Considering multiple outcomes with different weights informed the hierarchy of interventions in network-meta analysis

Dimitris Mavridis[1], Adriani Nikolakopoulou[2], Irini Moustaki[3], Anna Chaimani[4], Raphaël Porcher[4], Isabelle Boutron[4] and Philippe Ravaud[4,5]

[1] Department of Primary Education, University of Ioannina, Ioannina, Greece.
[2] Institute for Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Germany.
[3] London School of Economics and Political Science, London, UK.
[4] Université Paris Cité, Research Centre in Epidemiology and Statistics (CRESS-UMR1153), Inserm Paris, France.
[5] Department of Epidemiology, Columbia University Mailman School of Public Health, NY, USA.

Abstract

Objectives

Ranking metrics in network meta-analysis (NMA) are computed separately for each outcome. Our aim is to 1) present graphical ways to group competing interventions considering multiple outcomes and 2) use conjoint analysis for placing weights on the various outcomes based on the stakeholders' preferences.

Study design and setting

We used multidimensional scaling (MDS) and hierarchical tree clustering to visualize the extent of similarity of interventions in terms of the relative effects they produce through a random effects NMA. We reanalyzed a published network of 212 psychosis trials taking three outcomes into account: reduction in symptoms of schizophrenia, all-cause treatment discontinuation and weight gain.

Results

Conjoint analysis provides a mathematical method to transform judgements into weights that can be subsequently used to visually represent interventions on a two-dimensional plane or through a dendrogram. These plots provide insightful information about the clustering of interventions.

Conclusion

Grouping interventions can help decision makers not only to identify the optimal ones in terms of benefit-risk balance but also choose one from the best cluster based on other grounds such as cost, implementation etc. Placing weights on outcomes allows considering patient profile or preferences.

Keywords: clustering; conjoint analysis; multidimensional scaling; ranking

Running title: Ranking interventions considering multiple weighted outcomes

What is new

Key findings

- Methods were suggested to visually group interventions considering multiple outcomes placing weights on outcomes based on stakeholder's preferences

What this adds to what is known

- Provides a framework to rank/group interventions considering both benefits and harms, patients/stakeholders preferences and patient profiles

What is the implication, what should change now

- Grouping interventions allows us choosing alternative interventions or select one based on information outside the systematic review (e.g., cost, implementation, side-effects for which we do not have data for).
- Weighing outcomes allows a clinician to consider patient profiles and/or preferences

1. Introduction

Network meta-analysis (NMA) synthesizes both direct and indirect evidence resulting in more precise effect estimates and allowing making inference for the relative effectiveness between pairs of interventions that have never been compared head-to-head[1,2]. Two of the most important outputs of an NMA are the estimated relative effects and a ranking of the competing interventions. Suppose that we have T interventions. There is a total of $\binom{T}{2} = \frac{T(T-1)}{2}$ relative effects for all possible pairs of interventions that can be represented through a league table. Ranking of the competing interventions for any outcome of interest is very appealing but also very controversial and prone to misinterpretations (e.g., does not consider the risk of bias of the included trials and the confidence one should place on the NMA findings)[3,4]. Several graphical and quantitative metrics have been developed for ranking interventions. Such measures utilize the distribution of relative effects to estimate probabilities for any intervention assuming any possible rank. The most commonly used measures are SUCRA values and P-scores[4–6]. The two outputs, relative effects and ranking, are highly dependent and PRISMA checklist for reporting NMAs state that ranking metrics may exaggerate small differences in relative effects and should be considered along with the corresponding estimates of pairwise comparisons[7].

Ranking metrics have been developed for a single outcome. Systematic reviews are encouraged to report both on efficacy and safety outcomes. Chaimani et al suggested using multidimensional scaling techniques (MDS) to visualize the level of similarity among interventions for a single outcome[8]. Veroniki et al presented the rank-heat plot, a simple graphical approach to present treatment ranking including multiple outcomes[9]. Mavridis et al extended the P-score methodology to allow for multiple outcomes and modified

P-scores to measure the mean extent of certainty that each intervention is better by another intervention by a certain amount, e.g. the minimum important clinical difference[10]. This research focuses on treatment effects but does not consider that patients have different profiles and preferences, and the benefit-harm balance of treatment options may differ accordingly.

We focus on ranking visualization methods when multiple outcomes are considered, and we use multidimensional techniques to map interventions and represent them spatially with an aim to see how they cluster together. More specifically, we consider that the estimated relative treatment effects reflect distances between interventions and we use MDS to place interventions on a two-dimensional plane so that between-treatment distances are preserved satisfactorily. Chung and Lumley were the first ones to attempt a similar analysis in a NMA setting for investigating inconsistency between direct and indirect effect estimates[11]. This approach offers an important advantage compared to other methods in the literature since it provides a visual representation of the similarity of interventions. We can map these interventions so that the distances between them reveal their similarity in terms of magnitude of effects on multiple outcomes, allowing us to visualize which interventions group together. Additionally, we address the important question of weighting the various outcomes considering patients' profile and characteristics and incorporating these weights in the derived rankings. We present some ideas on how preferences can be quantified using regression methods that have been employed in the field of marketing.

