

This is a repository copy of *Decision curve analysis for personalized treatment choice between multiple options*.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/id/eprint/197265/>

Version: Published Version

Article:

Chalkou, Konstantina, Vickers, Andrew J., Pellegrini, Fabio et al. (2 more authors) (2022) Decision curve analysis for personalized treatment choice between multiple options. Medical Decision Making. ISSN: 1552-681X

<https://doi.org/10.1177/0272989X221143058>

Reuse

This article is distributed under the terms of the Creative Commons Attribution-NonCommercial (CC BY-NC) licence. This licence allows you to remix, tweak, and build upon this work non-commercially, and any new works must also acknowledge the authors and be non-commercial. You don't have to license any derivative works on the same terms. More information and the full terms of the licence here: <https://creativecommons.org/licenses/>

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.



Decision Curve Analysis for Personalized Treatment Choice between Multiple Options

Konstantina Chalkou, MSc^{}, Andrew J. Vickers, PhD^{}, Fabio Pellegrini, PhD, Andrea Manca, PhD, and Georgia Salanti, PhD

Background. Decision curve analysis can be used to determine whether a personalized model for treatment benefit would lead to better clinical decisions. Decision curve analysis methods have been described to estimate treatment benefit using data from a single randomized controlled trial. **Objectives.** Our main objective is to extend the decision curve analysis methodology to the scenario in which several treatment options exist and evidence about their effects comes from a set of trials, synthesized using network meta-analysis (NMA). **Methods.** We describe the steps needed to estimate the net benefit of a prediction model using evidence from studies synthesized in an NMA. We show how to compare personalized versus one-size-fit-all treatment decision-making strategies, such as “treat none” or “treat all patients with a specific treatment” strategies. First, threshold values for each included treatment need to be defined (i.e., the minimum risk difference compared with control that renders a treatment worth taking). The net benefit per strategy can then be plotted for a plausible range of threshold values to reveal the most clinically useful strategy. We applied our methodology to an NMA prediction model for relapsing-remitting multiple sclerosis, which can be used to choose between natalizumab, dimethyl fumarate, glatiramer acetate, and placebo. **Results.** We illustrated the extended decision curve analysis methodology using several threshold value combinations for each available treatment. For the examined threshold values, the “treat patients according to the prediction model” strategy performs either better than or close to the one-size-fit-all treatment strategies. However, even small differences may be important in clinical decision making. As the advantage of the personalized model was not consistent across all thresholds, improving the existing model (by including, for example, predictors that will increase discrimination) is needed before advocating its clinical usefulness. **Conclusions.** This novel extension of decision curve analysis can be applied to NMA-based prediction models to evaluate their use to aid treatment decision making.

Highlights

- Decision curve analysis is extended into a (network) meta-analysis framework.
- Personalized models predicting treatment benefit are evaluated when several treatment options are available and evidence about their effects comes from a set of trials.
- Detailed steps to compare personalized versus one-size-fit-all treatment decision-making strategies are outlined.
- This extension of decision curve analysis can be applied to (network) meta-analysis-based prediction models to evaluate their use to aid treatment decision making.

Corresponding Author

Konstantina Chalkou, MSc, Institute of Social & Preventive Medicine, University of Bern, Mittelstrasse 43, Bern, 3012, Switzerland.

Email: konstantina.chalkou@ispm.unibe.ch

Keywords

decision curve analysis, network meta-analysis, prediction model, net benefit, clinical usefulness

Date received: February 4, 2022; accepted: November 3, 2022

Randomized controlled trials (RCTs) and their meta-analyses have traditionally focused on inferences about treatment effects for the average patient.¹ Yet, what clinicians want to know is the treatment effect for the patient in front of them, and the effects of treatment may differ between individuals. To identify the best treatment option for an individual, researchers can use prediction models to evaluate the treatment effects on health outcomes as a function of patient-level characteristics.^{2–5}

Personalized prediction models could be used to identify groups of patients for which the benefits of treatment outweigh the harms. Doing so would require extensive validation, and such validation should include an evaluation of clinical utility. The latter refers to the ability of the model to guide treatment decisions at the point of care. While methods to evaluate a model's performance have been well studied and are described in the literature (e.g., calibration measures, the area under the curve, etc.),^{6,7} evaluation of the clinical utility of a model is a relatively new concept.

Decision curve analysis (DCA) has been proposed to evaluate the clinical utility of personalized prediction models.^{8,9} DCA can be applied to models that predict an absolute risk (such as a model to predict the risk of

cancer to guide decisions about biopsy) and those predicting treatment benefit (such as a model to predict the change in outcome associated with drug therapy).^{8,10,11} The data used to calculate the net benefit (NB) for a treatment strategy typically come from an RCT that compares 2 treatments: a reference treatment (such as no treatment or placebo) and an active treatment of interest. Papers, software, tutorials, and data sets on DCA methodology can be found at www.decisioncurveanalysis.org.

There are often several treatment options for a given condition. Unfortunately, there is often uncertainty about their relative benefits due to a lack of a direct head-to-head comparison in a single RCT. Evidence synthesis in the form of pairwise meta-analysis (PMA) and its extension, network meta-analysis (NMA), can be used both to structure the evidence base (summarizing direct and indirect comparisons) and to produce an estimate of the effects of any treatment against other available options. It has been found that prediction models based on (network) meta-regression of multiple individual patient data (IPD) can be used to identify the best treatment option for an individual patient.^{10,12–14}

Consider a patient diagnosed with relapsing-remitting multiple sclerosis (RRMS) who is contemplating starting a disease-modifying drug. The individual and her or his clinician may have access to the results of an NMA of aggregated data to inform their decision, but this evidence gives insight into only the expected health outcomes and the efficacy of the treatments being considered for the “average patient” in the model.^{15–17} Personalized treatment recommendations can be obtained if patient characteristics are taken into account when predicting the outcome under different treatment options. This can be achieved using network meta-regression with IPD data,¹² with the model indicating the optimal drug (in the case of RRMS treatment decision, this may be the one that minimizes the predicted risk to relapse over the time horizon of 2 y) for any given patient profile. Extending this idea to several outcomes and accounting for the tradeoff between safety and efficacy will result in a hierarchy of treatment options that is tailored to a participant's characteristics.^{18,19}

In this article, we extend the DCA methodology, as proposed by Vickers et al.,⁹ to evaluate the clinical

Institute of Social and Preventive Medicine, University of Bern, Bern, Switzerland (KC, GS); Graduate School for Health Sciences, University of Bern, Switzerland (KS); Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, NY, USA (AJV); BDH, Biogen Spain, Madrid, Spain (FP); Centre for Health Economics, University of York, York, UK (AM). The authors declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: KC, AJV, AM, and GS declare that they have no conflict of interest with respect to this article. FP is an employee of and holds stocks/stock options in Biogen. The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: KC, AM, and GS are funded by the European Union's Horizon 2020 research and innovation program under grant agreement No. 825162. The HTx project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 825162. This dissemination reflects only the author's view, and the Commission is not responsible for any use that may be made of the information it contains. AJV is funded in part by a Cancer Center Support Grant from the National Cancer Institute made to Memorial Sloan Kettering Cancer Center (P30-CA008748) and a SPORE grant from the National Cancer Institute to Dr. H. Scher (P50-CA092629).

