
Early version, also known as pre-print

Link to published version (if available):
10.1111/rssa.12341
10.1111/rssa.12341

Link to publication record in Explore Bristol Research
PDF-document

University of Bristol - Explore Bristol Research

General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available: http://www.bristol.ac.uk/pure/about/ebr-terms
This is the pre-peer reviewed version of the following article:


This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Self-Archiving.

This work was supported by the Centre for Clinical Practice, NICE, with funding to the Clinical Guidelines Technical Support Unit, University of Bristol, the MRC ConDuCT-II Hub (MR/K025643/1), and the MRC grant MR/M005232/1.
Sensitivity of treatment recommendations to bias in Bayesian Network Meta-Analysis

David M Phillippo¹, Sofia Dias¹, AE Ades¹, Vanessa Didelez²,³, Nicky J Welton¹

¹School of Social and Community Medicine, University of Bristol
²Leibniz Institute for Prevention Research and Epidemiology
³Department of Mathematics, University of Bremen

ABSTRACT
Network meta-analysis (NMA) pools evidence on multiple treatments to estimate relative treatment effects. Included studies are typically assessed for risk of bias, however this provides no indication of the impact of potential bias on a decision based on the NMA. We propose methods to derive bias adjustment thresholds which measure the smallest changes to the data that result in a change of treatment decision. The methods use efficient matrix operations, and can be applied to explore the consequences of bias in individual studies or aggregate treatment contrasts, in both fixed and random effects NMA models. Complex models with multiple types of data input are handled using an approximation to the hypothetical aggregate likelihood. The methods are illustrated with a simple NMA of thrombolytic treatments and a more complex example comparing social anxiety interventions. An accompanying R package is provided.

Keywords: Evidence synthesis, influence matrix, mixed treatment comparison, quality of evidence, risk of bias, threshold analysis.

1. INTRODUCTION
Network meta-analysis (NMA) compares the relative effectiveness of multiple treatments by combining the evidence from randomised controlled trials (RCTs), each of which only compares a subset of the treatments of interest (Lumley 2002, Caldwell et al. 2005, Lu and Ades 2006). NMA is increasingly being used by policy-makers to inform treatment recommendations, based on the joint posterior distribution of the treatment effect parameters. However, if some of the included trials are biased then there is a risk that results from the NMA will also be biased, which could lead to sub-optimal treatment recommendations.
There are numerous reasons why results from RCTs may be biased with respect to the target population for decision-making, which are typically dichotomised into: issues of “internal validity”, including poor study design or conduct, for example inadequate randomisation or blinding, or loss to follow-up (Schulz et al. 1995, Savovic et al. 2012b, Savovic et al. 2012a); and issues of “external validity”, affecting generalisation to or representativeness of the target population (Rothwell 2005). The potential for bias in an individual study can be assessed qualitatively using methods such as the Cochrane risk of bias tool (Higgins et al. 2011). The GRADE framework (Guyatt et al. 2011) can also be used to give an indication of the reliability of the evidence informing a pairwise meta-analysis. Recently, two methods to extend the GRADE framework to NMA were proposed (Puhan et al. 2014, Salanti et al. 2014). Whilst such approaches can produce valuable and necessary qualitative assessments, they cannot tell how deficiencies in internal or external validity might affect the treatment recommendation. For example, studies rated at high risk of bias due to issues with internal or external validity that have negligible influence on the treatment recommendation should be of little concern; whereas if they have a larger influence on the treatment recommendation then they should be scrutinised carefully.

Recently, Caldwell et al. (2016) proposed a method for assessing how adjustment for bias (of any source, type, or direction) in individual studies or contrasts would affect the treatment recommendations from a NMA. A form of “threshold analysis”, the methods of Caldwell et al. describe an iterative numerical method based on a two-stage Bayesian NMA to obtain thresholds for how large potential bias adjustments may be before the base-case treatment recommendation changes, and what the new recommendation would be. The information provided by such a threshold analysis is therefore highly relevant to decision makers and guideline developers. However, there are some limitations to the approach taken by Caldwell et al. (2016). Firstly, the two-stage NMA, where pairwise meta-analysis is performed in a first step and then each of the pairwise estimates combined to give consistent NMA estimates, is only an approximation to the preferred one-stage NMA where all studies on all comparisons are synthesised at once (Lu et al. 2011). Decision makers such as NICE recommend the one-stage method due to its accuracy and convenience when results are used in decision models (Dias et al. 2011, NICE 2013, NICE 2014). Secondly, the numerical method is limited in its flexibility, requires the original data and full model details to be available to the analyst, and can involve lengthy computation times. This paper
presents an approach to threshold analysis that can be readily used by decision makers and guideline developers, but avoids the limitations of the approach taken by Caldwell et al. (2016).

The remainder of this paper is structured as follows. In section 2 bias adjustment thresholds are derived algebraically and decision-invariant bias adjustment intervals are constructed, which identify precisely how large a bias adjustment can be before the recommended treatment changes. In section 3 the method is illustrated with examples and applied to two published NMAs. Finally, results are discussed and compared with other approaches. Additional material, including detailed mathematical derivations and proofs, notes on computation, and an accompanying R package, is provided in a web appendix.

2. METHODS

2.1 Network Meta-Analysis

Suppose that we have data from \( n \) studies on \( K \) treatments. Without loss of generality, treatment 1 is set as the reference against which other treatments are compared. Let \( A_j \) be the number of arms in study \( j \in \{1,\ldots,n\} \), so study \( j \) contributes \( A_j - 1 \) relative effects measures (data points) of the treatments in arms 2 to \( A_j \), compared with that in arm 1. There are therefore \( N = \sum_{j=1}^{n} (A_j - 1) \) data points in total, which are contained in the data vector \( y = (y_1, \ldots, y_N)^T \) where we assume that any multiple data entries from the same study are arranged contiguously within \( y \). Each element \( y_j : i \in \{1, \ldots, N\} \) is a relative effect that corresponds to a comparison between treatment \( t_i \) and comparator \( c_j \) (also known).

We assume a multivariate Normal likelihood for the data, so \( y \sim N(\delta, V) \) with \( \delta = (\delta_1, \ldots, \delta_N)^T \) where the covariance matrix \( V \) is assumed known. Studies are assumed independent so that \( V = \text{diag}(V_1, \ldots, V_n) \) is block diagonal, where \( V_j \) is the \((A_j - 1) \times (A_j - 1)\) covariance matrix for study \( j \). If study \( j \) has only two treatments (so only one comparison is made between treatments) then \( V_j \) is a single element giving the variance of the corresponding relative effect.

