Received: 22 January 2021 | Revised: 6 June 2021 | Accepted: 9 August 2021

DOI: 10.1002/jrsm.1522

RESEARCH ARTICLE

Research Synthesis Methods WILEY

Cumulative meta-analysis: What works

Elena Kulinskaya | **Eung Yaw Mah**

School of Computing Sciences, University of East Anglia, Norwich, UK

Correspondence
Elena Kulinskaya, School of Computing Sciences, University of East Anglia, Norwich Research Park, Norwich NR4 7TJ, UK.
Email: e.kulinskaya@uea.ac.uk

Funding information
Economic and Social Research Council, Grant/Award Number: ES/L011859/1

To present time-varying evidence, cumulative meta-analysis (CMA) updates results of previous meta-analyses to incorporate new study results. We investigate the properties of CMA, suggest possible improvements and provide the first in-depth simulation study of the use of CMA and CUSUM methods for detection of temporal trends in random-effects meta-analysis. We use the standardized mean difference (SMD) as an effect measure of interest. For CMA, we compare the standard inverse-variance-weighted estimation of the overall effect using REML-based estimation of between-study variance τ^2 with the sample-size-weighted estimation of the effect accompanied by Kulinskaya–Dollinger–Bjørkestøl (*Biometrics*. 2011; 67:203–212) (KDB) estimation of τ^2. For all methods, we consider Type 1 error under no shift and power under a shift in the mean in the random-effects model. To ameliorate the lack of power in CMA, we introduce two-stage CMA, in which τ^2 is estimated at Stage 1 (from the first 5–10 studies), and further CMA monitors a target value of effect, keeping the τ^2 value fixed. We recommend this two-stage CMA combined with cumulative testing for positive shift in τ^2. In practice, use of CMA requires at least 15–20 studies.

KEYWORDS
CUSUM charts, effective-sample-size weights, inverse-variance weights, power, type 1 error

Highlights

What is already known

- Cumulative meta-analysis is a popular method of evaluating and monitoring temporal changes in accumulating evidence. Statistical evaluation of CMA is especially relevant because of the growing popularity of living systematic reviews.
- Repeated testing of the overall effect in CMA results in inflation of the Type 1 error rate. However, the amount of this inflation was not yet sufficiently quantified; it depends on the particular effect measure and on the choice of estimators.

What is new

- In extensive simulations for CMA of SMD δ, we observed considerable, and continuously increasing with addition of further studies, negative biases of the inverse-variance-weighed overall effect. In contrast, the sample-size-weighted overall effect is almost unbiased.
- This bias produces much larger inflation of the Type 1 error rate for IV REML in comparison with SSW KDB. The use of inverse-variance-weighted estimation of SMD is not recommended in CMA.
- A shift in the mean effect results in a very slow change of the overall effect in CMA, but in a considerable and rapid change in the estimated between-study variance τ^2. This change reduces the power of CMA.
- To increase the power of CMA, we introduced a two-stage CMA, in which τ^2 is estimated in Stage 1 without further re-estimation in Stage 2. This two-stage CMA with the sample-size-weighted overall effect performs better than the standard CMA and is recommended.
- A new CMA method based on re-estimation of τ^2 using the QP-based test is also useful when τ^2 is small. We recommend use of both new methods simultaneously.
- We also provide tables enabling choice of significance levels.

Potential impact for RSM readers outside the authors' field

- For CMA of SMD, we recommend the two-stage analysis based on SSW, used simultaneously with a QP-based test for shift in between-study variance.
- CMA of SMD has rather low power, which is reduced by inertia of the cumulative estimation and by heterogeneity of the data. At least 15–20 studies are required for CMA to be useful in practice.

1 | INTRODUCTION

Meta-analysis, a statistical methodology that combines estimated effects from multiple studies on the same topic, to arrive at an evidence-based overall effect, had revolutionized many scientific fields, helping to establish evidence-based practices and to resolve seemingly contradictory research outcomes.[1] The conduct of meta-analysis provides a snapshot of evidence at one time point, but the evidence is not static: as it accumulates, new studies often challenge the results of previous studies. If the evidence changes over time, the conclusions of meta-analysis will strongly depend on when the review was conducted, and any policy-relevant recommendations derived from it will quickly go out of date.[1,2]

Numerous studies have shown that substantial changes over time in the magnitude, the statistical significance, and even the sign of the reported effects are common in numerous disciplines, from biomedical research[2–5] to social sciences,[6–9] ecology and evolutionary biology,[10–13] and information systems.[14] These temporal trends could be caused by changes in the true effect, changes in study characteristics that influence the effect (known as moderators in meta-analysis) or biases (e.g., delay in the publication of studies with nonsignificant results).[10,15]

Temporal trends are typically visually explored and often formally detected through cumulative meta-analysis (CMA), introduced by Lau et al.[16] CMA is a process of updating the results of an existing meta-analysis to incorporate new study results. It is one of the most popular ways to present time-varying evidence.[4,17,18] Other methods for detecting temporal trends are reviewed in Trikalinos and Ioannidis,[19] Kulinskaya and Koricheva,[20] and Koricheva et al.[21] Until recently, CMA was applied to systematic reviews identified to be in need of updating.[22,23] But the use of CMA has grown substantially because of the growing popularity of living systematic reviews, online summaries updated as new research becomes available.[24] CMA also provides a graphical representation of shifts in evidence associated with other factors such as sample size, precision, or study quality.[25,26]

It is well understood that the repeated testing inherent in CMA inflates Type 1 error. This inflation was studied by simulation by Whitehead,[27] Hu et al.,[28] and Thorlund et al.[29] among others; but, to our knowledge, it was not systematically quantified. A number of methods

addressing this issue are based on methodology originally developed for sequential clinical trials.[30–33] Another approach uses quality control methods, in particular CUSUM charts.[20] However, the use of the random-effects model in CMA requires consecutive re-estimation of the between-study variance τ^2 as well as the overall effect δ, and makes both classes of methods problematic.[34,35]

In this paper, we investigate the properties of CMA, suggest possible improvements, and provide an in-depth simulation study of the use of CMA and CUSUM methods for detecting temporal trends in random-effects meta-analysis. We use the standardized mean difference (SMD) as the effect measure. For CMA, we consider both the standard inverse-variance-weighted estimator of the overall effect (δ) with REML-based estimation of the between-studies variance (τ^2) and a sample-size-weighted (SSW) estimator of δ combined with the Kulinskaya–Dollinger–Bjørkestøl[36] (KDB) estimator of τ^2, recommended by Bakbergenuly et al.[37] For all methods, we consider Type 1 error under no shift and power under a shift in the mean.

The requisite statistical methods are described in Section 2. Section 3 examines the properties of CMA and identifies the problems caused by gross overestimation of τ^2 resulting from a shift in the mean. Therefore, we suggest cumulative testing of τ^2. Also, we modify CMA in the spirit of quality control methods. At Stage 1 (the first 5–10 studies), we estimate both δ and τ^2, and then we 'monitor' known or estimated in Stage 1 value of effect, keeping the estimated value of τ^2 fixed. Section 4 presents the design and results of our simulations for standard and two-stage CMA methods and for CUSUM charts. Section 5 applies the methods to data on species richness, and Section 6 concludes with discussion. Our full simulation results are available as e-print.[38] R procedures implementing the proposed methods of CMA are available in the file `CMA for SMD.txt` and examples and R scripts of their use are provided in Appendices S3 and S4.

2 | PRELIMINARIES

2.1 | Study-level estimation of SMD

Consider a meta-analysis of k comparative studies, each consisting of two arms, treatment (T) and control (C), with sample sizes n_{iT} and n_{iC}. The total sample size in study i is $n_i = n_{iT} + n_{iC}$, and the ratio of the control sample size to the total is denoted by $q_i = n_{iC}/n_i$. The subject-level data in each arm are assumed to be normally distributed with means μ_{iT} and μ_{iC} and equal variances σ_i^2. The sample means are $\overline{x}_{ij}$, and the sample variances are s_{ij}^2, for $i = 1,...,k$ and $j = C$ or T.

The SMD effect measure is

$$\delta_i = \frac{\mu_{iT} - \mu_{iC}}{\sigma_i}.$$

The unbiased estimator of δ_i, sometimes called Hedges's g or Hedges's d, is

$$g_i = J(m_i)\frac{\overline{x}_{iT} - \overline{x}_{iC}}{s_i}, \tag{1}$$

where $m_i = n_{iT} + n_{iC} - 2$, and $J(m) = \Gamma\left(\frac{m}{2}\right)/\sqrt{\frac{m}{2}}\Gamma\left(\frac{m-1}{2}\right)$, often approximated by $1 - 3/(4m-1)$, corrects for bias.[39] The standard deviation, σ_i is estimated by the square root of the pooled sample variance

$$s_i^2 = \frac{(n_{iT}-1)s_{iT}^2 + (n_{iC}-1)s_{iC}^2}{n_{iT} + n_{iC} - 2}.$$

The variance of g_i is $\mathrm{Var}(g_i) = \widetilde{n}_i^{-1} + \delta_i^2/2(n_{iT} + n_{iC})$, where $\widetilde{n}_i = n_{iT}n_{iC}/(n_{iT} + n_{iC})$ is the effective sample size in study i. An unbiased estimator of this variance is[39]

$$v_i^2 = \widetilde{n}_i^{-1} + \left(1 - \frac{(m_i - 2)}{m_i J(m_i)^2}\right)g_i^2. \tag{2}$$

The sample SMD g_i has a scaled non-central t-distribution on m_i d.f. and non-centrality parameter $[n_i q_i(1-q_i)]^{1/2}\delta_i$:

$$\frac{\sqrt{n_i q_i(1-q_i)}}{J(m_i)}g_i \sim t_{m_i}\left([n_i q_i(1-q_i)]^{1/2}\delta_i\right). \tag{3}$$

2.2 | Standard random-effects model

In meta-analysis of k studies, the standard random-effects model assumes approximately normal distributions of within- and between-study effects. For a generic measure of effect,

$$\hat{\delta}_i \sim N(\delta_i, \sigma_i^2) \text{ and } \delta_i \sim N(\delta, \tau^2), \tag{4}$$

resulting in the marginal distribution $\hat{\delta}_i \sim N(\delta, \sigma_i^2 + \tau^2)$. $\hat{\delta}_i$ is the estimate of the effect in Study i, and its within-study variance is σ_i^2, estimated by $\hat{\sigma}_i^2$, $i = 1,...,k$. τ^2 is the between-study variance, estimated (from k studies) by $\hat{\tau}_k^2$. The overall effect $\delta = \delta_{(k)}$ can be estimated from k studies by the weighted mean

$$\hat{\delta}_{\mathrm{IV}(k)} = \frac{\sum_{i=1}^{k} \hat{w}_i\left(\hat{\tau}_k^2\right)\hat{\delta}_i}{\sum_{i=1}^{k} \hat{w}_i\left(\hat{\tau}_k^2\right)}, \tag{5}$$

where the $\hat{w}_i\left(\hat{\tau}_k^2\right) = \left(\hat{\sigma}_i^2 + \hat{\tau}_k^2\right)^{-1}$ are inverse-variance (IV) weights. The fixed-effect (FE) model assumes that $\tau^2 \equiv 0$, and the estimator $\hat{\delta}_{\mathrm{FE}(k)}$ uses the IV weights $\hat{w}_i = \hat{w}_i(0)$.

The variance of $\hat{\delta}_{\mathrm{IV}(k)}$ is routinely estimated by

$$\hat{\mathrm{Var}}\left(\hat{\delta}_{\mathrm{IV}(k)}\right) = \left[\sum_{i=1}^{k} \hat{w}_i\left(\hat{\tau}_k^2\right)\right]^{-1}. \tag{6}$$

The standard confidence interval for the overall effect $\delta_{(k)}$ uses the IV point estimator as its centre, and its half-width equals the estimated standard deviation (square root of the variance (6)) times the critical value from the normal distribution.

