A critical analysis of multi-criteria models for the prioritisation of health threats

Gilberto Montibeller, Pratik Patel and Victor J. del Rio Vilas (2019). A Critical Analysis of Multi-Criteria Models for the Prioritisation of Health Threats. European Journal of Operational Research. doi.org/10.1016/j.ejor.2019.08.018

A Critical Analysis of Multi-Criteria Models for the Prioritisation of Health Threats

Gilberto Montibeller[a], Pratik Patel[b] and Victor J. del Rio Vilas[c]

[a] Management Science and Operations Group, School of Business and Economics, Loughborough University, UK (g.montibeller@lboro.ac.uk) (corresponding author)
[b] ICCO Cooperation, Bangladesh & Zen Consortium, South Africa
[c] School of Veterinary Medicine, University of Surrey & Centre on Global Health Security, Chatham House, London, UK

Abstract

Multi-criteria assessments are increasingly being employed in the prioritisation of health threats, supporting decision processes related to health risk management. The use of multi-criteria analysis in this context is welcome, as it facilitates the consideration of multiple impacts of health threats, it can encompass the use of expert judgment to complement and amalgamate the evidence available, and it permits the modelling of policy makers' priorities. However, these assessments often lack a clear multi-criteria conceptual framework, in terms of both axiomatic rigour and adequate procedures for preference modelling. Such assessments are *ad hoc* from a multi-criteria decision analysis perspective, despite the strong health expertise used in constructing these models. In this paper we critically examine some key assumptions and modelling choices made in these assessments, comparing them with the best practices of multi-attribute value analysis. Furthermore, we suggest a set of guidelines on how simulation studies might be employed to assess the impact of these modelling choices. We apply these guidelines to two relevant studies available in the health threat prioritisation domain. We identify severe variability in our simulations due to poor modelling choices, which could cause changes in the ranking of threats being assessed and thus lead to alternative policy recommendations than those suggested in their reports. Our results confirm the importance of carefully designing multi-criteria evaluation models for the prioritisation of health threats.

Key-words: decision processes, health threat prioritisation, health risk analysis, multi-criteria analysis, simulation.

1 INTRODUCTION

Multi-criteria evaluations are increasingly being employed in the prioritisations of health threats to support decision processes related to health risk management initiatives (Balabanova et al., 2011;

Cox et al., 2013; FAO, 2014; Havelaar et al., 2010; Humblet et al., 2012; Krause, 2008; O'Brien et al., 2016). The use of multi-criteria analysis is welcome in these contexts. Firstly, it facilitates the assessment of multiple impacts, many of them without a natural quantitative attribute (such as number of cases), which are common in health threat prioritisations (Del Rio Vilas, Voller, et al., 2013; Mourits& Oude Lansink, 2007). Secondly, it can encompass the use of expert judgment in the estimation of impacts, when hard evidence is not available or sufficiently reliable (Batz et al., 2005; Morgan, 2014). Thirdly, it permits the modelling of different priorities, reflecting the concerns of policy makers and experts when dealing with multiple impacts in their decision processes (Angelis et al., 2017; Baltussen et al., 2010; Del Rio Vilas et al., 2013b).

The decision analysis literature provides extensive guidance on how to build up these models, for instance on how to frame the decision problem (Barcus and Montibeller, 2008; Keeney, 1992), structure value models (Franco and Montibeller, 2011; Keeney, 2013), define attributes (Keeney and Gregory, 2005) and elicit preference parameters (von Winterfeldt and Edwards, 1986). In addition, it provides warnings about modelling mistakes (Dillon-Merrill et al., 2008; Keeney, 2002) and identifies cognitive and motivational biases that may affect preference elicitation of individuals (Montibeller and von Winterfeldt, 2015) and of groups (Montibeller and Winterfeldt, 2018).

Despite the availability of this body of literature, best practices are still ignored in a number of health threat prioritisation models (Del Rio Vilas et al., 2011). Some of these models may lack a clear conceptual multi-criteria framework, in terms of both axiomatic rigour and adequate procedures for preference elicitation and modelling. They are thus *ad hoc* models from a multi-criteria decision analysis perspective, despite the strong health expertise employed to build them. While the problem is widespread, examples of well-designed models do exist (e.g. Brookes et al. (2014); Cox et al. (2013)).

Concern about best practices is emerging within the health evaluation community, as the recent reviews on multi-criteria models for health threat assessments (O'Brien et al., 2016) and healthcare decisions (Marsh et al., 2016) attest. In addition, thoughtful critiques of weak modelling practices have recently been published in health technology assessment (Morton, 2017) and life-critical medical decision making (Kujawski et al., 2019). However, these articles do not provide a critical assessment of health threat prioritisation models.

To address this gap, we examine in this paper the key modelling assumptions of *ad hoc* multi-criteria health threat prioritisation models, discuss the main modelling choices adopted in the design of such models, and compare them with well-established best practices of multi-attribute value

analysis (Belton and Stewart, 2002; Keeney and Raiffa, 1993; von Winterfeldt and Edwards, 1986). We review several health-threat prioritisations that were published in recent years.

In addition, we suggest a set of guidelines on how simulation studies might be employed to assess the impact of these modelling choices. We apply these guidelines to simulate the impacts of modelling assumptions and design choices on the overall ranking of health threats of two relevant studies available in the health threats domain: the prioritisation of communicable diseases developed by the Robert Koch Institute (RKI) in Germany (Balabanova et al., 2011; Krause, 2008) and the ranking of foodborne parasites conducted by the Food and Agricultural Organization of the United Nations (FAO) (FAO, 2014). These studies were chosen because they provide extensive details of their multi-criteria models, were rigorous in gathering and aggregating data, and dealt with highly relevant health issues.

