Received: 23 July 2019 | Revised: 14 April 2021 | Accepted: 28 June 2021

DOI: 10.1002/sim.9145

RESEARCH ARTICLE

Statistics in Medicine WILEY

A forward search algorithm for detecting extreme study effects in network meta-analysis

Maria Petropoulou[1,2] | **Georgia Salanti**[3] | **Gerta Rücker**[1] | **Guido Schwarzer**[1] | **Irini Moustaki**[4] | **Dimitris Mavridis**[2,5]

[1]Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Germany

[2]Evidence Synthesis Method Team, Department of Primary Education, University of Ioannina School of Education, Ioannina, Greece

[3]Institute of Social and Preventive Medicine, University of Bern, Bern, Switzerland

[4]Department of Statistics, London School of Economics and Political Science, London, UK

[5]Faculté de Médecine, Université Paris Descartes, Paris, France

Correspondence
Maria Petropoulou, Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Germany.
Email: petropoulou@imbi.uni-freiburg.de

Funding information
Hellenic Foundation for Research and Innovation; General Secretariat for Research and Technology

In a quantitative synthesis of studies via meta-analysis, it is possible that some studies provide a markedly different relative treatment effect or have a large impact on the summary estimate and/or heterogeneity. Extreme study effects (outliers) can be detected visually with forest/funnel plots and by using statistical outlying detection methods. A forward search (FS) algorithm is a common outlying diagnostic tool recently extended to meta-analysis. FS starts by fitting the assumed model to a subset of the data which is gradually incremented by adding the remaining studies according to their closeness to the postulated data-generating model. At each step of the algorithm, parameter estimates, measures of fit (residuals, likelihood contributions), and test statistics are being monitored and their sharp changes are used as an indication for outliers. In this article, we extend the FS algorithm to network meta-analysis (NMA). In NMA, visualization of outliers is more challenging due to the multivariate nature of the data and the fact that studies contribute both directly and indirectly to the network estimates. Outliers are expected to contribute not only to heterogeneity but also to inconsistency, compromising the NMA results. The FS algorithm was applied to real and artificial networks of interventions that include outliers. We developed an R package (NMAoutlier) to allow replication and dissemination of the proposed method. We conclude that the FS algorithm is a visual diagnostic tool that helps to identify studies that are a potential source of heterogeneity and inconsistency.

KEYWORDS
Cook's distance, forward search, network meta-analysis, NMAoutlier, outliers

1 | INTRODUCTION

In most healthcare conditions, we have to evaluate several competing interventions. Network meta-analysis (NMA) is an extension of pairwise meta-analysis that allows for multiple treatment comparisons by synthesizing direct and indirect evidence.[1-5] Transitivity is a fundamental assumption in NMA, stating that the distribution of effect modifiers is similar across treatment comparisons.[1] The statistical manifestation of transitivity is the consistency assumption, implying that direct and indirect evidence agree.

An outlier is defined as a study with a markedly different intervention effect estimate for a given treatment comparison.[6] A study that influences aspects of the model such as parameter estimates, heterogeneity, inconsistency is defined as an influential study. A study that is an outlier is not necessarily an influential one (eg, an extremely large effect from a very small study has little influence on the results of the model) and vice versa.

In a pairwise meta-analysis, we can visually detect extreme study effects through forest and funnel plots. Several statistical methods have been suggested to accommodate the results from outliers within a meta-analysis by allowing for flexible distributions of the random effects. Lee and Thompson argued that normality might be a restrictive assumption for the random-effects model and they provided alternative distributions with heavier tails.[7] Baker and Jackson also suggested alternative distributions that downweigh outlying studies, such as long-tailed distributions[8] and marginal distributions with additional parameters to model skewness and heavier tails.[9] A random-effects variance shift outlier model is also capable of identifying and downweighing outliers.[10] Beath proposed a method that considers a mixture of outlying and nonoutlying studies and downweighs the former.[11] Most of the outlier detection techniques are extensions of methods that have been applied to regression models. Alternative heterogeneity measures in meta-analysis have recently been proposed by Lin et al that are robust in the presence of outliers.[12] Viechtbauer and Cheung extended standard outlier deletion diagnostic measures in the context of meta-analysis[13] and included them in the R package **metafor**.[14,15]

In NMA, the extreme study effect can be visualized with a comparison-adjusted funnel plot[16] (eg, if the study markedly differs from the others for a given treatment comparison). The effect size can be rendered as aberrant not only by its mere magnitude but also by its size conditional on the comparison of the study and/or the corresponding effect derived from indirect evidence. For example, a null effect might be aberrant if all other studies in the same comparison have large effects or if the indirect evidence for that comparison suggests a large effect. Outlying and influential studies may be responsible for large heterogeneity and/or inconsistency in NMA compromising the validity of results.

Within the Bayesian NMA framework, Lu and Ades proposed the use of residual deviance,[17] Zhang et al[18] provided four measures for the detection of outlying studies by fitting the Bayesian hierarchical NMA model, and Zhao et al[19] extended several outlier detection measures for generalized hierarchical models to detect influential and outlying studies in NMA. Within a frequentist framework, Noma et al recently provided outlier diagnostics for the NMA model using multivariate random-effects meta-regression.[20]

Backward algorithms are widely used to detect outlying observations and can be potentially used in NMA. They start by removing observations according to some criterion (eg, largest residual) and stop when some other criterion is met (eg, all residuals are smaller than a threshold value).[21] The main drawback of backward methods is that in the presence of a cluster of outliers it is likely that results would be affected to such a degree that outliers will not be identified as such (masking). According to Atkinson, there are several deletion methods employed in backward methods that fail to detect outlying observations due to masking.[22]

In this article, we propose a forward search (FS) algorithm to detect studies with extreme results in the NMA model. The FS algorithm was initially developed as an outlier detection tool for the estimation of covariance matrices[23] and regression models.[24,25] It was subsequently extended to standard multivariate methods,[26] factor analysis,[27] and item response theory models[28] and was recently applied in meta-regression.[29] FS starts by fitting the hypothesized data generating model to a subset of the data which is gradually incremented by adding the remaining studies according to their closeness to the postulated model. In each step of the FS algorithm, parameter estimates, measures of fit, and goodness-of-fit test statistics are monitored, and sharp changes indicate the outlying behavior of the studies or observations entering the initial subset. An R package (**NMAoutlier**)[30] has been developed that allows the reproduction of our results and the application of the method to other data.

The article is organized as follows: Section 2 discusses motivating examples; Section 3 discusses the random effects NMA model using graph-theoretical methods as introduced by Rücker[31]; Section 4 outlines the methodological extension of the FS algorithm to the NMA model; Section 5 presents an application of the proposed methodology in published NMAs and simulated datasets; Section 6 discusses the main findings and provides directions for using the proposed diagnostic methodology for NMA; and Section 7 contains our conclusion.

2 | MOTIVATING EXAMPLES

The first example comprises four interventions to aid smoking cessation.[17,32] Twenty-four studies ($N = 24$), including 22 two-arm trials and two three-arm trials, compared the relative effects of four smoking cessation counseling programs ($n = 4$): defined as no contact (A), self-help (B), individual counselling (C), and group counselling (D). The outcome was whether an individual successfully stopped smoking at 6 to 12 months (binary) and the odds ratio was used as a

summary measure. The dataset with arm-level data is included in the R package **netmeta**[33] and the corresponding R code to calculate odds ratios with the `pairwise` function is provided in Appendix A1. The study-level data and the odds ratios are provided in Table A1.

Figure 1 (left side) shows the comparison-adjusted funnel plot[16] with interventions within comparisons ordered according to effectiveness: (1) no contact (A), (2) self-help (B), (3) group counseling (D), and (4) individual counseling (C). We can see that studies 3 and 7 lie far away from the bulk of the data judging from the large effect sizes given their sizes. However, these deviations could be genuine or due to chance and heterogeneity. Figure 2 (left side) provides the network plot for smoking cessation data.

In the second example, Gupta and Paquet[34] compared placebo and eight active interventions (denoted as treatments 1-9) for actinic keratosis. Thirty-five studies ($N = 35$), including three three-arm trials, compared the relative effects of placebo and eight active interventions. The outcome was participant complete clearance or equivalent and the odds ratio was used as a summary measure. The dataset and the actual treatments are provided in Table A2. Figure 1 (right side) provides the comparison-adjusted funnel plot[16] with interventions within comparisons ordered from treatment 1 to 9. We can see that study 28 with treatment comparisons 1 vs 6 vs 8 has a large effect size given its size for the treatment comparison 1 vs 6. Figure 2 (right side) provides the network plot for the actinic keratosis data.