2.3 | Point and interval estimation of τ^2

Because the $\hat{w}_i\left(\hat{\tau}_k^2\right)$ in Equation (5) involve $\hat{\tau}_k^2$, we need to choose an estimator of τ^2. It is well known that the maximum likelihood estimator of τ^2 is biased, and the restricted maximum-likelihood (REML) estimator is a good choice.[40] We use $\hat{\tau}^2_{\mathrm{REML}}$ in the IV estimator of δ and the Q-profile (QP) method[41] for the accompanying interval estimator of τ^2. For SMD, this method is one of the best traditional methods.[37]

The Q-profile confidence interval can be obtained from the lower and upper quantiles of F_Q, the approximate cumulative distribution function of Cochran's Q statistic[42] $Q\left(\hat{\tau}^2\right) = \sum \hat{w}_i\left(\hat{\tau}^2\right)\left(\hat{\delta}_i - \hat{\delta}_{\mathrm{IV}}\right)^2$:

$$Q\left(\tau_L^2\right) = F_{Q;1-\alpha/2}, \quad Q\left(\tau_U^2\right) = F_{Q;\alpha/2}. \tag{7}$$

The lower and upper confidence limits, τ_L^2 and τ_U^2, can be calculated iteratively. This Q statistic is often assumed to have a Chi-square distribution with $k-1$ degrees of freedom when $\tau^2 = 0$; $F_Q = \chi^2_{k-1}$ in the original Q-profile method.[41] However, χ^2_{k-1} is an adequate approximation only for very large sample sizes.

For SMD, Kulinskaya et al.[36] derived $O(1/n)$ corrections to moments of Q and suggested using the Chi-square distribution with estimated degrees of freedom based on the corrected first moment of Q to approximate its distribution. Bakbergenuly et al.[37] proposed an estimator of τ^2, $\hat{\tau}^2_{\mathrm{KDB}}$, based on this improved approximation.

An accompanying KDB confidence interval for τ^2 combines the Q-profile approach and the improved approximation by Kulinskaya et al.[36] Bakbergenuly et al.[37] demonstrated by simulation that $\hat{\tau}^2_{\mathrm{KDB}}$ is less biased than REML for small sample sizes ($n < 100$), especially for $k \geq 10$, and both estimators become almost unbiased from $n = 100$. Both the QP and the KDB confidence intervals for τ^2 have too high coverage for τ^2 near zero. For larger τ^2, QP performs well, and the KDB interval may be somewhat too liberal for small n.

Confidence intervals for τ^2 are related to tests of the null hypotheses $\tau^2 = \tau_0^2$. Values of τ_0^2 beyond the estimated confidence limits $\left(\hat{\tau}_L^2, \hat{\tau}_U^2\right)$ are rejected by a two-sided α-level test, and values below the lower bound of the one-sided interval $\left[\hat{\tau}_L^2, \infty\right)$ are rejected by a one-sided $\alpha/2$-level test in favour of $\tau^2 > \hat{\tau}_L^2$. We refer to these tests as QP and KDB.

2.4 | Point and interval estimators of δ

Traditional CMA based on REML uses $\hat{\tau}^2_{\mathrm{REML}}$ in $\hat{\delta}_{\mathrm{IV}(k)}$ (5) and in its estimated variance (6), in combination with the critical values from the normal distribution. We refer to this method as IV REML.

For SMD, the estimated effects $\hat{\delta}_i$ (1) and their estimated variances v_i^2 (2) are not independent. Because of this, the IV estimates of the overall effect $\hat{\delta}_{\mathrm{IV}(k)}$ are biased. Use of non-random weights eliminates this bias. The use of effective-sample-size weights (SSW) for estimation of δ, suggested by Hedges and Olkin,[43,p.110] was explored and found superior to IV weights in a comprehensive simulation study by Bakbergenuly et al.[37]

These weights depend only on the studies' arm-level sample sizes: $w_i = \widetilde{n}_i = n_{iT}n_{iC}/(n_{iT} + n_{iC})$; $\widetilde{n}_i$ is the effective sample size in Study i. They coincide with the inverse-variance weights when $g_i = 0$ in (2). We refer to the estimator of δ of the form (5) with these weights as SSW and denote it by $\hat{\delta}_{\mathrm{SSW}}$. The interval estimator corresponding to SSW (SSW KDB) uses the SSW point estimator as its centre, and its half-width equals the estimated standard deviation of SSW under the random-effects model times the critical value from the t-distribution on $k-1$ degrees of freedom. The estimator of the variance of $\hat{\delta}_{\mathrm{SSW}}$ is

$$\hat{\mathrm{Var}}\left(\hat{\delta}_{\mathrm{SSW}}\right) = \frac{\sum \widetilde{n}_i^2\left(v_i^2 + \hat{\tau}^2_{\mathrm{KDB}}\right)}{\left(\sum \widetilde{n}_i\right)^2}, \tag{8}$$

in which v_i^2 comes from Equation (2).

Once more, we refer to the tests of the hypothesis $\delta = \delta_0$ based on the respective confidence intervals as

IV REML and SSW KDB. In particular, the α-level two-sided IV REML test compares its statistic $\widehat{\mathrm{Var}}\left(\hat{\delta}_{\mathrm{IV\,REML}}\right)^{-1/2}\left|\hat{\delta}_{\mathrm{IV\,REML}}-\delta_0\right|$ against the normal critical value $z_{1-\alpha/2}$, and the SSW KDB test compares its statistic $\hat{\mathrm{Var}}\left(\hat{\delta}_{\mathrm{SSW}}\right)^{-1/2}\left|\hat{\delta}_{\mathrm{SSW}}-\delta_0\right|$ against the critical value from the t-distribution on $k-1$ degrees of freedom.

2.5 | CUSUM charts and their use in sequential meta-analysis

Methods of statistical quality control (QC) were initially developed in industrial applications of statistics, and are now commonly used in medicine, epidemiology and public health (e.g., to detect a start of an epidemic or to monitor quality of hip implants). Their use in meta-analysis for detection of temporal trends was suggested by Kulinskaya and Koricheva.[20] A process that is operating with only random causes of variation is said to be in statistical control; otherwise, the process is out of control.[44] The standard QC application includes two stages: a set-up stage where the parameters of a process in control are estimated, and the subsequent monitoring stage. At the monitoring stage, the samples (of size n) are collected on a regular basis; and if the values of interest (say, the sample mean $\bar{x}_n$) fall within the control limits on the control chart and do not exhibit any systematic pattern, the process is considered to be in control.

The cumulative sum or CUSUM chart[45] is one of the most efficient QC tools. It is equivalent to a sequential likelihood ratio test of the null hypothesis H_0 (the process is in control) against an alternative H_1 (the process is out of control). The cumulative log likelihood ratio (LLR) of H_1 versus H_0 is plotted at every step i, $i=1,2,...$, and the test stops in favour of H_1 when the LLR is large.[46]

In standard applications, the null hypothesis $H_0: y_i \sim N(\mu_0,\sigma^2)$ is tested against the alternative $H_1: y_i \sim N(\mu_0+\Delta\sigma,\sigma^2)$ of a practically relevant amount of shift Δ (times the standard deviation) in the underlying mean. Usually, two one-sided CUSUM charts (for positive and negative deviations) are plotted simultaneously. An upper one-sided CUSUM testing the hypothesis $H_0: \Delta=0$ versus $H_1: \Delta>0$ can be written as a cumulative LLR: $x_0=0, x_i=\max(0,x_{i-1}+l_i)$, $i=1,2,...$ for the ith log-likelihood contribution $l_i=\Delta((y_i-\mu_0)/\sigma-\Delta/2)$.

In meta-analysis, when there is no temporal shift, the process is in control, and all effect estimates are approximately normally distributed with the same mean, $\hat{\delta}_i \sim N\left(\delta,\sigma_i^2\right)$. If a shift of size Δ occurs at some time point, the mean of the process deviates from δ, so that $\hat{\delta}_i \sim N\left(\delta+\Delta,\sigma_i^2\right)$, and the process can be considered out of control. Applying the CUSUM chart to meta-analysis, the LLR for study i is[20] $l_i=\Delta\left[w_i\left(\hat{\delta}_i-\delta-\Delta/2\right)\right]$, where the weights are the inverse variances of the $\hat{\delta}_i$. Thus, the upper CUSUM accumulates weighted deviations from the target value that are greater than $\Delta/2$. Under positive shift, the expected path of a CUSUM increases linearly with the number of positive deviations and with their average weight, until it crosses the chosen control limit.

An important notion associated with the chosen significance level in sequential testing is the average run length (ARL) of the control chart. The ARL is the expected number of points plotted before a point falls outside the control limits. The CUSUM signals as soon as $x_i>h$. The value of the control limit, h, is chosen to provide good ARL values. The standard choices are $h=4$ or $h=5$ standard deviations. These values correspond to an ARL of 168 and 465, respectively, when the process is in control, and to an ARL of 4.75 and 5.75, respectively, for a shift of 1.5σ.[44] We use the R package *qcc*[47] for CUSUM analysis.

The use of QC charts in meta-analysis is justified in the fixed-effect model; but in the random-effects model, re-estimation of τ^2 introduces dependence between the sequential effect estimates, and hence their distribution is not consistent with the standard assumption of independent increments in the QC charts.[35]

3 | ESTIMATION AND TESTING IN CMA

Consider K consecutive studies and a simple random-effects model for study-level effects, with a simple shift in the mean (at study k_1+1):

$$\hat{\delta}_i \sim G\left(\delta_i+\Delta I(i>k_1),\sigma_i^2\right) \text{ and } \delta_i \sim N\left(\delta,\tau^2\right),\ i=1,...,K, \quad (9)$$

where $G(\cdot)$ is an arbitrary distribution with mean $\delta_i+\Delta I(i>k_1)$ and variance σ_i^2, K is the maximum number of studies, $I(i>k_1)$ is the 0/1 indicator and $1<k_1<K$. In the shift-in-the-mean model (9), the true SMD shifts from δ in the first k_1 studies to $\delta+\Delta$ in studies k_1+1, ..., K. The standard REM (4) holds for the first k_1 studies and, separately, for the studies from k_1+1 to K when the distribution $G(\cdot)$ is normal. For SMD, the distribution $G(\cdot)$ is given by Equation (3).

In this section, we consider some general CMA patterns under this model and suggest several approaches to testing.

3.1 | Estimation of δ

Define cumulative overall mean effect $\delta_{(k)}$ of $k\le K$ studies to be a weighted mean of study-level effects δ_i, $i=1,\cdots,k$: $\delta_{(k)}=\sum_{i=1}^{k} w_i\delta_i/\sum_1^k w_i$. In CMA, this cumulative mean effect $\delta_{(k)}$ is estimated by $\hat{\delta}_{(k)}=\sum_1^k \hat{w}_i\hat{\delta}_i/\sum_1^k \hat{w}_i$ whenever the number of studies k increases.

If there is no shift in δ, the cumulative mean $\delta_{(k)} \equiv \delta$. However, if there is a shift in δ at study k_1+1, $\delta_{(k)}$ becomes a weighted average of δ and $\delta+\Delta$ values.

Given a set of weights w_i, $i=1,\cdots,K$, and denoting the running sum of weights $W_k=\sum_1^k w_i$, the cumulative mean effect for $k \le k_1$ studies is δ, and for $k > k_1$ studies it is

$$\delta_{(k)} = \delta + \Delta(W_k - W_{k_1})/W_k. \tag{10}$$

Consider, for simplicity, that the sample sizes are equal between arms and among studies, $n_{iC}=n_{iT}\equiv n/2$, so that effective sample sizes are $\widetilde{n}_i \equiv n/4$. Thus, the SSW estimator uses equal weights $w_i = n/4$, and the cumulative mean $\delta_{\mathrm{SSW}(k)} = \delta + \Delta(1-k_1/k)$ increases from δ at $k=k_1$ to $\delta_1+\Delta(1-k_1/K)$, reaching $\delta+\Delta$ only at $K=\infty$.