NMA estimates basic relative treatment effect parameters \( d_k, k = 2, \ldots, K \), for treatment \( k \) compared to the reference treatment 1, with \( d_1 = 0 \). Contrasts between any two treatments \( b \) and
a can then be formed using the consistency assumptions (Lu and Ades 2004) as \( d_{ab} := d_b - d_a \).

Diffuse Normal priors are usually given for the treatment effect parameters; however all results presented here hold for any multivariate Normal prior distribution \( d \sim \mathcal{N}(d_0, \Sigma_d) \), where \( d = (d_1, \ldots, d_k)^T \).

From here, we can proceed in two ways: a fixed effects (FE) or random effects (RE) model. In a fixed effect model, \( \delta_i = d_{\alpha_i} - d_{\beta_i} \) for each \( i = 1, \ldots, N \), which can be written concisely in matrix form as

\[
\begin{align*}
\text{Prior:} & \quad d \sim \mathcal{N}(d_0, \Sigma_d) \\
\text{Likelihood:} & \quad y \mid d \sim \mathcal{N}(\delta, V) \\
\text{FE Model:} & \quad \delta = Xd
\end{align*}
\]

for appropriate \( N \times (K - 1) \) design matrix \( X \) which picks out the corresponding treatment parameters for each study contrast; for example, a study contrast comparing treatments 2 and 4 in a NMA of 5 treatments would have corresponding row in \( X \) set to \((-1, 0, 1, 0)\).

For a random effects model with 2-arm trials, \( \delta_i \sim \mathcal{N}(d_{\alpha_i} - d_{\beta_i}, \tau^2) \) for each \( i = 1, \ldots, N \). The between-study variance \( \tau^2 \) is assumed to be homogeneous between all treatment contrasts. The FE model can be thought of as a special case of the RE model, where the between-study variance \( \tau^2 \) is set to zero.

If there are trials with more than two arms, then a multivariate Normal distribution is required to capture the correlations between the estimated relative effects from the same RCT (Higgins and Whitehead 1996). In general the RE model can be written as

\[
\begin{align*}
\text{Priors:} & \quad d \sim \mathcal{N}(d_0, \Sigma_d), \quad \tau \sim \pi \\
\text{Likelihood:} & \quad y \mid \delta \sim \mathcal{N}(\delta, V) \\
\text{RE Model:} & \quad \delta \mid d, \tau^2 \sim \mathcal{N}(Xd, \Sigma_{\tau^2})
\end{align*}
\]

with some prior \( \pi \) on \( \tau \) (or \( \tau^2 \)) such as \( \tau \sim N(0, 10000) \), and where the between studies covariance matrix \( \Sigma_{\tau^2} \) is of the form \( A \tau^2 \) where \( A \) is a block diagonal “design matrix”. Since \( \tau^2 \) is assumed to be the same between all contrasts, the block of \( A \) corresponding to a study reporting
relative effects with three or more arms will have 1s on the diagonal and 0.5s everywhere else (Higgins and Whitehead 1996).

Whilst we have considered only data in relative effects form here, all results apply easily to data in absolute effects (arm-level) form (or even mixtures of the two) simply by modifying the design and covariance matrices appropriately.

2.2 Decision rule

We assume that the decision is made on the basis of the estimated relative treatment effects from the joint posterior distribution of $d_2, \ldots, d_K$, and (without loss of generality) assume that a larger observed outcome (e.g. log odds of success) is preferable. The optimal treatment is chosen to be that which has the highest expected treatment effect, i.e. $k^*$ which satisfies $\mathbb{E}_{d_{ij}}(d_{k^*}) \geq \mathbb{E}_{d_{ij}}(d_k) \quad \forall k = 1, \ldots, K$. For brevity we write $\mathbb{E}(\cdot)$ in place of $\mathbb{E}_{d_{ij}}(\cdot)$, so that

$$k^* := \operatorname{argmax}_{k=1,\ldots,K} \mathbb{E}(d_k).$$

Note that other decisions rules could be considered, for example the treatment that maximises the expected net benefit from an economic model (see discussion).

2.3 Deriving bias adjustment thresholds at study level

We begin by considering bias adjustments to individual study estimates of treatment effect one at a time, i.e. for each data-point $y_m$. The methods described in sections 2.3 and 2.4 below are repeated for each $m \in \{1, \ldots, N\}$ separately. In section 2.5 we extend the methods to consider bias-adjustments for multiple data-points.

Suppose that some study data point $y_m$, instead of estimating the true value of $d_{i,m}$, is biased so that it estimates $d_{i,m} - \beta_m$. We aim to find threshold values for $\beta_m$ at which the overall decision based on equation (3) changes. To this end we consider hypothetical data that have been bias adjusted, $\tilde{y}$, on which we could perform the NMA to obtain the “true” treatment effect. We define the bias-adjusted data as $\tilde{y}(\beta_m) = y + \beta_m$, where the $i$-th component of the vector $\beta_m$ is

$$[\beta_m]_i = \begin{cases} \beta_m & \text{if } i = m \\ 0 & \text{if } i \neq m \end{cases} \quad \forall i \in \{1, \ldots, N\}.$$
We shall denote posterior expectation with respect to the bias-adjusted data by $\tilde{\mathbb{E}}(\cdot) := \mathbb{E}_{d, y, (\beta_m)}(\cdot)$.

### 2.3.1 General form of bias adjustment thresholds

We wish to find the smallest positive and negative values of the bias adjustment such that the optimal treatment $k^*$ given by equation (3) changes; we call these values bias adjustment thresholds, and denote them $\beta_m^{+\text{thresh}}$ and $\beta_m^{-\text{thresh}}$ respectively. At each threshold value there is a new treatment $\tilde{k}^*$ that achieves the maximum posterior expected treatment effect.

