UNIVERSITY OF IOANNINA

SCHOOL OF EDUCATION

DEPARTMENT OF PRIMARY EDUCATION

Exploring methodological challenges in network meta-analysis models and developing methodology for outlier detection

Maria Petropoulou

DOCTORAL THESIS

IOANNINA 2020, GREECE

ΜΑΡΙΑ ΠΕΤΡΟΠΟΥΛΟΥ

«Διερευνώντας μεθοδολογικές πτυχές των μοντέλων μετα-ανάλυσης δικτύων και ανάπτυξη μεθοδολογίας για τον εντοπισμό ακραίων μελετών» (ημερομηνία ορισμού θέματος 12-09-2018)

Διδακτορική Διατριβή

Υποβληθείσα στο Παιδαγωγικό Τμήμα Δημοτικής Εκπαίδευσης της Σχολής Επιστημών Αγωγής, του Πανεπιστημίου Ιωαννίνων, για την εκπλήρωση των προϋποθέσεων απονομής Διδακτορικού Διπλώματος

Τριμελής Συμβουλευτική Επιτροπή (ημερομηνία ορισμού 25-01-2017):

1. **Δημήτριος Μαυρίδης**, Επίκουρος Καθηγητής Στατιστικής στην Εκπαίδευση του Παιδαγωγικού Τμήματος Δημοτικής Εκπαίδευσης του Πανεπιστημίου Ιωαννίνων (Επιβλέπων).
2. **Γεωργία Σαλαντή**, Αναπληρώτρια Καθηγήτρια Βιοστατιστικής και Επιδημιολογίας στο Ινστιτούτο Κοινωνικής και Προληπτικής Ιατρικής του Πανεπιστημίου Βέρνης.
3. **Ειρήνη Μουστάκη**, Καθηγήτρια Στατιστικής του Τμήματος Στατιστικής στο London School of Economics του Λονδίνου.

Επταμελής Εξεταστική Επιτροπή (ημερομηνία ορισμού 04-03-2020):

1. **Δημήτριος Μαυρίδης**, Επίκουρος Καθηγητής Στατιστικής στην Εκπαίδευση του Παιδαγωγικού Τμήματος Δημοτικής Εκπαίδευσης του Πανεπιστημίου Ιωαννίνων (Επιβλέπων).
2. **Γεωργία Σαλαντή**, Αναπληρώτρια Καθηγήτρια Βιοστατιστικής και Επιδημιολογίας στο Ινστιτούτο Κοινωνικής και Προληπτικής Ιατρικής του Πανεπιστημίου Βέρνης (Μέλος της Τριμελούς Συμβουλευτικής Επιτροπής).
3. **Ειρήνη Μουστάκη**, Καθηγήτρια Στατιστικής του Τμήματος Στατιστικής στο London School of Economics του Λονδίνου (Μέλος της Τριμελούς Συμβουλευτικής Επιτροπής).
4. **Αναστάσιος Μικρόπουλος**, Καθηγητής Πληροφορικής στην Εκπαίδευση με έμφαση τις Εικονικές Πραγματικότητες στη Διδασκαλία του Παιδαγωγικού Τμήματος Δημοτικής Εκπαίδευσης του Πανεπιστημίου Ιωαννίνων.
5. **Ευάγγελος Ευαγγέλου**, Αναπληρωτής Καθηγητής Υγιεινής με έμφαση στην Κλινική και Μοριακή Βιολογία του Τμήματος Ιατρικής του Πανεπιστημίου Ιωαννίνων.
6. **Κωνσταντίνος Τσιλίδης**, Αναπληρωτής Καθηγητής Επιδημιολογίας του Τμήματος Ιατρικής του Πανεπιστημίου Ιωαννίνων.
7. **Κωνσταντίνος Γαβριλάκης**, Επίκουρος Καθηγητής Περιβαλλοντικής Εκπαίδευσης του Παιδαγωγικού Τμήματος Δημοτικής Εκπαίδευσης του Πανεπιστημίου Ιωαννίνων.

Έγκριση Διδακτορικής Διατριβής με βαθμό **«ΆΡΙΣΤΑ»** στις **19-03-2020**

Acknowledgments

The work presented in this dissertation was funded by General Secretariat for Research and Technology (GSRT) and the Hellenic Foundation for Research and Innovation (HFRI) (Scholarship Code: 82234).

I particularly want to thank my supervisor Dr. Dimitris Mavridis for his best effort to convey to me his professional expertise and knowledge and for the continuous support and guidance he provided to me. My sincere thanks also goes to Georgia Salanti, Irini Moustaki, Gerta Rücker and Guido Schwarzer for their scientific advice and their valuable input on some of the projects presented in this dissertation.

I would never have been able to finish my dissertation without the help and support from my parents, my brother, and my fiancé Agapios.

Contents

1 Introduction 1
1.1 Systematic review and evidence synthesis methods 1
1.2 Meta-analysis and Education 4
1.3 Objectives & outline of the Thesis 7
2 Network meta-analysis model and motivating examples 9
2.1 Introduction 9
2.2 Network meta-analysis model from graph theory 10
2.3 Motivating examples 14
2.3.1 Dataset comprises four interventions to aid smoking cessation 14
2.3.2 Dataset comprises nine interventions for actinic keratosis 14
2.3.3 Dataset with thrombolytic drugs 15
3 Characteristics of published networks of interventions 17
3.1 Introduction 17
3.2 Methods 18
3.2.1 Inclusion and exclusion criteria 18
3.2.2 Literature search and screening 18
3.2.3 Extraction of data characteristics and categorization 18
3.2.4 Statistical analysis 19
3.3 Results 19
3.3.1 General NMA and publication characteristics 20
3.3.2 Evaluation of transitivity and consistency assumption 21
3.3.3 Statistical synthesis of the data 22
3.4 Discussion 25
4 Methods to detect outliers in meta-analysis 27
4.1 Introduction 27
4.2 Outlying detection strategies in meta-analysis 28
A. Alternative distributions for the random-effects model 29
B. Robust heterogeneity measures 29
C. Likelihood methods 30
D. Backward/Deletion methods 30
E. Forward methods 30
F. Robust estimation 31

4.3 Outlier and influential case diagnostics measures for NMA....31
4.3.1 Outlier detection measures.... 31
4.3.2 Outlier detection measures considered deletion 35
4.4 Illustrative examples with outlier and influential diagnostics measures....40
4.4.1 Dataset comprises four interventions to aid smoking cessation 40
4.4.2 Dataset with thrombolytic drugs.... 44
4.5 Discussion 45
5 Forward Search Algorithm to detect outliers in network meta-analysis 47
5.1 Introduction 47
5.2 Methodological extension of the forward search algorithm in NMA 47
5.2.1 Choice of the initial subset.... 48
5.2.2 Processing in the search 49
5.2.3 Monitor the search 51
5.3 Illustrative examples 52
5.3.1 Artificial simulated outlier.... 52
5.3.2 Dataset comprises four interventions to aid smoking cessation 53
5.3.3 Dataset comparing interventions for actinic keratosis 57
5.4 Discussion 59
6 Random shift variance NMA model for outlier identification..... 63
6.1 Introduction 63
6.2 Shift variance NMA model to downweigh outliers 64
6.2.1 Shift variance NMA model (RVSOM NMA).... 64
6.2.2 Monitoring measures.... 66
6.2.3 Likelihood Ratio Test (LRT) 66
6.2.4 Extended RVSOM NMA.... 67
6.3 Illustrative example of interventions to aid smoking cessation 67
6.4 Discussion 70
7 Using the R package NMAoutlier..... 73
7.1 Introduction 73
7.2 Software description....74
7.3 Application of NMAoutlier in practice with smoking cessation data....78
7.3.1 Part 1: Simply outlier detection measures 79
7.3.2 Part 2: Outlier detection measures considered deletion.... 80
7.3.3 Part 3: Forward Search Algorithm - Detection Method.... 81

7.3.4 Part 4: Shift Variance Network Meta-analysis – Detection method and sensitivity analysis downweighing outlier 83
7.4 Discussion 85
8 Summary 87
8.1 Summary 87
8.2 Περίληψη 88
Appendix 91
Appendix Tables. 91
Appendix A. 94
Bibliography 95

1 Introduction

1.1 Systematic review and evidence synthesis methods

Systematic reviews and meta-analyses have been established as an integral part of comparative effectiveness research and are used worldwide for decision-making in health and social care. The World Health Organization (WHO, www.who.int) considers them as the most reliable Evidence-Based Medicine method. Different types of studies such as cohort studies, animal studies, observational studies, case-control studies or randomized control trials can be synthesized in a systematic review to answer a research question. There is a broad agreement that the randomized controlled trial (RCT) is the most valid (gold standard) type of clinical trials. In RCT, individuals are randomly assigned to two groups where one group (experimental group) takes the intervention and the other (control group) usually receives a placebo intervention. Observed differences between the two groups depend exclusively on interventions received because the participants have been randomized in both groups and probability theory assures that they will not differ in any other characteristics beyond the intervention they accept.

There are a plethora of individual studies in most problems in health, social or education. Drawing conclusions from these studies may be misleading. Individual studies may be biased (low-quality studies) or having conflicting results without being aware of whether these differences are true or random. Systematic reviews synthesize several different studies and provide conclusions to answer a research question. Meta-analysis is (not necessarily) a part of the systematic review process.

Meta-analysis is a statistical technique that synthesizes evidence form individual studies and provides the relative effectiveness between two interventions for a specific research question [1]. It might happen that the results from individual studies disagree. Meta-analysis can quantify and investigate the reasons of this disagreement [1]. Meta-analytical results are more powerful and provide more precise results compared to the results from individual studies [1]. A *pairwise meta-analysis* pools the results from individual studies that compare two interventions. Decision making commonly focuses on comparing more than two interventions.

More complicated evidence synthesis methods are used to investigate the relative effectiveness of three or more interventions. *Network meta-analysis* (NMA), also known as a *mixed-treatment comparison* or *multiple-treatment meta-analysis*, can provide the relative effectiveness of several competing interventions for the outcome of interest. It synthesizes *direct* (information of intervention effect from the comparison of two studies) and *indirect* (information of intervention effect via a connected path) evidence with the aim to give a summary estimate [2].

Let us consider three school-based interventions A, B, C and we are interested to compare their relative effectiveness. Figure 1.1 (a) provides the direct evidence (solid lines) in a public-school comparing A and B educational interventions and the direct evidence of A versus C in a private school. It is obvious that there is no direct evidence between B and C educational interventions. We can only have the indirect evidence for B versus C (dashed line) via the direct paths through intervention A. The graphical representation in Figure 1.1 (a) is termed as *network plot* with cycles of nodes denoting interventions and lines or edges denoting the interventions compared in the included studies.

The network plot can provide information about the shape of network and *network geometry*. The network should be connected, which means that there is a path with lines in network plot (studies) to move from each intervention to any other intervention. If all intervention nodes are compared with a common intervention node the network is a *star-shaped network*. For example, Figure 1.1 (a) a star network that apart from the educational intervention comparisons A versus B, A versus C. The paths that begin from an intervention node and end to the same node via two or more intermediate interventions (e.g. there is a direct evidence for comparison B versus C with path A → B → C → A) are *closed loops*. Networks comparing only three treatments (for example, A, B, and C) in a number of two-arm studies called *triangular networks*. Networks with at least one closed loop called *full* or *entire networks*.

NMA synthesizes direct and indirect evidence and offers the ability to estimate the relative effectiveness of interventions (termed as *network estimates*) that have never compared before by the individual studies (e.g. relative effectiveness of B versus C educational interventions). The relative effectiveness of several interventions is provided in comparison with a common intervention named *reference* treatment that is usually a placebo, usual care, no treatment or active treatment. Several frequentist (e.g. multivariate meta-regression [3], [4], graph theoretical method [5], etc.) and Bayesian approaches (e.g. hierarchical models) have been

developed to derive the indirect and/or network estimates. NMA summarizing the results by providing a hierarchy of the interventions and *treatment ranking* [6], [7] can inform decision making. The most commonly used ranking measures are the probability of being the best and the surface under the cumulative ranking curve (SUCRA) [6], [8], [9].

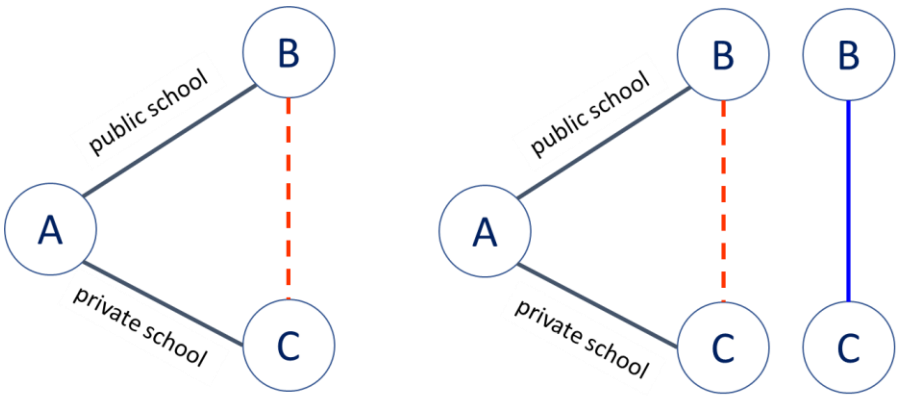

Figure 1.1 (a) Figure 1.1 (b)

Figure 1.1. Direct and indirect evidence between school-based interventions A, B and C (Figure 1.1 (a). Illustration of direct and indirect evidence for comparisons B versus C (Figure 1.1 (b)).

Methodological or clinical differences between studies may cause differences between the study-specific *true underlying effects*. This is a between-study variation, known as *heterogeneity*. The existence of heterogeneity may affect the summary estimate and its precision. Potential sources of heterogeneity can be investigated with *meta-regression*; a meta-analytical model including covariates [10], [11] but also with *sensitivity analyses* or *secondary analyses* such as *subgroup analysis*. that the authors performed to investigate potential sources of heterogeneity or inconsistency.

Individual studies usually differ due to their characteristics. For example, AB studies compare children from public schools while AC studies compare children from private schools. AB and AC studies may differ due to population characteristics. Private schools may have children from a high-income family with more educated and wealthy parents, highly-motivated children for knowledge and may offer a longer and demanding schedule, compared to public schools. Type of school can be an *effect modifier* which means that we cannot compare AB studies (children population in public schools) with AC studies (children population in private schools).

One clinically underlying fundamental assumption for the NMA model is the *transitivity assumption*; this implies that the effect modifiers are comparable across intervention

comparisons [2]. Transitivity assumes that the common comparator A is the same for AB and AC studies and the distribution of effect modifiers is balanced across the different comparisons.

The statistical manifestation of transitivity is *consistency*, that is when direct and indirect evidence agrees [2], [12]. Considering that there is a BC study in the network, we have a direct evidence between school-based interventions for comparison B versus C. The consistency assumption implies that the direct evidence of B versus C (solid line, Figure 1.1 (b)) is in agreement with the indirect evidence of B versus C comparison (dashed line, Figure 1.1 (b)). This implies that there are no differences between direct and indirect estimates in closed loops within networks. Several statistical methods have been developed to evaluate the consistency assumption in closed loops within networks such as the loop-specific approach [13], node-splitting approach [14], design-by-treatment interaction model [4] (a synopsis of several statistical tests for consistency are described in [15] and details are given).

It is common that a potential source of heterogeneity and inconsistency is the existence of extreme study effects. Extreme study effects may also be an *outlier* study. There are several definitions for outliers in the literature. A study with a markedly different intervention effect estimate or a study that does not explain by the assumed model is generally defined as outlying [16]. A study effect that lies far away from the bulk of the data can affect the summary effect, possible causing bias (especially if the study is large) and can lead to an increase in heterogeneity or inconsistency. An *influential* study can influence the model parameters, it might cause large heterogeneity and inconsistency and therefore give biased results. Moreover, small studies tend to give larger estimates than estimates from large studies (*small-study effects*). A frequent phenomenon in evidence synthesis is *publication bias* caused by the fact that small study effects without significant intervention effects are less likely to be published.

1.2 Meta-analysis and Education

Implementation of systematic reviews and meta-analyses is increased rapidly and there is a lot of evidence for several educational outcomes. The Campbell Collaboration (https://campbellcollaboration.org/) is an international network that published systematic reviews and meta-analyses and has a group called Education Coordinating Group (https://www.campbellcollaboration.org/contact/coordinating-groups/education.html) that focuses on education. Based on high-quality evidence synthesis methods, Campbell Collaboration provides a database of systematic reviews on educational outcomes that can inform policy-makers and stakeholders.

There are several published meta-analyses for autism [17], [18], bullying and cyberbullying [19], [20], educational technology outcomes [21] for teaching and learning in different subject matters (math, reading, writing, etc.) across a wide range of age children groups but also mental health-related outcomes in students such as anxiety and depression [22]–[25].

The first meta-analysis that evaluates computer-based scaffolding in science, technology, engineering and mathematics (STEM) education has recently been published on *Review of Educational Research* [21]. Three hundred thirty-three outcomes across 144 studies were included in the meta-analysis [21]. Computer-based scaffolding defined as the students' ability to generate and solve complex problems and goals [21]. According to Belland *et al.* there was a positive effect of computer-based scaffolding interventions on cognitive outcomes significant difference in STEM education [21]. If we are not interested only in the effectiveness of computer-based scaffolding interventions but we are interested in the comparison between different STEM disciplines (such as the comparison between science, technology, engineering or mathematics), multiple testing is needed and traditional meta-analyses can be conducted. Multiple testing with pairwise meta-analyses provides limitations such as increasing type I error rates. Network meta-analysis model can give more precise results for comparative treatment effectiveness and safety. A network meta-analysis that has recently been published in the *Review of Educational Research*, compares the influence of contexts of scaffolding used on cognitive outcomes in STEM education [26]. Effect sizes were reported with Hedge's g calculation [26]. Four different STEM disciplines, mathematics, technology, engineering and science, and control were compared [26]. Figure 1.2 provides the network plot of direct comparisons (solid lines) and indirect comparisons (dashed lines) comparing scaffolding used in the context of different STEM disciplines. Authors noticed an additional 70% of studies included in network meta-analysis than traditional pairwise meta-analyses [26]. Although the authors presented this as an NMA, it is actually a moderator analysis. Authors also ranked the disciplines in which STEM seems to be more effective by averaging the probability of being the best, the second e.t.c for mathematics (ranking 1.62), technology (ranking 2.23), engineering (ranking 3.23) and science (ranking 3.33) [26].

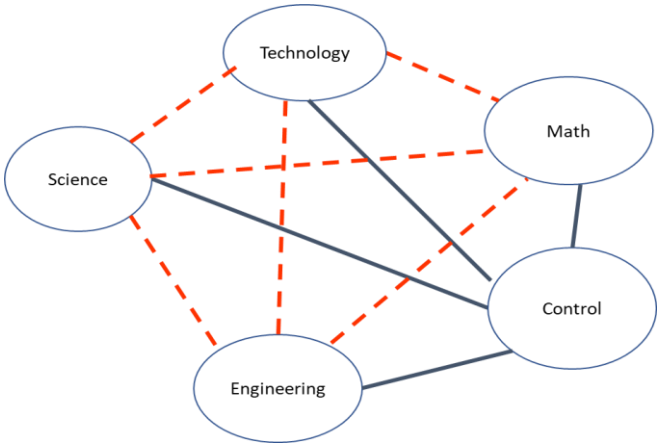

Figure 1.2. Network plot of direct comparisons (solid lines) and indirect comparisons (dashed lines) comparing scaffolding for the context of different STEM disciplines.

Caldwell *et al.* [27] have recently published in *Lancet Psychiatry* the first network meta-analysis to prevent anxiety and depression in children and young people. Several school-based interventions such as Behavioural therapy, Cognitive Behavioural Therapy, Third wave, Psychoeducation, Psychosupport, etc. were compared for the outcomes of anxiety and depression [27]. NMA model fitted using hierarchical models in the Bayesian framework. The results provided little evidence to suggest the effectiveness of school-based interventions for the prevention of anxiety or depression. The most included studies had an unclear risk of bias for random sequence generation and allocation concealment and there was evidence of small-study effects for self-report anxiety outcomes [27]. This phenomenon is usually identified in studies with mental health-related outcomes.

In bibliography, there are studies comparing educational interventions for several educational outcomes. Moreover, there are also many systematic reviews and meta-analyses in educational research. Based on this fact, systematic reviewers and meta-analysts can provide the evidence synthesis of studies in the field of education. Meta-analytical models can be a guidance for teachers providing which interventions could be the best in each case. The implementation of meta-analytical models can lead teachers in several aspects of educational system. For example meta-analytical models can answear the question ‘Which educational method is most effective for STEM education?’, but also they can inform special education teachers about the most effective educational process that would be helpful for students with dyslexia, communication disorders, physical disabilities e.t.c.

The currently published paper by Caldwell *et al.* [27] implements at first network meta-analysis in Education. This can be the evidence for a new step in a synthesis of studies expecting more

network meta-analyses to be conduced in the field of education. The proposed methodology on developing outier detection for network meta-analysis would be helpful in Educational research as outliers seems to be a usually problem when synthesize the results of studies comparing educational interventions.

1.3 Objectives & outline of the Thesis

The aim of this dissertation is to provide and extend several outliers and influential detection methodologies from pairwise meta-analysis to network meta-analysis model. Several statistical outliers and influential detection measures, the forward search algorithm, and the random variance shift outlier model are extended to NMA. All the proposed methodologies are focused on detecting outlying and influential measures 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 proposed outlier detection methodologies are well illustrated using motivated datasets of networks of interventions and simulation data. An R package **NMAoutlier** [28] was developed for reproducibility of the proposed outlier detection methods.

The Thesis is structured as follows. Chapter 2 provides the network meta-analysis model as was introduced by Rücker [5] using graph theory and motivating examples of networks of interventions. Chapter 3 provides an empirical study based on a collection of 456 published network meta-analyses by giving an empirical overview of NMA characteristics.. Chapter 4 outlines a synopsis of methodological strategies to detect outliers in a meta-analysis, gives an overview of several proposed statistical measures to detect outlier and influential cases in NMA with an application in a real dataset. Chapter 5 introduces the extended methodology of the forward search algorithm for identifying outliers and influential studies in NMA and provides applications of the proposed methodology in real and simulated datasets. Chapter 6 provides the extended methodology for outlier identification with the random shift variance NMA model and gives an application example on a published dataset. Chapter 7 describes and gives details of the R package **NMAoutlier** developed for the proposed outlier and influential detection methods and provides how to implement the package in real datasets of networks of interventions.

2 Network meta-analysis model and motivating examples

2.1 Introduction

Bucher *et al.* [13] were first to introduce the idea of indirect and mixed treatment comparisons. Having three treatments A, B, C the indirect summary relative effect of AB (i.e. A versus B) can be estimated indirectly by subtracting the direct relative effects of AC and BC [13] as $\hat{\mu}_{AB(indirect)} = \hat{\mu}_{AC(direct)} - \hat{\mu}_{BC(direct)}$ (where $\hat{\mu}$ is the estimate of relative treatment effects). Indirect estimates of relative treatment effects are also known as an *adjusted indirect comparison*. The variance of the indirect estimate is the sum of the variances of the two direct ones. The mixed estimate can be derived as a weighted average of direct and indirect treatment effects [13]. The Bucher method [13] is also known as an adjusted indirect comparison meta-analysis. This approach ignores correlations when multi-arm studies exist.

Extending the idea to larger networks, several indirect estimates of different comparing interventions can be derived from network estimates. Popular established NMA methods implemented in comparative effectiveness research adopt meta-regression, hierarchical modeling or a multivariate meta-analysis approach. The meta-regression approach was first proposed by Lumley [29] treating each treatment comparison as a covariate in a meta-regression model. Lu *et al.* [30] proposed a different approach based on a two-stage meta-regression. At the first stage, a meta-analysis is performed in each group of studies comparing the same treatments, (e.g. all two-arm studies comparing A versus B) offering the direct estimates on treatment comparisons. At the second stage, a weighted linear regression is performed with the direct estimates as dependent variables. In the Bayesian framework, the NMA model can be fitted as a hierarchical model with a multivariate normal likelihood assumed on the observed relative effects for each study [31], [32]. White *et al.* introduced the NMA model as a specific case of multivariate meta-regression (or multivariate meta-analysis) [3]. The multivariate meta-regression model based on multivariate distributions assumptions for the parameters representing random errors and random effects when multi-arm studies are included [3]. Rücker [5] introduced a frequentist network meta-analysis model with a graph-

theoretical approach by providing the correspondence between electrical networks and multivariate meta-regression models.

This Chapter is structured as follows: Section 2.2 provides the NMA model from graph theory and Section 2.3 offers examples of networks of interventions that motivating us to proposed outlier diagnostics methods provided in this dissertation.

2.2 Network meta-analysis model from graph theory

This section provides a brief description of the NMA approach and the reader can find more details in the relevant publications [5], [33], [34]. This NMA model is implemented in R package **netmeta** [35]. The notation for the NMA model is summarized in Table 2.1.

Suppose that we have N potentially multi-arm studies $i = 1, \dots, N$. For a study i, we denote with k the pairwise comparison, with S_i the set of treatments compared in study i with n_{S_i} to represent the cardinality of S_i and $k \in S_i$. Let m be the number of all possible pairwise comparisons (and hence $m = \sum_{i=1}^{N} \binom{S_i}{2}$ and $m = N$ if $S_i = 2, \forall i = 1, \dots, N$). We denote with n the total number of treatments and $\boldsymbol{\mu}$ to represent the vector with these n treatment effects. Let $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_N)'$ be the vector with the observed effect sizes and $\boldsymbol{s} = (\boldsymbol{s}_1, \boldsymbol{s}_2, \dots, \boldsymbol{s}_N)'$ be the vector with the corresponding standard errors. For a study $i = 1, \dots, N$, let $\boldsymbol{y_i} = \{y_{i,k}, k \in S_i\}$ be the observed effect size, $\boldsymbol{s_i} = \{s_{i,k}, k \in S_i\}$ be the observed standard error and $\boldsymbol{s_i^2} = \{s_{i,k}^2, k \in S_i\}$ the observed sampling variance (or else the *within-study variance*).

Having the variability of studies to be the within-study variance for each pairwise comparison in each study the *fixed-effect* (FE) network meta-analysis can be modeled. The fixed-effect network meta-analysis model is written as

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

where $\mathbf{S}$ is a diagonal form matrix with entries $\boldsymbol{s_i^2}$ and $\mathbf{X}$ is the $m \times n$ design matrix that describes the structure of the network with rows denoting the pairwise comparisons and with columns the treatments compared. Each row inputs one (1) in the column that corresponds to the first treatment (treatment group) and minus one (-1) in the column that belongs to the second treatment (control group). All other entries are equal to zero (0) for treatments not considered in the relevant comparison. Hence, each row of $\mathbf{X}$ sums up to zero. $\mathbf{X}$ matrix is not a full rank and it is not invertible.

Assuming a common heterogeneity variance τ^2 for each pairwise comparison, the *random-effects* (RE) network meta-analysis model can be modeled having the variability to be the within-study variance plus the between-study variance (heterogeneity). The random-effects network meta-analysis model is written as

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

where $\boldsymbol{\delta}$ represents the true random effects and $\boldsymbol{\Delta}$ is a block diagonal matrix with the heterogeneity variance τ^2. The between-study variance is estimated using a special case of the generalized DerSimonian–Laird estimator [36] given in [37]. In the random-effects model, the variance of each study is the heterogeneity estimator $\hat{\tau}^2$ plus the observed study variances $\boldsymbol{s_i^2}$.

Let $\mathbf{W}$ be a $m \times m$ diagonal *weight matrix* with a vector of weights in its diagonal to be the inverse study variance, $w_{i,k,FE} = 1/s_{i,k}^2, i = 1, \ldots, N, k \, \epsilon \, S_i$, for the fixed-effect model. Weight matrix can also be given by $\mathbf{W} = \mathbf{S^{-1}}$. Weight matrix for random-effects is diagonal with a vector of weights $w_{i,k,RE} = 1/(s_{i,k}^2 + \hat{\tau}^2), i = 1, \ldots, N, k \, \epsilon \, S_i$ or can be provided with $\mathbf{W} = (\mathbf{S} + \boldsymbol{\Delta})^{-1}$. Then, the variance-covariance matrix for the observed data under the random-effects model is $Cov(\boldsymbol{y}) = \mathbf{W}^{-1} = \mathbf{S} + \boldsymbol{\Delta}$.

In mathematical field of graph theory, Laplacian matrix, sometimes called admittance matrix or Kirchhoff matrix, is a matrix representation of a graph. Laplacian $n \times n$ matrix is given by $\mathbf{L} = \mathbf{X}'\mathbf{W}\mathbf{X}$, has $n - 1$ rank and it is not invertible [5], [33]. To estimate treatment effects, the Moore Penrose pseudoinverse $n \times n$ matrix $\mathbf{L}^+$ of the Laplacian matrix $\mathbf{L}$ is constructed as provided in [5], [33]. The Moore Penrose pseudoinverse is given by $\mathbf{L}^+ = (\mathbf{L} - \mathbf{J}/n)^{-1} + \mathbf{J}/n$ where $\mathbf{J}$ is $n \times n$ matrix with all elements equal to 1. In case of multi-arm studies, weights are adjusted and are reduced as introduced by Rücker and Schwarzer [33]. When multi-arm studies exist, the heterogeneity estimator $\hat{\tau}^2$ is added to the observed variance before reducing the weights.

Pairwise comparisons of the multi-arm study are correlated, so their variances need to be adjusted by a back-calculation method of the observed variances. Having a multi-arm study with S_i arms, variances can be artificially inflated by $\mathbf{L}^+ = -\frac{1}{2{S_i}^2}\mathbf{X}'\mathbf{X}\mathbf{V}'\mathbf{X}$, where $\mathbf{V}$ is $S_i \times S_i$ symmetric matrix with the observed variances of all comparisons [33].

