

# Multi-criteria benefit-risk assessment using network meta-analysis

Gert van Valkenhoef<sup>\*,a,b</sup>, Tommi Tervonen<sup>c</sup>, Jing Zhao<sup>a</sup>, Bert de Brock<sup>b</sup>, Hans L. Hillege<sup>a</sup>, Douwe Postmus<sup>a</sup>

<sup>a</sup>Department of Epidemiology, University Medical Center Groningen, The Netherlands

<sup>b</sup>Faculty of Economics and Business, University of Groningen, The Netherlands

<sup>c</sup>Econometric Institute, Erasmus University Rotterdam, The Netherlands

---

## Abstract

**Objective:** To enable multi-criteria benefit-risk assessment of any number of alternative treatments using all available evidence from a network of clinical trials.

**Study design and setting:** We design a general method for Multi-Criteria Decision Aiding (MCDA) with criteria measurements from Mixed Treatment Comparison (MTC) analyses. To evaluate the method, we apply it to benefit-risk assessment of four second-generation antidepressants and placebo in the setting of a published peer reviewed systematic review.

**Results:** The analysis without preference information shows that placebo is supported by a wide range of possible preferences. Preference information provided by a clinical expert showed that while treatment with antidepressants is warranted for severely depressed patients, for mildly depressed patients placebo is likely to be the best option. It is difficult to choose between the four antidepressants, and the results of the model indicate a high degree of uncertainty.

**Conclusions:** The designed method enables quantitative benefit-risk analysis of alternative treatments using all available evidence from a network of clinical trials. The preference-free analysis can be useful in presenting the results of an MTC considering multiple outcomes.

*Key words:* Benefit-risk analysis; Multi-criteria decision aiding; Stochastic multi-criteria acceptability analysis; Network meta-analysis; Mixed treatment comparison; Second-generation antidepressants

---

## 1. Introduction

The pharmaceutical regulatory authorities and pharmaceutical health care decision makers increasingly request an explicit Benefit-Risk (BR) analysis of drugs as it can provide a basis for rational decisions when choosing a particular therapy [1]. Drug BR analysis can be used to identify trade-offs between benefit and risk, where benefit is the efficacy of a drug and risk relates to its safety [2]. If there is only one measure of efficacy and one measure of safety, the BR analysis can be conducted by plotting the joint density of the benefit and risk criteria on a plane [3]. However, there is a growing need for evidence-based pharmacotherapy to consider more than two criteria, such as multiple safety criteria, the patient's quality of life, and costs. In these cases, the two-dimensional visualization technique cannot be applied.

Multi-Criteria Decision Aiding (MCDA) methods can help by structuring the decision problem and making the

underlying value trade-offs explicit. Specifically, Tervonen et al. [4] proposed a Stochastic Multi-criteria Acceptability Analysis (SMAA) model for analyzing BR. Their model allows taking into account the probability distributions of the criteria measurements and is able to quantify the uncertainty surrounding a decision. Moreover, measurements and value judgments (preferences) are clearly separated. However, the model relies on a single trial to evaluate the comparative BR profiles of the alternatives. In most cases, a BR assessment will need to be based on evidence synthesized from multiple trials or possibly a complex network of trials.

Although evidence synthesis is most often done through pair-wise meta-analyses, they are ill-suited as a basis for a computational BR method for a number of reasons. First, relative effects have to be assessed against a common comparator, and not all evidence structures have a single treatment against which all others are compared [5]. Second, choosing a common comparator introduces a selection bias by excluding studies that do not include the comparator. Sensitivity analyses would have to be carried out for every possible choice of comparator and even then some studies might be excluded. Finally, when a large number of treatments is available, the majority of evidence may be indirect regardless of the chosen common comparator. Traditional meta-analysis does not allow these indirect comparisons to

---

\*Corresponding author: Dept. of Epidemiology, University Medical Center Groningen, PO Box 30.001, 9700 RB Groningen, The Netherlands. Tel: +31 50 361 4522.

Email addresses: g.h.m.van.valkenhoef@rug.nl (Gert van Valkenhoef), tervonen@ese.eur.nl (Tommi Tervonen), j.zhao.5@student.rug.nl (Jing Zhao), e.o.de.brock@rug.nl (Bert de Brock), j.l.hillege@tcc.umcg.nl (Hans L. Hillege), d.postmus@epi.umcg.nl (Douwe Postmus)

be taken into account.

The recently proposed Mixed Treatment Comparison (MTC) method (also known as network meta-analysis) synthesizes all the available evidence through application of a Bayesian evidence network [6, 7]. The relative effects of all included treatments are estimated using both direct and indirect evidence. In this way, the results are consistent regardless of the chosen comparator, and it is not necessary that one of the treatments has been compared with all others. Graphical summaries of MTC results have been proposed as an informal decision aid in trading effectiveness against other factors [8]. To enable the formal BR analysis of a number of alternative treatments taking into account all relevant studies, this paper proposes to apply MTC for evidence synthesis in SMAA-based multi-criteria drug BR analysis. We call this method MTC/SMAA, and for illustration, we constructed a model to evaluate the comparative BR profiles of four second-generation antidepressants and placebo using 25 studies from the literature, selected on the basis of an existing systematic review [9].

## 2. Stochastic Multi-criteria Acceptability Analysis

SMAA-2 [10] considers a discrete, multi-criteria decision problem consisting of a set of  $m$  alternatives that are evaluated in terms of  $n$  criteria. The vector of criteria measurements corresponding to alternative  $i$  is denoted by  $\xi^i = (\xi_1^i, \dots, \xi_n^i)$ , where  $\xi_k^i$  is a random variable representing the performance of alternative  $i$  on criterion  $k$ , modeled using some density function. For each criterion, a partial value function  $v_k(\xi_k^i)$  is defined to normalize the criteria measurements, so that they are represented by values between zero (the worst value) and one (the best value). The overall value function is then defined as a weighted additive combination of the partial value functions:

$$v(\xi^i, \mathbf{w}) = \sum_{k=1}^n w_k \cdot v_k(\xi_k^i) ,$$

where  $v(\xi^i, \mathbf{w}) > v(\xi^j, \mathbf{w})$  implies that alternative  $i$  is preferred to alternative  $j$  given the weight vector  $\mathbf{w}$ . The weights define relative importances of the scale swings (changes from the worst to the best criterion values), and  $w_k > w_l$  implies that if the Decision Maker (DM) would have to choose between improving either criterion  $k$  or criterion  $l$  from the worst to the best value, he or she would increase the performance on criterion  $k$ .