There are $K - 1$ possible solutions $u_{ak^*,m}$, given by

For $k^* \neq 1$

$$u_{ak^*,m} = \frac{-\mathbb{E}(d_{ak^*})}{[H]_{k^*-1,m} - [H]_{a-1,m}}, \quad \text{for } a \in \{2, \ldots, K\} \setminus k^*$$

$$u_{1k^*,m} = \frac{-\mathbb{E}(d_{1k^*})}{[H]_{k^*-1,m}}, \quad \text{for } a = 1 \quad \text{(5)}$$

For $k^* = 1$

$$u_{1a,m} = \frac{-\mathbb{E}(d_{1a})}{[H]_{a-1,m}}, \quad \text{for } a \in \{2, \ldots, K\}$$

$H$ is the influence matrix of the data $y$ on the posterior estimates of the basic treatment effect parameters $d$:

$$\tilde{\mathbb{E}}(d) = \mathbb{E}(d) + H\beta$$

for any general vector $\beta$ changing the data $y$ to $\tilde{y}(\beta) = y + \beta$. The exact form of $H$ will depend on the model, and is described in the following sections for a number of typical NMA models. The influence matrix is related to the hat matrix (Konig et al. 2013, Krahn et al. 2013, Salanti et al. 2014), see discussion.

We note that each $u_{ak^*,m}$ in equation (5) reflects the amount of bias adjustment to data point $y_m$ required to change the sign of $\mathbb{E}(d_{ak^*})$ and make treatment $a$ more efficacious than the current optimal treatment $k^*$. This is determined by the expected difference in treatment effects,
\[-\mathbb{E}(d_{ak}^\ast)\] divided by the amount of influence \(y_m^\ast\) has on the expected difference, given by a linear combination of elements of the influence matrix \(H\). This influence can be thought of as the change in \(\mathbb{E}(d_{ak}^\ast)\) caused by a single unit change in \(y_m^\ast\).

For decisions of the form given in equation (3) based upon the joint posterior distribution of a fixed or random effects NMA, the threshold values \(\beta_m^{+\text{thresh}}\) and \(\beta_m^{-\text{thresh}}\) are then found simply by examining the set of possible solutions \(\{u_{ak^\ast,m}: a \in \{1,\ldots,K\} \setminus k^\ast\}\):

\[
\begin{align*}
\beta_m^{+\text{thresh}} &= u_{bk^\ast,m} \quad \text{where } b = \arg\min_{a \in \{1,\ldots,K\} \setminus k^\ast} \{u_{ak^\ast,m}: u_{ak^\ast,m} > 0\} \\
\beta_m^{-\text{thresh}} &= u_{bk^\ast,m} \quad \text{where } b = \arg\max_{a \in \{1,\ldots,K\} \setminus k^\ast} \{u_{ak^\ast,m}: u_{ak^\ast,m} < 0\}
\end{align*}
\] (7)

The new optimal treatment at the thresholds could be found using equation (3), which requires re-evaluating the joint posterior mean and taking a maximum for each \(\beta_m^{+\text{thresh}}\) and \(\beta_m^{-\text{thresh}}\). However, Lemma 1 (appendix A.1) shows that a more efficient approach is to simply note the new optimal treatment from the contrast whose posterior expectation changes sign at the bias adjustment threshold – treatment \(b\) from equation (7).

From the positive and negative bias adjustment thresholds, it is intuitive to think of constructing an interval \(\left(y_m + \beta_m^{-\text{thresh}}, y_m + \beta_m^{+\text{thresh}}\right)\) within which a bias-adjusted value of \(\tilde{y}_m^\ast\) can lie without changing the treatment decision. We refer to such an interval as the decision-invariant bias adjustment interval about \(y_m^\ast\), and visualise this as shown in Figure 1.

![Figure 1](image_url)

Figure 1: Example construction of a decision-invariant bias adjustment interval, which is shaded. The new treatment decisions at the negative and positive thresholds would be 2 and 3 respectively.

Thresholds and invariant intervals may be derived for more complex treatment decisions, as well as the simple “maximal efficacy” decisions described above, by examining the set of values \(\{u_{ab,m}: 1 \leq a < b \leq K, 1 \leq m \leq N\}\) from equation (5).
2.3.2 Bias adjustment thresholds for the fixed effect model

For the FE model with conjugate Normal prior distribution for the treatment effect parameters $d$ (equation (1)), appendix A.2 (see also Gelman et al. 2013, p. 71) shows that the posterior distribution is:

$$d \mid y \sim N\left(\Sigma_n^{-1}d_0 + X^TV^{-1}y, \Sigma_n\right)$$ (8)

where the posterior covariance matrix is $\Sigma_n = (\Sigma_d^{-1} + X^TV^{-1}X)^{-1}$.

The threshold values are found using equations (7) and (5), where the influence matrix is $H = \Sigma_nX^TV^{-1}$ (appendix A.3).

2.3.3 Bias adjustment thresholds for the random effects model

The RE model (equation (2)) is typically specified with a prior distribution over the between-studies standard deviation $\tau$ which, due to the hierarchical nature of the model, results in a joint posterior distribution that generally has no closed form solution. One approach in this situation would be to find bias adjustment thresholds numerically by iteratively changing the data until the decision changes; this is likely to be very computationally expensive. However, approximate algebraic bias adjustment thresholds can be obtained for the random effects model by considering the between-studies variance to be known, fixed, and unchanged after bias adjustment. Sensitivity analyses may then be performed to assess how the thresholds change for different values of $\tau^2$.

For the RE model given in equation (2) with $\tau^2$ assumed known and fixed, appendix A.4 (see also Gelman et al. 2013, p. 582) shows that the joint posterior distribution for $d$ and $\delta$ is

$$\begin{pmatrix} d \\ \delta \end{pmatrix} \mid y, \tau^2 \sim N\left(\Sigma_n^{-1}d_0 + \Sigma_{d\delta}^{-1}X^TV^{-1}y, \Sigma_n^{-1}\right)$$, where $\Sigma_n^{-1} = \begin{pmatrix} X^T\Sigma_{d\delta}^{-1}X + \Sigma_d^{-1} & -X^T\Sigma_d^{-1}X \\ -X^T\Sigma_{d\delta}^{-1}X & V^{-1} + \Sigma_{\delta\delta}^{-1} \end{pmatrix} = \begin{pmatrix} A & B \\ B^T & C \end{pmatrix}$, (9)

where the posterior covariance matrix $\Sigma_n$ is partitioned according to the dimensions of $d$ and $\delta$.

Under bias-adjusted data, it can be shown (appendix A.5) that the joint posterior mean becomes $\tilde{E}(d) = E(d) + B,V^{-1}\beta_m$. Following the same arguments as the basic FE case (appendix A.3), the thresholds are given by equations (7) and (5) where the influence matrix is now $H = B,V^{-1}$. 