2. **Methods**

2.1. Illustrative example

We will use a network of 212 randomized controlled trials (RCTs) and 43049 participants comparing 15 antipsychotic drugs and placebo[12]. The primary outcome was the mean overall change in symptoms of Schizophrenia measured in some standardized scale and secondary outcomes involve all-cause discontinuation (which is seen as a measure of acceptability) and weight gain. Efficacy and weight gain are continuous outcomes and are measured in the Standardized Mean Difference (SMD) scale whereas acceptability is dichotomous and is measured in the Odds Ratio (OR) scale. In instances where we had to combine the two outcomes, we transformed summary ORs to SMDs using $SMD = \frac{\sqrt{3}}{\pi}\log OR$[13]. More details about the ranking methods employed in this published NMA can be found in the original publication [12].

2.2. Placing weights on outcomes using conjoint analysis

Different stakeholders such as clinicians, policy makers and patients may have different views on the importance of outcomes and result in different intervention hierarchies. Additionally, physicians weigh outcomes differently depending on the patient's profile and baseline risk.

In marketing, a series of statistical methods have been developed to determine how people value different attributes of a product (feature, function, price etc.). Conjoint analysis has been suggested to determine a limited combination of attributes that are important for a product[14]. Such methods have also been used in health care; i.e. to determine which attributes of treatments for prostate cancer are more important to men[15] or to assess patients' preferences for a range of disease-modifying therapy attributes in multiple sclerosis[16].

We can use regression methods to determine which outcomes are most important. Suppose that the three outcomes in our example take three possible categorical values. For example, an antipsychotic can either have high, medium, or low efficacy, acceptability, and weight gain. A full-factorial design would consist of $3^3 = 27$ possible combinations. Stakeholders would be asked to put a preference score (e.g., from 1 to 10) or rank each of the 27 combinations. Table 1 shows such a hypothetical example from two individuals (one clinician and his/her patient) who put a preference score to the 27 designs.

Then, choosing one category (in the case "low" for Efficacy (E_{low}) and Acceptability (A_{low}) and "high" for Weight (W_{high})) as a reference category for each outcome, a regression model of the following type is fit

$$y = b_o + b_1 E_{medium} + b_2 E_{high} + b_3 A_{medium} + b_4 A_{high} + b_5 W_{low} + b_6 W_{medium}$$

The regression coefficients, b's, are called part-worth utilities and show the preference for a particular attribute of an outcome. All variables assume value 1 if a scenario has the relative attribute and zero otherwise. To compute the relative weight of each outcome, we estimate for each outcome the difference between the largest and smallest part-worth utility (note that for the reference category the utility is zero), then the weight of each outcome is the relative share of these numbers.

2.3. Multidimensional scaling (MDS), hierarchical tree clustering (HTS) and Individual Differences Multidimensional Scaling (IDMDS)

Multidimensional scaling (MDS) aims to reveal the structure of the data by plotting points preferably in one or two dimensions[17,18] The input data for MDS is in the form of a distance matrix representing the distances between pairs of objects. A classic example of its use is when we have a matrix of metric distances (e.g., in kilometers) among cities of a country and then MDS is used to reconstruct the actual map of the country in a two-dimensional plane. In this example, there is no ambiguity in the definition of a distance between two cities and in the requirement of a two-dimensional map. Where in general there is a degree of arbitrariness in the definition of a distance between pairs of objects and lack of knowledge on the number of dimensions needed[18]. In this work, we consider that the distance between two interventions reflects the relative effect for this pair of interventions. We can subsequently weigh this effect by its standard error. MDS is not a clustering algorithm per se, but a data reduction method. Once applied, we subsequently explore visually how points (treatments) cluster together. Technical information about the use of MDS in this context is given in the supplementary material.

Along with mapping treatments on a two-dimensional graph, we can also use hierarchical tree clustering (HTS) to build a hierarchy of clusters of the competing interventions and depict them using a dendrogram. For both MDS and HTS there are a variety of algorithms and software. Their comparison is beyond the scopes of this manuscript. For MDS we used the majorization algorithm as described in de Leeuw and Mair[19] and for HTC we used agglomerative complete linkage clustering.

To map the distances among objects perceived by different people, individual differences multidimensional scaling has been developed where instead of one distance matrix for a set of objects we have many[17,19]. We consider the same logic here but instead of one $T \times T$ league matrix for a single outcome we have several $T \times T$ matrices for the various outcomes of interest. We used R library SMACOF to conduct the analyses[19,20]. Note that the league table and the corresponding 95% confidence intervals

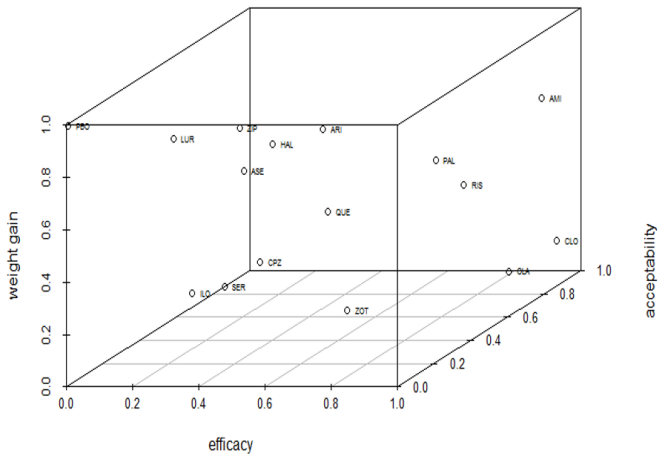
are standard outputs of the *netmeta* library in R[21,22]. More details are included in the supplementary material.