The DM's preferences may be unknown or partially known, and therefore the weights  $\mathbf{w}$  are also represented by a probability density. Total lack of preference information is represented by a uniform distribution in the feasible weight space. Partial information, such as importance ranking of the criteria, can easily be included by restricting the feasible weight space accordingly [10].

For given (exact) values of  $\xi$  and  $\mathbf{w}$ , the rank of each alternative is defined as an integer from the best rank ( $= 1$ ) to the worst rank ( $= m$ ) by means of a ranking function

$\text{rank}(i, \xi, \mathbf{w})$ . The main decision aiding measure is the *rank acceptability index*, denoted by  $b_r^i$ . It describes the share of all possible values of the weight vector  $\mathbf{w}$  and criteria measurements  $\xi$  for which  $\text{rank}(i, \xi, \mathbf{w}) = r$ . For example,  $b_2^5 = 0.3$  means alternative 2 has 5th-rank acceptability 0.3. The preferred (best) alternatives are those with high acceptabilities for the best ranks.

Instead of using the value function to rank the alternatives for an elicited weight vector  $\mathbf{w}$ , which is the traditional approach in multi-attribute value theory, the SMAA methods allow computing the weights a 'typical' DM supporting each alternative might have. This so-called *central weight vector*  $w_i^c$  can be presented to the DM to help him or her understand what kind of weights would favor a certain alternative  $i$ . The *confidence factor*  $p_i^c$  is the probability for alternative  $i$  to obtain the first rank when its central weight vector is chosen. The confidence factors indicate whether the criteria measurements are sufficiently accurate to discern the efficient alternatives. Low confidence factors ( $< 0.50$ ) should be interpreted with care, as then even if a DM finds the central weight vector corresponding to his or her preferences, there might be another alternative that achieves a higher first rank acceptability with those weights.

## 3. Mixed treatment comparison

The MTC method (also called network meta-analysis) synthesizes all available clinical evidence through application of a Bayesian hierarchical model [6, 7]. It enables the detection of heterogeneity (differences in studies comparing the same treatments) and inconsistency (differences between direct and indirect comparisons) in the evidence [6, 11, 12]. In this section, we briefly introduce the structure of a random effects MTC model for dichotomous data, as this type of model will be used in the case study (Section 5). For other model types and the handling of multi-arm trials, we refer to [6, 11].

Let  $i$  be a clinical trial. For each included treatment  $x$  we are given the sample size  $n_{i,x}$  and the number of events  $r_{i,x}$ , modelled as a binomial process:

$$r_{i,x} \sim \text{Bin}(p_{i,x}, n_{i,x}) ,$$

where  $p_{i,x}$  is the success probability (i.e. the *absolute risk* of an event). The risk  $p_{i,x}$  of an effect observed in the individual studies is transformed to log odds  $\theta_{i,x}$  through:

$$\theta = \text{logit}(p) = \log\left(\frac{p}{1-p}\right) .$$

The inverse transformation is given by:

$$p = \text{logit}^{-1}(\theta) = \frac{1}{1 + e^{-\theta}} .$$

The advantage of this transformation, also used in logistic regression, is that  $\theta_{i,x}$  can be assumed to be normally distributed. Moreover, if  $\theta_{i,x}$  and  $\theta_{i,y}$  are the log odds for  $x$

and  $y$ , then  $\theta_{i,x} - \theta_{i,y}$  is the log odds ratio of  $y$  compared to  $x$  in trial  $i$  (and  $e^{\theta_{i,y} - \theta_{i,x}}$  is the odds ratio).

Synthesis in MTC models is done in terms of treatment contrasts (relative effects) and not the absolute effects, as this leads to a more robust model that preserves the randomization in the trials [7]. To do this, we choose a baseline treatment  $b(i)$  for every trial  $i$ , and express the effect of  $b(i)$  as:

$$\theta_{i,b(i)} = \text{logit}(p_{i,b(i)}) = \mu_i \text{ ,}$$

and for every other treatment  $y \neq b(i)$  the effect is:

$$\theta_{i,y} = \text{logit}(p_{i,y}) = \mu_i + \delta_{i,b(i),y} \text{ ,}$$

where  $\delta_{i,b(i),y}$  is the random effect of  $y$  relative to  $b(i)$  in trial  $i$ . The random effects are related to the *relative effect* as follows:

$$\delta_{i,x,y} \sim \mathcal{N}(d_{x,y}, \sigma_{x,y}^2) \text{ ,}$$

where  $d_{x,y}$  is the relative effect of  $y$  compared to  $x$ , the parameter of interest, and  $\sigma_{x,y}^2$  is the *random effects variance*. If we set  $\sigma_{x,y}^2$  to be identical for all  $x$  and  $y$ ,  $\sigma_{x,y}^2 = \sigma^2$ , the model is a homogeneous variance model. Otherwise it is a heterogeneous variance model.

The model discussed so far is just a Bayesian formulation of pair-wise random effects meta-analysis. MTC enables the simultaneous synthesis of a network of trials through the additional assumption of *consistency*. Suppose we have three treatments, say A, B, and C, and studies comparing AB, AC, and BC. The consistency assumption then defines the relation between the relative treatment effects as

$$d_{AC} = d_{AB} + d_{BC} \text{ .}$$

A model that includes this assumption between all relative effects is a consistency model. Conclusions based on an MTC model are always derived using the consistency model. The model is estimated through stochastic simulation, e.g. using the BUGS [13] or JAGS [14] software. This enables the derivation of a point estimate and 95% credibility interval (CrI, the Bayesian analog to a confidence interval) for each of the relative effects, as well as the derivation of any other statistics of interest.

The assumption of consistency may be violated by the data at hand, in which case there exists *inconsistency*. As with pair-wise meta-analyses, the first step in dealing with inconsistency should be assessing whether the included studies are sufficiently similar to be combined. Statistical means of detecting inconsistency provide an additional safeguard against drawing conclusions from inconsistent datasets, though the lack of demonstrable inconsistency does not prove that the results are free of bias and diversity.