8
Note that the posterior covariance matrix $\Sigma_n$ is the inverse of a block matrix, and so can be calculated explicitly (see Bernstein 2005, p.45); it is however more likely that $\Sigma_n$ will have been estimated using Bayesian software such as WinBUGS (Lunn et al. 2000). We can then simply partition the posterior covariance matrix as in equation (9) to obtain $B$.

### 2.3.4 Extended models with additional parameters

We may wish to add additional parameters to the basic FE and RE models (sections 2.3.2 and 2.3.3), for example to include data as absolute effect measures (i.e. as one observation per study arm) where a nuisance study-level baseline parameter for arm 1 is included (Lu and Ades 2006). We denote the additional parameters by $\mu$, and give them a Normal prior distribution $\mu \sim N(\mu_0, \Sigma_\mu)$.

The simplest way to achieve this for the FE model is to extend the parameter vector to $\gamma = \begin{pmatrix} d \\ \mu \end{pmatrix}$. The design matrix $X$ is also extended to describe the model. Appendix A.6 shows that we obtain the threshold equations (7) and (5), where the influence matrix is now $H = \left[\Sigma_n X^T V^{-1}\right]_{\text{rows } 1:K-1}$.

For the RE model, the additional parameters have an associated design matrix $M$, and we impart further flexibility with a design matrix $L$ for $\delta$. Appendix A.7 shows that the thresholds are given by equations (7) and (5) with $H = \left(B^T L + D^T M^T\right) V^{-1}$. Here, analogously to equation (9), $B$, and $D$, are partitions of the posterior covariance matrix, corresponding to the covariance of $d$ with $\delta$ and of $d$ with $\mu$ respectively.

### 2.3.5 Class effect RE model

Class effect models are often utilised in NMAs where treatment effects may be assumed exchangeable within discrete classes, for example based upon common constituent compounds or modes of action (Dominici et al. 1999, Mayo-Wilson et al. 2014). In such models, treatment effects within the same class are assumed exchangeable and Normally distributed as $d \mid z \sim N(Zz, \Sigma_d)$, with class effect parameters $z$ and class design matrix $Z$ assigning a class to each treatment. The class effect parameters are given a Normal prior distribution. $\Sigma_d$ is the between-treatment covariance matrix, which may specify a common within-class variance or different within-class
variances for each class. If \( \Sigma_q \) is the zero matrix then the model is equivalent to fixed class effects. In order to proceed analytically we assume that the between-studies variance \( \tau^2 \) is fixed, known, and invariant to bias adjustment (as with the RE model in sections 2.3.3 and 2.3.4); we must also make the same assumptions about the within-class variances (for the random class effect model). In appendix A.8 we show that the influence matrix for a RE model including class effects is identical to that in the extended RE case in section 2.3.4; we may proceed exactly as in the extended RE case despite the presence of class effects.

2.4 Bias adjustment thresholds at the contrast level

In clinical guideline development, assessment of evidence quality is often directed at the entire body of evidence on a contrast rather than at individual studies. This is the method of evidence classification used in, for example, extensions of GRADE to NMA (Puhan et al. 2014, Salanti et al. 2014). We may therefore wish to examine the robustness of treatment decisions to bias in the combined body of evidence at contrast level, rather than for individual studies. In some cases it may only be possible to obtain decision invariant thresholds at the contrast level, i.e. when only the summary results (posterior means and covariance matrix for all parameters) from a NMA are available. Alternatively the NMA may entail a complex, hierarchical, or otherwise analytically intractable model but where the joint posterior distribution for the treatment effect parameters can be assumed to be approximately multivariate Normal.

Our approach is to consider a hypothetical dataset, consisting of a single independent data point for each contrast where there is direct evidence, which when pooled using a fixed effects NMA gives a posterior distribution that closely approximates the true posterior distribution as reported by the original NMA. Note that we are not suggesting independence of the original data, but that the posterior distribution could have arisen (at least approximately) from an alternative set of independent data points. Multi-arm trials, random effects, and other features are therefore handled as usual in the original NMA, and all correlations and uncertainty appropriately propagated into the joint posterior distribution upon which the contrast-level threshold analysis is based. We show that, in order to derive thresholds, we need only the covariance matrix of the hypothetical data, and not the hypothetical data points themselves. We then proceed to derive thresholds as for the basic FE model described in section 2.3.2.
We consider a hypothetical dataset consisting of single independent data points $y_{ab}$ with variances $v_{ab}$, representing the combined evidence on each contrast $d_{ab}$ where there is direct evidence, with multivariate Normal likelihood $y | d \sim N(Xd, V)$ where $X$ is a design matrix and $V$ is diagonal with elements $v_{ab}$. We design the hypothetical dataset so that pooling using a FE NMA gives a posterior distribution $N(\hat{\eta}, \hat{\Sigma})$ that closely approximates the true posterior distribution $N(\eta, \Sigma)$ reported by the original NMA. Thresholds can then be derived as for the basic FE model in section 2.3.2.

A full derivation of the contrast-level method is given in appendix A.9. We choose $V$ to solve $\hat{\Sigma} = \hat{\Sigma}$, where the covariance matrix of the reconstructed posterior distribution is $\hat{\Sigma} = (\Sigma^{-1} + X^T V^{-1} X)^{-1}$ (see section 2.3.2). When the evidence network is complete (that is, every treatment is joined to every other by direct evidence), there is a unique exact solution; otherwise an approximate solution is found using non-negative least squares (NNLS) (Lawson and Hanson 1995). In the latter case, the performance of the approximation may be assessed by examining the Kullback-Leibler (KL) divergence (Kullback and Leibler 1951) of the reconstructed posterior distribution from the true posterior distribution. Interpreting the KL divergence as a log Bayes factor, values less than 1 indicate negligible differences between the reconstructed posterior from the true posterior and a good approximation, whilst values greater than 3 indicate considerable differences and a poor approximation (Kass and Raftery 1995).

Once the hypothetical likelihood covariance matrix has been reconstructed, the thresholds are then evaluated as before using equations (7) and (5) with the influence matrix $H = \Sigma X^T V^{-1}$. Note that it would not be possible to re-evaluate the posterior means under the bias-adjusted data to obtain $\tilde{k}^{*}$ as we do not have the hypothetical data, but we can use the result of Lemma 1 (appendix A.1) to efficiently obtain the new optimal treatment, as before.

