THRESHOLD ANALYSIS IN NICE GUIDELINE DEVELOPMENT

David M. Phillippo, Sofia Dias, Nicky J. Welton, A. E. Ades
NICE Clinical Guidelines Technical Support Unit

March 24 2016

EXECUTIVE SUMMARY

This report addresses the use of threshold analysis in NICE guideline development for the assessment of sensitivity of treatment recommendations to bias adjustments. Threshold analysis is focussed around the derivation of *bias-adjustment thresholds*, which describe the smallest changes to the data that result in a change of treatment decision, and *decision-invariant bias-adjustment intervals*, within which changes to the data do not affect the decision. Using these, guideline developers can discern which evidence the treatment recommendation is sensitive to, and which evidence makes little difference.

We describe a new method for threshold analysis, based on the Bayesian joint posterior as arising from a Network Meta-Analysis (NMA), which is both highly flexible and easy to implement with the R functions provided in the technical appendix. Analysis of the effects of potential bias adjustment may be considered either to individual study estimates or to overall treatment contrasts. The method is illustrated with several examples, which demonstrate the results and implications of threshold analysis. In most cases the treatment recommendation was robust to plausible levels of bias in all but a small proportion of contrasts or studies. In larger, well connected networks with large numbers of trials, recommendations were robust against almost any plausible bias adjustments. We also present more complex applications of threshold analysis, where for example biases may be considered in groups of studies or groups of treatments.

The threshold method may be extended beyond decisions based on treatment efficacy, to decisions based on net benefit as the result of a probabilistic Cost-Effectiveness Analysis (CEA). We demonstrate the application to simple cases where the net benefit function is linear in the treatment parameters, and where log odds ratio treatment effects are related via an inverse logit transform to a net benefit function linear in probability. In the simple linear net benefit scenario, thresholds are derived in the same manner as for efficacy, as the net benefit posterior is multivariate normal; for net benefits involving an inverse logit transform, the posterior is logistic-normal and must be evaluated using numerical integration, leading to a numerical solution for the thresholds using standard root-finding tools. A general numerical solution for CEAs with complex, non-linear, or even unknown net benefit functions is described; firstly the net benefit function is emulated, allowing the posterior expectations to be evaluated efficiently using numerical integration, and then thresholds are found using numerical root-finding. This remains an area of active research.

In order to interpret the results of a threshold analysis, judgements must be made about the magnitude and direction of plausible biases, and whether such biases are present in each piece of evidence. We address the former question, summarising the literature on internal and external biases, including a meta-epidemiological study of the size of internal bias. In any case, judgements on the presence, magnitude, and direction of bias in any evidence must be clearly documented and backed up by reasoned argument or evidence.

We describe the role of threshold analysis in guideline development, and suggest possible uses. Threshold analysis may be used either reactively or pre-emptively, answering concerns raised about specific biases or studies in a retrospective manner as they arise, or to pre-empt such criticism ahead of time. Phrasing of the recommendation may be guided by the threshold analysis, either to strengthen recommendations which are seen to be robust, or to provide basis for necessary restraint in the light of the recommendation being sensitive to plausible bias. We show how threshold analysis may be used to reduce the number of laborious qualitative GRADE analyses required, suggesting that in-depth quality assessments might only need to be performed for evidence that the recommendation is shown to be sensitive to. In practice, plausible biases in only a small number of studies out of the total number included could have any effect on the treatment decision, leading to substantial time savings from the reduced number of GRADE analyses required.

In situations where the treatment recommendation is shown to be sensitive to plausible bias which the Guideline Development Group have reasonable grounds to believe is present, the recommended course of action is to perform an updated NMA which models and adjusts for any such bias. Treatment recommendations should then be made based on the updated NMA. We strongly suggest that the recommendation is not changed based on the threshold analysis alone, not least because this amounts to crudely changing the original data. In either case, any judgements and decisions should be clearly documented for the sake of transparency.

Threshold analysis is a powerful tool, which guideline developers may use to assess the robustness of treatment recommendations to bias adjustments. It should strengthen decisions that are shown to be robust to plausible bias, and provide proper cause for restraint where decisions are shown to be sensitive. Threshold analysis can both direct critical focus onto evidence that the recommendation is sensitive to, and allay concerns over biases in evidence that is non-influential. The flexibility of threshold analysis allows application to a wide range of scenarios of varying complexity, including decisions based on either efficacy or net benefit.

1. INTRODUCTION

Since Network Meta-Analysis (NMA) was first introduced as a method for routine application to multiple treatment decisions, doubts have been raised about its reliability and its assumptions. These have included a range of critiques. Among them there are empirical studies of the consistency of direct and indirect evidence [4-6]; claims that NMA was “mystifying” [7]; claims that indirect evidence was “observational” [8, 9]; a variety of statements about its assumptions [10-12]; and a wide range of papers suggesting that there were problems with the method, or that more research was needed [13, 14].

Taken together, this literature appears to present a formidable challenge to the credibility of NMA. Indeed, in spite of the many contradictions, inaccuracies, and lack of mathematical development in this body of work, it succeeds in focusing attention on one particular issue of over-riding importance: which is that the interpretation of an NMA analysis, like pair-wise meta-analysis, relies on there being no major effect modifiers in the included trials. A further difficulty is that this assumption is hard, if not impossible, to verify in practice.

In this report we examine the credibility of NMA in another way, by taking an *ex post facto* position. Given the results of an NMA, how robust is our treatment recommendation to plausible biases in the data?

The validity of conclusions drawn from network meta-analysis (NMA), like all statistical analyses, depends on whether the input data meet the assumptions required by the model. In the specific case of fixed effects NMA it has been shown that, just as with a pair-wise meta-analysis, the final estimate of any treatment effect is no more than a weighted average of all the study-level estimates [15]. Specifically, the FE NMA estimate of the effect of treatment *b* relative to treatment *a* is a sum of products of coefficients $\beta_{ab,XY}$ and the observed treatment effects D_{XY} in the original trials, summed over all the trials in the evidence base.

$$\hat{d}_{ab} = \sum_{XY} \beta_{ab,XY} D_{XY} \quad (1)$$

Equation (1) also holds approximately for random effects models. This is a key result because it means that if the evidence inputs are unbiased, meaning that they are unbiased estimates of the treatment effects in the target population, then the NMA estimates are also unbiased for that target population. These same coefficients also allow us to examine the influence that potential bias in any data element might have on any specific treatment effect. The matrix of coefficients $\beta_{ab,XY}$ has been called the *contributions matrix* [16, 17], and has been used to explore flow of evidence in a network of evidence as well as inconsistency.

Nevertheless, while the contribution matrix is a powerful tool in the analysis of influence of each data element on any treatment effect estimate, as we will see further below, it does not by itself tell us how the assess the "quality" of an NMA, or the credibility of conclusions that are based on it.

We review the work leading up to this project below, and then outline the rationale and scope of the present report.

1.1 PREVIOUS RESEARCH LEADING UP TO THE WORK IN THIS REPORT

1.1.1 GRADE-BASED APPROACHES TO QUALITY OF NMA

Two proposals aim to assess the credibility of an NMA by starting from a GRADE analysis of the quality of evidence attaching to each of the direct comparisons in the NMA, which we will refer to as "GRADE NMA" [18] and "CMI NMA" [14]. Briefly, a GRADE assessment rates the quality of evidence informing a pairwise meta-analysis as high, moderate, low or very low [19] across five domains - study limitations, imprecision, indirectness, inconsistency (in this context meaning heterogeneity), and publication bias. Evidence from randomised controlled trials starts as high

confidence and can be downgraded by a maximum of two levels per domain. A summative judgement of quality is formed across all five domains [20] and interpreted as shown in Table 1.

The GRADE NMA approach does not actually deliver an assessment of the credibility of the conclusions from the NMA, but instead delivers a set of unrelated assessments of the reliability of individual contrasts, based on a GRADE assessment of the direct and indirect evidence on each one. Because the GRADE NMA process comes to conclusions about the most reliable estimate of the A vs B effect without reference to the AC or BC effects, it is capable of reaching a set of conclusions that are internally incoherent.

The CMI NMA method delivers both an assessment of the reliability of the individual contrasts, and an assessment of the reliability of the posterior ranking of treatments generated by the entire network of evidence. The methods are statistically well-founded, and make extensive use of the contributions matrix, but they are complex and time-consuming to apply. Most importantly, although the CMI NMA method delivers an assessment of the reliability of the treatment rankings obtained from a NMA, this does not tell the decision maker whether the treatment decision derived from an NMA can be relied on, nor what the alternative decision might be.

1.1.2 THRESHOLD ANALYSIS BASED ON 2-STAGE NMA

Rather than assess the reliability of conclusions based on NMA by examining qualitative assessments of the "quality" of the body of evidence or parts of it, we can instead ask the question: "would bias in any of the data elements change the treatment recommendation based on an NMA?" This question can be formulated as a threshold analysis, which is a standard form of sensitivity analysis used in decision analysis or cost-effectiveness analysis. The decision maker considers a set of options S according to an objective function $F(S,\boldsymbol{\theta})$ of parameters $\boldsymbol{\theta}$ that includes the relative effects of the treatments estimated by the NMA. The decision maker then chooses the treatment $S*$ with the highest expected value on the objective function:

$$S* = \text{argmax}_S \, \mathbb{E}_{\boldsymbol{\theta}} \left[F\left(S,\boldsymbol{\theta}\right) \right] \tag{2}$$

where the expectation is taken over the joint distribution of the parameters. Various objective functions can be considered, including net benefit [21], which is used in NICE technical appraisals and clinical guidelines. Multi-criteria decision analysis (MCDA) [22] is a further possibility. However, the large majority of clinical guidelines, including those issued by US Colleges of medicine and by governmental bodies in jurisdictions outside the UK are based on efficacy alone.

Our initial work on this topic [23, 24] was based on a two-stage NMA [15]. This starts with the individual trial data as input to Stage 1. The output of the first stage is the set of pooled summaries of each of the pair-wise contrasts on which there is direct evidence. The second stage takes the output of the first stage as input and then produces an NMA analysis in which the consistency relationships are enforced. The reason for using this method was so that we could carry out an NMA analysis starting from the exact same data inputs as a GRADE NMA analysis. The threshold analysis would then proceed by going through each input data item (a pair-wise pooled summary) in turn, and adding

or subtracting a small amount in increments and then re-running the second stage of the two-stage NMA until the treatment recommendation changed. Caldwell et al. [24] implemented the two-stage method in WinBUGS, and, although it runs very quickly, it is a computation-intensive numerical solution.

Besides being cumbersome in the way it was implemented, the disadvantage of this approach is that the base-case treatment recommendation based on a two-stage analysis of this sort may not always be the same as the base-case recommendation that comes out of a one-stage NMA. There are two reasons for this: first the pooled pair-wise summaries are estimated separately, and may therefore have highly unstable between-study variances. Indeed, it has been shown that under the assumptions of an NMA, there are powerful constraints on the set of between-trials variance for the set of contrast in the network [25]. These cannot be realized if they are estimated independently. Contrasts informed by a single trial, in particular, will necessarily yield fixed effect estimates and have far less variance than is reasonable if other contrasts were estimated with random effects. A second reason is that the covariance introduced by multi-arm trials is not reflected. We have shown, however [24], that, if a two stage analysis is performed in ways that avoid these difficulties, it produces results that are almost identical to one-stage.

1.1.3 THRESHOLD ANALYSIS BASED ON THE BAYESIAN POSTERIOR DISTRIBUTION OF TREATMENT EFFECTS

However, a far better solution would be to work the threshold model from the Bayesian posterior distribution of treatment effects. This guarantees that the threshold analysis is based upon an identical foundation to the base-case treatment recommendation. Another reason for starting from the posterior, rather than from the data inputs is that in practical NMA applications, especially in the context of clinical guideline development at NICE, the NMA may incorporate a very complex mixture of data types. For example, in the Social Anxiety NMA [3, 26] the final NMA incorporated a mixture of response and recovery data, linked together by a regression model, as well as data in the form of odds ratios and data in the form of continuous scales. At the same time the NMA model itself may also incorporate special structures, such as class effects. The two-stage NMA approach adopted in our early work was adequate to illustrate the proof of principle, but it lacks the flexibility that would allow us to extend it to include multiple types of data input and additional model structures.

This report focuses on a method that starts from the joint posterior distribution of treatment effects generated by a Bayesian one-stage NMA. In practice, we characterize the input to the process as a multivariate normal distribution of the relative treatment effects, which can be recovered from a Bayesian NMA in WinBUGS by outputting the summary statistics for the relative treatment effects: that is the posterior means and the posterior standard deviations, together with the posterior correlations. The method is essentially based on a quantity we call the influence matrix, because its elements describe the influence each data point has on each parameter. This matrix is closely related to another quantity from classical statistics, known as the hat matrix, which describes the influence of each data point on each fitted value; the hat matrix may be derived from the influence matrix by pre-multiplication with a design matrix. What marks the method discussed here as different from standard applications of the hat matrix in classical statistics is the way it is recovered from the Bayesian posterior [27, 28].

Although the mathematical and computational basis for the threshold analysis based on a Bayesian posterior is fundamentally different from the two-stage analysis, both analyses ask exactly the same question "How much would the data input have to change before the treatment decision was changed?" and both give answers in exactly the same form.

1.2 THE SCOPE OF THIS REPORT

The original intent of this work was to:

1. Apply threshold analysis to a set of examples from completed NICE Clinical Guidelines, both at the trial-level and at the aggregate level
2. To further develop the method so that it could be applied to cost-effectiveness analyses as well as clinical efficacy analysis
3. To develop a set of questions that could be used to assess whether members of guideline development groups (GDGs) understood thresholds analyses and whether they found them useful in guideline development.

As the work progressed we were able to meet with the NMA Working group and participate in the Technical meetings. These discussions helped us steer the project in whatever direction seemed most likely to be productive. As a result of these discussions, and following further work on extension to CEAs, the objectives of the project were modified to:

1. Apply threshold analysis to a set of examples from completed NICE Clinical Guidelines, both at the trial-level and at the aggregate level
2. Develop threshold analyses that can be performed on cost-effectiveness analyses, initially with linear net benefit functions and then with increasingly complex or even unknown models, and illustrate with examples
3. Develop guidance on the circumstances under which the base-case recommendation can be changed as a result of threshold analyses
4. Develop a set of options on how threshold analyses might be used in guideline development.

At the time of writing the report we have not developed a consensus on how threshold analysis should be introduced into the guideline development process, so it would be premature to instigate a piloting exercise. However, we understand that the National Clinical Guideline Centre will be piloting threshold analyses into the guideline on deep vein thrombosis later this year.

1.3 OUTLINE OF THIS REPORT

In section 2 we set out some illustrative examples. These are designed to highlight the basic ideas and properties of threshold analysis, in both trial-level and contrast-level analyses. The exposition is designed to minimize the mathematics behind the method, although some of the key concepts are described in an appendix to this report. A paper covering the technical aspects of the method in detail has been submitted [27]. It will become apparent that the

way in which threshold analysis is used in practice, and the way in which it might impact on guideline development will be heavily context dependent.

In the following section (3) we illustrate some more complex threshold analyses, to demonstrate the flexibility of the method in addressing a wide range of practical concerns that could be raised in guideline development. Using the Social Anxiety guideline as an example, we begin by exploring the impact of adjusting for a generic bias in (a) all trials of drug treatments vs placebo, or (b) in all trials of psychological treatments vs waitlist or similar controls. We then look at a 2-dimensional threshold plane for both these kinds of bias. In a third analysis, we explore the impact of a generic bias in favour of the active treatment in all the trials in which a specific investigator or sponsor was involved.

Section 4 examines examples of threshold analysis applied to cost-effectiveness analysis, i.e. with net benefit as the objective function in equation (2) rather than treatment efficacy, on the scale of the linear predictor in the NMA model. Both trial level and contrast level thresholds are explored. The algebraic basis for threshold analysis as applied to clinical efficacy starts to break down in models where the net benefit is not linear in the efficacy parameters. However, we have made some important progress towards extending threshold analysis to increasingly complex net benefit functions, although this work is still in progress. In section 4 we illustrate with two examples: one where net benefit is linear in the efficacy parameters, and a second where it is linear in the inverse logit of the efficacy parameters, which is a particularly common configuration. We conclude section 4 by outlining a potentially fully general solution which has not yet been implemented, and by describing a computationally intensive method that could be adopted currently.

In section 5 we look at how large biases in the data can be expected to be. Part of this section is a summary of the meta-epidemiological findings on size of bias in relation to type of outcome (subjective / objective) and indicators of risk of bias, such as lack of allocation concealment and lack of blinding.

Section 6 sets out some options on how threshold analyses might be used in guideline development. We also make some suggestions about the circumstances under which a base-case recommendation could be overturned, and on what should be reported in guidelines where threshold analyses have been performed. Section 7 summarises what we believe are the current priorities for further research.

The mathematical details behind these methods [27, 28] are available on request. We provide a technical appendix that explains:

1. How to output the posterior summaries and correlations from WinBUGS
2. A set of R routines that perform the threshold analyses
3. A “how-to” guide on running theses routines to obtain the appropriate analysis with associated statistics and graphical aids.

2. ILLUSTRATIVE EXAMPLES: DECISIONS BASED ON CLINICAL EFFICACY ALONE

In this section, we illustrate threshold analysis based on the Bayesian posterior, for decisions based on clinical efficacy using four examples, which cover a varied range of use cases: fixed and random effects models, class effects, and both small and large networks of treatments. We begin with a simple fixed effects NMA of prophylactic treatments for headaches and migraines in over 12s, from the clinical guideline CG150.1 [29], with which we explain in detail the method and results of a threshold analysis. Three further examples are then presented, along with a discussion of the results and their implications on decision making.

2.1 THRESHOLD ANALYSIS EXPLAINED – HEADACHES EXAMPLE

We illustrate the method of threshold analysis using the headaches clinical guideline CG150.1 [29]. Eight prophylactic treatment regimens for primary headache were compared in a network meta-analysis of 11 studies. The network diagram showing how treatments are connected by study evidence is shown in Figure 1.

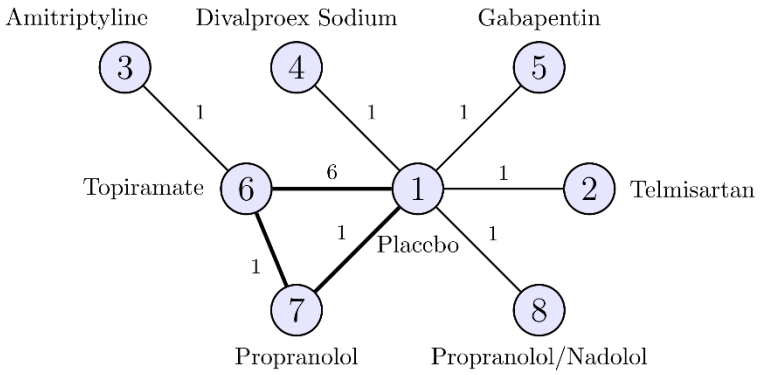


Figure 1: Treatment network for headaches example. Edges indicate study evidence between two treatments, and numbers on the edges show the number of studies making the comparison. Numbers inside the nodes are the treatment codings. The bold loop is formed by a single three-arm study.

The starting point for a threshold analysis is the Bayesian joint posterior arising from a network meta-analysis. This posterior may be briefly summarised by a table giving the efficacy of each comparator relative to the reference treatment (the basic treatment parameters) – which in this case is placebo (treatment 1). For the headaches NMA these treatment effect summaries are shown in Table 1, where we see that propranolol (treatment 7) is the recommended treatment based on maximum efficacy. We refer to the recommended treatment under the original analysis with no bias adjustment as the *base-case recommended treatment*.

Table 1: Treatment effect estimates vs. placebo, along with 95% credible intervals, from the original NMA. Based on efficacy alone propranolol (treatment 7) is base-case recommended treatment, having the greatest reduction in headache days per month (–1.19), followed by amitriptyline (3) and then topiramate (6).

Treatment		**Mean change in headache days per month compared to placebo (95% Credible Interval)**	
2	Telmisartan	–0.52	(–2.32, 1.27)
3	Amitriptyline	–1.14	(–2.45, 0.16)
4	Divalproex Sodium	0.13	(–0.99, 1.23)
5	Gabapentin	0.00	(–1.60, 1.58)
6	Topiramate	–1.04	(–1.52, –0.58)
7	Propranolol	–1.19	(–2.20, –0.20)
8	Propranolol/Nadolol	–0.60	(–1.65, 0.45)

From here we proceed to mathematically derive *bias-adjustment thresholds* – either for each individual study estimate, or for each aggregate treatment contrast. These thresholds describe exactly how much the evidence can change before the treatment recommendation changes.

To illustrate how the bias-adjustment thresholds are calculated, consider the headaches NMA with 8 treatments: one treatment has the greatest efficacy (propranolol), and there are 7 other treatments which are less efficacious. For a given data point (study estimate or aggregate contrast), we can attempt to make each of the 7 other treatments in turn have the greatest efficacy by adding or subtracting from the data point until another treatment has the greatest efficacy. Thus, for a given data point we obtain 7 values (positive or negative) which correspond to the smallest changes required to make the each of the remaining treatments have the highest efficacy. The smallest positive and negative changes out of these 7 values are the positive and negative bias thresholds respectively for the data point. Intuitively we can derive these threshold values by dividing the difference in efficacy between the base-case recommended treatment and each of the 7 possible new recommended treatments by the amount of influence the data point has on each difference, and then taking the smallest positive and negative values as the thresholds. In practice this is achieved directly and efficiently using a mathematical formula, without the need for any numerical methods (see [27, 28] for details).

The negative and positive bias thresholds for a study or contrast data point are then added to the value of the data point to create a *decision-invariant bias interval*, within which the data point can lie without changing the treatment recommendation (i.e. the recommended treatment is still the same). If the data point is changed beyond the limits of the invariant interval, the treatment recommendation will change.

CONTRAST LEVEL ANALYSIS

We perform a threshold analysis to determine how the quality of the combined body of evidence on each contrast might affect the treatment recommendation. The results of the threshold analysis are shown in the forest plot in Figure 2, where each row gives the results for a treatment contrast; we only include contrasts for which there is direct study evidence. We show the posterior mean for each relative effect, its credible interval, and the *decision-invariant bias interval* (shaded line). This is the region within which any change in the combined evidence on a contrast will not change the treatment recommendation. Beyond the invariant thresholds at each side of the interval a change in treatment recommendation would occur; the new recommendation treatment codes are indicated in Figure 2 at either side of the interval in column 4, or show "–" where there is no threshold in that direction. Larger invariant intervals therefore indicate that the recommendation is more robust to changes in the evidence on a contrast; smaller invariant intervals indicate that the recommendation is sensitive to the quality of evidence on a contrast. Contrasts where an invariant threshold lies within the 95% credible interval of the posterior mean are shown in bold, since the treatment recommendation is sensitive to the imprecision on these contrasts. In this case, all but the 4 vs. 1 contrast (divalproex sodium vs. placebo) have bias thresholds which lie within the 95% CrI.

