

A Markov chain approach for ranking treatments in network meta-analysis

Anna Chaimani, Raphaël Porcher, Emilie Sbidian, Dimitris Mavridis

▶ **To cite this version:**

Anna Chaimani, Raphaël Porcher, Emilie Sbidian, Dimitris Mavridis. A Markov chain approach for ranking treatments in network meta-analysis. Statistics in Medicine, 2020, Online ahead of print. 10.1002/sim.8784 . inserm-02982222

HAL Id: inserm-02982222

https://inserm.hal.science/inserm-02982222

Submitted on 17 Nov 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Received: 20 September 2019 | Revised: 10 September 2020 | Accepted: 29 September 2020

DOI: 10.1002/sim.8784

RESEARCH ARTICLE

Statistics in Medicine WILEY

A Markov chain approach for ranking treatments in network meta-analysis

Anna Chaimani[1,2] | **Raphaël Porcher**[1] | **Émilie Sbidian**[3,4] | **Dimitris Mavridis**[1,5]

[1]Université de Paris, Research Center of Epidemiology and Statistics (CRESS-U1153), INSERM, Paris, France

[2]Cochrane France, Paris, France

[3]Université Paris-Est Créteil, UPEC, EpiDermE EA 7379, Créteil, France

[4]AP-HP, Hôpitaux Universitaires Henri Mondor, Département de Dermatologie, UPEC, Créteil, France

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

Correspondence
Anna Chaimani, Université de Paris, Research Center of Epidemiology and Statistics (CRESS-U1153), INSERM, 1 place du Parvis Notre-Dame, 75004 Paris, France.
Email: anna.chaimani@inserm.fr

When interpreting the relative effects from a network meta-analysis (NMA), researchers are usually aware of the potential limitations that may render the results for some comparisons less useful or meaningless. In the presence of sufficient and appropriate data, some of these limitations (eg, risk of bias, small-study effects, publication bias) can be taken into account in the statistical analysis. Very often, though, the necessary data for applying these methods are missing and data limitations cannot be formally integrated into ranking. In addition, there are other important characteristics of the treatment comparisons that cannot be addressed within a statistical model but only through qualitative judgments; for example, the relevance of data to the research question, the plausibility of the assumptions, and so on. Here, we propose a new measure for treatment ranking called the *Probability of Selecting a Treatment to Recommend* (POST-R). We suggest that the order of treatments should represent the process of considering treatments for selection in clinical practice and we assign to each treatment a probability of being selected. This process can be considered as a Markov chain model that allows the end-users of NMA to select the most appropriate treatments based not only on the NMA results but also to information external to the NMA. In this way, we obtain rankings that can inform decision-making more efficiently as they represent not only the relative effects but also their potential limitations. We illustrate our approach using a NMA comparing treatments for chronic plaque psoriasis and we provide the Stata commands.

KEYWORDS

comparative effectiveness research, multiple treatments, selection probabilities, stochastic process, treatment hierarchy

1 | INTRODUCTION

Network meta-analysis (NMA) provides the highest possible level of evidence for the development of clinical guidelines and several health care organizations already incorporate NMA findings in their guidance. NMA provides estimates with increased precision and allows estimating the relative effectiveness for interventions that have never been compared head-to-head. Additionally, NMA provides a hierarchy for a specific outcome of all alternative treatments of a comparative

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2020 The Authors. *Statistics in Medicine* published by John Wiley & Sons Ltd.

effectiveness review as long as they form a connected network. Treatment ranking has gained much attention as well as a lot of criticism over the last years.[1–4] The main criticisms are: (a) methods widely used in the literature have focused on the probability of each treatment being the best without taking into account the whole ranking distribution and have produced misleading results,[5,6] (b) ranking is a very influential output and interpretation in isolation from relative effects may lead to spurious conclusions,[7] and (c) ranking of treatments most often is not interpreted in light of the limitations of the evidence base (such as risk of bias or insufficient evidence).

To date, the most appropriate approaches for treatment ranking in NMA are probably the Surface Under the Cumulative RAnking (SUCRA) curves values[8] and the more recent P-scores.[9] They express *the mean extent of certainty that a treatment is better than the other competing treatments*. The two measures are equivalent and differ only in that SUCRAs are obtained using resampling methods while P-scores are derived analytically.[9] Their main advantage is that they account for the variability in treatment ranks by considering not only the magnitude of relative effects but also their uncertainty and overlap of their confidence/credible intervals.

However, SUCRAs and P-scores are limited by the fact that they are based solely on the NMA relative effects and other factors that are often of interest to policy makers in treatment guidelines can only be incorporated in a qualitative way. For instance, if there are reasons to have less confidence in some relative effects than in others, then the usefulness of a conventional treatment ranking becomes questionable. In addition, treatments poorly connected to the rest of the network tend to appear at the top ranks although the evidence favoring such treatments is often insufficient or even biased.[10] Also, safety is a crucial aspect of treatment performance, though very often disregarded in the evidence synthesis setting due to absence of appropriate and adequate data. Last, the cost of the treatments is another important factor when forming clinical guidelines and should be taken into account at the country level.

To address some of these points, Salanti et al[11] developed a framework for evaluating the confidence in the treatment ranking from NMA, while Chaimani et al[12] suggested a graphical tool depicting the ranking of treatments based on two characteristics. The drawback of the former approach is that a low-confidence ranking would fail to answer its target question (ie, which are the most appropriate treatments) and efficiently inform medical decision-making. The limitation of the latter approach is that it does not provide a quantitative way to rank the available treatments considering jointly the two characteristics of interest.

In this article, we suggest the *Probability Of Selecting a Treatment to Recommend* (POST-R) as a new measure to rank treatments after conducting the systematic review and the NMA. Our approach is based on a Markov chain (MC) model and allows taking into account both the relative effects and the other aforementioned characteristics that may affect treatment choice in practice. MC models have already been employed in several fields for the development of hierarchies and optimization strategies, such as in social networks[13] and operations research.[14] Here we use a MC approach in which the conditional probability distribution of future states of the treatment selection process depends only on the current state (Markov property); the stationary distribution of this Markov process is regarded as the probabilities of recommending each treatment. The suggested algorithm allows incorporating information other than that of the relative effects through the initial probabilities assigned to each treatment. The rest of this article is structured as follows: in Section 2, we describe an exemplar NMA used to illustrate our ranking approach. Section 3, after a brief description of P-scores, introduces our ranking method and describes how some key characteristics on top of relative effects can be incorporated in treatment selection. Finally, in Section 4, we use the example dataset to illustrate the method and, in Section 5, we discuss the strengths and limitations of our approach.

