Title: Treatment recommendations based on Network Meta-Analysis: rules for risk-averse decision-makers

Running Title: Treatment recommendations for the risk-averse

Authors: A E Ades[1], Hugo Pedder[1], Annabel L Davies[1], H Thom[1] , David M Phillippo[1], Beatrice Downing[1], Deborah M Caldwell[1], Nicky J Welton[1].

Author Affiliations: [1] Population Health Sciences, Bristol University Medical School, Bristol, United Kingdom

Corresponding Author: A E Ades, PhD, Population Health Sciences, Bristol University Medical School, 37 Whately Road Bristol, BS2 8PS, United Kingdom.

t.ades@bristol.ac.uk

(44)-7879-401-276

Risk-averse recs
June 28 2024

ABSTRACT

Background: The treatment recommendation based on a Network Meta-analysis (NMA) is usually the single treatment with the highest Expected Value (EV) on an evaluative function. We explore approaches which recommend multiple treatments and which penalize uncertainty, making them suitable for risk-averse decision makers.

Methods: We introduce Loss-adjusted EV (LaEV) and compare it to GRADE and three probability-based rankings. We define the properties of a valid ranking under uncertainty and other desirable properties of ranking systems. A two-stage process is proposed: the first selects treatments superior to the reference treatment; the second identifies those that are also within a Minimal Clinically Important Difference (MCID) of the best treatment. Decision rules and ranking systems are compared on stylized examples and 10 NMAs used in NICE Guidelines.

Results: Only LaEV reliably delivers valid rankings under uncertainty and has all the desirable properties. In 10 NMAs comparing between 4 and 40 treatments, an EV decision maker would recommend 4-14 treatments, and LaEV 0-3 (median 2) fewer. GRADE rules give rise to anomalies, and, like the probability-based rankings, the number of treatments recommended depends on arbitrary probability cutoffs. Among treatments that are superior to the reference, GRADE privileges the more uncertain ones, and in 3/10 cases GRADE failed to recommend the treatment with the highest EV and LaEV.

Conclusions: A two-stage approach based on MCID ensures that EV- and LaEV-based rules recommend a clinically appropriate number of treatments. For a risk-averse decision maker, LaEV is conservative, simple to implement, and has an independent theoretical foundation.

Highlights

What is already known?

A risk-neutral decision-maker should make treatment decisions based on Expected Value (EV), meaning that the single treatment with the highest expected efficacy from a network meta-analysis should be recommended, regardless of uncertainty. In practice, decision makers may recommend several treatments, and take uncertainty into account on an *ad hoc* basis.

What is new?

We introduce Loss-adjusted EV (LaEV) as a mechanism for risk-averse decision making, and set out desirable properties of ranking systems. We define a ranking as valid under uncertainty if a higher EV is ranked above a lower one at the same uncertainty and a lower uncertainty above a higher one at the same EV. We compare LaEV to GRADE and probabilistic rankings. Of the methods examined, only LaEV provides a valid ranking under uncertainty and has all the desirable properties.

Implications

For a risk-averse decision maker, LaEV is a reliable, conservative, and easy-to-implement decision metric, with an independent theoretical foundation. Adoption of a risk-averse stance might focus attention on more accurate quantification of uncertainty, and encourage generation of better quality evidence.

Keywords. Network meta-analysis; decision-making; loss-adjustment; expected value; treatment ranking; GRADE.

1. INTRODUCTION

In decision theory a risk-neutral decision-maker bases their recommendations on the 'Expected Value' (EV) of a chosen evaluation function, without consideration of uncertainty in this. The evaluation function could be:

(i) a measure of treatment efficacy, for example probability of an event estimated from a network meta-analysis (NMA).
(ii) Net Benefit,[1] which is monetized lifetime health gain minus lifetime costs.
(iii) or any function of health improvements and adverse events, such as Multi-Criteria Decision Analysis.[2]

The choice of EV as a decision metric is based on a substantial statistical literature[3-6] going back to the 17th century.[7] In health economic evaluations EV is regarded as optimal at a societal level[8] as it delivers a maximally efficient allocation of resources, known as Pareto-optimality.

Faced with multiple options, an EV decision maker should therefore recommend the single treatment with the highest EV, regardless of uncertainty.[9] In this sense, the EV decision maker is 'risk-neutral'. In practice, however, decision makers often recommend multiple treatments, and are influenced by the degree of uncertainty in the evidence, suggesting that they are acting as risk-averse decision-makers who have a preference for more certain outcomes. In the UK, for example, multiple treatments have been recommended by NICE (National Institute of Health and Care Excellence) in both Multiple Technology Assessments,[10,11] and more often in clinical guidelines.[12-14] This seems to be done on an *ad hoc* basis usually when treatments have similar efficacy, reflecting a desire to keep clinical options open in case of patient differences in efficacy or side effects, factors that are seldom included in the formal decision model.

Uncertainty has also been treated in an *ad hoc* and even ambiguous manner in NICE's official documents. The 2022 NICE manual for health technology evaluation (Section 6.3.5) requires that "the degree of certainty or uncertainty around the ICER" (Incremental Cost-Effectiveness Ratio) be taken into account.[15] The general intention is that less should be paid for an uncertain technology (Section 6.2.34), representing a 'risk-averse' approach. However, if it is considered that better evidence is unlikely to be forthcoming, NICE may set a *higher* ICER threshold: this is regarded as appropriate in Highly Specialised Technology evaluations for rare diseases (Section 7.1). In this case more is paid for the more uncertain technology, representing a 'risk-seeking' stance. Thus, while the general decision-making position in NICE guidance is risk-neutral EV, the behaviour of NICE committees, and NICE's own documentation, departs from EV in *ad hoc* and seemingly unprincipled ways.

Uncertainty in treatment rankings has also attracted the attention of NMA methodologists.[16-19] Besides ranking by EV, properties of alternative 'treatment hierarchies', or treatment rankings, have been examined formally,[20] including: the probability of having the highest value, Pr(Best); the proportion of competitors that a treatment is superior to, also known SUCRA (Surface Under the Cumulative Ranking curve),[21] or its equivalent the P-Score.[22] The probability that the value of the evaluative function exceeds a certain threshold, abbreviated here as Pr(V>T), has also been studied.[23,24]

It has been proposed that these and other[25] probability-based metrics, which, unlike EV, take uncertainty into account, could help guide NMA treatment decisions.[20,26,27] However, by themselves ranking metrics do not define how many – or even if any – of the top-ranked treatments should be recommended. In an EV context, this can be addressed by a two-stage approach, suggested in earlier work on threshold analysis.[28] The first stage identifies treatments that are superior to a standard reference treatment, the second selects all those that are also within a Minimal Clinically Important Difference (MCID) margin of the best treatment. The GRADE Working Group adopted a similar scheme: in Stage 1 it picks out treatments where Pr(V>T) exceeds a standard probability

criterion such as 0.975.[29] Stage 2 identifies a subset of these treatments none of which are better than any other, on the same criterion.

It has been said that "each ranking metric … answer[s] a specific treatment hierarchy question, and … every ranking metric provides a valid treatment hierarchy for the corresponding question,"[20] a sentiment repeated in subsequent papers.[24,27,30] However, any number of rankings and decision schemes could be proposed: we therefore need to ask: what are the properties that would make a ranking "valid" under uncertainty? And what is the "treatment hierarchy question" that decision makers *should* be trying to answer? After all, both Pr(Best) and SUCRA can have the perverse effect of privileging treatments with more uncertain effects.[18-20]

In this paper we attempt to identify and evaluate an alternative to EV which provides a rational approach to multiple treatment recommendations, and at the same time penalizes uncertainty. We will propose a metric based on Bayesian statistical decision theory,[31,32] in which the Expected Loss arising from taking a decision under uncertainty is subtracted from the EV: we call this Loss-adjusted Expected Value (LaEV).

