Original Research Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2022, Vol. 31(5) 899–916

DOI: 10.1177/09622802211072512
journals.sagepub.com/home/smm
SAGE

A comparison of various aggregation functions in multi-criteria decision analysis for drug benefit–risk assessment

Tom Menzies[1,2], **Gaelle Saint-Hilary**[3,4], **and Pavel Mozgunov**[5]

Abstract

Multi-criteria decision analysis is a quantitative approach to the drug benefit–risk assessment which allows for consistent comparisons by summarising all benefits and risks in a single score. The multi-criteria decision analysis consists of several components, one of which is the utility (or loss) score function that defines how benefits and risks are aggregated into a single quantity. While a linear utility score is one of the most widely used approach in benefit–risk assessment, it is recognised that it can result in counter-intuitive decisions, for example, recommending a treatment with extremely low benefits or high risks. To overcome this problem, alternative approaches to the scores construction, namely, product, multi-linear and Scale Loss Score models, were suggested. However, to date, the majority of arguments concerning the differences implied by these models are heuristic. In this work, we consider four models to calculate the aggregated utility/loss scores and compared their performance in an extensive simulation study over many different scenarios, and in a case study. It is found that the product and Scale Loss Score models provide more intuitive treatment recommendation decisions in the majority of scenarios compared to the linear and multi-linear models, and are more robust to the correlation in the criteria.

Keywords

Aggregation function, benefit–risk, decision-making, loss score, multi-criteria decision analysis

1 Introduction

The benefit–risk analysis of a treatment consists of balancing its favourable therapeutic effects versus adverse reactions it may induce.[1] This is a process which drug regulatory authorities, such as EMA[2] and FDA[3] use when deciding whether a treatment should be recommended. Benefit–risk assessment (BRA) is mostly performed in a qualitative way.[4] However, this approach has been criticised for a lack of transparency behind the final outcome, in part due to large amounts of data considered for this assessment, and the differing opinions on what this data means. To counter this, quantitative approaches ensuring continuity and consistency across drug BRA, and making the decisions easier to justify and to communicate, were proposed.[5,6] While there is a number of methods to conduct the quantitative BRA, the multi-criteria decision analysis (MCDA) has been particularly recommended by many expert groups in the field.[7–10] MCDA provides a single score

[1]Clinical Trials Research Unit, Leeds Institute of Clinical Trials Research, University of Leeds, UK
[2]Department of Mathematics and Statistics, Lancaster University, UK
[3]Department of Biostatistics, Institut de Recherches Internationales Servier (IRIS), France
[4]Dipartimento di Scienze Matematiche (DISMA) Giuseppe Luigi Lagrange, Politecnico di Torino, Italy
[5]MRC Biostatistics Unit, University of Cambridge, UK

Corresponding author:
Gaelle Saint-Hilary, Department of Biostatistics, Institut de Recherches Internationales Servier, Suresnes, le-de-France, France.
Email: gsainthilary@gmail.com

(a utility or loss score) for a treatment, which summarises all the benefits and risks induced by the treatment in question. These scores are then used to compare the treatments and to guide the recommendation of therapies over others.

Mussen et al.[7] proposed to use a linear aggregation model in the MCDA, which takes into account all main benefits and risks associated with a treatment (as well as their relative importance) to generate a treatment utility score by taking a linear combination of all criteria. This utility score is then compared against the utility score of a competing treatment, and that with the highest score is recommended. This model appealed for numerous reasons, one of which was its simplicity. The proposed method, however, was deterministic, point estimates of the benefit and risk criteria were used, and no uncertainty around these estimates was considered. Yet, uncertainty and variance are expected in treatments' performances, and must therefore be accounted for in the decision making.

To resolve this shortcoming, probabilistic MCDA (pMCDA)[11] that accounts for the variability of the criteria through a Bayesian approach was proposed. Generalisations of pMCDA for the case of uncertainty in the relative importance of the criteria were developed, named stochastic multi-criteria acceptability analysis (SMAA)[12] or Dirichlet SMAA.[13] However, it was acknowledged that by accounting for several sources of uncertainty, these models become more complex and should be used primarily for the sensitivity analysis.

All the works discussed above concern a linear model for aggregation of the criteria, which is thought to be primarily due to its wider application in practice rather than its properties. One argument against the linear model is that a treatment which has either no benefit or extreme risk could be recommended over other alternatives without such extreme characteristics.[14–16] In addition, the linearity implies that the relative tolerance in the toxicity increase is constant for all levels of benefit that might not be the case for a number of clinical settings. To address these points, a Scale Loss Score (SLoS) model was developed. This model made it impossible for treatments with no benefit or extremely high risk be recommended. It also incorporates a decreasing level of risk tolerance relative to the benefits: where an increase in risk is more tolerated when benefit improves from 'very low' to 'moderate' compared to an increase from 'moderate' to 'very high'. SLoS model resulted in similar recommendations to the linear MCDA model when the one treatment is strictly preferred to another (i.e. has both lower risk and higher benefit), but resulted in more intuitive recommendations if one of the treatments has either extremely low benefit or extremely high risk.

Whilst other methods are discussed in the literature, the only application of a non-linear BRA model to the medical field is made by Saint-Hilary et al.,[14] and this only compares the linear and SLoS models. This paper shall build on this comparison by introducing various different aggregation models (AM) to analyse how each work compared to the other in the medical field (by conduction a case study and a simulation study), and allow an informed decision to be made as to which one should be used using the results of an extensive and comprehensive simulations study over a number of clinical scenarios. We will also use a case study to demonstrate the implication of the choice of AM on the actual decision making using the MCDA.

The rest of the paper proceeds as follows. The general MCDA methodology, the four different aggregation models considered, linear, product, multi-linear and SLoS, and the choice of the weights for them are given in Section 2. In Section 3, we revisit a case study conducted by Nemeroff[17] looking at the effects of Venlafaxine, Fluoxetine and a placebo on depression, applying the various aggregation models to a given dataset. In Section 4, a comprehensive simulation study comparing the four aggregation models in many different scenarios is presented, as well as the effects any correlation between criteria may have. We conclude with a discussion in Section 5.

2 Methodology

All of the aggregation models (referred to as to 'models' below) considered in this work are all classified within the MCDA family – they aggregate the information about benefits and risks in a single (utility or loss) score. Therefore, we would refer to each of the approaches by their models for the computation of the score. Below, we outline the general MCDA framework for the construction of a score using an arbitrary model. We consider the MCDA taking into account the variability of estimates, pMCDA.[11]

2.1 Setting

Consider m treatments (indexed by i) which are assessed on n criteria (indexed by j). To ensure continuity, we use the same notations as those of Saint-Hilary et al.[14]:

- $\xi_{i,j}$ is the performance of treatment i on criterion j, so that treatment i is characterised by a vector showing how it performed on each criterion: $\boldsymbol{\xi_{i,j}} = (\xi_{i,1}, \ldots, \xi_{in})$.

- The monotonically increasing partial value functions $0 \leq u_j(\cdot) \leq 1$ are used to normalise the criterion performances. Let ξ'_j and ξ''_j be the most and the least preferable values, then $u_j(\xi''_j) = 0$ and $u_j(\xi'_j) = 1$. The inequality $u_j(\xi_{ij}) > u_j(\xi_{hj})$ indicates that the performance of the treatment i is preferred to the performance of the treatment h on criterion j. In this work, we focus on linear partial value functions, one of the most common choice in treatment BRA[5,7,12,18,11] that can be written as

$$u_j(\xi_{ij}) = \frac{\xi_{ij} - \xi''_j}{\xi'_j - \xi''_j}. \tag{1}$$

- The weights indicating the relative importance of the criteria are known constants denoted by w_j. The vector of weights used for the analysis is denoted by $\boldsymbol{w} = (w_1, \ldots, w_n)$.
- The MCDA utility or loss scores of treatment i are obtained as

$$u(\boldsymbol{\xi}_i, \boldsymbol{w}) := u\big(w_j, u_j(\xi_{ij})\big), \quad j = 1, \ldots, n$$

and

$$l(\boldsymbol{\xi}_i, \boldsymbol{w}) := l\big(w_j, u_j(\xi_{ij})\big), \quad j = 1, \ldots, n$$

respectively, where $u(\cdot)$ and $l(\cdot)$ are the functions specifying how the criteria should be summarised in a single score, and are referred to as 'aggregation models'. The impact of this model's choice on the performance of treatment recommendation is the focus on this work. The higher the utility score, or lower the loss score, the more preferable the benefit–risk ratio. Then, the comparison of treatments i and h is based on

$$\Delta u(\boldsymbol{\xi}_i, \boldsymbol{\xi}_h, \boldsymbol{w}) := u(\boldsymbol{\xi}_i, \boldsymbol{w}) - u(\boldsymbol{\xi}_h, \boldsymbol{w}) \tag{2}$$

or

$$\Delta l(\boldsymbol{\xi}_i, \boldsymbol{\xi}_h, \boldsymbol{w}) := l(\boldsymbol{\xi}_i, \boldsymbol{w}) - l(\boldsymbol{\xi}_h, \boldsymbol{w}). \tag{3}$$

Within a Bayesian approach, the utility score $u(\boldsymbol{\xi}_i, \boldsymbol{w})$ and the loss score $l(\boldsymbol{\xi}_i, \boldsymbol{w})$ are random variables having a prior distribution. Given observed outcomes $\mathbf{x_i} = (x_{i1}, \ldots, x_{in})$ and $\mathbf{x_h} = (x_{h1}, \ldots, x_{hn})$ (corresponding to treatment performances $\boldsymbol{\xi}_i$ and $\boldsymbol{\xi}_h$, respectively) for i and h, one can obtain the posterior distribution of $\Delta u(\boldsymbol{\xi}_i, \boldsymbol{\xi}_h, \boldsymbol{w})$ or $\Delta l(\boldsymbol{\xi}_i, \boldsymbol{\xi}_h, \boldsymbol{w})$, respectively. The inference is based on the complete posterior distribution and the conclusion on the benefit–risk balance is supported by the probability of treatment i to have a greater utility score (or smaller loss score) than treatment h:

$$\mathcal{P}_u^{ih} = \mathbb{P}(\Delta u(\boldsymbol{\xi}_i, \boldsymbol{\xi}_h, \boldsymbol{w}) > 0 \mid \mathbf{x_i}, \mathbf{x_h}). \tag{4}$$

or

$$\mathcal{P}_l^{ih} = \mathbb{P}(\Delta l(\boldsymbol{\xi}_i, \boldsymbol{\xi}_h, \boldsymbol{w}) < 0 \mid \mathbf{x_i}, \mathbf{x_h}). \tag{5}$$

The probabilities (4) or (5) are used to guide a decision on taking/dropping a treatment. A possible way to formalise the decision based on this probability is to compare it to a threshold confidence level $0.5 \leq \psi \leq 1$. Then, $\mathcal{P}_u^{ih} > \psi$ (or $\mathcal{P}_l^{ih} > \psi$) would mean that one has enough evidence to say that treatment i has a better benefit–risk balance than h with a level of confidence ψ. Note that $\mathcal{P}_u^{ih} = 0.5$ (and $\mathcal{P}_l^{ih} = 0.5$) corresponds to the case where the benefit–risk profiles of i and h are equal according to the corresponding MCDA model.

