Reevaluation of statistically significant meta-analyses in advanced cancer patients using the Hartung–Knapp method and prediction intervals—A methodological study

Waldemar Siemens¹,²,³ | Joerg J. Meerpohl¹,² | Miriam S. Rohe³
Sabine Buroh⁴ | Guido Schwarzer⁵ | Gerhild Becker³

¹Institute for Evidence in Medicine, Medical Center, University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg, Germany
²Cochrane Germany, Cochrane Germany Foundation, Freiburg, Germany
³Clinic for Palliative Care, Medical Center, University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg, Germany
⁴Library of the Center of Surgery, University Medical Center, Freiburg, Germany
⁵Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Germany

Abstract

Using the Hartung–Knapp method and 95% prediction intervals (PIs) in random-effects meta-analyses is recommended by experts but rarely applied. Therefore, we aimed to reevaluate statistically significant meta-analyses using the Hartung–Knapp method and 95% PIs. In this methodological study, three databases were searched from January 2010 to July 2019. We included systematic reviews reporting a statistically significant meta-analysis of at least four randomized controlled trials in advanced cancer patients using either a fixed-effect or random-effects model. We investigated the impact of switching from fixed-effect to random-effects meta-analysis and of using the recommended Hartung–Knapp method in random-effects meta-analyses. Furthermore, we calculated 95% PIs for all included meta-analyses. We identified 6234 hits, of which 261 statistically significant meta-analyses were included. Our recalculations of these 261 meta-analyses produced statistically significant results in 132 of 138 fixed-effect and 114 of 123 random-effects meta-analyses. When switching to a random-effects model, 19 of 132 fixed-effect meta-analyses (14.4%) were no longer statistically significant. Using the Hartung–Knapp method in random-effects meta-analyses resulted in 34 of 114 nonsignificant meta-analyses (29.8%). In the full sample (N = 261), the null effect was included by the 95% PI in 195 (74.7%) and the opposite effect (e.g., hazard ratio 0.5, opposite effect 2) in 98 meta-analyses (37.5%). Using the Hartung–Knapp method and PIs substantially influenced the null effect inclusion.
interpretation of many published, statistically significant meta-analyses. We strongly encourage researchers to check if using the Hartung–Knapp method and reporting 95% PIs is appropriate in random-effects meta-analyses.

**KEYWORDS**
cancer, heterogeneity, meta-analysis, prediction interval, random-effects model, systematic review

**Highlights**

**What is already known**
- Assessing heterogeneity in meta-analyses by Cochran’s Q and I\(^2\) is associated with various limitations.
- The random-effects model is often more adequate than the fixed-effect model in medicine.
- The classic inverse variance method for calculating a random-effects model has been criticized.
- The reporting of 95% prediction intervals (95% PIs) has been recommended.

**What is new**
- Switching from fixed-effect to random-effects model resulted in a loss of statistical significance in 14.4% of 132 meta-analyses.
- Switching to the Hartung–Knapp method in random-effects meta-analyses resulted in a loss of statistical significance in 29.8% of 114 meta-analyses.
- The *null effect* was included by the 95% PI in 195/261 (74.7%) statistically significant meta-analyses in advanced cancer patients.
- The *opposite effect* was covered by the 95% PI in 98/261 (37.5%) meta-analyses statistically significant meta-analyses in advanced cancer patients.

**Potential impact for Research Synthesis Methods readers outside the authors’ field**
- The results underline the impact of the Hartung–Knapp method on statistical significance in random-effects meta-analyses.
- Reporting 95% PIs to enhance the understanding of heterogeneity in random-effects meta-analyses with ≥4 studies and may have an impact on clinical decision-making.

**1 | BACKGROUND**

Statistical methods employed in applications often lag behind recommended approaches in the statistical literature due to reasons like unawareness of the methods by researchers or unavailability of user-friendly software. In meta-analyses, the use of a random-effects method is nowadays typically preferred over fixed-effect approaches.\(^2\)–\(^4\) Within random-effects models, the Hartung–Knapp method has been shown to yield improved coverage probabilities in several simulations studies and is recommended by experts.\(^2\)–\(^4\) Finally, prediction intervals (PIs) have been introduced as a new approach to facilitate a clinically meaningful interpretation of heterogeneity in meta-analyses.\(^5\),\(^6\)

In meta-analyses, the observed effect of an intervention often varies across studies due to a combination of clinical heterogeneity (e.g., different patients or interventions), methodological heterogeneity (e.g., different risk of bias or study designs), and random variation.\(^1\),\(^7\) This variation in intervention effects may result in statistical heterogeneity, or simply heterogeneity, if the observed effects vary more than one would expect by random variation alone. The choice of the model (i.e., fixed-effect or random-effects) depends on the assumption of heterogeneity of a meta-analysis and should be made a priori.\(^7\) In medicine, the random-effects model is typically more adequate than the fixed-effect model, but the latter is still applied frequently usually leading to narrower 95%
confidence intervals (95% CI) than the random-effects model and thus to more statistically significant results.\(^5\)

The classic inverse variance method for calculating a random-effects model goes back to DerSimonian and Laird (1986). It assumes a standard normal distribution (=z-distribution) to calculate the random-effects confidence interval and has been criticized for inflating the Type 1 error rate.\(^2,8\) In contrast, the calculation of the random-effects confidence interval with the Hartung–Knapp method is based on a t-distribution typically resulting in wider confidence intervals and showed less inflated Type 1 error rates in simulation studies.\(^2–4\)

Published meta-analyses, especially with statistically significant results, play an important role in treatment guidelines, health policy, or best patient care. As described before, the use of the nowadays recommended Hartung–Knapp method is expected to produce wider confidence intervals and thus fewer statistically significant results compared to classic fixed-effect or random-effects meta-analysis.\(^5,10\) Accordingly, we were interested in the impact of using up-to-date statistical methods in meta-analyses in our medical field. First, we investigated whether switching to the classic random-effects method changed the statistical significance of the pooled effect estimate in statistically significant fixed-effect meta-analyses in advanced cancer patients. Second, we assessed how switching to the Hartung–Knapp method impacted the statistical significance of the pooled effect estimate in statistically significant random-effects meta-analyses of advanced cancer patients.

