Research Synthesis Methods

Received 24 August 2010, Revised 31 January 2011, Accepted 2 February 2011 Published online 8 April 2011 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/jrsm.29

Use of quality control charts for detection of outliers and temporal trends in cumulative meta-analysis

Elena Kulinskaya[a]*† and Julia Koricheva[b]

Cumulative meta-analysis (CMA) aims to aggregate accumulating evidence. Essentially a visual tool, CMA should be supplemented by formal statistical methods for assessment of the significance of the accumulating evidence, and for detection of temporal trends in effect sizes. These methods should also take into account multiple testing inherent in CMA. We review the existing methods for detection of temporal trends in effect sizes and suggest a new approach, namely the use of standard quality control (QC) charts, in particular $\bar{X}$ charts and CUSUM charts, to detect possible outliers and trends over time. We discuss the application of the QC charts to four popular measures of effect size: the odds ratios, the relative risks, the correlation coefficients and the standardized mean differences. Applications of QC charts are illustrated by three meta-analysis examples from medicine, ecology and evolutionary biology. Copyright © 2011 John Wiley & Sons, Ltd.

Keywords: cumulative meta-analysis; CUSUM charts; quality control charts; $\bar{X}$ charts; temporal changes in effect sizes

1. Introduction

The general aim of meta-analysis, as well as of any other form of research synthesis, is to combine scientific evidence scattered through a number of individual studies addressing the same topic in order to create generalizations and provide the evidence base for decision making in various scientific fields. Any given meta-analysis provides just a snapshot of the available evidence at more or less arbitrary point in time. However, scientific evidence is not static and tends to 'evolve' with time (Trikalinos and Ioannidis, 2005). New studies might either strengthen or challenge the conclusions of previous reports, resulting in changes in the mean effect size and its variance over time. As the magnitude and direction of the mean effect size and the breadth of its confidence interval largely determine the conclusions drawn from a meta-analysis, it is important to be aware of the extent of temporal variation in effect sizes. If the reported effect size changes over time, the result of meta-analysis would depend on when it was performed. Thus, temporal changes in effect sizes could impair the replication validity of a meta-analysis.

Temporal changes in effect sizes have been reported in many scientific fields including medicine (Gehr *et al*., 2006; Trikalinos *et al*., 2004), the social sciences (Grabe *et al*., 2008) and ecology and evolutionary biology (Jennions and Møller, 2002; Nykänen and Koricheva, 2004). Factors which may cause these changes include time-lag bias, publication bias, heterogeneity among studies, changes in research methods, paradigm shifts or real changes in effects over time due to temporal changes in baseline values or strength of causal agents (Gehr *et al*., 2006; Jennions and Møller, 2002; Trikalinos and Ioannidis, 2005). Most commonly earlier studies report larger estimates of the effect size than subsequent studies (Ioannidis *et al*., 2001; Jennions and Møller, 2002; Trikalinos *et al*., 2004) although increases in effect sizes with publication year have also been reported (Barto and Rillig, 2010; Grabe *et al*., 2008). Temporal changes in the magnitude of the effect sizes may be quite dramatic (several folds) and often lead to the loss (or gain) of the statistical significance or even changes in the sign of the cumulative mean effect

[a] *School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, U.K.*

[b] *School of Biological Sciences, Royal Holloway, University of London, Egham, Surrey, TW20 0EX, U.K.*

**Correspondence to: Elena Kulinskaya, School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, U.K.*

† E-mail: e.kulinskaya@uea.ac.uk

(Ioannidis and Trikalinos, 2005; Nykänen and Koricheva, 2004; Saikkonen *et al*., 2006). This suggests that results of meta-analyses conducted early in the process of research accumulation on the topic should be interpreted with caution and regularly updated. Detection of temporal trends in effect sizes is thus an important methodological issue and in the rest of the paper we review pros and cons of the existing methods for detection of temporal changes in effect sizes, suggest a new approach and illustrate its application on three meta-analytic data sets from medicine, ecology and evolutionary biology.

2. Existing methods for detection of temporal trends in effect sizes

Several different methods have been used so far to detect temporal changes in effect sizes. The first approach is to subdivide studies into groups based on the publication year (e.g. by decades or published before and after year X) and to compare mean effect sizes between the groups by using the homogeneity analysis (Gardner *et al*., 2003; Grabe *et al*., 2008) and tests of interaction between subgroups (Altman and Bland, 2003; Gurevitch *et al*., 2000). This relatively crude approach ignores likely gradual character of temporal changes and their possible occurrence within as well as between the studied groups. The second approach is based on the correlation or regression analysis of the relationship between the magnitude of the effect size and publication year. Pearson's product–moment or Spearman's rank correlations are used in Jennions and Møller (2002) and Torres-Vila and Jennions (2005). Linear regression is used in Gehr *et al*. (2006) and Kampichler and Bruckner (2009). Linearity of the relationship between effect size and publication year is seldom, if ever, tested and may be often violated (Nykänen and Koricheva, 2004). Regression applied to consecutive combined effects is proposed in Bagos and Nikolopoulos (2009). An exponential model of proportional decrease in effect is introduced in Baker and Jackson (2010). All these methods are applicable only when the magnitude of the effect exhibits a monotonic decrease or increase with time, which is not always the case (Ioannidis and Trikalinos, 2005; Nykänen and Koricheva, 2004).

The third group of methods involves cumulative meta-analysis (CMA). This method was initially introduced in medicine to detect the earliest date by which a treatment effect became statistically significant and a conclusion about its clinical efficiency could therefore be drawn (Lau *et al*., 1992). In CMA, data are sorted in a chronological order and combined estimates of the effect $\hat{\theta}_k$ and their confidence intervals are obtained consecutively for $k=1,\ldots,K$. The obtained trajectories are plotted and scrutinized visually for the possible temporal trends (Ioannidis and Trikalinos, 2005; Leimu and Koricheva, 2004). Ioannidis and Lau (2001) have suggested a modification of CMA, a recursive cumulative meta-analysis (RCMA), which shows the relative change in the magnitude of the treatment effect as a new study is added to the meta-analysis. The advantage of the CMA approach is that it allows to reveal uneven, irregular or nonlinear changes in the effect size as well as multiple shifts in opposite directions.

As all visual tools, CMA and RCMA plots might be subjected to misinterpretation and should be supplemented by formal statistical methods. Statistical methods for detection of temporal trends in effect sizes should be applicable to a variety of effect size metrics used in meta-analysis and should allow detection of nonlinear and uneven irregular changes in effects over time. These methods should also take into account multiple testing inherent in CMA. One recently proposed class of methods is based on statistical methods developed for sequential clinical trials (Brok *et al*., 2008; Wetterslev *et al*., 2008).