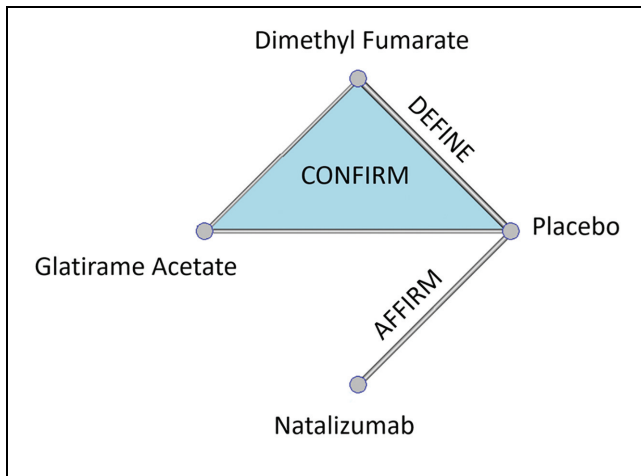


Figure 1 Net-graph: treatments compared in each one of the available randomized controlled trials: AFFIRM, DEFINE, and CONFIRM.^{26–28}

usefulness of a personalized prediction model that aims at recommending a treatment among many possible options according to individual characteristics, such as the one described above. The focus of the article is methodological, and we use an example with RRMS only to outline the developed methodology. This work is supported and funded by the HTx. The HTx is a Horizon 2020 project supported by the European Union lasting for 5 y from January 2019. The main aim of HTx is to create a framework for the next-generation health technology assessment to support patient-centered, societally oriented, real-time decision making on access to and reimbursement for health technologies throughout Europe. We describe the network meta-regression prediction model in the next section. In the “Methods” section, we describe ways to select the threshold values for each treatment option and how treatment recommendations can be made based on the results of a network meta-regression prediction model, and we describe the estimation of quantities in DCA methodology from PMA and NMA data sets. We show the results from the case study in the “Results” section, and we conclude with a discussion of the advantages and limitations of the proposed approach.

Case Study: Personalized Treatment Recommendation for Patients with RRMS

Multiple sclerosis (MS) is an immune-mediated disease of the central nervous system with several subtypes. The most common subtype is RRMS.²⁰ Patients with RRMS

present with acute or subacute symptoms (relapses) followed by periods of complete or incomplete recovery (remissions).²¹ Effective treatment of patients with RRMS can prevent disease progression and associated severe consequences, such as spasticity, fatigue, cognitive dysfunction, depression, bladder dysfunction, bowel dysfunction, sexual dysfunction, pain, and death.²² There are several available treatment options for RRMS, and their efficacy and safety profiles vary. For instance, natalizumab is more effective (on average) than dimethyl fumarate but associated with important side effects and increased risk of progressive multifocal leukoencephalopathy, which can cause death.^{23,24}

Recently, a 2-stage model was presented to predict the personalized probability of relapse within 2 y in patients diagnosed with RRMS.¹² Three phase III RCTs were used: AFFIRM, DEFINE, and CONFIRM.^{25–27} Patients were randomized into 3 active drugs (natalizumab, glatiramer acetate, dimethyl fumarate) and placebo, as shown in Figure 1.

In a first stage, the baseline risk score for relapse was developed, which is a score that summarizes the patient-level characteristics and indicates the severity of the baseline health condition. In a second stage, the baseline risk score was used as the only effect modifier, which has an impact on relative treatment effects, in a network meta-regression model to predict the risk to relapse within the next 2 y under the 3 drugs or placebo. The results are presented in Figure 2 as well as in an interactive R-Shiny application available at <https://cinema.ispm.unibe.ch/shinies/koms/>. A detailed description of the development of the RRMS personalized prediction model, which we use as an example here, has been previously given.¹²

Such models can be used to guide clinical decisions, assuming heuristically that relapse is the only health outcome of interest. For example, this prediction model would recommend dimethyl fumarate to patients whose baseline risk is lower than 25% and natalizumab to patients whose baseline risk is higher than 25%. However, even when a patient has baseline risk score equal to 30%, where natalizumab minimizes the predicted risk to relapse, the absolute predicted difference in relapse probability is only 5% compared with dimethyl fumarate. In addition, natalizumab is a drug with more serious side effects compared with dimethyl fumarate; hence, the doctor in discussion with the patient might decide to administer dimethyl fumarate.

We want to evaluate whether this personalized prediction model could guide the decision-making process. We will compare the treatment decisions that this model entails (“treat patients according to the prediction model”) to those from “one-size-fit-all” strategies: “treat

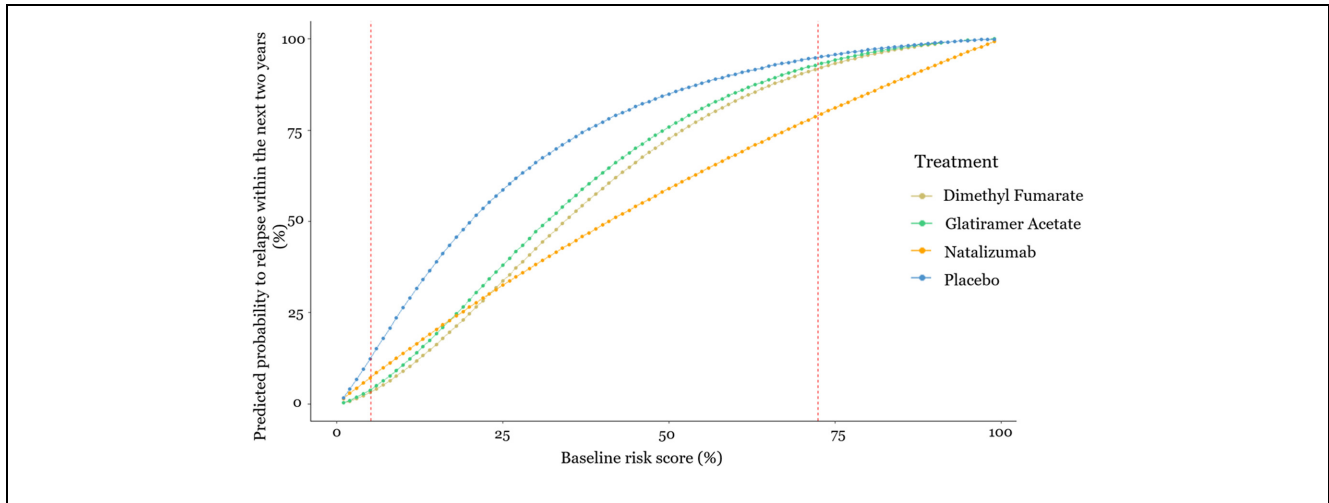


Figure 2 Estimated probability to relapse within the next 2 y as a function of the baseline risk score. The x -axis shows the baseline risk score of relapsing within the next 2 y, and the y -axis shows the estimated probability of relapsing within the next 2 y under each of the treatments. Between the 2 dashed vertical lines are the baseline risk values observed in the data used.

none,” “treat all patients with natalizumab,” “treat all patients with dimethyl fumarate,” and “treat all patients with glatiramer acetate.”