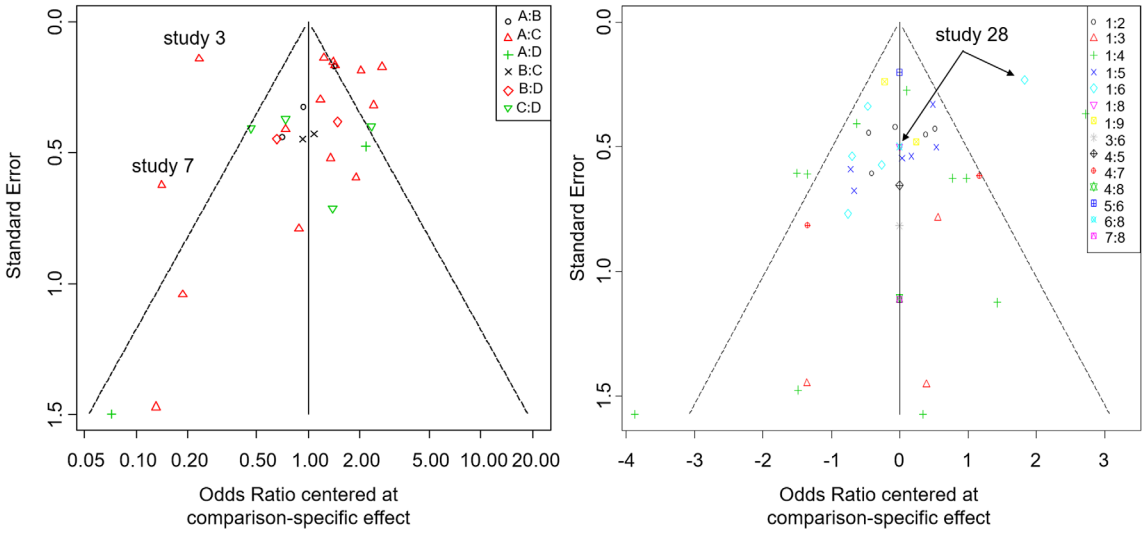

FIGURE 1 Comparison-adjusted funnel plot[16] for smoking cessation data (left)[17,32] and actinic keratosis[34] (right side). Comparison-adjusted funnel plot produced in R[15] from **netmeta** package.[33] The *y*-axis provides the SE, and the *x*-axis provides the odds ratio centered at comparison-specific effect [Colour figure can be viewed at wileyonlinelibrary.com]

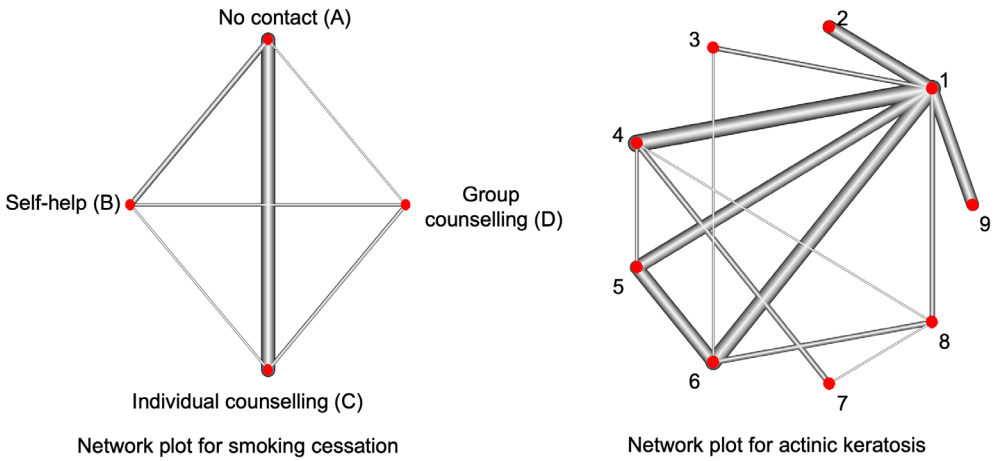

FIGURE 2 Network plot for smoking cessation data[17,32] (left side) and actinic keratosis[34] (right side) [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 1 Notation for the methodology of the FS algorithm in NMA

Data: Studies $i = 1, \dots, N$ Treatments $1, \dots, n$ Observed pairwise comparisons $1, \dots, m$ Observed pairwise effect sizes $\boldsymbol{y} = (y_1, y_2, \dots, y_m)'$ and corresponding observed standard errors $\boldsymbol{s} = (s_1^2, s_2^2, \dots, s_m^2)'$

Matrix of NMA model: $\mathbf{X}$ the design $m \times n$ matrix $\widetilde{\mathbf{X}}$ the reduced design matrix with dimensions $(n-1) \times n$ $\mathbf{W}$ the $m \times m$ diagonal weight matrix $\mathbf{L}^+$the Moore-Penrose pseudoinverse $n \times n$ matrix of $\mathbf{X}$ $\mathbf{L}^+ = (\mathbf{L} - \mathbf{J}/n)^{-1} + \mathbf{J}/n$ where $\mathbf{J}$ is a $n \times n$ matrix with all elements equal to 1 In the case of $d_i (> 2)$arms of study i $\mathbf{L}^+ = -\frac{1}{2d_i^2}\mathbf{X}'\mathbf{X}\mathbf{V}\mathbf{X}'\mathbf{X}$ where $\mathbf{V}$ is a $d_i \times d_i$ symmetric matrix with the observed variances of all comparisons. $\widetilde{\mathbf{X}}\mathbf{L}^+\widetilde{\mathbf{X}}'$ the $(n-1) \times (n-1)$ variance-covariance matrix of $n-1$ relative treatment estimates

Estimated model parameters Treatment effects $\widehat{\boldsymbol{\mu}}$ Heterogeneity variance $\widehat{\tau}^2$ $\widetilde{\boldsymbol{\mu}}$ the $n-1$ relative treatment estimates compared with the reference

FS algorithm notation: $l = \max(n, 0.2 \times N)$ the size of the initial subset Pa large number of randomly chosen initial subsets of size l $p = 1, \dots, P$ each candidate initial subset of size l $j = 1, \dots, N-l$ each iteration of the FS algorithm

Steps of FS algorithm: *For selecting the initial basic set:* D_p^l each candidate initial subset $p = 1, \dots, P$ of l studies $(\widehat{\boldsymbol{\mu}}_{D_p^l}, \widehat{\tau}^2_{D_p^l})$ estimates corresponding to the subset D_p^l median$(f(\boldsymbol{y}_i, \boldsymbol{s}_i, \boldsymbol{X}_i, \widehat{\boldsymbol{\mu}}_{D_p^l}, \widehat{\tau}^2_{D_p^l}))$ is the objective function with observations $\boldsymbol{y}_i, \boldsymbol{s}_i, \boldsymbol{X}_i$ of the entire dataset. *For the first iteration* $j = 1$: D^linitial basic set, $(D^l)^c$ nonbasic set $(\widehat{\boldsymbol{\mu}}_{D^l}, \widehat{\tau}^2_{D^l})$ subset-specific estimates for the initial basic set D^l $f(\boldsymbol{y}_i, \boldsymbol{s}_i, \boldsymbol{X}_i, \widehat{\boldsymbol{\mu}}_{D^l}, \widehat{\tau}^2_{D^l})$ objective function with observations $\boldsymbol{y}_i, \boldsymbol{s}_i, \boldsymbol{X}_i \in (D^l)^c$ *For iterations* $j = 2, \dots .N-l$: D^{l+j} basic set, $(D^{l+j})^c$ nonbasic set $(\widehat{\boldsymbol{\mu}}_{D^{l+j}}, \widehat{\tau}^2_{D^{l+j}})$ subset-specific estimates for the basic set D^{l+j} $f(\boldsymbol{y}_i, \boldsymbol{s}_i, \boldsymbol{X}_i, \widehat{\boldsymbol{\mu}}_{D^{l+j}}, \widehat{\tau}^2_{D^{l+j}})$ objective function with observations $\boldsymbol{y}_i, \boldsymbol{s}_i, \boldsymbol{X}_i \in (D^{l+j})^c$

Outlier diagnostics measures: *Standardized residuals* $\widehat{\varepsilon}_i^{\text{stand}} = \sqrt{\frac{1}{s_i^2 + \widehat{\tau}^2}}(y_i - \boldsymbol{X}_i\widehat{\boldsymbol{\mu}})$ for the ith two-arm study $\widehat{\varepsilon}_i^{\text{stand}} = \text{mean}(\widehat{\boldsymbol{\varepsilon}}_i^{\text{stand}})$, where $\widehat{\boldsymbol{\varepsilon}}_i^{\text{stand}} = \begin{pmatrix} \widehat{\varepsilon}_i^{1,\text{stand}}, \dots, \widehat{\varepsilon}_i^{\binom{d_i}{2},\text{stand}} \end{pmatrix}'$ in case of $d_i (> 2)$arms of study i *Cook's statistic*

$C_j = (\widetilde{\boldsymbol{\mu}}_{D^{l+j}} - \widetilde{\boldsymbol{\mu}}_{D^{l+j-1}})'(\widetilde{\mathbf{X}}_{D^{l+j}}\mathbf{L}^+_{D^{l+j}}\widetilde{\mathbf{X}}'_{D^{l+j}})^{-1}(\widetilde{\boldsymbol{\mu}}_{D^{l+j}} - \widetilde{\boldsymbol{\mu}}_{D^{l+j-1}})$ at iteration $j = 1, \dots, N-l$ *The ratio of the determinants of the variance-covariance matrix* $\text{COVRATIO}_j = \frac{\det(\widetilde{\mathbf{X}}_{D^{l+j}}\mathbf{L}^+_{D^{l+j}}\widetilde{\mathbf{X}}'_{D^{l+j}})}{\det(\widetilde{\mathbf{X}}_{D^{l+j-1}}\mathbf{L}^+_{D^{l+j-1}}\widetilde{\mathbf{X}}'_{D^{l+j-1}})}$ at iteration j to iteration $(j-1)$

Abbreviations: FS, forward search; NMA, network meta-analysis.