In this paper we propose the use of standard quality control (QC) charts, in particular $\bar{X}$ charts and CUSUM charts to detect possible outliers and trends over time in meta-analysis. We discuss the application of the QC charts to three popular effect measures: the odds ratio, the correlation coefficients and the standardized mean differences (SMD). The results are illustrated by three examples from medicine, ecology and evolutionary biology.

$\bar{X}$ and CUSUM QC charts are introduced in Section 3, their application to meta-analysis is described in Section 4. Three examples of CMA in medicine, ecology and evolutionary biology are in Section 5, and the summary and discussion are in Section 6.

3. Quality control charts

Methods of statistical QC were initially developed in industrial applications of statistics to assess whether the variability of a production process is due to chance or due to assignable causes. A process that is operating with only chance causes of variation is said to be in statistical control. Other kinds of variability may occasionally be present. For industrial applications, these will be improperly adjusted machines, operator errors or defective raw materials. Such variability is usually large in comparison to background noise. These causes of variability are referred to as assignable causes. A process that is operating in the presence of assignable causes is said to be out of control (Montgomery and Runger, 1994, Chapter 14). Nowadays, QC charts are commonly used in medicine, epidemiology and public health to detect a start of an epidemic or to control quality within the National Health Service (Grigg and Spiegelhalter, 2005; Marshall *et al*., 2004; Winkel and Zhang, 2007), in forensic applications from computing (Hadjidja *et al.*, 2009) to authorship analysis (Farringdon *et al.*, 1996) and in other areas. Similar problem also arises in meta-analysis, where we are aiming to find out whether variation in effect sizes is due to sampling error (background noise) or assignable causes (study-specific covariates or temporal trends).

Standard use of QC includes two stages: a set-up stage where the parameters in control of a process are ascertained, and the subsequent monitoring stage. In this section we discuss in detail two popular QC charts aimed at detecting changes in continuous outcomes. QC procedures are available in all major statistical packages. R package qcc (Scrucca, 2004) was used for analysis of all our examples.

3.1. $\bar{X}$ control chart

The QC control charts are used primarily for online process monitoring. The sample data (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 and do not exhibit any systematic pattern, the process is considered to be in control.

In general, if a normally distributed statistic W with the mean μ_W and the standard deviation σ_W is used to measure some quality characteristics of interest, the center line on $\bar{X}$ control chart is drawn at μ_W, and the upper and lower control limits are at $\mu_W \pm k\sigma_W$. Here, k is the distance of the control limits from the central line in the standard deviation units. There is a close connection between a control chart and significance testing. The points within the limits mean that the hypothesis that the process is in control was not rejected, and a point outside the limits means that the hypothesis is rejected. The general theory of these charts was developed by W.A. Shewhart, and these charts are often called Shewhart control charts.

The control limits on $\bar{X}$ charts are usually plotted at $k=3$. The limits are therefore called three-sigma control limits. These limits correspond to a significance level of $\alpha=0.0027$ when testing the null hypothesis $\mu=\mu_W$ against two-sided alternatives. This small significance level is chosen to adjust for multiple testing. An important notion associated with the chosen significance level is the average run length (ARL) of the control chart. The ARL is the expected number of points that are to be plotted before a point is outside the control limits. For any $\bar{X}$ chart, the ARL$=1/p$, where p is the probability that any point exceeds the control limits. When the process is in control the ARL$=1/\alpha$ and for a three-sigma chart, ARL$=370$.

The power of the test is reflected in the ARL for a particular magnitude of the process shift expressed in the multiples of σ. For example, for a shift of 1.5σ the ARL is 15, and for a shift of 3σ the ARL$=2$ when the sample sizes are 1 (Montgomery and Runger, 1994, Table 14-1)). ARL is considerably smaller for larger sample sizes because the probability p of a sample mean to exceed the control limits increases with the sample size N. For a shift of δ standard deviations, the probability p is given by $p=\Phi((\delta-3)N)+\Phi(-(3+\delta)N)$.

Examples of $\bar{X}$ charts applied to meta-analytic data are given on the middle plots in Figures 1–3. We discuss these examples in detail in Sections 4 and 5.

3.2. CUSUM chart

The main disadvantage of the $\bar{X}$ control charts is that they are relatively insensitive to small shifts of an order of 1.5σ or less. A very effective alternative to $\bar{X}$ charts is the cumulative sum or CUSUM chart (Page, 1954). This chart has much smaller ARL for detecting small shifts, but does not cause the in-control ARL to drop significantly, delivering an increase in the power while preserving type I error.

The CUSUM chart plots the cumulative sums of the deviations of the sample values from a target value. While the $\bar{X}$ chart tests the null hypothesis H_0 (the process is in control) against an alternative H_1 (the process is out of control) independently for each new observation, the CUSUM chart is equivalent to a sequential likelihood ratio test that uses all the accumulated information. The cumulative log likelihood ratio (LLR) of H_1 compared to H_0 is plotted at every step and the test stops in favour of H_1 when the LLR is large (Grigg and Spiegelhalter, 2008). The chart is restricted from falling below zero, and often two one-sided CUSUM charts (for positive and negative deviations) are plotted simultaneously.

For normal data $y_1,\ldots,y_t \sim N(\mu_t,1)$, Page (1954) proposed plotting a cumulative sum $S_t=\sum_1^t(y_i-\mu_0)$ on a chart, where μ_0 indicates some target mean level of a process. If the mean of the process μ_t is larger than μ_0, the mean path of S_t is upward sloping. If, on the other hand, the mean μ_t is less than μ_0, the trend is downward sloping. Page (1954) proposed to take action if $S_t-\min_{0\leqslant i\leqslant t} S_i \geqslant a$ or $\max_{0\leqslant i\leqslant t} S_i - S_t \geqslant a$ for some fixed value of a.

Equivalently, an upper one-sided CUSUM testing the hypothesis $H_0:\mu=0$ against $H_1:\mu=\delta>0$ can be written as cumulative LLR

$$x_0=0, \quad x_t=\max(0,x_{t-1}+v_t), \quad t=1,2,\ldots$$

where for the standard normal distribution with the density $\phi(\cdot)$, the value of LLR at time t is