2 | ILLUSTRATIVE EXAMPLE

To illustrate the use of our method and compare the results with the conventional ranking results according to SUCRA (or P-score), we use a previously published NMA. Our example is a Cochrane review[15] comparing the efficacy and safety of 19 different drugs and placebo for patients with moderate to severe psoriasis. The primary outcomes were the events of achieving clear or almost clear skin (ie, PASI 90) for efficacy and serious adverse events (SAE) for safety.

The authors report in their results for efficacy " … ixekizumab was the best treatment at drug level (versus placebo: RR 32.45, 95% CI 23.61 to 44.60; SUCRA = 94.3; high-certainty evidence), followed by secukinumab (versus placebo: RR 26.55, 95% CI 20.32 to 34.69; SUCRA = 86.5; high-certainty of evidence), brodalumab (versus placebo: RR 25.45, 95% CI 18.74 to 34.57; SUCRA = 84.3; moderate-certainty evidence), guselkumab (versus placebo: RR 21.03, 95% CI 14.56 to 30.38; SUCRA = 77; moderate-certainty evidence) … "; and for safety " … methotrexate was associated with the best safety profile at drug level in terms of serious adverse events (versus placebo: RR 0.23, 95% CI 0.05 to 0.99; SUCRA = 90.7; moderate-certainty evidence), followed by ciclosporin (versus placebo: RR 0.23, 95% CI 0.01 to 5.10; SUCRA = 78.2; very

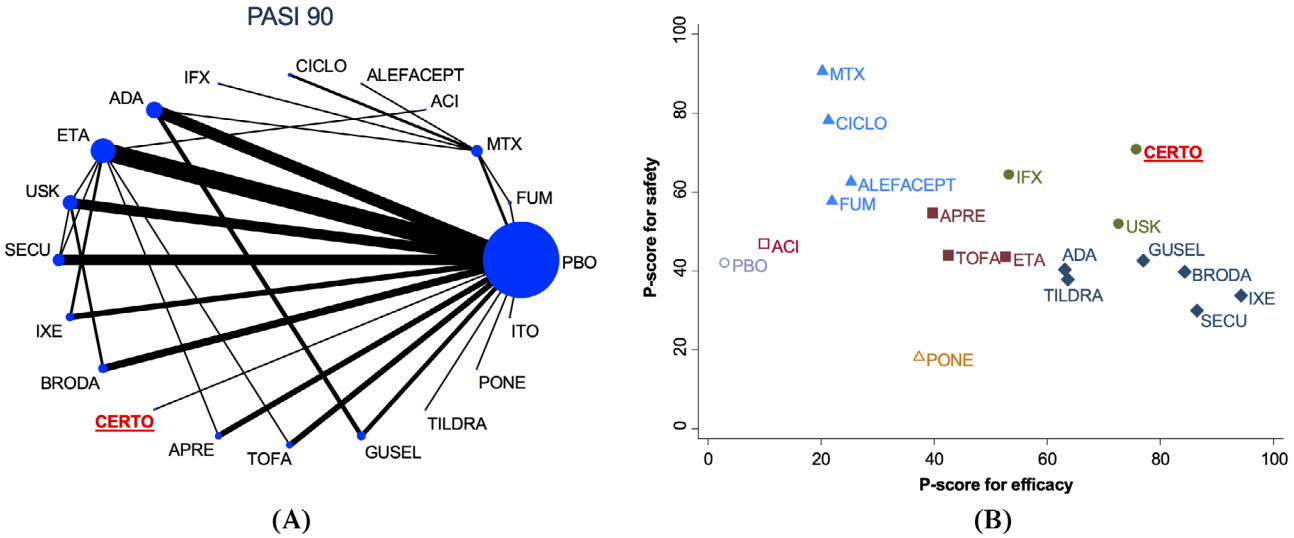


FIGURE 1 Data and ranking results of the psoriasis network. A, The network diagram for efficacy (PASI 90). B, The treatment ranking based on P-scores for efficacy and safety jointly. Different colors represent different clusters of treatments with respect to their ranking on both outcomes jointly. ACI = Acitretin, ADA = Adalimumab, ALEFACEPT = Alefacept, APRE = Apremilast, BRODA = Brodalumab, CERTO = Certolizumab, CICLO = Ciclopsorin, ETA = Etanercept, FUM = Fumaric acid esters, GUSEL = Guselkumab, IFX = Infliximab, ITO = Itolizumab, IXE = Ixekizumab, MTX = Methotrexate, PBO = Placebo, PONE = Ponesimod, SECU = Secukinumab, TILDRA = Tildrakizumab, TOFA = Tofacitinib, USK = Ustekinumab [Color figure can be viewed at wileyonlinelibrary.com]

low-certainty evidence) […] Nevertheless, analyses on serious adverse events were based on a very low number of events and were reduced to the short time frame of the trials".[15] These findings imply that (a) the confidence (or certainty) of the evidence is not consistent across drugs and (b) the authors are somewhat concerned about the safety results due to rare events and short follow-up of the trials. However, the ranking based on SUCRA provided in the published paper does not account for these factors.

In addition, considering the SUCRAs for both efficacy and safety the original review suggests that Certolizumab, a treatment connected rather weakly to the rest of the network, achieves the better compromise between the two outcomes (Figure 1). Due to the scarcity of the evidence for that treatment, though, several clinicians might be sceptic about this finding. In our application we show how treatments are ranked using the POST-R measure and compare results from each approach.

3 | METHODS

3.1 | Conventional ranking in NMA

Several different approaches have been used in the literature to rank competing interventions in a network of trials.[6] Here we describe briefly how P-score[9] (the analytical equivalent of SUCRA[8]), which is based on both the mean and the variance of the NMA relative effects, is computed.

Consider a network with N treatments (nodes). P-score is based on the probabilities $p_{i>j}$ that treatment i produces a better outcome than treatment j with $i, j = 1, \ldots, N$ and $p_{i>j} = 0$ for $i = j$. Ruecker and Schwarzer[9] showed that, assuming that the relative effects follow a normal distribution, these probabilities can be obtained as

$$p_{i>j} = \Phi\left(\frac{\hat{\mu}_i - \hat{\mu}_j}{\hat{\sigma}_{ij}}\right), \tag{1}$$

where Φ is the cumulative density function of the standard normal distribution, $\hat{\mu}_i - \hat{\mu}_j$ is the estimated difference in the studied outcome between the two treatments, and $\hat{\sigma}_{ij}$ is the estimated SE of this difference. Once we calculate the $p_{i>j}$ probabilities for every possible pair of treatments (i, j), then the P-score for treatment i is

$$\overline{P}_i = \frac{1}{N-1}\sum_{j=1}^{N} p_{i>j}.$$

3.2 | The POST-R measure

3.2.1 | Treatment ranking as a discrete stochastic process

We assume a Markov process (S) with a countable state space where every treatment (node) $i = 1, \dots, N$ within the network under investigation is a state s_i. S starts at time $t = 1$ by moving between the N treatment options with an aim to identify those with the best performance for the main outcome of interest (eg, efficacy). Each time-point $t = 1, 2, 3, \dots$ corresponds to a new step in S. A movement from i to j in this setting implies that we were not satisfied with i and we select j as a potentially more beneficial treatment. Since movements represent the preference between two treatments, they are not uniformly distributed but they are based on prior beliefs and actual data. So, let $l_i^{(t)}$ be the probability of selecting treatment i at time (ie, step) t and $\mathbf{Q^{(t)}} = (l_1^{(t)}, \dots, l_N^{(t)})'$. The vector $\mathbf{Q^{(0)}}$ is called the *initial state probability vector* with entries the probabilities $l_1^{(0)}, \dots, l_N^{(0)}$ of selecting each treatment i at time $t = 0$; hence without looking at the NMA relative effects for the outcome of interest. Different ways for defining $\mathbf{Q^{(0)}}$ are suggested in Section 3.3.

