International Journal of Engineering, Mathematical and Physical Sciences
ISSN: 2517-9934
Vol:9, No:1, 2015

# A Sequential Approach to Random-Effects Meta-Analysis

Samson Henry Dogo, Allan Clark, Elena Kulinskaya

***Abstract*-** The objective of meta-analysis is to combine results from several independent studies in order to create generalization and provide evidence base for decision making. But recent studies show that the magnitude of effect size estimates reported in many areas of research significantly changed over time and this can impair the results and conclusions of meta-analysis. A number of sequential methods have been proposed for monitoring the effect size estimates in meta-analysis. However they are based on statistical theory applicable only to fixed effect model (FEM) of meta-analysis. For random-effects model (REM), the analysis incorporates the heterogeneity variance, $\tau^2$ and its estimation create complications. In this paper we study the use of a truncated CUSUM-type test with asymptotically valid critical values for sequential monitoring in REM. Simulation results show that the test does not control the Type I error well, and is not recommended. Further work required to derive an appropriate test in this important area of applications.

***Keywords:*** Meta-analysis, random-effects model, sequential testing, temporal changes in effect sizes.

## I. Introduction

META-ANALYSIS is a statistical technique used to combine results from related but independent studies in order to estimate an overall treatment effect. It is used in numerous applications to synthesize and strengthen evidence about the treatment efficacy and provide evidence for decision making. Meta-analysis helps to decide when evidence of benefit or harm of a new intervention is statistically significant and scientifically convincing to adopt or reject the investigated treatment [4], [18], [21], [24]. It may also be used to decide whether enough evidence has been gathered so that further trials are unnecessary. By combining information from several studies meta-analysis allows the sample size to increase and achieve a higher statistical power for the outcome of interest compared to the less precise measures derived from single individual studies.

However recent findings have shown that effect size estimates used in combining results in meta-analysis may change significantly with the year of publication in many areas of research. For example, Hodgson, Parkinson and Karpf [12] found a significant decline in the sensitivity of chest X-rays in detecting hypersensitivity pneumonitis of about 1.4 % per annum, which they attributed to secular trends in knowledge and earlier diagnosis or changes in the disease itself. Nieuwkarm et al. [22] found a decrease in case fatality of aneurysmal sub-arachnoid haemorrhage during the period 1960-1995, which they attributed to improvement in early diagnostic and treatment strategies. Similar temporal changes have also been reported in education [14], medicine [6], psychology [2], [9], [25] to mention but a few. These changes in the trends can be dramatic and often lead to the loss or gain of the statistical significance [16]. Therefore if meta-analysis is conducted by ignoring temporal trends when trends are actually present, its results and conclusions can be impaired and any statistical inference about the treatment effect will be misleading. In order to address this problem, it is important to find appropriate statistical techniques that are able to detect any possible trends in the effect size estimates so that results and conclusions of meta-analysis can be interpreted based on the time it was conducted.

A number of sequential methods have been proposed for monitoring the trends in effect size estimates in meta-analysis, see [1], [11], [16], [19]–[21], [24], [27], [29]. The methods allow to gauge sufficiency of evidence [20], [24], [27] and can be used for monitoring the trends in effect size estimates [15], [16], [21]. However these methods of monitoring effect size estimates are based on the solid statistical theory only in the fixed effect model (FEM) of meta-analysis. For random-effects-model (REM), the analysis incorporates the heterogeneity variance, $\tau^2$ and its estimation creates complications.

In this paper we review the standard sequential methods for meta-analysis and propose the use of Gombay [7] truncated CUSUM-type test in which the heterogeneity parameter, $\tau^2$ is treated as a nuisance parameter- a parameter that is not of immediate interest but must be accounted for in the course of the analysis. The proposed method has solid statistical foundations and may therefore constitute a better and more efficient sequential approach to monitoring effect size estimates in random-effects meta-analysis. The rest of the paper is organised as follows. In section II, we review the existing sequential methods for meta-analysis. In section III, we formulate the Gombay test statistic for random effects model. In section IV, we report on a simulation study to evaluate the performance of the new method. Section V is the summary and conclusions.

## II. Standard Sequential Methods for Meta-Analysis

Before reviewing the existing standard sequential methods for meta-analysis used in monitoring temporal changes in effect sizes we present the models used in combining the results from studies.

### A. Fixed-Effect and Random-Effects Models

Consideration for a meta-analytic model is the choice between fixed- and random-effects models. Fixed effect model (FEM) in meta-analysis assumes that all the included studies investigate the same population and therefore share a single common parameter. Assume that $y_1$, $y_2$, ..., $y_K$ are the estimates of treatment effects derived from studies. The fixed effect model is given by

$$y_i = \theta + e_i, \tag{1}$$

where $\theta$ is the common parameter, $e_i \sim N(0, v_i^2)$ is the sampling error and $v_i^2$ is the variance. Estimates of $v_i^2$ values are easily

S. H. Dogo, A. Clark and E. Kulinkaya are with the University of East Anglia, Norwich, United Kingdom. (e-mail: s.dogo@uea.ac.uk; Allan.Clark@uea.ac.uk; E.Kulinskaya@uea.ac.uk).

calculated for all effect sizes used in meta-analysis and are treated as known constants [26]. In FEM, each study is assigned a weight proportional to the inverse of its variance which we denote by $w_i = 1/v_i^2$. The combined effect is estimated as a weighted mean of the individual effect estimates, $\hat{\theta}_{FEM} = \sum_i y_i w_i / \sum_i w_i$. Standard inference about the combined effect is based on normality of its distribution, $\hat{\theta}_{FEM} \sim N\left(\theta, (\sum_i w_i)^{-1}\right)$. To test the hypothesis for the presence of a treatment effect, the Wald's statistic is compared with the critical values for the standard normal distribution.

