K} w_i}\Big) \cdot (1,\ldots,1) \cdot \text{diag}(w_i)]_{jj} \\
&= [\Big(\frac{1}{\sum_{i=1}^{K} w_i}\Big) \cdot \begin{pmatrix} 1 & \ldots & 1 \\ \vdots & & \vdots \\ 1 & \ldots & 1 \end{pmatrix} \cdot \text{diag}(w_i)]_{jj} \\
&= \Big(\frac{1}{\sum_{i=1}^{K} w_i}\Big) \cdot \begin{pmatrix} w_1 & \ldots & w_K \\ \vdots & & \vdots \\ w_1 & \ldots & w_K \end{pmatrix}_{jj} = \frac{w_j}{\sum_{i=1}^{K} w_i}.
\end{aligned}$$

Thus

$$\begin{aligned}
\hat{\boldsymbol{\Sigma}}_\ell &= \frac{c_\ell}{(\sum_{i=1}^{K} w_i)^2} \boldsymbol{X'}\hat{\boldsymbol{W}} \cdot \hat{\boldsymbol{\Omega}}_\ell \cdot \hat{\boldsymbol{W}}\boldsymbol{X} \\
&= \frac{c_\ell}{(\sum_{i=1}^{K} w_i)^2} (1,\ldots,1) \cdot \text{diag}(w_i) \cdot \hat{\boldsymbol{\Omega}}_\ell \cdot \text{diag}(w_i) \cdot (1,\ldots,1)' \\
&= \frac{c_\ell}{(\sum_{i=1}^{K} w_i)^2} \cdot (w_1,\ldots,w_K) \cdot \hat{\boldsymbol{\Omega}}_\ell \cdot (w_1,\ldots,w_K)' \\
&= \frac{c_\ell}{(\sum_{i=1}^{K} w_i)^2} \sum_{j=1}^{K} w_j^2 \hat{\varepsilon}_j^2 (1 - x_{jj})^{\gamma_\ell},
\end{aligned}$$

with the diagonal elements of the hat matrix given by $x_{jj} = \frac{w_j}{\sum_{i=1}^{K} w_i}$. □

Remark 1. *Based on Proposition 1, the six HC-type variance estimators in the case of no moderators (meta-analysis) are given by*

$$\hat{\boldsymbol{\Sigma}}_0 = \frac{1}{(\sum_{i=1}^{K} w_i)^2} \sum_{j=1}^{K} w_j^2 \hat{\varepsilon}_j^2 (1 - x_{jj})^0 = \frac{1}{(\sum_{i=1}^{K} w_i)^2} \sum_{j=1}^{K} w_j^2 \hat{\varepsilon}_j^2 \tag{S2}$$

$$\hat{\boldsymbol{\Sigma}}_1 = \frac{K/(K-1)}{(\sum_{i=1}^{K} w_i)^2} \sum_{j=1}^{K} w_j^2 \hat{\varepsilon}_j^2 (1 - x_{jj})^0 = \frac{K/(K-1)}{(\sum_{i=1}^{K} w_i)^2} \sum_{j=1}^{K} w_j^2 \hat{\varepsilon}_j^2 \tag{S3}$$

$$\hat{\boldsymbol{\Sigma}}_2 = \frac{1}{(\sum_{i=1}^{K} w_i)^2} \sum_{j=1}^{K} w_j^2 \hat{\varepsilon}_j^2 (1 - x_{jj})^{-1} \tag{S4}$$

$$\hat{\mathbf{\Sigma}}_3 = \frac{1}{(\sum_{i=1}^K w_i)^2} \sum_{j=1}^K w_j^2 \hat{\varepsilon}_j^2 (1 - x_{jj})^{-2} \tag{S5}$$

$$\hat{\mathbf{\Sigma}}_4 = \frac{1}{(\sum_{i=1}^K w_i)^2} \sum_{j=1}^K w_j^2 \hat{\varepsilon}_j^2 (1 - x_{jj})^{-\delta_j},\ \delta_j = \min\{4, \frac{x_{jj}}{\bar{x}}\} \tag{S6}$$

$$\hat{\mathbf{\Sigma}}_5 = \frac{1}{(\sum_{i=1}^K w_i)^2} \sum_{j=1}^K w_j^2 \hat{\varepsilon}_j^2 (1 - x_{jj})^{-\alpha_j},\ \alpha_j = \min\{\frac{x_{jj}}{\bar{x}}, \max\{4, \frac{\eta x_{max}}{\bar{x}}\}\} \tag{S7}$$

with $x_{jj} = \frac{w_j}{\sum_{i=1}^K w_i}$ *and* $\bar{x} = \frac{1}{K} \sum_{i=1}^K x_{ii}$. *Furthermore* $x_{max} = \max\{x_1, \ldots, x_K\}$ *and* η *is a tuning parameter that we set equal to 0.7 based on recommendations in the literature*[3].

