Should Network Meta-Analysis Become the Standard in Evidence-Based Clinical Practice?

by

Romina Brignardello-Petersen

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy

Institute of Health Policy, Management and Evaluation
University of Toronto

Should Network Meta-Analysis Become the Standard in Evidence-Based Clinical Practice?

Romina Brignardello-Petersen

Doctor of Philosophy

Institute of health Policy, Management and Evaluation
University of Toronto

2016

Abstract

Network Meta-analysis (NMA) is a statistical tool that allows comparing and estimating the relative effects of multiple interventions at the same time. Although its popularity has increased in the latest years, many questions remain unsolved before it can be proposed as the standard method of evidence synthesis for informing evidence clinical practice. In this thesis, using systematic surveys of the literature, we aimed to determine 1) the extent to which NMA can be used to answer current clinical questions; 2) whether systematic reviews (SRs) using NMA report the same results as SRs using head-to-head comparisons (HTHC); and 3) the robustness of the rankings obtained from NMA to the exclusion randomized clinical trials from the network, and the impact of increasing decision thresholds on the ranking probabilities. We observed that only 25.3% (205 out of 809 Cochrane SRs published in a 1-year period) of the SRs had questions in which an NMA was necessary; that SRs using NMA to assess the effects to stents in patients undergoing percutaneous coronary intervention reported the same results as SRs using HTHC in 44.8% to 83.3% of the cases; that rank probabilities and the rankings remain reasonably constant when excluding trials from the analysis (with an overall mean absolute change of 4.3% in rank probabilities across NMAs), but there are cases in which dramatic increases or decreases of the rank probabilities, and switches in the treatments ranked first and second can be observed; and

that increasing the threshold to claim superiority may result in important changes in rank probabilities, which in some cases lead to the first treatment having extremely low probabilities of being distinguishable as the best. These issues suggest that because NMA is still in its infancy compared to HTHC, more research and guidance for its use are necessary before it can be claimed that NMA should become the standard for comparing treatment effectiveness.

Acknowledgments

This thesis was possible thanks to the contribution and support of many people. In particular, I am very grateful to my supervisor, Dr. George Tomlinson, who guided me through the whole process. His knowledge, dedication, assistance, and patience, made this learning experience very pleasant. I am also thankful for the help provided by my thesis committee members, Alejandro Jadad and Bradley Johnston. Their feedback and suggestions shaped this thesis, and made the final product much better than I expected. In addition, I cannot go without mentioning those who helped me discussing ideas that were important for the development of these studies, in particular Dr. Gordon Guyatt and Dr. Alonso Carrasco-Labra.

Finally, I should express my gratitude to the Bicentennial Becas Chile Scholarship program, from the Government of Chile, who provided the financial support for me to pursue this degree.

Table of Contents

List of Tables

List of Figures

List of Appendices

Chapter 1
Introduction and overview

1 Introduction and aims

For most clinical conditions, there are several possible treatments. Decisions regarding the optimal treatment for a given patient population are informed by studies assessing the available interventions.[1] For most clinical research questions there are often many available treatments to choose from. Systematic reviews (SRs) of well-designed and conducted randomized controlled trials (RCTs) have been recognized as providing the strongest evidence for supporting treatment decisions.[2] Unfortunately, most systematic reviews of RCTs focus on pair-wise comparisons of treatments, which makes it difficult to determine what is the best treatment for a given condition.[3]

SRs comparing multiple interventions and their quantitative synthesis, called "Network Meta-Analysis" (NMA), provide a broad and inclusive picture of the evidence, which allows inferences about all treatments available for a clinical condition and population.[4] An NMA is defined as an evidence base that consists of two or more RCTs connecting more than two interventions.[5] It is an extension of the traditional meta-analysis that includes multiple head-to-head comparisons (HTHC) and combines them statistically, to provide estimates about the relative effectiveness of the treatments.[4 5]

Among the advantages of NMA are the ability to compare interventions that have not been compared directly in RCTs, the improvement in the precision of the treatment effect estimates, and the capability of ranking the treatments.[4-7] On the other hand, there are disadvantages such as the observational nature of the indirect comparisons, technical difficulties related to the assumptions underlying the models and the statistical expertise needed, and the risk of combining studies that are too different from each other [3 4]. The potential disadvantages of NMAs have been a source of controversy, and methodologists have questioned if the NMA method of evidence synthesis is better than a traditional meta-analysis based on HTHC of two treatments.

Traditional meta-analysis faced similar criticisms when they were first introduced,[8] and so these disadvantages do not necessarily present insurmountable challenges. In recent years, NMA has become very popular, and this increased uptake has led to the publication of articles providing guidance on its reporting[9] and use.[10-12] Even more, it has been proposed that NMA should become the standard for comparing the effectiveness of multiple treatments.[13] However, the extra complexity in the conduct and interpretation NMA still makes its implementation more difficult for clinical decision-making than approaches to SRs that use the traditional HTHC.[12]

Therefore, the balance between the compatibility of NMA with the clinical decision making process and the level of training required to make use of it will be fundamental in its implementation. In clinical contexts in which there are more than two acceptable courses of action to choose from, it would be reasonable to think that NMA would be widely used. Because this is a relatively new technique, however, it is still under development and there are some methodological issues that have not been addressed yet. In addition, performing and interpreting the results of NMA appears to be a process more complex when compared to a traditional HTHC analysis. These two concerns suggest that NMA is not ready to become the standard in informing evidence-based clinical practice, and its future uptake may depend on how the two forces- the compatibility with clinical decision-making and the potential methodological and users issues- interact with each other.

To date, most of the research regarding NMA has focused on technical aspects.[14-28] Only a few studies have addressed practical issues and implications of these methods in the clinical setting.[29 30] This thesis has three aims, each of which will provide insight about potential issues that may affect the uptake of NMA including: (1) Identifying the extent to which NMA can be used to answer current clinical questions; (2) Assessing whether the results obtained when using NMA are similar to those observed in traditional SRs with HTHC; and (3) Exploring whether the rankings obtained from NMA are robust to the omission of a single trial from the analysis, or if the use of different thresholds to claim superiority between two treatments have impact the rankings.

2 Overview

Chapter 2 provides background information relevant to the thesis and its methods. First, we define NMA and describe its main features, advantages and disadvantages. Second, we describe how a SR with NMA is designed, conducted, and presented. The underlying assumptions of NMA and how violations of these assumptions may affect the results of NMA are also described. Third, we focus on the particular feature of rankings, their benefits, and the controversies around them. Chapter 3 provides information relevant to some of the specific methods of the thesis. It introduces systematic surveys, and it provides an overview on how to compare two treatment effects for the same comparison that were obtained from two different study designs.

Chapters 4-6 present the three papers of this thesis. Chapter 4 is a systematic survey of SRs published in the Cochrane Database of Systematic Reviews over a 1-year period, in which the clinical question of the SRs was assessed to determine whether it would have been necessary to use NMA to obtain an answer. Chapter 5 is a systematic survey, in which we compare the results reported by SRs using NMA and SRs using HTHC that aim to answer the same clinical question. Chapter 6 is a systematic survey in which we explored whether the probability of a treatment being ranked best changes when a single trial is excluded from the NMA, or when the threshold to define a difference between two treatments is different than a relative effect of 1.

Finally, in Chapter 7 we summarize the results, and discuss the strengths, limitations, implications for research, implications for practice, and conclusions drawn from the three papers.

Chapter 2
Background

3 Network meta-analysis

3.1 Definition

Network-meta-analysis (NMA), also known as “Multiple Treatment Comparisons” or “Mixed Treatment Comparisons” is a technique that allows quantitative summaries of the results of a systematic review (SR). It is an extension of a traditional meta-analysis, in which all included studies compare the same intervention against the same control, allowing a comparison of only two treatments at a time. In an NMA multiple head-to-head comparisons (HTHC) are done across multiple interventions in the same analysis.[31]

The term NMA encompasses any meta-analysis in the context of a network of evidence. That is, every time there are more than two randomized controlled trials (RCTs) connecting more than two interventions, an NMA can be performed.[5 9 32] The networks can have different shapes depending on the number of interventions and how they are connected, which depends on which interventions have been compared in RCTs. The simplest possible network is formed by two RCTs in which two interventions have been compared with the same control (Figure 2.1A). In other networks, there are more than three treatments, but all interventions have been compared against the same control (Figure 2.1B). If all treatments in the network have been compared against the others directly, they form a closed loop or connected network (Figure 2.1C). Finally, if some of the treatments have been compared directly with all the others, but some of them have not, there is a complex network (Figure 2.1D).

NMA combines both direct and indirect evidence to obtain estimates of the relative effectiveness of all pairwise comparisons formed by all the interventions from the network.[31] Some may argue that in some cases there would only be indirect evidence to inform a pairwise comparison (Figures 2.1A and 2.1B), therefore making them only adjusted indirect comparisons. However, since they can be displayed as a network of

evidence and analyzed using of NMA methods, these cases fall under the umbrella of NMA.[5 32]

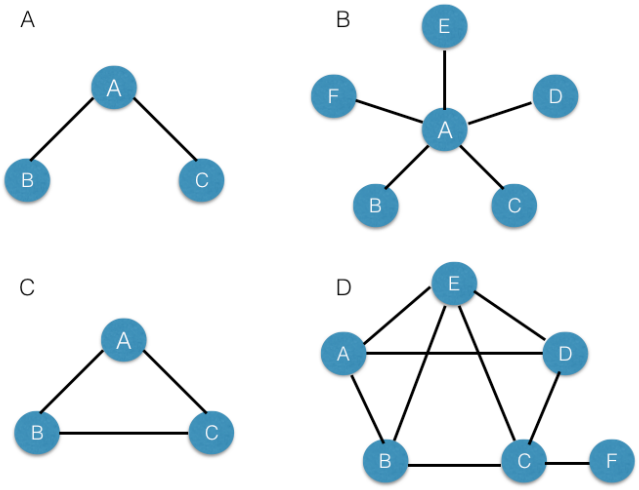


Figure 2.1: Basic shapes of networks.

The circles or nodes represent interventions and the solid lines represent studies that have compared them directly. A: Simplest possible network, with three interventions where two of them have been compared to the same control. B: Star-shaped network, in which many interventions have only been compared with the same control. C: Closed loop or connected network, with all the interventions directly compared against the others in RCTs. D: Complex network, in which some interventions have been compared directly and others have not.

3.2 Advantages and disadvantages

Among the benefits of NMA is the possibility of comparing two treatments that have not been compared directly in RCTs, a feature which is particularly useful in scenarios where most treatments have been compared to placebo or a well-established standard of care.[4 17 19] Second, from a statistical perspective, combining direct and indirect evidence may lead to more precise estimates. This would be useful in cases in which, even if there is direct evidence to compare two treatments, this evidence is not strong enough to yield

estimates with good precision.[33 34] Increased precision for the clinician using the results of NMA to inform clinical practice should give more confidence in a given treatment effect estimate. At the policy-making level, less uncertainty around the treatment effect measures used leads to more focused results in cost-effectiveness analyses.[35] Third, it has been claimed that NMA would be more compatible with clinical and policy-related decision-making by considering all the relevant options and providing a probability of each of the treatments being the best or worst for a specific outcome (rankings), which would facilitate the decision-making process.[31 32]

There are some disadvantages or limitations, however, to the use of NMA. First, the indirect comparisons that are performed are, by nature, observational.[6 34] Even though patients are randomized to interventions within an RCT, the interventions and patients are not randomized across trials.[36] Second, the usual and most useful NMA assumes that the direct and indirect evidence informing a specific comparison is consistent (that os, in agreement with each other), which may not always be the case.[20 28 30] Third, there is extra expertise needed to perform and use NMA, which means that SR authors and users need to acquire new skills to develop and use this method.[12] Finally, the field is evolving rapidly, and researchers are constantly proposing new approaches to deal with many issues, such as the assessment of bias, missing data, repeated measures, publication bias, and other issues that have already been well-explored in reviews using HTHC. This makes it harder for review authors and users to keep up to date with the methods to be able to conduct NMA and apply their results to inform clinical decisions.[23]

4 Conducting SRs with NMA

4.1 Formulating the question and searching for studies

NMAs are performed in the context of SRs, just like the traditional HTHC.[9] By allowing the reviewer to compare the relative effectiveness of more than two treatments, the clinical questions that can be answered using NMA are, by nature, broader than those that can be answered using a HTHC. The broad approach leads to an increase in the applicability of the results, but at the same time, it could lead to including studies in

which the participants, the interventions or the outcomes are too dissimilar, leading to important heterogeneity among the studies.[12]

From a theoretical perspective, NMA can be used to answer any clinical question in which there is need for a network of evidence. In the simplest case, SR authors may be interested in performing an indirect comparison between two treatments that have only been compared against a common comparator and not directly (Figure 2.1A). In this case, the NMA would allow calculation of adjusted indirect comparison, but would not be able to take advantage of the benefits that a more complex network would provide (such as the ones cited in Section 3.2). If authors are interested in comparing one treatment against many others, or many treatments against one control, they could also use an NMA (Figure 1B). In these cases, depending on their eligibility criteria for the studies, they could have a star-shaped network (if they only search for and include studies that compare directly the one treatment of interest –or the control- against the others) or a more complex, connected network (if they also search for indirect evidence, from studies that compare the other treatments against each other to inform the comparisons they are interested in). It can be argued, however, that when the interest is only in pairwise comparisons for which there is direct evidence available, even if it is possible to perform an NMA, it would not be necessary to do it.

Since in an NMA there are more interventions of interest, the searching and study selection process is likely to be more burdensome when compared to the same steps in a more focused review. Nevertheless, guidance with regards to how to conduct and report a search from a SR that uses an NMA do not propose any extra steps in the searching process: the date span, restrictions with regards to language and years searched, and eligibility criteria should all be justified.[12 37] In addition to these aspects, the databases searched, the search terms, supplementary searches and a flow diagram should be reported.[9 12 37] Therefore, the added burden in the process comes from the increased number of references to screen after the searches are performed, due to the broadness of the clinical question that the SR is trying to answer.[38 39]

Although some strategies have been proposed to make the searching more efficient, such as iterative searches that keep going on depending on the results of the current stage,[38] these have been suggested in the context of indirect comparisons and thus would be potentially useful to only some NMAs. In addition, proponents of these approaches still recognize the importance of including all the relevant evidence when doing NMAs.[39]

In order to use resources more efficiently, some authors have also proposed using SRs of the pairwise comparisons included in the network to identify primary studies, if these SRs are up to date.[40] There is no research determining whether this has an impact on the results of the NMA.

4.2 Data synthesis

4.2.1 Structure of the network

The first step to synthesizing the results of all the RCTs included is to create the network of evidence. The plot of the network shows what interventions have been compared directly with others and provides a general understanding of the amount of direct and indirect comparisons, what treatments have been studied more than others, and what comparisons have been performed more than others, among others.[36 41] Figure 2.2 illustrates a network plot. Some plots can include extra information, such as the overall risk of bias of the studies comparing two treatments, or the years in which the studies were published, by using different colors to depict the lines and nodes. Other plots use multiple lines to illustrate the number of studies addressing a comparison while using the thickness of the lines to illustrate the number of patients.[36]

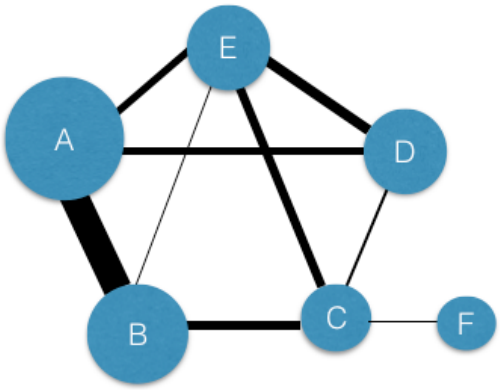


Figure 2.2: Network plot.

Each node represents a treatment included in the network. The solid lines represent direct evidence available comparing the treatments connected by the lines. The size of the nodes represent the number of patients randomized to such interventions, while the thickness of the lines represent the number of studies comparing directly the two treatments connected by the lines. Where there are no lines, the relative effectiveness of those two treatments is estimated using indirect evidence.

The geometry of the network informs the assessment of the confidence in the estimates of effect,[10 12] that is, how confident we are that the estimates obtained from the NMA are correct (or close to the "truth").[42] The presence of direct evidence, that tends to contribute more to the final relative effect estimate than indirect evidence, increases the confidence in such estimates of effect.[12]

4.2.2 Assumptions of NMA

NMA has the same assumptions as the traditional pairwise meta-analysis. These assumptions, however, must be met across the network of trials.[20] The two main underlying assumptions of NMA are transitivity (or similarity) and consistency (or coherence).[43]

Transitivity means that an estimate of the relative effectiveness of treatment A versus treatment B can be obtained through another treatment C, by using information of the

comparisons A versus C (AC trials) and B versus C (BC trials).[36] This can be evaluated by checking that potential effect modifiers are similarly distributed in AC and BC trials so that the estimate of A versus B resulting from these is valid.[4 36] An effect modifier is any characteristic of the studies that may vary across them, and thus, it is reasonable to question transitivity when effect modifiers are different between AC and BC trials. Examples of these are patient characteristics that may influence the treatment effect, such as mean age, sex distribution or severity of the disease.

Other interpretations of transitivity include assuming that the intervention C is similar in AC and BC trials (therefore, differences in the intervention C, such as different dosage or regimen, may cause intransitivity); that participants included in the studies could have been randomized to any of the treatments, A, B or C (and thus, that the three treatments have the same indications); and that the "missing" treatment in each trial (treatment B from AC trials and treatment A from BC trials) was missing at random.[4]

Judgment should be used when assessing the plausibility of the transitivity assumption.[43] When the transitivity assumption is not met, the indirect estimates of the relative effectiveness of A versus B are going to be biased.[11] If the potential effect modifiers are reported in the included RCTs, and there are enough RCTs, a network meta-regression can be used to adjust for such modifiers and improve the transitivity of the network.[36]

Consistency is the statistical agreement between the direct and indirect estimates of effect.[20 30 36 43] When the estimate of the relative effectiveness of A versus B is not statistically significantly different to the difference between the estimates AC and BC, then the loop ABC is consistent.[11] As a "statistical manifestation of transitivity",[43] checking the consistency assumption is an indirect method to check the transitivity assumption.

There are many methods to check the consistency at both, the "loop" level (for a specific comparison) and the "global" level (for the whole network).[36] If the consistency assumption does not hold, the estimates of effects obtained from the NMA are less

reliable and harder to interpret, however, consistency issues may affect only some parts of the network.[4]

4.2.3 Statistical methods of NMA

NMA can be conducted using fixed-effects or random-effects models with either frequentist or Bayesian frameworks.[44] The assumptions behind these are the same as in a traditional pairwise meta-analysis but extended to the NMA. Fixed-effects models assume that there is a constant relative treatment effects across studies for a pairwise comparison. In other words, any variation among the treatment effects estimated for one comparison is due only to chance. Therefore, the aim when using a fixed-effects model is to estimate the one "true" treatment effect.[31] A random-effects model assumes that the relative effects across studies comparing the same two treatments are a sample from a distribution of true treatment effects, where the mean is the pooled relative effect and the standard deviation is the heterogeneity across studies. Thus, when using a random-effects model, the aim is estimation of the average of the true treatment effects.[5 32]

NMA models can be implemented using three main approaches:[36]

1. Meta-regression: this approach can be used by either treating the different pairwise comparisons as covariates in a meta-regression model,[19] or by using a two stage method.[45] In the first stage, a set of regular pairwise meta-analysis is performed to obtain the direct estimates of treatment effects. In the second stage the direct estimates and their variances are used to find pooled effects parameters that satisfy consistency equations through a weighted linear regression.

2. Hierarchical model: these model the observed outcome data - not treatment effect- using an arm-based approach (for example, dichotomous outcomes can be modeled using the binomial distribution while continuous data can be modeled using a normal distribution and count data can be modeled using a Poisson distribution). The estimated arm-level parameters are then used to estimate overall treatment effects.[19]

3. Multivariate meta-analysis model: in these, the pairwise comparisons are treated as different outcomes, and multiple-outcome meta-analytical techniques are used to model these outcomes.[46]

4.2.4 Software to perform NMA

Depending on the framework to use, there are many choices of software to perform NMA. If using the frequentist framework, SR authors can use SAS (through the use of *proc glimmix* or *proc mcmc*) or STATA (with the *nvmeta* command). In the Bayesian framework, code exists for running NMA models in software such as WinBUGS, Open BUGS and JAGS. To facilitate the NMA process and minimize the burden of programming, some packages have been developed for use in the R environment that relieve the analyst of the need to program in JAGS or BUGS.[44]

The three packages specifically designed to perform NMA in R are *netmeta*, which allows implementation of frequentist methods based on graph theory;[47 48] *pcnetmeta*, which implements a multivariate meta-analysis using Bayesian methods;[49] and *gemtc*, that fits hierarchical generalized linear models in a Bayesian framework.[34] Both, *pcnetmeta* and *gemtc* interact with JAGS to perform the analyses. All three packages allow use of fixed and random-effects models and supply functions for plotting a network.[44]

Despite the fact that these R packages aim to eliminate the need to program the actual Bayesian models, there is still some expertise required for using them. Basic knowledge of the R environment[50] is necessary, and this is a skill that most clinical researchers do not have. NMA cannot be implemented in software that would be considered user-friendly by systematic reviewers without some background in biostatistics; there is nothing for NMA equivalent to the software developed by The Cochrane Collaboration (RevMan) for standard systematic reviews.[51]

4.2.5 Results from an NMA

In the results section, SRs using NMA should present the network plot (see Section 4.2.1 and Figure 2), the relative effects of all the pairwise comparisons, and the rankings. In

addition, results related to the assessment of assumptions and extra analyses such as sensitivity analyses can be found in many published NMAs.[9]

4.2.5.1 Results from pairwise comparisons

The results from pairwise comparisons can be presented either in tables or plots. In both, the point estimates and confidence or credible intervals must be presented.[9] Table 2.1 shows an example of a "League Table", based on the data published by Lu and Ades.[52] Although in the example the lower and upper triangles present reciprocal values (the outcome from the example is continuous and box B,A presents the relative effect of B compared to A, while the box A,B present the relative effect of A compared to B), authors can use the lower and upper triangles to display results of different outcomes, or results from the direct comparisons and network estimates as well.

Table 2.1: Example of a League Table. Mean differences and 95% credible intervals of relative-effects of 4 different smoking cassation treatments.

	A	**B**	**C**	**D**
A		0.49 (-0.3, 1.34)	0.83 (0.39, 1.35)	1.1 (0.27, 2.01)
B	-0.49 (-1.34, 0.3)		0.34 (-0.48, 1.17)	0.6 (-0.31, 1.58)
C	-0.83 (-1.35, -0.39)	-0.34 (-1.17, 0.48)		0.26 (-0.53, 1.12)
D	-1.1 (-2.01, -0.27)	-0.6 (-1.58, 0.31)	-0.26 (-1.12, 0.53)	

Figure 2.3 illustrates how the pairwise comparisons are presented in forest plots, in the same way as those in a traditional pairwise SR. The data comes from the same example of smoking cessation treatments published by Lu and Ades, this time for a dichotomous outcome[52].

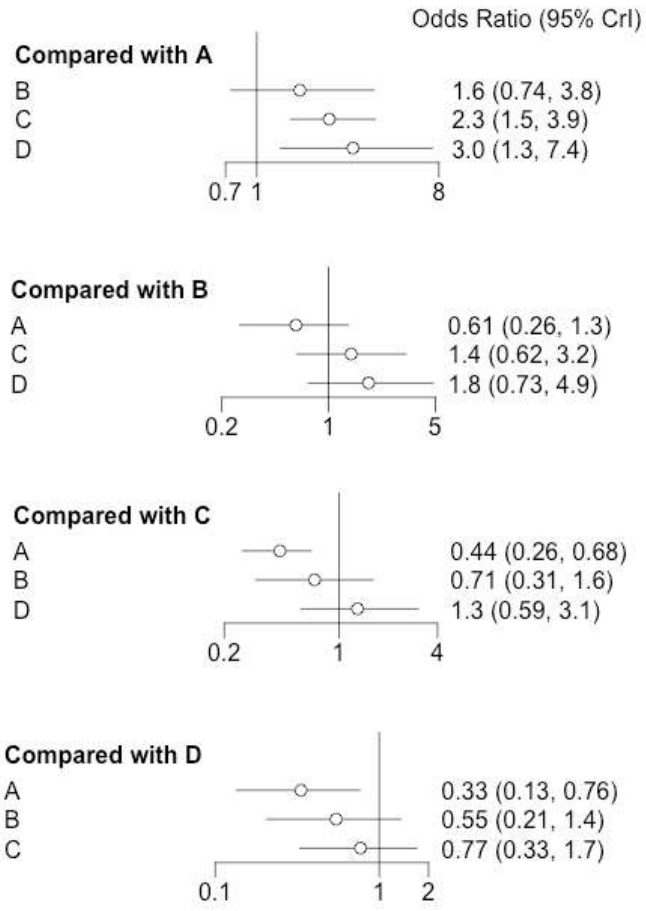


Figure 2.3: Forest plots presenting the results of pairwise comparisons of treatments for smoking cessation treatments.

4.2.5.2 Rankings

When using a Bayesian framework to analyze the data, NMA also provides rankings, based on the probability of each treatment being the best, second best, third best, etc. Ranking results can be presented as a table (Table 2.2) or as graphs (Figure 2.4).