Contrast	Posterior Mean	95% Credible Interval	Bias Invariant Interval with $\tilde{k}^*$		
2 vs. 1	–0.51	(–2.32, 1.27)	2	(–1.19, NT)	–
4 vs. 1	0.13	(–0.99, 1.23)	4	(–1.19, NT)	–
5 vs. 1	–0.00	(–1.60, 1.58)	5	(–1.18, NT)	–
†6 vs. 1	–1.04	(–1.51, –0.58)	3	(–1.20, –0.09)	8
†7 vs. 1	–1.19	(–2.20, –0.20)	–	(NT, –1.04)	3
8 vs. 1	–0.60	(–1.64, 0.45)	8	(–1.19, NT)	–
†6 vs. 3	0.10	(–1.13, 1.32)	–	(NT, 0.15)	3
†7 vs. 6	–0.15	(–1.11, 0.82)	–	(NT, –0.07)	3

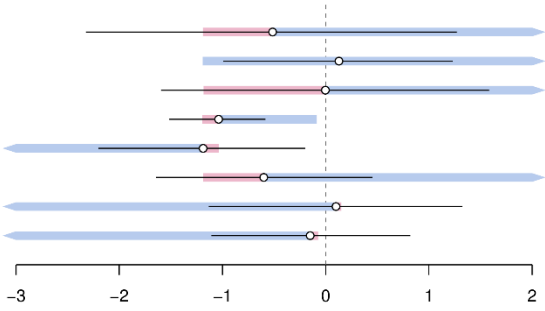


Figure 2: Forest plot showing the results of the contrast-level threshold analysis. A bias adjustment to the combined evidence on a contrast outside of the invariant interval (shaded) will result in a new treatment recommendation (based on the greatest reduction in headache days per month); the new recommended treatment code is shown at either end of the interval in column 4. Bold text indicates contrasts where bias thresholds lie inside the 95% CrI. The base-case recommended treatment is 7 (propranolol).

† indicates contrasts with bias-adjustment thresholds less than the minimally important difference (0.5 days).

NT = no threshold; no amount of bias adjustment in this direction will change the treatment recommendation.

Let us examine more closely the first row of Figure 2, which relates to the contrast of treatment 2 (telmistartan) vs. placebo. The threshold analysis for this contrast showed that no amount of bias adjustment in the positive direction could change the treatment recommendation, which is indicated by "NT". However, a negative bias threshold does exist: a bias-adjustment of –0.674 headache days per month would result in telmisartan (2) being recommended. To derive the (one sided) invariant interval from the bias thresholds we then simply add the thresholds to the contrast estimate, resulting in the bias invariant interval shown in column 4 of the table: $-1.19 = -0.52 - 0.67$ for the lower end of the invariant interval, and "NT" for the upper. The new treatment recommendations at either end of the bias invariant interval are shown to the left and right of the interval in column 4. The plot to the right of the table shows

the invariant interval as the shaded region, and the distance between the contrast estimate and the lower end of the shaded invariant interval is the negative bias-adjustment threshold. Note that the lower portion of the invariant interval is shaded red and the contrast label in column 1 is bold: the lower end of the invariant interval lies inside the lower end of the 95% CrI (shown as a thin black line), meaning that the treatment recommendation is sensitive to the imprecision of the evidence on this contrast. The remaining rows of the figure are interpreted in the same fashion.

Of particular note are bias thresholds which are also less than the minimally important difference of 0.5 days (as defined in the original guideline). Contrasts with such bias thresholds are identified in Figure 2 by a dagger (†): topiramate vs. placebo (6 vs. 1), propranolol vs. placebo (7 vs. 1), topiramate vs. amitriptyline (6 vs. 3), and propranolol vs. topiramate (7 vs. 6). For each of these contrasts, it is possible for a change in the combined evidence which would be considered clinically negligible to change the treatment recommendation – in all cases the new treatment recommendation at the threshold is amitriptyline (3). For the topiramate vs. amitriptyline contrast, a positive bias of just 0.05 in favour of amitriptyline would be enough to change the treatment recommendation – equivalent to an extra hour and a quarter of headache per month; for the propranolol vs. topiramate contrast, a positive bias of just 0.08 in favour of topiramate would change the treatment recommendation – an extra hour and fifty minutes of headache per month.

The threshold analysis gives insight into the robustness of the treatment recommendation to bias, and shows how the treatment recommendation may change. In this case, the treatment recommendation is sensitive to the level of imprecision in all but one contrast, and in four contrasts bias adjustments that might be considered clinically negligible could lead to a change in the treatment recommendation. As such, the evidence supporting each treatment comparison should be carefully assessed for potential bias; in particular those studies comparing topiramate to placebo, propranolol to placebo, topiramate to amitriptyline, and propranolol to topiramate, as these contrasts are highly sensitive to very small biases which would result in the base-case recommendation of propranolol changing to a new recommendation of amitriptyline.

STUDY LEVEL ANALYSIS

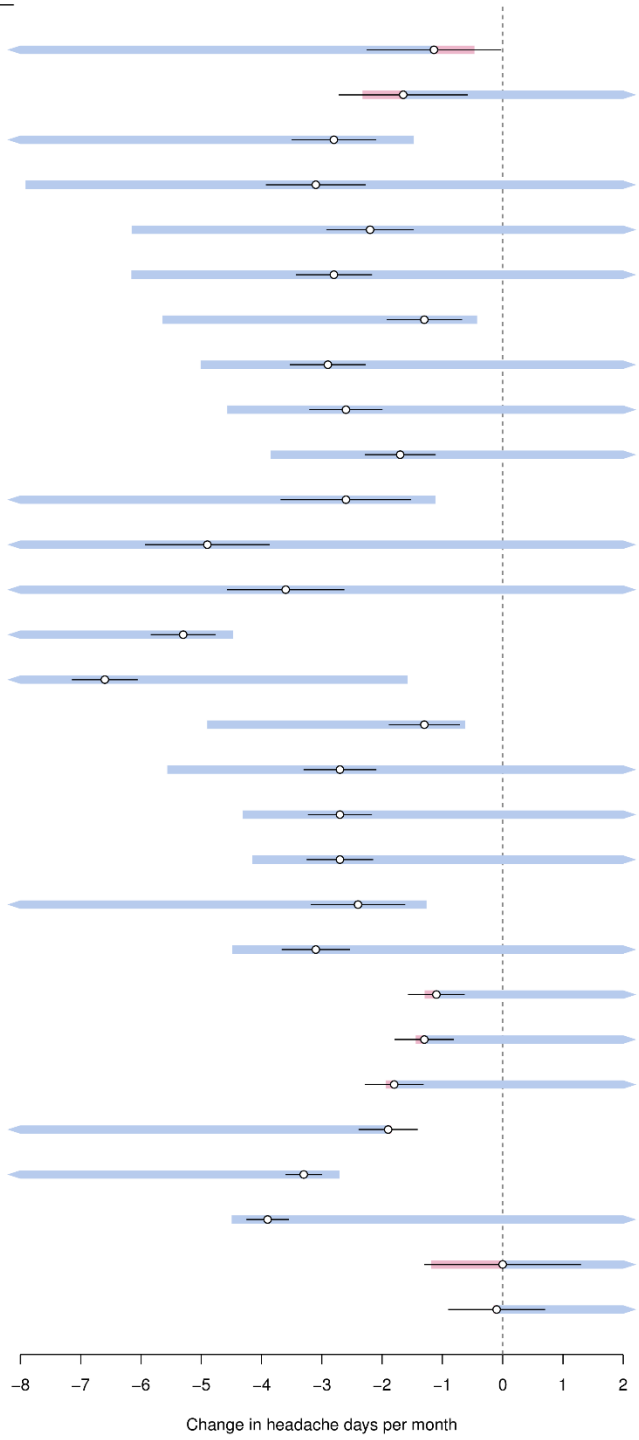
To examine the sensitivity of recommendations to the quality of evidence in more detail, we perform a threshold analysis at the study level; the effects of bias adjustment are considered on each study estimate – either absolute treatment effects or relative differences, depending on which the study reported – instead of on aggregate bodies of evidence. This allows us to determine which studies in particular are critical to the overall bias sensitivity of the treatment decision. The results of the study-level threshold analysis are presented in Figure 3. The interpretation of the forest plot is as in the contrast-level analysis, with one exception: the thin black lines about the study estimates are now 95% confidence intervals as reported by the studies (in the contrast-level analysis these indicate 95% credible intervals for the aggregate estimates from the Bayesian joint posterior).

The results show clearly that sensitivity of the treatment recommendation to plausible levels of bias is an issue in only a small number of studies: the treatment recommendation is sensitive to the level of imprecision in 4 out of 11 studies, which are indicated with bold labels in Figure 3. In two of these studies (Diener 2004 and Dodick 2009), bias

adjustments that would be considered clinically negligible (i.e. less than the minimally important difference of 0.5 days) could result in a change of treatment recommendation – in both cases to amitriptyline (treatment 3). The treatment recommendation is sensitive to adjustment in any arm of Diener 2004, the smallest threshold being a positive bias of 0.05 headache days per month in the propranolol (7) arm, equivalent to 1 hour 15 minutes more headache per month; the recommendation is also sensitive to the contrast reported in Dodick 2009, where the negative bias threshold is –0.05 headache days per month in the relative effect of amitriptyline vs. topiramate (3 vs. 6), equivalent to 1 hour 10 minutes less headache on amitriptyline per month. We also note that these two study results give evidence on treatments 3, 6, and 7, which together make up precisely the contrasts of concern highlighted by the previous contrast level analysis (see Figure 2). Two other studies (Diener 2009 and Silberstein 2013) have thresholds that lie within the 95% CIs, showing that the treatment recommendation is sensitive to the level of imprecision in these estimates also.

The results of this analysis should lead to further scrutiny of the study evidence which the treatment recommendation is sensitive to; additionally, the results may placate any concerns raised about studies with wide invariant intervals which the treatment recommendation is not so sensitive to.

Study (Trt.)	Posterior Mean	95% Confidence Interval	Bias Invariant Interval with $\tilde{k}^*$		
Diener 2009 (1)	−1.14	(−2.26, −0.02)	–	(NT, −0.47)	2
Diener 2009 (2)	−1.65	(−2.72, −0.58)	2	(−2.32, NT)	–
Apostol 2008 (1)	−2.80	(−3.50, −2.10)	–	(NT, −1.48)	4
Apostol 2008 (4)	−3.10	(−3.93, −2.27)	4	(−7.92, 6.08)	3
Apostol 2008 (4)	−2.20	(−2.93, −1.47)	4	(−6.15, 3.50)	3
Apostol 2008 (4)	−2.80	(−3.43, −2.17)	4	(−6.16, 16.23)	3
Brandes 2004 (1)	−1.30	(−1.93, −0.67)	8	(−5.64, −0.43)	3
Brandes 2004 (6)	−2.90	(−3.53, −2.27)	3	(−5.01, 6.96)	8
Brandes 2004 (6)	−2.60	(−3.21, −1.99)	3	(−4.57, 9.43)	8
Brandes 2004 (6)	−1.70	(−2.29, −1.11)	3	(−3.85, NT)	–
Lewis 2009 (1)	−2.60	(−3.68, −1.52)	8	(−12.65, −1.12)	3
Lewis 2009 (6)	−4.90	(−5.93, −3.87)	3	(−8.24, 10.40)	8
Lewis 2009 (6)	−3.60	(−4.57, −2.63)	3	(−8.22, 19.98)	8
Lipton 2011 (1)	−5.30	(−5.84, −4.76)	8	(−11.50, −4.47)	3
Lipton 2011 (6)	−6.60	(−7.14, −6.06)	3	(−8.38, −1.58)	8
Silberstein 2004 (1)	−1.30	(−1.89, −0.71)	8	(−4.90, −0.62)	3
Silberstein 2004 (6)	−2.70	(−3.30, −2.10)	3	(−5.56, 9.63)	8
Silberstein 2004 (6)	−2.70	(−3.23, −2.17)	3	(−4.31, 9.00)	8
Silberstein 2004 (6)	−2.70	(−3.25, −2.15)	3	(−4.15, 8.36)	8
Winner 2005 (1)	−2.40	(−3.18, −1.62)	2	(−12.07, −1.26)	3
Winner 2005 (6)	−3.10	(−3.67, −2.53)	3	(−4.49, 5.40)	8
†**Diener 2004 (1)**	−1.10	(−1.57, −0.63)	3	(−1.29, NT)	–
†**Diener 2004 (6)**	−1.30	(−1.79, −0.81)	3	(−1.44, NT)	–
†**Diener 2004 (6)**	−1.80	(−2.29, −1.31)	3	(−1.94, NT)	–
†**Diener 2004 (7)**	−1.90	(−2.39, −1.41)	–	(NT, −1.85)	3
Holroyd 2010 (1)	−3.30	(−3.60, −3.00)	–	(NT, −2.71)	8
Holroyd 2010 (8)	−3.90	(−4.25, −3.55)	8	(−4.49, NT)	–
Silberstein 2013 (5 vs. 1)	0.00	(−1.30, 1.30)	5	(−1.19, NT)	–
Dodick 2009 †(3 vs. 6)	−0.10	(−0.90, 0.70)	3	(−0.15, NT)	–



○ Posterior Mean — 95% Confidence Interval — Bias Invariant Interval

Figure 3: Forest plot showing the results of the threshold analysis at study level. A bias adjustment to the reported study estimate outside of the bias invariant interval (shaded) will result in a new treatment recommendation (based on the greatest reduction in headache days per month); the new recommended treatment code is shown at either end of the interval in column 4. Bold text indicates contrasts where bias thresholds lie inside the 95% CI. The base-case recommended treatment is 7 (propranolol).

† indicates contrasts with bias thresholds less than the minimally important difference (0.5 days).

NT = no threshold; no amount of bias adjustment in this direction will change the treatment recommendation.

2.2 FURTHER EXAMPLES

2.2.1 TOCOLYTICS FOR PRETERM LABOUR

Nineteen tocolytic treatments were grouped into 7 classes and considered for their effects on estimated gestational age (EGA) at delivery, combining evidence from 51 studies in an NMA for clinical guideline NG25 by the National Collaborating Centre for Women's and Children's Health (NCC-WCH) [1]. Prostaglandin inhibitors were seen to have the greatest improvement on EGA, with a mean (95% CrI) increase of 2.32 (1.25, 3.35) weeks. The treatment network is shown in Figure 4.

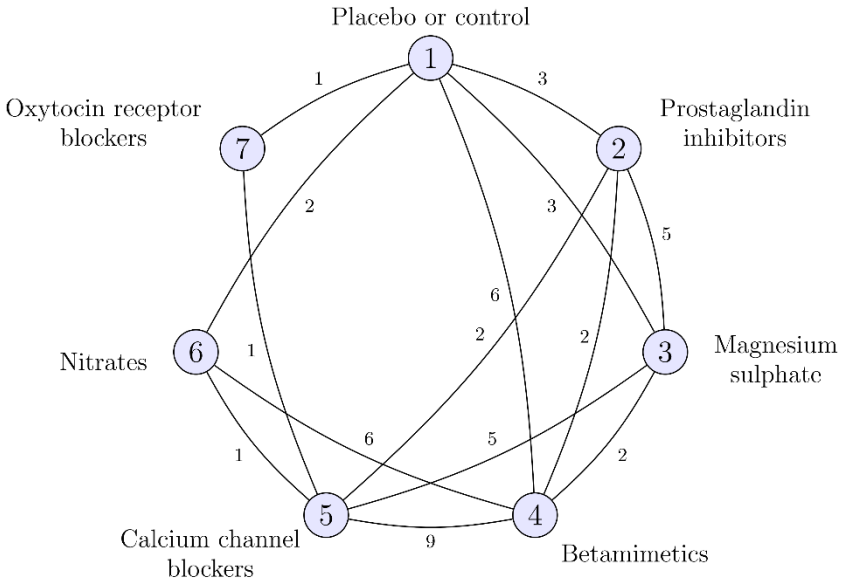


Figure 4: Network of tocolytic treatment classes. Nodes represent treatment classes and edges show study comparisons. Numbers inside the nodes are the treatment codings; numbers on the edges give the number of studies making that comparison. For full details see [1].

CONTRAST LEVEL ANALYSIS

Firstly, we consider the effects of bias adjustments to the combined body of evidence on individual contrasts. Figure 5 gives the results of such a contrast-level analysis, where invariant intervals (shaded lines) are presented about the posterior mean of each treatment contrast which has direct evidence. 95% credible intervals for the posterior means are also shown; none of the contrasts have thresholds which lie inside the 95% CrI, indicating that the treatment recommendation is robust to the level of imprecision in the combined evidence on each contrast. The majority of invariant intervals are wide; the smallest threshold is a positive change of 1.07 weeks in favour of treatment 6 on the 6 vs. 4 contrast, which would result in treatment 6 being recommended. The four contrasts comparing prostaglandin inhibitors (treatment 2) are one-sided – meaning that no amount of bias adjustment that increases the relative efficacy of treatment 2 will result in a new treatment recommendation. Although this seems intuitive, unless the base-case recommended treatment is located on a "spur" in the network with just one incident edge (i.e. only compared to one other treatment) then it is possible for *increases* (as well as decreases) in the efficacy of the base-case recommended

treatment to change the treatment recommendation; such a situation is more likely to occur when the decision is finely balanced between two or more treatments with similar efficacies, located close together in the network. Overall, the results of this analysis attest to the robustness of the treatment recommendation to bias in the combined evidence on each contrast.

Contrast	Posterior Mean	95% Credible Interval	Bias Invariant Interval with $\tilde{k}^*$		
2 vs. 1	2.32	(1.25, 3.35)	6	(–0.13, NT)	–
3 vs. 1	1.28	(0.27, 2.27)	6	(–6.76, 7.60)	3
4 vs. 1	1.25	(0.40, 2.08)	1	(–6.47, 5.43)	4
6 vs. 1	1.65	(0.51, 2.78)	1	(–29.13, 4.54)	6
7 vs. 1	0.68	(–1.34, 2.69)	1	(–71.48, 4.24)	7
3 vs. 2	–1.03	(–2.01, –0.04)	–	(NT, 0.84)	5
4 vs. 2	–1.07	(–2.08, –0.04)	–	(NT, 1.69)	6
5 vs. 2	–0.63	(–1.67, 0.43)	–	(NT, 1.86)	5
4 vs. 3	–0.04	(–0.99, 0.91)	3	(–12.45, 8.63)	6
5 vs. 3	0.40	(–0.50, 1.32)	3	(–4.14, 2.45)	5
5 vs. 4	0.44	(–0.32, 1.20)	4	(–3.09, 2.19)	5
6 vs. 4	0.40	(–0.55, 1.36)	4	(–13.58, 1.47)	6
6 vs. 5	–0.03	(–1.16, 1.09)	5	(–13.59, 4.50)	6
7 vs. 5	–1.01	(–2.99, 0.99)	5	(–21.29, 2.02)	7

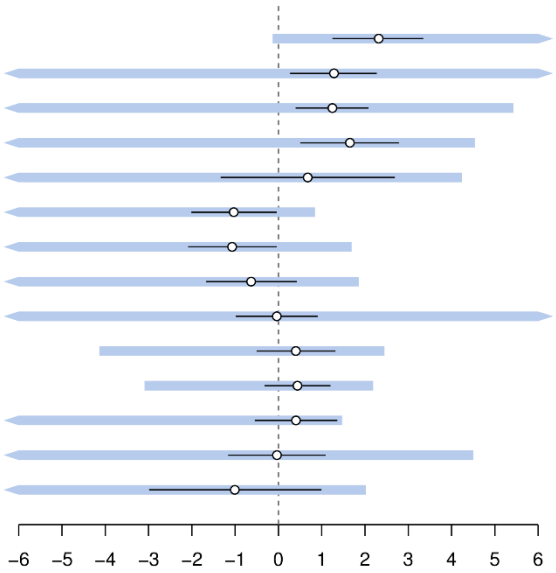


Figure 5: Results of the contrast-level threshold analysis for tocolytics treatments. Invariant intervals are plotted as thick shaded lines, and in the table are shown with the new recommended treatments at the thresholds at either side of the interval in column 4. The base-case treatment recommendation is prostaglandin inhibitors (2).

NT = no threshold; no amount of bias adjustment in this direction will change the treatment recommendation.

STUDY LEVEL ANALYSIS

We may also consider the effects of bias in individual study estimates. Figure 6 shows the results of the study-level threshold analysis, for study estimates with thresholds smaller than 10 weeks. As with the contrast-level analysis the invariant intervals are wide, indicating that the treatment recommendation is robust to bias adjustments in individual study estimates. The smallest threshold is in the treatment 5 arm of study 51, where a negative adjustment of –3.05 weeks to the estimated EGA at delivery would result in treatment 7 (oxytocin receptor blockers) becoming the recommended treatment. There are no study estimates with threshold that lie within the 95% confidence intervals, indicating that the treatment recommendation is robust to the level of imprecision in the estimates.