The within-study variances of Hedges's g are $\mathrm{Var}(g_i) = 4/n + \delta_i^2/2n$. Under the scenario of no shift in δ or τ^2, the population IV weights should be equal. Given a positive shift in the mean from δ to $\delta+\Delta$, the variances of the g_i, $i \ge k_1+1$ also increase, and the weights decrease accordingly. Therefore, with IV weights the cumulative mean effect (10) will be reduced, compared with SSW. This may reduce the power of the CMA based on the IV weights. Figure S1 in Appendix S2 illustrates changes in $\delta_{\mathrm{SSW}(k)}$ and in $\delta_{\mathrm{IV}(k)}$ under the scenario of equal sample sizes. The difference between the two cumulative mean effects is the largest when $\tau^2=0$; but it decreases in τ^2, and for $\tau^2>0$ it also decreases rapidly with n and is negligible at $n=100$.

However, the estimated variances v_i^2 given by Equation (2) and, therefore, the IV weights $\hat{w}_i$ are random variables and are not exactly equal. Additionally, these random weights are not independent from the estimated effects g_i, resulting in order $1/n$ bias of the estimated cumulative mean effect $\hat{\delta}_{\mathrm{IV}(k)}$, even under the no-shift scenario, as we demonstrate in Section 4.4.

3.2 | Estimation of τ^2

To understand the pattern of changes in the estimated value of τ^2 in CMA, denoted by $\hat{\tau}^2_{(k)}$, consider the running value of Cochran's Q statistic, $Q_{(k)} = \sum_{i=1}^k w_i\left(\hat{\delta}_i - \hat{\delta}_{(k)}\right)^2$ for IV weights $w_i = v_i^{-2}$. A number of estimators of τ^2, including KDB, are based on the first moment of Q. Consider the same simple scenario of equal sample sizes and a shift in the mean from δ to $\delta+\Delta$ at study k_1+1 given by (9).

Let $\xi_i = \hat{\delta}_i - \Delta I(i>k_1)$. The variables ξ_i coincide with $\hat{\delta}_i$ under the hypothesis of no shift. Write $\hat{\delta}_i = \xi_i + \Delta I(i>k_1)$, and $\hat{\delta}_{(k)} = \overline{\xi}_{(k)} + \Delta(W_k - W_{k_1})/W_k$ for $\overline{\xi}_{(k)} = \sum_1^k w_i\xi_i / \sum_1^k w_i$. The running value of the Q statistic is $Q_{(k)} = \sum_{i=1}^k w_i\left(\xi_i + \Delta I(i>k_1) - \overline{\xi}_{(k)} - \Delta(W_k - W_{k_1})/W_k\right)^2$.

It follows that

$$Q_{(k)} = \sum_{i=1}^k w_i\left(\xi_i - \overline{\xi}_{(k)}\right)^2 + 2\Delta W_{k_1}\left(\overline{\xi}_{(k)} - \overline{\xi}_{(k_1)}\right) + \Delta^2 W_{k_1}(W_k - W_{k1})/W_k,$$

The first term of $Q_{(k)}$ accounts for between-study heterogeneity of ξ_i. The mean of the second term is zero, and we expect it to be reasonably small. And the third term reflects the contribution of the shift Δ. Under no shift, only the first term is nonzero.

The moment-based estimators of τ^2 described in DerSimonian and Kacker,[48] which include the popular DerSimonian–Laird estimator,[49] and the recent SDL estimator[37] have the form $\hat{\tau}^2_M = \left(Q_{(k)} - E_0(Q)\right)/\left(W_k - W_k^{(2)}/W_k\right)$ where $E_0(Q)$ is the expected value of Q under homogeneity (i.e., when $\tau^2=0$), the denominator is the multiplier at τ^2 in the Taylor-series expansion of $E\left(Q_{(k)}\right)$, and $W_k^{(2)} = \sum_1^k w_i^2$.

Hence,

$$\hat{\tau}^2_M \approx \hat{\tau}^2_\xi + \Delta^2 \frac{W_{k_1}(W_k - W_{k1})}{\left(W_k^2 - W_k^{(2)}\right)}, \tag{11}$$

where $\hat{\tau}^2_\xi$ is the estimator of the heterogeneity variance under no shift. For the case of equal sample sizes, the weights are $w_i = n/4$, $W_k = kn/4$ and $W_k^{(2)} = kn^2/4^2$, so $\hat{\tau}^2_M \approx \hat{\tau}^2_\xi + \Delta^2\frac{k_1(k-k_1)}{k(k-1)}$. At $k=k_1+1$, $\hat{\tau}^2_M \approx \hat{\tau}^2_\xi + \Delta^2/(k_1+1)$, and then it increases in k almost linearly. This increase in estimated τ^2 does not depend on the sample size.

In our simulation study, we do not use $\hat{\tau}^2_M$, but both the KDB and the popular Paule–Mandel estimator[50] are also based on the first moment of $Q_{(k)}$, and we expect similar behaviour from these estimators. REML is not a moment estimator, but our simulation results, discussed in Section 4.3, demonstrate this behaviour for both KDB and REML (Figure 1).

3.3 | Testing for a shift in the mean in CMA

Consider consecutive testing for a shift at $k=1,\ldots,K$ in the model (9) based on the cumulative estimates $\left(\hat{\delta}_{(k)},\ \hat{\tau}^2_{(k)}\right)$ from K studies ordered over time. At each k, the standard CMA tests $H_0: \delta_{(k)} = \delta_0$ against $\delta_{(k)} \ne \delta_0$ for a known value of δ_0, at the same level α. We consider two types of tests, described in Section 2.4, IV REML and SSW KDB. These test statistics have the form $\hat{\mathrm{Var}}\left(\hat{\delta}_{(k)}\right)^{-1/2}\left(\hat{\delta}_{(k)} - \delta_0\right)$. The variance $\mathrm{Var}\left(\hat{\delta}_{(k)}\right)$ is of

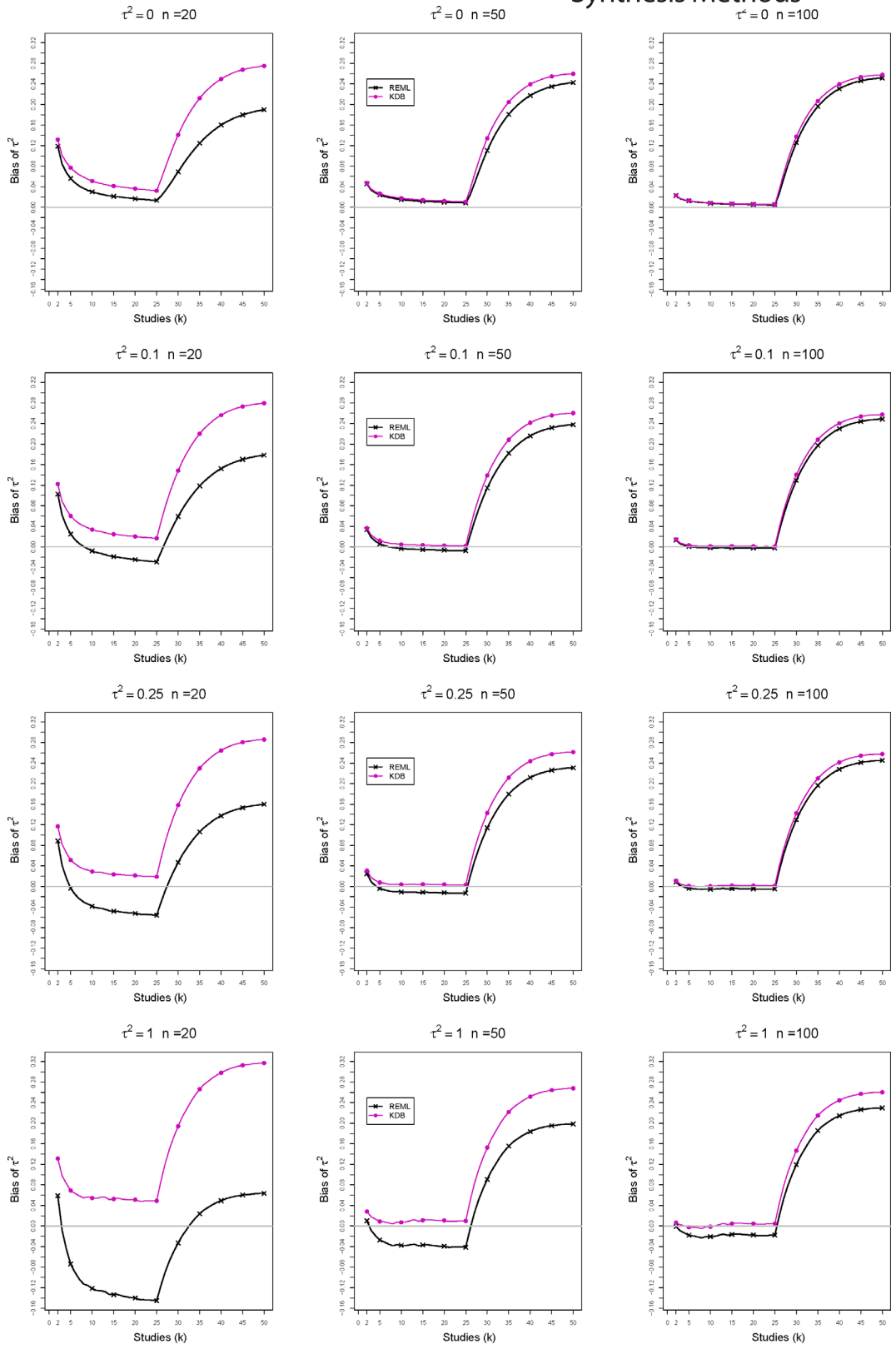

FIGURE 1 Bias of REML and KDB estimators of between-studies variance τ^2 when $\delta = 1$ for $k \leq 25$ followed by a shift to $\delta = 2$ for $k \geq 26$, $n = 20$, 50 and 100, $K = 50$ and $\tau^2 = 0,\ 0.1,\ 0.25$ and 1. Light grey line at 0 [Colour figure can be viewed at wileyonlinelibrary.com]

order $1/N_{(k)} + \tau^2_{(k)}/k$, where $N_{(k)}$ is the total sample size in the k studies. This is easier to see from Equation (8) for the variance of $\hat{\delta}_{\mathrm{SSW}}$, but it is also true for $\hat{\delta}_{\mathrm{IV\,REML}}$. Therefore, for sufficiently large sample sizes, these statistics are approximately equivalent to the ratios $\sqrt{k}\left(\hat{\delta}_{(k)} - \delta_0\right)/\hat{\tau}_{(k)}$. The power of these tests depends primarily on the value of $\sqrt{k}\left(\delta_{(k)} - \delta_0\right)/\tau_{(k)}$, but it may be distorted by biases of $\hat{\delta}_{(k)}$ and $\hat{\tau}_{(k)}$.

Visually, the CMA provides a forest plot of the $\hat{\delta}_{(k)}$ values with their confidence intervals for increasing k, and a δ_0 outside the kth confidence interval is rejected by the kth test. An example of these CMA plots is provided in Row 2 of Figure 7; see Section 5 for further details.

As discussed in Section 3.2, the cumulative estimate $\hat{\tau}^2_{(k)}$ increases with the shift in the mean. Because of this, we also consider two tests, QP and KDB, of $\tau^2_{(k)} \le \tau^2_0$ versus $\tau^2_{(k)} > \tau^2_0$ for a known value of τ^2_0 as possible tests for shift. These tests are also performed consecutively for increasing k. However, based on (11) these tests, described in Section 2.3, are likely to be more powerful for early shifts (small k). Similar to the standard CMA tests for $\hat{\delta}_{(k)}$, these tests are easily performed visually on a forest plot of cumulative $\hat{\tau}^2_{(k)}$ values. Examples of these plots appear in the first row of Figure 7.

From the findings in Section 3.2, it also follows that for a test of a shift based on $\hat{\delta}_{(k)}$, the accompanying increase in $\hat{\tau}^2_{(k)}$ may decrease its power. It would make sense to use a sufficiently well-estimated value of τ^2 before this increase occurs. Because of this, we consider the following modification to traditional CMA, which we refer to as the two-stage CMA.

