the british psychological society
promoting excellence in psychology

British Journal of Mathematical and Statistical Psychology (2022), 75, 1–22

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 $\hat{\tau}_0^2 = \frac{1}{K}\sum_{i=1}^{K}(y_i - \overline{y})^2$ and is then updated via the expression

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

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 $\hat{\rho}_{\text{pool}}$, the pooled sample correlation coefficient, we obtain an approximate CI for ρ as

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

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 $\tanh(\overline{z} \pm u_{1-\alpha/2}/\sqrt{N - 3K})$. 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 $w_i = (1/(n_i - 3) + \hat{\tau}^2)^{-1}$ (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 $w_i = (\sigma_i^2 + \tau^2)^{-1}$, 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 μin a random-effects meta-analysis (REMA):

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

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 $\hat{z}^*_1,\cdots,\hat{z}^*_B$,

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

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

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

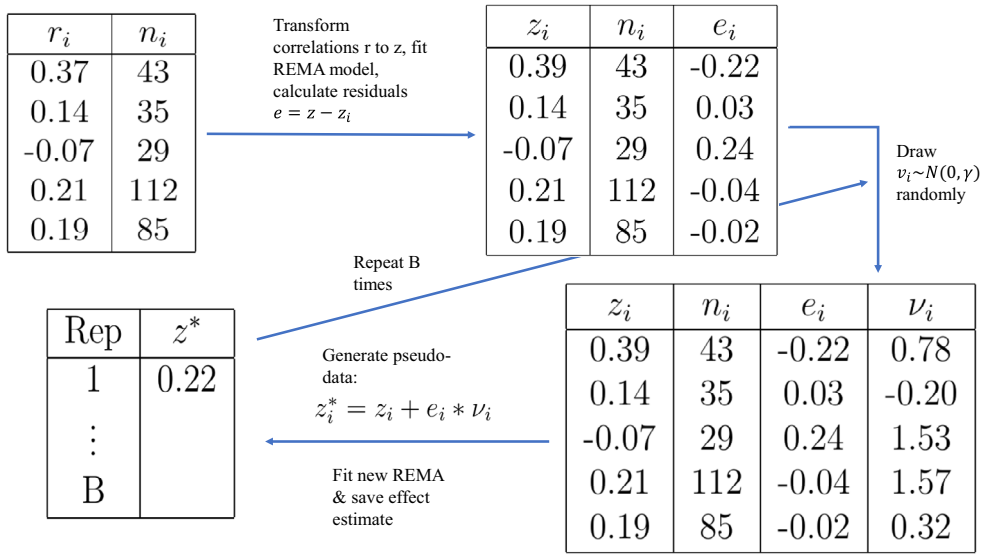

Figure 1. Visual illustration of the wild bootstrap procedure for generating B bootstrap samples of the main effect estimate on the z scale. REMA, random-effects meta-analysis.

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

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

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

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 $\xi : \mathcal{N}(\mathrm{artanh}(\rho), \sigma^2)$ 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.

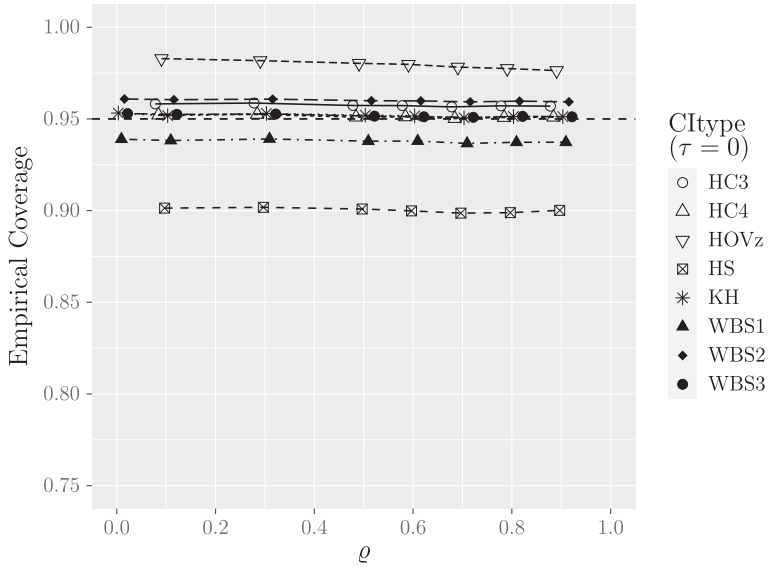

Figure 2. Mean Coverage for truncated normal distribution model with $\tau = 0$, aggregated across all number of studies and study size settings. HC, heteroscedasticity-consistent; HOVz, Hedges–-Olkin–Vevea Fisher z; HS, Hunter–Schmidt; KH, Knapp–Hartung; WBS, wild bootstrap

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%.