Heterogeneity in meta-analyses can be quantified statistically, for example, by a formal test for heterogeneity like Cochran’s Q and/or by the \(I^2\) statistic representing inconsistency across studies, but both are associated with various limitations.\(^1,11–13\) In order to facilitate a clinically meaningful interpretation of heterogeneity in meta-analyses, the reporting of 95% PIs in random-effects meta-analyses has been recommended by Cochrane and other meta-analysis experts.\(^1,5,6,14\) When using a random-effects model, the 95% PI indicates the 95% probability range in which the true effect of a similar, future study can be expected. A further interpretation of the 95% PI is to understand it as a summary of the distribution of the underlying study effects in a random-effects meta-analysis.\(^1,2,6\)

It is of particular relevance for the interpretation of statistically significant meta-analyses whether the null effect (e.g., hazard ratio [HR] = 1) or even the opposite effect is included in the 95% PI.\(^6\) For example, a meta-analysis with a pooled HR of 0.5 for overall survival and a 95% PI overlapping \(HR = 2\) indicates that patients of a future study could on average experience a doubling instead of a halving of the mortality risk. The 95% PI refers to the same scale as the outcome of the meta-analysis, which facilitates the clinical interpretation. Moreover, it provides information for the assessment of clinically relevant benefits or harms.\(^1,2,6\) Despite the valuable additional information on heterogeneity 95% of PIs are commonly not reported in meta-analyses.

The third and primary objective of our work was to calculate 95% PIs of statistically significant meta-analyses and assess whether they contain the null effect or even the opposite effect.

2  |  METHODS

2.1  |  Study design

We conducted a methodological study based on a systematic search and identification of systematic reviews to address our research questions. Although this study is not a typical systematic review, we followed the preferred reporting items for systematic reviews and meta-analyses (PRISMA) guideline as far as applicable.\(^15\) We preregistered the study in PROSPERO, an international prospective register of systematic reviews (PROSPERO-ID: CRD42019134904).

2.2  |  Eligibility criteria

Systematic reviews with one selected meta-analysis per review were the unit of analysis in this study. The pooled estimate of the selected meta-analysis had to be statistically significant based on at least four randomized controlled trials (RCTs) assuming that these meta-analyses potentially have a large clinical impact. If a systematic review contained more than one meta-analysis, the statistically significant meta-analysis that was reported first in the abstract or full text was selected.

Pharmacological, surgical, and radiotherapeutic interventions in patients with an advanced, incurable tumor with metastases or a life expectancy below 2 years were eligible.\(^16\) No restrictions were made regarding the control group intervention. We considered meta-analyses of continuous, dichotomous, and time-to-event outcomes. Applying these criteria, we were able to draw a manageable sample of systematic reviews in terms of screening and data extraction in a clinically relevant patient group.

Network meta-analyses, reviews on prognostic factors or prognostic models, validation or diagnosis, scoping reviews, outdated Cochrane reviews for which an update existed, and meta-analyses evaluating only one intervention (e.g., analysis of mutations or biomarkers) were excluded.
2.3 | Identification of systematic reviews

The databases Medline (via Ovid), the Cochrane Database of Systematic Reviews (via Wiley), and Web of Science (Science Citation Index Expanded) were searched from January 2010 to July 2019 (approximately 10-year period after PRISMA publication and after a landmark publication on PI’s by Higgins et al.).\textsuperscript{5,15} The Medline search strategy was developed first and then adapted for the other databases (see Data S1—Appendix 1). The search strategy was revised iteratively in cooperation with an experienced librarian (S.B.) and the precision and validity were enhanced by using a search filter for reviews\textsuperscript{17} as well as a test set that had to be identified by the Medline search.

2.4 | Study selection and data extraction

Two reviewers (W.S., M.R.) started with a pilot screening of about 50 references. Later in the screening process, results were reconciled between both reviewers first after screening 500 and then after every 1000 references. The obtained full texts were also assessed by both reviewers. Disagreements were resolved by a discussion between the two reviewers. A third reviewer (G.B.) was consulted if necessary. To quantify the agreement between reviewers, percent agreement and Cohen’s Kappa were calculated. Values of 0.40–0.59 reflect a fair agreement, 0.60–0.74 a good, and 0.75 or more an excellent agreement.\textsuperscript{1}

Variables that were relevant for data extraction were divided into two sets. Each set was extracted separately by one reviewer (W.S. or L.J.). In order to standardize the data extraction, the first three studies were extracted by both reviewers. Additionally, 5% of all included systematic reviews were randomly selected and extracted by both reviewers to enable a consistent extraction process. The necessary random sequence was generated with the statistical software R.\textsuperscript{18} A list of the extracted variables is part of the Data S1—Appendix 2.

2.5 | Classic (inverse variance) meta-analysis

Mathematically, a meta-analysis estimate $\hat{\mu}$ is a weighted average of study-specific effects $y_i$: $\hat{\mu} = \sum w_i y_i / \sum w_i$ with study weights $w_i$.\textsuperscript{19} An approximate 95% CI is given by

$$\hat{\mu} \pm z_{0.975} \text{SE}(\hat{\mu}),$$

with the 97.5% quantile of the standard normal distribution $z_{0.975}$ and standard error $\text{SE}(\hat{\mu}) = \sqrt{1/\sum w_i}$. A corresponding test for an overall effect can be constructed using the test statistic $\hat{\mu}/\text{SE}(\hat{\mu})$.