$$v_t=\log\phi(y_t-\delta)-\log\phi(y_t)=\delta(y_t-\delta/2)$$

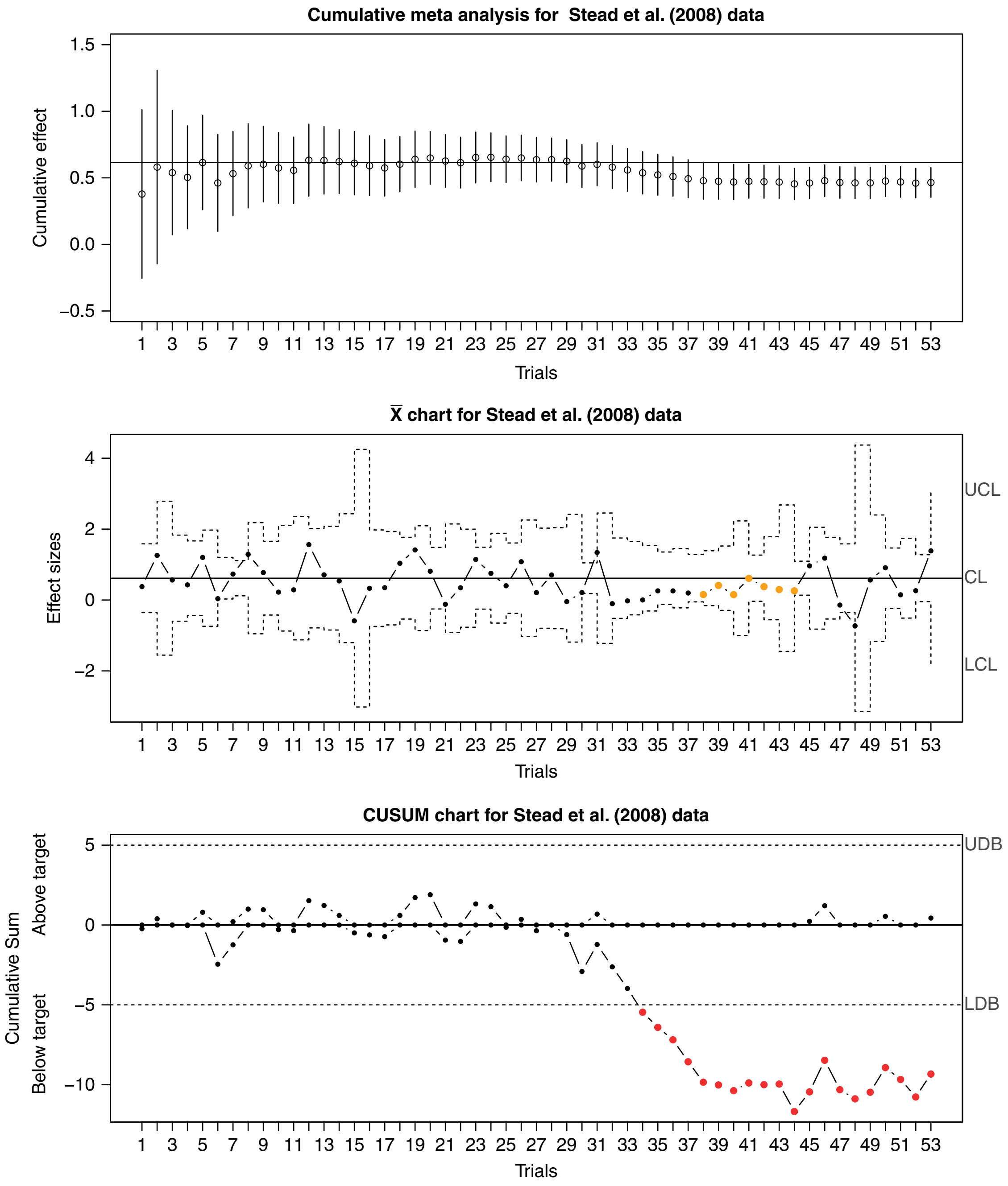

Figure 1. CMA, $\bar{X}$ and CUSUM charts for Stead *et al.* (2008) data. Random-effects model (REM) is used for CMA, and fixed-effect model for QC charts. Horizontal central lines on CMA and Xbar chart are at $0.615=\ln(1.850)$, corresponding to the REM combined effect of the first five studies on log odds ratio scale. Control limits (dashed lines) are at $\pm 3\sigma$ and at $\pm 5\sigma$ on $\bar{X}$ and on CUSUM charts, respectively. Out-of-control values are in red; run test violations are in orange.

Thus, the upper CUSUM accumulates deviations from the target value that are greater than $\delta/2$. When the variance $\sigma_t^2 \neq 1$, a more general null hypothesis $H_0: y_t' \sim N(\mu_0, \sigma_t^2)$ is tested against an alternative $H_1: y_t' \sim N(\mu_0+\delta\sigma_t, \sigma_t^2)$ of practically relevant amount of shift δ (in standard deviations) in the underlying mean μ. This procedure can easily be implemented by using the transformation $y_t=(y_t'-\mu_0)/\sigma_t$ and using the LLR v_t above. A constant h is chosen so that the CUSUM signals as soon as $x_t > h\sigma_t$. This value is chosen to provide good ARL values. The standard choices are $h=4$ or $h=5$. These values provide the ARL of 168 and 465, respectively, when the process is in control, and the ARL of 4.75 and 5.75, respectively, for a shift of 1.5σ, see (Montgomery and Runger, 1994, Table 14-9).

Typically, sample statistics from samples of sizes n_t such as sample means are plotted instead of individual observations. Then, σ_t is proportionate to $n_t^{-1/2}$. The target shift $\delta\sigma_t$ decreases as the inverse square root of the sample size n_t and the power of the CUSUM chart increases accordingly.