There are two competing methods for detecting inconsistency: inconsistency models [11] and node splitting models [12]. Inconsistency models assess inconsistency by adding inconsistency factors to closed loops in the evidence

graph, whereas in node splitting models a single comparison is chosen for which the direct and indirect evidence are contrasted. Inconsistency models have the advantage that only a single model needs to be run, but the results are often difficult to interpret. Node splitting models are easier to interpret, but require a different model to be run for each of the potentially inconsistent comparisons. Which method should be preferred is not yet clear and, in this paper, we will present the results of the node-splitting analysis because they are easier to interpret and verify them with an inconsistency model.

Inconsistency within an evidence network could reflect genuine diversity, bias or a combination of both [6]. If there is inconsistency, the reason for the inconsistency must be determined, and a clinically sound explanation must be given. If the explanation is sufficient, the offending studies are removed [11], a new inconsistency model is constructed and inconsistency evaluation is repeated until no relevant inconsistency remains. If there is considerable inconsistency that cannot be eliminated, the consistency model cannot be used. It is difficult to judge whether a certain amount of inconsistency should be considered relevant, and the debate on how to do this is ongoing [6, 11, 12].

#### 4. MTC/SMAA for BR analysis

The process of performing an MTC/SMAA analysis is shown in Figure 1. Analyzing BR based on clinical studies starts with a systematic review of the available studies relevant to the clinical domain for which BR should be assessed. In this step, which should be carried out with experts in the clinical domain, the relevant studies and important issues are identified. In the ideal case, a relevant high-quality systematic review can be found in the literature. Based on the review, the criteria to be considered are agreed upon and operationalized. Then, for each criterion, the relevant outcomes are extracted from the individual studies and inconsistency is evaluated. If there is no relevant inconsistency, a consistency model is subsequently constructed and used to create the measurements for the SMAA model. If no reasonable explanation of inconsistency is found, the whole process has to be terminated.

##### 4.1. Measurement scales

For reasons of statistical robustness, evidence synthesis methods estimate only relative effects, while absolute measures are more suitable for applying evidence to concrete decisions [15]. In a multi-criteria model, the use of absolute measures is desirable since explicit trade-offs must be made between unit increases in the scaled criteria. The problem is especially salient for dichotomous criteria, as the result in these cases is expressed as an odds ratio, which is difficult to interpret when assessing the relative importance of the scale swings between criteria. To solve this, the log odds ratio can be converted to (absolute) risk

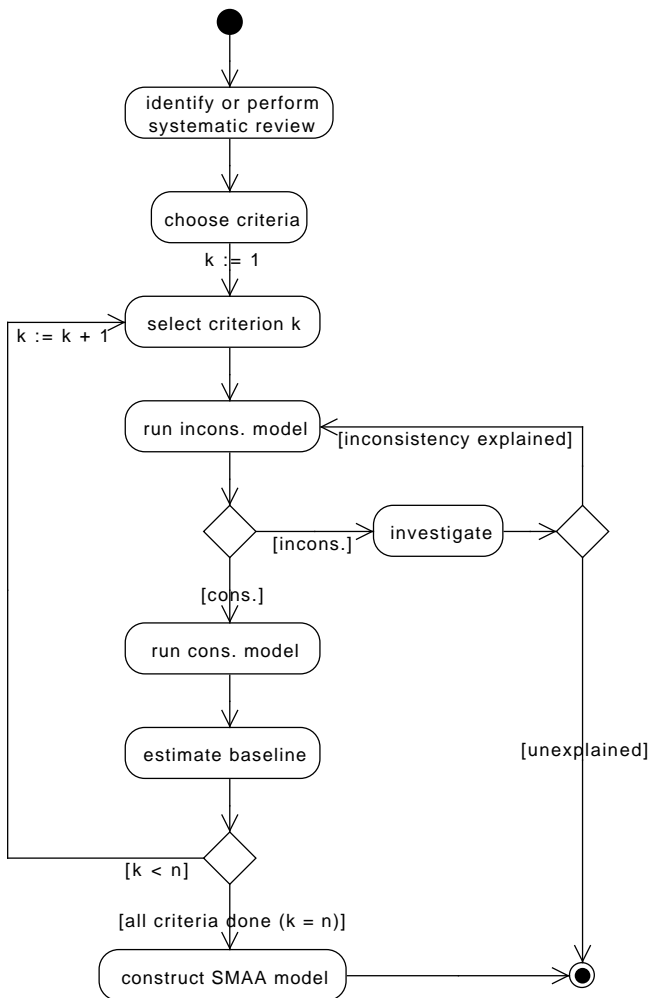


Figure 1: The process of performing an MTC/SMAA analysis (UML activity diagram notation).  $n$  is the number of criteria

by assuming a distribution for the log odds of a baseline treatment 1:

$$\theta_1 \sim \mathcal{N}(\mu, \sigma^2) .$$

Note that  $\theta_1$  is an overall estimate for treatment 1, and should not be confused with the trial-level log odds  $\theta_{i,1}$ . It does not matter which of the  $m$  included treatments is selected as the baseline. For every non-baseline treatment  $j \neq 1$ , the MTC analysis gives us the log odds ratio:

$$\begin{pmatrix} d_{1,2} \\ \vdots \\ d_{1,m} \end{pmatrix} \sim \mathcal{N} \left( \begin{pmatrix} \nu_2 \\ \vdots \\ \nu_m \end{pmatrix}, \Sigma \right) ,$$

which can be used to obtain the distribution of the non-baseline treatments' log odds conditional on  $\theta_1$ :

$$\begin{pmatrix} \theta_2 \\ \vdots \\ \theta_m \end{pmatrix} | \theta_1 \sim \mathcal{N} \left( \begin{pmatrix} \theta_1 + \nu_2 \\ \vdots \\ \theta_1 + \nu_m \end{pmatrix}, \Sigma \right) .$$

Then, for any treatment  $i$  the risk is

$$p_i = \text{logit}^{-1}(\theta_i) ,$$

as discussed in Section 3. The  $p_i$  are the measurements used in the SMAA analysis (thus  $\xi_k^i = p_i$ , where  $p_i$  is obtained for criterion  $k$ ). In the SMAA simulation, to obtain samples of the  $p_i$ 's, we first sample the baseline log odds  $\theta_1$  and then sample the log odds  $\theta_i$  for all other alternatives based on  $\theta_1$ , and transform them to risk, as given above. Note that ranking the treatments based on the  $p_i$  is equivalent to ranking them based on the  $d_{1,i}$  (with  $d_{1,1} = 0$ ), and will thus result in the same rank probabilities as from the MTC analysis if the  $d_{1,i}$  accurately reflect the posterior distribution. The rank probabilities in MTC [8] are calculated for a single criterion and are therefore distinct from the rank acceptabilities discussed in Section 2, which incorporate trade-offs between multiple criteria.