Meta-analysis methods differ in the weights $w_i$ used for individual studies. The classic fixed-effect meta-analysis method uses the inverse of the study-specific variances $\text{Var}(y_i)$ as weights: $w_i = 1/\text{Var}(y_i)$. Contrary, the classic random-effects method considers the between-study variance $\tau^2$ as a second variance component: $w_i = 1/(\text{Var}(y_i) + \tau^2)$.

Obviously, the between-study variance $\tau^2$ must be estimated from the studies. In our reanalyses, we will use the recently recommended Paule–Mandel method instead of the DerSimonian–Laird method to estimate $\tau^2$.\textsuperscript{1,20}

Using weights $w_i$ or $w_i^*$, we calculate meta-analysis estimates $\hat{\mu}_F$ and $\hat{\mu}_R$ and corresponding standard errors $\text{SE}(\hat{\mu}_F)$ and $\text{SE}(\hat{\mu}_R)$, respectively.

2.6 | Hartung–Knapp method

The Hartung–Knapp method is an alternative random-effects approach using the same random-effects estimate $\hat{\mu}_R$, however, calculating a different 95% CI\textsuperscript{21}:

$$\hat{\mu}_R \pm t_{k-1}^{0.975} \text{SE}_\text{HK}(\hat{\mu}_R)$$

The 95% CI differs in two aspects from the classic meta-analysis CI. First, in a meta-analysis with $k$ studies, the 97.5% quantile of a $t$-distribution with $k-1$ degrees of freedom is used instead of the quantile from the standard normal distribution. Second, a different standard error $\text{SE}_\text{HK}(\hat{\mu}_R)$ is utilized.

The Hartung–Knapp method typically results in wider confidence intervals than the classic random-effects method. However, in rare, very homogeneous cases, the Hartung–Knapp CI can be smaller.\textsuperscript{9} Adjustment methods for such cases, which are not considered in our reanalyses, have been suggested.\textsuperscript{8}

2.7 | Prediction intervals

CIs describe highly probable values for summary effects in a meta-analysis. In contrast, the PI is a measure of heterogeneity in random-effects meta-analyses. PIs are based on a $t$-distribution and provide information on the distribution of the underlying study effects by taking into account the between-study variance $\tau^2$ in the random-effects model. The 95% PI indicates where the true study effects are to be expected for 95% of similar (exchangeable) future studies.\textsuperscript{1,5,6}
The formula for the 95% PI used in our study is implemented in the R package *meta* and goes back to Higgins et al.\(^5\)

\[
\hat{\mu}_R \pm t_{k-2, 0.975} \sqrt{\frac{\hat{\tau}^2 + SE(\hat{\mu}_R)^2}{k}}.
\]

The pooled estimate of the random-effects model is \(\hat{\mu}_R\), \(t_{k-2, 0.975}\) is the 97.5% quantile of the \(t\)-distribution with \(k-2\) degrees of freedom and \(k\) studies, \(\hat{\tau}^2\) is the estimate for the between-study variance, and \(SE(\hat{\mu}_R)^2\) is the estimate of the variance of the pooled estimate of the random-effects meta-analysis.\(^1,5\) Again, we used the Paule–Mandel method to estimate the between-study variance \(\hat{\tau}^2.\(^1,20\)

In this study, we examined the relevance of heterogeneity of all included meta-analyses by assessing if the 95% PI included the null effect or the opposite effect as suggested by IntHout et al.\(^6\) The null effect was considered to be included if the 95% PI crossed a value of 1 for dichotomous outcomes (e.g., HR = 1) or zero for continuous outcomes. The opposite effect was considered to be included if the 95% PI crossed the reciprocal of the pooled dichotomous effect estimate (e.g., the effect of HR = 0.5; HR = 2 included in 95% PI) or the positive/negative counterpart of a pooled continuous effect estimate (e.g., mean difference 1 and \(-1\) included in 95% PI) using the random-effects model.

### 2.8 Sample size and statistical analysis

We expected to include a total of 250–300 reviews based on preliminary searches. This number of reviews is comparable to or even exceeds the sample size of other methodological studies\(^23–28\) and was considered sufficient for addressing the research questions.
To answer the three research questions, it was necessary to recalculate all included statistically significant meta-analyses, which required data extraction for each meta-analysis (see R code and data for the results of this article at https://osf.io/xr6c5/). We structured the analysis as follows:

1. We recalculated all included statistically significant fixed-effect meta-analyses using the inverse variance method for pooling.
2. After recalculating the fixed-effect meta-analyses sample (see 1.), we analyzed the impact of switching from the fixed-effect model to the random-effects model (research question 1). The Paule–Mandel method was applied for the random-effects model for estimating the between-study variance \( \tau^2 \). The Paule–Mandel estimator as well as the restricted maximum likelihood estimator are recommended over the DerSimonian–Laird estimator.\(^1\,^{20}\)
3. We recalculated all included statistically significant random-effects meta-analyses using the inverse variance method for pooling and the Paule–Mandel estimator for the between-study variance \( \tau^2 \).
4. After recalculating the random-effects meta-analyses sample (see 3.), we analyzed the impact of switching from the classic z-distribution-based meta-analytic model to the Hartung–Knapp method (research question 2). The impact of the Hartung–Knapp method on statistical significance was of great interest because it is based on a t-distribution and showed less inflated Type 1 error rates in simulation studies.\(^2\,^{4}\) The Hartung–Knapp method was not used in the original calculations of the included random-effects meta-analyses.
5. Finally, we calculated 95% PIs for all included meta-analyses according to Higgins et al.\(^5\) and we examined if the 95% PI included the null effect (e.g., HR = 1) or the opposite effect (e.g., the effect of HR = 0.5; HR = 2 included in 95% PI) (research question 3).\(^6\)

Descriptive statistics with absolute and relative frequencies for categorical outcomes and means and standard deviations for continuous outcomes were used to address the research questions. We used the statistical program R\(^18\) (version 3.6.3) and package meta for all calculations (version 4.11–0, function metagen).\(^22\,^{29}\)

### RESULTS

In total, 6234 references were identified by our search strategy (last update: July 2019). The elimination of duplicates resulted in 5608 unique references of which titles and abstracts were screened against our eligibility criteria.