Network estimates are weighted sums of the observed estimates with weights to come from the rows of $\mathbf{H}$; $\hat{\boldsymbol{y}}^{nma} = \mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^+\mathbf{X}'\mathbf{W}\boldsymbol{y} = \mathbf{H}\boldsymbol{y}$, where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^+\mathbf{X}'\mathbf{W}$ is the $m \times m$ hat

matrix. The variance-covariance matrix of network estimates is given by $Var(\hat{\boldsymbol{y}}^{nma}) = \mathbf{X}\mathbf{L}^{+}\mathbf{X}' = \mathbf{G}_{m\times m}$. Treatment effects can be estimated ($\hat{\boldsymbol{\mu}}$) using direct evidence and each piece of indirect by defining the vector $\hat{\boldsymbol{\mu}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{+}\mathbf{X}'\mathbf{W}\boldsymbol{y}$ of dimension n that represents the effects of the interventions. The variance of comparison between treatments A and B is defined as $\mathbf{V}_{AB} = \mathbf{L}^{+}_{AA} + \mathbf{L}^{+}_{BB} - 2\mathbf{L}^{+}_{AB}$ [38]. Having the design of treatment comparison $\mathbf{X}_i$ and $\hat{\boldsymbol{\mu}}$ the estimate of treatment effects, the predicted effect size for the study i is given by $\hat{\boldsymbol{y}}_i = \{\hat{y}_{i,k} = \mathbf{X}_i\hat{\boldsymbol{\mu}}, k \epsilon S_i\}, i = 1, \dots, N$.

Let denote with $\tilde{\boldsymbol{\mu}}_{\boldsymbol{i}}$ the relative treatment estimates compared with the treatment reference for study i with dimensions $(n-1)$ and with $\tilde{\mathbf{X}}$ to be the reduced design matrix with dimensions $(n-1)\times n$ of treatment comparisons with the reference (for each row denote treatment comparisons with the reference input zero entries, values 1 to the column corresponding the reference treatment and -1 for the treatment compared). Then, the variance-covariance matrix of $(n-1)$ relative treatment estimates $\tilde{\boldsymbol{\mu}}_{\boldsymbol{i}}$ is denoted with $\tilde{\mathbf{X}}\mathbf{L}^{+}\tilde{\mathbf{X}}'$ and has dimensions $(n-1)\times(n-1)$.

The restricted (residual) maximum log-likelihood (REML) function for random-effects NMA model is given by

$$\begin{aligned} LR(\boldsymbol{y};\, \tau^2) = &-\frac{1}{2}log(det|\mathbf{S}+\boldsymbol{\Delta}|) - \frac{1}{2}log(det|\mathbf{X}'(\mathbf{S}+\boldsymbol{\Delta})^{-1}\mathbf{X}|) \\ &-\frac{1}{2}(\boldsymbol{y}-\mathbf{X}\hat{\boldsymbol{\mu}})'(\mathbf{S}+\boldsymbol{\Delta})^{-1}(\boldsymbol{y}-\mathbf{X}\hat{\boldsymbol{\mu}}) \\ &= \frac{1}{2}log(det|\mathbf{W}|) - \frac{1}{2}log(det|\mathbf{X}'\mathbf{W}\mathbf{X}|) - \frac{1}{2}(\boldsymbol{y}-\mathbf{X}\hat{\boldsymbol{\mu}})'\mathbf{W}(\boldsymbol{y}-\mathbf{X}\hat{\boldsymbol{\mu}}) \end{aligned}$$

where $\mathbf{W} = (\mathbf{S}+\boldsymbol{\Delta})^{-\mathbf{1}}$. The restricted maximum estimation method minimizes the above likelihood function to obtain the parameter estimates.

Krahn *et al.* provided generalized Cochran's Q (Q^{total}) [39]. Based on the fixed-effect model and assuming homogeneity and consistency in the whole network, the generalized Cochran's Q statistic is given by

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

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

- 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_{FE}^{inc}

where the *design* of a study is called the set of treatments compared within the study [39].

The Q statistic (Q_{FE}^{inc}) assess consistency under the assumption of a full design-by-treatment interaction model with a fixed-effect. Having the design-by-treatment interaction model with random-effects, we can measure the inconsistency between studies with different design Q^{inc}. The full design-by-treatment interaction model looks for global inconsistency by allowing for both loop inconsistency and design inconsistency [40].

Table 2.1. Notation for network meta-analysis model.

Studies $i = 1, \dots, N$
Treatments $1, \dots, n$
k pairwise comparison
S_i the set of treatments compared in a study or else the number of arms in study i.
n_{S_i} represents the cardinality of S_i and $k \in S_i$.
Pairwise comparisons $k = 1, \dots, m$.
Observed effect size vector $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_N)'$ with $\boldsymbol{y_i} = \{y_{i,k}, k \in S_i\}, k \in S_i$.
Observed standard errors $\boldsymbol{s} = (\boldsymbol{s}_1, \boldsymbol{s}_2, \dots, \boldsymbol{s}_N)'$ with $\boldsymbol{s_i} = \{s_{i,k}, k \in S_i\}, k \in S_i$.
Design $m \times n$ matrix $\mathbf{X}$, X_i the design of treatment comparison.
$\hat{\tau}^2$ the Generalized DerSimonian–Laird heterogeneity estimator.
Weight $m \times m$ matrix $\mathbf{W}$ a diagonal matrix with weights of pairwise comparisons in its diagonal.
$w_{i,k,FE}$ the fixed-effect (FE) weight of pairwise comparison $$w_{i,k,FE} = 1/s_{i,k}^2 , i = 1, \dots, N , k \in S_i$$
$w_{i,k,RE}$ the random-effects (RE) weight of pairwise comparison $$w_{i,k,RE} = 1/(s_{i,k}^2 + \hat{\tau}^2), i = 1, \dots, N, k \in S_i$$
Laplacian $n \times n$ matrix $\mathbf{L} = \mathbf{X'WX}$
Moore Penrose pseudoinverse $n \times n$ matrix $\mathbf{L}^+$ $$\mathbf{L}^+ = (\mathbf{L} - \mathbf{J}/n)^{-1} + \mathbf{J}/n$$ where $\mathbf{J}$ is $n \times n$ matrix with all elements equal to 1. In case of k multi-arm studies $$\mathbf{L}^+ = -\frac{1}{2{S_i}^2}\mathbf{X'XVX'X}$$ where $\mathbf{V}$ is $S_i \times S_i$ symmetric matrix with the observed variances of all comparisons.

Hat $m \times m$ matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{+}\mathbf{X}'\mathbf{W}$
Network estimates $$\widehat{\boldsymbol{y}}^{nma} = \mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{+}\mathbf{X}'\mathbf{W}\boldsymbol{y} = \mathbf{H}\boldsymbol{y}$$ with variance-covariance $m \times m$ matrix $$Var(\widehat{\boldsymbol{y}}^{nma}) = \mathbf{X}\mathbf{L}^{+}\mathbf{X}'$$
Treatment effects estimates $\widehat{\boldsymbol{\mu}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{+}\mathbf{X}'\mathbf{W}\boldsymbol{y}$
Variances between treatments A and B, $\mathbf{V}_{AB} = \mathbf{L}^{+}_{AA} + \mathbf{L}^{+}_{BB} - 2\mathbf{L}^{+}_{AB}$
Predicted effect size $\widehat{\boldsymbol{y}}_i = \{\hat{y}_{i,k} = \mathbf{X}_i\widehat{\boldsymbol{\mu}}, k \epsilon S_i\}, i = 1, \dots, N.$
Relative treatment estimates $\widetilde{\boldsymbol{\mu}}_{\boldsymbol{i}}$ compared with the reference for study i with dimensions $(n-1)$.
$\widetilde{\mathbf{X}}$ the reduced design matrix with dimensions $(n-1) \times n$ of treatment comparisons with the reference treatment.
$\widetilde{\mathbf{X}}\mathbf{L}^{+}\widetilde{\mathbf{X}}'$ the $(n-1) \times (n-1)$ variance-covariance matrix of $(n-1)$ relative treatment estimates $\widetilde{\boldsymbol{\mu}}_{\boldsymbol{i}}$ compared with the reference treatment.

2.3 Motivating examples

This section provides three published examples of networks of interventions that motivate us to proceed and provide the proposed research in this dissertation; a synthesis of studies to aid smoking cessation dataset, dataset for actinic keratosis and a dataset with thrombolytic drugs.

2.3.1 Dataset comprises four interventions to aid smoking cessation

The first example comprises four interventions to aid smoking cessation [41] [42]. Twenty-four studies ($N = 24$), including twenty-two 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 counseling (C), and group counseling (D). The binary outcome was the number of individuals that successful stopped smoking at 6 to 12 months and the odds ratio was used as summary measure. The dataset with arm level data is a part of R package **netmeta** [35] and the corresponding R code to calculate odds ratios is provided in Appendix A. Data with odds ratios are provided in Appendix Table 1.

2.3.2 Dataset comprises nine interventions for actinic keratosis

Gupta and Paquet [43] compared eight interventions and placebo for actinic keratosis. Thirty-five studies ($N = 35$), including three three-arm trials, compared the relative effects of interventions ($n = 9$): placebo/vehicle (including placebo-PDT) (treatment 1), diclofenac 3% in 2.5% hyaluronic acid (DCF/HA) (treatment 2), 5-fluorouracil (5-FU) 0.5% (treatment 3), imiquimod (IMI) 5% (treatment 4), methyl aminolevulinate (MAL)-PDT (treatment 5), 5-

aminolaevulinic acid (ALA)-photodynamic therapy (PDT) (treatment 6), 5-fluorouracil (5-FU) 5.0% (treatment 7), cryotherapy (treatment 8), and ingenolmebutate (IMB) 0.015–0.05% (treatment 9). The binary outcome was participant complete clearance or an equivalent efficacy and the odds ratio was used as summary measure. The dataset is provided in Appendix Table 2.

2.3.3 Dataset with thrombolytic drugs

Boland *et al.* [44] compared eight thrombolytic drugs ($n = 8$) prescribed after acute myocardial infarction. Twenty-eight studies ($N = 28$), including two three-arm studies and twenty-six two-arm studies comparing interventions: streptokinase (treatment 1), accelerated alteplase (treatment 2), alteplase (treatment 3), streptokinase plus alteplase (treatment 4), tenecteplase (treatment 5), reteplase (treatment 6), urokinase (treatment 7), and nistreptilase (treatment 8). The binary outcome was the mortality within 30 to 35 days of hospital admission. The dataset is provided in Appendix Table 3.

3 Characteristics of published networks of interventions

3.1 Introduction

NMA has been considered as the 'new norm' in evidence synthesis [45]. However, there are still limitations that may cast doubt on the reliability of results. Such limitations are ignoring the underlying assumptions, potential biases, inadequate and not transparent reporting of methods used and the use of wrong synthesis models [8], [12], [46].

There are previous empirical studies exploring the characteristics of networks of interventions concluding to the need for improving the quality of NMA applications [47]–[52]. For example, Bafeta *et al.* resulted that reporting guidelines are necessary to reduce bias in NMA results while Nikolakopoulou *et al.* reported that 68% of the NMAs published by the end of 2012 used inappropriate or unspecified methods for the assessment of inconsistency [48], [49].

There are empirical studies that provide information about the choice of optimal methods used for the assessment of risk of bias in the included studies [53]–[55], the magnitude of heterogeneity [56]–[58], the relative advantages of different methods to evaluate publication bias and small-study effects [59]–[61], and the importance of a comprehensive search for relevant studies [62]. Song *et al.* evaluated the prevalence of inconsistency in networks with three treatments [63], [64], Veroniki *et al.* studied the assessment of inconsistency in NMAs that included at least four treatments using two alternative methods [65] while Chaimani *et al.* have provided empirical evidence about the impact of risk of bias and small study effects [66].

Several developments have been made in the field of NMA and many tutorial and guidance papers have been published [2], [12], [67]–[70]. Efthimiou *et al.* provided a review of methodological articles published until March 2014 [71] and found an increase of published articles with NMA methodology after 2011. In 2012 and 2013, 83 methodological articles were published compared to 58 articles published between 2005-2011. New NMA estimation methods have been provided; Rücker introduced the NMA model from graph theory [5], White *et al.* introduced the NMA model as a specific case of multivariate meta-regression, while

Higgins *et al.* provided a new test for the inconsistency assessment [3], [4]. Several statistical packages and codes have been developed to fit the models using frequentist software. R package **netmeta** offers an advantage for the implementation of the NMA model from graph theory [35]. Moreover, Stata routines are provided by White *et al.* and Chaimani *et al.* for the application of NMA models [72]–[74].

We aim to describe how methodological aspects of NMA and reporting quality of results have evolved over time, monitor the rate of adoption for the new methodologies and provide an overview of the characteristics of published networks of interventions. We aim to provide empirical studies based on the data of published NMA database with the future target to describe how often outliers are provided in NMA datasets and how their existence biased the final results.

3.2 Methods

We conducted an empirical study to collect a database of published NMAs as well as published methodological papers about NMA.

3.2.1 Inclusion and exclusion criteria

Networks were included if they evaluate at least four different interventions (defined as different drugs or other medical treatments, or different schedules, doses or formulations of the same treatment) including placebo, no treatment, waiting list or other control interventions. NMAs with observational or diagnostic test accuracy studies were excluded. NMAs with a smaller number of studies than the number of interventions and NMAs performed with naive indirect comparisons for pooling data were excluded.

3.2.2 Literature search and screening

The search was conducted in Medline, Embase and the Cochrane Database of Systematic Reviews from inception until April 14, 2015, without language restrictions. Titles and abstracts were screened for the eligibility criteria. Potentially relevant full-text articles were screened in the same manner.

3.2.3 Extraction of data characteristics and categorization

We extracted general publication characteristics of articles such as first author, year and journal of publication, residence country of the contact author. We recorded whether the primary outcome measured efficacy or safety and we categorize it into dichotomous, continuous, time-to-event or rate. The total number of interventions was extracted (termed nodes of the network plot) and the reference intervention. Each network is categorized according to the type of

treatment comparison; pharmacological versus placebo, pharmacological versus pharmacological or non-pharmacological versus any treatment. When the reference treatment was not reported, any of the following were selected as the reference treatment node: placebo, usual care, or no treatment. The network geometry was extracted (connected/disconnected network) and each network categorized to star-shaped or network with closed loops.

Characteristics for NMA methodology with an emphasis on statistical analysis and reporting were also extracted. We recorded whether and how the authors evaluated the plausibility of transitivity [2]. For networks including at least one closed-loop, we also recorded the use of inconsistency tests. We categorized the method used to derive indirect and/or network estimates, the effect measure employed to undertake the analysis (such as odds ratio or mean difference) and whether a fixed-effect, random-effects or both models are used. We also recorded any secondary analyses such as subgroup, network meta-regression, or sensitivity analyses that the authors performed to investigate potential sources of heterogeneity or inconsistency. We examined whether authors assessed small-study effects, whether they considered the potential for publication bias and the methods they applied to evaluate their impact on the results. Although, the aim of this dissertation focus on outlier diagnostics, we did not extracted any information because the methodology of outlier detection in network meta-analysis is new and has not been provided in practice yet.

We recorded whether the published article or even the supplementary material presented every possible relative effect estimate between the nodes of the network or if only a subset of them was provided. We also extracted if a ranking measure used for the treatment hierarchy.

3.2.4 Statistical analysis

For the extracted characteristics a descriptive statistical analysis was performed. We evaluated changes over time for several characteristics such as the use of appropriate methods to evaluate consistency or the use of frequentist NMA framework, and quality of reporting over the years. We used a X^2 test for time trend for dichotomous characteristics and the Cox-Stuart trend test for continuous characteristics [75]. All analyses were performed in R software [76] using the R package **trend** [77].

3.3 Results

We identified 3727 abstracts that resulted in 456 networks satisfying all inclusion criteria. Figure 3.1 provides the flow chart of the search strategy and the selection process.

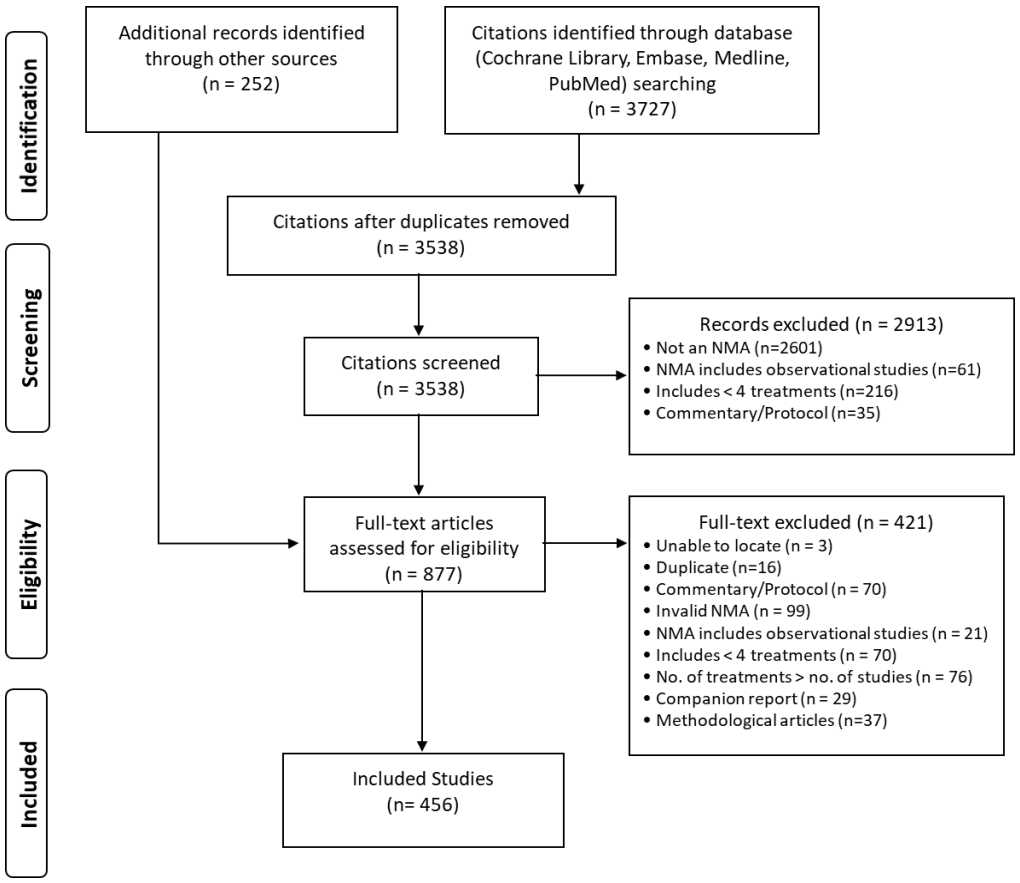

Figure 3.1. Flow chart of the selection process of published networks of interventions**.**

3.3.1 General NMA and publication characteristics

We first monitored that the number of published NMAs has been increasing over the last two decades. Only 6 NMAs were published for the period from 1999 to 2004.

The median number of studies per network was 21 (interquartile range (IQR) 13 to 40) and the median number of treatments was 7 (IQR of 5 to 9) (Table 3.1). Most articles were published in general medicine journals (183 NMAs, 40%). 234 NMAs (51%) had a contact author with affiliation from Europe and 140 NMAs (31%) from the United States.

The majority of NMAs provided pharmacological interventions and placebo treatment comparisons (299 NMAs, 66%). 88 NMAs (19%) provided only pharmacological interventions (19%) and 69 NMAs (15%) compared a mixture of pharmacological, non-pharmacological and control treatments (Table 3.1).

Regarding the network geometry, 73% of networks (331 NMAs) included at least one closed-loop and 27% of NMAs (125 NMAs) were star-shaped networks. Table 3.2 provides that all the NMAs published in 2005 were star-shaped networks but the percentage decreased to 19% in 2015. Moreover, the number of networks with at least one closed-loop has increased through time (p=0.01, Table 3.2).

Characteristics regarding the outcome indicated that the majority of NMAs provided a beneficial type of outcome (260 NMAs, 57%) while 43% of NMAs provided a harmful outcome. The most commonly measured as a dichotomous scale (267 NMAs, 59%), while only 30% of NMAs provide an outcome measured on a continuous scale (Table 3.1).

Table 3.1. Characteristics of 456 NMAs published until 2015. IQR: Interquartile range.

Characteristics of NMAs	**Median (IQR)**
Median number of included treatments	7 (5, 9)
Median number of included studies	21 (13, 40)
	Number of NMAs (%)
General publication characteristics	
Published in general medicine journals*	183 (40%)
Published in health services research journals**	56 (12%)
Published in specialty journals	217 (48%)
Contact author with affiliation in Europe	234 (51%)
Contact author with affiliation in the United States	140 (31%)
Treatment comparisons	
Compare Pharmacological treatments versus placebo	299 (66%)
Only pharmacological treatment comparisons	88 (19%)
A mixture of pharmacological, non-pharmacological and control treatments	69 (15%)
Network geometry	
Networks included at least one closed loop	331 (73%)
Star-shaped networks	125 (27%)
Outcome	
Beneficial outcome	260 (57%)
Dichotomous scale	267 (59%)
Continuous scale	135 (30%)
*Medicine, General & Internal, Pharmacology & Pharmacy, Multidisciplinary Sciences, Medicine, Research & Experimental, Primary Health Care. ** Health Care Sciences & Services, Health Policy & Services.	

3.3.2 Evaluation of transitivity and consistency assumption

The majority of NMAs (353 NMAs, 77%) did not report any statement regarding the transitivity assumption. This tendency changed over time as we found 77% of NMAs published

in 2015 discussing the transitivity assumption (p<0.01) (Table 3.2). We only found 5 NMAs (1%) in which authors reported concerns about potential intransitivity. We found 100 NMAs (22%) that did report how transitivity was evaluated and the majority provided study characteristics comparisons (76 NMAs).

331 NMAs included at least one closed-loop allowing assessment of inconsistency. Nearly half of the networks (150 NMAs, 45%) used appropriate statistical methods to assess consistency and their uptake has increased in the last years (p<0.01, Table 3.2). The most commonly used method for the assessment of inconsistency was the loop-specific approach [13] (59 NMAs, 18%) followed by the node-splitting approach [14] (39 NMAs, 12%). We found only 5 NMAs (2%) implemented the design-by-treatment interaction model [4] but the method was introduced in 2012. Almost 28% percent of NMAs (94 NMAs) did not report any method used to check the plausibility of the consistency assumption.

The proportion of NMAs considered transitivity or methods to evaluate the consistency increased over the years (p<0.01, Table 3.2) with a percentage of 17% of published NMAs in 2006 to 86% of published NMAs in 2015 discussing transitivity or inconsistency.

3.3.3 Statistical synthesis of the data

The most commonly used effect size for NMAs was the odds ratio (177 NMAs, 39%) for the dichotomous outcome and the mean difference (89 NMAs, 20%) for the continuous outcome. Trend test indicated that reporting quality was poor overtime of explaining the reason to choose between the fixed and random-effects model (p=0.01, Table 3.2). Half of the networks (230 NMAs) performed the analysis using the random-effects model. Among the 170 networks (37%) that used the fixed-effect model, the majority (141 NMAs, 83%) also applied the random-effects approach either as sensitivity analysis or with the aim to choose between the two models.

Only 24 NMAs (5%) did not report the synthesis NMA model used while the percentage of NMAs reporting the statistical method used to fit NMA has increased over time from 67% in 2005 to 100% in 2015 (p<0.01, Table 3.2). We found that the Bayesian hierarchical approach (302 NMAs, 64.5%) followed by the Bucher method (88 NMAs, 18.8%) were implemented more often for the statistical evidence synthesis (Table 3.3). Only 80 (18%) NMAs which included at least one multi-arm study employed a method to derive the treatment effect that ignored correlations (e.g. adjusted indirect comparison meta-analysis or Bucher method). We

found 5 NMAs that fit the NMA model as multivariate meta-analysis or multivariate meta-regression. We also found one NMA that employed the NMA approach form graph-theory.

Subgroup, meta-regression or sensitivity analysis were employed to investigate potential sources of heterogeneity or inconsistency by almost half on the NMAs (256 NMAs, 56%). 143 (31%) NMAs implemented methods and graphical tools for small-study effects and publication bias for pairwise comparisons in meta-analysis. Funnel plots (116 NMAs, 81%) and regression tests (82 NMAs, 57%) were the most commonly used methods for the assessment of publication bias while only 7 NMAs (5%) applied the trim and fill method. More complicated approaches, such as the comparison-adjusted funnel plot and the extended selection models [73], [78]–[80] were only implemented by 6 NMAs (4%).

3.3.4 Presentation of results

The presentation of outcome data decreased over time (p=0.03). All possible relative treatment effects are provided for half of the NMAs (234 NMAs, 51%). The rest NMAs present only a subset of relative treatment effects and one NMA (0.2%) did not report any relative treatment effect. 43% (195 NMAs) of NMAs provided the treatment hierarchy with the probability of being the best to be the most commonly used (166 NMAs, 85%) followed by SUCRA values (39 NMAs, 20%). The time trend indicated that the use of the probability of being the best has not changed significantly (p=0.86) but the use of SUCRA values has increased (p<0.01) (Table 3.2).

Table 3.2. The number of NMAs and percentages for characteristics of NMAs published between 2005 and 2015 (until 15 April).

Characteristics of NMAs	Total Number	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	p-value
Star Networks	125	6 (100%)	5 (42%)	2 (22%)	6 (50%)	7 (26%)	7 (23%)	14 (26%)	17 (29%)	19 (20%)	32 (31%)	8 (19%)	0.01
Compare pharmacological vs pharmacological	88	1 (17%)	2 (17%)	2 (22%)	3 (25%)	8 (30%)	3 (10%)	5 (9%)	5 (8%)	23 (24%)	29 (28%)	7 (16%)	0.15
Compare pharmacological vs placebo	299	5 (83%)	8 (67%)	6 (67%)	9 (75%)	14 (52%)	22 (73%)	43 (81%)	42 (71%)	62 (65%)	56 (54%)	26 (60%)	0.31
Compare non-plarmacological vs any	69	0 (0%)	2 (17%)	1 (11%)	0 (0%)	5 (19%)	5 (17%)	5 (9%)	12 (20%)	11 (11%)	18 (17%)	10 (23%)	0.05
No information or discussion on transitivity	353	6 (100%)	12 (100%)	7 (78%)	11 (92%)	23 (85%)	26 (87%)	46 (87%)	46 (78%)	67 (70%)	71 (69%)	33 (77%)	<0.01
Reported that transitivity is likely to hold	98	0 (0%)	0 (0%)	1 (11%)	1 (8%)	4 (15%)	4 (13%)	7 (13%)	13 (22%)	27 (28%)	30 (29%)	10 (23%)	<0.01
Use appropriate methods to test inconsistency* **	150	NA	1 (14%)	2 (29%)	2 (33%)	6 (30%)	4 (17%)	13 (33%)	16 (38%)	43 (56%)	36 (51%)	26 (74%)	<0.01
Discuss about transitivity or inconsistency (at least one of the two)	285	0 (0%)	2 (17%)	3 (33%)	5 (42%)	12 (44%)	17 (57%)	30 (57%)	40 (68%)	66 (69%)	72 (70%)	37 (86%)	<0.01
Clearly reported whether random or fixed effects are used	400	5 (83%)	10 (83%)	7 (78%)	10 (83%)	20 (74%)	25 (83%)	44 (83%)	53 (90%)	91 (95%)	93 (90%)	38 (88%)	0.01
Method for NMA reported	432	4 (67%)	8 (67%)	9 (100)	11 (92%)	23 (85%)	30 (100%)	51 (96%)	56 (95%)	95 (99%)	99 (96%)	43 (100)	<0.01
Use Bayesian hierarchical model to fit NMA	302	1 (17%)	3 (25%)	3 (33%)	4 (33%)	13 (48%)	19 (63%)	35 (66%)	43 (73%)	77 (80%)	71 (69%)	33 (77%)	<0.01
Formal exploration of heterogeneity	256	2 (33%)	9 (75%)	5 (56%)	6 (50%)	16 (59%)	20 (67%)	36 (68%)	32 (54%)	56 (58%)	51 (50%)	20 (47%)	0.1

All pairwise effects are presented	234	1 (17 %)	3 (25 %)	2 (22 %)	4 (33 %)	15 (56 %)	17 (57 %)	31 (58 %)	29 (49 %)	54 (56 %)	55 (53 %)	23 (53 %)	0.02
Available outcome data	308	4 (67 %)	8 (67 %)	8 (89 %)	10 (83 %)	23 (85 %)	24 (80 %)	36 (68 %)	38 (64 %)	55 (57 %)	71 (69 %)	27 (63 %)	0.03
Use only Pbest for ranking	137	1 (17 %)	2 (17 %)	3 (33 %)	1 (8 %)	10 (37 %)	13 (43 %)	16 (30 %)	20 (34 %)	33 (34 %)	32 (31 %)	6 (14 %)	0.86
Use SUCRA	39	0 (0%)	0 (0%)	0 (0 %)	0 (0 %)	0 (0 %)	0 (0%)	1 (2 %)	4 (7 %)	10 (10 %)	9 (9 %)	14 (33 %)	<0.01
Number of NMAs published	456*	6	12	9	12	27	30	53	59	96	103	43	0.04**

*There are 6 networks published before 2005 and are included in the total NMA group. ** In the test for trend for the total number of published NMAs we excluded the year 2015 as it is not complete. ***Here the denominator is the number of articles with at least one closed-loop (number of NMAs published minus the star-shaped NMA). P-values from a trend test.

Table 3.3. Statistical models used for evidence synthesis. The number of articles and percentages.

Characteristics of NMAs	**Number of NMAs (%)**
Bayesian hierarchical model	302 (64.5%)
Bucher method	88 (18.8 %)
Meta-regression	44 (9.4 %)
Not reported or unclear	25 (5.4 %)
Multivariate meta-analysis or meta-regression	5 (1.1 %)
Indirect synthesis method	3 (0.6 %)
NMA from graph theory	1 (0.2 %)

3.4 Discussion

The number of published NMAs increased substantially over the years. The importance of multiple treatment comparisons has now well known among researchers in various fields of health and education.

Reporting quality was also improved as we found all articles published in 2015 to include a description of the statistical methods used. The PRISMA (Preferred Reporting Items for Systematic reviews and Meta-analysis) statement was published only recently and we expect to have an impact on the improvements in reporting quality for NMA applications.