Random-effects model (REM) assumes that each effect size estimate $y_i$ estimate a different effect size parameter, $\theta_i$ with error $e_i$, and that the parameter, $\theta_i$ is sampled from a population of parameters with mean $\theta$. The random-effects model is a two level model given by

$$\begin{aligned} y_i &= \theta_i + e_i;\ e_i \sim N(0, v_i^2) \\ \theta_i &= \theta + \epsilon_i;\ \epsilon_i \sim N(0, \tau^2), \end{aligned} \tag{2}$$

where $v_i^2$ and $\tau^2$ are the within- and between-study variances, respectively. Combining the two equations in (2), the random effects model is defined by

$$y_i = \theta + \xi_i;\ \xi_i \sim N(0, \tau^2 + v_i^2). \tag{3}$$

The between-study variance, $\tau^2$ describes the degree of inconsistency among the effect estimates. The special case $\tau^2 = 0$ implies that the effect size estimates ($y_1$, $y_2$, ....) are homogeneous [26], and the resulting model reduces to FEM in (1). The weights assigned to studies in REM are calculated using a variance component that incorporates the between study variance in addition to the within-study variance used in fixed-effect model. We denote the weights in REM by $w_i^* = (\tau^2 + v_i^2)^{-1}$. The combined effect is estimated as weighted mean of the individual effect estimates, $\hat{\theta}_{REM} = \sum_i w_i^* y_i / \sum_i w_i^*$. As in FEM, standard inference about the combined effect is based on the normality of its distribution, $\hat{\theta}_{REM} \sim N\left(\theta, (\sum_i w_i^*)^{-1}\right)$. To test the hypothesis for the presence of a treatment effect, the Wald's statistic is compared with the critical values for the standard normal distribution.

Estimation of the between-study variance is crucial in random-effects meta-analysis. Consequently a number of methods have been proposed to estimate $\tau^2$, see [3], [5], [10]. The most commonly used methods include DerSimonian and Laird [4]; Mandel and Paule [23] and the restricted maximum likelihood (REML) methods which are described below along with the method by Higgins, Whitehead and Simmonds [11] proposed specifically for sequential testing in meta-analysis.

*1) DerSimonian and Laird [4] Method:* The DerSimonian and Laird [4] estimator is given by

$$\hat{\tau}_{DL}^2 = \frac{Q - (K-1)}{C}, \tag{4}$$

where $Q = \sum_{i=1}^{K} w_i (y_i - \hat{\theta})^2$ and $C = \sum_{i=1}^{K} w_i - \frac{\sum_{i=1}^{K} w_i^2}{\sum_{i=1}^{K} w_i}$.

*2) Higgins, Whitehead and Simmonds [11] Method:* The Higgins, Whitehead and Simmonds [11] estimator is a modification of the DerSimonian and Laird [4] method using semi-Bayes approach. It is defined by

$$\hat{\tau}_H^2 = \frac{2\lambda + K\hat{\tau}_{DL}^2}{2\eta + K - 2}, \tag{5}$$

where $\lambda$ and $\eta$ are parameters of a prior inverse gamma distribution for $\tau^2$.

*3) Mandel and Paule [23] Method:* The Mandel and Paule [23] estimator of $\tau^2$ is calculated from the solution of the estimating equation for the expected value of the $Q$ statistic under $H_0$ given by

$$Q(\hat{\tau}_{MP}^2) = \sum_{i=1}^{K} w_i^*(\hat{\tau}_{MP}^2) \left(y_i - \hat{\theta}(\hat{\tau}_{MP}^2)\right)^2 - (K-1) = 0, \tag{6}$$

where $\hat{\theta}(\hat{\tau}_{MP}^2)$ and $w_i^*(\hat{\tau}_{MP}^2)$ are functions of $\hat{\tau}_{MP}^2$.

*4) Restricted maximum likelihood Method:* The restricted maximum likelihood (REML) estimator of $\tau^2$ is given by

$$\hat{\tau}_{REML}^2 = \frac{\sum_{i=1}^{K} w_i^{*2} \left[(y_i - \hat{\theta})^2 - v_i^2\right]}{\sum_{i=1}^{K} w_i^{*2}} + \frac{1}{\sum_{i=1}^{K} w_i^*}. \tag{7}$$

Each of these methods differs in terms of precision and bias in estimating $\tau^2$, and thus can have a different effect on the sequential testing. We examine this using Monte Carlo simulations in Section IV.

### *B. Standard Methods for Monitoring Temporal Trends in Meta-analysis*

Several sequential methods for meta-analysis have been proposed for monitoring temporal changes in magnitude of effect sizes. The first is cumulative meta-analysis (CMA) which can be described as an open sequential test. It was initially proposed by Lau et al. [20] as a method to identify when a treatment effect in a clinical trial is statistically significant as early as possible. It is routinely used for monitoring temporal changes in effect sizes [15], [20], [21]. The technique involves pooling of effect size estimates in a cumulative manner as new trial results are published. When results from studies are arranged in a chronological sequence according to year of publication, the plotted values of the combined effects, $\hat{\theta}_k$ and confidence intervals calculated consecutively for k=1, 2, ..., K can reveal patterns, uneven irregular and non-linear shifts in opposite direction [16]. However the technique by definition involves multiple looks on the accumulating evidence and the continuing addition of new studies and multiple testing leads to the inflation of the overall Type I error in the analysis.

The second group of methods is the sequential meta-analysis (SMA). These methods use formal group-sequential boundaries to monitor cumulative meta-analysis. The method proposed by Pogue and Yusuf [24] is aimed to address the issue of inflated Type I error in CMA. The SMA approach involves calculating an optimum information size (OIS) and then determines the monitoring boundaries using alpha spending function and stochastic curtailment. But the calculation of the OIS is based on fixed effect model and thus the method cannot be used for REM. Wetterslev, Thorlund, Brok and Gluud [27] used a heterogeneity inflated OIS to account for heterogeneity in treatment effects, but this method is problematic [17]. Whitehead [28] describes the use of the standard stopping boundaries for random-effects meta-analysis. Bollen, Uiterwaal, Vught and Van der Tweel [1] used the double triangular test in a retrospective meta-analysis. Higgins, Whitehead and Simmonds [11] proposed a sequential method for random-effects meta-analysis that uses a semi-Bayes procedure to update evidence on the among-study variance, starting with an informative prior distribution that may be based on findings from a previous meta-analysis. Monitoring boundaries of formal group sequential methods are generally defined based on fixed effect approach and do not incorporate the presence of heterogeneity in treatment effects. Simulations on these methods have shown a considerable inflation of the Type I error when the values of $\tau^2$ are large, see [11], [27]. Therefore using such methods for random-effects model can lead to spurious statistical inference.