### Table 1 Characteristics of the included systematic reviews

| Characteristics           | Sample: N = 261 |
|---------------------------|-----------------|
| Tumor localization:       |                 |
| Lung cancer               | 68 (26.1%)      |
| Colorectal cancer         | 45 (17.2%)      |
| Breast cancer             | 31 (11.9%)      |
| Gastric cancer            | 30 (11.5%)      |
| Pancreatic cancer         | 19 (7.3%)       |
| Mixed types of cancer     | 17 (6.5%)       |
| Urogenital cancer         | 11 (4.2%)       |
| Brain metastases          | 10 (3.8%)       |
| Esophageal cancer         | 9 (3.4%)        |
| Hepatocellular carcinoma  | 6 (2.3%)        |
| Bone metastases           | 3 (1.1%)        |
| Other                     | 12 (4.6%)       |
| Type of intervention:     |                 |
| Pharmacological           | 244 (93.5%)     |
| Radiotherapeutic          | 10 (3.8%)       |
| Surgical                  | 7 (2.7%)        |
| Control group:            |                 |
| Standard care             | 156 (59.8%)     |
| Standard care and placebo | 41 (15.7%)      |
| Comparative intervention  | 57 (21.8%)      |
| Same intervention: variation of characteristics (e.g., dose) | 7 (2.7%) |
| Outcome of meta-analyses: |                 |
| Overall survival          | 78 (29.9%)      |
| Tumor progression         | 71 (27.2%)      |
| Complete response         | 62 (23.8%)      |
| Adverse event             | 35 (13.4%)      |
| Disease event             | 4 (1.5%)        |
| Symptoms                  | 3 (1.1%)        |
| Other                     | 8 (3.1%)        |
| Type of outcome in meta-analyses\(^4\): |               |
| Objective                 | 227 (87.0%)     |
| Objective decision-dependent | 10 (3.8%)   |
| Subjective                | 24 (9.2%)       |
| Effect measures:          |                 |
| HR                        | 128 (49.0%)     |
| RR                        | 74 (28.4%)      |
| OR                        | 57 (21.8%)      |
| SMD                       | 2 (0.8%)        |

Number of studies in meta-analysis, median (min, q1-q3, max) 6 (4, 4–9, 28)
Subsequently, full texts of 528 systematic reviews were assessed for eligibility (Figure 1). The three most common reasons for the exclusion of full texts were (i) less than four RCTs in meta-analyses, (ii) observational studies instead of RCTs, and (iii) mixed study designs included in the meta-analyses. Finally, 261 systematic reviews studying interventions in advanced cancer patients and reporting statistically significant results of a meta-analysis were included in the analysis. The percent agreement between the reviewers was 93.5% with a Cohen’s Kappa of 0.72 (95% CI 0.69–0.75), which indicates good agreement.

Patients with lung cancer were included in 26.1% of the systematic reviews (Table 1). Most reviews evaluated pharmacological interventions (244, 93.5%) and used standard care as control intervention (156, 59.8%), that is, no additional placebo and no other active comparator. Objective outcomes (e.g., overall survival, tumor progression, and complete response) constituted 87.0% resulting in very few outcomes classified as objective decision-dependent (3.8%) and subjective (9.2%). The methodological quality according to AMSTAR 2 was critically low in 88.1% of the reviews (details published elsewhere)36 with non-registration (222, 85.1%) and non-reporting of excluded full-texts and missing justifications for exclusion (218, 83.5%) being the main reasons for this classification (Table 1).

### 3.1 Impact of switching from fixed-effect to random-effects model

Six of the 138 fixed-effect meta-analyses (4.4%) lost statistical significance when recalculating the meta-analyses using the inverse variance method, so this analysis was performed with a sample of 132 meta-analyses. When switching from fixed-effect to random-effects model with Paule–Mandel estimator of the between-study variance, 19/132 meta-analyses (14.4%) were no longer statistically significant (Figure 2a).

### 3.2 Impact of switching from random-effects to Hartung–Knapp method

Nine of 123 random-effects meta-analyses (7.3%) lost statistical significance when recalculating them using the random-effects model with the Paule–Mandel estimator. Accordingly, this analysis was performed with a sample

| Characteristics     | Sample: N = 261 |
|---------------------|-----------------|
| Model of meta-analysis: |                |
| Fixed               | 138 (52.9%)     |
| Random              | 123 (47.1%)     |
| $r$, median (min, q1–q3, max) | 0.10 (0.00, 0.00–0.25, 1.60) |
| $r^2$, median (min, q1–q3, max) | 0.01 (0.00, 0.00–0.06, 2.56) |
| $I^2$ (%), median (min, q1–q3, max) | 31.8 (0.00, 0.00–60.8, 97.0) |

Methodological quality according to AMSTAR 2:
- Critically low: 230 (88.1%)
- Low: 11 (4.2%)
- Moderate: 8 (3.1%)
- High: 12 (4.6%)

Note: AMSTAR 2 categories: high (no or one noncritical weakness), moderate (more than one noncritical weakness), low (one critical flaw with or without noncritical weaknesses), very low (more than one critical flaw with or without noncritical weaknesses).30

Abbreviations: AMSTAR, a measurement tool to assess systematic reviews; HR: hazard ratio; OR, odds ratio; q1–q3, 1.-3. quartile; RR, risk ratio, SMD, standardized mean difference.