Different methods can be used to arrive at a sensible assumption for the baseline log odds  $\theta_1$ . One could use an observational effectiveness study with a suitable population, let a clinical expert provide estimates, or attempt to derive them from the included trials. In this paper, we will apply arm-based pooling of the placebo arms. This is supplemented by a visual assessment (through a forest plot) of the effects found in the individual studies.

Since the risk scale is bound to  $[0, 1]$ , either  $v_k(\xi_k^i) = \xi_k^i$  or  $v_k(\xi_k^i) = 1 - \xi_k^i$  can be used as the partial value function  $v_k$  for any dichotomous criterion  $k$ , respectively when more or less events are preferred (see Section 2). We will return to the advantages and disadvantages of this approach in the discussion.

## 5. Application to antidepressants

To illustrate the use of MTC/SMAA, we used an existing systematic review [9] to create a model for evaluating

the comparative BR profiles of four second-generation antidepressants (fluoxetine, paroxetine, sertraline and venlafaxine) and placebo. The application is meant as an example, and the results should be interpreted with care. A full BR analysis of antidepressants should ideally be based on a more recent systematic review that explicitly includes placebo-controlled studies. This is even more important in the light of recent doubt on the efficacy of antidepressants [16]. Even if we consider the efficacy of antidepressants to be proven, in the context of a multi-criteria decision model consideration of other factors may imply that placebo is the best option, as the placebo response in depression trials is considerable [17].

### 5.1. Previous work

The review included 46 studies comparing 10 second-generation antidepressants on the Hamilton Rating Scale for Depression (HAM-D) or Montgomery-Asberg Depression Rating Scale (MADRS). In total, 20 comparisons were made in the included studies (out of 45 possible comparisons). Meta-analysis was applied for just 3 comparisons using 16 studies in total. All meta-analyses assessed efficacy (50% or greater improvement from baseline on the HAM-D or MADRS scale) relative to fluoxetine, and studies between the other drugs (paroxetine, sertraline and venlafaxine) were not considered. Meta-analysis yielded risk ratios relative to fluoxetine, with a significant but small additional effect for sertraline and venlafaxine. The authors concluded that the four antidepressants did not differ substantially for treatment of major depressive disorder. A more recent review [18] used an MTC analysis to show that there are differences among second-generation antidepressants in terms of efficacy and the proportion of patients completing the study.

### 5.2. Methods

An MTC/SMAA analysis was performed to compare fluoxetine, paroxetine, sertraline, venlafaxine, and placebo on one benefit criterion (efficacy) and five risk criteria. Efficacy was assessed by means of treatment response, defined as a 50% or greater improvement on the HAM-D rating scale for depression. The five risk criteria corresponded to the most common Adverse Drug Reactions (ADRs): diarrhea, dizziness, headache, insomnia, and nausea. All of the criteria were measured in terms of absolute risk, based on dichotomous data from the included trials.

As [9] did not include sufficient information to construct the MTC models, we did not take the measurements directly from the review, but used the included individual studies to perform our own analysis. Although the review did not consider placebo, sufficient studies with a placebo arm were present to include it in the analysis. The papers included in the review were retrieved and the data extracted. We used the drugis.org MTC software (<http://drugis.org/mtc>) [19] to generate MTC models for the 25 studies (see Table 1 and Figure 2) comparing fluoxetine, paroxetine, sertraline, venlafaxine, and placebo.

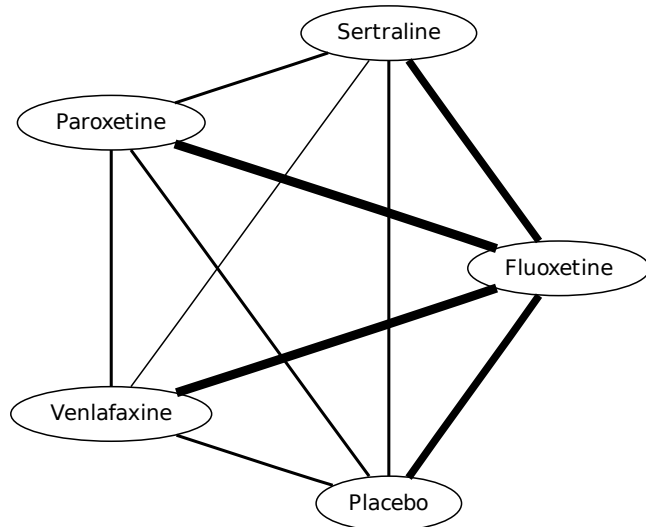


Figure 2: Evidence network of studies comparing the four included second-generation antidepressants and placebo. The width of the lines indicates the number of studies that include that comparison (the minimum is 1 and the maximum 6)

We used the homogeneous variance assumption and specified a uniform prior  $\sigma \sim \mathcal{U}(0, 4)$  for the random effects variance. For the trial baseline effects  $\mu_i$  and random effects  $\delta_{i,b(i),y}$  we specified a  $\mathcal{N}(0, 10^3)$  prior. Markov Chain Monte Carlo simulation with 4 parallel chains of 30,000 tuning and 20,000 simulation iterations each was used to estimate each MTC model, and the computations were done using JAGS [14] and R [20]. Inconsistency was primarily assessed using node-splitting models [12] and inconsistency models [11] were run as a secondary analysis. Convergence was assessed using the Brooks-Gelman-Rubin diagnostic [21], where a potential scale reduction factor of 1.05 or lower was considered sufficient if visual inspection of the convergence plots and time-series also indicated convergence.

We constructed a SMAA model with the measurements derived from the consistency models, and baseline estimates derived from the trials and discussed with an expert. The SMAA model was computed using R with 10,000 Monte Carlo iterations giving sufficient accuracy for the indices [22]. The SMAA analyses were performed for three scenarios: one with missing preference information and two with a criteria ranking elicited from the expert: mild and severe depression. The data files are available online at <http://drugis.org/network-br>. There we also provide a JSMAA [23] v0.8.4 model that allows the reader to explore the trade offs in an interactive graphical user interface.