We begin by defining three ranking and three decision methodologies, and illustrate their properties through stylized examples. We define the validity of ranking systems under uncertainty, and suggest some desirable properties. The methodologies are then applied to ten NMAs conducted by NICE guideline developers and published in NICE guidelines and associated publications.

2. METHODS: DECISION RULES AND RANKING SYSTEMS

In this section we outline a range of existing decision rules and ranking systems, and propose a new metric, LaEV. We begin by defining the standard risk-neutral EV approach, and a 2-stage extension that allows for multiple recommendations. We then define the GRADE method for a 'minimally contextualised framework',[29] followed by LaEV. Finally, we define three probability-based ranking systems, all familiar from previous literature, and present a 2-stage version of these so that they can be used as decision rules, and to facilitate comparisons with the other methods.

2.1 The NMA model and its relation to decision outcomes

We assume a standard reference treatment 1, and 'new' treatments $2...k...K$. The NMA estimates $(K-1)$ relative treatment effect parameters: $\{\delta_2...\delta_k...\delta_K\}$ defined on the linear predictor scale. Given an estimate of the outcome on the reference treatment, μ, we can obtain estimates of the absolute efficacy: $\mu+\delta_k$ for all K treatments. By convention $\delta_1=0$. We assume here that the joint probability distribution of these parameters is informed by a Bayesian or frequentist NMA, along with a baseline model for the target population.[37] The link function $H(\mu,\delta_k)$ maps the linear predictor parameters onto values on the natural scale. For example, for a continuous outcome $H(\mu,\delta_k)=\mu+\delta_k$, and for a probability outcome $H(\mu,\delta_k)=\text{logit}^{-1}(\mu+\delta_k)$.

2.2. Risk-neutral Expected Value

2.2.1 Expected Value (Stage 1)

At Stage 1 $F_1(\mu,\delta_k)=H(\mu,\delta_k)-H(\mu,\delta_1)$ reflects differences between each treatment and the standard treatment 1 on the natural scale. Treatments are selected which are expected to be better than treatment 1, that is if $EV_1(k)>0$, where $EV_1(k)=E_{\mu,\delta_k}[F_1(\mu,\delta_k)]$**.** The best treatment is $k^*_{EV}=Argmax_k\{EV_1(\mu,\delta_k)\}$. Note that throughout we assume that $F_1()$ measures the "good"

outcome, so that $F_1(\mu,\delta_{k^*_{EV}}) > F(\mu,\delta_k) > F(\mu,\delta_1)$. All treatments where $EV_1(k) > 0$ are then considered in Stage 2.

2.2.2 Expected Value (Stage 2)

The Stage 2 evaluative function compares each treatment to the best treatment and sets the difference against a threshold, T: $F_2(\mu,\delta_{k^*_{EV}},\delta_k) = T - \left(F_1(\mu,\delta_{k^*_{EV}}) - F_1(\mu,\delta_k)\right)$. This function increases to a maximum value at T as treatment k approaches the best treatment in efficacy, and becomes negative if k is worse than k^*_{EV} by more than T. T is therefore the maximum amount by which k can be inferior to k^*_{EV} and still be recommended. The evaluative metric is the expectation over all uncertain parameters:

$$EV_2(k) = E_{\mu,\delta_{k^*_{EV}},\delta_k}\left[T - \left(F_1(\mu,\delta_{k^*_{EV}}) - F_1(\mu,\delta_k)\right)\right]$$

The decision rule is: adopt k if $EV_1(k) > 0$ and $EV_2(k) > 0$.

The MCID is a natural choice for T, as suggested in earlier work on threshold analysis[28] and by the GRADE Working Group.[29]

For probability outcomes, the threshold MCID might be expressed as the maximum Relative Risk by which k^*_{EV} can exceed k. Under these circumstances, $F_1(\mu,\delta_{k^*_{EV}}) - F_1(\mu,\delta_1) = RR\left(F_1(\mu,\delta_k) - F_1(\mu,\delta_1)\right)$, and therefore $T = \left(1 - \frac{1}{RR}\right)H(\mu,\delta_{k^*_{EV}})$. Values like 1.25 would be typical.[33] Note that the estimated expected value of $H(\mu,\delta_{k^*_{EV}})$ is treated as a constant for this purpose. Equivalent transformations will be required if the MCID is expressed as a hazard ratio, odds ratio, log odds ratio, or probit difference.

2.3 Loss-adjusted Expected Value

The decision rules based on our proposed LaEV metric are also two-staged, and parallel the rules for EV.

2.3.1 LaEV Stage 1

Under uncertainty, the expected value of a (Stage 1) decision to adopt treatment k rather than the reference treatment on current evidence is $EV_1(k)$. If however we knew the parameters μ,δ_k exactly, then we would be able to select whether treatment 1 or k is the best treatment based on $F_1(\mu,\delta_k)$. Over the values where $F_1(\mu,\delta_k)$ is positive treatment k is best, and so there is no loss to adopting treatment k over treatment 1. However, over the region where $F_1(\mu,\delta_k)$ is negative, we would obtain a higher payoff, $-F_1(\mu,\delta_k)$, if we adopted the current standard treatment. The Expected Loss $EL_1(k)$ from selecting treatment k rather than the reference treatment is therefore:

$$EL_1(k) = E_{\mu,\delta_k}\left[Max\left(0,-F_1(\mu,\delta_k)\right)\right]$$

We now define the Loss-adjusted Expected Value by subtracting $EL_1(k)$ from $EV_1(k)$: $LaEV_1(k) = EV_1(k) - EL_1(k)$.

The Expected Loss of making a decision under uncertainty is equivalent to the EV of a decision made with perfect information (EVPI), an established concept in Bayesian decision theory.[31] However, these concepts usually refer to the value of decisions between multiple treatments, whereas here interest is focussed on the value of a decision to choose a single treatment over the reference. Like the EV, the LaEV of each treatment is therefore independent of the value of the others.[34]

2.3.2 LaEV Stage 2

As with EV, we introduce Stage 2 to prevent recommending treatments that are worse than the best treatment, $k^*_{LaEV} = Argmax_k\{LaEV_1(k)\}$, by too large a margin. The new evaluation function, $F_2(\mu, \delta_{k^*_{LaEV}}, \delta_k)$ defined above for EV is applied to treatments that pass Stage 1. The Stage 2 LaEV parallels the Stage 2 EV:

$$LaEV_2(k) = EV_2(\mu, \delta_{k^*_{LaEV}}, \delta_k) - E_{\mu, \delta_{k^*_{LaEV}}, \delta_k}\left[Max_{\delta_k}\left\{0, -F_2(\mu, \delta_{k^*_{LaEV}}, \delta_k)\right\}\right]$$

The decision rule is: adopt k if $LaEV_1(k) > 0$ and $LaEV_2(k) > 0$. From here, we use k^* because in all the real examples below $k^*_{LaEV} = k^*_{EV}$, although this will not always be the case.

2.4 GRADE Working Group method

The GRADE two-stage process for drawing conclusions from an NMA, within what they term a 'minimally contextualised framework',[29] first picks out the set of 'Category 1' treatments that are superior to the reference treatment, by a threshold margin T, with probability P; in other words, all treatments k conforming to $\Pr(F_1(\mu, \delta_k) > T) > P$, for example with P set at the standard benchmark $P = 0.975$, and T set to the MCID. On the second step, any Category 1 treatment k is promoted to Category 2 if it is superior to at least one other Category 1 treatment by the same criteria. The process continues to Category 3 or more, until we are left with a set of treatments none of which are superior to any other by the margin T with probability P. Finally, we assume that the decision rule is to recommend all treatments in the highest category. (In GRADE's own procedures, checks for evidence inconsistency and certainty ratings may intervene before recommendations are made). The values of T and P can be changed but are assumed to stay the same within each evaluation.