*Outcome classification modified after Savović et al.31,32 Objective outcomes: overall survival, laboratory values, and tumor size (e.g., progression-free survival, complete response) assessed by imaging techniques and consolidated criteria.33–35 Objective decision-dependent outcomes: objectively measured, but could potentially be influenced by the clinician’s judgment (e.g., hospital admission or discharge, or initiation of treatment in the opinion of the treating physician). Subjective outcomes: based on a subjective assessment by patients or physicians (e.g., quality of life, symptoms or adverse events).
of 114 statistically significant random-effects meta-analyses. After using the Hartung–Knapp method, 34/114 meta-analyses (29.8%) were no longer statistically significant (Figure 2b).

### 3.3 | Prediction intervals

PIs were not considered in any of the 261 published meta-analyses in advanced cancer patients reporting statistically significant results. The calculation of 95% PIs showed that the null effect was included in 195/261 cases (74.7%). The null effect was included in 65/67 of meta-analyses (97.0%) with $I^2$ values of 60%–100% (Table 2). However, in meta-analyses with an $I^2$ of zero ($n = 92$), the 95% PIs still included the null effect in 47 cases (51.1%).

The opposite effect was covered by the 95% PIs in 98/261 reportedly statistically significant meta-analyses (37.5%). Fifty of 67 meta-analyses (74.6%) with $I^2$ values between 60% to 100% included the opposite effect. There were four of 92 meta-analyses (4.3%) with an $I^2$ value of zero that included the opposite effect (Table 2).

### 4 | DISCUSSION

In this methodological study of published statistically significant meta-analyses in advanced cancer patients, switching from fixed-effect to random-effects model resulted in 14.4% of meta-analyses that were no longer statistically significant. The impact of using the Hartung–Knapp method for random-effects meta-analysis was even larger with 29.8% of the meta-analyses changing from significant to nonsignificant. The 95% PI included the null effect in 74.7% and the opposite effect in 37.5% of published meta-analyses reporting statistically significant results.

### 4.1 | Explanation of results and implications

#### 4.1.1 | Impact of switching from fixed-effect to the random-effects model

Six of 138 fixed-effect meta-analyses lost statistical significance when recalculating them using the inverse variance method (function metagen in R package meta instead of the often-used Mantel–Haenszel method, which led to small differences in the recalculations compared to the original meta-analyses.

Switching from fixed-effect to a random-effects model resulted in 19/132 meta-analyses (14.4%) that were no longer statistically significant. According to Borenstein et al., the choice of the model should depend on the assumption whether the studies included in the meta-analysis are based on a common true effect ($= fixed-effect$) or whether a distribution of true effects is assumed for the included studies ($= random-effects$). In most cases, a random-effects model seems more appropriate for medical interventions, since the assumptions of the fixed-effect model do not seem realistic due to clinical and methodological heterogeneity between most studies.

#### 4.1.2 | Impact of the Hartung–Knapp method in random-effects meta-analyses

In the recalculation, nine of 123 meta-analyses (7.3%) were no longer statistically significant due to the use of the Paule–Mandel method for estimating the between-study variance instead of the DerSimonian–Laird estimator for $r^2$.

Switching to the Hartung–Knapp method resulted in 34 of 114 random-effects meta-analyses (29.8%) with advanced cancer patients that were no longer statistically significant. A comparable analysis with 689 Cochrane reviews including a wider range of patients and using the

| $I^2$ | 0% | >0%–<30% | 30%–<60% | 60%–100% | Total |
|-------|----|----------|----------|----------|-------|
|       | n = 92 | n = 35 | n = 67 | n = 67 | N = 261 |
| Null effect: | | | | | |
| Included | 47 (51.1%) | 25 (71.4%) | 58 (86.6%) | 65 (97.0%) | 195 |
| Excluded | 45 (48.9%) | 10 (28.6%) | 9 (13.4%) | 2 (3.0%) | 66 |
| Opposite effect: | | | | | |
| Included | 4 (4.3%) | 10 (28.6%) | 34 (50.7%) | 50 (74.6%) | 98 |
| Excluded | 88 (95.7%) | 25 (71.4%) | 33 (49.3%) | 17 (25.4%) | 163 |

Note: $I^2$: 0%–40%: might not be important; 30%–60%: may represent moderate heterogeneity; 50%–90%: may represent substantial heterogeneity; 75%–100%: considerable heterogeneity (Higgins, Thomas, Chandler, Cumpston, Li, Page, Welch, 2019); Categories of 0%–<30%, 30%–<60%, and 60%–100% according to IntHout et al. (2016); the reported $I^2$ values are based on the reanalysis of the extracted data.
DerSimonian–Laird estimator for $I^2$ showed that 25.1% of statistically significant random-effects meta-analyses were no longer statistically significant when the Hartung–Knapp method was applied.\textsuperscript{2} The classic meta-analysis method is considered to be anti-conservative, since a standard normal distribution is used to calculate the random-effects confidence interval. The Hartung–Knapp method usually yields wider confidence intervals, since its calculation is based on a $t$-distribution.\textsuperscript{8} In extensive simulations under the assumption of no effect (5% false-positive results allowed), the alpha error rate of the Hartung–Knapp method proved to be superior to the classic meta-analysis method. However, the authors point out that in the case of five or less studies and unequal sample sizes in the studies, the error rate for false-positive results can double to about 10% for the Hartung–Knapp method, while it can increase sixfold to about 30% for the classic meta-analysis method.\textsuperscript{5}