A key aspect of the model is the definition of the *transition probabilities* $p_{i\to j}^{(1)}$ at the beginning of S; that corresponds to the probability of selecting treatment j at $t = 2$ after having selected treatment i when $t = 1$. These probabilities form the component of the MC model that incorporates the NMA relative effects for updating the initial probabilities. We suggest that $p_{i\to j}^{(1)}$'s can be driven by the probabilities $p_{i>j}$ in Equation (1) implying that at $t = 1$ movements between treatments are solely based on the probabilities that each treatment i produces a better outcome than j. Given that the transition probabilities at every t should satisfy $\sum_{j=1}^{N} p_{i\to j}^{(t)} = 1$, we define

$$p_{i\to j}^{(1)} = \frac{p_{j>i}}{\sum_{i=1}^{N} p_{j>i}}. \tag{2}$$

It follows from Equations (1) and (2) that $p_{i\to i}^{(1)} = 0$ while for any pair (i, j) $p_{i\to j}^{(1)} > 0$. As a result, the transition matrix at the beginning of S is

$$\mathbf{A^{(1)}} = \begin{pmatrix} 0 & p_{1\to 2}^{(1)} & \cdots & p_{1\to N}^{(1)} \\ p_{2\to 1}^{(1)} & 0 & \cdots & p_{2\to N}^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ p_{N\to 1}^{(1)} & p_{N\to 2}^{(1)} & \cdots & 0 \end{pmatrix}$$

while the transition matrix at each t is the product of transition matrices until t, $(\mathbf{A^{(1)}})^t$.[16] For simplicity, in the rest of the manuscript we suppress the superscript (1) when $t = 1$. Equation (2) implies that $\mathbf{A}^t$ corresponds to an irreducible (ie, $p_{i\to j}^{(t)} > 0$ for some $t > 0$) and ergodic (since for all states $p_{i\to i}^{(t)} = 0$ for $t = 1, 2, 3, \dots$) MC.[16] When the MC is ergodic it has a unique stationary distribution $\boldsymbol{\pi}$ to which it converges for large values of t, namely $\lim_{t\to\infty} \mathbf{Q^{(t)}} = \boldsymbol{\pi}$. It has been shown that

$$\boldsymbol{\pi} = \mathbf{A}\boldsymbol{\pi}, \tag{3}$$

and thus $\boldsymbol{\pi}$ is an eigenvector of $\mathbf{A}$ with eigenvalue 1.[17] For a treatment network, this stationary probability distribution of the MC contains the probabilities $l_i^{(t\to\infty)}$ of selecting each treatment after many steps $\boldsymbol{\pi} = (l_1^{(t\to\infty)}, \dots, l_N^{(t\to\infty)})'$ and offers an intuitive way to rank the treatments of a NMA.

According to Equation (3), treatment ranking could be easily obtained as an eigenvector (with eigenvalue 1) of matrix $\mathbf{A}$; however, it is informed only by the relative effects (via the transition probabilities in $\mathbf{A}$) and therefore it shares the same limitation with the conventional ranking approaches discussed earlier.

3.2.2 | Incorporating the initial probability distribution in ranking

To build our model, we employ a special case of an MC previously used for the development of the PageRank algorithm that ranks websites by Google search results.[13] The difference between Google's model and other MC models is that the former considers that there is always a probability of starting again the stochastic process according to the initial

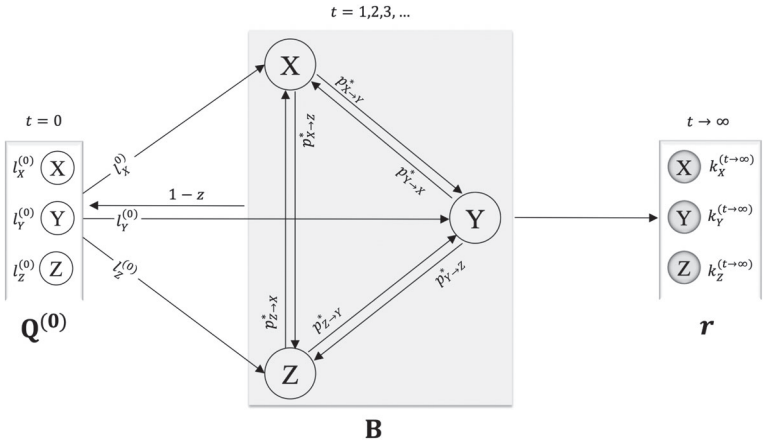


FIGURE 2 Graphical representation of the Markov Chain model for treatment selection using a hypothetical example of a three-treatment network. Arrows represent the movement from one treatment to the other that happens when a change in the preference between the two treatments occurs

probability distribution. So, let $z \in (0, 1)$ be the probability that the treatment selection process continues with updating at each step the probability distribution according to **A** and $(1-z)$ the probability that the process starts again from $\mathbf{Q^{(0)}}$. It should be noted that such a scenario reflects situations where the NMA fails to adequately update the existing evidence (eg, due to few data) and conclusions are based to some extent on previous knowledge or other characteristics (such as data quality). More discussion on the definition of z is presented in Section 3.4.

This MC configuration leads to the modified transition probabilities

$$p^*_{i\to j} = z p_{i\to j} + (1-z) l^{(0)}_j, \tag{4}$$

and the modified transition matrix

$$\mathbf{B} = \begin{pmatrix} p^*_{1\to 1} & p^*_{1\to 2} & \cdots & p^*_{1\to N} \\ p^*_{2\to 1} & p^*_{2\to 2} & \cdots & p^*_{2\to N} \\ \vdots & \vdots & \ddots & \vdots \\ p^*_{N\to 1} & p^*_{N\to 2} & \cdots & p^*_{N\to N} \end{pmatrix},$$

with $p^*_{i\to i} = (1-z) l^{(0)}_i$.