A method recently introduced by Kulinskaya and Koricheva [16] is based on the use of quality control charts for detection of outliers and temporal trends in meta-analysis. The use of QC charts in meta-analysis is straightforward once the distribution of the effect estimates can be approximated by the normal distribution. However the method has so far been used in fixed effect model, for random-effects model the estimation of $\tau^2$ can introduce dependency between the variance estimates and the sequential estimates of the effects [16], which is not consistent with the standard assumptions of the QC charts.

Another interesting approach is the 'penalized Z test' introduced by Lan, Hu and Cappelleri [19] as an alternative way to address the issue of inflated Type I error in cumulative meta-analysis. The method is using the law of iterated logarithm to 'penalize' for the multiple testing in CMA. The usual Z-statistic is adjusted and at the kth interim analysis is defined by

$$Z^*(k)\frac{S(k)}{\sqrt{\lambda\Gamma_k \log\log(\Gamma_k)}}, \tag{8}$$

where $\lambda$ is the adjustment factor determined using simulation, $S(k)$ is the sum of the estimates of treatment effects up to the kth interim analysis and $\Gamma_k$ is the sum of weights assigned to studies. The 'penalized Z test' exhibits a good control of the Type I error in CMA both in FEM and REM when a reasonable value of $\lambda$ is used. For example the value of $\lambda = 1.5$ is found to control the Type I error in FEM while the value of $\lambda = 2$ is found to control the Type I error in REM when relative risks, odds ratio and risks difference effect sizes are used to combine results of up to 25 studies [13]. The constant $\lambda$ is an important factor in controlling the Type I error, however its value varies according to the type of effect size, number of studies, average studies size and amount of heterogeneity in the treatment effects. Therefore the determination of the 'reasonable value of $\lambda$' can be difficult in practice .

## III. Formulation of the Gombay Test Statistic

In this section we briefly describe the Gombay method and formulate the Gombay test statistic for random-effects meta-analysis on which the sequential methods are based.

### A. Gombay Method

The Gombay method described in [7], [8] is a truncated CUSUM-type test used for sequential change detection in parametric models involving a nuisance parameter. For simplicity we shall refer to this method as the Gombay test. Consider the sequence of variables $X_1$, $X_2$, .... $\sim f_{\theta_i,\eta_i}$, where $f$ is a probability density function, $\theta$ is a parameter of interest and $\eta$ is a nuisance parameter. The Gombay test is a test for the composite hypothesis

$H_0$: $\theta_i = \theta$, $\eta_i = \eta$; $i = 1, 2, ....$

$$H_1\text{: } \begin{cases} \theta_i = \theta_0, \eta_i = \eta; & i = 1, 2, ...r \\ \theta_i = \theta_1, \eta_i = \eta; & i \geq r+1 \end{cases};$$

where r is an unknown time of change, and the values of $\theta_1$ and $\eta$ are also unknown. In order to define a test statistic for the hypotheses, a Fisher information matrix, I is partitioned as

$$I = \begin{pmatrix} I_{\theta\theta} & I_{\theta\eta} \\ I_{\eta\theta} & I_{\eta\eta} \end{pmatrix},$$

where

$$I_{11} = \left(-\mathrm{E}\tfrac{\partial^2}{\partial\theta^2}\log f_{\theta\eta}\right),\ I_{22} = \left(-\mathrm{E}\tfrac{\partial^2}{\partial\eta^2}\log f_{\theta\eta}\right) \text{ and}$$
$$I_{12} = I_{21}^t = \left(-\mathrm{E}\tfrac{\partial^2}{\partial\theta\partial\eta}\log f_{\theta\eta}\right).$$

Denote $\psi = (\theta, \eta)$. The efficient score vector for $\theta$ and $\eta$ at the $k^{th}$ interim analysis is defined by

$$V_k(\theta,\eta) = \sum_{i=1}^{k} \frac{\partial}{\partial\psi}\log f_{\theta_i\eta_i} \tag{9}$$

Replacing the nuisance parameter, $\eta$ with its restricted maximum likelihood estimate, $\hat{\eta}_k$ obtained from the solution of

$$\sum_{i=1}^{k} \frac{\partial}{\partial\eta}\log f(X_i : \theta_0, \eta) = 0. \tag{10}$$

The efficient score vector, $V_k$ is given by

$$V_k(\theta,\hat{\eta}_k) = \sum_{i=1}^{k} \frac{\partial}{\partial\theta}\log f_{\theta_i\hat{\eta}_k}. \tag{11}$$

Under some regularity conditions in $H_0$, Gombay and Serbian [8] showed that as $k \to \infty$, the efficient score vector

$$\begin{aligned} V_k(\theta,\hat{\eta}_k) =& \sum_{i=1}^{k} \frac{\partial}{\partial\theta}\log f_{\theta\hat{\eta}_k} \\ =& \sum_{i=1}^{k}\left\{\frac{\partial}{\partial\theta}\log f_{\theta_0\eta}\right\} \\ & - \sum_{i=1}^{k}\left\{\frac{\partial}{\partial\eta}\log f_{\theta_0\eta} I_{22}^{-1}(\theta_0,\eta) I_{21}(\theta_0,\eta)\right\} \\ & + O(\log\log k) \\ =& \sum_{i=1}^{k} Z_i + O(\log\log k), \end{aligned} \tag{12}$$

where $Z_i$ are i.i.d.r.v's with expected value $\mathrm{E}[Z_i] = 0$ and $\mathrm{cov}(Z_i) = \Gamma_k(\theta_0,\eta)$ for $\Gamma_k(\theta_0,\eta) = I_{11} - I_{12}I_{22}^{-1}I_{21}$. It follows that the statistic