Methods

In the next section, we describe how treatment recommendations via a prediction model are reached when we have multiple treatment options. In the subsequent section, we introduce the proposed extension of the DCA methodology when considering several competing treatment strategies. In the “Comparing Different Treatment Strategies via DCA” section, we describe the implementation and software used to evaluate the model on predicting the optimal treatment to prevent relapsing within the next 2 y in RRMS.

Threshold Values

Let us consider that there are several treatment options available for a health condition. Each available treatment option is denoted with j , where $j = 1, 2, \dots, J$. Each treatment is associated with different side effects, cost, and inconvenience. For a dichotomous outcome, the threshold value T_j for treatment j is defined as the minimum risk difference compared with control treatment that renders treatment j worth taking. T_j sums into a single value—the harms, costs, and inconvenience of treatment j —and expresses how much benefit would be expected to outweigh the harm that treatment j might cause. At a population level, setting $T_j = 20\%$ means

that we would be willing to treat up to 5 patients with j to prevent 1 patient relapsing; 4 patients will be unnecessarily taking the drug (and hence subjected to its toxicity) and are traded against 1 patient with prevented relapse. In the RRMS example, it would be reasonable to set a lower threshold for dimethyl fumarate and glatiramer acetate compared with natalizumab ($T_{DF} = T_{GA} = 10\%$ and $T_N = 20\%$) because of their different side effect profiles.

Note that specifying T_j is not a novel feature of our proposed methodology but rather a routine and necessary aspect of traditional clinical trial methodology. It is required both to determine sample size and to evaluate the clinical relevance of the findings: if the difference in event rates between treatment j and control is statistical significant but less than T_j , we infer that while better than control, j should not be used in practice. As different patients might weight differently the risk of an event and risks associated with each treatment, a clinically relevant range of threshold values for all treatment options may be indicated.²⁸ In the RRMS example, we used a range of threshold values based on discussions with 2 experienced MS neurologists (see the acknowledgments) on the drugs’ side effects and toxicity to illustrate how the suggested methodology could be applied.

Reaching Treatment Recommendations when We Have Multiple Options via a Model

Let us consider a personalized prediction model for the probability of an event, $R_{i,j}$, for each patient i , where

Table 1 Reaching the Recommended Treatment, via a Prognostic Model, between 4 Options: Placebo, Glatiramer Acetate, Dimethyl Fumarate, and Natalizumab (Hypothetical Example in Relapsing-Remitting Multiple Sclerosis)

Treatment	Placebo	Glatiramer Acetate	Dimethyl Fumarate	Natalizumab
Predicted risk to relapse within 2 y ($R_{i,j}$)	75%	66%	52%	44%
Predicted risk difference versus placebo ($RD_{i,j}$)	-	9%	23%	31%
Threshold value for treatment j (T_j)		10%	10%	20%
$RD_{i,j} - T_j$		-1%	12%	11%
Recommended treatment via the prediction model		Dimethyl fumarate		

The bold font indicates the maximum difference between the risk difference of treatment j , $RD_{i,j}$, and its threshold value T_j , based on which the optimal treatment via the prediction model is recommended.

$i = 1, 2, \dots, N$, under each available treatment option j , where $j = 1, 2, \dots, J$. Then, the risk difference, $RD_{i,j}$, for patient i between treatment j and the control treatment (or placebo) is the difference between the patient's predicted probabilities under these 2 options: $RD_{i,j} = R_{i,control} - R_{i,j}$. Whether patient i will be prescribed treatment j depends on several factors. First, treatment j needs to be effective, $RD_{i,j} > 0$; that is, it must decrease the predicted probability of a harmful outcome compared with the control treatment. Second, the benefits of treatment need to outweigh its harms. For example, natalizumab is a drug with important side effects and is associated with increased mortality.^{23,24} Now, imagine an RRMS patient, whose predicted risk to relapse within 2 y is decreased by $RD_{i,N} = 3\%$ under natalizumab compared with placebo. It is possible that, given the side effects of treatment, this patient will not choose natalizumab for such a small reduction in the predicted probability of relapse.

We define the threshold value T_j as the minimum risk difference compared with control that renders treatment j worth taking. T_j depends on the risks, harms, costs, and inconvenience of treatment j . For a patient i , the recommended treatment j under the prediction model is the one that satisfies $\max\{RD_{i,j} - T_j\}$, between those treatments with $RD_{i,j} \geq T_j$. When all active treatments lead to $RD_{i,j} < T_j$, then the control treatment is recommended for patient i . In Table 1, we present a fictional example showing how treatment recommendation is made via a prognostic model, with assumed threshold values $T_{DF} = T_{GA} = 10\%$ for dimethyl fumarate and $T_N = 20\%$ for natalizumab.

While the model makes personalized predictions under each treatment j , the threshold values T_j are not based on individual preferences.²⁸ To evaluate the clinical usefulness of the model, we first need to understand the typical range of preferences of patients, with respect to the

possible tradeoff between the harms and benefits of each treatment. Then, these preferences will determine the range of thresholds over which the clinical utility of the model comparing the various competing strategies should be assessed.²⁸

Comparing Different Treatment Strategies via DCA

In the case of medical treatments, there are several decision strategies that can be evaluated and compared. Consider a treatment strategy s that refers to the choice between $j = 0, 1, \dots, J$ treatments, with 0 denoting the control. That strategy recommends a treatment for each patient and can be “treat all with drug j ” ($s = j$), “treat none” ($s = 0$), or a more nuanced strategy suggested by a prediction model. A strategy associated with a prediction model was discussed in detail in the previous section; now assume that the recommended treatment for a patient is well defined in each of the s competing strategies. Control could be any treatment or combination of treatments used as reference (e.g., standards of care, placebo, or no treatment at all). From now on, we will assume placebo as control and strategy $s = 0$ as the “treat none” strategy.

The measure of performance of each strategy is the NB. The NB is the benefit that a decision entails minus the relevant harms weighted by a tradeoff preference value. In the case of medical treatments, benefit could be measured as the reduction in a harmful health outcome (e.g., relapses) with the treatment. Harms include all disbenefits of treatment, including side effects, risks, financial cost, and inconvenience. Vickers et al. described in detail the DCA methodology and defined the net treatment benefit for a single treatment.⁹ The NB estimation involves counterfactuals, the unobserved outcome if a particular strategy is employed. Consequently, the estimation of NB for a model predicting treatment benefit is best estimated using RCT data.⁹

We generalize the idea to the NB of a strategy s , NB_s , for 2 or more treatment options, and we show how to estimate it in a PMA and NMA of RCTs. We define NB_s as the benefit (decrease in event rate using strategy s) minus the treatment rates multiplied by a set of treatment-specific threshold values T_j . The threshold values T_j are measured on a risk scale (from 0 to 1), which identifies which reduction in risk will justify the use of each treatment. Notice that the value of T_j may vary from patient to patient depending on personal preferences and other medical considerations (such as comorbidities). The strategy s with the highest NB, for specific threshold values T_j , is chosen as leading to better clinical decisions.⁹

More specifically, we define the NB of each strategy s compared with strategy $s = 0$ (“treat none”) as

$$NB_s = \varepsilon_0 - \varepsilon_s - \sum_j \pi_{s,j} \times T_j,$$

where ε_0 denotes the event rate under no treatment, ε_s the event rate under strategy s , and $\pi_{s,j}$ the proportion of patients treated with treatment j under strategy s .