3. Results

Employing ranking methods to the Schizophrenia dataset

We start by focusing on each outcome separately and Figure 1 shows the three-way scatter plot for the SUCRA values for the three outcomes, as these were derived from the original publication[12]. If we focus on reduction in symptoms and all-cause discontinuation, we see that antipsychotics clozapine, amisulpride, olanzapine, risperidone and paliperidone form a distinct class of drugs taking the five top ranks in both outcomes. It is also noteworthy that although haloperidol performs satisfactorily on efficacy (7th rank), it performs poorly on acceptability (15th rank). If we include weight gain, things are getting blurred. Clozapine and olanzapine perform poorly on weight gain and only amisulpride performs well in all three outcomes.

Figure 1: Three-way scatter plot for the SUCRA values for efficacy, acceptability, and weight gain.

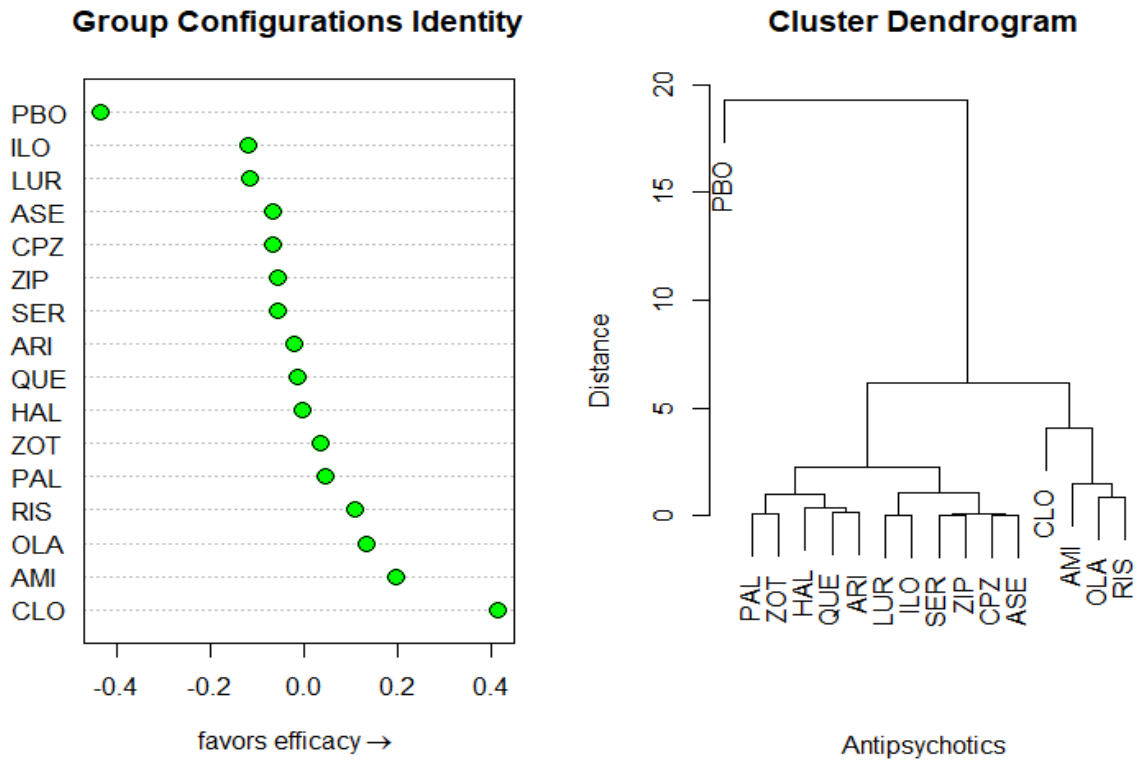


Antipsychotic	Sucra for Efficacy	Sucra for all-cause discontinuation	Sucra for weight
Clozapine (CLO)	0.99	0.83	0.17
Amisulpride (AMI)	0.92	0.92	0.69
Olanzapine (OLA)	0.85	0.89	0.05
Risperidone (RIS)	0.79	0.73	0.44
Paliperidone (PAL)	0.64	0.85	0.49
Zotepine (ZOT)	0.61	0.40	0.10
Haloperidol (HAL)	0.53	0.15	0.86
Quetiapine (QUE)	0.49	0.56	0.42
Aripiprazole (ARI)	0.46	0.56	0.72
Sertaline (SER)	0.35	0.23	0.28
Ziprasidone (ZIP)	0.35	0.33	0.84
Chlorpromazine (CPZ)	0.33	0.47	0.27
Asenapine (ASE)	0.32	0.40	0.65
Lurasidone (LUR)	0.19	0.25	0.84
Iloperidone (ILO)	0.17	0.40	0.19
Placebo (PLA)	0	0.01	0.99

We present a unidimensional configuration of the antipsychotics' efficacy in the left-hand side graph of Figure 2. Horizontal axis can be safely interpreted as showing efficacy with horizontal differences among drugs representing their differences in efficacy. We see that placebo lies far away from the rest of the drugs and that four drugs (CLO, AMI, OLA, RIS) seem to be slightly more effective than most of the remaining drugs. This configuration clearly reflects differences in treatment effects and/or ranking metrics such as SUCRA and p-score values. The stress value (a goodness-of-fit measure for MDS with values less than 0.2 considered acceptable) was 0.01, an indication that observed and expected distances are almost identical. In the right-hand side of Figure 2 we present the output from a hierarchical clustering in the form of a dendrogram where we see that drugs group in three or four clusters. Placebo is placed alone, an

indication that antipsychotics work as a whole. There is a group formed by clozapine, amisulpride, olanzapine and risperidone (or one may consider that clozapine forms a cluster on its own) and the rest of the antipsychotics formed a third group. In this Figure, we used complete-linkage clustering but we repeated the analysis using various agglomerative hierarchical clustering methods (those available in R function hclust[23]) and all methods produced either the same three distinct classes or considered clozapine to form a cluster on its own.