$$T_k = \frac{\sum_{i=1}^{k}\frac{\partial}{\partial\theta}\log f_{\theta_0,\hat{\eta}_k}}{\sqrt{\Gamma_k(\theta_0,\eta)}} \tag{13}$$

is asymptotically ($k \to \infty$) the sum of independent random variables with mean 0 and variance equal to 1, and thus can be approximated by a standard Wiener's process. In order to use the statistic $T_k$ for testing hypotheses the covariance $\Gamma_k(\theta_0,\eta)$ is replaced with its estimate $\Gamma_k(\theta_0,\hat{\eta}_k)$. Gombay [7] introduced a sequential change detection test defined using statistic $T_k$ in (13) as follows. For $k = 2, 3, \cdots, K$, where $K$ is a truncation point, reject $H_0$ if

$$G(K) = {}^{\max}_{1<k\leq K}\ \frac{1}{\sqrt{k}}T_k \geq C(\alpha) \tag{14}$$

and if no such k, $k \leq K$, exists do not reject $H_0$. The critical values of the 1-sided test are given by

$$\begin{aligned} C(\alpha) =& (2\log\log K)^{-\frac{1}{2}}(-\log(-\log(1-\alpha)) + 2\log\log K \\ & + \frac{1}{2}\log\log K - \frac{1}{2}\log\pi). \end{aligned} \tag{15}$$

For a two-sided test based on $|T_k|$, the critical values are given by

$$\begin{aligned} C^*(\alpha) =& (2\log\log K)^{-\frac{1}{2}}(-\log(-\frac{1}{2}\log(1-\alpha)) + 2\log\log K \\ & + \frac{1}{2}\log\log K - \frac{1}{2}\log\pi). \end{aligned} \tag{16}$$

See [7], [8] for a detailed derivation and discussion on the Gombay method.

### *B. Application of the Gombay Method to Random Effects Model*

To apply the Gombay method in random effects model of meta-analysis, consider a sequence of independent studies conducted over time. Each study estimates a treatment effect, $y_i$ for i=1, 2, .... with variance $v_i^2$. We assume that there is no correlation between the effect size estimates and the variances. Under the null hypothesis, $H_0$, each effect estimate is normally distributed with the same mean, $y_i \sim N\left(\theta, (\hat{w}_i^*)^{-1}\right)$, where $\hat{w}_i^* = (\tau^2 + v_i^2)^{-1}$ is the estimate of the weight in random effects model. The mean parameter, $\theta$ is the population treatment effect and it is estimated as weighted mean of the individual effect estimates, $\hat{\theta}_k = \sum_{i=1}^{k} \hat{w}_i^* y_i / \sum_{i=1}^{k} \hat{w}_i^*$, k=1, 2, ..... Let $\theta = \theta_0$ be the target value of the effect parameter. As more studies are conducted and results are continually combined, the goal is to determine when the combined effect, $\hat{\theta}_k$ changes significantly from the target value, $\theta_0$ and stop further studies.

The log likelihood function of $y_i$ required to define the Gombay test statistic is given by

$$L\left(y_i : \theta, \tau^2\right) = \frac{1}{2}\Big\{ \log \hat{w}_i^* - \hat{w}_i^*(y_i - \theta_0)^2 + C \Big\}, \qquad (17)$$

where $C$ is a constant. Equation (13) results in the statistic

$$T_k = \frac{\sum\limits_{i=1}^{k} \hat{w}_i^*(y_i - \theta_0)}{\sqrt{\sum\limits_{i=1}^{k} \mathrm{E}[\hat{w}_i^*]}}, \qquad (18)$$

which, as follows from (12) is asymptotically a sum of independent random variables with mean 0 and variance equal to 1 and can be approximated as a standard Wiener's process. Because the probability distribution of $\hat{\tau}^2$ is unknown, the expected value of the weight estimate, $\hat{w}_i^*$ in (18) needs to be approximated. Assuming that the expected value, $\mathrm{E}[\hat{\tau}_i^2] = \tau^2$ for i=1, 2, ..., K, the expected value of the weight estimates can be approximated by the first term in its Taylor series expansion, $\mathrm{E}[\hat{w}_i^*] = w_i^*(\tau^2)$. The between-study variance component $\tau^2$ in its term is estimated using the best estimate available from all K studies, $\hat{\tau}_K^2$. Therefore we use $\mathrm{E}[\hat{w}_i^*] = w_i^*(\hat{\tau}_K^2)$ in (18).

Therefore if it is desired at the beginning of the sequential meta-analysis that a decision to accept or reject the existence of a shift (which may be equivalent to the existence of a treatment effect) is to be made after combining a maximum of $K$ studies, the one-sided Gombay test statistic from (14) is defined for random-effects meta-analysis by

**Test**: For k=2, 3, ..., K, reject $H_0$ if

$$T_k = \frac{\sum\limits_{i=1}^{k} w_i^*(\hat{\tau}_k^2)(y_i - \theta_0)}{\sqrt{\sum\limits_{i=1}^{k} w_i^*(\hat{\tau}_K^2)}} \geq \sqrt{k}C(\alpha) \qquad (19)$$

and if no such k, $k \leq K$, exists do not reject $H_0$. The critical values are given in (15) or (16) for 1-sided or 2-sided test, respectively. The 2-sided test is based on $|T_k|$ values. As mentioned earlier we assess the behaviour of this test when $\tau^2$ is estimated by one of the methods by DerSimonian and Laird [4]; Higgins, Whitehead and Simmonds [11]; Mandel and Paule [23] and the REML. In what follows, the Gombay tests for REM based on the four above estimators are denoted by $GDL$, $GH$, $GMP$ and $GREML$, respectively.

