Dimitris Mavridis, Irini Moustaki, Melanie Wall, Georgia Salanti

Detecting outlying studies in meta-regression models using a forward search algorithm

Article (Accepted version) (Refereed)

Original citation: Mavridis, Dimitris and Moustaki, Irini and Wall, Melanie and Salanti, Georgia (2017) Detecting outlying studies in meta-regression models using a forward search algorithm. Research Synthesis Methods, 8 (2). pp. 199-211. ISSN 1759-2887

DOI: 10.1002/jrsm.1197

Detecting outlying studies in meta-regression models using a forward search algorithm

Dimitris Mavridis[1,2], Irini Moustaki[3], Melanie Wall[4], Georgia Salanti[1,5]

[1] Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece.

[2] Department of Primary Education, University of Ioannina, Ioannina, Greece.

[3] Department of Statistics, London School of Economics, United Kingdom.

[4] Department of Psychiatry, Columbia University, New York, USA

[5] Institute of Social and Preventive Medicine (ISPM) & Berner Institut für Hausarztmedizin (BIHAM), University of Bern, Bern, Switzerland.

The paper has been accepted in Research Synthesis Methods.

Abstract

When considering data from many trials, it is likely that some of them present a markedly different intervention effect or exert an undue influence on the summary results. We develop a forward search algorithm for identifying outlying and influential studies in meta-analysis models. The forward search algorithm starts by fitting the hypothesized model to a small subset of likely outlier-free studies and proceeds by adding studies into the set one-by-one that are determined to be closest to the fitted model of the existing set. As each study is added to the set, plots of estimated parameters and measures of fit are monitored to identify outliers by sharp changes in the forward plots. We apply the proposed outlier detection method to two real data sets; a meta-analysis of 26 studies that examines the effect of writing-to-learn interventions on academic achievement adjusting for three possible effect modifiers, and a meta-analysis of 70 studies that compares a fluoride toothpaste treatment to placebo for preventing dental caries in children. A simple simulated example is used to illustrate the steps of the proposed methodology and a small scale simulation study is conducted to evaluate the performance of the proposed method.

Keywords : backward methods, Cook’s distance, masking, meta-analysis, swamping, outliers

1. Introduction

There is an acknowledged need for better evidence-based practice to inform public policy (Oxman *et al.,* 2010). Meta-analysis is one of the most influential and powerful techniques underpinning evidence-based practice (Patsopoulos *et al.,* 2005). It synthesizes quantitatively evidence from many studies addressing the same research hypothesis and provides summary results that have increased precision and power compared to individual studies (Hedges, 1985). It may be seen as a two-stage procedure (Deeks *et al.,* 2011). In the first stage, data are extracted from the relevant studies and an effect size along with its variance (or standard error) are computed for each study. In the second stage, a weighted average of the estimated effect sizes is computed. There are two popular models for the second stage; the fixed effect and the random effects model (Borenstein *et al.,* 2010;Nikolakopoulou *et al,.* 2014). The fixed effect model assumes that all included studies share a common effect size. This assumption is not always realistic since studies are likely to be heterogeneous in various characteristics such as in their design and conduct as well as in participants, interventions and outcomes (Higgins and Thompson, 2008). The random-effects model assumes each study has its own study-specific effect size coming from a distribution of effects (conventionally assumed to be a normal distribution). The expected value of that distribution represents the overall mean effect size and its variance represents study heterogeneity, also known as between-study variance.

In some cases, a random-effects model will fail to explain all variation between true effects because of outlying studies. In a collection of studies, we may observe an extreme effect size that lies far away from the bulk of the data. An outlying study is defined as a study with a markedly different intervention effect estimate (Sterne *et al.,* 2008). This can be the case when a study's setting is substantially different from the settings in other studies which can have an impact on the intervention effect.

When outlying studies are present, it is not advised to simply exclude them from the meta-analysis. Instead, investigators should try to understand those patient, intervention and methodological characteristics of the outliers that modify the effect and produce aberrant results (Deeks *et al.,* 2011). When the effect modifiers are unobserved, unreported or unknown, the Cochrane Handbook suggests to apply a random-effects meta-analysis both with and without outlying studies as a part of a sensitivity analysis (Deeks *et al.,* 2011) where outliers are identified by a funnel plot (scatter plot of effect size vs inverted standard error). In simple meta-analysis with no covariates, it is likely that a funnel plot will reveal outlying studies, but in more complex models this is not always possible as interpretation of a funnel plot in the presence of explanatory variables (also called effect modifiers) is challenging. Meta-regression is widely used to account for heterogeneity by examining the impact of explanatory variables on effect size. A study appearing to be outlying in a funnel plot may actually be well explained by its regressor values.

Various methods have been suggested for accommodating and identifying outlying studies in meta-analysis. To account for the presence of possible outliers alternative distributions for the random effects can be used (Lee and Thompson, 2008). Baker and Jackson (2008) suggested a random-effects model with a long-tailed distribution for the underlying true effects in order to downweight the impact of outlying studies. Gumedze and Jackson (2011) suggested a method that identifies and downweights studies that have an inflated variance as part of a sensitivity analysis. Beath developed a method similar in spirit that assumes studies to be a mixture of standard and outlier studies allowing any number of outlier studies and then downweights outliers when estimating the summary effect (Beath, 2014). A popular class of methods for

identifying outliers are deletion diagnostics in which one or multiple observations are deleted at the same time and deletion statistics are computed that measure the effect of those deleted observations on parameter estimates or residuals. Viechtbauer and Cheung (2010) presented a method based on the impact of excluding studies on various statistics such as the summary estimate, heterogeneity and residuals.

In this paper, we propose the use of a Forward Search (FS) algorithm to identify outlying and influential studies. The FS was initially developed for robust estimation of covariance matrices (Hadi, 1992), regression models (Atkinson, 1994) and later applied to multivariate methods (Atkinson *et al.*, 2004) including factor analysis and item response theory models (Mavridis and Moustaki, 2008; Mavridis and Moustaki, 2009). In this paper we focus on a meta-regression model. To illustrate the forward search algorithm, we use two real datasets. The first data set involves 26 studies examining the effect of writing-to-learn interventions on academic achievement (Bangert-Drowns *et al.*, 2004) and the second dataset compares fluoride toothpaste to placebo for dental carries (Marinho *et al.*, 2003). . The second dataset has been widely used for outlier detection in meta-analysis. The FS algorithm sorts studies by their closeness to the hypothesized model and estimates non-parametrically the distribution of various statistics throughout the search allowing us to explore if a change is due to an outlying study or due to random error. The FS is compared with a backward search method for detecting outliers and it shows a clear advantange over existing methods.

The paper is organised as follows: section 2 provides an overview of the meta-analysis model; section 3 discusses the proposed forward search algorithm for the detection of outlying studies; section 4 presents the results from the two real data analyses and a simulation study and finally section 5 concludes. MATLAB code and documentation are available at www.mtm.uoi.gr and also provided as supplementary material through the publisher's website.

2. Meta-analysis model

Suppose that we have n Randomized Control Trials (RCT's) each comparing two interventions A and B. Each study i $(i = 1, \dots, n)$ yields an effect size y_i with a corresponding standard error s_i. The study-specific y_i could be any measure of the relative treatment effects such as the mean differences (MD), standardized mean differences (SMD), logarithm of risk ratio (logRR) or logarithm of odds ratio (logOR).