2.5 Probability-based ranking systems

We examine three approaches: the probability of being best, Pr(best),[35] the Surface Under the Cumulative Ranking curve (SUCRA),[21] and the probability that the value exceeds a threshold, Pr(V>T). In the latter case the decision maker ranks treatments according to the probability that their incremental value exceeds a given threshold, T.[24] The three ranking metrics are defined as follows:

$$Pb(k) = \Pr\left(F_1(\mu, \delta_k) > F_1(\mu, \delta_j) \forall_{j \neq k}\right)$$

$$Su(k) = \frac{1}{K-1}\sum_{j \neq k} I\left(F_1(\mu, \delta_k) - F_1(\mu, \delta_j)\right), \quad I(c) = 1 \ if \ c > 0, \ else \ 0$$

$$Pv(k, T) = \Pr\left(F_1(\mu, \delta_k) > T\right)$$

(To implement Pr(V>T) where T is a Relative Risk MCID, the RR is relative to reference treatment 1).

The three probability-based ranking systems are not decision rules, but the rankings can be compared to rankings generated by EV, LaEV, and GRADE. To help readers assess how they might

perform as decision rules, and to aid comparison with EV-based decisions, we reported the N most highly ranked treatments in each NMA, where N is the number recommended by the EV decision rule.

3. Illustration of properties of ranking methods in stylized examples

In the following, we present a set of four hypothetical scenarios to illustrate, compare and contrast the properties of the alternative decision rules and ranking methods. The scenarios are explained alongside the results. WinBUGS code for each illustration is available in the Supplementary Materials.

3.1 Impact of uncertainty on EV, LaEV, and Pr(V>T).

Consider a one-stage two-choice decision involving relative treatment effects of a single new treatment against a standard, and evaluation functions with distribution $F_1(k) \sim N(1,\sigma^2)$. As we vary σ **,** there is no effect on the EV, but LaEV declines, slowly at first until σ is about 1, at which point it falls off in a roughly linear fashion, reaching half its value at σ=2.3, and turning negative at σ=3.6. (Fig 1a). At this point the decision maker would choose the standard treatment.

Pr(V>T) also declines as σ increases, but only when T<EV. Otherwise, it rises if T>EV, or remains constant at 0.50 if T=EV (Figure 1b). Pr(V>T) therefore does not generate a ranking suitable for routine use. GRADE rules, which take the form: 'select if Pr(V>T)>P' are similarly limited.

Illustration 2: counter-intuitive properties of Pr(V>T).

Even when EV>T, Pr(V>T) can deliver counter-intuitive rankings. Figure 2 portrays the value distributions of three treatments, A,B, and C, with EVs 1.0, 2.0, 3.0. While A has the lowest EV, the uncertainty in A is negligeable, and the probability that V>0 is virtually 1. However, B and C both have an SD that is exactly one half of their EV, so the probability that V>0 is equal at 0.977. Pr(V>0) therefore ranks them (best to worst) A,B=C. In contrast, a LaEV decision maker, would rank them C,B,A with metrics (2.99, 1.99, 1.0), the same ranking as an EV-based decision, and with almost identical metrics. Metrics need to reflect the *extent* of gain or loss, not just its probability.

Illustration 3: Anomalies in GRADE decision rules.

Figure 3 portrays three scenarios where GRADE rules are implemented with MCID=1 and probability threshold P=0.975. In Scenario 1 the highly uncertain treatment B is recommended along with A, while in Scenario 2, the much more certain treatment C is *not* recommended although it has the same EV as B. A treatment that reaches Stage 2 is therefore more likely to be recommended if it is uncertain.

In Scenario 3, all three treatments are compared. In contrast to Scenario 1, where both A and B are recommended, in Scenario 3 only A is recommended. The recommendation of treatment B depends on the presence or absence of treatment C, even though C would not be recommended in any of these scenarios.

Illustration 4: Properties of a valid ranking system in response to uncertainty.

Here we consider a (one-stage) ranking of 25 treatments with evaluation functions distributed $F_1(k) \sim N(\mu_k, \sigma_k^2)$ arranged in a five-by-five grid with mean $\mu = 1.1, 1.2, 1.3, 1.4, 1.5$ and $\sigma = 1, 2, 3, 4, 5$. The rankings of the 25 treatments by EV, LaEV, SUCRA, Pr(Best), Pr(V>0.6), Pr(V>1.3), and Pr(V>2.3) decision rules are presented in a series of grid plots (Figure 4), in which arrows point from highest ranked treatment to the 2nd, then the 3rd, and so on. For a ranking system to be valid, the arrows must start at the lower right corner and end at the top left. Further, treatments with a higher EV must be ranked above those with a lower EV and the same SD; and those with a lower SD must be ranked above treatments with a higher SD and the same EV.

Based on this simple test, EV, SUCRA, Pr(Best), Pr(V>1.3) and Pr(V>2.0) all generate invalid rankings under uncertainty. Only LaEV and Pr(V>0.6) generate exclusively valid rankings.

4. PREFERRED PROPERTIES AND ATTRIBUTES OF TREATMENT RANKINGS

Before turning to real examples, we summarise some preferred properties of decision rules and the treatment rankings under uncertainty, based partly on the illustrative examples. The results are set out in Table 1.

5. RESULTS ON NICE GUIDELINES

We ran the original WinBUGS code, data, and initial values, discarding the same number of burn-in samples. Additional code generated results for decision rules and rankings (see Supplementary Materials). Results were based on at least 500,000 samples from the Bayesian posterior distribution.

5.1 Smoking cessation

The 2021 NICE Guidelines *Tobacco: prevention of uptake, promoting quitting and treating dependence*[14] included an NMA of 13 classes of treatments for smoking cessation against placebo. The trial outcome was the probability of cessation. The results of both Stage 1 and Stage 2 calculations appear in Table 2. Caterpillar plots (Figure 5) show the mean (EV) and uncertainty (95%CrI) in the Stage 1 and Stage 2 evaluation functions. Also shown are the LaEV of each treatment. We have applied the EV and LaEV to all treatments at both stages for illustrative purposes: in practice only treatments with $EV_1(k) > 0$ and $LaEV_1(k) > 0$ would go on to Stage 2.

In Stage 1, all but one of the 13 treatments could be recommended as better than placebo based on EV. Note that loss adjustment has virtually no impact on the Stage 1 valuations. This is because, although there is considerable uncertainty in the expected treatment effects, there is very little decision uncertainty: the EVs are so far from zero that the Expected Loss attaching to choosing each treatment over the reference treatment is negligeable. Accordingly, LaEV picked out the same treatments as EV (see Table 2 and Figure 3). In Stage 2, based on EV, the best treatment is joined by 5 other treatments that were not worse than the best treatment by more than the MCID (RR=1.50), while LaEV picks out 4 of these.

Application of the GRADE decision rules with the same MCID and a 0.975 cut-off resulted in 9 treatments reaching Category 1. In Stage 1, GRADE ranks treatments 9,6,7 highest because they have exceptionally low SD. Note that if a P=0.50, 'balance of evidence' probability had been employed, instead of 0.975, then the effect of GRADE would be identical to EV. This is what would be expected unless the distributions of the evaluative functions are highly asymmetrical. As none of the 9 Category 1 treatments were significantly better than any others by an RR of 1.50, none were

promoted to Category 2, and all would therefore be recommended. However, while the ranking by GRADE at Stage 1 was quite different to the ranking by EV, at Stage 2 the 9 treatments recommended by GRADE were among the 10 most highly ranked on EV.

SUCRA delivers a ranking that is very close to EV, while the Pr(Best) ranking departs from EV quite markedly. However, if SUCRA and Pr(Best) decision makers were to recommend the same number of treatments as an EV decision maker, they would choose the same 6 treatments. A Pr(V>T) decision maker would recommend only 4 of the treatments recommended by EV.