2.2 Aggregation models

Below, we consider four specific forms of aggregation models, namely, linear, product, multi-linear and SLoS, that were argued by various authors to be used in the MCDA to support decision making.

2.2.1 Linear model

A linear aggregation of treatment's effects on benefits and risks remains the most common choice for the treatment development.[7,12,19,18,13] Under the linear model, the utility score is computed as

$$u^L(\boldsymbol{\xi}_i, \boldsymbol{w}^L) : = \sum_{j=1}^{n} w_j^L u_j(\xi_{i,j}) \tag{6}$$

where $w_j^L > 0 \ \ \forall j$ and $\sum_{j=1}^{n} w_j^L = 1$, the superscript L referring to the linear model. The expression (6) is used in equations (2) and (4) to compare the associated linear scores for a pair of treatments.

As an illustration of all considered aggregation models, we will use the following example with two criteria: one benefit indexed by 1, one risk indexed by 2. The linear utility score for treatment i at fixed parameter values θ_{i1}, θ_{i2} takes the form

$$u^L(\theta_{i1}, \theta_{i2}, w^L) := w^L u_1(\theta_{i1}) + (1 - w^L) u_2(\theta_{i2}). \tag{7}$$

As values $u_1(\theta_{i1}), u_2(\theta_{i2}) \in (0, 1)$, one can interpret $u_1(\theta_{i1})$ as a probability of benefit and $1 - u_2(\theta_{i2})$ as a probability of risk. This utility score can be transformed into a loss score by subtracting it from one

$$l^L(\theta_{i1}, \theta_{i2}, w^L) := 1 - u^L(\theta_{i1}, \theta_{i2}, w^L) \tag{8}$$

We do this as, historically, the concept of a loss function is preferred both in statistical decision theory and Bayesian analysis for parameter estimation.[20] The contours of equal linear loss score for all values of $u_1(\theta_{i1})$ and $(1 - u_2(\theta_{i2}))$ are given in panel A of Figure 1 using $w^L = 0.5$ (top row) and $w^L = 0.25$ (bottom row).

The contours represent the loss score for each benefit–risk pair. Lower values of $l^L(\theta_{i1}, \theta_{i2}, w^L)$ correspond to better treatment benefit–risk profiles. It is minimised (right bottom corner) when the maximum possible benefit is reached ($u_1(\theta_{i1}) = 1$) with no risk ($1 - u_2(\theta_{i2}) = 0$). The contours are linear, with a constant slope $w^L/(1 - w^L)$. This implies that if one treatment has an increased probability of risk of x% compared to another, its benefit probability should be increased by $(1 - w^L)/w^L \times x$% to have the same utility score, and this holds for all values of benefit and risk. This figure allows for an illustration of the penalisation of various benefit–risk criteria and for an illustrative comparison between treatments with different criteria. For example, any pairwise comparison that lies on a contour line shows that the two treatments are seen as equal.

The major advantage of the linear model is its intuitive interpretation: a poor efficacy can be compensated by a good safety, and vice-versa. However, the linear utility score can result in the recommendation of highly unsafe or poorly effective treatment[21,6] and, consequently, in a counter-intuitive conclusion. Moreover, the linearity implies that the relative tolerance in the toxicity increase is constant for all levels of benefit.[14] These pitfalls could be avoided (or at least reduced) by using non-linear models.[6,22] Specifically, Saint-Hilary et al.[14] advocated introducing two principles a desirable benefit-risk analysis aggregation model should have:

1. One is not interested in treatments with extremely low levels of benefit or extremely high levels of risks (regardless of how the treatment performs on other criteria).

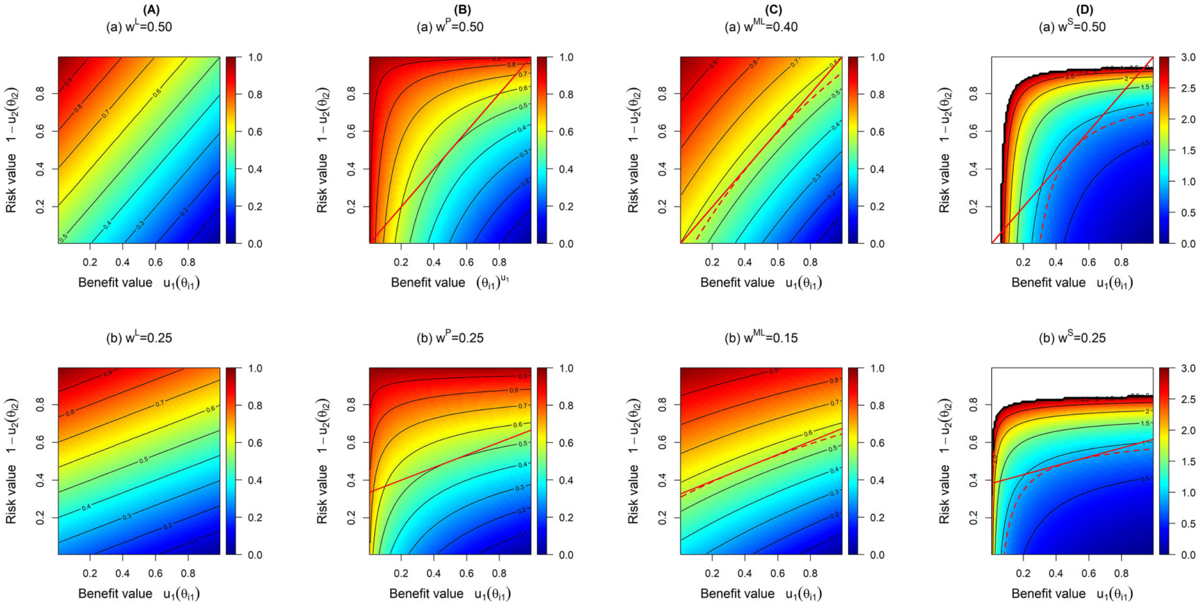

Figure 1. Contour plots for linear (A), product (B), multi-linear (C) and SLoS (D) models with (i) two equally important criteria (top row), and (ii) the risk criterion being twice as important (on average for non-linear model) as the benefit criterion (bottom row). Red lines on panels B to D represent the tangents at the middle point (0.5, 0.5).

2. Decreasing level of risk tolerance relative to benefits: an increase in risk could be more tolerated when benefit improves from 'very low' to 'moderate', compared to from 'moderate' to 'very high'.

Below, we consider three models having one or both of these properties.

2.2.2 Product model

A multiplicative aggregation (known as a product model) is an alternative method of comparing treatment's effects on benefits and risks.[23] Under the product model, the utility score is computed as

$$u^P(\boldsymbol{\xi_i}, \boldsymbol{w}^P) := \prod_{j=1}^{n} u_j(\xi_{i,j})^{w_j^P} \tag{9}$$

where the superscript P refers to the product model. The expression (9) is used in equations (2) and (4) to compare the associated product scores for a pair of treatments.

The product utility score for treatment i with two criteria at fixed parameter values θ_{i1}, θ_{i2} takes the form

$$u^P(\theta_{i1}, \theta_{i2}, w^P) := u_1(\theta_{i1})^{w^P} \times u_2(\theta_{i2})^{(1-w^P)}. \tag{10}$$

Similarly as for the linear model, this utility score can be transformed into a loss score by subtracting it from one

$$l^P(\theta_{i1}, \theta_{i2}, w^P) := 1 - u^P(\theta_{i1}, \theta_{i2}, w^P) \tag{11}$$

The contours of equal product loss score for all values of $u_1(\theta_{i1})$ and $(1 - u_2(\theta_{i2}))$ are given in panel B of Figure 1 using $w^P = 0.5$ (top row) and $w^P = 0.25$ (bottom row).

One advantage the product model has over the linear model is that it cannot recommend treatments with either zero benefit or extreme risk. This is because either of these two options would result in a score of zero for the utility function, and as such would make it impossible for such a treatment to be recommended. The contour lines in panel B in Figure 1 demonstrate how the product model penalises undesirable values compared to the linear model. These contours are curved, and are bunched together tightest at points where benefit values are low and where risk values are high. This shows how the penalisation differs this model from the linear model, as under the linear model, an increase/decrease in benefit–risk is treated equally regardless of the marginal values of these criteria, whereas the values of these criteria often have an effect on our decision making under the product model.