Table 2.2: Example of a table presenting rank probabilities and rankings. Probability of each treatment of being the best, second best, third best and worst (based on smoking cessation data example from Lu and Ades)

Treatment and corresponding ranking probabilities	**Ranking**			
	1	**2**	**3**	**4**
A	0.00	0.00	0.11	0.89
B	0.06	0.18	0.66	0.10
C	0.23	0.6	0.17	0.00
D	0.71	0.22	0.06	0.00

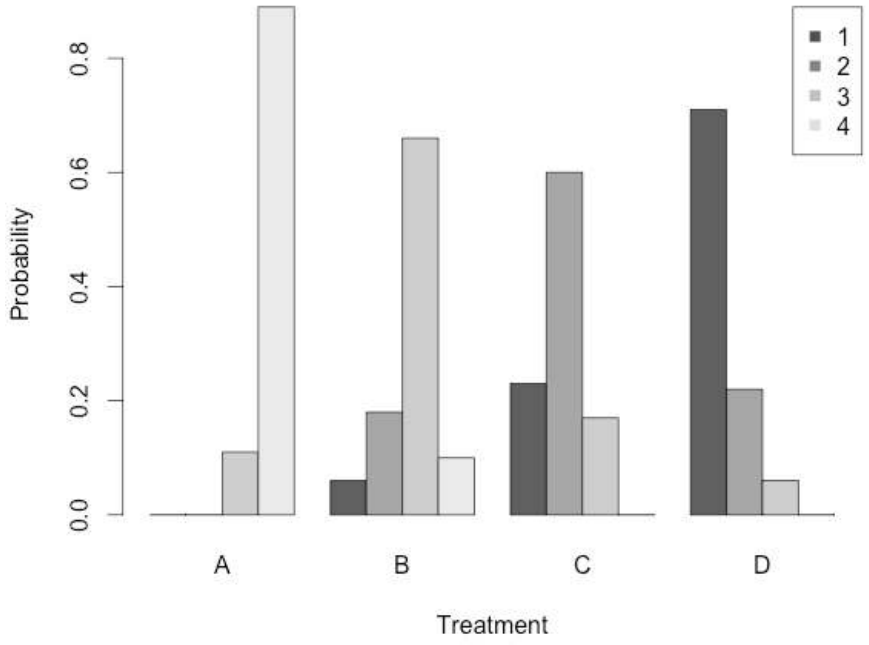


Figure 2.4: Plot of rankings based on smoking cessation data example from Lu and Ades. Bars represents, respectively, the probability of being the best, second best, third best and worst treatment for the specific outcome.

Although rankings are frequently cited as one of the main advantages of NMA, because they can facilitate decision-making by providing another piece of evidence that seems to be compatible with the way clinicians think,[4 5 13] some concerns have been raised with regards to the use of rankings.[53] First, there are concerns with respect to the fact that when rankings are presented, these are based on any differences among the treatment effects, no matter how small. Therefore, it is possible that a treatment with a high probability of being better than another treatment is better only by a small, not clinically meaningful amount.[12 40] Second, the probability of a treatment being the best may change if a new trial or treatment is introduced into the network, especially when the network is sparse.[12 54]

Chapter 3
Methods used in the thesis

5 Systematic surveys

Systematic surveys, also called methodological surveys, have been used to study many questions related to research methodology. Some examples include systematic surveys to explore issues around missing data,[55-57] randomization,[58 59] reporting of absolute estimates of effect,[60] the robustness of statistically significant results,[61] and subgroup analyses,[62] among others.

Although there are no established methods or guidance to perform these surveys, what most of them have in common is that they follow most of the methods of SRs. A search strategy is constructed and implemented in the databases of interest, and the screening of the resulting references is done in duplicate and independently. The main difference with a traditional systematic review is that systematic surveys are not trying to answer a clinical question and estimate a treatment effect, but to provide an overview and synthesis of aspects related to research methodology.

Authors have summarized the usefulness of systematic surveys in three main areas:[63]

1. Analyze and summarize a methodological argument: these would follow the methodology used in effectiveness SRs strictly; however, they would not need to be as exhaustive as these SRs.

2. Determine whether there are differences when studying the same question using different methods: these use the literature as a source of data to assess whether different effects in outcomes are observed depending on the approach used to estimate these effects (for example, whether outcomes seem to be different in observational studies versus RCTs). The focus of these would be in the magnitude, precision and/or statistical significance of the effect estimate.

3. Assess the relative merits of different methods: in these, new data is collected either by surveying the literature or conducting primary research. These would focus in advantages and disadvantages, in general, of using the different methods.

Despite this categorization, it is recognized that there are not well-defined methods to perform methodological research, and that systematic surveys are seen as a middle point between effectiveness SRs and narrative reviews, in the sense that they are looking for new ideas in a systematic way.[63]

6 Comparison of estimates of treatment effects obtained using two different methods

With increasing volumes of clinical research being published, it is common to find several reports of studies that aim to answer the same clinical question. Even though systematic reviews address this issue by summarizing all the available literature regarding a clinical question,[64] it is not unusual to find more than one systematic review answering the same question. Despite the great amount of guidance to users of the scientific literature [42 65 66], there seems to be no explicit guidance available for readers who wish to reach their own conclusions regarding the similarity between the results of different studies.

Acknowledging the fact that different study features may cause the estimates of effects to differ among studies answering the same clinical question, researchers have performed systematic surveys that formally compared the results obtained when using different study designs [67-69], and different aspects within a study design, such as the statistical methods to analyze the results of the study [70], or a particular strategy to minimize the risk of bias [71 72]. Unfortunately, the methods and criteria used for making these comparisons are diverse and there is no unified approach.

6.1 Methods used by authors of systematic surveys to compare the results from different study designs

A scoping review was performed as a complement to this thesis. In that review, we included studies of dichotomous outcomes in which authors compared the estimates of

effects from different study designs. We performed searches in electronic databases and in the list of references of relevant studies. Two reviewers independently selected studies and abstracted data. We created a list of the criteria used to compare estimates of effects between study designs, described their main features, and classified them using a clinical perspective.

We included 26 studies,[67-69 73-95] from which we identified 24 criteria to compare the estimates of treatment effects from two study designs. Most of the studies focused on comparing estimates from observational studies and randomized controlled trials (n=19). The most common criteria aimed to determine whether there was a difference or not (n=18), provided guidance for such a judgment (n=16), and were based on the point estimates (n=11).

We classified all of the criteria that had been used as either appropriate or inappropriate. A criterion was judged inappropriate either because it was not useful or because it could be misleading, when used as by itself. Criteria deemed not useful were those that did not provide explicit guidance to determine that two treatment effects were different. Potentially misleading criteria were defined as those that, although appropriate in some cases, would lead to claim that two effect estimates are different when they may not be. We judged fourteen criteria to be appropriate, and classified them as either statistically-related or clinically-related (Table 3.1).

Table 3.1: Classification of the 14 criteria for comparing two estimates of treatment effects judged to be appropriate

Category	Subcategory	Criterion
Statistically-related	Based on statistical significance	**1.** Z test to determine whether the estimates from each design, per topic, are different beyond chance
		2. Z test to determine whether the estimates from each design, across topics, are different beyond chance
		3. Test for interaction between the effect estimate and the study design
		4. Binomial (sign) test to determine whether one design is more likely to report beneficial effects than the other
		5. Meta-regression to determine whether study design is a statistically significant predictor of size of effect estimate
	Based on interpretation of statistical significance or other statistical measure or coefficient	**6.** Disagreement regarding statistical significance between estimates from each design, per topic
		7. Change in I^2 from meta-analysis using all estimates versus meta-analysis that accounts for study design using subgroups is higher than 10%
		8. Coefficients that compare the estimates of effect from each design
Clinically-related	Provides rules for clinical interpretation	**9.** Estimate from one design, per topic (in the OR or RR scale) is at least double or half the estimate from the other design
		10. Difference in direction of effect
		11. Visual comparison of estimates from each design, per topic, and consensus about difference
		12. ROR or RRR is lower than 0.7 or higher than 1.43
	Requires interpretation	**13.** Pooled estimate from a meta-analysis of the ROR or RRR calculated per topic
		14. Difference in RR reduction or increase between designs, per topic

6.2 Limitations of the methods to compare two estimates of treatment effect

We identified 14 criteria that could be appropriate when used individually, but among the inappropriate criteria we found some that have been used which may be misleading for claiming that two estimates of effect are different. For example, the use of the overlap of confidence intervals as a requirement for similarity may be misleading in a hypothetical scenario in which we see two very precise pooled effect estimates of reduction on an outcome (we can imagine one OR 95% CI ranging from 0.40 to 0.43 and the other ranging from 0.45 to 0.48, which in many circumstances, on the basis of clinical expertise, would be interpreted as very similar effect estimates). We also came across criteria that were not useful without being accompanied by other pieces of relevant information. For instance, the description of the difference between relative risk reductions or increases with each design needs extra information to be interpreted, such as the knowledge of the minimally important difference (for continuous outcomes) or decision threshold (for dichotomous outcomes) for the specific outcome under consideration.

The main limitation related to the criteria judged to be appropriate and classified as clinically-related has to do with their applicability to different clinical areas. Some of these criteria did not provide specific rules for their interpretation, making it necessary to have a deeper understanding of the clinical problem and the minimal important difference or decision threshold (if there is any established) to determine whether two estimates are different or not. Furthermore, it is necessary to decide whether the clinically-related criteria that do provide rules for clinical interpretation, such as thresholds to judge that two estimates are similar, are applicable to most clinical areas or whether these rules should be revised depending on the clinical context.

Some of the statistically-related appropriate criteria also present some challenges to their use. For example, most of the criteria that are based on the results of a statistical test assume that the two estimates of treatment effect are completely independent. Therefore, such criteria can be used in scenarios in which the interest is to compare treatment effects that come from completely different bodies of evidence (such as RCTs versus

observational studies), but not to compare estimates of effect obtained from bodies of evidence that are related somehow (for instance, estimates of effects obtained with NMA and HTHC in which there are studies that are included for the estimation of both effects, as in one of the chapters of this thesis).

Considering these limitations, we concluded that the use of one criterion would not by itself be enough to judge the similarity of the estimates of treatment effects, and that some of the criteria would need to be modified depending on the clinical context.

Chapter 4
To what extent can network meta-analysis be used to answer current clinical questions? - Systematic survey of Cochrane reviews

7 Abstract

Objective: To determine the extent to which network meta-analysis is used, and should be used, to pool the results of current systematic reviews.

Methods: Systematic survey of Cochrane reviews published between July 2014 and June 2015. We included all systematic reviews of interventions that included only randomized clinical trials. We classified these reviews according to whether they aimed to answer a question that should use a network meta-analysis or not, based on the description of the objective, type of interventions, and searching process described by the authors. For the reviews that were aiming to answer a network meta-analysis question, we determined whether they performed or could have performed such analysis.

Results: Out of the 809 systematic reviews included, 74.7% were not aiming to answer a network meta-analysis question. From the 25.3% of reviews that were answering a network meta-analysis question, only 4 performed this type of analysis, while 95 could have used it but did not.

Conclusions: Network meta-analysis is used, and could be used, to answer a small proportion of the current clinical questions addressed in SRs.

8 Introduction

Since its introduction in the in mid 1990s,[33] network meta-analysis (NMA) has become very popular and its use has increased over the years. This statistical technique, which allows comparisons of the relative effectiveness of multiple treatments[4], has increased its use over time,[96] and there have been many publications addressing the methodology[4 16 17 20 29] and guidelines for its use.[5 11 12]

One of the main advantages of NMA is that by comparing more than two treatments in the same analysis, authors of systematic reviews (SRs) are addressing questions that are more relevant from a clinical perspective, where clinicians are usually faced with choosing among multiple courses of action.[4-6] For this same reason, some authors are suggesting that NMA should become the norm when combining the results of clinical trials.[13]

A trustworthy NMA has to meet many requirements. Some of these are related to the methodology of the SR performed to collect the data to be used in the analysis, but others are particular to the NMA technique itself.[12] In an ideal scenario, the network would be well connected (that is, most pairs of treatments would have been directly compared) and the results would be consistent between direct and indirect comparisons within the network. However, at minimum, having only three interventions in two RCTs provides enough data to perform an NMA.

The fact that, from a clinical perspective, there are usually more than two alternative courses of action would make NMA suitable to perform when the clinical question is: which course of action is most appropriate? However, it is not clear whether this is the research question that authors of systematic reviews are currently addressing. In addition, there are no published guidelines on how to determine the aim and scope of a SR, and the Cochrane website only establishes that on choosing the topic and scope authors should address a question that is important to consumers, health professionals and policy-makers; that the proposed review does not duplicate any (Cochrane) work either

published or under development, and that “there may be some discussion with the editors to clarify or change the scope of the proposed review”.[97]

Use of NMA in most SRs would provide a broad view of the clinical topic and yield all the advantages of this technique, such as the possibility of comparing interventions that have not been directly compared in randomized clinical trials, the improvement in the precision of the effect estimates because of the use of both direct and indirect evidence; and the ability to know which treatment is most likely to be the best, and the worst, for a specific outcome.[6 7 19 29]

Nevertheless, using NMA in most SRs could also increase the burden of reviewers and users. Authors of SRs would need to invest more time and effort to include all the relevant evidence, and acquire new competencies specific to performing, summarizing, interpreting, and reporting the results of NMA. Literature users may also need to acquire new skills to judge the trustworthiness and learn how to use a SR in which an NMA was performed.[12]

Taking all of this into consideration, we aimed to gain a deeper understanding of the issue and study to what extent NMA could be used to answer current clinical questions, by examining the clinical questions in SRs published in the Cochrane Database of Systematic Reviews, as Cochrane SRs are widely known and accepted as one of the most trustworthy sources of evidence summaries relevant to clinical practice.[98]

9 Methods

We performed a systematic survey of the literature.

9.1 Search and study selection

We included all the SRs of interventions that included only RCTs, published in the Cochrane Database of Systematic Reviews between July 1, 2014 and June 30, 2015. We searched all the articles using the journal and “publication date” filter in PubMed Medline. Two reviewers screened the titles and abstracts to determine the eligibility of

the studies. We excluded editorials, SRs with both RCTs and observational studies, SRs of diagnostic test studies, protocols, overviews of reviews, and methodological reviews and SRs that were withdrawn after publication.

9.2 Data abstraction

We abstracted information regarding the medical area of the review (according to the Canadian Medical Association classification, see Appendix 1), the number of trials included, the number of interventions included, and the type of primary outcome (dichotomous, continuous, or other). We judged that authors had a question best answered by NMA methods when they were interested in comparing the relative effectiveness of all treatment options against each other. Based on the description of the review objective and study selection criteria, we classified the SRs in the following categories:

1) SRs that were not aiming to answer a question that should be answered using an NMA. There were three subcategories:

    a. the authors were interested in comparing only two interventions against each other;

    b. the authors were interested in more than two interventions, but they were interested in comparing one intervention against many others (for example, they were interested in the effect of intervention A versus interventions B, C, and D, but not in the effect of B versus C, B versus D, and C versus D);

    c. the authors were interested in more than two interventions, but they were interested in comparing many interventions against one control (for example, they were interested in the effects of A, B and C versus placebo or the standard of care, but not in the effect of A, B or C versus the other two active treatments)

Although in both, 1b and 1c an NMA could be done, it is not necessary to answer the question. In addition, when the clinical question is like the ones stated, it is unlikely that the authors of the SR would include indirect evidence to inform their comparison of interest through an NMA.

2) SRs that were aiming to answer a question that should be answered using an NMA. In this category, we further categorized studies as those that

    a. performed an NMA.

    b. did not perform an NMA. Here, we determined whether it would have been possible to perform it or not (either because of lack of data or because clinical heterogeneity among the studies included in the review, according to the authors, prevented from performing an NMA).

For this classification, we used the primary outcome since we considered this to be the most clinically relevant outcome of the review. To determine the primary outcome, we used the information provided by the authors in the methods section of the review. If more than one primary outcome was listed, we used the first-listed primary outcome that met the eligibility criteria. If no primary outcome was specified, we assumed that the primary outcome was the first outcome presented in the results of the review.

We used the intervention descriptions as reported by the authors in the methods and meta-analyses, assuming that they were experts in their specific clinical fields. That is, depending on the review, the intervention could be a class of drugs, a combination of drugs, specific drugs, or a drug with a specific dose or mode of administration. The same principles were used for non-pharmacological interventions.

We used two criteria to judge whether clinical heterogeneity prevented the authors from performing an NMA: 1) Explicit: the authors of the SR reported explicitly that trials were too heterogeneous to perform any type of meta-analysis, or 2) Implicit: the HTHC reported for the primary outcome differed with respect to the type of intervention, comparator or outcome (either the outcome definition or the time-point where it was

measured) sufficiently that this was one of the characteristics used by the authors to separate different groups of HTHC.

For the SRs that performed an NMA, we also abstracted information regarding the shape of the network and the quality of the evidence, as reported by the authors of the review.

9.3 Outcomes

Our primary outcome was the overall proportion of SRs that were answering a clinical question that should be answered using an NMA. Our secondary outcomes were:

- the proportion of SRs that used NMA
- the proportion of SRs that were answering a clinical question that should be answered using an NMA and where an NMA could have been performed, but it was not
- the proportion of SRs in which the authors were interested in only two interventions
- the proportion of SRs in which the authors were interested in more than two interventions, but in which NMA was not necessary to answer the question

9.4 Data analysis

We used proportions to summarize the information for the outcomes of interest. We used mean, medians and interquartile range to describe the main characteristic of the SRs included per category. All calculations were done using the software R.[50]

10 Results

10.1 Study selection

The search resulted in 961 references, from which 809 were eligible and included in our study. The main reasons for exclusion were that the reviews included both randomized

controlled trials and observational studies (55 reviews) and that the review was withdrawn (34 reviews). Figure 4.1 summarizes the study selection process and provides more details regarding the reason for exclusion.

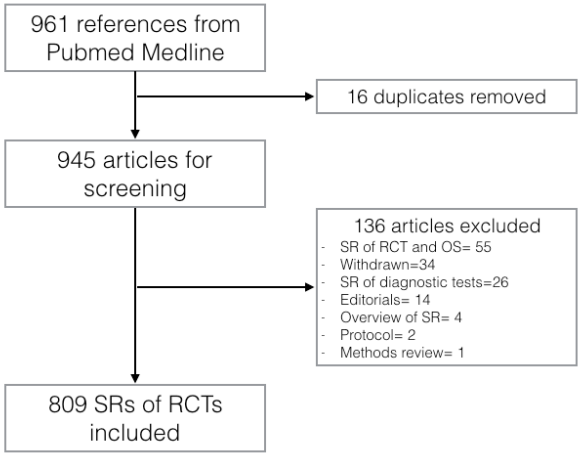


Figure 4.1: Study selection

10.2 Included studies

The reviews included a range of 0 to 182 RCTs, with a median number of 7 and an IQR from 3 to 14. Most of the reviews had a dichotomous primary outcome (n= 521, 64.4%). The medical areas in which there were the most SRs published were obstetrics/gynecology (n= 110, 13.6%), neurology (n= 59, 7.3%), and psychiatry (n= 53, 6.6%) (Appendix 2). The number of SRs according to the specified categories is summarized in Figure 4.2, and described below.

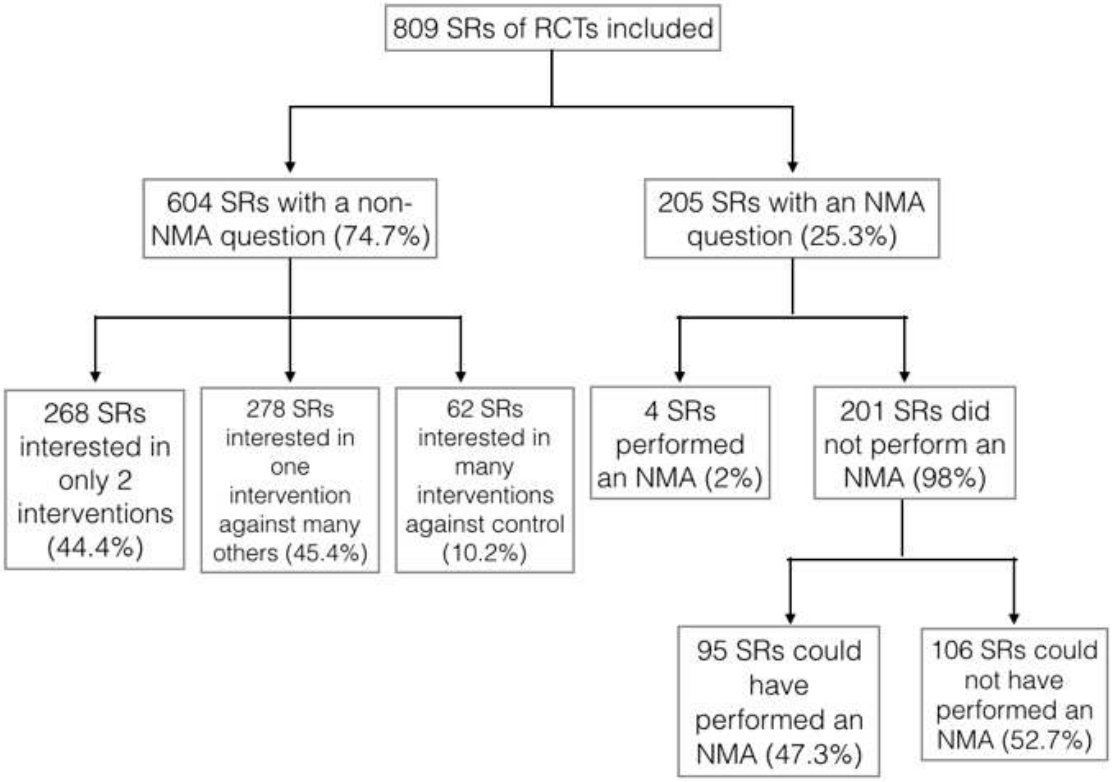


Figure 4.2: SRs according to their question and the use of NMA

10.3 SRs that were not aiming to answer NMA questions

Most of the SRs were not aiming to answer an NMA question (n= 604, 74.7%). From these, 45.4% (n=274) were reviews in which the authors were interested in comparing one intervention against many others, 44.4% (n=268) were SRs in which the authors were interested in comparing only two interventions, and 10.2% (n=62) were SRs in which the authors were interested in comparing many different interventions against one control. Table 4.1 shows details regarding the number of RCTs included and the type of primary outcome of the SRs in each of these categories.

Table 4.1: Main characteristics of the SRs that were not aiming to answer an NMA question, according to the focus of the clinical question

	One intervention against many others	**Two interventions against each other**	**Many interventions against one control**	**Total**
n	274	268	62	604
%	45.4	44.4	10.2	100
Number of RCTs				
Mean	12.4	8.9	13.2	10.9
Median	8	5	8	7
IQR	3-16	2-10	2-14.5	2-14
Type of outcome (n, %)				
Dichotomous	165 (60.2)	185 (69.0)	42 (67.7)	392 (64.9)
Continuous	96 (35.0)	68 (25.4)	20 (32.3)	184 (30.5)
Other	13 (4.8)	15 (5.6)	0	28 (4.6)

In most of the medical areas the majority of reviews were not aiming to answer an NMA question; the exceptions were dermatology (23%, 2 out of 9 SRs), and dentistry (25%, 5 out of 20 SRs) (Appendix 2).

10.4 SRs that were aiming to answer an NMA question

Only 25.3% of the included SRs were aiming to answer an NMA question (n= 205). This is 37.9% of all the SRs in which the authors were interested in assessing the effects of more than two interventions (205 out of 541). From these, only 4 reviews (2%) performed an NMA. These reviews were from the medical areas of general surgery, nephrology, obstetrics/gynecology, and ophthalmology (1 SR with NMA in each area). They included a median number of 41 RCTs (range 13 to 137), and they all had a dichotomous primary outcome. Table 4.2 shows the main characteristics of these studies.

Table 4.2: Characteristics of the SRs that performed an NMA

Reference	Medical area	Number of RCTs included in the review	Type of primary outcome	Number of interventions included in the network	Number of RCTs included in the network	Network shape	Quality of the evidence
Dumville, 2015**[99]**	General surgery	13	Dichotomous	6	10	Complex	Low
Wilhelmus, 2015**[100]**	Ophthalmology	137	Dichotomous	8	83	Complex	Moderate to low
Palmer, 2014**[101]**	Nephrology	56	Dichotomous	6	11	Star with some other treatments connected	Moderate to very low
Le Cleach, 2014**[102]**	Obstetrics/ gynecology	26	Dichotomous	4	18	Complex	Not assessed

In 47.3% of the reviews that did not perform an NMA, this analysis could have been done (n=95). These SRs included a median number of 12 RCTs (IQR 8 to 24, range 3 to 137). There was not enough data to perform an NMA in 77 SRs (38.3%), and the clinical heterogeneity did not allow performing an NMA in 29 SRs (14.4%).

In most of the 77 SRs in which the authors were aiming to answer an NMA question, but could not have done it because there was not enough data, there were no RCTs included (n=19, 24.7%). The median number of RCTs included in the SRs with lack of data for performing an NMA was 3 (IQR 1 to 5). There was one SR in these category in which there were 157 RCTs included;[103] however, even though there was enough data to perform an NMA for secondary outcomes, the RCTs included did not report enough data to perform this analysis for the primary outcome.

SRs in which, according to the authors, the studies were too heterogeneous to perform meta-analyses were spread between the different clinical areas. These included a median number of 7 RCTs (IQR 5 to 14). There was one SR in the area of public health and preventive medicine that included 182 RCTs;[104] nevertheless, the authors claimed that the interventions were too complex and diverse to allow classification into a small enough number of meaningful intervention types.