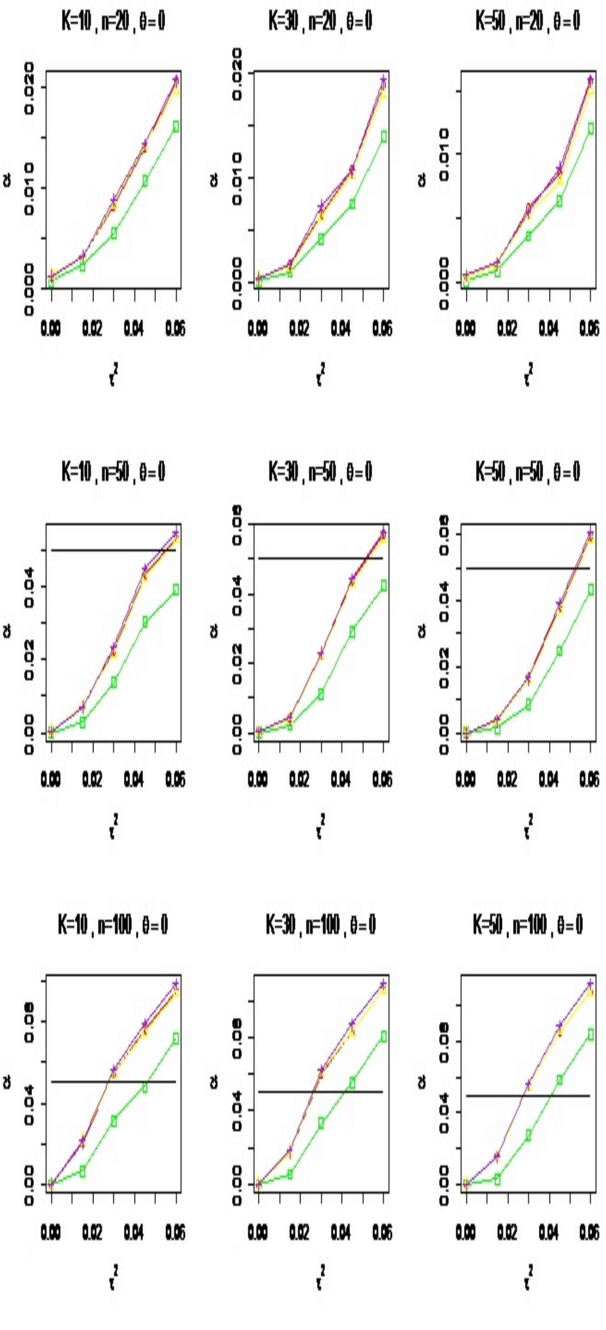

Fig. 1. Overall Type I error achieved by the Gombay tests for REM based on DerSimonian and Laird [4]; Higgins, Whitehead and Simmonds [11]; Mandel and Paule [23] and the REML estimators of $\tau^2$ (GDL -red line, GH - green, GMP - yellow and GREML - purple line, respectively). $K$ is the number of studies included in the meta-analysis; $n$ is the average sample size of studies; $\theta$ is the value of the effect parameter, $\tau^2$ is the value of the between-study variance. The black straight line represents the nominal 0.05 level of the test.

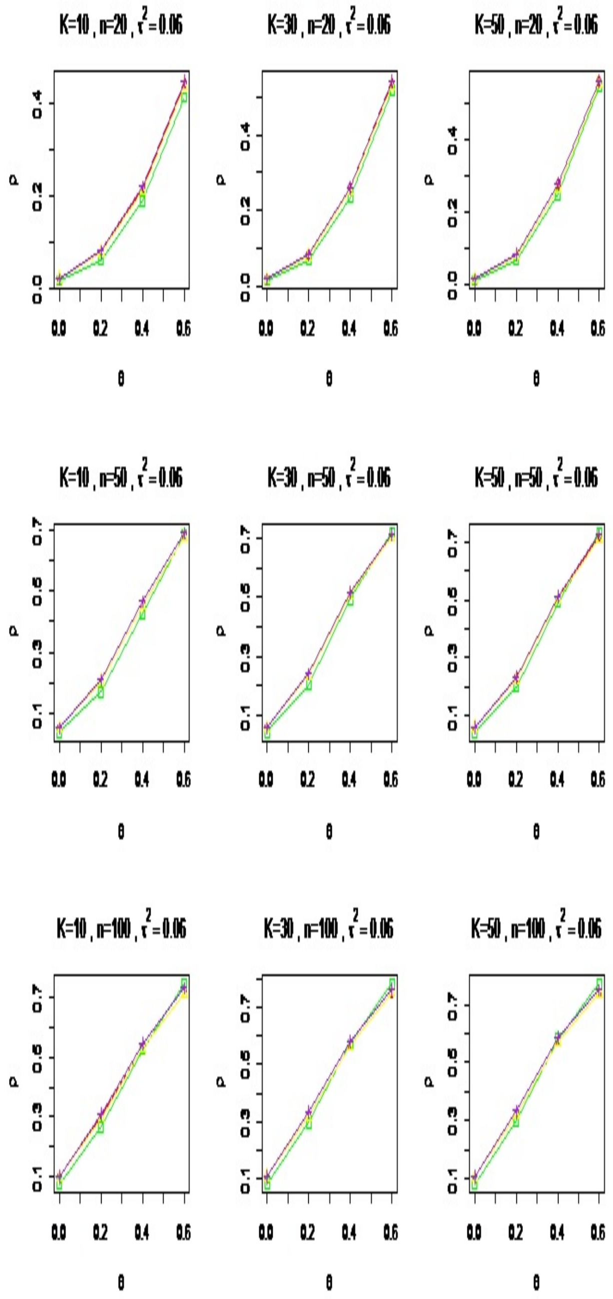

Fig. 2. Comparison of power of the Gombay tests for REM based on DerSimonian and Laird [4]; Higgins, Whitehead and Simmonds [11]; Mandel and Paule [23] and the REML estimators of $\tau^2$ (GDL -red line, GH - green, GMP - yellow and GREML - purple line, respectively). $K$ is the number of studies included in the meta-analysis; $n$ is the average sample size of studies; $\rho$ on the y-axis is the power while $\theta$ on the x-axis is the value of the effect parameter; $\tau^2 = 0.06$ is the between-study variance.

