PreTA: A network meta-analysis ranking metric measuring the probability of being preferable than the average treatment

Adriani Nikolakopoulou[*,1], Dimitris Mavridis[2,3], Virginia Chiocchia[1], Theodoros Papakonstantinou[1], Toshi A Furukawa[4] and Georgia Salanti[1]

[1] Institute of Social and Preventive Medicine (ISPM), University of Bern, Bern, Switzerland

[2] Department of Primary Education, University of Ioannina, Ioannina, Greece

[3] Faculté de Médecine, Université Paris Descartes, Paris, France

[4] Departments of Health Promotion and Human Behavior and of Clinical Epidemiology, Kyoto University Graduate School of Medicine/School of Public Health, Kyoto, Japan

[*]Corresponding author: e-mail: adriani.nikolakopoulou@ispm.unibe.ch, Phone: +41 31 631 35 15

Abstract

Background: Network meta-analysis (NMA) produces complex outputs as many comparisons between interventions are of interest and a treatment ranking is often included in the aims of the evidence synthesis. The estimated relative treatment effects are usually displayed in a forest plot or in a league table and several ranking metrics are calculated and presented, such as the median and mean treatment ranks.

Methods: We estimate relative treatment effects of each competing treatment against a fictional 'average' treatment using the 'deviation from the means' coding that has been used to parametrize categorical covariates in regression models. Based on this alternative parametrization of the NMA model, we present a new ranking metric (PreTA: Preferable Than Average) interpreted as the probability that a treatment is better than a fictional treatment of average performance.

Results: We compare PreTA with existing probabilistic ranking metrics in 232 networks of interventions. We use two networks of interventions, a network of 18 antidepressants for acute depression and a network of four interventions for heavy menstrual bleeding, to illustrate the methodology. The agreement between PreTA and existing ranking metrics depends on the precision with which relative effects are estimated.

Conclusions: PreTA is a viable alternative to existing ranking metrics which can be interpreted as the probability of being better than the 'average' treatment. It enriches the decision-making arsenal with a ranking metric which is interpreted as a probability and considers the entire ranking distributions of the involved treatments.

Keywords: Alternative parametrization; Deviation from means; Indirect evidence; Probabilistic ranking; Treatment hierarchy

1 Introduction

Results from network meta-analysis (NMA) are often used to inform health-care decision making and their presentation in a coherent and understandable way is of critical importance (1,2). The main output of NMA is a set of relative effects between all treatments, which has been produced by combining direct and indirect evidence in a network of trials comparing different treatments (3,4). A very informative way of presenting the NMA relative treatment effects is in a league table, where the names of the treatments are presented in the diagonal and each cell contains the relative treatment effect (5). Such a table allows for the simultaneous presentation of two outcomes, or of the results from pairwise and network meta-analysis, below and above the diagonal.

Although the set of relative effects contains all the information produced from NMA, a treatment hierarchy is often of interest to decision makers and end users. To this aim, alongside treatment effect estimates, several ranking metrics have been proposed to present NMA results. Ranking probabilities of each treatment being at each possible rank are calculated using simulation or resampling techniques either in a Bayesian or in a frequentist framework. Other ranking metrics include the surface under the cumulative ranking curve (SUCRA), that averages across all ranking probabilities for each treatment, and its frequentist analogue, P-score, which is calculated analytically (6,7). SUCRA and P-score can be interpreted as the mean extent of certainty that a treatment is better than all the other treatments. As authors of (6) point out, however, *"it is impossible to tell what constitutes a modest or large difference in SUCRA between two treatments, either statistically or clinically"*. An alternative way to produce a treatment hierarchy is to simply rank treatments according to the relative effects versus placebo, or another reference treatment. However, this hierarchy either does not take into account uncertainty (by considering only point estimates) or depends a lot on the uncertainty around the reference treatment.

In this paper, we develop a probabilistic ranking metric that naturally incorporates uncertainty and is a viable alternative to existing ranking metrics. We re-parametrize the NMA model to derive treatment effects against a fictional treatment of average performance using the deviation of means coding that has been used to parametrize categorical covariates in regression models (8). Then, we use the derived treatment effects to compute the probability of each treatment being better than the 'average' treatment. This ranking metric

aids the interpretation of NMA results in classifying treatments as superior, equivalent and inferior to an imaginary 'average' treatment.

2 Reparametrization of the NMA model

2.1 Deviation from means coding in regression models

We start with a short description of the deviation from means coding in regression models as described by Hosmer and Lemeshow (8). This is an alternative parametrization to the most common 'reference cell coding' in order to avoid the use of a reference level. According to the reference cell coding, a categorical independent variable with C categories is expressed through $C-1$ dummy/indicator variables.

Consider, for example, that we aim to estimate the effect of a covariate with four groups on the probability of an event. We fit a logistic regression model

$$g(p(\boldsymbol{x})) = \gamma_0 + \gamma_1 \boldsymbol{x_1} + \gamma_2 \boldsymbol{x_2} + \gamma_3 \boldsymbol{x_3}$$

where $\boldsymbol{x} = (\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3})'$ are the dummy variables for the covariate and $g(p(\boldsymbol{x}))$ is the logit link function $g\big(p(\boldsymbol{x})\big) = logit\big(p(\boldsymbol{x})\big) = log(p(\boldsymbol{x})/(1-p(\boldsymbol{x})))$ with $p(\boldsymbol{x})$ indicating the probability of event.

According to the reference cell coding, the indicator variables are parametrized as shown in Table 1 and result into estimating logarithms of the relative odds ratios (logOR) between the categories represented by the values 0 and 1 in these indicator variables.

According to the alternative deviation from means coding, the indicator variables express effects as deviations between each category mean (here the logit of the outcome in that category) from the overall (grand) mean (here the average logit outcome over all categories). We re-write the model as

$$g(p(\boldsymbol{x})) = \gamma_0^* + \gamma_1^* \boldsymbol{x_1^*} + \gamma_2^* \boldsymbol{x_2^*} + \gamma_3^* \boldsymbol{x_3^*}$$

where the indicator variables x_1^*, x_2^*, x_3^* are defined as shown in Table 1. The model results in estimating the coefficients $\hat{\gamma}_1^*, \hat{\gamma}_2^*, \hat{\gamma}_3^*$, interpreted as the relative effects among groups versus the average effect across all groups. Note that the exponential of the coefficients $\hat{\gamma}_1^*, \hat{\gamma}_2^*, \hat{\gamma}_3^*$ are not odds ratios because in the denominator is the average odds that includes the odds of the numerator.

For example, the logit of group 2 versus average logit over all groups is derived as

$$\log(OR) = \hat{\gamma}_0^* + \hat{\gamma}_1^*(1) + \hat{\gamma}_2^*(0) + \hat{\gamma}_3^*(0) - \left(\hat{\gamma}_0^* + \hat{\gamma}_1^*(0) + \hat{\gamma}_2^*(0) + \hat{\gamma}_3^*(0)\right) = \hat{\gamma}_1^*$$

For further information and examples on the deviation from means coding, see (8).

2.2 Notation for the NMA model

In this section, we introduce some general notation for the NMA model. Let the entire evidence base consist of $i = 1, \ldots, n$ studies forming a set of treatments, denoted as $k = 1, \ldots, K$. The number of treatments in study i is denoted as K_i. Index j denotes a treatment contrast. A core assumption in NMA is that of transitivity, which implies that in a network of K treatments, and subsequently $\binom{K}{2}$ possible relative treatment effects, only $K - 1$ need to be estimated and the rest are derived as linear combinations of those (9,10). The target parameter is therefore a vector $\boldsymbol{\mu}$ of $K - 1$ relative treatment effects $\mu_2, \mu_3, \ldots \mu_K$, called the vector of basic parameters (11,12).

With arm-level data we can model arm level parameters, for example the event probability for a binary outcome, in study i and treatment arm k denoted as y_{ik}(13). A link function $g(y_{ik})$ maps the parameters of interest onto a scale ranging from minus to plus infinity and u_i are the trial-specific baselines. For an overview of commonly used link functions in meta-analysis see (14). All arm-level parameters y_{ik} across studies are collected in a vector $\boldsymbol{y^a}$ of length $\sum_{i=1}^{n} K_i$, where superscript a stands for 'arm-level'.

With contrast-level data we model trial specific summaries, for example logOR, log risk ratio, mean difference or standardized mean difference (13). Let y_{ij} be the observed effect size for treatment contrast j in study i. The vector of the estimated contrasts across all studies is denoted as $\boldsymbol{y^c}$ and is of length $\sum_{i=1}^{n}(K_i - 1)$. The superscript $\boldsymbol{c}$ indicates the fact that 'contrast-level' data are modeled.

We will first describe the arm-level and then the contrast-level NMA models using reference cell coding and the equivalent alternative deviation from the means parametrization, which allows estimation of all treatments versus a fictional treatment of average performance. We will exemplify the models using a hypothetical network of three treatments, A, B and C examined in four studies, one comparing A and B, one comparing A and C, one comparing B and C and one three-arm study comparing treatments A, B and C. The target vector of basic parameters is usually taken to include the relative effects of all treatments versus an arbitrary

reference, here treatment A, and hence is $\boldsymbol{\mu} = \begin{pmatrix} \mu_{AB} \\ \mu_{AC} \end{pmatrix}$. The transitivity assumption implies consistency between relative treatment effects; in particular, it holds that

$$\mu_{BC} = \mu_{AC} - \mu_{AB}.$$

2.3 NMA with arm-level data

2.3.1 Reference cell coding

The model for study 1, comparing treatments A and B is shown in Table 2; $\delta_{1,AB}$ denotes the random effect of study 1 for the comparison AB and τ^2 denotes heterogeneity. It is customary to assume that heterogeneity is common across comparisons. The model is straightforwardly generalized for the other three studies (Table 2).