11 Discussion

In our systematic survey, we reviewed 809 SRs of RCTs published in the Cochrane Database of Systematic Reviews in a one-year time frame. Surprisingly, we found that most of the authors of such reviews did not aim to answer a clinical question where an NMA should be used (74.7%), and that in the majority (52.7%) of SRs asking an NMA question, an NMA could not have been done either because of a lack of data or because of clinical heterogeneity among the included studies.

11.1 Discussion of main findings

NMA has become very popular, with some authors claiming that they expect it to become "the new norm for combination of results of clinical trials".[13] The main argument for this is that decision-makers usually consider more than two treatment options for most the clinical conditions, and NMA offers the possibility of comparing the relative effectiveness of all these options, even when they have not been compared directly against each other in RCTs.[4-7 13] We observed that being equally interested in more than two treatment options may not always be the case, since even though the majority of the authors of the SRs were interested in assessing the effects of more than two interventions (66.9% of all the SRs), in less than half of these SRs (37.9% of the SRs in which the authors were interested in more than two interventions) the objective was to compare all interventions against each other.

From a clinical perspective, there are three scenarios in which authors of a SR would not need to do an NMA even if they could: they are only interested in comparing two interventions, they are interested in comparing one intervention against many others, or they are interested in comparing many interventions against one control. We saw that in 268 SRs (which represented 44.4% of the SRs that were not aiming to answer an NMA question, and 33.1% of all the SRs) the authors were interested in comparing only two interventions.

We also observed that in 336 SRs (that is, 55.6% of the SRs that were not aiming to answer an NMA question, and 41.5% of all the SRs) the authors were interested in assessing the effects of more than two interventions, but their objective was to make specific comparisons between pairs of interventions, as opposed to comparing all available treatment options against each other. Authors of most of these reviews were interested in comparing one intervention against many different controls (274 out of the 336). For example, Rezale et al.[105] aimed to determine the effects of budenisone for induction of remission in patients with Crohn's disease. In the methods section, they state that they included studies in which the active treatment was oral budenisone and the comparator was placebo or any other active agent, such as corticosteroids and sulfazalazine. In the other SRs, authors were interested in comparing many interventions against one control (62 out of the 336). One example of such reviews is the one published by Maldonado-Fernandez et al.[106], where the authors were interested in evaluating the effects of pharmacological treatments for preventing vestibular migraine, by comparing beta-blockers, calcium antagonists, anticonvulsants and others, against placebo or no treatment.

Even though in both of these cases authors could, from a technical perspective, do an NMA; when examining their question, we realized that they were only interested in specific comparisons. Even more, by restricting the eligibility of studies to RCTs that report only those comparisons as they do, the data collected would generate a star-shaped network (with the common treatment at the hub and the comparators at the points of the star, see Appendix 3) where they could only do indirect comparisons among the other treatments. This shape would result from searching only for some of the studies that could create a network in which most treatments are connected to each other, and not because all the evidence relevant to the interventions results in the star-shaped network. In other words, the network would be incomplete and the comparisons between some pairs of treatments could not benefit from the indirect evidence, which is one of the advantages of NMA. More importantly, it seems like the authors of these SRs were not interested in the indirect comparisons. For instance, Rezale et al.[105] were not interested in comparing corticosteroids versus sulfasalazine, and that Maldonado-Fernandez et

al.[106] were not interested in comparing beta-blockers versus calcium antagonists or anticonvulsants. For these reasons, performing an NMA would not be necessary.

It is interesting that such a large proportion of the SRs do not aim to answer an NMA question. One of the reasons we decided to survey only Cochrane SRs is that the proposals for new Cochrane SRs undergo a process of prioritization according to their need, and we assumed it would publish reviews answering clinical questions that are current and relevant. One interesting finding when exploring these SRs was that some reviews were closely related to others. For example, Derry et al. published two SRs in treatments for adults with neuropathic pain. In the first one, they aimed to assess the effects of topical lidocaine versus placebo or any other active treatment,[107] while in the second one, they wanted to compare the effects of nortriptyline against placebo or any other active treatment. Another example is the SRs published by Dumville et al., in which the focus was on treatments for pressure ulcers. In the first review they wanted to assess the effects of hydrogel dressings,[108] in the second one they were interested in the effects of negative pressure,[109] and in the third one the intervention of interest was alginate dressings.[110] Instead of undertaking different systematic reviews, with different searches, screening, data abstraction and analysis; the authors could have conducted one broad review covering all the interventions of interest using direct and indirect evidence to inform the comparisons of interest. This would have been more informative because of the fact that in both examples, and in other cases not described here, many of the interventions included in the review as the comparators are repeated across SRs.

In the situation just described, it could be argued that there is one NMA question underlying these SRs closely related to each other. We could not avoid wondering whether the decision to do separate reviews is related to other issues, such as knowledge of how to do NMA, the need of working with narrower and more manageable topics, or the desire to publish more than one SR. However, considering the fact that all of these reviews went through an editorial process in which the scope was approved and accepted as relevant, and that their results were published as many SRs that would be found as single pieces of information, we classified them as not having an NMA question. In

addition, we decided to stick to our initial assumption that Cochrane is an organization in which groups are led by experts in their respective fields, who would prioritize and make decisions regarding the scope of a SR based on clinical relevance, despite the fact that there is no specific guidance to assess the relevance of a topic and that the different editorial groups have the autonomy to use their own criteria to make such decisions.

We were also surprised with the finding that from the SRs that ask an NMA question, only a few of them had actually performed an NMA (2% of the SRs that aimed to answer an NMA question, 4.2% of the SRs that aimed to answer an NMA and could have actually done it). Only a few other reviews that aimed to answer an NMA question described explicitly in their methods that they had considered using NMA but that this was not possible, but since this was not one of our aims we did not keep a register of exactly how many these were.

There were a considerable number of SRs that were aiming to answer an NMA question but could not have performed an NMA, either because of lack of data or because of the clinical heterogeneity among the included studies (52.7% of the SRs that aimed to answer an NMA question). In the case of lack of data (72.6% of the SRs that could not have done an NMA), it is reasonable to assume that authors of SRs in which there is not much evidence available would tend to ask broader questions (even if the availability of RCTs should not drive the decisions with regards of the scope of the SR more than the clinical relevance), as opposed to performing and publishing a series of narrowly focused SRs in which each result shows that there is no evidence available to answer the question. In the second case – clinical heterogeneity (27.4% of the SRs that could not have done an NMA)- the clinical expertise of the authors judging that the patients, interventions and outcomes are too diverse across studies to even perform a head-to-head meta-analysis is an appropriate reason to not combine the results of these studies.

Based on a close examination of the multiple head-to-head comparisons presented in the results and data analysis sections of the SRs, there were 95 SRs in which the authors were aiming to answer an NMA question and could have performed such analysis, yet they did not do it (11.7% of the total of SRs). From the description of the analysis in the methods

section, it is not clear whether this was in the authors' plans at any stage of the review development, since there was no mention of NMA in the data-analysis description, and thus we can only speculate about the reasons this may have happened. One of the reasons may be the lack of familiarity with the statistical methods to perform an NMA. A survey published in 2012[111] shows that among Cochrane review authors, 87% had heard of or had some knowledge regarding indirect and mixed-treatment comparisons methods; however, only 18% had considered these methods when planning their reviews. Although the reasons for not doing so varied, 40% declared that it was because their lack of knowledge of the methods, and 89% said that they would like to receive more training on these methods. Only 9% of the surveyed authors considered that the evidence from indirect comparisons is not valid.

Even though there is no published data about this, we speculate another reason for underutilization of NMA when it is appropriate and possible: since Cochrane has standardized and facilitated the review process so much, by providing tools such as RevMan,[51] the groups of authors of Cochrane SRs are not encouraged to, and do not need to, include a biostatistician among their members. RevMan allows authors to perform many types of head-to-head meta-analyses by entering the necessary information in a friendly software interface, and thus a basic understanding of the principles of meta-analysis is more than enough to do the analyses in a Cochrane review. The lack of close support by a biostatistician, the difficulty of providing this support to all groups of authors, and the fact that RevMan does not support NMA yet, may make it difficult to transition from a set of head-to-head comparisons to an NMA.

11.2 Agreement with previous research

To date, we are not aware of any other study addressing our research question. There is only one study that is related to this topic, in which the authors aimed to review the methods used to synthesize the evidence in public health evaluations.[112] They characterized the 39 National Institute for Health and Care Excellence public health appraisals published in a 6-year period and evaluated how the information was synthesized. Their findings showed that 23% of the reports included a meta-analysis, and only 1 report (2.6%) included an NMA. They make the case that more modern statistical

methods, such as NMA, could be used. Nevertheless, the authors of this study only looked at the methods actually used by the authors of the report they reviewed, and did not make any judgments regarding the type of question that the report was trying to answer and how this relates to the possibility and appropriateness of performing an NMA.

11.3 Strengths and limitations

This systematic survey has two main strengths. First, it takes a broad overview of all the Cochrane SRs in a one-year time frame and secondly, it uses use explicit methods and a clinical perspective to judge the extent to which NMA could be used to answer current clinical questions. We decided to include all the reviews published in the one-year period, to provide the most comprehensive assessment of the issue. At the same time, we always took a clinical perspective to establishing the methods and, in particular, the categories into which the SRs would be classified. This led us to consider the objectives, selection criteria, data analysis and results description of each SR as a whole. This stands in contrast to an approach that would simply examine whether the SR included more than two interventions and two or more RCTs, conditions that would, technically allow an NMA. By looking past these simple minimal technical conditions, and considering if, from a clinical problem perspective, NMA was appropriate, we aimed to be as fair as possible when claiming that an NMA could have been done but was not.

This study suffers from the two weaknesses of most systematic surveys, mainly, the need to make some decisions that may seem arbitrary yet make the process more feasible, and the need to rely on the descriptions that the authors provide with regards to their methods and results. For example, we had to restrict our survey to the primary outcome of the SRs, and if there was more than one objective, the primary objective of the SRs. There were many cases where an NMA could have been performed to achieve one of the secondary aims of the review or for a secondary outcome. One example of the first case is the review published by Otasowie et al.[113] The authors aimed to study the effects of tricyclic antidepressants in children and adolescents with attention deficit hyperactivity disorder. Although they were interested in comparing different types of antidepressants against each other, which is a question that could be answered using an NMA, they

explicitly state that for their primary analyses of antidepressants versus placebo or versus any other active agent they grouped all antidepressants as one single treatment. An example of the second case is the review published by Van Zuuren et al.[103] on interventions for hirsutism. Despite the fact that there were 157 eligible trials included in this SR, few studies reported data on the primary outcome "participant-reported improvement of hirsutism", while most of the studies reported data on the secondary outcome "clinician-reported improvement of hirsutism".

One reason we chose to survey Cochrane SRs was to overcome the potential limitation of having to rely on the author's description of their objectives, data analysis and results to make our judgments. Cochrane SRs have high reporting standards, and they provide many details with regards to their methodology.

11.4 Implications for research

The findings of our study have several implications for systematic reviewers and for methodological research. First and foremost, and despite the fact of its increasing popularity, it appears that NMA would not be necessary in most of the SRs being performed currently. Therefore, reviewers would be able to keep developing their SRs without the extra burden of the need to get training and special support in the use and interpretation of results from this approach. Secondly, it would be interesting to determine whether the clinical questions that we assumed to be relevant because of their appearance in the Cochrane's catalogue of reviews are also relevant from the clinicians' perspective, to actually validate our main finding. Thirdly, it would be interesting to explore how common it is to find NMA questions that are answered in a series of related SRs addressing only head-to-head comparisons; something we could not do because we sampled only Cochrane reviews and only reviews published over a one-year period. Finally, researchers could explore with more depth how the question of a review is formulated, and whether the knowledge of NMA methods, or the possibility of using them in a user-friendly platform such as RevMan, would influence the study question, providing insights on whether NMA could eventually become a norm for evidence synthesis.

11.5 Implications for practice

For clinicians who use evidence to inform their clinical practice, in particular those to rely on Cochrane SRs, our study shows that as of mid-2015, they will rarely be faced with an NMA to answer their questions, and thus no extra efforts in terms of critically appraising studies that used these methods would be needed. Unfortunately, it also means that for most of their clinical questions they would find themselves with a set of head-to-head comparisons that might give them an incomplete picture with which to inform their practice.

12 Conclusions

Despite its increasing popularity, NMA can be used to answer a small proportion of the current clinical questions addressed in SRs. Most of the current SRs are designed to answer clinical questions for which there is no need to perform an NMA, and in the majority of the SRs in which an NMA question is asked, the heterogeneous characteristics and amount of data from the available studies would not allow such an analysis.

Chapter 5
Do systematic reviews performing network meta-analysis report the same results as systematic reviews using head-to-head comparisons? – The case of stents in patients undergoing percutaneous coronary intervention

13 Abstract

Objective: Among systematic reviews (SRs) evaluating stents in patients with percutaneous coronary intervention (PCI), to determine whether the results of a network meta-analysis (NMA) as the method of statistical analysis are different to those that use head-to-head comparisons (HTHC).

Study design and setting: We searched Medline for all the SRs with NMA assessing the relative effectiveness of stents in patients undergoing PCI. We abstracted the effect estimates for all the comparisons for the primary outcome and matched them with estimates from SRs using HTHC answering the same clinical question. We created perfectly and imperfectly matched pairs of SRs and assessed whether they were similar using a series of clinically-oriented criteria.

Results: We included 12 SRs with NMA, which were paired with 20 SRs with HTHC, giving a total of 42 pairs of estimates. In perfectly matched pairs (n=12), the effect estimates were similar in 66.7% to 83.3% of the cases depending on the criteria used. Two thirds of the pairs met both criteria to claim that they were similar. The proportions ranged from 44.8% to 75.9% in the imperfectly matched pairs, with 44.8% meeting both criteria of similarity.

Conclusions: Clinicians that aim to inform their practice with evidence may find dissimilar answers depending on whether they choose to use SRs with HTHC or NMA in an important proportion of cases in which they can choose between NMA and HTHC. Critical appraisal skills are fundamental for using the results from SRs with NMAs.

14 Introduction

Systematic reviews (SRs) that use network meta-analysis (NMA) as a tool for comparing the relative effectiveness of more than two treatments of interest are being published with increasing frequency.[96] These SRs provide information that cannot be obtained from the traditional head-to-head comparisons (HTHC), such as the probability that a treatment is the best option for a specific outcome and estimates comparing treatments that have not been directly compared in randomized clinical trials.[4 6 7] Furthermore, they also provide the estimates of effect that can be found in the traditional SRs that perform HTHC, but they obtain these estimates by using both the direct and the indirect evidence that goes into the NMA.[4-6]

Therefore, for many clinical situations, clinicians looking to inform their practice with the best and most updated evidence will be faced with the choice of using a traditional SR with only HTHC or a SR with NMA. Using a SR with NMA would have summary information from all available treatment options in a single source document, including the use of clinician friendly rankings.[4 6 7 29] Although this seems attractive, using a SR with NMA presents additional challenges to the clinician, such as the need to be familiar with the NMA methodology and basic principles to be able to judge the trustworthiness of the SR using NMA.[12] In addition SRs with NMA may provide information that the clinician is not interested in, for example when his clinical question concerns only two treatment options.

On the other hand, clinicians aiming to inform their clinical practice with the traditional SR with HTHC may find it easier and faster, as it is a process that many clinicians know and feel comfortable with. Nevertheless, using SRs with HTHC would not allow the clinician to take advantage of the potential benefits of using indirect evidence, such as obtaining narrower confidence intervals, and having a broader view of the clinical topic.[5 19] Furthermore, using a SR with HTHC may not even be an option for some clinical scenarios in which the two treatments of interest have not been compared directly in randomized clinical trials, or when trialists only compare new agents versus an established standard of care or placebo for regulatory purposes.[114]

An additional challenge may arise when clinicians are trying to decide whether to use a SR with HTHC or NMA. Since the many steps and procedures leading to the final results vary between the two approaches, their results may differ for the same clinical question. These differences may arise not only because NMA uses both direct and indirect evidence to calculate the relative effectiveness of two or more treatments, but also because of features of the SR itself. For example, eligibility criteria may be broader in SRs using NMA in order to obtain networks with as many direct comparisons as possible; the SR using NMA may perform searches that are more exhaustive to capture all evidence available, or in contrast, they may perform less exhaustive searches that are based on updates of previously published SRs;[12] or they may include or exclude trials or interventions in a specific analysis for an outcome based on the availability and reporting of all the other trials.

These are only a few examples of design and analysis features that may create differences between the results that a SR with NMA and a SR with HTHC report for a given clinical question. Since there may be many other reasons for differences in results, it is unlikely that literature users with basic training in critical appraisal will be able to detect such nuances. To understand how big of a concern this may be, it is necessary to study the potential extent of this problem. The objective of this study was to determine whether the results of SRs that use NMA are different to those from SRs that use HTHC. Since the field of cardiovascular medicine has the most published SRs using NMA methods,[96] we focused on SRs evaluating the effects of stents in patients with percutaneous coronary intervention.

15 Methods

We performed a systematic survey of the literature.

15.1 Eligibility criteria

We included all SRs assessing the relative effects of different types of stents in patients undergoing percutaneous coronary intervention that use NMA as the method of analysis. In order to be included, the SRs needed to meet the following eligibility criteria:

1. SRs of RCTs: defined as a review that performed a systematic search of RCTs or quasi RCTs in at least one electronic database, and where authors selected articles based on specific and explicit selection criteria. Updates of SRs, overviews of reviews, and SRs reported in an HTA, that meet this definition were included. We also included SRs in which the authors included other type of study designs, as long as they pooled the results of RCTs and quasi RCTs separately from the results of the observational studies.

2. The question of interest was about more than two types of stents (of any type) compared to each other, which should have been explicit in the aim, methods or results of the review.

3. Use of NMA as the method of analysis (defined as an evidence network that involves two or more RCTs and more than two interventions) as the primary or secondary method of analysis. The use of NMA in a SR was determined from the description of the statistical analysis in the methods section of the SRs. We judged that an NMA had been used when at least one of the following criteria was met: the authors described using an NMA; the analysis included all trials in one single statistical model to estimate treatment effects; or the SR cited references that were relevant to NMA in the methods section where there was a description of the type of analysis used.

4. Focus on a dichotomous primary outcome. The primary outcome was established according to what was specified by the authors in the aim or methods; otherwise the first outcome presented in the results was considered the primary outcome

We excluded SRs that performed NMA based on individual-patient data, because these usually have less available data for the analysis when compared to grouped-data meta-

analyses. We also excluded SRs in which the authors compared stents against any other intervention, and SRs that were not published in the English language.

15.2 Study searching and selection

We constructed an electronic search for OVID Medline, which was run from the database inception to November 2015 (See Appendix 4). After removing duplicates, we screened the titles and abstracts of all the citations, and gathered the full-text screening of all the references that seemed relevant. All these articles were screened in full-text to confirm eligibility. Two reviewers performed the both screening stages, screening independently and in duplicate. All final decisions were reached through consensus, and expert advice was sought in those cases were there was doubt regarding some of the criteria (in particular, whether the method of analysis performed corresponded to an NMA).

15.3 Matching SRs that used NMA with SRs that use HTHC

For each of the SRs included, we searched for SRs with HTHC that aimed to answer the same clinical question in terms of patients, interventions, comparison, and primary outcome (PICO). We used the following procedure:

1) For each SR with NMA, we extracted the PICO questions corresponding to the primary outcome. Since the NMA would have more than two interventions and more than two comparisons, there were at least three PICO questions per SR that used NMA. For the outcome component of the question, we considered the clinical definition of the outcome, the follow-up time, and the measure of association used (relative risk, odds ratio, rate ratio or hazard ratio).

2) We constructed a search strategy for retrieving SRs with HTHC assessing the effects of stents in patients undergoing percutaneous coronary intervention (See Appendix 4). We included SRs of RCTs with HTHC that were answering at least one of the PICO questions corresponding to any of the SRs using NMA, either as the primary or a secondary research question. Two independent reviewers assessed eligibility of the SRs with HTHC at title and abstract and full-text screening stages. There were no restrictions regarding the studies included in both

SRs, that is, it was not necessary that the SRs using HTHC included the same studies as the SR using NMA.

3) We matched the SRs using NMA with the SRs using HTHC. Only one HTHC per combination of comparisons of the NMA was selected. Thus, the maximum number of SRs with HTHC paired with a SR with NMA was the maximum number of comparisons reported in the NMA. When we found more than one SR using HTHC matched with an NMA on the same question, the reviews that had the most similar (1) dates and (2) databases searched, were selected. These two criteria were used hierarchically.

4) We considered a matched pair to be a perfect match when all the components of the PICO question were the same. Since the outcomes in this clinical area are infrequent, and since a clinician would interpret them the same, we considered relative risks and odds ratios to be the same.[115] We considered a matched pair to be imperfect when there was some small difference in one of the PICO components, or when the information provided by the authors did not allow making a judgment regarding some of the components. We allowed imperfect matches in which: a) only RCTs including patients undergoing PCI de novo were included in one of the SRs while the other did not specify this, b) one of the stent types was subdivided into two types and treated as two different interventions in one SR while they were grouped together in the other, c) the definition of the outcome was not detailed enough to judge whether the outcome was exactly the same, and d) one of the SRs presented the comparison using relative risk or odds ratio while the other used rate ratio or hazard ratio.

In the cases where there was an imperfect match because one SR divided a stent type into two types, whereas the matching one treated it as one single intervention, we matched the two estimates of one SR with the one estimate of the other, obtaining two pairs.

15.4 Data abstraction

We extracted the following data from the SRs using NMA methods: number of RCTs included, number of treatments, number of direct comparisons, and structure of the network.

We extracted the following data for each comparison, from both the SR using NMA and the SR using HTHC: number of trials included in the comparison (counting only the trials that directly compared the treatments in the case of the network), number of patients included in the comparison, the point estimate of the treatment effect, the confidence or credible interval of the estimate of effect, the direction of the effect (using the confidence or credible interval), and whether the difference was statistically significant.

15.5 Outcomes

15.5.1 Main outcome measure

Proportion of SRs with NMAs reporting results that are similar to those from the matching HTHC. To do this, we determined whether both of the following two criteria were met: the ratio of the point estimates was between 0.8 and 1.25; and the point estimate of the HTHC was contained in the credible interval of the NMA. We did this assessment at a pair level, that is, using the estimate from the NMA and HTHC for a specific PICO as the unit of analysis.

15.5.2 Secondary outcomes

1) Proportion of pairs in which the ratio of the point estimates was within a similarity threshold. We used a conservative threshold of 0.8 to 1.25 (which is 1/0.8) and a less conservative threshold of 0.7 to 1.43.[116]

2) Proportion of pairs in which the point estimate of the HTHC was contained in the credible interval of the NMA.

3) Proportion of pairs in which the point estimate of the NMA was contained in the confidence interval of the HTHC.

4) Proportion of pairs in which both estimates benefit the same intervention, different interventions, or none- by means of the confidence interval (in other words, both confidence intervals completely below 1, completely above 1, or crossing 1).

5) Proportion of pairs in which both estimates benefit the same intervention or different interventions- by means of the point estimate.

6) Proportion of pairs in which both estimates were statistically significant. For this we used the p-value from frequentist analyses and posterior probabilities from Bayesian analysis.

7) Agreement on statistical significance, using the Kappa chance corrected agreement.

15.6 Statistical analysis

All analyses were done separately for perfect and imperfect matched pairs. All analyses and plots were done using R software,[50] version 3.2.4.

We calculated proportions for most of the primary analyses. We used Cohen's Kappa to evaluate the agreement on statistical significance, and used Landis and Koch's guidelines for its interpretation.[117] We assessed the association between the primary outcome (similar results within a pair) and the specific clinical outcome, and the time period between the publication of the SR with HTHC and the SR with NMA within a pair (categorized as ≤2 years or > 2 years) using Fisher's exact test. We also assessed the association between the primary outcome and the number of direct comparisons used in the NMA using logistic regression. In addition, for imperfect matches we evaluated the association between primary outcome and the reason for being an imperfect match, categorized as described above, using Fisher's exact test. We used a level of significance of 5%.

Finally, we performed an exploratory analysis in which we included the pairs that had been discarded because they were addressing the exact same comparison as another pair (that is, the pairs excluded in stage 3 of the matching procedure described above).

16 Results

Our search for SRs using NMAs yielded 41 references, from which 24 were reviewed in full text and 12 were deemed eligible.[118-129] From the 1654 references obtained when running the search for SRs with HTHC, 139 were reviewed in full text and 22 were eligible to be paired with the SRs using NMA. After choosing only one HTHC per comparison within an NMA, we included 20 SRs with HTHC[130-149] and obtained 42 pairs of comparisons.

16.1 Characteristics of the included SRs using NMA

The included SRs using NMA were published between 2007 and 2015, although most of them were published between 2012 and 2014 (9 out of the 12).[119-127] The SRs included a mean of 61 RCTs, with a minimum of 22 and a maximum of 126. The NMAs had a mean of 6 interventions included (minimum= 3, maximum= 8), and a mean of 11 direct comparisons. Most of the NMAs had a complex structure (10 out of 12), except for 2 NMAs that included only 3 interventions.[128 129] All authors used Bayesian random effects models.

Most of the SRs included studies of all patients undergoing PCI, but some of them limited these to other clinical conditions such as myocardial infarction[121 124 128 129] and diabetes.[126] The outcomes measured in these SRs were target vessel revascularization at different time points,[123 124 126 127] mortality at different time points,[119 121 122 128 129] and stent thrombosis (either definite[118 125] or definite or probable[120]).

16.2 Characteristics of the included SRs using HTHC

The included SRs using HTHC were published between 2005 and 2015. They included a mean of 10 RCTs (minimum= 2, maximum=33), and most of them used a frequentist fixed effect model approach for performing the HTHC (60%).