2.5 Thresholds for multiple biases

Thus far we have been concerned with the effects of bias adjustment for a single data point at a time. However it is possible that we may wish to consider the impact of bias adjustment in multiple studies or contrasts simultaneously; for example all the relative effects estimates from a multi-arm
study, or perhaps multiple studies that are of concern. Such analyses are possible at both study and contrast level, though are more likely to be motivated by knowledge of individual trials and their characteristics. Equation (6) shows how a general bias adjustment $\beta$ would change the posterior mean of the treatment effect parameters. We extend the approach taken in appendix A.3 to let two elements of $\beta$ be non-zero in equation (4), allowing for bias adjustment in two data points $y_{m_1}$ and $y_{m_2}$ simultaneously. We end up solving $K-1$ equations in two unknowns:

$$0 = \mathbb{E}(d_{ak}) = \mathbb{E}(d_{ak}) + \left( [H]_{k-1,m_1} - [H]_{a-1,m_1} \right) \beta_{m_1} + \left( [H]_{k-1,m_2} - [H]_{a-1,m_2} \right) \beta_{m_2}$$

$$\forall a \in \{1, \ldots, K\} \setminus k^*$$

where $m_1, m_2$ are the indices of the two data points to be bias-adjusted. We see that, instead of threshold points, we have $K-1$ threshold lines in two dimensions. By rearranging (10) and using the definition of $u_{ak,m}$ from equation (5), we can use the set of previously calculated $u_{ak,m}$ to arrive at the equation for each threshold line $\beta_{ak}^{\text{thresh}}$: $\beta_{m_2} = u_{ak,m_2} - \left( u_{ak,m_2} / u_{ak,m_1} \right) \beta_{m_1}$. The

Figure 2: Example of threshold lines in two dimensions, with the invariant region about the origin (no bias adjustment) shaded. Any simultaneous bias adjustment $(\beta_{m_1}, \beta_{m_2})$ to data points $y_{m_1}$ and $y_{m_2}$ which remains within the invariant region does not change the optimal treatment. At the boundary of the invariant region formed by threshold line $\beta_{ak}^{\text{thresh}}$ the new optimal treatment is $\tilde{k}^* = a$. 
intersection of these threshold lines creates a bias invariant region, for example as portrayed in Figure 2.

It is simple both mathematically and computationally to carry on extending such a technique to higher dimensions: allowing \( r \leq N \) components of \( \beta \) to be non-zero results in \( r \)-dimensional threshold hyperplanes \( \rho_{ak}^{\text{thresh}} \) with equations \( \beta^T \mathbf{w}_{ak} = 1 \) for \( a \in \{1, \ldots, K\} \setminus k^* \), where the \( i \)-th component of \( \mathbf{w}_{ak}^* \) is

\[
\begin{bmatrix}
\mathbf{w}_{ak}^*
\end{bmatrix}_i = \begin{cases} 
\mathbf{u}_{ak}^i \mathbf{1} & \text{if } i \in \{m_1, \ldots, m_r\} \\
0 & \text{otherwise}
\end{cases},
\forall i \in \{1, \ldots, N\}.
\]

However, beyond two (or possibly three) dimensions it becomes impossible to visualise and effectively analyse these threshold hyperplanes and the resulting invariant hypervolume formed by their intersection. As such, any analysis of simultaneous bias adjustment is likely best approached in a targeted manner, identifying a small number of data points on which to focus attention.

An alternative approach is to report the vectors \( \mathbf{p}_{ak}^\text{min} = \mathbf{w}_{ak}^* \cdot \left\| \mathbf{w}_{ak}^* \right\|^2 \) for \( a \in \{1, \ldots, K\} \setminus k^* \) giving the point on each threshold hyperplane \( \rho_{ak}^{\text{thresh}} \) which lies closest to the origin, and so minimises the amount of overall bias adjustment required to change the optimal treatment decision to \( \tilde{k}^* = a \).

3. EXAMPLES

We apply the threshold method to two examples: firstly, a NMA of thrombolytic treatments (Caldwell et al. 2005) to demonstrate study- and contrast-level analyses on a simple FE model, along with simultaneous bias adjustment in two data points; secondly, a large class effects RE NMA comparing treatments for social anxiety (NCCMH 2013, Mayo-Wilson et al. 2014) to demonstrate the power of a contrast-level analysis when applied to complex models. Notes on practical computation are included in appendix A.10. Code is provided in the supplementary materials, along with an R package for general use.

3.1 Example: Thrombolytics

Figure 3 shows the network of treatment comparisons for \( K = 6 \) thrombolytic treatments based on \( n = 14 \) studies, taken from two systematic reviews (Boland et al. 2003, Keeley et al. 2003).
Previous work has shown that a fixed effects model is appropriate for the data (Caldwell et al. 2005, Dias et al. 2013d). All studies have two arms apart from one with three, so the number of data points (log odds ratios) is $N = 15$.

![Figure 3: Thrombolysis example network, showing how the six treatments are connected by study evidence. Nodes represent treatments and edges show comparisons made by studies. Numbers inside the nodes are the treatment codings. Numbers on the edges give the number of studies making that comparison. The bold triangle is the loop formed by the three-arm study.](image)

### 3.1.1 Study-level fixed effects analysis

A fixed effect model was fitted to the data using WinBUGS 1.4.3 (Lunn et al. 2000) and code from Dias et al. (2013d). The treatment effect parameters $d_k$ are interpreted as the log odds ratio of mortality between treatment $k$ and the reference treatment 1, and $d_1 = 0$. In this example the optimum treatment is the one which minimises the log odds ratio (OR) of mortality, here $k^* = \arg\min_k \mathbb{E}(d_k) = 3$ (full results of the NMA are available in the accompanying R package).