5.2 Other NICE Guidelines

Detailed results, references, and commentary for a further 9 NMAs from NICE Guidelines are given in the Supplementary Materials, and all 10 are summarized in Table 3. The 10 NMAs compared between 4 and 40 treatments to the reference treatment. Some of the NMAs incorporate class models, and in some the guideline developers decided between classes of treatments. To improve network connectivity, NMA datasets sometimes include treatments that are excluded from the decision set. In these cases, we have applied rankings and decision rules only to the decision set.

EV decision makers would recommend between 2 and 14 interventions (median 5), while LaEV would recommend between 3 and 11 (median 3), between zero and 3 (median 2) fewer than EV. GRADE rules with a 0.975 probability cut-off recommend between zero and 24 treatments (median 2.5), between 9 fewer and 17 more than EV. In 3/10 cases the treatment which was ranked best by EV (and LaEV) was not among the treatments recommended by GRADE. At Stage 1, GRADE privileges more certain treatments at the expense of better EV, as seen in illustration 2. However, at Stage 2, the more *un*certain treatments are recommended as they are less likely to be 'significantly' different from the best treatment.

The rankings produced by Pr(Best) and Pr(V>T) tend to differ from the EV and LaEV rankings, while SUCRA rankings are closer to EV. If we look at the N top-ranked treatments, where N is the number recommended by EV, SUCRA decision makers would recommend the same treatments in all 10 cases, Pr(Best) decision makers in 7, and Pr(V>T) decision makers 5.

DISCUSSION

This paper attempts to define a rational basis for recommending more than one treatment on the basis of NMA evidence, while penalizing uncertainty. This represents a risk-averse decision-making position, in contrast to the standard EV approach. Using stylized illustrations and real examples from NICE Guidelines, we have compared EV, LaEV, and GRADE decision rules. The performance of rankings systems based on Pr(Best), SUCRA, and Pr(V>T) has also been documented, in view of the growing literature proposing that probabilistic rankings can help inform recommendations.[20,23-27,36]

We defined a ranking as valid under uncertainty if a treatment with a higher EV must always be ranked above a treatment with a lower EV and the same uncertainty, and a treatment with less uncertainty must always be ranked above a treatment with more uncertainty at the same EV. Of the methods examined only LaEV provides a valid ranking on this definition. Pr(V>T) is valid only if EV>T *for all treatments*, a property which blocks its use in routine applications, and which is inherited by the GRADE Working Group rules for a minimally contextualised framework. Although SUCRA usually generates a ranking close to EV in real examples, except when treatments differ substantially in uncertainty, it cannot be relied on to produce valid rankings under uncertainty, and it possesses none of the preferred properties. The probabilistic ranking systems and GRADE all take uncertainty into account, sometimes in irrational ways, but they do not always penalize it. They may register the

probability of loss, but not its extent. Their fundamental drawback is that they do not distinguish between uncertainty in treatment effects from uncertainty in decision.

Because the EV and LaEV decision metrics are on the same scale as the evaluative function, they can access the MCID. This provides a natural basis for deciding how many treatments besides the best treatment should be recommended. MCID has been used in this way in NMA threshold analyses[28] and has had a similar role in Bayesian sensitivity analyses more generally.[37] Because Expected Loss is always positive (for treatments better than the reference), LaEV decision rules cannot recommend more treatments than EV, and any approach that penalizes uncertainty should have this property. The number of treatments recommended by GRADE sometimes exceeded EV, and is effectively arbitrary, subject only to the choice of probability cutoff. SUCRA delivers rankings close to the EV ranking in real examples, Pr(Best) and Pr(V>T) less so, but arbitrary cutoffs would again be required to control the number of treatments recommended by all three probabilistic ranking methods.

Adoption of any risk-averse decision rule would put a new spotlight on uncertainty and its sources. Much of the uncertainty in model parameters originates in sampling error in their estimation, but variation arising from random effects models also contributes, representing, perhaps, the uncertain relevance[38] of evidence from trials with widely dispersed treatment effects. These sources of uncertainty are 'within' the decision model and can therefore engage risk-averse methods for decision making. On the other hand, the use of GRADE certainty ratings[39] and Risk of Bias tools[40] identifies further sources of uncertainty which tend to be treated as external or contextual factors that are 'taken into account' alongside the results of formal modelling. Model structure and choice of data sources represent further sources of uncertainty outside the decision model, often addressed by sensitivity analyses. Adopting decision rules that penalize uncertainty would encourage investigators to bring all such sources of uncertainty *into* the decision model, and would place a premium on statistical methods that reduce between-study heterogeneity, including: informative priors on variance parameters;[41] bias modelling;[42] and methods that increase precision such as multi-level network meta-regression.[43] Bias models are already in common use in NICE guidelines.

The circumstances under which a risk-averse posture is appropriate remain a matter of debate, and beyond the scope of this paper. Briefly, an EV (risk-neutral) position is considered appropriate for a decision maker making large numbers of decisions under uncertainty,[44] for example a national re-imbursement agency. Put simply, the risks 'average out'. However, for individual patients making a one-time decision, a risk averse stance – penalizing uncertainty – would be justified. Risk aversion is also appropriate for institutional decision makers if costs or benefits are born by individuals and cannot be transferred,[44] or where payers have limited budgets.[34,44] Although clinical guidelines may apply to large numbers of patients, guideline development groups typically take one-time decisions. There is empirical evidence that both patients[45-47] and clinicians[48] are risk averse when facing health care decisions.

A limitation of this paper is that we have not discussed other approaches to risk aversion in the literature, including: mean-variance trade-offs, methods setting a maximum probability of a poor outcome, and methods where risk aversion is a parameter input. These alternatives have seen limited uptake[34] and none have been considered in the NMA literature. In most cases, fair comparisons would be difficult to contrive, as additional parameters are required whose values are to some extent arbitrary. A possibly more serious short-coming is our focus on risk aversion, excluding the potential role of a risk-seeking stance. Prospect Theory asserts that risk posture depends on baseline risk,[49] and there is evidence that, in health care decisions, individuals are risk-seeking at low levels of baseline health.[46,50-53] In the context of Net Benefit analysis this has been addressed by Generalized Risk-Adjusted Cost-Effectiveness (GRACE), in which willingness-to-pay varies with baseline risk.[54,55] It may be, therefore, that LaEV as elaborated here is not suited to life-threatening conditions, or where the baseline life expectancy or quality-adjusted life expectancy is low. Whether our proposals can be extended to allow risk posture to depend on baseline health status, and more generally to evaluations based on Net Benefit, are topics of on-going research.

LaEV appears to constitute a relatively conservative methodology for risk-averse decision makers. In the 10 examples, it recommended only 0-3 fewer (median 2) treatments than EV. It requires an SD of 2.3 units to halve a single unit of EV, and an SD of 3.6 units to entirely neutralise it (Illustration 1). We can therefore anticipate that if LaEV was to replace EV-based decision making, the impact would be no more than moderate. A more substantial impact would be expected where highly uncertain evidence is used, for example evaluations based on non-randomised evidence, or 'unanchored' comparisons.[56] This underscores the importance of properly representing uncertainty within the decision model: if this was implemented, routine use of risk-averse decision-making methods might incentivize the production of better quality data,[57] reversing the trend towards accepting evidence from non-randomised and one-arm studies.

Methods used by guideline developers need to be acceptable to key stakeholders, including professional colleges, manufacturers, health care workers, and patients. Stakeholders require a degree of certainty regarding which methods for health technology assessment are acceptable, and how they are to be applied. To achieve this, methods have to meet criteria for transparency and consistency across conditions.[58] This weighs against methods where parameters can be set in arbitrary ways, therefore against GRADE and against decision rules based on SUCRA or Pr(V>T) rankings, if they were to be proposed. Also problematic are ranking methods that combine efficacy with other outcomes such as adverse effects, costs, or GRADE certainty ratings, using arbitrary, condition-dependent weightings,[23] even if they were able to reliably produce valid rankings under uncertainty. More fundamentally, GRADE and the probabilistic ranking systems, and indeed other novel ranking approaches,[25,36] stand outside the standard theory and practice of health evaluation. Indeed, no theoretical basis has been proposed in which any of these methods would represent an optimal basis for decision making.