2.2.3 Multi-linear model

A multi-linear model for the aggregation of treatments' benefits and risks provides a one more alternative for the comparison of two treatments.[22] This model can be seen as attempt to combine the linear and product model. Under the multi-linear model, the utility score is computed as

$$\begin{aligned} u^{ML}(\boldsymbol{\xi_i}, \boldsymbol{w}^{ML}) := {} & \textstyle\sum_{j=1}^{n} w_j^{ML} u_j(\xi_{i,j}) + \sum_{j=1,k>j}^{n} w_{j,k}^{ML} u_j(\xi_{i,j}) u_k(\xi_{i,k}) \\ & + \textstyle\sum_{j=1,l>k>j}^{n} w_{j,k,l}^{ML} u_j(\xi_{i,j}) u_k(\xi_{i,k}) u_l(\xi_{i,l}) + \cdots + w_{1,2,\ldots,n}^{ML} u_1(\xi_{i,1}) u_2(\xi_{i,2}) \cdots u_n(\xi_{i,n}) \end{aligned} \tag{12}$$

where the superscript ML refers to the multi-linear model, and the weight criteria $w_{i,j,\ldots}^{ML}$ refer to the weight criteria given to the interaction term between criteria $i, j, \ldots$ We require all the weights in the ML model to sum up to 1. The expression (12) is used in equations (2) and (4) to compare the associated multi-linear scores for a pair of treatments.

Considering the example with two criteria, the multi-linear utility score for treatment i at fixed parameter values θ_{i1}, θ_{i2} takes the form

$$u^{ML}(\theta_{i1}, \theta_{i2}, w_1^{ML}, w_2^{ML}, w_{1,2}^{ML}) := w_1^{ML} u_1(\theta_{i1}) + w_2^{ML} u_2(\theta_{i2}) + w_{1,2}^{ML}(u_1(\theta_{i1}) u_1(\theta_{i2})). \tag{13}$$

Note that the even under the constraint of the sum of the weights to be equal to one, there is one more weight parameter than for the linear and product models. This immediately can make the weight elicitation procedure more involving for all stakeholders. To link the weights of the ML model with the rest of the competing approaches (see more details in Section 2.3), we set up one more constraint, so that the number of weight parameters is the same in all considered model (for the purpose of the comparison in this manuscript). Specifically, we fix $w_{1,2}^{ML} = c$ where $0 \le c \le 1$, implying that we fix the effect of the interaction term. Similarly as for the linear and product models, this utility score can be transformed into a loss score by subtracting it from one:

$$l^{ML}(\theta_{i1}, \theta_{i2}, w^{ML}) := 1 - u^{ML}(\theta_{i1}, \theta_{i2}, w^{ML}) \tag{14}$$

The contours of equal linear loss score for all values of $u_1(\theta_{i1})$ and $(1 - u_2(\theta_{i2}))$, $c = 0.20$ are given in panel C of Figure 1 using $w_1^{ML} = 0.40$ (top row) and $w_1^{ML} = 0.15$ (bottom row).

The contour lines demonstrate the almost linear trade-off between benefit and risk, but that there is a slight curvature (which becomes more prominent as it moves further away from more desirable values), indicating a moderate penalisation of extreme values. This shows that while this model attempts to penalise the undesirable criteria values, this effect does not seem to be as strong as in the product model, admittedly due to the chosen value of the weight, $w_{1,2}^{ML}$, given to the interaction term. A moderate level of penalisation for the chosen value of the weight corresponding to the interaction term allows for treatments to be recommended when there is no benefit or extreme risk, as is the case in the linear model. The more the weight of the interaction terms, the less likely this would happen.

2.2.4 SLoS model

An alternative to the models proposed above is the Scale Loss Score (SLoS) model, which was proposed by Saint-Hilary *et al.*[14] to satisfy the two desirable properties for an aggregation method. First of all, in contrast to the three models above, SLoS considers a loss score, rather than a utility score, as the output. Therefore, lower values are more desirable. Under the SLoS model, the loss score is computed as

$$l^S(\boldsymbol{\xi_i}, \boldsymbol{w}^S) := \sum_{j=1}^{n} \left(\frac{1}{u_j(\xi_{i,j})} \right)^{w_j^S} \tag{15}$$

where the superscript S refers to the SLoS model. The expression (15) is used in equations (3) and (5) to compare the associated SLoSs for a pair of treatments. The loss score could theoretically be transformed into a utility score as $u^S(\boldsymbol{\xi_i}, \boldsymbol{w}^S) = -l^S(\boldsymbol{\xi_i}, \boldsymbol{w}^S)$. However, this form is usually not used because it provides negative utility values, which is not intuitive for a utility concept.

Coming back to the example with two criteria, the loss score for treatment i at fixed parameter values θ_{i1}, θ_{i2} takes the form

$$l^s(\theta_{i1}, \theta_{i2}, w^S) := \left(\frac{1}{u_1(\theta_{i1})} \right)^{w^S} + \left(\frac{1}{u_2(\theta_{i2})} \right)^{(1-w^S)}. \tag{16}$$

The contours of equal scale loss score for all values of $u_1(\theta_{i1})$ and $(1 - u_2(\theta_{i2}))$ are given in panel D of Figure 1 using $w^S = 0.5$ (top row) and $w^S = 0.25$ (bottom row).

As is the case with the product model, this penalisation makes it impossible for treatments with either no benefit or extreme risk to be recommended over other potential treatments, compared to the linear and multi-linear models (which can recommend such treatments). This is because a treatment that had either of these would return a loss score of infinity (regardless of the values of any other criteria) and would therefore be non-recommendable. On the figure, the white colour at extreme undesirable values (either very low benefit or very high risk) corresponds to very high to infinite loss scores and demonstrate the penalisation effect.

Of note, Figure 1 displays the contours of equal *loss* score for all the models, so all the plots on this figure could be interpreted in the same way, with lower scores (in blue) corresponding to more desirable benefit–risk profiles. Even when these contour plots concern the same *values* of weights in the models, the weights themselves are different in each model (represented by different indices). Therefore, when to provide a fair comparison of these models, it is important to ensure that the models carry (approximately) the same relative importance of the criteria defined through the slope of the contour lines. We propose an approach to match the relative importance of the models below.

2.3 Weight elicitation and mapping

Methods for quantifying subjective preferences, for example, Discrete Choice Experiment and Swing-Weighting, have been widely studied in the literature.[6,7,24,25] Applied to drug BRA, the majority of the weight elicitation methods concern the linear model. In the linear model framework, the weight assigned to one criterion is interpreted as a scaling factor which relates one increment on this criterion to increments on all other criteria.

Note that each of the aggregation models use the individual weights, w^L, w^P, w^{ML} and w^S. However, in the actual analysis, regardless of the aggregation model used, one can expect only one underlying level of the relative importance of the considered benefit and risk criteria, as the stakeholders' preferences between the criteria should not depend on the methodology used for the decision making. Therefore, it is crucial to make sure when applying different models to the same problem that they reflect the same stakeholders' preferences. We adapt the approach proposed by Saint-Hilary et al.[14]

to achieve that. Since comprehensive work has been published and is currently being continued on the weight elicitation for the linear model, we will map the weights w_j^L (hypothetically) elicited for the linear model to the weights w^P, w^{ML} and w^S such that they reflect the same trade-off preferences between the criteria.

2.3.1 Mapping for two criteria

As described in Saint-Hilary et al.,[14] formally, the trade-off between the criteria could be represented by the slope of the tangent of the contour lines where the contour line passes through the point (0.5, 0.5) (see the red lines in the contour plot of panels B to D in Figure 1). Therefore, the expressions for the mapping of the linear weight to the competitive models are found through the equality of the slopes of the tangents to the corresponding contour lines.

We start from the setting with two criteria. As stated above, even for the two criteria setting, the multi-linear model requires one more weight to be specified. Therefore, we impose a constraint on the weight corresponding to the interaction term to obtain the unique solution for the mapped weight w^{ML}, specifically $w_{1,2}^{ML} = 1 - w_1^{ML} - w_2^{ML} = c$, where $0 \leq c \leq 1$. Note that for $c = 0$, the multi-linear model reduces to the linear one, and for $c = 1$ it becomes the product of the two criteria values.

Using the utility/loss scores z^P, z^{ML}, z^S obtained at point $(u_1(\theta_{i1}), u_2(\theta_{i2}))$, the expressions of the equality of the tangents with two criteria take the form

$$\begin{aligned}
\frac{w^L}{1-w^L} &= \frac{w^P}{1-w^P}\left(\frac{1}{u_1(\theta_{i,1})}\right)\left(\frac{z^P}{u_1(\theta_{i,1})^{w^P}}\right)^{\frac{1}{1-w^P}}, \\
\frac{w^L}{1-w^L} &= \frac{w_1^{ML}w_2^{ML}+z^{ML}-z^{ML}(1-c)}{(w_2^{ML}+cu_1(\theta_{i,1}))^2}, \\
\frac{w^L}{1-w^L} &= \frac{w^S}{1-w^S}\left(z^S - u_1(\theta_{i1})^{-w^S}\right)^{\frac{w^S-2}{1-w^S}} \times u_1(\theta_{i1})^{-(w^S+1)}.
\end{aligned} \tag{17}$$

where the slope for the linear model is given in the left hand size, and the slopes for the product, multi-linear and SLoS models are given in the right hand side, respectively.

Note, however, that the slope of the tangent of the contours for the linear model are *constant* for all values of parameters and defined by the weights w_j^L only, while the slopes for the competitive models change with the values of the criteria. For the purpose for the weights mapping, we would interpret w_j^L as an *average* relative importance of each criterion over the others, and would match the slopes of the tangents to the corresponding contours in the middle point, $u_1(\theta_{i1}) = u_2(\theta_{i2}) = 0.5$.[14] Then, the equalities above reduce to

$$\begin{aligned}
w^L &= w^P, \\
w^L &= w^{ML} + c/2, \\
\frac{w^L}{1-w^L} &= \frac{w^S}{1-w^S} \cdot 2^{(2w^S-1)}.
\end{aligned} \tag{18}$$

Therefore, the product weight coincides with the linear weight in the given middle mapping point. For the SLoS model, the weight mapping does not have an analytical solution, but the approximate value of w^S can be obtained by line search.