Borrowing the quality control concept of two-stage monitoring, in Stage 1 we aim to estimate τ^2. Therefore, we test for $\delta_{(k)} = \delta_0$ for $k = 5, ..., 10$ using IV REML or SSW KDB. If that test rejects at $k_1 + 1$, we take $\hat{\tau}^2_{k_1}$ (estimated by REML or KDB, respectively) as the 'true' value τ^2_0; otherwise we take $\tau^2_0 = \hat{\tau}^2_{(10)}$. The rationale for this choice of $5 \le k \le 10$ studies for estimating τ^2 in Stage 1 is that an estimate of τ^2 is not sufficiently precise for $k < 5$, and is reasonably accurate by $k = 10$ under no shift (see Section 4.3). However, we do not want to use an inflated value of τ^2 if an early shift does occur. In Stage 2, we monitor the (known) mean effect δ_0, using the τ^2_0 value estimated in Stage 1 as the known between-study variance $\tau^2_{(k)}$ without re-estimating it. We refer to the respective tests and estimators as IV REML (τ^2_0) and SSW KDB (τ^2_0).

We also consider a more realistic scenario of monitoring the estimated value of δ (estimated in Stage 1 as described above at $k_1 + 1$ or at $k = 10$) instead of the 'known' δ_0. We refer to these tests as IV REML ($\hat{\delta}_0$) and SSW KDB ($\hat{\delta}_0$) when further testing proceeds as in the standard CMA (i.e., with τ^2 re-estimated each time), and IV REML ($\hat{\delta}_0$, τ^2_0) and SSW KDB ($\hat{\delta}_0$, τ^2_0) when also using the value of τ^2_0 estimated in Stage 1.

For comparison, we also use two-stage CUSUM charts with $h = 4$, 5 and 6 to monitor the known mean δ_0 for the effects $\hat{\delta}_i$ with variances $v_i^2 + \tau^2_0$. Table 1 gives a comprehensive summary of methods and abbreviations.

TABLE 1 Methods and abbreviations

Weights in estimation of δ	IV	Inverse-variance weights $w_i = 1/v_i^2$
	SSW	Effective sample size weights $\tilde{n} = n_C n_T / n$
Point estimators of τ^2	REML	Restricted maximum-likelihood estimator
	KDB	Kulinskaya–Dollinger–Bjørkestøl estimator[36,37]
Point estimators of δ	IV REML	IV weights with REML-estimated τ^2
	SSW KDB	SSW weights, KDB-estimated τ^2 in variance calculation
Interval estimators of τ^2	QP	Q-profile interval based on χ^2_{K-1} distribution of Q[41]
	KDB	Q-profile interval based on improved approximation[36]
Critical values in intervals and tests	IV REML	Standard normal distribution
	SSW KDB	t-Distribution with $K - 1$ d.f.
Methods of CMA	Standard CMA	$\hat{\delta}_{(k)}$ and $\hat{\tau}^2_{(k)}$ are re-estimated at each $k = 2, ..., K$; known δ_0 value is used for testing $H_0 : \delta_{(k)} = \delta_0$ at each k
	Two-stage CMA (τ^2_0)	$\tau_0 = \hat{\tau}^2$ is estimated in Stage 1 from 10 or fewer studies; this τ^2_0 value is constant in Stage 2; δ_0 is known
	Standard CMA ($\hat{\delta}$)	$\hat{\delta}_{(k)}$ and $\hat{\tau}^2_{(k)}$ are re-estimated at each $k = 2, ..., K$; $\hat{\delta}_0$ estimated in Stage 1 used for testing $H_0 : \delta_{(k)} = \hat{\delta}_0$ in Stage 2
	Two-stage CMA ($\hat{\delta}_0$, τ^2_0)	$\hat{\delta}_0$ and $\tau_0 = \hat{\tau}^2$ both estimated in Stage 1 are used for testing $H_0 : \delta_{(k)} = \hat{\delta}_0$ in Stage 2
CUSUM analysis	Two-stage CUSUM (τ^2_0)	$\tau_0 = \hat{\tau}^2$ is estimated in Stage 1 from 10 or fewer studies; this τ^2_0 value is constant in stage 2; δ_0 is known

4 | SIMULATION STUDY

4.1 | Simulation design

In the majority of our simulations we use $K = 50$ as the maximum number of studies, though in some scenarios we also use $K = 100$ and 1000. For each combination of parameters, we use equal sample sizes from $n = 20$ to 1000 in all K studies. The sample sizes in the treatment and control arms are equal.

We use a total of 10,000 repetitions for each combination of parameters. Thus, the simulation standard error for estimated coverage of δ or τ^2 at the 95% confidence level, or testing at 0.05 level is roughly $\sqrt{0.95 \times 0.05/10{,}000} = 0.00218$. The simulations were programmed in R version 3.3.2.

As the number of studies increases in CMA, with $k \le K$, we examine bias and coverage in estimation of δ and τ^2, and the accumulating rejection rates (Type 1 error or power) of tests for the shift in the mean effect and in the between-study variance τ^2. We also consider the cumulative signalling rate and the ARL in CUSUM analysis.[44] We consider both a constant value of δ (the null hypothesis of no shift in the mean) and a shift from $\delta = 1$ to $\delta = 2$ at 0.25, 0.5 and 0.75 of the total number of studies. Our summaries of results in Sections 4.3–4.7 include illustrative figures and are based on examination of the figures in our ePrint.[38] We vary five parameters: the overall true SMD (δ) and the between-studies variance (τ^2), in addition to the maximum number of studies (K), the point of shift (if any) (k_1), and the studies' total sample size (n). Table 1 lists the values of each parameter.

We generate the true effect sizes δ_i from a normal distribution: $\delta_i \sim N(\delta, \tau^2)$. We generate the values of Hedges's estimator g_i directly from the appropriately scaled non-central t-distribution, given by Equation (3), and obtain their estimated within-study variances from Equation (2).

We study two approaches to point and interval estimation and testing of τ^2 (REML/QP and KDB) and two resulting approaches to point and interval estimation and testing of δ (IV REML and SSW KDB). Each of these two approaches was studied in the four CMA settings listed in Table 1: traditional CMA setting with known δ_0 value, CMA with estimated in Stage 1 value of δ_0, the two-stage CMA with known value of δ_0, and the two-stage CMA with the estimated in Stage 1 value of δ_0 (Table 2).

4.2 | Outcomes recorded in simulations

In all simulations, we assumed the shift-in-the-mean model (9) and, for the CMA methods of interest, we studied the bias of the point estimators $\hat{\delta}_{(k)}$ of the cumulative mean $\delta_{(k)}$ (10) and (for the standard CMA) the bias of $\hat{\tau}^2_{(k)}$ in estimating τ^2 for $5 \le k \le K$. We also investigated coverage of the $\delta_{(k)}$ and of τ^2 by the relevant interval estimators and empirical levels of the accompanying two-sided tests for the null hypothesis of no shift in δ and (separately) of one-sided tests of no shift in τ^2. We also investigated cumulative Type I errors of these tests at the 0.05, 0.01 and 0.005 levels for δ and at the 0.025, 0.005 and 0.0025 levels for one-sided tests for τ^2.

TABLE 2 Data patterns in simulations

Parameter	Input values
δ (true value of the SMD)	0, 0.5, 1
Δ (shift in δ)	± 0.5 and ± 1, both for $\delta = 1$
τ^2 (variance of random effect)	0, 0.1, 0.25, 1
K (maximum number of studies)	50, 1000 (two-stage CMA)
k_1 (point of shift in δ)	$i\lceil K/4 \rceil$, $i = 1,\ 2,\ 3$ for shift from 1 to 2; $K = 50$ and $k_1 = 26$ for shifts from 1 to 1.5, 1 to 0.5, and 1 to 0
n (sample size: total of the two arms)	20, 50, 100, 500, 1000 (shift in the mean) 20, 50, 100, 500 (no shift in the mean)
α (two-sided significance level)	0.05, 0.01, 0.005
M (number of repetitions)	10,000 1000 (two-stage CMA with $K = 1000$)

4.3 | Bias of $\hat{\tau}^2_{(k)}$

When there is no shift in δ, both estimators of τ^2 (REML and KDB) have non-negligible positive bias for small k, especially for small sample sizes ($n \le 50$), Figure 1. KDB retains small positive bias for larger values of k, whereas the bias of REML becomes negative when $\tau^2 > 0$. Biases do not depend visibly on the value of δ, but they increase in k and increase considerably in τ^2. The bias of REML is about -0.04, and the bias of KDB is $+0.04$ when $n = 20$, $\tau^2 = 0.25$ and $k = 10$, compared with -0.10 and $+0.07$ when $\tau^2 = 1$. The biases decrease in n; when $n = 100$, KDB is practically unbiased, and REML has small negative bias of about 2%.

From the point of a shift, both estimators of τ^2 increase rapidly, KDB somewhat faster than REML. However, for larger n, the behaviour of REML converges

to that of KDB, and the difference between the estimators is negligible at $n = 100$.

4.4 | Bias and coverage of $\hat{\delta}_{(k)}$

The cumulative effect estimated by SSW is almost unbiased under all simulated conditions, regardless of the value or a shift in δ. In Figure 2, SSW coincides with its expected value, given by Equation (10). IV REML is also unbiased when $\delta = 0$ (not shown). However, IV REML has a small negative bias, up to about 5%–7%, when $n = 20$ and $\delta = 1$, Figure 2. The bias increases in δ and in k. It also increases, though rather slowly, in τ^2. The bias decreases for larger sample sizes; when $n = 100$ and $\delta = 1$, the bias is about 1.5%. After the shift in δ, IV REML is somewhat lower than SSW, and it deviates from its nominal mean (10), but these differences decrease in sample size and are practically eliminated by $n = 100$.

As illustrated by Figure S2 in Appendix S2, coverage of SSW KDB is rather conservative (i.e. above nominal) for small numbers of studies, but it improves for larger values of k and τ^2. When $n = 20$ and $k \le 5$, coverage of IV REML is somewhat conservative when $\tau^2 = 0$ and $\delta = 0$, but it drops below nominal for larger k when $\delta = 1$. For larger sample sizes, IV REML provides stable, if somewhat conservative, coverage when $\tau^2 = 0$. When $\tau^2 > 0$, IV REML has very low coverage for $k \le 10$, and it does not improve much in n. Coverage at the nominal 95% level is about 85%–90% when $k = 20$ and $\tau^2 \ge 0.25$, and it remains below nominal when $n = 100$. Coverage is visibly reduced for $\delta > 0$. As we shall see in the next section, this liberal coverage translates into higher Type 1 error in CMA.

4.5 | Level and power of tests for δ in CMA

Because of multiple testing over the increasing number of studies k, the empirical levels of SSW KDB and IV REML at the same nominal level are increasing in k, but the empirical levels of SSW KDB are considerably lower. The difference between the two methods is more pronounced for larger values of δ; see Appendix S2, Figures S3 and S4 for $\delta = 0$ and $\delta = 0.5$ up to $K = 50$ and Figure S5 for $\delta = 1$ and K up to 1000. Tables S1–S3 in Appendix S1 provide empirical levels at selected values of k at the nominal 0.05, 0.01 and 0.005 levels for $\delta = 0$, 0.5 and 1. As an example, at the nominal 0.05 level, these levels for $\delta = 1$ are 0.048 for SSW KDB versus 0.118 for IV REML at $k = 12$ and 0.089 versus 0.187 at $k = 25$ for $n = 20$ and $\tau^2 = 0$. These levels increase further in τ^2 (0.079 vs. 0.177 at $k = 12$ for $n = 20$ and $\tau^2 = 0.1$); and they increase somewhat in n, in this example to 0.150 versus 0.253 at $k = 12$ for $n = 1000$ and $\tau^2 = 0.1$. Testing at the lower nominal levels makes sense for larger values of k, τ^2 and/or n, see Tables S1–S3 for some guidance.