Study (Trt.)	Study Estimate	95% Confidence Interval	Bias Invariant Interval with $\tilde{k}^*$		
3 (2)	31.80	(30.92, 32.68)	5	(27.94, NT)	–
3 (3)	31.20	(30.37, 32.03)	–	(NT, 39.85)	3
3 (5)	31.80	(30.94, 32.66)	–	(NT, 36.51)	5
4 (1)	31.20	(30.88, 31.52)	–	(NT, 37.72)	6
4 (2)	36.40	(36.08, 36.72)	6	(30.41, NT)	–
5 (1)	33.00	(32.39, 33.61)	–	(NT, 38.04)	6
5 (2)	35.20	(34.64, 35.76)	6	(29.43, NT)	–
7 (1)	33.00	(32.89, 33.11)	3	(2.94, 39.25)	6
14 (1)	34.10	(32.64, 35.56)	6	(27.82, NT)	–
14 (6)	35.20	(34.08, 36.32)	–	(NT, 41.84)	6
15 (1)	38.30	(37.75, 38.85)	6	(32.60, NT)	–
15 (6)	37.80	(36.88, 38.72)	–	(NT, 43.11)	6
16 (1)	30.10	(29.10, 31.10)	7	(26.54, NT)	–
16 (7)	31.00	(30.01, 31.99)	–	(NT, 34.45)	7
18 (2)	35.50	(34.93, 36.07)	5	(28.87, NT)	–
18 (3)	35.70	(34.91, 36.49)	–	(NT, 42.70)	5
21 (2)	35.50	(34.28, 36.72)	6	(28.19, NT)	–
21 (4)	33.80	(32.13, 35.47)	–	(NT, 41.09)	6
22 (2)	36.70	(36.41, 36.99)	6	(32.66, NT)	–
22 (4)	35.20	(34.91, 35.49)	–	(NT, 39.87)	6
23 (2)	35.20	(34.51, 35.89)	5	(31.01, NT)	–
23 (5)	34.10	(33.25, 34.95)	–	(NT, 38.22)	5
24 (2)	34.90	(33.85, 35.95)	5	(25.65, NT)	–
24 (3)	34.80	(33.51, 36.09)	–	(NT, 43.63)	5
25 (2)	35.30	(34.65, 35.95)	5	(28.60, NT)	–
25 (3)	34.70	(33.91, 35.49)	–	(NT, 41.79)	5
27 (5)	35.60	(34.64, 36.56)	3	(14.34, 45.53)	5
28 (5)	34.30	(33.95, 34.65)	3	(18.27, 43.56)	5
30 (3)	35.80	(35.11, 36.49)	5	(27.87, 59.31)	3
30 (5)	36.00	(35.39, 36.61)	3	(13.86, 43.68)	5
40 (4)	37.40	(36.40, 38.40)	6	(31.05, NT)	–
40 (6)	37.10	(36.08, 38.12)	–	(NT, 43.82)	6
41 (4)	37.30	(36.53, 38.07)	6	(30.44, NT)	–
41 (6)	37.30	(36.44, 38.16)	–	(NT, 42.97)	6
42 (4)	35.20	(34.49, 35.91)	6	(29.02, NT)	–
42 (6)	35.10	(34.37, 35.83)	–	(NT, 41.00)	6
43 (4)	30.00	(28.16, 31.84)	6	(20.70, NT)	–
43 (6)	35.10	(33.39, 36.81)	–	(NT, 44.24)	6
46 (4)	36.30	(35.64, 36.96)	6	(30.82, NT)	–
46 (6)	36.50	(35.96, 37.04)	–	(NT, 41.62)	6
48 (4)	35.20	(34.48, 35.92)	6	(28.73, NT)	–
48 (6)	35.80	(35.05, 36.55)	–	(NT, 41.94)	6
50 (5)	34.30	(33.67, 34.93)	6	(29.67, 47.52)	5
50 (6)	35.60	(35.12, 36.08)	5	(18.54, 40.14)	6
51 (5)	36.40	(35.77, 37.03)	7	(33.35, 54.24)	5
51 (7)	35.20	(34.50, 35.90)	5	(14.63, 38.31)	7

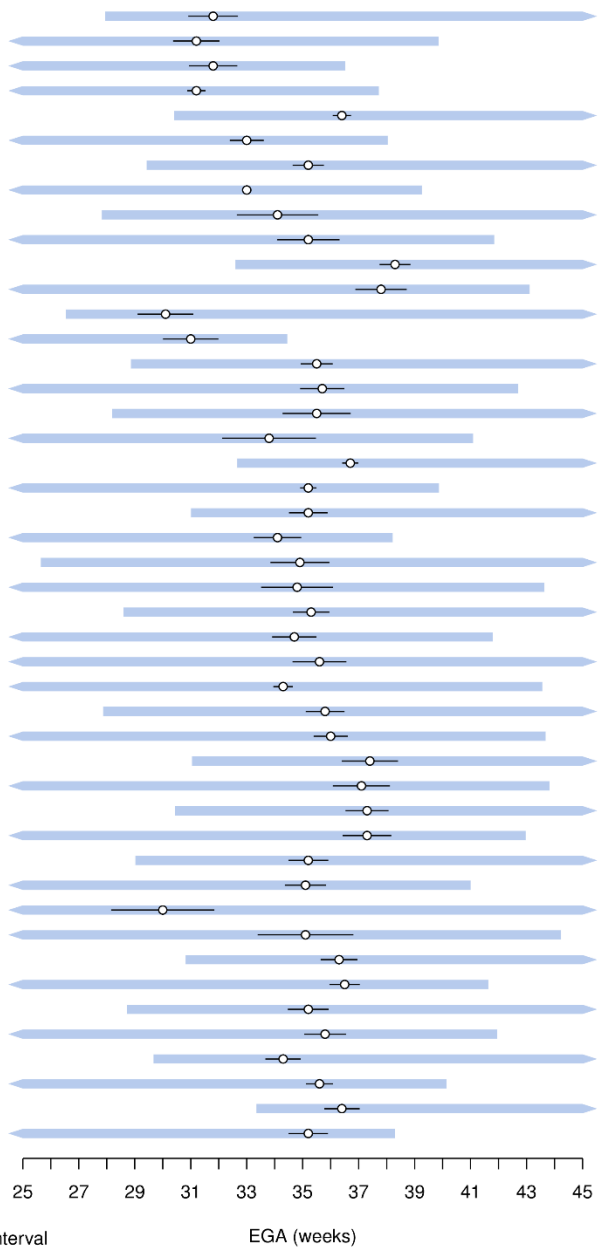


Figure 6: Results of the study-level threshold analysis for tocolytics treatments. Each study has two or more estimates, one for each arm; the treatment on each arm is shown in brackets. Only study estimates with thresholds smaller than 10 weeks are shown for brevity. Invariant intervals are plotted as thick shaded lines, and in the table are shown with the new recommended treatments at the thresholds at either side of the interval in column 4. The base-case treatment recommendation is prostaglandin inhibitors (2).

NT = no threshold; no amount of bias adjustment in this direction will change the treatment recommendation.

Overall, this analysis demonstrates the robustness of the treatment decision to bias adjustments at both study and contrast level. It should alleviate concerns about any one study or evidence base, and may reduce the need for laborious critical appraisal of each piece of evidence.

2.2.2 URINARY INCONTINENCE

The National Collaborating Centre for Women's and Children's Health (NCC-WCH) [2] performed a NMA of 14 treatments for the management of urinary incontinence in adult women for clinical guideline CG171, combining evidence from 22 studies; the resulting treatment network is shown in Figure 7. Immediate release (IR) oxybutynin was considered the reference treatment, as it was the primary recommendation from a previous analysis. The results of the NMA showed that all active treatments were significantly more effective than placebo, though there was no evidence for significant differences between the active treatments. Based on efficacy alone, oxybutynin IR was the first-ranked treatment.

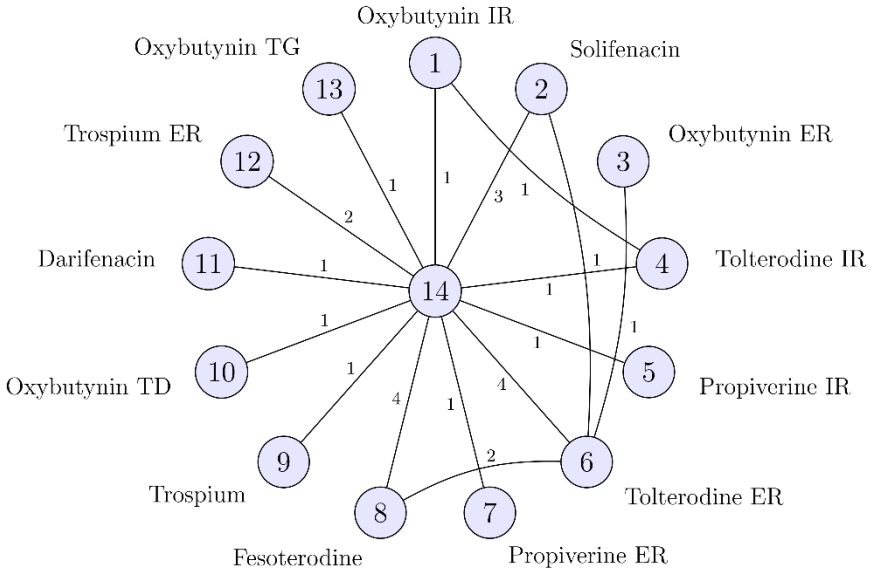


Figure 7: Treatment network for urinary incontinence. Nodes represent treatments and edges show study comparisons. Numbers inside the nodes are the treatment codings; numbers on the edges give the number of studies making that comparison. Treatment 14 is placebo. For full details see [2].

CONTRAST LEVEL ANALYSIS

The results of a contrast-level threshold analysis are shown in Figure 8. As we would expect, with the relatively small amount of evidence and the lack of statistical significance of the treatment effects, the first-place ranking of oxybutynin IR is sensitive to the imprecision on 7 out of 16 contrasts; these contrasts have thresholds that lie within the 95% credible intervals for the estimates. However, none of the new first-ranked treatments at any of the thresholds is placebo, which supports the conclusion of the original analysis that any active treatment is better than placebo. The smallest threshold is for contrast 4 vs. 1, where a positive adjustment of 0.10 to the LOR of continence in favour of treatment 4 (tolterodine IR) results in tolterodine IR becoming the first ranked treatment.

Contrast	Posterior Mean	95% Credible Interval	Bias Invariant Interval with $\tilde{k}^*$		
4 vs. 1	−0.08	(−0.70, 0.54)	–	(NT, 0.02)	4
14 vs. 1	−1.10	(−1.86, −0.34)	–	(NT, −0.27)	11
6 vs. 2	−0.29	(−0.42, −0.16)	2	(−1.09, 2.27)	3
14 vs. 2	−0.68	(−0.82, −0.54)	2	(−1.54, NT)	–
6 vs. 3	−0.35	(−0.72, 0.02)	3	(−0.71, 241.30)	5
14 vs. 4	−1.02	(−1.82, −0.25)	4	(−1.44, −0.36)	11
14 vs. 5	−0.80	(−1.64, 0.03)	5	(−1.11, 70.49)	4
8 vs. 6	0.18	(0.08, 0.27)	3	(−0.85, 1.61)	8
14 vs. 6	−0.39	(−0.49, −0.29)	3	(−1.10, 23.33)	4
14 vs. 7	−0.54	(−0.89, −0.20)	7	(−1.10, NT)	–
14 vs. 8	−0.57	(−0.67, −0.46)	8	(−1.49, 19.12)	11
14 vs. 9	−0.76	(−1.30, −0.23)	9	(−1.11, 23.97)	4
14 vs. 10	−0.61	(−1.47, 0.22)	10	(−1.11, 61.67)	3
14 vs. 11	−0.86	(−1.41, −0.33)	11	(−1.11, NT)	–
14 vs. 12	−0.67	(−0.91, −0.43)	12	(−1.11, 7.09)	4
14 vs. 13	−0.61	(−0.96, −0.27)	13	(−1.11, 23.44)	11

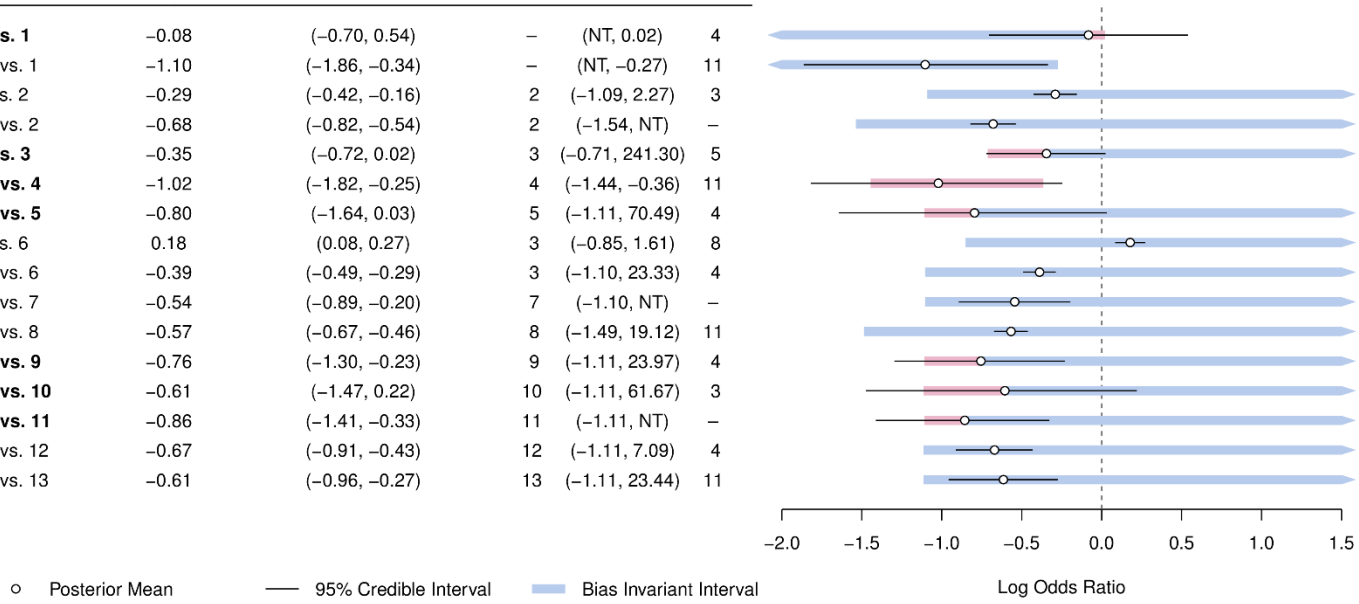


Figure 8: Results of the contrast-level threshold analysis for urinary incontinence. Invariant intervals are plotted as thick shaded lines, and in the table are shown with the new first ranked treatments at the thresholds at either side of the interval in column 4. Contrasts where thresholds lie within the 95% CrI are shown in bold. The base-case rank 1 treatment is oxybutynin IR (treatment 1).
NT = no threshold; no amount of bias adjustment in this direction will change the first place ranking.

STUDY LEVEL ANALYSIS

To identify which studies the first place treatment ranking is sensitive to, we perform a study-level threshold analysis. The results of this are shown in Figure 9. An in the contrast-level analysis, we see that the first place ranking is sensitive to the level of imprecision in some estimates; however, this sensitivity is seen in only 8 out of 22 studies. The other 14 studies have wider invariant intervals, and thus may not require such an in-depth assessment of risk of bias. The influential studies provide evidence on the same contrasts as identified as influential in the contrast-level analysis, and indeed show the same thresholds and new first ranked treatments; for example the smallest threshold is for the one study estimating the 4 vs. 1 contrast (study 7), where a positive adjustment of 0.10 to the LOR of continence results in treatment 4 being ranked first. This is because the network is sparse and contains few loops, and most contrasts are only formed by one study. In this analysis, where there are multiple studies estimating the same contrast, or the network is better connected and so provides indirect evidence as well as direct, we see that the first place ranking is more robust to these studies and contrasts.

Study (Timepoint)	LOR	95% Confidence Interval	Bias Invariant Interval with $\tilde{k}^*$		
1 (1): 6 vs. 2	−0.19	(−0.43, 0.05)	2	(−1.85, 5.11)	3
1 (2): 6 vs. 2	−0.37	(−0.59, −0.14)	2	(−1.90, 4.55)	3
2 (1): 14 vs. 5	−0.78	(−1.60, 0.05)	5	(−1.08, 67.55)	4
3 (1): 14 vs. 7	−0.54	(−0.89, −0.19)	7	(−1.09, NT)	–
4 (1): 8 vs. 6	0.11	(−0.11, 0.32)	3	(−12.61, 3.44)	8
4 (1): 14 vs. 6	−0.59	(−0.87, −0.31)	3	(−3.54, 877.46)	10
4 (2): 8 vs. 6	0.24	(0.03, 0.46)	3	(−17.11, 3.42)	8
4 (2): 14 vs. 6	−0.44	(−0.71, −0.18)	3	(−3.02, 813.12)	10
5 (1): 8 vs. 6	0.16	(−0.02, 0.34)	3	(−11.27, 2.40)	8
5 (1): 14 vs. 6	−0.30	(−0.53, −0.08)	3	(−2.17, 560.11)	10
5 (2): 8 vs. 6	0.18	(0.00, 0.36)	11	(−13.40, 2.42)	8
5 (2): 14 vs. 6	−0.20	(−0.42, 0.02)	3	(−1.96, 563.36)	10
6 (1): 14 vs. 12	−0.64	(−1.05, −0.24)	12	(−1.89, 21.31)	4
6 (2): 14 vs. 12	−0.69	(−1.07, −0.31)	12	(−1.79, 18.57)	4
7 (1): 4 vs. 1	0.05	(−0.63, 0.74)	–	(NT, 0.15)	4
8 (1): 14 vs. 1	−1.31	(−2.28, −0.34)	–	(NT, −0.91)	11
9 (1): 14 vs. 2	−0.62	(−0.97, −0.26)	2	(−3.18, NT)	–
10 (1): 14 vs. 2	−0.78	(−1.12, −0.44)	2	(−3.13, NT)	–
11 (1): 14 vs. 2	−0.67	(−1.04, −0.30)	2	(−3.52, NT)	–
12 (1): 6 vs. 3	−0.34	(−0.71, 0.03)	3	(−0.71, 239.13)	5
13 (1): 14 vs. 4	−0.70	(−1.67, 0.26)	4	(−1.07, −0.14)	11
14 (1): 14 vs. 6	−0.30	(−0.93, 0.33)	3	(−14.10, 457.91)	4
15 (1): 14 vs. 6	−0.54	(−0.94, −0.14)	3	(−6.11, 184.45)	4
16 (1): 14 vs. 8	−0.49	(−0.84, −0.14)	8	(−6.53, 128.84)	11
17 (1): 14 vs. 8	−1.02	(−1.45, −0.58)	8	(−10.36, 199.04)	11
18 (1): 14 vs. 9	−0.74	(−1.27, −0.22)	9	(−1.08, 23.12)	4
19 (1): 14 vs. 10	−0.58	(−1.41, 0.25)	10	(−1.07, 58.87)	3
20 (1): 14 vs. 11	−0.84	(−1.37, −0.30)	11	(−1.08, NT)	–
21 (1): 14 vs. 12	−0.66	(−1.14, −0.19)	12	(−2.38, 29.33)	4
22 (1): 14 vs. 13	−0.61	(−0.95, −0.27)	13	(−1.11, 23.31)	11

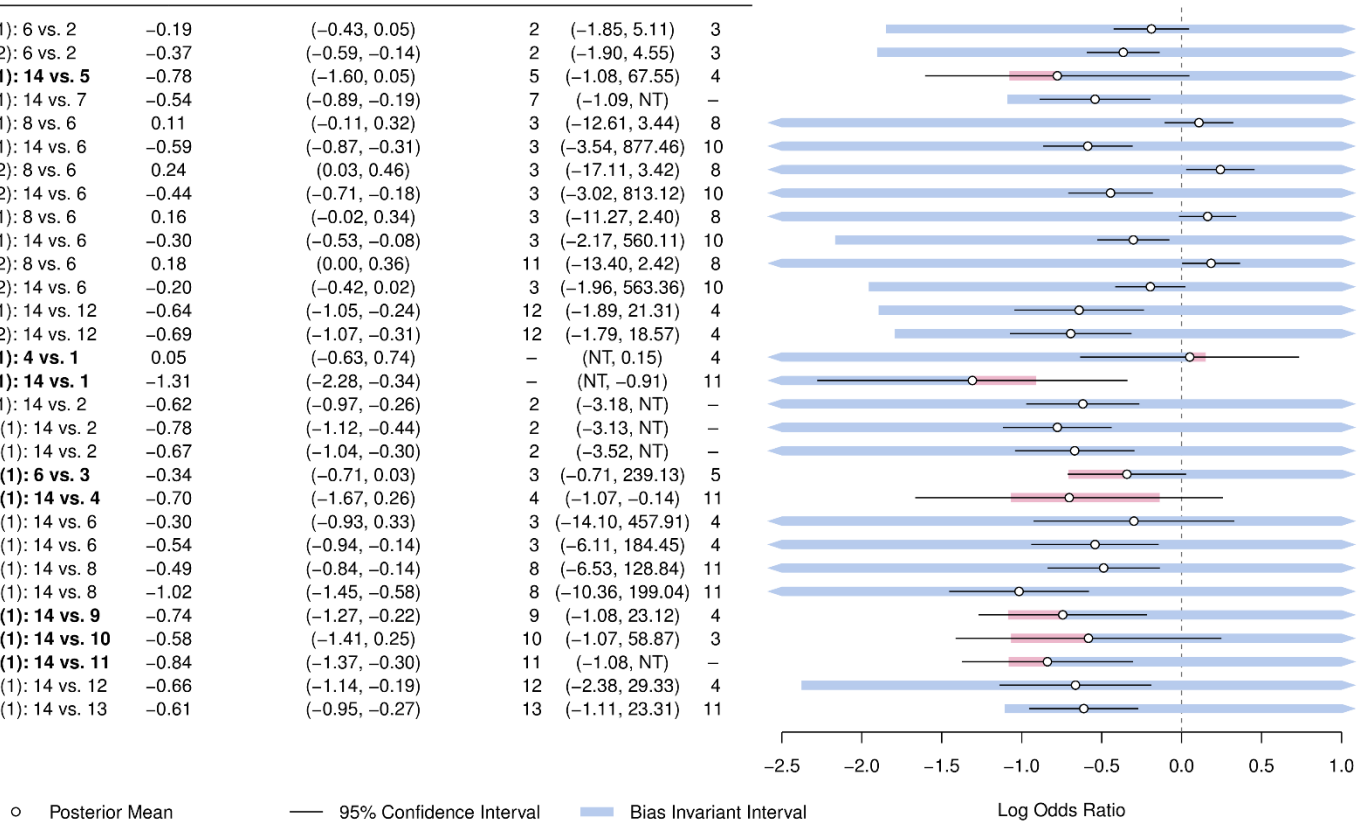


Figure 9: Results of the study-level threshold analysis for urinary incontinence. Invariant intervals are plotted as thick shaded lines, and in the table are shown with the new first ranked treatments at the thresholds at either side of the interval in column 4. Study estimates where thresholds lie within the 95% CrI are shown in bold. The base-case rank 1 treatment is oxybutynin IR (treatment 1).

NT = no threshold; no amount of bias adjustment in this direction will change the first place ranking.

2.2.3 SOCIAL ANXIETY

We now consider a more complex example where analysis is greatly simplified using the contrast-level approach. Figure 10 shows the network for an NMA of 41 treatments for social anxiety from 100 studies, performed by the National Collaborating Centre for Mental Health for clinical guideline CG159 [3]. The original analysis uses a random effects model with treatment response as an outcome. The model includes class effects for 17 different treatment classes, incorporates data in the form of standardised mean differences (SMDs) and log odds ratios, and includes data on recovery based on a regression calibration. The resulting model is highly complex, and the influence of individual pieces of evidence cannot be determined intuitively.

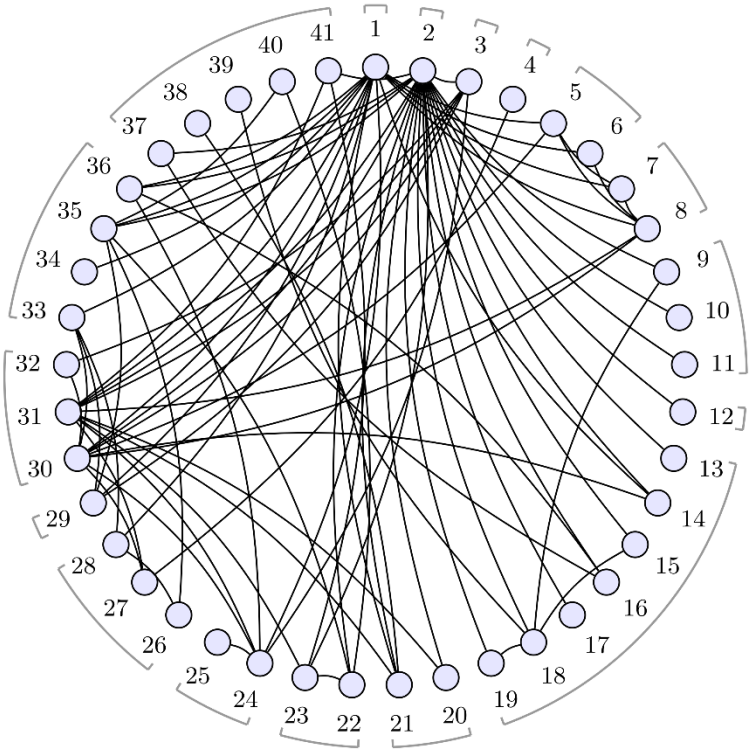


Figure 10: Social Anxiety treatment network. Nodes represent treatments and edges show study comparisons. Numbers around the edge are the treatment codings. Treatment classes are indicated by the braces, some classes contain a single treatment only. Treatment 1 is waitlist, treatment 2 is pill placebo, and treatment 3 is psychological placebo. For full details see National Collaborating Centre for Mental Health [3].