16.3 Perfectly matched pairs

Twelve of the 42 pairs were perfect matches (28.6%). These pairs corresponded to 8 out of the 12 NMAs.[119-122 124 125 127 129]

Eight pairs (66.7%) satisfied the two criteria for similarity of results from the NMA and HTHC.

Most of the pairs had ratios of point estimates contained within the established thresholds; point estimates contained in the other's estimate confidence or credible interval, and were concordant with regards to the direction of the effect and statistical significance. Cohen's Kappa showed moderate agreement with regards to statistical significance (Table 5.1).

Table 5.1: Outcomes for perfect matches

Criterion	**n (of 12)**	**%**
Primary outcome		
Ratio of point estimates between 0.8 to 1.25 and point estimate of HTHC contained in credible interval of NMA	8	66.7
Secondary outcomes		
Ratio of point estimates - Between 0.8 and 1.25 - Between 0.7 and 1.43	 8 10	 66.7 83.3
Point estimate of HTHC contained in credible interval of NMA	9	75
Point estimate of NMA contained in confidence interval of HTHC	9	75
Concordance of the direction of the effect - Using the point estimates - Using the confidence/ credible intervals	 9 10	 75 83.3
Concordance of statistical significance - Proportion of pairs - Cohen's Kappa (p value)	 10 0.43	 83.3 (0.07)

There was no evidence of an association between the between the specific clinical outcome and the similarity between the NMA and HTHC estimate (Fisher test p-value= 0.86), between the number of direct comparisons of the NMA and the similarity between the NMA and HTHC estimate (OR=0.84, p=0.25), or between the time period between the publication of the SRs of each pair (Fisher test p-value=1).

When doing the sensitivity analysis for the primary outcome with all the pairs (that is, including pairs that were addressing the same comparison), 13 pairs (76.5%) met the two similarity criteria.

16.4 Imperfectly matched pairs

Thirty of the pairs were imperfect matches (71.4%). These pairs corresponded to 11 out of the 12 NMAs.[118-120 122-129]

The main two reasons for pairs being an imperfect match were differences in the interventions (n=11, 36.7%), and differences in the measure of association used to calculate the relative effect (n=9, 30%). There were 4 imperfect matches (13.3%) due to differences in patients and 4 imperfect matches due to uncertainty about the outcome, and 2 imperfect matches that combined two of the reasons.

Thirteen pairs (44.8%) satisfied the two criteria for similarity of results from the NMA and HTHC.

Most of the pairs had point estimates contained in the other's estimate confidence or credible interval, and were concordant with regards to the direction of the effect and statistical significance. Cohen's Kappa showed fair agreement with regards to statistical significance. The ratio of point estimates was within the less conservative threshold for similarity for most of the pairs, but it was only within this threshold in 44.8% of the pairs when using the conservative threshold (Table 5.2).

Table 5.2: Outcomes for imperfect matches

Criterion	n (of 29)	%
Primary outcome		
Ratio of point estimates between 0.8 to 1.25 + point estimate of HTHC contained in credible interval of NMA	13	44.8
Secondary outcomes		
Ratio of point estimates - Between 0.8 and 1.25 - Between 0.7 and 1.43	 13 20	 44.8 69
Point estimate of HTHC contained in credible interval of NMA	20	69
Point estimate of NMA contained in confidence interval of HTHC	22	75.9
Concordance of the direction of the effect - Using the point estimates - Using the confidence/ credible intervals	 18 21	 62.1 72.4
Concordance of statistical significance* - Proportion of pairs - Cohen's Kappa (p value)	 20 0.33	 66.7* (0.03)

*n=30 for this outcome. One NMA did not report the point estimate of the comparison because the results were not statistically significant so its corresponding pair could only be included in this outcome

There was no evidence of an association between the reason for the pair being imperfect and the similarity between the NMA and HTHC estimate (Fisher test p-value= 0.37), between the specific outcome and the similarity between the NMA and HTHC estimate (Fisher test p-value= 0.24), between the number of direct comparisons of the NMA and the similarity between the NMA and HTHC estimate (OR=0.92, p=0.22), or between the time between the publication of the SRs of the pair (Fisher test p-value=0.71).

In the sensitivity analysis with all the pairs (that is, included pairs that were addressing the same comparison), 19 pairs (52.8%) satisfied the two criteria for similarity of results from the NMA and HTHC.

17 Discussion

We compared the results reported by SRs using NMA and SRs using HTHC for a specific PICO question. Starting with a sample of 12 SRs using NMA assessing the effects of stents in patients undergoing percutaneous coronary intervention, we matched the effect estimates from the NMA with effect estimates reported by SRs using HTHC for the same population, comparison and outcome. Based on specific characteristics of the SRs, we identified 12 perfectly matched pairs and 30 imperfectly matched pairs. While the majority of perfect matched pairs met the criteria to claim that the effect estimate is similar in the SR with NMA and the SR with HTHC (66.7%), less than half of the imperfect matched pairs (44.6%) were found to report similar results.

17.1 Discussion of main findings

The results of this study show that in many situations, clinicians would be faced with different effect estimates when using SRs that provide information for the same clinical question, depending on whether the clinicians searched for evidence in an NMA or a HTHC. Depending on how conservative we were with the threshold for similarity, either 66.7% (when using the conservative threshold) or 83.3% (when using the less conservative threshold) of the perfectly matched pairs presented similar results (primary outcome). The outcomes these SRs were looking at were mortality, stent thrombosis and target vessel revascularization at different time points. Although the decision threshold (that is, the size in the difference in effect to claim that one treatment is better than another) for these outcomes has not been established in the literature, the low prevalence of these outcomes and the severity of their consequences suggest that some clinicians would accept the conservative threshold (or perhaps demand an even more stringent threshold), and that they would be faced with different results one third of the time. When the SRs seem to be reporting results about the same comparison, but differ in specific aspects that are difficult to detect because of reporting issues, or because the authors used a different measure of association (in this study, the imperfect matches), clinicians would find estimates of effect that differ with regards to the point estimate by more than a relative risk reduction of 20% and a relative risk increase of 25% in 55.2% of the cases.

When looking at other measures of similarity used in this study, even though there seem to be fewer differences between the results reported by SRs with NMA and SRs with HTHC for a specific comparison, clinicians would be faced with different results in approximately 17 to 25% of the findings when there is complete certainty that the SRs are answering the same question (perfect matches), and approximately 25 to 40% of the findings when there may be some slight difference in the question the SRs are answering (imperfect matches).

These potential differences could have arisen for a number of reasons. The fact that we constructed matched pairs of estimates of effect by using two different SRs makes it possible that some of the design and conduct features, besides the method of analysis, had influenced the estimates of effect. The eligibility criteria may be more or less restrictive despite the aim of two SRs being to answer the exact same clinical question. For example, 4 of the NMAs had as an eligibility criterion to have enrolled more than 50[123 124 126] or 100[127] patients and followed them up for at least 6 months, despite the fact that the primary outcome of interest for the authors was taken over the longest follow-up available. This could potentially introduce differences in the effect estimates obtained from these SRs when compared to others that did not have such eligibility restriction, as some trials may be excluded from the former and not the latter. Similarly, the searching process, mainly with regards to the electronic datasets (and search strategies) or other resources used might not have been exactly the same in two SRs, which is also a potential source of differences. Other potential differences may stem from the intervention or outcome definition, and the measure of association used, as observed in our imperfect matched pairs.

Nevertheless, it is unlikely that clinicians using evidence to inform their clinical practice would perform such detailed analyses to determine the source of the differences in the results they have found, when faced with conflicting results. In addition, when we explored the association between similarity and other factors such as the specific clinical outcome, the number of direct comparisons that a network used, the time of publication between the two SRs and the reason for imperfect matching (in the case of imperfect matched pairs), we did not find any statistically significant relationships. Our small

sample size, however, would have power to only detect strong relationships. Therefore it is difficult to speculate about how each of these factors may play a role individually or in combination, and the influence they would have in the difference in effect estimates may be specific to each clinical scenario.

17.2 Choice of study design

We designed this study from a pragmatic, clinical perspective. By this, we meant to approach the problem by trying to replicate what a clinician would find in the literature. Since our research question was whether traditional head-to-head meta-analysis and NMAs report similar results, we decided to look at results that were already published and available to clinicians, and that clinicians would find when looking for evidence to inform their clinical questions. That is why we decided to pair results from SRs using NMA with results from SRs using HTHC for the same clinical question, despite the fact that there may be more differences between these than just the method of statistical analysis.

An alternative approach to this answering this question could have to calculate direct and network (that is, direct + indirect) estimates of effects from the NMA, without the need to pair SRs. We discarded the use of this approach because 1) it is unlikely that a clinician would be faced with the choice between these two estimates, since we assumed that when using an NMA a clinician would look at the reported network estimate, and 2) the potential differences we would have found would have been due to the extra trials that are included in the estimation of the network versus the direct effect (that is, the addition of the indirect evidence). This approach would answer a statistical question about the additional information that can be learned by using NMA instead of a set of HTHC for a given set of trials. But it does not answer the larger, more pragmatic question about whether results vary in an important way if NMA or HTHC is used to guide the whole SR study design and analysis process.

We chose SRs with NMA from cardiovascular medicine because after performing a survey of all the NMAs published, and as reported in bibliometric studies,[96] this is the field with the highest frequency of SRs using NMAs. Within the field of cardiovascular

medicine, stent use in patients undergoing percutaneous coronary intervention was the type of intervention most frequently studied. While all the SRs with NMA were answering similar questions, and the RCTs included repeated across SRs, none of the SRs using NMA was answering exactly the same question as any other SR using NMA.

17.3 Strengths and limitations

This is the first study that addresses this increasingly relevant question using an approach focused on literature users as opposed to statistical aspects. Our methods for collecting data and constructing pairs of estimates of effects to be compared were developed with that in mind, and making sure that all the information we used was information that clinicians using evidence would also have access to. In addition, we used multiple criteria to assess whether there were differences between the estimates of effect from SRs with NMA and SRs with HTHC for the same comparison. Some criteria focused on the magnitude of the point estimates such as the ratio of estimates, while other focused on the confidence interval, the direction of the effect, and the statistical significance of the effect.[116] In this way, we tried to cover most of the aspects that can be looked at when appraising the results of a SR.

Feasibility concerns meant that, in this first approach, we had to pick a subset of all SRs using NMA. In the strictest sense, our results are only applicable to SRs using NMAs in stents in patients undergoing percutaneous coronary intervention. Fortunately, we were able to include all the SRs with NMAs from that area. With this broad and detailed view of a specific clinical topic, we are positive that our results and conclusions are applicable to all the NMAs published on this topic.

We did not assess the methodological quality or risk of bias of the SRs to determine whether this was related to the differences between the estimates of effects. To date, there are two tools proposed to assess the methodological quality of SRs and none of them fit our needs. Despite the fact that the AMSTAR tool[150] has undergone a validation process, it has still received criticisms for including questions that seem to be about reporting rather than methodological quality. The ROBIS tool[151] that has been developed recently is composed of 4 domains and 21 questions and it has not been widely

used or accepted yet. In addition, neither of these explores issues that may be relevant exclusively to NMA and therefore decisions with regards of the quality of the SRs within a pair (such as which SR had a higher methodological quality) would have been harder to make and would have been quite subjective in many cases. Similarly, and although two approaches have been proposed,[10 11] there is still not a widely accepted method to assess the quality of the evidence for a specific outcome.

Lastly, we could not characterize the observed differences between effect estimates to explore whether NMAs tend to overestimate or underestimate treatment effects with respect to HTHC. This was not possible because we were faced with a series of HTHC for which there was no established treatment that is better than the others. Because we did not know the direction in which the effect was supposed to go, we could not determine whether any particular difference was an overestimation or underestimation of the treatment effect.

17.4 Agreement with previous research

To our knowledge, three other studies have aimed to answer a research question similar to ours. In the first study, Song et al.[152] assessed discrepancies between direct and indirect comparisons using 11 comparisons from a systematic review of antibiotic prophylaxis in patients undergoing colorectal surgery. They calculated the absolute difference in the log odds ratio between the direct comparisons and adjusted indirect comparisons, and between the direct comparisons and simple indirect comparisons. They found statistically significant differences in 5 cases for the simple indirect comparisons and in 2 cases for the adjusted indirect comparisons. In the second study, Song et al. quantified the discrepancies between direct and indirect estimates of effects using 44 comparisons reported in 28 Cochrane SRs published up to the year 2000.[153] They observed that 3 comparisons were statistically significantly different. They also reported a moderate agreement between the statistical conclusions of the direct and indirect estimates (weighted Kappa 0.51). Finally, Song et al.[30] expanded on the previous two studies and compared the direct and indirect estimates of effects of 112 comparisons from 85 Cochrane SRs published up to 2008. They report that in 14% of the comparisons there were statistically significant differences between the direct and the adjusted indirect

estimate. While in the first study the authors concluded that more research was needed to explore the validity of indirect comparisons, and in the second study they concluded that adjusted indirect comparisons agree with direct comparisons often but not always; in the third study they claim that there is significant inconsistency between direct and indirect comparisons more often than previously observed.

While overall the results of the studies described above seem to go in a similar direction to ours, comparisons between those results and ours are difficult because of all the differences in the research question, design, and analysis. First, the authors of such studies were aiming to compare the effect estimates obtained with direct HTHC and indirect comparisons, while we aimed to compare the effect estimates obtained with direct HTHC and NMA. Therefore, in our approach, we are considering the estimates obtained using both direct and indirect evidence as opposed to only indirect evidence. Secondly, although their first study was very focused on a specific clinical topic,[152] the other two studies used SRs of different clinical areas. In contrast, we took all NMAs published on a specific clinical topic. Third, while the authors of these studies used as a sample SRs that reported enough data to perform both direct and indirect comparisons within the same review and made sure that the same trials had not been used in both comparisons; we obtained our effect estimates from different SRs answering the same question. Fourth, the authors of the studies had to calculate the indirect estimates of effects most of the time as they had not been reported in the original SR. Therefore, it is unlikely that clinicians looking for evidence with regards to a comparison would have been able to access to such estimates to inform their practice. On the other hand, we only used estimates of effect that were already published and that clinicians could easily use when looking for evidence. Finally, the authors of these studies use as a main measure or difference between the estimates of effect either the absolute difference or a z-test, while we focus in outcomes that have a clinical rather than a statistical interpretation.

17.5 Implications for research

Our study is the first exploring the issue of potential differences between estimates of effect obtained with NMA and HTHC from a clinical perspective, and such, it opens the door to many other research questions. First, this study could be extended to other clinical

topics either within cardiovascular medicine or other clinical areas where NMA are becoming common. A second direction of future research is to explore the differences between the estimates of effect reported on SRs with NMA and HTHC in continuous outcomes.

More importantly, researchers could focus on development of clinician-friendly guidance with regard to how to choose an estimate of effect when faced with situations where they see discrepancies between two of them, without the need to do a detailed and time-consuming exploration of what the cause of the differences may be.

17.6 Implications for practice

The results of our study show that clinicians searching for information in SRs with HTHC and in SRs with NMA could frequently be faced with conflicting evidence regarding to the same clinical question, which they may find alarming. Therefore, clinicians need to be cautious when using evidence from SRs with either NMA or HTHC. This highlights once again the need for training in critical appraisal methods to be able to determine the trustworthiness of a body of evidence. It also implies, however, that new skills and knowledge would have to be gained when it comes to using evidence from NMAs.

18 Conclusions

Clinicians that aim to inform their practice with evidence may find dissimilar answers depending on whether they choose to use SRs with HTHC or NMA. This situation may happen about half of the time when there is no absolute certainty about the SRs answering the exact same question, and one third of the times where the SRs answer the same question. Critical appraisal skills are fundamental to overcome the challenges that this may bring.

Chapter 6
How robust are the rankings from network meta-analysis to changes in the trials included in the network and decision thresholds to rank the treatments?

19 Abstract

Objective: To explore the robustness of the rank probabilities and the rankings obtained from network meta-analysis (NMA) when excluding trials and using increasing thresholds to define a difference.

Methods: Systematic survey of the literature. We included all systematic reviews with NMA from the field of cardiovascular medicine, with trial-level data available. We reran all the NMAs and determined the probabilities of each treatment being the best and second best, and the rankings, when excluding each trial in turn from the analysis. We also examined the effect of increasing the decision threshold required to declare two treatments different on the probability of each treatment being the best.

Results: We included 14 systematic reviews, with a median of 20 randomized trials and 9 treatments. The best treatments had probabilities of being best that ranged from 38% to 85.3%. The mean absolute change in the probabilities when a single trial was dropped was 4.3% (range 1.7% to 8.4%). We observed decreases and increases in the probability of being the best treatment as large as 51.9% and 18.5%, respectively. On average, when a trial was dropped, the best treatment changed 5% of the time, and the second best changed 13% of the time. The effect of different thresholds on the probability of a treatment being best varied across scenarios.

Conclusions: Rank probabilities and rankings of treatments in NMA can be fragile to changes in the RCTs included in the network, and to increases in decision thresholds to claim that one treatment is more effective than the other. Rankings in NMA should be used with caution.

20 Introduction

Network meta-analysis (NMA) is a statistical technique used to combine the results of a series of head to head comparisons, in order to estimate the relative effectiveness of all the interventions included in the network.[34] NMA has become very popular in recent years, with fewer than 10 of these studies published each year from 2003 to 2008, but approximately 85 in 2013.[96] In the same way, the number of hits when searching for NMA and NMA related terms in databases like PubMed, has increased steadily,[96] a sign that that this technique is being developed, studied and discussed more often.

NMA is attractive for informing clinical decision-making because of its ability to provide a broad view of all the available treatments for a specific clinical condition.[13 32] Other advantages of NMA include the possibility of comparing treatments that have not been directly compared in randomized clinical trials (RCTs) or systematic reviews (SRs), and the increased precision that is obtained when using both direct and indirect evidence to obtain estimates that compare the relative effectiveness of two treatment options.[4 7 36]

An advantage of the Bayesian implementation of NMA, frequently highlighted by the proponents of this technique, is the possibility of obtaining the probabilities of each treatment being the best treatment (or second best, or worse)- also known as rank probabilities- for a specific outcome, which allows ranking the treatments from the best to the worst.[4 7 31 36] In each of the Markov Chain Monte Carlo simulation iterations of the Bayesian implementation of an NMA, a baseline treatment is used as a reference to rank every other treatment, based on its relative effect with respect to the baseline treatment. Then, the probability of each treatment being the best (or second best, etc.) is calculated based on the proportion of simulation iterations in which each treatment was ranked first (that is, using the number of iterations in which the treatment was first divided by the total number of iterations). Finally, the treatment with the highest probability of being the best is ranked as the best treatment; while the ranking with the second highest rank probability of being the best is ranked as the second best treatment, and so on.

The rankings convey a different piece of information from the relative treatment effects of all possible pairwise comparisons. While ranking results from the network as a whole, relative treatment effects are estimated by weighting the direct and indirect evidence, with some evidence contributing to the final estimate more than other evidence. Thus, the confidence in the rankings may be different from that of the relative treatment effects[10] and their use to inform clinical decision-making should account for that.

Rankings obtained from NMA are potentially misleading.[154 155] They could be misinterpreted by emphasizing the probability of a treatment being the best, without considering the difference in treatment effects. That is, the rankings do not take into account how much better a treatment is compared to the next best treatment, since the standard comparisons to obtaining the rankings considers any difference bigger than 0 to be relevant.[53 155] In addition, the probability of a treatment being the best may be fragile and change when including new trials and treatments in the NMA.[53] It has been shown that removing one treatment from an NMA can change the top three treatments in up to half of the networks.[54] To date, unfortunately, there is no further insight into how the difference in treatment effects, or the exclusion of trials may affect the probabilities of treatments being the best, and the rankings obtained from NMA.

The aim of this study was to explore the robustness of the probabilities of each treatment being the best, and the rankings to two factors: 1) excluding an RCT from the analysis; and 2) using different thresholds to declare two treatment effects different and therefore, to calculate the probabilities of being the best. We did this through a survey of all the NMAs that had trial-level data available in the field of cardiovascular medicine, the medical specialty with the most NMAs published.[96]

21 Methods

We performed a systematic survey of the literature.

21.1 Eligibility criteria

We included a sample of SRs that use NMA as the method of analysis. In order to be included, the SRs needed to meet the following eligibility criteria:

1. The SR included RCTs: This was defined as a review that performed a systematic search of RCTs or quasi RCTs in at least one electronic database, and where authors selected articles based on specific and explicit selection criteria. Updates of SRs, overviews of reviews, and SRs reported in an HTA that met this definition were included. We also included SRs in which the authors included observational study designs, as long as they pooled the results of RCTs and quasi RCTs separately.

2. The question of interest concerned more than two interventions (any drugs, administered by any route, or stents), which should be explicit in the aim, methods or results of the review. We also included SRs where the primary aim was to illustrate the process of an NMA if the authors used a SR as described in (1).

3. The SR used NMA as the primary or secondary method of analysis (defined as an evidence network that involves two or more RCTs and at least three interventions). The use of a network was determined through the description of the statistical analysis in the methods section of the SRs; we looked at whether the authors described use of NMA, had included all trials in one single statistical model to estimate treatment effects, or cited one ore more references (in the methods section) that were relevant only to NMA.

4. The SR had a dichotomous primary outcome with a pooled estimated reported as a RR, OR, or HR. The primary outcome was established according to what the authors specified in the aim or methods; otherwise, the first outcome presented in the results was considered the primary outcome.

5. The SR had to be in the field of cardiovascular medicine: interventions or outcomes needed to be relevant primarily to cardiovascular medicine

6. The SRs provided data from all the included RCTs to enable us to re-run the NMA: SRs had to report the interventions assessed in each trial, and the number of patients and events per arm.

We excluded SRs with NMAs where patients had a cardiovascular disease, yet both the interventions and outcomes were relevant to another medical field (for example, an NMA addressing the effects of drugs for treating diabetes on diabetes related outcomes, in patients with hypertension), and studies that were not published in the English language.

21.2 Study searching and selection

We constructed an electronic search for OVID Medline, which was run from the database's inception to February 2015 (Appendix 5). After removing duplicates, we screened the titles and abstracts of all the citations, and gathered the full-text screening of all the references that seemed relevant. All these articles were screened in full-text to confirm eligibility. Two reviewers screened articles at both stages independently, and in duplicate. All final decisions were reached through consensus, and expert advice was sought in those cases were there was doubt regarding some of the criteria (in particular, whether the method of analysis performed was indeed an NMA).

21.3 Data abstraction

We created datasets with the data from the primary studies included in each NMA. For this, we used the data reported by the SR authors in the tables of the articles or, when necessary, online supplementary material. When there was inconsistency between the data presented in figures and tables, or when information was missing (such as the number of events for a specific trial) we referred to the primary study to find the information.

The following decisions were made when abstracting the data to re-run the NMAs:

- We used the names of the interventions as reported in the network figures and results, even if the table reporting the data was more specific with regards to the

intervention classification. In these cases, we combined the information from two arms of a study to match the single arm used in the NMA.

- When data for some of the study arms were reported together, yet the treatments were separated in the network, we split the number of events and number of patients in half to create the two arms, unless the authors reported a different splitting proportion.

- When studies reported the proportion of patients having the event, as opposed to the number of events, we calculated the number of events based on the exact proportion of patients having the event reported by the authors.

- When there was arm-level data for some treatments that were not included in the primary NMA, according to what was reported in the NMA figure and results, these arms were not included. Data from treatments included only in sensitivity analysis was not included.

- For each trial, we combined groups receiving no treatment and placebo into one single arm in all datasets.

21.4 Data analysis and outcomes

We re-ran the NMA for all the SRs, using the datasets we constructed with the data reported in the SR publication. We excluded from these NMAs RCTs in which the number of events was 0 in all arms. We were interested in both the rankings (best, second best, third best, etc.) and the probabilities of each treatment being the best for the primary outcome, and how the rankings and probabilities changed when excluding RCTs from the NMA and when changing the threshold to declare two treatment effects different and therefore, to calculate the probabilities of being the best (see details below). We used the odds ratio (OR) as the measure of effect, and calculated the rank probabilities using the proportion of simulation iterations that each of the treatments was ranked first

(or second, or third, etc.) when comparing all the treatments to a reference treatment by means of the OR. The rankings of each iteration were assigned based on any difference between treatments > 0, that is, a relative effect comparison the treatments > OR= 1.

21.4.1 Excluding RCTs from the analysis

We reran each of the NMAs and subsequently excluded one RCT one at a time. That is, we ran each NMA as many times as there were RCTs included in the network. In each analysis with one RCT excluded, we obtained the probability of each treatment being the best, and the rankings. We then compared each of these with the results obtained when doing the NMA including all trials. When an NMA included K trials, this gave K complete sets of NMA results. The outcomes of interest were computed across these K sets for each NMA: the mean of the absolute change in the probability of being the best treatment, the range of the change of the probability of being the best treatment, and the proportion of cases in which the best, the second best, and either or both of them changed. Each of these outcomes gave us a single value for each NMA and these results were summarized across NMAs (See Appendix 6 for more details).

21.4.2 Increasing the threshold to calculate the probabilities of being the best treatment

We calculated the rank probabilities using the proportion of MCMC simulation iterations in which each of the treatments was ranked first (or second, or third, etc.) when comparing all the treatments to a reference treatment by means of the OR, when using an increasing set of thresholds for superiority (OR 1- value used in the original model, and ORs of 1.1, 1.2, 1.3, and 1.4). For example, if the top two treatments are A and B, and the OR comparing their effect is 1.25, A would be better than B for a good outcome with a threshold of OR=1, 1.1 and 1.2; however, A would not be better than B if the threshold is OR= 1.3 and 1.4. The outcome of interest was the value of these probabilities for each of the thresholds. For each NMA, we identified the best treatment when the threshold was at the default value of 1 and plotted the probability of being best or in the top two treatments versus the threshold.