Estimation of ε_0 . When data from 1 RCT with placebo are available, ε_0 is directly quantifiable from the data as the observed proportion of participants with an event in the placebo arm $\hat{\varepsilon}_0 = e_0^{Data}$, where $Data$ is the data set of all available RCTs.⁹ However, when we have several RCTs instead of one, the estimation needs to account for the fact that patients are randomized within trials but not across them. Hence, when estimating event rates, we cannot simply pool treatment arms together or results will be biased (Simpson’s paradox).^{29,30} In this case, we first need to perform a meta-analysis of all placebo event rates in $Data$ across trials to obtain an estimate of the pooled event rate in the placebo $\hat{\varepsilon}_0$

Estimation of $\pi_{s,j}$ and ε_s . The interest now lies in the estimation of ε_s and $\pi_{s,j}$ with strategy s when several RCTs are available that compare different subsets of the treatments. This is accomplished by considering the congruent data set for strategy s , $Data_s$. A congruent data set for s is the subset of $Data$ including those patients for whom the recommended treatment coincides with the actual given treatment. Using $Data_s$, we estimate all $\pi_{s,j}$ as the observed proportion of participants under each treatment j in strategy s , $\hat{\pi}_{s,j} = p_{s,j}^{Data_s}$.

To derive ε_s we need the following steps. First, we need to estimate the event rate $\varepsilon_{s,j}$ for each treatment as

recommended by strategy s . Then, the weighted average event rate under strategy s can be estimated as

$$\hat{\varepsilon}_s = \sum_{j=0}^J p_{s,j}^{Data_s} \times \hat{\varepsilon}_{s,j}.$$

The quantity $\hat{\varepsilon}_{s,j}$ depends on the strategy and the available data.

1. When we have only 1 RCT, then $\hat{\varepsilon}_{s,j} = e_j^{Data_s}$, that is, $\varepsilon_{s,j}$ is estimated as the observed proportion of events under arm j in $Data_s$.
2. When we have several RCTs, we first need to perform a meta-analysis of all placebo arms in $Data_s$ to obtain an estimate of the pooled placebo event rate $\hat{\varepsilon}_{s,0}$. Then, we perform a synthesis of all studies in $Data_s$ to estimate the pooled risk ratio of each treatment versus the control, $RR_j^{Data_s}$. Then, the treatment-specific event rates are $\hat{\varepsilon}_{s,j} = \hat{\varepsilon}_{s,0} \times RR_j^{Data_s}$.
 - a. In case $Data_s$ does not include placebo arms (e.g., when the treatments are highly effective or when the threshold values set are very low, it is more likely for the model to recommend an active treatment rather than placebo, and therefore, the congruent data set will include only active treatment arms), we could estimate the pooled event rate $\hat{\varepsilon}_{s,k}$ for another treatment k (instead of $\hat{\varepsilon}_{s,0}$), designated as the reference treatment, included in the congruent data set. Then, we again perform a synthesis of all studies in $Data_s$ to estimate the pooled risk ratio of each treatment versus k treatment, $RR_j^{Data_s}$. Then, the treatment-specific event rates are $\hat{\varepsilon}_{s,j} = \hat{\varepsilon}_{s,k} \times RR_j^{Data_s}$.
 - b. When $Data_s$ includes only 1 treatment arm, we could estimate the event rate $\hat{\varepsilon}_{s,j}$ as a meta-analysis of all j arms in $Data_s$.
3. When the strategy s is treat all with treatment $j = x$, with $x \neq 0$, the event rate $\hat{\varepsilon}_{s,x}$ can be estimated from the entire data set $Data$ as $\hat{\varepsilon}_{s,x} = \hat{\varepsilon}_x = \hat{\varepsilon}_0 \times RR_x^{Data}$. The observed proportion $p_{s,x}^{Data_s}$ is equal to 1, whereas the observed proportion $p_{s,j \neq x}^{Data_s}$ is equal to 0.
4. When the strategy s is “treat none,” then the NB is 0 as $\hat{\varepsilon}_{s,0} = \hat{\varepsilon}_0$, and the $\hat{\pi}_{s,j} = p_{s,j}^{Data_s}$ is 0 for all the available treatments j .

Hence, considering the nature of the strategies, the congruent data set, $Data_s$, is mainly used when the NB of a personalized model needs to be estimated. For the NB estimation of all other “fit all” strategies, the entire

data set, *Data*, is used. Considering also the nature of the available data, when several RCTs comparing several treatments are available, NMA and/or meta-analysis must be performed for the NB estimation; however, when only 1 RCT is available, the observed proportion of the event can be directly estimated.

NB and comparisons of strategies. We define the NB, which can be applied to all strategies and settings (i.e., 1 RCT, several RCTs, single treatment comparison, and several treatment comparisons) as

$$NB_s = \varepsilon_0 - \sum_{j=0}^J \pi_{s,j} \times \varepsilon_{s,j} - \sum_{j=0}^J \pi_{s,j} \times T_j.$$

The NB_s ranges between $-\max\{T_j\}$ and 1. It is $-\max\{T_j\}$ when there is no decrease in event rate compared with “treat none,” and at the same time, all patients take the drug with the highest threshold value T_j . NB has a theoretical maximum of 1 for the impossible case in which the decrease in event rate is 100% and none of the patients takes any treatment.

The advantage of any strategy $s = w$ compared with a strategy $s = m$, for specific threshold values T_w and T_m , can be calculated as the difference between the NB_w and the NB_m , and can be interpreted in terms of the decrease in event rate as follows: the use of strategy w compared with strategy m leads to $NB_w - NB_m$ fewer events for a constant treatment rate in each treatment j .

All of the required steps to calculate the NB for several strategies into an NMA of the RRMS example are presented in detail in Table 2. Following these steps, the NB for each strategy s is estimated for each combination of threshold values. If the personalized model has the highest NB across the entire range of threshold values, then its clinical relevance compared with the default strategies can be argued. If the optimal approach depends on the threshold values, then the typical conclusion would be that the personalized model is of unproven benefit.²⁸

Application in Comparison of Treatment Strategies in RRMS

We used NB to evaluate the clinical usefulness of the 2-stage personalized prediction model (described briefly in the “Case Study” section). As natalizumab is a treatment with serious side effects and is less safe than the other 2 options, patients would be prescribed natalizumab only when their benefit (here the predicted risk difference) is high. Dimethyl fumarate and glatiramer acetate, on the

other hand, are similar in terms of side effects and considered safer than natalizumab. In line with 2 consulting MS neurologists (see the acknowledgments), the threshold for natalizumab was set higher than those for dimethyl fumarate and glatiramer acetate.