CONTRAST-LEVEL ANALYSIS

Firstly we present the results of a contrast-level threshold analysis, which, despite the complexity of the original analysis, is straightforward to perform. Due to the large number of contrasts, Figure 11 shows the results of the threshold analysis only for the 48 contrasts with thresholds less than 3 SMD, out of the total 84 contrasts. The base-case treatment recommendation is treatment 41, group cognitive behavioural therapy (CBT) with phenelzine. All but five contrasts have thresholds greater than 0.8 SMD in magnitude, a difference which Cohen [30] considered to be large for behavioural sciences. The smallest threshold is a positive change of 0.46 in the SMD of the 41 vs. 31 contrast, at which point treatment 36 (cognitive therapy) becomes the recommended treatment. For all five of the contrasts with

Contrast	Posterior Mean	95% Credible Interval	Bias Invariant Interval with $\tilde{k}^*$		
41 vs. 31	−0.88	(−1.27, −0.49)	–	(NT, −0.42)	36
41 vs. 23	−0.41	(−0.80, −0.02)	–	(NT, 0.06)	36
36 vs. 1	−1.56	(−1.86, −1.27)	36	(−2.05, 13.61)	7
36 vs. 16	−0.69	(−1.04, −0.35)	36	(−1.26, 4.91)	16
36 vs. 24	−0.73	(−1.06, −0.40)	36	(−1.42, 5.61)	25
41 vs. 2	−1.21	(−1.60, −0.83)	–	(NT, −0.39)	36
17 vs. 2	−0.47	(−0.68, −0.26)	17	(−1.33, 50.71)	12
13 vs. 2	−0.45	(−0.65, −0.25)	13	(−1.33, 139.66)	4
39 vs. 18	−0.36	(−0.91, 0.18)	39	(−1.31, 3.60)	36
38 vs. 21	−0.22	(−0.70, 0.27)	38	(−1.16, 3.71)	36
18 vs. 2	−0.53	(−0.65, −0.40)	39	(−1.48, 8.79)	36
23 vs. 2	−0.81	(−1.05, −0.56)	23	(−2.69, 0.17)	36
36 vs. 2	−1.09	(−1.41, −0.77)	36	(−2.11, 34.05)	18
16 vs. 2	−0.40	(−0.65, −0.14)	36	(−1.43, 17.40)	12
19 vs. 2	−0.49	(−0.67, −0.32)	19	(−1.55, 29.73)	23
25 vs. 24	−0.05	(−0.49, 0.39)	25	(−1.16, 554.74)	39
31 vs. 8	0.07	(−0.17, 0.32)	31	(−13.67, 1.22)	36
34 vs. 1	−1.02	(−1.42, −0.63)	34	(−2.21, 101.29)	5
31 vs. 1	−0.80	(−1.02, −0.58)	31	(−16.38, 0.40)	36
32 vs. 30	−0.25	(−0.61, 0.12)	32	(−1.57, 5.66)	36
11 vs. 2	−0.42	(−0.89, 0.04)	11	(−1.84, 856.05)	34
9 vs. 2	−0.25	(−0.51, 0.01)	9	(−1.68, 46.63)	36
40 vs. 35	−0.05	(−0.52, 0.43)	40	(−1.53, 3.12)	36
30 vs. 24	−0.02	(−0.27, 0.22)	30	(−8.80, 1.46)	36
15 vs. 2	−0.41	(−0.63, −0.19)	15	(−1.92, 49.40)	23
22 vs. 2	−0.27	(−0.48, −0.06)	22	(−1.81, 5.92)	18
8 vs. 1	−0.88	(−1.04, −0.71)	8	(−2.86, 0.68)	36
8 vs. 6	−0.21	(−0.46, 0.04)	8	(−13.40, 1.36)	6
37 vs. 30	−0.10	(−0.46, 0.25)	37	(−1.84, 7.10)	36
8 vs. 7	−0.03	(−0.34, 0.28)	8	(−17.13, 1.72)	7
31 vs. 23	0.47	(0.21, 0.74)	36	(−1.32, 2.58)	23
21 vs. 2	−0.60	(−0.95, −0.25)	21	(−2.41, 10.57)	36
12 vs. 2	−0.33	(−0.93, 0.28)	12	(−2.30, 59.55)	36
35 vs. 1	−1.19	(−1.48, −0.89)	35	(−3.27, 2.62)	36
33 vs. 30	−0.12	(−0.37, 0.13)	33	(−2.23, 8.91)	30
35 vs. 22	−0.45	(−0.77, −0.13)	35	(−2.66, 7.94)	22
32 vs. 3	−0.47	(−0.85, −0.09)	32	(−2.68, 13.60)	3
31 vs. 21	0.27	(−0.09, 0.63)	31	(−17.42, 2.60)	21
30 vs. 1	−0.85	(−1.04, −0.67)	30	(−7.21, 1.50)	36
27 vs. 4	−0.04	(−0.62, 0.53)	29	(−1303.22, 2.32)	4
31 vs. 3	−0.17	(−0.44, 0.10)	31	(−13.04, 2.20)	36
18 vs. 15	−0.12	(−0.35, 0.11)	39	(−8.32, 2.35)	15
33 vs. 1	−0.97	(−1.21, −0.74)	33	(−3.55, 6.18)	36
31 vs. 24	0.03	(−0.26, 0.31)	31	(−38.05, 2.61)	36
31 vs. 27	−0.41	(−0.83, 0.01)	31	(−33.53, 2.32)	27
30 vs. 3	−0.22	(−0.48, 0.03)	36	(−3.01, 3.90)	3
8 vs. 5	−0.03	(−0.27, 0.20)	8	(−7.95, 2.77)	5
30 vs. 5	−0.01	(−0.27, 0.25)	30	(−11.91, 2.87)	5

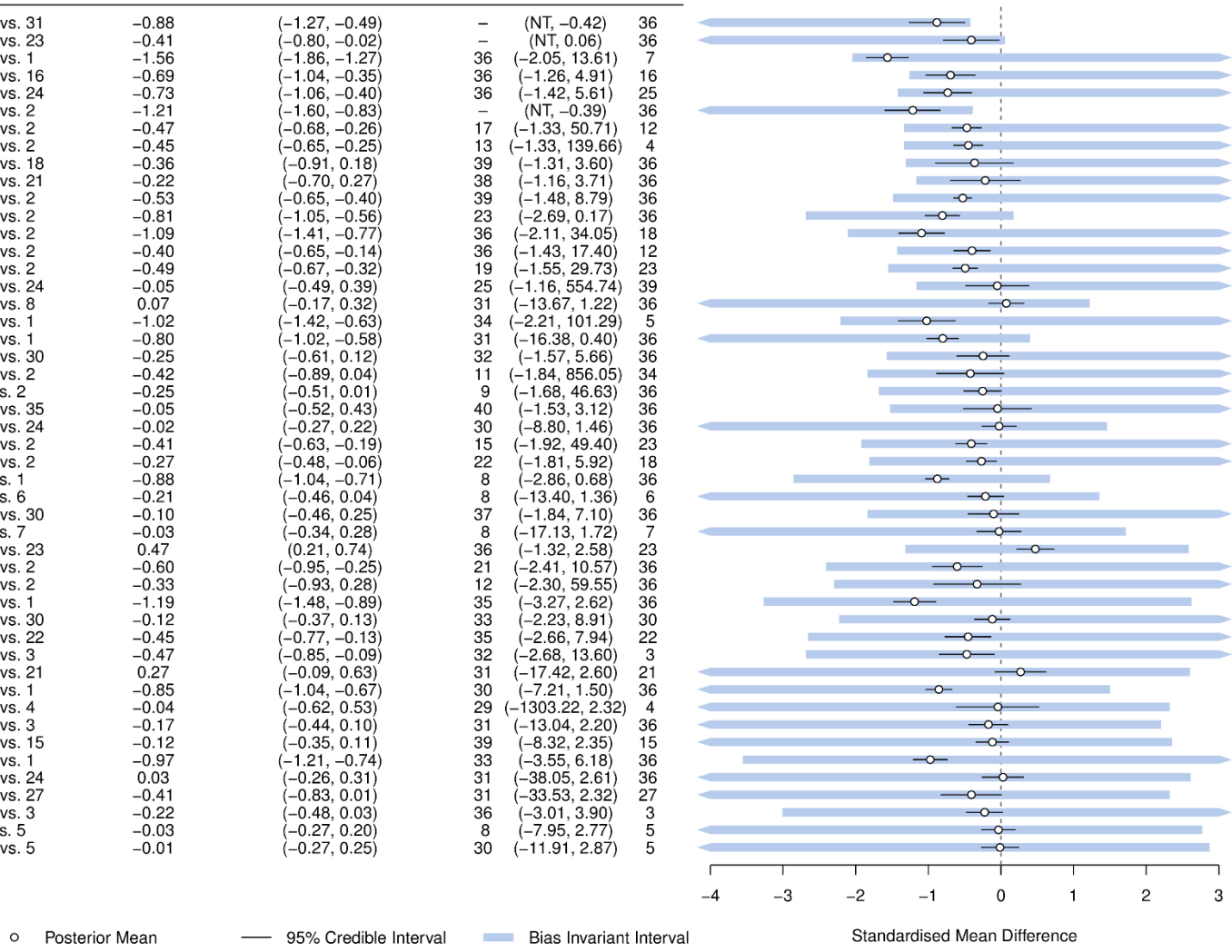


Figure 11: Contrast-level forest plot for the Social Anxiety example showing results of the threshold analysis, sorted with smallest thresholds first. Invariant intervals are plotted as thick shaded lines, and in the table are shown with the new recommended treatments at the thresholds at either side of the interval in column 4. Only contrasts with a threshold smaller than 3 SMD are shown here for brevity. The base-case recommended treatment is 41.

NT = no threshold; no amount of bias adjustment in this direction will change the first place ranking.

the smallest thresholds, cognitive therapy is the new treatment recommendation at the threshold. Two of these five contrasts involve treatment 41 (ranked first in the base-case analysis) and the other three involve treatment 36 (ranked second); adjustments that reduce the efficacy of treatment 41 or increase that of treatment 36 both result in treatment 36 being the recommended treatment.

The large majority of invariant intervals in the contrast-level analysis are wide, and there are no thresholds which lie inside the respective 95% CrI which would indicate sensitivity to the imprecision on that contrast. Altogether, this suggests that the first place ranking on efficacy of group CBT with phenelzine is largely robust to changes in the combined evidence on contrasts, with sensitivity to plausible bias adjustments on only a small number of contrasts. In

particular, the evidence on only a small subset of all 84 contrasts need be assessed qualitatively in detail for possible bias.

STUDY-LEVEL ANALYSIS

We also perform a threshold analysis where bias adjustments are considered for individual study estimates. Due to the large number of studies involved, Figure 12 shows the results of the threshold analysis only for the 44 study data points with thresholds less than 3 SMD, out of the total 146 study data points. Echoing the contrast-level analysis, the large majority of study estimates have wide invariant intervals, and there are only six data points with thresholds less than 0.8 SMD in magnitude. Of particular note are the 41 vs. 2 contrast of study 96 (BLANCO2010), in which a positive adjustment of 0.22 to the SMD reducing the efficacy of treatment 41 compared to treatment 2 results in treatment 36 becoming recommended, and the 36 vs. 2 contrast of study 81 (CLARK2003), in which a negative adjustment of –0.50 to the SMD increasing the efficacy of treatment 36 compared to treatment 2 also results in treatment 36 becoming recommended. These thresholds lie within the 95% confidence intervals for the study estimates, indicating that the first place rank of treatment 41 is sensitive to the imprecision in these data points. Indeed, study 96 is the only one comparing treatment 41, and then only on 32 patients, so this is not surprising.

The large majority of invariant intervals in the study-level analysis are wide. This suggests that the first place ranking on efficacy of group CBT with phenelzine is insensitive to the results of most studies, which should allay concerns (either reactively or pre-emptively) raised for many studies regarding their perceived quality or bias, instead focussing discussion on the evidence which is most influential.

Study (Contrast)	SMD	95% Confidence Interval	Bias Invariant Interval with $\tilde{k}^*$		
96 (41 vs 2)	−1.28	(−1.73, −0.84)	–	(NT, −1.06)	36
96 (31 vs 2)	−0.25	(−0.68, 0.18)	36	(−0.72, NT)	–
81 (36 vs 2)	−0.74	(−1.26, −0.21)	36	(−1.23, 7.39)	16
72 (36 vs 1)	−1.85	(−2.40, −1.31)	36	(−2.45, 8.65)	25
50 (39 vs 18)	−0.43	(−1.08, 0.21)	39	(−1.15, 2.65)	36
91 (40 vs 22)	−0.68	(−1.24, −0.13)	40	(−1.46, 1.64)	36
18 (12 vs 2)	−0.33	(−0.79, 0.13)	12	(−1.21, 47.47)	36
100 (37 vs 2)	−0.53	(−0.84, −0.22)	37	(−1.42, 3.78)	36
81 (16 vs 2)	−0.13	(−0.65, 0.39)	16	(−4.95, 0.83)	36
52 (38 vs 21)	−0.11	(−0.62, 0.41)	38	(−1.11, 3.88)	36
96 (23 vs 2)	−0.52	(−0.95, −0.09)	36	(−1.57, NT)	–
79 (36 vs 1)	−1.91	(−2.39, −1.43)	36	(−2.96, 177.49)	1
72 (24 vs 1)	−0.94	(−1.47, −0.42)	25	(−6.43, 0.15)	36
17 (11 vs 2)	−0.49	(−0.91, −0.08)	11	(−1.65, NT)	–
79 (36 vs 1)	−1.63	(−2.10, −1.16)	36	(−2.79, 17.26)	34
89 (32 vs 3)	−0.51	(−0.85, −0.16)	32	(−1.68, 19.16)	39
95 (20 vs 2)	−0.12	(−0.80, 0.57)	20	(−1.30, 28.73)	23
91 (35 vs 22)	−0.85	(−1.39, −0.30)	36	(−2.90, 0.65)	40
100 (16 vs 2)	−0.67	(−0.99, −0.35)	36	(−2.20, 4.55)	37
12 (34 vs 1)	−0.45	(−0.91, −0.00)	34	(−2.07, 22.43)	39
49 (31 vs 8)	0.23	(−0.08, 0.54)	31	(−26.00, 1.91)	36
94 (3 vs 2)	−0.26	(−0.71, 0.19)	36	(−2.03, 7.40)	23
36 (21 vs 2)	−0.91	(−1.32, −0.50)	21	(−2.71, 15.43)	18
71 (31 vs 1)	−1.06	(−1.80, −0.32)	31	(−30.68, 0.77)	36
53 (25 vs 24)	−0.01	(−0.45, 0.43)	25	(−1.86, 221.32)	39
88 (35 vs 2)	−0.34	(−0.89, 0.20)	35	(−2.23, 13.08)	22
89 (30 vs 3)	−0.17	(−0.52, 0.19)	36	(−9.32, 1.80)	32
73 (28 vs 1)	−0.44	(−0.84, −0.04)	28	(−2.42, 13.62)	36
94 (23 vs 2)	−1.40	(−1.87, −0.94)	23	(−3.54, 0.61)	36
16 (10 vs 2)	−0.40	(−1.27, 0.48)	10	(−2.52, 85.91)	39
87 (23 vs 2)	−1.53	(−2.02, −1.04)	23	(−4.97, 0.66)	36
44 (27 vs 4)	−0.15	(−0.70, 0.39)	19	(−5412.72, 2.06)	4
51 (31 vs 21)	0.11	(−0.39, 0.60)	31	(−15.93, 2.39)	21
40 (23 vs 2)	−0.33	(−0.81, 0.15)	23	(−6.27, 2.02)	36
13 (34 vs 1)	−1.70	(−2.35, −1.04)	34	(−4.08, NT)	–
98 (7 vs 5)	0.14	(−0.32, 0.59)	7	(−2.35, 42.53)	8
74 (29 vs 1)	−0.87	(−1.14, −0.59)	29	(−3.39, 38.84)	33
71 (24 vs 1)	−1.50	(−2.30, −0.70)	36	(−4.06, 64.60)	31
87 (22 vs 2)	−0.95	(−1.43, −0.48)	36	(−3.55, 7.44)	23
80 (9 vs 2)	−0.04	(−0.33, 0.25)	9	(−2.68, 4.86)	36
99 (18 vs 2)	−0.45	(−0.65, −0.25)	39	(−7.45, 2.24)	15
23 (16 vs 2)	0.00	(−0.44, 0.44)	36	(−2.89, 20.96)	39
60 (31 vs 27)	−0.77	(−1.24, −0.30)	31	(−24.36, 2.13)	27
95 (31 vs 2)	−0.12	(−0.75, 0.51)	31	(−10.44, 2.81)	20

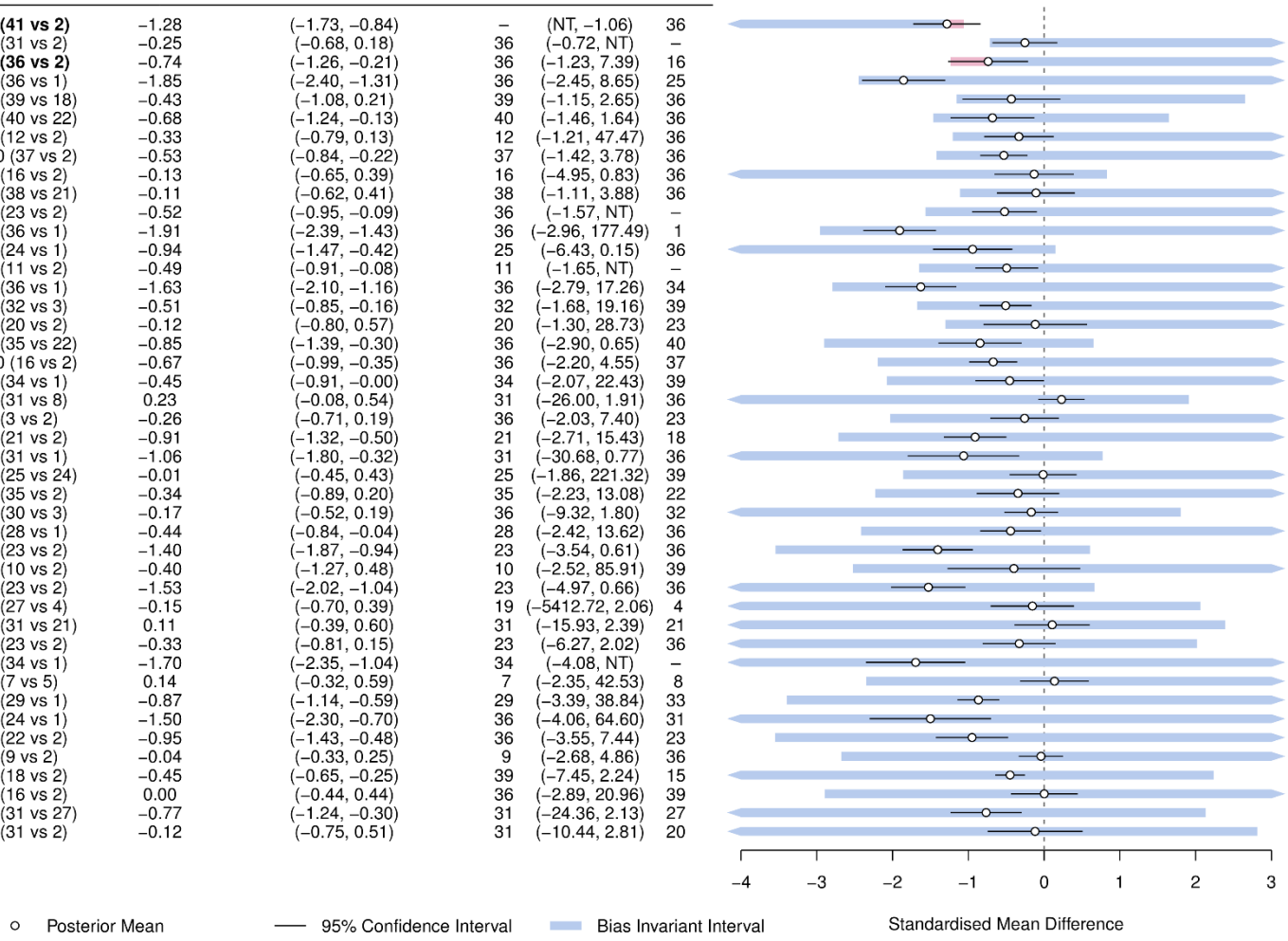


Figure 12: Study-level forest plot for the Social Anxiety example showing results of the threshold analysis, sorted with smallest thresholds first. Invariant intervals are plotted as thick shaded lines, and in the table are shown with the new recommended treatments at the thresholds at either side of the interval in column 4. Only contrasts with a threshold smaller than 3 SMD are shown here for brevity. The base-case recommended treatment is 41.

NT = no threshold; no amount of bias adjustment in this direction will change the first place ranking.

2.3 SUMMARY OF THE SECTION

In this section we have seen the implementation of the threshold method in practice, explained the ideas behind the derivation of bias-adjustment thresholds and invariant intervals, and presented and interpreted the results of threshold analyses for NMAs of varying types. A more detailed background to the threshold method is given in [27, 28]; technical details are available on request. We have also suggested how the results might be used as part of the decision making process, and the potential uses and impacts they could have. The role of threshold analysis in guideline development is explored further in section 6.

3. MORE COMPLEX EXAMPLES

In this section we continue with the social anxiety NMA (section 2.2.3), and consider more complex uses of threshold analysis where, rather than considering bias in individual studies or contrasts, we consider the effects of biases that might apply to entire groups of contrasts or studies at the same time. Such analyses are likely to be considered reactively to explore concerns raised about specific biases to treatment classes or sets of studies. Here we consider generic biases to pharmacological treatments when compared to an inactive treatment, and similarly for psychological treatments. We then consider these two biases simultaneously, to examine how the presence of these specific biases jointly might affect the treatment recommendation. We also explore the sensitivity of the treatment recommendation to potential bias in studies carried out by a particular investigator. Finally, we repeat our analysis where the treatment decision is restricted to a subset of the treatments in the full NMA. It should be noted that no new techniques or methods are required to carry out any of these analyses; they may be performed using the same tools and ideas as used in earlier sections [27, 28].

3.1 GENERIC BIAS IN PHARMACOLOGICAL VS. INACTIVE CONTRASTS

Firstly we consider a generic bias adjustment for pharmacological vs. inactive contrasts (treatments numbers 9-23 and 1-3, respectively). Such an adjustment may arise for example when it is thought that treatment effects are being exaggerated for active pharmacological treatments, perhaps due to lack of blinding. Such biases are believed to occur quite commonly for subjective reported outcomes, such as those in social anxiety studies (see section 5). We assess how much bias adjustment may be introduced to the evidence on pharmacological vs. inactive contrasts before the treatment decision changes, deriving an invariant interval which is presented in Figure 13.