In a random-effects meta-analysis model, the observed contrast in study i is modelled as

$$y_i = \mu + \delta_i + \varepsilon_i,$$

where μ is the true effect of treatment B relative to A, δ_i is a random effect with $\delta_i \sim N(0, \tau^2)$ where the between-study variance (heterogeneity) τ^2 denotes how effectiveness varies across studies, and ε_i is a sampling error term with $\varepsilon_i \sim N(0, s_i^2)$ and s_i^2 is known and usually taken to be the observed standard error for the contrast in the i_{th} study. In a fixed effects meta-analysis model, τ is assumed to be zero.

More generally, meta regression analysis extends the meta-analysis model to include study level covariates. In matrix notation we have

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\mu} + \boldsymbol{\delta} + \boldsymbol{\varepsilon} \qquad (1)$$

where $\boldsymbol{y}$ is a n-vector including the observed contrasts for each study, $\boldsymbol{\varepsilon}$ is a n-vector of normally distributed sampling errors with $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \mathbf{S})$ and known diagonal covariance matrix $\mathbf{S} = diag(s_i^2)$, and $\boldsymbol{\delta}$ is a n-vector of normally distributed random effects, $\boldsymbol{\delta} \sim N_n(\mathbf{0}, \boldsymbol{\Delta})$ with a diagonal covariance matrix $\Delta = diag(\tau^2)$, and $\mathbf{X}$ has $p + 1$ columns with the first column being a vector of ones and the rest are the values of the p regressors. The design matrix can be

accustomed appropriately to address multiple outcomes (Mavridis and Salanti, 2013) and/or multiple interventions (Salanti, 2012).

The summary estimates are computed using weighted least squares (WLS) giving $\widehat{\boldsymbol{\mu}} = \left(\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{X}'\widehat{\mathbf{W}}\boldsymbol{y}$ with weights given by $\widehat{\mathbf{W}} = \left(\mathbf{S} + \widehat{\boldsymbol{\Delta}}\right)^{-1}$ and $var(\widehat{\boldsymbol{\mu}}) = \left(\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X}\right)^{-1}$. There are many methods suggested for estimating Δ including the method of moments, maximum likelihood and restricted maximum likelihood (Sidik and Jonkman, 2007). Predicted values are computed as $\widehat{\boldsymbol{y}} = \mathbf{X}\widehat{\boldsymbol{\mu}} = \mathbf{X}\left(\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{X}'\widehat{\mathbf{W}}\boldsymbol{y} = \mathbf{H}\boldsymbol{y}$ where $\mathbf{H} = \mathbf{X}\left(\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{X}'\widehat{\mathbf{W}}$ is known as the hat matrix and its diagonal points are called leverages. The hat matrix maps the observed values to the fitted ones and the leverage points are widely used in regression models to identify observations that lie far away from their fitted value. Standardized residuals can be computed for each study using the formula $\hat{\varepsilon}_i = \sqrt{\widehat{w}_i}(y_i - \boldsymbol{X}_i\widehat{\boldsymbol{\mu}})$ where $\boldsymbol{X}_i$ is the i_{th} row of the $\mathbf{X}$ matrix and $\widehat{w}_i$ is the i_{th} diagonal element of W. Studentized residuals can be computed as

$\hat{\varepsilon}_i^{stud} = \sqrt{\frac{\widehat{w}_i}{1-h_i}}(y_i - \boldsymbol{X}_i\widehat{\boldsymbol{\mu}})$ where h_i is the i_{th} diagonal element of $\mathbf{H}$.

Standard outputs from most meta-analyses are the summary estimate $\boldsymbol{\mu}$, the heterogeneity variance estimate τ^2, and the chi-squared statistic Q that assess whether observed differences in treatment effects are explained by chance alone (Higgins and Thompson, 2002). To estimate heterogeneity and compute the chi-squared statistic, we can use the DerSimonian and Laird estimator (DerSimonian and Laird, 1986) which yields $Q = \boldsymbol{y}'\widehat{\mathbf{P}}\boldsymbol{y}$ and $\hat{\tau}^2 = (Q - (n - p - 1))/trace(\widehat{\mathbf{P}})$ where $\widehat{\mathbf{P}} = \widehat{\mathbf{W}} - \widehat{\mathbf{W}}\mathbf{X}\left(\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{X}'\widehat{\mathbf{W}}$ and $trace(\widehat{\mathbf{P}})$ denotes the trace of matrix $\widehat{\mathbf{P}}$; that is the sum of its diagonal elements. The Q statistic is used to test for a zero-heterogeneity, $H_0: \tau = 0$, and it follows asymptotically a chi-squared distribution with $n - p - 1$ degrees of freedom. This test has been routinely used for deciding between a fixed and a random effect model though it has been argued that is has low power and decision should not be based on it (Nikolakopoulou *et al.*, 2014; Higgins and Thompson, 2002; Borenstein *et al.*, 2010)

3. Selection algorithms to detect outliers: the Forward Search and Backward Search algorithms

3.1. Forward Search

The forward search starts with a small initial clean subset of the studies that is most likely outlier-free and proceeds by adding studies until all are included. Eventually, it results in an ordering of the studies according to their fit to the hypothesized model and thus provides a way to identify outlying studies. The stages of the FS can be summarized as follows:

Stage 1. Choose an initial clean (i.e. likely outlier-free) subset of m studies from the $\binom{n}{m}$ possible subsets. Those m studies are used to initialize the search and constitute the initial subset or the '*basic set*' at the beginning of the search while the remaining $n - m$ studies constitute the '*non-basic set.*'

Stage 2. Progress in the FS by adding studies from the '*non-basic set*' to the '*basic set*' until all studies are finally included in the '*basic set*'. This stage of the algorithm has $n - m$ steps where

at each step the number of studies in the '*basic set*' increases by one and the number of studies in the '*non-basic set*' decreases by one. More details for stage 2 (e.g. which studies to add from the $n - m$ studies in the '*non-basic set*') are presented in section 3.2.2.

Stage 3. Monitor various statistics of interest such as the true effect estimates, heterogeneity, the Q statistic and other measures explained in more detail in Section 3.2.3 during the $n - m$ steps of stage 2. Studies that fall outside of monitoring bounds are identified as potential outliers.

Once the potential outliers have been identified from the forward search algorithm, the researcher may then follow the Cochrane Report recommendation and fit the meta-analysis model removing these studies as a part of a sensitivity analysis.

3.2. Detail for each step of the algorithm

3.2.1. Choice of the initial clean subset (Stage 1)

3.2.1.1. Selecting the size of the initial clean subset:

The number of studies included in the initial clean subset varies according to model complexity. At a minimum there needs to be at least as many studies as the numbers of parameters to estimate, which in a simple random-effects meta-analysis is 2 (i.e. the true effect and heterogeneity) and in a meta-regression with p covariates is p+2. We note the Cochrane handbook suggests that a meta-regression is meaningless with less than ten studies and at least three studies should be used to estimate heterogeneity (Deeks *et al.,* 2011). Hence, following this guideline, we recommend the initial clean subset to start with at least 10 studies in a meta-regression model (or with $p + 2$ in the unlikely case that $p > 10$) and with 3 studies in a simple meta-analysis model.