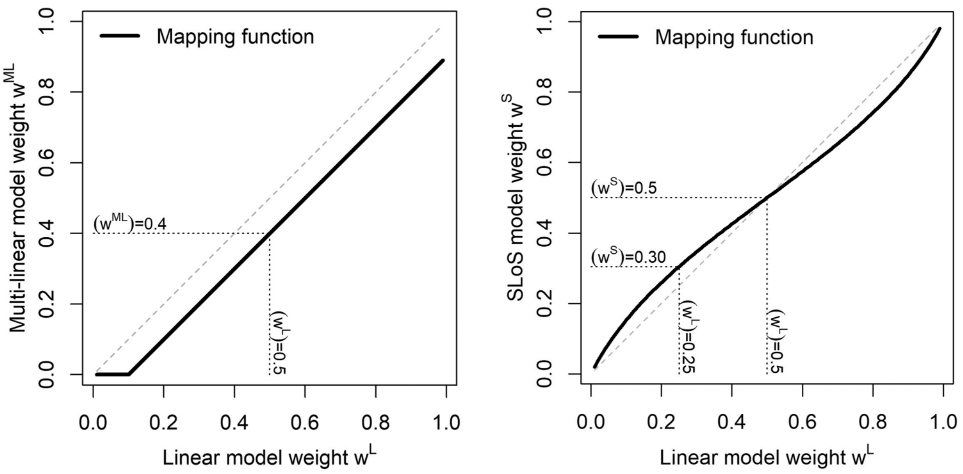

Figure 2. Weight mapping from the linear model to the multi-linear model (left) and to the SLoS model (right).

Figure 2 shows the mapping from the linear model to the multi-linear and SLoS models. It demonstrates how the value for the linear model (x-axis) can be used to find the respective weights for the multi-linear and SLoS models on the y-axes.

One can note that for the multi-linear model, the proposed mapping process may result in the obtained negative mapped values of weight. This is because of how the weight mapping function is elicited in the two criteria case: if the value of a weight under the linear model is less than half the value of c, then this will map to a negative value (which, in theory, gives our criteria a negative importance – which is impossible) to reflect the same relative importance as induced by the linear model. Intuitively, if the interaction terms already contributes more to the importance of the one of the criterion in the interaction, the model needs to subtract the 'excessive' importance from the weight corresponding to this criterion standing alone. Whilst this effect can be negated by setting an upper limit of the values c can take, this in term limits the effect the interaction terms have, and can make the model more similar to the linear model. This is demonstrated in Figure 2 for $c = 0.2$, where any weights for the linear model that are given a value of 0.1 or less would be mapped to 0 in the multi-linear model, rather than a negative value.

The mapped weights for the multi-linear and SLoS models do not have a direct interpretation, and should be back-transformed to linear weights to be interpreted. For instance, with two criteria, a weight of 0.3 for SLoS corresponds to a weight of 0.25 for the linear model. Therefore, it still means that the risk criterion is twice as important, *on average*, as the benefit criterion. Since the values are different but the underlying interpretation remains the same, mapping the weights permits to provide the fairest comparison between the models.

Proof for the above workings is given in the Supplemental Material.

2.3.2 *Mapping for setting with more than two criteria*

The derivation above concerns the setting with two criteria only but could be directly extended for the product and SLoS models. Specifically, one can apply the proposed mapping function to each of the weights in the setting with more than two criteria marginally. This would imply that the weights are mapped with respect to the importance of all other criteria rather than a single benefit (or risk).[14] The extension for the multi-linear model, however, is less straightforward. Generally, it would be a much more involving procedure to elicit weights for all the interactions terms as their number increases noticeably if more than two criteria are considered. Specifically, in the case study considered in Section 3, there are four criteria resulting in 11 interaction terms. Following the two criteria setting, we suggest to fix the total weight attributed to *all* the interactions to be equal to $c = 0.2$. Then, the ML model for the setting with four criteria takes the form

$$\begin{gathered}u^{ML}(\boldsymbol{\xi_i}, \boldsymbol{w}^{ML}) := \sum_{j=1}^{n} w_j^{ML} u_j(\xi_{i,j}) + \frac{c}{2^n-n-1} \sum_{j=1,k>j}^{n} u_j(\xi_{i,j}) u_k(\xi_{i,k}) + \\ \frac{c}{2^n-n-1} \sum_{j=1,l>k>j}^{n} u_j(\xi_{i,j}) u_k(\xi_{i,k}) u_l(\xi_{i,l}) + \cdots + \frac{c}{2^n-n-1} u_1(\xi_{i,1}) u_2(\xi_{i,2}) \cdots u_n(\xi_{i,n})\end{gathered} \tag{19}$$

where the fraction $\frac{c}{2^n-n-1}$ ensures that the sum of all the interaction terms equals c and this is split equally between all interaction terms. To calculate the individual weights $w_j, j = 1, \ldots, n$, again, a mapping to the linear weights can be used. In order for the weights to sum up to 1, the transformation $w^{ML} = w^{ML} - c/n$ could be applied. For $n = 2$, this translates into the corresponding mapping in equation 18. While this procedure does not guarantee the equality of the slopes of the tangents, it, however, emphasises the potential challenge associated with the use of the multi-linear model that should be taken into account when considering it.

3 Case study

In this section, the performance of the four aggregation models is illustrated in the setting of an actual case study. This will provide an insight on how the various models perform, and what difference in the decision making they induce when applied to real-life data. The case study in question analyses the effects of two treatments (Venlafaxine and Fluoxetine) compared to a placebo, on the effects of treating depression. This study uses data from Nemeroff,[17] and expands on the studies conducted by Tervonen et al.[12] and Saint-Hilary et al.[13]

Fluoxetine and Venlafaxine are both treatments used to treat depression. Here, the benefit criterion is the treatment response (an increase from baseline score of Hamilton Depression Rating Scale of at least 50%), and the three risk criteria are nausea, insomnia and anxiety.

Table 1 shows the outcomes of the trial for the two treatments and the placebo.

For all criteria, we approximate the distributions of the event probabilities by Beta distributions $\mathcal{B}(a, b)$, with a = number of occurrences and b = (number of patients − number of occurrences) of the considered event (response or adverse event), assuming Beta(0,0) priors. We generated 100,000 samples from each distribution. These samples are then used to

Table 1. Number of events and number of patients for each criteria for Venlafaxine, Fluoxetine and Placebo.

	Venlafaxine	Fluoxetine	Placebo
Treatment response	51/96	45/100	37/101
Nausea	40/100	22/102	8/102
Insomnia	22/100	15/102	14/102
Anxiety	10/100	7/102	1/102

Table 2. Mean (95% credible interval) of the Beta posterior distributions of benefit and risk parameters and of corresponding PVFs for Venlafaxine, Fluoxetine and Placebo (with values in **bold** corresponding to those that leading to significant differences between models).

	Venlafaxine	Fluoxetine	Placebo
Treatment response			
$\xi_{i,1}$	0.52 (0.42, 0.62)	0.45 (0.35, 0.55)	0.37 (0.28, 0.46)
$u_1(\xi_{i,1})$	0.53 (0.37, 0.70)	0.42 (0.26, 0.58)	**0.28**(**0.13**, **0.44**)
Nausea			
$\xi_{i,2}$	0.40 (0.31, 0.50)	0.22 (0.14, 0.30)	0.08 (0.04,0.14)
$u_2(\xi_{i,2})$	**0.20**(**0.00**, **0.39**)	0.57 (0.40, 0.72)	0.84 (0.72, 0.93)
Insomnia			
$\xi_{i,3}$	0.22 (0.15, 0.31)	0.15 (0.09, 0.22)	0.14 (0.08, 0.21)
$u_3(\xi_{i,3})$	0.56 (0.39, 0.71)	0.71 (0.56, 0.83)	0.73 (0.58, 0.84)
Anxiety			
$\xi_{i,4}$	0.10 (0.05, 0.17)	0.07 (0.03, 0.13)	0.01 (0.00, 0.04)
$u_4(\xi_{i,4})$	0.80 (0.67, 0.90)	0.86 (0.75, 0.94)	0.98 (0.93, 1.00)

approximate the distributions of the linear partial value functions (PVFs) as defined in equation (1) for all criteria and all treatment arms, with the following most and least preferred probabilities of occurrence ξ'_j and ξ''_j:

- Most and least preferable values of $\xi'_j = 0.8$ and $\xi''_j = 0.2$ for the response.
- Most and least preferable values $\xi'_j = 0$ and $\xi''_j = 0.5$ for the adverse events.

This case study considers three different weighting combinations, which were used under the linear model by Saint-Hilary et al.[13] These sets of weights correspond to three different scenarios of the relative importance of the criteria for the stakeholders. The first scenario reflects the case when all four criteria are equally important. The second scenario corresponds to the benefit criterion having more relative importance than all risk criteria together. The third scenario can be considered as a 'safety first' scenario, in which each risk criterion has a higher weight than the benefit criterion. As discussed in Section 2.3, the weights of the criteria for the product, multi-linear and SLoS models are obtained by mapping. Note, again, that while the multi-linear model might not exactly induce the same average relative importance of the criteria, the proposed procedure suggests to control the contribution of the interaction terms in the decision at the given level of $c = 0.20$, and therefore is used for the sake of simplicity. The mapped weights for each of the three scenarios are presented in Table 3.

Three pairwise comparisons are made: Venlafaxine against Fluoxetine, Venlafaxine against Placebo and Fluoxetine against Placebo. We consider that one treatment is recommended over another if the probabilities defined in (4) or (5) are greater than $\psi = 0.8$. The probabilities of recommendations under all three scenarios and for each aggregation model are given in Table 4.

Under the first scenario with the equal weights for all criteria, the treatment with preferable risk criteria values was more likely to be recommended as the three risk criteria altogether have a greater weight than the one benefit criterion. For the comparison between Venlafaxine and Fluoxetine, the probability that Venlafaxine has better benefit–risk characteristics is around 1.7%–1.8% under all four models. For the comparison between Venlafaxine and the placebo, there is only a minor difference in the probability that Venlafaxine has better benefit–risk characteristics (<0.1% in the linear and multi-linear models, 1.6% in the product model and 3.7% in the SLoS model), not enough of a difference to change the recommendation. However, when comparing Fluoxetine to the placebo, a notable difference is observed. Under the linear and multi-linear models, the probability of Fluoxetine having the better benefit–risk characteristics is around 7%–10% (suggesting the placebo is much more preferable), whilst this rises to 37% under the product model and 47.3% under the SLoS model

Table 3. Table of mapped weights for each of the three scenarios.