Improvements might be due to statisticians becoming more experienced with the NMA methodology. There are several educational published articles for NMA methodology that might provide some impact for the improvement of NMA methodology and the reporting quality [2], [12], [67], [81], [82]. Suggestions for accompanied detailed protocols on which authors should base their NMAs have started to be applied [2], [12]. NMA protocol registration can improve the reporting quality of NMAs and it can help to define a priori the assessment of transitivity, inconsistency and several methodologies used for the NMA analysis.

It is notable that the Bayesian hierarchical model found to be the most popular approach for NMA model synthesis as only five articles employed NMA using frequentist approaches. The use of new methodological frequentist developments is expected to be increased after 2015, such as the NMA model from graph theory with **netmeta** R package [35], Stata routines for the multivariate random-effects meta-analysis model [3], [74], and design-by-treatment interaction model to test the consistency assumption [4].

Many NMAs in this database provide important methodological limitations. Accordingly, the use of appropriate methods improved over the years. For example, an increasing number of NMAs addressed transitivity or inconsistency, as three-quarters of networks published in 2015 used appropriate methods to test the plausibility of the consistency assumption. It was not noticed a change regarding the discussion of the transitivity assumption but it is a need to change that aiming to take valid NMA results.

To the best of our knowledge, this is the largest collection of published NMAs up to date. It includes nearly three times the data included in Bafeta *et al.* [47], more than twice the data included in Nikolakopoulou *et al.* [49] and about 40% more data compared to the collection by Chambers *et al.* [83]. This is also the first study to formally investigate the changes in methodology and reporting quality of NMAs over time.

This empirical evidence could inform simulation scenarios (e.g. median number of studies or treatments) conducted in the NMA field. Based on this database of NMAs, several empirical projects can be conducted in several aspects of NMA methodology. For the aim of the research of this dissertation, the extracted data of NMA database could be used to provide a description on how often outliers are provided in NMA datasets and how their existence biased the final results.

4 Methods to detect outliers in meta-analysis

4.1 Introduction

There are many definitions for outliers in the literature and several methods to detect outliers in regression models have been provided. Meta-analysis is actually a weighted regression model. Assumption that no outliers exist in the data is the basis for weighted least squares estimation or for normal likelihood maximization. Outlier detection is crucial as weighted least squares estimates are sensitive to outliers and their existence may be bias the model parameter estimation. Several outlier detection methods have been extended from regression models to a pairwise meta-analysis.

Outlier diagnostics measures fitted in the fixed-effect meta-analysis were firstly provided in a Chapter by Hedges [84]. Viechtbauer and Cheung offered outlier diagnostics measures considering the effect deletion of study have on fixed- and random-effects meta-analysis [85].

The detection of outliers is not an integral part of NMA. Outlier detection in NMA is much more challenging compared to a pairwise meta-analysis as outlying studies may have an impact on the underlying model and may not be easily identified visually. In addition, data are multivariate and an effect can be suspicious not only by its mere size but also by its size conditional on the comparison of the study and/or the corresponding effect derived from indirect evidence.

As Zhang *et al.* [86] and Zhao *et al*. [87] remarked previously, outliers may be the primary source of heterogeneity or inconsistency and may affect the validity of NMA results. Only four methodological papers have addressed how to detect outliers in NMA evidence structures up to date. Lu and Ades proposed the use of residual deviance to detect outliers in the Bayesian hierarchical model [41]. Zhang *et al.* [86] provided four measures for the detection of outliers in the Bayesian hierarchical model while Zhao *et al.* [87] offered outlier detection measures for generalized Bayesian hierarchical models to detect outliers at observation data and not at

aggregated data. Noma *et al.* [88] have recently provided four measures to detect outliers in frequentist NMA model based on multivariate random-effects meta-regression method.

In this Chapter, we provide a brief synopsis of the several outlier detection strategies that have been proposed in regression and meta-analytical context. We propose and evaluate several measures and graphical tools that seek to accomodate influential studies and outliers in network meta-analysis. These procedures are logical extensions from pairwise meta-analysis and regression to the NMA model. The proposed outlier and influential detection measures and visual tools can be implemented to any NMA dataset with our developed R package **NMAoutlier** [28].

This Chapter is organized as follows: Section 4.2 provides an overview of outlying identification in the meta-analytical context; Section 4.3 outlines the proposed measures extended from pairwise meta-analysis to network meta-analysis; Section 4.4 provides an application of the several outlier and influential measures in real datasets of networks of interventions and Section 4.5 discusses the main findings and provides conclusions.

4.2 Outlying detection strategies in meta-analysis

Outliers can affect model parameters possibly causing bias. For example, the arithmetic mean, is known to be particularly sensitive to outlying observations and the presence of even one outlier unduly influences the results derived.

There are two different ways to interpret outliers; geometrically and probabilistically. With the former interpretation, outliers are extreme values that lie far away from the other observations while with the latter, outliers are those observations that are most unlikely to occur under the hypothesized model [89]. Fitting of the model should be based on a *clean* dataset.

There are two interpretation mistakes if characterizing a study as an outlier or not; *the masking effect* and the *swamping effect*. If there is a cluster of outliers, it is likely that results would be affected to such a degree that outliers will not be identified. This is known as masking effect and it is similar to a false negative. Barnett and Lewis define the masking effect as the inability to identify even a single outlier in the presence of several suspected values [90]. Thus, the presence of a single outlier masks the appearance of other outliers. Moreover, we may have a false-positive result. Outliers may affect the summary effect to a such a degree that non-outlying values may falsely appear to be outlying. This is known as the swamping effect.

Based on our bibliographic knowledge, we outline a synopsis of the strategies to accommodate outliers in meta-analysis. The several strategies to detect and adjust for outliers in meta-analytical models can be classified in five general categories: A. Use of distribution other than normal (alternative distributions) for random effects; B. Robust heterogeneity measures; C. Likelihood methods; D. Deletion/backward methods; E. Forward methods and F. Robust estimation.

A. Alternative distributions for the random-effects model

Meta-analysis typically assumes a normal distribution for the random-effects model. Jackson and White provided several situations in which normality is questioned [91]. It has been suggested that more flexible distributions should be taken more frequent in practice. Alternative long-tailed random effects distributions reduce the weight given to more extreme study effects (outliers). Lee and Thompson argued that normality might be a restrictive assumption for the random-effects model and they provided alternative distributions with heavier tails [92]. They suggested the t distribution for random effects u_j with density function

$$p(u_j/mean, scale, df) = \frac{\Gamma((df+1)/2)}{\Gamma(df/2)\sqrt{\pi df}}\left(1+\frac{(u_j-mean)^2}{dfscale}\right)^{-(df+1)/2}$$

where $mean$ is the mean, $scale$ is the scale parameter and df degrees of freedom and they offered skewed extensions for normal and t distribution. Baker and Jackson suggested alternative distributions to downweigh outlying studies such as long-tailed distributions, arcsin distribution, beta distribution, Subbotin distribution and alternative vague priors in Bayesian analysis [93]. Baker and Jackson proposed two new marginal distributions with additional parameters to model skewness and heavier tails [94].

B. Robust heterogeneity measures

Outliers can have an impact on the estimation of heterogeneity causing bias to meta-analytical results. Lin *et al.* proposed alternative heterogeneity measures in the meta-analysis that are robust in the presence of outliers [95]. They provided two alternative Dersimonian and Laird heterogeneity estimators using the weighted average and the weighted median instead of the standard weighted mean. Yu *et al.* proposed a robust to outliers version of maximum likelihood (ML) estimation method based on two loss functions for log-likelihood; Huber's rho function and Tukey's biweight function [96].

C. Likelihood methods

Gumedze and Jackson, based on the likelihood function, proposed a model that shifts the random-effects variance of each included study, separately [97]. A random-effects shift variance model is capable to identify and downweigh studies with inflated variance (outliers) [97]. Monitoring is deemed helpful as sharp changes in detection measures can be an indication for outliers. Beath [98] proposed a method that considers two classes of outlying and non-outlying studies and based on a finite mixture approach can detect and downweigh the outlying cases.

D. Backward/Deletion methods

Viechtbauer and Cheung extended several outlying and influence diagnostic measures developed for the linear regression model in the context of meta-analysis [85] and included them in the R package **metafor** [99]. The diagnostic measures provided the influence of a study to model parameters considering its deletion [85]. Deleted residuals, Cook's distance, the change in the variance-covariance matrix of the parameter estimates when a study deleted and R_i statistic are some of the measures provided [85]. Shi *et al.* provided an updating formula of the measures using case deletion diagnostic method and local influence analysis under the DerSimonian and Laird and maximum likelihood estimation, respectively [100]. Backward algorithms have been developed for outlier diagnosis in meta-analysis [101]. This type of algorithm removes observations according to some criterion (e.g. the largest residual) and stops when some criterion is met (e.g. all residuals are smaller than a threshold value) [100]. The main drawback of backward methods is that may have masking and swamping effects due to the fact that all observations, including outliers, are used and conclusion may be affected.

E. Forward methods

Forward Search algorithm is an outlying identification method and it has recently been implemented in meta-regression [101]. It starts with an initial subset of studies that is ideally assumed to be outlier-free and it gradually adds the remaining studies according to their closeness to the set of selected studies under the hypothesized model. In each iteration, parameter estimates, measures of fit and test statistics can be monitored. Sharp changes in monitoring measures can indicate potential outlying studies. In contrast to deletion methods, forward methods are unaffected of masking and swamping effects and this is the main advantage of their implementation.

F. Robust estimation

The classical least squares regression is sensitive in the presence of outliers. The robust estimation has been developed for regression models and several methods have been suggested. Yu *et al.* provided a review of robust estimation methods that have been proposed in regression and conducted a simulation comparison of the detection methods [96]. Huber [102] introduced the *M-estimator* which is a solution of the normal equation with appropriate weight functions. Rousseeuw provided the *least trimmed squares* (LTS) estimates [103] which minimize the trimmed sum of squared residuals. Rousseeuw [104] also introduced the *least median squares* (LMS) estimation which minimize the median sum of squares. Extension of robust statistics to meta-analytical models is innovative as there is no methodological work up to date. It can be an area for future work in the view that there is large bibliographic research for robust estimation statistics in regression models and it is not sensitive in the presence of outliers.

4.3 Outlier and influential case diagnostics measures for NMA

This Section provides the extension of the several outlying and influential detection measures from regression or pairwise meta-analytical models to the frequentist random-effects NMA model [5]. Details about the fitted model and useful notation can also be found in Chapter 2 of this dissertation.

4.3.1 Outlier detection measures

Table 4.1 provides an overview of the proposed outlier detection measures; contribution to the Q statistic (Mahalanobis distance), residuals, and leverage.

Contribution to the Q_i (Mahalanobis distance)

The most commonly used tool to assess the presence of outliers for multivariate data is the Mahalanobis distance. An analogy to Mahalanobis distance for a pairwise comparison (or two-arm study) is the contribution of the study i to Cochran's Q statistic [105]. More specifically, the formula is give by

$$D^2_{Q_{i,k}} = Q_{i,k} = w_{i,k,FE}\left(y_{i,k} - \hat{y}_{i,k}\right)^2$$

where $w_{i,k,FE} = 1/s^2_{i,k}$, $i = 1, \ldots, N$, $k \in S_i$ is the fixed-effect weight of pairwise comparison of a study.

In the case of multi-arm study, the squared of contribution to the Q for a study i with k pairwise comparisons is given by the arithmetic average

$$D_{Q_i}^2 = Q_i = \frac{1}{n_{S_i}} \sum_{k \in S_i} Q_{i,k} = \frac{1}{n_{S_i}} \sum_{k \in S_i} w_{i,k,FE} \left(y_{i,k} - \hat{y}_{i,k}\right)^2$$

where n_{S_i} the cardinality of S_i and in matrix form

$$D_{Q_{i,k}}^2 = \boldsymbol{Q}_{i,k} = (\boldsymbol{y_i} - \hat{\boldsymbol{y}}_{\boldsymbol{i}})' \left(Cov(\boldsymbol{y_i})\right)^{-1} (\boldsymbol{y_i} - \hat{\boldsymbol{y}}_{\boldsymbol{i}})$$

$$= (\boldsymbol{y_i} - \boldsymbol{X}_i \hat{\boldsymbol{\mu}})' \left(\mathbf{S}_{\boldsymbol{i,adj}}\right)^{-1} (\boldsymbol{y_i} - \boldsymbol{X}_i \hat{\boldsymbol{\mu}})$$

$$= (\boldsymbol{y_i} - \boldsymbol{X}_i \hat{\boldsymbol{\mu}})' \mathbf{W}_i (\boldsymbol{y_i} - \boldsymbol{X}_i \hat{\boldsymbol{\mu}})$$

where $\mathbf{W}_i$ is the fixed-effect weight matrix of a study i with $\mathbf{W}_i = \left(\boldsymbol{S_{i,adj}}\right)^{-1}$, with $\boldsymbol{S_{i,adj}}$ a matrix with adjusted squares of standard errors for a study i.

Leverage

Influential studies are observations that have a large impact on the model parameters. The study with extreme value and moderate to large weight in the model parameters is called the *leverage* point. Detection of such observations in regression can easily be observed with the leverage score. For the k^{th} pairwise comparison in the i^{th} study, the leverage score is the k^{th} diagonal of the hat matrix, defined as

$$h_{i,k} = (\mathbf{H})_{i,kk}$$

where $\mathbf{H}$ is the block diagonal hat matrix of random-effects model $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}$ with blocks referring to different studies. Large leverage points indicate the existence of influential studies.

Raw residuals

A more formal approach is to examine the residuals. The *raw pairwise residual* for the k^{th} pairwise comparison in i^{th} the study is the difference between the observed effect size $y_{i,k}$ and the predicted effect size based on the random-effects model $\hat{\boldsymbol{y}}_{\boldsymbol{i}} = \left\{\hat{y}_{i,k} = \boldsymbol{X_i}\hat{\boldsymbol{\mu}}, k \, \epsilon \, S_i\right\}$ given by

$$\hat{\varepsilon}_i^{k,pair,raw} = y_{i,k} - \hat{y}_{i,k}$$

$$= y_{i,k} - \boldsymbol{X_i}\hat{\boldsymbol{\mu}}$$

For each two-arm study, the raw residual (*raw study residual*) is the same with its raw pairwise residual. In case of a multi-arm study i with $k\epsilon \, S_i$ pairwise comparisons, we define the *raw*

study residual to be equal to the square root of the average of squared *raw residuals within the trial.*

$$\hat{\varepsilon}_i^{k,study,raw} = \sqrt{\frac{1}{n_{S_i}} \sum_{k \in S_i} \left(y_{i,k} - \hat{y}_{i,k}\right)^2}$$

Standardized residuals

The *standardized pairwise residual* (equivalent to the squared of Mahalanobis distance with random-effects weights) for the k^{th} pairwise comparison in i^{th} study is the standardized difference between the observed effect size $y_{i,k}$ and the predicted effect size $\hat{y}_{i,k}$ given by

$$\hat{\varepsilon}_i^{k,pair,stand} = \frac{y_{i,k} - \hat{y}_{i,k}}{\sqrt{s_{i,k}^2 + \hat{\tau}^2}}$$

where $\hat{\tau}^2$ is the Generalized DerSimonian and Laird heterogeneity estimator. For each two-arm study, the standardized residual (*standardized study residual*) is the same with its standardized pairwise residual. In case of a multi-arm study i with $k \epsilon\, S_i$ pairwise comparisons, we define the *standardized study residual* to be the squared root of the average of squared *standardized pairwise residuals*

$$\hat{\varepsilon}_i^{k,study,stand} = \sqrt{\frac{1}{n_{S_i}} \sum_{k \in S_i} \left(\frac{y_{i,k} - \hat{y}_{i,k}}{\sqrt{s_{i,k}^2 + \hat{\tau}^2}}\right)^2}$$

where n_{S_i} represents the cardinality of S_i .

Studentized residuals

The *studentized pairwise residual* for the k^{th} pairwise comparison in i^{th} study is given by

$$\hat{\varepsilon}_i^{k,pair,stud} = \frac{y_{i,k} - \hat{y}_{i,k}}{\sqrt{(1 - h_i)\left(s_{i,k}^2 + \hat{\tau}^2\right)}}$$

where $h_{i,k} = (\mathbf{H})_{kk}$ is the i^{th} diagonal of hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}$ in the NMA model. For each two-arm study, the studentized residual (*studentized study residual*) is the same with its studentized pairwise residual. In case of a multi-arm study i with k pairwise comparisons,

we define the *studentized study residual* to be the squared root of the average of squared *studentized pairwise residuals*

$$\hat{\varepsilon}_i^{k,study,stud} = \sqrt{\frac{1}{n_{S_i}} \sum_{k \in S_i} \left(\frac{y_{i,k} - \hat{y}_{i,k}}{\sqrt{(1-h_i)\left(s_{i,k}^2 + \hat{\tau}^2\right)}} \right)^2}$$

We can use a boundary of 1.96 or 2 for the value of standardized and studentized study residuals.

Table 4.1 Overview of outlier detection measures in NMA.

Outlier detection measures	**Formula**	**Cut-offs**
Contribution to the Q (Mahalanobis distance)	$D_{Q_i}^2 = \frac{1}{n_{S_i}} \sum_{k \in S_i} Q_{i,k}$ $D_{Q_i}^2 = \frac{1}{n_{S_i}} \sum_{k \in S_i} w_{i,k,FE} \left(y_{i,k} - \hat{y}_{i,k} \right)^2$	
Leverage	$h_{i,k} = (\mathbf{H})_{i,kk}$ $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}$	large value
Raw study residual	$\hat{\varepsilon}_i^{k,study,raw} = \sqrt{\frac{1}{n_{S_i}} \sum_{k \in S_i} \left(y_{i,k} - \hat{y}_{i,k} \right)^2}$	
Standardized study residual	$\hat{\varepsilon}_i^{k,study,stand} = \sqrt{\frac{1}{n_{S_i}} \sum_{k \in S_i} \left(\frac{y_{i,k} - \hat{y}_{i,k}}{\sqrt{s_{i,k}^2 + \hat{\tau}^2}} \right)^2}$	1.96 or 2
Studentized study residual	$\hat{\varepsilon}_i^{k,study,stud} = \sqrt{\frac{1}{n_{S_i}} \sum_{k \in S_i} \left(\frac{y_{i,k} - \hat{y}_{i,k}}{\sqrt{(1-h_i)\left(s_{i,k}^2 + \hat{\tau}^2\right)}} \right)^2}$	1.96 or 2

4.3.2 Outlier detection measures considered deletion

"*Leave one out" parameters for NMA model*

Viechtbauer and Cheung suggested using residuals after study deletion [85]. We provide by analogy the corresponding measures in NMA context. Table 4.2 summarizes the notation for "*leave one out*" model parameters. We denote the NMA model parameters excluding study i; the summary estimate of treatment effects "leave one out" $\widehat{\boldsymbol{\mu}}_{(-\boldsymbol{i})}$ and the between-study variance estimator $\hat{\tau}^2_{(-i)}$ termed as heterogeneity estimator "leave one out". $\boldsymbol{X_i}$ is the raw of the design matrix for study i from the whole network that provides the treatment comparison of the study i. Based on model estimation, let us define the predicted value "leave one out" for study i (that is actually excludes the study i) in the k^{th} pairwise comparison $\widehat{\boldsymbol{y}}_{i,k(-i)} = \boldsymbol{X}_i'\widehat{\boldsymbol{\mu}}_{(-\boldsymbol{i})}$. We define the random-effects weight "leave one out" to be the weight excluding the study i of the k^{th} pairwise comparison

$$w_{i,k(-i)} = \frac{1}{s^2_{i,k} + \hat{\tau}^2_{(-i)}}$$

The weight "leave one out" for a multi-arm study is given with the arithmetic mean by

$$w_{i(-i)} = \frac{1}{n_{S_i}} \sum_{k \in S_i} w_{i,k(-i)}$$

Having $w_{i,k(-i)}$ in a diagonal for the k^{th} pairwise comparison in i^{th} study, we denote the weight matrix $\mathbf{W}_{(-i)}$ to be a diagonal with $w_{i,k(-i)}$ entries. Then, the hat matrix "leave one out" can be defined as $\mathbf{H}_{(-\boldsymbol{i})} = \mathbf{X}_i\left(\mathbf{X}_i{}'\mathbf{W}_{(-i)}\mathbf{X}_i\right)^{-1}\mathbf{X}_i{}'\mathbf{W}_{(-i)}$ with the leverage "leave one out" to be the k^{th} diagonal of the hat matrix $h_{(-i)} = \left(\mathbf{H}_{(-\boldsymbol{i})}\right)_{kk}$.

Table 4.2. "Leave one out" parameters for the NMA model.

"leave one out" model parameters	**Symbol or Formula**
Summary estimate of treatment effects "leave one out"	$\widehat{\boldsymbol{\mu}}_{(-\boldsymbol{i})}$
Heterogeneity estimator "leave one out"	$\hat{\tau}^2_{(-i)}$
Predicted value "leave one out"	
Weight "leave one out"	$w_{i(-i)} = \frac{1}{n_{S_i}} \sum_{k \in S_i} w_{i,k(-i)}$

	$w_{i,k(-i)} = \frac{1}{s_{i,k}^2 + \hat{\tau}_{(-i)}^2}$
Leverage "leave one out"	$h_{(-i)} = \left(\mathbf{H}_{(-i)}\right)_{kk}$ $\mathbf{H}_{(-i)} = \mathbf{X}_i\left(\mathbf{X}_i'\mathbf{W}_{(-i)}\mathbf{X}_i\right)^{-1}\mathbf{X}_i'\mathbf{W}_{(-i)}$ $\mathbf{W}_{(-i)} = diag(\frac{1}{s_{i,k}^2 + \hat{\tau}_{(-i)}^2})$

Raw, Standardized, Studentized residuals considered deletion

The *raw study deleted residual* $\hat{\varepsilon}_{i(-i)}^{k,study,raw}$ can be provided as

$$\hat{\varepsilon}_{i(-i)}^{k,study,raw} = \sqrt{\frac{1}{n_{S_i}} \sum_{k \in S_i} \left(y_{i,k} - \hat{y}_{i,k(-i)}\right)^2}$$

and the *standardized study deleted residual* is given by

$$\hat{\varepsilon}_{i(-i)}^{k,study,stand} = \sqrt{\frac{1}{n_{S_i}} \sum_{k \in S_i} \left(\frac{y_{i,k} - \hat{y}_{i,k(-i)}}{\sqrt{s_{i,k}^2 + \hat{\tau}_{(-i)}^2}}\right)^2}$$

where $\hat{\tau}_{(-i)}^2$ is the heterogeneity "leave one out" and *studentized study deleted residual* is provided by

$$\hat{\varepsilon}_{i(-i)}^{k,study,stud} = \sqrt{\frac{1}{n_{S_i}} \sum_{k \in S_i} \hat{\varepsilon}_{i(-i)}^{k,pair,stud^2}}$$

$$= \sqrt{\frac{1}{n_{S_i}} \sum_{k \in S_i} \left(\frac{y_{i,k} - \hat{y}_{i,k(-i)}}{\sqrt{Var\left(y_{i,k} - \hat{y}_{k(-i)}\right)}}\right)^2}$$

$$= \sqrt{\frac{1}{n_{S_i}} \sum_{k \in S_i} \left(\frac{y_{i,k} - \hat{y}_{i,k(-i)}}{\sqrt{s_{i,k}^2 + \hat{\tau}^2 + h_{(-i)} \frac{1}{s_{i,k}^2 + \hat{\tau}_{(-i)}^2}}}\right)^2}$$

where n_{S_i} represents the cardinality of S_i.

We can use a boundary of 1.96 or 2 for the value of standardized and studentized study deleted residuals.

Cook's distance

To examine the influence of the deletion of study i, Viechtbauer and Cheung provided an analogy to Cook's statistic in meta-analytical context [85]. Let us consider again the relative treatment estimates $\tilde{\boldsymbol{\mu}}_{\boldsymbol{i}}$ compared with the reference for study i with dimensions $(n-1)$ as introduced in Chapter 2. We denote with $\tilde{\boldsymbol{\mu}}_{i(-i)}$ the relative treatment estimates after considered deletion of study i. Based on the formula that Viechtbauer and Cheung provided for Cook's statistic in meta-analytical context. Noma *et al.* [88] extended the Cook's distance measure in a multivariate meta-regression model. The analogy for the *Cook's distance* or *Cook's statistic* for study i in NMA model from graph theory is given by

$$C_i = \left(\tilde{\boldsymbol{\mu}}_i - \tilde{\boldsymbol{\mu}}_{i(-i)}\right)' \left(\tilde{\mathbf{X}}\mathbf{L}^{+}\tilde{\mathbf{X}}'\right)^{-1} \left(\tilde{\boldsymbol{\mu}}_i - \tilde{\boldsymbol{\mu}}_{i(-i)}\right)$$

where $\tilde{\mathbf{X}}$ is the reduced design matrix and $\tilde{\mathbf{X}}\mathbf{L}^{+}\tilde{\mathbf{X}}'$ is the $(n-1)\times(n-1)$ variance-covariance matrix of $(n-1)$ relative treatment estimates $\tilde{\boldsymbol{\mu}}_{\boldsymbol{i}}$ as introduced in Chapter 2. A general rule provided in the bibliography for a cut off value for Cook's statistic is that the study i is considered outlier and/or influential if $C_i > 1$ [106], [107]. It has been suggested that C_i larger than the 50% of F distribution with n and $m-n$ degrees of freedom, $F(n, m-n)$, can indicates an influential study [106]. According to Chatterjee and Hadi, a graphical plot with all Cook's distance values can be examined rather than using a cut off value [106].

COVRATIO

Viechtbauer and Cheung propose the *ratio of the determinants of the variance-covariance matrix of treatment estimates (COVRATIO)* when excluding the i^{th} study from model fitting [85]. Noma *et al.* [88] extended the COVRATIO measure in a multivariate meta-regression model.

For graph-theoretical NMA model is given by

$$COVRATIO_i = \frac{det\left(\mathrm{Cov}(\tilde{\boldsymbol{\mu}}_{(-i)})\right)}{det(\mathrm{Cov}(\tilde{\boldsymbol{\mu}}))}$$

where det denotes the determinant of a matrix. More analytically, the ratio of the determinant of the variance-covariance matrix of treatment estimates is defined as

$$COVRATIO_i = \frac{det\left(\widetilde{\mathbf{X}}_{(-i)}\, \mathbf{L}^{+}_{(-i)}\, \widetilde{\mathbf{X}}'_{(-i)}\right)}{det\left(\widetilde{\mathbf{X}}\mathbf{L}^{+}\, \widetilde{\mathbf{X}}'\right)}$$

where $\widetilde{\mathbf{X}}_{(-i)}$ is the reduced design matrix and $\mathbf{L}^{+}_{(-i)}$ the Laplacian matrix "leave one out" by fitting the model without the study i. When the ratio of determinants of the variance-covariance matrix of treatment estimates is lower than the value 1 indicates that the removal of study i yields to more precise treatment estimates [85].

R_i statistic

Large changes in the estimate of between-study heterogeneity can provide the presence of a potential outlier. Viechtbauer and Cheung provided the R_i statistic is given by

$$R(\hat{\tau}^2)_i = 100 \times \frac{\hat{\tau}^2 - \hat{\tau}^2_{(-i)}}{\hat{\tau}^2}$$

$R(\hat{\tau}^2)_i$ statistic quantifies the change in the estimate of the heterogeneity estimator with the exclusion of the study.

Hedges and Olkin suggested also to examine changes in Cochran's Q statistic [84]. We provide the R_i statistic for monitoring changes for generalized Cochran's Q (Q^{total}) defined as

$$R(Q^{total})_i = 100 \times \frac{Q^{total} - Q^{total}_{(-i)}}{Q^{total}}$$

The analogy of R_i statistic for Q statistic within designs (Q^{het}) as the presence of outlier can influence the amount of heterogeneity

$$R(Q^{het})_i = 100 \times \frac{Q^{het} - Q^{het}_{(-i)}}{Q^{het}}$$

and the analogy of R_i statistic for Q statistic between designs (Q^{inc}) as the presence of outlier can influence the inconsistency

$$R(Q^{inc})_i = 100 \times \frac{Q^{inc} - Q^{inc}_{(-i)}}{Q^{inc}}$$

Large positive values of R_i statistics indicate that the removal of study i provides large changes in heterogeneity or inconsistency measures which we expect it when study i is outlier and/or influential study [85].

$DFBETAS$ statistic

The influence of deletion of study i can also be examined with the *$DFBETAS$ statistic* that is given in NMA by

$$DFBETAS_i = \tilde{\boldsymbol{\mu}}_i - \tilde{\boldsymbol{\mu}}_{i(-i)} = (\tilde{\boldsymbol{\mu}}_i - \tilde{\boldsymbol{\mu}}_{i(-i)}) \sqrt{\frac{1}{n_{S_i}} \sum_{k \in S_i} w_{i,k(-i)}}$$

Where $w_{i,k(-i)} = 1/\left(s_{i,k}^2 + \hat{\tau}_{(-i)}^2\right)$. When $DFBETAS_i > 1$ then the study i is considered influential for small to medium datasets [85], [108]. Table 4.3 summarizes the "leave one out" detection measures.

Table 4.3. Overview of outlying detection measures considered study deletion "leave one out" measures.