3.2.1.2. Selecting the studies to include in the initial clean subset:

Recall the goal is to identify a clean subset of studies that is free of outliers to use as the initial set. Exhaustively searching all $\binom{n}{m}$ subsets of size *m* and identifying the most likely subset to be free of outliers is a viable option when the meta-analysis includes a small but sizeable number of studies (e.g. 15), but with many studies, an exhaustive analysis is prohibitive and practically unnecessary. Indeed by examining a large number of candidate initial subsets of size m, say B (e.g. $B = 1000$) we are highly likely to find a clean subset of studies as long as the number of outliers is small proportionate to the size of the data which we would assume by the very nature or definition of outliers. Specifically, for each candidate initial subset b ($b = 1, \dots, B$), of m studies denoted by D_b^m we obtain initial subset-specific estimates for the parameters of interest namely the mean effect and the between-study variability ($\widehat{\boldsymbol{\mu}}_{D_b^m}, \hat{\tau}_{D_b^m}$) and we calculate an objective function $f\left(y_i, s_i, \boldsymbol{X}_i, \widehat{\boldsymbol{\mu}}_{D_b^m}, \hat{\tau}_{D_b^m}\right)$ that measures the fit of the entire dataset $(y_i, s_i, \boldsymbol{X}_i)$ to the parameters estimates from the initial subset candidate. The subset that optimizes the objective function is chosen to be the initial clean subset. A typical strategy that has been extensively used in the FS literature (Atkinson and Riani 2000) and is similar in rationale to the least median of

squares regression (Rousseew, 1984) is to choose the initial subset D_b^m that minimizes the median of the absolute standardized residuals, i.e.

$$f\left(y_i, s_i, \boldsymbol{X}_i, \hat{\boldsymbol{\mu}}_{D_b^m}, \hat{\tau}_{D_b^m}\right) = median\left(\left|\hat{\varepsilon}_{i,D_b^m}\right|\right), \quad (2)$$

where

$$\hat{\varepsilon}_{i,D_b^m} = \sqrt{\frac{1}{s_i^2 + \hat{\tau}_{D_b^m}^2}}\left(y_i - \boldsymbol{X}_i\hat{\boldsymbol{\mu}}_{D_b^m}\right), i = 1, \ldots, n$$

and s_i is the standard error for the effect size y_i. In Equation 2 we may omit the heterogeneity parameter, $\hat{\tau}_{D_b^m}$, if the subset is too small (e.g. smaller than 5) to provide an accurate estimate of heterogeneity.

Other objective functions to select the initial dataset include other types of residuals or log-likelihood contributions. An easily understood measure is to select that subset with the smallest value for the Q statistic or for the heterogeneity parameter defined in Section 2.

3.2.2. Progressing in the search: Stage 2

We have now found the initial clean subset D^m as described in Stage 1 that will constitute the '*basic set*' at the beginning of the search while its complementary set $(D^m)^c$ constitutes the '*non-basic set*'. The observations in the '*non-basic set*' are sorted by their 'closeness' to the '*basic set*'. This is again achieved by fitting the hypothesized model to the '*basic set*', estimating the parameters $\hat{\boldsymbol{\mu}}_{D^m}$ and $\hat{\tau}_{D^m}$ and sorting the observations according to their closeness to the *'basic set'* as measured by an objective function $f(y_i, s_i, \boldsymbol{X}_i, \hat{\boldsymbol{\mu}}_{D^m}, \hat{\tau}_{D^m})$.

Specifically, the algorithm proceeds as follows:

Step 1: For all observations in the 'non-basic set', $y_i, s_i, X_i \in (D^m)^c$, calculate a closeness measure $f(y_i, s_i, \boldsymbol{X}_i, \hat{\boldsymbol{\mu}}_{D^m}, \hat{\tau}_{D^m})$ where $\hat{\boldsymbol{\mu}}_{D^m}$ and $\hat{\tau}_{D^m}$ are estimated from the initial clean basic set. The closeness measure f is taken to be the median of the absolute standardized squared residuals as in Equation 2. Then after identifying the study from the non-basic set with optimal closeness to the basic set, re-define the 'basic set' D^{m+1} to include this optimal study and have $m + 1$ studies.

Step j: Add to the *'basic set'* D^{m+j-1}, the study with the optimal $f\left(y_i, s_i, \boldsymbol{X}_i, \hat{\boldsymbol{\mu}}_{D^{m+j-1}}, \hat{\tau}_{D^{m+j-1}}\right)$ for $y_i, s_i, \boldsymbol{X}_i \in$ in the $\left(D^{m+j-1}\right)^c$.Then re-define the 'basic set' D^{m+j} to have $m + j$ studies. Proceed with the algorithm until $j = n - m$ and all studies are included in the basic data set.

3.2.3. Statistics to be monitored during the forward search: Stage 3

At each of the $n - m$ steps of Stage 2 various statistics associated with a meta-analysis model can be computed and monitored. Forward plots are drawn for each statistic at each step *j* to monitor its behaviour during the search. For example, estimates of the summary effects (e.g. the true intervention effect) at each step *j* with their associated confidence intervals are plotted and monitored for how they change as additional studies are added.

The heterogeneity parameter and *Q* statistic described in Section 2 also are plotted at each step j. A sudden increase in either of these statistics after a study is included may be indicative of that study being an outlier. A measure of the total change in all parameter estimates that has been widely used to identify outliers is the Cook's statistic (Cook and Weisberg, 1982). The Cook's statistic is computed at the j^{th} step of the search as

$$\boldsymbol{C_j} = \left(\widehat{\boldsymbol{\mu}}_{D^{m+j}} - \widehat{\boldsymbol{\mu}}_{D^{m+j-1}}\right)' \left(\mathbf{X}'_{\mathrm{D^{m+j}}} \mathbf{W}_{\mathrm{D^{m+j}}} \mathbf{X}_{\mathrm{D^{m+j}}}\right)^{-1} \left(\widehat{\boldsymbol{\mu}}_{D^{m+j}} - \widehat{\boldsymbol{\mu}}_{D^{m+j-1}}\right). \qquad (3)$$

Another measure that is used in regression diagnostics and is useful to monitor for outliers is the change in uncertainty in the estimation of the summary estimates measured as the ratio of the determinants of the total variance at step $\boldsymbol{m+j}$ to that at step $\boldsymbol{m+j-1}$

$$\boldsymbol{R_j} = \frac{\left|\left(\mathbf{X}'_{\mathrm{D^{m+j}}} \mathbf{W}_{\mathrm{D^{m+j}}} \mathbf{X}_{\mathrm{D^{m+j}}}\right)^{-1}\right|}{\left|\left(\mathbf{X}'_{\mathrm{D^{m+j-1}}} \mathbf{W}_{\mathrm{D^{m+j-1}}} \mathbf{X}_{\mathrm{D^{m+j-1}}}\right)^{-1}\right|} \qquad (4)$$

where $\mathbf{W}_{D^{m+j}}$ is the weight matrix for the D^{m+j} subset and $\left|\left(\mathrm{X}'_{\mathbf{D^{m+j}}} \mathrm{W}_{\mathbf{D^{m+j}}} \mathrm{X}_{\mathbf{D^{m+j}}}\right)^{-1}\right|$ denotes the determinant of matrix $\left(\mathrm{X}'_{\mathbf{D^{m+j}}} \mathrm{W}_{\mathbf{D^{m+j}}} \mathrm{X}_{\mathbf{D^{m+j}}}\right)^{-1}$. Monitoring this statistic helps us understand how much more (or less) precise results become when a study enters the search. Similar statistics have been used to explore how much uncertainty is introduced by assuming a random effects model in multivariate meta-analysis (Jackson *et al.*, 2012)

Other measures that can be monitored during the FS are the standardized and studentized residuals computed for each study (presented in Section 2). Non-smooth forward plots of residuals are indicative of outlying values.