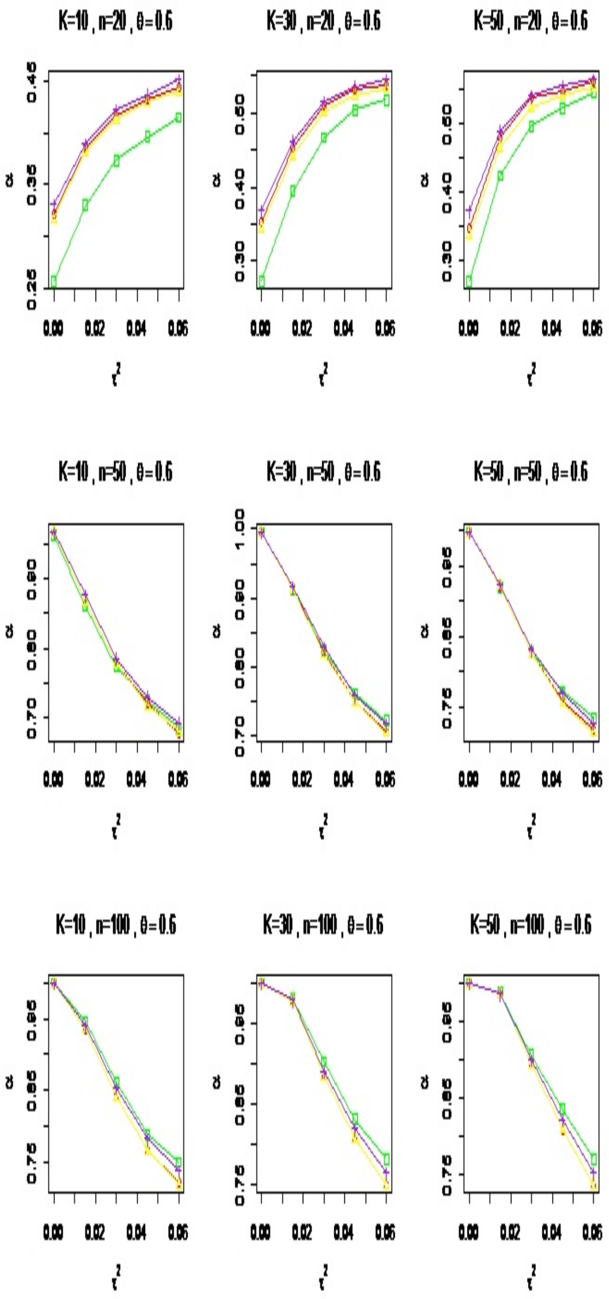

Fig. 3. Comparison of power of the Gombay tests for REM based on DerSimonian and Laird [4]; Higgins, Whitehead and Simmonds [11]; Mandel and Paule [23] and the REML estimators of $\tau^2$ (GDL -red line, GH - green, GMP - yellow and GREML - purple line, respectively). $K$ is the number of studies included in the meta-analysis; $n$ is the average studies size; $\rho$ on the y-axis is the power while $\tau^2$ on the x-axis is the between-study variance. $\theta = 0.6$ is the value of the effect parameter.

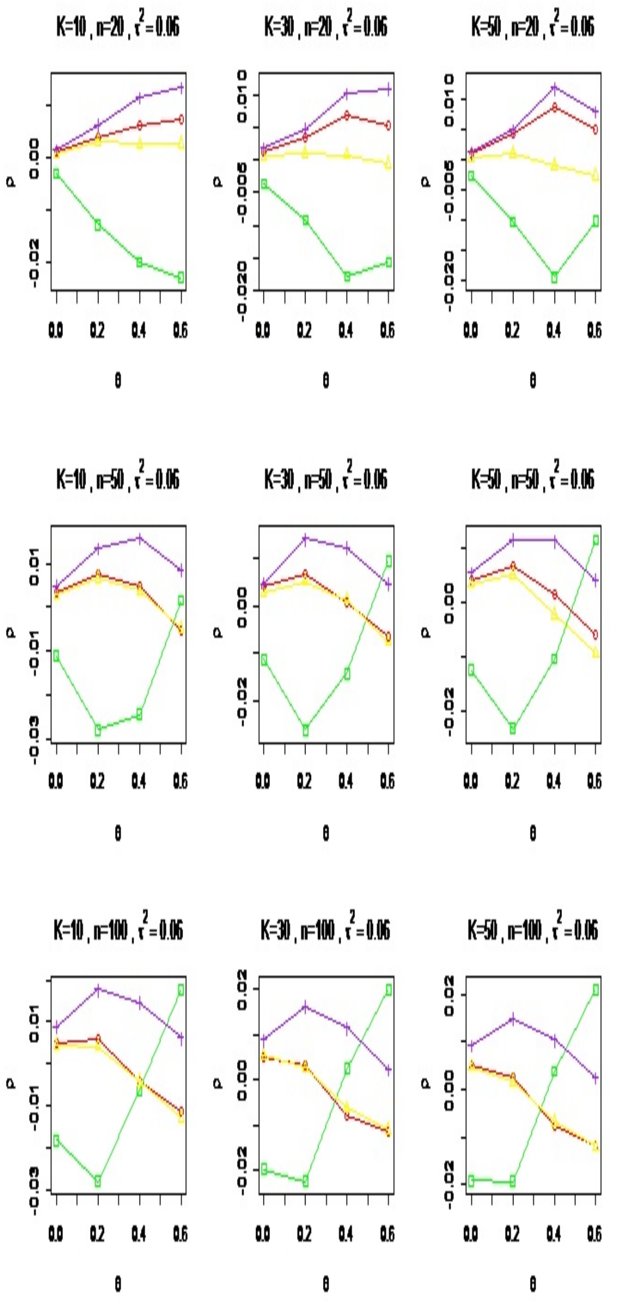

Fig. 4. Comparison of power of the Gombay tests for REM based on DerSimonian and Laird [4]; Higgins, Whitehead and Simmonds [11], Mandel and Paule [23] and the REML estimators of $\tau^2$ (GDL -red line, GH - green, GMP - yellow and GREML - purple line, respectively). $K$ is the number of studies included in the meta-analysis; $n$ is the average studies size; $\rho$ on the y-axis is the deviation in power from the average power of the four tests while $\theta$ on the x-axis is the value of the effect parameter; $\tau^2 = 0.06$ is the between-study variance.