Numerical Example (HC estimators in a random-effects meta-analysis setting)

Consider the following hypothetical data set of five studies containing study id's, as well as effect ($\hat{\theta}_i$) and variance (σ_i^2) estimates: We first need to calculate the

study i	effect $\hat{\theta}_i$	variance σ_i^2
1	3.40	0.34
2	2.70	0.13
3	2.50	0.10
4	2.90	0.17
5	4.10	0.43

weights $w_i = (\sigma_i^2 + \hat{\tau}^2)^{-1}$. Using the DerSimonian-Laird estimator for the between study heterogeneity τ^2, we obtain $\hat{\tau}_{DL}^2 = 0.0894492$ and thus get the weights $w_1 = 2.328564, w_2 = 4.556863, w_3 = 5.278460, w_4 = 3.854319$ and $w_5 = 1.925116$ respectively. (Here we could of course have used any of the other estimators for τ^2.) This yields the main effect estimate $\hat{\theta} = \frac{\sum_{i=1}^K w_i \hat{\theta}_i}{\sum_{i=1}^K w_i} = 2.925172$. The squared residuals $\hat{\varepsilon}_j^2$ are thus given by $\hat{\varepsilon}_1^2 = 0.225462, \hat{\varepsilon}_2^2 = 0.050702, \hat{\varepsilon}_3^2 = 0.180771, \hat{\varepsilon}_4^2 = 0.000634$ and $\hat{\varepsilon}_5^2 = 1.380221$ respectively. After calculating the standardized weights x_{jj} we are able to calculate the different HC estimators. Table 7 contains the mentioned estimates.

id j	effect $\hat{\theta}_j$	variance σ_j^2	weight w_j	$\hat{\varepsilon}_j^2$	standardized weight x_{jj}
1	3.40	0.34	2.33	0.23	0.12977
2	2.70	0.13	4.56	0.05	0.25396
3	2.50	0.10	5.28	0.18	0.29417
4	2.90	0.17	3.85	0.00	0.21481
5	4.10	0.43	1.93	1.38	0.10729

We obtain the following values for the various HC-type variance estimators (as well as the Knapp-Hartung variance estimate for comparison):

$$\begin{aligned}
\hat{\mathbf{\Sigma}}_0 &= 0.0386275 \\
\hat{\mathbf{\Sigma}}_1 &= 0.0482844 \\
\hat{\mathbf{\Sigma}}_2 &= 0.0487442 \\
\hat{\mathbf{\Sigma}}_3 &= 0.0622734 \\
\hat{\mathbf{\Sigma}}_4 &= 0.0519365 \\
\hat{\mathbf{\Sigma}}_5 &= 0.0519365 \\
\hat{\mathbf{\Sigma}}_{KH} &= 0.0608829
\end{aligned}$$

We observe that $\hat{\mathbf{\Sigma}}_0$ has the smallest value, followed by $\hat{\mathbf{\Sigma}}_1$ and $\hat{\mathbf{\Sigma}}_2$. This is in alignment with the behavior observed in our simulation study, i.e. the liberal behavior of the estimators HC_0–HC_2. The estimators HC_4 and HC_5 are equal in this example because $\delta_i = \alpha_i \; \forall i \in \{1, \ldots, 5\}$, as in both cases the minima in the equations correspond to $\frac{x_{jj}}{\bar{x}}$.

Theorem 1. *Suppose there exists $K_0 \in \mathbb{N}$ such that $(K^{-1}\boldsymbol{X}'\boldsymbol{X})^{-1}$ exists and is uniformly bounded element-wise for all $K \geq K_0$. Furthermore assume that $M_K := \frac{1}{K}\sum_{i=1}^{K}(X_i' X_i)$ is non-singular for all K sufficiently large, where X_i is the i^{th} row of the matrix $\boldsymbol{X}$. If $\hat{\tau}^2$ is a consistent estimator for τ^2, HC_0–HC_5 are asymptotically equivalent and for any choice within ϕ we have $\mathbb{E}(\phi) \longrightarrow \alpha$ under $H_0 : \{\beta_j = 0\}$ as well as $\mathbb{E}(\phi) \longrightarrow 1$ under $H_1 : \{\beta_j \neq 0\}$ as $K \to \infty$.*