3.2.4. Characterizing changes in a statistic as important and identification of outliers

Forward plots of the statistics presented in Section 3.2.3 show how they are affected by the inclusion of studies in the '*basic set*'. The important question is whether the observed fluctuations in the forward plots are due to the inclusion of a study with outlying effect size or due to random variation. As the monitored statistics do not have known asymptotic distributions, we propose the bootstrap technique for estimation of confidence bounds.
Simulation envelopes using parametric bootstrap (Efron and Tibshirani 1993) are constructed to determine the limits of a change in the statistic that could be attributed to chance. The steps for constructing simulation envelopes are:

(a) The hypothesized meta-analysis model is fitted to the n studies and the model parameters are estimated
(b) M datasets of n studies each are generated using the parameter estimates from step a (parametric bootstrapping).
(c) A FS is applied to each of the M generated data producing M estimations for the changes in the monitored statistic at step $j, j = 1, \ldots, n-m$, where m is the size of the initial basic set.
(d) The confidence bounds for the monitored statistic at each step $j = 1, \ldots, n-m$ are the the $(1-\alpha)$ quantiles of the M values.

3.2.5. Sensitivity analysis to evaluate robustness of the forward search

It is possible that the ordering of the studies to enter the search during Stage 2 may be influenced by the studies chosen in the initial subset or generally by the studies in the '*basic set*' in general. We suggest repeating the algorithm a number of times from random starting points to evaluate robustness of the ordering.

3.2.6. Backward Search

We provide a description of the backward search because we intend to compare it with the FS algorithm. Backward search algorithms start with the full data set and remove sequentially outlying observations until all outliers have been removed. The backward search algorithm starts by fitting the hypothesized model to the entire data set that includes all studies. Then the objective function that measures outlyingess for each study (e.g. residuals, likelihood contributions) is computed . The study with the the worst value of the objective function is deleted. The above process is iterated until some criterion is met (e.g. all residuals are less than 2 in absolute value).

These methods can be useful when there are a few outlying studies, but they also can be negatively affected by the observations they are supposed to identify (Hadi and Simonoff, 1993). Backward methods can suffer from what is known as masking and swamping effects (Atkinson, 1986). Masking refers to outlying studies going undetected probably because outliers are clustered together and identification of a single outlier is affected by the presence of other outliers already included in the dataset. Swamping refers to studies erroneously detected as outliers. This may happen because true outliers shift parameter estimates towards them and away from non-outlying observations. Masking is akin to a 'false negative' while swamping is akin to a 'false positive'. It should also be noted that omitting studies leads to a truncated distribution of test statistics for which asymptotics do not hold. Hence, asymptotic distributions for residuals or the *Q* statistic are no longer valid. Hence, it is not safe to omit studies and then look for statistical significance without accounting for multiple testing.

4. Examples

4.1. Simple Simulated example

To acquaint readers with the forward search we will use a simple simulated example with 7 studies, one of which is an outlier by construction. The variances for the 7 studies are generated from $s_i^2 \sim x_{0.5}^2/8$. The summary effect is chosen to be zero and the between-study heterogeneity is $\tau^2 = 0.2^2$. Six effect sizes are generated from $y_i \sim N(0, s_i^2 + 0.2^2)$, $i = 1, \dots, 6$ and the effect size for the seventh study is generated from $y_7 \sim N(3\max(\boldsymbol{s}), s_7^2 + 0.2^2)$ where $\boldsymbol{s} = (s_1, \dots, s_7)'$ is the vector with the simulated standard errors. The effect sizes and the standard errors are shown in column 2 of Table 1. The size of the initial subset is taken to be 3 and all 35 possible subsets, $\binom{7}{3}$, are studied to find the one which yields the smallest chi-squared statistic (*Q*) value. The 'basic set' at step 1 of the search consists of studies 1, 4 and 5 (gave the smallest *Q* value =1.27) (shown in bold in column 2 in Table 1) and the the remaining studies (2, 3, 6 and 7) constitute the '*non-basic set'*. We compute the residual values for the studies in the '*non-basic*

set' using the parameters estimated from the '*basic set'*. The study with the smallest residual in absolute value (study 3 with residual value -2.21, shown in bold in column 3 in Table 1) is the next study to enter the '*basic set'* which now consists of studies 1, 3, 4 and 5 in step 2 (shown in bold). The FS proceeds until the '*basic set'*includes all seven studies. We notice that the artificial outlier (study 7) is the last to enter the FS incurring also a large increase in the Q value from 10.10 to 32.12.

Table 1: Simple simulated example with 7 studies. The five steps of the FS together with the effect sizes and standard errors. FS starts with three studies (those three studies that give the minimum chi-squared statistic (Q) out of the possible $\binom{7}{3} = 35$ subsets, bold letters refer to studies belonging to the '*basic'* set). The study with the smallest residual (in absolute values) is the next to enter in the following step. The smallest residual is with bold letters.

Study All studies $Q = 32.12$	Step 1 Studies 1, 4 and 5 $Q = 1.27$		Step 2 Studies 1, 3-4 and 5 $Q = 5.61$	Step 3 Studies 1, 3-6 $Q = 5.91$	Step 4 Studies 1-6 $Q = 10.10$
	y_i (s_i)	Residuals in the '*non basic*' set	Residuals in the '*non basic*' set	Residuals in the '*non basic*' set	Residuals in the '*non basic*' set
1	**0.58** **(0.46)**				
2	-0.11 (0.04)	-10.38	-0.63	**-1.04**	
3	-1.04 (0.63)	**-2.21**			
4	**0.18** **(0.27)**				
5	**0.93** **(0.73)**				
6	0.06 (0.09)	-3.12	**-0.24**		
7	2.42 (0.53)	3.92	3.25	4.02	**4.39**

4.2.Writing to learn interventions on academic achievement

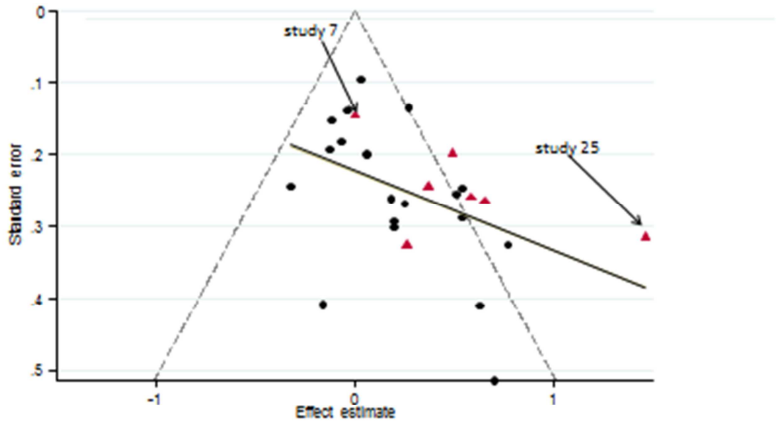
As already explained in the introduction, the 26 studies examine the effect of writing-to-learn interventions on academic achievement. The effect of the intervention was measured as a Standardized Mean Difference (SMD) between writing-to-learn interventions and conventional instructions. Table 2 gives their standard errors and three potential effect modifiers, namely whether the sample consisted of high-school or college students (dummy variable coded as college=1, high-school=0), the length of the intervention (in weeks), and whether the intervention incorporated prompts for metacognitive reflection' (dummy variable coded as yes=1, no=0).