Leave one out detection measure	**Formula**	**Cut-offs**
Raw study deleted residual	$\hat{\varepsilon}_{i(-i)}^{k,study,raw} = \sqrt{\frac{1}{n_{S_i}} \sum_{k \in S_i} \left(y_{i,k} - \hat{y}_{i,k(-i)}\right)^2}$	
Standardized study deleted residual	$\hat{\varepsilon}_{i(-i)}^{k,study,stand} = \sqrt{\frac{1}{n_{S_i}} \sum_{k \in S_i} \left(\frac{y_{i,k} - \hat{y}_{i,k(-i)}}{\sqrt{s_{i,k}^2 + \hat{\tau}_{(-i)}^2}}\right)^2}$	1.96 or 2
Studentized study deleted residual	$\hat{\varepsilon}_{i(-i)}^{k,study,stud} = \sqrt{\frac{1}{n_{S_i}} \sum_{k \in S_i} \left(\frac{y_{i,k} - \hat{y}_{i,k(-i)}}{\sqrt{s_{i,k}^2 + \hat{\tau}^2 + h_{(-i)} \frac{1}{s_{i,k}^2 + \hat{\tau}_{(-i)}^2}}}\right)}$	1.96 or 2
Cook's distance C_i	$C_i = \left(\tilde{\boldsymbol{\mu}}_i - \tilde{\boldsymbol{\mu}}_{i(-i)}\right)' \left(\tilde{\mathbf{X}}\mathbf{L}^+ \tilde{\mathbf{X}}'\right)^{-1} \left(\tilde{\boldsymbol{\mu}}_i - \tilde{\boldsymbol{\mu}}_{i(-i)}\right)$	$C_i > 1$

Ratio of the determinants of the variance-covariance matrix ($COVRATIO_i$) of treatment estimates	$COVRATIO_i = \frac{det\left(\widetilde{\mathbf{X}}_{(-i)}\, \mathbf{L}^{+}_{(-i)}\, \widetilde{\mathbf{X}}'_{(-i)}\right)}{det\left(\widetilde{\mathbf{X}}\mathbf{L}^{+}\, \widetilde{\mathbf{X}}'\right)}$	$COVRATIO_i < 1$
R_i statistic for heterogeneity	$R(\hat{\tau}^2)_i = 100 \times \frac{\hat{\tau}^2 - \hat{\tau}^2_{(-i)}}{\hat{\tau}^2}$	Large positive values
R_i statistic for relative treatment estimates	$R(\tilde{\mu}_i)_i = 100 \times \frac{\widetilde{\boldsymbol{\mu}}_i - \widetilde{\boldsymbol{\mu}}_{i(-i)}}{\widetilde{\boldsymbol{\mu}}_i}$	Large positive values
R_i statistic for generalized Cochran's Q (Q^{total})	$R(Q^{total})_i = 100 \times \frac{Q^{total} - Q^{total}_{(-i)}}{Q^{total}}$	Large positive values
R_i statistic for Q statistic within designs (Q^{het})	$R(Q^{het})_i = 100 \times \frac{Q^{het} - Q^{het}_{(-i)}}{Q^{het}}$	Large positive values
R_i statistic for Q statistic between designs (Q^{inc})	$R(Q^{inc})_i = 100 \times \frac{Q^{inc} - Q^{inc}_{(-i)}}{Q^{inc}}$	Large positive values
$DFBETAS_i$ statistic	$DFBETAS_i = (\widetilde{\boldsymbol{\mu}}_i - \widetilde{\boldsymbol{\mu}}_{i(-i)}) \sqrt{\frac{1}{n_{S_i}} \sum_{k \in S_i} w_{i,k(-i)}}$	$DFBETAS_i > 1$ for small to medium datasets

4.4 Illustrative examples with outlier and influential diagnostics measures

4.4.1 Dataset comprises four interventions to aid smoking cessation

We performed an outlier detection analysis based on several proposed outlying and influential detection measures in the dataset comprises four interventions to aid smoking cessation [41], [42] (details for dataset is given provided in Chapter 2). Figure 4.1 indicates the contribution to the Q statistic (Mahalanobis distance) for each study computed with function `NMAoutlier_measures()` and plotted with `plot_NMAoutlier_measures()`. We monitored that study 3 has the largest contribution to the Q statistic with value 117.39 while the rest values ranged from 0.34 to 26.40. Hence, study 3 is potential an outlying study.

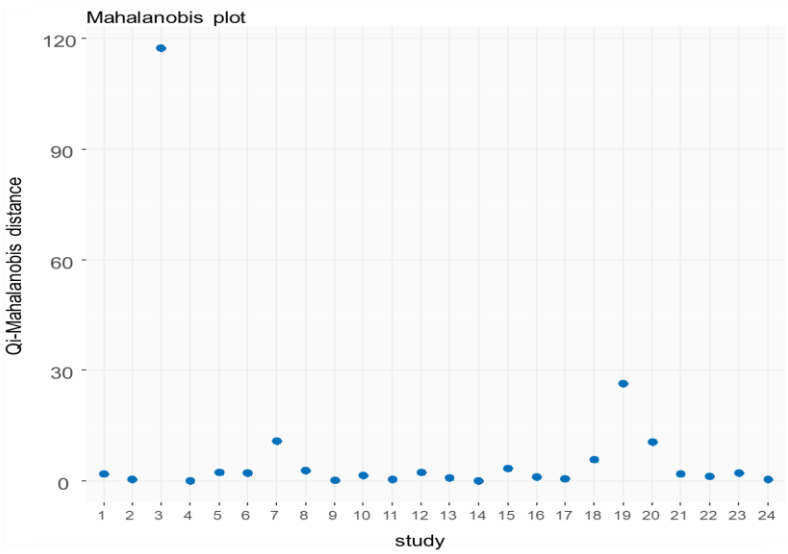

Figure 4.1. Contribution to the Q (Mahalanobis distance) for each study for the smoking cessation dataset.

Figure 4.2 shows the standardized study deleted residuals for each study. For all included studies in the network, standardized study deleted residuals values range inside the $(0, 2)$ interval except study 3 that is far away with a large value (4.14) and study 7 that is close to the boundary of two with value (2.11) (Figure 4.2).

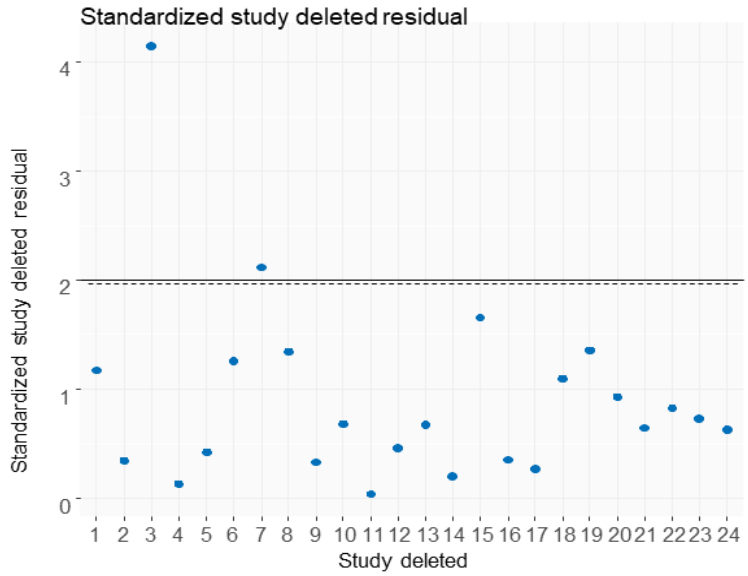

Figure 4.2. Standardized study deleted residuals for the smoking cessation dataset.

Figure 4.3 depicts that study 3 has the largest Cook's distance (value 1.51) and is the only study that exceeds the cut off value and satisfies $C_3 > 1, i = 3$. All the included studies have

$COVRATIO_3, i = 3$, close to 1 or larger but study 3 is the only one that provides the smallest $COVRATIO_3$ with value 0.06 satisfying $COVRATIO_3 < 1$ (Figure 4.3, right-hand side).

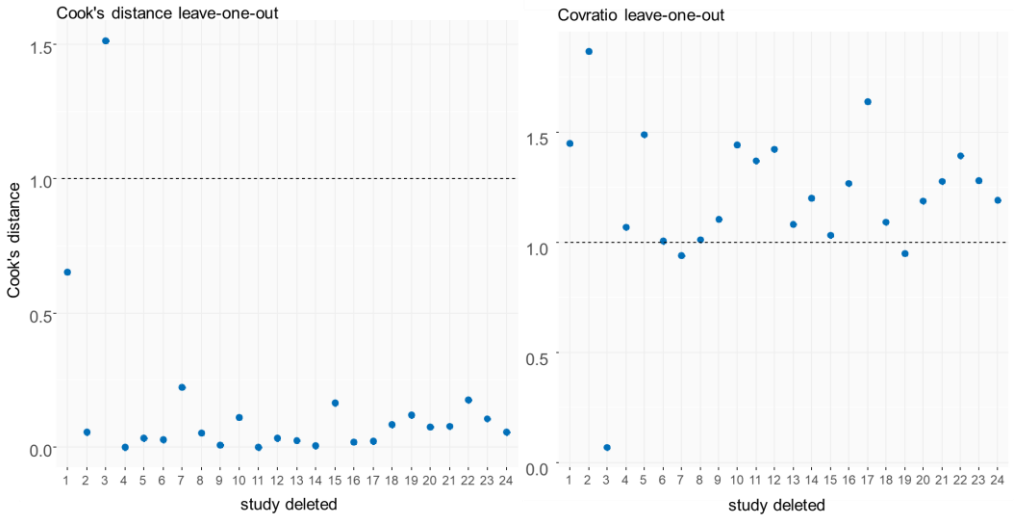

Figure 4.3. Cook distance (left-hand side) and COVRATIO (right-hand side) for the smoking cessation dataset.

Leaving the study 3 out of the NMA model fitting, "leave-one-out" model parameters providing a large change. Study 3 provides a large impact in model parameters as its deletion creates a large change to "leave-one-out" model parameters. We monitored the heterogeneity "leave-one-out" for study 3 to be dramatically decreased affecting the weight "leave one out" to be increased and the relative treatment effects to be influenced (Figure 4.4).

Figure 4.4. Heterogeneity (left-hand side) and weight "leave-one-out" (right-hand side) for smoking cessation dataset.

Figure 4.5 shows the $DFBETAS$ statistic for each treatment considering the deletion of a study. $DFBETAS_3, i = 3$ statistic (deletion of study 3) has the largest change for each treatment B, C, and D.

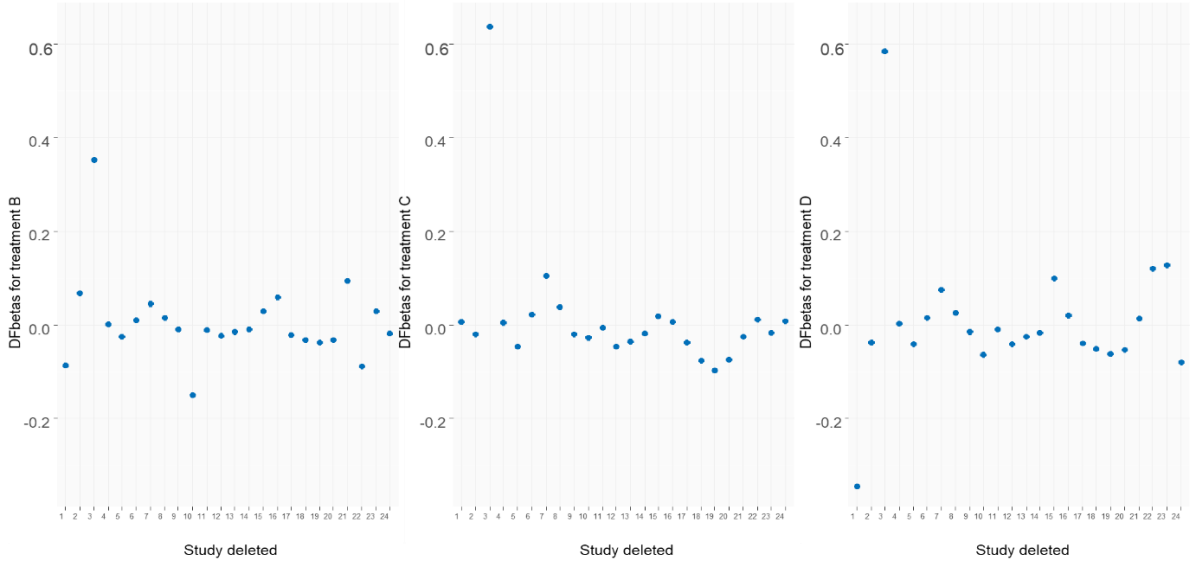

Figure 4.5. $DFBETAS$ statistic for each treatment considering deletion of a study for smoking cessation dataset.

Table 4.4 provides the changes in R_i statistic $R(\hat{\tau}^2)$ for heterogeneity estimator $\hat{\tau}^2$, total Q statistic Q^{total}, Q statistic within designs Q^{het} and between designs Q^{inc}. We monitored large changes in R_i statistic for Q^{total}, Q^{het} and $R(\hat{\tau}^2)$ as values for $R_i, i = 3$ statistic has been increased ($R_3(Q^{total}) = 69.79,\ R_3(Q^{het}) = 75.89,\ R_3(\hat{\tau}^2) = 74.20$). The deletion of study 3 reduced heterogeneity. Large changes provided also in R_i statistic for Q^{inc} with a large decreased ($R_3\left(Q^{inc}\right) = -105.29$) can provide us that the deletion of study 3 increased the inconsistency. Hence, "leave one out" measures providing large changes when removing study 3 from NMA model fitting. Based on several outliers and influential detection measures conducted, we can conclude that study 3 is an influential study and outlier.

Table 4.4 "Leave one out" detection measures for $R\left(Q^{total}\right), R\left(Q^{inc}\right), R(Q^{het}), R(\hat{\tau}^2)$.

Study deletion	$R(Q^{total})$	$R(Q^{inc})$	$R(Q^{het})$	$R(\hat{\tau}^2)$
1	3.82	42.52	0.00	-0.95
2	0.57	3.74	0.00	-5.30
3	69.79	-105.29	75.89	74.20

4	0.03	0.57	0.03	-1.11
5	1.33	8.29	1.52	-14.42
6	1.03	-0.77	1.11	0.43
7	5.37	-4.53	5.77	4.59
8	1.43	-1.15	1.53	0.71
9	0.08	1.25	0.09	-1.85
10	1.76	4.52	1.42	-5.44
11	0.20	2.68	0.41	-4.37
12	1.36	7.21	1.55	-12.28
13	0.41	0.96	0.46	-1.14
14	0.02	2.91	0.03	-4.79
15	1.76	53.61	0.00	1.24
16	0.63	-1.06	0.90	-2.10
17	0.38	10.44	0.46	-19.02
18	3.01	0.97	3.31	-0.73
19	14.73	-2.21	16.24	5.51
20	5.87	3.11	6.50	-4.08
21	1.08	6.89	0.00	-1.79
22	0.84	10.14	0.00	-1.64
23	1.40	-3.94	0.92	-1.08
24	0.22	-5.12	0.92	-1.08

4.4.2 Dataset with thrombolytic drugs

We applied the proposed outlier detection measures in a network of interventions of eight thrombolytic drugs for acute myocardial infarction [44] provided in Chapter 2. We selected this dataset partly because inconsistency has been detected by others due to studies 22 and 23 and partly because Zhao *et al.* [87] have provided an outlying diagnosis and concluded that the above studies are indeed outliers. We are interested to investigate if outliers are responsible for this inconsistency and if this can be detected with our proposed detection measures.

Figure 4.6 depicts that studies 22 and 23 have the largest values in the contribution to the Q statistic (Mahalanobis distance) plot. Study 22 has the largest contribution to the Q statistic followed by study 23. Based on our detection outlier analysis, we can conclude that studies 22 and 23 are indeed outliers which comes in agreement with Zhao *et al.* [87].

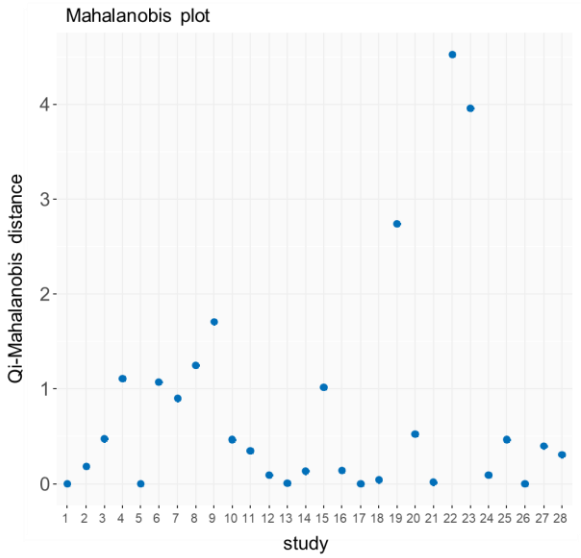

Figure 4.6. Contribution to the Q for each study in the thrombolytic drugs dataset.

4.5 Discussion

We propose and extend several measures and visual tools to detect outlier and influential cases in NMA. The outlier and influential diagnostics presented are logical extensions from regression or pairwise meta-analysis to the NMA context. Several graphical methods provided in this Chapter can be used to figure out studies that are far away for the rest of the data and to find the studies that are responsible for large changes in model parameters, heterogeneity and inconsistency measures.

Conclusions for outlying and influential cases can be made if there are sharp changes and/or if the proposed cut-offs no satisfied. Following bibliographic recommendations, several cut-offs for the detection measures are provided but this should not strictly be used as this offered as empirical rules to make conclusions for outlyingness and influential cases. For example, Viechtbauer and Cheung [85] provided values 1.96 and 2 for the absolute studentized residuals while Noma *et al.* [88] following a parametric bootstrap method to obtain the sampling distribution for studentized residual. There is no subjective rule in the diagnosis of outlyingness as conclusions made due to sharp changes or empirical cut-offs in proposed measures. The proposed measures and visual tools can detect and visualize extreme study effects that are outliers and influential studies and studies responsible for heterogeneity and inconsistency existence.

Meta-analytical models are sensitive to outliers and identification of such cases needs further investigation. Deletion of studies should not be routinely done in meta-analyses as this may be

problematic to omit nonoutlier extreme study effects due to their large sampling errors [109]. Following the recommendation in systematic reviews, potential outliers can be excluded in sensitivity analyses to explore the robustness of results [110], [85]. Results should be stated with caution when outlier analysis provides different conclusions than the primary analysis [85]. If outlier analysis does not alter the results, we can be more confident that conclusions are robust to outliers [85]. In the smoking cessation example, there is enough evidence that studies 3 and 7 are potential outliers. However, results from these studies can be valid and genuine. It could be the case that characteristics of these trials may explain the differences in observed effect sizes. Generally, caution is needed in deleting outliers and hereby extension in methodological aspects for downweighing the effects of influential and outlying studies may be an alternative choice for future work and motivate us to provide the proposed research provided in Chapter 6.

The proposed measures fitted in the frequentist framework with NMA model from graph theory but they can also be implemented in Bayesian framework. To date, Zhang *et al.* [86] have provided some outlier detection measures for the Bayesian NMA model. Matsushima *et al.* fitted also some measures to detect outliers in the meta-analysis of diagnostic accuracy studies [111]. The proposed measures and visual tools in this Chapter can also be extended in the meta-analysis of diagnostics accuracy studies or in meta-analysis with individual participant data.

5 Forward Search Algorithm to detect outliers in network meta-analysis

5.1 Introduction

The Forward Search (FS) algorithm was initially developed for the estimation of covariance matrices [112] and regression models [113], [114]. It was subsequently extended for outlier detection to multivariate data methods [115], factor analysis and item response theory models [116], [117]. Mavridis *et al*. recently extended the FS algorithm in meta-regression [101]. This algorithm starts with an initial subset of studies that is ideally assumed to be outlier-free and it gradually adds the remaining studies according to their closeness to the set of selected studies under the hypothesized model. Parameter estimates, measures of fit and test statistics can be monitored during the search. During the search, sharp changes denote the existence of influential studies and/or outliers and can detect studies responsible for heterogeneity and inconsistency existence.

This Chapter provides the methodological extension of the FS algorithm in network meta-analysis. As part of this research, the R package **NMAoutlier** [28] was developed to perform FS to any NMA evidence datasets. The structure of this Chapter is organized as follows: Section 5.2 outlines the methodological extension of the FS algorithm to NMA model; Section 5.3 presents applications of the proposed FS methodology in real and simulated datasets of networks of interventions and Section 5.4 discusses the main findings and provides directions of the proposed diagnostic methodology in NMA applications.

5.2 Methodological extension of the forward search algorithm in NMA

The FS algorithm is a diagnostic iterative method for outlier detection. It starts with a small subset of the included studies that is ideally considered to be outlier-free. The initial subset of studies constitutes the *basic set*. The studies not included in this basic set constitute the *non-basic set*. These two sets are not constant throughout the search but they are continuously

changed. In each iteration, the method adds a study from the non-basic set to the basic set. The study from the non-basic set was chosen based on how close the study is to the hypothesized model fitted to the basic set. The process is repeated until all studies are included in the basic set. We monitored model parameter estimates and other statistics of interest during the search. Monitoring is helpful to identify the studies that have an impact on model parameters or/and in statistics measurements and are responsible for heterogeneity and inconsistency. We categorized the FS procedure in three steps: (1) the choice of the initial subset; (2) the processing of search and (3) monitoring. The steps in relation to NMA and details are presented as follow:

5.2.1 Choice of the initial subset

The choice of the initial subset is the first crucial point of the FS. Network meta-analysis is a regression model with the number of treatments defining the number of columns of the design matrix. During the search, the number of columns of the design matrix may increase if a new treatment is added to the basic set. Hence, when contrasting the basic to the non-basic set we may compare different models. Also, we need to make sure that there is a path between each pair of vertices in the network. In a nutshell, the requirements for the initial subset are:

- to include all n treatments, otherwise, the design matrix **X** will not have the same number of columns throughout the search and
- the network to be connected.

For the choice of the initial subset, we need to define how to select the size of the initial subset and how to select the studies that constitute the initial subset.

5.2.1.1 Selecting the size of the initial subset

In a network meta-analysis with n treatments, the number of model parameters to estimate is n ($n-1$ effect estimates and heterogeneity – assumed equal heterogeneity estimator across treatment comparisons). Also, a minimum of $n-1$ two-arm studies is necessary to create a connected network graph with n treatments (nodes). We require the size l of the initial subset to include all n treatments, $l = n$ studies.

For networks with a large number of trials, a bigger size l of the initial subset can be considered to save time and allow the search to start with a more robust initial subset. We chose to start with a size equal to the maximum between the number of treatments and the 20% of the total number of studies, $l = \max(n\ studies, 0.2 \times N\ studies)$ aiming to have better parameter estimation in the early iterations of the search.

5.2.1.2 Selecting the studies to include in the initial subset

Search a large number (P) of candidate initial subsets of size l (e.g. $P = 100$). Ideally, the initial subset is clean of outliers. Following the typical strategy of literature in the FS algorithm [113], we fit an objective function to each candidate's initial subset. We choose to minimize the median of the absolute standardized residuals. We can also assume other objective functions such as maximizing the median of the absolute log-likelihood contributions. The candidate subset that optimizes the objective function (minimize the median of standardized residuals or maximize the median of the absolute log-likelihood contributions) is considered to be the initial subset.

Let us denote with D_p^l each candidate initial subset $p = 1, .., P$ of l studies, then we obtain the subset-specific estimates $(\widehat{\boldsymbol{\mu}}_{D_p^l}, \hat{\tau}^2_{D_p^l})$ of each subset D_p^l and we calculate the objective function $f\left(\boldsymbol{y}_i, \boldsymbol{s}_i, \boldsymbol{X}_i, \widehat{\boldsymbol{\mu}}_{D_p^l}, \hat{\tau}^2_{D_p^l}\right)$ with observations $\boldsymbol{y}_i, \boldsymbol{s}_i, \boldsymbol{X}_i, i = 1, \ldots., N$ of all entire data set. Objective functions of the median of the absolute standardized residuals and of the median of the absolute log-likelihood contributions are provided with the equations (1) and (2), respectively:

$$f\left(\boldsymbol{y_i}, \boldsymbol{s_i}, \boldsymbol{X_i}, \widehat{\boldsymbol{\mu}}_{D_p^l}, \hat{\tau}^2_{D_p^l}\right) = median\left(\left|\hat{\varepsilon}^{\,k,study,stand}_{i,D_p^l}\right|\right) \qquad (1)$$

$$f\left(\boldsymbol{y_i}, \boldsymbol{s_i}, \boldsymbol{X_i}, \widehat{\boldsymbol{\mu}}_{\boldsymbol{D_p^l}}, \hat{\tau}^2_{D_p^l}\right) = median\left(\left|-\log(\boldsymbol{w}_i) - \left(\hat{\varepsilon}^{\,k,study,stand}_{i,D_p^l}\right)^2\right|\right) \qquad (2)$$

where $\boldsymbol{w_i} = 1/\left(\boldsymbol{s_i^2} + \hat{\tau}^2_{D_p^l}\right)$ and $\hat{\varepsilon}^{\,k,study,stand}_{i,D_p^l}$ the standardized residual of study i as introduced in Chapter 4 by replacing $\widehat{\boldsymbol{\mu}}$ with $\widehat{\boldsymbol{\mu}}_{\boldsymbol{D_p^l}}$ and $\hat{\tau}^2$ with $\hat{\tau}^2_{D_p^l}$; $\hat{\varepsilon}^{\,k,study,stand}_{i,D_p^l} = \sqrt{\frac{1}{n_{S_i}}\sum_{k\in S_i}\left(\frac{\boldsymbol{y}_i - \boldsymbol{X}_i\,\widehat{\boldsymbol{\mu}}_{\boldsymbol{D_p^l}}}{\sqrt{\boldsymbol{s_i^2} + \hat{\tau}^2_{D_p^l}}}\right)^2}$

5.2.2 Processing in the search

Let us denote the initial basic set (for $j = 1$) with D^l and the complementary non-basic set as $(D^l)^c$. For each study in non-basic set $\boldsymbol{y}_i, \boldsymbol{s}_i, \boldsymbol{X}_i \in (D^l)^c$, calculate the objective function $f\left(\boldsymbol{y}_i, \boldsymbol{s}_i, \boldsymbol{X}_i, \widehat{\boldsymbol{\mu}}_{D^l}, \hat{\tau}^2_{D^l}\right)$ that measures the closeness between D^l and $(D^l)^c$, where $\widehat{\boldsymbol{\mu}}_{D^l}, \hat{\tau}^2_{D^l}$ are estimated from the basic set D^l. The study lies closer to the basic set is the next to enter it. Bear

in mind that if a S_i - arm study enters the FS algorithm, all $\binom{S_i}{2}$ possible treatment comparisons enter at once.

Proceed with the algorithm for $j = 2, \ldots . N - l$ iterations until all studies are included in the basic set. For each iteration define the basic set D_j^l and the non-basic $\left(D_j^l\right)^c$. Compute the objective function $f\left(\boldsymbol{y}_i, \boldsymbol{s}_i, \boldsymbol{X}_i, \widehat{\boldsymbol{\mu}}_{D_j^l}, \hat{\tau}^2_{D_j^l}\right)$, where $(\widehat{\boldsymbol{\mu}}_{D_j^l}, \hat{\tau}^2_{D_j^l})$ are subset-specific estimates for the basic set D_j^l with observations $\boldsymbol{y}_i, \boldsymbol{s}_i, \boldsymbol{X}_i \epsilon \left(D_j^l\right)^c$. Then, re-define the basic set and the non-basic, order studies during the FS procedure by monitoring parameter estimates, outlier and influential diagnostics, ranking measures, heterogeneity and inconsistency measures (as described in Section 4.2.3). Forward plots for statistical measures are developed during the procedure aiming to monitor changes in the statistical measures. Table 5.1 summarizes the notation for the methodology of the FS algorithm in NMA.

Table 5.1. Notation for the methodology of the FS algorithm in NMA.

FS algorithm notation:
l the size of the initial subset
P a large number of candidate initial subsets of size l (e.g. $P = 100$)
$p = 1, .., P$ each candidate initial subset of l studies
$j = 1, \ldots, N - l$ each iteration of the FS algorithm
Steps of FS algorithm:
For the initial subset:
D_p^l each candidate initial subset p of l studies
$(\widehat{\boldsymbol{\mu}}_{D_p^l}, \hat{\tau}^2_{D_p^l})$ subset-specific estimates of each subset D_p^l
$f\left(\boldsymbol{y}_i, \boldsymbol{s}_i, \boldsymbol{X}_i, \widehat{\boldsymbol{\mu}}_{D_p^l}, \hat{\tau}^2_{D_p^l}\right)$ objective function with observations $\boldsymbol{y}_i, \boldsymbol{s}_i, \boldsymbol{X}_i$ of all entire data set.
For the first iteration $j = 1$:
D_1^l initial basic set, $\left(D_1^l\right)^c$ non-basic set
$(\widehat{\boldsymbol{\mu}}_{D_1^l}, \hat{\tau}^2_{D_1^l})$ subset-specific estimates for the initial basic set D_1^l
$f\left(\boldsymbol{y}_i, \boldsymbol{s}_i, \boldsymbol{X}_i, \widehat{\boldsymbol{\mu}}_{D_1^l}, \hat{\tau}^2_{D_1^l}\right)$ objective function with observations $\boldsymbol{y}_i, \boldsymbol{s}_i, \boldsymbol{X}_i \epsilon \left(D_1^l\right)^c$.
For iterations $j = 2, \ldots . N - l$*:*
D_j^l basic set, $\left(D_j^l\right)^c$ non-basic set

$(\widehat{\boldsymbol{\mu}}_{D_j^l}, \hat{\tau}^2_{D_j^l})$ subset-specific estimates for the basic set D_j^l

$f\left(\boldsymbol{y}_i, \boldsymbol{s}_i, \boldsymbol{X}_i, \widehat{\boldsymbol{\mu}}_{D_j^l}, \hat{\tau}^2_{D_j^l}\right)$ objective function with observations $\boldsymbol{y}_i, \boldsymbol{s}_i, \boldsymbol{X}_i \in \left(D_j^l\right)^c$.