For performing the NMAs, we used random-effects Bayesian hierarchical consistency models with uninformative priors.[34] We used Markov Chain Monte Carlo with 4 parallel chains for assessment of convergence with an adaptation phase of 5000 samples and 20000 simulation samples. We checked convergence of the models using the Gelman-Rubin statistic.[156] If models had not converged, we ran more iterations until convergence was reached. If convergence was not reached after 400,000 iterations, we explored the cause. After identifying the parameters and RCTs that caused the issue, we either excluded the RCT if it had no events in any of the arms (because it was not providing any extra information for the estimation of the parameter), or added one event to both arms if one arm had 0 events and the other had 1 or more. We explored if any convergence issues remained after doing this and if so, reported the results accordingly.

We did all NMA analyses using the *gemtc* package[157] in the R software.[50] The probabilities of each treatment being the best under varying thresholds were obtained using a modification of the *rank.probability* built-in function in the package, and the revised rankings were obtained from these probabilities. Appendix 7 shows the code for performing all analyses.

22 Results

The search resulted in 975 references, from which 131 articles were screened in full-text and 14 were included. The SRs were published between 2003 and 2015, and covered a wide range of patients, interventions, and outcomes from the field of cardiovascular medicine (see Table 6.1). These NMAs for the primary outcomes included a median of 20 RCTs (range 11-57), and 9 treatments (range 4-12).

Table 6.1: Main characteristic of the included SRs

Study	Patients	Interventions	Primary outcome	Numbers		
				RCTs	Treatments	Direct comparisons
Dogliotti, 2014[158]	Atrial fibrillation	Antithrombotics	Stroke	20	8	11
Dooley, 2014[159]	Hospitalized patients	Low-molecular weight heparins	Mortality	14	9	11
Castellucci, 2013[160]	Venous thromboembolism	Antiplatelet or oral anticoagulant	Recurrent venous thromboembolism	11	9	11
Landoni, 2013[161]	Undergoing cardiac surgery	Anaesthetic drugs	Mortality	36	4	5
Navarese 2013[162]	Undergoing treatment with statins	Statins	Diabetes	17	12	14
Wu, 2013[163]	Diabetes	Antihypertensives	Mortality	57	8	14
Bash, 2012[164]	Atrial fibrillation	Interventions to achieve cardioversion	Successful cardioversion	20	10	13
Harenberg, 2012[165]	Undergoing total hip or knee replacement	New oral anticoagulants	Venous thromboembolism	16	5	5
Phung, 2011[166]	Hospitalized, at risk of venous thromboembolism	Thromboprophylaxis	Deep venous thrombosis	13	4	5
Sciarretta, 2011[167]	Hypertension	Antihypertensives	Heart failure	26	8	16
Roskell, 2010[168]	Atrial fibrillation	Anticoagulants	Stroke	13	12	17
Coleman, 2008[169]	Undergoing treatment with anihypertensives	Antihypertensives	Cancer	27	6	10
Cooper, 2006[170]	Non rheumatic atrial fibrillation	Stroke prevention agents	Stroke	19	9	14
Psaty, 2003[171]	Undergoing treatment with antihypertensives	Antihypertensives	Coronary heart disease	27	11	18

22.1 Rankings and probabilities of being the best treatment from the NMAs including using all available data and a threshold OR of 1

The treatments that ranked first had a mean probability of being the best treatment for the primary outcome of 58.5% (range 38.0% to 85.3%). For the treatments that ranked second best, the mean probability of being the best treatment was 20.5% (median 18.5, range 10.3% to 31.5%)(Table 2). The mean difference between the best and second best treatments in the probabilities of being the best treatment was 38.0% (range 2.5%[168] to 75%[164]). The best and second best treatments had a mean total probability of being in the top two of 79.0% (range 58.2% in an NMA with 11 treatments[171] to 99.2% in a network with five treatments[165]).

22.2 Excluding RCTs from the analyses

The mean across NMAs of the mean absolute change in the probability of being the best treatment when excluding an RCT from the analysis was 4.3% (range 1.7%[170] to 8.4%[165]). Table 6.2 shows the mean absolute change in the probability of being the best treatment when excluding trials from the analysis for the treatment ranked first, and its corresponding range. The mean change in the probability of being the best treatment, across treatments, for each NMA, is also shown. The NMA with the largest observed range of changes in the rank probabilities of the first treatment had, at one extreme, a decrease of 51.9% in the probability, and at the other extreme an increase of 2.7%.[163] The second largest observed range of changes in rank probabilities went from a decrease of 35.5% to an increase of 18.5%.[166] The smallest range of changes for an NMA went from a decrease of 7.4% to an increase of 3.4%.[170]

Table 6.2: Treatments ranked first and second in each of the NMAs, and their probabilities of being the best treatment in the analyses with the complete dataset and when excluding RCTs.

Study	Best treatment	Probability (%)	Second best treatment	Probability (%)	Mean of the absolute change in probability of being the best treatment when excluding trials for the best treatment (%)**	Range of the change in probability of being the best treatment for the best treatment (%)	Overall mean change in probability of being the best across treatments (%)**
Dogliotti, 2014[158]	Dabigatran 150 mg	64.6	Rivaroxaban	19.8	3.1	-6.9; 10.8	1.2
Dooley, 2014[159]	Fondaparinux	64.7	Enoxaparin 40 mg	15.3	3.6	-6.3; 10.1	1.5
Castellucci, 2013[160]	Standard dose vitamin K antagonist	57.1	Dabigatran 150 mg tid	18.5	7.1	-25.5; 5.8	1.9
Landoni, 2013[161]	Desflurane	67.4	Isoflurane	31.5	3.2	-12.7; 10.1	1.7
Navarese 2013[162]	Pravastatin 20 mg	40.6	Lovastatin	27.8	3.4	-9.5; 10.3	1.2
Wu, 2013*[163]	ACE inhibitor+ Calcium channel blocker	73.4	Diuretics	13.6	2.4	-51.9; 2.7	0.9
Bash, 2012[164]	Vernakalant iv	85.3	Flecainide oral	10.3	5.4	-5.6; 8.7	1.1
Harenberg, 2012[165]	Rivaroxaban	66.5	Apixaban	32.7	8.4	-26.7; 18.3	3.4
Phung, 2011[166]	Unfraccionated heparin bid	75.6	Low molecular weight heparin	15.0	7.5	-35.5; 18.5	4.1
Sciarretta, 2011[167]	ACE inhibitors	49.3	Angiotensin receptor blockers	15.9	5.0	-41.6; 10.0	2.0
Roskell, 2010[168]	Ximelagatran	34.8	Dabigatran 150 mg tid	32.3	3.6	-2.7; 18.7	1.1
Coleman, 2008[169]	Diuretics	38.3	Beta-blockers	23.2	3.6	-15.9; 10.2	2.1
Cooper, 2006[170]	Alternate day aspirin	63.0	Low dose warfarin	11.0	1.7	-7.4; 3.4	1.4
Psaty, 2003[171]	Beta-blockers/ diuretics	38.0	Ace inhibitors	20.2	2.7	-14.8; 4.3	1.3

* There were convergence issues when performing this NMA; however, these were observed in the estimation of treatment effects different from the best treatment.

** Calculation of these outcomes is described in Appendix 6

After removing RCTs from each included NMA one at a time, the proportion of times that the best and second best treatment, and either one or both of them changed is presented in Table 6.3. The treatment ranked first changed on average in 5.8% of the cases that an RCT was excluded from the analysis (range 0[161 166] to 12.5%[165]). In the case of the second best treatment, this changed in 13% of cases on average (range 0[161] to 23.1%[166]).

Table 6.3: Proportion of times that the rankings changed when excluding RCTs from the analysis

Study	Number of RCTs	Proportion of times that the best treatment changed (%)	Proportion of times that the second best treatment changed (%)	Proportion of times that either one or both treatments changed (%)
Dogliotti, 2014[158]	20	5.0	10.0	10.0
Dooley, 2014[159]	14	7.1	14.3	14.3
Castellucci, 2013[160]	11	9.1	9.1	9.1
Landoni, 2013[161]	36	0.0	0.0	0.0
Navarese 2013[162]	17	5.9	17.6	17.6
Wu, 2013[163]	57	1.7	5.1	5.1
Bash, 2012[164]	20	5.0	10.0	10.0
Harenberg, 2012[165]	16	12.5	12.5	12.5
Phung, 2011[166]	13	0.0	23.1	23.1
Sciarretta, 2011[167]	26	3.8	15.4	15.4
Roskell, 2010[168]	13	7.7	15.4	15.4
Coleman, 2008[169]	27	7.4	14.8	18.5
Cooper, 2006[170]	19	5.3	15.8	15.8
Psaty, 2003[171]	27	11.1	18.5	18.5

22.3 Increasing the threshold to calculate the probabilities of the treatments being the best

We observed four different scenarios when increasing the OR threshold for declaring a difference and recalculating the probability of the best treatment being the best or being in the top two treatments: A) the change in this probability was small,[161 164 170] B) the probability decreased in a constant manner, with a moderate size change,[159 163 166 168] C) the probability decreased in a constant manner, with a large size change,[158 160 165] and D) the probability decreased rapidly, reaching very low values.[122 167 169 171] Figure 6.1, which shows an example of each of the four scenarios, illustrates the change in the probabilities of the best treatment being the best or in the top two treatments when increasing the OR threshold. Appendix 8 shows these changes for each of the NMAs.

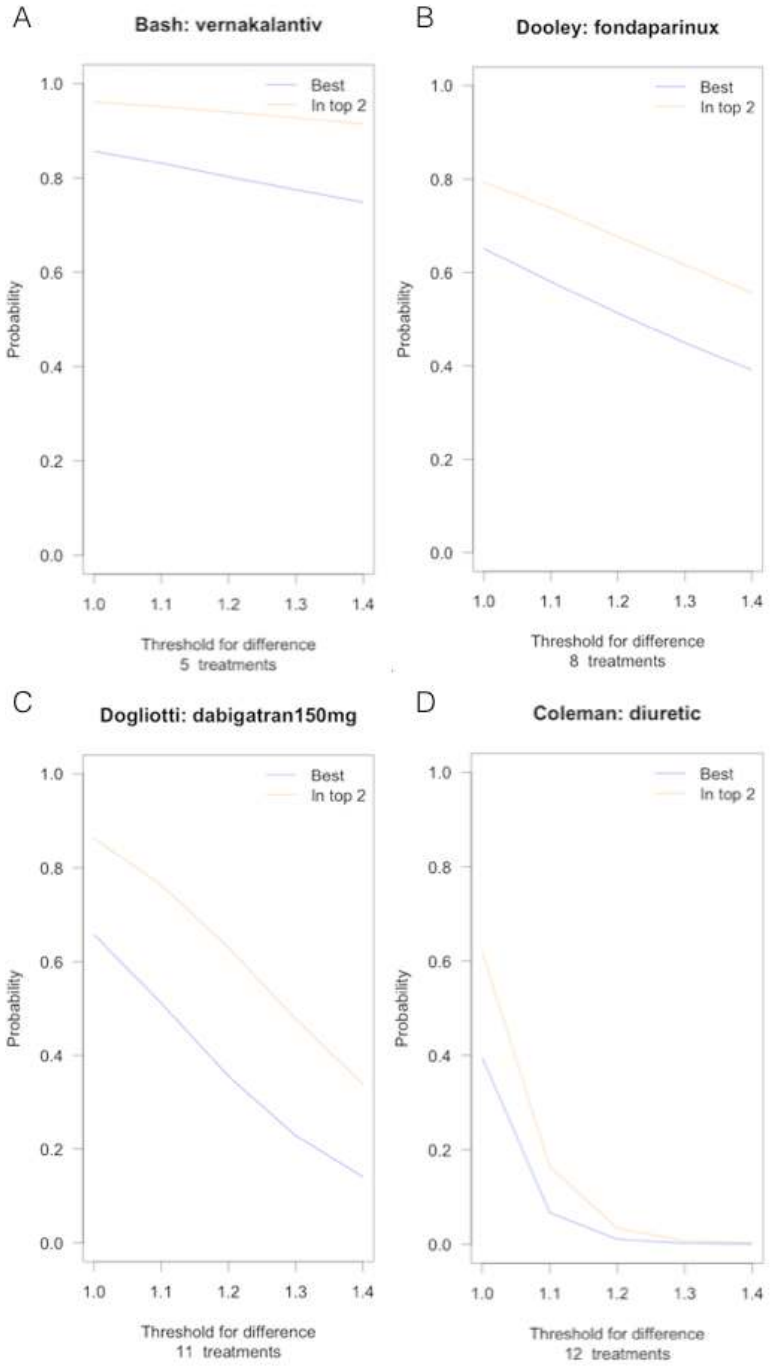


Figure 6.1: Change in the probabilities of the best treatment when increasing the OR thresholds to calculate these probabilities.

A: very small change; B: constant moderate decrease in probabilities; C: constant large decrease in probabilities; D: rapid decrease to very small probabilities.

Within each of these groups, the overall change in the probability of the best treatment being the best, and the slope of this change depended on the specific case. For the NMAs in which the best treatment had small changes when increasing the OR threshold, the range in the total probability change was less than 20%. In these cases, the best treatments started with probabilities higher than 60%. In the NMAs in which the change was moderate, the change was less than 35%, and the treatments started with probabilities from 35% to 75%. In the NMAs in which the change was constant and large this change was up to 50%, approximately, with the treatments starting with probabilities of 60% to 70%, approximately. In the NMAs with rapid changes, the probabilities changed from approximately 40% at the threshold of OR=1, to almost 0% at the threshold of OR= 1.4, with all of them having very low probabilities of being the best treatments when the threshold was OR 1.2.

23 Discussion

In this study, we aimed to explore the robustness of rank probabilities and the rankings to the exclusion of RCTs from NMAs as well as the impact of increasing the decision thresholds (from OR 1, to 1.1, 1.2, etc.) on the probabilities of the treatments being the best. We performed a systematic survey of NMAs in cardiovascular medicine that reported trial-level data and re-analyzed the NMAs, exploring changes in the results when excluding one RCT at a time from the analysis and when increasing the thresholds to calculate the rank probabilities. The mean absolute change in the rank probabilities was 4.3%, and the best treatment changed in 5% of the cases when a single trial was dropped, while the second best treatment changed 13% of the time, on average. Increases in the threshold to calculate the rank probabilities decreased the size of the probability that a given treatment was best, with the magnitude of this change depending on the specific case.

The rankings are among one of the most cited advantages of NMA.[4 32 36] The attractiveness of the rankings lies in their simplicity to illustrate which treatment is the best for a specific outcome, which is easy to understand, especially in cases in which the alternative treatments are numerous. By summarizing the comparative effectiveness based on rankings, the potential to facilitate the decision-making process is increased.[13 32] Nevertheless, rankings can be

misleading since they do not provide information about the size of the difference in effects,[53 154 155] and because they could be fragile to changes in the network.[53]

23.1 Discussion of main findings

We observed that the rank probabilities could suffer important changes when a single RCT was excluded from the analysis. Even though the mean absolute change in the rank probabilities of the best treatment across NMAs was, on average, 4.3% (with a range of 1.7% to 8.4%), there were specific cases in which excluding one RCT from the analysis resulted in decreases in the rank probability up to 51.9%[163] and increases up to 18.7%.[168] In some cases, the variation in the change of the rank probabilities had a range larger than 50%,[163 166 167] with decreases always larger than increases. Wide ranges were observed in the NMAs with different numbers of of RCTs (small to large). To sum up: 1) the mean absolute change in the rank probabilities was relatively small; and 2) the ranges of the changes in the rank probabilities show that there were cases where dropping a single RCT led to a change that was very large. Based on this, we can claim that although rank probabilities may remain reasonably constant, small changes in the structure of the network (which in this case was excluding one single RCT from the analysis) can result in drastic changes in these probabilities. This is supported by the fact than the NMAs in which the largest range of change in rank probabilities had 13,[166] 26,[167] and 57[163] RCTs included.

There were changes in the best treatment in 5.8% of the cases that an RCT was excluded, while the second best treatment changed 13% of the times. Both treatments remained the same in 86.8% of the cases that an RCT was excluded from the analysis. These results suggest that despite potential changes in rank probabilities, the treatments ranked first and second tend to remain the same.

Despite providing insightful information on the fragility of the rankings to minor changes in the content of a network, the change in the rank probability and the change in the treatments ranked first or second do not say anything about the actual size of the difference in effects between these treatments and the ones ranked below them. The second valuable contribution of this thesis is the exploration of the relationship between effect size (or decision threshold to claim that one treatment is better than the other) and the subsequent rank probabilities. We calculated the rank probabilities using the conventional decision threshold of a relative effect of 1 (that is, one

treatment is better than the other if the odds ratio comparing the two treatments for a bad outcome is <1 or >1 for a good outcome), and using decision thresholds of relative effects larger than 10, 20, 30 and 40%.

As expected, the rank probabilities of the best treatment decreased with the increase in the decision threshold. The magnitude and pattern of this decrease, however, was specific to each NMA. We could see cases with small decreases across increasing thresholds in one extreme, and cases with rapid decreases that went down to 0% in others. This supports the notion that the interpretation of the rankings must be accompanied by a careful interpretation of the pairwise estimates comparing the best treatment with the other treatments.[12 155] In fact, the recalculation of rankings using a set of thresholds relevant to the particular NMA could become a standard part of the analyses of networks of evidence.

We observed minor convergence issues in the Markov Chain Monte Carlo sampling, in particular with the NMA published by Wu et al.,[163] which did not converge after 400,000 simulation iterations in 4 of the cases in which an RCT was excluded from the analysis, even after excluding 3 RCTs contributing data to a problematic parameter in which there were no events for both arms, and after adding one event in each arm of 2 RCTs contributing data for a problematic parameter after the previously mentioned change was made. Since these convergence issues were observed in only 4 cases (out of more than 50 RCTs included in this NMA), and did not affect the parameters related to the treatment ranked first, we decided to still use the corresponding results.

23.2 Agreement with previous research

To our knowledge, and despite the fact that potential issues with the use of rankings have been raised,[53 155] there is only one study that has explored this in more depth.[54] Mills et al. performed a systematic survey of 18 NMAs that had 5 or more treatments. They analyzed each network with the complete datasets and also excluding the trials that had specific treatments (that is, excluding one or more treatment nodes), and calculated the changes in the best, and the three top treatments. Their results showed that after removing the treatment node with the highest impact in the results, the top three treatments changed in half of the networks. This study differs

from ours because the authors were interested in the impact of excluding treatments from the network; in other words, they excluded all the trials that reported data on the treatment they were excluding from the analysis. In contrast, we explored the impact of excluding only one trial from the analysis at a time. It should also be noted that Mills et al used the conventional threshold to calculate the rank probabilities, while we explored the impact of changes in these thresholds as well. Our approach uses only minor modifications of the network and represents the kinds of changes in rank probabilities and rankings that might occur as more trials slowly appear in a given area, or as a result of decisions about the inclusion or exclusion of a particular trial from a SR.

23.3 Strengths and limitations

Our study has several strengths. As a first attempt to explore the issues around the rankings in NMA, we designed our study to follow methods as systematic as possible, while at the same time keeping it feasible. We chose to use cases from the field of cardiovascular medicine because this is the field in which NMAs are most frequently published.[96] Since the outcomes studied in these NMAs are those common in this field, our results could be applicable to the majority of the NMAs in this area.

Our study also has limitations. We must acknowledge that one of our inclusion criteria was the availability of primary data, and due to the variation observed in the results, our results must be interpreted and applied carefully. However, there is no reason to believe that availability of primary data within a SR should be related to the fragility of a network. The results here may give an optimistic estimate of the impact of omitting a single trial. Given the relatively large numbers of trials (median 20) in most of our eligible networks, perhaps because trials in cardiovascular medicine are common, the observed fragility associated with excluding a single trial is likely lower than would be seen in fields of research where NMAs may contain fewer trials.

A second limitation of our study is related to the reporting of the trial-level data in the NMAs. As stated above, this was one of the eligibility criteria for the SRs, in order to make this study feasible. In addition to the applicability concerns discussed in the previous paragraph, this also led to us constructing the datasets using data as reported by the authors of the SRs, which caused some issues that had to be dealt with during the data abstraction process. Some of these issues

included the need to calculate number of events based on proportion of events, which may result in inexact numbers of events because the proportions are reported with numbers rounded to a specific decimal (but this results in only small potential errors when numerators are large); and having to split arms equally when they were reported combined in the study level data but had been used separately in the NMA analysis. These issues may explain any existing differences between the results reported in the original SR and the ones we obtained when running the NMAs, but they do not affect our overall conclusions regarding the robustness of rankings.

23.4 Implications for research

In addition to illustrating how fragile rankings can be, our study has several implications for research concerning NMA, systematic reviewers using NMAs, and clinicians who may inform their practice using SRs that report NMAs. First, it would be interesting to explore how changes in RCTs included and decision thresholds affect the surface under the cumulative ranking curve (SUCRA), a measure recently introduced to report the information about rank probabilities.[17] Methodologists could also explore how excluding RCTs and changing the decision thresholds affect the rank probabilities and ranking in SRs from other medical fields, including NMAs with continuous outcomes.

Second, authors of SRs that use NMAs, peer reviewers and journal editors should be careful when interpreting rank probabilities and rankings in their articles. They must acknowledge that although this information may be useful and attractive, they must complement it and draw conclusions about treatment effects that also take into account the size of the difference in effectiveness, in a transparent manner. Authors of SRs using NMAs focused on dichotomous outcomes could also consider establishing a decision threshold for each outcome that they are assessing, and estimate the rank probabilities and rankings based on those. Assessing and reporting the confidence in the rankings[10] could be useful for readers and should be encouraged.

23.5 Implications for practice

Finally, our results highlight the need to interpret rankings and rank probabilities with caution when using SRs with NMA to inform clinical practice. Although rankings are highly attractive, they could be potentially misleading if they are used as a stand-alone piece of information.

Literature users are encouraged to consider both the relative effects of the pairwise comparisons and the rankings to make their conclusions with regards to the effectiveness of a set of treatments, even if this requires a bigger effort on their part.

24 Conclusions

Rank probabilities and rankings of treatments in NMA can be fragile to changes in the RCTs included in the network, and to increases in decision thresholds to claim that one treatment is more effective than others. Although most of the time the changes in rank probabilities were of modest size, and there was a proportion small of cases in which we observed a change in the best and second treatment, there are cases in which these changes could be large. It is not always apparent from the standard NMA results which category an NMA falls into. In addition, changing the decision threshold could decrease the rank probability to an important degree. All of this highlights the need for reporting, interpreting and using rankings together with the pairwise comparison estimates. Modifying the way in which rank probabilities and rankings are estimated by including thresholds may facilitate their interpretation and use, and avoid the need to combine two pieces of information (the rankings and the pairwise estimates).

Chapter 7
Conclusion and Implications

Network meta-analysis (NMA) is a statistical tool that allows pooling of the results from a series of head-to-head comparisons (HTHC) connected in a network of evidence,[33] and the estimation of the relative effectiveness of different interventions based on direct and indirect evidence. In recent years, NMA has received a great deal of attention;[96 172] several publications have addressed the methodology,[32 34 38 41 46 94] reporting[9] and use of NMA to inform clinical practice,[5 12] and authors of systematic reviews (SR) use it more frequently.[96] Although the advantages of this evidence synthesis technique are obvious, leading some researchers to question whether it should become the norm when comparing the effectiveness of interventions, [13 173] others remain cautious and warn about the methodological issues that have arisen and which should be addressed before such claims can be made.[40 53]

25 Summary of methods and findings

In an attempt to address some of the issues that could influence whether NMA becomes a standard when performing SRs, this thesis aimed to determine: 1) the extent to which NMA can be used to answer current clinical questions; 2) whether SRs using NMA report the same results as SRs using HTHC; and 3) the robustness of the rank probabilities and rankings obtained from NMA to the omission of a single randomized clinical trial (RCT) from the network, and increasing decision thresholds.

To determine the extent that NMA can be used to answer current clinical questions, we performed a systematic survey of all the SRs of RCTs published in the Cochrane Database of Systematic Reviews in a one-year period (July 2014 to June 2015). Based on the description of the methods and results of the reviews, we quantified the proportion of SRs that posed research questions requiring the use of NMA techniques. The reviews in this category were subcategorized according to whether they had done or not done an NMA, and the ones that had not were classified according to whether NMA could have been done. We also subcategorized the SRs that had questions in which using NMA was not strictly necessary according to whether

the authors were interested in comparing only two interventions against each other, or doing a series of HTHC.

We were surprised by our findings that only 25.3% (205 out of 809 SRs included in the survey) of the SRs had questions in which an NMA was necessary, and that only 4 of those 205 had actually performed such an analysis. Almost half of the SRs in which an NMA was necessary (95 SRs) could have performed this type of analysis but failed to do so. We were also surprised by the large proportion of SRs in which an NMA would not be needed, since the authors of the SRs were aiming to perform (and followed searching and screening methods appropriate to) a series of specific direct HTHC (340 SRs).