Proof. **Theorem 1**
Let $K \in \mathbb{N}$ be the number of studies, $\boldsymbol{\beta} \in \mathbb{R}^{m+1}$, $\mathbf{X} \in \mathbb{R}^{K \times (m+1)}$ and $\hat{\tau}^2$ be a

consistent estimator for τ^2, i.e. with $\mathbb{P}(|\frac{\hat{\tau}^2}{\tau^2} - 1| > \varepsilon) \xrightarrow[K\to\infty]{p} 0$ for all $\varepsilon > 0$. Then consider the test statistic

$$T_j = \frac{\hat{\beta}_j}{\sqrt{\hat{\boldsymbol{\Sigma}}_{jj}}} = \frac{(\boldsymbol{X'}\hat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X'}\hat{\boldsymbol{W}}\boldsymbol{Y_j}}{\sqrt{(\boldsymbol{X'}\hat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X'}\hat{\boldsymbol{W}}\hat{\boldsymbol{E}}^{\boldsymbol{2}}\hat{\boldsymbol{W}}\boldsymbol{X}(\boldsymbol{X'}\hat{\boldsymbol{W}}\boldsymbol{X})^{-1}_{jj}}}.$$

Obviously $\hat{\boldsymbol{W}} = \text{diag}\Big((\sigma_i^2 + \hat{\tau}^2)^{-1}\Big)$ is symmetric because it is a diagonal matrix. Furthermore it holds that $\hat{\boldsymbol{E}}^2 = \Big(\text{diag}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})\Big)^2 \cdot \text{diag}\Big((1 - h_{jj})^{-\gamma}\Big)$, with $\gamma \in \mathbb{R}$ fixed, converges towards $\Big(\text{diag}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})\Big)^2$ for $K \to \infty$ because $h_{jj} := \mathbf{H}(j,j)$ are the diagonal elements of the hat matrix $\mathbf{H} = \mathbf{X}(\boldsymbol{X'}\hat{\boldsymbol{W}}\boldsymbol{X})^{-1}\mathbf{X'}\hat{\boldsymbol{W}}$, which has trace $m+1$. This can be easily verified by utilizing the commutative property of the trace:

$$\begin{aligned} tr(\mathbf{H}) &= tr(\mathbf{X}(\boldsymbol{X'}\hat{\boldsymbol{W}}\boldsymbol{X})^{-1}\mathbf{X'}\hat{\boldsymbol{W}}) \\ &= tr(\mathbf{X'}\hat{\boldsymbol{W}}\mathbf{X}(\boldsymbol{X'}\hat{\boldsymbol{W}}\boldsymbol{X})^{-1}) \\ &= tr(\mathbf{I}_{(m+1)\times(m+1)}) = m+1. \end{aligned}$$

Due to the regularity condition that there exists a $K_0 \in \mathbb{N}$ such that $\forall K \geq K_0$: $(K^{-1}\mathbf{X'X})^{-1}$ exists and is uniformly bounded element wise, we have that for every $j \in \{1,\dots,K\}$: $h_{jj} \underset{K\to\infty}{\longrightarrow} 0$. The given regularity condition is also enough in the weighted least squares context (consider the weight matrix $\hat{\boldsymbol{W}} = \text{diag}(\varphi_i)$ with $\varphi_i := (\sigma_i^2 + \hat{\tau}^2)^{-1}$ contained in the hat matrix $\mathbf{H}$) because of the bounded variances $0 < \sigma_i^2 + \tau^2 < \infty$:

$$\begin{aligned} (\boldsymbol{X'}\hat{\boldsymbol{W}}\boldsymbol{X}) &= \boldsymbol{X'}\text{diag}(\varphi_i)\boldsymbol{X} \\ &= \begin{pmatrix} x_{1,1} & \dots & x_{K,1} \\ \vdots & \ddots & \vdots \\ x_{1,m+1} & \dots & x_{K,m+1} \end{pmatrix} \cdot \begin{pmatrix} \varphi_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \varphi_K \end{pmatrix} \cdot \begin{pmatrix} x_{1,1} & \dots & x_{1,m+1} \\ \vdots & \ddots & \vdots \\ x_{K,1} & \dots & x_{K,m+1} \end{pmatrix} \\ &= \begin{pmatrix} \varphi_1 \cdot x_{1,1} & \dots & \varphi_K \cdot x_{K,1} \\ \vdots & \ddots & \vdots \\ \varphi_1 \cdot x_{1,m+1} & \dots & \varphi_K \cdot x_{K,m+1} \end{pmatrix} \cdot \begin{pmatrix} x_{1,1} & \dots & x_{1,m+1} \\ \vdots & \ddots & \vdots \\ x_{K,1} & \dots & x_{K,m+1} \end{pmatrix} \\ &= \sum_{i=1}^{K} \varphi_i \begin{pmatrix} x_{i,1}^2 & \dots & x_{i,m+1}x_{i,1} \\ \vdots & \ddots & \vdots \\ x_{i,1}x_{i,m+1} & \dots & x_{i,m+1}^2 \end{pmatrix} \end{aligned}$$