Table 2: Effect size y_i, standard error s_i and values of covariates for 26 studies examining the effect of writing-to-learn interventions on academic achievement.

Study	college	length	Meta-cognition	Effect size y_i	Standard error s_i
1	1	15	1	0.65	0.265
2	1	9	0	-0.04	0.138
3	1	1	0	0.03	0.095
4	0	4	1	0.26	0.326
5	1	4	0	0.06	0.200
6	1	15	0	0.77	0.327
7	1	15	1	0.00	0.145
8	1	4	0	0.54	0.288
9	1	14	0	0.20	0.293
10	1	15	0	0.20	0.302
11	1	4	0	-0.16	0.409
12	1	3	0	0.51	0.255
13	0	19	0	0.54	0.247
14	0	12	1	0.37	0.245
15	0	1	0	-0.13	0.192
16	0	1	0	0.18	0.263
17	0	1	0	0.27	0.134
18	1	11	0	-0.32	0.245
19	0	1	0	-0.12	0.152
20	1	15	0	-0.07	0.182
21	1	15	0	0.70	0.515
22	1	2	1	0.49	0.198
23	0	24	1	0.58	0.259
24	1	15	0	0.63	0.410
25	0	15	1	1.46	0.315
26	1	15	0	0.25	0.268

Figure 1 shows the funnel plot; a scatterplot of effect size y_i vs standard error s_i centered at the line of no-effect, 95% pseudo-confidence intervals that include all trials pointing to a null effect and a regression line of effect size vs standard error weighted by inverse variance overlayed. We see from the regression line with negative slope that the funnel plot is asymmetric with larger effects being found in smaller studies, a common small study effect. It is also evident that Study 25, with the largest effect size, lies far away from the bulk of the data, far outside of the 95% confidence bounds and may be an outlier although it is one of the smaller studies. When covariates are considered, it becomes apparent that the near-zero effect found for Study 7, is actually at the low end of effects found for studies with an intervention prompting for metacognitive reflection which all have effect sizes larger than the summary effect with several reaching statistical significance. The summary effects and the 95% confidence intervals (CI) using all 26 studies are $\hat{\mu}_0 = 0.25[0.01,0.49]$ for the intervention effect, $\hat{\mu}_1 = -0.10[-0.37,0.16]$ for college compared to high school, $\hat{\mu}_2 = 0.09[-0.04,0.23]$ for intervention length and $\hat{\mu}_3 = 0.24(-0.06,0.54)$ for metacognition. Heterogeneity is estimated to be $\tau^2 = 0.0472$ with a chi-squared statistic Q equal to 44.14 (p-value 0.0034 on a x^2_{22} distribution) suggesting studies are heterogeneous.

Figure 1 : Funnel plot centered at the point of no effect with 95% confidence intervals. The solid line corresponds to the regression line. Triangles refer to studies where the intervention incorporated prompts for meta-cognitive reflection).



4.2.1. Simple meta-analysis with no covariates

We conducted a FS starting from an initial sample of $m = 3$ studies that yielded the smallest value of the Q statistic out of a total of $B = 100$ subsets. Figure 2 shows the forward plots for heterogeneity τ^2, Cook's distances (expression (3)), the chi-squared statistic and the ratio of variances R_h (expression (4)). The 95 and 97.5 quantiles were estimated using 10000 bootstrap samples. Study 7 is not indicated as an outlier which was expected since its effect size was unlikely only under a model including a covariate for metacognitive reflection. In the forward plot of the Cook's distances and the ratio of variances, study 25 is clearly an outlier because it enters in the last steps of the search and incurs changes that can not be explained by chance. We

should note that the FS algorithm is a stochastic method that is based on many parameters (choice of the initial subset, size of the initial subset, number of initial samples explored, method of adding studies). Not all methods yield the same ordering of studes. For that reason we suggest repeating the FS a number of times (e.g 10 times) from different starting points and also run it a couple of times using different methods of choosing the initial subset or of progressing in the search. In all repetitions that we run, study 25 appearred to be an outlier and it entered at the last step (unless it was included in the initial subset). In few of the replications, there were other studies entering at the beginning or the mid of the search that incurred changes in statistics that were lying outside the 95% confidence bounds. These studies are potential outliers when considering that subset of studies and not when considering all the studies. The FS was robust in identifying study 25 as outlying. If we re-run the meta-analysis model deleting study 25, the summary estimate reduces from 0.24 (95% CI 0.11 to 0.37) to 0.18 (95% CI 0.07 to 0.30).

Since we do not allow studies to interchange between the '*basic set*' and the '*non-basic set*', a question arising is how the FS algorithm behaves in the unlikely case that a suspicious case is included in the '*basic set*' or early in the search. Figure 3 shows the forward plot of the residual for study 25 during a FS for simple meta-analysis without covariates where we forced study 25 to be included in the initial subset. It starts with a small residual value but it keeps increasing as more studies are added indicating that this study differs from the rest of the studies. This pattern is typical of suspicious studies when these either are included in the initial subset or enter the '*basic set*' early in the search. Hence, even if we do not allow studies to leave the basic set, if they are outliers, the residual values associated with them are expected to deteriorate as clean data enter the search. We therefore strongly suggest monitoring residuals for all studies across the FS. We also applied a backward search algorithm that deleted those studies that had a large impact on the chi-squared statistic. The first study to be deleted was study 25 followed by studies 1, 6, 22 and 23.

Figure 2 : Forward plots of heterogeneity, Cook distances, chi-square statistic and R accompanied by the 95th (dotted line) and 97.5th (solid line) quantiles for the meta-analysis model.

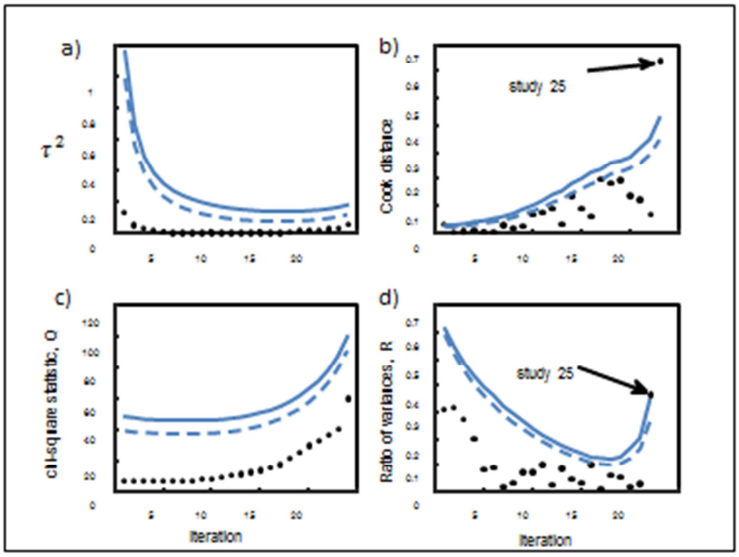
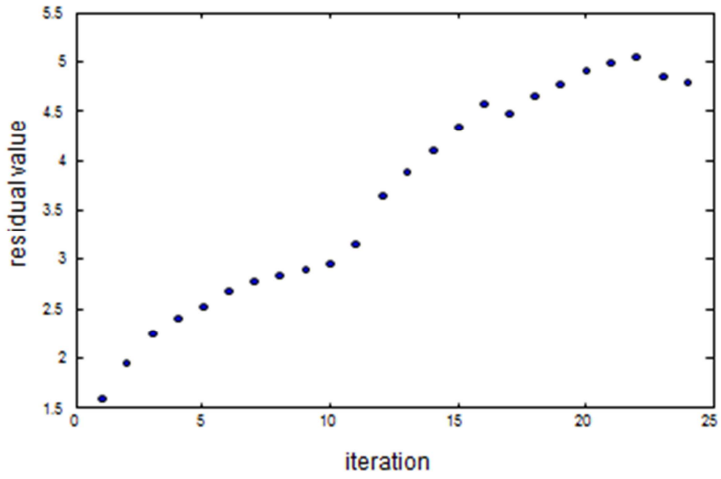