Given that $\sum_{j=1}^N p_{i\to j} = 1$ and $\sum_{j=1}^N l_j = 1$ it holds that $\sum_{j=1}^N p^*_{i\to j} = 1$. We denote the unique stationary probability distribution (always existing for $z \in (0, 1)$) of the transition matrix **B** with the vector $\boldsymbol{r} = (k_1^{(t\to\infty)}, \ldots, k_N^{(t\to\infty)})'$. Note that this matrix can be written also in the form:

$$\mathbf{B} = (z\mathbf{A}' + (1-z)\mathbf{Q^{(0)}}\mathbf{1}')', \tag{5}$$

where **1** is a $N \times 1$ vector with all elements equal to 1.[18] It follows from ergodic theory for the new MC (S^*) that S^* has a unique stationary distribution $\boldsymbol{r}$ with $\boldsymbol{r} = \mathbf{B}\boldsymbol{r}$ and thus the probability distribution $\boldsymbol{r}$ is an eigenvector of **B** with eigenvalue 1.

A graphical representation of our MC model for treatment selection is provided in Figure 2. The graph shows a hypothetical example of a network with three alternative treatment options. The process S starts from $\mathbf{Q^{(0)}}$ and based on existing knowledge about each treatment we move to one of the X_1, X_2, X_3 with probability $l^{(0)}_{X_1}, l^{(0)}_{X_2}, l^{(0)}_{X_3}$, respectively. Let us assume that we first select X_1 (at $t = 1$) where we are informed about the NMA relative effects. At the next step we have three options:

(i) to select X_2 with probability $p^*_{X_1\to X_2} = z p_{X_1\to X_2} + (1-z) l^{(0)}_{X_2}$;
(ii) to select X_3 with probability $p^*_{X_1\to X_3} = z p_{X_1\to X_3} + (1-z) l^{(0)}_{X_3}$;
(iii) to go back to $\mathbf{Q^{(0)}}$ and start again from X_1 with probability $(1-z) l^{(0)}_{X_1}$.

After several steps, we will have tried all treatments and will have formed the final selection probabilities $k^{(t\to\infty)}_{X_1}, k^{(t\to\infty)}_{X_2}, k^{(t\to\infty)}_{X_3}$.

3.3 | Defining the $\mathbf{Q^{(0)}}$ vector

The initial state probability vector in $\mathbf{Q^{(0)}}$ is the key feature that differentiates our method with the existing ranking approaches (see Section 3.1), and allows incorporating information other than NMA relative effects for treatment recommendation. It is, though, challenging to find a reasonable way to define the initial probabilities $l_i^{(0)}$. Investigators should bear in mind that each $l_i^{(0)}$ represents the probability of selecting every treatment i in clinical practice without knowing the NMA relative effects for the outcome of interest and consequently neither the probabilities $p_{i>j}$. In such a case, treatment choice depends on other characteristic(s) expressed via $\mathbf{Q^{(0)}}$. Here we consider the following characteristics, which we believe are among the most important for forming treatment guidelines:

- Confidence in the evidence: evidence on some of the treatments might be less "trustworthy" than for others. It should be noted that confidence in the evidence from NMA is based on the network structure, precision, the validity of the assumptions, assessment of risk of bias, and so on, so it is dependent on the NMA setting but independent of the actual values of the relative effects.
- Clinical experience: prior information from clinical practice is important and is not always in agreement with study results as the latter may lack power, have a short follow-up period, and so on.
- Cost of the treatments: the relative outcomes and costs of available interventions should be assessed, cheaper treatments might be preferable if they yield similar outcomes to slightly more effective, but expensive, ones.
- Treatment safety: efficacy and safety should always be considered jointly when forming recommendations.

It is evident that the definition of the probabilities $l_i^{(0)}$ is subjective to some extent and depends on the medical field and research question. In the following sections, we suggest ways for defining $\mathbf{Q^{(0)}}$ using the four characteristics mentioned above.

3.3.1 | Confidence in the evidence as a criterion for $\mathbf{Q^{(0)}}$

All recent reporting guidelines for NMA recommend including in the manuscript a table with the confidence (called also quality or certainty in the literature) for every treatment comparison in the network.[7,19] To date, two approaches are available for preparing such a table.[11,20] Here we consider the approach by Salanti et al[11] and the recent updates implemented in the online tool CINeMA (available from http://cinema.ispm.ch/).[21,22] Note that such an evaluation typically is conducted after completing the systematic review and the NMA and not at the beginning of the process. However, it still can be used to form "prior" information on treatment ranks.

We focus on the final report of the full evaluation where there are four possible levels of confidence assigned at each comparison i vs j: high, moderate, low, and very low. We translate these levels into probabilities that the evidence is trustworthy. Reasonable values might be $P(trust)_{i-j} = 0.9, 0.7, 0.3$, and 0.1, respectively. Then, we can obtain the following confidence matrix

$$\mathbf{C} = \begin{pmatrix} 0 & c_{2-1} & \dots & c_{N-1} \\ c_{1-2} & 0 & \dots & c_{N-2} \\ \vdots & \vdots & \ddots & \vdots \\ c_{1-N} & c_{2-N} & \dots & 0 \end{pmatrix},$$

where c_{i-j} is the probability of "trusting" the use of treatment i more than treatment j. We define c_{i-j} as follows:

- If the mean relative effect between i and j favours i, then $c_{i-j} = P(trust)_{i-j}$ and $c_{j-i} = 0$.
- If the mean relative effect between i and j favours j, then $c_{j-i} = P(trust)_{i-j}$ (note that $P(trust)_{i-j} = P(trust)_{j-i}$) and $c_{i-j} = 0$.
- If the mean relative effect between i and j does not favour any treatment, then $c_{i-j} = P(trust)_{i-j}/2$ and $c_{j-i} = P(trust)_{i-j}/2$.

It is important to note that what relative effect would lead to favor one treatment over the other depends on the clinical setting. For example, the investigators might define a priori the value of a clinically important effect to be used (such as an estimated risk ratio of 1.1, 1.2, etc.). Using the columns of **C**, we can obtain an estimate of the average confidence for using each treatment i as

$$\overline{C}_i = \frac{1}{N-1}\sum_{j=1}^{N} c_{i-j},$$

which is the analogous to the P-score ($\overline{P}_i$). Since $\mathbf{Q^{(0)}}$ is a probability vector, its elements should add up to 1. Therefore, we define

$$l_i^{(0)} = \overline{C}_i / \sum_{i=1}^{N} \overline{C}_i. \tag{6}$$

This is the probability to select treatment i among all treatments in the network given the matrix **C**.