We first assume a threshold value $T_N = 20\%$ for natalizumab and an equal (and lower) threshold value for the other 2 treatments, $T_{GA} = T_{DF} = 10\%$. These threshold values reflect the drugs’ profiles. Different patients might weight in differently the risk to relapse and the risks associated with each treatment. Therefore, we consider a range of threshold values for natalizumab (19%–40%) in combination with a range of common threshold values for dimethyl fumarate and glatiramer acetate (4%–25%). These ranges were selected based on safety concerns for each drug and on the congruent data set’s limitations; for lower threshold values, the NB could not be calculated because the congruent data set had only single-arm studies; hence, NMA could not be conducted. Then we plot NB_s as a function of threshold values T_j , to identify which treatment strategy leads to better clinical decision under different preferences on threshold values.

All of our analyses were done in R,³¹ using version 3.6.2. We made the code available in the following GitHub library: https://github.com/htx-r/Reproduce-results-from-papers/tree/master/DCA_NMA. The analysis code uses the `metaprop` command to estimate the event rate in control arm and the `netmeta` command to estimate the relative risk for each active treatment versus placebo.

Results

The results from comparing the 5 competing strategies with treatment thresholds $T_N = 20\%$, $T_{DF} = T_{GA} = 10\%$ are presented in Table 3, following all the steps presented in Table 2. A more detailed description of these estimations is presented in the appendix. Using these thresholds, the strategy based on the prediction model will recommend dimethyl fumarate to 1251 patients, natalizumab to 740 patients, and placebo to 9 patients. No patient is recommended to take glatiramer acetate. For the example threshold values ($T_N = 20\%$, $T_{DF} = T_{GA} = 10\%$), the congruent data set with the model strategy includes 652 patients: 4 patients in placebo, 418 in dimethyl fumarate, and 230 in natalizumab. The NB_s values are presented in Table 2 and show that treating RRMS patients using the strategy “treat patients according to the prediction model” results in higher NB_s compared with the default “one-size-fits-all” strategies. The NB_s for the “treat patients according to the prediction model” strategy is equal to 0.17.

Table 2 Detailed Description of the Net Benefit Estimation for “Treat All with Treatment j ” and “Treat Patients according to the Prediction Model” Strategies on the Relapsing-Remitting Multiple Sclerosis Example^a

Treat All with					Treat according to the Model				
	Placebo (“Treat None”)	Glatiramer Acetate	Dimethyl Fumarate	Natalizumab	Placebo (“Treat None”)	Glatiramer Acetate	Dimethyl Fumarate	Natalizumab	Total
Treatment rate	0%	100%	100%	100%	$p_0^{Data_s}$	$p_1^{Data_s}$	$p_2^{Data_s}$	$p_3^{Data_s}$	$\sum_{j=0}^J p_j^{Data_s}$
Event rate	\hat{e}_0 as a meta-analysis of all placebo arms in $Data$	$\hat{e}_1 = \hat{e}_0 \times RR_1^{Data}$	$\hat{e}_2 = \hat{e}_0 \times RR_2^{Data}$	$\hat{e}_3 = \hat{e}_0 \times RR_3^{Data}$	$\hat{e}_{s,0}$ as a meta-analysis of all placebo arms in $Data_s$	$\hat{e}_{s,1} = \hat{e}_{s,0} \times RR_1^{Data_s}$	$\hat{e}_{s,2} = \hat{e}_{s,0} \times RR_2^{Data_s}$	$\hat{e}_{s,3} = \hat{e}_{s,0} \times RR_3^{Data_s}$	$\hat{e}_s = \sum_{j=0}^J p_j^{Data_s} \times \hat{e}_{s,j}$
Decrease in event rate	0	$\hat{e}_0 - \hat{e}_1$	$\hat{e}_0 - \hat{e}_2$	$\hat{e}_0 - \hat{e}_3$			$\hat{e}_0 - \sum_{j=0}^J p_j^{Data_s} \times \hat{e}_{s,j}$		
Net strategy benefit	0	$NB_1 = \hat{e}_0 - \hat{e}_1 - T_{GA}$	$NB_2 = \hat{e}_0 - \hat{e}_2 - T_{DF}$	$NB_3 = \hat{e}_0 - \hat{e}_3 - T_N$		$NB_{model} = \hat{e}_0 - \sum_{j=0}^J p_j^{Data_s} \times \hat{e}_{s,j} - \sum_{j=1}^J p_j^{Data_s} \times T_j$			

^aThe threshold values for glatiramer acetate, dimethyl fumarate, and natalizumab are noted as T_{GA} , T_{DF} , and T_N respectively.

Table 3 Net Benefit (NB) Estimation for Each Strategy in the Multiple Sclerosis Example: “Treat None,” “Treat All Patients with Glatiramer Acetate,” “Treat All Patients with Dimethyl Fumarate,” “Treat All Patients with Natalizumab,” and “Treat Patients according to the Prediction Model”^a

Treat All with					Treat according to the Model				
	Placebo (“Treat None”)	Glatiramer Acetate	Dimethyl Fumarate	Natalizumab	Placebo (“Treat None”)	Glatiramer Acetate	Dimethyl Fumarate	Natalizumab	Total
Treatment rate	0%	100%	100%	100%	4/652 = 0.6%	0/652 = 0%	418/652 = 64.1%	230/652 = 35.3%	100%
Risk ratio from congruent data set	—	0.68	0.59	0.52	—	—	0.24	0.40	
Event rate	$\hat{e}_0 = 53\%$	$\hat{e}_1 = 36\%$	$\hat{e}_2 = 31\%$	$\hat{e}_3 = 28\%$	$\hat{e}_{s,0} = 75\%$	—	$\hat{e}_{s,2} = 18\%$	$\hat{e}_{s,3} = 30\%$	$\hat{e}_s = 23\%$
Decrease in event rate	0	17%	22%	25%			30%		
Net strategy benefit	0	$NB_1 = 0.07$	$NB_2 = 0.12$	$NB_3 = 0.05$			$NB_{model} = 0.17$		

^aThe threshold values used for the NB estimation are $T_{GA} = T_{DF} = 10\%$ and $T_N = 20\%$, respectively.