Table 1: The number of studies included in the network meta-analysis for each criterion

Criterion	Placebo	Fluoxetine	Paroxetine	Sertraline	Venlafaxine	Total
HAM-D	8	18	9	8	9	24
Diarrhea	5	11	7	8	5	17
Dizziness	3	9	4	5	6	12
Headache	5	12	8	8	6	19
Insomnia	7	12	8	6	6	18
Nausea	6	15	9	8	8	22
Total	8	18	10	9	9	25

### 5.3. Results

*Inconsistency analysis.* The node-splitting analysis of inconsistency revealed two potential problems at the  $\alpha = 0.05$  significance level, though given that there were 56 comparisons, it is to be expected that some are significant due to chance. However, we chose not to correct the threshold a priori, but rather to investigate these two cases. One occurred in the headache network, where one split node was significant, and the other in the nausea network, where two directly related split nodes were significant. In neither of these cases could we identify any systematic differences between the studies, and as the number of significant findings is compatible with chance, we decided to continue on the basis of consistency models including all studies. The studies involved in these comparisons did not lead to inconsistencies in the other evaluated networks, and the secondary analysis using inconsistency models did not indicate any inconsistencies.

*Consistency analysis.* The results of the consistency analysis are visualized as forest plots for the odds ratio relative to placebo in Figure 3. Including indirect evidence leads to somewhat smaller 95% credibility intervals for treatment response than pair-wise meta-analysis. Therefore the evidence from the studies additionally included in the MTC model discriminate the drugs better with respect to efficacy.

*Preference-free model.* Baseline estimates were derived by random-effects pooling of the placebo arms (Table 2) and discussed with an expert, who compared them to sources known to him and did not contest the values or the method used to derive them. He did note that these values are expected to vary greatly between trials, and that this fact is reflected in the width of the confidence intervals.

The rank-acceptabilities with missing preferences are shown in Figure 4. There is a large share of the possible preferences for which placebo attains rank 1. From the central weights (Figure 5), it is estimated that the preference scenarios that are favorable to placebo have a low weight for efficacy, and that a ‘typical’ DM that would choose placebo implicitly finds each of the ADRs to be about twice as important as efficacy. Placebo is also the only alternative to attain a confidence factor close to 1 (Table 5).

Table 2: Baseline measurements derived from the placebo trials, given as mean  $\pm$  standard error for the log-odds and the corresponding median and 95% CrI of the resulting logit-normal distribution for the absolute risk

Criterion	Parameters	Risk (95% CrI)
HAM-D	$-0.17 \pm 0.11$	0.46 (0.40, 0.51)
Diarrhea	$-2.19 \pm 0.21$	0.10 (0.07, 0.14)
Dizziness	$-2.23 \pm 0.61$	0.10 (0.03, 0.26)
Headache	$-1.20 \pm 0.29$	0.23 (0.15, 0.35)
Insomnia	$-2.61 \pm 0.19$	0.07 (0.05, 0.10)
Nausea	$-2.02 \pm 0.19$	0.11 (0.08, 0.16)

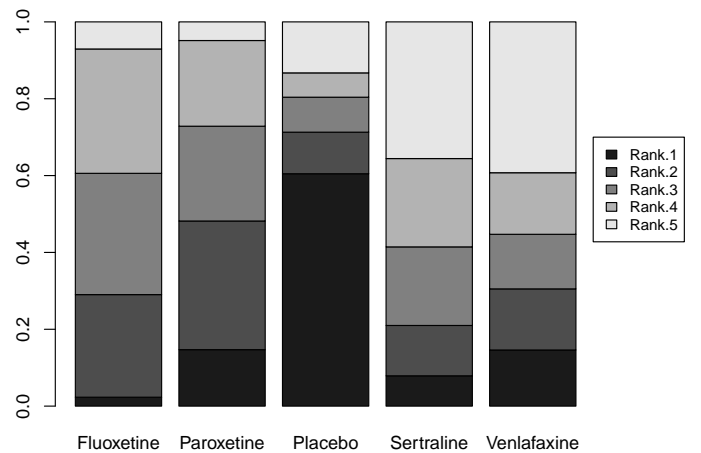
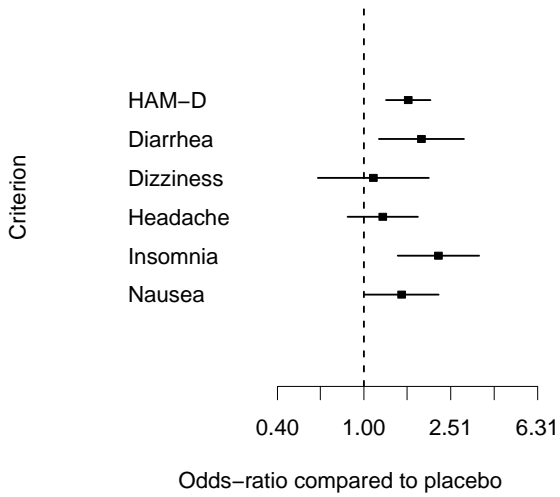
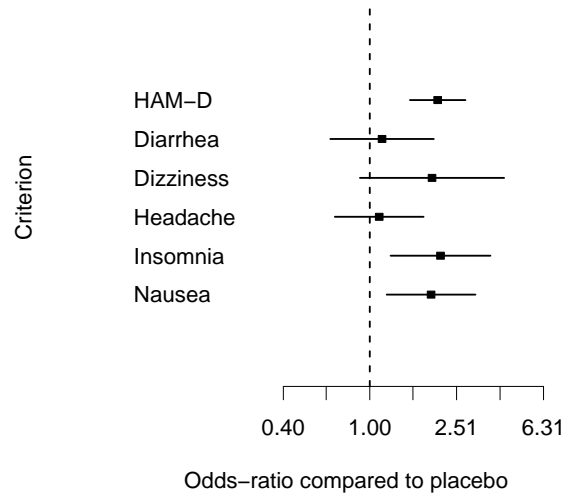