In the third example, Sciarretta et al[35] provided a synthesis of 26 studies ($N = 26$), comparing antihypertensive strategies for heart failure prevention (Figure A1, left side).

3 | NMA MODEL

We use the frequentist random-effects NMA model as presented by Rücker,[31] which uses all pairwise comparisons within multiarm trials by reducing their weight in the NMA.[36] We briefly describe the approach that has been implemented in the R package **netmeta**[33]; for more details see articles.[31,36,37] The notation used is summarized in Table 1.

Suppose that we have N studies and each study has d_i arms, $i = 1, \dots, N$. Let m denote the number of observed pairwise comparisons ($m = \sum_{i=1}^{N}\binom{d_i}{2}$ and $m = N$ if $d_i = 2, i = 1, \dots, N$). Let us denote with n the total number of treatments. Let $\boldsymbol{\mu}$ represent the vector with the n absolute treatment effects. Let $\boldsymbol{y} = (y_1, y_2, \dots, y_m)'$ be the vector with the observed effect sizes from the N studies and $\boldsymbol{s} = (s_1^2, s_2^2, \dots, s_m^2)'$ the vector with the corresponding observed standard errors.

Assuming a common heterogeneity variance τ^2 across pairwise comparisons, the random effects NMA model is written as

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\mu} + \boldsymbol{\delta} + \boldsymbol{\varepsilon}, \boldsymbol{\delta} \sim N(0, \boldsymbol{\Delta}), \boldsymbol{\varepsilon} \sim N(0, \mathbf{S})$$

where $\mathbf{S}$ is a block diagonal within-study variance-covariance matrix with data entries $s_1^2, \dots, s_m^2$ in the diagonal and $\mathbf{X}$ is the $m \times n$ design matrix that describes the structure of the network, with rows denoting the observed pairwise comparisons and columns the treatments being compared within each comparison.[31,36] We consider the true variances to be equal to the observed sample variances, an assumption that holds when sample sizes are reasonably large. $\boldsymbol{\Delta}$ denotes a block diagonal between-study variance-covariance matrix with the heterogeneity variances τ^2 in the diagonal and is estimated from the data. The between-study variance is estimated using a special case of the generalized DerSimonian-Laird estimator.[38,39]

Let $\mathbf{W}$ be a $m \times m$ diagonal *weight matrix* with a vector of weights in its diagonal to be the observed inverse study variance of all existing comparisons. The Laplacian $n \times n$ matrix is given by $\mathbf{L} = \mathbf{X}'\mathbf{W}\mathbf{X}$.[31,36] To estimate treatment effects, the Moore Penrose pseudoinverse $n \times n$ matrix $\mathbf{L}^+$of the Laplacian matrix $\mathbf{L}$ is constructed.[31,36] In the case of multiarm studies ($d_i > 2$), standard errors are recalculated (increased) with a back-calculation adjustment as described in Rücker and Schwarzer,[36] $s^2_{\text{adjusted},i}$, and new reduced weights are derived. In that case, the Laplacian matrix is given with $\mathbf{L}^+ = -\frac{1}{2d_i^2}\mathbf{X}'\mathbf{X}\mathbf{V}\mathbf{X}'\mathbf{X}$ having $\mathbf{V}$ to be a $d_i \times d_i$ symmetric matrix with the observed variances of all comparisons.

We define the vector $\widehat{\boldsymbol{\mu}}$ of dimension n that represents the effects of the interventions and a vector $\widetilde{\boldsymbol{\mu}}$ of dimension $n-1$ that represents the relative effects of the interventions to a reference treatment. The $(n-1) \times (n-1)$ variance-covariance matrix of $\widetilde{\boldsymbol{\mu}}$ is $\widetilde{\mathbf{X}}\mathbf{L}^+\widetilde{\mathbf{X}}'$ where $\mathbf{L}^+$ is the Moore-Penrose pseudoinverse $n \times n$ matrix of $\mathbf{L}$ and $\widetilde{\mathbf{X}}$ is the reduced design matrix of dimensions $(n-1) \times n$ referring to the interventions reported in $\widetilde{\boldsymbol{\mu}}$ (all but the reference one).

Table 1 provides the notation for the FS algorithm in NMA.

4 | EXTENSION OF THE FS ALGORITHM TO NMA

Most methods used for outlier detection opt to divide the data into two parts: a large clean part and the outliers. FS starts by selecting candidate subsets of *likely* outlier-free studies and proceeds by adding one-by-one studies until all are included. FS consists of three stages (choice of the initial subset, progression of the search, monitoring of the search).

In the first stage, FS chooses the initial subset of studies by selecting a candidate subset of *likely* outlier-free studies. We conventionally refer to this subset as the initial subset or the "basic" set at the beginning of the search. Studies not included in this basic set constitute the "nonbasic" set. A data generating (hypothesized) model is assumed to fit the data in the initial subset.

In the stage of progression of the search, the method gradually adds studies, one-by-one, from the nonbasic to the basic subset based on how close the study in the nonbasic set is to the hypothesized model in the basic set using some objective functions. This process is repeated until all studies are included in the basic set.

In the monitoring stage, estimated model parameters, measures of model fit, and goodness-of-fit test statistics are monitored in each step/iteration. A sharp change in the monitoring measures can be an indication of an outlying study. Moreover, ordering the studies based on how close they are to the basic set makes outlying studies more likely to be entered in the last iterations.

Below we present each step of the algorithm in detail.

4.1 | Choice of the initial subset

When selecting studies for the initial subset we need to ensure that all n treatments are included and that the resulting network is connected. The requirement of network connectivity for each candidate subset of studies is evaluated with the `netconnection` function in the **netmeta** package.[33]

Selecting the size l of the initial subset. The number of parameters in a NMA with n treatments is n ($n-1$ relative treatment effects estimates and a single heterogeneity parameter). We require the initial subset to include all n treatments. Inclusion of the number of studies equal to the number of treatments or the number of treatments minus 1 suffices if there are only two-arm studies included. The requirement can be satisfied with fewer studies in the case of multiarm studies and for some network structures with two-arm studies (eg, consider a network of studies that compare the treatments A, B, C with study comparisons A vs B and A vs C). Large initial sets can save computation time and prevent large fluctuations in the parameter estimation during the first steps of the search, but at the same time increase the chance of including outliers in the initial subset. This is not necessarily a drawback but, in such cases, it is useful to repeat the search a couple of times from random starting points. We choose to set the size equal to the maximum of the number of treatments and 20% of the total number of studies; that is, $l = \max(n, 0.2 \times N)$. Other rules can be adopted.

Selecting the studies to include in the initial subset. We start with a subset of studies that ideally is outlier-free to use as the initial subset. We consider a large number of potential sets (P) of randomly chosen initial subsets of studies each of size l. We require each chosen initial subset of studies to be a connected subnetwork including all comparative interventions. If the total number of potential subsets $\binom{N}{l}$ is not very large, we can provide an exhaustive search of all subsets of studies aiming to identify the subset that is the most likely subset to be outlier-free. Alternatively, for large

networks, an exhaustive analysis is prohibitive and practically unnecessary. In such cases, we may explore a large number of initial subsets (the larger the network, the larger the number of subsets to investigate for example, 100). We can measure the fit of the NMA model for each candidate initial subset of studies using an objective function. The objective function evaluates candidate subsets and returns a measure of their fit. The better the fit of a subset, the more likely it is outlier-free.