5.2.3 Monitor the search

5.2.3.1 Outlier and influential case diagnostics measures

During the forward search algorithm, several outlier and influential case diagnostics measures are monitored. Standardized residual for i study $\hat{\varepsilon}_i^{k,study,stand}$ (provided in Chapter 4) is calculated in each iteration. We monitored the Cook's statistic for NMA introduced in Chapter 4 at $j-1$ iteration to j iteration by replacing $\widetilde{\boldsymbol{\mu}}_{\boldsymbol{i}}$ with the relative treatment estimates $\widetilde{\boldsymbol{\mu}}_{D_j^l}$ at D_j^l basic set and the relative treatment estimates after considered deletion of study $\widetilde{\boldsymbol{\mu}}_{i(-i)}$ with the relative treatment estimates at the basic set D_{j-1}^l, $\tilde{\mu}_{D_{j-1}^l}$. For Cook's statistic, we denote the $(n-1)\times(n-1)$ variance-covariance matrix $\widetilde{\mathbf{X}}_{D_j^l}\mathbf{L}^+_{D_j^l}\widetilde{\mathbf{X}}'_{D_j^l}$ of $(n-1)$ relative treatment estimates $\tilde{\mu}_{D_j^l}$ compared with the reference. The ratio of determinants of the variance-covariance matrix of treatment estimates (provided in Chapter 4) was also calculated at iteration j to iteration $j-1$ denoting the variance-covariance matrix with $\widetilde{\mathbf{X}}_{D_j^l}\mathbf{L}^+_{D_j^l}\widetilde{\mathbf{X}}'_{D_j^l}$ and $\widetilde{\mathbf{X}}_{D_{j-1}^l}\mathbf{L}^+_{D_{j-1}^l}\widetilde{\mathbf{X}}'_{D_{j-1}^l}$ for basic sets D_j^l and D_{j-1}^l respectively.

5.2.3.2 Ranking measures

We monitored P-scores [7] to identify potential studies that influence the ranking of treatments. The P-score (P_A) of a treatment A can be interpreted as the proportion of treatments inferior to treatment A. That is, a large P-score indicates a good treatment option for treatment A instead of the other competing treatments.

5.2.3.3 Heterogeneity and inconsistency measures

Large heterogeneity will challenge the interpretation of summary results whereas inconsistency may lead to biased treatment estimates. During the FS procedure, we calculated the generalized Cochran's Q (Q^{total}), Q statistic within designs (Q^{het}) and Q statistic between-designs with full design-by-treatment interaction model (Q^{inc}) [39] as provided in Chapter 2. The assumption of consistency can be tested by comparison of direct and indirect evidence [118]. König *et al.* introduced a back-calculation method to derive indirect estimates from direct

pairwise comparisons and network estimates [119]. We monitored the $z-values$ of disagreement between direct and indirect evidence for each comparison.

5.3 Illustrative examples

5.3.1 Artificial simulated outlier

For illustration, we simulated a single NMA data set with four treatments (A, B, C, and D) and eight pairwise comparisons (studies), $i = 1, \ldots.8$. Treatment A is chosen as reference treatment, the true effects are chosen with values $\mu_{AB} = 0.3, \mu_{AC} = 0.4, \mu_{AD} = 0.5$ and the between-study variance is $\tau^2 = 0.1^2$. Following Kontopantelis and Reeves [120] and Brockwell and Gordon [121], we generated the study variances from ${\sigma_i}^2 \sim X_1^2/4$, $i = 1, \ldots.8$ with values restricted to the interval $(0.009, 0.6)$. Seven effect sizes are generated from $y_{i,t1t2} \sim N(\mu_{t1t2}, {\sigma_i}^2 + \tau^2), i = 1, \ldots, 7$ where $t1t2$ are the treatment comparisons $t1t2 = (AB, AC, BC, BD, AD, CD, CD)$, μ_{t1t2} the true effects generated from consistency equations for $t1t2 \neq (AB, AC, AD)$, i.e $\mu_{BC} = \mu_{AC} - \mu_{AB}$. Following Filzmoser [122], Knight and Wang [123], and Hardin and Rocke [124], we artificially generated a *shift outlier* for the eighth study $(i = 8)$ with treatment comparison C versus D and with observed effect size to follow the formula $y_{8,CD} \sim N(\mu_{CD} + 4\mathrm{SD}(y), {\sigma_8}^2 + \tau^2)$, where $\mathrm{SD}(y)$ is the sample standard deviation of values $y = (y_{1,AB}, \ldots\ldots, y_{7,CD})$.

Study 8 provides a markedly different intervention effect compared to the rest of the simulated data (Table 5.2). The artificial simulated dataset is provided in Table 5.2. We used R function `NMAoutlier` from R package **NMAoutlier** [28] with the criterion of the smallest absolute standardized residuals; see equation (1). We considered $P = 100$ candidate initial subsets and the size l to be equal with the number of treatments, $l = max(4, 0.2 \times 8) = 4$ studies. The subset with studies 3, 5, 1 and 7 minimized the criterion (the median of the absolute standardized residuals) among the candidate subsets and set as initial set. Then, forward search gradually added the study that minimized the median of the absolute standardized residuals, until all studies entered after four iterations. FS algorithm was completed after a total of five iterations.

Table 5.2. Effect size $\boldsymbol{y_i}$, standard error $\boldsymbol{s_i}$ and treatment comparisons for each study of the artificial simulated dataset.

Effect size y_i	Standard error s_i	treat1	treat2	study label
-0.0820	0.5091	A	B	1
0.3198	0.0125	A	C	2
0.2171	0.2437	B	C	3
0.2100	0.0153	B	D	4
0.4926	0.1928	A	D	5
-0.8612	0.4800	C	D	6
0.4115	0.1007	C	D	7
2.7639	0.4604	C	D	8

Figure 5.1 provides the forward plot of standardized residuals for each iteration produced with `fwdplot` from R package **NMAoutlier** [28]. Study 8 entered last in the forward selection procedure. Figure 5.1 shows that study 8 has a very large standardized residual in comparison with other studies and, thus, clearly detected as outlying.

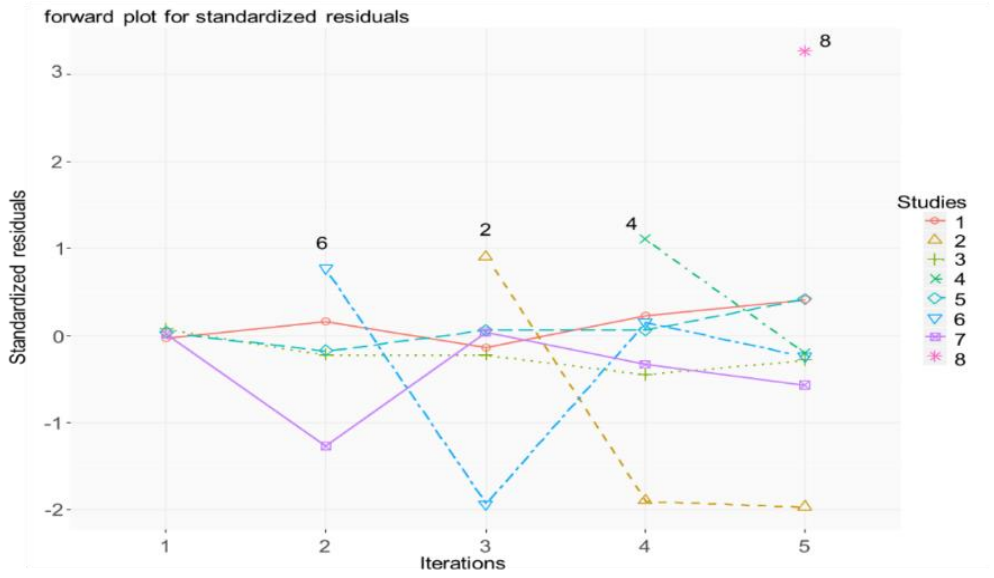

Figure 5.1. Forward plot for standardized residuals of basic set in each iteration of the FS algorithm.

5.3.2 Dataset comprises four interventions to aid smoking cessation

5.3.2.1 Assessing outlying cases

Figure 5.2 provides the comparison-adjusted funnel plot [78] for the smoking cessation data by choosing the order from least effective to most effective treatment, (1) no contact (A), (2) self-help (B), (3) group counseling (D) and (4) individual counseling (C). We can see that study 3 lies far away from the bulk of the data. This is because study 3 has a large effect size given its size. We conducted the FS algorithm starting with $P = 100$ candidate initial subsets of size

$l = 5$ and used the criterion of the smallest absolute residual. The FS steps were completed in 27 seconds. Table 5.3 summarizes the studies constitute the initial set and which study entered each iteration of the FS algorithm but also heterogeneity and inconsistency measures. We noticed that study 3 entered in the last iteration.

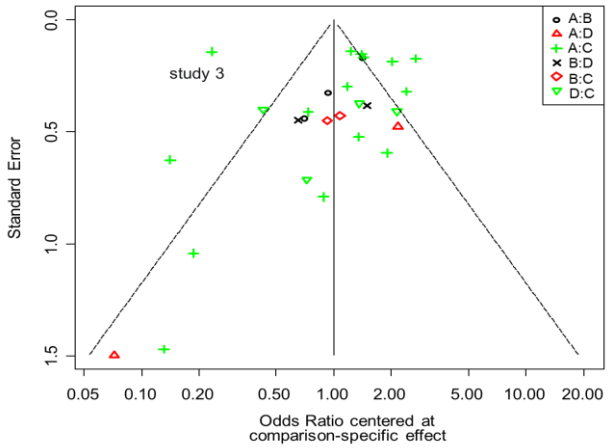

Figure 5.2. Comparison-adjusted funnel plot [78] for smoking cessation data. Comparison-adjusted funnel plot produced in R [76] from **netmeta** package [35]. $y - axis$ provides the standard error and the $x - axis$ provides the odds ratio centered at comparison-specific effect.

Table 5.3. Initial set and **s**tudy entered into the basic set of FS algorithm. Q statistics ($Q^{total}, Q^{inc}, Q^{het}$) and heterogeneity estimator $\hat{\tau}^2$ for each iteration of the FS algorithm. Results are given from **NMAoutlier** [35] package.

iterations	Study entering	Q^{total}	Q^{inc}	Q^{het}	$\hat{\tau}^2$
1	9, 14, 17, 11, 15 (initial set)	0.06	0.00	0.06	0.00
2	4	0.22	0.00	0.22	0.00
3	16	0.50	0.00	0.50	0.00
4	21	0.65	0.16	0.50	0.00
5	13	1.16	0.13	1.03	0.00
6	5	1.46	0.08	1.38	0.00
7	12	1.69	0.06	1.64	0.00
8	6	4.17	0.06	4.11	0.00
9	18	7.42	0.03	7.38	0.00
10	8	11.05	0.04	11.01	0.00

11	20	15.03	0.00	15.02	0.02
12	10	18.46	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	53.45	6.84	43.42	0.16
16	24	53.45	6.85	43.42	0.15
17	2	55.40	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

Confidence intervals of summary estimates for the treatments B and C broaden in the last iteration (Figure 5.3) because the between-study heterogeneity estimator increased substantially in this iteration (Table 5.3). As the forward plot (Figure 5.4, right panel) shows, the ratio of variances increased rapidly in the last iteration. We monitored a dramatic increase for heterogeneity estimator, Q^{het} and Q^{net} but a reduction for Q^{inc} in this final iteration (Table 5.3). Thus, inconsistency in the whole network is masked due to the large heterogeneity.

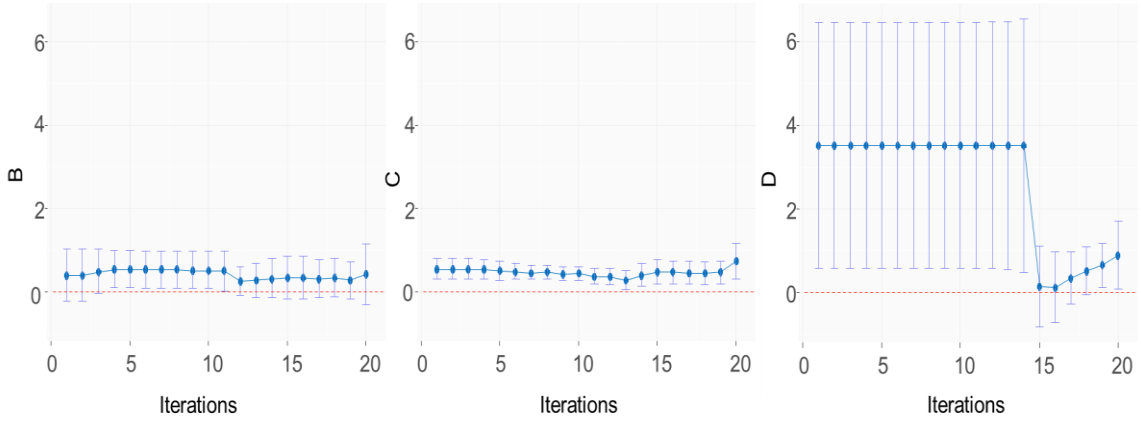

Figure 5.3. Forward plots for summary estimates and their 95% confidence intervals for each treatment B, C, D versus control A. Forward plots produced in R [76] with function `forwardest()` from **NMAoutlier** package [28].

In conclusion, study 3 is considered an outlier and influential study. Study 3 entered the last FS iteration, produced sharp changes in outlying measures and influenced the model parameters (heterogeneity and summary estimates). Table 5.3 provides that the whole network has large values for heterogeneity estimator $\hat{\tau}^2 = 0.59$ and $Q^{het} = 187.40$ (last iteration 20 of

the FS algorithm). Before adding study 3, at iteration 19, heterogeneity measures are decreased with values $\hat{\tau}^2 = 0.15, Q^{het} = 45.17$ (Table 5.3). By entering of study 3, we monitored the 95% confidence intervals of summary estimates to broaden (Figure 5.4). Inclusion of study 3 provide different summary estimates with values (last iteration 18 of FS) $\hat{\mu}_B = 0.42\ (-0.30, 1.13)$, $\hat{\mu}_C = 0.73\ (0.30, 1.16)$, $\hat{\mu}_D = 0.90\ (0.09, 1.70)$ in comparison with summary estimates at iteration 19, $\hat{\mu}_B = 0.27\ (-0.17, 0.71), \hat{\mu}_C = 0.47(0.19, 0.73)$, $\hat{\mu}_D = 0.65\ (0.12, 1.19)$. Moreover, study 7, a study that compares A versus C treatments, provides closer effect estimate with study 3 that the rest studies with treatment comparisons A versus C. It enters at iteration 14 and occurring an increase in heterogeneity but also a sharp change in Cook distance (Figure 5.4 and Table 5.3).

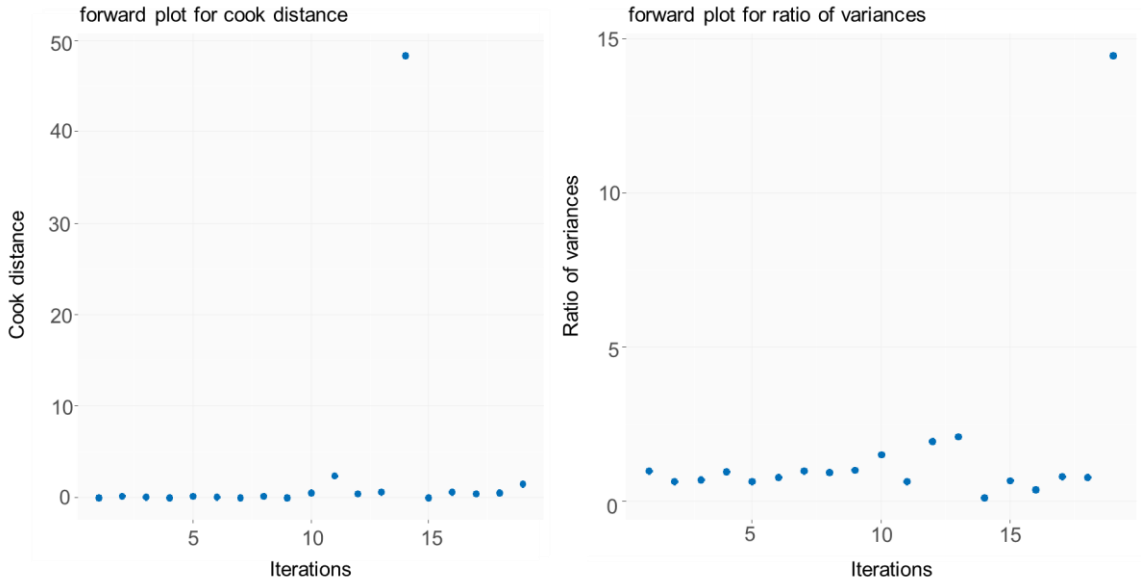

Figure 5.4. Forward plots for cook distance (left) and the ratio of variances (right) for each iteration of the FS algorithm. Forward plots are provided by function `fwdplot()` in package **NMAoutlier** [28].

Repeating the process of FS algorithm for smoking cessation data for 100 times, conclusions about the robustness of study ordering indicates that study 3 entered in the last FS iteration in 82 times, in the initial set 15 times and 3 times at an intermediate iteration.

5.3.2.2 Assessing inconsistency in smoking cessation data

Higgins *et al.* [40] initially applied the full design-by-treatment interaction model to the smoking cessation data. For the whole network, the full interaction model does not provide evidence for inconsistency ($Q^{inc} = 4.66, p = 0.7$). Monitoring the full interaction model in each iteration of the FS algorithm, we noticed a sharp increase in Q^{inc} when study 1 entered at

iteration 15 (Table 5.3). Searching for local inconsistency, entering of study 1 provides changes of differences between direct and indirect comparisons. Forward plot of $z - values$ (Figure 5.5) depicts that at iteration 15, differences between direct and indirect evidence are large for 'A versus D' and 'C versus D' ($z_{A\ versus\ D} = 1.50, z_{C\ versus\ D} = 2.20$). Study 1 is a triangle three-arm study with treatment arms A, C, and D. It was the first time that a study comparing interventions A, C and D enters the search. Conclusions from forward plot of $z - values$ but also changes of Q^{inc} show that study 1 is influential for the design inconsistency in 'A versus B' and 'A versus D' effect sizes between the two-arm and three-arm studies. We monitored also changes in inconsistency measures when the other three-arm, study 2 with treatment arms B, C and D, entered (iteration 17). This comes in agreement with the conclusion given by Higgins *et al.* [40] that there is a design inconsistency in effect sizes between two-arm and three-arm studies in smoking cessation data.

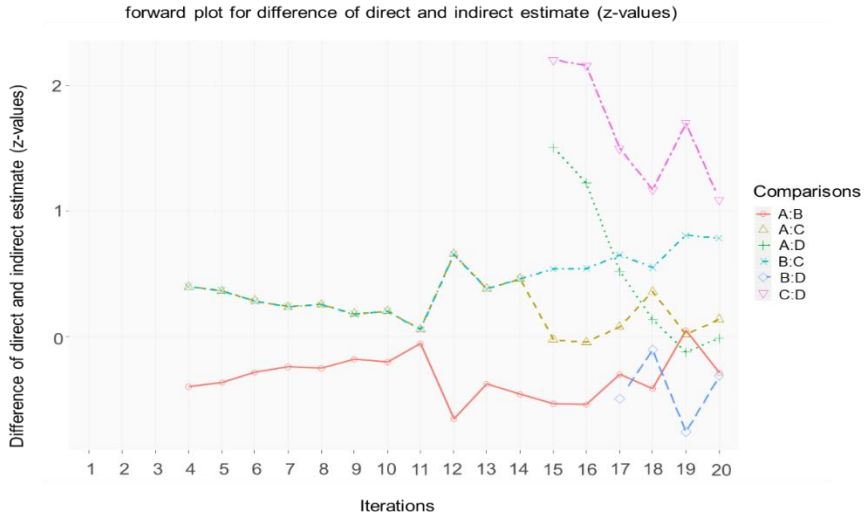

Figure 5.5. The forward plot of $z - values$ of disagreement between direct and indirect evidence for each comparison (back-calculation method) for each iteration of the FS algorithm. Forward plots are provided by function `fwdplot()` in package **NMAoutlier** [28].

5.3.3 Dataset comparing interventions for actinic keratosis

We identified statistically significant inconsistency in the whole dataset for actinic keratosis [43] using the design-by-treatment interaction model ($Q^{inc} = 23.05, df = 7, p = 0.001$). Table 5.4 provides the between-designs Q^{inc} statistic after detaching a single design for actinic keratosis dataset. The between-designs Q^{inc} statistic indicates that the dataset satisfies the consistency assumption only when the design 1 versus 6 versus 8 was detached ($Q^{inc} = 10.18, df = 5, p = 0.07$) (Table 5.4). Study 28 is the unique study with treatment design 1

versus 6 versus 8. We are interested to investigate if study 28 is a potential source of inconsistency. Therefore, we are interested to check whether study 28 is outlying or influential.

Table 5.4. Between-designs Q^{inc} statistic after detaching of single designs for actinic keratosis dataset.

Detached design	Between-designs Q statistic	Degrees of freedom (df)	p-value
1 versus 3	103.72	6	<0.0001
1 versus 4	86.80	6	<0.0001
1 versus 5	93.89	6	<0.0001
1 versus 6	98.61	6	<0.0001
3 versus 6	103.72	6	<0.0001
4 versus 5	101.94	6	<0.0001
4 versus 7	95.78	6	<0.0001
1 versus 5 versus 6	73.77	5	<0.0001
1 versus 6 versus 8	10.18	5	0.0704
4 versus 7 versus 8	73.52	5	<0.0001

We conducted the FS algorithm starting with $P = 100$ candidate initial subsets of size $l = 9$ using the criterion of the smallest absolute residual. Study 28 entered in the last step of the FS algorithm which can be the first indication that study 28 is outlying or influential. Fitting the full interaction model, we monitored a sharp increase in Q^{inc} statistic (from 3.68 to 23.05) when study 28 entered (iteration 27) (Figure 5.6). Thus, the FS algorithm confirms that study 28 is a source of inconsistency.

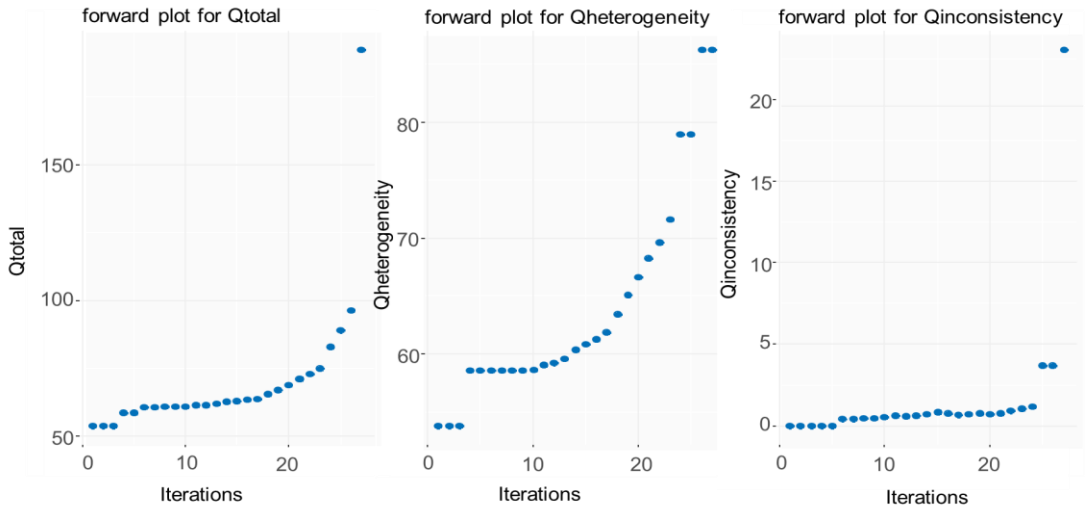

Figure 5.6. Forward plots for generalized Cochran's Q $\left(Q^{total}\right)$, Q statistic within designs (Q^{het}) Q statistic between-designs with full design-by-treatment interaction model Q^{inc}. Forward plots produced in R [76] with function `forwardplot()` from **NMAoutlier** package [28].

Sharp changes in the last step of FS in forward plots for Cook distance and the ratio of variances lead to the conclusion that study 28 is an outlier (Figure 5.7). After removing the outlying study 28, i.e., the source of inconsistency, from the dataset, the inconsistency problem was overcome (design-by-treatment interaction model, $Q = 3.68, df = 5, p = 0.59$).

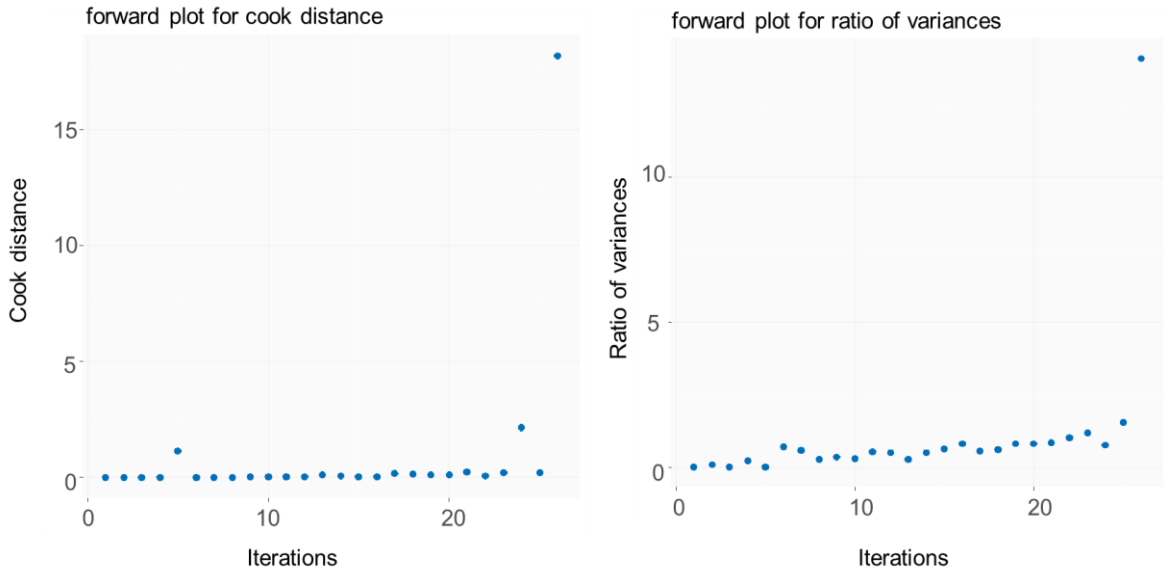

Figure 5.7. Forward plots for Cook distance (left) and ratio of variances (right). Forward plots produced in R [76] with function `forwardplot()` from **NMAoutlier** package [28].

5.4 Discussion

We propose a diagnostic method with the FS algorithm that detects studies having a disproportionate effect on summary effects, heterogeneity, and inconsistency in NMA. The novel tool allows us to identify outlying and influential studies on the basis of observing sharp changes in the chosen monitoring measures.

There are some limitations to the FS methodology in NMA. FS methodology can be a time-consuming method as it requires a lot of time for very large NMAs. FS application in the real data sets indicated that computation time increased in accordance with the total number of studies in the network (27 seconds for 24 studies with smoking cessation outcome in comparison with 59 seconds for 35 studies with actinic keratosis outcome). FS methodology

should be used with caution as different criteria to initiate and progress in the search may result in a different ordering of studies.

Even if the same criteria are selected, different ordering of studies can be provided, but sharp changes in monitoring measures conclude to the same conclusions. The initial subset is likely considered outlier-free but there is a possibility of outlier entering in the initial subset. For example, in smoking cessation data, we monitored 15 out of 100 times the entry of outlier (study 3) in the first FS iteration. Entering of an outlying study in the initial subset would be sure if this study is the only one with a specified treatment comparison in the network structure and therefore it will be entered in the initial subset due to the methodological requirements (connectivity of initial subset including all treatments) of FS procedure. Entering an outlier in the initial subset can cause abnormalities; that can be the larger heterogeneity estimator for the first iteration compared to other iterations.

The FS algorithm may be impaired by abnormalities in the search and for this reason, we suggest to rerun the forward search 5-10 times from random starting points to explore the robustness of the ordering. Moreover, if abnormalities are still provided in the initial subset, even if some repeats of forward search conducted, we advise monitoring the FS methodology in accordance with simple measures, such as residuals or contribution to the Q statistic for each study (see Chapter 4 for measure details).

Another issue is how to detect if a change in a statistic is due to the inclusion of an outlying study or can be attributed to random variation. For this reason, it has been suggested to accompany forward plots with simulation envelopes that set the boundaries of changes in fit statistics that can be attributed to random variation. These simulations envelopes can be generated from many forward searches applied to data sets generated using parameters, summary effects, and heterogeneity, equal to those observed in the dataset in question. Such a process is time-consuming and future research may find alternative methods of computing the acceptable changes in fit statistics that can be attributed to random variation.

The forward search algorithm is a promising diagnostic tool for extreme study effects which is unaffected from masking and swamping effects. Robustness of results indicates that we get reliable conclusions by using different criteria and there is the robustness of ordering the studies. The FS algorithm can be extended to the meta-analysis of diagnostic accuracy studies, the meta-analysis with multiple outcomes or individual patient data meta-analysis.

In conclusion, we argue that the method should be employed as a diagnostic tool and may reveal important information about the data. It is particularly useful for detecting studies responsible for heterogeneity and inconsistency.

6 Random shift variance NMA model for outlier identification

6.1 Introduction

Several outlier detection measures, such as deletion measures, allow a shift in the mean for a single observation (study result) known as a *mean shift*. An outlier is typically observed as an inflated (shifted) effect variance [97]. Cook *et al.* [125], based on *shift variance,* introduced an alternative approach to detect outliers in which each data observation was considered with inflated variance (shift variance) in the fixed-effect linear model. All parameters were estimated with the maximum likelihood estimation method [125]. Cook and Weisberg [107] used the term *variance shift* for this model. Thompson [126] considered the same model using restricted maximum likelihood estimation instead of maximum likelihood. Harville [127] and Thompson [126] recommended to use the restricted maximum estimation method rather than the maximum estimation method. Gumedze and Jackson [97] extended the variance shift model in a random-effects meta-analysis model for identifying and downweighing outlying studies. For brevity, we denote the Random-effects Variance Shift Outlier Model with RVSOM. The shift variance model initially allows the identification of outliers and if any study or studies are identified as outliers, the RVSOM model allows their downweighing [97].