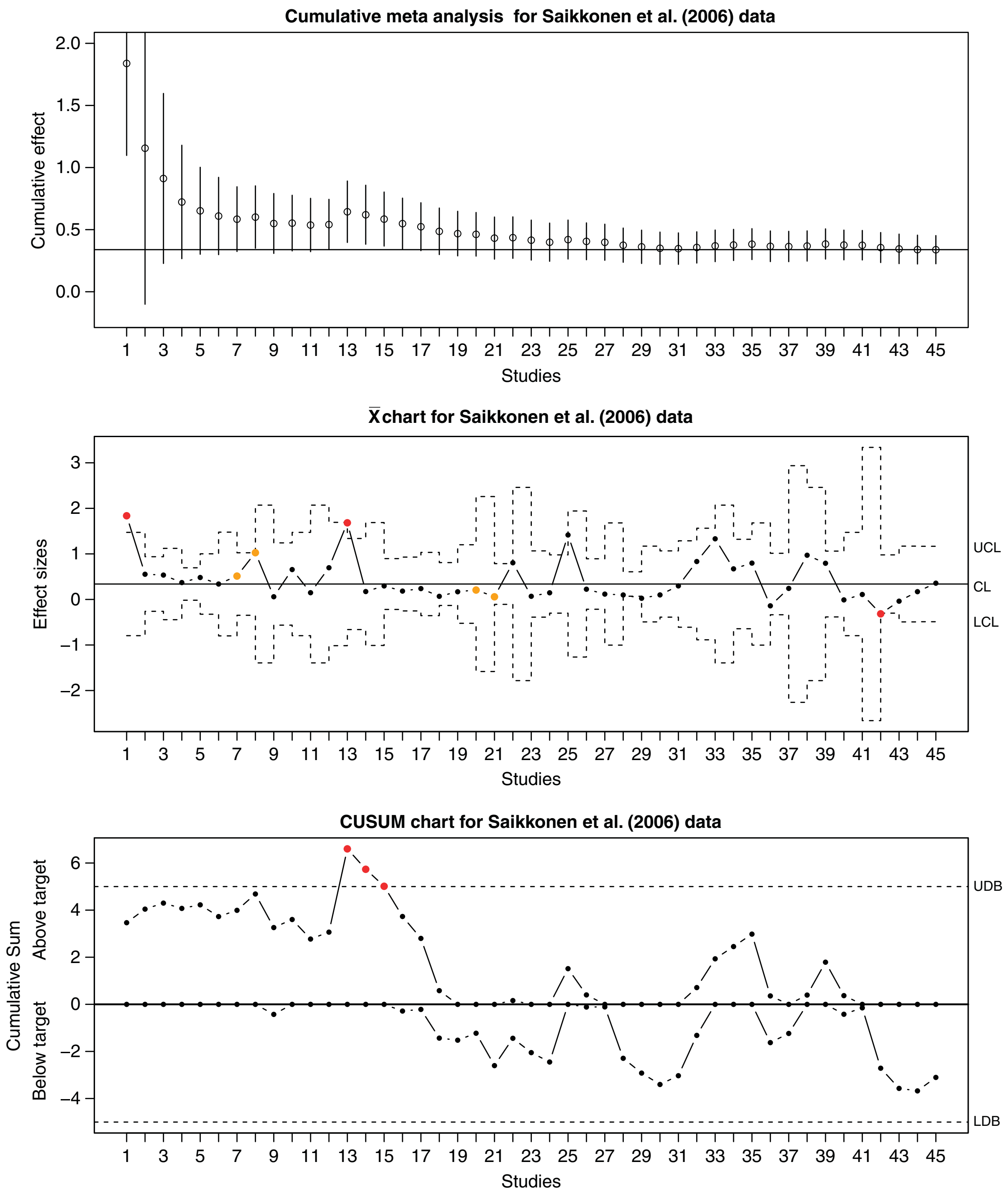

Figure 2. CMA, $\bar{X}$ and CUSUM charts for Saikkonen *et al.* (2006) data. REM is used for CMA, and fixed-effect model for QC charts. Horizontal central lines on CMA and $\bar{X}$ chart are at the target value of 0.339, REM combined effect on the *z*-transformed scale. Control limits (dashed lines) are at $\pm 3\sigma$ and at $\pm 5\sigma$ on Shewhart and on CUSUM charts, respectively. Out-of-control values are in red; run test violations are in orange.

For further statistical properties of the CUSUM statistic see Grigg and Spiegelhalter (2008) and Hawkins and Olwell (1997) and the references therein. Examples of CUSUM charts for detection of temporal changes in meta-analysis are given in the bottom plots in Figures 1–3. Details of these applications are provided in Sections 4 and 5.

4. Application of QC charts to meta-analysis

Accumulation of scientific evidence is a process where new studies on a given topic appear over time. Each consecutive study estimates an effect of interest θ_t, $t = 1, 2, \ldots$. Effect estimates $\hat{\theta}_t$ reported by the study t (or their transformations) are approximately normally distributed. When there is no temporal shift, the process is in control and all effect estimates are normally distributed with the same mean, $\hat{\theta}_t \sim N(\theta, \sigma_t)$. If a shift happens

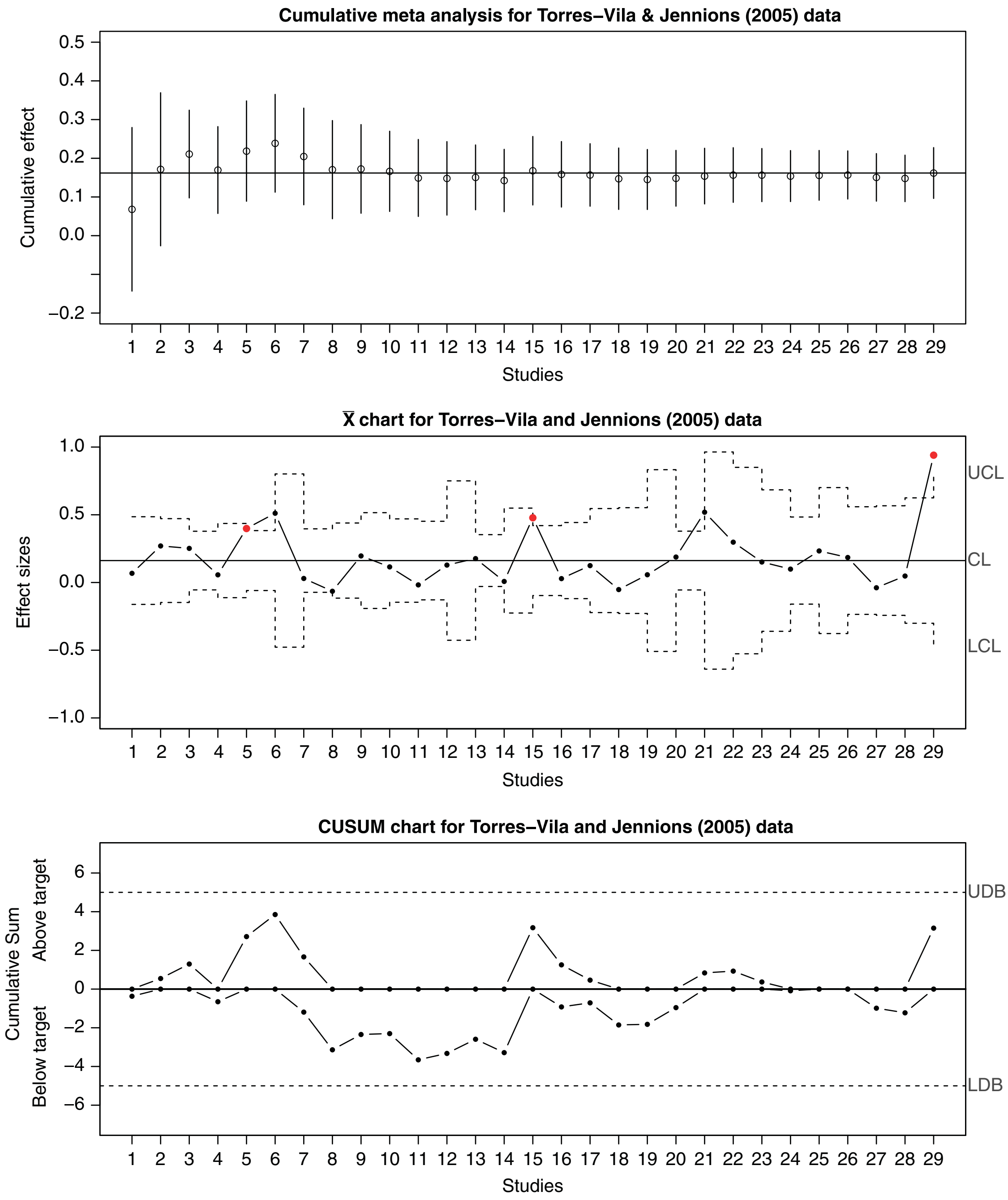

Figure 3. CMA, $\bar{X}$ and CUSUM charts for Torres-Vila and Jennions (2005) data. REM is used for CMA, and fixed-effect model for QC charts. Horizontal central lines on CMA and $\bar{X}$ chart are at the target value of 0.162, REM combined effect on the variance-stabilized scale. Control limits (dashed lines) are at $\pm 3\sigma$ and at $\pm 5\sigma$ on $\bar{X}$ and on CUSUM charts, respectively. Out-of-control values are in red.