The results of each study and the study-level threshold analysis are shown in Figure 4: the table on the left hand side displays the estimated log OR from each study comparison along with a 95% confidence interval, and a decision-invariant bias adjustment interval about the estimate, showing how far any bias adjustment can be made before the optimal treatment changes. The new optimal treatments $\tilde{k}^*$ are reported alongside either end of the invariant interval. The right-hand side of the figure displays these graphically, with points and lines for estimated log ORs and their confidence intervals and shaded bands for the invariant regions. Where a 95% CI extends beyond the invariant interval the study label is bold, indicating that the treatment recommendation is sensitive to the level of imprecision in this study estimate. In this example, the treatment recommendation is sensitive to the level of imprecision in studies 11 and 35 ($\tilde{k}^* = 5$), and study 34 ($\tilde{k}^* = 6$). For
example, the estimated log OR of 0.005 for treatment 6 vs. 3 in study 34 has an invariant interval of (−0.00, 18.51); a negative bias adjustment to the log OR of only −0.006 in favour of treatment 6 is enough to change the optimal treatment from $k^* = 3$ to $k^* = 6$. Looking at the network of treatments (Figure 3), this is not surprising. Treatment 6 is only compared directly to treatment 3, and only in study 34, which found no evidence of a significant difference between the two treatments (the 95% CI for the log OR contains zero). Since $k^* = 3$, adjusting the log OR to be in favour of treatment 6 means that the network of evidence behind treatment 3 now points to $k^* = 6$. Similarly, negative bias adjustments of −0.062 and −0.144 to the 5 vs. 3 and 5 vs. 1 comparisons of studies 35 and 11 respectively both result in $k^* = 5$ becoming optimal. Thus our method highlights that the treatment recommendation is sensitive to bias adjustments in certain studies.

The thresholds and invariant intervals should be further interpreted in light of the expected direction of bias, for example novelty bias favouring a new treatment, or sponsorship bias favouring a manufacturer’s own treatment in their own trial (Song et al. 2008, Chaimani and Salanti 2012). Our method also reveals that no changes in other studies – no matter how large – can ever plausibly lead to changes in the treatment recommendation, particularly studies 2-4 and 7-10, since the bias adjustment thresholds are infeasibly large on the log OR scale.
3.1.2 Contrast-level analysis

We also perform a contrast-level analysis to examine sensitivity to changes in the aggregate bodies of evidence on each contrast. We do not need the original data to do this; we use only the posterior means and covariance matrix from the joint posterior distribution of the treatment effect parameters \( \mathbf{d} = (d_2, \ldots, d_6)^T \). We treat the posterior distribution as if it arose from a NMA on seven independent data points \( \mathbf{y} = (y_{12}, y_{13}, y_{14}, y_{15}, y_{34}, y_{35}, y_{36})^T \) — the number of direct comparisons in the treatment network (i.e. the edges in Figure 3). Each data point \( y_{ab} \) represents the combined direct study evidence on a treatment contrast \( d_{ab} \).

Following the methods in section 2.4, we construct an approximate hypothetical likelihood using non-negative least squares (NNLS) — see appendix A.11. The Kullback-Leibler divergence of the reconstructed posterior distribution from the true posterior distribution is very small at \( 6.76 \times 10^{-5} \), indicating that the hypothetical data are a good approximation.

Figure 5 presents results of the NMA and the contrast-level analysis, which echoes the study-level analysis: it is possible to make plausible adjustments for bias that result in either treatment 5 or treatment 6 becoming optimal. Notably the thresholds for contrasts where there is a single two-arm study making the comparison (6 vs. 3 and 5 vs. 1) match almost exactly with the thresholds for the corresponding studies in the study-level analysis (section 3.1.1), as expected. Furthermore, we clearly see the effects of bias adjustment on entire bodies of evidence in comparison with the study-level approach: individually, studies making the 2 vs. 1 comparison have little influence on the treatment decision, shown by wide invariant intervals (Figure 4); when the evidence from these studies is considered collectively for bias adjustment, the combined invariant interval becomes
narrower. Note here that the black lines in Figure 5 correspond to the 95% credible intervals for each contrast estimate resulting from the NMA, instead of confidence intervals for each study estimate as in the study-level analysis (Figure 4).

The threshold analysis gives very small thresholds for the combined evidence on treatment contrasts 5 vs. 3 and 6 vs. 3, which is symptomatic of the lack of evidence for significant differences between these treatments. A likely treatment decision in such a scenario (in the absence of issues surrounding cost or adverse events) would therefore be to recommend any of these three treatments.

3.1.3 Simultaneous bias adjustment in two data points

We shall now consider analysing bias adjustment in two data points simultaneously. Such an analysis is possible at both study- and contrast-level, though it is more likely to be motivated by knowledge of individual trials and their characteristics; thus we shall return to the study-level scenario for this example. In the thrombolitics dataset, study 1 was a three-armed study comparing treatments 1, 3, and 4, resulting in two log odds ratio estimates against the reference treatment 1. The two log ORs are not independent, and so if bias adjustment is required it is possible that both estimates will need to be bias-adjusted together – if the trial failed to blind patients, for example. Figure 6 presents the invariant region for simultaneous bias adjustments in the two log ORs estimated by study 1, formed by the polygon of intersecting threshold lines about the origin, which can either be closed (threshold lines in every direction) or open (bias adjustment in some direction will never cross a threshold line). In this example, three threshold lines form an invariant region for bias adjustment, with new optimal treatments at the thresholds $\hat{k}^* = 4$, 5, and 6. Notice that the points where the boundaries of the invariant region intersect the axes correspond to the one dimensional invariant intervals presented in Figure 4, since setting one of the two bias adjustments to zero returns us to analysing a bias adjustment in one data point only. Of particular interest are the threshold lines for $\hat{k}^* = 4$ and $\hat{k}^* = 5$ which lie closest to the origin. In the one-dimensional case we saw that, individually, bias adjustments of +0.120 in the log OR of treatment 3 vs. 1 or –0.107 in the log OR of treatment 4 vs. 1 were needed to change the optimal treatment to $\hat{k}^* = 4$, and a bias adjustment of +5.277 in the log OR of treatment 4 vs. 1 was needed to change the optimal treatment to $\hat{k}^* = 5$ (Figure 4). Now allowing both estimates to be bias-adjusted simultaneously, we see that it is possible to arrive at $\hat{k}^* = 5$ with much smaller amounts of bias
adjustment than this; for example with bias adjustments of just +0.144 to the 3 vs. 1 log OR and +0.021 to the 4 vs. 1 log OR we cross the invariant threshold and would recommend treatment 5.

![Diagram](image)

Figure 6: Invariant region formed from threshold lines for bias adjustment to the two relative effect estimates from Study 1. The invariant region is shaded, and the new optimal treatments on the boundary are indicated by $k_+$. The optimal treatment without bias adjustment is $k_+ = 3$.