In this Chapter, the extension of the RVSOM to detect outliers from pairwise meta-analysis to the NMA model is provided (RVSOM NMA). We provide the methodological challenges to extend the model in a network of interventions with the presence of multi-arm studies and we will discuss the technicalities of model fitting. As part of this research, R package **NMAoutlier** was developed for performing the RVSOM NMA model and offering visual tools with the proposed methodology to any NMA data. We implement the proposed outlier detection method to the smoking cessation dataset which was introduced in Chapter 2.

The Chapter is organized as follows: Section 6.2 outlines the RVSOM model by generalizing the method from pairwise to network meta-analysis model [97]. Section 6.3 presents an application of the RVSOM NMA model in published data of networks of interventions; Section

6.4 outlines the main findings and provides guidance and advice for the use of RVSOM outlier detection methodology in NMA.

6.2 Shift variance NMA model to downweigh outliers

We employed the shift variance model to the NMA model presented by Rücker and used in R library netmeta [5] (details of the model provided also in Chapter 2). The shift variance model assumes three sources of variance; the within-study variance, the between-study variance, and the shift variance. Within-study variances are data and the two other sources of variance, between and shift variance of each study, need to be estimated based on likelihood methods. The RVSOM model is fitted using the restricted maximum estimation method. Restricted maximum estimation is widely used to estimate the variance as it provides estimates with less downwards bias than maximum estimation. RVSOM outlier detection method is a repeatable procedure. RVSOM model is fitted by shifting the variance for each included study sequentially in NMA and therefore the number of RVSOM model fitting is equal to the number of studies. RVSOM provides an estimate of the shift in the error variance associated with that study [97]. A large shift may indicate a possible outlier and, if desired, can be downweighed [97].

6.2.1 Shift variance NMA model (RVSOM NMA)

The RVSOM model initially allows the identification of outliers [97]. It allows inflating variance for the study i. If the variance for a treatment effect y_i needs to be inflated, that means that the standard variance considerations of the typical random effects model are not enough to explain the study treatment effect and this study probably gives an outlying effect estimate. The model RVSOM NMA takes the form

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

which adds an extra term $\boldsymbol{u_i}$ compared to the standard NMA model. The term $\boldsymbol{u_i}$ is an unknown random effect with $\boldsymbol{u_i} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Omega_i})$ that adds the additional variance term associated. $\boldsymbol{\Omega_i}$ denotes a $m \times m$ block diagonal shift variance-covariance matrix with shift variance estimators $\widehat{\boldsymbol{w}_i}^2$ of shift variances $\boldsymbol{w_i}^2$ in its diagonal. For the study i, the vector $\boldsymbol{w_i}^2$ has length m and values $\boldsymbol{w_i}^2 \geq 0$, with zero value to denote a non-inflated study variance and with $\boldsymbol{w_i}^2 > 0$ to denote an inflated study variance in i^{th} position (or positions of pairwise comparisons in case of a S_i-

arm study, $i\,(1), \ldots.,\ i\left(\binom{S_i}{2}\right)$). For a multi-arm study, the variance was shifted for all pairwise comparisons $\binom{S_i}{2}$ within the study. The vector $\boldsymbol{w_i}^2 = \left(w_i{}^{2\,(1)}, \ldots, w_i{}^{2\,\binom{S_i}{2}}\right)'$ denotes the shift variance values for each pairwise comparison within a S_i-arm study. Assuming a common shift-variance within a study with d_i arms, the block diagonal shift variance-covariance sub-matrix $\boldsymbol{\Omega}_{S_i}$ with dimensions $\binom{S_i}{2} \times \binom{S_i}{2}$ of $\boldsymbol{\Omega_i}$ is given by

$$\boldsymbol{\Omega}_{S_i} = \begin{bmatrix} \widehat{w}_i^2 & \cdots & \frac{\widehat{w}_i^2}{2} \\ \vdots & \ddots & \vdots \\ \frac{\widehat{w}_i^2}{2} & \cdots & \widehat{w}_i^2 \end{bmatrix}_{\binom{S_i}{2} \times \binom{S_i}{2}}$$

The total variance for the study i in RVSOM NMA model is decomposed to the within-study variance $\boldsymbol{s}_i^2$, the shift variance $\boldsymbol{w_i}^2$ and the between-study heterogeneity τ^2. The weight matrix for the RVSOM model is $\mathbf{W}_{shift} = (\mathbf{S} + \boldsymbol{\Delta} + \boldsymbol{\Omega_i})^{-1}$. REML offers us estimates for the two unknown variance model parameters, the heterogeneity τ^2 and the shift variance $\boldsymbol{w_i}^2$ as well as summary estimates $\widehat{\boldsymbol{\mu}}_{\boldsymbol{shift}} = \mathbf{X}\left(\mathbf{X}'\mathbf{W}_{\boldsymbol{shift}}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W}_{\boldsymbol{shift}}\boldsymbol{y}$.

The restricted (residual) maximum log-likelihood function for RVSOM NMA model multiplied with $-\frac{1}{2}$ is given by

$$LR_{(i)}(\boldsymbol{y};\ \tau^2, \boldsymbol{w_i}^2)$$

$$= -\frac{1}{2} log(det|\mathbf{S} + \boldsymbol{\Delta} + \boldsymbol{\Omega_i}|)$$

$$-\frac{1}{2} log(det|\mathbf{X}'(\mathbf{S} + \boldsymbol{\Delta} + \boldsymbol{\Omega_i})^{-1}\mathbf{X}|)$$

$$-\frac{1}{2}\left(\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\mu}}_{\boldsymbol{shift}}\right)'(\mathbf{S} + \boldsymbol{\Delta} + \boldsymbol{\Omega_i})^{-1}\left(\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\mu}}_{\boldsymbol{shift}}\right)$$

$$= \frac{1}{2} log\left(det\left|\mathbf{W}_{\boldsymbol{shift}}\right|\right) - \frac{1}{2} log\left(det\left|\mathbf{X}'\mathbf{W}_{\boldsymbol{shift}}\mathbf{X}\right|\right)$$

$$-\frac{1}{2}\left(\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\mu}}_{\boldsymbol{shift}}\right)'\mathbf{W}_{\boldsymbol{shift}}\left(\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\mu}}_{\boldsymbol{shift}}\right)$$

where $Cov(\boldsymbol{y}) = \mathbf{W}_{\boldsymbol{shift}}^{-1} = \mathbf{S} + \boldsymbol{\Delta} + \boldsymbol{\Omega}_{\boldsymbol{i}}$ is the variance-covariance matrix for $\boldsymbol{y}$ under the RVSOM NMA model.

6.2.2 Monitoring measures

Downweighing an outlying study may provide large changes in summary estimates and heterogeneity. For each RVSOM model fitted, model parameters, heterogeneity, inconsistency measures, ranking measures and likelihood ratio test are monitored. Sharp changes in monitoring measures can be an indication for outlier existence. Plotting the changes of the several monitoring measures visually conveys the possibility of each study to have an inflated variance or else the possibility to be an outlier.

Summary estimates and their 95% confidence intervals for each treatment, REML heterogeneity estimator $\hat{\tau}^2$ and shift variance estimator $\widehat{\boldsymbol{w}_{\boldsymbol{i}}}^2$ are monitored. Standardized residual $\hat{\varepsilon}_i^{k,study,stand}$ from the RVSOM model for i study is calculated by replacing the standard NMA estimates with RVSOM NMA estimates in formula provided for standardized residual (see Chapter 4) and ranking measure with P-scores [7] are monitored.

We monitored the generalized total of Cochran's Q statistic Q^{total}, Cochran's Q statistic for heterogeneity Q^{het} and inconsistency Q^{inc} [39] as provided in Chapter 2 by replacing the standard NMA estimates with RVSOM NMA estimates. Having the treatment estimates $\widehat{\boldsymbol{\mu}}_{\boldsymbol{shift}}$ and the weight matrix $\mathbf{W}_{\boldsymbol{shift}}$ for RVSOM NMA model, the total Cochran's Q statistic is given by $Q^{total} = \left(\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\mu}}_{\boldsymbol{shift}}\right)' \mathbf{W}_{\boldsymbol{shift}} \left(\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\mu}}_{\boldsymbol{shift}}\right)$ [39],[128], [129] [130]. We also monitored the $z - values$ of disagreement between direct and indirect evidence for each comparison to derive the mixed estimates [118], [119].

6.2.3 Likelihood Ratio Test (LRT)

Likelihood ratio test (LRT) can be an objective measure to evaluate if a study i can be considered outlier and if downweighing is needed with the RVSOM model. Having the null hypothesis $H_0: \boldsymbol{w_i}^2 = 0$ against the alternative $H_a: \boldsymbol{w_i}^2 > 0$, the LRT_i statistic for study i is given by

$$LRT_i = 2\{LR(\boldsymbol{y};\, \tau^2) - LR_{(i)}(\boldsymbol{y};\, \tau^2, \boldsymbol{w_i}^2)\}$$

and can evaluate if standard NMA model (when shift variance for study i is zero) is fitted better for study i than the RVSOM NMA model (when shift variance for study i is larger than zero). If study i is an outlier, the RVSOM NMA model provides a better fit and, in this case, the null hypothesis is rejected. Gumedze and Jackson [97] proposed an empirical distribution for the likelihood ratio test using a parametric bootstrap procedure following the steps:

1) Under the assumption of no outlier exist, fit the standard NMA model and obtain the estimates $\widehat{\boldsymbol{\mu}}$ and $\hat{\tau}^2$.
2) Generate new simulated data with the standard NMA, $\boldsymbol{y}^*\sim\mathrm{N}(\mathbf{X}\widehat{\boldsymbol{\mu}}, \hat{\tau}^2\mathrm{I}_m + diag(\mathbf{S}))$, where m denotes the pairwise comparisons.
3) Compute the likelihood ratio test LRT_i for each study fitting the RVSOM NMA model with the simulated data $\boldsymbol{y}^*$ and save the order of LRT_i.
4) For each order statistic, repeat the steps 2) and 3) for a large number of replications R, (e.g. $R = 5000$).
5) For a level of significance a ($a = 0.05$) compute an $(1-a)\%$ percentile for each ordered likelihood statistic (e.g. the largest likelihood ratio test).

The empirical distribution of likelihood ratio test statistic can provide the threshold for identifying outliers with the LRT test under the null hypothesis that no outliers are present.

6.2.4 Extended RVSOM NMA

The RVSOM model can be extended to detect more than one outlier. Having a subset of $\Xi = \{1, 2, \dots, \xi\}$ studies considered as outliers, RVSOM model can be extended by allowing different inflated variances $\boldsymbol{w}_{\Xi}{}^2 = \left({w_1}^2, \dots, {w_\xi}^2\right)'$ for more than one study. Allowing for Ξ inflated variances of studies, the vector $\boldsymbol{w}_i{}^2$ has length m with values $\boldsymbol{w}_i{}^2 \geq 0$, where zero values denoting the studies with non-inflated variances ($m - \xi$ studies) and with $\boldsymbol{w}_{i=\Xi}{}^2 > 0$ the studies $i = \Xi = \{1, 2, \dots, \xi\}$ with Ξ different inflated variances.

6.3 Illustrative example of interventions to aid smoking cessation

We implement the RVSOM NMA model to the smoking cessation dataset which was introduced in Chapter 2. We used R function `NMAsvr()` from R package **NMAoutlier** [5] to fit the RVSOM NMA model. Number of RVSOM model fitting is equal to the number of studies ($N = 24$). The variance is shifted for each study and 24 overdispersion parameter estimators were calculated (Figure 6.1). Study 3 has the largest overdispersion parameter estimator and this is an indication that study 3 is an outlier (Figure 6.1). Study 7 provides the second largest value of overdispersion parameter estimator when study 7 was shifted.

Figure 6.2 depicts the LRT test when fitting the RVSOM NMA model for each study. It is clearly provided that study 3 has the largest value for the LRT test and afterward study 7 provides the second largest value of LRT. A large LRT value can claim that the null hypothesis of LRT test $H_0: {w_3}^2 = 0$ can be rejected and it can be a promise that study 3 is a potential outlier. A large LRT value for study 7 can also be evidence that study 7 is a potential outlier.

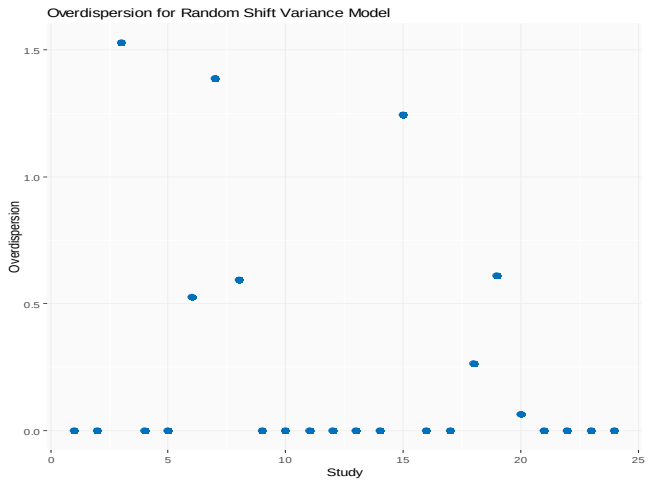

Figure 6.1. Over-dispersion parameter of the random shift variance model for each study for smoking cessation data.

We monitored a sharp decrease of the REML heterogeneity estimator by downweighing the study 3 (Table 6.2) with value $\hat{\tau}^2 = 0.16$ compared to the rest REML heterogeneity estimator values ranging from 0.34 to 0.42.

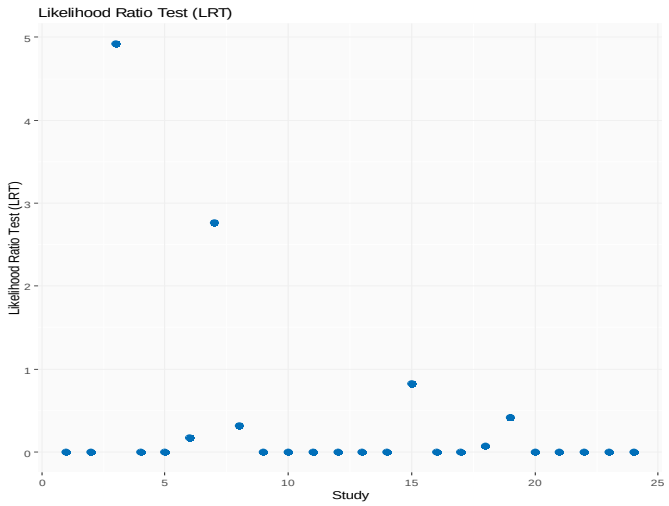

Figure 6.2 Likelihood ratio test (LRT) values of random shift variance model for each study for smoking cessation data.

Shifting the variance of study 3 has a large effect on the estimated summary odds-ratios. RVSOM NMA model resulted in $\hat{\mu}_{AB}$ =1.32 (0.83, 2.09), $\hat{\mu}_{AC}$ = 1.62 (1.23, 2.15), $\hat{\mu}_{AD}$ = 1.96 (1.13, 3.39) with REML heterogeneity estimator $\hat{\tau}^2 = 0.16$ when the variance of study 3 is downweighed while the results from standard NMA model are $\hat{\mu}_{AB}$ = 1.48 (0.79, 2.77), $\hat{\mu}_{AC}$ = 2.02 (1.39, 2.93), $\hat{\mu}_{AD}$ = 2.36 (1.15, 4.83) and REML heterogeneity estimator $\hat{\tau}^2 = 0.42$. Shifting the variance of study 7 provides also a large effect on the estimated summary odds-ratios (Figure 6.3). We can conclude that study 3 is an outlier followed by study 7. Conclusion with RVSOM method for outlier detection is the same with the conclusion from the FS outlier detection method for smoking cessation data (see also Chapter 5). This can confirm that our proposed methods are reliable for outlier detection. Both methods are iterative with 27 seconds of completion for FS method and 29 seconds for RVSOM model.

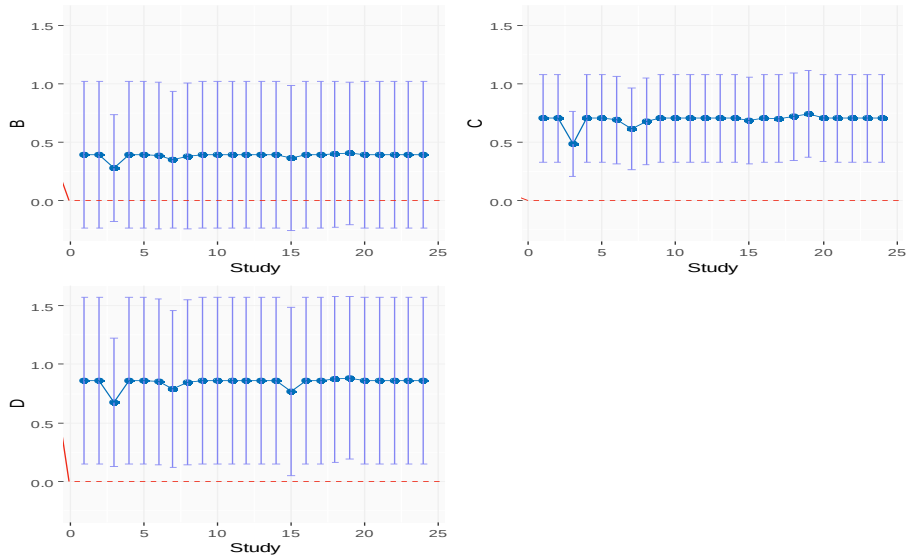

Figure 6.3 Summary estimates and their 95% confidence intervals for each treatment fitting the random shift variance model for each study in smoking cessation data.

Table 6.2. Study downweighed fitting the RVSOM NMA model. Q statistics ($Q^{total}, Q^{inc}, Q^{het}$) and restricted maximum likelihood heterogeneity estimator $\hat{\tau}^2$ for the RVSOM NMA model fitted by downweighed a study for each application. Results are given from **NMAoutlier** [35] package.

Study downweighed	Q^{total}	Q^{het}	Q^{inc}	$\hat{\tau}^2$
1	202.6188	187.3985	4.663504	0.4224061

2	202.6189	187.3985	4.663506	0.4225465
3	62.45292	46.43391	9.504379	0.1688964
4	202.6189	187.3985	4.663506	0.422459
5	202.6189	187.3985	4.663505	0.4225125
6	201.6597	186.4441	4.683338	0.4169078
7	192.7938	177.6191	4.848289	0.3454409
8	200.8982	185.6879	4.69686	0.4099748
9	202.6189	187.3985	4.663506	0.422459
10	202.6189	187.3985	4.663506	0.4225465
11	202.6189	187.3985	4.663505	0.4225034
12	202.6189	187.3985	4.663505	0.4225125
13	202.6189	187.3985	4.663506	0.4224101
14	202.6189	187.3985	4.663506	0.4224838
15	200.1248	185.6879	3.087227	0.4101728
16	202.6189	187.3985	4.663506	0.4225034
17	202.6188	187.3985	4.663481	0.4221723
18	198.3849	183.1005	4.647703	0.4135832
19	174.4131	158.6404	4.778086	0.389938
20	197.6053	182.2837	4.617336	0.4210417
21	202.6189	187.3985	4.663506	0.4224417
22	202.6189	187.3985	4.663506	0.4224379
23	202.6188	187.3985	4.663505	0.4225674
24	202.6189	187.3985	4.663506	0.4224101

6.4 Discussion

RVSOM model is an outlier diagnostic method that identifies and downweighs outliers in NMA. We fit an alternative model from standard random effects NMA that take into account outlyingness by shifting (downweighing) the variance of identified outliers.

Based on observed sharp changes in monitoring measures, the novel RVSOM tool allows the identification of outlying studies. A large over-dispersion parameter estimator can be an indication of an outlier. Conclusions on outlyingness should be based on the LRT test. Following bibliographical suggestions, a threshold for outlyingness of the LRT test can be provided by an empirical distribution with the parametric bootstrap method. However, empirical distribution with the parametric bootstrap method requires a lot of iterations and

makes the RVSOM model be a computationally intensive method. This is a limitation for the RVSOM NMA model and our future research is to provide alternative methods to obtain a threshold for the LRT test.

RVSOM method concluded to the same results with FS method after the application in smoking seccation data. This can confirm that both FS and RVSOM methods are reliable for outlier detection. Both methods indicated that computation time increased in accordance with the total number of studies in the network and thresolds with parametric bootstrap in RVSOM method and simulation envelopes in FS method can make the methods more computational intensive.

Caution is needed in deleting/omitting outliers and hereby extension of downweighing outliers may be an alternative choice. Shifting (downweighing) the variance of an outlier is more conservative than a simple study deletion. Exploring the robustness of results with the proposed downweighing method may be an alternative choice, possibly reducing bias, compared to deletion methods. RVSOM NMA estimates can be provided by downweighing outlier as secondary or sensitivity analysis. In any case, primary analysis with standard NMA is on real importance and attention of outlyingness needed when results from primary analysis and sensitivity differed. In conclusion, we argue that the method can be employed as an outlier diagnostic tool.

7 Using the R package NMAoutlier

7.1 Introduction

In this Chapter, we provide a description of how to use the R package **NMAoutlier** [28] for the implementation of the proposed methodologies presented in Chapters 4-6.

The aim of the package is to detect outliers in network meta-analysis. For transparency and reproducibility purposes, our proposed methods and visual tools have now offered in R statistical package **NMAoutlier** [28]. **NMAoutlier** [28] was developed with the aim to detect outlying and/or influential studies in NMA datasets with several outlier detection measures (such as Chapter 3) including the methods of the forward search algorithm and the shift variance random effects NMA model presented in Chapters 4 and 5 respectively. The package contains published NMA datasets that can be used for illustration issues of the detection methods.

There are several approaches [85], [92]–[95], [97], [98], [101] that have been developed for outlier detection in a pairwise meta-analysis (see for more details Section 3). Viechtbauer provided deletion outlier diagnostic measures for meta-analysis [85] and offered with the R package **metafor** [99]. Viechtbauer and Cheung [85] and Hedges and Olkin [84] provided the function metaoutliers in package **altmeta** [131] for the calculation of standardized residuals for each study. Beath developed an R package **metaplus** [132] for the implementation of the outlier detection method of finite mixture method of outliers and non-outliers.

To our knowledge, there does not exist any statistical software to offer the advantage of an outlier and influential detection in NMA. The need for the implementation of the outlier and influential diagnostical tools in NMA models motivating us to develop the R package **NMAoutlier** [28]. The Chapter is organized as follows: Section 7.2 outlines the **NMAoutlier** package description and Section 7.3 applies the package in R [76] with a real NMA dataset.

7.2 Software description

R package **NMAoutlier** [28] implements several outlier detection measures, the FS algorithm and the random shift variance model in NMA datasets. **NMAoutlier** [28] package employs the following:

• Outlying and influential detection measures

- Several outlier detection measures (function: `NMAoutlier_measures`) are provided: (a) Raw, (b) Standardized, (c) Studentized residuals; (d) Contribution to the Q statistic (Mahalanobis distance) and (c) leverage.
- Plots for outlier and influential measures (function: `plot_NMAoutlier_measures`): (a) Raw, (b) Standardized, (c) Studentized residuals; (d) contribution to the Q statistic (Mahalanobis distance) and (c) leverage
- Several outlier and influential detection measures considered deletion (function: `NMAoutlier_measures_deletion`): (a) Raw, (b) Standardized, (c) Studentized deleted residuals; (d) Cook distance; (e) COVRATIO; (f) weight “leave one out”; (g) leverage “leave one out”; (h) heterogeneity “leave one out”; (i) R heterogeneity; (k) R Qtotal; (l) R Qheterogeneity (m); R Qinconsistency; (n) DFBETAS.
- Plots for outlier and influential detection measures considered deletion (function: `plot_NMAoutlier_measures_deletion`). In plots, the y-axis provides the monitoring outlier detection measure considered deletion (measures (a)-(n) in function `NMAoutlier_measures_deletion)` the x-axis provides the study deleted.

• The forward search algorithm in network meta-analysis (function `NMAoutlier`)

• Forward plots (`fwdplot`) for the monitoring statistics in each iteration of forward search algorithm: (a) P-scores, Rücker G & Schwarzer G (2015) [33]; (b) z-values for difference of direct and indirect evidence with back-calculation method, König (2013) [119], Dias (2010) [118]; (c) Standardized residuals; (d) heterogeneity variance estimator; (e) cook distance; (f) ratio of variances; (g) Q statistics, Krahn et al. (2013) [39] .

• Forward plot (`fwdplotest`) for summary estimates and their confidence intervals for each treatment in each iteration of the forward search algorithm.

• Random shift variance NMA model, RVSOM NMA (function: `NMAsvr`)

• Plots for the monitoring measures for random shift variance model (function: `svrplot`)

• Plots for the monitoring measures for random shift variance model (function: `svrplotest`)

The package apart from the functions; `NMAoutlier_measures`, `plot_NMAoutlier_measures`, `NMAoutlier_measures_deletion`, `plot_NMAoutlier_measures_deletion`, `NMAoutlier`, `fwdplot`, `fwdplotest`, `NMAsvr`, `svrplot`, `svrplotest`. An overview of functions in **NMAoutlier** [28] package and a short description for the implementation of each function is provided in Table 7.1.

Table 7.1. Overview of functions in **NMAoutlier** [28] package and a short description for the implementation of each function.

Function	Implementation
`NMAoutlier_measures`	Several outlier detection measures: (a) Raw residuals (b) Standardized residuals (c) Studentized residuals (d) Contribution to the Q statistic (c) leverage
`plot_NMAoutlier_measures`	Plots for outlier and influence measures provided with function `NMAoutlier_measures`
`NMAoutlier_measures_deletion`	Several outlier and influence detection measures considered deletion: (a) Raw deleted residuals (b) Standardized deleted residuals (c) Studentized deleted residuals (d) Cook distance (e) COVRATIO (f) weight “leave one out” (g) leverage “leave one out” (h) heterogeneity “leave one out” (i) R heterogeneity (k) R Qtotal (l) R Qheterogeneity (m) R Qinconsistency (n) DFBETAS

plot_NMAoutlier_measures_deletion	Plots for outlier and influence measures provided with function NMAoutlier_measures_detetion
NMAoutlier	The forward search algorithm in network meta-analysis.
fwdplot	Forward plots for the monitoring statistics in each step of forward search algorithm: (a) P-scores, Rücker G & Schwarzer G (2015) [33]; (b) z-values for difference of direct and indirect evidence with back-calculation method, König (2013) [119], Dias (2010) [118]; (c) Standardized residuals; (d) heterogeneity variance estimator; (e) cook distance; (f) ratio of variances; (g) Q statistics, Krahn *et al.* (2013) [39]
fwdplotest	Forward plot for summary estimates and their confidence intervals for each treatment in each step of the forward search algorithm.
NMAsvr	Random shift variance network meta-analysis model.
svrplot	Plots for the monitoring measures for the random shift variance model.
svrplotest	Plot for summary estimates and their confidence intervals for each treatment for the random shift variance model.

NMAoutlier_measures function provides several outlier and influential measures and NMAoutlier_measures_deletion offers the ability to compute several outliers and influential measures considered deletion. NMAoutlier function employs the forward search algorithm and NMAsvr function employs the random shift variance model. The proposed detection measures and methods can be diagnostic tools for detection outlying and/or influential studies. They can also be used to detect studies that are responsible for heterogeneity and inconsistency.

An overview of the arguments and their descriptions of the components of all functions in the **NMAoutlier** package is provided in the Appendix (Appendix Tables 4-14). A brief overview of the package can also be provided with `help` function by typing

```
R> help(NMAoutlier)
```

`NMAoutlier_measures`, `NMAoutlier_deletion_measures`, `NMAoutlier`, and `NMAsvr` functions calculate the outlier detection methodologies for network meta-analysis model from graph theory [5] fitted (`netmeta` function) with R package **netmeta** [35]. The researcher can choose the reference treatment (`reference`) fitted in the NMA model.

Monitoring is helpful to identify outlying and/or influential studies. Monitoring statistical measures for the basic set in each FS iteration and for RVSOM NMA model can be:

- **Likelihood statistics** (for `NMAsvr` function only). The heterogeneity estimation method is conducted under the Restricted Maximum likelihood estimator and likelihood statistics offered from the calculation, the twice of maximum log-likelihood, the convergence diagnostic, and the Likelihood Ration test (LRT) test.

- **Outlier and influential case diagnostics measures**. Standardized residuals (arithmetic mean in case of multi-arm studies); Cook statistic; Ratio of determinants of the variance-covariance matrix

- **Ranking measures** (for `NMAoutlier` function only). P-scores for ranking of treatments [33] with implementation of (`netrank` function) from R package **netmeta** [35]. Therefore, argument `small.values` is an argument for function NMAoutlier with options "`good`" or "`bad`" if small values considered beneficial or harmful on the outcome, respectively.

- **Heterogeneity and inconsistency measures**. Overall heterogeneity/inconsistency Q statistic (Q) is the design-based decomposition of Cochran Q as provided by Krahn *et al.* [39]; Overall heterogeneity Q statistic (Q); Between-designs Q statistic (Q), based on a random-effects model with square-root of between-study variance estimated embedded in a full design-by-treatment interaction model. Implementation with (`decomp.design` function) from R package **netmeta** [35]; Z-values for comparison between direct and indirect evidence for each iteration of the forward search algorithm.

By monitoring the difference between direct and indirect evidence, potential sources of consistency can be detected with the implementation of (`netsplit` function) from R package **netmeta** [35]. Based on the methodology with the back-calculation method to derive indirect estimates from direct pairwise comparisons and network estimates (Dias *et al.* [118], König *et al.* [119]).

The development version of the package is available on the GitHub repository: https://github.com/petropouloumaria/NMAoutlier.

7.3 Application of NMAoutlier in practice with smoking cessation data

This example comprises four interventions to aid smoking cessation [41], [42] introduced in Chapter 2. Smoking cessation data is part of **netmeta** [35] package with arm level data. We load the dataset by typing

```
R> data(smokingcessation, package = "netmeta")
```

Before conducting the analysis, the R packages **netmeta** and **NMAoutlier** should be installed. The function `install.packages()` can be used to install the packages that the user needs. The two above packages can be installed by typing

```
R> install.packages(c("netmeta", "NMAoutlier"))
```

The function `library` can be used to make the library available

```
R> library(netmeta)
```

The **NMAoutlier** package performs outlier and influential detection methodologies to NMA datasets with contrast level data. The transformation is needed if arm level data provided and can be conducted with function `pairwise` from the **netmeta** package.