The power of SSW KDB and IV REML is comparatively low. Figure 3 shows empirical levels for shift from 1 to 2 at Study 26. For both methods, the power is highest when $\tau^2 = 0$ and deteriorates considerably in τ^2. Taking into account its lower level, SSW KDB is more powerful than IV REML. The power increases in n, and by $n = 100$ both methods reach power 80% at 31 studies when $\tau^2 = 0$ and at 36 studies for $\tau^2 = 0.25$. Choosing the nominal level of 0.01 safeguards the empirical levels about 0.05 at $k = 25$ and reduces the power of CMA accordingly. Table S6 provides the number of studies needed to reach power of 80% and 90% for a shift from $\delta = 1$ to $\delta = 2$ at $k = 13$, 26 and 38. As expected, the power of all tests is lower at smaller shifts, but the direction of shift does not seem to matter (Figures S9–S20 in Appendix S2).

4.6 | Level and power of tests for τ^2 in CMA

We studied one-sided tests for $\tau^2_{(k)} > \tau^2_0$, and typical results for $\Delta = 1$ at $k = 26$ are depicted in Figure 4 for nominal levels 0.025, 0.005 and 0.0025. Multiple testing inflates empirical levels, more so for KDB than for QP. Table S4 in Appendix S1 provides empirical levels at selected values of k at the nominal 0.025, 0.005 and 0.0025 levels for $\delta = 1$. The power increases in n and decreases in τ^2. When $\tau^2 = 0$, the power is quite high from $n = 50$; but when $\tau^2 = 0.25$, the power reaches 80% for both tests only at $k = 41$ for $n = 100$. Power is extremely low when $\tau^2 = 1$, even for very large sample sizes; for shift at $k = 26$, power barely reaches 30% at $k = 50$ when $n = 1000$ (not shown).

4.7 | Comparing tests for shift in δ in one- and two-stage CMA

When there is no shift in δ, two-stage CMA and the standard one-stage CMA have very similar inflation of the empirical levels. However, two-stage CMA is somewhat more powerful under the shift. This difference in power is clear for KDB SSW (τ^2_0) from $n = 20$, and for IV REML (τ^2_0) from $n = 50$, Figure 5 and Figure S6. This difference in power is explained by inflation in the estimated $\tau^2_{(k)}$ in the standard CMA, as discussed in Section 3.3.

For comparison, Figure 5 and Figures S6–S20 in Appendix S2 also include CUSUM-based CMA with $h = 4$ and 5 along with the CMA tests at the 0.05 and 0.01

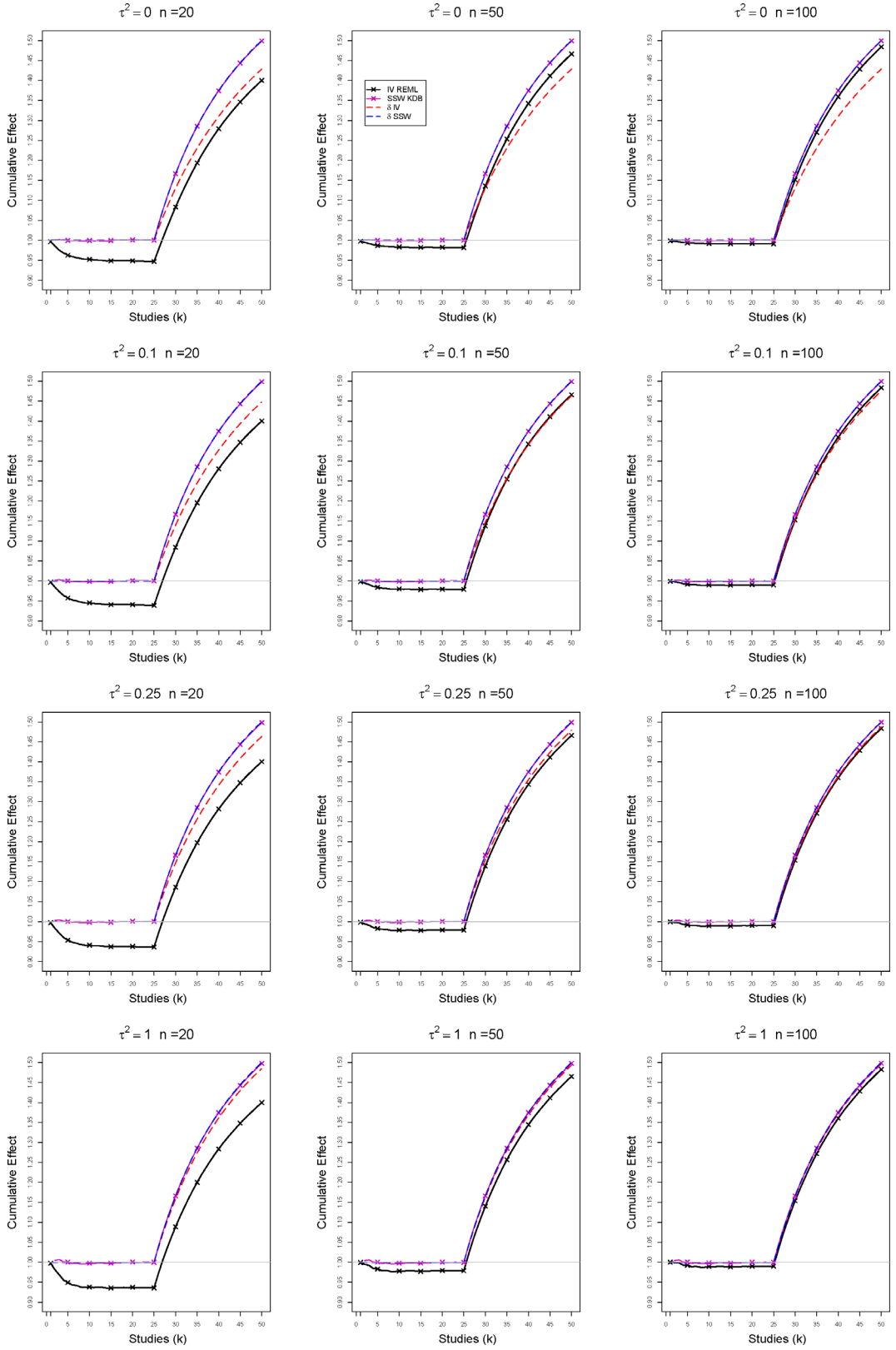

FIGURE 2 Weighted cumulative effects $\delta_{SSW(k)}$ and $\delta_{IV(k)}$ given by Equation (10) (dashed lines) and estimated by SSW and IV REML cumulative effects $\hat{\delta}_{(k)}$ (solid lines) when $\delta = 1$ for $k \leq 25$ followed by a shift to $\delta = 2$ for $k \geq 26$, $n = 20$, 50 and 100, $K = 50$ and $\tau^2 = 0,\ 0.1,\ 0.25$ and 1. Light grey line at 1 [Colour figure can be viewed at wileyonlinelibrary.com]

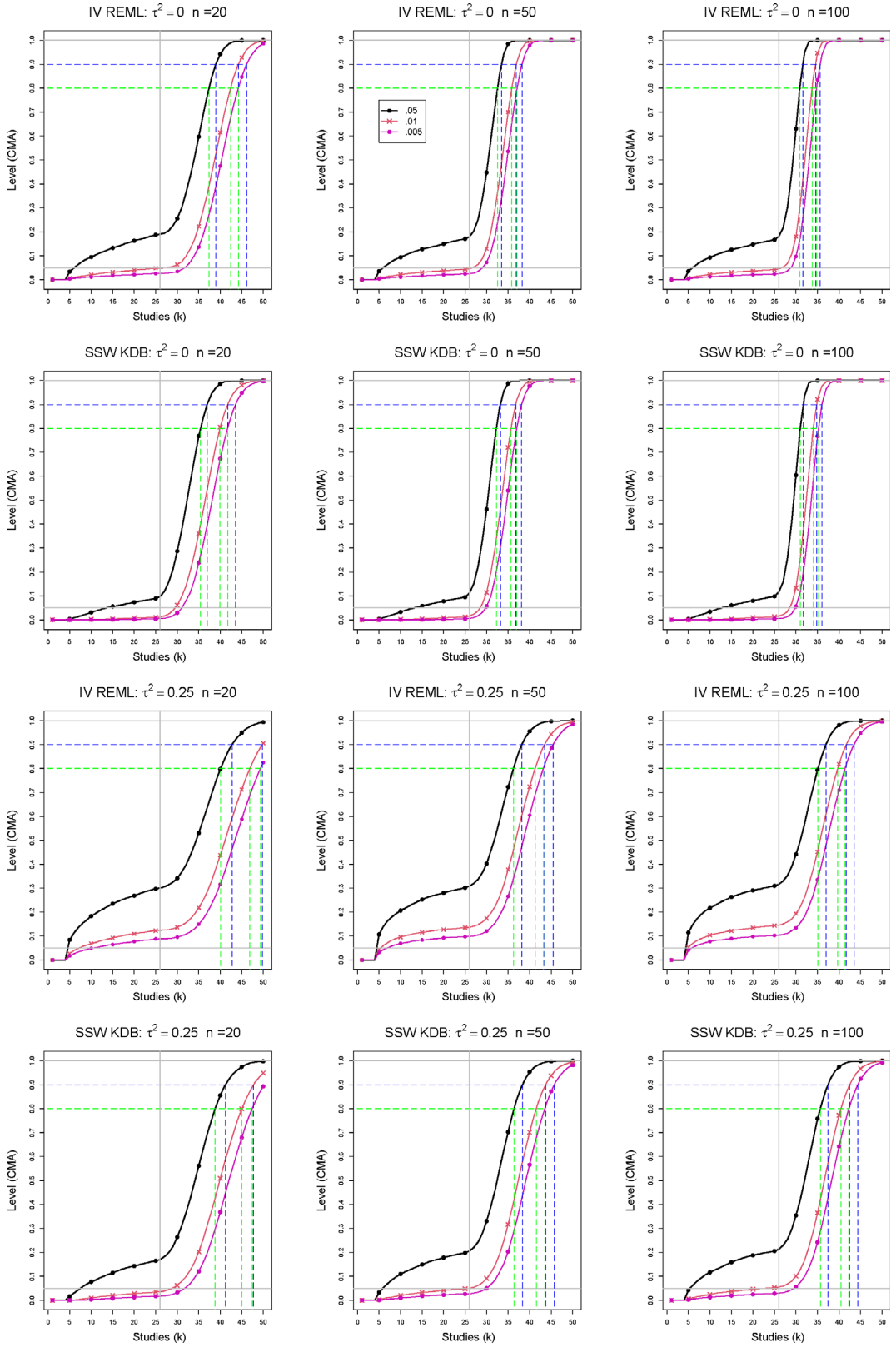

FIGURE 3 Empirical levels of CMA tests for shift in δ based on SSW KDB and IV REML at nominal levels 0.05, 0.01 and 0.005 for equal sample sizes $n_{iC} + n_{iT} = n = 20,\ 50$ and 100, $\tau^2 = 0$ and 0.25 and a shift from $\delta = 1$ to $\delta = 2$ at Study 26. Light grey line at 0.05. Green and blue dashed lines correspond to power 80% and 90%, respectively [Colour figure can be viewed at wileyonlinelibrary.com]