at some time point, the mean of the process deviates from θ and the process can be considered out of control. Thus, an application of QC charts to temporal trends in meta-analysis is rather straightforward, a minor variation being the use of non-constant variances σ_t.

Applying the CUSUM charts to meta-analysis, the LLR for study t is

$$v_t = \delta[(\hat{\theta}_t - \theta_0)/\sigma_t - \delta/2] = \delta[\sqrt{w_t}(\hat{\theta}_t - \theta_0) - \delta/2]$$

where the weights w_t are the inverse variances of $\hat{\theta}_t$, typically proportionate to $\sqrt{n_t}$. For a positive shift of δ standard deviations $\theta_t = \theta_0 + \delta\sigma_t$, the expected value $E(v_t) = \delta^2/2 > 0$. Denote the cumulative sum of inverse standard deviations $W_t = \sum_1^t \sqrt{w_i}$. Consider, for simplicity, the case when all deviations $\sqrt{w_i}(\theta_i - \theta_0) - \delta/2$ are positive. Then the upper CUSUM value at study t is

$$x_t = \delta\left(\sum_1^t \sqrt{w_i}(\hat{\theta}_i - \theta_0) - t\delta/2\right) = \delta t[\overline{\sqrt{w}}_t(\bar{\theta}_{\sqrt{w},t} - \theta_0) - \delta/2]$$

where $\bar{\theta}_{\sqrt{w},t}$ is the weighted mean effect for studies 1 to t with weights $\sqrt{w_i}$, and $\overline{\sqrt{w}}_t = W_t/t$ is the average of the inverse standard deviations proportionate to the weighted average of $\sqrt{n_i}$. When some of the deviations are negative, only $t^* < t$ positive deviations are added in x_t and contribute to average effect $\bar{\theta}^+_{\sqrt{w},t*}$ and average weight $\overline{\sqrt{w}}^+_{t*}$, the above formula becomes

$$x_t = \delta t^*[\overline{\sqrt{w}}^+_{t*}(\bar{\theta}^+_{\sqrt{w},t*} - \theta_0) - \delta/2].$$

Expected path increases linearly with the number of positive deviations and with their average weight, until it crosses the chosen limit. Similarly, when there is a negative shift $\theta_t = \theta_0 - \delta\sigma_t$, x_t accumulates negative deviations, and the lower CUSUM decreases linearly. Both upper and lower one-sided CUSUMs are plotted on standard CUSUM charts.

All standard QC software provides procedures able to plot QC charts for samples of independent normally distributed variables. In meta-analysis some popular effect measures are not normally distributed, and the variances often depend on the underlying effects. In these cases variance stabilization provides convenient means of transforming the data into a scale with constant variance. After this transformation, the standard QC software can be readily used. In the following subsections we show how QC methodology can be applied to odds ratios, relative risks, correlation coefficients and SMD.

4.1. Odds ratios and relative risks

Consider an experiment with a binary outcome (success or failure) studied under two experimental conditions: Treatment and Control, with sample sizes n_C and n_T, respectively. The data consist of the numbers of outcome of interest under these two conditions (X_T and X_C). It is assumed that the data are binomially distributed $X_T \sim B(p_T, n_T)$, $X_C \sim B(p_C, n_C)$. The probabilities (risks) p_C and p_T are estimated by $\hat{p} = (X+g)/(n+2g)$, where an optional constant g is added in case of zeros and to reduce bias. Usually, $g = 0.5$.

The odds ratio $\mathrm{OR} = p_T(1-p_C)/(1-p_T)p_C$ or its log, $\psi = \ln(\mathrm{OR})$, and relative risk $\mathrm{RR} = p_T/p_C$ or its log, $\phi = \ln(\mathrm{RR})$, are popular effect metrics. To estimate the OR or the RR, the risk estimates $\hat{p}_T$ and $\hat{p}_C$ are substituted into the above formulas. The sample log odds ratio $\hat{\psi}$ and log relative risk $\hat{\phi}$ are approximately normally distributed with variances $\mathrm{Var}(\hat{\psi}) = [n_C p_C(1-p_C)]^{-1} + [n_T p_T(1-p_T)]^{-1}$ and $\mathrm{Var}(\hat{\phi}) = p_C^{-1} + p_T^{-1} - [n_C]^{-1} - [n_T]^{-1}$. Inverse variance weights are often used in meta-analysis.

Given that the sample log odds ratios or log relative risks from individual studies are approximately normally distributed with the above variances, we can use the standard QC software, specifying the sample sizes as 1, and the standard deviations as the square root of the variances. The QC charts can be used in a similar way for any other approximately normally distributed effect measures.

4.2. Correlation coefficients

The correlation coefficient ρ from a bivariate normal population with unknown mean and covariance matrix is a common measure of the effect size in meta-analyses. It is estimated by Pearson's correlation coefficient r. This statistic has a complicated non-null distribution when $\rho \neq 0$ (Fisher, 1915), and in particular the variance of r depends on its value: $\mathrm{Var}(r) = (1-r^2)^2/(N-1)$. This is the reason why the standard meta-analytic approach of using inverse variance weights to combine evidence does not work for r. Instead, Fisher's z-transform is widely used.

Fisher's z-transform $z = \frac{1}{2}\log((1+r)/(1-r))$ is a variance-stabilizing transformation with variance given by $\mathrm{Var}(z) \approx (N-3)^{-1}$ and applicable for $N \geqslant 4$. Thus, transformed correlation coefficients, $z(r_i)$ can be added with known weights, $N_i - 3$, $i = 1, \ldots, t$ in meta-analysis.

Importantly, $z(r_i)$ is approximately normally distributed, and its distribution is the same as that of a mean of $N_i - 3$ independent normal random variables with variance 1. CUSUM chart procedure with $\sigma = 1$ and sample sizes $N_i - 3$ can be applied to $z(r_i)$, $i = 1, \ldots, t$.