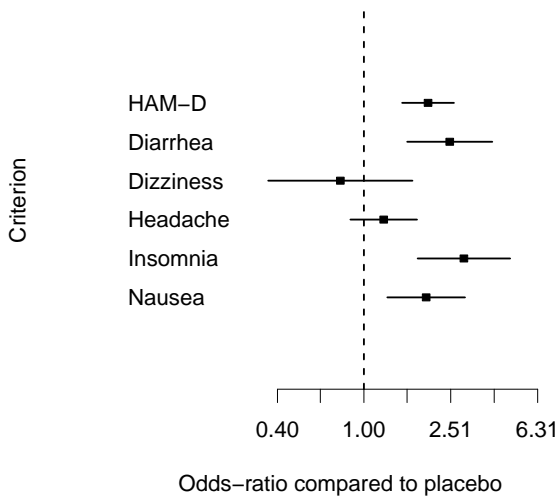
Figure 4: Rank acceptabilities for the preference-free model



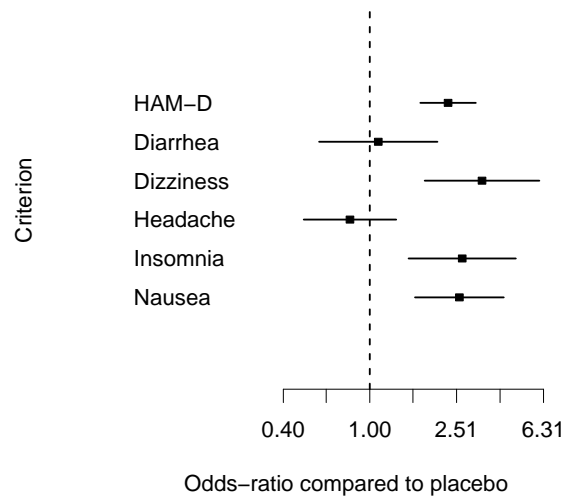
(a) Fluoxetine



(b) Paroxetine



(c) Sertraline



(d) Venlafaxine

Figure 3: Network meta-analysis results: odds ratios relative to placebo, with 95% CrI. Results to the right of the no-effect line indicate a higher incidence for the active treatment

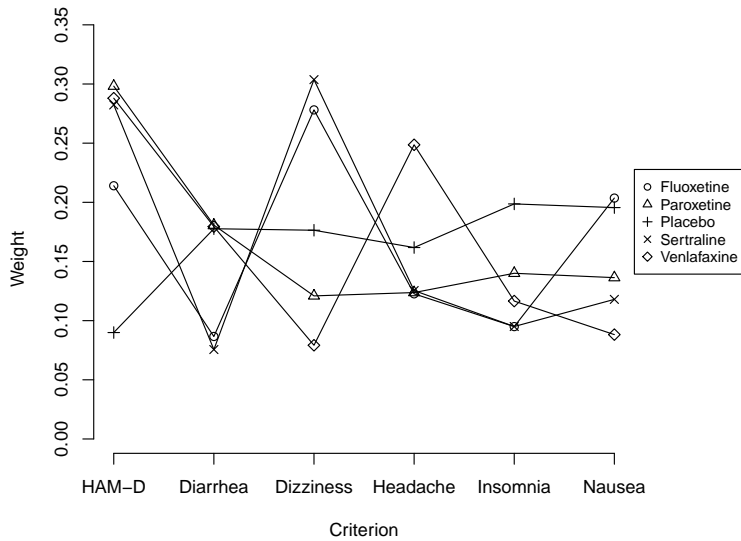


Figure 5: Central weights for the preference-free model

Fluoxetine has a low confidence factor (0.12) for its central weights, and in fact given its central weights, other alternatives have a higher first-rank acceptability. Thus, fluoxetine is likely to be dominated by the other alternatives. In general, if efficacy is highly valued, placebo is unlikely to be the best option, but it is difficult to choose a drug based on the data.

*Mild depression.* Preferences for the mild depression scenario were elicited from the expert using ordinal swing weighting. This resulted in the following ranking of the criteria: Insomnia  $\succ$  HAM-D  $\succ$  Dizziness  $\succ$  Nausea  $\succ$  Diarrhea  $\succ$  Headache. The rank acceptabilities for this scenario are shown in Figure 6. Placebo obtains the highest first-rank acceptability (0.56), followed by paroxetine (0.28), while venlafaxine has the highest last-rank acceptability (0.62), followed by sertraline (0.24). Clearly, the high incidence of both insomnia and dizziness are unfavorable to venlafaxine given the preferences. Only placebo, fluoxetine, and paroxetine have  $> 0.5$  probability of being among the best 3, and only placebo and paroxetine have  $< 0.5$  probability of being among the worst 3.

*Severe depression.* The preferences elicited for this scenario differed only in that the Insomnia and HAM-D criteria were swapped. The rank acceptabilities for severe depression are shown in Figure 7. As would be expected based on the central weights analysis with missing preferences, ranking HAM-D as the most important criterion reverses the situation for placebo, which now has only 0.09 first-rank acceptability, and 0.56 last-rank acceptability. Placebo is also the only alternative to have  $< 0.5$  probability of being among the best 3. Paroxetine has the highest first-rank acceptability (0.47) and paroxetine and sertraline are the only alternatives that have  $< 0.5$  probability of being among the worst 3.