	Scenario 1				Scenario 2				Scenario 3			
Model	w_1	w_2	w_3	w_4	w_1	w_2	w_3	w_4	w_1	w_2	w_3	w_4
Linear	0.25	0.25	0.25	0.25	0.58	0.11	0.15	0.15	0.18	0.28	0.25	0.29
Product	0.25	0.25	0.25	0.25	0.58	0.11	0.15	0.15	0.18	0.28	0.25	0.29
Multi-linear	0.20	0.20	0.20	0.20	0.53	0.06	0.10	0.10	0.13	0.23	0.20	0.24
SLoS	0.30	0.30	0.30	0.30	0.56	0.16	0.21	0.21	0.24	0.33	0.30	0.34

Table 4. Probability of treatment being recommended as the best treatment against another for the three pairwise comparison, using each of the four aggregation models, for each of the three weighting scenarios.

Probability of treatment being recommended as best treatment	Venlafaxine over Fluoxetine	Venlafaxine over Placebo	Fluoxetine over Placebo
		Scenario 1	
Linear	1.7%	<0.1%	7.2%
Product	1.7%	1.6%	37.0%
Multi-linear	1.7%	<0.1%	9.1%
SLoS	1.8%	3.7%	47.3%
		Scenario 2	
Linear	48.0%	64.7%	66.9%
Product	42.6%	74.9%	80.4%
Multi-linear	46.3%	63.0%	66.3%
SLoS	36.6%	72.5%	81.4%
		Scenario 3	
Linear	0.6%	0%	2.1%
Product	0.5%	0.1%	18.5%
Multi-linear	0.6%	0%	3.0%
SLoS	0.6%	0.6%	30.1%

(suggesting near-parity of treatments). This occurs due to the penalisation of low benefit criterion values for the placebo, where the 95% credible interval includes values close to zero (in bold in Table 2). These low values are harshly penalised under the product and SLoS models, as they suggest that the placebo induces no treatment benefit with a non-neglectable probability. The linear model does not account for this and strongly favours the placebo, while the multi-linear does not penalise these values strongly.

Under the second scenario, the treatment response is considered as the most important factor, and is given a weighting greater than that of the three risk criteria combined. For the comparison between Venlafaxine and Fluoxetine, both the product and SLoS models say that Venlafaxine has inferior benefit–risk characteristics (42.6% and 36.6% probability of being better, respectively). More average results are observed with both the linear model, which gives a probability of 48.0%, and the multi-linear model, which gives a probability of 46.3%. Again, the difference between the probability of the linear model and those of the product and SLoS models is due to the penalising effects of the latter. This occurs because of the nausea risk criterion interval contains zero for Venlafaxine (in bold in Table 2), which causes the product and SLoS models to recommend Fluoxetine more often than Venlafaxine, despite the weighing criteria giving preference to the treatment response (which is greater with Venlafaxine). With the multi-linear model, the penalisation of the undesirable nausea criterion is not as strong as in the product or SLoS models, as the weight mapping induces a drop from 0.11 to 0.06 in the weight given to the corresponding individual term, and the effect of the interaction terms is not enough to overcome this.

For the comparison between Venlafaxine and the placebo, the probability that Venlafaxine has better benefit-risk characteristics is between 63% and 75% across the four models. The product and SLoS models both penalise the low benefit value of the placebo, which is why they are both more likely to recommend Venlafaxine than the other two models. Additionally, the product and SLoS models both also penalise the nausea criterion value of Venlafaxine, and due to the

		Probability of benefit θ_{i1}								
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Probabiltiy of risk θ_{i2}	0.1	○■	○	○	○	○	○	○	○	○■
	0.2	○	○	○	○	○	○	○	○	○
	0.3	○	○	○■	○	○	○	○■	○	○
	0.4	○	○	○	○	○	○	○	○	○
	0.5	○	○	○	○	○■	○	○	○	○
	0.6	○	○	○	○	○	○	○	○	○
	0.7	○	○	○■	○	○	○	○■	○	○
	0.8	○	○	○	○	○	○	○	○	○
	0.9	○■	○	○	○	○	○	○	○	○■
■ = T1										
○ = T2										

Figure 3. Simulation scenarios for the trial with two criteria.

increase weighting given to it by the SLoS model mapping, this causes the product model to be more likely to recommend Venlafaxine than the SLoS model.

For the comparison between Fluoxetine and the placebo, the probability that Fluoxetine has better benefit-risk characteristics is around 65%–80% under all four models, with the probability of Fluoxetine being preferable increasing as the methods increase the penalisation applied to the placebo's lack of benefit effect. The stronger penalisation occurs under the product and SLoS models, hence why they are both more likely to recommend Fluoxetine.

Across all three comparisons, the multi-linear model is always slightly less likely to recommend the treatment with the greater benefit value than the linear model. As this is the scenario where the benefit criterion is considered to be the most important, this shows that the weight splitting with the multi-linear model induces a loss of the preferences that were given when the weights were originally set out for the linear model, illustrating some of the problems theorised in the methods section.

Under the third scenario, a 'safety first' approach is adopted, giving the risk factors a higher weighting. The probability that Venlafaxine has better benefit–risk characteristics is around 0.5%–0.6% when it is compared to Fluoxetine and around 0%–0.6% when it is compared to placebo, under all four models. For the comparison between Fluoxetine and the placebo, the probability that Fluoxetine has better benefit–risk characteristics is around 2.1%–3.0% for the linear and multi-linear models, whilst this increases to 18.5% under the product model and 30.1% under the SLoS model. This increase occurs for the same reasons outlined for the same comparison in scenario 1: The penalisation of the benefit criterion for the placebo, with its 95% credible interval including low values (in bold in Table 2). The linear model does not account for this and strongly favours the placebo, while the multi-linear does not penalise these values sufficiently and still favours the placebo.

Overall, this case study provides us with a number of important observations shedding a light on the differences in the aggregation performances. Firstly, the effects of extremely undesirable outcomes (those highlighted in bold in Table 2) are more significantly and consistently penalised in the product and SLoS models (the penalisation is stronger in the SLoS model than the product model, although they give the same recommendation for every comparison). These examples also help to show that the models provide similar recommendations when one treatment is clearly preferable than its competitor. Lastly, the weight splitting in the multi-linear model induces a change in the relative importance between criteria that may not always reflect the choices of weights as well as other models, highlighted in scenario 2. This makes it less appealing than other models.

To draw further conclusions regarding the differences between models, we conduct a comprehensive simulation study under various scenarios and under their many different realisations.

4 Simulation study

4.1 Setting

To evaluate the performances of the four aggregation models, a comprehensive simulation study covering a wide range of possible clinical cases is conducted. This allows us to investigate many scenarios and their various realisations rather than a single dataset as in the case study. The simulation study is preformed in a setting with two treatments, named T_1 and T_2, that are compared in randomised clinical trials with $N = 100$ patients allocated to each treatment. Each treatment is evaluated based on two criteria: one benefit ($j = 1$) and one risk ($j = 2$). We assume that benefit events are desirable (e.g. treatment response), while risk events should be avoided (e.g. adverse event), with θ_{ij} being their true probability of occurrence for each treatment $i = 1, 2$. The PVFs are defined as $u_1(\theta_{i1}) = \theta_{i1}$ and $u_2(\theta_{i2}) = 1 - \theta_{i2}$. The two criteria are deemed equally important and therefore are given equal weighting criteria. We start with the case of uncorrelated criteria and explore the effect of the presence of correlations in Section 4.4. The range of true values of the benefit and risk criteria and the corresponding simulation scenarios are given in Figure 3.

Figure 3 shows all the different values that the benefit and risk criteria can take for both T_1 and T_2, where black squares correspond to the pairs of criterion values for T_1 and white circles correspond to the pairs of criterion values for T_2. For each of the nine fixed characteristics of T_1, all 81 possible values of T_2, with $\theta_{2,1}, \theta_{2,2} \in (0.1, 0.2, \ldots, 0.9)$ are considered, resulting in 729 scenarios. The fixed characteristics for T_1 are referred to as follows:

Scenario 1: T_1=($\theta_{1,1}$=0.5,$\theta_{1,2}$=0.5) Scenario 2: T_1=($\theta_{1,1}$=0.3,$\theta_{1,2}$=0.7)
Scenario 3: T_1=($\theta_{1,1}$=0.7,$\theta_{1,2}$=0.3) Scenario 4: T_1=($\theta_{1,1}$=0.1,$\theta_{1,2}$=0.1)
Scenario 5: T_1=($\theta_{1,1}$=0.9,$\theta_{1,2}$=0.9) Scenario 6: T_1=($\theta_{1,1}$=0.3,$\theta_{1,2}$=0.3)
Scenario 7: T_1=($\theta_{1,1}$=0.7,$\theta_{1,2}$=0.7) Scenario 8: T_1=($\theta_{1,1}$=0.9,$\theta_{1,2}$=0.1)
Scenario 9: T_1=($\theta_{1,1}$=0.1,$\theta_{1,2}$=0.9)) where $\theta_{1,j}$ is the true value of criterion j for T_1.

4.2 Data generation and comparison procedure

The following Bayesian procedure is used for the simulation study:

- Step 1: Simulate randomised clinical trials with two treatments T_1 and T_2, each with two uncorrelated criteria, and the sample size of $N = 100$ in each treatment arm.
- Step 2: Derive the posterior distributions using the simulated data assuming a degenerate prior, Beta(0,0), to reduce the influence of the prior distribution. Draw 2000 samples from each posterior distribution of the criteria and obtain the corresponding empirical distribution for the PVF.
- Step 3: Use the posterior distributions of the PVF in each of the aggregation models as given in equations 2 and 3 to compute the probability in equations 4 and 5 that treatment T_1 has better benefit–risk profile, $\mathcal{P}_X^{1,2}$ (for some model X), and compare to the threshold value $\psi = 0.8$. If $\mathcal{P}_X^{1,2} > 0.8$, then treatment T_1 is recommended. If $\mathcal{P}_X^{1,2} < 0.2$, then treatment T_2 is recommended. If $0.2 \leq \mathcal{P}_X^{1,2} \leq 0.8$, then neither treatment is recommended.
- Step 4: Repeat steps 1–3 for 2500 simulations trials.
- Step 5: Estimate the probability that each treatment is recommended ($\mathbb{P}(\mathcal{P}_X^{1,2} > 0.8)$) by its proportion over 2500 simulated trials.