In its general form, the NMA model using arm-based analysis can be written as

$$\boldsymbol{g(y^a)} = \boldsymbol{Zu} + \boldsymbol{X^a\mu} + \boldsymbol{W\delta}$$

Equation 1

where $\boldsymbol{u}$ is the vector of baselines u_i of length n, which can be assumed to be either fixed and unrelated to each other, or exchangeable drawn from a normal distribution (15). We assume fixed and unrelated baseline effects for the remainder of this paper. Vector $\boldsymbol{\delta}$ includes the study random effects $\delta_{i,j}$ and follows the multivariate normal distribution

$$\boldsymbol{\delta} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$$

Matrix $\boldsymbol{\Sigma}$ is a block-diagonal between-study variance-covariance matrix of dimensions $\{\sum_{i=1}^{n}(K_i - 1)\} \times \{\sum_{i=1}^{n}(K_i - 1)\}$. The matrices $\boldsymbol{Z}, \boldsymbol{X^a}, \boldsymbol{W}$ are design matrices linking the vector of baselines, basic parameters and random effects respectively with $\boldsymbol{g(y^a)}$. The construction of these design matrices depends on the modeled arm-level parameters y_{ik} and is exemplified in the following example.

For the example of Table 2, Equation 1 takes the form

$$\begin{pmatrix} g(y_{1A}) \\ g(y_{1B}) \\ g(y_{2A}) \\ g(y_{2C}) \\ g(y_{3B}) \\ g(y_{3C}) \\ g(y_{4A}) \\ g(y_{4B}) \\ g(y_{4C}) \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \\ -1 & 1 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_{AB} \\ \mu_{AC} \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \delta_{1,AB} \\ \delta_{2,AC} \\ \delta_{3,BC} \\ \delta_{4,AB} \\ \delta_{4,AC} \end{pmatrix}$$

with

$$\begin{pmatrix} \delta_{1,AB} \\ \delta_{2,AC} \\ \delta_{3,BC} \\ \delta_{4,AB} \\ \delta_{4,AC} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau^2 & 0 & 0 & 0 & 0 \\ 0 & \tau^2 & 0 & 0 & 0 \\ 0 & 0 & \tau^2 & 0 & 0 \\ 0 & 0 & 0 & \tau^2 & \tau^2/2 \\ 0 & 0 & 0 & \tau^2/2 & \tau^2 \end{pmatrix} \right)$$

Matrix $\boldsymbol{X^a}$ indicates which elements of $\boldsymbol{\mu}$ are estimated by each $g(y_{ik})$. It contains one row per study arm and one column per basic parameter. The first row corresponds to treatment arm A of the first study taking the value 0 both for μ_{AB} and μ_{AC}. The second row indicates that μ_{AB} is estimated in treatment arm B of the first study. Similarly, the construction of the next rows of $\boldsymbol{X^a}$, as well as that of $\boldsymbol{Z}$ and $\boldsymbol{W}$, is implied by the arm-level data included in each study and the subsequent elements of $\boldsymbol{\mu}$ to be estimated (Table 2).

2.3.2 Deviation from means coding

The above model in Equation 1 can be modified using the deviation from means coding (8). The model will be parametrized in such a way to estimate the effects of each treatment versus the 'average' treatment. The target parameter of this model is a vector $\boldsymbol{b}$ that includes $K-1$ parameters b_k with $k = 2, \ldots, K$ which are the effects of treatment k versus the average effect over all treatments. One of the treatments – here treatment 1 – is arbitrarily chosen to be excluded for identifiability. Results do not depend on the choice of this 'reference' treatment.

For the deviation from means coding, the model will be

$$\boldsymbol{g(y^a)} = \boldsymbol{Zu} + \boldsymbol{X^{a^*}b} + \boldsymbol{W\delta}$$

Equation 2

with $\boldsymbol{X^{a^*}}$ denoting the modified design matrix. The matrices $\boldsymbol{Z}$ and $\boldsymbol{W}$ remain unchanged. The new design matrix $\boldsymbol{X^{a^*}}$ will take values -1 for the arbitrarily chosen treatment that is not included in vector $\boldsymbol{b}$; all other entries in the matrix are as in $\boldsymbol{X^a}$.

Consider the example of Table 1 and the first two rows of the $\boldsymbol{X^a}$ matrix, $\begin{pmatrix}0 & 0\\ 1 & 0\end{pmatrix}$, corresponding to the first study. According to the deviation from means coding as illustrated in Table 1, we chose a treatment (here treatment A) for which $\boldsymbol{X^{a^*}}$ will take -1 for both dummy variables (both columns of the design matrix) and the group corresponding to treatment B takes 1 and 0 for the two columns of the design matrix, as in $\boldsymbol{X^a}$. Thus, the respective part of the new design matrix will be $\begin{pmatrix}-1 & -1\\ 1 & 0\end{pmatrix}$. The model for study 1 with the alternative parametrization is

$$g(y_{1A}) = u_1 - b_B - b_C$$

$$g(y_{1B}) = u_1 + b_B + \delta_{1,AB}$$

$$\delta_{1,AB} \sim N(0, \tau^2)$$

where the parameters b_B and b_C denote the effects of B versus average treatment and C versus average treatment respectively. The effect of A versus the average treatment is $-b_B - b_C$ and the relative effect of B versus A for the study 1 is derived as

$$g(y_{1B}) - g(y_{1A}) = 2b_B + b_C + \delta_{1,AB}$$

The models for all studies are given in Table 2 and the full model is written as

$$\begin{pmatrix} g(y_{1A})\\ g(y_{1B})\\ g(y_{2A})\\ g(y_{2C})\\ g(y_{3B})\\ g(y_{3C})\\ g(y_{4A})\\ g(y_{4B})\\ g(y_{4C}) \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0\\ 1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\\ 0 & 0 & 0 & 1\\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} u_1\\ u_2\\ u_3\\ u_4 \end{pmatrix} + \begin{pmatrix} -1 & -1\\ 1 & 0\\ -1 & -1\\ 0 & 1\\ 1 & 0\\ 0 & 1\\ -1 & -1\\ 1 & 0\\ 0 & 1 \end{pmatrix} \begin{pmatrix} b_B\\ b_C \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 & 0 & 0\\ 1 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0\\ 0 & 1 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 1 & 0 & 0\\ 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \delta_{1,AB}\\ \delta_{2,AC}\\ \delta_{3,BC}\\ \delta_{4,AB}\\ \delta_{4,AC} \end{pmatrix}$$

Note that the reparametrization described using the deviation from the means coding should not be confused with different parametrizations of the NMA model to produce relative treatment effects of all treatments versus each other. We present in the Additional file 1 an

example of different parametrizations for specifying the means using reference cell coding and deviation from means coding using arm-level data.

2.4 NMA with contrast-level data

2.4.1 Reference cell coding

In the contrast-level NMA, data from $K_i - 1$ contrasts for each study are modeled. The model for study i and treatment contrast j is written as

$$y_{ij} = \mu_j + \varepsilon_{ij} + \delta_{ij}$$

$$\varepsilon_{ij} \sim N\left(0, s_{ij}^2\right)$$

$$\delta_{ij} \sim N(0, \tau^2)$$

with ε_{ij} being the random error for study i and treatment contrast j where s_{ij}^2 is the sample variance of y_{ij}. The random effect δ_{ij} is defined as in the NMA with arm-level data. For example, for the first study the model is

$$y_{1,AB} = \mu_{AB} + \varepsilon_{1,AB} + \delta_{1,AB}$$

$$\varepsilon_{1,AB} \sim N\left(0, s_{1,AB}^2\right)$$

$$\delta_{1,AB} \sim N(0, \tau^2)$$

and, similarly, for the other studies the models are given in Table 2.

The contrast-based NMA model in its general form is then written as

$$\boldsymbol{y^c} = \boldsymbol{X^c \mu} + \boldsymbol{\delta} + \boldsymbol{\varepsilon}$$

Equation 3

with the vector of random effects $\boldsymbol{\delta}$ having the distribution given in the arm-level NMA model and the vector of random errors being distributed as

$$\boldsymbol{\varepsilon} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{S})$$

where $\boldsymbol{S}$ is the block-diagonal within-study variance-covariance matrix of the same dimensions as $\boldsymbol{\Sigma}$. The design matrix $\boldsymbol{X^c}$ has dimensions $\sum_{i=1}^{n}(K_i - 1) \times (K - 1)$. The entries

in each row describe the relationship between the vector of basic parameters $\boldsymbol{\mu}$ and the vector of observed contrast-level data $\boldsymbol{y^c}$.

For example, in the illustrative network of three treatments and four studies, the full model is written as

$$\begin{pmatrix} y_{1,AB} \\ y_{2,AC} \\ y_{3,BC} \\ y_{4,AB} \\ y_{4,AC} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_{AB} \\ \mu_{AC} \end{pmatrix} + \begin{pmatrix} \delta_{1,AB} \\ \delta_{2,AC} \\ \delta_{3,BC} \\ \delta_{4,AB} \\ \delta_{4,AC} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1,AB} \\ \varepsilon_{2,AC} \\ \varepsilon_{3,BC} \\ \varepsilon_{4,AB} \\ \varepsilon_{4,AC} \end{pmatrix}$$

The first row of the $\boldsymbol{X^c}$ matrix indicates that the first two-arm study estimates μ_{AB}. Note that the arm-level model using reference cell coding for study 1 implies that

$$g(y_{1B}) - g(y_{1A}) = \mu_{AB} + \delta_{1,AB}$$

and, consequently, the first row of the $\boldsymbol{X^c}$ matrix results as the subtraction of the second minus the first row of $\boldsymbol{X^a}$.