### 3.2 Example: Social Anxiety

We now consider a more complex example where analysis is greatly simplified using the contrast-level approach. Figure 7 shows the network for a NMA of 41 interventions for social anxiety from 100 studies (NCCMH 2013, Mayo-Wilson et al. 2014). The original analysis uses a random effects model which includes class effects for 17 different treatment classes and a secondary network of studies for a regression calibration on recovery. No single common outcome measure was used across the included studies, so instead treatment effects were transformed into standardised mean differences (SMDs) for the purposes of NMA. Table A1 in appendix A.12 lists the treatment codes and classes, and full results of the NMA are available in the accompanying R package.
3.2.1 Contrast-level analysis

Despite the complexity of the original analysis, a contrast-level threshold analysis is straightforward. We consider the joint posterior distribution as if it arose from a NMA on 84 independent data points, each representing the aggregate direct evidence available on a single treatment contrast. Following the methods in section 2.4, we construct an approximate hypothetical likelihood using NNLS. The fitted hypothetical likelihood covariance matrix includes a single infinite variance for one contrast (7 vs. 1), meaning that the direct evidence on this contrast is estimated to have no influence on the posterior distribution. The Kullback-Leibler divergence of the reconstructed posterior distribution from the true posterior distribution is 1.55, indicating that the hypothetical data are a reasonable approximation (interpreted as a log Bayes factor, greater than 1 but less than 3).
Due to the large number of contrasts, Figure 8 shows only the results of the threshold analysis for contrasts with thresholds $< 2$ standardised mean difference (SMD). The optimal treatment under the original analysis is $k^* = 41$, group cognitive behavioural therapy (CBT) with phenelzine. No contrasts have invariant intervals which lie inside the 95% CrI, meaning that the treatment recommendation is robust to the level of imprecision in the contrast-level data. The smallest threshold is a positive change of 0.46 in the estimate of $-0.88$ SMD for the 41 vs. 31 contrast (the upper limit of the corresponding invariant interval is $-0.88 + 0.46 = -0.42$), at which point treatment 36 (cognitive therapy) becomes optimal. Cohen (1988) considered a standardised mean difference...
difference of more than 0.8 to be large in the context of behavioural sciences; all but five thresholds are larger than this, and for each of these the new optimal treatment is treatment 36. An important observation from this analysis is that the treatment recommendation is insensitive to changes in the combined evidence on the large majority of contrasts. Rather than performing a long and laborious qualitative assessment of all 84 contrasts and 100 studies, attention can be focused on the smaller number of contrasts (for example the 5 studies with thresholds smaller than 0.8 SMD) where plausible adjustments to the data may cause a change in treatment recommendation.

### 3.2.2 More complex analyses: pharmacological and psychological treatment bias

The methods described in section 2 are easily extended to more complex decision rules and bias adjustment scenarios, simply by manipulating the set of $u_{ak,m}$ values. Here, we have considered the effects of adjusting for a potential common bias (at the contrast level) amongst all pharmacological treatments (including combination therapies) and similarly amongst all psychological treatments.

The results of these analyses are shown in Figure 9. In each case, at the smallest threshold, treatment 36 becomes optimal: with an adjustment of $+0.67$ SMD for all pharmacological treatments compared to inactive control (i.e. reducing their efficacy), or with an adjustment of $-0.66$ SMD for all psychological treatments compared to an inactive control (i.e. increasing their efficacy). The magnitude of these thresholds is large (Cohen 1988) – likely much larger than any plausible common bias. We might also consider the effects of adjusting for these common biases.

![Figure 9](image_url)
simultaneously by examining the set of $u_{ak,m}$ values. The resulting two-dimensional invariant region is shown in Figure 10. The size of the invariant region would likely reassure decision makers that adjustment for common pharmacological and/or psychological treatment effect biases (if they exist) would not affect the treatment recommendation.

Figure 10: Invariant region for simultaneous adjustments for common biases in all psychological and all pharmacological treatments. The new treatment recommendation at the boundary is shown as $k^*$; the optimal treatment without bias adjustment is $k^* = 41$.

4. DISCUSSION

The threshold method presented in this paper enables researchers and decision makers to quantify the robustness of their conclusions to potentially biased evidence. Current approaches based around the GRADE framework (Puhan et al. 2014, Salanti et al. 2014) give a thorough qualitative evaluation of the quality of evidence behind such decisions, but fall short of describing the impact on treatment recommendations of any bias present in the evidence. Providing bias adjustment thresholds and invariant regions can attest to the robustness of conclusions despite poor quality evidence, or can highlight areas where the evidence should be carefully assessed for bias since bias adjustments of plausible magnitude could change the optimal treatment decision. Although our method gives quantitative results, their interpretation still requires qualitative judgements to
determine which evidence might plausibly be biased and to what extent. We also note that, although we particularly highlight bias adjustment thresholds that lie inside the confidence/credible interval for an estimate as a cause for concern, it is entirely possible for larger bias adjustments to be plausible; for example large studies yielding precise estimates may be biased beyond the range of their 95% CI. In this case it must be judged whether any necessary bias adjustment is larger than the invariant threshold.

It must be clear that threshold analyses do not seek to test for the presence or absence of bias, nor do they make any assumptions about the source, type, or expected magnitude or direction of any bias. Rather, if any such bias was present, then subsequent adjustment would only alter the treatment decision if it were larger than the given thresholds. Knowledge of the likely nature of possible biases should be used in the planning and – most importantly – interpretation of threshold analyses.

Threshold analysis has previously been proposed by Caldwell et al. (2016), using a numerical method to derive thresholds based on a two-stage Bayesian NMA. A particular feature of the method proposed in this paper is that it starts from the one-stage Bayesian posterior distribution of relative treatment effects and manipulates it algebraically, rather than iteratively modifying the data. Not only can algebraic solutions be reached almost instantaneously using matrix operations rather than lengthy and computationally expensive numerical techniques, but this confers considerable flexibility; in practice, treatment recommendations are often based on complex models with multiple types of data input which would be difficult to fit into the two-stage framework. Furthermore, the original data are not required for the threshold analysis to be performed, provided that posterior means and covariance matrix of the parameters are available. Although not frequently published at present, this level of summary data is likely to be much easier to obtain on request than the full original dataset; this significantly widens the scope of threshold analysis, compared to numerical methods.