In 2001, the Institute of Medicine identified 'patient centred medicine' as an objective for improved health in the 21st century,[59] and this was widely endorsed by research funders and organisations delivering health care. Patient-centric decision making was seen as an essential component. Given that individuals are generally risk averse when facing health care decisions, a risk-averse methodology by guideline developers would be a step towards patient-centred medicine. For this purpose, the two-stage LaEV method can be recommended as reliable, conservative, theoretically well-motivated, and simple to implement.

AUTHOR CONTRIBUTIONS

A E Ades: Conceptualisation; analysis; software; writing - original draft; writing – review and editing. **Hugo Pedder**: conceptualisation; writing – review and editing. **Annabel L. Davies**: conceptualisation; writing – review and editing. **H Thom**: conceptualisation; writing – review and editing**. David M Phillippo**: conceptualisation; writing – review and editing. **Beatrice Downing**: writing – review and editing. **Deborah M Caldwell**: conceptualisation; writing – review and editing. **Nicky J Welton**: conceptualisation; writing – review and editing

ACKNOWLEDGEMENTS

CONFLICT OF INTEREST STATEMENT: HT owns shares in the consulting company Clifton Insight which has received fees from Amicus, Argenx, Baxter, Bayer, Daiichi-Sankyo, Eisai, Kalvista, Merck, Novartis, Novo Nordisk, Pfizer, Roche, and UCB. No other authors have any conflicts of interest to declare.

DATA AVAILABILITY: No new data were created or analysed in this study. The original NMA data and WinBUGs code are available from the cited guidelines.

FUNDING: None

ORCID: A. E. Ades https://orcid.org/0000-0001-7822-3552
Hugo Pedder https://orcid.org/0000-0002-7813-3749
Annabel L. Davies https://orcid.org/0000-0003-2320-7701
H. Thom https://orcid.org/0000-0001-8576-5552
David M. Phillippo https://orcid.org/0000-0003-2672-7841
Beatrice Downing https://orcid.org/0000-0002-7106-1033
Deborah M. Caldwell https://orcid.org/0000-0001-8014-7480
Nicky J. Welton https://orcid.org/0000-0003-2198-3205

REFERENCES

1. Stinnett A, Mullahy J. Net health benefits: a new framework for the analysis of uncertainty in cost-effectiveness analyses. *Med Decis Making* 1998; **18**: S68-S80.
2. Tervonen T, Lahdelma R. Implementing stochastic multi-criteria acceptibility analysis. *European Journal of Operations Research* 2007; **178**(2): 500-13.
3. Raiffa H. Decision analysis: introductory lectures on choices under uncertainty. Reading, Mass: Addison-Wesley; 1961.
4. Lindley DV. Making Decisions. 2nd ed. London: Wiley; 1985.
5. Berger JO. Statistical Decision Theory and Bayesian Analysis. 2nd ed. New York: Springer-Verlag; 1975.
6. von Neumann J, Morgenstern O. Theory of Games and Economic Behavior. 2nd ed. Princeton, NJ: Princeton University Press; 1947.
7. Wikipedia. Problem of points, https://en.wikipedia.org/wiki/Problem_of_points. (accessed February 10, 2024.
8. Claxton K, Lacey LF, Walker SG. Selecting treatments: a decision theoretic approach. *Journal of the Royal Statistical Society (A)* 2000; **163**: 211-26.
9. Claxton K. The irrelevance of inference: a decision-making approach to the stochastic evaluation of heath care technologies. *J Health Econ* 1999; **18**: 341-64.
10. National Institute for Health and Clinical Excellence. Etanercept, infliximab and adalimumab for the treatment of psoriatic arthritis [TA199]. London, 2010.
11. National Institute for Health and Care Excellence. Bisphosphonates for treating osteoporosis [TA464]. London, 2017.
12. National Institute for Health and Care Excellence. Depression in adults: treatment and management. NICE Guideline [NG 222]. London, 2022.
13. National Institute for Health and Care Excellence. Acne vulgaris: management. NICE Guideline [NG 198]. London, 2021.
14. National Institute for Health and Care Excellence. Tobacco: preventing uptake, promoting quitting and treating dependence: update [NG209]. London, 2021.
15. National Institute for Health and Clinical Excellence. NICE health technology evaluations: the manual [PMG36]. London, 2022.
16. Trinquart L, Attiche N, Bafeta A, Porcher R, Ravaud P. Uncertainty in Treatment Rankings: Reanalysis of Network Meta-analyses of Randomized Trials. *Ann Intern Med* 2016; **164**(10): 666-73.
17. Veroniki AA, Straus SE, Rücker G, Tricco A. Is providing uncertainty intervals in treatment ranking helpful in a network meta-analysis? *J Clin Epidemiol* 2018; **100**: 122-9.
18. Kibret T, Richer D, Beyene J. Bias in identification of the best treatment in a Bayesian network meta-analysis for binary outcome: a simulation study. *Clin Epidemiol* 2014; **6**: 451-60.
19. Davies AL, Galla T. Degree irregularity and rank probability bias in network meta-analysis. *Research Synthesis Methods* 2012; **12(3)**: 316-22.
20. Salanti G, Nikolakopoulou A, Efthimiou O, Mavridis D, Egger M, White IR. Introducing the Treatment Hierarchy Question in Network Meta-Analysis. *Am J Epidemiol* 2021; **191(5)**: 930-38.

21. Salanti G, Ades AE, Ioannidis JPA. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial. *J Clin Epidemiol* 2011; **64**: 163-71.
22. Rücker G, Schwarzer G. Ranking treatments in frequentist network meta-analysis works without resampling methods. *BMC Med Res Methodol* 2015; **15(58)**.
23. Mavridis D, Porcher R, Nikolakopoulou A, Salanti G, Ravaud P. Extensions of the probabilistic ranking metrics of competing treatments in network meta-analysis to reflect clinically important relative differences on many outcomes. *Biometrical Journal* 2020; **62**: 375-85.
24. Papakonstantinou T, Salanti G, Mavridis D, Rucker G, Schwarzer G, Nikolakopoulou A. Answering complex hierarchy questions in network meta-analysis. *BMC Med Res Methodol* 2022; **22(47)**.
25. Chaimani A, Porcher R, Sbidian E, Mavridis D. A Markov chain approach for ranking treatments in network meta-analysis. *Stat Med* 2020; **40**(2): 451-64.
26. Mbuagbaw L, Rochwerg B, Jaeschke R, et al. Approaches to interpreting and choosing the best treatments in network meta-analyses. *Systematic Reviews* 2017; **6(79)**.
27. Chiocchia V, Nikolakopoulou A, Papakonstantinou T, Egger M, Salanti G. Agreement between ranking metrics in network meta-analysis: an empirical study. *BMJ Open* 2020; **10(8)**: e037744.
28. Phillippo DM, Dias S, Welton NJ, Caldwell DC, Taske N, Ades AE. Threshold Analysis as an Alternative to GRADE for Assessing Confidence in Guideline Recommendations Based on Network Meta-analyses. *Ann Intern Med* 2019; **170**: 538-46.
29. Brignardello-Petersen R, Florez ID, Izcovich A, et al. GRADE approach to drawing conclusions from a network meta-analysis using a minimally contextualised framework. *BMJ* 2020; **371**: m3900.
30. Chiocchia V, White IR, Salanti G. The complexity underlying treatment rankings: how to use them and what to look at. *BMJ Evidence-Based Medicine* 2023; **28**: 180-2.
31. Raiffa H, Schlaiffer R. Applied statistical decision theory. Wiley Classics Library ed. New York: Wiley Interscience; 1967.
32. Pratt JW, Raiffa H, Schlaiffer R. Introduction to Statistical Decision Theory. Cambridge MA: Massachusetts Institute of Technology; 1995.
33. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 6. Rating the quality of evidence—imprecision. *J Clin Epidemiol* 2011; **64**(12): 1283-93.
34. Kirwin E, Paulden M, McCabe C, Round J, Sutton M, Meacock R. The risk-based price: incorporating uncertainty and risk attitudes in health technology pricing (June 16 2023). [Available at SSRN: https://ssrn.com/abstract=3956084 or http://dx.doi.org/10.2139/ssrn.3956084]. *Social Science Research Network* 2023.
35. Caldwell DM, Ades AE, Higgins JPT. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. *BMJ* 2005; **331**: 897-900.
36. Nikolakopoulou A, Mavridis D, Chiocchia V, Papakonstantinou T, Furukawa TA, Salanti G. Network meta-analysis results against a fictional treatment of average performance: Treatment effects and ranking metric. *Research Synthesis Methods* 2021; **12**: 161-75.
37. Felli JC, Hazen G. A Bayesian Approach to Sensitivity Analysis. *Health Econ* 1999; **8**: 263-8.
38. Du Mouchel WH, Harris JE. Bayes methods for combining the results of cancer studies in humans and other species. *Journal Of The American Statistical Association* 1983; **78**: 293-307.
39. Puhan MA, Schünemann HJ, Murad MH, et al. A GRADE Working Group approach for rating the quality of treatment effect estimates from network meta-analysis. *BMJ* 2014; **349**: g5630.
40. Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. *Br Med J* 2019; **366**: 14898.
41. Lilienthal J, Sturtz S, Schürmann C, et al. Bayesian random-effects meta-analysis with empirical heterogeneity priors for application in health technology assessment with very few studies. *Research Synthesis Methods* 2023; **15**(2): 275-87.
42. Dias S, Ades AE, Welton NJ, Jansen JP, Sutton AJ. Network meta-analysis for decision making. Hoboken NJ: Wiley; 2018.