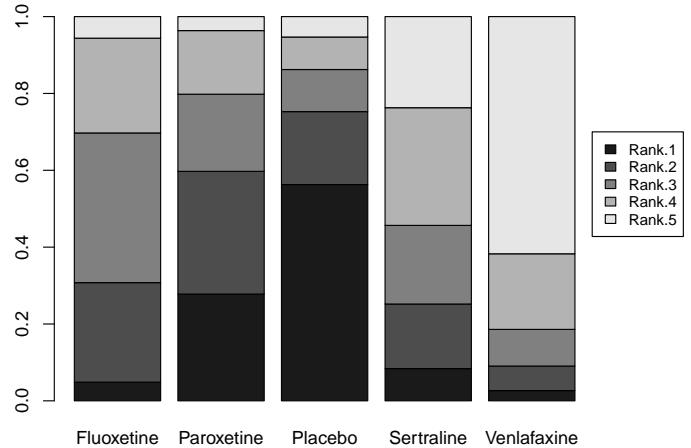


Figure 6: Rank acceptabilities for the mild depression scenario

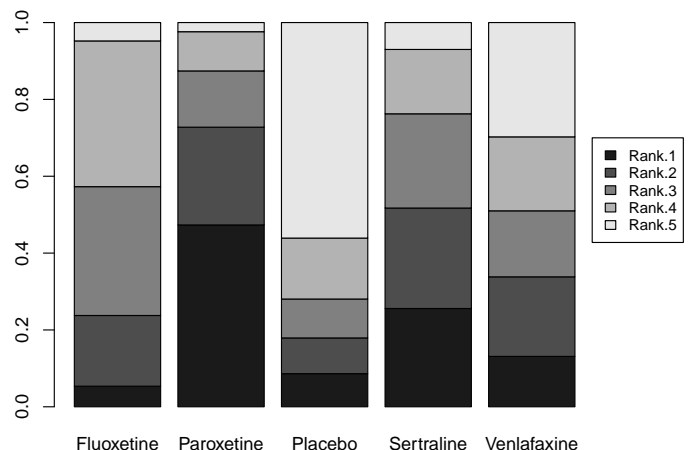


Figure 7: Rank acceptabilities for the severe depression scenario

Table 3: Central weights and confidence factors (CFs) for the preference-free model

Alternative	CF	HAM-D	Diarrhea	Dizziness	Headache	Insomnia	Nausea
Fluoxetine	0.12	0.21	0.09	0.28	0.12	0.09	0.20
Paroxetine	0.57	0.30	0.18	0.12	0.12	0.14	0.14
Placebo	0.99	0.09	0.18	0.18	0.16	0.20	0.20
Sertraline	0.55	0.28	0.08	0.30	0.13	0.10	0.12
Venlafaxine	0.63	0.29	0.18	0.08	0.25	0.12	0.09

## 6. Discussion

Pharmacological decision making is a complex domain in which decisions regarding multiple criteria are informed by complex evidence networks consisting of heterogeneous clinical studies. This paper introduced MTC/SMAA, which uses the MTC evidence synthesis method together with SMAA to assess multi-criteria BR trade-offs while taking into account all available evidence from clinical trials.

The MTC/SMAA method has four main advantages. First, MTC/SMAA allows taking into account all the available evidence no matter whether the treatments are directly or indirectly related. Second, a group of treatments without a common comparator can be analyzed, which is impossible with pair-wise evidence synthesis methods. Third, inconsistencies in the evidence structure due to incompatible study design can be detected early in the analysis and systematically removed if the inconsistency is judged to be clinically relevant. Fourth, application of SMAA enables explicit assessment of trade-offs that exist between the criteria and provides valuable insights even if the DMs are not willing or able to provide exact preferences.

### 6.1. Case study

We illustrated the MTC/SMAA method with a case study on second-generation antidepressants. Although the case study is indicative of the method’s feasibility, further work should evaluate the model in other therapeutic groups.

As we demonstrated in the example, a preference-free analysis of the central weight vectors can provide substantial insight into trade-offs between the treatments under consideration. As such, a SMAA central weights analysis of the most important outcomes could be a valuable addition to any (network) meta-analysis. It allows drawing firmer conclusions on which treatments are likely to be most suited to specific situations, and which treatments are unlikely to be the best in any situation. The mild and severe preference scenarios showed that for severe depression, treatment with an antidepressant is warranted, but for mild depression this is not clear. Recent research suggests that placebo may be effective even without deception [24] (in irritable bowel syndrome), so it may be worthwhile to explore this option for mildly depressed patients. The analyses also suggest that fluoxetine is unlikely to be the best among the five alternatives.

However, the data do not conclusively distinguish the alternatives, especially the active treatments, and given the amount of data it is likely that much of this uncertainty is inherent to the field, especially when distinguishing the active treatment options. Some improvement may be possible by eliciting more precise weights. However, except for placebo in the mild depression scenario, making the weights more precise within the constraints imposed by the ordinal preferences elicited from the expert would not allow much more conclusive results as the data have a high degree of uncertainty.

Compared to the systematic review on which we based the case study, the MTC/SMAA analysis explicitly takes into account the ADRs in addition to efficacy and gives a clearer picture of the strengths and weaknesses of the alternatives. Including placebo in the analysis provides further insight into the trade-offs. Moreover, the model can quantitatively support the statement, also made in the original review, that it is difficult to choose among the four considered antidepressants.

### 6.2. Limitations and future work

The main challenges in applying MTC/SMAA are the evaluation of inconsistency and estimation of baseline effects. Assessing inconsistency is especially difficult in cases where many potential inconsistencies have to be considered (large evidence networks or many different criteria) since significant results may also arise by chance. Different methods to assess inconsistency have been proposed [6, 11, 12], and general consensus on the best method has not yet been reached. The second concern is the scale employed for the criteria measurements. We developed a procedure for converting the relative scales from evidence synthesis to absolute ones to be used in decision making using minimal information. However, baseline effects have to be estimated, and further work is necessary to identify the best way to do this.

Another consideration is the scale on which criteria are evaluated in preference elicitation. In contrast to the previous work on SMAA for BR analysis [4], we choose to use the full  $[0, 1]$  scale instead of the hull of the 95% confidence intervals. This has the advantage that trade offs are easier to evaluate, and that introducing additional alternatives does not require re-eliciting the preferences. The disadvantage of this approach is that a stronger linearity assumption on the partial value functions is required (see [4]). This limitation is especially important when the

observed frequencies differ greatly, e.g. when a trade off between high efficacy and rare but serious adverse events needs to be made. In those cases the scales should be assessed using the confidence interval hull. Of course, for scales that do not have natural bounds (e.g. weight gain in kg) the confidence interval hull approach is the only viable option.

In the current work we applied a SMAA decision model based on additive value functions. Although the additive model is widely applied and reasonably easy to understand, we acknowledge that other approaches are possible. For example, Data Envelopment Analysis (DEA) models have been commonly applied in cost-benefit analyses outside the area of healthcare, and there is also a SMAA variant for DEA, the SMAA-D [25]. Future work should assess whether other simulation-based decision models are applicable in the context of drug BR analysis.

Finally, our model is based only on criteria that are measured in clinical trials, which is appropriate in the context of health policy decision making. However, other criteria may need to be considered, such as cost in reimbursement decisions, or the route of administration in prescription decisions. While we did not consider such criteria, they would not be difficult to include in an MTC/SMAA analysis.

## Acknowledgements

This study was performed in the context of the Escher project (T6-202), a project of the Dutch Top Institute Pharma.