To assess whether SRs using NMA report the same results as SRs using HTHC, we performed a systematic survey of all the SRs with NMA that were published in journals indexed in Medline, and which compared the effects of stents in patients undergoing coronary percutaneous intervention. Then, we determined all the specific questions addressed by these SRs, in terms of patients, interventions and primary outcome (PICO questions). Next, we performed a systematic search of SRs using HTHC addressing the PICO questions found in the NMAs. Subsequently, we constructed pairs of effect estimates for each specific PICO question - the NMA estimate and the HTHC estimate - and compared them using various similarity criteria. These similarity criteria were chosen based on a scoping review of the literature and discussion among the experts involved in this thesis.[116]

The 12 SRs with NMA allowed us to construct 42 pairs of estimates, 12 perfectly matched on the PICO question and 30 with slight differences in one of the PICO question components. Depending on the similarity criteria, SRs using NMA reported the same results as SRs using HTHC in 66.7% to 83.3% of the perfectly matched pairs. These proportions were smaller in the imperfect matches, in which we observed that similarity criteria were satisfied in 44.8% to 75.9% of pairs.

To explore the robustness of the probabilities and rankings to changes in the RCTs included in the network and the decision threshold used to declare treatment effects different and calculate the rank probabilities, we performed a systematic survey of all the NMAs from the field of cardiovascular medicine that reported trial-level data. We used these data to re-analyze the NMA and assess the changes in rank probabilities and rankings when modifying the RCTs included in

the NMA (by excluding one trial at a time from the analysis) and when using different thresholds to calculate the rank probabilities. In this latter analysis, in addition to comparing the OR between treatments to the default threshold of 1 to claim superiority of one treatment over another, we used stricter thresholds of 1.1 to 1.4.

We observed that although, on average, the rank probabilities and the rankings remain reasonably constant when excluding RCTs from the analysis (with an overall mean absolute change of 4.3% in rank probabilities across NMAs), there are cases in which excluding one RCT can result in dramatic increases or decreases of the rank probabilities, and switches in the treatments ranked first and second. We also observed that increasing the threshold to claim superiority may result in important changes in rank probabilities, which in some cases lead to the first treatment having extremely low probabilities of being distinguishable as the best.

26 Choice of study designs

When designing and conducting these three studies, we aimed to use sound methods, while at the same time ensuring the feasibility of what was, to our knowledge, a first attempt to address these methodological questions. Acknowledging the fact that we had to choose a specific sample to answer each question, we selected each sample based on relevance and quality. For our first study, we decided to select Cochrane SRs because these reviews are accepted as one of the most trustworthy sources of evidence summaries relevant to clinical practice, and the topics and scope that they cover undergo a process of prioritization and approval by formal editorial groups.[98] In addition, these SRs report their methods and results with much more detail than SRs not published in the Cochrane database.[174 175] This thorough reporting was key in allowing us, with the most certainty possible, to categorize the type of question posed by the review, and to identify other characteristics explored in our study.

For our second and third studies, we chose to sample SRs with NMAs from the field of cardiovascular medicine. This is the field with the largest published number of NMAs, according to the results of our systematic search and classification, a finding that was supported by a bibliometric study.[96] For our second study we chose SRs using NMA to assess the effects of stents in patients undergoing percutaneous coronary intervention because our search identified these as the most common interventions addressed in the reviews. An alternative would have been to choose one SR from each of the topics or groups of interventions, but this would have

meant arbitrary selection of specific SRs from each topic and a result that would have had limited generalizability to other topics within cardiovascular medicine. Our approach provides a complete view of all the SRs with NMA in the chosen topic and avoids biases due to the selective inclusion of specific SRs. For our third study, the selection of our sample was based on the availability of trial level data that would allow us to reproduce the NMAs, without having to perform data abstraction of numerous primary studies. We acknowledge, however, that despite the fact that we strived for relevant samples, the applicability of our results may be threatened. Specific applicability concerns were discussed in each of the chapters. On the other hand, it must be noted that studies addressing methodological questions are typically selected based on specific, purposive samples, even in the context of NMA.[39 41 54]

26.1 NMA as a standard for evidence summaries

Each of our three studies provides information related to our initial question - the extent to which NMA should become a standard for evidence summaries to inform evidence-based clinical practice. One of the main advantages highlighted by proponents of this technique is that it provides a broad view of the evidence base, in which all the relevant treatments (usually numbering more than two) for a specific condition are considered. This makes it more compatible with the clinical decision-making process. [5 13 173] Even though this claim appears to be very sensible, it stands in contrast to the results of our first study, where we found that only a quarter of Cochrane reviews published over the course of a year aim to answer such broad questions, where all the interventions are considered to be equally relevant. In the chapter describing our first study, we speculate about possible reasons, such as little familiarity with NMA methods, the desire to publish more SRs by splitting one overarching question into several smaller ones, and technological limitations such as NMA not being yet implemented in friendly software to perform meta-analysis (for example, RevMan); three factors that could influence the selection of the scope of the review. These potential reasons are only speculation; given the stated scope of the reviews in our study, the only firm conclusion that we can draw is that NMA is not needed in about 75% of current systematic reviews.

A factor that could influence the use of NMA by clinicians is whether there are important differences between the estimates obtained for a specific PICO question when a SR was undertaken using NMA versus a more traditional HTHC. NMA is still a relatively new technique

and as such, there may be low penetration of the skills needed for executing or understanding such an analysis. The use of NMA as the main analytic approach when undertaking SRs, interpreting the results, and informing clinical practice requires the systematic reviewer to acquire a new set of skills. Clinicians who aim to inform their practice by evidence in SRs with NMA also need to learn new critical appraisal skills. If the main interest is in a specific comparison between a pair of treatments, it is reasonable to question why literature users should make the extra effort to use SRs with NMA. If SRs with NMA report essentially the same results as SRs with HTHC, the choice between one and the other would not make a big difference. Matters are less clear, however, if the results differ depending on the type of SR, as literature users should be able to identify which of the estimates (the one from the HTHC or the one from the NMA) is more trustworthy. In our study we observed that, more frequently than expected, results differ, and this opens the door for research questions to identify reasons for these differences, and suggests a need for guidance for literature users. Even more importantly, it highlights the need to perform a detailed critical appraisal of the methods, results and applicability of SRs whether they use HTHC or NMA, as it is not clear which of the designs is more valid and this could also vary depending on the specific case.

Another key advantage of NMA is the ability to rank treatments for a specific outcome, that is, identify which treatment is the best, second best, third best, and so on.[36] This piece of information is easy to understand and user-friendly, attributes which may make NMA attractive and increase its uptake. Nevertheless, users have been warned to be cautious when interpreting rankings as they reflect only the ordering of a set of treatment effects, and say nothing about the magnitudes of the differences between them and the quality of the evidence behind them. Furthermore, the rank probabilities and rankings may be fragile in response to exclusion of RCTs or treatments from the analysis.[53 54 154 155] Our study supports claims of fragility by showing that even though on average the change in rank probabilities when excluding a single trial from the analysis is small, there are cases in which the change can be large and important. It also supports claims about the need to exercise caution when interpreting the rankings; in some cases, there are dramatic decreases in the probability that the “best” treatment is in fact the best when the threshold to claim superiority is increased. This highlights the need to judge the confidence in the rankings and interpret them together with the pairwise estimates of relative

effectiveness, or switch to an approach in which the thresholds are used to estimate the rank probabilities and the rankings either as the main analysis or as sensitivity analysis.

26.2 Implications for future work

Our studies have several implications for research on NMA methods, and for authors and users of SRs. These were discussed in detail in each of the chapters. In general terms, our findings provide insight with regards to some of the issues that may hinder NMA use, contribute to the growing amount of research concerning NMA methods, and emphasizes how important it is to remain cautious when interpreting and using results from such analysis. Our findings also are a call to answer key research questions concerning NMA: Is NMA used so scarcely, relative to all the SRs published, because it is not compatible with current clinical questions or because the questions and scope of current SRs is formulated based on knowledge of SR methods and technological limitations? What are the key methodological features that SRs users should appraise when deciding how to resolve differences in a specific treatment effect that may be found in SRs with NMAs versus SRs with HTHC? Is it feasible to incorporate sensitivity analyses to explore the impact of increasingly stringent thresholds to claim superiority (at both the software level and the user level) to estimate the rank probabilities and rankings, and thereby solve the issue of the usual rankings being potentially misleading?

26.3 Conclusion

In conclusion, the uptake of NMA may be hampered by the current nature of clinical questions, the differences in treatment effect estimates that may arise when using NMA as opposed to HTHC, and the fragility of the rankings and rank probabilities to small changes in the included RCTs and to varying thresholds to claim superiority. These issues also suggest that because NMA is still in its infancy compared to HTHC, more research and guidance for its use are necessary before it can be claimed that NMA should become the standard for comparing treatment effectiveness. On the other hand, the fast-growing popularity of NMA makes it very likely that clinicians will be faced with this type of analysis when using evidence to inform their practice, a situation that highlights the need for research to help overcome the issues explored in this thesis.

The issues raised in this thesis do not lead to a claim that NMA will not, or should not, become the standard for evidence summaries that inform clinical practice. Instead they should be interpreted as a call for further research to address the main implications of our three studies.

References

1. Rawlins M. In pursuit of quality: the National Institute for Clinical Excellence. Lancet 1999;**353**(9158):1079-82

2. Guyatt GH, Haynes B, Jaeschke R, et al. The philosophy of evidence-based medicine. In: Guyatt GH, Rennie D, Meade MO, et al., eds. Users' guides to the medical literature: a manual for evidence based clinical practice. 2nd ed: McGraw Hill, 2008:9-16.

3. Caldwell DM, Ades AE, Higgins JP. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. Bmj 2005;**331**(7521):897-900

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

5. Jansen JP, Fleurence R, Devine B, et al. 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 2011;**14**(4):417-28

6. Lumley T. Network meta-analysis for indirect treatment comparisons. Stat Med 2002;**21**(16):2313-24

7. Sutton A, Ades AE, Cooper N, et al. Use of indirect and mixed treatment comparisons for technology assessment. Pharmacoeconomics 2008;**26**(9):753-67

8. Hunt M. *How science takes tool*. New York: Russel Sage Found, 1997.

9. Hutton B, Salanti G, Caldwell DM, et al. The PRISMA extension statement for reporting of systematic reviews incorporating network meta-analyses of health care interventions: checklist and explanations. Annals of internal medicine 2015;**162**(11):777-84

10. Salanti G, Del Giovane C, Chaimani A, et al. Evaluating the quality of evidence from a network meta-analysis. PloS one 2014;**9**(7):e99682

11. Puhan MA, Schunemann HJ, Murad MH, et al. A GRADE Working Group approach for rating the quality of treatment effect estimates from network meta-analysis. Bmj 2014;**349**:g5630

12. Mills EJ, Ioannidis JP, Thorlund K, et al. How to use an article reporting a multiple treatment comparison meta-analysis. Jama 2012;**308**(12):1246-53

13. Higgins JP, Welton NJ. Network meta-analysis: a norm for comparative effectiveness? Lancet (London, England) 2015;**386**(9994):628-30

14. Giovane CD, Vacchi L, Mavridis D, et al. Network meta-analysis models to account for variability in treatment definitions: application to dose effects. Stat Med 2013;**32**(1):25-39

15. Mavridis D, Sutton A, Cipriani A, et al. A fully Bayesian application of the Copas selection model for publication bias extended to network meta-analysis. Stat Med 2013;**32**(1):51-66

16. Jansen JP, Schmid CH, Salanti G. Directed acyclic graphs can help understand bias in indirect and mixed treatment comparisons. J Clin Epidemiol 2012;**65**(7):798-807

17. Salanti G, Ades AE, Ioannidis JP. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial. J Clin Epidemiol 2011;**64**(2):163-71

18. Salanti G, Marinho V, Higgins JP. A case study of multiple-treatments meta-analysis demonstrates that covariates should be considered. J Clin Epidemiol 2009;**62**(8):857-64 19. Salanti G, Higgins JP, Ades AE, et al. Evaluation of networks of randomized trials. Stat Methods Med Res 2008;**17**(3):279-301

20. Dias S, Welton NJ, Caldwell DM, et al. Checking consistency in mixed treatment comparison meta-analysis. Stat Med 2010;**29**(7-8):932-44

21. Lu G, Ades A. Modeling between-trial variance structure in mixed treatment comparisons. Biostatistics 2009;**10**(4):792-805

22. Caldwell DM, Welton NJ, Ades AE. Mixed treatment comparison analysis provides internally coherent treatment effect estimates based on overviews of reviews and can reveal inconsistency. J Clin Epidemiol 2010;**63**(8):875-82

23. Efthimiou O, Mavridis D, Cipriani A, et al. An approach for modelling multiple correlated outcomes in a network of interventions using odds ratios. Stat Med 2014;**33**(13):2275-87

24. Mavridis D, Chaimani A, Efthimiou O, et al. Addressing missing outcome data in meta-analysis. Evidence-based mental health 2014;**17**(3):85-9

25. Mavridis D, Chaimani A, Efthimiou O, et al. Missing outcome data in meta-analysis. Evidence-based mental health 2014 doi: 10.1136/eb-2014-101899

26. Mavridis D, Welton NJ, Sutton A, et al. A selection model for accounting for publication bias in a full network meta-analysis. Stat Med 2014;**33**(30):5399-412

27. Mavridis D, White IR, Higgins JP, et al. Allowing for uncertainty due to missing continuous outcome data in pairwise and network meta-analysis. Stat Med 2015;**34**(5):721-41

28. Veroniki AA, Mavridis D, Higgins JP, et al. Characteristics of a loop of evidence that affect detection and estimation of inconsistency: a simulation study. BMC medical research methodology 2014;**14**:106

29. O'Regan C, Ghement I, Eyawo O, et al. Incorporating multiple interventions in meta-analysis: an evaluation of the mixed treatment comparison with the adjusted indirect comparison. Trials 2009;**10**:86

30. Song F, Xiong T, Parekh-Bhurke S, et al. Inconsistency between direct and indirect comparisons of competing interventions: meta-epidemiological study. Bmj 2011;**343**:d4909

31. Jansen JP, Crawford B, Bergman G, et al. Bayesian meta-analysis of multiple treatment comparisons: an introduction to mixed treatment comparisons. Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research 2008;**11**(5):956-64.

32. Hoaglin DC, Hawkins N, Jansen JP, et al. Conducting indirect-treatment-comparison and network-meta-analysis studies: report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices: part 2. Value Health 2011;**14**(4):429-37

33. Higgins JP, Whitehead A. Borrowing strength from external trials in a meta-analysis. Stat Med 1996;**15**(24):2733-49

34. Lu G, Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons. Statistics in medicine 2004;**23**(20):3105-24

35. Thorlund K, Zafari Z, Druyts E, et al. The impact of incorporating Bayesian network meta-analysis in cost-effectiveness analysis - a case study of pharmacotherapies for moderate to severe COPD. Cost effectiveness and resource allocation : C/E 2014;**12**(1):8

36. Efthimiou O, Debray TP, van Valkenhoef G, et al. GetReal in network meta-analysis: a review of the methodology. Res Synth Methods 2016 doi: 10.1002/jrsm.1195

37. Laws A, Kendall R, Hawkins N. A comparison of national guidelines for network meta-analysis. Value Health 2014;**17**(5):642-54.

38. Hawkins N, Scott DA, Woods B. How far do you go? Efficient searching for indirect evidence. Medical decision making : an international journal of the Society for Medical Decision Making 2009;**29**(3):273-81

39. Hawkins N, Scott DA, Woods BS, et al. No study left behind: a network meta-analysis in non-small-cell lung cancer demonstrating the importance of considering all relevant data. Value Health 2009;**12**(6):996-1003

40. Li T, Puhan MA, Vedula SS, et al. Network meta-analysis-highly attractive but more methodological research is needed. BMC medicine 2011;**9**:79

41. Salanti G, Kavvoura FK, Ioannidis JP. Exploring the geometry of treatment networks. Annals of internal medicine 2008;**148**(7):544-53

42. Balshem H, Helfand M, Schunemann HJ, et al. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol 2011;**64**(4):401-6

43. Cipriani A, Higgins JP, Geddes JR, et al. Conceptual and technical challenges in network meta-analysis. Annals of internal medicine 2013;**159**(2):130-7

44. Neupane B, Richer D, Bonner AJ, et al. Network meta-analysis using R: a review of currently available automated packages. PloS one 2014;**9**(12):e115065

45. Lu G, Welton NJ, Higgins JP, et al. Linear inference for mixed treatment comparison meta-analysis: A two-stage approach. Res Synth Methods 2011;**2**(1):43-60

46. White IR, Barrett JK, Jackson D, et al. Consistency and inconsistency in network meta-analysis: model estimation using multivariate meta-regression. Res Synth Methods 2012;**3**(2):111-25

47. Krahn U, Binder H, Konig J. A graphical tool for locating inconsistency in network meta-analyses. BMC medical research methodology 2013;**13**:35

48. Rucker G. Network meta-analysis, electrical networks and graph theory. Res Synth Methods 2012;**3**(4):312-24

49. Zhang J, Carlin BP, Neaton JD, et al. Network meta-analysis of randomized clinical trials: reporting the proper summaries. Clinical trials (London, England) 2014;**11**(2):246-62

50. R: A Language and Environment for Statistical Computing [program]. Vienna, Austria: R Foundation for Statistical Computing, 2008.

51. Review Manager (RevMan) [program]. 5.3 version. Copenhagen, 2014.

52. Lu G, Ades A. Assessing Evidence Inconsistency in Mixed Treatment Comparisons. Journal of the American Statistical Society 2006;**101**(474):447-59

53. Mills EJ, Thorlund K, Ioannidis JP. Demystifying trial networks and network meta-analysis. Bmj 2013;**346**:f2914

54. Mills EJ, Kanters S, Thorlund K, et al. The effects of excluding treatments from network meta-analyses: survey. Bmj 2013;**347**:f5195

55. Akl EA, Carrasco-Labra A, Brignardello-Petersen R, et al. Reporting, handling and assessing the risk of bias associated with missing participant data in systematic reviews: a methodological survey. BMJ open 2015;**5**(9):e009368

56. Akl EA, Kahale LA, Agoritsas T, et al. Handling trial participants with missing outcome data when conducting a meta-analysis: a systematic survey of proposed approaches. Systematic reviews 2015;**4**:98 doi: 10.1186/s13643-015-0083-6[published Online First: Epub Date]|.

57. Akl EA, Shawwa K, Kahale LA, et al. Reporting missing participant data in randomised trials: systematic survey of the methodological literature and a proposed guide. BMJ open 2015;**5**(12):e008431

58. Mills EJ, Kelly S, Wu P, et al. Epidemiology and reporting of randomized trials employing re-randomization of patient groups: a systematic survey. Contemporary clinical trials 2007;**28**(3):268-75

59. Wu D, Akl EA, Guyatt GH, et al. Methodological survey of designed uneven randomization trials (DU-RANDOM): a protocol. Trials 2014;**15**:33

60. Alonso-Coello P, Carrasco-Labra A, Brignardello-Petersen R, et al. Systematic reviews experience major limitations in reporting absolute effects. J Clin Epidemiol 2015 doi: 10.1016/j.jclinepi.2015.11.002

61. Evaniew N, Files C, Smith C, et al. The fragility of statistically significant findings from randomized trials in spine surgery: a systematic survey. The spine journal : official journal of the North American Spine Society 2015;**15**(10):2188-97

62. Sun X, Briel M, Busse JW, et al. Subgroup Analysis of Trials Is Rarely Easy (SATIRE): a study protocol for a systematic review to characterize the analysis, reporting, and claim of subgroup effects in randomized trials. Trials 2009;**10**:101

63. Lilford RJ, Richardson A, Stevens A, et al. Issues in methodological research: perspectives from researchers and commissioners. Health technology assessment (Winchester, England) 2001;**5**(8):1-57

64. Murad H, Jaeschke R, Devereaux P, et al. The process of a systematic reciew and meta-analysis. In: Guyatt G, Rennie D, Meade M, et al., eds. Users' guide to the medical literature A manual for evidence-based clinical practice 3rd ed, 2015.

65. Guyatt G, Rennie D, Meade M, et al. *Users' guide to the medical literature. A manual for evidence-based clinical practice*. 3rd ed: Mc Graw Hill Education, 2015.

66. Higgins J, Altman D, Sterne J. Chapter 8: Assesing risk of bias in included studies. In: Higgins J, Green S, eds. Cochrane Handbook for Systematic Reviews of Interventions: The Cochrane Collaboration, 2011.

67. Dahabreh IJ, Sheldrick RC, Paulus JK, et al. Do observational studies using propensity score methods agree with randomized trials? A systematic comparison of studies on acute coronary syndromes. European heart journal 2012;**33**(15):1893-901

68. Guyatt GH, DiCenso A, Farewell V, et al. Randomized trials versus observational studies in adolescent pregnancy prevention. J Clin Epidemiol 2000;**53**(2):167-74

69. Lonjon G, Boutron I, Trinquart L, et al. Comparison of treatment effect estimates from prospective nonrandomized studies with propensity score analysis and randomized controlled trials of surgical procedures. Annals of surgery 2014;**259**(1):18-25

70. Hong H, Carlin BP, Shamliyan TA, et al. Comparing Bayesian and frequentist approaches for multiple outcome mixed treatment comparisons. Medical decision making : an international journal of the Society for Medical Decision Making 2013;**33**(5):702-14

71. Schulz KF, Chalmers I, Hayes RJ, et al. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. Jama 1995;**273**(5):408-12

72. Savovic J, Jones H, Altman D, et al. Influence of reported study design characteristics on intervention effect estimates from randomised controlled trials: combined analysis of meta-epidemiological studies. Health Technol Assess 2012;**16**(35):1-82.

73. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. The New England journal of medicine 2000;**342**(25):1878-86

74. Shikata S, Nakayama T, Noguchi Y, et al. Comparison of effects in randomized controlled trials with observational studies in digestive surgery. Annals of surgery 2006;**244**(5):668-76

75. Zhang Z, Ni H, Xu X. Observational studies using propensity score analysis underestimated the effect sizes in critical care medicine. J Clin Epidemiol 2014;**67**(8):932-9

76. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. The New England journal of medicine 2000;**342**(25):1887-92

77. MacLehose RR, Reeves BC, Harvey IM, et al. A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies. Health technology assessment (Winchester, England) 2000;**4**(34):1-154

78. Anglemyer A, Horvath HT, Bero L. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. Cochrane Database Syst Rev 2014;**4:MR000034.**

79. Edwards JP, Kelly EJ, Lin Y, et al. Meta-analytic comparison of randomized and nonrandomized studies of breast cancer surgery. Can J Surg 2012;**55**(3):155-62.

80. Furlan AD, Tomlinson G, Jadad AA, et al. Examining heterogeneity in meta-analysis: comparing results of randomized trials and nonrandomized studies of interventions for low back pain. Spine (Phila Pa 2008;**33**(3):339-48.

81. Mueller D, Sauerland S, Neugebauer EA, et al. Reported effects in randomized controlled trials were compared with those of nonrandomized trials in cholecystectomy. J Clin Epidemiol 2010;**63**(10):1082-90.

82. Papanikolaou PN, Christidi GD, Ioannidis JP. Comparison of evidence on harms of medical interventions in randomized and nonrandomized studies. CMAJ 2006;**174**(5):635-41.

83. Sipahi I, Celik S, Tozun N. A comparison of results of the US food and drug administration's mini-sentinel program with randomized clinical trials: the case of gastrointestinal tract bleeding with dabigatran. JAMA Intern Med 2014;**174**(1):150-1.

84. Bhandari M, Tornetta P, 3rd, Ellis T, et al. Hierarchy of evidence: differences in results between non-randomized studies and randomized trials in patients with femoral neck fractures. Arch Orthop Trauma Surg 2004;**124**(1):10-6.

85. Chou R, Carson S, Chan BK. Gabapentin versus tricyclic antidepressants for diabetic neuropathy and post-herpetic neuralgia: discrepancies between direct and indirect meta-analyses of randomized controlled trials. Journal of general internal medicine 2009;**24**(2):178-88

86. Tzoulaki I, Siontis KC, Ioannidis JP. Prognostic effect size of cardiovascular biomarkers in datasets from observational studies versus randomised trials: meta-epidemiology study. BMJ 2011;343:d6829**.**

87. Cappelleri JC, Ioannidis JP, Schmid CH, et al. Large trials vs meta-analysis of smaller trials: how do their results compare? JAMA 1996;**276**(16):1332-8.

88. Golder S, Loke YK, Bland M. Meta-analyses of adverse effects data derived from randomised controlled trials as compared to observational studies: methodological overview. PLoS Med 2011;**8**(5):e1001026.

89. Ioannidis JP, Haidich AB, Pappa M, et al. Comparison of evidence of treatment effects in randomized and nonrandomized studies. JAMA 2001;**286**(7):821-30.

90. Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. BMJ 1998;**317**(7167):1185-90.

91. Kuss O, Legler T, Borgermann J. Treatments effects from randomized trials and propensity score analyses were similar in similar populations in an example from cardiac surgery. J Clin Epidemiol 2011;**64**(10):1076-84.

92. Chou R, Fu R, Huffman LH, et al. Initial highly-active antiretroviral therapy with a protease inhibitor versus a non-nucleoside reverse transcriptase inhibitor: discrepancies between direct and indirect meta-analyses. Lancet 2006;**368**(9546):1503-15.

93. Song F, Harvey I, Lilford R. Adjusted indirect comparison may be less biased than direct comparison for evaluating new pharmaceutical interventions. J Clin Epidemiol 2008;**61**(5):455-63.

94. Song F, Altman DG, Glenny AM, et al. Validity of indirect comparison for estimating efficacy of competing interventions: empirical evidence from published meta-analyses. BMJ 2003;**326**(7387):472.