The aggregation models will be compared using $\mathbb{P}(\mathcal{P}_X^{1,2} > 0.8)$, which is the probability that the model X recommends T_1 over T_2, and $\phi_{X-Y} = \mathbb{P}(\mathcal{P}_X^{1,2} > 0.8) - \mathbb{P}(\mathcal{P}_Y^{1,2} > 0.8)$, which is the difference between the probability that the model X recommends T_1 and the probability that the model Y recommends T_1. The value of ϕ represents a difference between two probabilities, and can therefore take the range of values $-1 \leq \phi \leq 1$. If $0 < \phi \leq 1$, then the model X recommends T_1 more often than model Y. If $-1 \leq \phi < 0$, then the model Y recommends T_1 more often than model X. If $\phi = 0$, then the two models make the recommendations with the same probability. Note that, for the ML model, we adopt $c = 0.20$ as in the case study above.

4.3 Results

The results are presented on Figures 4 and 5. The first seven scenarios referred to above for treatment T_1 are presented in the rows labelled 1–7. Each graph corresponds to fixed expected probabilities of event for treatment T_1, and each cell corresponds to a combination of expected probabilities of benefit and risk for T_2. When reference is made to the ‘diagonal’, this refers to the diagonal line that runs from the bottom left corner of the graph to the top right. In all scenarios, all models agree to recommend T_1 when it is undoubtedly better than T_2, i.e., when T_1 is more effective and less harmful than T_2 (or to recommend T_2 when T_1 is indisputably worse, i.e., less effective and more toxic). For this reason, the results for scenarios 8 and 9 are not presented here, but are included in the Supplemental Material for completeness.

The probabilities $\mathbb{P}(\mathcal{P}_L^{1,2} > 0.8)$ (red), $\mathbb{P}(\mathcal{P}_P^{1,2} > 0.8)$ (purple), $\mathbb{P}(\mathcal{P}_{ML}^{1,2} > 0.8)$ (orange) and $\mathbb{P}(\mathcal{P}_S^{1,2} > 0.8)$ (blue) are shown in Figure 4, and all six pairwise comparisons in these probabilities are given in Figure 5. From left to right, Figure 5 shows ϕ_{P-L}, ϕ_{ML-L}, ϕ_{ML-P}, ϕ_{S-L}, ϕ_{S-P} and ϕ_{S-ML}.

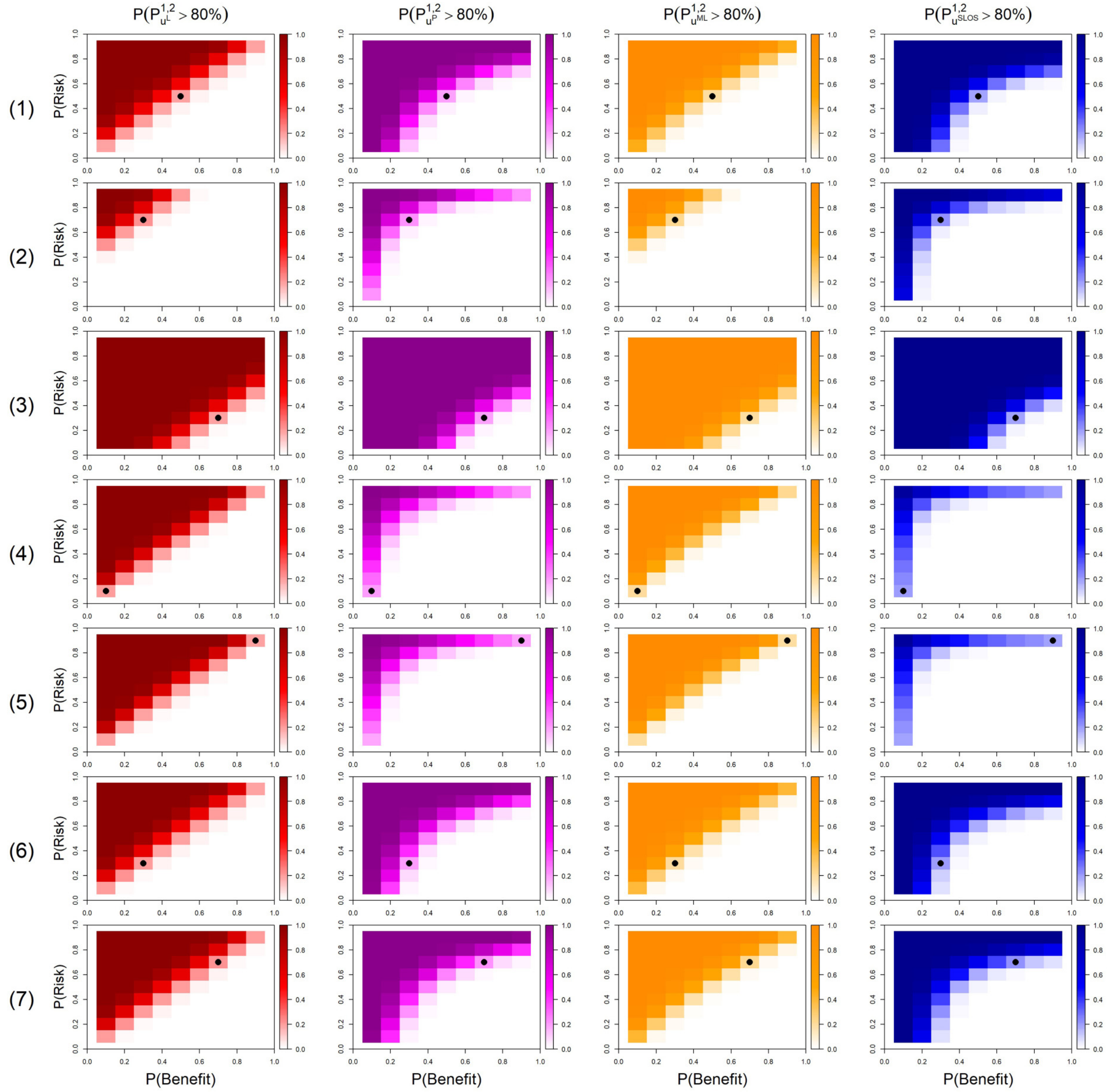

Figure 4. Probability that the model recommends T_1 over T_2, $\mathbb{P}(\mathcal{P}^{1,2} > 0.8)$, for scenarios 1 to 7 for the linear (red), product (purple), multi-linear (orange) and SLoS (blue) models.

In Figure 5, a colour of a cell corresponds to the aggregation model of this colour to recommend treatment T_1 with higher probability than another method. For instance, red cells in the first column of Figure 5 showing (ϕ_{P-L}) indicate that, when T_2 characteristics take the corresponding value, the linear model recommends T_1 more often than the product one.

In scenario 1, the four models are in agreement to recommend T_1 when T_2 corresponds to less benefit and more risk. On the diagonal, the product and SLoS models both favour T_1 over T_2 when T_2 has either extremely high benefit and risk (top right corner), or extremely low benefit and risk (left bottom corner), compared to either the linear or multi-linear models. This occurs due to the penalisation of extremely low benefit and extremely high risk by the product and SLoS models. Comparing product and SLoS models for these values of benefit–risk, SLoS favours T_1 over T_2 more often for low but not boundary values of the criteria. This occurs due to the SLoS model penalising the undesirable qualities more than the product one (this is similar to trends observed in the case study). Compared to the linear model, the multi-linear model recommends T_1 over T_2 with higher probability when T_2 has either higher benefit and higher risk, or lower benefit and lower risk due to the interaction term providing mild penalisation of extremely high risk or extremely low

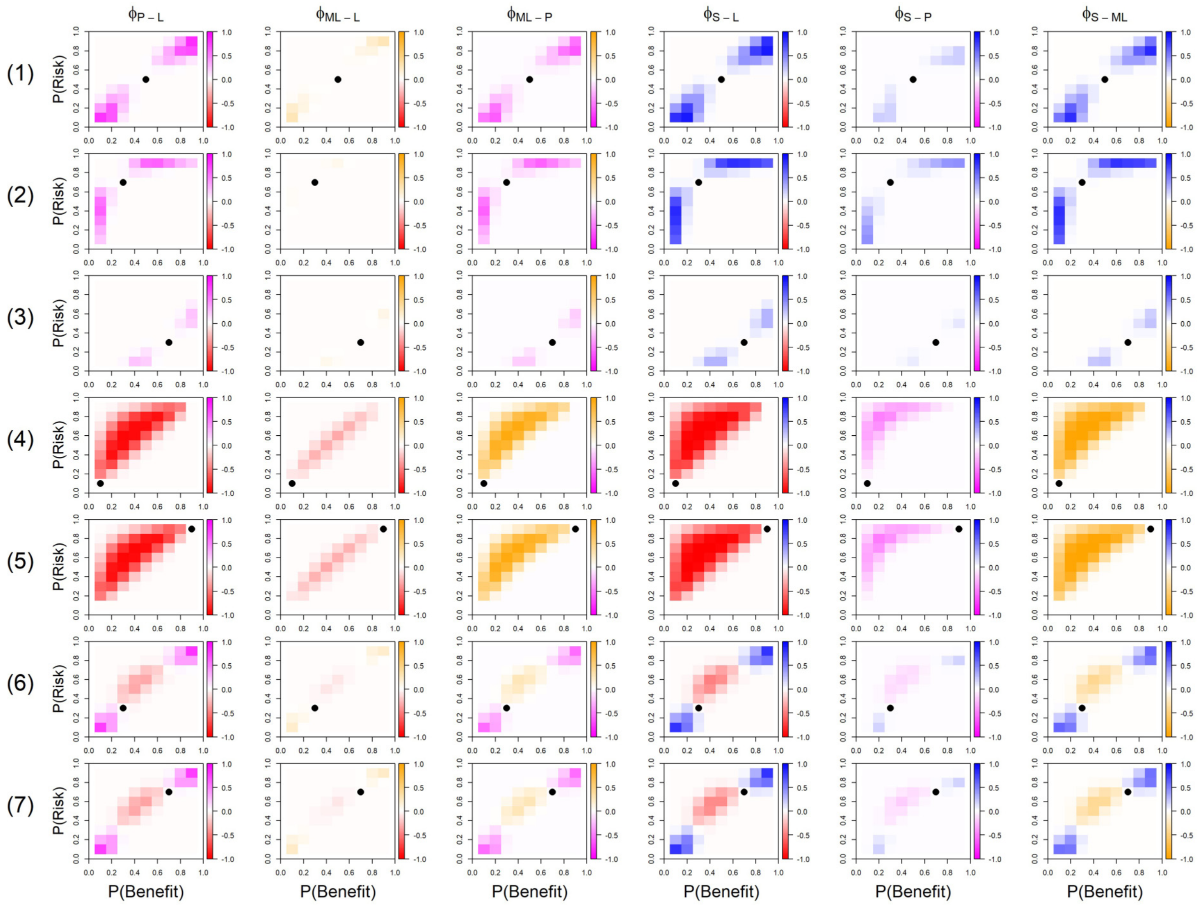

Figure 5. Results of the six pairwise comparisons of the four AM, where a cell being a colour indicates that AM recommended T_1 more than the comparative AM (the deeper the colour, the greater the difference in recommendation).