Figure 3 : Forward plot for the residual of the 25th study for the meta-analysis model when this study is included in the initial subset



4.2.2. Meta-regression model

Backward search

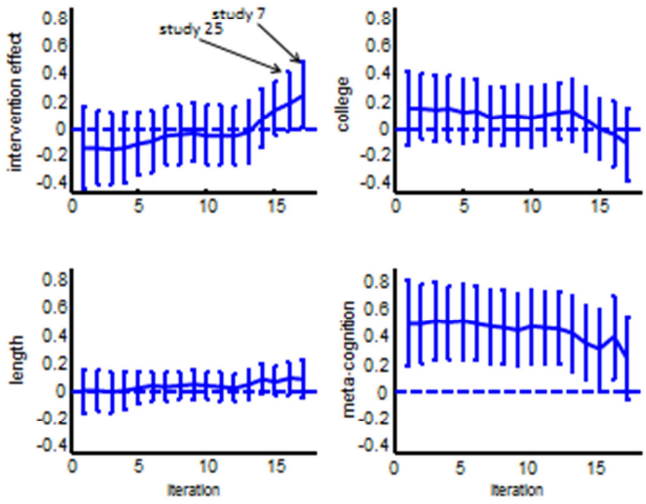
Viechtbauer and Cheung (Viechtbauer and Cheung, 2010) analysed this dataset accounting for all covariates. They adapted a backward search deleting each study in turn and identified studies 7 and 25 as potentially outlying and influential since they incurred large changes in the Q statistic, the Cook distance, studentized residuals and R statistic. It should be noted that differences in summary estimates obtained from deleting either study 7 or study 25 are not statistically significant. Hedges and Olkin suggested examining changes in the Q statistic when each study is omitted in turn. Omitting studies 7 and 25 yields a Q statistic equal to 26.26 that is non-significant (p-value 0.16 on a x_{20}^2 distribution) indicating no evidence of heterogeneity (Viechtbauer and Cheung, 2010). Thus, heterogeneity seems to be caused by these two studies. Also, if we omit study 7 then prompting for metacognitive reflection becomes significant ($\hat{\mu}_3 = 0.40(0.09,0.70)$). However, omitting studies leads to a truncated distribution for which asymptotics do not hold and the assumed x_{20}^2 distribution under which statistical heterogeneity is not present is not valid. We adopted a backward search where we start with the whole data set and computed weighted least squares residuals for all studies as a measure to quantify how aberrant a study is. We then omitted the study that incurred a large decrease to the Q statistic. The first study to leave the search was study 25 followed by 7, 18, 17 and finally 20.

Forward search

The FS started with an initial subset of 10 studies. We explored 1000 initial subsets using the median criterion of absolute standardised residuals (Equation 2). The simple version of the algorithm is used where once a study is included in the '*basic set*' it stays until the end of the search and only one study is added at each step. Studies 25 and 7 enter in the last two steps of the

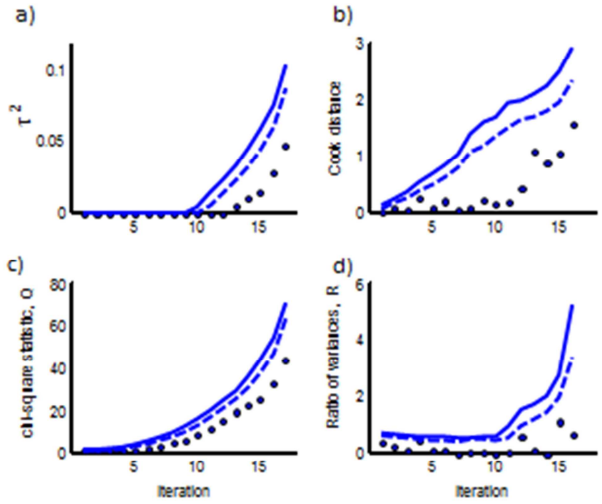
search. Forward search aims at providing a natural ordering of the data under the hypothesized model and although study 7 does not have an extreme SMD it enters at the last steps because it is very unlikely to observe its effect size conditional on its covariate values. Figure 4 shows the 95% confidence intervals of parameter estimates throughout the search. The estimated coefficient for meta-cognition (μ_3) becomes marginally non-significant at the last step of the search. The estimated summary effect μ_0 increases during the last four steps of the search and turns from negative to positive.

Figure 4 : Forward plot of summary estimates with 95% confidence intervals



We emphasize that we are not looking for changes in significance or direction of effects but on changes that cannot be explained by chance alone when the hypothesized model is corret. Figure 5 shows the forward plots for heterogeneity τ^2, Cook's distances (expression (3)), the chi-squared statistic and the ratio of variances R_h (expression (4)). The 95 and 97.5 quantiles were estimated using 10000 bootstrap samples. Study 25 that was identified as an outlier by inspecting the funnel plot (Figure 1) and using backward methods incur changes to the statistics being monitored that are fully explained by random variation. The change that occurs when study 25 or 7 is added is approximately half the magnitude of the biggest accepted change due to random variation. One possible reason that study 25 or study 7 are not identified as outliers could be that the model is very sparse (26 studies – 3 covariates) and as a result there is no power to detect outliers and sharp changes in the forward plots are expected.

Figure 5 : Forward plots of heterogeneity, Cook distances, chi-square statistic and R accompanied by the 95th (dotted line) and 97.5th (solid line) quantiles for the meta-regression model.



A FS is repeated considering 100 different initial subsets chosen randomly to explore how robust the results are to the choice of the initial subset. Study 25 enters the search in the last two steps in 54% of the times whereas study 7 enters the search in the last two itereations in 81% of the times (percentages are computed out of the subsets that the particular studies were not included in the initial subset). In the remaining 46%, study 25 enters between steps 15-18 incuring changes in the statistics of interest and until that point heterogeneity was estimated to be zero.

4.2.3. Conclusions from the learning interventions example

Overall, we conclude that study 25 is an outlier in a simple meta-analysis model without covariates. If we account for covariates then study 25 does not seem to be aberrant. On the other hand, study 7 is not an outlier in the simple meta-analysis model but its effectiveness is considered very low for a study that prompts for metacognitive reflection. However, its value is not unexpected if we account for metacognitive reflection. We should also note that small-study effects are evident in this dataset (p-value<0.01 for Egger's test, Egger 1997) if we apply a model with no covariates that accounts for small-study effects none of the studies appear to be outliers and all changes lie within the bounds from the simulation envelopes. This stresses the fact that results that are unusual under one model may not be unusual under a different model.