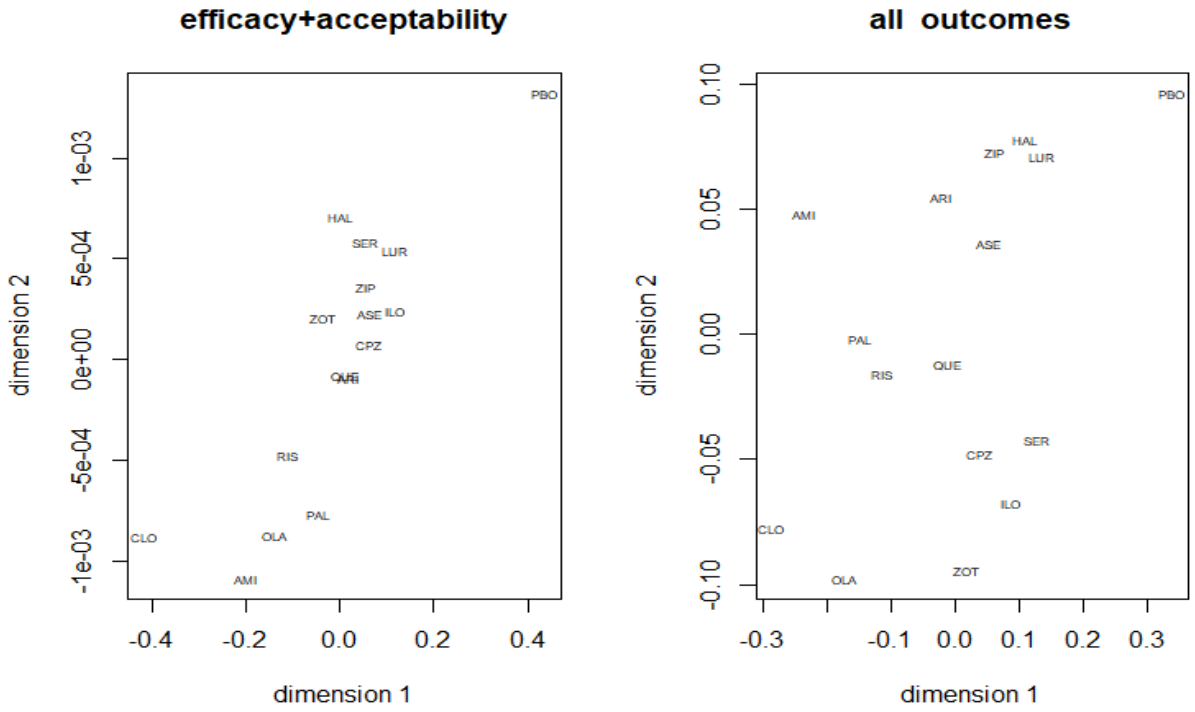
Figure 2: Unidimensional representation of efficacy using MDS (left-hand side plot). Horizontal axis refers to efficacy. Hierarchical tree clustering of the interventions considering efficacy alone (right-hand side plot).



We can use IDMDS to visualize antipsychotics considering multiple outcomes. Figure 3 shows the group configuration when only efficacy and acceptability are considered (left-hand side plot) and for all three outcomes (right-hand side plot). In the former case, we can safely assume that the horizontal axis shows differences in efficacy with small values showing more efficacious antipsychotics while the vertical axis shows differences in acceptability with smaller values showing better antipsychotics. The stress value was estimated to be 0.02. Efficacy and acceptability are positively correlated, and results are not much different than those derived from considering only efficacy. Five interventions (clozapine, amisulpride, olanzapine, risperidone and paliperidone) seem to form a cluster, placebo is alone, and the rest of antipsychotics have similar performance, forming a third cluster. Things become more complicated when we consider weight (right-hand side plot). In this case, differences among the five best antipsychotics have increased with olanzapine moving far away from the rest of the cluster. As expected, it is more difficult to interpret axes, though we may argue that small values in the horizontal axis show antipsychotics that score

high on efficacy and acceptability and differences in the vertical axis reflect differences in effect estimates in weight gain. Ideally, we are looking for antipsychotics placed on the left top corner (e.g. AMI). The stress value was estimated to be 0.21 suggesting that the fit has deteriorated when all outcomes are considered.

Figure 3: Group configuration using two dimensions for the efficacy and acceptability (left-hand side plot) and all three outcomes (right-hand side plot)



We repeat the analysis assuming different weights for each outcome. We consider the weights derived from the conjoint analysis of the ranking of two hypothetical stakeholders depicted in Table 1. Table 2 shows the estimated part-worth utilities for each attribute. We used the R library "*radiant*"[24] to estimate and produce Figures (Figure 4) of the estimated part-worth utilities. The relative importance stakeholder 1 (the clinician) places on each outcome are 57.4% for efficacy, 28.7% for acceptability and 13.8% for weight gain. The corresponding relative importance for the second stakeholder are 40.6%, 14.6% and 44.8% respectively. The cluster dendrograms of the antipsychotics under the two scenarios are presented in Figure 5. The left-hand side plot shows the dendrogram for the first stakeholder (clinician) whereas the right-hand side plot shows dendrogram for the second stakeholder (patient). In the left-hand side plot, clozapine, olanzapine, amisulpride, risperidone and paliperidone are close together although clozapine and olanzapine do much worse in weight. This is because the first stakeholder puts little emphasis on weight (13.8%). In the right-hand side plot, amisulpride moves away from its previous cluster and it is closer to antipsychotics who perform well in weight and moderately well in the other two outcomes.