An extension of this work is to embed the threshold method into a probabilistic cost-effectiveness analysis (CEA) (Doubilet et al. 1985, Critchfield and Willard 1986, Dias et al. 2013c), where the optimal treatment is found not by maximising the posterior expected treatment effect as in equation (3), but by maximising the posterior expected value of some net benefit function instead. A CEA seeks to weigh up the improvements in quality of life and life expectancy against the total costs
for each treatment regimen, and this is achieved by the use of a net benefit function (Stinnett and Mullahy 1998). Such analyses are used extensively by reimbursement agencies and threshold analysis would be useful to determine how bias adjustments can affect the outcome of a CEA. When the net benefit function is linear in treatment efficacy (or can be approximated as such), the threshold equations (5) can be easily transformed onto the net benefit scale. However, CEA models can be complex and often involve net benefit functions that are non-linear; as such it would be useful to extend the threshold methodology to deal with non-linear decision functions.

Other decision rules besides maximum efficacy or net benefit may be considered, for example recommending any active treatment if better than placebo, recommending a group of treatments whose efficacies are clinically equivalent (e.g. within some minimum clinically important difference or non-inferiority margin), or restricting a recommendation to currently available treatments. More complex threshold analyses are also possible, for example to examine generic bias in a class of treatments or studies sharing given characteristics. All of these analyses are possible directly by examining the set of values $u_{ab,m}$ (see equation (5)), giving the amount of adjustment to data point $y_m$ which would see the posterior expectation of the contrast between treatments $a$ and $b$ change sign (so treatment preference between $a$ and $b$ switches).

Derivation of algebraic thresholds for the random effects model is hindered by the analytic intractability of the joint posterior distribution when the between studies variance is given a prior distribution; instead we make the assumption that this variance is fixed and known so that conjugacy is preserved. This assumption should be tested by sensitivity analyses substituting plausible values of $\tau^2$, for example from the upper and lower limits of the 95% credible interval obtained from the NMA or from predictive distributions derived from similar meta-analyses (Rhodes et al. 2015, Turner et al. 2015). There is empirical evidence that heterogeneity is greater in biased evidence bases (Savovic et al. 2012b, Savovic et al. 2012a), so it might be expected that $\tau^2$ would reduce after adjusting for bias (possibly beyond the lower credible limit).

Further applications of the threshold method are to meta-regression and bias adjustment models (Dias et al. 2013b). The approach would follow from section 2.3.4, where the additional parameters are regression covariates. Particular care should be taken to define an appropriate decision rule and in the interpretation of the treatment effect parameters since decisions can be
different at different covariate values. The interpretation of the thresholds in this case is in terms of the adjustment for residual biases not accounted for in the model.

We have seen that the influence matrix $H$ of the data $y$ on the posterior mean of the treatment effect parameters $\mathbb{E}(d)$ is central to the derivation of thresholds, and describes how changes in individual data points affect the posterior means of the basic treatment effect parameters. The role of a related quantity, the hat or contributions matrix, has been highlighted by several authors previously in the context of influence analysis (Konig et al. 2013, Krahn et al. 2013, Salanti et al. 2014). The hat matrix describes how changes in individual data points affect the predicted values (as opposed to the treatment parameters), and in the basic FE model is derived from the design and influence matrices as $XH$. Krahn et al. (2013) and Konig et al. (2013) use the hat matrix to visualise the “flow of evidence” in a NMA and to analyse and detect inconsistency, which Salanti et al. (2014) utilise within the context of GRADE applied to NMA.

A key contribution of this paper has been the reconstruction of the influence matrix (and therefore the hat matrix) from the Bayesian posterior distribution. This allows considerable flexibility because, in practice, treatment recommendations may be based on complex models, for example including class effects, and may incorporate several types of data. The social anxiety guideline (NCCMH 2013), for example, incorporates data on both response and recovery rates, and synthesises trials reporting either odds ratios or outcomes on continuous scales. It is, of course, the need for flexible computation methods in the face of irregular and complex data that has made Bayesian MCMC the method of choice in practical applications (Dias et al. 2013a).

One practical limitation of the threshold method is that there is no satisfactory way to display the results of simultaneous bias adjustments in more than three contrasts or data points. As we have shown in section 2.5 the problem lies not with deriving bias adjustment thresholds in higher dimensions but in visualising and interpreting them. We have given examples of how to visualise invariant regions in two dimensions (e.g. Figure 6, section 3.1.3), and a similar approach is possible in three dimensions. For more than three contrasts or data points, a graphical representation of this kind is not possible, making interpretation difficult. However, in practice, if bias adjustment is to be considered for a large number of studies, it may be preferable to estimate the study-level bias adjustment within the hierarchical NMA analysis, either by regression (Dias et al. 2010, Salanti et al. 2010, Naci et al. 2014) or by giving bias terms informative priors based on expert opinion.
(Turner et al. 2009, Welton et al. 2009, Dias et al. 2010). A potential avenue for future research into the effects of multiple simultaneous bias adjustments lies with the influence matrix. By examining this matrix it should be possible to identify whether bias adjustment in a given combination of data points may lead to wider invariant intervals, due to the influences of multiple data points partially cancelling out, or smaller invariant intervals, due to the combined influence increasing additively.

Importantly, threshold analysis of complex, hierarchical, or otherwise atypical NMA models may always be performed at the contrast level, provided that the joint posterior distribution of the treatment effect parameters is available (either first hand from an analysis, or sufficiently reported in a published NMA), and that this joint posterior distribution is at least approximately Normal. Under such conditions we can apply the methods proposed in section 2.4 to derive bias adjustment thresholds and invariant intervals, regardless of the manner in which the joint posterior distribution arose. As such, the threshold method proposed is applicable to a wide range of situations that may be encountered by decision makers, and has the potential to focus discussion on the risk of bias in particular studies or comparisons to which the final treatment recommendation is most sensitive.

SUPPLEMENTARY MATERIALS

Appendices: All appendices are contained in a separate document providing technical derivations, statements and proofs of theorems and lemmas, and notes on computation. (PDF document)

Computer Code: An R package nmathresh is provided that implements the threshold method, along with the R code and data used to perform the example analyses. (Source tarball)

BIBLIOGRAPHY


R Core Team (2016) R: A Language and Environment for Statistical Computing, Vienna, Austria.


P. M. Rothwell (2005) External validity of randomised controlled trials: “To whom do the results of this trial apply?”* The Lancet*, 82-93.