3.3.2 | Clinical experience as a criterion for $\mathbf{Q^{(0)}}$

Several methods have been suggested in the literature for expert opinion elicitation on probability distributions.[23] Possibly one of the simplest approaches that can be easily understood by clinicians is the so-called *roulette method* which has been implemented also in the MATCH tool (http://optics.eee.nottingham.ac.uk/match/uncertainty.php).[24] Briefly, we adapted the method to our setting as follows: We provide to experts a grid with N equally sized bins each corresponding to a treatment in the network. Subsequently, we ask from the experts to allocate 100 chips (representing 100 patients) to the N bins based on their experience about the performance of the treatments in terms of a particular characteristic; for instance, by considering their safety alone. Then, the initial probability $l_i^{(0)}$ of selecting each treatment i is the proportion of chips allocated to the respective bin. Inevitably, expert opinions are subjective to some degree but asking several experts may mitigate this subjectivity. A simple (although not perfect) way, for instance, to combine opinions from 2 or more experts is to use as $l_i^{(0)}$ the weighted average of the expert-specific probabilities with weights the years of experience.

3.3.3 | Cost as a criterion for $\mathbf{Q^{(0)}}$

Consider that T is the most expensive treatment in the market among the N options. We define

$$d_i = \begin{cases} 0, & if\ i = T \\ 1 - \frac{cost\ of\ i}{cost\ of\ T}, & if\ i \neq T \end{cases}.$$

This means that the probability of selecting treatment i based on its cost is related to the percentage of the cost of T that i has. Then, the probabilities $l_i^{(0)}$ are obtained by Equation (6), replacing $\overline{C}_i$ by d_i.

3.3.4 | Safety as a criterion for $\mathbf{Q^{(0)}}$

Although we focus on treatment safety here, the approach applies equally to any other outcome. The combination of two outcomes in the treatment selection process can be made through a two-stage approach:

- At the first stage we rank the treatments for the "less" important outcome between the two (eg, safety if efficacy is considered more important) using the POST-R approach.
- Then, at the second stage the stationary probability distribution r obtained for this outcome becomes the initial probability vector $\mathbf{Q^{(0)}}$ in Equation (5) to be combined with A for the overall treatment selection probabilities.

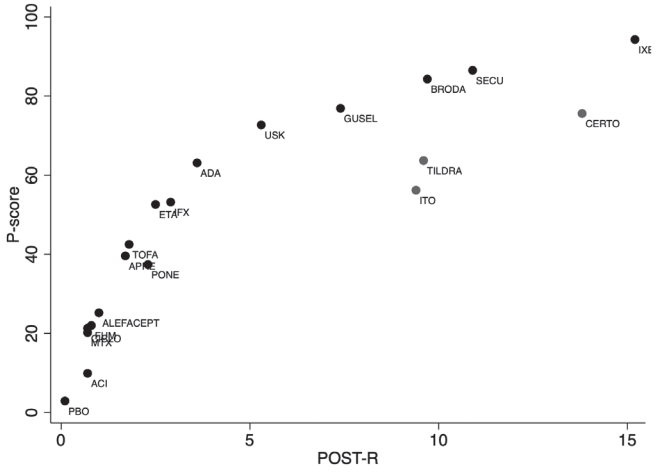


FIGURE 3 Relationship between P-scores and POST-R values when only the relative effects are considered (ie, $z = 1$). Red points correspond to drugs not presenting a good agreement between the two measures and r is the Pearson correlation coefficient

3.4 | Defining the probability z

Considering that an NMA is (or should be) conducted when there is a need for updating and extending the existing evidence, we believe that z should generally take values larger than 0.5. We suggest that values between 0.6 and 0.9 might be reasonable. It is not straightforward, though, how to select the optimal z value as this might depend also on the clinical setting and the available data. Running a sensitivity analysis on a range of values for z could show how robust ranking results are under different scenarios. The value of z can also be informed by expert opinion. For example, we can ask several experienced clinicians to rate (eg, in a scale) the usefulness/necessity of updating the evidence and clinical practice using the NMA relative effects. In simpler words, we can ask experts or even patients to indicate "how likely it would be to try a drug," for instance, with only low confidence evidence available (when $\mathbf{Q^{(0)}}$ represents confidence in the evidence) or known to cause serious adverse effects (when $\mathbf{Q^{(0)}}$ represents drug safety).

4 | APPLICATION TO PSORIASIS NETWORK

4.1 | Conventional treatment ranking

Following the original Cochrane review, we performed a random effects NMA using a multivariate meta-analysis approach assuming a common heterogeneity across comparisons.[25] Table 1 shows the P-score percentages and the respective ranks for all treatments for the two primary outcomes (PASI 90 and SAE). These were obtained using the *network* and *network graph* packages in Stata.[25,26] The dataset and the Stata script used to obtain the results can be found in the Supplementary files. The league table of the relative effects and the respective league table with the probabilities $p_{i>j}$ are available in the Appendix (Appendix Tables 1 and 2, respectively). In Figure 3, we show the association between P-scores and POST-R for efficacy based only on the relative effects, hence assuming $z = 1$. Although P-scores and POST-R values are different (since they are measuring different probabilities), there is a good agreement in terms of the treatment ranks. Differences appear mainly for treatments with large uncertainty, such as certolizumab, itolizumab, and tildrakizumab (Appendix Figure 1).

4.2 | Treatment ranking combining efficacy and confidence in the evidence

Using the CINeMA application, we obtained the domain-specific judgments for each comparison in the network and then we summarized them into an overall confidence for every comparison (Appendix Table 3). Note that the original Cochrane review used similar methodology to that implemented in CINeMA[11] but not exactly the same for all domains as the tool was only released in 2017 (ie, the publication year of the review). The last column of the table presents the probability $P(trust)$ for each comparison. We considered overall in CINeMA a risk ratio of 2 as a clinically important relative effect between two treatments to justify selection of one treatment over the other. This clinically important effect was used for

TABLE 1 Ranking of treatments considering only one characteristic at a time without combining them