4.3. SMD

In some applications, scales of measurement differ between the studies. This is especially widespread in environmental sciences, epidemiology, psychology and social sciences. Different tools are used, for example, to measure pain or cognition. Then, the SMD is chosen as an effect measure:

$$\delta = (\mu_T - \mu_C)/\sigma$$

Standard deviation σ is assumed to be equal in treatment (T) and control (C) arms of a study. Let μ_T and μ_C be the means in the treatment and control arms, respectively, and denote the sample sizes by n_T and n_C.

The naïve estimator $\hat{\delta}=(\bar{X}_T-\bar{X}_C)/s_p$ (where $\bar{X}_T$ and $\bar{X}_C$ are the sample means, and s_p is the pooled sample standard deviation) is slightly biased, and the unbiased estimator of δ is defined by

$$\hat{d}=\left[1-\frac{3}{4N-9}\right](\bar{X}_T-\bar{X}_C)/s_p$$

where $N=n_T+n_C$ is the total sample size, Hedges and Olkin (1985). This estimator is sometimes referred to as Hedges d. This correction is negligible for $N\geqslant 30$.

Variance of the SMD estimator is

$$\text{Var}(\hat{d})\approx N^{-1}\left[\frac{1}{q(1-q)}+\frac{\delta^2}{2(1-3.94N^{-1})}\right] \quad \text{for } q=n_C/N.$$

The SMD is not normally distributed, it has scaled non-central t distribution: $S_N=\sqrt{Nq(1-q)}\hat{d}\sim t_{N-2}(\sqrt{Nq(1-q)}\delta)$. Additionally, variance is a quadratic function of the effect δ. To be able to apply the CUSUM chart in CMA of the SMD, variance stabilization needs to be used.

Variance-stabilizing transformation for SMD is given in formula (20.3) of Kulinskaya *et al.* (2008) as

$$h(\hat{\delta})=\sqrt{2}\ln\left(\sqrt{\frac{q(1-q)}{2}}\hat{\delta}+\sqrt{1+\frac{q(1-q)}{2}\hat{\delta}^2}\right)$$

The transformed effect $h(\hat{\delta}_i)$ has approximately normal distribution with variance $1/N_i$. This distribution is the same as that of a mean of N_i independent normal random variables with variance 1. CUSUM chart procedure with $\sigma=1$ and sample sizes N_i can be applied to variance-stabilized effects $h(\hat{\delta}_i)$.

5. Examples

The following three examples from medicine, ecology and evolutionary biology demonstrate the use and interpretation of QC charts for detection of temporal trends in effect sizes. Within each meta-analytic data set individual studies were sorted in a chronological order. Studies published in the same year were entered into the analysis in a random order. When several estimates of effect size were available per study, as was the case in Saikkonen *et al.* (2006), we averaged effects within a study. The variance of such an average effect depends on the values of correlations between the outcomes within a study. We considered correlations ρ from 0.1 to 0.9. For illustrative purposes we assume $\rho=0.5$ in example 5.2 below. Cumulative meta-analyses were conducted using the R package meta (Schwarzer, 2010) and QC charts were produced by using the qcc R package (Scrucca, 2004). The shift $\delta=1$ was used for all CUSUM charts.

5.1. Nicotine replacement therapy for smoking cessation

The first example is based on a systematic review by Stead *et al.* (2008) testing the effectiveness of nicotine replacement therapy for smoking cessation. We selected trials using nicotine gum, constituting 53 trials published from 1979 to 2005. The effect measure is the odds ratio of smoking cessation. The analysis is based on log odds ratios ψ_i, which are approximately normally distributed. The full data set and its analysis are provided in Web Table 1.[‡] For a new intervention we would consider the null hypothesis of no effect of the gum ($\psi=0$). The CMA based on random-effects model (REM) provides significant result ($p=0.024$) at trial 3. For the CUSUM chart with the target value of zero and the shift $\delta=1$ consecutive CUSUM values are obtained from consecutive values of log OR ψ_i and their standard deviations SD$_i$ as $x_t=\sum_{i=1}^{t}\psi_i/\text{SD}_i-t/2$. The lower CUSUM values for this example are all zero, with the only exception of trial 48. The first five values of the upper CUSUM are 0.67, 1.91, 2.81, 3.53 and 5.69. Thus, the CUSUM chart with the decision limits at $h=5$ signals at trial 5, where the CMA p-value is 0.0007 and the combined odds ratio is 1.850 [1.298; 2.637]. The CUSUM chart is more conservative than the CMA to account for multiple testing.

At this stage, the positive effect of nicotine gum in smoking cessation can be considered established, and we may wish to monitor it prospectively. The CMA based on the REM and QC charts for the Stead *et al.* data are given in Figure 1. The centerline at the CMA and $\bar{X}$ plots is placed at $0.615=\ln(1.850)$. The CMA shows gradual decrease in the magnitude of the effect over time starting from trial 32 but the difference is significant only from trial 39. The $\bar{X}$ chart shows significant run test violations from trial 38, published in 1992. The CUSUM chart signals negative shift at trial 34, a year earlier. The final combined odds ratio is 1.593 [1.425; 1.780].

[‡] *Supporting information may be found in the online version of this article.*

5.2. *Endophyte-mediated resistance to herbivores in grasses*

The second example explores temporal changes in reported magnitude of endophyte-mediated resistance to herbivores in grasses (Saikkonen *et al*., 2006). Endophytic fungi live asymptomatically within plant tissues and the degree of infection varies both temporally and spatially. Endophytes may produce various toxins, including alkaloids, which have been shown to have negative effects on livestock and insect pests. Therefore, it has been hypothesized that endophytic fungi in grasses act as plant mutualists and protect the plants from herbivore attack. The meta-analysis by Saikkonen *et al*. (2006) combines the results of 47 studies published between 1985 and 2005 and assesses the relationship between the degree of endophytic infection in grasses and plant resistance to herbivores. Fisher's *z*-transformed Pearson's product–moment correlation coefficient is used as a metric of the effect size and positive correlation supports the hypothesis that the presence of endophytes increases plant resistance to herbivores. Saikkonen *et al*. (2006) have conducted a CMA of the above data and showed that although overall magnitude of the effect was positive and significantly different from 0, it decreased with the publication year, suggesting that more recent studies report weaker relationship between endophyte infection and herbivore resistance in grasses as compared with studies published in the 1980s. Their conclusion is tested here retrospectively.

The data for 45 studies with sample size greater than 2 and the analysis of this dataset are provided in Web Table 2, and the CMA and the QC charts for this data are given in Figure 2. The centerline is chosen at 0.339, the combined effect on the *z*-transformed scale in the REM of meta-analysis. The $\bar{X}$ chart shows two high outliers, studies 1 and 13, one low outlier, study 42, and two runs rules violations (two-out-of-three runs outside two sigma limits). CUSUM chart of the above data confirms the conclusion by Saikkonen *et al*. (2006), the combined mean of the early studies is above average, the first out-of-control signal corresponding to study 13 (published in 1992). CUSUM chart clearly shows a negative slope from study 13 (1992) to study 19 (1995) meaning a significant shift in the mean effect size. Decrease in the magnitude of the effect sizes in the more recent studies could be due to broadening of the scope of studied species.