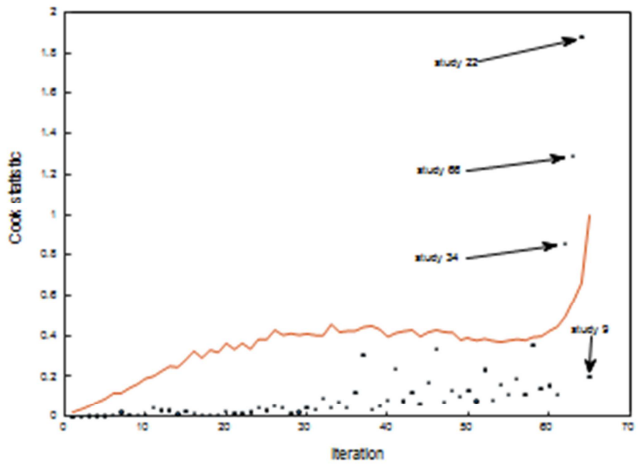
4.3. Fluoride toothpaste for preventing dental caries

This meta-analysis comprises 70 trials comparing a fluoride toothpaste to placebo or another treatment for preventing dental caries in children and adolescents (Marinho *et al.*, 2003). Studies report the standardized mean difference between treatment and control of tooth areas with carries. Negative values support that the fluoride intervention is beneficial. This data set has

some obvious extreme effect sizes and has been used for outlier identification (Baker and Jackson 2008, Gumedze and Jackson 2011). It has been suggested that there is no evidence of publication bias in the original review though Baker and Jackson argue that this may not be true. Egger's test (Egger 1997) clearly showed that there is evidence for small-study effects (p-value<0.01) and a selection model (Mavridis *et al.*, 2013) showed that the correlation between the probability of publication and magnitude of effect is not zero ($\rho = -0.47$ $(95\%\ CI -0.82$ to $-0.11)$ suggesting that there is publication bias. This finding complicates the analysis. A forward search analysis had studies 34, 22, 9 and 66 entering in the last four steps and the same studies were found to be outlying if we accounted for small study effects. The same four studies were identified as potential outliers using other outlier identification techniques (Baker and Jackson 2008, Gumedze and Jackson 2011 but it should be noted that in these papers a different numbering of studies is used, we used the one suggested in the original publication). A backward search suggested deleting studies 22, 66, 34 and 21.

Figure 6 shows the forward plot of the Cook statistic accompanied by 99% confidence bounds. Three studies (studies 34 ,66 and 22) that enter in thefour last steps of the search seem to have incurred a change that is not explained by chance.

Figure 6: Forward plot for the Cook statistic for the Fluoride dataset accompanied by 99% confidence bounds



4.4. Simulation study

It is difficult to conduct a large simulation study to evaluate the FS algorithm because there are many aspects of the algorithm that need to be taken into consideration (e.g. initial sample size, method of initializing and progressing in the search e.t.c.). We also advise repeating the search a small number of times and constructing simulation envelopes. These are difficult to control in a simulation study and are omitted here. We simulated 100 forward searches for a simple meta-analysis model under various scenarios. In all scenarios we used residual values for progressing

in the search and we started with an initial subset of 5 studies selected by inspecting 1000 subsets and choosing that with the lowest value of the chi-squared statistic *Q*. Effect sizes were drawn from a normal distribution with zero mean and heterogeneity standard deviation τ. To create variances for the effect sizes, we generated values from a chi-squared distribution with 0.5 degrees of freedom and divide it by 8. We created 1, 2 or 3 outiers by generating effect sizes from a normal distribution with a mean value equal to three times the largest generated standard error and heterogeneity standard deviation τ. We counted how many times out of the 100 simulated data set the outlying *k* values entered in the last *k* steps (*k=1,2,3*). Results are shown in Table 3. It seems that the method is quite powerful in identifying outliers but it gets worse as heterogeneity increases. It should also be noted that power is underestimated because if we could monitor forward plots and repeat the search a small number of times we would most probably have detected outliers in some of the cases where these were not included in the last steps of the search.

Table 3: Simulation Study results for three sample sizes and two values of the heterogeneity variance parameter: percentage of times the *k* outliers entered the search in the last steps of the FS

	$\mu = 0, \tau = 0$			$\mu = 0, \tau = 0.2$		
Number of outliers k	$n = 20$	$n = 20$	$n = 100$	$n = 20$	$n = 20$	$n = 100$
1	97%	89%	84%	79%	84%	85%
2	88%	89%	87%	52%	60%	73%
3	80%	84%	82%	49%	52%	60%

5. Discussion

The FS algorithm is a robust, diagnostic, graphical method that is easily extended to meta-analyses models. Forward plots may reveal important information and unsuspected structure in the data. In systematic reviews, it is advised to repeat the analysis excluding outlying studies (Deeks *et al.,* 2011) and backward methods are routinely used to identify outliers. Asymptotic distributions do not hold for truncated samples and it is difficult to infer if a change in parameter estimates is statistically significant. Also, unlike backward methods, the FS does not suffer from the masking and swamping effects because it is not affected by the studies it is supposed to identify. Other methods suggested in the literature also use all observed effect sizes and potential outliers in their attempt to identify or downweight outlying studies. Since, the FS algorithm has been primarily developed for identifying outliers in regression models (Atkninson 1994, Atkinson and Riani 2000), we argue that it naturally extends to meta-analysis models. It can also

be employed to more complicated models, not presented here, such as inconsistency models in network meta-analysis where it can be used to detect studies that may cause inconsistency between direct and indirect estimates.

The benefits compared to a backward search have been stressed in the literature (Atkinson and Riani 2000) and these are summarized mainly in the fact that outliers are not used in the method employed to identify them. There were no differences between the results of a FS and a backward search in the examples employed as the studies entering in the end of a FS are the same with the studies that were first omitted in a backward searc. The construction of simulation envelopes provides a visual tool about the significance of changes as studies enter the search in the FS algorithm. Both in a FS and in a backward method, different criteria to quantify outlyingness would give a different ordering of data. The FS algorithm proposes a straightforward method to determine if a change in any of the statistics being monitored is outside the limits defined by random variation and allows us to judge on the arbitrariness of a study irrespective of the point at which it enters the search. Unlike backward methods, a FS algorithm does not give the same ordering each time it is applied even if the same criteria are used, though differences are usually mild. We noticed that when the initial subset represents 10% of the original number of studies (or with at least 10 studies in a meta-regression problem), the method is very robust. It is also easy to spot aberrant patterns caused by additions of studies in the search or an aberrant pattern of a study included in the '*basic set*' no matter which method is used for selecting the initial subset and for progressing in the search. If an outlier is included in the '*basic set*' in the early steps, this may cause some abnormalities in the search. To overcome this problem we suggest running the FS 5-10 times from random starting points and explore how robust the ordering is. In all repetitions of the FS algorithm in the writing-to-learn interventions example, study 25 was identified as suspicious when a simple meta-analysis model was assumed and no other study was identified as suspicious in more than 10% of the repetitions.
In most meta-analysis models, the small number of studies does not allow for much power to detect outlying studies. Studies that may stick out in a scatter plot are not necessarily outliers and the FS algorithm provides an alternative way to detect studies with a disproportionate effect on various components of the model. Outlying studies should not be dropped merely because of large intervention effects. A different model such as a meta-regression may accommodate these studies and the FS algorithm can be used to evaluate the overall fit of the model to the data.
There is a plethora of different objective functions that can be used to initialize and progress in the search and our analyses (not reported here) show that the FS is robust to the choice of the objective function. The FS algorithm can be applied to any meta-analysis model and is expected to be more helpful in multivariate problems (e.g. many regressors, many outcomes, and many treatments). In case those suspicious studies are detected, a close examination of them may identify potential covariates that affect the results.