2.4.2 Deviation from means coding

The reparametrized model will differ from that presented in Equation 3 in two ways; the target parameter to be estimated, which again are the relative effects $\boldsymbol{b}$ against an 'average' treatment, and the design matrix $\boldsymbol{X^{c^*}}$. The matrix $\boldsymbol{X^{c^*}}$ can be easily obtained from $\boldsymbol{X^{a*}}$ by subtracting its rows within each study contrast.

In its general form, the model is

$$\boldsymbol{y^c} = \boldsymbol{X^{c^*}}\boldsymbol{b} + \boldsymbol{\delta} + \boldsymbol{\varepsilon}$$

Equation 4

Consider in our example the part of $\boldsymbol{X^{a*}}$ corresponding to study 1, $\begin{pmatrix} -1 & -1 \\ 1 & 0 \end{pmatrix}$, then the row of $\boldsymbol{X^{c^*}}$ corresponding to that first study will be $(2 \quad 1)$, which is the subtraction of the two rows. This is also evident considering that

$$g(y_{1B}) - g(y_{1A}) = 2b_B + b_C + \delta_{1,AB}$$

according to the arm-based model using the deviation from means coding.

The models for studies 1 to 4 are given in Table 2 and can be written as

$$\begin{pmatrix} y_{1,AB} \\ y_{2,AC} \\ y_{3,BC} \\ y_{4,AB} \\ y_{4,AC} \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \\ -1 & 1 \\ 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} b_B \\ b_C \end{pmatrix} + \begin{pmatrix} \delta_{1,AB} \\ \delta_{2,AC} \\ \delta_{3,BC} \\ \delta_{4,AB} \\ \delta_{4,AC} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1,AB} \\ \varepsilon_{2,AC} \\ \varepsilon_{3,BC} \\ \varepsilon_{4,AB} \\ \varepsilon_{4,AC} \end{pmatrix}$$

The estimation of $\boldsymbol{b}$ in the contrast-based NMA model using deviation from means coding (Equation 4) is

$$\hat{\boldsymbol{b}} = \left(\left(\boldsymbol{X^{c^*}} \right)' \left(\boldsymbol{S} + \hat{\boldsymbol{\Sigma}} \right)^{-1} \boldsymbol{X^{c^*}} \right)^{-1} \left(\boldsymbol{X^{c^*}} \right)' \left(\boldsymbol{S} + \hat{\boldsymbol{\Sigma}} \right)^{-1} \boldsymbol{y^c}$$

with variance-covariance matrix

$$\boldsymbol{var}(\hat{\boldsymbol{b}}) = \left(\left(\boldsymbol{X^{c^*}} \right)' \left(\boldsymbol{S} + \hat{\boldsymbol{\Sigma}} \right)^{-1} \boldsymbol{X^{c^*}} \right)^{-1}$$

Vector $\hat{\boldsymbol{b}}$ includes the estimation of the $K-1$ parameters b_k for $k = 2, \ldots, K$. The estimation of the effect of treatment $k = 1$, which was chosen to be excluded for identifiability, versus the average effect is given as

$$\hat{b}_1 = \sum_{k=2}^{K} \left(-\hat{b}_k \right)$$

with variance $\sum_{k=2}^{K} var\left(\hat{b}_k\right) + \sum_{k \neq l,\, k<l, k>1, l>1}^{K} 2cov\left(\hat{b}_k, \hat{b}_l\right)$. Note that results do not depend on the choice of reference treatment.

Network estimates $\hat{\boldsymbol{\mu}}^{\boldsymbol{N}}$ can be derived as linear combinations of $\hat{\boldsymbol{b}}$

$$\hat{\boldsymbol{\mu}}^{\boldsymbol{N}} = \boldsymbol{Y}^* \hat{\boldsymbol{b}}$$

with variance-covariance matrix

$$\boldsymbol{var}(\hat{\boldsymbol{\mu}}^{\boldsymbol{N}}) = \boldsymbol{Y}^* \left(\left(\boldsymbol{X^{c^*}} \right)' \left(\boldsymbol{S} + \hat{\boldsymbol{\Sigma}} \right)^{-1} \boldsymbol{X^{c^*}} \right)^{-1} (\boldsymbol{Y}^*)'$$

and are equivalent to the network estimates derived using reference cell coding. Matrix $\boldsymbol{Y}^*$ of dimensions $\binom{K}{2} \times (K-1)$ is constructed similarly to $\boldsymbol{X^{c^*}}$ and connects $\hat{\boldsymbol{b}}$ with network estimates $\hat{\boldsymbol{\mu}}^{\boldsymbol{N}}$. We can use several methods for estimating $\boldsymbol{\Sigma}$ such as likelihood-based methods and an extension of the DerSimonian and Laird method (11,16). For the worked example, it holds that

$$\begin{pmatrix} \hat{\mu}_{AB}^{N} \\ \hat{\mu}_{AC}^{N} \\ \hat{\mu}_{BC}^{N} \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} \hat{b}_B \\ \hat{b}_C \end{pmatrix} = \begin{pmatrix} 2\hat{b}_B + \hat{b}_C \\ \hat{b}_B + 2\hat{b}_C \\ -\hat{b}_B + \hat{b}_C \end{pmatrix}$$

The contrast-level NMA model can be written as a two-stage model, as first described in (11,17,18), where results of separate pairwise meta-analyses are used instead of $\boldsymbol{y^c}$ in the model described in Equation 3. Constructing the respective design matrix follows the logic of constructing $\boldsymbol{X^c}$ and its modification to parametrize the model using the deviation from means coding is straightforward.

3 PreTA: Probability of a treatment being preferable than the average treatment

Applying the deviation from means coding in NMA models results into the derivation of the effects of each treatment against a fictional treatment of 'average' performance. In this section we use the K estimated parameters $\hat{b}_k$ to compute the probability of each treatment being better than the average treatment. To do so, we follow similar steps as those followed by Rücker and Schwarzer who derived the frequentist analogue of SUCRA, P-score (7).

Intermediate to the calculation of P-scores is the derivation of the probability that treatment k is better than treatment l, calculated as

$$P_{kl} = P(\hat{\mu}_{kl}^{N} > 0) = \Phi\left(\frac{\hat{\mu}_{kl}^{N}}{\sqrt{var(\hat{\mu}_{kl}^{N})}}\right)$$

assuming that higher values represent a better outcome. Accordingly, the probability that treatment k is better than the fictional treatment of average performance (PreTA) can be derived as

$$PreTA_k = P(\hat{b}_k > 0) = \Phi\left(\frac{\hat{b}_k}{\sqrt{var(\hat{b}_k)}}\right)$$

The range of values for $PreTA_k$ is (0.5, 1) if $\hat{b}_k > 0$, and (0, 0.5) if $\hat{b}_k < 0$. As it is the case with P-scores, the mean of $beta_k$ across all treatments is 0.5. Alternatively, the z-score

$\frac{\hat{b}_k}{\sqrt{var(\hat{b}_k)}}$ can be used to classify treatments according to their 'distance' from the average treatment.

Of note is that the above calculations assume normality of the estimated parameters $\hat{b}_k$. However, as $\hat{b}_k$ are not effect sizes expressed for example as logOR or mean differences, using them for hypothesis testing is not meaningful. Despite that, drawing $\hat{b}_k$ along with the associated 95% confidence intervals can be useful in capturing uncertainty around the ranking produced by relative treatment effects. Furthermore, by making some extra assumptions, $\hat{b}_k$ can be translated into absolute effects; for example, performing a meta-analysis of all treatment arms in the network can give an estimate of the expected outcome in the 'average' treatment. This estimate combined with $\hat{b}_k$ will then give absolute effects of each treatment.

3.1 Comparison of PreTAs with existing ranking metrics: theoretical considerations and empirical analysis

The, usually called, probability of being the best is a popular ranking metric which is interpreted as the probability of producing the best value in the outcome (pBV) in a network of interventions (e.g. large effects for a beneficial outcome, or small effects for a harmful outcome). While its derivation might be sensible in some cases, we should not overlook the fact that it only takes into account one tail of the treatment effects' distributions; e.g. it does not account for the probability to produce a small effect on a beneficial outcome. SUCRAs and P-scores are useful summaries of the entire ranking distributions; suggested interpretations include *"the average proportion of competing treatments, which produce outcome values worse than treatment k"* and *"the mean extent of certainty that treatment k produces better values than all other treatments*" (7,19).

We performed an empirical comparison of the treatment hierarchies obtained with PreTA, pBV and SUCRA, calculated using parametric bootstrap in a frequentist framework. The agreement between ranking metrics was measured using Kendall's tau. We used a previously described database of NMAs published until 2015 including networks of four or more interventions (1). We included networks with available outcome data in arm-level format, for which the primary outcome was analysed either as binary or as continuous. We used the effect measure used in the original review. Details about the inclusion criteria of the NMAs included in the database can be found in (1). The empirical analysis was performed with the use of the nmadb package in R (20).

In the following section, we illustrate our method in two networks of interventions, for which at least some disagreements between pBV, SUCRAs and PreTAs occur.

4 Worked examples

4.1 Network of antidepressants

We illustrate the derivation of the method using as an example a recently published NMA comparing the effectiveness of antidepressants for major depression (21). The primary efficacy outcome was response measured as 50% or greater reduction in the symptoms scales between baseline and 8 weeks of follow up and results were presented as ORs. The authors aimed at comparing active antidepressants and considered the inclusion of both head-to-head and placebo-controlled trials. The network comprised 522 double-blind, parallel, RCTs comparing 21 antidepressants or placebo. However, in line with previous empirical evidence (22,23), the authors have found evidence that the probability of receiving placebo decreases the overall response rate in a trial and dilutes differences between active compounds (24). Based on this ground, authors of this NMA (21) synthesized only head-to-head studies separately to estimate the relative efficacy of active interventions. Here, we will focus on the latter network that included 179 head-to-head studies comparing 18 antidepressants (Figure 1a).