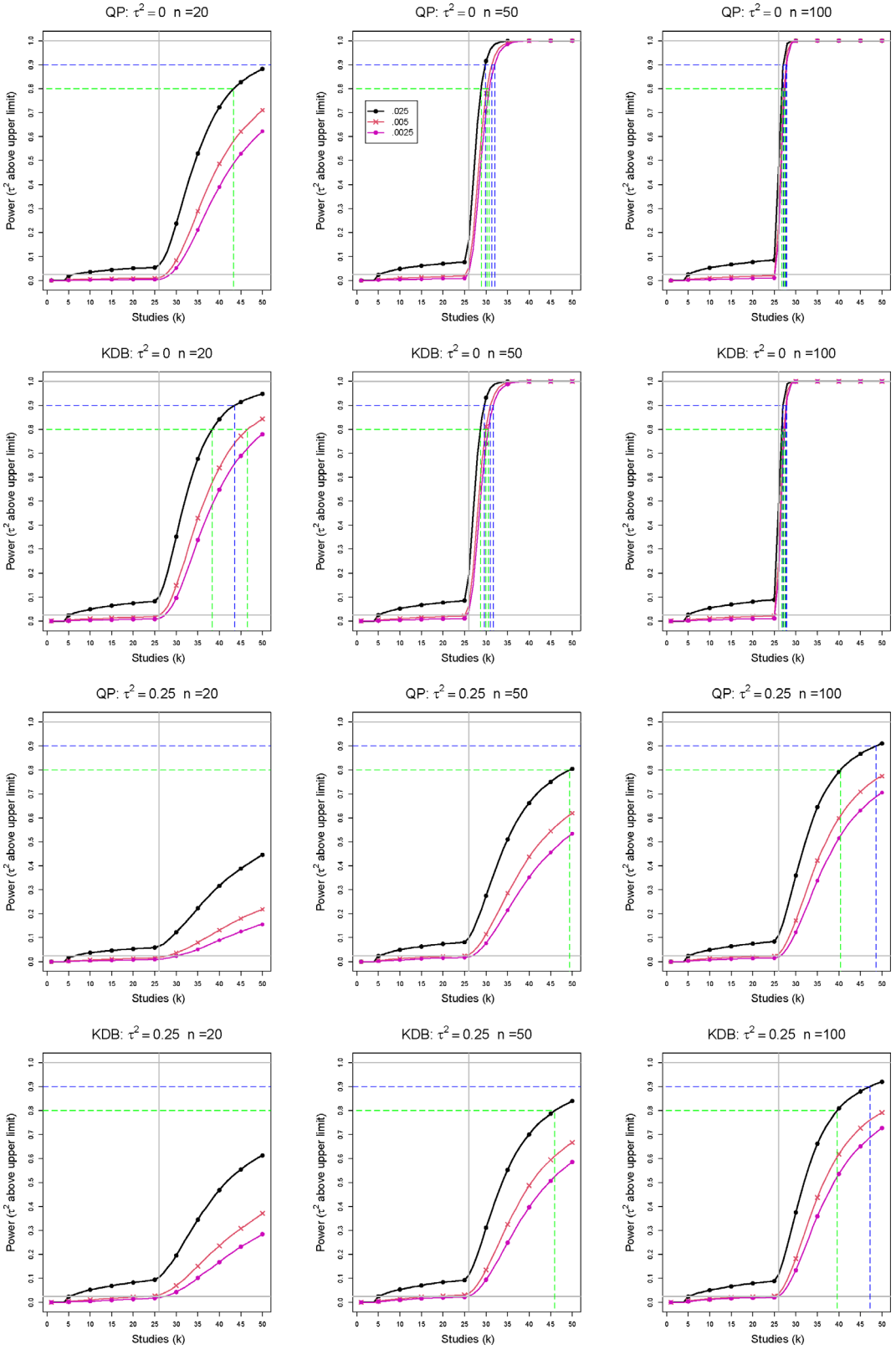

FIGURE 4 Empirical levels of CMA tests for positive shift in τ^2 based on KDB and QP at nominal levels 0.025, 0.005 and 0.0025 for equal sample sizes $n_{iC} + n_{iT} = n = 20,\ 50$ and 100, $\tau^2 = 0$ and 0.25 and a shift from $\delta = 1$ to $\delta = 2$ at Study 26. Light grey line at 0.025. Green and blue dashed lines correspond to power 80% and 90%, respectively [Colour figure can be viewed at wileyonlinelibrary.com]

nominal levels, respectively. Under no shift, CUSUM-based analysis results in greater inflation of the empirical levels, but equally, it has more power under shift. Similarly to other tests, its power increases in n and decreases in τ^2.

Figure 5 and Figures S6–S20 also include the two-stage methods using an estimated in Stage 1 mean effect $\hat{\delta}_0$. In this scenario, the 'standard' CMA methods, which re-estimate $\tau^2_{(k)}$, have especially inflated levels, and the two-stage methods, which use the estimated τ^2_0, are clearly the better choice. Unexpectedly, methods that use two parameters estimated at Stage 1 $(\hat{\delta}_0, \tau^2_0)$ have somewhat lower Type 1 error and somewhat more power than the comparative methods using known δ_0, especially for IV REML.

Table S5 in Appendix S1 provides empirical levels of two-stage CUSUM analysis of k studies with $h = 4,\ 5$ and 6 when $\delta = 1$. Table S7 provides empirical levels of two-stage CMA of k studies at nominal levels 0.05, 0.01 and 0.005 when $\delta = 1$, and Table S8 gives the number of studies (k) required for 80%/90% power for detecting a shift from $\delta = 1$ to $\delta = 2$.

5 | EXAMPLE

As an example, we use data by Batáry et al.[51] on the role of agri-environment management schemes in conservation and environmental management. We use the data on species richness from 39 studies published from 1992 to 2010, with mostly small to medium sample sizes, ranging from 2 to 37 per arm, though one study has 152 observations in each arm. The effect measure is SMD, and positive values correspond to higher species richness in the extensive (organic) than in the intensive (conventional) fields. The majority of the studies originated from European countries and compared conventional with organic management. The original meta-analysis did not take chronology into account. Observations of multiple taxa and/or of different geographical regions in an individual study were included separately in the dataset, resulting in 109 records in total. We chose a single sub-study with median value of $\hat{\delta}$ from each study. Figure 6 provides the raw data and forest plot, depicting the 39 sub-study effects with corresponding 95% confidence intervals.

Visual examination of the forest plot shows that the effects in the first 18 studies seem to hover somewhat above zero, and Study 19 is a high outlier. The next subset of effects (Studies 20–33) is somewhat more positive, and Study 33 is another high outlier. The last subset of effects (Studies 34 to 39) seems to drift back toward zero. These observations are clearly confirmed by the plots of cumulative τ^2 values in the top row of Figure 7, using both QP and KDB confidence intervals. Heterogeneity is very high in the first eight studies, but then it settles down and is relatively low up to Study 18. It jumps at Study 19, but then decreases from Study 20–33, indicating that Study 19 is just an outlier and not the start of a shift. The same happens at Study 33. In this example, the KDB values of $\hat{\tau}^2$ are higher than the REML values.

In the second row of Figure 7, IV REML provides higher estimates of $\hat{\delta}_{(k)}$ and narrower confidence intervals than KDB SSW. As we expected from the simulations, IV REML is less conservative and shows a significantly positive effect $\hat{\delta}_{(8),\,\text{IV REML}} = 0.959\ (0.014, 1.905)$ at the 0.01 level at Study 8 (discounting Study 1), compared with Study 36 for SSW KDB (with $\hat{\delta}_{(36),\,\text{SSW}\,KDB} = 0.952\ (0.005,\ 1.898)$). Both methods show wider confidence intervals because of increased τ^2 values at the outlier Study 19.

In two-stage CMA, the values $\hat{\tau}^2_{\text{REML},(7)} = 1.04$ and $\hat{\tau}^2_{\text{KDB},(10)} = 0.836$ are used in two-stage IV REML and KDB SSW, respectively. The value of $\hat{\delta}_{(10),\,\text{SSW}} = 0.664$. The KDB test for τ^2 shows significant increase in τ^2 at Study 19, compared with Study 10. The confidence intervals for $\hat{\delta}_{(k)}$ are somewhat wider for two-stage CMA than for standard CMA at the start, and somewhat lower by the end, for both methods. Two-stage SSW KDB results in a significant effect at Study 25, where $\hat{\delta}_{(25),\text{SSW}\,KDB} = 0.872(0.022,\ 1.721)$, considerably faster than the one-stage SSW KDB CMA.

In the CUSUM plots, the first 7 points (for REML) and first10 points (for KDB) are obtained with changing, cumulative values of $\hat{\tau}^2_{(k)}$, but thereafter use the fixed values throughout. The CUSUM based on the value of $\hat{\tau}^2_{\text{REML},(7)} = 1.04$ does not reach significance, but the CUSUM based on the smaller value of $\hat{\tau}^2_{\text{KDB},(10)} = 0.836$ does at Study 19, and both CUSUMs quickly react to any changes in effects.

In this example, which involves no significant shifts in effects, SSW KDB CMA appears to be too conservative. However, as our simulations demonstrate, this method would result is a much lower false-positive rate. Different methods provide complementary information, and in practice we therefore recommend the use of multiple plots, including the forest plot, the CMA plot for τ^2 and the two-stage CMA plot for δ.

6 | DISCUSSION: PRACTICAL IMPLICATIONS FOR CMA

Cumulative meta-analysis is a well-established and popular method of evaluating and monitoring accumulating evidence. This method is especially widely used in health and environmental applications where multiple

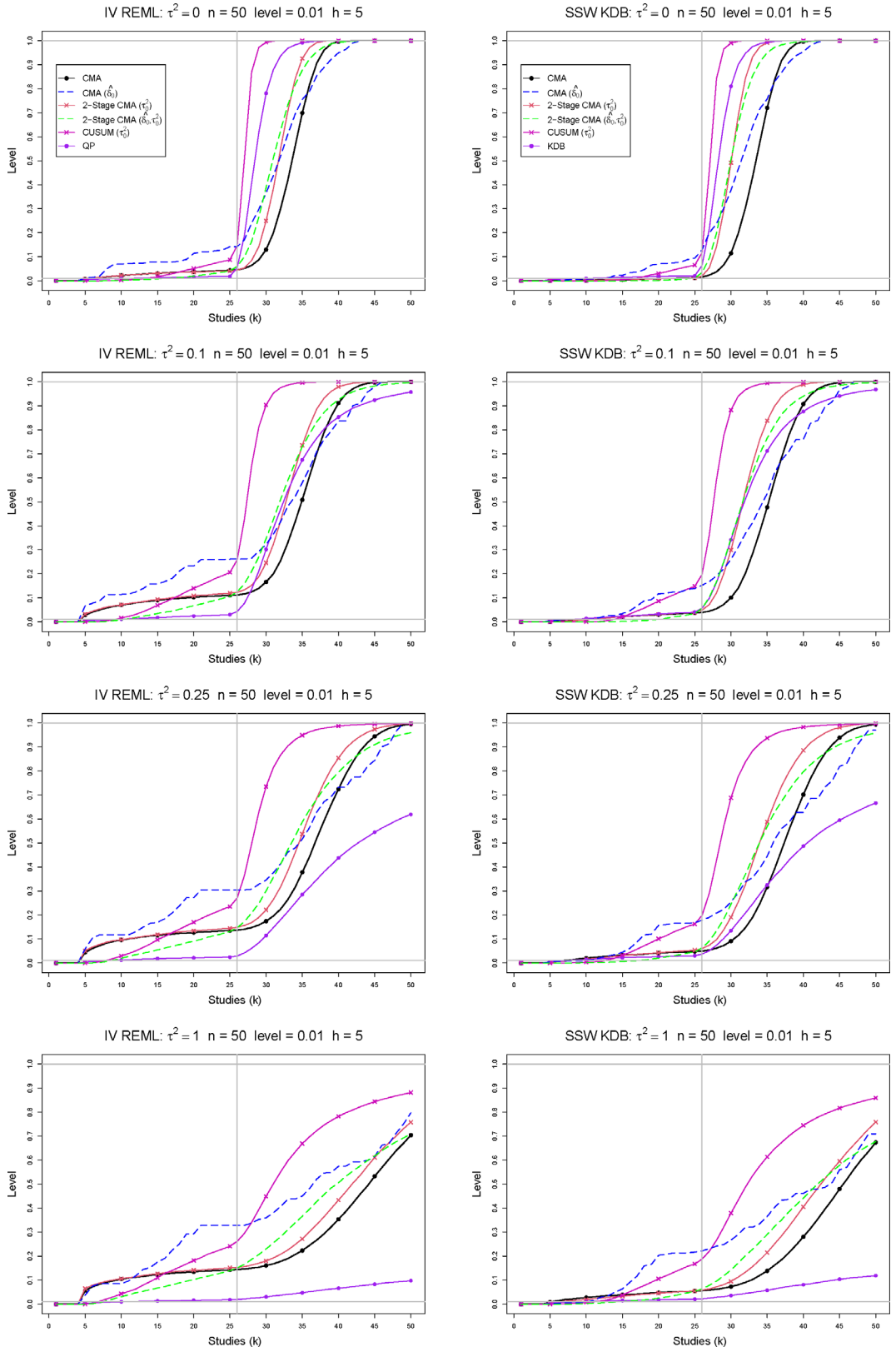

FIGURE 5 Empirical levels of one- and two-stage CMA tests for shift in δ at nominal level 0.01, shift in τ^2 at nominal levels 0.005 and of CUSUM with $h = 5$ for equal sample sizes $n_{iC} + n_{iT} = n = 50$, $\tau^2 = 0,\ 0.1, 0.25$ and 0.1 and a shift from $\delta = 1$ to $\delta = 2$ at Study 26. Light grey line at 0.01 [Colour figure can be viewed at wileyonlinelibrary.com]