Let us denote with D_p^l each candidate initial subset $p = 1, \dots, P$ of size l. We obtain the subset-specific estimates $(\widehat{\boldsymbol{\mu}}_{D_p^l}, \widehat{\tau}^2_{D_p^l})$ of each subset D_p^l and calculate the objective function median$(f(\boldsymbol{y}_i, \boldsymbol{s}_i, \boldsymbol{X}_i, \widehat{\boldsymbol{\mu}}_{D_p^l}, \widehat{\tau}^2_{D_p^l}))$ with observations $\boldsymbol{y}_i, \boldsymbol{s}_i, \boldsymbol{X}_i, i = 1, \dots, N$ from the complete dataset.

Examples of objective functions can be defined as the median of the absolute standardized residuals or the median of the absolute log-likelihood contributions given by the median $(f(\boldsymbol{y}_i, \boldsymbol{s}_i, \boldsymbol{X}_i, \widehat{\boldsymbol{\mu}}_{D_p^l}, \widehat{\tau}^2_{D_p^l}))$ with where $\widehat{\varepsilon}_{i,D_p^l}$ is the standardized residual for each study defined below, $\log(w_i) - (\widehat{\varepsilon}_{i,D_p^l})^2$ is the log-likelihood contribution (a proof of Equation (2) is given in Appendix B1), $w_i = 1/(s_i^2 + \widehat{\tau}^2_{D_p^l})$ is the weight for each comparison in each study and s_i^2 is adjusted to take account of a multiarm study ($d_i > 2$). Alternatively, we may consider the mean or some other quantile of f. We considered the median because it resembles the median least of squares regression suggested by Rousseeuw[40] (and it is a robust alternative to the classical least squares estimator) and it was also considered by Atkinson and Riani[24] in the FS development. Either way, our goal is to optimize the objective function defined.

The standardized residual of a pairwise comparison for a two-arm study is given by.

$\widehat{\varepsilon}_{i,D_p^l} = \sqrt{\frac{1}{s_i^2 + \widehat{\tau}^2_{D_p^l}}}(y_i - \boldsymbol{X}_i \widehat{\boldsymbol{\mu}}_{D_p^l}), i = 1, \dots, N$. For a multiarm study, we take the arithmetic mean of the standardized residuals or the log-likelihood contributions of all pairwise comparisons in this study, that is, for standardized residuals we take $\widehat{\varepsilon}_{i,D_p^l} = \text{mean}(\widehat{\boldsymbol{\varepsilon}}_{i,D_p^l})$ with $\widehat{\boldsymbol{\varepsilon}}_{i,D_p^l} = \begin{pmatrix} \widehat{\varepsilon}^1_{i,D_p^l}, \dots, \widehat{\varepsilon}^{\binom{d_i}{2}}_{i,D_p^l} \end{pmatrix}'$ denoting the vector of all standardized residual terms within a d_i-arm study. For log-likelihood contributions, in the case of a multiarm study, we take $\log(w_i) - (\widehat{\varepsilon}_{i,D_p^l})^2 = \text{mean}(log(\boldsymbol{w}_i) - (\widehat{\boldsymbol{\varepsilon}}_{i,D_p^l})^2)$ with $\boldsymbol{w}_i = \begin{pmatrix} w_i^1, \dots, w_i^{\binom{d_i}{2}} \end{pmatrix}'$ and $\widehat{\boldsymbol{\varepsilon}}_{i,D_p^l} = \begin{pmatrix} \widehat{\varepsilon}^1_{i,D_p^l}, \dots, \widehat{\varepsilon}^{\binom{d_i}{2}}_{i,D_p^l} \end{pmatrix}'$.

Among the P candidate subsets D_p^l, the subset that optimizes the objective function is considered as the initial subset (eg, minimize the median of Equation 1 or maximize the median of Equation 2).

4.2 | Progressing in the search

For brevity, we drop the subindex p from the initial set D_p^l and we denote the initial basic set with D^{l+j} and the complementary nonbasic set with $(D^{l+j})^c$ at iteration $j = 1, 2, \dots N - l$. In the first step of the algorithm ($j = 1$), we calculate the objective function median$(f(\boldsymbol{y}_i, \boldsymbol{s}_i, \boldsymbol{X}_i, \widehat{\boldsymbol{\mu}}_{D^{l+1}}, \widehat{\tau}^2_{D^{l+1}}))$ for each study in the initial nonbasic set $\boldsymbol{y}_i, \boldsymbol{s}_i, \boldsymbol{X}_i \in (D^{l+1})^c$ using $\widehat{\boldsymbol{\mu}}_{D^{l+1}}, \widehat{\tau}^2_{D^{l+1}}$ estimated from the basic set D^{l+1}. This measures the closeness between the basic set D^{l+1} and each study of the nonbasic set that is a candidate for addition to the basic set. The study optimizing the objective function (the median of Equation 1 or the median of Equation 2) is added to the basic set.

We proceed with the algorithm for $j = 2, \dots N - l$ until all studies are included in the basic set. At iteration j, there are $l + j$ studies in the enlarged basic set denoted as D^{l+j} and $N - l - j$ studies in the nonbasic set denoted by $(D^{l+j})^c$. For the basic set D^{l+j}, the subset-specific estimates are denoted by $(\widehat{\boldsymbol{\mu}}_{D^{l+j}}, \widehat{\tau}^2_{D^{l+j}})$. For each iteration j, we compute the objective function median$(f(\boldsymbol{y}_i, \boldsymbol{s}_i, \boldsymbol{X}_i, \widehat{\boldsymbol{\mu}}_{D^{l+j}}, \widehat{\tau}^2_{D^{l+j}}))$, the median of absolute standardized residuals (1) or the median of absolute log-likelihood contributions (2), for each observation $\boldsymbol{y}_i, \boldsymbol{s}_i, \boldsymbol{X}_i \in (D^{l+j})^c$.

4.3 | Monitoring the search

In each iteration, parameter estimates, model diagnostic statistics, ranking metrics that provide treatment hierarchy, heterogeneity, and inconsistency are monitored using a plot (*forward plot*). Forward plots visually convey the influence of each study.

4.3.1 | Outlier case diagnostics measures

The standardized residual for the pairwise comparison of a two-arm study i is given by $\hat{\varepsilon}_i^{\text{stand}} = \sqrt{\frac{1}{s_i^2+\hat{\tau}^2}}(y_i - \boldsymbol{X_i}\hat{\boldsymbol{\mu}})$. In the case of a multiarm study with d_i (> 2) arms, $i = 1, \ldots, N$, the standardized residual is calculated as an arithmetic mean of the standardized residuals of all pairwise treatment comparisons, $\hat{\varepsilon}_i^{\text{stand}} = \text{mean}(\hat{\boldsymbol{\varepsilon}}_i^{\text{stand}})$, where $\hat{\boldsymbol{\varepsilon}}_i^{\text{stand}} = \left(\hat{\varepsilon}_i^{1,\text{stand}}, \ldots, \hat{\varepsilon}_i^{\binom{d_i}{2},\text{stand}}\right)'$ denoting the vector of all standardized residual terms within a d_i-arm study.

To explore the impact of adding a study on summary relative treatment estimates we define modified Cook's statistics for NMA (in analogy to those described in pairwise meta-analysis[13]) as

$$C_j = (\widetilde{\boldsymbol{\mu}}_{D^{l+j}} - \widetilde{\boldsymbol{\mu}}_{D^{l+j-1}})'(\widetilde{\mathbf{X}}_{D^{l+j}}\, \mathbf{L}^{+}_{D^{l+j}}\, \widetilde{\mathbf{X}}'_{D^{l+j}})^{-1}(\widetilde{\boldsymbol{\mu}}_{D^{l+j}} - \widetilde{\boldsymbol{\mu}}_{D^{l+j-1}})$$

where $\widetilde{\boldsymbol{\mu}}_{D^{l+j}}$ and $\widetilde{\boldsymbol{\mu}}_{D^{l+j-1}}$ are the relative treatment estimates at iteration $j, j-1$, respectively. A general rule provided in the bibliography for a cut-off value of Cook's statistic is that the study j is considered an outlier and/or influential if $C_j > 1$.[41,42]

The influence of a study can also be assessed by the change that incurs to model fitting. We can compute the ratio of the determinants of the variance-covariance matrix of relative treatment estimates at iteration j to iteration $(j-1)$[13] for NMA as

$$\text{COVRATIO}_j = \frac{\det(\widetilde{\mathbf{X}}_{D^{l+j}}\, \mathbf{L}^{+}_{D^{l+j}}\, \widetilde{\mathbf{X}}'_{D^{l+j}})}{\det(\widetilde{\mathbf{X}}_{D^{l+j-1}} \mathbf{L}^{+}_{D^{l+j-1}} \widetilde{\mathbf{X}}'_{D^{l+j-1}})}.$$

A proof showing that these definitions (Cook's distance, ratio of determinants of the variance-covariance matrix of treatment estimates) generalize the classical measures to NMA is given in Appendix B2.