5.3. *Male mating history and fecundity of female Lepidoptera*

The third example addresses the hypothesis that fecundity of female Lepidoptera (moths and butterflies) is higher when they mate with virgin males as compared with sexually experienced males (Torres-Vila and Jennions, 2005). This hypothesis is based on the fact that Lepidoptera females use nutrients transferred during the mating to increase their lifetime fecundity and Lepidoptera males with a greater number of previous matings tend to produce smaller spermatophores (capsules containing spermatozoa which are transferred to the female during copulation). Torres-Vila and Jennions (2005) have conducted a meta-analysis of 29 studies conducted on 25 Lepidoptera species and published between 1971 and 2003. SMD between the reproductive output of females mating with virgin versus sexually experienced males was used as a metric of the effect size. The data and analysis are provided in Web Table 3. In their original meta-analysis, Torres-Vila and Jennions (2005) have found that females which have mated with virgin males had significantly higher fecundity than females which have mated with sexually experienced males. They have also examined temporal trends in effect sizes by calculating Spearman's correlation between the year of publication and effect size. No significant correlation between effect size and publication year was found. Their data are reanalyzed here retrospectively. The QC charts are given in Figure 3. The centerline is chosen at 0.162, the combined effect on the variance-stabilized scale in the REM of meta-analysis. The $\bar{X}$ chart shows three high outliers, studies 5, 15 and 29. The CUSUM chart shows lack of temporal trends, in agreement with the conclusions by Torres-Vila and Jennions (2005).

6. Summary and discussion

The recent evidence indicates that temporal changes in magnitude, statistical significance and even sign of the effect are common in various scientific disciplines and may threaten the replication validity of meta-analysis (Gehr *et al*., 2006; Grabe *et al*., 2008; Jennions and Møller, 2002; Nykänen and Koricheva, 2004; Trikalinos and Ioannidis, 2005; Trikalinos *et al*., 2004). Among the existing methods for detection of temporal trends in effect sizes, only cumulative meta-analysis allows to reveal uneven, irregular or nonlinear changes as well as multiple shifts in opposite directions. However, this graphical tool has to be supplemented by the formal statistical methods for assessment of accumulating evidence and detection of temporal trends in effect sizes. Appropriate statistical methods should be applicable to a variety of effect size metrics used in meta-analysis and should also take into account multiple testing inherent in CMA. In this paper we proposed the use of QC charts for detection of outliers and assessment of accumulating evidence. We demonstrated applications of the QC charts to three popular metrics of effect size: the odds ratios, the correlation coefficients and the SMD, but they can be applied to any other approximately normally distributed effect metrics. As we have seen, the use of QC charts in combination with CMA allows more objective evaluation of cumulative evidence and temporal trends in effect sizes. Significant changes in effect with time have been detected in two out of three examined meta-analytical datasets. The $\bar{X}$ charts allowed the detection of outliers and the CUSUM charts were sensitive to temporal changes in the effect. Moreover, the CUSUM charts have allowed earlier detection of shifts as compared to the CMA approach as well as successfully detected significant deviation from the target value even quite late in the process of data accumulation. Such changes are unlikely to be detected by the CMA approach which is purely visual and has more 'inertia' (i.e. it is more difficult to shift the value of the cumulative mean when many studies have been added to the analysis). Other advantages of the QC methods include the ease of interpretation and readily available software.

This study is the first demonstration of application of QC charts to meta-analytic data and we can envisage several ways in which the use of QC charts in meta-analysis can be developed further. We suggest a two-stage approach in use of the QC methods in meta-analysis, similar to their industrial applications. At the first set-up stage the CUSUM charts are to be used to establish significance and stability of the accumulated evidence. At the subsequent monitoring stage the established effect is to be monitored for temporal changes. This usage of the CUSUM techniques was demonstrated in Section 5.1. We find the $\bar{X}$ charts to be a useful aid to interpretation of changes on CUSUM charts.

Another recently proposed class of statistical methods for assessment of accumulating evidence is based on trial sequential methods (Brok *et al*., 2008; Wetterslev *et al*., 2008). Both QC charts and trial sequential methods are assuming normality of the effect sizes. It would be of interest to compare both methods on a large variety of meta-analytical data, and also by simulation. It may be necessary to develop modifications more applicable to particular effect sizes. As one of the examples, we discussed an application of QC charts to log odds ratios. This approach may have some deficiencies when the risks are low, and the data are sparse. It is well known that in these circumstances the inverse variance weights do not work well and the Mantel–Hansel methods are to be preferred (Sweeting *et al*., 2004). The quality of the resulting QC charts is to be studied elsewhere. It may be possible to develop Mantel–Hansel-type QC charts.

In all our examples cumulative meta-analyses were based on REM. In contrast, all QC charts were based on the variances $\hat{\sigma}_i^2$ of the effects that correspond to fixed-effect model of meta-analysis. Although we used the combined effect obtained by REM as the centerline in examples 2 and 3, we have not used the REM variances of the effects given by $\hat{\sigma}_i^2 + \hat{\tau}^2$, where τ^2 is the random variance component. Such a choice would introduce dependencies between variance estimates and may affect the statistical properties of the QC charts. Further research is required to develop the random-effects QC charts.

Another important direction of further research is the use of QC charts with an adjustment for covariates/moderators. Risk-adjusted QC charts were developed in medical outcomes monitoring context with a goal of adjusting for patient risk (Grigg and Farewell, 2004). An application to meta-analysis is to be studied elsewhere.

In general, we recommend that research synthesists always explore temporal patterns in effect sizes when conducting a meta-analysis, particularly when the total heterogeneity of effect sizes is significant and the data set offers a sufficient temporal span (at least 10 years). QC charts can be used along with the CMA plots to assess the significance of the accumulating evidence, or to detect temporal changes in effect sizes. Examination of temporal trends in effect sizes should become a routine procedure in meta-analyses and is useful for several reasons. First, exploration of temporal trends in meta-analytic data set is crucial for the assessment of stability of the results of the analysis and sufficiency of the data. Second, examination of temporal trends in effect sizes is a good diagnostic tool for detection of sources of heterogeneity. It might also allow early detection of the point in time when the evidence is sufficient and stable enough to provide the basis for practical recommendations. It may also facilitate timely detection of temporal changes in effect sizes, which might indicate the need for changes in previously accepted policies. The lack of stability in effect sizes over time could also be used as justification for the need for more research on the topic.

Acknowledgements

The authors are grateful to Kari Saikkonen for providing the effect size data from their meta-analysis.