43. Phillippo DM, Dias S, Ades AE, et al. Validating the Assumptions of Population Adjustment: Application of Multilevel Network Meta-regression to a Network of Treatments for Plaque Psoriasis. *Med Decis Making* 2023; **43**(1): 53-67.
44. Arrow KJ, Lind RC. Uncertainty and the evaluation ofpublic health investment decisions. *Am Econ Rev* 1970; **60(3)**: 364-78.
45. Rosen AB, Tsai JS, Downs SM. Variations in Risk Attitude across Race, Gender, and Education. *Med Decis Making* 2003; **20**(6): 511-7.
46. Rouyard T, Attema A, Baskerville R, Leal J, Gray A. Risk attitudes of people with 'manageable' chronic disease: An analysis under prospect theory. *Soc Sci Med* 2018; **214**: 144-53.
47. Ortendahl M. Shared decision-making based on different features of risk in the context of diabetes mellitus and rheumatoid arthritis. *Ther Clin Risk Manag* 2022; **3**(6): 1175-80.
48. Lawton L, Robinson O, Harrison R, Mason S, Conner M, Wilson B. Are more experienced clinicians better able to tolerate uncertainty and manage risks? A vignette study of doctors in three NHS emergency departments in England. *BMJ Quality and Safety* 2019; **28**: 382-88.
49. Kahneman D, Tversky A. Prospect Theory: An Analysis of Decision under Risk. *Econometrica* 1979; **47**(2): 263-92.
50. Lakdawalla DN, Romley JA, Sanchez Y, Maclean JR, Penrod JR, Philipson T. How cancer patients value hope and the implications for cost-effectiveness assessments of high-cost cancer therapies. *Health Aff (Millwood)* 2012; **31**(4): 676-82.
51. Attema AE, Brouwer WB, l'Haridon O, Pinto JL. An elicitation of utility for quality of life under prospect theory. *J Health Econ* 2016; **48**: 121-34.
52. Shafrin J, Schwartz TT, Okoro T, Romley JA. Patient versus physician valuation of durable survival gains: implications for value framework assessments. *Value Health* 2017; **20**(2): 217-23.
53. Mulligan K, Baid D, Doctor JN, Phelps CE, Lakdawalla DN. Risk preferences over health: Empirical estimates and implications for medical decision-making. *J Health Econ* 2024; **94**.
54. Lakdawalla DN, Phelps CE. Health technology assessment with diminishing returns to health: the Generalized Risk-Adjusted Cost Effectiveness (GRACE) approach. *Value Health* 2021; **24**(2): 244-9.
55. Lakdawalla DN, Phelps CE. The Generalized Risk-Adjusted Cost-Effectiveness (GRACE) model for measuring the value of gains in health: an exact formulation. *Journal of Benefit-Cost Analysis* 2023; **14**(1): 44-67.
56. Phillippo DM, Ades AE, Dias S, Palmer S, Abrams KR, Welton NJ. Methods for Population-Adjusted Indirect Comparisons in Health Technology Appraisal. *Med Decis Making* 2018; **38**(2): 200-11.
57. Claxton K, Briggs A, Buxton MJ, et al. Value based pricing for NHS drugs: an opporrtunity not to be missed ? *Br Med J* 2008; **336**: 251-4.
58. Dias S, Welton NJ, Sutton AJ, Ades AE. Evidence Synthesis for Decision Making 1: Introduction. *Med Decis Making* 2013; **33**: 597-606.
59. Institute of Medicine. Crossing the Quality Chasm: A New Health System for the 21st Century. Washington DC: National Academy Press; 2001.

Table 1. Performance of alternative ranking methods regarding preferred properties. Properties marked with an asterisk are considered essential. GRADE Working Group minimally contextualized framework.; Pr(best) Probability Best; SUCRA Surface Under the Cumulative Ranking curve; Pr(V>T) probability that evaluative function exceeds threshold T; LaEV Loss-adjusted Expected Value

Property or attribute	GRADE	Pr(best)	SUCRA	Pr(V>T)	LaEV	Comments
Valid ranking in the face of uncertainty *	N	N	N	N	Y	Illustration 4
Method should generate recommendations, not just rankings*	Y	N	N	N	Y	The probabilistic rankings by themselves do not specify how many, or even if any, treatments should be recommended
Methods that penalize uncertainty should only recommend as many, or fewer treatments, than EV, never more.	N	N	N	N	Y	The probabilistic rankings do not specify upper or lower limits on how many treatments would be recommended
Methods should not depend on arbitrary probability cutoffs	N	N	N	N	Y	
Metric should be in same units as evaluation function	N	N	N	N	Y	This facilitates the use of clinically interpretable benchmarks, such as MCID
Must reflect extent of loss, not just its probability*	N	N	N	N	Y	Illustration 2. Also see.[34]
Metric for each treatment should be independent of the number of alternative treatments, and the value of metrics for other treatments *	?	N	N	Y	Y	The Pr(V>T) metric underlying GRADE is independent, but its decision rule makes GRADE decisions dependent on the presence or absence of treatments that would not be recommended (Illustration 3). See[34] for a similar argument for independence.
Methods should have independent theoretical support*	N	N	N	N	Y	Only EV and LaEV have independent theoretical justifications, in the contexts of risk-neutral and -averse decision making, respectively.[34,44]

Table 2. NICE Guideline Smoking Cessation.[14] Outcome is Risk of cessation relative to Placebo. MCID based on RR=1.50. All the ranks are those generated by an EV ranking. Treatments meeting the decision criteria are shaded. For the Ranking systems in Stage 2 we have highlighted the 6 highest-rank treatments, because 6 treatments are recommended by EV.