4.3.2 | Heterogeneity and inconsistency measures

Based on the fixed effects model and assuming homogeneity and consistency in the whole network, the generalized Cochran's Q statistic is given by Krahn et al[43]

$$Q^{\text{total}} = (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\mu}})'\mathbf{W}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\mu}})$$

Q^{total} can be decomposed into two parts[43]:

- a part coming from *within designs* (heterogeneity between studies that compare the same set of treatments), Q^{het}
- a part coming from *between designs* (inconsistency between studies that compare different sets of treatments), $Q^{\text{inc}}_{\text{FE}}$

where the *design* of a study is called the set of treatments compared within the study in the context of NMA.[2,44] For the FS procedure, we monitor generalized Cochran's Q (Q^{total}) and the Q statistic within designs (Q^{het}). Moreover, the Q statistic (Q^{inc}) is monitored to assess consistency under the assumption of a full design-by-treatment interaction model with random effects.[45]

The assumption of consistency can also be tested by comparison of direct and indirect estimates of the relative treatment effects.[46] We monitor the $z-$*values* of disagreement between direct and indirect evidence for each comparison to derive indirect estimates.[3]

4.4 | Backward search

We briefly describe the backward search method, which is compared with the FS method in the examples. The backward search starts by fitting the complete network and gradually deletes studies until some criterion is met. For instance, it starts by fitting the hypothesized model to all studies, calculates an objective function given by the median of

(Equation 1) or (Equation 2) (eg, median absolute standardized residuals) and the study with the worst value (maximum of the median absolute standardized residual) is deleted. We proceed until some criterion is met (eg, all absolute standardized residuals are less than 2).

5 | ILLUSTRATIVE EXAMPLES

We study the performance of the FS in detecting outlying studies using a simulated dataset as well as two real data examples.

5.1 | Simulated dataset

We simulate a single NMA dataset with $n = 4$ treatments (A, B, C, and D) and $N = m = 8$ two-arm studies (Table A3). Treatment A is chosen as the reference treatment, the true relative effects are set $\mu_{\alpha\beta} = 0.3, \mu_{\alpha C} = 0.4, \mu_{\alpha D} = 0.5$ and the between-study variance is $\tau^2 = 0.1^2$. Following Kontopantelis and Reeves[47] and Brockwell and Gordon,[48] variances of individual studies are generated from $\sigma_i^2 \sim X_1^2/4, i = 1, \ldots 8$ with values restricted to the interval $(0.009, 0.6)$. Results from seven studies are generated from $y_{i,XY} \sim N(\mu_{XY}, \sigma_i^2 + \tau^2), i = 1, \ldots ., 7$ where $XY = (AB, AC, BC, BD, AD, CD, CD)$, and according to the assumption of consistency, that is, $\mu_{XY} = \mu_{AY} - \mu_{AX}$. We then create a study with extreme effect size that compares the treatments C and D, $y_{8,CD} \sim N(\mu_{CD} + 4\text{SD}(y), \sigma_8^2 + \tau^2)$, where SD($y$) is the sample SD of the effect sizes from the first seven studies $y = (y_{1,AB}, \ldots, y_{7,CD})$.[49-51]

The FS is conducted using R function `NMAoutlier` in R package **NMAoutlier.**[30] The median of absolute standardized residuals and the absolute standardized residuals (Equation 1) are used for choosing the initial basic subset and for progressing in the FS, respectively. The initial basic subset was selected among $P = 100$ candidate subsets of size l each, equal to the number of treatments, $l = \max(4, 0.2 \times 8) = 4$ studies. The initial subset consisting of studies 1, 3, 5, and 7, gave the lowest median absolute standardized residual. Table 2 gives the steps of the FS until all studies are included in the basic set. Based on the absolute value of the residuals, the studies entered in the following order: study 6 with an absolute residual of 2.64, study 2, study 4, and finally study 8. Figure A2 (left side) in Appendix provides the forward plot of standardized residuals for each iteration produced with `fwdplot()`. Study 8 has a large, standardized residual compared with the other studies and, thus, was detected as outlying. The backward search was also conducted and study 8 was the only one deleted.

We also added two more studies with extreme effect sizes (studies: 9, 10) which were generated with $y_{9,AB} \sim N(\mu_{AB} + 4\text{SD}(y), \sigma_9^2 + \tau^2)$ and $y_{10,CD} \sim N(\mu_{CD} + 6\text{SD}(y), \sigma_{10}^2 + \tau^2)$ with $\sigma_9^2, \sigma_{10}^2 \sim X_1^2/4$ restricted to the interval $(0.009, 0.6)$. For this simulation scenario (artificial extreme studies 8, 9, and 10 included in the data), FS was conducted using the same criteria with the case only one artificially outlier was included. Study 8 entered at iteration 5, study 10 at iteration 6, and study 9 at the last iteration. Moreover, studies 8, 9, and 10 provide large, standardized residuals compared with the other studies (Figure A2, right side) and, thus, were detected as outliers.

5.2 | Application 1: Interventions to aid smoking cessation

We applied the proposed FS to the network comparing interventions to aid smoking cessation.[17,32] The corresponding R code with the **NMAoutlier**[30] package is provided in Appendix A2 allowing the reproducibility of results. The initial basic subset was selected among $P = 100$ possible subsets of size $l = 5$ each using the absolute residual criterion. The FS steps were completed in 27 seconds*. Table 3 summarizes which studies were part of the initial basic subset (studies: 18, 21, 9, 20, 15) and the progression steps. The FS method was completed in 20 iterations and study 3 entered in the last iteration.

Confidence intervals of summary relative treatment effects between treatments B and C broaden in the last iteration (Figure A3) due to the estimated τ^2, which increased substantially in this iteration (Table 3). The forward plot (Figure 3, right side) shows that the ratio of variances increased rapidly in the last iteration. However, the full interaction model does not provide evidence for inconsistency ($Q^{\text{inc}} = 4.66, p = 0.7$). We monitored a large increase in estimated τ^2, Q^{het}, and Q^{net}, but a reduction in Q^{inc} in the final iteration (Table 3); inconsistency in the whole network is masked due to the large heterogeneity.

TABLE 2 Initial set and study entered into the basic set of FS algorithm, simulated dataset

Studies	y_i (s_i)	Iteration 1	Iteration 2 Residual values of the nonbasic set	Iteration 3 Residual values of the nonbasic set	Iteration 4 Residual values of the nonbasic set	Iteration 5
1	−0.0820 (0.5091)	**study 1 entered**				
2	0.3198 (0.0125)		18.47	**0.27 study 2 entered**		
3	0.2171 (0.2437)	**study 3 entered**				
4	0.2100 (0.0153)		26.67	0.47	**3.86 study 4 entered**	
5	0.4926 (0.1928)	**study 5 entered**				
6	−0.8612 (0.4800)		**2.64 study 6 entered**			
7	0.4115 (0.1007)	**study 7 entered**				
8	2.7639 (0.4604)		5.11	5.24	5.57	**study 8 entered**

Note: The study with the smallest residual (in absolute value) is the next to enter. The smallest residual is denoted with bold letters.
Abbreviation: FS, forward search.

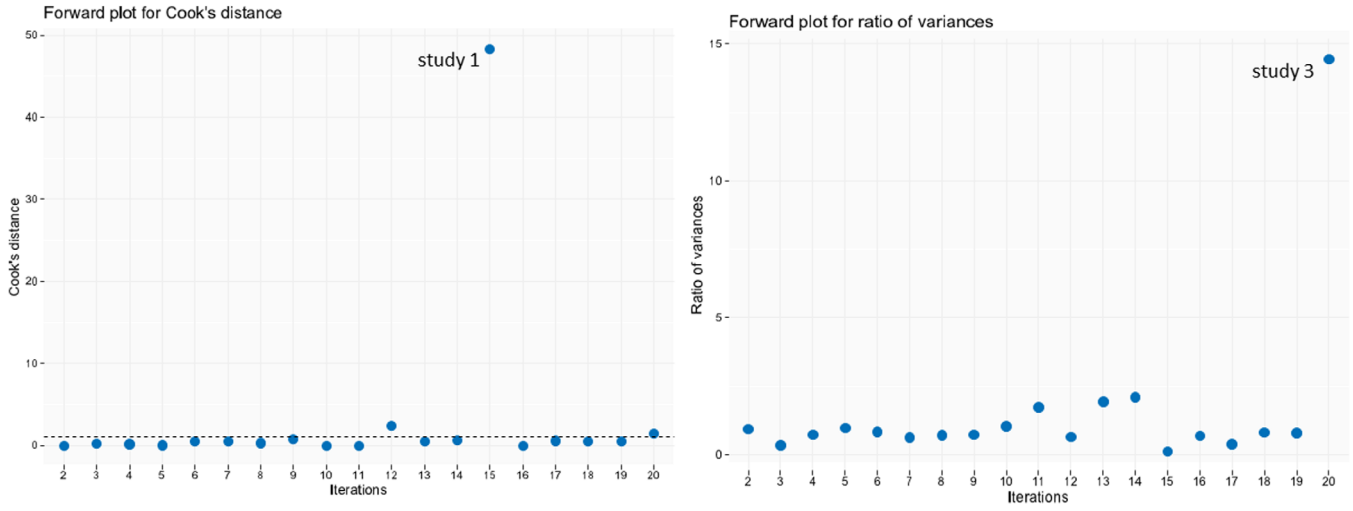

FIGURE 3 Forward plots for Cook's distance (left side) and the ratio of variances (right side) for smoking cessation data[17,32] [Colour figure can be viewed at wileyonlinelibrary.com]