We transform the dataset from arm to contrast level data with odds ratios using the function `pairwise` from the **netmeta** package [35] (the same information is also provided in Appendix A) (this is the reason that we load the **netmeta** package).

```
R> p1 <- pairwise(list(treat1, treat2, treat3),
+   list(event1, event2, event3),
+   list(n1, n2, n3),
```

```
+   data=smokingcessation,
+   sm="OR")
```

We denoted with `p1` the object that assigned the data with smoking cessation in contrast level.

7.3.1 Part 1: Simply outlier detection measures

We can calculate some simple outlier detection measures for NMA. The function `NMAoutlier_measures` calculate several outlier detection measures for each study.

```
R> measures <- NMAoutlier_measures(p1)
```

The object `measures` apart from the calculation of several measures for each study offered from function `NMAoutlier_measures`; raw residuals, standardized residuals, studentized residuals, contribution to the Q statistic, and leverage. We can see the contribution to the Q statistic (Mahalanobis distance) for each study by typing

```
R> measures$Mahalanobis.distance
```

Function `plot_NMAoutlier_measures` generates plot(s) to monitor selected outlier and influential statistical measure(s). The function creates a plot of the selected outlier detection measure of each study in the network. An object of class function `NMAoutlier_measures` (for this example object `measures`) is mandatory for running this function. Candidate statistics to be monitored (argument `stat`) can be raw residuals; standardized residuals; studentized residuals; contribution to the Q statistic and leverage.

We can plot the contribution to the Q statistic (Mahalanobis distance) measure.

```
R> plot_NMAoutlier_measures(measures, stat = "mah")
```

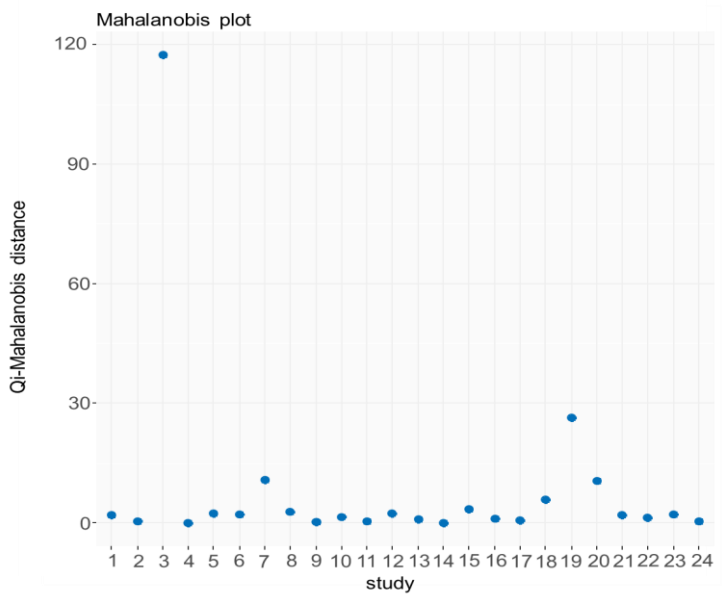

Figure 7.1. Contribution to the Q statistic (Mahalanobis distance) values for each study for smoking cessation data.

7.3.2 Part 2: Outlier detection measures considered deletion

We can calculate some outlier detection measures considering the deletion of a study. The function `NMAoutlier_deletion_measures` calculates several outlier detection measures considering study deletion.

```
R> deletion <- NMAoutlier_deletion_measures(p1)
```

Measures that provided in this function are raw, standardized and studentized deleted residuals, Cook's distance, COVRATIO, weight "leave one out" etc. We can see the standardized deleted residuals for each study

```
R> deletion$stand.deleted
```

We can see the values of COVRATIO when considering deletion for each study

```
R> deletion$covratio
```

Function `plot_NMAoutlier_deletion_measures` generates plot(s) to monitor selected outlier and influential statistical measure(s) after considered a deletion of a study. The function creates a plot of the selected outlier detection measure after the deletion of a study. An object of class function `NMAoutlier_deletion_measures` is mandatory (here object `deletion`) for running the function. Candidate statistics to be monitored (argument

`stat`) can be raw deleted residuals; standardized deleted residuals; studentized deleted residuals; COVRATIO; Cook distance; R statistic for heterogeneity; R statistic for Qtotal; R statistic for Qheterogeneity; R statistic for Qinconsistency.

We can display the R statistic for Qinconsistency by typing

```
R> plot_NMAoutlier_deletion_measures(deletion, stat = "rqinc")
```

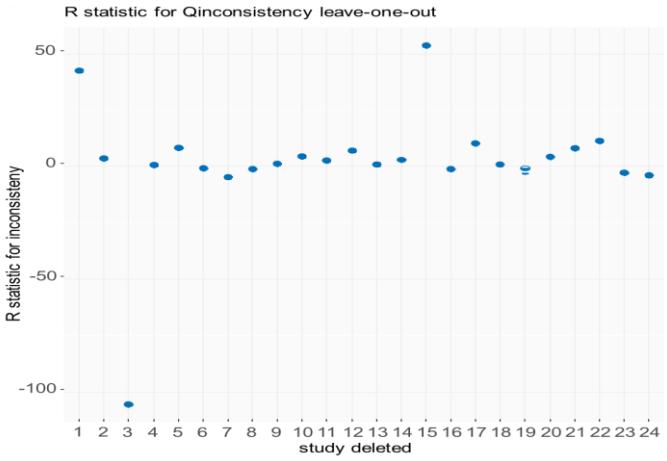

Figure 7.2. R statistic for Qinconsistency for smoking cessation data.

7.3.3 Part 3: Forward Search Algorithm - Detection Method

`NMAoutlier` function employs the forward search algorithm in network meta-analysis. During the search, several measures calculated and among them, P-scores can be monitored. A researcher should take into account outcome is beneficial or harmful as this is a need for P-scores calculation. The default value considered a beneficial outcome and therefore small values specified to be "`good`". In smoking cessation data, the outcome is harmful and we should specify the argument `small.values` with "`bad`".

We can conduct the forward search algorithm in this dataset with the criterion of the smallest absolute standardized residuals (default value that the researcher should not specify) as follows:

```
R> FSresult1 <- NMAoutlier(p1, small.values = "bad")
```

We can see the basic set for each iteration of the forward search algorithm

```
R> FSresult1$basic
```

We can conduct the forward search algorithm with function `NMAoutlier` by taking the criterion of the maximum of median absolute likelihood contributions to select the initial subset and the study entered from non-basic set to basic set.

```
R> FSresult2 <- NMAoutlier(p1, crit1 = "L", crit2 = "L",
+   small.values = "bad")
```

We can see the basic set for each iteration of the forward search algorithm

```
R> FSresult2$basic
```

Function `fwdplot` generates forward plot(s) to monitor selected statistic(s) and/or method(s). The function creates a plot of the selected statistic throughout the iterations of the forward search algorithm. An object of class function `NMAoutlier` is mandatory for running the function. Candidate statistics to be monitored (argument `stat`) can be P-score; z-values by back-calculation method to derive indirect estimates from direct pairwise comparisons and network estimates; standardized residuals; heterogeneity variance estimator; cook distance; ratio of variances; Q statistics (Overall heterogeneity/ inconsistency Q statistic (Q), overall heterogeneity Q statistic (Q), between-designs Q statistic (Q) based on a random-effects design-by-treatment interaction model).

We can see the forward plot to monitor z-values by the back-calculation method to derive indirect estimates from direct pairwise comparisons and network estimates

```
R> fwdplot(FSresult1, stat = "dif")
```

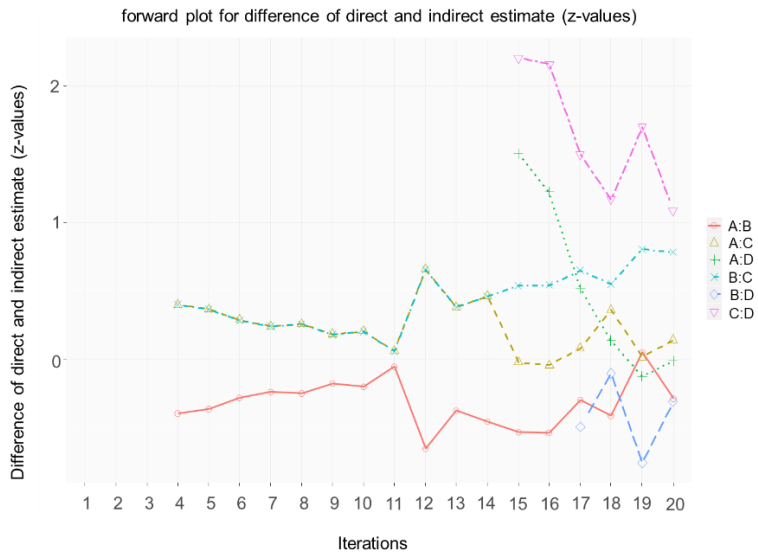

Figure 7.3. Forward plot for z-values by the back-calculation method in smoking cessation data.

The researcher has the choice to provide forward plots for a selected statistical measure (argument `select.st`) for P-scores/z-values by the back-calculation method to derive indirect

estimates from direct pairwise comparisons and network estimates/standardized residuals for selected treatment/comparisons/study, respectively.

We can see the forward plot

```
R> fwdplot(FSresult1, stat = "", select.st)
```

Function `fwdplotest` generates forward plots for summary estimates with a 95 percent confidence interval for each treatment. An object of class function `NMAoutlier` is the only argument and it is mandatory for running the function.

We can see the forward plots for summary estimates with 95 percent confidence interval for each treatment

```
R> fwdplotest(FSresult1)
```

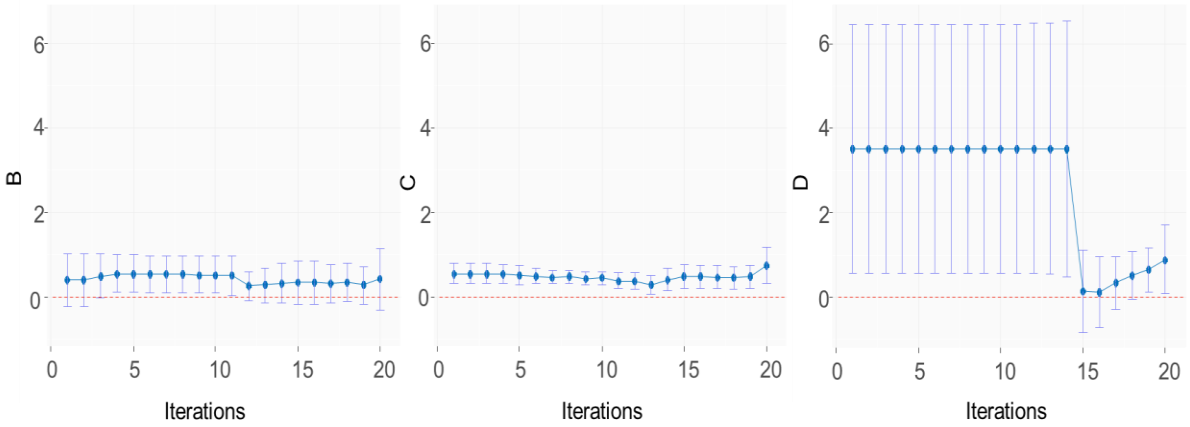

Figure 7.4. Forward plot for summary estimates with 95 percent confidence interval for each treatment in smoking cessation data.

7.3.4 Part 4: Shift Variance Network Meta-analysis – Detection method and sensitivity analysis downweighing outlier

`NMAsvr` function employs the RVSOM NMA model for the detection of outlying and influential studies. We can implementthe model for each study in smoking seccation dataset as follows:

```
R> SVRresult1 <- NMAsvr(p1, small.values = "bad")
```

Some measures for the random shift variance model are outlined, such as; the variance estimator of shift variance model (over-dispersion) and likelihood statistics such as the twice maximum log-likelihood, its convergence diagnostic, and the likelihood ratio test (LRT).

Moreover, values of statistical monitoring measures of random shift variance model are given, such as; standardized residuals, P scores, Q statistics, heterogeneity estimator, etc.

We can see the LRT with random shift variance model of each study with

```
R> SVRresult1$LRT
```

We can see the over-dispersion with random shift variance model of each study with

```
R> SVRresult1$over_disp
```

Function `svrplot` generates plots for monitoring measures. An object of function `NMAsvr` is the first mandatory argument for running the function and the statistic to be monitored should be the second argument of the function. We can figure out a plot of LRT of the random shift variance model for each study with

```
R> svrplot(SVRresult1, "LRT")
```

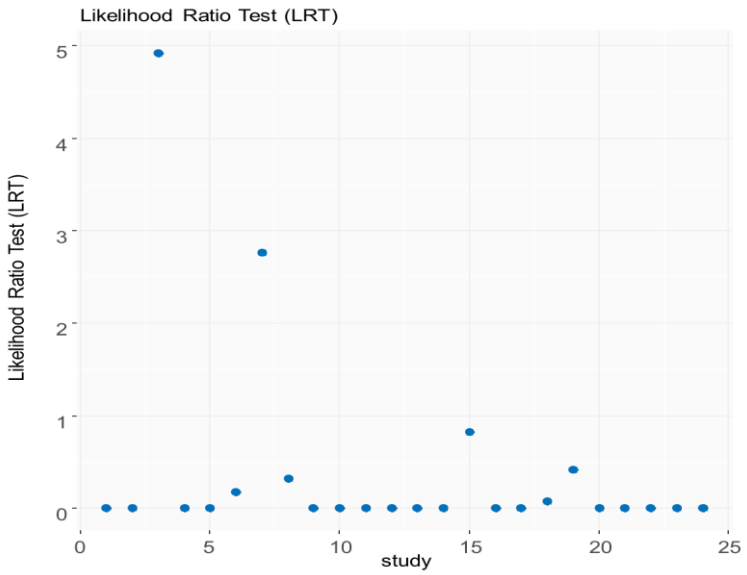

Figure 7.5. Likelihood Ratio Test (LRT) of the random shift variance model for each study for smoking cessation data.

We can draw a plot of the over-dispersion parameter of the random shift variance model for each study with

```
R> svrplot(SVRresult1, "over_disp")
```

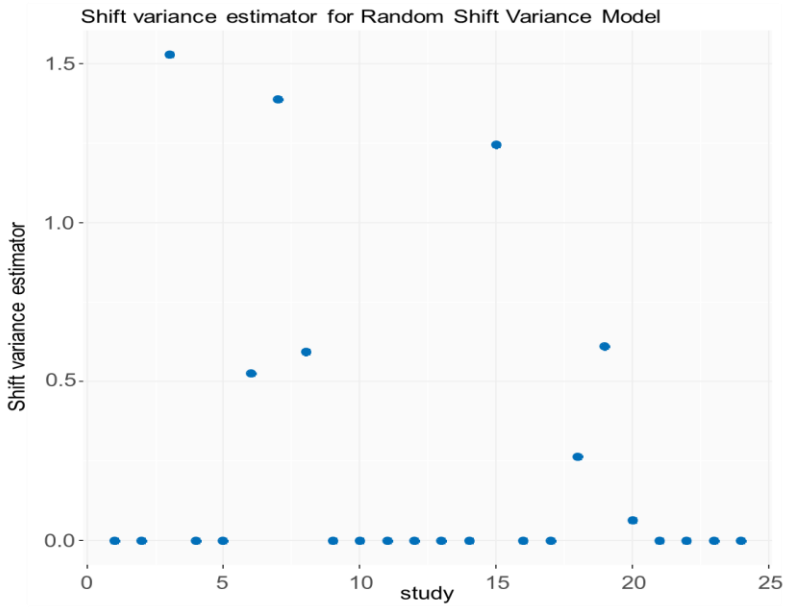

Figure 7.6. Over-dispersion parameter of the random shift variance model for each study for smoking cessation data.

7.4 Discussion

Network meta-analysis is the most popular evidence synthesis method and there are several statistical packages available for the implementation of meta-analytical models up to date. The R package **netmeta** implements the network meta-analysis model in a frequentist framework and it is the most comprehensive R package for NMA [35].

The proposed package **NMAoutlier** [28] is the first package that implements outlier diagnostics measures, methods, and tools in NMA evidence structures. It offers the ability to calculate several outlier and influential measures for NMA but also two methods for outlier diagnosis; the FS algorithm and the RVSOM NMA model. In this Chapter, we described details about the R package **NMAoutlier** [28] and an overview of the methods offered. For illustration reasons, a working example of smoking cessation data is provided to give an insight on how to use the R package **NMAoutlier** [28].

8 Summary

8.1 Summary

Systematic reviews and meta-analyses have been established as an integral part of comparative effectiveness research. The increasing number of different educational and psychological interventions in the educational system has led to the need for comparative effectiveness research with the aim to identify the best intervention. Network meta-analysis synthesizes both direct and indirect evidence, gives more powerful results and provides estimates with increased precision compared to pairwise estimates. NMA has become a popular statistical tool in evidence synthesis. Based on a database of published NMA from the onset until 14 April 2015, the time trend indicates the increasing number of published NMAs and the tendency for the use of appropriate methods. Moreover, the overview of the characteristics of published NMAs is a useful resource of information for methodologists that aim to update the current knowledge on appraising NMA methods. This collection of 456 published NMAs indicates that many NMAs provide important methodological limitations, but the comprehensive use of appropriate methodologies and completeness of reporting (such as the description of the statistical methods used) improved over the years. For example, an increasing number of NMAs used appropriate methods to test the plausibility of the consistency assumption and in recent years around 90% of articles clearly reported whether a random-effects of the fixed-effect model was used.

A common problem in the synthesis of studies is the existence of outlying or/and influential studies. Outlying and influential studies may bias the results but little work has been done for outlying identification in NMA. For this reason, this Thesis focuses on developing several methodologies for the identification of outliers and influential studies in network meta-analysis. Heterogeneity and inconsistency can be seen as differences in the potential effect modifiers within and across the pairwise comparisons in a network of interventions. It is common that a potential source of heterogeneity and inconsistency is provided due to the existence of extreme study effects. Extreme study effects may be an outlier or influential study. A study that is far away from the rest of the data and does not explain by the assumed model

defined as an outlier and a study that influences the model parameters (network estimates and heterogeneity estimator) defined as influential. Several methods for outlier and influential identification have been developed in a pairwise meta-analysis considering addition or deletion of studies, based on the likelihood or by taking alternative distributions for heterogeneity. In this dissertation, several simple measures for outlier and influential studies detection are provided. Measures considering the deletion of a study for outlier and influential studies detection are extended in NMA. A forward search algorithm, considered the addition of studies, has recently been developed in meta-regression. This algorithm starts with a subset of studies that considered outlying-free and it gradually adds studies until all studies entered. Sharp changes in monitoring measures during the search are considered potential outlying and/or influential studies. In this Thesis, the methodology with the forward search algorithm for outlying identification has been developed in the NMA model. Additionally, a novel model with shifting the variance taking into account outlying studies from meta-analysis to network meta-analysis model is extended. The advantage of the random shift variance model is that it offers the ability of down-weighting studies and therefore can be used as a sensitivity analysis.

The several outlier and influential measures and two proposed methods in NMA for outlying identification, forward search algorithm and shift random NMA model, applied in real and in simulated datasets. Results of measures and methods indicate the potential source of outlying and influential cases in datasets. The methods are promising tools for the identification of outlying and influential cases and sources of heterogeneity and/or inconsistency. For the implementation of the several detection measures and methods a flexible and user-friendly software, an R package, called **NMAoutlier**, was developed with a description and details to provide guidance on how to use the R package through real datasets.

8.2 Περίληψη

Οι συστηματικές ανασκοπήσεις και μετα-αναλύσεις έχουν καθιερωθεί ως αναπόσπαστο κομμάτι της έρευνας για τη σχετική αποτελεσματικότητα μεταξύ παρεμβάσεων. Σήμερα, η λήψη αποφάσεων και η ιεράρχηση μεταξύ ανταγωνιστικών παρεμβάσεων σε πολλούς τομείς, βασίζονται στην ανάπτυξη του μετα-ανάλυσης δικτύων (ΜΑΔ). Ο ολοένα αυξανόμενος αριθμός διαφορετικών εκπαιδευτικών και ψυχολογικών παρεμβάσεων στο εκπαιδευτικό σύστημα οδηγεί στην ανάγκη σύγκρισης τους με στόχο την εύρεση της καταλληλότερης παρέμβασης. Η μετα-ανάλυση δικτύων συνθέτει τόσο άμεσες όσο και έμμεσες πληροφορίες

έτσι ώστε να παρέχει πιο ισχυρά αποτελέσματα και εκτιμήσεις με αυξημένη ακρίβεια σε σχέση με τις εκτιμήσεις ανά ζεύγη. Η ΜΑΔ έχει γίνει ένα δημοφιλές στατιστικό εργαλείο στη σύνθεση στοιχείων. Βάσει μιας συλλογής δεδομένων με δημοσιευμένες ΜΑΔ από την αρχή έως τις 14 Απριλίου 2015, η τάση δείχνει τον αυξανόμενο αριθμό δημοσιευμένων της ΜΑΔ και της εφαρμογής ολοένα και καταλληλότερων μεθοδολογιών. Επιπλέον, η επισκόπηση των χαρακτηριστικών των δημοσιευμένων ΜΑΔ είναι μια χρήσιμη πηγή πληροφόρησης για ερευνητές που στοχεύουν να αναβαθμίσουν την υπάρχουσα γνώση σχετικά με την αξιολόγηση των μεθόδων ΜΑΔ. Αυτή η συλλογή από 456 δημοσιευμένων ΜΑΔ υποδεικνύει ότι πολλές ΜΑΔ παρέχουν σημαντικούς μεθοδολογικούς περιορισμούς, αλλά η εκτεταμένη χρήση των κατάλληλων μεθοδολογιών και της πληρότητας των εκθέσεων (όπως η περιγραφή των χρησιμοποιούμενων στατιστικών μεθόδων) έχει βελτιωθεί με την πάροδο των ετών. Για παράδειγμα, ένας αυξανόμενος αριθμός ΜΑΔ χρησιμοποίησε κατάλληλες μεθόδους για να ελέγξει την αξιοπιστία της υπόθεσης της συνέπειας και τα τελευταία χρόνια γύρω στο 90% των άρθρων ανέφερε σαφώς αν χρησιμοποιήθηκε το μοντέλο τυχαίων ή σταθερών επιδράσεων.

Παρόλο που η βιβλιογραφική μελέτη έδειξε βελτιωμένη στατιστική μεθοδολογία, εξακολουθούν να υπάρχουν μεθοδολογικές πτυχές στα μοντέλα μετα-αναλύσεων δικτύων που χρειάζονται ακόμα περαιτέρω ανάπτυξη. Ένα κοινό πρόβλημα στη σύνθεση των μελετών είναι η ύπαρξη ακραίων και / ή επηρεάζουσων μελετών. Παρόλο που οι ακραίες και οι επηρεάζουσες μελέτες ενδέχεται να οδηγήσουν σε μεροληπτικά αποτελέσματα, ελάχιστη ερευνητική δουλειά έχει πραγματοποιηθεί για τη διερεύνηση τέτοιων μελετών στη ΜΑΔ. Για το λόγο αυτό, η παρούσα διδακτορική διατριβή επικεντρώνεται στην ανάπτυξη μεθοδολογίας για τη διερεύνηση ακραίων και επηρεάζουσων μελετών. Η ετερογένεια και η ασυνέπεια μπορούν να θεωρηθούν ως διαφορές στους τροποποιητές του αποτελέσματος σε ένα δίκτυο παρεμβάσεων. Μια πιθανή πηγή ετερογένειας και ασυνέπειας αποτελεί η ύπαρξη ακραίων ή επηρεάζουσων μελετών. Ως ακραία ορίζεται η μελέτη που απέχει πολύ από τα υπόλοιπα δεδομένα και δεν προβλέπεται ικανοποιητικά από το μοντέλο που έχουμε υποθέσει, ενώ ως επηρεάζουσα η μελέτη που επηρεάζει τις παραμέτρους του μοντέλου, δηλαδή τις εκτιμήσεις του δικτύου και την ετερογένεια. Αρκετές μεθοδολογίες για την εύρεση ακραίων και επηρεάζουσων μελετών έχουν αναπτυχθεί στη μετα-ανάλυση δύο παρεμβάσεων, μεθοδολογίες που θεωρούν την είσοδο ή έξοδο μελετών, μεθοδολογίες που στηρίζονται στη συνάρτηση πιθανοφάνειας ή μεθοδολογίες που βασίζονται στη λήψη εναλλακτικών κατανομών για ετερογένεια. Στη παρούσα διδακτορική διατριβή παρουσιάζονται πολλά απλά μέτρα εύρεσης ακραίων και

επηρεάζουσων μελετών. Μέτρα θεωρώντας τη διαγραφή μελέτης για την εύρεση ακραίων και επηρεάζουσων μελετών επεκτάθηκαν στο μοντέλο ΜΑΔ. Ο προς τα εμπρός αλγόριθμος αναζήτησης αναπτύχθηκε πρόσφατα στη μετα-παλινδρόμηση. Ο αλγόριθμος βασίζεται στη σταδιακή προσθήκη των μελετών, ξεκινά με ένα υποσύνολο μελετών που θεωρείται απαλλαγμένο από ακραίες μελέτες και προσθέτει σταδιακά τις μελέτες μέχρι να εισέλθουν όλες οι μελέτες. Οι έντονες αλλαγές των μέτρων παρακολούθησης κατά τη διάρκεια της αναζήτησης αποτελεί ένδειξη για πιθανές ακραίες ή / και επηρεάζουσες μελέτες. Στην διδακτορική διατριβή, η μεθοδολογία με τον προς τα εμπρός αλγόριθμο αναζήτησης για τη διερεύνηση ακραίων ή/και επηρεάζουσων μελετών αναπτύχθηκε στο μοντέλο ΜΑΔ από το μοντέλο της μετα-παλινδρόμησης. Επιπλέον, επέκτεινα ένα νέο μοντέλο τυχαίων επιδράσεων με τη μετατόπιση της διακύμανσης, λαμβάνοντας υπόψη τις ακραίες μελέτες από το μοντέλο της απλής μετα-ανάλυσης δύο παρεμβάσεων στο μοντέλο της ΜΑΔ. Το πλεονέκτημα του μοντέλου τυχαίων επιδράσεων με τη μετατόπιση της διακύμανσης είναι ότι προσφέρει την ικανότητα μείωσης του βάρους των ακραίων μελετών και συνεπώς μπορεί να χρησιμοποιηθεί ως ανάλυση ευαισθησίας.

Τα διάφορα μέτρα και οι δύο προτεινόμενες μεθοδολογίες στη ΜΑΔ για διερεύνηση των ακραίων και επηρεάζουσων μελετών, με τον προς τα εμπρός αλγόριθμο αναζήτησης και το μοντέλο τυχαίων επιδράσεων με μετατόπιση της διακύμανσης, εφαρμόστηκαν σε δημοσιευμένα δίκτυα μετα-αναλύσεων και σε προσομοιωμένα δεδομένα. Τα αποτελέσματα από τι εφαρμογές υποδεικνύουν την εύρεση ακραίων και επηρεάζουσων μελετών στα δεδομένα. Οι προτεινόμενες μεθοδολογίες αποτελούν καλά υποσχόμενα εργαλεία για τον εντοπισμό ακραίων και επηρεάζουσων μελετών και την εύρεση πηγών δημιουργίας υψηλής ετερογένειας και / ή ασυνέπειας. Για την υλοποίηση των διάφορων μέτρων εύρεσης ακραίων και επηρεάζουσων μελετών αλλά και των δύο προτεινόμενων μεθοδολογιών σε ένα ευέλικτο και φιλικό προς το χρήστη λογισμικό, αναπτύχθηκε το στατιστικό πακέτο **NMAoutlier** στην R που περιγράφεται στη παρούσα διδακτορική διατριβή παρέχοντας λεπτομέρειες και οδηγίες για τον τρόπο χρήσης του πακέτου μέσω της εφαρμογής του σε πραγματικά δεδομένα.

Appendix

Appendix Tables.

Appendix Table 1. Effect size $\boldsymbol{y_i}$, standard error $\boldsymbol{s_i}$ and treatment comparisons for each study of the smoking cessation dataset.

Effect size $\boldsymbol{y_i}$	Standard error $\boldsymbol{s_i}$	treat1	treat2	study label
-1.0513	0.4132	A	C	1
-0.1285	0.4760	A	D	1
0.9228	0.3998	C	D	1
-0.0012	0.4504	B	C	2
-0.2253	0.3839	B	D	2
-0.2241	0.3723	C	D	2
-2.2023	0.1430	A	C	3
-0.8704	0.7911	A	C	4
-0.4156	0.1557	A	C	5
-2.7797	1.4698	A	C	6
-2.7054	0.6252	A	C	7
-2.4252	1.0423	A	C	8
-0.4436	0.5220	A	C	9
0.0160	0.1699	A	B	10
-0.3935	0.3266	A	B	11
-0.3904	0.1680	A	C	12
-0.1063	0.5956	A	C	13
-0.5834	0.2983	A	C	14
-3.5225	1.4970	A	D	15
-0.6796	0.4411	A	B	16
-0.5397	0.1401	A	C	17
0.1255	0.3200	A	C	18
0.2400	0.1737	A	C	19
-0.0390	0.1874	A	C	20
0.1517	0.4290	B	C	21
-1.0435	0.4490	B	D	22
-0.6807	0.4092	C	D	23

0.4055	0.7139	C	D	24

Appendix Table 2. Effect size $\boldsymbol{y_i}$, standard error $\boldsymbol{s_i}$ and treatment comparisons for each study of the actinic keratosis dataset.