## References

- [1] W. L. Holden, Benefit-risk analysis: a brief review and proposed quantitative approaches, *Drug Safety* 26 (2003) 853–862. doi:10.2165/00002018-200326120-00002.
- [2] N. Victor, J. Hasford, Risk-benefit analyses of drugs: fundamental considerations and requirements from the point of view of the biometrician. Problems in the assessment of the combination of trimethoprim with sulfamethoxazole, *Infection* 15 (1987) 236–240. doi:10.1007/BF01643196.
- [3] L. D. Lynd, B. J. O'Brien, Advances in risk-benefit evaluation using probabilistic simulation methods: an application to the prophylaxis of deep vein thrombosis, *Journal of Clinical Epidemiology* 57 (8) (2004) 795–803. doi:10.1016/j.jclinepi.2003.12.012.
- [4] T. Tervonen, G. van Valkenhoef, E. Buskens, H. L. Hillege, D. Postmus, A stochastic multi-criteria model for evidence-based decision making in drug benefit-risk analysis, *Statistics in Medicine* 30 (12) (2011) 1419–1428. doi:10.1002/sim.4194.
- [5] G. Salanti, F. K. Kavvoura, J. P. A. Ioannidis, Exploring the geometry of treatment networks, *Annals of Internal Medicine* 148 (7) (2008) 544–553.
- [6] G. Salanti, J. P. T. Higgins, A. E. Ades, J. P. A. Ioannidis, Evaluation of networks of randomized trials, *Statistical Methods in Medical Research* 17 (3) (2008) 279–301. doi:10.1177/0962280207080643.
- [7] G. Lu, A. E. Ades, Combination of direct and indirect evidence in mixed treatment comparisons, *Statistics in Medicine* 23 (20) (2004) 3105–3124. doi:10.1002/sim.1875.
- [8] G. Salanti, A. E. Ades, J. P. A. Ioannidis, Graphical methods and numerical summaries for presenting results from multiple-treatments, *Journal of Clinical Epidemiology* 64 (2) (2011) 163–171. doi:10.1016/j.jclinepi.2010.03.016.
- [9] R. A. Hansen, G. Gartlehner, K. N. Lohr, B. N. Gaynes, T. Carey, Efficacy and safety of second-generation antidepressants in the treatment of major depressive disorder, *Annals of Internal Medicine* 143 (6) (2005) 415–426.
- [10] R. Lahdelma, P. Salminen, SMAA-2: Stochastic multicriteria acceptability analysis for group decision making, *Operations Research* 49 (3) (2001) 444–454. doi:10.1287/opre.49.3.444.11220.
- [11] G. Lu, A. E. Ades, Assessing evidence inconsistency in mixed treatment comparisons, *Journal of the American Statistical Association* 101 (474) (2006) 447–459. doi:10.1198/016214505000001302.
- [12] S. Dias, N. J. Welton, D. M. Caldwell, A. E. Ades, Checking consistency in mixed treatment comparison meta-analysis, *Statistics in Medicine* 29 (7-8, Sp. Iss. SI) (2010) 932–944. doi:10.1002/sim.3767.
- [13] D. Spiegelhalter, A. Thomas, N. Best, D. Lunn, WinBUGS User Manual, version 1.4 (January 2003). URL <http://www.mrc-bsu.cam.ac.uk/bugs>
- [14] M. Plummer, JAGS Version 1.0.3 manual (April 2009). URL <http://www.fis.iarc.fr/~martyn/software/jags>
- [15] M. Egger, G. D. Smith, A. N. Phillips, Meta-analysis: principles and procedures, *BMJ* 315 (7121) (1997) 1533–1537.
- [16] H. E. Pigott, A. M. Leventhal, G. S. Alter, J. J. Boren, Efficacy and effectiveness of antidepressants: current status of research, *Psychotherapy and Psychosomatics* 79 (2010) 267–279. doi:10.1159/000318293.
- [17] J. G. Storosum, A. J. Elferink, B. J. van Zwieten, W. van den Brink, J. Huyser, Natural course and placebo response in short-term, placebo-controlled studies in major depression: a meta-analysis of published and non-published studies, *Pharmacopsychiatry* 37 (2004) 32–36. doi:10.1055/s-2004-815472.
- [18] A. Cipriani, T. A. Furukawa, G. Salanti, J. R. Geddes, J. P. T. Higgins, R. Churchill, N. Watanabe, A. Nakagawa, I. M. Otori, H. McGuire, M. Tansella, C. Barbui, Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis, *The Lancet* 373 (9665) (2009) 746 – 758. doi:10.1016/S0140-6736(09)60046-5.
- [19] G. van Valkenhoef, T. Tervonen, B. de Brock, H. L. Hillege, Algorithmic parametrization of mixed treatment comparisons, *Statistics and Computing* (in press).
- [20] R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0 (2008). URL <http://www.R-project.org>
- [21] S. P. Brooks, A. Gelman, General methods for monitoring convergence of iterative simulations, *Journal of Computational and Graphical Statistics* 7 (4) (1998) 434–455.
- [22] T. Tervonen, R. Lahdelma, Implementing stochastic multicriteria acceptability analysis, *European Journal of Operational Research* 178 (2) (2007) 500–513. doi:10.1016/j.ejor.2005.12.037.
- [23] T. Tervonen, JSMAA: an open source software for SMAA computations, in: C. Henggeler Antunes, D. Rios Insua, L. Dias (Eds.), Proceedings of the 25th Mini-EURO conference on Uncertainty and Robustness in Planning and Decision Making (URPDM2010), Coimbra, Portugal, 2010. URL <http://drugis.org/files/tervonen-urpdm2010.pdf>
- [24] T. J. Kaptchuk, E. Friedlander, J. M. Kelley, M. N. Sanchez, E. Kokkotou, J. P. Singer, M. Kowalczykowski, F. G. Miller, I. Kirsch, A. J. Lembo, Placebos without deception: A randomized controlled trial in irritable bowel syndrome, *PLoS ONE* 5 (12) (2010) e15591. doi:10.1371/journal.pone.0015591.
- [25] R. Lahdelma, P. Salminen, Stochastic multicriteria acceptability analysis using the data envelopment model, *European Journal of Operational Research* 170 (1) (2006) 241–252. doi:10.1016/j.ejor.2004.07.040.