Study 3 entered in the last iteration of the FS and, most importantly, produced sharp changes in the estimated heterogeneity. Furthermore, study 3 has an important impact on the estimated summary odds ratios; its inclusion resulted in $\widehat{\mu}_{AB}$ =1.52 (0.74, 3.09), $\widehat{\mu}_{AC}$ =2.07 (1.34, 3.18), and $\widehat{\mu}_{AD}$ = 2.45(1.09, 5.47) (iteration 20 of the FS algorithm) in comparison to $\widehat{\mu}_{AB}$ = 1.30 (0.84, 2.03), $\widehat{\mu}_{AC}$ = 1.59 (1.20, 2.07),and $\widehat{\mu}_{AD}$ = 1.91 (1.12, 3.28) when study 3 is not included (iteration 19 of the FS algorithm). We observed sharp changes in the monitoring statistics through the FS search for study 3 (Figure 3).

Although the overall Q^{inc} statistic did not suggest any inconsistency in the whole network, we noticed a sharp increase in Q^{inc} when study 1, which compares A, C, and D, enters the basic set at iteration 15 (Table 3). A sharp change in Cook's distance was detected when study 1 entered at iteration 15 (Figure 3, left-hand side). The forward plot of

TABLE 3 Initial set and progression of the FS algorithm for smoking cessation data[17,32]

Iterations	Study entering	Q^{total}	Q^{inc}	Q^{het}	$\hat{\tau}^2$
1	18, 21, 9, 20, 15 (initial basic set)	0.86	0.00	0.86	0.00
2	13	0.87	0.00	0.87	0.00
3	11	1.00	0.12	0.87	0.00
4	16	1.48	0.33	1.14	0.00
5	4	2.48	0.27	2.21	0.00
6	14	4.71	0.11	4.60	0.00
7	12	5.70	0.03	5.67	0.00
8	5	6.42	0.00	6.42	0.00
9	17	8.48	0.00	8.48	0.00
10	6	11.17	0.00	11.16	0.00
11	8	15.03	0.00	15.02	0.02
12	10	18.45	0.42	17.69	0.03
13	19	29.12	0.14	28.67	0.07
14	7	43.96	0.21	43.42	0.13
15	1	53.44	6.84	43.42	0.16
16	24	53.45	6.84	43.42	0.15
17	2	55.39	7.61	43.42	0.14
18	23	58.44	7.92	45.17	0.15
19	22	61.21	9.57	45.17	0.15
20	3	202.62	4.66	187.40	0.59

Note: Q statistics (Q^{total}, Q^{inc}, Q^{het}) and heterogeneity estimator $\hat{\tau}^2$ for each iteration of the FS algorithm.
Abbreviation: FS, forward search.

$z-$values (Figure 4) shows that study 1 is associated with large differences between direct and indirect evidence for "A vs D" and "C vs D" comparisons ($z_{A \text{ vs } D} = 1.50, z_{C \text{ vs } D} = 2.20$, at iteration 15). We conclude that study 1 influences the model substantially as it is responsible for design inconsistency in "A vs D" and "C vs D" effect sizes between the two-arm and three-arm studies. We observed negligible changes in inconsistency measures when the other three-arm study, study 2 with treatment arms B, C, and D, entered (iteration 17). This agrees with the conclusion given by Higgins et al[2] that there is a design inconsistency in effect sizes between two-arm and three-arm studies.

The changes incurred by studies 1 and 3 in the monitoring measures differ substantially from the changes incurred in the FS process by the other studies in the smoking cessation data. We also conducted a backward search method which completed within one iteration by deleting study 3. Study 1 was not identified as an aberrant study by backward methods. This gives a nice example of how the aberrant studies can be identified even if they do not have an extreme effect size or do not enter in the last iterations of the FS. It is common practice in the FS literature to check the robustness of results by repeating the FS search from random starting points (initial subsets). We repeated the FS 100 times from random starting points using $P = 1$ for each run. During monitoring, we noticed that study 3 entered in the last iteration of the FS 82 times, it was included in the initial subset 15 times and entered in an intermediate iteration 3 times. In these three instances, we noticed sharp changes in the monitoring measures when study 3 entered the search. When study 3 is included in the initial subset, we observed peculiar patterns in the monitoring statistics (such as the heterogeneity estimator) in the FS procedure. For example, the estimated heterogeneity for the initial subset was large and was subsequently reduced as the FS progresses (Figure A4, left side). Moreover, Figure A4 right side shows that the standardized residual for study 3 decreased and got far away as other studies entered the search. For completeness, we employed variations of the FS algorithm (different methods for selecting the initial subset, progressing and statistics monitored) but all methods led to the same conclusions. In addition, repeating the FS whilst including study 1 in the initial set did not affect the outlying diagnosis for study 3.

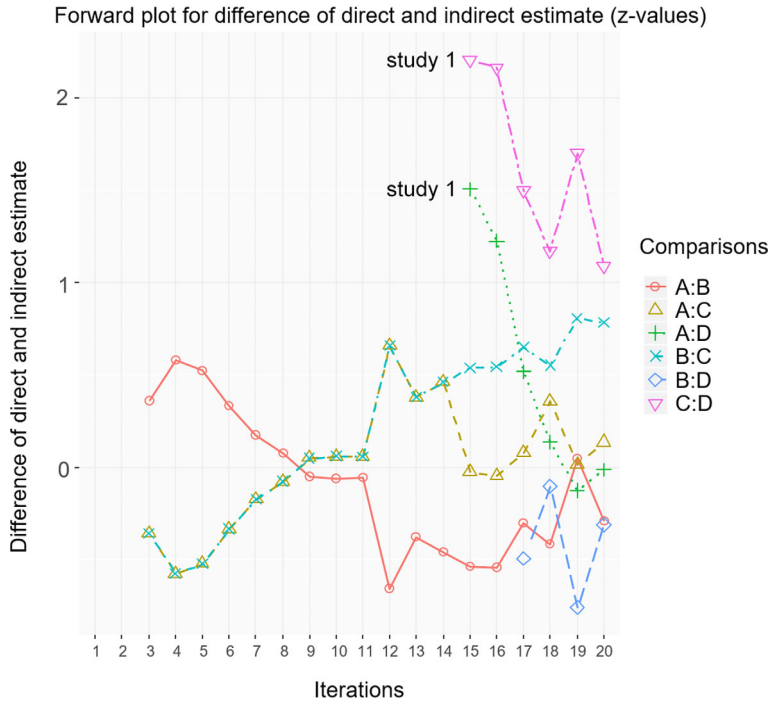

FIGURE 4 Forward plot of z-values that compare relative treatment effects estimated from direct and indirect evidence for smoking cessation data[17,32] [Colour figure can be viewed at wileyonlinelibrary.com]

5.3 | Application 2: Interventions for actinic keratosis

The FS is also applied to the network of 35 studies for actinic keratosis.[34] The design-by-treatment interaction model ($Q^{\text{inc}} = 23.05, df = 7, p = 0.001$) showed statistically significant inconsistency. The between-designs Q^{inc} statistic indicated that the dataset provides evidence of consistency when the design including treatments 1 vs 6 vs 8 (observed only in study 28) was detached ($Q^{\text{inc}} = 10.18, df = 5, p = 0.07$) (Table A4). The initial subset was selected among $P = 100$ subsets of size $l = 9$ each using the smallest absolute residual criterion. The FS was completed after 27 iterations at 59 seconds. A sharp increase in the Q^{inc} statistic (from 3.68 to 23.05) occurred when study 28 entered in the last iteration (Figure A5) indicating that study 28 is a potential source of inconsistency. Sharp changes occurred in the forward plots for Cook's distance and the ratio of variances when study 28 entered the search (Figure A6). After removing study 28 from the dataset, the design-by-treatment interaction model ($Q^{\text{inc}} = 3.68, df = 5, p = 0.59$) indicated no statistically significant inconsistency. The FS and the backward search led to different conclusions this time. The backward search removed studies 24, 23, 22, and 21 in turn until all included trials had absolute standardized residuals less than 2. This is an example of a case where forward and backward methods give different results. Study 28 is mainly responsible for inconsistency and a study with effects different than those estimated indirectly for the respective comparisons does not necessarily have a large residual and cannot, therefore, be detected by backward methods.