Effect size $\boldsymbol{y_i}$	Standard error $\boldsymbol{s_i}$	treat1	treat2	study label
-0.7069	0.4287	1	2	1
-1.2933	0.4227	1	2	2
-1.6319	0.6076	1	2	3
-0.8391	0.4543	1	2	4
-5.1527	1.4500	1	3	5
-4.0763	1.4747	1	4	6
-3.2321	0.7862	1	3	7
-1.1632	1.1247	1	4	8
-1.6802	0.4443	1	2	9
-2.4849	0.5381	1	5	10
0.0000	0.8165	3	6	11
-3.3998	1.4527	1	3	12
-3.2241	0.4078	1	4	13
-3.3266	0.6755	1	5	14
-2.6210	0.5475	1	5	15
-3.0888	0.5388	1	6	16
-4.1017	0.6070	1	4	17
-2.4902	0.2764	1	4	18
-2.2548	1.5713	1	4	19
-3.9478	0.6104	1	4	20
0.1367	0.3700	1	4	21
-0.2448	0.6159	4	7	22
-2.1370	1.1131	7	8	22
-2.3817	1.1078	4	8	22
-2.7642	0.8146	4	7	23
-2.1145	0.5023	1	5	24
-2.2003	0.4821	1	9	25
-6.4754	1.5717	1	4	26
-3.1540	0.7690	1	6	27

-0.5564	0.2344	1	6	28
2.5020	0.5005	6	8	28
1.9456	0.4999	1	8	28
-1.6112	0.6282	1	4	29
-1.8099	0.6269	1	4	30
-3.3759	0.5914	1	5	31
-2.6507	0.5733	1	6	32
-2.1698	0.3323	1	5	33
-2.8571	0.3413	1	6	33
-0.6873	0.2033	5	6	33
-2.6659	0.2410	1	9	34
-1.2164	0.6562	4	5	35

Appendix Table 3. Effect size $\boldsymbol{y_i}$, standard error $\boldsymbol{s_i}$ and treatment comparisons for each study of the thrombolytics dataset.

Effect size $\boldsymbol{y_i}$	Standard error $\boldsymbol{s_i}$	treat1	treat2	study label
0.1575	0.0486	1	2	1
0.04521	0.0471	1	4	1
-0.1123	0.0558	2	4	1
0.0260	0.0394	1	3	2
0.0048	0.0392	1	8	2
-0.0211	0.0394	3	8	2
0.3718	0.5427	1	3	3
0.8988	0.8571	1	3	4
-0.0162	0.8361	1	3	5
-0.0532	0.0491	1	3	6
0.6096	0.6464	1	3	7
0.5463	0.4908	1	3	8
0.7323	0.5618	1	3	9
-0.4054	0.6603	1	4	10
0.0603	0.0891	1	6	11
0.3976	0.7068	1	7	12
-0.1013	0.9302	1	8	13
-0.2816	0.7281	1	8	14

0.7672	0.7750	1	8	15
-0.1818	0.4491	1	8	16
-0.0054	0.0638	2	5	17
-0.0341	0.0667	2	6	18
0.7508	0.4826	2	6	19
-0.5806	0.8401	2	7	20
-0.0217	0.3908	2	7	21
-1.2789	0.5185	2	8	22
-1.4737	0.6520	2	8	23
0	0.6318	3	7	24
0.5490	0.5285	3	7	25
0.1652	0.6194	3	7	26
0.2623	0.4360	3	8	27
0.3247	0.6051	3	8	28

Appendix A.

R code calculates odds ratios for smoking cessation data. The dataset is a part of R package **netmeta** [35] that compared the relative effects of four smoking cessation counseling programs ($n = 4$): defined as no contact (A), self-help (B), individual counseling (C), and group counseling (D). The binary outcome was the number of events that successful stopped smoking at 6 to 12 months. Arm level data (number of events, total sample size in each arm and treatments compared) can be found in **netmeta** [35] package. Here is provided the code for calculating odds ratios.

```
library(netmeta)

data("smokingcessation")

pm <- pairwise (list (treat1, treat2, treat3),
                event = list (event1, event2, event3),
                n = list (n1, n2, n3),
                data = smokingcessation,
                sm = "OR")
```

Bibliography

1. Higgins, J.P.T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M.J., Welch, V.A. (editors). (2019). Cochrane Handbook for Systematic Reviews of Interventions. 2nd Edition. Chichester (UK): John Wiley & Sons.
2. Salanti, G. (2012). Indirect and mixed-treatment comparison, network, or multiple-treatments meta-analysis: many names, many benefits, many concerns for the next generation evidence synthesis tool. Research Synthesis Methods, **3**(2), 80–97.
3. White, I.R., Barrett, J.K., Jackson, D., Higgins, J.P.T. (2012). Consistency and inconsistency in network meta-analysis: model estimation using multivariate meta-regression. Research Synthesis Methods, **3**(2), 111–125.
4. Higgins, J. P. T., Jackson, D., Barrett, J.K., Lu, G., Ades, A.E., White, I.R. (2012). Consistency and inconsistency in network meta-analysis: concepts and models for multi-arm studies. Research Synthesis Methods, **3**(2), 98–110.
5. Rücker, G. (2012). Network meta-analysis, electrical networks and graph theory. Research Synthesis Methods, **3**(4), 312–324.
6. Salanti, G., Ades, A.E., Ioannidis, J.P. (2011). Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial. Journal of Clinical Epidemiology, **64**(1878-5921), 163–171.
7. Rücker, G., Schwarzer, G. (2015). Ranking treatments in frequentist network meta-analysis works without resampling methods. BMC Medical Research Methodology, **15**, 58.
8. Mavridis, D., Giannatsi, M., Cipriani, A., Salanti, G. (2015). A primer on network meta-analysis with emphasis on mental health. Evidence Based Mental Health, **18**(2), 40–46.
9. Rücker, G., Schwarzer, G. (2015). Ranking treatments in frequentist network meta-analysis works without resampling methods. BMC Medical Research Methodology, **15**, 58,
10. Borenstein, M., Hedges, L.V., Higgins, J.P.T., Rothstein H.R. (2009). Introduction to meta-analysis (first). Chichester, UK: Wiley.
11. Thompson, S.G., Sharp, S.J. (1999). Explaining heterogeneity in meta-analysis: a comparison of methods. Statistics in Medicine, **18**(20), 2693–2708.
12. Cipriani, A, Higgins, J.P.T., Geddes, J.R., Salanti, G. (2013). Conceptual and Technical Challenges in Network Meta-analysis. Annals of Internal Medicine, **159**(2), 130–137,

13. Bucher, H.C., Guyatt, G.H., Griffith, L.E., Walter, S.D. (1997). The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. Journal of Clinical Epidemiology, **50**(0895-4356), 683–691.
14. Dias, S., Welton, N.J., Caldwell, D.M., Ades, A.E. (2010). Checking consistency in mixed treatment comparison meta-analysis. Statistics in Medicine, **29**(1097–0258), 932–944.
15. Donegan, S., Williamson, P., D'Alessandro, U., Tudur Smith, C. (2013). Assessing key assumptions of network meta-analysis: a review of methods. Research Synthesis Methods, **4**(4), 291-323.
16. Sterne, J.A.C., Egger, M., Moher, D. (editors). (2011). Sterne, J.A.C., Egger, M., Moher, D. (editors). Chapter 10: Addressing reporting biases. In: Higgins, J.P.T., Green, S. (editors). Cochrane Handbook for Systematic Reviews of Intervention. Version 5.1.0 (updated March 2011). The Cochrane Collaboration. Available from www.handbook.cochrane.org.
17. Nevill, R.E., Lecavalier, L., Stratis, E.A. (2018). Meta-analysis of parent-mediated interventions for young children with autism spectrum disorder. Autism, **22**(2), 84–98.
18. Tachibana, Y., Miyazaki, C., Ota, E., Mori, R., Hwang, Y., Kobayashi, E., *et al.* (2017). A systematic review and meta-analysis of comprehensive interventions for pre-school children with autism spectrum disorder (ASD). PLOS ONE, **12**(12), e0186502.
19. van Geel, M., Vedder, P, Tanilon, J. (2014). Relationship Between Peer Victimization, Cyberbullying, and Suicide in Children and Adolescents: A Meta-analysis. JAMA, **168**(5), 435–442.
20. Jadambaa, A., Hannah, T.J., Scott, J.G., Graves, N., Brain, D., Pacella, R. (2019). Prevalence of traditional bullying and cyberbullying among children and adolescents in Australia: A systematic review and meta-analysis. Aust. N. Z. J. Psychiatry, **53**(9), 878–888.
21. Belland, B.R., Walker, A.E., Kim, N.J., Lefler, M. (2017). Synthesizing Results From Empirical Research on Computer-Based Scaffolding in STEM Education: A Meta-Analysis. Review of Educational Research, **87**(2), 309–344.
22. Werner-Seidler, A., Perry, Y., Calear, A., Newby, J., Christensen, H. (2017). School-based depression and anxiety prevention programs for young people: A systematic review and meta-analysis. Clinical Psychology Review, **51**, 30–47.
23. Lawrence, P., Murayama, K., Creswell, C. (2019). Systematic Review and Meta-Analysis: Anxiety and Depressive Disorders in Offspring of Parents With Anxiety

Disorders. Journal of the American Academy of Child and Adolescent Psychiatry, **58**(1), 46–60.

24. Barker, F., de Lusignan, S., Cooke, D. (2016). Improving Collaborative Behaviour Planning in Adult Auditory Rehabilitation: Development of the I-PLAN Intervention Using the Behaviour Change Wheel. Annals of Behavioral Medicine: a Publication of the Society of Behavioral Medice, **52**(6), 489-500.
25. Stapersma, L., van den Brink, G., Szigethy, E., Escher, J., Utens, E. (2018). Systematic review with meta-analysis: anxiety and depression in children and adolescents with inflammatory bowel disease. Alimentary Pharmacology & Therapeutics, **48**(5), 496–506.
26. Belland, B.R., Walker, A.E., Kim, N.J. (2017). A Bayesian Network Meta-Analysis to Synthesize the Influence of Contexts of Scaffolding Use on Cognitive Outcomes in STEM Education. Review of Educational Research, **87**(6), 1042–1081.
27. Caldwell, D., Davies, S.R., Hetrick, S.E., Palmer J.C., Caro, P., José A López-López J.A., *et al*. (2019). School-based interventions to prevent anxiety and depression in children and young people: a systematic review and network meta-analysis. Lancet Psychiatry, **S2215-0366**(19), 30403–1.
28. Petropoulou, M., Schwarzer, G., Panos, A., Mavridis, D. (2019). NMAoutlier: Detecting outliers in network meta-analysis. R package version 0.1.7. URL `https://cran.r-project.org/web/packages/NMAoutlier`.
29. Lumley, T. (2002). Network meta-analysis for indirect treatment comparisons. Statistics in Medicine, **21**(16), 2313–2324.
30. Lu, G., Welton, N.J., Higgins, J.P.T., White, I.R., Ades, A.E. (2011). Linear inference for mixed treatment comparison meta-analysis: A two-stage approach. Research Synthesis Methods, **2**(1), 43–60
31. Lu, G., Ades, A.E. (2004). Combination of direct and indirect evidence in mixed treatment comparisons. Statistics in Medicine, **23**(20), 3105–3124.
32. Salanti, G., Higgins, J.P.T., Ades, A.E., Ioannidis, J.P.A. (2008). Evaluation of networks of randomized trials, Statistical Methods in Medical Research, **17**(3), 279–301.
33. Rücker, G., Schwarzer, G. (2014). Reduce dimension or reduce weights? Comparing two approaches to multi-arm studies in network meta-analysis. Statistics in Medicine, **33**(25), 4353–4369.

34. Schwarzer, G., Carpenter, J.R., Rücker, G. (2015). Meta-Analysis with R. Springer International Publishing.
35. Rücker, G., Krahn, U., König, J., Efthimiou, O., Schwarzer, G. (2019). netmeta: Network Meta-Analysis using Frequentist Methods. R package version 1.2-0. URL `https://CRAN.R-project.org/package=netmeta`.
36. DerSimonian, R., Laird, N. (1986). Meta-analysis in clinical trials. Controlled Clinical Trials, **7**(3), 177–188.
37. Jackson, D., White, I.R., Riley, R.D. (2012). Quantifying the impact of between-study heterogeneity in multivariate meta-analyses. Statistics in Medicine, **31**(29), 3805–3820.
38. Bollobas, B. (2002). Modern Graph Theory. Springer, Heidelberg/New York.
39. Krahn, U., Binder, H., König, J. (2013). A graphical tool for locating inconsistency in network meta-analyses. BMC Medical Research Methodology, **13**, 35.
40. Higgins, J.P.T., Jackson, D., Barrett, J.K., Lu, G., Ades, A.E., White, I.R. (2012). Consistency and inconsistency in network meta-analysis: concepts and models for multi-arm studies. Research Synthesis Methods, **3**(2), 98–110.
41. Lu, G., Ades, A.E. (2006). Assessing Evidence Inconsistency in Mixed Treatment Comparisons. Journal of the American Statistical Association, **101**(474), 447–459.
42. Hasselblad, V. (1998). Meta-analysis of multitreatment studies. Medical Decision Making: An International Journal of Society for Medical Decision Making, **18**(1) 37–43.
43. Gupta, A.K., Paquet, M. (2013). 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 Journal of Dermatology, **169**(2), 250–259.
44. Boland, A., Dundar, Y., Bagust, A., Haycox, A., Hill, R., Mujica Mota, R., *et al.* (2003). Early thrombolysis for the treatment of acute myocardial infarction: a systematic review and economic evaluation. Health Technology Assessment Winchester England, **7**(15), 1–136.
45. Higgins, J.P.T, Welton, N.J. (2015). Network meta-analysis: a norm for comparative effectiveness?. Lancet, **386**(9994), 628–630.
46. Caldwell, D.M, Gibb, D.M., Ades, A.E. (2007). Validity of indirect comparisons in meta-analysis. Lancet, **369**(1474–547X), 270.

47. Bafeta, A., Trinquart, L., Seror, R., Ravaud, P. (2013). Analysis of the systematic reviews process in reports of network meta-analyses: methodological systematic review. BMJ, **347**, f3675.
48. Bafeta, A., Trinquart, L., Seror, R., Ravaud, P. (2014). Reporting of results from network meta-analyses: methodological systematic review. BMJ, **348**, g1741.
49. Nikolakopoulou, A., Chaimani, A., Veroniki, A.A., Vasiliadis, H.S., Schmid, C.H., Salanti, G. (2014). Characteristics of Networks of Interventions: A Description of a Database of 186 Published Networks. PLOS ONE, **9**(1), e86754.
50. Song, F., Loke, Y.K., Walsh, T., Glenny, A.M., Eastwood, A.J., Altman, D.G. (2009). Methodological problems in the use of indirect comparisons for evaluating healthcare interventions: survey of published systematic reviews. BMJ, **338**, b1147.
51. Glenny, A.M., Altman, D.G., Song, F., Sakarovitch, C., Deeks, J.J., D'Amico, R., Bradburn, M., Eastwood, A.J. International Stroke Trial Collaborative Group. (2005). Indirect comparisons of competing interventions. Health Technology Assessment Winchester England., **9**(26), 1–134.
52. Donegan, S., Williamson, P., Gamble, C., Tudur-Smith, C. (2010). Indirect Comparisons: A Review of Reporting and Methodological Quality. PLOS ONE, **5**(11), e11054.
53. Wood, L., Egger, M., Gluud, L.L., Schulz, K.F., Jüni, P., Altman, D.G., *et al.* (2008). Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. BMJ, **336**(7644), 601–605.
54. Savović, J., Jones, H.E., Altman, D.G., Harris, R.J., Jüni, P., Pildal, J., *et al.* (2012). Influence of reported study design characteristics on intervention effect estimates from randomized, controlled trials. Annals of Internal Medicine, **157**(6), 429–438.
55. Sterne, J.A.C., Jüni, P., Schulz, K.F., Altman, D.G., Bartlett, C., Egger, M. (2012). Statistical methods for assessing the influence of study characteristics on treatment effects in 'meta-epidemiological' research. Statistics in Medicine. **21**(11), 1513–1524.
56. Rhodes, K.M., Turner, R.M., Higgins, J.P.T. (2015). Empirical evidence about inconsistency among studies in a pair-wise meta-analysis. Research Synthesis Methods. **7**(4), 346-370.
57. Turner, R.M., Davey, J., Clarke, M.J., Thompson, S.G., Higgins, J.P.T. (2012). Predicting the extent of heterogeneity in meta-analysis, using empirical data from the Cochrane Database of Systematic Reviews. International Journal of Epidemiology, **41**(3), 818–827.

58. Higgins, J.P.T., Thompson, S.G., Deeks, J.J., Altman, D.G. (2003). Measuring inconsistency in meta-analyses. BMJ, **327**(7414), 557–560.
59. Nüesch, E., Trelle, S., Reichenbach, S., Rutjes, A.W., Tschannen, B., Altman, D.G., Egger, M., Jüni, P. (2010). Small study effects in meta-analyses of osteoarthritis trials: meta-epidemiological study. BMJ, **341**, c3515.
60. Kjaergard, L.L., Villumsen, J., Gluud, C. (2001). Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. Annals of Internal Medicine, **135**(11), 982–989.
61. Dechartres, A., Trinquart, L., Boutron, I., Ravaud, P. (2013). Influence of trial sample size on treatment effect estimates: meta-epidemiological study. BMJ, **346**, f2304.
62. Egger, M., Juni, P., Bartlett, C., Holenstein, F., Sterne, J. (2003). How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study. Health Technology Assessment Winchester England, **7**(1), 1–76.
63. Song, F., Altman, D.G., Glenny, A.M., Deeks, J.J. (2003). Validity of indirect comparison for estimating efficacy of competing interventions: empirical evidence from published meta-analyses BMJ, **326**(1756-1833), 472.
64. Song, F., Xiong, T., Parekh-Bhurke, S., Loke, Y.K., Sutton, A.J., Eastwood, A.J., *et al.* (2011). Inconsistency between direct and indirect comparisons of competing interventions: meta-epidemiological study. BMJ, **343**(1756-1833), d4909.
65. Veroniki, A.A., Vasiliadis, H.S., Higgins, J.P.T., Salanti, G. (2013). Evaluation of inconsistency in networks of interventions. International Journal of Epidemiology, **42**(1464-3685), 332–345.
66. Chaimani, A., Vasiliadis, H.S., Pandis, N., Schmid, C.H., Welton, N.J., Salanti, G. (2013). Effects of study precision and risk of bias in networks of interventions: a network meta-epidemiological study. International Journal of Epidemiology, **42**(1464-3685), 1120–1131.
67. Jansen, J.P., Naci, H. (2013). Is network meta-analysis as valid as standard pairwise meta-analysis? It all depends on the distribution of effect modifiers. BMC Medicine, **11**, 159.
68. Dias, S., Sutton, A.J., Ades, A.E., Welton, N.J. (2013). Evidence synthesis for decision making 2: a generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials. Medial Decision Making, **33**(1552–681X), 607–617.

69. Jansen, J.P., Fleurence, R., Devine, B., Itzler, R., Barrett, A., Hawkins, N., *et al.* (2011). Interpreting indirect treatment comparisons and network meta-analysis for health-care decision making: report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices: part 1. Value Health: The Journal of International Society of Pharmacoeconomics and Outcomes Research, **14**(4), 417–428.
70. Jansen, J.P., Fleurence, R., Devine, B., Itzler, R., Barrett, A., Hawkins, N., *et al.* (2014). Indirect treatment comparison/network meta-analysis study questionnaire to assess relevance and credibility to inform health care decision making: an ISPOR-AMCP-NPC Good Practice Task Force report. Value Health: The Journal of International Society of Pharmacoeconomics and Outcomes Research, **17**(2), 157–173.
71. Efthimiou, O., Debray, T.P., van Valkenhoef, G., Trelle, S., Panayidou, K., Moons, K.G., *et al.* (2016). GetReal in network meta-analysis: a review of the methodology. Reserch Synthesis Methods, **7**(3), 236-63.
72. van Valkenhoef, G., Lu, G., de Brock, B., Hillege, H., Ades, A.E., Welton, N.J. (2012). Automating network meta-analysis. Research Synthesis Methods, **3**(4), 285–299.
73. Chaimani, A., Salanti, G. (2015). Visualizing assumptions and results in network meta-analysis: The network graphs package. Stata Journal, **15**(4), 905–950.
74. White, I.R. (2015). Network meta-analysis. Stata Journal, **15**(4), 951–985.
75. Cox, D.R., Stuart, A. (1955). Some Quick Sign Tests for Trend in Location and Dispersion. Biometrika, **42**(1–2), 80–95.
76. R Development Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL `http://www.R-project.org`.
77. Pohlert, T. (2020). trend: Non-Parametric Trend Tests and Change-Point Detection. R package version 1.1.2. URL `https://CRAN.R-project.org/package=trend`.
78. Chaimani, A., Salanti, G. (2012). Using network meta-analysis to evaluate the existence of small-study effects in a network of interventions. Research Synthesis Methods, **3**(2), 161–176.
79. Mavridis, D., Efthimiou, O., Leucht, S., Salanti, G. (2015). Publication bias and small-study effects magnified effectiveness of antipsychotics but their relative ranking remained invariant. Journal of Clinical Epidemiology, **69**, 161-9.

80. 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**(1), 51–66.
81. Li, T., Puhan, M.A., Vedula, S.S., Singh, S., Dickersin. K.; Ad Hoc Network Meta-analysis Methods Meeting Working Group. (2011). Network meta-analysis-highly attractive but more methodological research is needed. BMC Medicine, **9**(1), 79.
82. Dias, S., Welton, N.J., Sutton, A.J., Caldwell, D.M., Lu, G. Ades, A.E. (2013). Evidence synthesis for decision making 4: inconsistency in networks of evidence based on randomized controlled trials. Medical Decision Making: An International Journal of Society for Medical Decision Making, **33**(5), 641–656.
83. Chambers, J.D., Naci, H., Wouters, O.J., Pyo, J., Gunjal, S.J.D., *et al.* (2015). An assessment of the methodological quality of published network meta-analyses: a systematic review. PLOS ONE, **10**(4), e0121715.
84. Hedges, L.V., Olkin, I. (2014). Statistical Method for Meta-Analysis (first Edition). Orlando, FL: Academic Press.
85. Viechtbauer, W., Cheung M.W.-L. (2010). Outlier and influence diagnostics for meta-analysis. Research Synthesis Methods, **1**(2), 112–125.
86. Zhang, J., Fu, H., Carlin, B.P. (2015). Detecting outlying trials in network meta-analysis. Statistics in Medicine, **34**(19), 2695–2707.
87. Zhao, H., Hodges, J.S., Carlin, B.P. (2017). Diagnostics for generalized linear hierarchical models in network meta-analysis. Research Synthesis Methods, **8**(3), 333–342.
88. Noma, H., Gosho, M., Ishii, R., Oba, K., Furukawa, T.A. (2019). Outlier detection and influence diagnostics in network meta-analysis. ArXiv191013080 Stat.
89. Copas, J.B. (1988). Binary Regression Models for Contaminated Data. Journal of the Royial Statistical Society Series B (Statistical Methodology), **50**(2), 225–265.
90. Pincus, R., Barnett, V., Lewis, T. (1995). Outliers in Statistical Data (third edition). J. Wiley & Sons 1994, Biometrical Journal, **37**(2), 256–256.
91. Jackson, D., White I.R. (2018). When should meta-analysis avoid making hidden normality assumptions?. Biometrical Journal, **60**(6), 1040–1058.
92. Lee, K.J., Thompson, S.G. (2008). Flexible parametric models for random-effects distributions, Statistics in Medicine, **27**(3), 418–434.
93. Baker, R., Jackson, D. (2008). A new approach to outliers in meta-analysis. Health Care Management Science, **11**(2), 121–131.

94. Baker, R. Jackson, D. (2016). New models for describing outliers in meta-analysis. Research Synthesis Methods, **7**(3), 314–328.
95. Lin, L., Chu, H., Hodges, J.S. (2017). Alternative measures of between-study heterogeneity in meta-analysis: Reducing the impact of outlying studies. Biometrics, **73**(1), 156–166.
96. Yu, D., Ding, C., He, N., Wang, R., Zhou, X., Shi, L. (2019). Robust estimation and confidence interval in meta-regression models. Computational Statistics and Data Analysis, **129**, 93–118.
97. Gumedze, F.N., Jackson, D. (2011). A random effects variance shift model for detecting and accommodating outliers in meta-analysis. BMC Medical Research Methodology, **11**, 19.
98. Beath, K.J. (2014). A finite mixture method for outlier detection and robustness in meta-analysis. Research Synthesis Methods, **5**(4), 285–293.
99. Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. Journal of Statistical Software, **36**(3), 1–48.
100. Shi, L., Zuo, S., Yu, D., Zhou, X. (2017). Influence diagnostics in meta-regression model. Research Synthesis Methods, **8**(3), 343–354.
101. Mavridis, D., Moustaki, I., Wall, M., Salanti, G. (2016). Detecting outlying studies in meta-regression models using a forward search algorithm. Research Synthesis Methods, **8**(2), 199–211.
102. Huber, P. (1973). Robust Regression: Asymptotics, Conjectures and Monte Carlo. Annals of Statisitcs, **1**, 799–821.
103. Rousseeuw, P.J. (1985). Multivariate estimation with high breakdown point. The Annals of Mathematical Statistics.
104. Rousseeuw P.J. (1984). Least Median of Squares Regression. Journal of American Statistical Association, **79**(388), 871–880.
105. Cochran, W. (1954). The combination of estimates from different experiments. Biometrics, **10**, 101–129.
106. Chatterjee, S. Hadi A.S. (2013). Regression Analysis by Example (fifth Edition). ISBN: 978-1-118-45624-8. Wiley.
107. Cook R.D., Weisberg, S. (1982). Residuals and Influence in Regression. New York: Chapman and Hall.
108. Neter, J., Kutner, M., Nachtsheim, C., Wasserman, W. (1996). Applied Linear Statistical Models. Irwin: Chicago.

109. Hunter, J.E. (2004). Methods of Meta-Analysis: Correcting Error and Bias in Research Findings (second edition). Thousand Oaks, Calif: SAGE Publications, Inc.
110. Deeks, J.J., Higgins, J.P.T., Altman, D.G. (editors). (2008). Chapter 9: Analysing data and undertaking meta-analyses. In: Higgins, J.P.T., Green, S. (editors). Version 5.1.0 (updated March 2011). The Cochrane Collaboration. Available from: www.handbook.cochrane.org.
111. Matsushima, Y., Noma, H., Yamada, T., Furukawa, T.A. (2019). Bayesian influence diagnostics and outlier detection for meta- analysis of diagnostic test accuracy. arXiv:1906.10445.
112. Hadi, A.S. (1992). Identifying multiple outliers in multivariate data. Journal of the Royal Statistical Society: Series B (Statistical Methodology), **54**, 761–771.
113. Atkinson, A., Riani, M. (2000). Robust Diagnostic Regression Analysis. New York: Springer.
114. Atkinson, A. (1994). Fast Very Robust Methods for the Detection of Multiple Outliers. Journal of the American Statistical Association, 89(428), 1329–1339.
115. Atkinson, A., Riani, M., Cerioli, A. (2004). Exploring Multivariate Data with the Forward Search. New York: Springer.
116. Mavridis, D., Moustaki, I. (2008). Detecting Outliers in Factor Analysis Using the Forward Search Algorithm. Multivariate Behavioral Research, **43**(3), 453–475.
117. 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**(4), 1016–1034.
119. Dias, S., Welton, N.J., Caldwell, D.M., Ades, A.E. (2010). Checking consistency in mixed treatment comparison meta-analysis. Statistics in Medicine, **29**(7–8), 932–944.
119. König, J., Krahn, U., Binder, H. (2013). Visualizing the flow of evidence in network meta-analysis and characterizing mixed treatment comparisons. Statistics in Medicine, **32**(30), 5414–5429.
120. Kontopantelis, E., Reeves, D. (2012). Performance of statistical methods for meta-analysis when true study effects are non-normally distributed: A simulation study. Statistical Methods in Medical Research, **21**(4), 409–426.
121. Brockwell, S.E., Gordon, I.R. (2001). A comparison of statistical methods for meta-analysis. Statistics in Medicine, **20**(6), 825–840.
122. Filzmoser, P. (2005). Identification of Multivariate Outliers: A Performance Study. Austrian Journal of Statistics, **34**(2), 127–138.

123. Knight, N.L., Wang, J. (2009). A Comparison of Outlier Detection Procedures and Robust Estimation Methods in GPS Positioning. Journal of Navigation, **62**(4), 699–709.
124. Hardin, J., Rocke, D.M. (2004). Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator. Computational Statistics and Data Analysis, **44**(4), 625–638.
125. Cook, R., Holschuh, N., Weisberg, S. (1982). A Note on an Alternative Outlier Model. Journal of the Royal Statistical Society: Series B (Statistical Methodology), **44**(3), 370–376.
126. Thompson, R. (1985). A Note on Restricted Maximum Likelihood Estimation with an Alternative Outlier Model. Journal of the Royal Statistical Society: Series B (Statistical Methodology), **47**(1), 53–55.
127. Harville, D.A. (1977). Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. Journal of American Statistical Association, **72**(358), 320–338.
128. Higgins, J.P.T., Jackson, D., Barrett, J.K., Lu, G., Ades, A.E., White, I.R. (2012). Consistency and inconsistency in network meta-analysis: concepts and models for multi-arm studies. Research Synthesis Methods, **3**(2), 98–110.
129. White, I.R., Barrett, J.K., Jackson, D., Higgins, J.P.T. (2012). Consistency and inconsistency in network meta-analysis: model estimation using multivariate meta-regression. Research Synthesis Methods, **3**(2), 111–125.
130. Jackson, D., Barrett, J.K., Rice, S., White, I.R., Higgins, J.P.T. (2014). A design-by-treatment interaction model for network meta-analysis with random inconsistency effects. Statistics in Medicine, **33**(21), 3639–3654.
131. Lifeng, L., Haitao, C. (2020). altmeta: Alternative Meta-Analysis Methods. R Package Version 2.3. `URL https://CRAN.R-project.org/package=altmeta`.
132. Beath, K.J. (2016). metaplus: An R Package for the Analysis of Robust Meta-Analysis and Meta-Regression. The R Journal, **8**(1), 5–16.