Treatment (numbering as in NICE guidelines)		**STAGE 1 Decision Rules**							**Ranking systems**			**STAGE 2 Decision Rules**					
		EV			**LaEV**		**GRADE** (0.975) Category 1					**EV**			**LaEV**		**GRADE** (0.975) Category 1
		Rk	**EV**	**Sd**	**Rk**	**LaEV**	**Rk**	**Pr(V>T)**	**P(Best)**	**SUCRA**	**Pr(V>T)**	**Rk**	**EV**	**sd**	**Rk**	**LaEV**	
Buproprion+NRT L&S	11	1	0.30	0.12	1	0.30	9	1.000	1	1	9	1(R)	0.14	0 (R)	1(R)	0.14	9
E-cigarette +NRT L/S	14	2	0.23	0.10	2	0.23	6	1.000	2	2	6	2	0.07	0.15	2	0.042	6
E-cigarette	9	3	0.21	0.07	3	0.21	7	1.000	4	3	7	3	0.05	0.13	3	0.020	7
Varenicline+Buproprion	13	4	0.21	0.07	4	0.21	3	0.997	3	4	3	4	0.05	0.14	4	0.014	3
Vareniciline+NRT L/S	12	5	0.19	0.06	5	0.19	4	0.992	5	6	4	5	0.03	0.13	6	-0.011	4
NRT long & short	6	6	0.19	0.04	6	0.19	5	0.991	6	5	5	6	0.03	0.12	5	-0.014	5
Varencline	8	7	0.15	0.02	7	0.15	1	0.991	12	7	1	7	-0.01	0.12	7	-0.066	1
Buproprion+NRT L / S	10	8	0.11	0.03	8	0.11	10	0.987	7	8	10	8	-0.05	0.12	8	-0.131	10
NRT long/short	5	9	0.10	0.01	9	0.10	2	0.976	8	9	2	9	-0.06	0.12	9	-0.147	2
Buprioprion	7	10	0.09	0.01	10	0.09	8	0.964	11	10	8	10	-0.07	0.12	10	-0.168	
No Drug Treatment	2	11	0.05	0.02	11	0.05	12	0.250	13	11	12	11	-0.11	0.12	11	-0.236	
Wait List	3	12	0.03	0.05	12	0.02	11	0.226	9	12	11	12	-0.13	0.13	12	-0.269	
Usual Care	4	13	-0.04	0.01	13	-0.07	13	0.000	10	13	13	13	-0.20	0.12	13	-0.395	

Abbreviations: NRT Nicotine Replacement Therapy; L&S Long & Short acting; L/S Long/Short acting; Rk Rank; R Reference.

Table 3. Summary results on 10 NMAs from NICE Guidelines. Treatment recommendations from Decision Rules (EV, LaEV, GRADE) at Stages 1 and 2, and results from ranking systems, Pr(Best, SUCRA, Pr(V>T). The numbers listed are the treatment rankings under EV. For ranking systems, the N highest ranked treatments are listed, where N is the number recommended by EV. The summary statistics for GRADE assume a 0.975 probability cutoff throughout.

GUIDELINE	MCID	Max	Treatments recommended (number recommended)				Ranking systems: N Top-Ranked treatments		
			(N) EV	LaEV	GRADE	(P)	Pr(Best)	SUCRA	Pr(V>T)
Smoking Cessation	RR 1.5	13	(6) 1-6	(4) 1,3,2,4	(9) 1-7,9,10	(0.975)	1-6	1-6	3-7,9
Moderate to Severe Acne	CfB 25%	26	(14) 1-14	(11) 1-11	(5) 1,6-9	(0.975)	1-7,10,11,14,15,18,23,26	1-14	1-14
Mild to Moderate Acne	CfB 25%	40	(3) 1-3	(2) 1-2	(1) 2	(0.975)	1-3	1-3	1-3
More severe Depression	SMD 0.5	26	(5) 1-5	(3) 1,3,2	(6) 1,3,6,7,9,10	(0.85)	1-5	1-5	1,3,6,9,10
Joint Replacement	RR 1.5	4	(2) 1-2	(2) 1-2	(1) 2	(0.975)	1-2	1-2	1-2
Headache	Days 0.5	6	(3) 1-3	(3) 1-3	(1) 3	(0.975)	1-3	1-3	1-3
Social Anxiety (Treatment)	SMD 0.5	40	(7) 1-7	(5) 1-5	(24) 1-16,19-22,24,25,27,29	(0.975)	1-4,6,8,28	1-7	1,2,5,7,11,19,22
Social Anxiety (Class)	SMD 0.5	16	(9) 1-9	(6) 1-6	(4) 1,2,5,6	(0.975)	1-5,7,8,10,13	1-9	1-9
Urinary incontinence	RR 1.25	13	(5) 1-5	(2) 1-2	(8) 1,3,6-8,10-11,13	(0.975)	1-4,9	1-5	6-8,11,13
Tocolytics	Weeks 1	6	(3) 1-3	(3) 1-3	(1) 1	(0.975)	1-3	1-3	1-3
SUMMARY STATISTICS									
Number recommended									
Mean		19	5.7	4.1	5.5				
Range (median)		4-40	2-14 (5)	2-11 (3)	(0-24) (2.5)				
% of Max: Range (median)			7-56 (38)	5-50 (34)	(0-69) (22)				

Abbreviations: MCID=Minimal Clinically Important Difference; Max=Maximum number of treatments that could be recommended; N=number of treatments recommended by EV; P=GRADE probability cutoff; RR Relative Risk; CfB Change from Baseline; SMD Standardized Mean Difference

Figure 1. Evaluative function with mean 1.0 and SD varying from 0.1 to 5. (a) Impact of uncertainty on Expected Value with and without Loss-adjustment (b) Impact of uncertainty on Probability that the value exceeds a threshold, T.

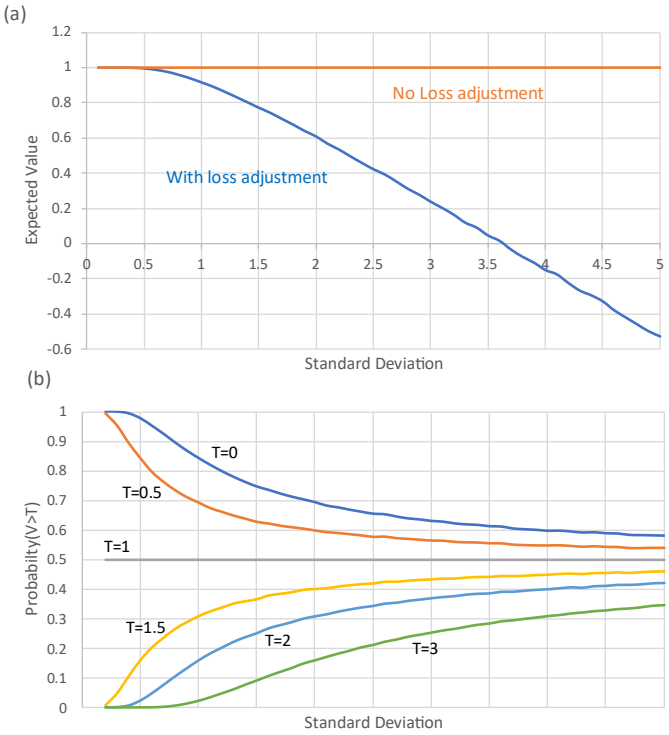

Figure 2. Forest plot showing Expected Value and 95% Credible intervals of three treatments, A, B, C. The probability that the value of A exceeds zero is virtually 1, while the probability that the value of B and C exceed 1 is equal at 0.977. Pr(V>0) would rank them A,B=C, with metrics (1,0.977, 0.977). An LaEV decision maker would rank them C,B,A with metrics (2.99, 1.99, 1.0), almost identical to an EV decision maker (3.0, 2.0, 1.0).