benefit. However, there is (in most cases), a greater magnitude of difference between the SLoS and product models than between the linear and multi-linear models.

For example, when T_2 has criteria values $\theta_{2,1} = 0.2$ (benefit), $\theta_{2,2} = 0.1$ (risk) (lower benefit, lower risk), T_1 is recommended in 2% of the trials under the linear model, in 70% under the product, in 8% under the multi-linear and in 90% under SLoS. This tells us that the product and SLoS models do not permit that the decrease in risk is worth the decrease in benefit that comes with it (the SLoS model more than the product model), whilst the linear and multi-linear models both consider it acceptable. Considering the case when $\theta_{2,1} = 0.7$, $\theta_{2,2} = 0.7$ (higher benefit and higher risk compared to T_1), T_1 is recommended in 20% of the trials under the linear model compared to 49% for product model, 25% for the multi-linear model and 61% for SLoS model. This tells us that the product and SLoS models do not permit that the increase in benefit is worth the increase in risk that comes with it (again, this effect is stronger in the SLoS model than the product model), whilst the linear and multi-linear both consider it acceptable (the linear model more-so than the multi-linear model). Similar observations can be made in scenarios 2 and 3.

However, a distinguishing difference between the designs under scenario 1 can be found when T_2 has the criteria $\theta_{2,1} = 0.9$, $\theta_{2,2} = 0.7$. In this comparison, T_1 is recommended in 0% of the trials under the linear model compared to 11% for product model, 0% for the multi-linear model and 30% for SLoS model. Meanwhile, T_2 is recommended in 92% of the trials under the linear model compared to 32% for product model, 84% for the multi-linear model and 13% for SLoS model. This shows that the linear, product and multi-linear models are all more likely to recommend T_2, whilst only the SLoS model is more likely to recommend T_1. This occurs due to the different strengths of penalisation between the models, and only the SLoS model does not consider this an acceptable trade-off. This shows that the product model and the SLoS model do not always make the same recommendations, and that these differences can sometimes be quite large.

In scenario 4, where T_1 has extremely low benefit and risk, it is very rarely recommended by either the product of SLoS models, whereas it recommended by both the linear and multi-linear models, in cases where T_2 has some increase in benefit, but a higher increase in risk. This occurs because the SLoS and product models penalise extremely low benefit so severely that the level of risk has almost no impact on the recommendation. The multi-linear model also penalises the extreme low benefit, but on a much smaller scale. For example, for T_2 with criteria values $\theta_{2,1} = 0.6$, $\theta_{2,2} = 0.7$, T_1 is recommended with probability 68% under the linear model, never recommended under the product model, 41% under the multi-linear model and never recommended under the SLoS model. This shows that the product and SLoS models reflect the desirable properties outlined above: that we are not interested in the risk criterion value of a treatment if the benefit criterion value is small/zero, whilst both the linear and multi-linear models do not reflect this (although the multi-linear model does somewhat penalise this). Similar results are observed in scenario 5, where T_1 has extreme risk and extreme benefit. The SLoS and product models will recommend T_2 if it has lower risk than T_1 as long as it has some benefit, whereas the linear model and the multi-linear model will recommend T_1 over T_2 if the benefit of T_2 decreases by a greater amount than the risk.

It should be noted that poor recommendations can be made under the product and SLoS models if both T_1 and T_2 have a risk criterion value of 0.9, as the strength of the penalisation of the undesirable criteria overpowers the effect of the benefit. For example, in scenario 5 where T_2 has criteria values $\theta_{2,1} = 0.8$, $\theta_{2,2} = 0.9$ (same risk criterion value as T_1 but a lower benefit criterion value), T_1 is recommended with probability 75% under the linear model, 27% under the product model, 68% under the multi-linear model and 23% under the SLoS model (this effect is stronger in the SLoS model than in the product model due to its harsher penalisation of the undesirable criteria). They both did recommend T_2 with probabilities 13% and 17%, respectively, showing that they still recommend the better treatment T_1 more often than T_2, but that these two models hardly discriminate very unsafe drugs (for comparison, both the linear and multi-linear models only recommended T_2 with probability 1% each).

In scenarios 6 and 7, all AM recommend T_1 over T_2 when T_2 is unarguably worse (similarly they all recommend T_2 over T_1 when T_1 is unarguably worse). Along the diagonal, the SLoS model recommends T_1 over T_2 more often than the other AM when T_2 has either extreme low benefit and extreme low risk, or extreme high benefit and extreme high risk, compared to T_1 (although the product model recommends T_1 only a slightly smaller proportion of times than the SLoS model). Again, this is the result of the penalisation of extremely low benefit or extremely high risk criteria. Similarly, the multi-linear model recommends T_1 over T_2 more often than the linear model in the same circumstances. For example, in Scenario 6, when T_2 has criteria values $\theta_{2,1} = 0.2$, $\theta_{2,2} = 0.2$ (lower benefit and lower risk), T_1 is recommended with probability 21% under the linear model, 59% under the product model, 28% under the multi-linear model and 68% under the SLoS model. This shows how the different levels of penalisation affect the recommendations, where the stronger the penalisation of the undesirable low benefit criterion value, the more likely an AM is to recommend T_1, and is the reason why there is such a large difference between the linear and SLoS models recommendations.

Overall, the simulation study has shown that, for the two criteria having an equal relative importance, SLoS penalises extremely low benefit and extremely high risk criteria the most, whilst the product model penalises these moderately, acting as a sort of middle ground between the linear and SLoS models. The multi-linear model offers a small amount of penalisation (less than the product model), but due to the added complexity of this model when more criteria are added, it should not be recommended over either the SLoS model or the product model. The linear and multi-linear models both recommend treatments with no benefit/high risk over other viable alternatives, which contradicts conditions set out by Saint-Hilary et al.[14] Therefore we can provisionally conclude that the two models that appeal most at this point are the product and SLoS models.

4.4 Sensitivity analysis: Correlated criteria

The results above concerned the case with the two criteria being uncorrelated. However, it might be reasonable to assume that the criteria for one treatment might be correlated. In this section, we study how robust the recommendation by each of the four models are to the correlation between the benefit and risk criteria. We consider two cases of the correlation: a strong positive correlation ($\rho = 0.8$) and a strong negative correlation ($\rho = -0.8$) between the criteria. The correlated outcomes were generated using a procedure laid out in Mozgunov et al.[26]

We study how likely the correlated outcomes are to change the final recommendation of one of the treatments. Specifically, we study the proportion of cases under each of the scenarios in which the difference in the probability of recommending treatment T_1, $\mathbb{P}(\mathcal{P}_X^{1,2} > 0.8)$, changes by more than 2.5% and by 5%. Table 5 shows the number of cases (out of 81) under each of nine scenarios, in which the differences in the probabilities to recommend T_1 over T_2 changes by at least 2.5% and 5% comparing the positively correlated and uncorrelated criteria. The case investigating the effects of negative correlation shows similar results to those presented here, and is included in the Supplemental

Table 5. Number of times (%) when the difference in recommending T_1 changes by at least 2.5% or 5% between the positively correlated criteria and the non-correlated criteria.

		Linear model	Product model	Multi-linear model	SLoS model
Scenario 1	≥2.5%	30/81 (37.0%)	24/81 (29.6%)	29/81 (35.8%)	22/81 (27.2%)
	≥5%	22/81 (27.2%)	15/81 (18.5%)	21/81 (25.9%)	12/81 (14.8%)
Scenario 2	≥2.5%	10/81 (12.3%)	15/81 (18.5%)	15/81 (18.5%)	12/81 (14.8%)
	≥5%	5/81 (6.2%)	3/81 (3.7%)	5/81 (6.2%)	3/81 (3.7%)
Scenario 3	≥2.5%	16/81 (19.8%)	13/81 (16.0%)	17/81 (21.0%)	14/81 (17.3%)
	≥5%	6/81 (7.4%)	5/81 (6.2%)	4/81 (4.9%)	4/81 (4.9%)
Scenario 4	≥2.5%	22/81 (27.2%)	3/81 (3.7%)	22/81 (27.2%)	0/81 (0%)
	≥5%	17/81 (21.0%)	0/81 (0%)	15/81 (18.5%)	0/81 (0%)
Scenario 5	≥2.5%	23/81 (28.4%)	4/81 (4.9%)	22/81 (27.2%)	0/81 (0%)
	≥5%	17/81 (21.0%)	0/81 (0%)	15/81 (18.5%)	0/81 (0%)
Scenario 6	≥2.5%	29/81 (35.8%)	21/81 (25.9%)	30/81 (37.0%)	17/81 (21.0%)
	≥5%	20/81 (24.7%)	12/81 (14.8%)	16/81 (19.8%)	4/81 (4.9%)
Scenario 7	≥2.5%	25/81 (30.9%)	19/81 (23.5%)	24/81 (29.6%)	13/81 (16.0%)
	≥5%	14/81 (17.3%)	9/81 (11.1%)	15/81 (18.5%)	2/81 (2.5%)
Scenario 8	≥2.5%	0/81 (0%)	0/81 (0%)	0/81 (0%)	0/81 (0%)
	≥5%	0/81 (0%)	0/81 (0%)	0/81 (0%)	0/81 (0%)
Scenario 9	≥2.5%	1/81 (1.2%)	1/81 (1.2%)	1/81 (1.2%)	1/81 (1.2%)
	≥5%	0/81 (0%)	0/81 (0%)	0/81 (0%)	0/81 (0%)
Total	≥2.5%	156/729 (21.4%)	100/729 (13.7%)	160/729 (21.9%)	79/729 (10.88%)
	≥5%	101/729 (13.9%)	91/729 (12.5%)	44/729 (6.0%)	25/729 (3.4%)

Material. For example, the first entry in Table 5 shows that in 37% cases under scenario 1, the probability to recommend T_1 changes by at least 2.5% if the linear model is used.