Since all pharmacological vs. inactive contrasts are considered for bias adjustment by the same amount simultaneously, there is a single invariant interval. A bias adjustment to account for an exaggeration of treatment effect would need to be larger than 2.23 SMD before the treatment recommendation changes, at which point treatment 2 (drug placebo) is recommended. Not only is this adjustment large but it would also mean that, in truth, pharmacological treatments are significantly worse than an inactive comparator. The threshold for bias adjustment in the opposite direction, to account for an underestimation of effect, is also reasonably large at –1.22 SMD, at which point treatment 39 (paroxetine with clonazepam) would be the recommended treatment. As such, we may consider the treatment recommendation to be robust to such generic bias in the pharmacological vs. inactive contrasts.

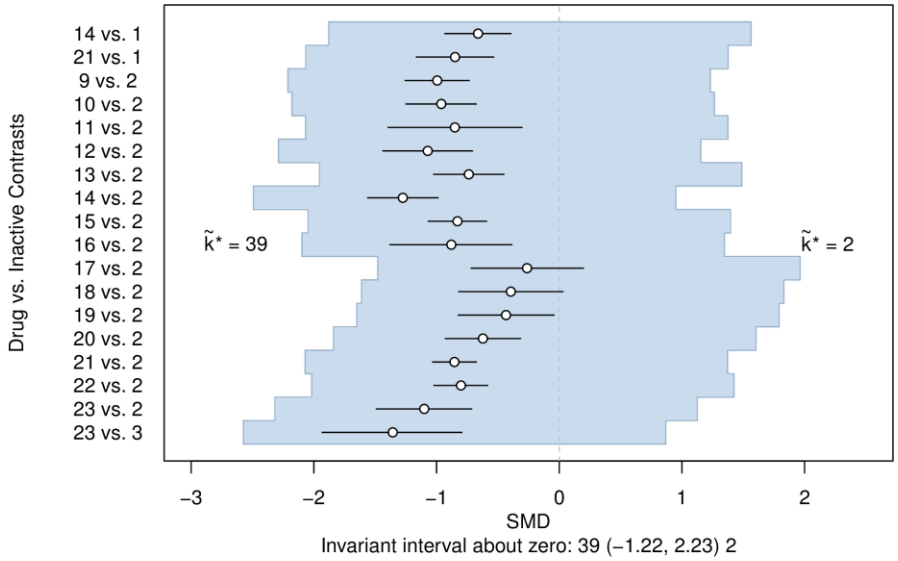


Figure 13: The invariant interval for all pharmacological treatments against an inactive comparator, considered to be bias adjusted by the same amount simultaneously. The base-case recommended treatment is 41.

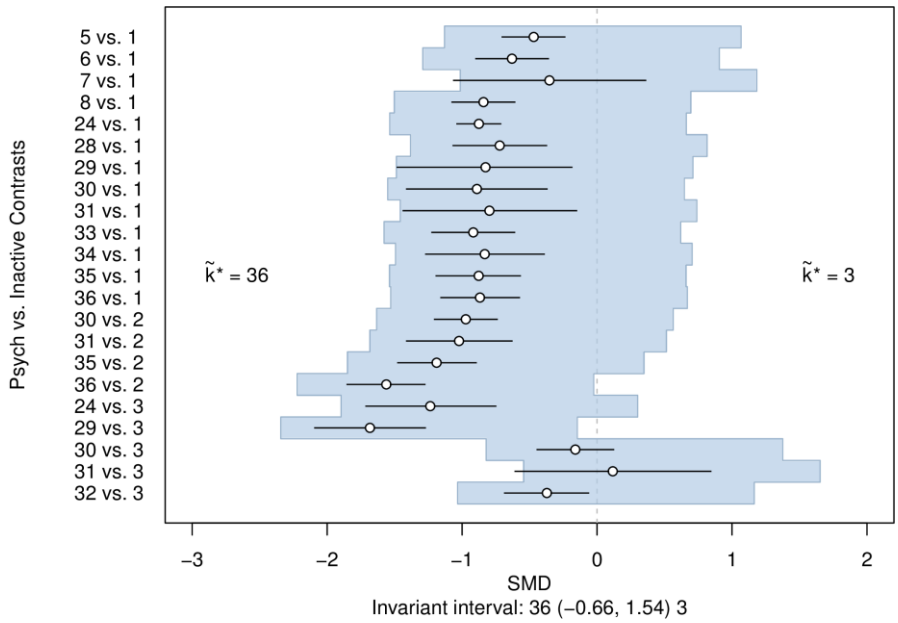


Figure 14: The invariant interval for all psychological treatments against an inactive comparator, where all contrasts are considered to be bias adjusted by the same amount simultaneously. The base-case recommended treatment is 41.

3.2 GENERIC BIAS IN PSYCHOLOGICAL VS. INACTIVE CONTRASTS

In the same manner to the pharmacological vs. inactive contrasts, we may consider the psychological vs. inactive (treatments numbers 4-8, 24-36 and 1-3, respectively) contrasts for generic bias adjustment. Again, since all psychological treatment vs. inactive contrasts are considered for bias adjustment by the same amount simultaneously, there is a single invariant interval which is presented in Figure 14. If the true effect of all psychological treatments is understated when compared to an inactive comparator, then a bias adjustment for this of more than –0.66 SMD would result in treatment 36 (cognitive therapy) being recommended. On the other hand, in the perhaps more likely scenario, an adjustment to correct for the effect of psychological treatment being exaggerated when compared to an inactive treatment would need to be 1.54 SMD or larger to change the treatment recommendation, at which point psychological placebo (treatment 3) would be recommended. Not only is this adjustment relatively large, but would require most psychological treatments to be significantly worse than an inactive comparator. We may therefore consider the treatment recommendation to be robust to generic bias in the psychological vs. inactive contrasts also.

3.3 SIMULTANEOUS PHARMACOLOGICAL AND PSYCHOLOGICAL TREATMENT BIAS

The threshold method is not restricted to analysis in one dimension; we may assess sensitivity to bias in multiple studies or contrasts simultaneously. Here we examine the effects of simultaneous bias adjustments for the generic pharmacological and psychological treatment biases discussed above. Figure 15 shows the 2-dimensional invariant region (shaded) formed by threshold lines; bias adjustments within this region do not change the treatment recommendation, however crossing a threshold line will result in a new treatment being recommended. The invariant

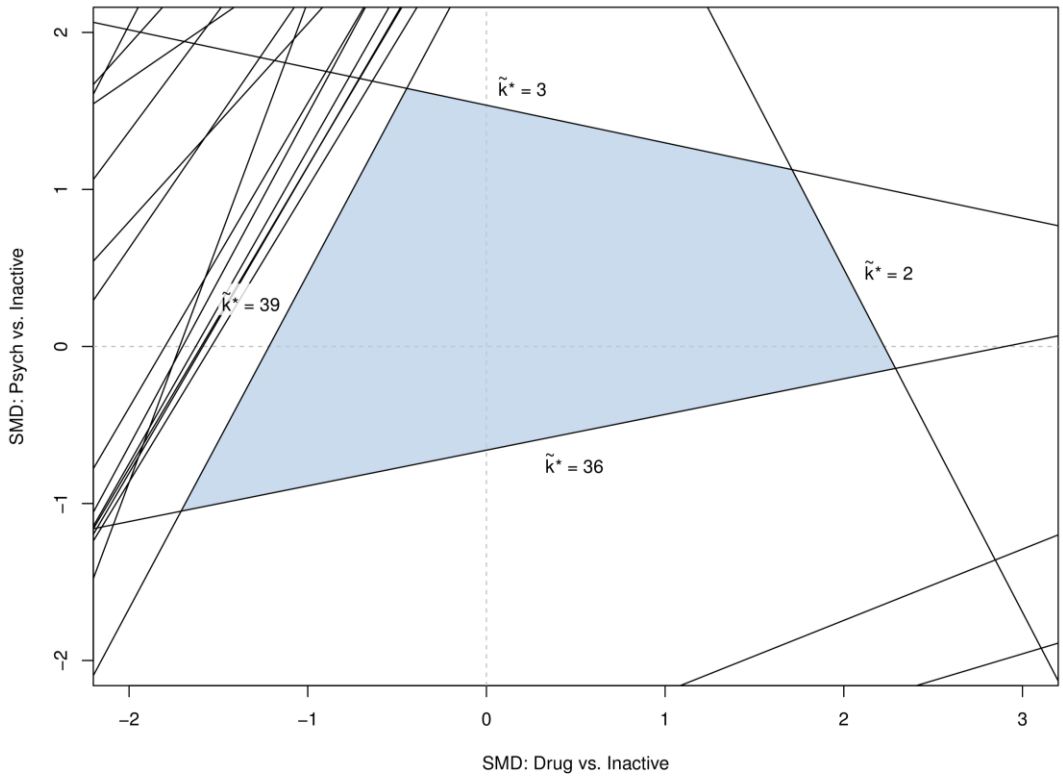


Figure 15: Invariant region (shaded) for simultaneous adjustment for psychological and pharmacological biases. The new treatment recommendations at the threshold lines are shown; the base-case treatment recommendation is treatment 41.

region is large, requiring large bias adjustments to cross a threshold line. The most notable feature is that increasing the amount of positive bias adjustment for pharmacological vs. inactive contrasts reduces the amount of negative bias adjustment for psychological vs. inactive contrasts required to cross the threshold line for treatment 36 to become optimal. For example, with zero bias adjustment to the pharmacological vs. inactive contrasts, the threshold value for bias adjustment to the psychological vs. inactive contrasts is –0.66 SMD; with a 1 SMD adjustment to the pharmacological contrasts this reduces to –0.43 SMD; and with a 2 SMD adjustment to –0.21 SMD.

3.4 BIAS IN STUDIES BY A PARTICULAR INVESTIGATOR

Professor David M. Clark developed treatment 36 (cognitive therapy, or CT) and also authored three trials which compare cognitive therapy to other active and inactive treatments. A critic of CT might argue that patients undergoing CT in a Clark trial achieve better outcomes than they otherwise would with a different PI in another CT trial. With threshold analysis we do not attempt to detect whether any such bias exists, rather whether it could plausibly make any difference to the treatment decision.

Figure 16 shows the invariant interval for bias adjustment to the CT arms of Clark trials. A bias adjustment of –0.18 SMD in favour of treatment 36 (CT) would result in CT becoming the recommended treatment. Such an adjustment is small and lies within the 95% CI of each study estimate, however the direction of this adjustment would be to correct an *underperformance* of CT in Clark trials – not an exaggeration of effect as suggested. In the opposite direction, there is no plausible bias adjustment for overestimation of CT efficacy in Clark trials that could result in a new treatment recommendation. In the context of a base-case recommendation for CBT plus phenelzine, such a result should provide defence from specific criticisms relating to the Clark trials and the possibility of over-performance.

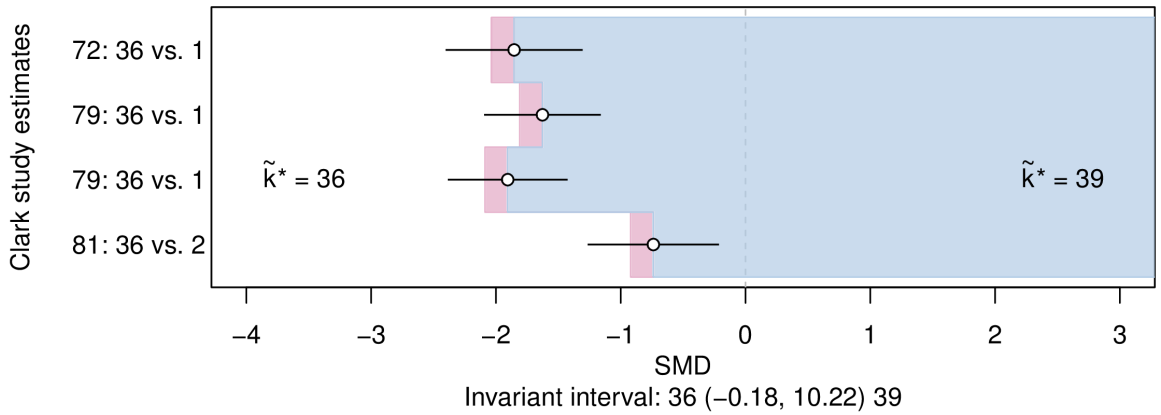


Figure 16: The invariant interval for bias adjustment in the CT arms of Clark trials. The base-case treatment recommendation is treatment 41.

3.5 THRESHOLD ANALYSES FOR A SUBSET OF TREATMENTS

Using inputs from the social anxiety network meta-analysis, a cost-effectiveness analysis (CEA) was performed to arrive at a final treatment recommendation. However, not all treatments were considered for inclusion in the CEA; for example if there was insufficient evidence on the efficacy of a treatment, or if a treatment was not available in the UK. Of particular note is the decision to remove treatment 41 (group CBT with phenelzine) from the CEA, which was the highest ranked treatment for efficacy in the NMA. As a consequence, treatment 36 (cognitive therapy) which ranked second amongst all treatments in the NMA, was ranked first amongst those included in the CEA. We therefore also consider the robustness of the treatment decision when ranking is restricted to treatments included in the CEA, which in the base case is treatment 36.

Contrast	Posterior Mean	95% Credible Interval	Bias Invariant Interval with $\tilde{k}^*$		
17 vs. 2	−0.47	(−0.68, −0.26)	17	(−1.21, 12.65)	23
13 vs. 2	−0.45	(−0.65, −0.25)	13	(−1.21, 26.15)	23
18 vs. 2	−0.53	(−0.65, −0.40)	18	(−1.35, 18.78)	23
23 vs. 2	−0.81	(−1.05, −0.56)	23	(−1.65, 6.16)	18
19 vs. 2	−0.49	(−0.67, −0.32)	19	(−1.38, 14.25)	23
34 vs. 1	−1.02	(−1.42, −0.63)	34	(−2.02, 24.96)	23
31 vs. 23	0.47	(0.21, 0.74)	31	(−5.64, 1.57)	23
36 vs. 1	−1.56	(−1.86, −1.27)	–	(NT, −0.44)	23
9 vs. 2	−0.25	(−0.51, 0.01)	9	(−1.50, 368.36)	26
36 vs. 16	−0.69	(−1.04, −0.35)	–	(NT, 0.60)	23
15 vs. 2	−0.41	(−0.63, −0.19)	15	(−1.72, 16.86)	23
22 vs. 2	−0.27	(−0.48, −0.06)	22	(−1.60, 5.02)	18
35 vs. 1	−1.19	(−1.48, −0.89)	35	(−2.57, 15.31)	7
8 vs. 6	−0.21	(−0.46, 0.04)	8	(−9.88, 1.19)	6
8 vs. 1	−0.88	(−1.04, −0.71)	8	(−2.29, 13.09)	34
8 vs. 7	−0.03	(−0.34, 0.28)	8	(−12.88, 1.48)	7
36 vs. 24	−0.73	(−1.06, −0.40)	–	(NT, 0.85)	23
12 vs. 2	−0.33	(−0.93, 0.28)	12	(−2.02, 1207.51)	17
33 vs. 30	−0.12	(−0.37, 0.13)	33	(−2.01, 5.94)	30
35 vs. 22	−0.45	(−0.77, −0.13)	35	(−2.46, 4.74)	23
33 vs. 1	−0.97	(−1.21, −0.74)	33	(−2.99, 20.32)	7
18 vs. 15	−0.12	(−0.35, 0.11)	18	(−6.74, 2.03)	15
16 vs. 2	−0.40	(−0.65, −0.14)	16	(−2.97, 1.87)	23
36 vs. 2	−1.09	(−1.41, −0.77)	–	(NT, 1.21)	23
8 vs. 5	−0.03	(−0.27, 0.20)	8	(−6.83, 2.36)	5
30 vs. 5	−0.01	(−0.27, 0.25)	30	(−8.66, 2.56)	5
31 vs. 27	−0.41	(−0.83, 0.01)	23	(−10.65, 2.24)	27
30 vs. 24	−0.02	(−0.27, 0.22)	23	(−3.54, 2.63)	24
23 vs. 22	−0.54	(−0.83, −0.25)	23	(−3.30, 8.22)	22
7 vs. 5	−0.01	(−0.34, 0.33)	7	(−2.84, 7.06)	5
33 vs. 29	−0.35	(−0.68, −0.03)	33	(−4.55, 2.50)	29
19 vs. 18	0.03	(−0.16, 0.23)	19	(−2.83, 5.60)	18
31 vs. 8	0.07	(−0.17, 0.32)	23	(−2.86, 5.23)	8

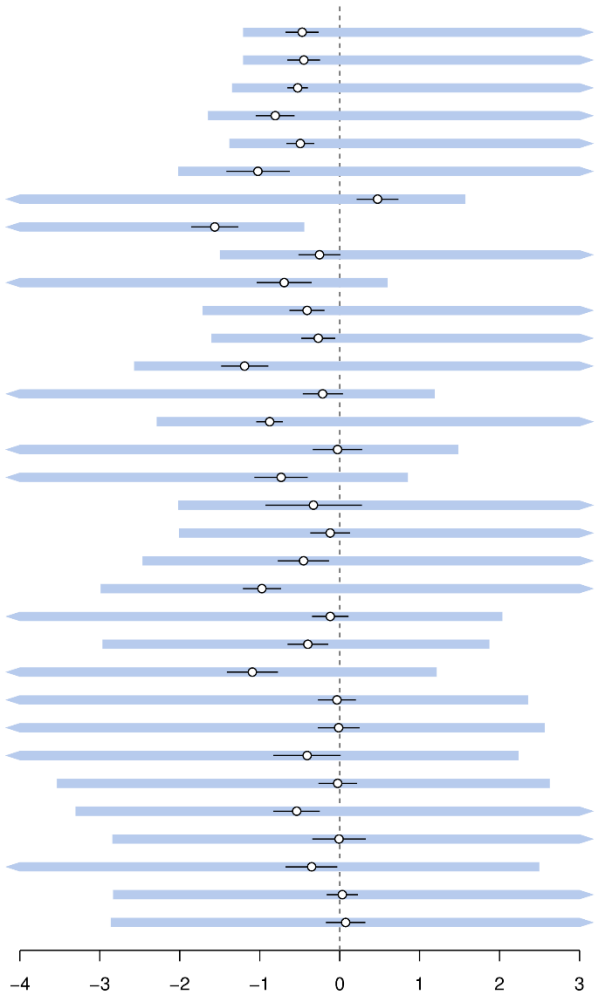


Figure 17: Contrast-level forest plot for the Social Anxiety example showing results of the threshold analysis for treatments included in the CEA only, sorted with smallest thresholds first. Only contrasts with a threshold smaller than 3 SMD are shown here for brevity. The base-case treatment recommendation is 36.

NT = no threshold; no amount of bias adjustment in this direction will change the first place ranking.

Figure 17 shows the contrast-level results of the threshold analysis restricted to treatments included in the CEA only. The results are broadly similar to those including all treatments, and indicate that the treatment recommendation for CT is largely robust to bias in aggregate data on the contrasts. There are only two contrasts with thresholds less than 0.8 SMD in magnitude, and none of the contrasts have thresholds that lie within the 95% CrI. The smallest two thresholds are negative adjustments of –0.74 and –0.76 SMD to the 17 vs. 2 and 13 vs. 2 contrasts, at which point treatments 17 and 13 are recommended respectively.

The results of the study level threshold analysis for CEA-included treatments in Figure 18, parallel the contrast level results: we again see that the restricted decision is again largely robust to biases, this time in the individual study

Study (Contrast)	SMD	95% Confidence Interval	Bias Invariant Interval with $\tilde{k}^*$		
18 (12 vs 2)	−0.33	(−0.79, 0.13)	12	(−1.08, 284.21)	17
81 (36 vs 2)	−0.74	(−1.26, −0.21)	–	(NT, 0.35)	23
94 (23 vs 2)	−1.40	(−1.87, −0.94)	23	(−2.55, 46.02)	29
96 (23 vs 2)	−0.52	(−0.95, −0.09)	23	(−1.82, 17.57)	31
12 (34 vs 1)	−0.45	(−0.91, −0.00)	34	(−1.77, 25.34)	23
72 (36 vs 1)	−1.85	(−2.40, −1.31)	–	(NT, −0.47)	23
87 (23 vs 2)	−1.53	(−2.02, −1.04)	23	(−3.17, 6.99)	22
88 (35 vs 2)	−0.34	(−0.89, 0.20)	35	(−2.02, 4.94)	23
73 (28 vs 1)	−0.44	(−0.84, −0.04)	28	(−2.20, 14.06)	33
13 (34 vs 1)	−1.70	(−2.35, −1.04)	34	(−3.72, 78.17)	23
100 (16 vs 2)	−0.67	(−0.99, −0.35)	16	(−2.81, 4.31)	30
98 (7 vs 5)	0.14	(−0.32, 0.59)	7	(−2.01, 29.58)	8
80 (9 vs 2)	−0.04	(−0.33, 0.25)	9	(−2.20, 21.79)	18
81 (16 vs 2)	−0.13	(−0.65, 0.39)	23	(−2.36, NT)	–
74 (29 vs 1)	−0.87	(−1.14, −0.59)	29	(−3.16, 20.78)	33
75 (35 vs 1)	−0.06	(−0.72, 0.60)	35	(−2.36, 8.06)	29
94 (3 vs 2)	−0.26	(−0.71, 0.19)	30	(−17.90, 2.10)	23
91 (35 vs 22)	−0.85	(−1.39, −0.30)	35	(−3.21, 7.21)	23
40 (23 vs 2)	−0.33	(−0.81, 0.15)	23	(−2.71, 17.27)	17
95 (23 vs 2)	−0.61	(−1.28, 0.07)	23	(−3.02, 32.63)	31
72 (24 vs 1)	−0.94	(−1.47, −0.42)	23	(−3.42, NT)	–
79 (36 vs 1)	−1.63	(−2.10, −1.16)	–	(NT, 0.86)	23
99 (18 vs 2)	−0.45	(−0.65, −0.25)	18	(−5.95, 2.04)	15
96 (31 vs 2)	−0.25	(−0.68, 0.18)	31	(−5.42, 2.27)	23
79 (36 vs 1)	−1.91	(−2.39, −1.43)	–	(NT, 0.82)	23
60 (31 vs 27)	−0.77	(−1.24, −0.30)	23	(−10.96, 2.00)	27
19 (13 vs 2)	−0.37	(−0.54, −0.20)	13	(−3.14, 28.31)	34
92 (5 vs 1)	−0.40	(−0.74, −0.05)	5	(−3.19, 14.56)	30
22 (15 vs 2)	−0.25	(−0.44, −0.06)	15	(−3.10, 18.09)	23
93 (2 vs 2)	−0.13	(−0.37, 0.12)	34	(−23.82, 2.77)	13
21 (14 vs 2)	−0.37	(−1.06, 0.31)	14	(−3.29, 24.35)	23
28 (17 vs 2)	−0.42	(−0.63, −0.21)	17	(−3.38, 26.69)	23
93 (13 vs 2)	−0.40	(−0.64, −0.15)	13	(−3.40, 171.61)	35

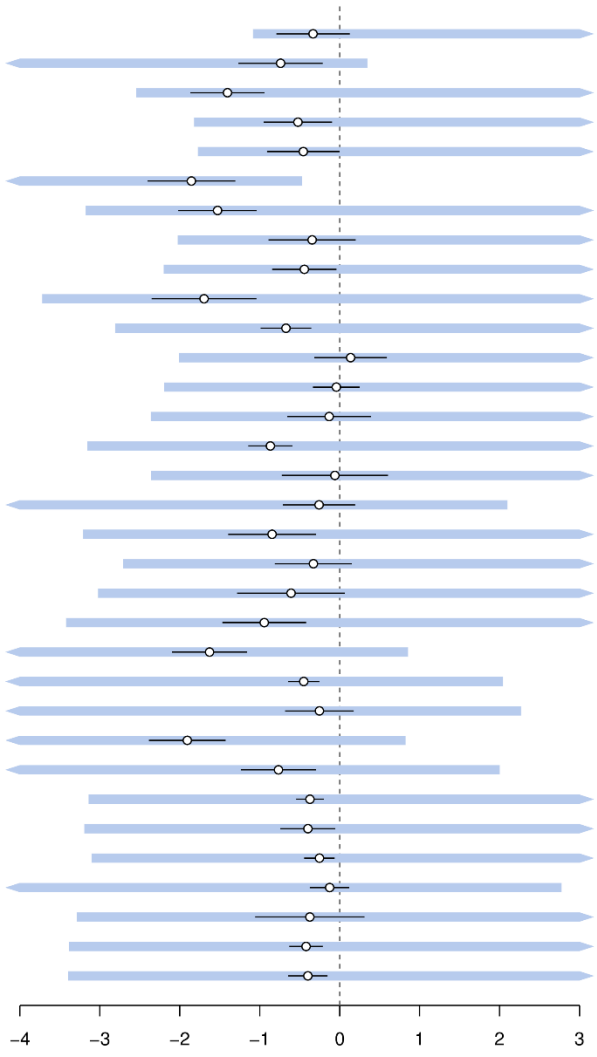


Figure 18: Study-level forest plot for the Social Anxiety example showing results of the threshold analysis for treatments included in the CEA only, sorted with smallest thresholds first. Only contrasts with a threshold smaller than 3 SMD are shown here for brevity. The base-case treatment recommendation is 36.