Scenarios	Efficacy	Acceptability	Weight	Clinician	Patient
Profile 1	High	High	Low	10	10
Profile 2	High	High	Medium	10	9
Profile 3	High	High	High	9	4
Profile 4	High	Medium	Low	9	10
Profile 5	High	Medium	Medium	9	8
Profile 6	High	Medium	High	7	3
Profile 7	High	Low	Low	7	9
Profile 8	High	Low	Medium	5	6
Profile 9	High	Low	High	4	3
Profile 10	Medium	High	Low	8	8
Profile 11	Medium	High	Medium	7	7
Profile 12	Medium	High	High	6	2
Profile 13	Medium	Medium	Low	7	7
Profile 14	Medium	Medium	Medium	7	5
Profile 15	Medium	Medium	High	6	2
Profile 16	Medium	Low	Low	5	6
Profile 17	Medium	Low	Medium	4	5
Profile 18	Medium	Low	High	3	2
Profile 19	Low	High	Low	3	5
Profile 20	Low	High	Medium	3	4
Profile 21	Low	High	High	2	1
Profile 22	Low	Medium	Low	2	4
Profile 23	Low	Medium	Medium	2	3
Profile 24	Low	Medium	High	1	1
Profile 25	Low	Low	Low	1	3
Profile 26	Low	Low	Medium	1	1
Profile 27	Low	Low	High	1	1

Table 1: Full factorial design and preference scores for two hypothetical respondents (a clinician and a patient) who scored each of the possible 27 profiles with a score from one to ten.

Figure 4: Part-worth utilities as derived from the ratings of the clinician and the patient.

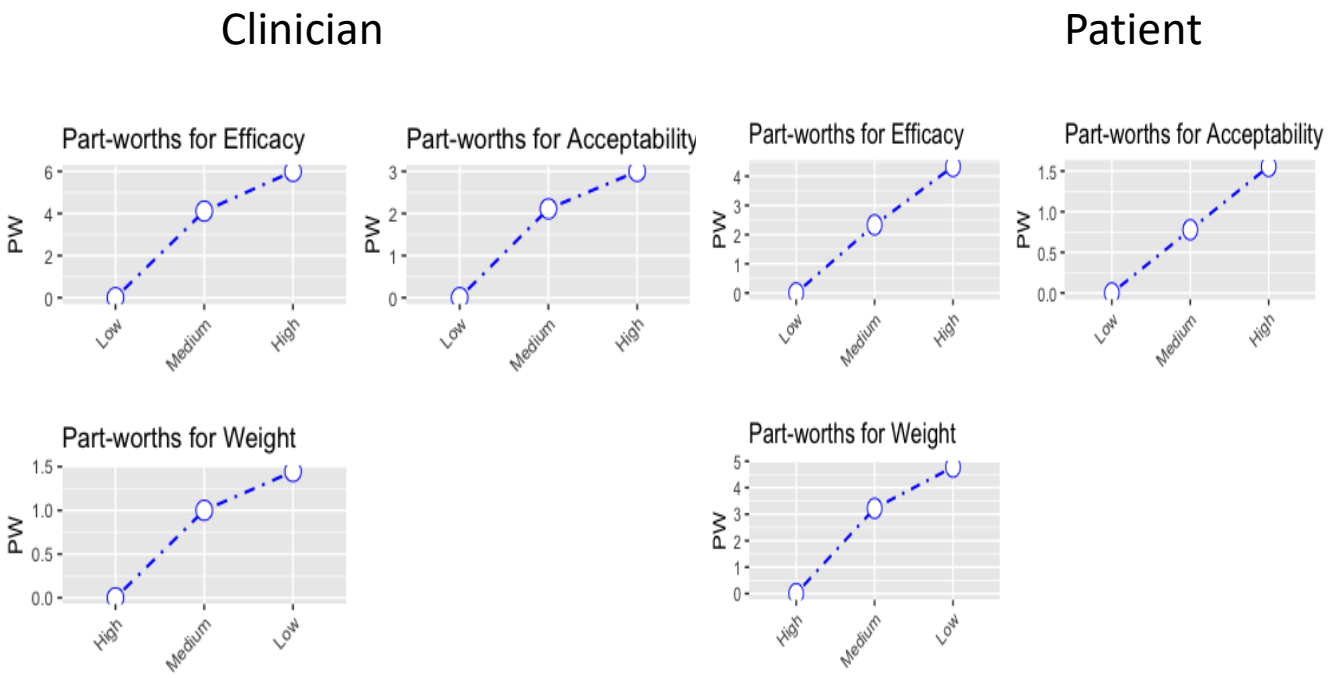
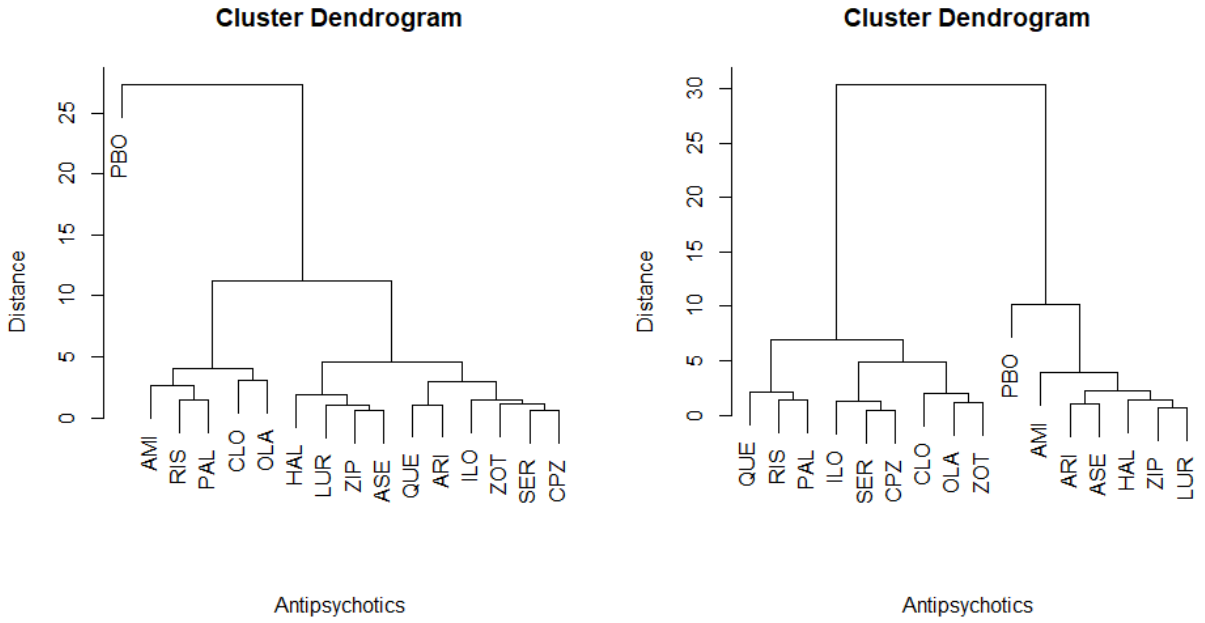