The FS should be used as a diagnostic tool that aids clinical judgement and interpretation of results and not as a strict rule for omitting or downweighting studies. We strongly suggest accompanying forward plots by simulation envelopes to explore if changes can be attributed to random variation and also to repeat the search a small number of times (e.g. 10) to explore its robustness. Studies may be spotted as outlying because of small study effects. Mega-trials are expected to influence heavily summary estimates and therefore observe large Cook distances. These trials are rarely outliers but render identification of outliers more difficult since they dominate the analysis. Overall, the FS is a robust diagnostic method that may reveal important

information about the studies in a meta-analysis but it should not be overinterpreted. A study lying outside the simulation envelopes is an outlier when the studies at the '*basic set*' are considered and not necessarily an outlier when all studies are considered. The method is not time consuming and it can be run within a few seconds even for large datasets (e.g. the Fluoride example). While it is possible to alter the algorithm to allow studies to leave the 'basic set', we recommend not to because it ensures smoother plots in the beginning of the search. We noticed that we get smoother and more easily interpretable plots. The small number of studies results in large permutations in the studies in the 'basic' and 'non-basic' sets. It is possible that this will not be a problem with large number of studies. Even if an outlying effect size is included in the initial subset forward plots of residuals will identify it. Also, with a small number of studies, if many studies leave or enter the search it is very difficult to isolate the influence of a specific study.

Reference List

Atkinson AC. 1986. Masking unmasked. *Biometrika* 73: 533-541.

Atkinson AC. 1994. Fast very robust methods for the detection of multiple outliers. *Journal of the American Statistical Association* 89: 1329-1339.

Atkinson AC, Riani M. 2000. Robust Diagnostic Regression Analysis. New York: Springer.

Atkinson AC, Riani M, Cerioli A. 2004. Exploring Multivariate Data with the Forward Search. New York: Springer.

Baker R, Jackson D. 2008. A new approach to outliers in meta-analysis. *Health Care Management Science* 11: 121-131.

Bangert-Drowns RL, Hurley MM, Wilkinson B. 2004. The effects of school-based writing-to-learn interventions on academic achievement: a meta-analysis. *Review of Educational Research* 74: 29-58.

Beath KJ. 2014 A finite mixture method for outlier detection and robustness in meta-analysis. *Research Synthesis Method* 5(4):285-293.

Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. 2010. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods* 1(2): 97-111.

Cook RD, Weisberg S. 1982. Residuals and Influence in Regression. London: Chapman and Hall.

Deeks JJ, Higgins JPT, Altman DG. 2011. Analysing data and undertaking meta-analyses. In *Cochrane Handbook for Systematic Reviews of Interventions* (eds J.P.T. Higgins and S. Green).The Cochrane collaboration.

DerSimonian RD, Laird N. 1986. Meta-analysis in clinical trials. *Controlled Clinical Trials* 7: 177-188.

Efron B, Tibshirani RJ. 1993. An ntroduction to the Bootstrap. Chapman & Hall.

Egger M, Smith GD, Scheider M, Minder C. 1997. Bias in meta analysis detected by a simple graphical test. *British Medical Journal* 315:629-634.

Gumedze FN, Jackson D. 2011. A random effects variance shift model for detecting and accommodating outliers in meta-analysis. *BMC Medical Research Methodology* 11.

Hadi AS. 1992. Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society, Series B* 54:761-771.

Hadi AS, Simonoff JS. 1993. Procedures for the identification of multiple outliers in linear models. *Journal of the American Statistical Association* 88:1264-1272.

Hedges LV, Olkin I. 1985. Statistical methods for Meta-Analysis. New York: Acedemic Press.

Hedges LV. 1992. Meta-Analysis. Journal of Educational and Behavioral Statistics 17:279-296.

Higgins JPT, Thompson SG. 2002. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* 21:1539-1558.

Higgins JPT, Green S. 2008. Cochrane Handbook for Systematic Reviews of Interventions.
Jackson D, White IR, Riley RD. 2012. Quantifying the impact of between-study heterogeneity in multivariate meta-analyses. *Statistics in Medicine* 31:3805-3820.

Lee KJ, Thompson SG. 2008. Flexible parametric models for random-effects distributions. *Statistics in Medicine* 27:418-434.

Marinho V, Higgins JPT, Logan S, Sheiham A 2003. Fluoride toothpastes for preventing dental caries in children and adolescents. *Cochrane Database of Systematic Reviews (1):CD002278*

MATLAB, The MathWorks, Inc., Natick, Massachusetts, United States.

Mavridis D, Moustaki I. 2008. Detecting outliers in factor analysis using the forward search algorithm. *Multivariate Behavioral Research* 43:453-475.

Mavridis D, Moustaki I. 2009. The forward search algorithm for detecting aberrant response patterns in factor analysis for binary data. *Journal of Computational and Graphical Statistics* 18:1016-1034.

Mavridis D, Salanti G. 2013. A practical introduction to multivariate meta-analysis. *Statistical Methods in Medical Research* 22:133-158.

Mavridis D, Sutton A, Cipriani A, Salanti G. 2013. A fully Bayesian application of the Copas selection model for publication bias extended to network meta-analysis. *Statistics in Medicine* **32:** 51-66.

Nikolakopoulou A, Mavridis D, Salanti G. 2014. Demystifying fixed and random effects meta-analysis . *Evidence Based Mental Health* 17(3):53-57.

Oxman AD, Bjorndal A, Becerra-Posada F, Gibson M, Block MA, Haines A, Hamid M,Odom CH, Lei H, Levin B, Lipsey MW, Littell JH, Mshinda H, Ongolo-Zogo P, Pang T, Sewankambo N, Songane F, Soydan H, Torgerson C, Weisburd D, Whitworth J, Wibulpolprasert S. 2010. A

framework for mandatory impact evaluation to ensure well informed public policy decisions. *Lancet* 375:427-431.

Patsopoulos NA, Analatos AA, Ioannidis JP 2005. Relative citation impact of various study designs in the health sciences. *JAMA* 293:2362-2366.

Rousseeuw PJ. 1984. Least median of squares regression. *Journal of the American Statistical Association* 79:871-880.

Salanti G. 2012. Indirect and mixed-treatment comparison, network, or multiple-treatment meta-analysis: many names, many benefits, many concerns for the next generation evidence synthesis tool. Research Synthesis Methods 3(2):80-97.

Sidik K, Jonkman JN. 2007. A comparison of heterogeneity variance estimators in combining results of studies. *Statistics in Medicine* 26:1964-1981.

Sterne JAC, Egger M, Moher D. 2008. Addressing reporting biases. In *Cochrane Handbook for Systematic Reviews of Interventions* (eds J.P.T. Higgins and S. Green).The Cochrane collaboration.

Viechtbauer W, Cheung MWL. 2010. Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods* 1(2):112-125.