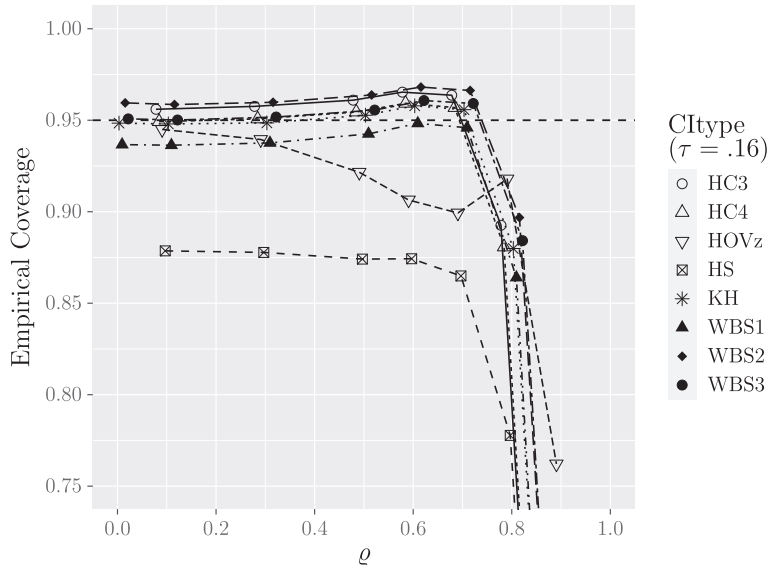

Figure 3. Mean coverage for truncated normal distribution model with $\tau = 0.16$, aggregated across all number of studies and study size settings. HC, heteroscedasticity-consistent; HOVz, Hedges–-Olkin–Vevea Fisher z; HS, Hunter–Schmidt; KH, Knapp–Hartung; WBS, wild bootstrap

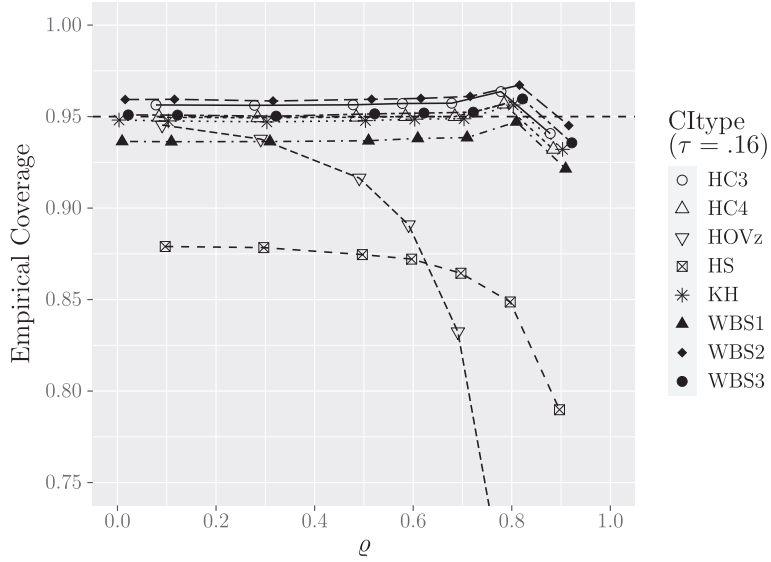

Figure 4. Mean coverage for transformed beta distribution model with $\tau = 0.16$, aggregated across all number of studies and study size settings. HC, heteroscedasticity-consistent; HOVz, Hedges–-Olkin–Vevea Fisher z; HS, Hunter–Schmidt; KH, Knapp–Hartung; WBS, wild bootstrap

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 $r_{FE} = .130$, $r_{RE} = .154$ and $\hat{\tau}^2_{SJ} = 0.012$. 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 $r_{FE} = .168$ and $r_{RE} = .170$ resulted and slightly lower values for the prospective studies ($r_{FE} = .108$, $r_{RE} = .147$). Heterogeneity estimates were $\hat{\tau}^2_{SJ} = 0.007$ (cross-sectional) and $\hat{\tau}^2_{SJ} = 0.016$ (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.