Table 2: Conjoint analysis results (part-worth utilities) for the two hypothetical individuals.

Coefficient/stakeholder	Clinician (respondent 1)			
	estimate	Standard error	t-value	p-value
β_0 (intercept)	-0.74	0.39	-1.93	0.07
β_1 (efficacy-medium)	4.11	0.36	11.55	<0.001
β_2 (efficacy-high)	6.00	0.36	16.85	<0.001
β_3 (acceptability-medium)	2.11	0.36	5.93	<0.001
β_4 (acceptability-high)	3.00	0.36	8.43	<0.001
β_5 (weight gain-medium)	1.00	0.36	2.81	0.01
β_6 (weight gain-high)	1.44	0.36	4.06	<0.001
Coefficient/stakeholder	**Patient (respondent 2)**			
	estimate	Standard error	t-value	p-value
β_0 (intercept)	-0.89	0.46	-1.94	<0.001
β_1 (efficacy-medium)	2.33	0.43	5.50	<0.001
β_2 (efficacy-high)	4.33	0.43	10.21	<0.001
β_3 (acceptability-medium)	0.78	0.43	1.83	0.08
β_4 (acceptability-high)	1.56	0.43	3.66	0.002
β_5 (weight gain-medium)	3.22	0.43	7.59	<0.001
β_6 (weight gain-high)	4.78	0.43	11.25	<0.001

Figure 5: Hierarchical tree clustering assuming different weights for each outcome. Relative weights for the hypothetical clinician (let-hand side plot) are 57.4% for efficacy, 28.7% for acceptability and 13.8% for weight gain. Relative weights for the hypothetical patient (right-hand side plot) are 40.6% for efficacy, 14.6% for acceptability and 44.8% for weight gain.



4. Discussion

Recommending an intervention is a complex issue in which both efficacy and safety should be considered. Multidimensional scaling offers a series of graphical methods that help us identify groups of interventions with similar performance.

We can borrow techniques used in marketing for evaluating the importance of products' attributes to assess the importance of outcomes. These methods provide a formal method by which preferences can be transformed to mathematical quantities.

There are certain risks and limitations associated with the suggested methodology. Using MDS and HTC for mapping treatments considering multiple outcomes is straightforward. However, relative effects and ranking metrics, do not include assessments of the quality of the evidence. Even interventions with large and precise effects scoring high on ranking metrics are not necessarily preferable if most information about them comes from low quality trials. Researchers using NMA results should consider the confidence placed on the relative effects[25,26]. Ideally, we are interested in treatments that not only rank high on the outcomes of interest, but they are also informed mainly by studies at low risk of bias[26].

A mere mapping of the similarity of treatments across multiple outcomes makes the unrealistic assumption that all outcomes are equally important for all patients. Assigning weights to outcomes is much trickier and should be applied with caution after having completed the systematic review. There are many subjective elements in the process that make the use of outcome weights within a systematic review problematic. Different researchers are expected to have different views and, additionally, these views

cannot be universally applied as patients have different profiles and baseline characteristics. Hence, we would not be able to generalize the results of the systematic review had we weighed outcomes within the systematic review.

Health professionals are encouraged to consider patient preferences to increase patient satisfaction and get better health outcomes. Our method provides a tool to consider patient outcome preferences for treatment recommendation but is not free of risks. Patient preferences may be unreasonable, differ considerably from those of their physician or with evidence-based guidelines and clinical practice[27]. Moreover, in clinical practice, it is not sufficient to give a questionnaire to patients to elicit their preferences. Patients should be sufficiently and unbiasedly informed about their condition, available treatment, and possible risks before expressing their preferences.