Authors presented relative treatment effects between all pairs of the 18 antidepressants in a league table (figure 4 in (21)). When effect sizes are used to rank treatments, selecting a reference treatment against which to draw a forest plot of NMA effects is of particular importance. Although the choice of reference does not affect the estimates obtained, the uncertainty around NMA effects depends on the precision with which the selected reference treatment is associated. Figure 2 shows the relative treatment effects against fluoxetine and vortioxetine, the treatments that have been studied most and least respectively. While results are equivalent, choosing to present one over the other forest plot might implicitly lead to different interpretations on the similarity between the drugs based on visually inspecting the overlap of the confidence intervals.

Figure 2 also shows the derived odds of each treatment versus the odds of a fictional treatment of average response with their confidence intervals. The line of no effect is included in the graph for illustration reasons, although $e^{\hat{b}_k}$ are not suited for hypothesis testing. The amount of uncertainty around the relative effects versus the average treatment is between the

amount of uncertainty around the relative effects of fluoxetine and that of vortioxetine. In fact, presenting $e^{\hat{b}_K}$ with their confidence intervals offers a solution to the ambiguity of selecting a reference treatment, in terms of the uncertainty around them and the consequent conclusions about similarity of treatments. Moreover, Figure 2 shows the approximated absolute responses for each treatment, assuming the expected risk of the average treatment calculated as the meta-analytic effect of all treatment responses.

Table 3 summarizes the ranking metrics for the network of antidepressants; pBV, the SUCRA and PreTAs are presented (6,25). Escitalopram, which is the first treatment according to PreTA, ranks second according to SUCRA and third according to pBV. The disagreement between PreTA and pBV is explained by the fact that pBV favours vortioxetine and bupropion over escitalopram because their effects are estimated with greater uncertainty. The small disagreement between PreTA and SUCRA reflects their different interpretations: vortioxetine, ranked first according to SUCRA, beats on average a larger proportion of treatments compared to escitalopram (0.90 versus 0.83) but escitalopram has a larger probability to be better than the fictional average treatment compared to vortioxetine (0.93 versus 0.87). Similarly, fluoxetine ranks last according to PreTA whereas it is followed by trazodone according to SUCRA.

Figure 3 shows the PreTAs for the 18 antidepressants; treatments around 0.5 are the treatments closest to the average treatment. Vortioxetine has the largest point estimate against the average treatment but its estimation comes with great uncertainty. Escitalopram versus average is more precisely estimated in favor of escitalopram and it is associated with the greatest PreTA (97%). Duloxetine and milnacipran are the treatments closest to the average treatment. The point estimate of nefazodone versus the average treatment is slightly larger than that of duloxetine. Due to the associated uncertainty, however, there is 34% probability that nefazodone is superior to the fictional average treatment, compared to 52% of duloxetine. Fluoxetine, clomipramine, fluvoxamine, trazodone and reboxetine are among the worst treatments in the network, either because of their point estimates against the average treatment or because of the respective precision in the estimation.

4.2 Network of interventions for heavy menstrual bleeding

We use as a second example a network of interventions for the treatment of heavy menstrual bleeding. The following four interventions were compared: levenorgestel-releasing intrauterine system (Mirena), first generation endometrial destruction, second generation

endometrial destruction and hysterectomy (26). The primary outcome was patients' dissatisfaction at 12 months and the network included 20 studies (Figure 1b).

Figure 4 shows the treatment effects of the four treatments compared to a fictional average treatment and Appendix Figure 1 illustrates the relative position of each treatment according to its probability of being superior (green) or inferior (red) than the average treatment. There is a clear advantage of hysterectomy compared to the other three treatments with no treatment lying close to the 'average treatment area' (0.5 of PreTA).

In this example, hysterectomy outperforms the other three treatments and ranks first according to all ranking metrics. Similarly, all ranking metrics agree that first generation endometrial destruction is the least preferable option (Figure 4). The disagreement between ranking metrics occurs for the second and third position between Mirena and second generation endometrial destruction. The two interventions are similar according to the point estimates but second generation is more precise. This leads to a greater certainty that second generation is worse than the average treatment compared to Mirena, resulting in a smaller PreTA. However, second generation beats on average more treatments than Mirena does since the relative effect of second generation is larger than that of Mirena; this results in a larger SUCRA for second generation than for Mirena.

5 Results of the empirical analysis

We ended up with 232 networks to be included in the empirical analysis. There was strong agreement between hierarchies obtained by PreTAs and SUCRAs, shown by a median Kendall's tau (in the following called 'correlation') of 0.94 with interquartile range (IQR) 0.86 to 1.00). Almost half of the networks (101, 44%) had correlation of 1 while only two networks (1%) had correlation less than 0.6. The network with the smallest correlation (0.4) is shown in Appendix Figure 2 (27); while the hierarchy itself does not change much, PreTA and SUCRA disagree in proximity of treatments with similar point estimates and different precision. The agreement between PreTAs and pBV was lower with a median correlation of 0.74 (IQR 0.61 to 0.89) and 49 networks (21%) having correlation less than 0.6 (Appendix Figure 3).

As with all ranking metrics, any disagreements between PreTAs and pBV or SUCRAs are attributed to the different ways they incorporate uncertainty in the estimation. pBV favors treatments associated with uncertainty, as the tail of the distribution of treatments with uncertain effects is larger compared to the tail of the distribution for treatments with similar

point estimate but high precision. The probability P_{kl} tends to 0.5 with increased $var(\hat{\mu}_{kl}^{N})$; consequently, the greater the uncertainty associated with a treatment, the more its P-score tends to 0.5. A research paper describing theoretically the interpretation and the role of uncertainty in the various ranking metrics, as well as a detailed empirical analysis are in preparation (19,28).

6 Discussion

In this paper, we developed a new ranking metric, PreTA, interpreted as the probability of each treatment being preferable than a fictional treatment of average performance. The notion of the average treatment refers to the average absolute efficacy among the treatments included in the systematic review. Thus, as with all ranking metrics, the interpretation of PreTAs is subject to the set of treatments compared. PreTAs can be produced in all NMAs as long as the eligibility of treatments is well justified. The usefulness of the interpretation of the $\hat{b}_k$ coefficients, however, depends on whether the the notion of an 'average' treatment makes sense.

In the presence of a reference treatment, e.g. placebo, a simple and intuitive non-probabilistic ranking metric can be obtained by ranking all relative effects against placebo. Authors of NMA often present estimated treatment effects against placebo or standard care in a forest plot, providing implicitly or explicitly a treatment hierarchy. While such a hierarchy might be appropriate in many settings, they assume that treatment effects against placebo are of primary interest for the analysis. This might not be the case in other healthcare areas where one or more established therapies exists (29) or where researchers are concerned about the quality of the evidence from placebo-controlled studies (30–32) and choose to, exclusively or complementary, analyse a network without placebo. Moreover, it should be taken into account that the amount of data associated with the reference treatment might have an impact on the judgement regarding the similarity of the treatments, when such a judgement is made by visually inspecting a forest plot of NMA effects. Point estimates against the fictional average treatment provide a solution to this ambiguity. Furthermore, data from registries can be assumed to approximate the response of an average treatment, as participants may take any of the available interventions. Thus, using such external data, absolute effects can be approximated using the point estimates against the average.

Alternative methods to avoid the reference group coding have been suggested in the literature. The application of quasi-variances (33), independently proposed as 'floating absolute risks' in epidemiology (34), do avoid setting a reference group. However, the scope of their use pertains to approximating a set of variances of the model contrasts such that the variances between any linear combination of contrasts can be derived without the disposal of the covariance matrix (35). Thus, quasi-variances approaches target a different problem from the model described in this paper and the relevance of the estimated quantities to NMA is not clear.

Producing a treatment hierarchy in NMA is popular, with 43% of published NMAs presenting at least one ranking metric (1), but also debatable. Recent developments tackle common criticisms against ranking metrics, pertaining to arguments that they are unstable (36,37), uncertain (38), do not differentiate between clinically important and unimportant differences (4,39), do not account for multiple outcomes (40) and are not accompanied by a measure of uncertainty (41). In particular, recent developments include extensions of P-scores for two or more outcomes (42), incorporation of clinically important values in their calculation (42), application of multiple-criteria decision analysis (43) and partial ordering of interventions according to multiple outcomes (44). PreTAs can be easily extended to incorporate clinically important values as shown in (42); such probabilities will then be interpreted as the probability of a treatment being better than the average by at least a certain value.

PreTA is a viable alternative to existing ranking metrics, that can be interpreted as a probability and takes into account the entire ranking distribution. As it is also the case with PreTA, all existing ranking metrics use the distribution of NMA treatment effects to produce a hierarchy of the treatments. This hierarchy can be based either on probabilities like "which is the probability that each treatment produces the best outcome value" or "which is the probability of treatment A beating treatment B" or summaries of these probabilities. Rankograms visualise the entire ranking distributions for each treatment and SUCRAs, P-scores and mean ranks summarise these probabilities in a single number for each treatment. The interpretation of these summaries is, however, not always straightforward. The development of PreTAs enriches the decision-making arsenal with a presentational and ranking tool, which can be interpreted in a clinically meaningful way.

7 Declarations

Data Availability Statement

Outcome data and the code for applying our methods are available in https://github.com/esm-ispm-unibe-ch/alternativenma.