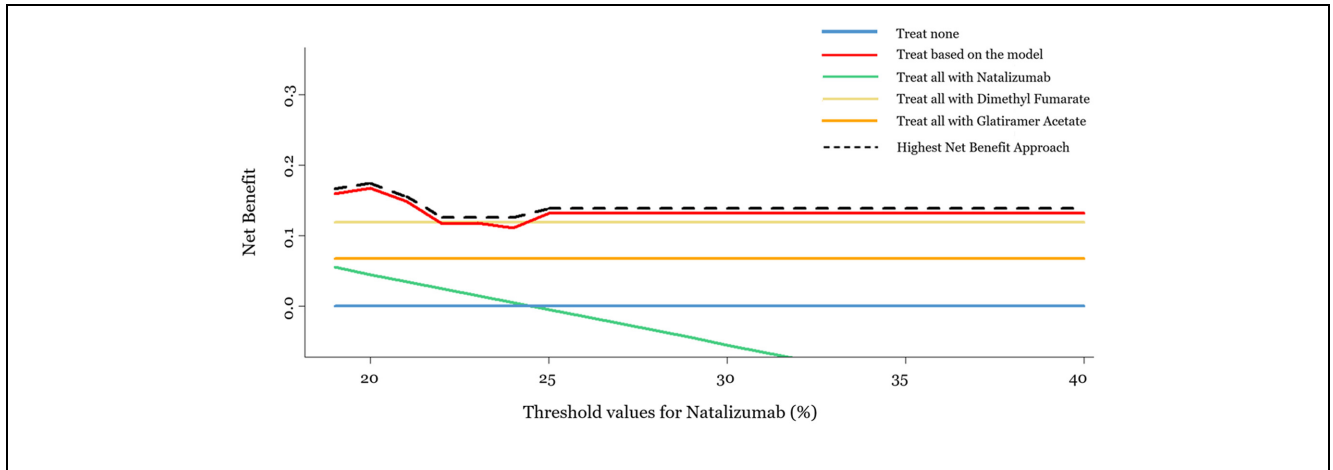


Figure 3 Decision curve analysis plot for a range of threshold values for natalizumab (19%–40%) and equal constant threshold values for dimethyl fumarate and glatiramer acetate (10%). The x-axis represents the range of threshold values for natalizumab, and the y-axis represents the net benefit for each of the 5 strategies: “treat none,” “treat based on the model,” “treat all with natalizumab,” “treat all with dimethyl fumarate,” and “treat all with glatiramer acetate.” The dashed black line represents the highest net benefit.

In Figure 3, we present the NB_s for each strategy when $19\% \leq T_N \leq 40\%$ and $T_{GA} = T_{DF} = 10\%$. This is an introductory plot that connects the traditional way of presenting the DCA results with its suggested extension and shows that the strategy “treat patients according to the prediction model” has the highest NB_s compared with the other strategies, almost in the whole range of natalizumab threshold values. We were restricted to using as the minimum threshold a natalizumab value of 19%, as lower threshold values result in a congruent data set consisting of only single-arm studies, and hence, NMA cannot be conducted. For threshold values higher than 40%, the results remain the same, and the strategy “treat based on the model” outperforms the others.

In Figure 4, we also present a heat plot showing the strategy with the highest NB_s when T_N is between 19% and 40% in combination with $T_{GA} = T_{DF}$ ranging between 4% and 25%. The empty gray cells in Figure 4 correspond to $T_{GA} = T_{DF} > T_N$, which is deemed clinically irrational. The numbers in the cells are differences in NB between the 2 strategies (multiplied by 100). As our focus is the clinical utility of the personalized prediction model, Figure 4 presents the NB from the model versus the highest NB from the default strategies. For instance, when $T_{GA} = T_{DF} = 20\%$ and $T_N = 25\%$, the “treat all with dimethyl fumarate” strategy outperforms all other strategies with an NB difference (multiplied by 100) versus “treat patients according to the prediction model” strategy of 0.2. This means that treating everyone with dimethyl fumarate would lead to 0.2% fewer relapse

events compared to choosing the treatment based on the model. The strategy “treat patients according to the prediction model” performs either better than or close to the one-size-fit-all treatment strategies (based on the NB differences). However, even small differences may be important in clinical decision making. The strategy “treat patients according to the prediction model” leads to better clinical decisions, when the thresholds for dimethyl fumarate and glatiramer acetate are low ($< \sim 10\%$) or when the threshold value for natalizumab is low ($< \sim 22\%$). The “treat none” strategy seems to outperform the others when all threshold values are high (i.e., for natalizumab $> 25\%$, for dimethyl fumarate and glatiramer acetate $> 20\%$). The “treat all with dimethyl fumarate” strategy seems to lead to better clinical decisions when the thresholds for dimethyl fumarate and glatiramer acetate are intermediate (between 10% and 20%) and at the same time the threshold for natalizumab is high ($> 25\%$). The strategy “treat all patients with glatiramer acetate” does not lead to the largest NB for any of the examined threshold combinations. Our methodology raises some questions about the universal applicability of the current personalized model and indicates that a better personalized model may be needed to be universally applicable for decision making.

Discussion

We extended the DCA methodology to an NMA framework to evaluate the clinical usefulness of a prediction

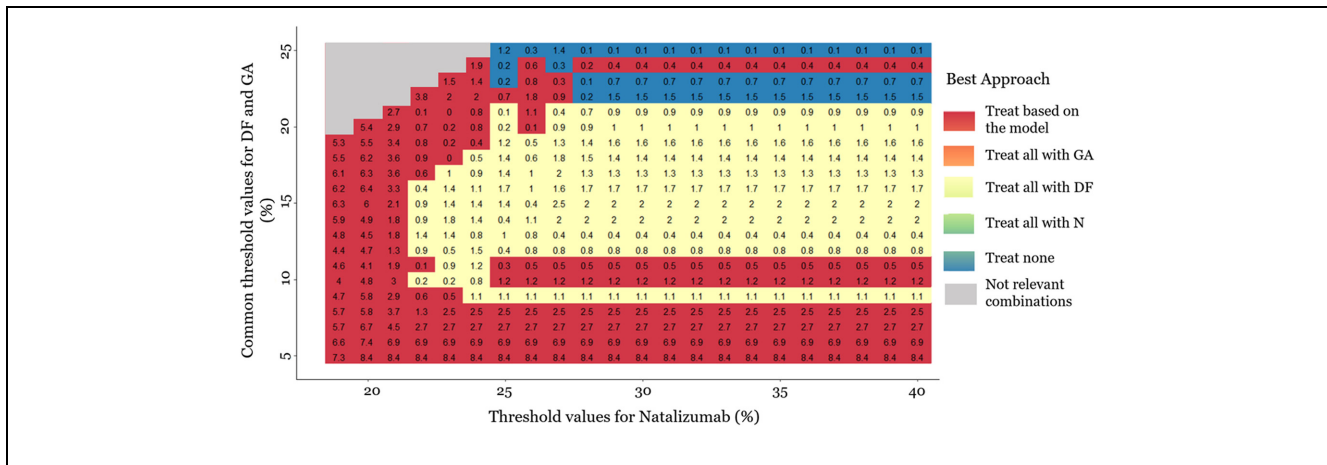


Figure 4 Heat plot for the decision curve analysis, in a range of threshold values. The same threshold is assumed for dimethyl fumarate (DF) and glatiramer acetate (GA) (4%–25%). The threshold values for natalizumab range between 19% and 40%. The plot shows which approach has the highest net benefit between all possible approaches: (a) treat all patients with placebo, (b) treat all patients with natalizumab (N), (c) treat all patients with dimethyl fumarate, (d) treat all patients with glatiramer acetate, and (e) treat patients based on the prediction model. The empty gray cells present the threshold value combinations that are not clinically possible. The numbers in the cells are differences in net benefit (NB) between the 2 strategies. When the “treat patients based on the prediction model” strategy is the best, the number in the cell (i.e., red cells) is the difference between its NB and the NB of the second-best strategy. Otherwise, we present the difference between the NB of the best strategy and the NB of the “treat patients based on the prediction model” strategy. The presented NB estimations are multiplied by 100.