95. Song F, Xiong T, Parekh-Bhurke S, et al. Inconsistency between direct and indirect comparisons of competing interventions: meta-epidemiological study. BMJ 2011;343:d4909

96. Greco T, Biondi-Zoccai G, Saleh O, et al. The attractiveness of network meta-analysis: a comprehensive systematic and narrative review. Heart, lung and vessels 2015;**7**(2):133-42

97. The Cochrane Collaboration. Proposing and registering new reviews. Secondary Proposing and registering new reviews 2015. https://community.cochrane.org/cochrane-reviews/proposing-new-reviews.

98. The Cochrane Collaboration. Our vision, mission, and principles. Secondary Our vision, mission, and principles 2016. http://www.cochrane.org/about-us/our-vision-mission-and-principles.

99. Dumville JC, McFarlane E, Edwards P, et al. Preoperative skin antiseptics for preventing surgical wound infections after clean surgery. Cochrane Database Syst Rev 2015;**4**:CD003949

100. Wilhelmus KR. Antiviral treatment and other therapeutic interventions for herpes simplex virus epithelial keratitis. Cochrane Database Syst Rev 2015;**1**:CD002898

101. Palmer SC, Saglimbene V, Mavridis D, et al. Erythropoiesis-stimulating agents for anaemia in adults with chronic kidney disease: a network meta-analysis. Cochrane Database Syst Rev 2014;**12**:CD010590

102. Le Cleach L, Trinquart L, Do G, et al. Oral antiviral therapy for prevention of genital herpes outbreaks in immunocompetent and nonpregnant patients. Cochrane Database Syst Rev 2014;**8**:CD009036

103. van Zuuren EJ, Fedorowicz Z, Carter B, et al. Interventions for hirsutism (excluding laser and photoepilation therapy alone). Cochrane Database Syst Rev 2015;**4**:CD010334

104. Nieuwlaat R, Wilczynski N, Navarro T, et al. Interventions for enhancing medication adherence. Cochrane Database Syst Rev 2014;**11**:CD000011

105. Rezaie A, Kuenzig ME, Benchimol EI, et al. Budesonide for induction of remission in Crohn's disease. Cochrane Database Syst Rev 2015;**6**:CD000296

106. Maldonado Fernandez M, Birdi JS, Irving GJ, et al. Pharmacological agents for the prevention of vestibular migraine. Cochrane Database Syst Rev 2015;**6**:CD010600

107. Derry S, Wiffen PJ, Moore RA, et al. Topical lidocaine for neuropathic pain in adults. Cochrane Database Syst Rev 2014;**7**:CD010958.

108. Dumville JC, Stubbs N, Keogh SJ, et al. Hydrogel dressings for treating pressure ulcers. Cochrane Database Syst Rev 2015;**2**:CD011226

109. Dumville JC, Webster J, Evans D, et al. Negative pressure wound therapy for treating pressure ulcers. Cochrane Database Syst Rev 2015;**5**:CD011334.

110. Dumville JC, Keogh SJ, Liu Z, et al. Alginate dressings for treating pressure ulcers. Cochrane Database Syst Rev 2015;**5**:CD011277

111. Abdelhamid AS, Loke YK, Parekh-Bhurke S, et al. Use of indirect comparison methods in systematic reviews: a survey of Cochrane review authors. Res Synth Methods 2012;**3**(2):71-9

112. Achana F, Hubbard S, Sutton A, et al. An exploration of synthesis methods in public health evaluations of interventions concludes that the use of modern statistical methods would be beneficial. Journal of clinical epidemiology 2014;**67**(4):376-90

113. Otasowie J, Castells X, Ehimare UP, et al. Tricyclic antidepressants for attention deficit hyperactivity disorder (ADHD) in children and adolescents. Cochrane Database Syst Rev 2014;**9**:CD006997

114. Falissard B, Zylberman M, Cucherat M, et al. Real medical benefit assessed by indirect comparison. Therapie 2009;**64**(3):225-32

115. Prasad K, Jaeschke R, Wyer P, et al. Tips for teachers of evidence-based medicine: understanding odds ratios and their relationship to risk ratios. Journal of general internal medicine 2008;**23**(5):635-40

116. Brignardello-Petersen R, Carrasco-Labra A, Jadad AR, et al. Diverse criteria and methods are used to compare treatment effect estimates: a scoping review. J Clin Epidemiol 2016 doi: 10.1016/j.jclinepi.2016.02.001

117. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;**33**(1):159-74

118. Palmerini T, Benedetto U, Biondi-Zoccai G, et al. Long-Term Safety of Drug-Eluting and Bare-Metal Stents: Evidence From a Comprehensive Network Meta-Analysis. J Am Coll Cardiol 2015;**65**(23):2496-507

119. Palmerini T, Biondi-Zoccai G, Della Riva D, et al. Clinical outcomes with bioabsorbable polymer- versus durable polymer-based drug-eluting and bare-metal stents: evidence from a comprehensive network meta-analysis. J Am Coll Cardiol 2014;**63**(4):299-307

120. Kang SH, Park KW, Kang DY, et al. Biodegradable-polymer drug-eluting stents vs. bare metal stents vs. durable-polymer drug-eluting stents: a systematic review and Bayesian approach network meta-analysis. Eur Heart J 2014;**35**(17):1147-58

121. Palmerini T, Biondi-Zoccai G, Della Riva D, et al. Clinical outcomes with drug-eluting and bare-metal stents in patients with ST-segment elevation myocardial infarction: evidence from a comprehensive network meta-analysis. J Am Coll Cardiol 2013;**62**(6):496-504

122. Navarese EP, Tandjung K, Claessen B, et al. Safety and efficacy outcomes of first and second generation durable polymer drug eluting stents and biodegradable polymer biolimus eluting stents in clinical practice: comprehensive network meta-analysis. Bmj 2013;**347**:f6530

123. Bangalore S, Toklu B, Amoroso N, et al. Bare metal stents, durable polymer drug eluting stents, and biodegradable polymer drug eluting stents for coronary artery disease: mixed treatment comparison meta-analysis. Bmj 2013;**347**:f6625

124. Bangalore S, Amoroso N, Fusaro M, et al. Outcomes with various drug-eluting or bare metal stents in patients with ST-segment-elevation myocardial infarction: a mixed treatment comparison analysis of trial level data from 34 068 patient-years of follow-up from randomized trials.[Erratum appears in Circ Cardiovasc Interv. 2013 Dec;6(6):e80]. Circ 2013;**6**(4):378-90

125. Palmerini T, Biondi-Zoccai G, Della Riva D, et al. Stent thrombosis with drug-eluting and bare-metal stents: evidence from a comprehensive network meta-analysis. Lancet 2012;**379**(9824):1393-402

126. Bangalore S, Kumar S, Fusaro M, et al. Outcomes with various drug eluting or bare metal stents in patients with diabetes mellitus: mixed treatment comparison analysis of 22,844 patient years of follow-up from randomised trials. Bmj 2012;**345**:e5170

127. Bangalore S, Kumar S, Fusaro M, et al. Short- and long-term outcomes with drug-eluting and bare-metal coronary stents: a mixed-treatment comparison analysis of 117 762 patient-years of follow-up from randomized trials. Circulation 2012;**125**(23):2873-91

128. Stettler C, Allemann S, Wandel S, et al. Drug eluting and bare metal stents in people with and without diabetes: collaborative network meta-analysis. Bmj 2008;**337**:a1331

129. Stettler C, Wandel S, Allemann S, et al. Outcomes associated with drug-eluting and bare-metal stents: a collaborative network meta-analysis. Lancet 2007;**370**(9591):937-48

130. Toyota T, Shiomi H, Morimoto T, et al. Meta-analysis of long-term clinical outcomes of everolimus-eluting stents. Am J Cardiol 2015;**116**(2):187-94

131. Li P, Liu JP. Long-term risk of late and very late stent thrombosis in patients treated with everolimus against paclitaxel-eluting stents: an updated meta-analysis. Coron Artery Dis 2014;**25**(5):369-77

132. Zhang X, Xie J, Li G, et al. Head-to-head comparison of sirolimus-eluting stents versus paclitaxel-eluting stents in patients undergoing percutaneous coronary intervention: a meta-analysis of 76 studies. PLoS ONE 2014;**9**(5):e97934

133. Qiao Y, Bian Y, Yan X, et al. Efficacy and safety of sirolimus-eluting stents versus bare-metal stents in coronary artery disease patients with diabetes: a meta-analysis. Cardiovasc 2013;**24**(7):274-9

134. Fan J, Du H, Yin Y, et al. Efficacy and safety of zotarolimus-eluting stents compared with sirolimus-eluting stents in patients undergoing percutaneous coronary interventions--a meta-analysis of randomized controlled trials. Int J Cardiol 2013;**167**(5):2126-33

135. Sethi A, Bahekar A, Bhuriya R, et al. Zotarolimus-eluting stent versus sirolimus-eluting and paclitaxel-eluting stents for percutaneous coronary intervention: a meta-analysis of randomized trials. Arch Cardiovasc Dis 2012;**105**(11):544-56

136. Piccolo R, Cassese S, Galasso G, et al. Long-term clinical outcomes following sirolimus-eluting stent implantation in patients with acute myocardial infarction. A meta-analysis of randomized trials. Clin 2012;**101**(11):885-93

137. de Waha A, Cassese S, Park DW, et al. Everolimus-eluting versus sirolimus-eluting stents: an updated meta-analysis of randomized trials. Clin 2012;**101**(6):461-7

138. de Waha A, Dibra A, Byrne RA, et al. Everolimus-eluting versus sirolimus-eluting stents: a meta-analysis of randomized trials. Circ 2011;**4**(4):371-7

139. Zhang F, Dong L, Qian J, et al. Clinical safety and efficacy of everolimus-eluting stents compared to paclitaxel-eluting stents in patients with coronary artery disease. Ann Med 2011;**43**(1):75-9

140. Pan XH, Chen YX, Xiang MX, et al. A meta-analysis of randomized trials on clinical outcomes of paclitaxel-eluting stents versus bare-metal stents in ST-segment elevation myocardial infarction patients. J Zhejiang Univ Sci B 2010;**11**(10):754-61

141. Li YL, Wan Z, Lu WL, et al. Comparison of sirolimus- and paclitaxel-eluting stents in patients undergoing primary percutaneous coronary intervention for ST-elevation myocardial infarction: a meta-analysis of randomized trials. Clin Cardiol 2010;**33**(9):583-90

142. Juwana YB, Rasoul S, Ottervanger JP, et al. Efficacy and safety of rapamycin as compared to paclitaxel-eluting stents: a meta-analysis. J Invasive Cardiol 2010;**22**(7):312-6

143. De Luca G, Valgimigli M, Spaulding C, et al. Short and long-term benefits of sirolimus-eluting stent in ST-segment elevation myocardial infarction: a meta-analysis of randomized trials. J Thromb Thrombolysis 2009;**28**(2):200-10

144. Pan XH, Zhong WZ, Xiang MX, et al. Clinical outcomes of serolimus-eluting stents versus bare metal stents in ST-segment elevation myocardial infarction patients: a meta-analysis. Chin Med J 2009;**122**(1):88-92

145. Mahmud E, Bromberg-Marin G, Palakodeti V, et al. Clinical efficacy of drug-eluting stents in diabetic patients: a meta-analysis. J Am Coll Cardiol 2008;**51**(25):2385-95

146. Hill RA, Boland A, Dickson R, et al. Drug-eluting stents: a systematic review and economic evaluation. Health Technol Assess 2007;**11**(46):iii, xi-221

147. Schomig A, Dibra A, Windecker S, et al. A meta-analysis of 16 randomized trials of sirolimus-eluting stents versus paclitaxel-eluting stents in patients with coronary artery disease. J Am Coll Cardiol 2007;**50**(14):1373-80

148. Kastrati A, Dibra A, Eberle S, et al. Sirolimus-eluting stents vs paclitaxel-eluting stents in patients with coronary artery disease: meta-analysis of randomized trials. Jama 2005;**294**(7):819-25

149. Bavry AA, Kumbhani DJ, Helton TJ, et al. What is the risk of stent thrombosis associated with the use of paclitaxel-eluting stents for percutaneous coronary intervention?: a meta-analysis. J Am Coll Cardiol 2005;**45**(6):941-6

150. Shea BJ, Hamel C, Wells GA, et al. AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews. Journal of clinical epidemiology 2009;**62**(10):1013-20

151. Whiting P, Savovic J, Higgins JP, et al. ROBIS: A new tool to assess risk of bias in systematic reviews was developed. Journal of clinical epidemiology 2016;**69**:225-34

152. Song F, Glenny AM, Altman DG. Indirect comparison in evaluating relative efficacy illustrated by antimicrobial prophylaxis in colorectal surgery. Controlled clinical trials 2000;**21**(5):488-97

153. Song F, Altman DG, Glenny AM, et al. Validity of indirect comparison for estimating efficacy of competing interventions: empirical evidence from published meta-analyses. Bmj 2003;**326**(7387):472

154. Neumann I, Brignardello-Petersen R, Guyatt G. ACP Journal Club. Review: Novel oral anticoagulants reduce stroke more than ASA in nonvalvular atrial fibrillation. Annals of internal medicine 2014;**160**(4):Jc3

155. Brignardello-Petersen R, Rochwerg B, Guyatt GH. What is a network meta-analysis and how can we use it to inform clinical practice? Polskie Archiwum Medycyny Wewnetrznej 2014;**124**(12):659-60

156. Brooks SP, Gelman A. Alternative methods for monitoring convergence of iterative simulations. J Comput Graph Stat 1998;**7**:434-45

157. van Valkenhoef G, Lu G, de Brock B, et al. Automating network meta-analysis. Research synthesis methods 2012;**3**(4):285-99

158. Dogliotti A, Paolasso E, Giugliano RP. Current and new oral antithrombotics in non-valvular atrial fibrillation: a network meta-analysis of 79 808 patients. Heart 2014;**100**(5):396-405

159. Dooley C, Kaur R, Sobieraj DM. Comparison of the efficacy and safety of low molecular weight heparins for venous thromboembolism prophylaxis in medically ill patients. Curr Med Res Opin 2014;**30**(3):367-80

160. Castellucci LA, Cameron C, Le Gal G, et al. Efficacy and safety outcomes of oral anticoagulants and antiplatelet drugs in the secondary prevention of venous thromboembolism: systematic review and network meta-analysis. Bmj 2013;**347**:f5133

161. Landoni G, Greco T, Biondi-Zoccai G, et al. Anaesthetic drugs and survival: a Bayesian network meta-analysis of randomized trials in cardiac surgery. Br J Anaesth 2013;**111**(6):886-96

162. Navarese EP, Buffon A, Andreotti F, et al. Meta-analysis of impact of different types and doses of statins on new-onset diabetes mellitus. Am J Cardiol 2013;**111**(8):1123-30

163. Wu HY, Huang JW, Lin HJ, et al. Comparative effectiveness of renin-angiotensin system blockers and other antihypertensive drugs in patients with diabetes: systematic review and bayesian network meta-analysis. Bmj 2013;**347**:f6008

164. Bash LD, Buono JL, Davies GM, et al. Systematic review and meta-analysis of the efficacy of cardioversion by vernakalant and comparators in patients with atrial fibrillation. Cardiovasc Drugs Ther 2012;**26**(2):167-79

165. Harenberg J, Marx S, Dahl OE, et al. Interpretation of endpoints in a network meta-analysis of new oral anticoagulants following total hip or total knee replacement surgery. Thromb Haemost 2012;**108**(5):903-12

166. Phung OJ, Kahn SR, Cook DJ, et al. Dosing frequency of unfractionated heparin thromboprophylaxis: a meta-analysis. Chest 2011;**140**(2):374-81

167. Sciarretta S, Palano F, Tocci G, et al. Antihypertensive treatment and development of heart failure in hypertension: a Bayesian network meta-analysis of studies in patients with hypertension and high cardiovascular risk. Arch Intern Med 2011;**171**(5):384-94

168. Roskell NS, Lip GY, Noack H, et al. Treatments for stroke prevention in atrial fibrillation: a network meta-analysis and indirect comparisons versus dabigatran etexilate. Thromb Haemost 2010;**104**(6):1106-15

169. Coleman CI, Baker WL, Kluger J, et al. Antihypertensive medication and their impact on cancer incidence: a mixed treatment comparison meta-analysis of randomized controlled trials. J Hypertens 2008;**26**(4):622-9

170. Cooper NJ, Sutton AJ, Lu G, et al. Mixed comparison of stroke prevention treatments in individuals with nonrheumatic atrial fibrillation. Arch Intern Med 2006;**166**(12):1269-75

171. Psaty BM, Lumley T, Furberg CD, et al. Health outcomes associated with various antihypertensive therapies used as first-line agents: a network meta-analysis. Jama 2003;**289**(19):2534-44

172. Nikolakopoulou A, Chaimani A, Veroniki AA, et al. Characteristics of networks of interventions: a description of a database of 186 published networks. PloS one 2014;**9**(1):e86754

173. Crequit P, Trinquart L, Yavchitz A, et al. Wasted research when systematic reviews fail to provide a complete and up-to-date evidence synthesis: the example of lung cancer. BMC medicine 2016;**14**(1):8

174. Jadad AR, Cook DJ, Jones A, et al. Methodology and reports of systematic reviews and meta-analyses: a comparison of Cochrane reviews with articles published in paper-based journals. Jama 1998;**280**(3):278-80

175. Shea B, Moher D, Graham I, et al. A comparison of the quality of Cochrane reviews and systematic reviews published in paper-based journals. Evaluation & the health professions 2002;**25**(1):116-29

176. Canadian Medical Association. Specialties. Canadian specialty profiles. Secondary Specialties. Canadian specialty profiles 2016. https://www.cma.ca/En/Pages/specialty-profiles.aspx.

Appendices

Appendix 1: List of specialties used to classify the systematic reviews

(based on the Canadian Medical Association Website)[176]

1. Anatomical pathology
2. Anesthesiology
3. Cardiology
4. Cardiovascular/thoracic surgery
5. Clinical immunology/allergy
6. Dentistry
7. Dermatology
8. Diagnostic radiology
9. Emergency medicine
10. Endocrinology/metabolism
11. Family medicine
12. Gastroenterology
13. General Internal Medicine
14. General/clinical pathology
15. General surgery
16. Geriatric medicine
17. Hematology
18. Medical biochemistry
19. Medical genetics
20. Medical oncology
21. Medical microbiology and infectious diseases

22. Nephrology
23. Neurology
24. Neurosurgery
25. Nuclear medicine
26. Obstetrics/gynecology
27. Occupational medicine
28. Ophthalmology
29. Orthopedic Surgery
30. Otolaryngology
31. Pediatrics
32. Physical medicine and rehabilitation
33. Plastic surgery
34. Psychiatry
35. Public health and preventive medicine
36. Radiation oncology
37. Respiratory medicine/respirology
38. Rheumatology
39. Urology
40. Vascular Surgery
41. Other

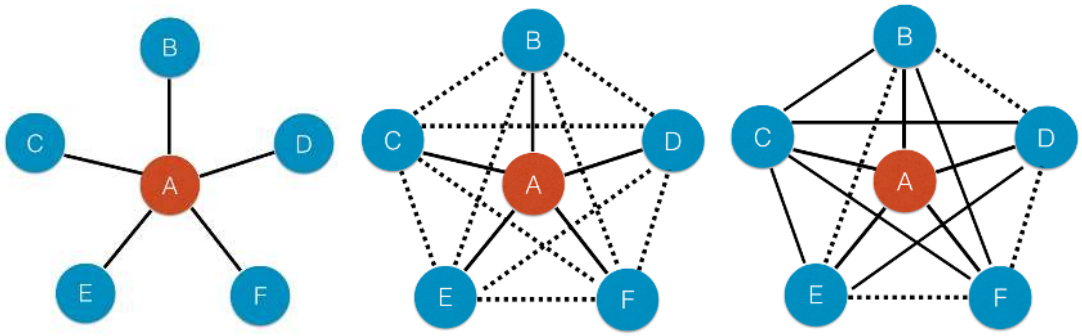
Appendix 2: SRs according to whether they had a network meta-analysis question per medical area

Medical area	Not aiming to answer an NMA question (n,% from area)	Aiming to answer an NMA question (n,% from area)	Total (n, % from all SRs)
Anesthesiology	**8 (88.9)**	**1 (11.1)**	**9 (1.1)**
Cardiology	**11 (91.7)**	**1 (8.3)**	**12 (1.5)**
Cardiovascular/Thoracic Surgery	**7 (70)**	**3 (30)**	**10 (1.2)**
Clinical Immunology/Allergy	**3 (50)**	**3 (50)**	**6 (0.7)**
Dentistry	**5 (25)**	**15 (75)**	**20 (2.5)**
Dermatology	**2 (22.2)**	**7 (77.8)**	**9 (1.1)**
Emergency Medicine	**8 (80)**	**2 (20)**	**10 (1.2)**
Endocrinology/Metabolism	**2 (50)**	**2 (50)**	**4 (0.5)**
Family Medicine	**21 (75)**	**7 (25)**	**28 (3.5)**
Gastroenterology	**21 (75)**	**7 (25)**	**28 (3.5)**
General Internal Medicine	**22 (84.6)**	**4 (15.4)**	**26 (3.2)**
General Surgery	**17 (77.3)**	**5 (22.7)**	**22 (2.7)**
Geriatric Medicine	**6 (85.7)**	**1 (14.3)**	**7 (0.9)**
Hematology	**14 (51.9)**	**13 (48.1)**	**27 (3.3)**
Medical Biochemistry	**2 (100)**	**0**	**2 (0.2)**
Medical Genetics	**3 (100)**	**0**	**3 (0.4)**
Medical Oncology	**28 (87.5)**	**4 (12.5)**	**32 (4)**
Medical Microbiology And Infectious Diseases	**2 (66.7)**	**1 (33.3)**	**3 (0.4)**
Nephrology	**9 (50)**	**9 (50)**	**18 (2.2)**
Neurology	**44 (74.6)**	**15 (25.4)**	**59 (7.3)**
Neurosurgery	**6 (85.7)**	**1 (14.3)**	**7 (0.9)**
Obstetrics/Gynecology	**81 (73.6)**	**29 (26.4)**	**102 (13.6)**
Occupational Medicine	**2 (66.7)**	**1 (33.3)**	**11 (0.4)**
Ophthalmology	**21 (70)**	**9 (30)**	**30 (3.7)**
Orthopedic Surgery	**16 (66.7)**	**8 (33.3)**	**24 (3)**
Otolaryngology	**11 (91.7)**	**1 (8.3)**	**12 (1.5)**
Pediatrics	**40 (80)**	**10 (20)**	**50 (6.2)**
Physical Medicine And Rehabilitation	**23 (88.5)**	**3 (11.5)**	**26 (3.2)**
Plastic Surgery	**2 (100)**	**0**	**2 (0.2)**
Psychiatry	**45 (84.9)**	**8 (15.1)**	**53 (6.6)**

Public Health And Preventive Medicine	**26 (81.2)**	**6 (18.8)**	**32 (4)**
Radiation Oncology	**3 (60)**	**2 (40)**	**5 (0.6)**
Respiratory Medicine/Respirology	**36 (72)**	**14 (28)**	**50 (6.2)**
Rheumatology	**16 (94.1)**	**1 (5.9)**	**17 (2.1)**
Urology	**8 (72.7)**	**3 (27.3)**	**11 (1.4)**
Vascular surgery	**8 (72.7)**	**3 (27.3)**	**11 (1.4)**
Other	**25 (80.6)**	**6 (19.4)**	**31 (3.8)**

Appendix 3: Diagram of a network in which authors are interested in comparing one treatment against many others or many treatments against control

A B C



Each circle represents one treatment. If SR authors are interested in comparing treatment A against all the other treatments, they could perform an NMA but they do not need to do so. When authors do not have an NMA approach in mind they would only search for and include the trials that compare A against the other treatments directly (solid lines). Even though they could perform an NMA, they would have a star-shaped network (A) and comparisons of other treatments against each other would be only indirect (B, dashed lines). When authors have an NMA in mind, the search is likely to result in including more direct comparisons (C); however, since the clinical question does not cover the comparison of the other treatments against each other, it would not be necessary to use an NMA approach.

Appendix 4: Search strategies for systematic reviews with network meta-analysis and head-to-head comparisons assessing the effects of stents in patients undergoing percutaneous coronary intervention

Systematic review section (1)

1. Meta-Analysis as Topic/
2. meta analy$.tw.
3. metaanaly$.tw.
4. Meta-Analysis/
5. (systematic adj (review$1 or overview$1)).tw.
6. exp Review Literature as Topic/
7. metanaly$.tw.
8. or/1-7
9. cochrane.ab.
10. embase.ab.
11. (psychlit or psyclit).ab.
12. (psychinfo or psycinfo).ab.
13. (cinahl or cinhal).ab.
14. science citation index.ab.
15. bids.ab.
16. cancerlit.ab.
17. medline.ab.
18. pubmed.ab.
19. lilacs.ab.
20. scopus.ab.
21. "web of science".ab.
22. bibliographic database?.ab.
23. electronic database?.ab.
24. or/9-23
25. reference list$.ab.
26. bibliograph$.ab.

27. hand-search$.ab.

28. relevant journals.ab.

29. manual search$.ab.

30. handsearch*.ab.

31. manually search*.ab.

32. or/25-31

33. selection criteria.ab.

34. data extraction.ab.

35. inclusion criteria.ab.

36. exclusion criteria.ab.

37. or/33-36

38. Review/

39. 37 and 38

40. Comment/

41. Letter/

42. Editorial/

43. animal/

44. human/

45. 43 not (43 and 44)

46. or/40-42,45

47. 8 or 24 or 32 or 37

48. 47 not 46

Network meta-analysis section (2)

49. (multiple treatment* adj3 (comparison* or meta analy* or meta-analy* or metanaly* or metaanaly*)).mp.