Conflicts of interest

TAF reports personal fees from Mitsubishi-Tanabe, MSD and Shionogi and a grant from Mitsubishi-Tanabe, outside the submitted work; TAF has a patent 2018-177688 pending.

Funding

AN, VC, TP and GS were supported by project funding (Grant No. 179158) from the Swiss National Science Foundation.

Authors' contributions

AN conceived the idea, contributed to the modelling, produced the results and wrote the R code and the first draft of the manuscript. VC contributed to the analysis. TP contributed to the modelling and to the R code. DM, TAF and GS contributed to the modelling, reviewed the R code and contributed to the writing. All authors read and approved the final manuscript.

8 Figure legends

Figure 1. Panel a: Network plot of head-to-head randomized control trials comparing 18 antidepressants. Panel b: Network plot of head-to-head randomized control trials comparing 4 interventions for heavy menstrual bleeding. First and second generation interventions refer to endometrial destruction. Nodes and edges are unweighted.

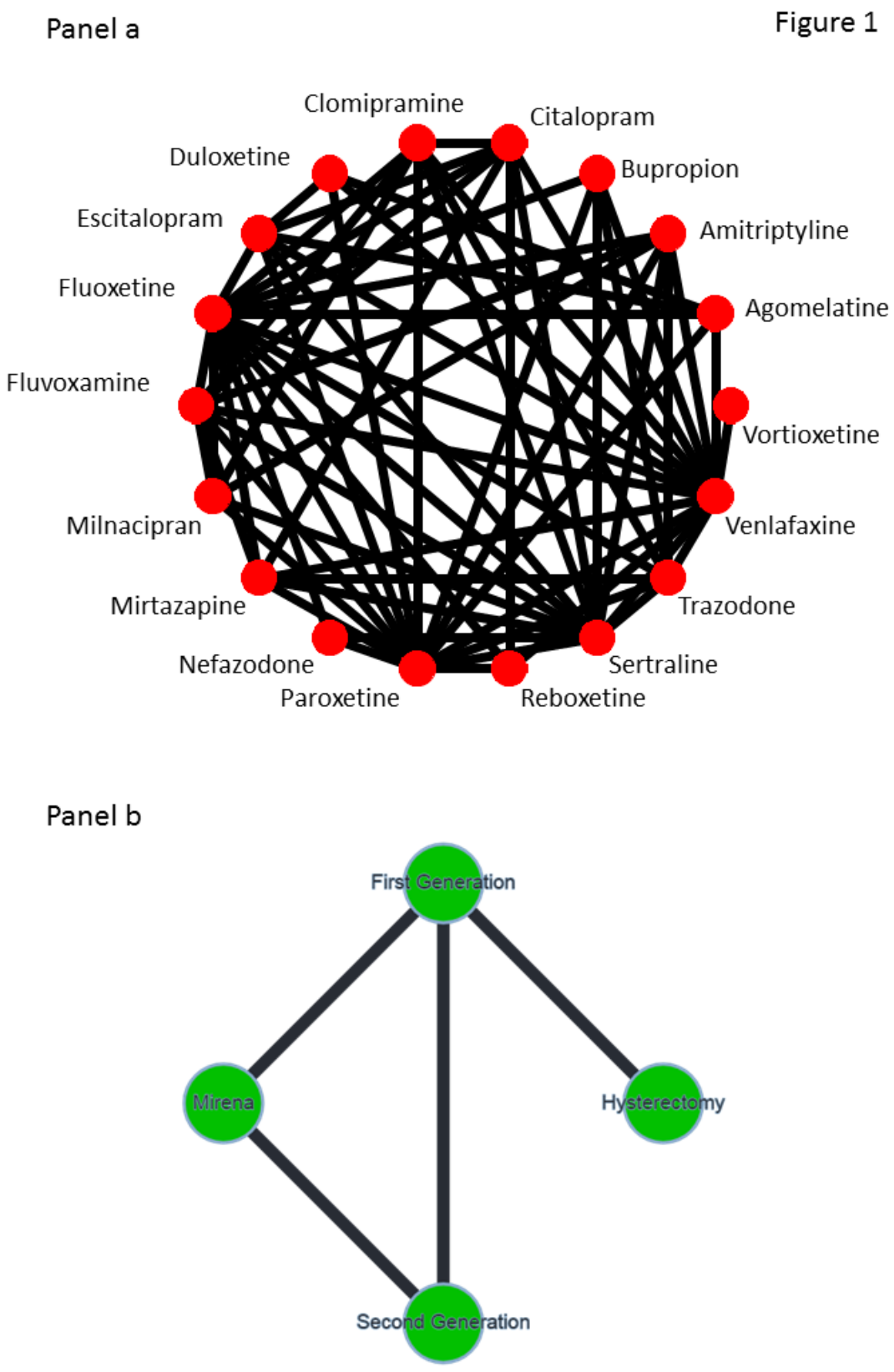


Figure 2. Odds ratios of each treatment versus fluoxetine, odds of each treatment versus odds of a fictional treatment of average response $exp(\widehat{b}_k)$ and odds ratios versus vortioxetine in the network of head-to-head studies comparing 18 antidepressants. OR: odds ratio; CI: confidence interval.

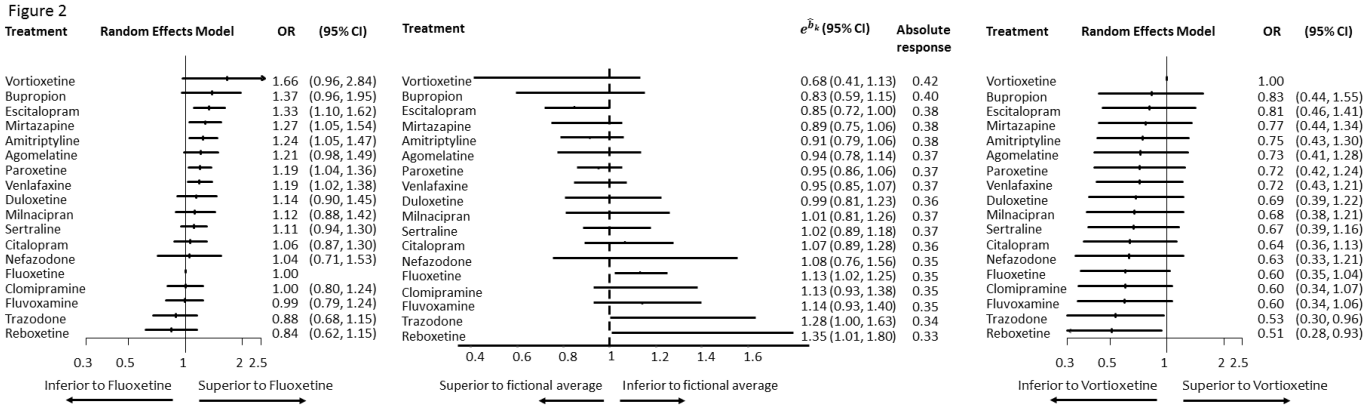


Figure 3. Classifier of interventions for the network of 18 antidepressants according to the probability of being preferable than average (PreTA).