4.1.3 | Prediction intervals—inclusion of the null effect

Our overall result regarding the inclusion of the null effect in 95% PIs is very much in line with IntHout et al.\textsuperscript{6} (73.5%, our study: 74.7%), who evaluated the first statistically significant meta-analysis of the “data and analyses” part with a dichotomous or continuous outcome of 920 Cochrane reviews. Interestingly, in our sample of statistically significant meta-analyses with an $I^2$ of zero, the 95% PIs still included the null effect in 51.1% of the reviews, which is substantially less compared to IntHout et al. (74.6%).\textsuperscript{6} These large percentages can be explained by the use of $t$-quantiles for calculating PIs combined with the low number of studies in most meta-analyses, which inflates the $t$-quantiles. The larger proportion in IntHout's sample can mainly be explained by arbitrarily assuming an $I^2$ of 20% in the PI calculation resulting in wider PIs whereas we calculated PIs with the actual $I^2$ of 0%. Another difference is the use of a $t$-distribution with $k$-1 and $k$-2 degrees of freedom in IntHout's and our analyses, respectively. We chose $k$-2 degrees of freedom in line with Higgins et al.\textsuperscript{8} Taking the median number of studies per meta-analysis in the two empirical evaluations into account (IntHout four vs. six in this article), we see that the median degrees of freedom were three and four, respectively, contributing to wider PIs in IntHout's meta-analyses.

4.1.4 | Prediction intervals—inclusion of the opposite effect

The opposite effect was included by the 95% PI in 37.5% of meta-analyses for $I^2$ values of 0% and above 0%. IntHout et al. did not report PI results regarding the opposite effect for meta-analyses with an $I^2$ of 0. In meta-analyses with an $I^2$ above 0%, there was a relevant difference to IntHout et al. regarding the inclusion of the opposite effect: 20.3% versus 55.6% in our study. One reason for this difference might be that we used the same meta-analysis model (fixed or random) as the authors of the original publications including confidence intervals based on the $z$-distribution to select statistically significant meta-analyses. In contrast, IntHout et al. included meta-analyses from a sample of 3263 only after recalculating them using the Hartung–Knapp method for random-effects meta-analyses based on a $t$-distribution, even in originally performed fixed-effect meta-analyses.\textsuperscript{6} In order to obtain a statistically significant result with the Hartung–Knapp method, treatment effects must be either large or treatment estimates must be very precise. Larger treatment effects mean that the meta-analysis estimate is further away from the null effect resulting in a more extreme opposite effect. In our sample, especially fixed-effect meta-analyses became significant for smaller treatment effects, and using a quantile from a $t$- instead of $z$-distribution probably had a large impact on the inclusion of the opposite effect.

IntHout's meta-analyses were sampled from the Cochrane Database of Systematic Reviews (2009–2013) and not from a wide range of journals like in our sampling procedure, that is, via Medline, the Cochrane Database of Systematic Reviews, and Web of Science. The higher methodological quality\textsuperscript{37–39} and higher precision (i.e., smaller standard errors, narrower confidence intervals)\textsuperscript{37} of Cochrane reviews support the differences regarding the inclusion of the opposite effect, while the tendency of lower effects in Cochrane reviews\textsuperscript{37,40} and absence of differences in precision between Cochrane reviews and non-Cochrane reviews in another study\textsuperscript{40} question this explanation.

Additionally, we can hardly assess if differences in the clinical characteristics may have contributed to the different results because they were only provided for the overall sample ($N = 3263$)\textsuperscript{10} but not for the sample for which the opposite effect was analyzed ($n = 479$).\textsuperscript{6}

Moreover, while IntHout et al. focused on showing the advantages of routinely reporting the PI in meta-analyses for a better understanding of heterogeneity, we aimed at examining conclusion-changing results of PIs and additionally investigated the impact of how statistical significance would change if different methods were applied.

4.2 | Limitations

It should be noted that a 95% PI applies only to very similar, future studies. Furthermore, the study effects are
assumed to be normally distributed, which cannot be verified, especially when only few studies are included in a meta-analysis.\(^5\)

The sample in the present study is not representative for systematic reviews in general or in the field of oncology but rather refers to reviews including patients with advanced, incurable cancer with metastases or a life expectancy below 2 years because of our specific medical background. In addition, interventions were pharmacological in 93.5\% and lung cancer patients were included in 26.1\% of included reviews, which further limits generalizability.

Some may argue that IntHout et al. used more stringent criteria to select significant studies by selecting all meta-analyses, which showed a significant result for the Hartung–Knapp method. In contrast, we included all studies that were either statistically significant using the fixed-effect or random-effects model. However, by using the originally published meta-analytic results in our approach we believe that the PI results presented in this article draw an important and critical picture of the currently published evidence for the treatment of advanced cancer patients.

We included only statistically significant meta-analyses with at least four RCTs in order to allow a clinically meaningful analysis regarding the inclusion of no effect and the opposite effect by the 95\% PI. The first statistically significant meta-analysis reported in the abstract or full text was selected. This somewhat arbitrary decision may reflect selection bias within systematic reviews with more than one meta-analysis, and the impact on our analyses is unclear. However, primary outcomes are typically reported first in publications and we aimed to identify meta-analyses with a high clinical relevance as such meta-analyses might have an impact on treatment guidelines. Despite our selection procedure, the values for the between-study variance \(\tau^2\) were comparable to the total sample of IntHout et al.\(^10\) \((N = 3263)\), while \(I^2\) differed moderately: median of our data versus IntHout's data: \(\tau^2\) 0.01 versus 0 and \(I^2\) 31.8\% versus 0. The median of tau was comparable to medians of predictive distributions in meta-analyses with objective outcomes from the Cochrane Database of Systematic Reviews: median of 0.10 in our sample and median of 0.12 and 0.13 comparing a pharmacological intervention with another pharmacological intervention or placebo.\(^31\) The reported limitations along with the small differences to the samples of other methodological studies\(^10,41\) should be considered when interpreting the generalizability of our findings.

On a general note, the presence of some publication bias in the included meta-analyses should be taken into account.\(^42,43\) Resolution of publication bias would probably result in lower overall effect estimates and in increased statistical heterogeneity in the meta-analyses due to the extended range of study effects.