50. (mixed treatment* adj3 (comparison* or meta analy* or meta-analy* or metanaly* or metaanaly*)).mp.

51. (indirect adj4 (comparison* or meta analy* or meta-analy* or metanaly* or metaanaly*)).mp.

52. (network adj3 (meta analy* or meta-analy* or metanaly* or metaanaly*)).mp.

53. (network adj2 evidence).mp.

54. ((evidence or comparison) adj3 (combination of direct and indirect)).mp.

55. ((evidence or treatment* comparison or combination) adj4 (direct and indirect)).mp.

56. ((multi-treatment or multitreatment) adj3 (meta-analy* or meta-analy* or metanaly* or metaanaly*)).mp.

57. 49 or 50 or 51 or 52 or 53 or 54 or 55 or 56

58. 48 and 57

Stents in patients undergoing percutaneous coronary intervention section (3)

1 Stents/

2 stent$.tw.

3 or/1-2

4 drug elut$.tw.

5 exp Rapamycin/

6 sirolimus.tw.

7 Paclitaxe l/

8 paclitaxel .tw.

9 exp immunosuppressive agents/

10 coat$ stent$.tw.

11 rapamycin.tw.

12 exp Taxoids/

13 taxane$.tw.

14 taxol.tw.

15 qp2.tw.

16 hexanoyltaxol.tw.

17 everolimus.tw.

18 abt-578.tw.

19 Tacrolimus/

20 Dactinomycin/

21 actinomycin.tw.

22 batimastat.tw.

23 exp Dexamethasone/

24 dexamethasone.tw.

25 exp Estradiol/

26 estradiol.tw.

27 or/4-26

28 3 and 27

29 eluting stent$.tw.

30 28 or 29

Search strategy for systematic reviews using network meta-analysis

Section 1 AND Section 2 AND Section 3

Search strategy for systematic reviews using head-to-head comparisons

Section 1 AND Section 3

Appendix 5: Search strategy for chapter 6

Ovid MEDLINE(R) 1946 to Feb Week 2 2015

1. Meta-Analysis as Topic/
2. meta analy$.tw.
3. metaanaly$.tw.
4. Meta-Analysis/
5. (systematic adj (review$1 or overview$1)).tw.
6. exp Review Literature as Topic/
7. metanaly$.tw.
8. or/1-7
9. cochrane.ab.
10. embase.ab.
11. (psychlit or psyclit).ab.
12. (psychinfo or psycinfo).ab.
13. (cinahl or cinhal).ab.
14. science citation index.ab.
15. bids.ab.
16. cancerlit.ab.
17. medline.ab.
18. pubmed.ab.
19. lilacs.ab.
20. scopus.ab.
21. "web of science".ab.
22. bibliographic database?.ab.
23. electronic database?.ab.
24. or/9-23
25. reference list$.ab.
26. bibliograph$.ab.
27. hand-search$.ab.
28. relevant journals.ab.
29. manual search$.ab.
30. handsearch*.ab.

31. manually search*.ab.

32. or/25-31

33. selection criteria.ab.

34. data extraction.ab.

35. inclusion criteria.ab.

36. exclusion criteria.ab.

37. or/33-36

38. Review/

39. 37 and 38

40. Comment/

41. Letter/

42. Editorial/

43. animal/

44. human/

45. 43 not (43 and 44)

46. or/40-42,45

47. 8 or 24 or 32 or 37

48. 47 not 46

49. (multiple treatment* adj3 (comparison* or meta analy* or meta-analy* or metanaly* or metaanaly*)).mp.

50. (mixed treatment* adj3 (comparison* or meta analy* or meta-analy* or metanaly* or metaanaly*)).mp.

51. (indirect adj4 (comparison* or meta analy* or meta-analy* or metanaly* or metaanaly*)).mp.

52. (network adj3 (meta analy* or meta-analy* or metanaly* or metaanaly*)).mp.

53. (network adj2 evidence).mp.

54. ((evidence or comparison) adj3 (combination of direct and indirect)).mp.

55. ((evidence or treatment* comparison or combination) adj4 (direct and indirect)).mp.

56. ((multi-treatment or multitreatment) adj3 (meta-analy* or meta-analy* or metanaly* or metaanaly*)).mp.

57. 49 or 50 or 51 or 52 or 53 or 54 or 55 or 56

58. 48 and 57

Appendix 6: Calculation of outcomes of interest

Excluding RCTs from the NMA

1. For each of the NMAs, we obtained the probabilities of each treatment being the best when using the complete set of data reported by the authors of the SR. For example, if a network was composed of treatments A, B and C, and trials W, X, Y and Z, we used all four RCTs to run the NMA.

2. We run the NMA as many times as RCTs included in the complete set of data. Each of these times we excluded one of the RCTs from the analysis. We obtained the probabilities of each treatment being the best in each of these iterations. For instance, we obtained the probability of A being the best treatment when excluding trial W, trial X, Y and Z (four probabilities). In the same way, we obtained the probability of B, and C being the best treatment when excluding each of the trials. The following outcomes were used:

Mean in the absolute change of each treatment being the best

3. We calculated the difference between the probabilities of each treatment being the best in each of the iterations and the probability of each treatment being the best when using the complete set of data. This allowed us to obtain the change in the probability of each treatment being the best, when excluding one RCT from the analysis. For instance, if the probability of A being the best treatment using all four RCTs was 75%, and the probability of A being the best treatment when excluding trial W was 80%, the change in this probability was 5%. This change was calculated for all the treatment/RCTs combinations.

4. Since the probability of being the best treatment could increase or decrease, depending on which RCT was excluded from the analysis, and we were interested in the size of this change, we used the absolute change in the probability of being the best treatment. For example, if the best treatment of an NMA had a probability of being the best of 75%, and excluding one trial either increased this probability to 80% or decreased it to 70%, the absolute change was still 5%

5. We calculated the mean in the absolute change of the probability of each treatment being the best treatment, across RCTs excluded. For example, if excluding W resulted in a change of 5% in the probability of A being the best treatment, and excluding X, Y and Z resulted in changes of 3%, 7% and 6%, respectively, the mean absolute change of the probability of A being the best treatment was the average of these numbers, that is, 5.3%.

6. We calculated the mean across interventions to obtain the average of the mean absolute change, at the NMA level. For instance, if the mean absolute change of the probabilities of A, B and C being the best treatment was 5.3%, 7% and 9%, the overall change in the probabilities of each treatment being the best was 7.1%.

Range of the change of the probability of each treatment being the best

We recorded the range of the change of the probability of each treatment being the best, using the values obtained in (3). We present this value for the best treatment.

Proportion of iterations in which the best, second best, and either or both of them changed

From (1) and (2), we obtained the name of the treatment ranked first and second. We compared the treatments obtained in each iteration of (2) with the ones obtained when analyzing the complete set of data (1), and determined whether the best, second best, and both treatments together were the same or changed. Then we calculated the proportion of iterations in which:

a. the best treatment changed

b. the second best treatment changed

c. either the best, the second or both treatments changed

Increasing the thresholds to calculate the probabilities of being the best treatment

1. We established different decision thresholds to estimates the probabilities of each treatment being the best: OR= 1 (which indicates that both treatment effects are the same,

and any difference bigger than 0 makes a treatment better than the other) and OR= 1.1, 1.2, 1.3 and 1.4.

2. We reran each of the NMAs and recorded the best treatment when OR=1 (default threshold to calculate the probability of each treatment being the best) and its probability of being the treatment.

3. We calculated the probability of the best treatment being the best when using each of the thresholds (that is, better than all the other treatments in the network by a difference bigger than the threshold).

4. We calculated the probability of the best treatment being in the top two treatments when using each of the thresholds (that is, better than all the treatments of the network except for one other).

5. We plotted these thresholds against the probabilities of being the best treatment

Appendix 7: Code for performing the analyses

```
#Setting work directory, loading data and libraries
setwd("~/Dropbox/Onedrive files/PhD Thesis/Chapter 2/Analysis")
data<- read.csv("Primary data from included studies2.csv")

#Loading libraries and functions
library("gemtc")
library("rjags")

#Loading custom written functions
source("my.rank.probability.R")

#Setting a seed
set.seed(123)

#Setting the number of simulation iterations
k<- 1000

#Identifying how many NMAs there are and what are they called
papers<- as.character(unique(data$Paper))

#Creating vectors to store the results
meanchangebest<- vector(length = length(papers),mode="list")
names(meanchangebest)<-papers
meanchangesecbest<- meanchangebest
propchangerank<- meanchangebest
bestlist<- meanchangebest
secbestlist<- meanchangebest
bothbestlist<- meanchangebest
gdvalue<-meanchangebest
converged<- rep(TRUE,length(papers))
names(converged)<- papers
gdvaluedrop<- meanchangebest
hasconvergeddrop<- meanchangebest

#Setting up the number of iterations for the models
n.iters <- c(20,20,20,120,20,20,20,20,20,20,400,20,20,20)*k
names(n.iters) <- papers
iters.in.ranks <- min(n.iters)


#Setting the direction for treatment comparisons
dir<- c(-1,-1,-1,-1,-1,-1,1,-1,-1,-1,-1,-1,-1,-1)
names(dir)<- papers

dirfun<- c(T,T,T,T,T,T,F,T,T,T,T,T,T,T)
names(dirfun)<- papers

#getting the number of studies per paper to put in the graphs
n.treatments<- tapply(data$treatment, data$Paper, function(x) length(unique(x)))
names(n.treatments)<- papers


#Loop for each SR (variable called "Paper")

for (papername in papers){

  #Loading the data
```

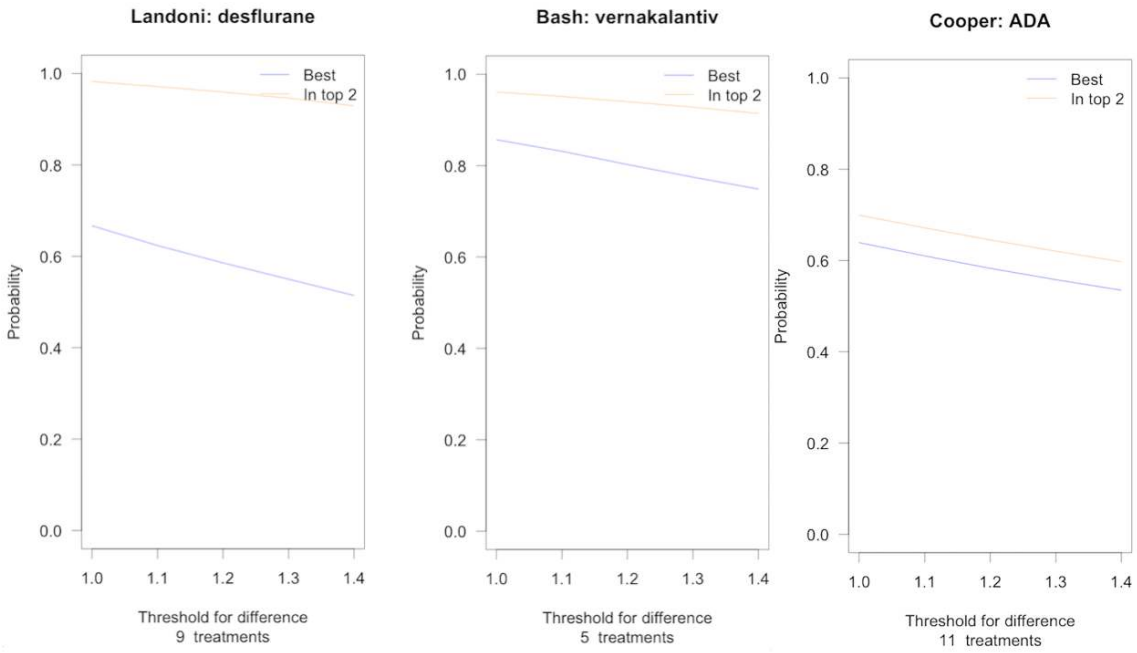
```
  d<- data[data$Paper==papername,]


  #Run NMA
  #First need to create the network.
  network<- mtc.network(d)

  #Setting up the model
  model.all<- mtc.model(network, type="consistency", factor=2.5, n.chain=3, link="logit")
  #Obtaining the samples
  sampled.values <- mtc.run(model.all, n.adapt=5000, n.iter=n.iters[papername])

  #Checking that the models have converged
  gd<- gelman.diag(sampled.values)
  gdvalue[[papername]]<-gd
  hasconverged<- all(gd$psrf[,2]<1.1)

  #Seeing whether all models have converged
  if(!hasconverged){
    converged[papername]<- FALSE
  }


  #Getting the probabilities of treatments being the best
  genrank<- my.rank.probability(sampled.values, preferredDirection = dir[papername])
  probbest.all<- genrank[,1]
  probsecbest.all<- genrank[,2]
  best2all<- names(sort(probbest.all,decreasing =T)[1:2])

#### INCREASING THRESHOLD FOR CALCULATING PROBABILITIES###

  #Finding the most common reference group and using that to
  #obtain relative effects

  parms <- model.all$monitors$enabled
  parms <- parms[-length(parms)]
  parms <- substring(parms,3)
  ref.groups <- substring(parms, 1,regexpr(",",parms,fixed=T)-1)
  #We always sort the treatments from most common to least common and choose   #the most common
one as the reference
  ref.tmt <- names(sort(table(ref.groups),decreasing = T))[1]


  #This obtains the log(ORs) between the other treatments and a single #reference.
  releff <- relative.effect(sampled.values,t1=ref.tmt)


  # need to specify the maximum number of iterations to use iters.in.ranks
  thresholds <- c(1, 1.1, 1.2, 1.3, 1.4)
  RanksWithThresholds <- rank.threshold(releff, threshold=thresholds,
  reference = ref.tmt, n.iter = iters.in.ranks, direction = dir[papername])


  #Pick the rankings matrix with OR=1: use this to define best and 2nd best
  NoThreshold <- RanksWithThresholds[[1]]

  #Best is the one with the highest probability in column 1 - take the name and store in
bestTreatment
  bestTreatment <- names(which.max(NoThreshold[,1]))
  #same for second best - after leaving out the best, find the largest remaining prob
  secondbestTreatment <- names(which.max(NoThreshold[row.names(NoThreshold)!=bestTreatment,1]))
```

```
  #Plots for best and second best
  par(mfrow=c(1,2))
  x<-do.call("rbind",RanksWithThresholds)
  tmts <- c(bestTreatment, secondbestTreatment)
  x <- x[is.element(row.names(x), tmts),]

  X1 <- x[rownames(x)==bestTreatment,]
  X2 <- x[rownames(x)==secondbestTreatment,]

plot(thresholds,X1[,1],type="l",ylim=0:1,col="blue",ylab="Probability",
       main=paste(papername,": ",bestTreatment,sep=""),yaxt="n",
               sub=paste(n.treatments[papername], " treatments"),
       xlab="Threshold for difference")
  axis(2,las=2)
  lines(thresholds,X1[,1]+X1[,2],col="orange")
  legend("topright",col=c("blue","orange"),legend=c("Best","In top 2"),
         lty=1, bty="n")

  plot(thresholds,X2[,1],type="l",ylim=0:1,col="blue",ylab="Probability",
       main=paste(papername,": ",secondbestTreatment,sep=""),
       #want to add the numbers of treatments here, based on the var n.treatments
       sub=paste(n.treatments[papername], "treatments"),
       yaxt="n",
       xlab="Threshold for difference")
  axis(2,las=2)
  lines(thresholds,X2[,1]+X2[,2],col="orange")
  legend("topright",col=c("blue","orange"),legend=c("Best","In top 2"),
         lty=1, bty="n")




  ##### DROPPING TRIALS #####


  if(rundrop){
    studies<- as.character(unique(d$study))

  #matrices to store the probabilities
  tmt<- sort(unique(d$treatment))
  probbestdrop<- matrix(ncol=length(studies), nrow=length(tmt))
  dimnames(probbestdrop)<- list(tmt,studies)
  probsecbestdrop<- probbestdrop
  dimnames(probsecbestdrop)<- dimnames(probbestdrop)
  #matrix to store the rankings
  #rankdrop<- matrix(ncol=length(studies),nrow=length(tmt))
  #dimnames(rankdrop)<-dimnames(probbestdrop)
  best2drop<- matrix('',ncol=2, nrow=length(studies))
  dimnames(best2drop)<- list(studies,c("best","secondbest"))
  gdvaluedrop[[papername]]<- vector(length = length(studies),mode="list")
  names(gdvaluedrop[[papername]])<- studies
  hasconvergeddrop[[papername]]<- rep(TRUE,length(studies))
  names(hasconvergeddrop[[papername]])<- studies


#Loop for each trial
for(dropstudy in studies){
      # Drop the trial and run NMA
      chdata<- d$study!=dropstudy
      d.drop<-d[chdata,]
      network<- mtc.network(d.drop)
      #plot(network, main=paste(papername,"drop",dropstudy))
```

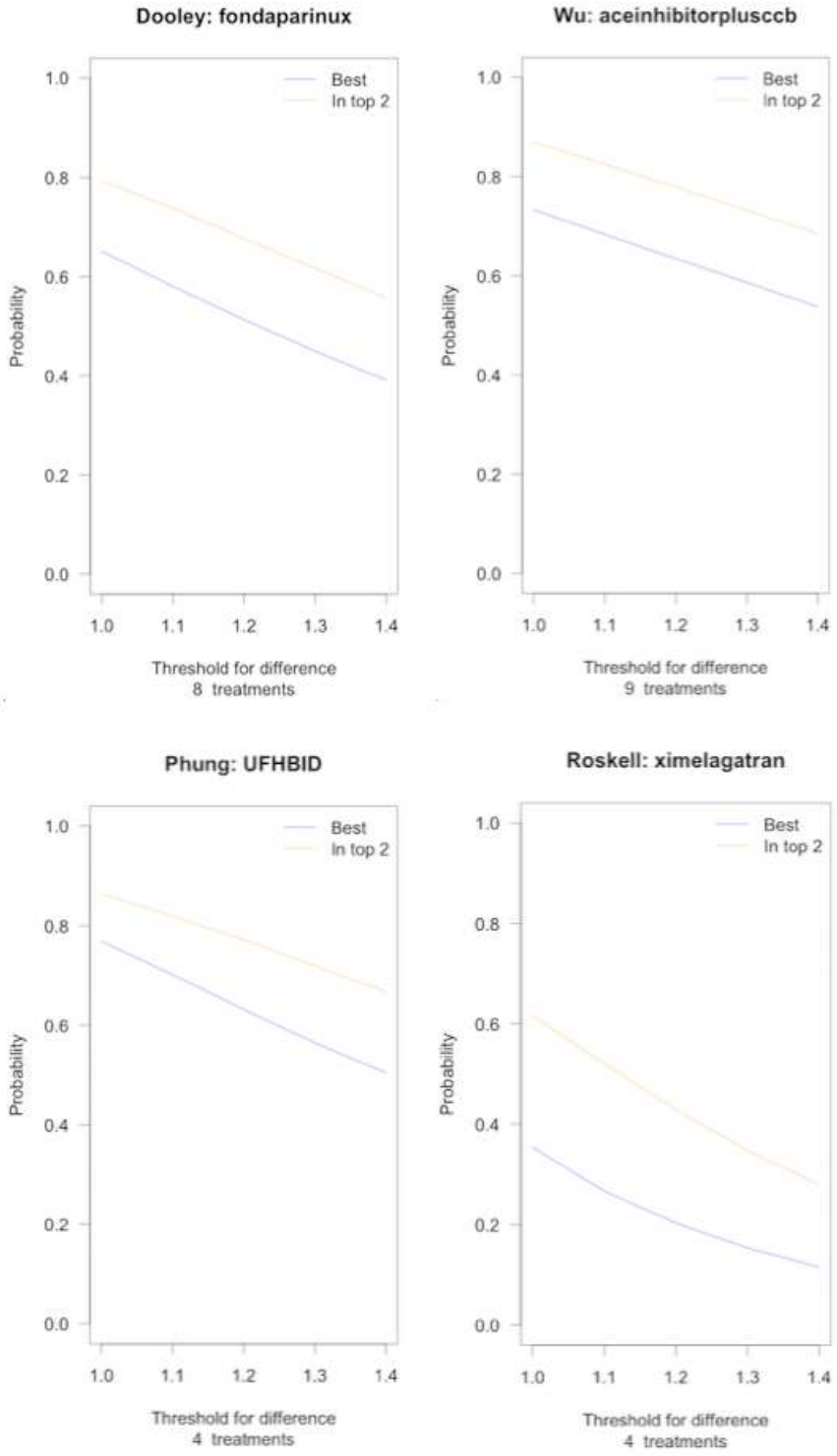
```
      model<- mtc.model(network, type="consistency", factor=2.5, n.chain=3, link="logit")
      result<- mtc.run(model, n.adapt=5000, n.iter=n.iters[papername])

      #Check they converged
      gd<- gelman.diag(result)
      gdvaluedrop[[papername]][[dropstudy]]<-gd
      convergeddrop<- all(gd$psrf[,2]<1.1)


      #Seeing whether all models have converged
      if(!convergeddrop){
        hasconvergeddrop[[papername]][dropstudy]<- FALSE
      }


      # Obtain and save p of being best and second best per tmt when MID is 0
      # Obtain and save a ranking per number per treatment
      #ranking_j<- c(NA, number of studies in data)
      prob<- my.rank.probability(result, preferredDirection=dir[papername])
      pbest<-prob[,1]
      psecbest<-prob[,2]
      probbestdrop[names(pbest),dropstudy]<-pbest
      probsecbestdrop[names(psecbest),dropstudy]<-psecbest
      #ranking<-rank(-pbest)
      #rankdrop[names(ranking),dropstudy]<-ranking
      best2drop[dropstudy,]<- names(sort(pbest,decreasing =T)[1:2])
    }

#Calculating the change in prob of best, second best and ranking per study dropped
changebestdrop<- probbestdrop-probbest.all
meanchangebestdrop<- rowMeans(abs(changebestdrop), na.rm = T)
rangechangebestdrop<- t(apply(changebestdrop,1,range,na.rm=T))

changesecbestdrop<-probsecbestdrop-probbest.all
meanchangesecbestdrop<- rowMeans(abs(changesecbestdrop),na.rm = T)
rangechangesecbestdrop<- t(apply(changesecbestdrop,1,range,na.rm=T))


#Proportion of droppings in which the first stays the same, the second and both
bestlist[[papername]]<- prop.table(table(best2all[1]==best2drop[,'best']))
secbestlist[[papername]]<-prop.table(table(best2all[2]==best2drop[,'secondbest']))
bothbestlist[[papername]]<-prop.table(table(best2all[1]==best2drop[,'best'] &
                                            best2all[2]==best2drop[,'secondbest']))

#summary for the SR
meanchangebest[[papername]]<-cbind(meanchangebestdrop,rangechangebestdrop)
meanchangesecbest[[papername]]<-cbind(meanchangebestdrop,rangechangesecbestdrop)
#propchangerank[[papername]]<- changerank
}
}
```

Appendix 8: Plots of probabilities of best treatments being the best when changing the decision thresholds.

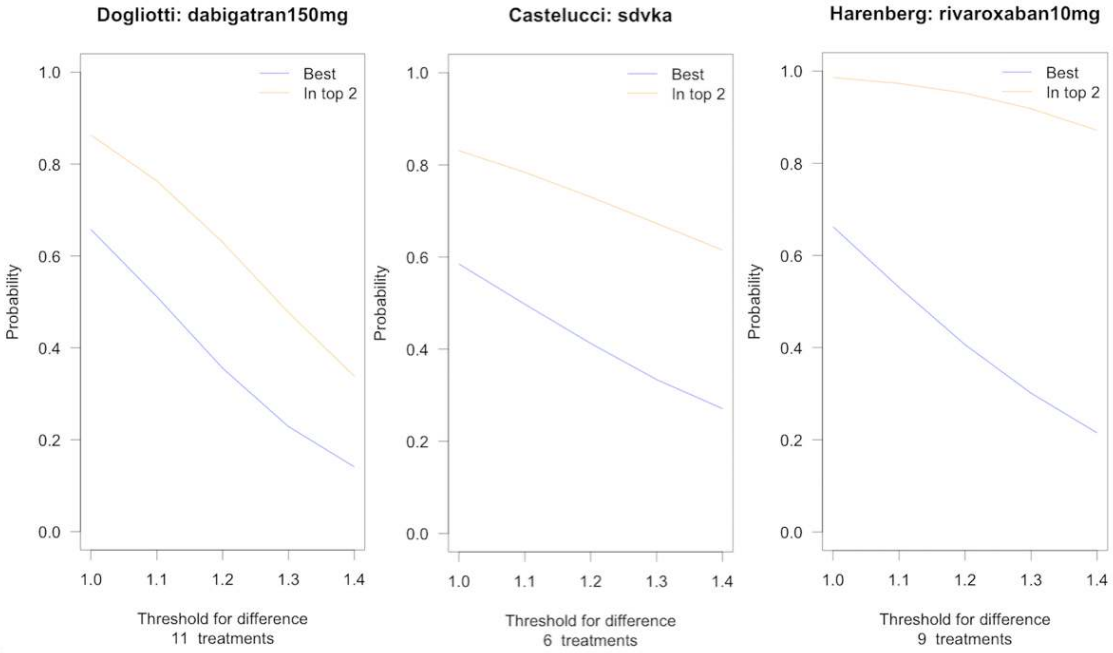
NMAs in which the change was small



NMAs in which the change was constant and moderate



NMAs in which the change was constant and large



NMAs with a rapid decrease