There are also some mathematical limitations. When employing the IDMDS, it is plausible that there is some dependence across outcomes. In most NMAs, relative effects are estimated separately for each outcome and between outcomes correlation is unknown unless one has the individual patient data. This is a general problem that concerns the vast majority of NMAs. We should also bear in mind that the resulting configuration is arbitrary, and axes have subjective interpretations. As a result, there is a subjective component on determining the clustering of interventions. Caution is needed when outcomes that favor different interventions are equally weighted as this was the case with our example and Figure 5. In such cases, interventions that do well in one of the two outcomes will cluster together even if they support different outcomes. Therefore, one should look at the actual relative effects (as PRISMA guidelines suggest) to understand which clusters of interventions are giving the best results.

Outcomes may also be heavily correlated (for example systolic and diastolic blood pressure) and in such a case we kind of use the same information twice. We expect outcomes to be correlated but caution is needed to avoid using very similar outcomes when employing these methods. One could argue that this was the case with efficacy and acceptability in our example and using one of them would suffice.

Conjoint analyses requires a careful design to elicit the appropriate information from each stakeholder. There are various methods in marketing that are used to elicit customers' preferences (e.g., conjoint analysis, discrete choice experiments). A comparison or description of those is beyond the scope of this manuscript. Ideally, the method should include questions tailored to each health condition. Here, for illustration purposes, we considered a simple example, with straightforward outcomes and attributes, but, in practice, attributes of outcomes or questions being put forward to stakeholders must be carefully designed. Some outcome scales are not easily understood by patients and any categorization of outcome values cannot hold universally as people have different baseline values (e.g., for example a clinician may consider that weight gain is not important, but it can be very important to an obese person with serious health problems). Once we have established a clustering of interventions one could decide between competing interventions on other grounds (e.g., cost, other adverse effects for which we do not have data, patient's reaction to an intervention).

5. References

1. Seitidis G, Nikolakopoulos S, Hennessy E, Tanner-Smith E, Mavridis D. Network Meta-Analysis Techniques for Synthesizing Prevention Science Evidence. *Prevention Science 2021*. 2021;1:1-10. doi:10.1007/S11121-021-01289-6

2. Salanti G. Indirect and mixed-treatment comparison, network, or multiple-treatments meta-analysis: many names, many benefits, many concerns for the next generation evidence synthesis tool. *Research Synthesis Methods*. 2012;3(2):80-97. doi:10.1002/jrsm.1037

3. Salanti G, Nikolakopoulou A, Efthimiou O, Mavridis D, Egger M, White IR. Introducing the Treatment Hierarchy Question in Network Meta-Analysis. *American Journal of Epidemiology*. 2022;191(5):930-938. doi:10.1093/AJE/KWAB278

4. Trinquart L, Attiche N, Bafeta A, Porcher R, Ravaud P. Uncertainty in Treatment Rankings: Reanalysis of Network Meta-analyses of Randomized Trials. *Annals of Internal Medicine*. 2016;164(10):666. doi:10.7326/M15-2521

5. Rücker G, Schwarzer G. Ranking treatments in frequentist network meta-analysis works without resampling methods. *BMC Medical Research Methodology*. 2015;15(58). doi:10.1186/s12874-015-0060-8

6. Salanti G, Ades AE, Ioannidis JPA. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial. *Journal of Clinical Epidemiology*. 2011;64(2):163-171. doi:10.1016/j.jclinepi.2010.03.016

7. Hutton B, Salanti G, Caldwell DM, et al. The PRISMA Extension Statement for Reporting of Systematic Reviews Incorporating Network Meta-analyses of Health Care Interventions: Checklist and Explanations. *Annals of Internal Medicine*. 2015;162(11):777. doi:10.7326/M14-2385

8. Chaimani A, Higgins JPT, Mavridis D, Spyridonos P, Salanti G. Graphical Tools for Network Meta-Analysis in STATA. *PLoS One*. 2013. doi:10.1371/journal.pone.0076654

9. Veroniki AA, Straus SE, Fyraridis A, Tricco AC. The rank-heat plot is a novel way to present the results from a network meta-analysis including multiple outcomes. *Journal of Clinical Epidemiology*. 2016;76:193-199. doi:10.1016/j.jclinepi.2016.02.016

10. 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. *Biometrical Journal*. 2019:bimj.201900026. doi:10.1002/bimj.201900026

11. Chung H, Lumley T. Graphical exploration of network meta-analysis data: The use of multidimensional scaling. *Clinical Trials*. 2008;5(4):301-307. doi:10.1177/1740774508093614

12. Leucht S, Cipriani A, Spineli L, et al. Comparative efficacy and tolerability of 15 antipsychotic drugs in schizophrenia: A multiple-treatments meta-analysis. *Lancet*. 2013. doi:10.1016/S0140-6736(13)60733-3

13. Chinn S. A simple method for converting an odds ratio to effect size for use in meta-analysis. *Statistics in Medicine*. 2000;19(22):3127-3131. doi:10.1002/1097-0258(20001130)19:22<3127::AID-SIM784>3.0.CO;2-M

14. Gustafsson A, Herrmann A, Huber F. *Conjoint Measurement : Methods and Applications*. Springer

Berlin Heidelberg; 2007, eds(4), doi: https://doi.org/10.1007/978-3-540-71404-0

15. Sculpher M, Bryan S, Fry P, de Winter P, Payne H, Emberton M. Patients' preferences for the management of non-metastatic prostate cancer: discrete choice experiment. BMJ. 2004 Feb 14;328(7436):382. doi: 10.1136/bmj.37972.497234.44..