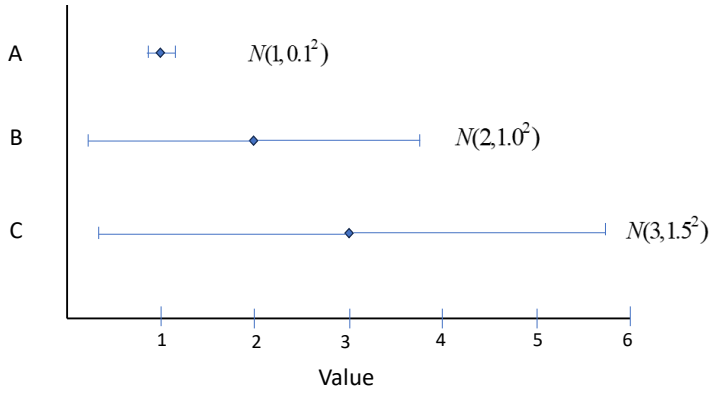

Figure 3. Forest plot showing Expected Value and 95% Credible intervals of three treatments, A, B, C. In Scenario 1, treatments A and B have reached GRADE Category 1 because Pr(V>1)>0.975, The MCID being 1. Because A is not superior to B by 1 with Probability 0.975, both A and B remain in Category 1 and are recommended. In Scenario 2, A is superior to C: A is promoted to Category 2 and is recommended, but C is not. In Scenario 3, A is superior to C and is promoted, while B is not.

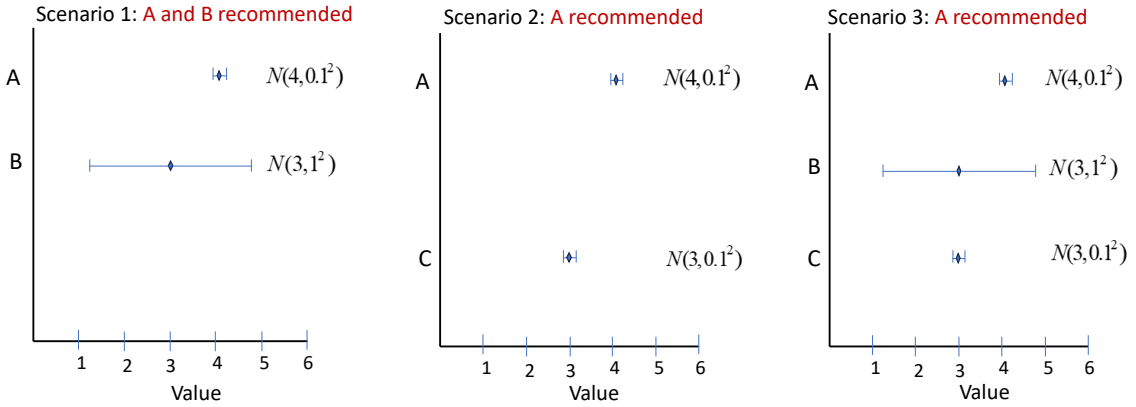

Figure 4. 25 treatments in a 5x5 grid with EVs 1.1,1.2,1.3,1.4,1,5, and SDs 1,2,3,4,5. Rankings generated by 7 metrics: EV, LaEV, SUCRA, Pr(Best), Pr(V>0.6), Pr(B>1.3), Pr(V>2.3). Arrows start from the highest ranked treatment, marked with a red blob, and point to the 2nd ranked, then the 3rd ranked, and so on. Treatments linked by a blue line are of equal rank. Valid rankings (coloured purple, see Panel 8) must start at the bottom right and end at the top left. Further, they can only point Leftwards, Upwards, bottom-left to top-right, or top-right to bottom-left. Arrows pointing downwards (red) are invalid because they imply a higher ranking for a more uncertain treatment with the same EV. Likewise, arrows pointing Rightwards are invalid as they imply a higher ranking for a treatment with a lower EV at the same SD. Arrows running top-left to bottom-right imply higher ranking for treatments with both lower EV *and* higher SD. Arrows pointing bottom-right to top-left are also invalid because they skip over treatments that either have higher EV, or lower SD, or both.

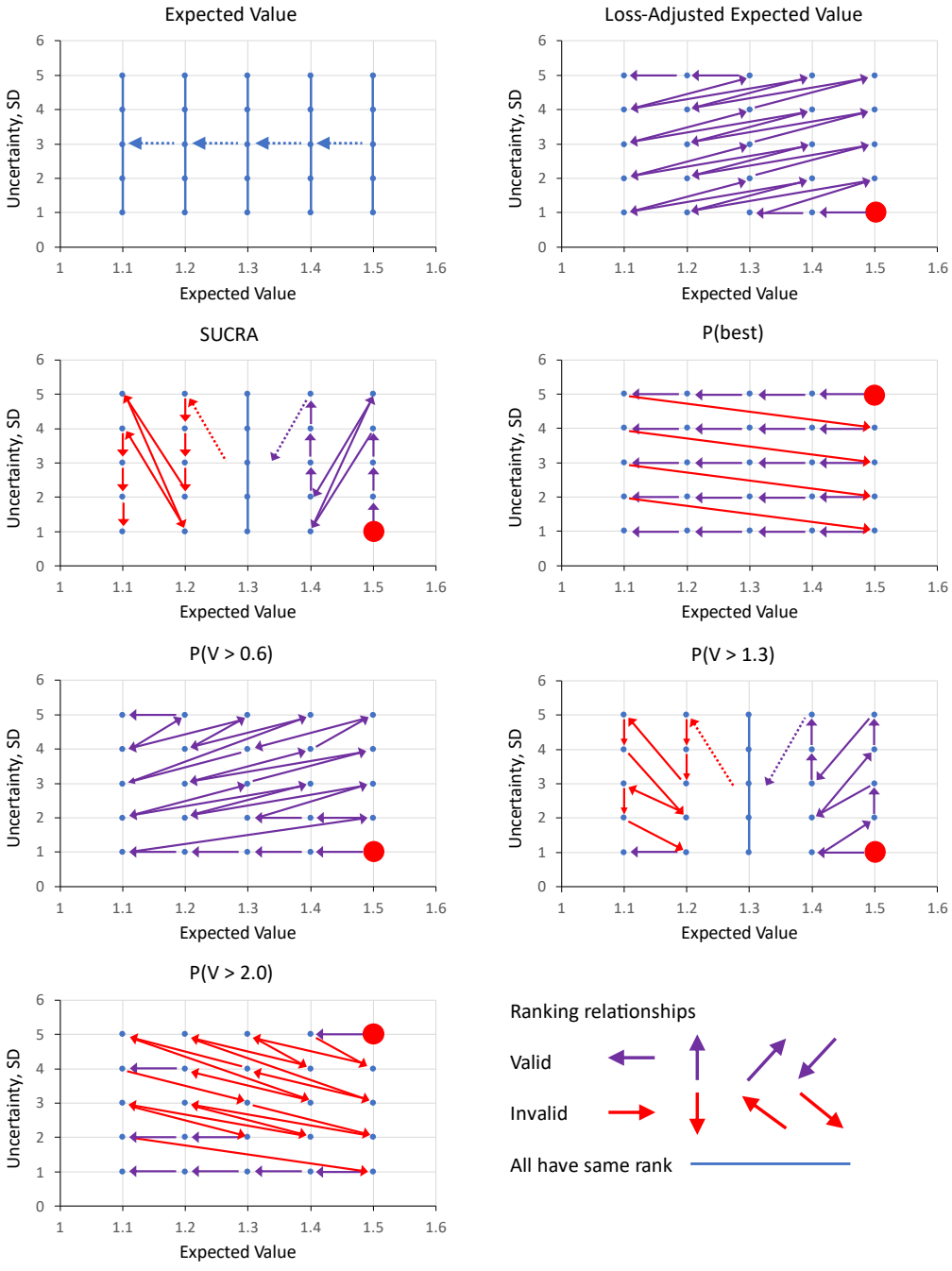

Figure 5. Smoking cessation. Caterpillar plots of the EV (blue dots) and its 95%CrI, and LaEV (red circles) of the Stage 1 and Stage 2 evaluation functions. Also shown: the coding of treatments in NICE guidelines; the MCID at Stage 2 (dotted line). Treatments recommended are those with EV, or LaEV, to the right of the zero (dashed) line in Stage 2. GRADE recommended treatments are those in bold and marked with asterisks.