5.4 | Application 3: Antihypertensive strategies for heart failure

We applied the FS method to the network of 26 studies comparing antihypertensive strategies for heart failure.[35] Noma et al[20] applied their proposed outlier diagnostics to this dataset and found three studies with aberrant behavior (studies 23, 24, and 26). According to the FS method, study 26 entered at the last iteration (iteration 19), study 24 entered at iteration 16, and study 23 at iteration 17. Sharp changes in the ratio of variances are seen when studies 23 and 26 entered the FS (Figure A1, right side). Therefore, studies 23 and 26 have an impact as they increase the variance and influence the model parameters by giving less precise results.

6 | DISCUSSION

In NMA, there is a lack of visual tools to identify extreme study effects. Overall, the FS provides a practical set of visual diagnostic tools that can help us not only identify outliers but also those studies responsible for differences between direct and indirect estimates. For transparency and reproducibility purposes, we developed the R package **NMAoutlier.**[30] Many decisions are required to apply the algorithm: the size of the initial subset, the number of initial subsets to be examined, criteria for choosing the initial subset, criteria of progressing in the search, and the statistics to be monitored. We ran many FSs for the examples presented in the article, using several combinations of the methods available, and we found results to be robust.

Three common criticisms to the method are: (i) why not employ the more popular backward selection methods, (ii) what happens if an outlier is included in the initial subset or does not enter in the last iterations, and (iii) how do we know a change in the monitored statistics is not due to chance?

Regarding the first criticism, backward methods are known to behave poorly in the presence of multiple outliers that may affect mean values to such a degree that they do not seem to be outliers anymore (masking effect). In the meta-analysis literature, a common strategy is to exclude all studies, one at a time, and see the impact on results (parameter estimates, heterogeneity, inconsistency, and so on). Hence, the computational burden is bigger than the FSs, the method is sensitive to masking, and it is not known whether monitored changes are due to chance or not. In addition, we saw in the actinic keratosis example that the backward method failed to identify the study responsible for the inconsistency in the network. Deletion diagnostics that are not based on residual values could have potentially located the problematic study, but it is overall time-consuming and complicated to apply several deletion strategies. The FS provides a breadth of information regarding the structure of the data and the impact of each study on various aspects of the NMA model. A careful investigation of the search and its repetition from random starting points can help identify atypical studies irrespectively of the stage at which they enter the search.

Regarding the second criticism, we argue that even if outlying studies do not enter towards the last iterations of the search, we may be still able to spot them through sharp changes in the forward plots of estimated heterogeneity and standardized residuals. For example, when an outlier enters the initial subset, it is common to start with large heterogeneity estimates that gradually decrease. Another benchmarking technique we can use concerning this criticism is to run the FS several times from random starting points as a sensitivity analysis. The running time for the FS algorithm depends on how large the network and the initial subset are. We also suggest that, once we consider some studies to be aberrant, the FS should be repeated but this time forcing the aberrant studies to be included in the initial subset. When outliers are included in the initial subset, it is typical to see some "undesirable" statistics at the beginning (eg, large heterogeneity/inconsistency) that then improve as studies are included in the "basic" set.

The most serious concern is the third: how do we know that changes in the forward plots are not due to chance? One method suggested in the literature is the construction of simulation envelopes that give, at each iteration of the FS, the bands within which we expect statistics to lie if there are no outliers in the data. To construct these bands through simulation, we would need to run the FS hundreds of times, increasing the computational time. There is currently a lot of work in constructing these bands without simulation and we aim to equip the FS for NMA with this capability in the future. Johansen et al provided the asymptotic distribution of scaled forward residuals[52] offered in R package **ForwardSearch**.[53] Of course, the same problems apply also to the backward methods.

A challenging issue is to delineate the relationship of outlying studies with heterogeneity and/or inconsistency. Outliers may cause heterogeneity/inconsistency, but they can also be masked by them. If a comparison is not informed by both direct and indirect evidence, then it is judged merely by the magnitude of the reported effect sizes.

Throughout this article, we have focused on detecting outliers at the study level. Studies give aggregate measures, which may have been influenced by the presence of outliers or data extraction errors within the study. The FS algorithm can also be extended to meta-analysis with multiple outcomes, a meta-analysis of diagnostic accuracy studies, or individual participants' data meta-analysis.

6.1 | Recommendations

The proposed method aims to detect outlying and influential studies and should be used cautiously, recognizing that it is a diagnostic method and not to be used for throwing out studies depending on results. We give some guidance on how to interpret results from applying the FS algorithm.

Make a priori decisions for technical aspects of the FS algorithm. Technical aspects of the FS algorithm should be defined in advance. These are: the size of the initial subset, the number of initial subsets P to be examined, criteria for choosing the initial subset, criteria of progressing in the search, and the statistics to be monitored.

Run the FS algorithm and monitor the forward plots. Proceed in applying the FS algorithm with the a priori decisions taken. **NMAoutlier** package has sensible defaults for some of these technical settings (these are: the number of initial subsets to be $P = 100$, the median of the absolute standardized residuals for the criteria for choosing the initial subset and the criteria of progressing in the search). The main results for the FS algorithm are the ordering of the studies that enter the search and several monitoring statistics. Depending on the technical aspects of the search, it is likely that the method can produce different output results. But even if we proceed with the same criteria (those for choosing the initial subset, criteria of progressing in the search), FS can provide different ordering of the studies entering at each iteration and accordingly different results of the monitoring measures in each run. We have observed that in most cases outliers enter at the last iterations. However, there are exceptions and outliers may enter at any iteration (even included in the initial subset). By monitoring various statistics (eg, Cook distance, the ratio of variance, residuals, Q statistics, model parameters), we can look for sharp changes or other patterns that will be indicative of the same potential outliers at any FS run at any step of the search. We can also look at forward plots of residual values. If an outlier enters early in the search or is included in the initial subset, we expect that its residual values will keep increasing as studies that are very different (and not outlying) enter the search. We also expect that heterogeneity will keep decreasing in such a case.

Run the FS algorithm from different random starting points and compare the results. We suggest rerunning the FS 5 to 10 times from random starting points (initial subsets) to explore the robustness of the ordering and to avoid results driven by chance. Even when outliers are included in the initial subset, it is typical to see large changes in statistics at the beginning of the search (eg, large heterogeneity/inconsistency) that then improve as studies are included in the "basic" set.

Run the FS algorithm using different criteria. We suggest rerunning the FS with smaller or larger initial sample sizes or using a different method for progressing in the search to explore the robustness of results.

Have a closer look at potentially outlying studies that the FS algorithm identified and interpret the results. There is not a definite *"decision pathway"* for the interpretation of the results; instead, one should critically review the potentially outlying studies as indicated by the FS algorithm and have a look at the original data and/or the graphical tools (eg, comparison-adjusted funnel plot). One should think of a broad range of possible explanations, from a possible data extraction error (for data or eligibility) to factors that could introduce heterogeneity or inconsistency. We can further use inconsistency diagnostics (eg, node/side-splitting or net heat plot) and explore whether outliers are responsible for inconsistency.

7 | CONCLUSION

In conclusion, we argue that the method can be employed as a diagnostic tool to provide a comprehensive outlier detection analysis as it can offer information about the data and detect extreme study effects responsible for heterogeneity and inconsistency.

ACKNOWLEDGEMENTS

This work formed part of the first author's PhD dissertation at the Department of Primary Education at the University of Ioannina. This research has been financially supported by the General Secretariat for Research and Technology (GSRT) and the Hellenic Foundation for Research and Innovation (HFRI) (Scholarship Code: 82234).