5 | CONCLUSION

A relevant percentage of published meta-analyses in advanced cancer patients with statistically significant results were no longer statistically significant when switching from fixed-effect to random-effects model or using the recommended Hartung–Knapp method in random-effects meta-analyses. This empirical result is important for the interpretation of many meta-analyses having an impact on treatment guidelines, health policy, and the best patient care.

If review authors choose the random-effects model in their meta-analyses, which is typically more adequate than the fixed-effect model in medicine, we strongly encourage them to check if using the Hartung–Knapp method is appropriate.

Clinical decision-making solely based on the pooled estimate of a meta-analysis and its 95\% CI can be misleading. Our results showed that the opposite effect was included by the 95\% PI in 37.5\% of statistically significant meta-analyses in advanced cancer patients. This means that the average effect in a similar, future study may shift from halving to doubling of the mortality risk, which would have a major impact on patients. Accordingly, in order to see the whole picture of heterogeneity in a meta-analysis and facilitate clinical decision-making, the reporting of 95\% PIs in meta-analyses is highly recommended.

ACKNOWLEDGMENTS

We are very grateful to Lisa Joos (Psychologist, M.Sc.) for her great help regarding the data extraction in this project.

CONFLICT OF INTEREST

This work was carried out in the scope of the PhD thesis of Waldemar Siemens. Waldemar Siemens was employed at Roche Pharma AG, Grenzach-Wyhlen, Germany, from April 2020 to June 2021. Data analysis was finalized in December 2019. Roche Pharma AG was not involved in the project and had no influence at any time on the project. Waldemar Siemens is employed at the Institute for Evidence in Medicine (Medical Center, University of Freiburg, Germany) and Cochrane Germany in Freiburg since September 2021. Guido Schwarzer is an external statistical consultant of Roche Pharma AG, Grenzach-Wyhlen, Germany, since August 2019. All other authors have no competing interests to declare.
AUTHOR CONTRIBUTIONS
Conception and design: Waldemar Siemens, Joerg J. Meerpohl, Guido Schwarzer, and Gerhild Becker. Collection and assembly of data: Waldemar Siemens. Data assembly, analysis and interpretation: Waldemar Siemens, Joerg J. Meerpohl, Guido Schwarzer, and Gerhild Becker. Manuscript writing: Waldemar Siemens, Guido Schwarzer. Critical revision and final approval of manuscript: All authors. Accountable for all aspects of the work: All authors.

DATA AVAILABILITY STATEMENT
The R code and data of this study are available at https://osf.io/xr6c5/.

ORCID
Waldemar Siemens https://orcid.org/0000-0003-4238-5327
Joerg J. Meerpohl https://orcid.org/0000-0002-1333-5403
Sabine Buroh https://orcid.org/0000-0001-6474-4656
Guido Schwarzer https://orcid.org/0000-0001-6214-9087
Gerhild Becker https://orcid.org/0000-0001-9060-0285

REFERENCES
1. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA. Cochrane Handbook for Systematic Reviews of Interventions: version 6.2 (updated February 2021). Accessed March 15, 2021. www.training.cochrane.org/handbook
2. IntHout J, Ioannidis JPA, Born G. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol. 2014;14:25. doi:10.1186/1471-2288-14-25
3. Hartung J, Knapp G. On tests of the overall treatment effect in meta-analysis with normally distributed responses. Stat Med. 2001;20(12):1771-1782. doi:10.1002/sim.791
4. Hartung J, Knapp G. A refined method for the meta-analysis of controlled clinical trials with binary outcome. Stat Med. 2001;20(24):3875-3889. doi:10.1002/sim.1009
5. Higgins JPT, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. J R Stat Soc Ser A Stat Soc. 2009;172(1):137-159. doi:10.1111/j.1467-985X.2008.00552.x
6. IntHout J, Ioannidis JPA, Rovers MM, Goeman JJ. Plea for routinely presenting prediction intervals in meta-analysis. BMJ Open. 2016;6(7):e010247. doi:10.1136/bmjopen-2015-010247
7. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. Introduction to Meta-Analysis. Wiley; 2009.
8. Jackson D, Law M, Rücker G, Schwarzer G. The Hartung-Knapp modification for random-effects meta-analysis: a useful refinement but are there any residual concerns? Stat Med. 2017;36(25):3923-3934. doi:10.1002/sim.7411
9. Wiksten A, Rücker G, Schwarzer G. Hartung-Knapp method is not always conservative compared with fixed-effect meta-analysis. Stat Med. 2016;35(15):2503-2515. doi:10.1002/sim.6879
10. IntHout J, Ioannidis JPA, Born GF, Goeman JJ. Small studies are more heterogeneous than large ones: a meta-meta-analysis. J Clin Epidemiol. 2015;68(8):860-869. doi:10.1016/j.jclinepi.2015.03.017
11. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ. 2003;327(7414):557-560. doi:10.1136/bmj.327.7414.557
12. Hoaglin DC. Misunderstandings about Q and ‘Cochran’s Q’ test in meta-analysis. Stat Med. 2016;35(4):485-495. doi:10.1002/sim.6632
13. Borenstein M, Higgins JPT, Hedges LV, Rothstein HR. Basics of meta-analysis: I2 is not an absolute measure of heterogeneity. Res Syn Meth. 2017;8(1):5-18. doi:10.1080/19460042.2016.11929568
14. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med. 2009;6(7):e1000097. doi:10.1371/journal.pmed.1000097
15. American Cancer Society. Cancer Facts & Figures; 2019. Accessed June 25, 2020. https://www.cancer.org/content/dam/cancer-organization/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2019/cancer-facts-and-figures-2019.pdf
16. McMaster University, Health information research unit. Search Filters for MEDLINE in Ovid Syntax and the PubMed Translation. Accessed June 25, 2020. https://hiru.mcmaster.ca/hiru/HIRU_Hedges_MEDLINE_Strategies.aspx
17. 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(1):55-79. doi:10.1002/jrsm.1164
18. Hartung J. An alternative method for meta-analysis. Biom J. 1999;41(8):901-916. doi:10.1002/(SICI)1521-4036(199912)41:8<901::AID-BIMJ901>3.0.CO;2-W
19. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. Res Syn Meth. 2010;1(2):97-111. doi:10.1002/jrsm.13
20. 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(1):55-79. doi:10.1002/jrsm.1164
21. Bühn S, Mathes T, Prengel P, et al. The risk of bias in systematic reviews published in high-impact journals neglected to register the protocols: a meta-epidemiological study. J Clin Epidemiol. 2017;84:50-58. doi:10.1016/j.jclinepi.2017.02.008
22. Tian J, Zhang J, Ge L, Yang K, Song F. The methodological and reporting quality of systematic reviews from China and the USA are similar. J Clin Epidemiol. 2017;85:50-58. doi:10.1016/j.jclinepi.2016.12.004
23. Tao Y, Sun R, Chen Y-L, et al. The quality of evidence in Chinese meta-analyses needs to be improved. J Clin Epidemiol. 2016;74:73-79. doi:10.1016/j.jclinepi.2016.01.003
24. Wood L, Egger M, Gluud LL, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. BMJ. 2008;336(7644):601-605. doi:10.1136/bmj.39465.451748.AD
25. Bühn S, Mathes T, Prengel P, et al. The risk of bias in systematic reviews tool showed fair reliability and good construct validity. J Clin Epidemiol. 2017;91:121-128. doi:10.1016/j.jclinepi.2017.06.019
28. Ge L, Tian J-H, Li Y-N, et al. Association between prospective registration and overall reporting and methodological quality of systematic reviews: a meta-epidemiological study. *J Clin Epidemiol.* 2018;95:45-55. doi:10.1016/j.jclinepi.2017.10.012