$$\leq \max\Big\{\varphi_i : 1 \leq i \leq K\Big\} \cdot \sum_{i=1}^{K} \begin{pmatrix} x_{i,1}^2 & \dots & x_{i,m+1}x_{i,1} \\ \vdots & \ddots & \vdots \\ x_{i,1}x_{i,m+1} & \dots & x_{i,m+1}^2 \end{pmatrix},$$

with all finite elements. Thus $h_{jj} \underset{K\to\infty}{\longrightarrow} 0$ for every $1 \leq j \leq K$ and therefore $\widehat{\boldsymbol{E}}^2 \underset{K\to\infty}{\longrightarrow} \Big(\text{diag}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})\Big)^2$. The latter is an estimator for $\text{diag}\Big(\text{Var}(\mathbf{y}_i)\Big)$. As we also have convergence in distribution of each component of the standardized least squares estimator to a standard normal distribution under the given assumptions (cf.[4] Lemma 2) it follows that

$$T_j \approx \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \underset{K\to\infty}{\overset{d}{\longrightarrow}} Z,\ Z \sim N(0,1)$$

by Slutzky's Lemma. As $\phi = \mathbb{1}_{\{|T_j| > t_{K-m-1,1-\alpha/2}\}}$ and $t_{K-m-1,1-\alpha/2} \underset{K\to\infty}{\longrightarrow} z_{1-\alpha/2}$, where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution, it follows that ϕ is a consistent, level-α test for testing $H_0 : \{\beta_j = 0\}$ vs. $H_1 : \{\beta_j \neq 0\}$. □

The REML estimator of τ^2 (Details):

The REML estimator $\hat{\tau}^2_{REML}$ maximizes the restricted log-likelihood

$$l_{RE} = -\frac{1}{2}K\ln(2\pi) + \frac{1}{2}\ln(|\mathbf{X'X}|) - \frac{1}{2}\ln(|\tau^2\boldsymbol{I} + \boldsymbol{V}|) - \frac{1}{2}\ln(|\mathbf{X'WX}|) - \frac{1}{2}\mathbf{y'Py}, \tag{S8}$$

where **P** is defined as

$$\mathbf{P} = \boldsymbol{W} - \boldsymbol{W}\boldsymbol{X}(\boldsymbol{X'W}\boldsymbol{X})^{-1}\boldsymbol{X'W}.$$

Here, $\mathbf{V} = diag(\sigma_i^2)$ and $\mathbf{W} = diag(w_i)$ is the diagonal weight matrix with $w_i = (\sigma_i^2 + \hat{\tau}^2_{current})^{-1}$.

The estimate $\hat{\tau}^2_{REML}$ can then be computed iteratively via Fisher's scoring algorithm with step halving.[1,5] The method is implemented in the `R` function `rma` from the `metafor` package, where the default starting value is set as the non-iterative Hedges estimator.[6] In principle though, any of the non-iterative heterogeneity estimators for τ^2 may be chosen as a suitable starting value.

References

[1] 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.

[2] Jackson D, White IR. When should meta-analysis avoid making hidden normality assumptions? *Biom J.* 2018;60.

[3] Cribari-Neto F, Souza TC, Vasconcellos KL. Inference under heteroskedasticity and leveraged data *Commun Stat Theory Methods.* 2007;36:1877–1888.

[4] White H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity *Econometrica.* 1980:817–838.

[5] Jennrich RI, Sampson P. Newton-Raphson and related algorithms for maximum likelihood variance component estimation *Technometrics.* 1976;18:11–17.

[6] Viechtbauer W. Conducting meta-analyses in R with the metafor package *J Stat Softw.* 2010;36:1–48.

Name: Thilo Welz Matrikelnummer: 727081

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Dissertation mit dem Titel

"Robust Covariance Estimation in Mixed-Effects Meta-Regression Models"

selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet sowie die wörtlich oder inhaltlich übernommenen Stellen als solche kenntlich gemacht habe und die Satzung der Technischen Universität Dortmund zur Sicherung guter wissenschaftlicher Praxis in der jeweils gültigen Fassung beachtet habe. Ich versichere außerdem, dass ich die beigefügte Dissertation nur in diesem und keinem anderen Promotionsverfahren eingereicht habe und dass diesem Promotionsverfahren keine endgültig gescheiterten Promotionsverfahren vorausgegangen sind. Ferner erkläre ich, dass keine Aberkennung eines bereits erworbenen Doktorgrades vorliegt. Ich versichere an Eides statt, dass ich nach bestem Wissen die reine Wahrheit erklärt und nichts verschwiegen habe.

Dortmund, den .....

Thilo Welz