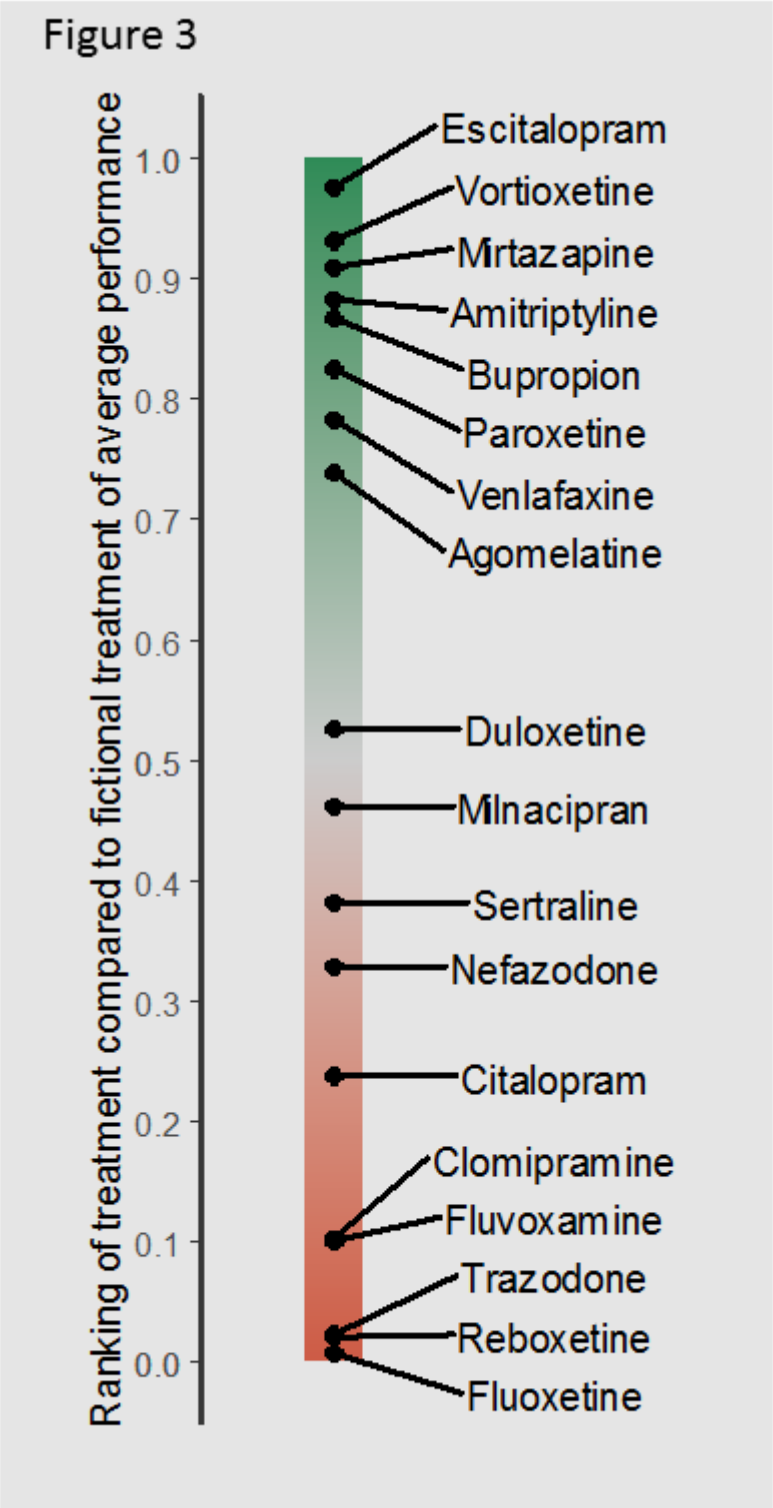


Figure 4. Odds of each treatment versus odds of a fictional treatment of average response $exp(\hat{b}_k)$, probability of each treatment being better than the average (PreTA), probability of producing the best value (pBV) and SUCRA in the network of head-to-head studies comparing 4 interventions for heavy menstrual bleeding. Numbers in parentheses under PreTA, pBV and SUCRA represent ranks. CI: confidence interval; PreTA: preferable than average; pBV: probability of producing the best value; SUCRA: surface under the cumulative ranking curve.

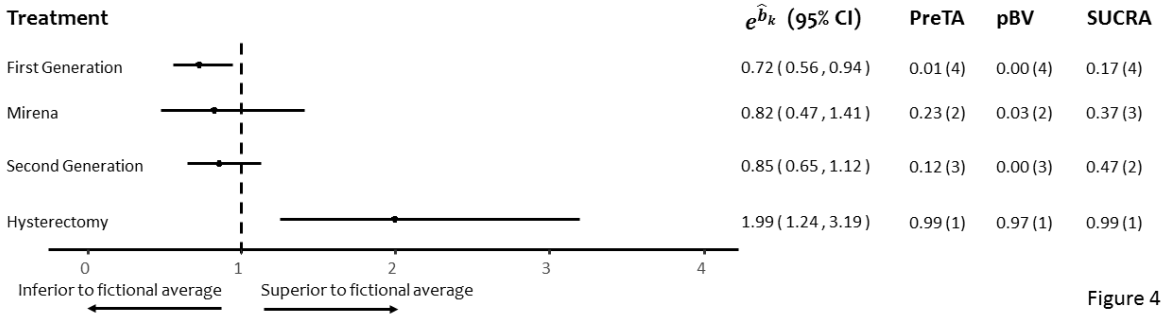
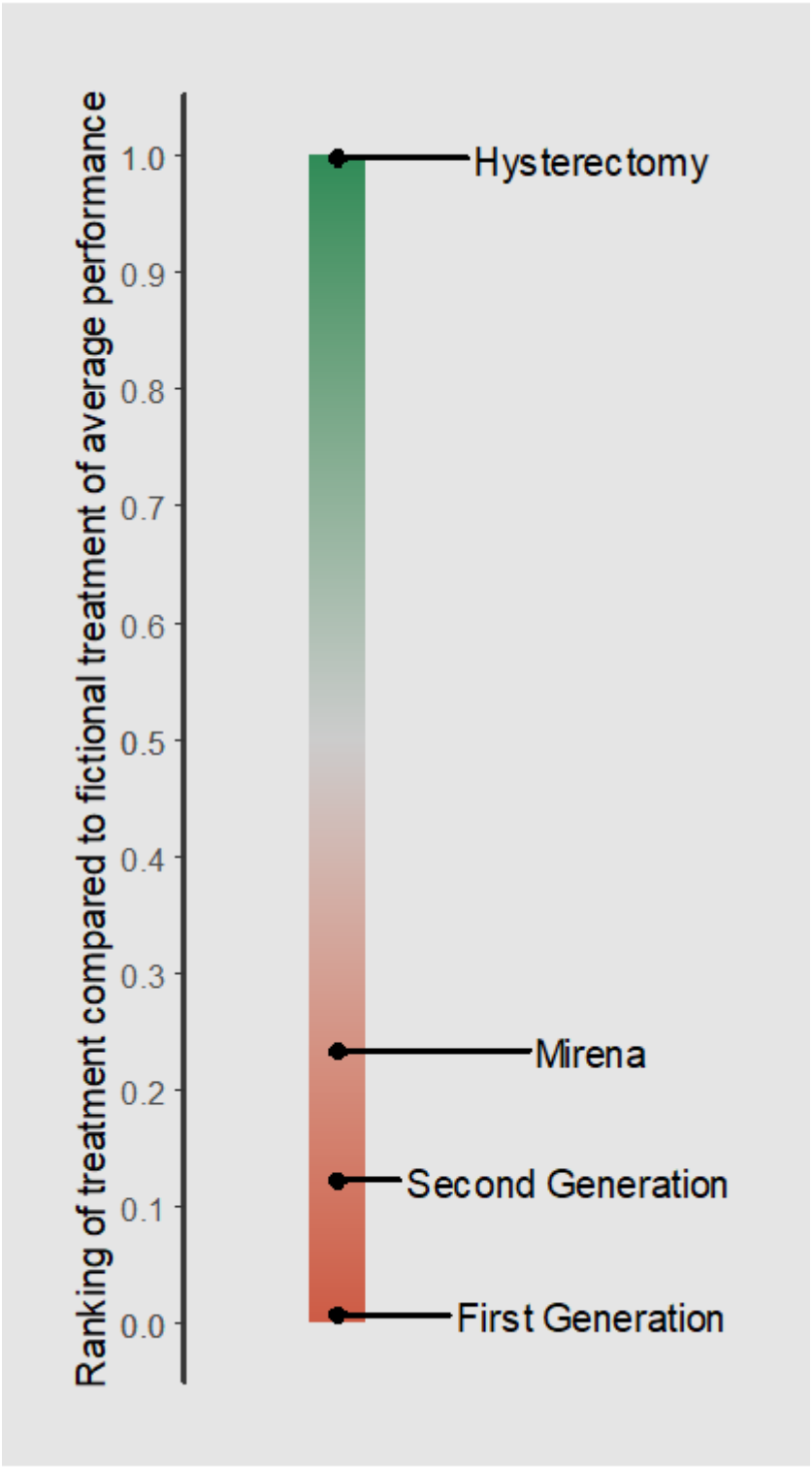
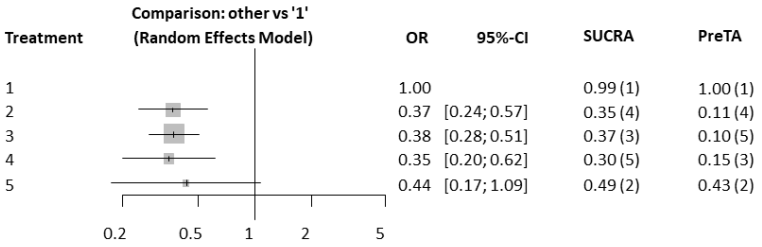


Figure 4

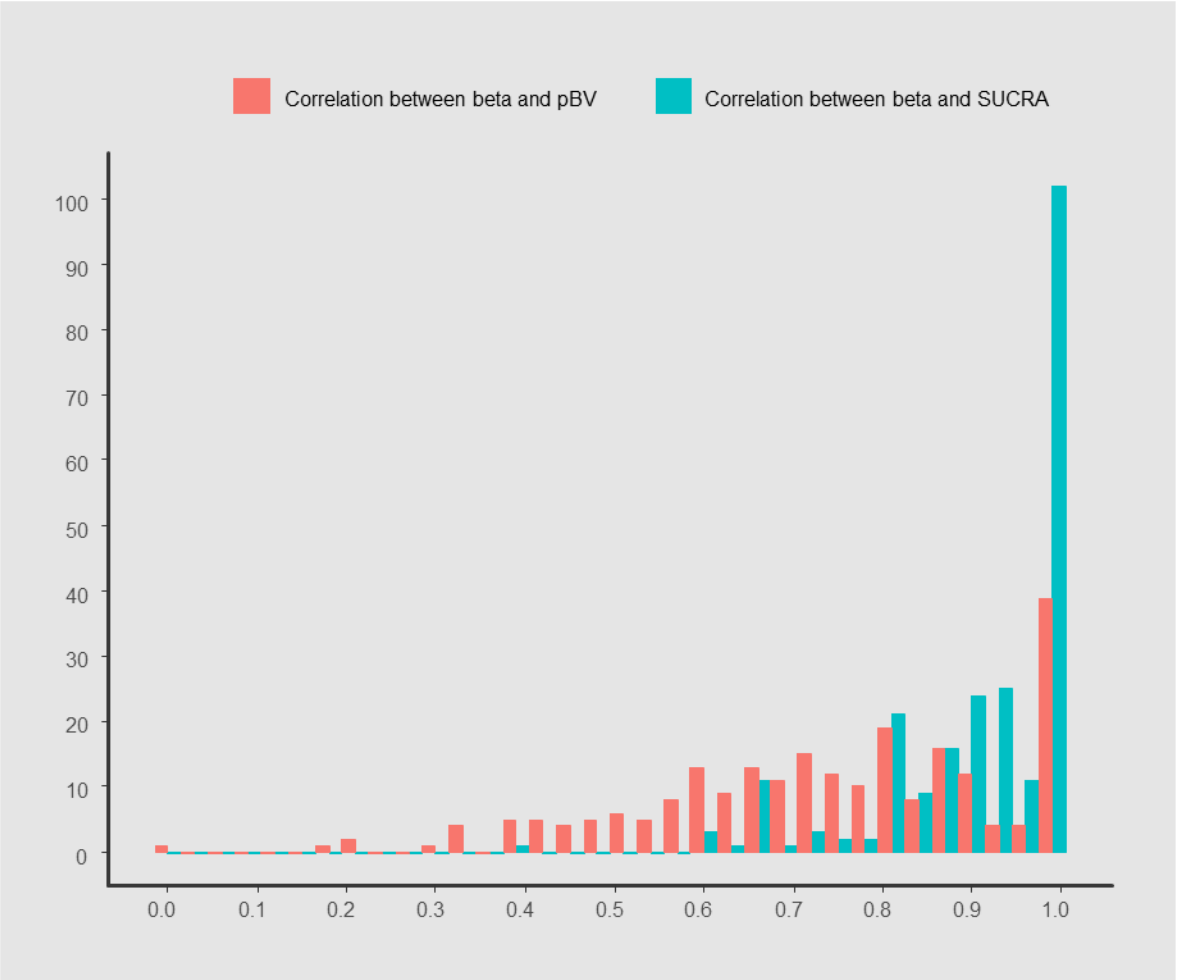
Appendix Figure 1. Classifier of interventions for the network of four interventions for heavy menstrual bleeding according to the probability of being preferable than average (PreTA).