NT = no threshold; no amount of bias adjustment in this direction will change the first place ranking.

estimates. There is only one study estimate with a threshold less than 0.8 SMD, in the 12 vs. 2 contrast of study 18. Here a negative bias of –0.75 SMD in favour of treatment 12 results in treatment 12 being recommended. There are no study estimates with thresholds that lie within the 95% CI, indicating that the treatment decision restricted to treatments included in the CEA is not sensitive to the imprecision of any estimates.

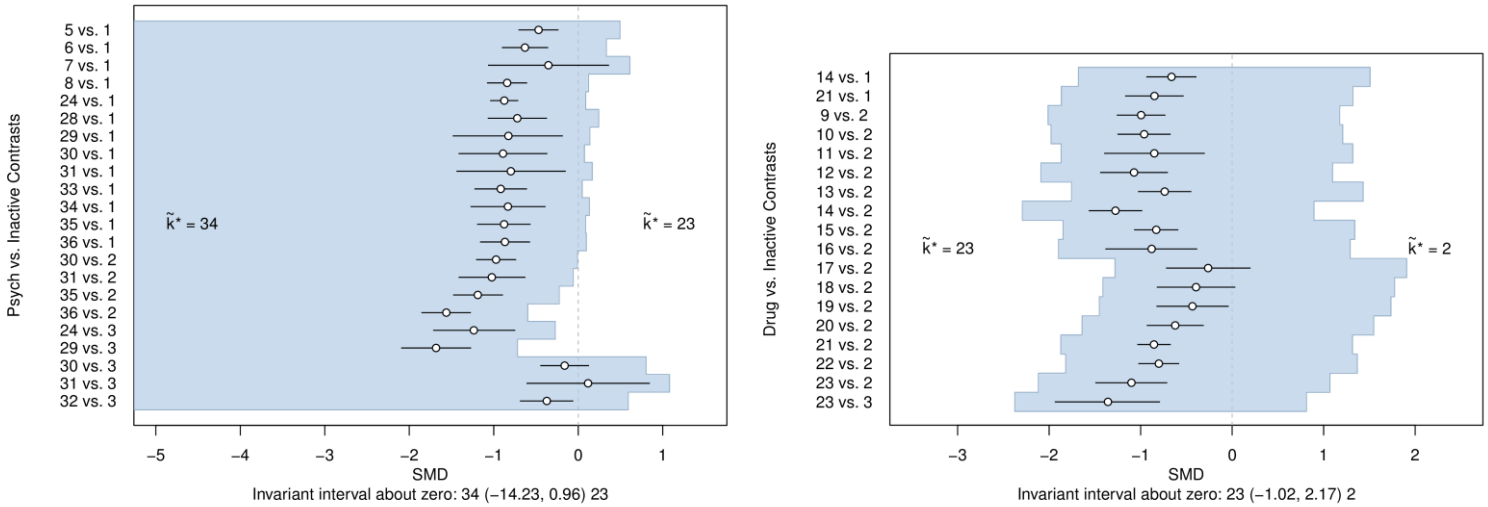


Figure 19: (L) The invariant interval for all psychological treatments against an inactive comparator, where all contrasts are considered to be bias adjusted by the same amount simultaneously. (R) The invariant interval for all pharmacological treatments against an inactive comparator, considered to be bias adjusted by the same amount simultaneously. The base-case recommended treatment is 36 when restricted to treatments included in the CEA.

We also repeat the analyses for generic pharmacological and psychological treatment biases, when the treatment decision is restricted to treatments included in the CEA. These results are displayed in Figure 19. A bias adjustment reducing the efficacy of all psychological treatments compared with inactive by 0.96 SMD, or a bias adjustment increasing the efficacy of all pharmacological treatments compared with inactive by 1.02 SMD both result in treatment 23 (phenelzine) being recommended. In order for treatment 2 to become recommended from adjusting for a pharmacological treatment bias, a reduction in efficacy of 2.17 SMD of all pharmacological treatments is required – which is likely implausible, as this would result in all drugs being deemed detrimental compared to an inactive treatment. In general, the treatment recommendation restricted to CEA treatments is less sensitive to these bias adjustments than the treatment recommendation based on all treatments.

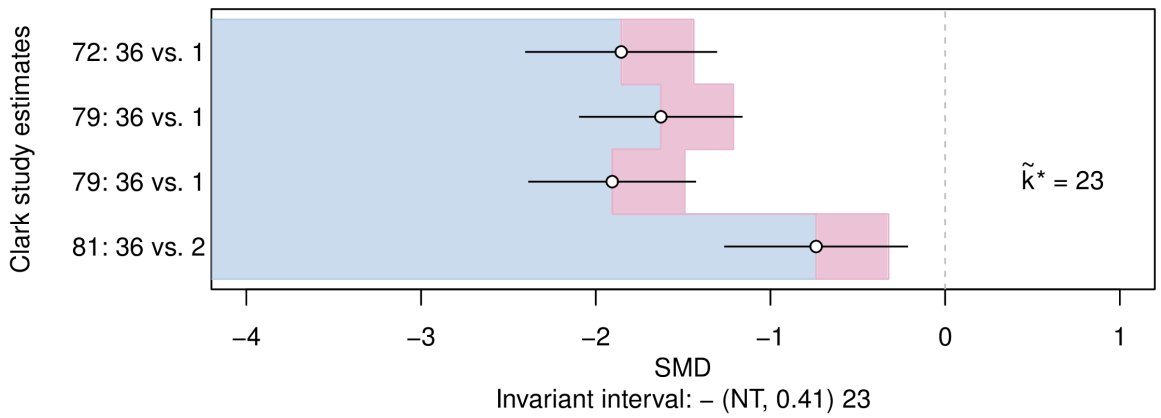


Figure 20: The invariant interval for a treatment decision based on treatments included in the CEA only for bias adjustment in the CT arms of Clark trials. The base-case treatment recommendation is 36.

NT = no threshold; a bias adjustment in this direction can never change the treatment decision.

Figure 20 shows the invariant interval for the restricted treatment decision for bias adjustment in the cognitive therapy arms of studies carried out by Professor David M. Clark. There is no negative threshold in this case, meaning that adjustments increasing the efficacy of CT (treatment 36) compared to inactive treatment can never result in a new treatment decision. However, we now see that a positive bias adjustment of 0.41 SMD, reducing the efficacy of CT compared to inactive treatment, results in treatment 23 (phenelzine) becoming the new recommendation. This threshold lies inside the 95% CI for each study estimate, meaning that the treatment recommendation is sensitive to the imprecision in these estimates. If the effect of CT is exaggerated in Clark studies by this amount or greater, any bias adjustment to account for this would result in phenelzine being recommended.

4. THRESHOLDS FOR NET BENEFIT

In the preceding sections we have explored the effects of bias adjustments on treatment recommendations based on efficacy, where the recommendation may be formed directly from the joint Bayesian posterior of the treatment parameters resulting from the NMA. However it is commonplace in guideline development that treatment recommendations are not based upon efficacy alone; instead, any potential gains in efficacy are considered alongside incurred costs, and evaluated together in a Cost-Effectiveness Analysis (CEA) [31]. Treatment recommendations are then made based on expected (incremental) net benefit. That is, if NB_k is the net benefit of the k -th treatment, then in the notation of equation (2) we have $F\left(k,\boldsymbol{\theta}\right)=\mathrm{NB}_k$ and decisions are made based on $\mathbb{E}_{\boldsymbol{\theta}}\left(\mathrm{NB}_k\right)$.

In this section we explore how threshold method is extended to assess sensitivity to bias adjustments for treatment decisions made based on net benefit, as the result of a probabilistic CEA. We focus on probabilistic rather than deterministic cost-effectiveness analyses, since the former fully accounts for all parameter uncertainty [32, 33]. A probabilistic CEA is based on the posterior distribution of the treatment effect parameters, and effectively transforms the posterior distribution over efficacy into a distribution over net benefit using the CEA model. In practice this

transformation may be coded directly in WinBUGS at the same time as the NMA is performed, or written in Excel and applied to realisations of the MCMC chains from WinBUGS; both methods provide samples from the joint posterior distribution of net benefit, upon which treatment recommendations are then made.

Firstly, we explain the impact of linearity in the CEA model: in fully linear cases, the transformation results in a posterior for net benefit which is of a known, analytically tractable form, and thresholds may be derived directly; linear models involving odds ratios also result in a net benefit posterior which is in a known form, however numerical methods must be used to evaluate some quantities in order to evaluate thresholds; when the CEA model is complex, non-linear, or even unknown / impractical to write down, a generalised numerical approach may be employed. Technical details are available on request. We then illustrate these cases in practice using examples.

4.1 EXTENDING TO CEA AND THE IMPACT OF MODEL LINEARITY

4.1.1 NET BENEFIT LINEAR IN TREATMENT EFFECT PARAMETERS

When the net benefit NB_k parameter for a treatment k can be written as a linear function of the treatment effect parameter d_k, that is

$$\mathrm{NB}_k = \alpha_k d_k + \gamma_k$$

for some parameters α_k and γ_k, then the posterior distribution of the net benefit parameters is known and we can derive thresholds algebraically using the same ideas as for efficacy decisions.

4.1.2 NET BENEFIT LINEAR IN PROBABILITY WHEN TREATMENT EFFECT IS AN ODDS RATIO

Suppose that the net benefit NB_k parameter for a treatment k can be written as a linear function of a probability p_k of which the treatment effect parameter d_k is the log odds ratio versus the reference treatment, that is

$$\mathrm{NB}_k = \alpha_k p_k + \gamma_k, \qquad d_k = \log\left(\frac{p_k(1-p_1)}{(1-p_k)p_1}\right)$$

In this case the posterior distribution of the net benefit parameters is known, however it is not analytically tractable; evaluation of the expected net benefit for a treatment requires numerical integration. We can therefore find thresholds for treatment decisions based on this form of net benefit function using numerical methods. This is not a brute-force approach requiring expensive re-evaluation of the entire NMA and CEA models – this would in many cases be so computationally expensive and time consuming as to be infeasible – but instead utilises the analytic properties of the posterior to inform a fast, efficient calculation of thresholds by finding roots of the net benefit contrasts numerically.

4.1.3 A GENERAL SOLUTION FOR COMPLEX, NON-LINEAR, OR UNKNOWN NET BENEFIT FUNCTIONS

Frequently in cost-effectiveness analysis the net benefit function is non-linear, perhaps involving several non-independent parameters, or is either analytically intractable or infeasible to write down, possibly due to the size and complexity of the model. In such scenarios we cannot take the analytic and semi-analytic approaches detailed in sections 4.1.1 and 4.1.2.

One possibility is a brute-force solution where the NMA and CEA models are re-run many times changing the data slightly each time, along the same lines of our previous work on threshold analysis (see section 1.1.2 and [24]). Although possible in theory, this may be far too computationally expensive and time consuming to be feasible in practice. If it is to be attempted, the CEA model should be run in, for example, R, calling WinBUGS to carry out the NMA. It would be much more difficult if the CEA is implemented separately from the NMA in Excel. A general solution to perform threshold analyses for treatment recommendations based on net benefit is therefore highly desirable.

We have been able to make considerable progress towards a general solution in which the analysis is carried out in two stages. The first stage uses Generalised Additive Models [34-36] or Gaussian Process Emulators [37, 38] to fit the net benefit functions. In the second stage numerical quadrature is employed to find the expected net benefit, followed by efficient numerical root-finding. This general method is an ongoing area of research.

4.2 EXAMPLES

We apply these ideas to examples of CEAs in order to derive bias-adjustment thresholds and invariant intervals. In the first example we continue with the headaches NMA from section 2.1, which has a simple linear CEA to which we apply the technique of section 4.1.1. We then return to the social anxiety example from section 2.2.3, for which a more complex CEA was performed; to this we apply the technique of section 4.1.2.

4.2.1 HEADACHES: NET BENEFIT LINEAR IN ALL TREATMENT EFFECT PARAMETERS

The headaches clinical guideline introduced as an example in section 2.1 included a probabilistic cost-effectiveness analysis, to reach a treatment recommendation based on net benefit. Not every treatment from the NMA was included in the CEA; only treatments which the guidelines committee judged to have sufficient evidence of clinical effectiveness from the NMA were included in the CEA. The resulting treatment network is shown in Figure 21. The base-case treatment recommendation is propranolol (treatment 7), with an expected (incremental) net benefit of £405 compared to placebo.

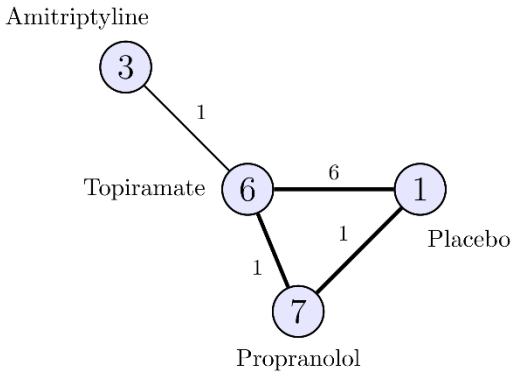


Figure 21: Treatment network for headaches example, restricted to treatments included in the CEA. Edges indicate study evidence between two treatments, and numbers on the edges show the number of studies making the comparison. Numbers inside the nodes are the treatment codings. The bold loop is formed by a single three-arm study.

The CEA model is a simple linear function of the treatment effects:

$$\text{incNB}_k = \lambda \cdot \text{incQALY}_k - \text{incCost}_k$$

which can be rearranged into the form $\text{incNB}_k = \alpha \cdot d_k + \gamma_k$. We therefore have the situation described in section 4.1.1 and derive thresholds algebraically.

Figure 22 shows the results of a contrast-level threshold analysis for the net benefit treatment recommendation. As with net benefit as the decision function, the treatment recommendation is highly sensitive to small changes in the data; indeed, the cost effectiveness analysis is even more sensitive to bias than the treatment efficacy recommendation. In the worst case, a positive bias in the relative number of headache days per month on topiramate vs. amitriptyline (6 vs. 3) of 0.01 results in amitriptyline being recommended on the basis of net benefit, equivalent to an increase in headache duration of just 20 minutes per month on topiramate compared to amitriptyline. The bias thresholds for the combined evidence on the other three contrasts are of similar magnitude: –0.05 for topiramate vs. placebo (6 vs. 1), 0.04 for propranolol vs. placebo (7 vs. 1), and 0.02 for propranolol vs. topiramate (7 vs. 6). At each of these thresholds the new treatment with greatest net benefit is amitriptyline (3). Each of these thresholds lies within the 95% CrI for

Contrast	Posterior Mean	95% Credible Interval	Bias Invariant Interval with $\tilde{k}^*$		
†**6 vs. 1**	–1.04	(–1.51, –0.58)	3	(–1.08, 0.81)	1
†**7 vs. 1**	–1.19	(–2.20, –0.20)	–	(NT, –1.14)	3
†**6 vs. 3**	0.10	(–1.13, 1.32)	–	(NT, 0.11)	3
†**7 vs. 6**	–0.15	(–1.11, 0.82)	–	(NT, –0.13)	3

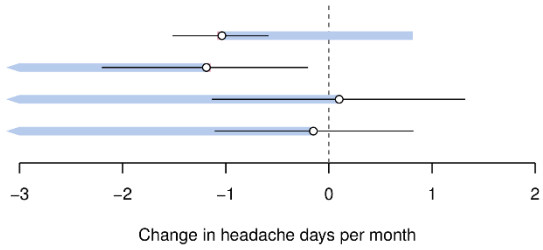


Figure 22: Forest plot showing the results of the contrast-level net benefit threshold analysis, based on the reduced network. Invariant intervals are shaded, and the new recommended treatment is shown at either end of the interval in column 4. All contrasts have thresholds that lie within the 95% CrI, and are smaller than the minimally important difference (0.5 days). The base-case recommended treatment is 7 (propranolol).

the contrast and is smaller than the minimally important difference of 0.5 days; as such the net benefit treatment recommendation is highly sensitive to bias adjustments. This is perhaps not surprising, as propranolol has only 47% probability of being most cost-effective, alluding to the uncertainty present in the base-case result.

A study-level threshold analysis, shown in Figure 23, tells the same story: the treatment recommendation based on efficacy is highly sensitive to bias adjustments and to the level of imprecision in every piece of evidence. In this case, such sensitivity likely arises due to the small number of studies.

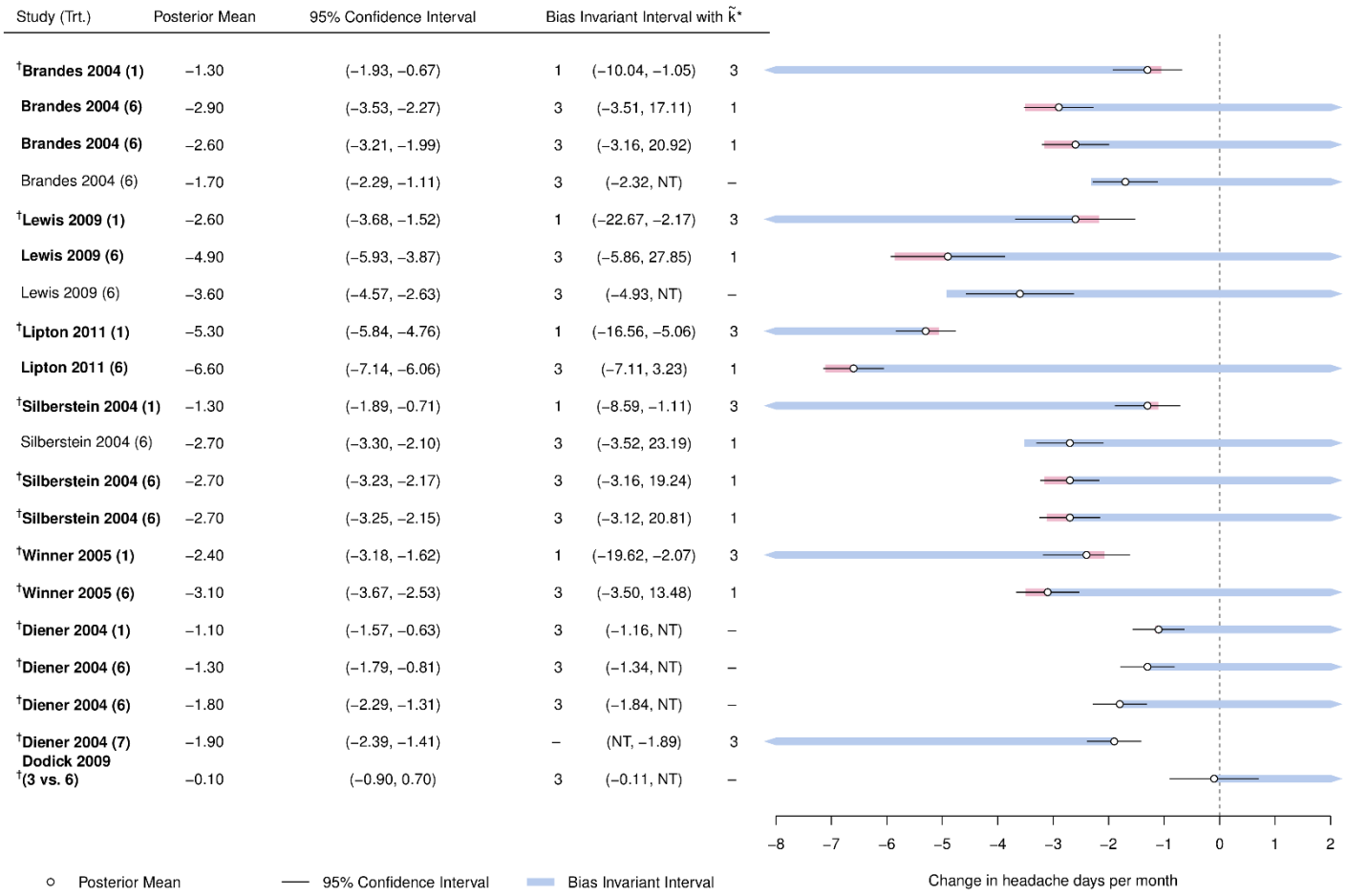


Figure 23: Forest plot showing the results of the study-level net benefit threshold analysis, based on the reduced network. Invariant intervals are shaded, and the new recommended treatment is shown at either end of the interval in column 4. Bold text indicates contrasts where bias thresholds lie inside the 95% CI. The base-case recommended treatment is 7 (propranolol).

† indicates contrasts with bias thresholds less than the minimally important difference (0.5 days).

NT = no threshold; no amount of bias adjustment in this direction will change the treatment recommendation.

4.2.2 SOCIAL ANXIETY: 1 YEAR DECISION TREE WITH LOG ODDS RATIO TREATMENT EFFECT

The social anxiety clinical guideline presented as an example in section 2.2.3 includes a probabilistic cost-effectiveness analysis of 28 treatments out of the full network; to arrive at a final treatment recommendation based on net benefit. Thirteen treatments were excluded from the CEA; for example if there was insufficient evidence on the

efficacy of a treatment, or if a treatment was not available in the UK. The base-case recommendation from the 1-year decision tree CEA was treatment 23 (phenelzine).