29. Schwarzer G, Carpenter JR, Rücker G. *Meta-Analysis with R.* Springer; 2015.

30. Shea BJ, Reeves BC, Wells G, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *BMJ.* 2017;358:j4008. doi:10.1136/bmj.j4008

31. Savović J, Harris RJ, Wood L, et al. Development of a combined database for meta-epidemiological research. *Res Syn Methods.* 2010;1(3-4):212-225. doi:10.1002/jrsm.18

32. Savović J, Jones H, Altman D, et al. Influence of reported study design characteristics on intervention effect estimates from randomised controlled trials: combined analysis of meta-epidemiological studies. *Health Technol Assess.* 2012;16(35):1-82. doi:10.3310/hta16350

33. Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer.* 2009;45(2):228-247. doi:10.1016/j.ejca.2008.10.026

34. Schwartz LH, Liticher S, de Vries E, et al. RECIST 1.1—update and clarification: from the RECIST committee. *Eur J Cancer.* 2016;62:132-137. doi:10.1016/j.ejca.2016.03.081

35. Therasse P, Arbuck SG, Eisenhauer EA, et al. New guidelines to evaluate the response to treatment in solid tumors. *J Natl Cancer Inst.* 2000;92(3):205-216. doi:10.1093/jnci/92.3.205

36. Siemens W, Schwarzer G, Rohe MS, Buroh S, Meerpolh JJ, Becker G. Methodological quality was critically low in 9/10 systematic reviews in advanced cancer patients—a methodological study. *J Clin Epidemiol.* 2021;136:84-95. doi:10.1016/j.jclinepi.2021.03.010

37. Useem J, Brennan A, LaValley M, et al. Systematic differences between Cochrane and non-Cochrane meta-analyses on the same topic: a matched pair analysis. *PLoS One.* 2015;10(12): e0144980. doi:10.1371/journal.pone.0144980

38. Moseley AM, Elkins MR, Herbert RD, Maher CG, Sherrington C. Cochrane reviews used more rigorous methods than non-Cochrane reviews: survey of systematic reviews in physiotherapy. *J Clin Epidemiol.* 2009;62(10):1021-1030. doi:10.1016/j.jclinepi.2008.09.018

39. Hacke C, Nunan D. Factors underpinning discrepant findings from reviews on the same topic: a systematic analysis of Cochrane and non-Cochrane systematic reviews. *bioRxiv;* 2019. doi:10.1101/574046

40. Hacke C, Nunan D. Physical activity interventions for major chronic disease: a matched-pair analysis of Cochrane and non-Cochrane systematic reviews. *bioRxiv;* 2019. doi:10.1101/571901

41. Turner RM, Davey J, Clarke MJ, Thompson SG, Higgins JP. Predicting the extent of heterogeneity in meta-analysis, using empirical data from the Cochrane database of systematic reviews. *Int J Epidemiol.* 2012;41(3):818-827. doi:10.1093/ije/dys041

42. Hopewell S, Loudon K, Clarke MJ, Oxman AD, Dickersin K. Publication bias in clinical trials due to statistical significance or direction of trial results. *Cochrane Database Syst Rev.* 2009;2009(1):MR000006. doi:10.1002/14651858.MR000006.pub3

43. Guyatt GH, Oxman AD, Montori V, et al. GRADE guidelines: 5. Rating the quality of evidence—publication bias. *J Clin Epidemiol.* 2011;64(12):1277-1282. doi:10.1016/j.jclinepi.2011.01.011

**SUPPORTING INFORMATION**

Additional supporting information may be found in the online version of the article at the publisher’s website.

**How to cite this article:** Siemens W, Meerpolh JJ, Rohe MS, Buroh S, Schwarzer G, Becker G. Reevaluation of statistically significant meta-analyses in advanced cancer patients using the Hartung–Knapp method and prediction intervals—A methodological study. *Res Syn Meth.* 2022;13(3):330-341. doi:10.1002/jrsm.1543