16. Arroyo R, Sempere AP, Ruiz-Beato E*, et al*. Conjoint analysis to understand preferences of patients with multiple sclerosis for disease-modifying therapy attributes in Spain: a cross-sectional observational study. *BMJ Open* 2017;**7:**e014433. doi: 10.1136/bmjopen-2016-014433

17. Everitt B, Rabe-Hesketh S. *The Analysis of Proximity Data*. Kendall's advanced theory of statistics, Arnold; 1997.ISBN:0340677767.

18. Bartholomew DJ, Steele F, Galbraith J, Moustaki I. Analysis of Multivariate Social Science Data, Second Edition (Chapman & Hall/Crc Statistics in the Social and Behavioral Scie). June 2008. http://www.amazon.co.uk/dp/1584889608. Accessed March 4, 2022.

19. Leeuw JANDE, Mair P. Multidimensional scaling using majorization: SMACOF in R. *J Stat Softw*. 2009;31(3):1-30.

20. Mair P, De Leeuw J, Groenen PJF, Borg I, Maintainer]. Package “smacof” Title Multidimensional Scaling. 2021.

21. Rücker G, Schwarzer G, Krahn U, König J. Package “netmeta” Title Network Meta-Analysis using Frequentist Methods. 2018. doi:10.1007/978-3-319-21416

22. Team RC. R: A language and environment for statistical computing. 2013. http://www.r-project.org.

23. Murtagh F, Legendre P. Ward’s Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward’s Criterion? *Journal of Classification 2014 313*. 2014;31(3):274-295. doi:10.1007/S00357-014-9161-Z

24. CRAN - Package radiant. https://cran.r-project.org/web/packages/radiant/index.html.

25. 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;98:62-69. doi:10.1016/j.jclinepi.2018.02.008

26. Nikolakopoulou A, Higgins JPT, Papakonstantinou T, et al. CINeMA: An approach for assessing confidence in the results of a network meta-analysis. *PLOS Med*. 2020;17(4):e1003082. doi:10.1371/JOURNAL.PMED.1003082

27. Say RE, Thomson R. The importance of patient preferences in treatment decisions—challenges for doctors. *BMJ*. 2003;327(7414):542. doi:10.1136/BMJ.327.7414.542

Supplementary material

Description of how multidimensional scaling is used for clustering interventions in network meta-analysis (NMA)

MDS is used to represent the distance matrix on a configuration (e.g., Cartesian coordinate system) in a small number of dimensions such as that distances on the configuration represent approximately, the original distance matrix. For N=1,2 and 3, we can visualize the resulting points in a scatter plot.

Suppose that we have a connected network of T interventions and we employ NMA to estimate the $\frac{T(T-1)}{2}$ relative effects μ_{ij} with $i, j = 1, \dots, T \ \& \ j > i$. We will use MDS to group the T interventions in terms of the $\frac{T(T-1)}{2}$ relative effects and their corresponding 95% confidence intervals or standard errors s_{ij}. A league table can be viewed as a distance matrix by considering the absolute estimated relative effects $|\hat{\mu}_{ij}|$, making it symmetrical ($|\hat{\mu}_{ij}|=|\hat{\mu}_{ji}|$) and placing zeros in the diagonal ($|\hat{\mu}_{ii}| = 0 \ \forall \ i = 1, \dots, T$) . We have fitted a NMA model assuming consistency. The consistency equation ensures that $\hat{\mu}_{ij} = \hat{\mu}_{ik} + \hat{\mu}_{kj} \leftrightarrow |\hat{\mu}_{ij}| = |\hat{\mu}_{ik} + \hat{\mu}_{kj}| \leq |\hat{\mu}_{ik}| + |\hat{\mu}_{kj}| \ \forall \ i, j, k = 1, \dots, T, i \neq j \neq k$. Hence, relative effects define a metric that measures the distance between the relevant interventions. We use MDS to map the interventions in a sub-dimensional space (e.g., two-dimensional) and cluster them in groups. We find the configuration (set of coordinate values) m_{ij} by minimizing $\frac{\sum_{i<j} w_{ij}(m_{ij}-\hat{\mu}_{ij})^2}{\sum_{i<j} m_{ij}^2}$, also known as stress function, and using $w_{ij} = 1/s_{ij}$. Stress values lower than 0.2 are generally considered good with the lower the stress value the lower the difference between the observed distances (relative effect sizes) and the expected ones derived from the MDS configuration.

How to use the suggested methodology in practice

The whole methodology depends on methods that have been established in the statistical literature. For example, for researchers familiar with R, one can use

- The *netmeta* library to take the league matrix with all pairwise relative effects and the corresponding standard errors. If we have an object of class *netmeta*, the component *TE.random* gives the estimated overall relative treatment effects for all pairs of treatments for the random-effects model and the component *seTE.random* gives the corresponding standard errors (replacing "*random*" with "*fixed*" would give the corresponding results for the fixed effect model).
- The *smacof* library to conduct multidimensional scaling (command *mds*) or individual differences multidimensional scaling (command *smacofIndDiff*).
- The command hclust to conduct hierarchical cluster analysis on a set of dissimilarities
- The *conjoint* library to conduct a conjoint analysis to estimate part-worth utilities and subsequently use them to estimate the relative importance of outcomes.