The CEA model has two stages: a decision tree for the first year after treatment, using probabilities of recovery derived from the SMD treatment effects via a conversion to log odds ratios, after which a two-state Markov model was used for four further years of follow-up. The guideline makes its final recommendations based upon the full 5-year model; however we shall consider only the first stage of the model here. Considering just the decision tree results in a net benefit function that is a linear function of the recovery probability, obtained by linearly transforming the original treatment effects, which are SMDs, into LORs. As such we have the situation described in section 4.1.2.

Figure 24 shows the results of the contrast-level threshold analysis. All of the invariant intervals are wide; the smallest threshold is for the 8 vs. 5 contrast, where a positive bias adjustment of 1.57 to the SMD in favour of treatment 5 (book self-help, no support) results in treatment 5 being recommended. Notice that at the negative threshold for this contrast treatment 5 again becomes optimal – this highlights the non-linear nature of these net benefit functions, which may not even be monotonic in the treatment parameters. Several contrasts had neither negative nor positive thresholds, and made no contribution to the net benefit decision; this occurs because treatments have been excluded from the CEA.

The results of the study-level threshold analysis are shown in Figure 25. Once again, all of the invariant intervals are wide. At the study level, the smallest threshold is for the 12 vs. 2 estimate of study 18, where a negative adjustment of –1.75 SMD in favour of treatment 12 will result in treatment 12 (mirtazapine) being recommended.

Altogether, the contrast- and study-level threshold analyses support the robustness of a treatment recommendation made based on net benefit from the 1-year decision tree CEA. There are no individual studies or combined evidence on contrasts to which the decision is particularly sensitive; this should defend the recommendation from concerns raised about individual studies or combined contrast evidence. Although not presented here, it is entirely possible for the extended analyses described in sections 3.1–3.4 to be considered, with no greater difficulty than before. Indeed, the numerical methods used in this section need only be performed once, as the threshold calculations can then be reused for the extended analyses.

Contrast	Posterior Mean	95% Credible Interval	Bias Invariant Interval with $\tilde{k}^*$		
5 vs. 1	−0.84	(−1.08, −0.60)	5	(−3.15, 7.99)	5
6 vs. 1	−0.66	(−0.94, −0.39)	6	(−9.17, NT)	–
8 vs. 1	−0.88	(−1.04, −0.71)	5	(−2.58, 5.66)	5
14 vs. 1	−0.83	(−1.27, −0.39)	14	(−8.51, 7.60)	14
24 vs. 1	−0.83	(−1.07, −0.59)	5	(−17.10, NT)	–
28 vs. 1	−0.43	(−0.83, −0.04)	27	(−20.21, NT)	–
30 vs. 1	−0.85	(−1.04, −0.67)	5	(−6.42, 9.85)	16
31 vs. 1	−0.80	(−1.02, −0.58)	18	(−13.09, 15.64)	18
33 vs. 1	−0.97	(−1.21, −0.74)	27	(−16.53, NT)	–
35 vs. 1	−1.19	(−1.48, −0.89)	27	(−11.79, 17.33)	18
36 vs. 1	−1.56	(−1.86, −1.27)	16	(−17.11, 16.63)	16
12 vs. 2	−0.33	(−0.93, 0.28)	12	(−4.27, 3.63)	12
13 vs. 2	−0.45	(−0.65, −0.25)	13	(−2.17, 1.64)	13
14 vs. 2	−0.36	(−0.76, 0.05)	14	(−7.20, 7.15)	14
15 vs. 2	−0.41	(−0.63, −0.19)	15	(−3.31, 2.99)	15
16 vs. 2	−0.40	(−0.65, −0.14)	16	(−5.18, 5.19)	16
17 vs. 2	−0.47	(−0.68, −0.26)	17	(−4.06, 3.34)	17
18 vs. 2	−0.53	(−0.65, −0.40)	18	(−2.34, 1.90)	18
19 vs. 2	−0.49	(−0.67, −0.32)	19	(−3.42, 2.86)	19
21 vs. 2	−0.60	(−0.95, −0.25)	–	(NT, 19.14)	18
22 vs. 2	−0.27	(−0.48, −0.06)	12	(−13.06, 9.81)	18
23 vs. 2	−0.81	(−1.05, −0.56)	24	(−9.38, 11.46)	18
30 vs. 2	−0.38	(−0.62, −0.16)	18	(−17.25, 12.22)	18
31 vs. 2	−0.33	(−0.57, −0.10)	18	(−14.95, 10.60)	18
29 vs. 3	0.01	(−0.36, 0.38)	30	(−17.45, NT)	–
7 vs. 5	−0.01	(−0.34, 0.33)	7	(−3.44, 4.55)	5
8 vs. 5	−0.03	(−0.27, 0.20)	5	(−6.05, 1.54)	5
30 vs. 5	−0.01	(−0.27, 0.25)	5	(−7.57, 1.96)	5
8 vs. 6	−0.21	(−0.46, 0.04)	8	(−18.57, 2.19)	6
8 vs. 7	−0.03	(−0.34, 0.28)	–	(NT, 1.80)	7
31 vs. 8	0.07	(−0.17, 0.32)	18	(−12.78, 13.91)	5
30 vs. 14	−0.02	(−0.47, 0.41)	14	(−12.55, 11.38)	14
18 vs. 15	−0.12	(−0.35, 0.11)	15	(−5.74, 4.67)	15
30 vs. 16	0.01	(−0.27, 0.29)	16	(−6.60, 5.67)	16
36 vs. 16	−0.69	(−1.04, −0.35)	16	(−10.61, 7.78)	16
37 vs. 16	−0.09	(−0.48, 0.30)	–	(NT, 18.21)	16
19 vs. 18	0.03	(−0.16, 0.23)	19	(−9.48, 10.94)	19
35 vs. 22	−0.45	(−0.77, −0.13)	–	(NT, 16.31)	18
31 vs. 23	0.47	(0.21, 0.74)	12	(−18.59, 9.80)	24
30 vs. 24	−0.02	(−0.27, 0.22)	5	(−12.07, NT)	–
28 vs. 26	−0.17	(−0.60, 0.25)	–	(NT, 4.72)	27
35 vs. 26	−0.93	(−1.39, −0.47)	–	(NT, 4.70)	27
30 vs. 27	−0.46	(−0.88, −0.04)	–	(NT, 19.27)	27
31 vs. 27	−0.41	(−0.83, 0.01)	–	(NT, 15.20)	27
33 vs. 28	−0.54	(−0.95, −0.13)	–	(NT, 8.79)	27
33 vs. 29	−0.35	(−0.68, −0.03)	–	(NT, 7.90)	30
33 vs. 30	−0.12	(−0.37, 0.13)	–	(NT, 11.47)	5

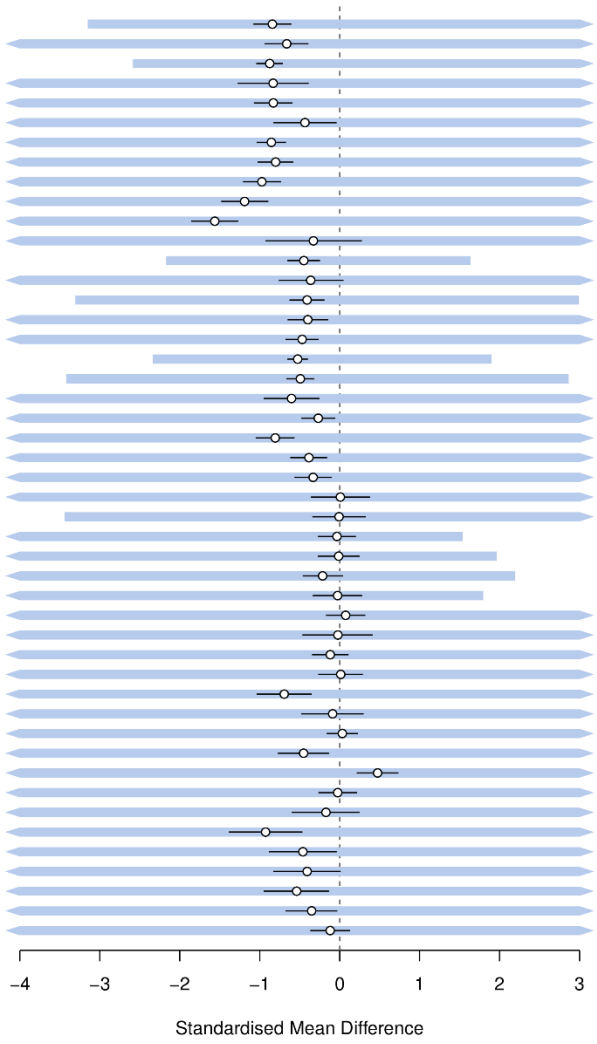


Figure 24: Contrast-level forest plot for the Social Anxiety example showing results of the threshold analysis for net benefit. Invariant intervals are plotted as thick shaded lines, and in the table are shown with the new recommended treatments at the thresholds at either side of the interval in column 4. Only contrasts that have at least one threshold (i.e. affect the recommendation) are shown. The base-case recommended treatment based on net benefit is treatment 23 (phenelzine).

NT = no threshold; no amount of bias adjustment in this direction will change the treatment recommendation.

Study (Contrast)	SMD	95% Confidence Interval	Bias Invariant Interval with $\tilde{k}^*$		
2 (7 vs 1)	−0.96	(−1.72, −0.19)	7	(−7.12, NT)	–
3 (8 vs 1)	−0.77	(−1.01, −0.52)	5	(−9.90, NT)	–
6 (8 vs 1)	−0.94	(−1.31, −0.57)	5	(−10.66, NT)	–
18 (12 vs 2)	−0.33	(−0.79, 0.13)	12	(−2.09, 1.43)	12
19 (13 vs 2)	−0.37	(−0.54, −0.20)	13	(−6.52, 7.09)	13
20 (13 vs 2)	−0.91	(−1.16, −0.66)	13	(−8.50, 8.30)	13
21 (14 vs 2)	−0.37	(−1.06, 0.31)	14	(−6.94, 6.84)	14
22 (15 vs 2)	−0.25	(−0.44, −0.06)	15	(−6.42, 6.99)	15
46 (8 vs 6)	−0.28	(−0.56, 0.00)	–	(NT, 5.84)	6
57 (28 vs 26)	−0.08	(−0.56, 0.39)	–	(NT, 5.50)	27
58 (35 vs 26)	−1.15	(−1.66, −0.65)	–	(NT, 3.05)	27
66 (5 vs 1)	−0.53	(−1.24, 0.19)	5	(−5.63, 10.90)	7
66 (7 vs 1)	−0.58	(−1.27, 0.10)	7	(−4.73, NT)	–
67 (5 vs 1)	−0.67	(−1.05, −0.28)	5	(−3.19, 8.99)	5
67 (8 vs 1)	−0.63	(−1.01, −0.24)	–	(NT, 7.28)	5
68 (6 vs 1)	−0.42	(−0.86, 0.03)	6	(−6.27, NT)	–
69 (14 vs 1)	−0.93	(−1.95, 0.09)	14	(−7.94, 6.77)	14
73 (28 vs 1)	−0.44	(−0.84, −0.04)	27	(−4.95, NT)	–
74 (29 vs 1)	−0.87	(−1.14, −0.59)	30	(−7.85, NT)	–
81 (16 vs 2)	−0.13	(−0.65, 0.39)	16	(−8.70, 9.89)	16
92 (5 vs 1)	−0.40	(−0.74, −0.05)	5	(−2.37, 7.14)	5
92 (30 vs 1)	−0.52	(−0.86, −0.18)	–	(NT, 6.88)	5
92 (30 vs 1)	−0.69	(−1.03, −0.34)	–	(NT, 7.86)	5
93 (2 vs 2)	−0.13	(−0.37, 0.12)	13	(−7.95, 6.33)	13
93 (13 vs 2)	−0.30	(−0.54, −0.05)	13	(−7.20, 8.08)	13
93 (13 vs 2)	−0.40	(−0.64, −0.15)	13	(−7.39, 8.08)	13
98 (7 vs 5)	0.14	(−0.32, 0.59)	7	(−2.46, NT)	–
98 (8 vs 5)	0.08	(−0.37, 0.53)	–	(NT, 6.60)	7
98 (8 vs 5)	0.41	(−0.04, 0.86)	–	(NT, 6.74)	7
99 (15 vs 2)	−0.30	(−0.50, −0.11)	15	(−8.19, 8.95)	15
99 (15 vs 2)	−0.57	(−0.77, −0.37)	15	(−8.65, 8.91)	15
99 (15 vs 2)	−0.39	(−0.58, −0.19)	15	(−8.53, 9.17)	15
99 (18 vs 2)	−0.45	(−0.65, −0.25)	15	(−7.00, 5.13)	15
100 (16 vs 2)	−0.67	(−0.99, −0.35)	16	(−4.39, 3.68)	16
100 (30 vs 2)	−0.50	(−0.81, −0.19)	16	(−9.51, 7.21)	16
100 (30 vs 2)	−0.61	(−0.93, −0.30)	16	(−10.10, 7.50)	16

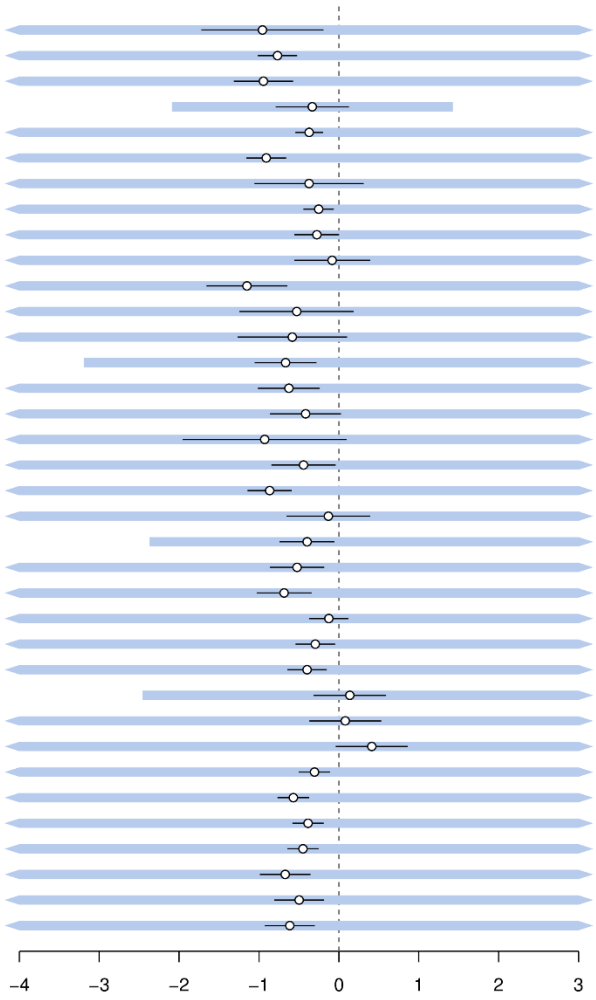


Figure 25: Study-level forest plot for the Social Anxiety example showing results of the threshold analysis for net benefit. Invariant intervals are plotted as thick shaded lines, and in the table are shown with the new recommended treatments at the thresholds at either side of the interval in column 4. Only contrasts with a threshold smaller than 10 SMD are shown here for brevity. The base-case recommended treatment based on net benefit is 23 (phenelzine).

NT = no threshold; no amount of bias adjustment in this direction will change the treatment recommendation.

5. PLAUSIBLE MAGNITUDE OF BIASES

The threshold analysis is designed simply to calculate whether changes in the data, of a size and direction that might reasonably be attributed to bias or error, would result in changes in treatment recommendation, whether based on clinical efficacy alone or cost-effectiveness. The question of how large a bias is "reasonable" is a totally separate question, although of central importance if the threshold analysis is to be clinically interpretable, and especially if it leads to an updated NMA model to include bias modelling and adjustment, which might then lead to changes in the recommendation (see Section 6.2). In this next section we set out what is known about the sizes of biases.

5.1 INTERNAL AND EXTERNAL BIAS

A key reference giving a very clear account of internal and external bias is by Turner [39]. It is important to emphasise that the "bias" which is of concern to the meta-analysis, is bias relative to the target parameter of interest.

To study internal bias in a trial report, we consider whether the way the trial was conducted and analysed could introduce bias in its estimate of the treatment effect in its chosen target population, whatever that might be. Among the potential causes of internal bias are: inadequate randomization; failure to conceal randomization, which might lead to allocation bias and confounding; lack of blinding, which could affect outcome assessment (performance bias); missing data, which could be different in each arm (attrition bias) and lack of complete outcome reporting. Thus, internal biases are considered with respect to the trial's target parameter, which may not be the target parameter for the NMA.

External bias reflects whether the target parameter for the trial differs from the target parameter for the evidence synthesis. External biases may have their origins in factors such as: a difference between the trial population and the target population; differences between the treatments in the trial – either active treatment or the control – and the treatment of interest in the NMA; differences in outcome definition, including different follow-up times.

In the conception of bias-adjustment discussed by Turner [39], a trial's results would be first adjusted for internal bias, and then if there were issues about the trial's target population, also for external bias.

5.2 META-EPIDEMIOLOGICAL FINDINGS OF INTERNAL BIAS

There have been several attempt to study internal biases quantitatively, but the databases of several of the earlier studies were combined in a large and systematic exercise under the BRANDO initiative (Bias in Randomised and Observational studies) [40, 41], and we outline the key findings from this study here.

1. Lack of allocation concealment and lack of blinding were both independently associated with exaggerated treatment effects of active treatments relative to control.
2. The extent of this bias was highest with subjective outcomes (patient- and clinician-reported outcomes) and lowest with mortality. Objective outcomes such as blood pressure or cholesterol measurements were intermediate but closer to mortality.

3. The extent of bias in trials with lack of allocation concealment or lack of blinding varied across trials, subjective outcomes were the most variable and mortality the least. There was variation both between trials within meta-analyses, and between meta-analyses.

Table 2 gives the main results. With subjective outcomes, inadequate or unclear allocation concealment decrease the risk of a negative outcome by an odds ratio of 0.85, lack of double blinding or unclear double blinding by 0.82, and both together by 0.70.

Allocation concealment has virtually no impact on objective outcomes, but lack of double blinding reduces the estimated risk of negative outcomes by about 0.9.

Table 2: Estimated relative odds ratios and 95% credible intervals for the influence of reported study design characteristics on average intervention effects. Reproduced from [41] (Table 18, p.30).

Model, study design characteristic and outcome	**ROR**	**95% CrI**
Inadequate or unclear allocation concealment (vs adequate)		
All	0.93	0.87 to 1.00
Mortality	1.00	0.89 to 1.13
Objective	0.97	0.84 to 1.13
Subjective	0.85	0.76 to 0.96
Lack of double blinding or unclear double blinding (vs double blind)		
All	0.88	0.79 to 0.97
Mortality	0.92	0.80 to 1.06
Objective	0.90	0.71 to 1.15
Subjective	0.82	0.68 to 0.96
Implied average bias in trials with high risk of bias for both characteristics		
All	0.83	0.74 to 0.92
Mortality	0.92	0.78 to 1.09
Objective	0.87	0.68 to 1.12
Subjective	0.70	0.57 to 0.84

5.3 POSSIBLE EXTENT OF EXTERNAL BIAS

Much less work has been done on quantifying external biases, which by definition are highly dependent on the context. The degree of heterogeneity typically found in meta-analyses might give some kind of guide, and there are publications providing quantitative estimates from meta-analyses undertaken by the Cochrane Collaboration [42], but these estimates include variation due to internal biases as well as external biases, and so they would represent very extreme upper bounds. In addition, the sets of trials that are used to generate NMAs in the context of guideline development at NICE generally have a far more tightly defined PICO specification – particularly regarding the Population, Intervention and Comparator aspects – than evidence syntheses published by the Cochrane Collaboration. The fact that NMAs allow different doses and different members of a class of treatments to be considered as *separate* treatments, rather than being "lumped" [43] also greatly reduces the potential for external biases in NICE CGs. In considering external biases, it is essential that guideline developers observe the advice on NMA analyses, which stresses that the initial review process should include careful consideration of known effect modifiers, which may have been examined in the analyses of individual trials.

Nevertheless, external biases may exist. Arguments that specific trials or contrasts are more likely than not to be biased, *in a specific direction*, must be based on the specifics of the trial and the target population, backed up by reasoned argument or evidence.

6. ROLE OF THRESHOLD ANALYSIS IN GUIDELINE DEVELOPMENT

This section is restricted to a consideration of how, in general terms, threshold analysis could enter into guideline development. We begin by making the fundamental observation that threshold analysis, like all forms of sensitivity analysis, is not undertaken with a view to changing base-case treatment recommendations. The purpose of sensitivity analysis in general is to find out exactly what is "driving" results. This can be seen as part of a general "reality check" of the model, as well as providing an opportunity to review the model and the way treatment recommendations are derived from it in a fresh way, throwing particular emphasis on sensitivity to possible biases in the trial evidence. The threshold analyses proposed here were originally developed as a more coherent way of asking questions about the reliability of conclusions drawn from NMA, and particularly questions about the reliability of the data inputs. Whatever the result of the threshold analysis, there is no implication that the base-case recommendation should necessarily be changed.

Above all, threshold analysis must not be seen as an opportunity to change the treatment recommendation to conform to any initial preconceptions or preferences on the part of those making a treatment recommendation.

In this section, we firstly outline a series of ways in which threshold analysis could contribute to guideline development. We then make some suggestions regarding the conditions under which threshold analysis can lead to a change in the base-case treatment recommendation. Recognizing the overriding need for transparency in guideline development, we place particular emphasis on how the results should be reported.

6.1 USES OF THRESHOLD ANALYSIS

6.1.1 REACTIVE USE

One of the most common criticisms of a draft guideline at consultation, or after publication, and particularly of the NMA analyses, relates to the inclusion of specific trials or sets of trials. A threshold analysis concerning individual trials, or sets of trials as exemplified in Section 3, can address this directly. Furthermore, threshold analysis may be used in reaction to more complex criticisms. For example, in sections 3.1–3.3 we demonstrated that threshold analysis could be used to allay concerns over the impact of possible generic pharmacological or psychological treatment biases, and in section 3.4 we used threshold analysis to establish the robustness of the treatment recommendation to possible exaggeration of effect in studies by a particular investigator.

6.1.2 PRE-EMPTIVE USE

In a similar vein, GDGs collectively should be aware of any vulnerability that their recommendations might have to criticism, and should be able to anticipate the kinds of criticism that are likely to be levelled at the trial evidence during and after consultation. Ideally, if these criticisms relate to inclusion of one or more trials, or to generic biases in sets of trials, they could be pre-empted by threshold analysis. Or, if a degree of sensitivity is detected, it can be made exactly clear how large biases could be without impacting on a decision.

6.1.3 PHRASING THE RECOMMENDATIONS

In cases where the base-case recommendation A would be over-turned in favour of a second treatment B following slight changes to the data, a GDG might consider phrasing their recommendation in terms such as "Choose A as a first option, but switch to B if patients do not do well on A". Typically it would already be obvious from the original analysis that the NMA failed to clearly establish superiority of A over B and would show B as the second best option.