Table 5 shows that all four models are the most affected by correlation under scenario 1 with the characteristics of T_1 being in the middle of the unit interval. This effect is, however, less prominent for the product and SLoS models. At the same time, under scenarios 2 to 7, the correlation has a larger effect on the linear and multi-linear models than on the other two models. Scenarios 8 and 9 are hardly affected by any correlation, and the effect is similar across all four models.

Overall, the SLoS model is the least affected by correlation between the criteria, the product model is the second least affected whereas the multi-linear (for the threshold 2.5%) and the linear model (for the threshold 5%) are the most affected ones.

5 Discussion

In this article, four potential AM are investigated for use in benefit–risk analyses: The linear model, product model, multi-linear model and the SLoS model. The differences of these models were highlighted in a case-study and a simulation study.

In most clear cases (i.e. when one treatment has more benefit and less risk than the competitor), all AM gave similar recommendations. However, in cases where one treatment had either no benefit or extreme risk, the models which penalised undesirable values more (the product and SLoS models) gave more desirable recommendations: non-effective or extremely unsafe treatments are never recommended. Furthermore, with these models, more risk is accepted in order to increase benefit when the amount of benefit is small than when it is high (or less benefit is desirable to reduce risk when the amount of risk is high than when it is small), which is consistent with the well established assumption of non-linearity of human preferences.[20] It should be noted that these models hardly discriminate two treatments that slightly differ but have both extremely undesirable properties. However, in this case, none of the treatments should be recommended anyway.

The effects of correlations between criteria was also investigated in this study. The overall effect of correlations was small to negligible in the product and SLoS models, showing these AM are not much affected by correlations between the criteria. However, the linear and multi-linear models were more likely to see a 2.5% or 5% change in the probability of recommending one treatment over another, showing that they are more affected by correlations between the criteria.

A simple mapping was applied to obtain multi-linear and SLoS weights from linear weights, so that the models could be fairly compared while preserving the weight interpretation. However, since the mapping is not far from an identity transformation, omitting it would not have a major impact on the results, as demonstrated in Saint-Hilary et al. 2018[14] for SLoS.

Overall, the two models to recommend from this investigation are the product model and the SLoS model, depending on how severely the decision maker whish to penalise treatments with either no benefit or extreme risk (moderate penalisation: product model, strong penalisation: SLoS model). The multi-linear model, whilst acting as a middle ground between the linear model and the product and SLoS models in the simulation study, involves an increased complexity behind the model. These include the increased complexity involved with adding additional terms and increased difficulty in weight mapping. This model also struggled to truly reflect the weightings given in the case study, especially in scenario 2. Because of this, we do not recommend this AM over the product or SLoS models. Additionally, the linear and multi-linear models should not be recommended as both of these models do not contain the two desirable properties outlined in Saint-Hilary et al.[14]: That treatments with no benefit/extreme risk should not be recommended, and that a larger increase in risk is accepted in order to increase the benefit if the benefit is small compared to if the benefit is high – both of which are present in the product and SLoS models.

Acknowledgements

This report is independent research supported by the National Institute for Health Research (NIHR Advanced Fellowship, Dr Pavel Mozgunov, NIHR300576). The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health and Social Care (DHCS). Dr Pavel Mozgunov also received funding from UK Medical Research Council (MC UU 00002/14).

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Tom Menzies https://orcid.org/0000-0003-3896-7228
Gaelle Saint-Hilary https://orcid.org/0000-0003-1643-3348
Pavel Mozgunov https://orcid.org/0000-0001-6810-0284

Supplemental material

Supplemental material available at: https://github.com/Tom-Menzies/Code-Menzies-2020.

References

1. Chuang-Stein C, Entsuah R and Pritchett Y. Measures for conducting comparative benefit: Risk assessment. *Drug Inf J* 2008; **42**: 223–233.
2. EMA. Guidance document on the content of the <co->rapporteur day 80 critical assessment report. https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/day-80-assessment-report-clinical-guidance_en.pdf (2013, accessed 30th August 2019).
3. FDA. Providing postmarket periodic safety reports in the ICH E2C(R2) format. *U.S. Food and Drug Administration* https://www.fda.gov/regulatory-information/search-fda-guidance-documents/providing-postmarket-periodic-safety-reports-ich-e2cr2-format-periodic-benefit-risk-evaluation?source=govdelivery, journal=U.S. Food and Drug Administration (2013, accessed 30th August 2019).
4. Hughes D, Bayoumi A and Pirmohamed M. Current assessment of risk-benefit by regulators: Is it time to introduce decision analyses?. *Clin Pharmacol Ther* 2007; **82**: 123–127.
5. Thokala P, Devlin N, Marsh K, et al. Multiple criteria decision analysis for health care decision making-an introduction: Report 1 of the ISPOR MCDA emerging good practices task force. *Value Health* 2016; **19**: 1–13.
6. Marsh K, IJzerman M, Thokala P, et al. Multiple criteria decision analysis for health care decision making-emerging good practices: Report 2 of the ISPOR MCDA emerging good practices task force. *Value Health* 2016; **19**: 125–137.
7. Mussen F, Salek S and Walker S. A quantitative approach to benefit–risk assessment of medicines – part 1: the development of a new model using multi-criteria decision analysis. *Pharmacoepidemiol Drug Saf* 2007; **16**: S2–S15.
8. IMI PROTECT. Work package 5: Benefit-risk integration and representation. http://www.imi-protect.eu/wp5.shtml (2013, accessed 30th August 2019).
9. NICE Decision Support Unit. Multi-criteria decision analysis (MCDA), http://scharr.dept.shef.ac.uk/nicedsu/wp-content/ uploads/sites/7/2016/03/MCDA-for-HTA-DSU.pdf (2011, accessed 1 February 2020).

10. Marsh K, Lanitis T, Neasham D, et al. Assessing the value of healthcare interventions using multi-criteria decision analysis: a review of the literature. *PharmacoEconomics* 2014; **32**: 345–365.
11. Waddingham E, Mt-Isa S, Nixon R, et al. A Bayesian approach to probabilistic sensitivity analysis in structured benefit-risk assessment. *Biom J* 2016; **58**: 28–42.
12. Tervonen T, van Valkenhoef G, Buskens E, et al. A stochastic multicriteria model for evidence-based decision making in drug benefit-risk analysis. *Stat Med* 2011; **30**: 1419–1428.
13. Saint-Hilary G, Cadour S, Robert V, et al. A simple way to unify multicriteria decision analysis (MCDA) and stochastic multicriteria acceptability analysis (SMAA) using a dirichlet distribution in benefit-risk assessment. *Biom J* 2017; **59**: 567–578.
14. Saint-Hilary G, Robert V, Gasparini M, et al. A novel measure of drug benefit–risk assessment based on scale loss score. *Stat Methods Med Res* 2018; **1**: 2738–2753.
15. Morton A. Treacle and smallpox: two tests for multicriteria decision analysis models in health technology assessment. *Value Health* 2017; **20**: 512–515.
16. Marsh KD, Sculpher M, Caro J, et al. The use of MCDA in HTA: great potential, but more effort needed. *Value Health* 2017; **21**: 394–397.
17. Nemeroff C. Prevalence and management of treatment-resistant depression. *J Clin Psychiatry* 2007; **68**: 17–25.
18. Marcelon L, Verstraeten T, Dominiak-Felden G, et al. Quantitative benefit-risk assessment by MCDA of the quadrivalent HPV vaccine for preventing anal cancer in males. *Expert Rev Vaccines* 2016; **15**: 139–148.
19. Nixon R, Dierig C, Mt-Isa S, et al. A case study using the PrOACT-URL and BRAT frameworks for structured benefit risk assessment. *Biometric J* 2016; **58**: 8–27.
20. Berger JO. *Statistical decision theory and Bayesian analysis*. 2nd ed. New York, NY: Springer-Verlag, 1985.
21. Morton A. Treacle and smallpox: two tests for multicriteria decision analysis models in health technology assessment. *Value Health* 2017; **20**: 512–515.
22. Raiffa H and Keeney RL. Decision analysis with multiple conflicting objectives, preferences and value tradeoffs. http://pure.iiasa.ac.at/id/eprint/375/ (1975, accessed 30th June 2019).
23. Cobb C and Douglas P. A theory of production. *Am Econ Rev* 1928; **18**: 139–165.
24. Tervonen T, van Valkenhoef G, Buskens E, et al. A stochastic multicriteria model for evidence-based decision making in drug benefit-risk analysis. *Statist Med* 2011; **30**: 1419–1428.
25. Broekhuizen H, IJzerman MJ, Hauber AB et al. Weighing clinical evidence using patient preferences: An application of probabilistic multi-criteria decision analysis. *PharmacoEconomics* 2017; **35**: 259–269.
26. Mozgunov P, Jaki T and Paoletti X. A benchmark for dose finding studies with continuous outcomes. *Biostatistics* 2018; **59**: 567578.