Appendix Figure 2. Odds ratios, probability of each treatment being better than the average (PreTA) and SUCRA in the network with the smallest correlation between PreTA and SUCRA. Numbers in parentheses under PreTA, pBV and SUCRA represent ranks. OR: odds ratio; CI: confidence interval; PreTA: preferable than average; pBV: probability of producing the best value; SUCRA: surface under the cumulative ranking curve.



Appendix Figure 3. Histogram of correlation measured as Kendall's tau between probability of being better than average (PreTA) and probability of producing the best value (red) and between PreTA and surface under the cumulative ranking curve (SUCRA) (blue). pBV: probability of producing the best value; SUCRA: surface under the cumulative ranking curve; PreTA: preferable than average.



9 Highlights

What is already known: A treatment hierarchy is often of interest to users of network meta-analysis. However, interpretation of ranking metrics is often challenging.

What is new: We present a new ranking metric (PreTA: Preferable Than Average) interpreted as the probability that a treatment is better than a fictional treatment of average performance.

Potential impact for Review Synthesis Methods readers outside the authors' field: The proposed ranking metric uses the entire ranking distribution and it is interpreted as a

probability. It can be used as a viable alternative to existing ranking metrics in systematic reviews with multiple interventions.

10 References

1. Petropoulou M, Nikolakopoulou A, Veroniki A-A, Rios P, Vafaei A, Zarin W, et al. Bibliographic study showed improving statistical methodology of network meta-analyses published between 1999 and 2015. J Clin Epidemiol. 2016 Nov 15;

2. Kanters S, Ford N, Druyts E, Thorlund K, Mills EJ, Bansback N. Use of network meta-analysis in clinical guidelines. Bull World Health Organ. 2016 Oct 1;94(10):782–4.

3. Higgins JPT, Welton NJ. Network meta-analysis: a norm for comparative effectiveness? Lancet Lond Engl. 2015 Aug 15;386(9994):628–30.

4. Cipriani A, Higgins JPT, Geddes JR, Salanti G. Conceptual and technical challenges in network meta-analysis. Ann Intern Med. 2013 Jul 16;159(2):130–7.

5. Tan SH, Cooper NJ, Bujkiewicz S, Welton NJ, Caldwell DM, Sutton AJ. Novel presentational approaches were developed for reporting network meta-analysis. J Clin Epidemiol. 2014 Jun;67(6):672–80.

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

7. Rücker G, Schwarzer G. Ranking treatments in frequentist network meta-analysis works without resampling methods. BMC Med Res Methodol. 2015 Jul 31;15:58.

8. Hosmer DW, Lemeshow S. Interpretation of the Fitted Logistic Regression Model. In: Applied Logistic Regression [Internet]. John Wiley & Sons, Ltd; 2005. p. 47–90. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/0471722146.ch3

9. Salanti G. Indirect and mixed-treatment comparison, network, or multiple-treatments meta-analysis: many names, many benefits, many concerns for the next generation evidence synthesis tool. Res Synth Methods. 2012 Jun;3(2):80–97.

10. Jansen JP, Naci H. Is network meta-analysis as valid as standard pairwise meta-analysis? It all depends on the distribution of effect modifiers. BMC Med. 2013;11:159.

11. Lu G, Welton NJ, Higgins JPT, White IR, Ades AE. Linear inference for mixed treatment comparison meta-analysis: A two-stage approach. ResSynthMeth. 2011;2(1):43–60.

12. Lu G, Ades AE. Assessing Evidence Inconsistency in Mixed Treatment Comparisons. J Am Stat Assoc. 2006 Jun 1;101(474):447–59.

13. Salanti G, Higgins JPT, Ades AE, Ioannidis JPA. Evaluation of networks of randomized trials. Stat Methods Med Res. 2008 Jun;17(3):279–301.

14. Dias S, Sutton AJ, Ades AE, Welton NJ. Evidence synthesis for decision making 2: a generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials. Med Decis Mak Int J Soc Med Decis Mak. 2013;33(5):607–17.

15. Jackson D, Law M, Stijnen T, Viechtbauer W, White IR. A comparison of seven random-effects models for meta-analyses that estimate the summary odds ratio. Stat Med. 2018 Mar 30;37(7):1059–85.

16. van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. Stat Med. 2002 Feb 28;21(4):589–624.

17. Rücker G. Network meta-analysis, electrical networks and graph theory. Res Synth Methods. 2012 Dec;3(4):312–24.

18. Rücker G, Schwarzer G. Reduce dimension or reduce weights? Comparing two approaches to multi-arm studies in network meta-analysis. Stat Med. 2014 Nov 10;33(25):4353–69.

19. Salanti G, Nikolakopoulou A, Efthimiou O, Egger M, Mavridis D, White IR. What works best? Obtaining a treatment hierarchy from network meta-analysis (in preparation).

20. Papakonstantinou T. nmadb: Network Meta-Analysis Database API [Internet]. 2019. Available from: https://CRAN.R-project.org/package=nmadb

21. Cipriani A, Furukawa TA, Salanti G, Chaimani A, Atkinson LZ, Ogawa Y, et al. Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: a systematic review and network meta-analysis. Lancet Lond Engl. 2018 07;391(10128):1357–66.

22. Papakostas GI, Fava M. Does the probability of receiving placebo influence clinical trial outcome? A meta-regression of double-blind, randomized clinical trials in MDD. Eur Neuropsychopharmacol J Eur Coll Neuropsychopharmacol. 2009 Jan;19(1):34–40.

23. Sinyor M, Levitt AJ, Cheung AH, Schaffer A, Kiss A, Dowlati Y, et al. Does inclusion of a placebo arm influence response to active antidepressant treatment in randomized controlled trials? Results from pooled and meta-analyses. J Clin Psychiatry. 2010;71(3):270–9.

24. Salanti G, Chaimani A, Furukawa TA, Higgins JPT, Ogawa Y, Cipriani A, et al. Impact of placebo arms on outcomes in antidepressant trials: systematic review and meta-regression analysis. Int J Epidemiol. 2018 01;47(5):1454–64.

25. Rücker G, Schwarzer G. Ranking treatments in frequentist network meta-analysis works without resampling methods. BMC Med Res Methodol. 2015 Jul 31;15:58.

26. Middleton LJ, Champaneria R, Daniels JP, Bhattacharya S, Cooper KG, Hilken NH, et al. Hysterectomy, endometrial destruction, and levonorgestrel releasing intrauterine system (Mirena) for heavy menstrual bleeding: systematic review and meta-analysis of data from individual patients. BMJ. 2010 Aug 16;341:c3929.

27. Wang L, Baser O, Kutikova L, Page JH, Barron R. The impact of primary prophylaxis with granulocyte colony-stimulating factors on febrile neutropenia during chemotherapy: a systematic review and meta-analysis of randomized controlled trials. Support Care Cancer Off J Multinatl Assoc Support Care Cancer. 2015 Nov;23(11):3131–40.

28. Chiocchia V, Nikolakopoulou A, Papakonstantinou T, Egger M, Salanti G. Empirical evaluation of the agreement between ranking metrics in network meta-analysis (in preparation).

29. Batra S, Howick J. Empirical evidence against placebo controls. J Med Ethics. 2017 Aug 9;

30. Turner EH, Knoepflmacher D, Shapley L. Publication Bias in Antipsychotic Trials: An Analysis of Efficacy Comparing the Published Literature to the US Food and Drug Administration Database. PLOS Med. 2012;9(3):e1001189.

31. Ioannidis JPA, Karassa FB. The need to consider the wider agenda in systematic reviews and meta-analyses: breadth, timing, and depth of the evidence. BMJ. 2010 Sep 13;341:c4875.

32. Lee K, Bacchetti P, Sim I. Publication of clinical trials supporting successful new drug applications: a literature analysis. PLoS Med. 2008 Sep 23;5(9):e191.

33. Ridout MS. Summarizing the Results of Fitting Generalized Linear Models to Data from Designed Experiments. In: Statistical Modelling [Internet]. Springer, New York, NY; 1989 [cited 2017 Sep 7]. p. 262–9. (Lecture Notes in Statistics). Available from: https://link.springer.com/chapter/10.1007/978-1-4612-3680-1_30

34. Easton DF, Peto J, Babiker AG a. G. Floating absolute risk: An alternative to relative risk in survival and case-control analysis avoiding an arbitrary reference group. Stat Med. 1991 Jul 1;10(7):1025–35.