ENDNOTE

*FS algorithm was completed in 27 seconds using Windows 10 operating system running under a 6 core AMD Ryzen 2600 with 8-gigabyte random access memory.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Maria Petropoulou https://orcid.org/0000-0002-7147-3644
Georgia Salanti https://orcid.org/0000-0002-3830-8508

Gerta Rücker https://orcid.org/0000-0002-2192-2560
Guido Schwarzer https://orcid.org/0000-0001-6214-9087
Irini Moustaki https://orcid.org/0000-0001-8371-1251
Dimitris Mavridis https://orcid.org/0000-0003-1041-4592

REFERENCES

1. Salanti G. Indirect and mixed-treatment comparison, network, or multiple-treatments meta-analysis: many names, many benefits, many concerns for the next generation evidence synthesis tool. *Res Synth Methods*. 2012;3(2):80-97.
2. Higgins JPT, Jackson D, Barrett JK, Lu G, Ades AE, White IR. Consistency and inconsistency in network meta-analysis: concepts and models for multi-arm studies. *Res Synth Methods*. 2012;3(2):98-110.
3. König J, Krahn U, Binder H. Visualizing the flow of evidence in network meta-analysis and characterizing mixed treatment comparisons. *Stat Med*. 2013;32(30):5414-5429.
4. Mavridis D, Giannatsi M, Cipriani A, Salanti G. A primer on network meta-analysis with emphasis on mental health. *Evid Based Ment Health*. 2015;18(2):40-46.
5. Salanti G, Higgins JPT, Ades AE, Ioannidis JPA. Evaluation of networks of randomized trials. *Stat Methods Med Res*. 2008;17(3):279-301.
6. Sterne JAC, Egger M, Moher D, eds. Chapter 10. Addressing reporting biases. *Cochrane Handbook for Systematic Reviews of Interventions. Version 5.1.0 (updated March 2011)*. Chichester: John Wiley & Sons; 2011. http://handbook.cochrane.org/chapter_9/9_8_chapter_information.html.
7. Lee KJ, Thompson SG. Flexible parametric models for random-effects distributions. *Stat Med*. 2008;27(3):418-434.
8. Baker R, Jackson D. A new approach to outliers in meta-analysis. *Health Care Manag Sci*. 2008;11(2):121-131.
9. Baker R, Jackson D. New models for describing outliers in meta-analysis. *Res Synth Methods*. 2016;7(3):314-328.
10. Gumedze FN, Jackson D. A random effects variance shift model for detecting and accommodating outliers in meta-analysis. *BMC Med Res Methodol*. 2011;11:19.
11. Beath KJ. A finite mixture method for outlier detection and robustness in meta-analysis. *Res Synth Methods*. 2014;5(4):285-293.
12. Lin L, Chu H, Hodges JS. Alternative measures of between-study heterogeneity in meta-analysis: reducing the impact of outlying studies. *Biometrics*. 2017;73(1):156-166.
13. Viechtbauer W, Cheung MWL. Outlier and influence diagnostics for meta-analysis. *Res Synth Methods*. 2010;1(2):112-125.
14. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *J Stat Softw*. 2010;36(3):1-48.
15. R Development Core Team. *R: A language and Environment for Statistical Computing Version 3.5.1*. Vienna, Austria: R Foundation for Statistical Computing; 2018. http://www.R-project.org.
16. Chaimani A, Higgins JPT, Mavridis D, Spyridonos P, Salanti G. Graphical tools for network meta-analysis in STATA. *PLoS One*. 2013;8:10.
17. Lu G, Ades AE. Assessing evidence inconsistency in mixed treatment comparisons. *J Am Stat Assoc*. 2006;101(474):447-459.
18. Zhang J, Fu H, Carlin BP. Detecting outlying trials in network meta-analysis. *Stat Med*. 2015;34(19):2695-2707.
19. Zhao H, Hodges JS, Carlin BP. Diagnostics for generalized linear hierarchical models in network meta-analysis. *Res Synth Methods*. 2017;8(3):333-342.
20. Noma H, Gosho M, Ishii R, Oba K, Furukawa TA. Outlier detection and influence diagnostics in network meta-analysis. *Res Synth Methods*. 2020;11(6):891-902.
21. Shi L, Zuo S, Yu D, Zhou X. Influence diagnostics in meta-regression model. *Res Synth Methods*. 2017;8(3):343-354.
22. Atkinson AC. Masking unmasked. *Biometrika*. 1986;73(3):533-541.
23. Hadi AS. Identifying multiple outliers in multivariate data. *J Royal Stat Soc Ser B*. 1992;54:761-771.
24. Atkinson AC, Riani M. *Robust Diagnostic Regression Analysis*. New York, NY: Springer Publishing Company; 2000.
25. Atkinson AC. Fast very robust methods for the detection of multiple outliers. *J Am Stat Assoc*. 1994;89(428):1329-1339.
26. Atkinson AC, Riani M, Cerioli A. *Exploring Multivariate Data with the Forward Search*. New York, NY: Springer Publishing Company; 2004.
27. Mavridis D, Moustaki I. Detecting outliers in factor analysis using the forward search algorithm. *Multivar Behav Res*. 2008;43(3):453-475.
28. Mavridis D, Moustaki I. The forward search algorithm for detecting aberrant response patterns in factor analysis for binary data. *J Comput Graph Stat*. 2009;18(4):1016-1034.
29. Mavridis D, Moustaki I, Wall M, Salanti G. Detecting outlying studies in meta-regression models using a forward search algorithm. *Res Synth Methods*. 2016;8(2):199-211.
30. Petropoulou M, Schwarzer G, Panos A, Mavridis D. NMAoutlier: detecting outliers in network meta-analysis, R package version 0.1.17; 2021. https://cran.r-project.org/packages/NMAoutlier/NMAoutlier.
31. Rücker G. Network meta-analysis, electrical networks and graph theory. *Res Synth Methods*. 2012;3(4):312-324.
32. Hasselblad V. Meta-analysis of multi-treatment studies. *Med Decis Making Int J Soc Med Decis Making*. 1998;18(1):37-43.
33. Rücker G, Krahn U, König J, Efthimiou O, Schwarzer G. netmeta: network meta-analysis using frequentist methods, R package version 1.3–0; 2021. https://CRAN.R-project.org/package=netmeta.
34. Gupta AK, Paquet M. Network meta-analysis of the outcome 'participant complete clearance' in nonimmunosuppressed participants of eight interventions for actinic keratosis: a follow-up on a Cochrane review. *British J Dermatol*. 2013;169(2):250-259.
35. Sciarretta S, Palano F, Tocci G, Baldini R, Volpe M. Antihypertensive treatment and development of heart failure in hypertension: a Bayesian network meta-analysis of studies in patients with hypertension and high cardiovascular risk. *Arch Int Med*. 2011;171(5):384-394.

36. Rücker G, Schwarzer G. Reduce dimension or reduce weights? Comparing two approaches to multi-arm studies in network meta-analysis. *Stat Med*. 2014;33(25):4353-4369.
37. Schwarzer G, Carpenter JR, Rücker G. *Meta-Analysis with R (Use-R!)*. Cham, Switzerland: Springer International Publishing; 2015.
38. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials*. 1986;7(3):177-188.
39. Jackson D, White IR, Riley RD. Quantifying the impact of between-study heterogeneity in multivariate meta-analyses. *Stat Med*. 2012;31(29):3805-3820.
40. Rousseeuw PJ. Least median of squares regression. *J Am Stat Assoc*. 1984;79(388):871-880.
41. Chatterjee S, Hadi AS. *Regression Analysis by Example*. 5th ed. Hoboken, NJ: Wiley; 2013.
42. Cook RD, Weisberg S. *Residuals and Influence in Regression*. New York, NY: Chapman & Hall; 1982.
43. Krahn U, Binder H, König J. A graphical tool for locating inconsistency in network meta-analyses. *BMC Med Res Methodol*. 2013;13:35.
44. White IR, Barrett JK, Jackson D, Higgins JPT. Consistency and inconsistency in network meta-analysis: model estimation using multivariate meta-regression. *Res Synth Methods*. 2012;3(2):111-125.
45. Jackson D, Barrett JK, Rice S, White IR, Higgins JPT. A design-by-treatment interaction model for network meta-analysis with random inconsistency effects. *Stat Med*. 2014;33(21):3639-3654.
46. Dias S, Welton NJ, Caldwell DM, Ades AE. Checking consistency in mixed treatment comparison meta-analysis. *Stat Med*. 2010;29(7–8):932-944.
47. Kontopantelis E, Reeves D. Performance of statistical methods for meta-analysis when true study effects are non-normally distributed: a simulation study. *Stat Methods Med Res*. 2012;21(4):409-426.
48. Brockwell SE, Gordon IR. A comparison of statistical methods for meta-analysis. *Stat Med*. 2001;20(6):825-840.
49. Filzmoser P. Identification of multivariate outliers: a performance study. *Austr J Stat*. 2005;34(2):127-138.
50. Knight NL, Wang J. A comparison of outlier detection procedures and robust estimation methods in GPS positioning. *J Navigat*. 2009;62(4):699-709.
51. Hardin J, Rocke DM. Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator. *Comput Stat Data Anal*. 2004;44(4):625-638.
52. Johansen S, Nielsen B. Asymptotic analysis of the forward search. *Soc Sci Res Netw Electron J*. 2013. https://ideas.repec.org/p/nuf/econwp/1302.html.
53. Nielsen B. ForwardSearch: forward search using asymptotic theory, R package version 1.0; 2014. https://CRAN.R-project.org/package=ForwardSearch.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Petropoulou M, Salanti G, Rücker G, Schwarzer G, Moustaki I, Mavridis D. A forward search algorithm for detecting extreme study effects in network meta-analysis. *Statistics in Medicine*. 2021;1–15. https://doi.org/10.1002/sim.9145