References

Altman DG, Bland JM. 2003. Interaction revisited: the difference between two estimates. *Br Med J* **326**: 219.

Bagos PG, Nikolopoulos GK. 2009. Generalized least squares for assessing trends in cumulative meta-analysis with applications in genetic epidemiology. *J Clin Epidemiol* **62**: 1037–1044.

Baker R, Jackson D. 2010. Inference for meta-analysis with a suspected temporal trend. *Biometrical J* **52**: 538–551.

Barto EK, Rillig MC. 2010. Does herbivory really suppress mycorrhiza? A meta-analysis. *J Ecol* **98**: 745–753.

Brok J, Thorlund K, Gluud C, Wetterslev J. 2008. Trial sequential analysis reveals insufficient information size and potentially false positive results in many meta-analyses. *J Clin Epidemiol* **61**: 763–769.

Farringdon JM. Morton AQ, Farringdon MG, Baker MD. 1996. *Analysing for Authorship*: *A Guide to the CUSUM Technique*. University of Wales Press, Cardiff.

Fisher RA. 1915. Frequence distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* **10**: 507–521.

Gardner TA, Côté IM, Gill JA, Grant A, Watkinson AR. 2003. Long-term region-wide decline in Carribean corals. *Science* **301**: 958–960.

Gehr BT, Weiss C, Porzsolt F. 2006. The fading of reported effectiveness. A meta-analysis of randomised controlled trials. *BMC Med Res Methodol* **6**: 25.

Grabe S, Ward LM, Hyde JS. 2008. The role of the media in body image concerns among women: a meta-analysis of experimental and correlational studies. *Psychol Bull* **134**: 460–476.

Grigg OA, Farewell V. 2004. An overview of risk-adjusted charts. *J R Stat Soc Ser A* **167**: 523–539.

Grigg OA, Spiegelhalter DJ. 2005. Random-effects CUSUMs to monitor hospital mortality. *Statistical Solutions to Modern Problems*: *Proceedings of the 20th International Workshop on Statistical Modelling*. University of Western Sydney Press: Sydney, Australia, 239–246.

Grigg OA, Spiegelhalter DJ. 2008. An empirical approximation to the null unbounded steady-state distribution of the cusum statistic. *Technometrics* **50**: 501–511.

Gurevitch J, Morrison JA, Hedges LV. 2000. The interaction between competition and predation: a meta-analysis of field experiments. *Am Nat* **155**: 435–453.

Hadjidja R, Debbabi M, Lounisa H, Iqbala F, Szporera A, Benredjema D. 2009. Towards an integrated E-mail forensic analysis framework. *Digit Invest* **5**: 124–137.

Hawkins DM, Olwell DH. 1997. *Cumulative Sum Charts and Charting for Quality Improvement*. Springer, Berlin.

Ioannidis JP, Lau J. 2001. Evolution of treatment effects over time: empirical insight from recursive cumulative metaanalyses. *Proc Natl Acad Sci* **98**: 831–836.

Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG. 2001. Replication validity of genetic association studies. *Nat Genet* **29**: 281–291.

Ioannidis JP, Trikalinos TA. 2005. Early extreme contradictory estimates may appear in published research: the proteus phenomenon in molecular genetics research and randomized trials. *J Clin Epidemiol* **58**: 543–549.

Jennions MD, Møller AP. 2002. Relationships fade with time: a meta-analysis of temporal trends in publication in ecology and evolution. *Proc R Soc Lond B* **269**: 43–48.

Kampichler C, Bruckner A. 2009. The role of microarthropods in terrestrial decomposition: a meta-analysis of 40 years of litterbag studies. *Biol Rev* **84**: 375–389.

Kulinskaya E, Morgenthaler S, Staudte RG. 2008. *Meta Analysis*: *A Guide to Calibrating and Combining Statistical Evidence*. Wiley, New York.

Lau J, Antman EM, Jimenez-Silva J, Kupelnick B, Mosteller F, Chalmers TC. 1992. Cumulative meta-analysis of therapeutic trials for myocardial infarction. *N Engl J Med* **327**: 248–254.

Leimu R, Koricheva J. 2004. Cumulative meta-analysis: a new tool for detection of temporal trends and publication bias in ecology. *Proc R Soc Lond B* **271**: 1961–1966.

Marshall C, Best N, Bottle A, Aylin P. 2004. Statistical issues in the prospective monitoring of health outcomes across multiple units. *J R Stat Soc A* **167**(3): 541–559.

Montgomery DC, Runger GC. 1994. *Applied Statistics and Probability for Engineers*. Wiley, New York, Chichester.

Nykänen H, Koricheva J. 2004. Damage-induced changes in woody plants and their effects on insect herbivore performance: a meta-analysis. *Oikos* **104**: 247–268.

Page ES. 1954. Continuous inspection schemes. *Biometrika* **41**: 100–115.

Saikkonen K, Lehtonen P, Helander M, Koricheva J, Faeth SH. 2006. Model systems in ecology: dissecting the endophyte-grass literature. *Trends Plant Sci* **11**: 428–433.

Schwarzer G. 2010. meta, v.1.5-0. CRAN, July 2010. R package. Fixed and random effects meta-analysis. Functions for tests of bias, forest and funnel plot.

Scrucca L. 2004. qcc: an R package for quality control charting and statistical process control. *R News* **4/1**: 11–17. Available from: http://CRAN.R-project.org/doc/Rnews/.

Stead LF, Perera R, Bullen C, Mant D, Lancaster T. 2008. Nicotine replacement therapy for smoking cessation. *Cochrane Database of Syst Rev*, Issue 1, CD000146.

Sweeting MJ, Sutton AJ, Lambert PC. 2004. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. *Stat Med* **23**: 1351–1375.

Torres-Vila LM, Jennions MD. 2005. Male mating history and female fecundity in the lepidoptera: do male virgins make better partners? *Behav Ecol Sociobiol* **57**: 318–326.

Trikalinos TA, Churchill R, Ferri M, Leucht S, Tuunainen A, Wahlbeck K, Ioannidis JPA. 2004. Effect sizes in cumulative meta-analyses of mental health randomized trials evolved over time. *J Clin Epidemiol* **57**: 1124–1130.

Trikalinos TA, Ioannidis JPA. 2005. Assessing the evolution of effect sizes over time. *Publication Bias in Meta-analysis—Prevention, Assessment and Adjustments* (Rothstein HR, Sutton AJ, Borenstein M eds). pp 241–259. Wiley, New York,

Wetterslev J, Thorlund K, Brok J, Gluud C. 2008. Trial sequential analysis may establish when firm evidence is reached in cumulative meta-analysis. *J Clin Epidemiol* **61**: 64–75.

Winkel P, Zhang NF. 2007. *Statistical Development of Quality in Medicine*. Wiley, New York.