	Efficacy (only relative effects) alone		Safety (only relative effects) alone		Confidence in the evidence for efficacy alone		Clinical experience for efficacy alone		Clinical experience for safety alone		Safety (relative effects and clinical experience for safety) alone		Treatment cost alone	
Drug	**P-score**	**rank**	**P-score**	**rank**	$l_i^{(0)}$	**rank**	$l_i^{(0)}$	**rank**	$l_i^{(0)}$	**rank**	$l_i^{(0)}$[a]	**rank**	$l_i^{(0)}$	**rank**
ACI	09.9	19	46.9	9	0.002	18	0	9	0	10	0.045	**7**	**0.162**	**1**
ADA	63.1	8	40.4	14	0.075	6	**0.100**	**3**	**0.050**	**3**	0.042	10	0.05	8
ALEFACEPT	25.3	15	**62.6**	**5**	0.017	17	0	9	0	10	0.045	7	0	10
APRE	39.7	13	54.7	7	0.029	14	0	9	0	10	0.038	13	**0.106**	**5**
BRODA	**84.3**	**3**	39.8	15	**0.096**	**4**	**0.050**	**4**	**0.050**	**3**	0.041	12	0	10
CERTO	**75.7**	**5**	**70.9**	**3**	0.062	7	**0.050**	**4**	**0.050**	**3**	**0.072**	**5**	0	10
CICLO	21.3	17	**78.2**	**2**	0.019	16	0	9	0	10	**0.075**	**3**	**0.162**	**1**
ETA	52.6	11	43.6	11	0.062	7	**0.050**	**4**	**0.150**	**2**	**0.073**	**4**	0.073	6
FUM	21.9	16	57.7	6	0.015	17	0	9	0	10	0.042	10	**0.162**	**1**
GUSEL	**77.0**	**4**	42.6	12	**0.104**	**2**	**0.150**	**2**	**0.050**	**3**	0.044	9	0	10
IFX	53.2	10	**64.4**	**4**	0.051	9	**0.050**	**4**	**0.050**	**3**	0.059	6	0.073	6
ITO	56.0	9	-	-	0.029	12	0	9	-	-	0	20	0	10
IXE	**94.3**	**1**	33.7	17	**0.117**	**1**	**0.050**	**4**	**0.050**	**3**	0.037	14	0	10
MTX	20.2	18	**90.7**	**1**	0.029	12	**0.100**	**3**	**0.150**	**2**	**0.112**	**2**	**0.162**	**1**
PBO	02.9	20	42.0	13	0.002	18	0	9	0	10	0.027	18	0	10
PONE	37.3	14	18.1	19	0.028	14	0	9	0	10	0.014	19	0	10
SECU	**86.5**	**2**	29.9	18	**0.100**	**3**	**0.050**	**4**	**0.050**	**3**	0.034	15	0	10
TILDRA	63.6	7	37.8	16	0.035	11	0	9	0	10	0.033	16	0	10
TOFA	42.5	12	44.0	10	0.046	10	0	9	0	10	0.030	17	0	10
USK	72.6	6	52.0	8	**0.082**	**5**	**0.350**	**1**	**0.350**	**1**	**0.136**	**1**	0.05	8

Note: P-score values for efficacy and safety and initial probability distributions of treatment selection based on confidence in the evidence, clinical experience, treatment cost, and safety. Gray cells correspond to the five top positions in the respective ranking.

[a]These values are the POST-R probabilities combining relative effects and clinical experience for safety.

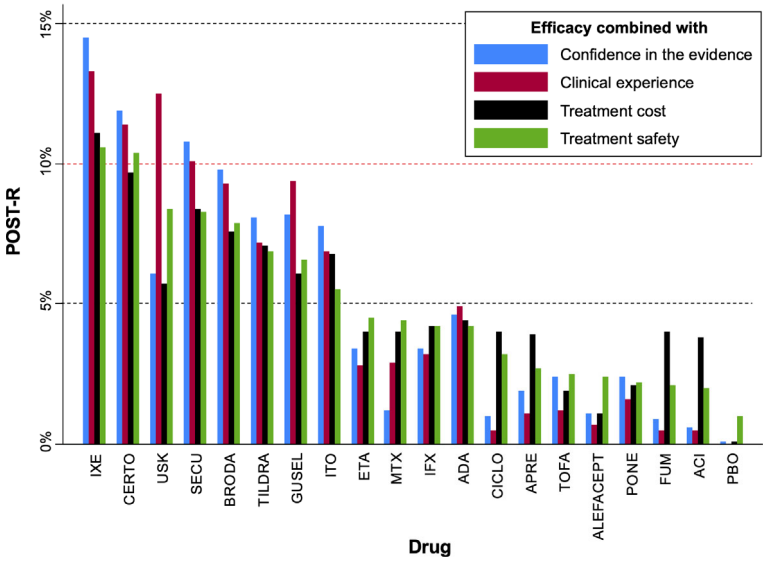


FIGURE 4 Ranking results of the psoriasis network using the POST-R measure and considering the relative effects for efficacy in the transition probabilities and different characteristics in the initial probability distribution as indicated by the legend. Drugs have been ordered according to the ranking for efficacy and safety (green bars) [Color figure can be viewed at wileyonlinelibrary.com]

the evaluation of imprecision, heterogeneity and incoherence (more details available in CINeMA's documentation page http://cinema.ispm.ch/#doc) as well as for the definition of probabilities c_{i-j} (see Section 3.3.1). The matrix C is available in the Appendix (Appendix Table 4), while the probabilities $l_i^{(0)}$ (Table 1) were obtained using Equation (6). Interestingly, ixekizumab that appears to be the most efficacious treatment for PASI 90 is also at the top of the ranking with respect to the confidence of the evidence followed by guselkumab (Table 1). We considered that $z = 0.75$ is a good compromise between the two sources of information (ie, relative effects for efficacy and confidence of the evidence for efficacy). It seems that the incorporation of the confidence in the evidence resulted in small changes in the overall picture of treatment ranking for efficacy (Figure 4). That was expected given that treatments that appeared to perform better with respect to efficacy were also at the top ranks in terms of the confidence in the evidence (Table 1).

4.3 | Treatment ranking combining efficacy and clinical experience for efficacy

We elicited the clinical experience data on efficacy in terms of treatment selection probabilities using the MATCH tool as Appendix Figure 1(a) shows. Each chip represented five hypothetical patients from the population of the original Cochrane review. For illustration purposes in the present study, one dermatologist was asked to allocate 100 patients to the 20 possible treatment options (including placebo) considering how participants are on average allocated to the different drugs in clinical practice. In the presence of prescription data (eg, from national registries), these could also be used to form some kind of clinical experience data. We used again $z = 0.75$ considering clinical experience for this outcome needs only be a Supplementary source of information. Figure 4 implies that the incorporation of clinical experience materially affects the ranks of the treatment. In particular, ustekinumab is favored and appears at the second position now with more than 12% probability of being selected among all treatments.

4.4 | Treatment ranking combining efficacy and safety informed by clinical experience

Here we obtained the final r vector in two steps:

- First, we combined the clinical experience probabilities for safety with the relative effects for safety to obtain the safety ranking using the POST-R approach. The data for SAE that gave the relative effects for safety are given in the Supplementary Material and the results of the roulette method for eliciting the clinical experience are provided in Appendix Figure 2. Considering the concerns expressed previously about the SAE data (see Section 2), we used $z = 0.7$ that reduced slightly more the influence of the relative effects in treatment ranking for safety. The resulted probability vector of this step (r_{safety}), which is shown in Table 1, was then used at the second step as the initial probability distribution

$\mathbf{Q^{(0)}}$. We can see that ranking for safety obtained from the relative effects (alone) and the clinical experience (alone) are somewhat different. This might reflect to some extent the limitation of the data due to insufficient follow-up of the studies and the presence of rare events.