## IV. Simulation Study

The objectives of the simulation is to evaluate the overall Type I error and power the four Gombay tests introduced in Section II in relation to the number of studies $K$ in the meta-analysis, average studies sizes $n$, the amount of heterogeneity in the treatment effects $\tau^2$ and how the four different estimators of $\tau^2$ affect the test. To generate $K$ studies of average size $n$, the sample sizes of the studies $n_i$ (the sample size of the study $i$), $i = 1, \cdots, K$ are generated from the normal distribution, $n_i \sim N\left(n, \frac{n}{4}\right)$ rounded to the nearest integer and truncated at 3. The estimates of sample variances, $v_i^2$ are generated from the Chi-squared distribution, $v_i^2 \sim \frac{v^2}{(n-1)}\chi^2_{n-1}$. The effect sizes are generated from the normal distribution, $y_i \sim N\left(\theta, \sqrt{v_i^2 + \tau^2}\right)$. We calculate the critical values of the test based on an alpha level of 5 % and the null value of the effect parameter set at $\theta_0 = 0$. The sequential testing starts with a minimum of two studies and stops as soon as a boundary value is reached or after the $K^{th}$ interim analysis. For each combination of the following variables: $v^2 = 1$, $\theta = (0.00, 0.20, 0.40, 0.60)$, $n = (20, 50, 100)$, $K = (10, 30, 30, 50)$ and $\tau^2 = (0.00, 0.025, 0.075, 0.10)$ we conducted a total of 10,000 simulations, calculated the power of the test to reject $H_0$ and recorded the results.

### A. Type I Error

Achieved Type I error is the most important issue in the evaluation of any test. Fig. 1 shows the overall Type I errors achieved by the four proposed Gombay tests for sequential random effects meta-analysis based on DerSimonian and Laird [4]; Higgins, Whitehead and Simmonds [11], Mandel and Paule [23] and the REML estimators of $\tau^2$ (GDL, GH, GMP and GREML, respectively). When n=20, the values of Type I errors achieved in the test based on all the four estimators is below the nominal level of 0.05. The tests are much too conservative. As n increases to 50, the GDL, GMP and GREML tests cross the nominal 5% level for larger values of $\tau^2$. The achieved level of the GH test is still below the nominal level for all studied values of heterogeneity. When n=100, the Type I errors achieved by the tests based on all the four estimators of $\tau^2$ increase and cross the nominal level when $\tau^2 = 0.025$ for GDL, GMP and GREML and when $\tau^2 = 0.04$ for GH. For all values on $n$ and $\tau^2$, GDL, GMP and GREML produce higher Type I errors compared to GH. In general, the Type I errors increase with increase in $K$, $n$ and $\tau^2$. This is comparable to the performance of the Lan, Hu and Cappelleri [19] method, except that it controls the Type I error in FEM while the Type I errors achieved by the Gombay test are practically zero when $\tau^2 = 0$. Overall, the performance of all four studied tests is disappointing and they are not recommended for use.

### B. Statistical Power

Fig. 2 shows the power of the four proposed Gombay tests based on DerSimonian and Laird [4]; Higgins, Whitehead and Simmonds [11], Mandel and Paule [23] and the REML estimators of $\tau^2$. As expected, the power increases with increase in the number of studies $K$, average study size $n$ and the value of the population treatment effect $\theta$. The differences in the power between the four tests are very small. Fig. 3 demonstrates that the power decreases with increase in heterogeneity $\tau^2$. This should be expected as the increase in variability makes the detection of an effect more difficult. However, counter-intuitively the power increases when n=20. The reason for this is the extreme conservativeness of the Gombay test when $n$ is relatively small, see Fig. 1. To be able to distinguish differences in the power, Fig. 4 compares the power of the tests based on four different estimators of $\tau^2$ when $\tau^2 = 0.06$. When $n = 20$ GREML is more powerful, followed by GDL, GMP and GH is the least powerful. To some extend this is also true for larger values of $n$, however as the value of $\theta$

increases, the power of GH increases and it eventually becomes more powerful compared to the other three tests.

## V. Summary and Conclusions

Meta-analysis is generally accepted as the standard statistical technique for research synthesis that allows generalization of individual studies results and provides evidence base for decision making. However temporal changes reported in many areas of research [2], [6], [12], [14], [22], [25] can be dramatic and lead to the impairment of results and conclusions of meta-analysis. The sequential methods previously proposed for monitoring the trends in effect size estimates (Cumulative meta-analysis, Sequential meta-analysis, the use of QC charts and the penalized Z-testing of CMA) are only effective when dealing with FEM. In sequential random-effects meta-analysis, the analysis incorporates the heterogeneity variance, $\tau^2$ and its estimation create problems, [17]. As an example, Hu, Cappellari and Lan [13] comment that the sequential boundaries obtained via the usual standard Wiener's process approach could inflate the Type I error.

In this paper we proposed the use of [8] truncated CUSUM-type test in which $\tau^2$ is treated as a nuisance parameter. Unfortunately, our simulations show that the test does not control the Type I error. In our simulation results, we have seen that the achieved level is close to zero when the values of $\tau^2$ are extremely small, and in contrast the larger values of $\tau^2$ lead to considerable inflation of the Type I error. Therefore we do not recommend this test for use in practice.

Without the control of type I error, the comparison of power of the tests based on different estimators of $\tau^2$ is not valid, though the test based on REML estimator of $\tau^2$ appears to result in the higher statistical power compared to the tests based on other three estimators considered.

The lack of control of the type I error by the proposed tests is explained by the use of asymptotic approximations based on Wiener's process to obtain the critical values of the tests. However the Gombay [7] method provides a basis for sequential approach to random-effects meta-analysis that can be improved upon. In our future research, we intend to derive bootstrap-based critical values for use with the Gombay tests.