6.1.4 THRESHOLD ANALYSIS AS A STIMULUS FOR A NEW NMA

It is possible that a threshold analysis might lead guideline developers to reconsider the possibility of bias in one or more data elements. One possibility, rather than issue recommendations based on the original NMA and the threshold analysis, would be to run a new NMA in which the biases were deliberately modelled and adjusted for in the NMA.

6.1.5 RESTRICTING THE NEED FOR APPLICATION OF GRADE TO SOURCE DATA

A different kind of application of threshold analysis would be in limiting the need for a GRADE analysis of all the trial data; in-depth GRADE assessments may be confined to the studies or contrasts in which threshold analysis reveals a sensitivity to bias adjustment. It is worth noting that with a small evidence base, the base-case decision is more likely to be vulnerable to relatively small changes in the data, so there would be little saving. However, as the evidence base is increased, the robustness of the decision to small changes in evidence sources becomes more and more apparent, and the potential saving in time becomes very substantial. In addition, while GRADE is usually performed separately

for different outcomes and outcomes at different times, the more these can be synthesized together in a single model, the more robust the recommendations will be, leading to still more impressive saving of time. This is illustrated in Table 3. Here, for illustrative purposes we have adopted a set of quite conservative estimates of how the time required for GRADE analyses might be reduced, based upon criteria for the size of plausible bias. The criteria used here are: ±2.5 MCIDs; ±0.5 SMDs, which is described by Cohen [30] as a "moderate" effect, and ±0.9 on a log OR, log relative risk, or log hazard ratio scale, which is equivalent to a SMD of ±0.5. These are only suggested criteria used for illustration; guideline developers should make an informed judgement on the size of plausible bias for their specific treatment scenario.

Taking advantage of this possibility would obviously entail only undertaking the GRADE exercise *after* the NMA and threshold analyses were completed.

Table 3: Reductions in the number of GRADE analyses required after a threshold analysis, based on conservative criteria for the magnitude of plausible bias. Sensitive thresholds [a]smaller than ±2.5 MCID; [b]smaller than ±0.9 LOR; [c]smaller than ±0.5 SMD.

Analysis	Number of studies	Number sensitive	Proportional reduction in GRADE analyses required after threshold analysis
Headaches (Efficacy)[a]	11	9	0.18
Headaches (CEA)[a]	11	8	0.27
Tocolytics[a]	51	0	1.00
Urinary Incontinence[b]	22	10	0.55
Social Anxiety (Efficacy)[c]	100	2	0.98
Social Anxiety (CEA)[c]	100	0	1.00

6.2 CHANGING THE BASE-CASE RECOMMENDATION

If concerns have been raised about the presence of bias to which a threshold analysis reveals the base-case treatment recommendation would be sensitive to, there are two possible courses of action. Firstly, the GDG could decide, based on the balance of evidence that such a bias exists, to change the treatment recommendation to the alternative given by the threshold analysis. Alternatively, an updated NMA can be performed which models and accounts for the suggested bias, and decisions are then made based on the updated analysis. The latter option is strongly preferred on statistical grounds, as the former is equivalent to simply changing the original data and re-running the NMA, whereas the bias modelling approach accounts for the (substantial) uncertainties in the magnitude of bias. Furthermore, the new treatment recommendations at the thresholds are only valid at that precise value of bias adjustment; the recommendation may well change again beyond the threshold value, but such changes are not presented in this analysis. Despite these shortcomings, and against our recommendation that changes to treatment decisions should be

based on a new NMA, we lay out guidelines for changing the base-case recommendation based on the threshold analysis alone. As noted above, a change in base-case treatment recommendation following a threshold analysis would require the most careful consideration, and the rationale for such a step would have to be fully documented (see section 6.3).

The threshold analysis, in effect, allows the decision makers to isolate a set of studies or contrasts in which a plausible change to the data would change the recommendation. However, establishing that the change is "plausible" is not in itself a sufficient reason, even if a small change within the level of uncertainty of the data would lead to a change in decision – in such a scenario it is likely that the GDG would give a compound recommendation anyway due to the lack of sufficient evidence for recommending one treatment over another exclusively. In order to establish the case for changing the base-case recommendation, it would be necessary for the GDG to agree that not only was a bias of such a size and direction "plausible", but, on the balance of the evidence, the GDG's opinion was that the data element in question was in fact biased by that amount, or more.

To carry this argument through it would be necessary to produce and document a convincing argument that the data involved were indeed biased. This could be based on one or both of two types of consideration:

First, the GDG could appeal to the kind of meta-epidemiological data presented in Section 5, to the effect that bias of such a size or more was more likely than not. However, if this was a type of bias that would be expected to affect other data elements in the study (e.g. novelty bias for newer treatments), then a further threshold analysis for this specific type of bias (see sections 3 and 6.1.1), or an updated NMA with such bias modelled and adjusted for (see section 6.1.4), would need to be considered.

A second form of argument that would refer to specific features the particular study or studies in question, such as their inclusion criteria, outcome assessment, or missing data. Critical to a claim along these lines is that it must be impartially applied. If a certain study is believed to be vulnerable to a specific form of bias, all studies of the same type should be considered vulnerable to the same bias – leading perhaps to a generic bias adjustment. Similarly, one might attribute bias to a particular feature of a trial population or setting. But again, any adjustment would have to be applied to all trials with that kind of population or setting, leading perhaps to a generic adjustment by meta-regression.

6.3 REPORTING THRESHOLD ANALYSES

However threshold analysis is used in guideline development, it needs to be fully reported. The plots used in this report could simply be included in the appendices of the full guideline document, and conclusions to be drawn from the analysis could be briefly noted. The existence of the threshold analyses and any key findings and conclusions would be reported in the main text. By the same token, unless or until threshold analysis becomes routine in guideline development, the reasons why it was undertaken should be discussed in advance and documented in the full guideline document.

To the extent that threshold analysis is simply a form of sensitivity analysis, the use of threshold analysis in guideline development, and documentation of findings, should follow the same principles as for any other form of sensitivity analysis. Every use of the threshold analysis, whether it results in no change to recommendation, to a new NMA, or to amended recommendation must be fully reported in order to ensure transparency.

7. FURTHER RESEARCH

There is still a great deal of research to be done on threshold analysis for NMA. At the technical end there is a need to develop rapid and reliable methods for applying threshold analysis to any CEA, however complex the net benefit function, and indeed for the cases where the net benefit function cannot be readily written down (see section 4.1.3). At the less technical end there is also a need for a better understanding of how to present this powerful method to GDGs, and how it is best used in guideline development. To some extent this might be "packaged" within a more general framework for sensitivity analysis.

A. TECHNICAL APPENDIX

A.1 OUTPUTTING POSTERIOR SUMMARIES AND CORRELATIONS FROM WINBUGS

Threshold analysis is based upon the joint posterior distribution of the treatment effect parameters, and requires that this is sufficiently specified, both in terms of mean and variance of each parameter, as well as the correlation/covariance between parameters. There are two options for obtaining these from WinBUGS:

1. Saving the means, variances, and correlations directly from WinBUGS and then reading these into R, or
2. Importing the CODA to R and calculating the posterior there.

In our experience, the second option is preferable, requiring less effort and with greater ease of reproducibility. To export the CODA once the model has been run, select "CODA" from the Inference > Samples menu. Save the resulting output as .txt files, which may then be read into R using the `coda` package [44]. Means, medians, quantiles, and other statistics may be calculated using the `summary` function. The covariance matrix is calculated using `cov`. Since reading the CODA and calculating the required quantities may take a few minutes, we recommend saving the computed outputs to an R data file using `save`; any subsequent analyses using the posterior need only read this data file using `load`, which is instant. Whichever of the two methods is chosen, we recommend that long MCMC samples are used to reduce the Monte Carlo Error, in order that the covariance structure which is critical to the derivation of thresholds may be distinguished from residual MCMC correlations; Monte Carlo Errors on the order of at least 10^{-3} or smaller are advised.

A.2 R FUNCTIONS FOR PERFORMING AND PRESENTING THE RESULTS OF THRESHOLD ANALYSES

We have developed R functions to perform threshold analysis and to present the results; these are available in a separate attachment to this report. We describe these functions and their use in the following appendix A.3.

A.3 PERFORMING THRESHOLD ANALYSIS IN R

A.3.1 STUDY LEVEL ANALYSIS: FE MODELS

To perform a threshold analysis on fixed effects models, we provide the function `nma_thresh`. The syntax is:

```
nma_thresh(
      X,       # = Design matrix.
      mean.dk, # = Posterior means of basic treatment parameters d_k
      lhood,   # = Likelihood (data) covariance matrix.
      post,    # = Posterior covariance matrix of all parameters (including
               #   nuisance study means if study level).
      opt.max, # = Should the optimal decision be the maximal treatment effect
               #   (TRUE, default) or the minimum (FALSE).
      trt.rank,# = Rank of the treatment to derive thresholds for. Defaults to 1,
               #   thresholds for the optimum treatment.
      trt.code,# = Treatment codings of the reference trt and in the parameter
               #   vector d_k. Use if some are missing (perhaps excluded) or
               #   re-ordered. Default is equivalent to 1:K.
      trt.sub, # = Only look at thresholds in this subset of treatments in trt.code,
               #   e.g. if some are excluded from the ranking. Default is equivalent
               #   to 1:K.
      verbose  # = Print intermediate matrices? Default is FALSE.
)
```

The four main inputs are the design matrix X, the posterior means of the treatment effect parameters $\mathbb{E}(\boldsymbol{d})$, and the likelihood and posterior covariance matrices V and Σ_n. The result is a data frame containing the negative and positive threshold values for each data point, along with the new treatment recommendations at the thresholds. This function works for both relative and absolute effects data, and models with extra parameters, so long as the treatment effect parameters are the first $K-1$ parameters in the covariance matrix.

A.3.2 STUDY LEVEL ANALYSIS: RE MODELS

To perform a threshold analysis on random effects models, we provide the function `nma_threshRE`. The syntax is similar to the function `nma_thresh`, and also outputs a data frame with the positive and negative thresholds and the new treatment recommendations at each.

```
nma_threshRE(
      X,       # = Design matrix.
      mean.dk, # = Posterior means of basic treatment parameters d_k
      lhood,   # = Likelihood (data) covariance matrix.
      post,    # = Posterior covariance matrix of all parameters (including
               #   nuisance study means if study level).
      mu.design,    # = Design matrix for any extra parameters. Defaults to NULL (no
                    #   extra parameters).
      delta.design, # = Design matrix for random effects delta, defaults to the NxN
                    #   identity matrix (random effect for every data point).
      opt.max, # = Should the optimal decision be the maximal treatment effect
               #   (TRUE, default) or the minimum (FALSE).
      trt.rank,# = Rank of the treatment to derive thresholds for. Defaults to 1,
               #   thresholds for the optimum treatment.
      trt.code,# = Treatment codings of the reference trt and in the parameter
               #   vector d_k. Use if some are missing (perhaps excluded) or
               #   re-ordered. Default is equivalent to 1:K.
      trt.sub, # = Only look at thresholds in this subset of treatments in trt.code,
               #   e.g. if some are excluded from the ranking. Default is equivalent
               #   to 1:K.
      verbose  # = Print intermediate matrices? Default is FALSE.
)
```

There are two additional parameters over the function `nma_thresh`: `mu.design` and `delta.design`. The first is for the matrix M in [27], which if specified is the design matrix for any additional parameters in the model (for example study-level baseline parameters for absolute effects data). The second is for the matrix L in [27], and is a design matrix for the random effects parameters; by default this is the identity matrix so that all data points have a random effect, but changing this allows specification of other models for example including absolute effects data where only the non-reference arms have random effects.

A.3.3 CONTRAST LEVEL ANALYSIS: ALL MODELS WITH NORMAL POSTERIOR

Contrast-level threshold analysis may be applied to any model, regardless of data type(s), complexity, hierarchical nature, or even whether the study-level data is known, as long at the posterior distribution of the treatment effect parameters is sufficiently specified and may be assumed (approximately) multivariate normal. We provide R functions which make contrast-level threshold analysis straightforward. Firstly, the hypothetical likelihood covariance matrix must be reconstructed (see [27]). The function `recon_vcov` achieves this using non-negative least squares, and returns the approximated covariance matrix. The syntax is:

```
recon_vcov(
      post,       # = Posterior covariance matrix of the treatment effect
                  #   parameters.
      prior.prec, # = Prior precision. Defaults to .0001 which is a common flat
                  #   prior for NMA. Not used if prior.vcov is specified.
      prior.vcov, # = Prior covariance matrix. Defaults to a diagonal matrix of the
                  #   same size as post, with elements 1/prior.prec
      X,          # = Contrast design matrix. If omitted a complete network is
                  #   assumed.
      verbose     # = Print intermediate matrices? Default is FALSE.
)
```

The parameters `prior.prec` and `prior.vcov` are mutually exclusive; if `prior.vcov` is specified it overrides the value of `prior.prec`. A single value for `prior.prec` may be given, specifying the prior precision of the treatment effect parameters. The parameter `prior.vcov` takes a full prior covariance matrix for the treatment effect parameters, and may be used to specify differing (co)variances for the treatment parameters. The design matrix X is for the hypothetical independent data points representing the combined direct evidence on each contrast for which there is study data (i.e. all edges in the treatment network); it is particularly important that care is taken to specify this correctly, in order that the approximation fits the correct underlying covariance structure. A common mistake is to not specify all the contrasts on which a multi-arm trial provides direct evidence; for example, a three-arm trial on A, B, and C provides direct evidence on *three* contrasts: AB, AC, and BC, all of which should be included in the design matrix. In our experience, cross-checking with the network diagrams at this stage is valuable.

Once the hypothetical likelihood covariance matrix has been reconstructed with `recon_vcov`, analysis proceeds under the fixed effects model using the function `nma_thresh`.

A.3.4 PRESENTING THE RESULTS OF THRESHOLD ANALYSIS

To present the results of threshold analyses clearly and succinctly, we provide a function thresh_forest which outputs forest plots as seen throughout this report. The main function inputs are the results of nma_thresh or nma_threshRE, along with study or contrast estimates and confidence/credible intervals. Other parameters provide easy customisation of the output figure. The syntax is:

```
thresh_forest(
      means,        # = Study estimates or posterior means of contrasts
      t.lo,         # = Negative threshold values
      t.hi,         # = Positive threshold values
      kstar.lo,     # = New k* optimal treatments for negative threshold
      kstar.hi,     # = New k* optimal treatments for positive threshold
      CI.lo,        # = Lower values of CIs (optional)
      CI.hi,        # = Upper values of CIs (optional)
      CI.title,     # = Title for CI column, default "95% Credible Interval"
      label,        # = Row labels
      label.title,  # = Column title for labels (e.g. "Contrast" or "Study")
      mean.title,   # = Column title for means, default "Posterior Mean"
      xlab,         # = Label for x-axis
      xlim,         # = View limits for x-axis
      xlim.p,       # = Instead of setting xlim explicitly, this tuning parameter in
                    #   (0,1] sets the view to display roughly a proportion of the
                    #   invariant intervals. Default 0.75.
      main,         # = Main title for plot (optional)
      sigfig,       # = Significant figures to display in the table. Default is 3.
      digits,       # = Decimal places to display in the table. Default is NULL,
                    #   displaying using sigfig instead. If set overrides sigfig.
      refline,      # = X position for reference line, or NULL to not show line.
                    #   Default NULL.
      greyscale,    # = Use greyscale colour palette? Default is FALSE, use vibrant
                    #   full colour.
      clinsig,      # = Set the clinical significance level. Mark rows with a dagger
                    #   that have thresholds less than this. Default NULL (none
                    #   marked).
      cutoff,       # = A single value or vector pair. Thresholds larger than this
                    #   value or outside the interval pair will be cut off and
                    #   display NT. Default is NULL (no cut off).
      ...           # = Additional parameters to pass to plotting functions.
)
```

REFERENCES

1. National Collaborating Centre for Women's and Children's Health, *Preterm labour and birth.* 2015.

2. National Collaborating Centre for Women's and Children's Health, *Urinary incontinence in women: management.* 2013.

3. National Collaborating Centre for Mental Health, *Social Anxiety Disorder: Recognition, Assessment and Treatment*. 2013, Leicester and London: The British Psychological Society and the Royal College of Psychiatrists.

4. Song, F., et al., *Validity of indirect comparison for estimating efficacy of competing interventions: evidence from published meta-analyses.* British Medical Journal, 2003. **326**: p. 472-476.

5. Song, F., et al., *Inconsistency between direct and indirect comparisons of competing interventions: meta-epidemiological study.* BMJ, 2011. **343**: p. d4909.

6. Veroniki, A.A., et al., *Evaluation of inconsistency in networks of interventions.* International Journal of Epidemiology, 2013. **42**: p. 332-345.

7. Pocock, S., *Safety of drug-eluting stents: demystifying network meta-analysis.* Lancet, 2007. **370**(9605): p. 2099-2100.

8. Higgins, J.P.T. and S. Green, *Cochrane Handbook for Systematic Reviews of Interventions Version 5.0.0 [updated February 2008]*, ed. J. Higgins and S. Green. 2008, Chichester: The Cochrane Collaboration, Wiley.

9. Efthimiou, O., et al., *GetReal in network meta-analysis: a review of the methodology.* Research Synthesis Methods, 2016: p. (early view).

10. Song, F., et al., *Methodological problems in the use of indirect comparisons for evaluating healthcare interventions: survey of published systematic reviews.* British Medical Journal, 2009. **338**(31): p. b1147.

11. 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**: p. 80–97.

12. Linde, K., et al., *Questionable assumptions hampered interpretation of a network meta-analysis of primary care depression treatments.* Journal of Clinical Epidemiology, 2016. **71**: p. 86-96.

13. Li, T., et al., *Network meta-analysis - highly attractive but more methodological research is needed.* BMC Medicine, 2011. **9**: p. 79.

14. Salanti, G., et al., *Evaluating the Quality of Evidence from a Network Meta-Analysis.* PLoS ONE, 2014. **9**(7): p. e99682.

15. Lu, G., et al., *Linear inference for Mixed Treatment Comparison Meta-analysis: A Two-stage Approach.* Research Synthesis Methods, 2011. **2**: p. 43-60.

16. Konig, J., U. Krahn, and H. Binder, *Visualizing the flow of evidence in network meta-analysis and characterizing mixed treatment comparisons.* Stat Med, 2013. **32**(30): p. 5414-29.

17. Krahn, U., H. Binder, and J. Konig, *A graphical tool for locating inconsistency in network meta-analyses.* BMC Med Res Methodol, 2013. **13**: p. 35.

18. Puhan, M.A., et al., *A GRADE Working Group approach for rating the quality of treatment effect estimates from network meta-analysis.* BMJ, 2014. **349**: p. g5630.

19. Balshem, H., et al., *GRADE guidelines: 3. Rating the quality of evidence.* Journal of Clinical Epidemiology, 2011. **64**(4): p. 401-406.

20. Guyatt, G., et al., *GRADE guidelines: 11. Making an overall rating of confidence in effect estimates for a single outcome and for all outcomes.* Journal of Clinical Epidemiology, 2013. **66**(2): p. 151-157.

21. Stinnett, A. and J. Mullahy, *Net health benefits: a new framework for the analysis of uncertainty in cost-effectiveness analyses.* Medical Decision Making, 1998. **18**: p. S68-S80.

22. Tervonen, T., et al., *Applying Multiple Criteria Decision Analysis to Comparative Benefit-Risk Assessment: Choosing Among Statins in Primary Prevention.* Medical Decision Making, 2015. **35**: p. 859-871.

23. Ades, A.E., et al., *Threshold analysis as an alternative to GRADE for assessing the reliability of recommendations based on Network Meta-analysis (NMA)*, in *G-I-N (Guidelines International Network)*. 2015: Amsterdam, Netherlands.

24. Caldwell, D.M., et al., *Threshold analysis can assess the credibility of conclusions from network meta-analysis.* Submitted to J Clin Epi, 2016.

25. Lu, G. and A.E. Ades, *Modelling Between-trial Variance Structure in Mixed Treatment Comparisons.* Biostatistics, 2009. **10**: p. 792-805.

26. Mayo-Wilson, E., et al., *Psychological and pharmacological interventions for social anxiety disorder in adults: a systematic review and network meta-analysis.* Lancet Psychiatry, 2014. **1**: p. 368-376.

27. Phillippo, D.M., et al., *Quantifying the effects of bias: deriving bias adjustment thresholds for Bayesian network meta-analysis.* In preparation.

28. Phillippo, D.M., *Masters Dissertation: Bias Threshold Analysis for Reliability of Network Meta-Analysis*, in *School of Mathematics*. 2015, University of Bristol.

29. National Institute for Health and Care Excellence, *Addendum to Clinical Guideline 150, Headaches in over 12s: diagnosis and management.* 2015.

30. Cohen, J., *Statistical Power Analysis for the Behavioral-Sciences.* Perceptual and Motor Skills, 1988. **67**(3): p. 1007-1007.

31. National Institute for Health and Clinical Excellence, *The guidelines Manual (November 2012). Available from* http://publications.nice.org.uk/the-guidelines-manual-pmg6. 2012, National Institute of Health and Clinical Excellence: London.

32. Doubilet, P., et al., *Probabilistic sensitivity analysis using Monte Carlo simulation: a practical approach.* Medical Decision Making, 1985. **5**: p. 157-177.

33. Critchfield, G.C. and K.E. Willard, *Probabilistic analysis of decision trees using Monte Carlo simulation.* Medical Decision Making, 1986. **6**: p. 85-92.

34. Hastie, T. and R. Tibshirani, *Generalized Additive Models.* Statistical Science, 1986. **1**(3): p. 297-310.

35. Hastie, T.J. and R.J. Tibshirani, *Generalized Additive Models*. 1990: Taylor & Francis.

36. Wood, S., *Generalized Additive Models: An Introduction with R*. 2006: Taylor & Francis.

37. Bhattacharya, S., *A simulation approach to Bayesian emulation of complex dynamic computer models.* 2007: p. 783-815.

38. Conti, S., et al., *Gaussian process emulation of dynamic computer codes.* Biometrika, 2009. **96**(3): p. 663-676.

39. Turner, R.M., et al., *Bias modelling in evidence synthesis.* Journal of the Royal Statistical Society (A), 2009. **172**: p. 21-47.

40. Savovic, J., et al., *Influence of Reported Study Design Characteristics on Intervention Effect Estimates From Randomized, Controlled Trials.* Annals of Internal Medicine, 2012. **157**: p. 429-438.

41. Savovic, J., et al., *Influence of study design characteristics on intervention effect estimates from randomised controlled trials: combined analysis of meta-epidemiological studies.* Health Technology Assessment, 2012. **16**(35).

42. Turner, R.M., et al., *Predictive distributions for between-study heterogeneity and simple methods for their application in Bayesian meta-analysis.* Statistics in Medicine, 2015. **34**(6): p. 984-998.

43. Caldwell, D.M., A.E. Ades, and J.P.T. Higgins, *Simultaneous comparison of multiple treatments: combining direct and indirect evidence.* BMJ, 2005. **331**: p. 897-900.

44. Plummer, M., et al., *CODA: Convergence Diagnosis and Output Analysis for MCMC.* R News, 2006. **6**(1): p. 7--11.