- At the second step, we combined the $\mathbf{Q^{(0)}}$ as defined above with the relative effects for efficacy to rank treatments based on both outcomes using $z = 0.65$.

Again, ustekinumab seems to be the most favored treatment in comparison with the conventional ranking results based on SUCRAs. However, ixekizumab and certolizumab remained at the top of the hierarchy.

4.5 | Treatment ranking combining efficacy and cost

In the psoriasis example, the 19 different systemic treatments are very diverse in terms of their cost; therefore we considered treatment cost as an important characteristic to be incorporated in treatment ranking. Specifically, the cost of each treatment per person per year in France was used to form the $\mathbf{Q^{(0)}}$ vector following the approach in Section 3.3.3. Note that treatment cost might be different across countries and therefore our findings might not be directly applicable outside France. Appendix Table 5 shows the average cost of each drug. Secukinumab, ixekizumab, and brodalumab are the most expensive treatments, while some of the drugs have not yet been commercialized; for such drugs, we considered that they have zero probability of being selected considering their cost only. We used a probability $z = 0.8$ implying that treatment cost should not be considered as a key component of treatment ranking but only for selection between drugs with overall similar performance on other important factors (such as the outcomes of interest). Figure 4 shows that, after incorporating in ranking the cost information, the highly efficacious drugs remained at the top positions.

Overall, according to Figure 4, a group of eight drugs achieved more 5% probability of being selected under all four considerations and one drug more than 10% (ixekizumab). Ustekinumab seems to be the drug with most unclear performance as it was favoured when clinical experience was considered but not so much by the actual data. This possibly means that the benefits of this drug need further investigation. A sensitivity analysis on a range of values for z could also be very informative. For example, Appendix Figure 3 shows how sensitive the POST-R scores are for every drug on a range of values for the probability z. For some drugs, such as ixekizumab and secukinumab, their score does not seem to be affected a lot by changes in z, while for other drugs (eg, certolizumab and guselkumab) as z changes their POST-R as well as their ranking change materially. This means that drugs with lower confidence of the evidence are "penalized" as we move away from $z = 1$, whereas those with higher confidence are favored.

5 | DISCUSSION

Treatment ranking is a key and potentially very informative output of NMA but inappropriate use and misinterpretation of ranking, which are regularly encountered in published NMAs, have made several researchers being skeptical about its usefulness.[1,5,6] A key point prior to drawing conclusions from ranking results is understanding what the estimated order of the treatments may represent. Currently available approaches provide rankings based either only on the mean of the relative effects (eg, ranked forest plots) or on the mean and a part of the distribution of the relative effects (eg, the probability of being best/worst) or at best on the mean and the full distribution of the relative effects (ie, SUCRAs/P-scores). As a result, these methods produce rankings that are as reliable as the estimated means and/or variances of the relative effects.

In this article, we move forward from ranking measures that just provide a summary of the relative effects and introduce a new concept of treatment ranking. A drawback of our approach is that the definition of the vector $\mathbf{Q^{(0)}}$ and the probability z is subjective to some degree. It is important that one describes in detail in the protocol of the analysis how (s)he will define these quantities. We can also mitigate this subjectivity by incorporating views from many clinicians with experience in the condition under investigation. Eliciting expert opinion about z is not covered in this article as our aim is to formulate the method. Certainly, developing approaches for informing the model parameters through expert opinion could help substantially improving and popularizing the method. This is something we intend to work upon in the near future. In addition, although this article focuses on a frequentist framework, the method can also be applied in Bayesian

framework where $\mathbf{Q^{(0)}}$ and z can be modeled through distributions allowing for uncertainty in their values. It should be noted that both $\mathbf{Q^{(0)}}$ and z are dependent on the characteristics being considered, the medical field and the research question and therefore the criteria for their definition are specific to every NMA. Although such an approach seems to require additional time and resources than current practices in NMA, recent directions in evidence synthesis that encourage living systematic reviews and living network meta-analyses within the context of a research community are well compatible to our method.[27–31]

An important limitation of our application is that probabilities based on clinical experience were elicited after the publication of the original NMA. Hence, the clinician who provided expertise could have been influenced by the published NMA results. It should be noted that we used a possibly extreme value of the minimally clinical important difference in CINeMA (ie, risk ratio of 2) which might not apply in clinical practice. Moreover, in the present article we surveyed only one clinician for simplicity; articles aiming to address clinical questions should opt for several experts to inform their POST-R model. In addition, there might be other important characteristics affecting treatment selection that were not considered here. It is important to note that our results are only illustrative of the methods and do not aim to draw clinical inferences; the latter would require careful consideration of the definition for $\mathbf{Q^{(0)}}$ and z at the beginning of the study. To avoid data-driven decisions and selective reporting of results when using our ranking approach, it is important that a clear and transparent description of the criteria to be used for the definition of $\mathbf{Q^{(0)}}$ and z are available in the protocol.

Our approach offers a more intuitive way of thinking treatment ranking as it accommodates the different considerations being made before selecting a treatment in practice. It should be acknowledged that this is an artificial MC process and that the time points t do not have a real meaning. However, we conceptualized the "ideal" scenario (in terms of exploring efficacy of interventions) of someone going through the whole evidence in order to decide upon a final hierarchy of all the treatment options at hand and we assumed that changing treatment is an evidence-based choice driven by the results of the NMA as well as by some prior/external information. In the presence of important aggregate covariate(s) that are substantially different across studies, the probabilities $p_{i>j}$ can be obtained by a network meta-regression model adjusting for these covariate(s). The POST-R approach has been implemented in the *sucra* command of the *network graphs* package[26] in Stata[32] which facilitates its use once $\mathbf{Q^{(0)}}$ has been specified. Further implementation in R is one of our next steps. In the present paper we only provide methods for combining efficacy with one of the other characteristics (confidence, clinical experience, etc.). Additional work is necessary to extend the method into a generalized ranking framework that will allow multiple characteristics to be combined simultaneously with the relative effects.

A key issue, very often disregarded in the literature, when obtaining the ranking of treatments is to acknowledge that it is always accompanied with a level of uncertainty.[2,4,11] The POST-R approach naturally captures the uncertainty in treatment selection by allowing the end-user of NMA to "move" between the different treatment options (or between groups of treatments) with some probability. It should be noted that there is no unique best way to rank treatments in NMA. All approaches have limitations and therefore NMA conclusions should never be driven only by ranking results. Of course, using POST-R to draw conclusions (as for any other ranking approach) assumes that the NMA results are valid. Hence, if the model assumptions (eg, homogeneity and coherence) are violated ranking of treatments would give misleading findings. A sensitivity analysis using different approaches to ranking might be useful to assess the robustness of the findings. Also, empirical evaluation of the performance of POST-R along with other ranking approaches using simulations and many published NMAs under different network structures with various properties and different clinical settings is one of our main future goals.