## References

[1] Casper W Bollen, Cuno SPM Uiterwaal, Adrianus J van Vught, and Ingeborg van der Tweel. Sequential meta-analysis of past clinical trials to determine the use of a new trial. *Epidemiology*, 17(6):644–649, 2006.

[2] S. Brugger, J.M. Davis, S. Leucht, and J.M. Stone. Proton magnetic resonance spectroscopy and illness stage in schizophreniaa systematic review and meta-analysis. *Biological psychiatry*, 69(5):495–503, 2011.

[3] Rebecca DerSimonian and Raghu Kacker. Random-effects model for meta-analysis of clinical trials: an update. *Contemporary clinical trials*, 28(2):105–114, 2007.

[4] Rebecca. DerSimonian and Nan. Laird. Meta-analysis in clinical trials. *Controlled clinical trials*, 7(3):177–188, 1986.

[5] Lynn Friedman. Estimators of random effects variance components in meta-analysis. *Journal of Educational and Behavioral Statistics*, 25(1):1–12, 2000.

[6] B. Gehr, C. Weiss, and F. Porzsolt. The fading of reported effectiveness. a meta-analysis of randomised controlled trials. *BMC medical research methodology*, 6(1):25, 2006.

[7] E Gombay. Sequential change-point detection and estimation. *Sequential Analysis*, 22(3):203–222, 2003.

[8] E. Gombay and D. Serbian. An adaptation of Pages CUSUM test for change detection. *Periodica Mathematica Hungarica*, 50(1):135–147, 2005.

[9] Shelly Grabe, L Monique Ward, and Janet Shibley Hyde. The role of the media in body image concerns among women: a meta-analysis of experimental and correlational studies. *Psychological bulletin*, 134(3):460, 2008.

[10] Larry V Hedges and Jack L Vevea. Fixed-and random-effects models in meta-analysis. *Psychological methods*, 3(4):486, 1998.

[11] Julian Higgins, Anne Whitehead, and Mark Simmonds. Sequential methods for random-effects meta-analysis. *Statistics in medicine*, 30(9):903–921, 2011.

[12] M. J. Hodgson, D. K. Parkinson, and M. Karpf. Chest x-ray in hypersensitivity pneumonities: A meta analysis of secular trends. *American journal of industrial medicine*, 16(1):45–53, 1989.

[13] Mingxiu Hu, Joseph C Cappelleri, and KK Gordon Lan. Applying the law of iterated logarithm to control type I error in cumulative meta-analysis of binary outcomes. *Clinical Trials*, 4(4):329–340, 2007.

[14] J.S. Hyde, E. Fennema, and S.J. Lamon. Gender differences in mathematics performance: a meta-analysis. *Psychological bulletin*, 107(2):139, 1990.

[15] John Ioannidis and Thomas A Trikalinos. Early extreme contradictory estimates may appear in published research: The proteus phenomenon in molecular genetics research and randomized trials. *Journal of clinical epidemiology*, 58(6):543–549, 2005.

[16] E. Kulinskaya and J. Koricheva. Use of quality control charts for detection of outliers and temporal trends in cumulative meta-analysis. *Research Synthesis Methods*, 2010.

[17] Elena Kulinskaya and John Wood. Trial sequential methods for meta-analysis. *Research Synthesis Methods*, 2013.

[18] Sofie Kuppens and Patrick Onghena. Sequential meta-analysis to determine the sufficiency of cumulative knowledge: The case of early intensive behavioral intervention for children with autism spectrum disorders. *Research in Autism Spectrum Disorders*, 6(1):168–176, 2012.

[19] KK Gordon Lan, Mingxiu Hu, and Joseph C Cappelleri. Applying the law of iterated logarithm to cumulative meta-analysis of a continuous endpoint. *Statistica Sinica*, 13(4):1135–1146, 2003.

[20] J. Lau, E.M. Antman, J. Jimenez-Silva, B. Kupelnick, F. Mosteller, and T.C. Chalmers. Cumulative meta-analysis of therapeutic trials for myocardial infarction. *New England Journal of Medicine*, 327(4):248–254, 1992.

[21] R. Leimu and J. Koricheva. Cumulative meta–analysis: a new tool for detection of temporal trends and publication bias in ecology. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 271(1551):1961–1966, 2004.

[22] Dennis J Nieuwkamp, Larissa E Setz, Ale Algra, Francisca HH Linn, Nicolien K de Rooij, and Gabriël JE Rinkel. Changes in case fatality of aneurysmal subarachnoid haemorrhage over time, according to age, sex, and region: a meta-analysis. *The Lancet Neurology*, 8(7):635–642, 2009.

[23] Robert C. Paule and John Mandel. Consensus values and weighting factors. *Journal of Research of the National Bureau of Standards*, 87(5):377–385, 1982.

[24] J.M. Pogue and S. Yusuf. Cumulating evidence from randomized trials: utilizing sequential monitoring boundaries for cumulative meta-analysis. *Controlled clinical trials*, 18(6):580–593, 1997.

[25] Jean M. Twenge., Sara. Konrath, Joshua D. Foster., W Keith Campbell., and Brad J. Bushman. Egos inflating over time: A cross-temporal meta-analysis of the narcissistic personality inventory. *Journal of personality*, 76(4):875–902, 2008.

[26] Wolfgang Viechtbauer. Confidence intervals for the amount of heterogeneity in meta-analysis. *Statistics in medicine*, 26(1):37–52, 2007.

[27] J. Wetterslev, K. Thorlund, J. Brok, and C. Gluud. Trial sequential analysis may establish when firm evidence is reached in cumulative meta-analysis. *Journal of Clinical Epidemiology*, 61(1):64–75, 2008.

[28] Anne Whitehead. A prospectively planned cumulative meta-analysis applied to a series of concurrent clinical trials. *Statistics in medicine*, 16(24):2901–2913, 1997.

[29] John Whitehead. *The design and analysis of sequential clinical trials*. John Wiley & Sons, 1997.