model that aims at recommending a treatment among many possible options according to individual characteristics.^{9,32} The personalized prediction models are used to inform patients and decision makers about the most appropriate treatment for each patient and hence contribute to personalized medicine.^{12,13,33} Such models need to be evaluated for their ability to guide treatment decisions at the point of care. For this purpose, Vickers et al. proposed DCA, which is the tool to evaluate such prediction models by comparing the benefit–risk tradeoffs they entail to those of other default treatment strategies or other available personalized prediction models.⁹ The data used to evaluate such prediction models typically come from an RCT that compares 2 treatments: a reference treatment and the treatment of interest. As the treatment options for each condition are numerous and their effects are evaluated in multiple RCTs, the extended proposed DCA approach could contribute to evaluating the ability of the widely used personalized prediction models to guide treatment decisions. We applied our methodology for RRMS to evaluate the strategy of choosing between 3 disease-modifying drugs (natalizumab, dimethyl fumarate, glatiramer acetate) and placebo using a personalized prediction model.¹²

The methods and their application in the data set of treatments for RRMS have several limitations. The personalized prediction model compares only 3 active drugs

among all available options (more than 15 available). The same approach can be applied to personalized prediction models that compare all relevant competing drugs, assuming that studies that compare them are available. The main limitation of our application is the inefficient data set’s sample size; the estimation of the parameters needed to estimate the NB in our approach needs a large amount of data to ensure that the sample size of the congruent data set will be large enough to conduct NMA. Confidence intervals around the estimated NB could be shown to present uncertainty due to the limited sample size³⁴; however, they are not typically used within a classical decision-making approach.²⁸

Another technical issue is that it is possible that the congruent data set for some thresholds includes many single-arm studies. In our application, we omitted the single-arm studies from the NMA in the congruent data set to establish causal effects of the treatments, but this resulted in discarding potentially relevant information. When the congruent data set consists of single arms, (network) meta-analysis cannot be conducted at all. Consequently, NBs cannot be estimated for some threshold combinations, and researchers have to calculate the lower and upper bounds for the thresholds examined to ensure that they would lead to enough data to estimate NB. In our application, the lower bounds were outside the range of thresholds indicated by the expert

neurologists as relevant. In practice, however, the lack of suitable data to estimate NB for relevant thresholds can limit the applicability of DCA. The issue of single-arm studies in the congruent data set should be the subject of further research. Models that include single-arm studies in the meta-analysis could be considered, although the risk of bias in the estimates they provide is not to be underestimated.^{35–39} Finally, the strategies need to be evaluated for a relevant range of threshold values for all treatment options, as different patients might weight differently the risk of an event and risks associated with each treatment.²⁸ In our application, we defined equal threshold values for dimethyl fumarate and glatiramer acetate and higher threshold values for natalizumab according to the expert opinion of 2 MS neurologists based on the drugs' safety profiles. In practical application, the integration of utilities across a distribution of patients' preferences might be used to justify the range of relevant threshold values.²⁸

To our knowledge, this is the first attempt to use DCA to evaluate a prediction model that refers to multiple treatments and, consequently, uses evidence from several studies that compare subsets of the competing treatments, relying on the assumptions underlying NMA and prediction models (transitivity, consistency, correct model specification, etc.).^{40–43} The proposed approach can be used to compare several treatment strategies, and we show how to estimate the NB of a treatment strategy using causal treatment effects. If the strategy based on the personalized model is shown to be clinically useful compared with the default “treat all patients with X” strategy, this does not necessarily mean that it should be implemented in practice. In many clinical areas, the treating physician evaluates the patient and determines the treatment strategy without using a guiding tool. This state-of-the-art strategy needs to be compared with the strategy based on the model in a randomized clinical trial, to inform about the health benefits, patient experiences, and costs associated with clinical implementation of the decision tool.^{32,44,45} The original formulations of DCA were intended to supplement, rather than replace, other decision analytic techniques. For instance, a diagnostic test might be evaluated using a decision curve, with utilities determined (implicitly) by a range of threshold probabilities, or by a decision tree, in which utilities are assessed more formally, such as by using data from the literature. A cost-effectiveness analysis would incorporate economic costs obtained by additional research. The advantage of DCA is that it can be implemented without the need for specifying a large number of parameters that must be obtained from sources other than the current data set; the disadvantage is that it depends

on the assumption that clinicians are using threshold probabilities that are rational. Comparably, our proposed method aims to supplement, not replace, other decision-analytic methods for evaluating treatments and shares the similar advantage of practicability and disadvantage of the assumption of rational thresholds.

Personalized prediction models for treatment recommendation have recently gained ground, and their popularity will increase with the availability of more data. It is therefore important that such models are evaluated for their performance before they are ready to be used by decision makers. The traditional biostatistical metrics of calibration and discrimination can be useful for analysts to determine how to build and evaluate a model but cannot determine its clinical value.^{8,9,28} We have contributed to the existing methodological arsenal by providing a method to infer about a prediction model's clinical utility in a (network) meta-analysis framework. With the proposed approach, and assuming that enough data from several randomized trials would be available, the evaluation of clinical relevance will now be possible for several prediction models comparing many treatment options.

Authors' Note


Meetings at which this work was presented:


- Conference of the Austro-Swiss Region (RoeS) of the International Biometric Society (Salzburg, Austria, 7–10 September 2021)
- Royal Statistical Society (RSS) international conference 2021 (Manchester, UK, 6–9 September 2021)
- 42nd (virtual) Annual Conference of the International Society for Clinical Biostatistics (ISCB; online event 18–22 July 2021)

Acknowledgments

The authors thank Johannes Lorscheider and Jens Kuhle for their expert opinions on the drugs' safety profiles.

ORCID iDs

Konstantina Chalkou  <https://orcid.org/0000-0001-9718-021X>

Andrew J. Vickers  <https://orcid.org/0000-0003-1525-6503>

Supplemental Material

Supplementary material for this article is available on the *Medical Decision Making* website at <http://journals.sagepub.com/home/mdm>.

Data Availability Statement

The data that support the findings of this study were made available from Biogen International GmbH. Restrictions apply to the availability of these data, which were used under license for this study.