Generally, despite the various and well-founded criticisms, estimation of ranking remains one of the main objectives for several NMAs implying that clinicians and decision-makers are interested in it. Thus, it is necessary to produce NMA summaries that, instead of just acknowledging any data limitations and/or availability of external evidence, they formally integrate these considerations with the results in the league table to make valid recommendations. Such an effort might be less important in dense and coherent networks with many high-quality studies for every treatment comparison. Our method may target primarily NMAs stating "more well-conducted studies are necessary" with an aim to assist such reviews drawing more useful conclusions until new studies come up.

DATA AVAILABILITY STATEMENT

The dataset and the Stata script used to obtain the results can be found in the Supplementary files.

ORCID

Anna Chaimani https://orcid.org/0000-0003-2828-6832
Dimitris Mavridis https://orcid.org/0000-0003-1041-4592

REFERENCES

1. 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;6:451-460.
2. 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;164(10):666-673.
3. Mbuagbaw L, Rochwerg B, Jaeschke R, et al. Approaches to interpreting and choosing the best treatments in network meta-analyses. *Syst Rev*. 2017;6:79.
4. Veroniki AA, Straus SE, Rücker G, Tricco AC. Commentary: is providing uncertainty intervals in treatment ranking helpful in a network meta-analysis? *J Clin Epidemiol*. 2018;100:122-129.
5. Chaimani A, Salanti G, Leucht S, Geddes JR, Cipriani A. Common pitfalls and mistakes in the set-up, analysis and interpretation of results in network meta-analysis: what clinicians should look for in a published article. *Evid Based Ment Health*. 2017;20(3):88-94.
6. Petropoulou M, Nikolakopoulou A, Veroniki A-A, et al. Bibliographic study showed improving statistical methodology of network meta-analyses published between 1999 and 2015. *J Clin Epidemiol*. 2017;82:20-28.
7. Chaimani A, Caldwell DM, Li T, Higgins JPT, Salanti G. Additional considerations are required when preparing a protocol for a systematic review with multiple interventions. *J Clin Epidemiol*. 2017;83:65-74.
8. Salanti G, Ades AE, Ioannidis JP. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial. *J Clin Epidemiol*. 2011;64:163-171.
9. Rücker G, Schwarzer G. Ranking treatments in frequentist network meta-analysis works without resampling methods. *BMC Med Res Methodol*. 2015;15:58.
10. Chaimani A, Salanti G. Investigating the impact of treatments with scarce evidence in network meta-analysis. Paper presented at: 36th ISCB Annual Conference; Utrecht; 2015.
11. Salanti G, Del Giovane C, Chaimani A, Caldwell DM, Higgins JPT. Evaluating the quality of evidence from a network meta-analysis. *PLoS One*. 2014;9(7):e99682.
12. Chaimani A, Higgins JPT, Mavridis D, Spyridonos P, Salanti G. Graphical tools for network meta-analysis in STATA. *PLoS One*. 2013;8(10):e76654.
13. Brin S, Page L. The anatomy of a large-scale hypertextual web search engine. *Comput Netw*. 1998;30:107-117.
14. Blanchet J, Gallego G, Goyal V. A Markov chain approximation to choice modeling. *Oper Res*. 2016;64(4):886-905.
15. Sbidian E, Chaimani A, Garcia-Doval I, et al. Systemic pharmacological treatments for chronic plaque psoriasis: a network meta-analysis. *Cochrane Database Syst Rev*. 2017;22(12):CD011535.
16. Ibe OC. Chap. 4—Discrete-time Markov chains. In: Ibe OC, ed. *Markov Process Stoch Model*. Second ed. Oxford: Elsevier; 2013:59-84.
17. Grinstead CM, Snell JL. *Introduction to Probability*. Minneapolis, MN: American Mathematical Society; 1997. https://open.umn.edu/opentextbooks/textbooks/21.
18. Langville AN, Meyer CD. Deeper inside PageRank. *Internet Math*. 2004;1(3):335-380.
19. Hutton B, Salanti G, Caldwell DM, et al. The PRISMA extension statement for reporting of systematic reviews incorporating network meta-analyses of healthcare interventions: checklist and explanations. *Ann Intern Med*. 2015;162(11):777-784.
20. Puhan MA, Schünemann HJ, Murad MH, et al. A GRADE Working Group approach for rating the quality of treatment effect estimates from network meta-analysis. *BMJ*. 2014;349:g5630.
21. CINeMA: Confidence in Network Meta-Analysis [Software]. Institute of Social and Preventive Medicine, University of Bern; 2017. http://cinema.ispm.ch
22. Nikolakopoulou A, Higgins JPT, Papakonstantinou T, et al. CINeMA: An approach for assessing confidence in the results of a network meta-analysis. *PLOS Medicine*. 2020;17(4):e1003082. http://dx.doi.org/10.1371/journal.pmed.1003082.
23. Garthwaite PH, Kadane JB, O'Hagan A. Statistical methods for eliciting probability distributions. *J Am Stat Assoc*. 2005;100(470):680-700.
24. Morris DE, Oakley JE, Crowe JA. A web-based tool for eliciting probability distributions from experts. *Environ Model Software*. 2014;52:1-4.
25. White IR. Network meta-analysis. *Stata J*. 2015;15(4):951-985.
26. Chaimani A, Salanti G. Visualizing assumptions and results in network meta-analysis: the network graphs package. *Stata J*. 2015;15(4):905-950.
27. Créquit P, Trinquart L, Yavchitz A, Ravaud P. Wasted research when systematic reviews fail to provide a complete and up-to-date evidence synthesis: the example of lung cancer. *BMC Med*. 2016;14:8.
28. Créquit P, Trinquart L, Ravaud P. Live cumulative network meta-analysis: protocol for second-line treatments in advanced non-small-cell lung cancer with wild-type or unknown status for epidermal growth factor receptor. *BMJ Open*. 2016;6(8):e011841.
29. Nikolakopoulou A, Mavridis D, Furukawa TA, et al. Living network meta-analysis compared with pairwise meta-analysis in comparative effectiveness research: empirical study. *BMJ*. 2018;360:k585.
30. Nikolakopoulou A, Mavridis D, Egger M, Salanti G. Continuously updated network meta-analysis and statistical monitoring for timely decision-making. *Stat Methods Med Res*. 2018;27(5):1312-1330.

31. Elliott JH, Synnot A, Turner T, et al. Living systematic review: 1. Introduction—the why, what, when, and how. *J Clin Epidemiol*. 2017;91:23-30.
32. StataCorp. Stata Statistical Software. College Station, TX: StataCorp LP; 2013.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Chaimani A, Porcher R, Sbidian É, Mavridis D. A Markov chain approach for ranking treatments in network meta-analysis. *Statistics in Medicine*. 2020;1–14. https://doi.org/10.1002/sim.8784