35. Firth D, Menezes D, X R. Quasi-variances. Biometrika. 2004 Mar 1;91(1):65–80.

36. Kibret T, Richer D, Beyene J. Bias in identification of the best treatment in a Bayesian network meta-analysis for binary outcome: a simulation study. Clin Epidemiol. 2014 Dec 3;6:451–60.

37. Mills EJ, Kanters S, Thorlund K, Chaimani A, Veroniki A-A, Ioannidis JPA. The effects of excluding treatments from network meta-analyses: survey. BMJ. 2013 Sep 5;347:f5195.

38. Trinquart L, Attiche N, Bafeta A, Porcher R, Ravaud P. Uncertainty in Treatment Rankings: Reanalysis of Network Meta-analyses of Randomized Trials. Ann Intern Med. 2016 May 17;164(10):666–73.

39. Brignardello-Petersen R, Johnston BC, Jadad AR, Tomlinson G. Using decision thresholds for ranking treatments in network meta-analysis results in more informative rankings. J Clin Epidemiol. 2018 Jun;98:62–9.

40. Mbuagbaw L, Rochwerg B, Jaeschke R, Heels-Andsell D, Alhazzani W, Thabane L, et al. Approaches to interpreting and choosing the best treatments in network meta-analyses. Syst Rev. 2017 12;6(1):79.

41. Veroniki AA, Straus SE, Rücker G, Tricco AC. Is providing uncertainty intervals in treatment ranking helpful in a network meta-analysis? J Clin Epidemiol. 2018 Aug;100:122–9.

42. Mavridis D, Porcher R, Nikolakopoulou A, Salanti G, Ravaud P. Extensions of the probabilistic ranking metrics of competing treatments in network meta-analysis to reflect clinically important relative differences on many outcomes. Biom J Biom Z. 2019 Oct 29;

43. Tervonen T, Naci H, van Valkenhoef G, Ades AE, Angelis A, Hillege HL, et al. Applying Multiple Criteria Decision Analysis to Comparative Benefit-Risk Assessment: Choosing among Statins in Primary Prevention. Med Decis Mak Int J Soc Med Decis Mak. 2015 Oct;35(7):859–71.

44. Rücker G, Schwarzer G. Resolve conflicting rankings of outcomes in network meta-analysis: Partial ordering of treatments. Res Synth Methods. 2017 Dec;8(4):526–36.

11 Tables

Reference cell coding				**Deviation from means coding**			
	Dummy variables				Dummy variables		
Covariate	x_1	x_2	x_3	Covariate	x_1^*	x_2^*	x_3^*
Group 1	0	0	0	Group 1	-1	-1	-1
Group 2	1	0	0	Average*	0	0	0
Group 3	0	1	0	Group 2	1	0	0
Group 4	0	0	1	Group 3	0	1	0
				Group 4	0	0	1

Table 1. Illustration of construction of dummy variables for modelling a categorical variable with four groups in regression using reference cell coding and deviation from means coding.

Study number, treatments compared	**Arm-based NMA**		**Contrast-based NMA**	
	Reference cell coding	Deviation from means coding	Reference cell coding	Deviation from means coding
Study 1, AB	$g(y_{1A}) = u_1$ $g(y_{1B}) = u_1 + \mu_{AB} + \delta_{1,AB}$ $\delta_{1,AB} \sim N(0, \tau^2)$	$g(y_{1A}) = u_1 - b_B - b_C$ $g(y_{1B}) = u_1 + b_B + \delta_{1,AB}$ $\delta_{1,AB} \sim N(0, \tau^2)$	$y_{1,AB} = \mu_{AB} + \varepsilon_{1,AB} + \delta_{1,AB}$ $\varepsilon_{1,AB} \sim N(0, s^2_{1,AB})$ $\delta_{1,AB} \sim N(0, \tau^2)$	$y_{1,AB} = 2b_B + b_C + \varepsilon_{1,AB} + \delta_{1,AB}$ $\varepsilon_{1,AB} \sim N(0, s^2_{1,AB})$ $\delta_{1,AB} \sim N(0, \tau^2)$
Study 2, AC	$g(y_{2A}) = u_2$ $g(y_{2C}) = u_2 + \mu_{AC} + \delta_{2,AC}$ $\delta_{2,AC} \sim N(0, \tau^2)$	$g(y_{2A}) = u_2 - b_B - b_C$ $g(y_{2C}) = u_2 + b_C + \delta_{2,AC}$ $\delta_{2,AC} \sim N(0, \tau^2)$	$y_{2,AC} = \mu_{AC} + \varepsilon_{2,AC} + \delta_{2,AC}$ $\varepsilon_{2,AB} \sim N(0, s^2_{2,AC})$ $\delta_{2,AC} \sim N(0, \tau^2)$	$y_{2,AC} = b_B + 2b_C + \varepsilon_{2,AC} + \delta_{2,AC}$ $\varepsilon_{2,AB} \sim N(0, s^2_{2,AC})$ $\delta_{2,AC} \sim N(0, \tau^2)$
Study 3, BC	$g(y_{3B}) = u_3$ $g(y_{3C}) = u_3 - \mu_{AB} + \mu_{AC} + \delta_{3,BC}$ $\delta_{3,BC} \sim N(0, \tau^2)$	$g(y_{3B}) = u_3 + b_B$ $g(y_{3C}) = u_3 + b_C + \delta_{3,BC}$ $\delta_{3,BC} \sim N(0, \tau^2)$	$y_{3,BC} = -\mu_{AB} + \mu_{AC} + \varepsilon_{3,BC} + \delta_{3,BC}$ $\varepsilon_{3,BC} \sim N(0, s^2_{3,BC})$ $\delta_{3,BC} \sim N(0, \tau^2)$	$y_{3,BC} = -b_B + b_C + \varepsilon_{3,BC} + \delta_{3,BC}$ $\varepsilon_{3,BC} \sim N(0, s^2_{3,BC})$ $\delta_{3,BC} \sim N(0, \tau^2)$
Study 4, ABC	$g(y_{4A}) = u_4$ $g(y_{4B}) = u_4 + \mu_{AB} + \delta_{4,AB}$ $g(y_{4C}) = u_4 + \mu_{AC} + \delta_{4,AC}$ $\delta_{4,AB} \sim N(0, \tau^2)$ $\delta_{4,AC} \sim N(0, \tau^2)$	$g(y_{4A}) = u_4 - b_B - b_C$ $g(y_{4B}) = u_4 + b_B + \delta_{4,AB}$ $g(y_{4C}) = u_4 + b_C + +\delta_{4,AC}$ $\delta_{4,AB} \sim N(0, \tau^2)$ $\delta_{4,AC} \sim N(0, \tau^2)$	$y_{4,AB} = \mu_{AB} + \varepsilon_{4,AB} + \delta_{4,AB}$ $y_{4,AC} = \mu_{AC} + \varepsilon_{4,AC} + \delta_{4,AC}$ $\varepsilon_{4,AB} \sim N(0, s^2_{4,AB})$ $\delta_{4,AB} \sim N(0, \tau^2)$ $\varepsilon_{4,AC} \sim N(0, s^2_{4,AC})$ $\delta_{4,AC} \sim N(0, \tau^2)$	$y_{4,AB} = 2b_B + b_C + \varepsilon_{4,AB} + \delta_{4,AB}$ $y_{4,AC} = b_B + 2b_C + \varepsilon_{4,AC} + \delta_{4,AC}$ $\varepsilon_{4,AB} \sim N(0, s^2_{4,AB})$ $\delta_{4,AB} \sim N(0, \tau^2)$ $\varepsilon_{4,AC} \sim N(0, s^2_{4,AC})$ $\delta_{4,AC} \sim N(0, \tau^2)$

Table 2. Arm-level and contrast-level NMA models using reference cell coding and deviation from means coding for a fictional network of three treatments examined in four studies.

	pBV	SUCRA	PreTA
Agomelatine	0.01 (6)	0.64 (6)	0.74 (8)
Amitriptyline	0.01 (7)	0.71 (5)	0.88 (4)
Bupropion	0.20 (2)	0.80 (3)	0.87 (5)
Citalopram	0.00 (17.5)	0.37 (13)	0.24 (13)
Clomipramine	0.00 (15)	0.26 (14)	0.10 (14.5)
Duloxetine	0.01 (9)	0.52 (9)	0.52 (9)
Escitalopram	0.07 (3)	0.83 (2)	0.97 (1)
Fluoxetine	0.00 (17.5)	0.23 (16)	0.01 (18)
Fluvoxamine	0.00 (12.5)	0.25 (15)	0.10 (14.5)
Milnacipran	0.01 (8)	0.48 (10)	0.46 (10)
Mirtazapine	0.03 (4)	0.75 (4)	0.91 (3)
Nefazodone	0.02 (5)	0.38 (12)	0.33 (12)
Paroxetine	0.00 (10)	0.62 (7)	0.82 (6)
Reboxetine	0.00 (15)	0.09 (18)	0.02 (16.5)
Sertraline	0.00 (11)	0.46 (11)	0.38 (11)
Trazodone	0.00 (15)	0.12 (17)	0.02 (16.5)
Venlafaxine	0.00 (12.5)	0.61 (8)	0.78 (7)
Vortioxetine	0.64 (1)	0.90 (1)	0.93 (2)

Table 3. Ranking metrics for the network of antidepressants and ranks according to each ranking metric in parentheses. pBV: probability of producing the best value; SUCRA: surface under the cumulative ranking curve; PreTA: preferable than average.