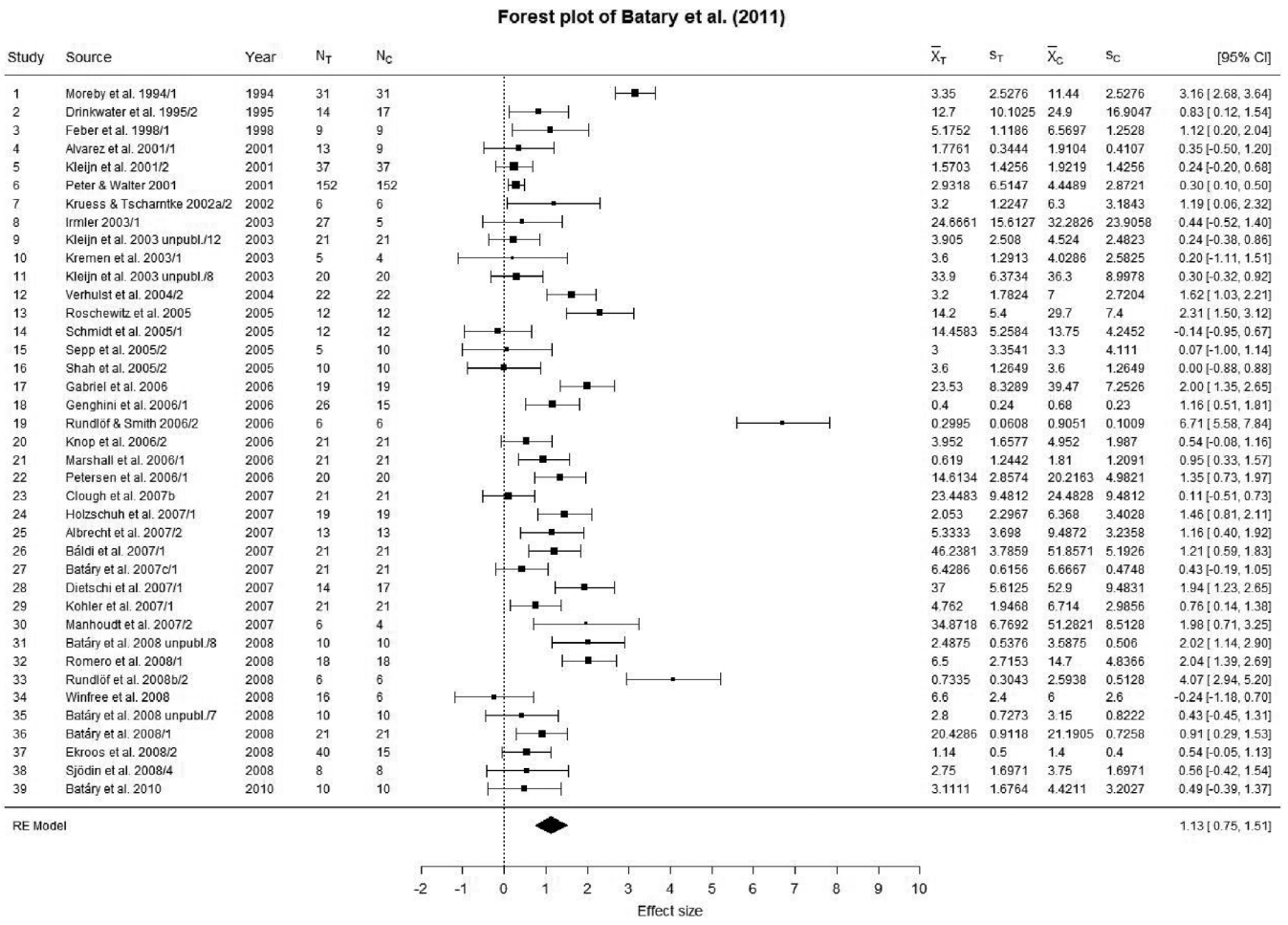

FIGURE 6 Forest plot for species richness data from Batáry et al. (2011) with 95% confidence interval. The effect measure is SMD, and positive values correspond to higher species richness in the intervention arm

publications on the same topic are available over a number of years. The multiplicity problems inherent in CMA are well known, and a number of alternative statistical methods aimed at resolving these problems are available. However, this does not seem to hinder the popularity of CMA in applied research.

Therefore we investigated, theoretically and by simulation, the level and power of CMA and how to improve both. For the popular effect measure SMD, we compared two approaches for CMA: the first (IV REML) is based on the popular REML estimation of the between-study variance τ^2 and the inverse-variance method for combining the evidence, whereas the second (SSW KDB) is based on the effective-sample-size weights and the KDB estimator of τ^2.[36] Our simulations clearly demonstrate that the SSW KDB analysis is a much better option when $\delta \neq 0$.

From theoretical consideration of CMA in Section 3, we recognized the issues with variance inflation in CMA when a shift in the mean occurs and suggested therefore a two-stage approach to CMA, as well as testing for a shift in τ^2. Our simulations show that the two-stage CMA performs better than the standard one-stage CMA on both Type 1 error and power. Testing for τ^2 also works well for small-to-moderate values of $\tau^2 \leq 0.25$; the Q-profile method[41] is the preferred option. However, this method has very low power for larger τ^2. For all studied methods, smaller shifts or higher heterogeneity results in lower power, but the direction of shift does not seem to matter. Power increases considerably in sample size n. However, even for $n = 1000$ and a large shift in δ from 1 to 2, at least three to five studies after the shift are needed to achieve 80% power at the 0.01 level when $\tau^2 = 0.1$, and seven to nine studies are needed when $\tau^2 = 0.25$. Overall, at least 15–20 studies are required to use any version of CMA.

In our simulations, we also considered CUSUM charts, suggested by Kulinskaya and Koricheva,[20] and modified them for random-effects MA by adding estimated at Stage 1 between-study variance τ_0^2. However, this resulted in too high a Type 1 error rate, and we do not recommend this method.

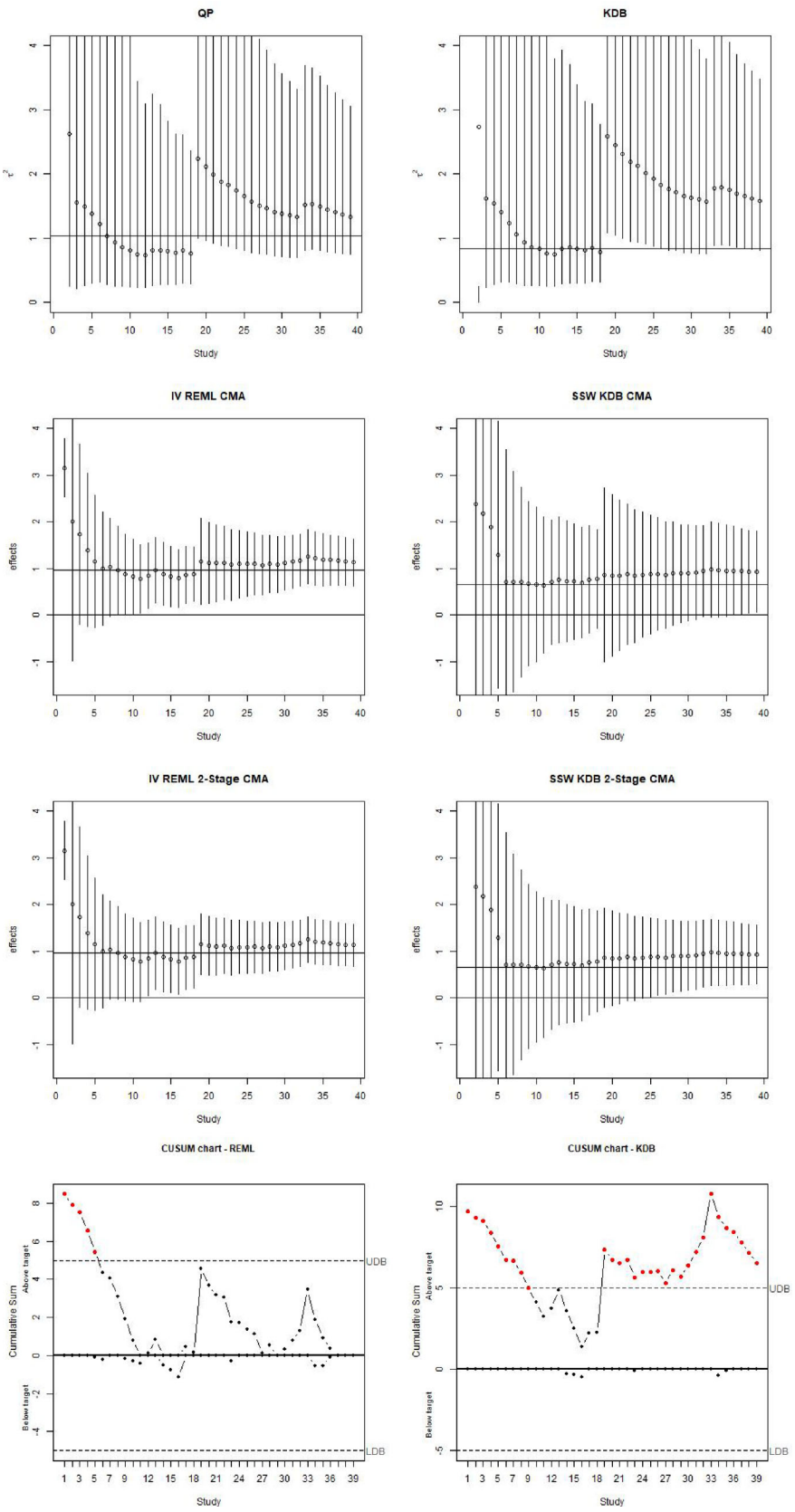

FIGURE 7 CMA plots for species richness data from Batáry et al. (2011). QP and KDB confidence intervals for τ^2 at 99% confidence level; one- and two-stage CMA intervals at 99% level, additional horizontal lines at $\hat{\delta}_{(8),\text{IV REML}} = 0.959$ and at $\hat{\delta}_{(10),\text{SSW KDB}} = 0.664$; CUSUM plots with $h = 5$. The effect measure is SMD, and positive values correspond to higher species richness in the intervention arm [Colour figure can be viewed at wileyonlinelibrary.com]

A practical recommendation is to run simultaneously two analyses: testing for cumulative τ^2 at the 0.005 level using the Q-profile method, and the two-stage testing for shift in the mean effect at the 0.01 level using SSW KDB, with either known or estimated in Stage 1 target value of δ, and using the constant value of τ_0^2 estimated in Stage

1 by KDB. The suggested levels guarantee an overall level close to 0.05 for $k \leq 26$ studies, as the two tests at levels α_1 and α_2, with rejection if at least one of them rejects the null hypothesis, result in approximately an $\alpha_1 + \alpha_2$ level. Somewhat higher levels would be possible for lower numbers of studies and/or lower between-study variances.