References

- Kent DM, Paulus JK, van Klaveren D, et al. The Predictive Approaches to Treatment effect Heterogeneity (PATH) statement. *Ann Intern Med.* 2020;172(1):35–45.
- Rekkas A, Paulus JK, Raman G, et al. Predictive approaches to heterogeneous treatment effects: a scoping review. *BMC Med Res Methodol.* 2020;20:264.
- Rothwell PM. Can overall results of clinical trials be applied to all patients? *Lancet Lond Engl.* 1995;345:1616–9.
- Kent DM, Hayward RA. Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification. *JAMA.* 2007;298:1209–12.
- Varadhan R, Segal JB, Boyd CM, et al. A framework for the analysis of heterogeneity of treatment effect in patient-centered outcomes research. *J Clin Epidemiol.* 2013;66:818–25.
- Harrell FE, Califf RM, Pryor DB, et al. Evaluating the yield of medical tests. *JAMA.* 1982;247:2543–6.
- Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiol Camb Mass.* 2010;21:128–38.
- Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making.* 2006;26:565–74.
- Vickers AJ, Kattan MW, Sargent DJ. Method for evaluating prediction models that apply the results of randomized trials to individual patients. *Trials.* 2007;8:14.
- Kent DM, Steyerberg E, van Klaveren D. Personalized evidence based medicine: predictive approaches to heterogeneous treatment effects. *BMJ.* 2018;363:k4245.
- Baker T, Gerdin M. The clinical usefulness of prognostic prediction models in critical illness. *Eur J Intern Med.* 2017;45:37–40.
- Chalkou K, Steyerberg E, Egger M, Manca A, Pellegrini F, Salanti G. A two-stage prediction model for heterogeneous effects of treatments. *Stat Med.* 2021;40:4362–75.
- Seo M, White IR, Furukawa TA, et al. Comparing methods for estimating patient-specific treatment effects in individual patient data meta-analysis. *Stat Med.* 2021;40:1553–73.
- Belias M, Rovers MM, Reitsma JB, et al. Statistical approaches to identify subgroups in meta-analysis of individual participant data: a simulation study. *BMC Med Res Methodol.* 2019;19:183.
- Tramacere I DGC, Salanti G DR, Filippini G. Immunomodulators and immunosuppressants for relapsing-remitting multiple sclerosis: a network meta-analysis. *Cochrane Database Syst Rev.* 2015;(9):CD011381.
- Lucchetta RC, Tonin FS, Borba HHL, et al. Disease-modifying therapies for relapsing-remitting multiple sclerosis: a network meta-analysis. *CNS Drugs.* 2018;32:813–26.
- Fogarty E, Schmitz S, Tubridy N, et al. Comparative efficacy of disease-modifying therapies for patients with relapsing remitting multiple sclerosis: systematic review and network meta-analysis. *Mult Scler Relat Disord.* 2016;9:23–30.
- Naci H, Fleurence R. Using indirect evidence to determine the comparative effectiveness of prescription drugs: do benefits outweigh risks? *Health Outcomes Res Med.* 2011;2:e241–9.
- van Valkenhoef G, Tervonen T, Zhao J, et al. Multicriteria benefit-risk assessment using network meta-analysis. *J Clin Epidemiol.* 2012;65:394–403.
- Ghasemi N, Razavi S, Nikzad E. Multiple sclerosis: pathogenesis, symptoms, diagnoses and cell-based therapy. *Cell J Yakhteh.* 2017;19:1–10.
- Goldenberg MM. Multiple sclerosis review. *Pharm Ther.* 2012;37:175–84.
- Crayton HJ, Rossman HS. Managing the symptoms of multiple sclerosis: a multimodal approach. *Clin Ther.* 2006;28:445–60.
- Rafiee Zadeh A, Askari M, Azadani NN, et al. Mechanism and adverse effects of multiple sclerosis drugs: a review article. Part 1. *Int J Physiol Pathophysiol Pharmacol.* 2019;11:95–104.
- Hoepner R, Faissner S, Salmen A, et al. Efficacy and side effects of natalizumab therapy in patients with multiple sclerosis. *J Cent Nerv Syst Dis.* 2014;6:41–9.
- Polman CH, O'Connor PW, Havrdova E, et al. A randomized, placebo-controlled trial of natalizumab for relapsing multiple sclerosis. *N Engl J Med.* 2006;354:899–910.
- Gold R, Kappos L, Arnold DL, et al. Placebo-controlled phase 3 study of oral BG-12 for relapsing multiple sclerosis. *N Engl J Med.* 2012;367:1098–107.
- Fox RJ, Miller DH, Phillips JT, et al. Placebo-controlled phase 3 study of oral BG-12 or glatiramer in multiple sclerosis. *N Engl J Med.* 2012;367:1087–97.
- Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. *Diagn Progn Res.* 2019;3:18.
- Simpson EH. The interpretation of interaction in contingency tables. *J R Stat Soc Ser B Methodol.* 1951;13:238–41.
- Ameringer S, Serlin RC, Ward S. Simpson's paradox and experimental research. *Nurs Res.* 2009;58:123–7.
- R Core Team. *A Language and Environment for Statistical Computing.* Vienna: R Foundation for Statistical Computing. 2020. Available from: <https://www.R-project.org/>
- Vickers AJ, Calster BV, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ.* 2016;352:i6.
- Kent DM, Rothwell PM, Ioannidis JP, et al. Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. *Trials.* 2010;11:85.
- Vickers AJ, Cronin AM, Elkin EB, et al. Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. *BMC Med Inform Decis Mak.* 2008;8:53.
- Schmitz S, Maguire A, Morris J, et al. The use of single armed observational data to closing the gap in otherwise disconnected evidence networks: a network meta-analysis in multiple myeloma. *BMC Med Res Methodol.* 2018;18:66.

36. Thom HHZ, Capkun G, Cerulli A, et al. Network meta-analysis combining individual patient and aggregate data from a mixture of study designs with an application to pulmonary arterial hypertension. *BMC Med Res Methodol*. 2015;15:34.
37. Signorovitch JE, Wu EQ, Yu AP, et al. Comparative effectiveness without head-to-head trials: a method for matching-adjusted indirect comparisons applied to psoriasis treatment with adalimumab or etanercept. *Pharmacoeconomics*. 2010;28:935–45.
38. Caro JJ, Ishak KJ. No head-to-head trial? Simulate the missing arms. *Pharmacoeconomics*. 2010;28:957–67.
39. Lin L, Zhang J, Hodges JS, et al. Performing arm-based network meta-analysis in R with the pnetmeta package. *J Stat Softw*. 2017;80:5.
40. Donegan S, Williamson P, D'Alessandro U, et al. Assessing key assumptions of network meta-analysis: a review of methods. *Res Synth Methods*. 2013;4:291–323.
41. Rouse B, Chaimani A, Li T. Network meta-analysis: an introduction for clinicians. *Intern Emerg Med*. 2017;12: 103–11.
42. 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.
43. Caldwell DM. An overview of conducting systematic reviews with network meta-analysis. *Syst Rev*. 2014;3:109.
44. Van Calster B, Wynants L, Verbeek JFM, et al. Reporting and interpreting decision curve analysis: a guide for investigators. *Eur Urol*. 2018;74:796–804.
45. Vickers AJ, Cronin AM. Traditional statistical methods for evaluating prediction models are uninformative as to clinical value: towards a decision analytic framework. *Semin Oncol*. 2010;37:31–8.