We studied only simple scenarios of equal sample sizes within and between studies, but we anticipate CMA to have even lower power in more realistic unbalanced settings. A study of the use of CMA for other effect measures would also be of interest for practitioners, though we do not believe that the power would improve. We considered only simple alternatives of a shift in δ at one point. Other realistic alternatives may include linear or nonlinear trends in effects and other more complicated options. Lai[52] provides a comprehensive review of the use of sequential methods for a wide class of alternatives. A critical review of these methods for applications in meta-analysis would be very useful.

In the case of high heterogeneity, the power of all CMA methods is very low. It would be of interest to consider the use of runs tests, which are routinely used in a similar quality control context,[53] and which may increase the power of CMA. Another important extension of CMA would be methodology for cumulative analysis when the heterogeneity is reduced through meta-regression. We shall address these possible improvements in further research.

ACKNOWLEDGMENTS

We are grateful to Peter Batáry who provided raw data from Batáry et al. (2011) for our example, to Julia Koricheva and Elizabeth Brisco for interesting discussions of cumulative meta-analyses in ecology and to David C. Hoaglin for useful comments which greatly improved the text. Work by E.K. was supported by the Economic and Social Research Council [grant number ES/L011859/1].

CONFLICT OF INTEREST

The authors declare that they have no competing interests.

AUTHOR CONTRIBUTIONS

Elena Kulinskaya and Eung Yaw Mah participated in the initial conceptualization. Elena Kulinskaya developed the required derivations. Eung Yaw Mah wrote the R code, implemented the simulations, and produced the examples, tables, and figures. Elena Kulinskaya and Eung Yaw Mah contributed to writing the manuscript, participated in revisions, and read and approved the final manuscript.

DATA AVAILABILITY STATEMENT

Our full simulation results are available as e-print [38]. R procedures implementing the proposed methods of CMA are available at https://mfr.osf.io/render?url=https\%3A\%2F\%2Fosf.io\%2F8bdvf\%2Fdownload and an example of their use is provided in Web Appendix S3.

ORCID

Elena Kulinskaya https://orcid.org/0000-0002-9843-1663

REFERENCES

1. Gurevitch J, Koricheva J, Nakagawa S, Stewart G. Meta-analysis and the science of research synthesis. *Nature*. 2018;555:175-182.
2. Shojania K, Sampson M, Ansari M, Ji J, Doucette S, Moher D. How quickly do systematic reviews go out of date? A survival analysis. *Ann Intern Med*. 2007;147(4):224-233.
3. Gehr B, Weiss C, Porzsolt F. The fading of reported effectiveness. A meta-analysis of randomised controlled trials. *BMC Med Res Meth*. 2006;6(1):25.
4. Fanshawe TR, Shaw LF, Spence GT. A large-scale assessment of temporal trends in meta-analyses using systematic review reports from the Cochrane library. *Res Synth Methods*. 2017;8(4):404-415. https://doi.org/10.1002/jrsm.1238
5. Monsarrat P, Vergnes JN. The intriguing evolution of effect sizes in biomedical research over time: smaller but more often statistically significant. *GigaScience*. 2018;7(1):1-10.
6. Brugger S, Davis J, Leucht S, Stone J. Proton magnetic resonance spectroscopy and illness stage in schizophrenia—a systematic review and meta-analysis. *Biol Psychiatry*. 2011;69(5):495-503.
7. Twenge JM, Konrath S, Foster JD, Keith Campbell W, Bushman BJ. Egos inflating over time: a cross-temporal meta-analysis of the narcissistic personality inventory. *J Pers*. 2008;76(4):875-902.
8. Johnsen T, Friborg O. The effects of cognitive behavioral therapy as an anti-depressive treatment is falling: a meta-analysis. *Psychol Bull*. 2015;141(4):747-768.
9. Grabe S, Ward LM, Hyde JS. The role of the media in body image concerns among women: a meta-analysis of experimental and correlational studies. *Psychol Bull*. 2008;134(3):460-476.
10. Jennions M, Møller A. Relationships fade with time: a meta-analysis of temporal trends in publication in ecology and evolution. *Proc Roy Soc Lond B*. 2002;269:43-48.
11. Nykänen H, Koricheva J. Damage-induced changes in woody plants and their effects on insect herbivore performance: a meta-analysis. *Oikos*. 2004;104:247-268.
12. Barto E, Rillig M. Dissemination biases in ecology: effect sizes matter more than quality. *Oikos*. 2012;121(2):228-235.
13. Sánchez-Tójar A, Nakagawa S, Sánchez-Fortún M, et al. Meta-analysis challenges a textbook example of status signalling and demonstrates publication bias. *eLife*. 2018;7:e37385.

14. Agogo D, Hess T. Blind to time? Temporal trends in effect sizes in IS research. In: ISIS2016; 2016. url=https://aisel.aisnet.org/icis2016/Methodological/Presentations/10/
15. Koricheva J, Kulinskaya E. Temporal instability of evidence base: a threat to policy making? *Trend Ecol Evol*. 2019;34(10):895-902. https://doi.org/10.1016/j.tree.2019.05.006
16. Lau J, Antman E, Jimenez-Silva J, Kupelnick B, Mosteller F, Chalmers T. Cumulative meta-analysis of therapeutic trials for myocardial infarction. *New Engl J Med*. 1992;327(4):248-254.
17. Trikalinos T, Churchill R, Ferri M, et al. Effect sizes in cumulative meta-analyses of mental health randomized trials evolved over time. *J Clin Epidemiol*. 2004;57:1124-1130.
18. Clarke M, Brice A, Chalmers I. Accumulating research: a systematic account of how cumulative meta-analyses would have provided knowledge, improved health, reduced harm and saved resources. *PLoS One*. 2014;9(7):e102670.
19. Trikalinos T, Ioannidis J. Assessing the evolution of effect sizes over time. In: Rothstein H, Sutton A, Borenstein M, eds. *Publication Bias in Meta-Analysis—Prevention, Assessment and Adjustments*. Wiley; 2005:241-259.
20. Kulinskaya E, Koricheva J. Use of quality control charts for detection of outliers and temporal trends in cumulative metaanalysis. *Res Synth Methods*. 2010;1:297-307.
21. Koricheva J, Jennions M, Lau J. Temporal trends in effect sizes: causes, detection, and implications. In: Koricheva J, Gurevitch J, Mengersen K, eds. *Handbook of Meta-Analysis in Ecology and evolution*. Princeton University Press; 2013:237-254.
22. Moher D, Tsertsvadze A, Tricco A, et al. When and how to update systematic reviews. *Cochrane Database Syst Rev*. 2008;1:MR000023.
23. Garner P, Hopewell S, Chandler J, et al. When and how to update systematic reviews: consensus and checklist. *BMJ*. 2016;354. https://doi.org/10.1136/bmj.i3507
24. Elliott J, Turner T, Clavisi O, et al. Living systematic reviews: an emerging opportunity to narrow the evidence-practice gap. *PLoS Med*. 2014;11(2):e1001603.
25. Rothstein H, Sutton J, Borenstein M. *Publication Bias in Meta-Analysis - Prevention, Assessment and Adjustments*. John Wiley and Sons; 2006.
26. Atakpo P, Vassar M. Cumulative meta-analysis by precision as a method to evaluate publication bias. *Dermatol Sci*. 2016;83(3):251-253.
27. Whitehead A. A prospectively planned cumulative meta-analysis applied to a series of concurrent clinical trials. *Stat Med*. 1997;16(24):2901-2913.
28. Hu M, Cappelleri JC, Lan KG. Applying the law of iterated logarithm to control type I error in cumulative meta-analysis of binary outcomes. *Clin Trials*. 2007;4(4):329-340.
29. Thorlund K, Imberger G, Walsh M, et al. The number of patients and events required to limit the risk of overestimation of intervention effects in meta-analysis—a simulation study. *PLoS One*. 2011;6(10):e25491. https://doi.org/10.1371/journal.pone.0025491
30. Pogue J, Yusuf S. Cumulating evidence from randomized trials: utilizing sequential monitoring boundaries for cumulative meta-analysis. *Control Clin Trials*. 1997;18(6):580-593.
31. Brok J, Thorlund K, Gluud C, Wetterslev J. Trial sequential analysis reveals insufficient information size and potentially false positive results in many meta-analyses. *J Clin Epidemiol*. 2008;61:763-769.
32. Wetterslev J, Thorlund K, Brok J, Gluud C. Trial sequential analysis may establish when firm evidence is reached in cumulative meta-analysis. *J Clin Epidemiol*. 2008;61(1):64-75.
33. Higgins J, Whitehead A, Simmonds M. Sequential methods for random-effects meta-analysis. *Stat Med*. 2011;30(9):903-921.
34. Kulinskaya E, Wood J. Trial sequential methods for meta-analysis. *Res Synth Methods*. 2014;5(3):212-220.
35. Dogo SH, Clark A, Kulinskaya E. Sequential change detection and monitoring of temporal trends in random-effects metaanalysis. *Res Synth Methods*. 2017;8(2):220-235. https://doi.org/10.1002/jrsm.1222
36. Kulinskaya E, Dollinger MB, Bjørkestøl K. Testing for homogeneity in meta-analysis I. the one-parameter case: standardized mean difference. *Biometrics*. 2011;67(1):203-212.
37. Bakbergenuly I, Hoaglin DC, Kulinskaya E. Estimation in meta-analyses of mean difference and standardized mean difference. *Stat Med*. 2020;39(2):171-191.
38. Kulinskaya E, Mah EY. Simulation results on the performance of statistical methods in cumulative meta analysis. *MetaArXiv*. 2021;464. https://doi.org/10.31222/osf.io/8t4pf
39. Hedges LV. A random effects model for effect sizes. *Psychol Bull*. 1983;93(2):388-395.
40. Viechtbauer W. Bias and efficiency of meta-analytic variance estimators in the random-effects model. *J Educ Behav Stat*. 2005;30(3):261-293.
41. Viechtbauer W. Confidence intervals for the amount of heterogeneity in meta-analysis. *Stat Med*. 2007;26(1):37-52.
42. Cochran WG. The combination of estimates from different experiments. *Biometrics*. 1954;10(1):101-129.
43. Hedges LV, Olkin I. *Statistical Methods for Meta-Analysis*. Academic Press; 1985.
44. Montgomery DC. Introduction to Statistical Quality Control. *Wiley*. 2000.
45. Page E. Continuous inspection schemes. *Biometrika*. 1954;41:100-115.
46. Grigg O, Spiegelhalter D. An empirical approximation to the null unbounded steady-state distribution of the CUSUM statistic. *Dent Tech*. 2008;50:501-511.
47. Scrucca L. Qcc: an R package for quality control charting and statistical process control. *R News*. 2004;4(1):11-17.
48. DerSimonian R, Kacker R. Random-effects model for meta-analysis of clinical trials: an update. *Contemp Clin Trials*. 2007;28(2):105-114.
49. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials*. 1986;7(3):177-188.
50. Paule RC, Mandel J. Consensus values and weighting factors. *J Res Natl Bur Stand*. 1982;87(5):377-385.
51. Batáry P, Báldi A, Kleijn D, Tscharntke T. Landscape-moderated biodiversity effects of Agri-environmental management: a meta-analysis. *Proc Roy Soc B Biol Sci*. 2011;278(1713):1894-1902. https://doi.org/10.1098/rspb.2010.1923
52. Lai TL. Sequential changepoint detection in quality control and dynamical systems. *J Roy Stat Soc*. 1995;57(4):613-644. https://doi.org/10.1111/j.2517-6161.1995.tb02052.x

53. Koutras M, Bersimis S, Maravelakis P. Statistical process control using Shewhart control charts with supplementary runs rules. *Meth Comput Appl Prob*. 2007;9:207-224.

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: Kulinskaya E, Mah EY. Cumulative meta-analysis: What works. *Res Syn Meth*. 2021;1-20. doi:10.1002/jrsm.1522