This paper makes two main contributions. From a conceptual perspective, policy analysts and health experts involved in health threat prioritisations may benefit from an improved understanding of the consequences of adopting *ad hoc* modelling practices, ignoring axiomatic properties, and eliciting preferences parameters with inadequate protocols. The modelling assumptions that we scrutinize can serve as warnings, as they are, in our experience, frequently made in this context. From a methodological perspective, we suggest a set of guidelines for conducting simulation studies to assess modelling issues that can be extended to other classes of multi-criteria evaluations employed in risk-related assessments (Greenberg et al., 2012). In doing so we go beyond prescriptive advices about what should be done (e.g. Dillon-Merrill et al. (2008); Keeney (2002)) and provide a framework to assess the impact of modelling assumptions on the prioritisation results.

The remaining of the paper has the following structure. Section 2 describes briefly multi-criteria value analysis and formalises the components of this type of model. For each component we discuss conceptual modelling assumptions and provide examples of health threat multi-criteria assessments available in the literature. Section 3 details the simulation analyses conducted on the two selected studies, where we assess the range of changes in the rankings of health threats due to their modelling assumptions and choices. We conclude the paper by discussing the findings of our simulation analyses, mentioning the limitations of the research design, and providing directions for further exploration on this topic.

2 MODELING ASSUMPTIONS IN MULTI-CRITERIA HEALTH THREAT PRIORITISATIONS

We searched for multi-criteria health threat assessments and identified sixteen studies covering a variety of human and animal health threats[1] (see first and second columns in Table 1). We included only studies that: i) prioritised health-related threats (diseases, pathogens, etc.); ii) were published from 2000 onwards; iii) used a multi-criteria method (or claimed to use a multi-criteria assessment) or, else, iv) employed an additive model that resembles a multi-attribute value assessment (scores for the threats and weights for the impact criteria).[2]

A number of multi-criteria health threat prioritisations models reported in the literature do not employ any conceptual multi-criteria framework or provide references to specific multi-criteria decision analysis methodologies, nor show any attempt of establishing links to the field (Del Rio Vilas et al., 2011). In our search, several of these applications do not make any explicit reference to the multi-criteria method employed (see third column in Table 1 which describes what method was employed in each study).

Several features of these *ad hoc* multi-criteria models seem to indicate that they could be compared against a standard multi-attribute value analysis (MAVA), namely: i) the format of the evaluation, with riskless health threats (e.g. parasites, diseases); ii) the creation of the scoring systems, with functions mapping the impacts on attributes into value; and iii) the method by which partial performances are aggregated, with the use of weighted sums. In this section we thus briefly formalize MAVA, describe the best practices employed and axiomatic properties required within these models, and illustrate the modelling issues that we observed in the models for health threat prioritisation that we analysed when compared with a MAVA framework.

2.1 Multi-Attribute Value Analysis

In decision analysis there is a conceptual distinction between value and utility, the former referring to situations of riskless choices, the latter to decisions with uncertain outcomes (Dyer and Sarin, 1979; Keeney and Raiffa, 1993). Here we will focus on multi-attribute value analysis (Keeney and

[1] The analysis follows the relevant components of the PROACT-URL framework (Hammond et al., 1999): the *prioritisation* **Pr**oblem that each study addresses; the *attributes* to assess **O**bjectives employed in the model; the type of **A**lternatives *(health threats*) being assessed; the *value functions* that measure **C**onsequences; the *attribute weights* that represent value **T**rade-offs; and the sensitivity analysis performed to assess **U**ncertainties.

[2] We analysed the papers listed in previous literature reviews (ECDC, 2015; Mehand et al., 2018; O'Brien et al., 2016) and conducted an extensive search using Google Scholar (https://scholar.google.co.uk) with the following combinations of key-words: ("multiple criteria decision analysis" OR "multi attribute decision analysis" OR "multi attribute value analysis" OR "MCDA") AND ("diseases" OR "pathogens" OR "health threats" OR "health risks" OR "zoonoses" OR "emerging health threats" OR "emerging zoonoses").

Raiffa, 1993; von Winterfeldt and Edwards, 1986), given the way that these *ad hoc* multi-criteria health threat prioritisations have been typically defined.

Let a set of N attributes x_i (i = 1, 2, …, N), be bounded between $[x_i^*, x_i^0]$; and V_i a monotonic function, which measures the partial dis-value of a threat on the i-th attribute. (Health threat prioritisations have a ranking system in which the highest ranked threat receives the highest score. To be compatible with value theory we needed to introduce the concept of dis-value, in line with the concept of dis-utility widely employed in Economics.)

Each dis-value function is usually normalized between $V_i(x_i^*) = 100$ (maximum dis-value) and $V_i(x_i^0) = 0$ (minimum dis-value). Let the impact caused by a k-th health threat on the i-th attribute be denoted by $V_i(x_i^k)(k = 1, 2, …, M)$ with M being the number of threats to be ranked. (In the paper we will use V(.) to denote generic dis-value functions and those used in the simulations and C(.) to denote dis-value functions taken from the models retrieved from the literature.)

The dis-value functions V_i should be constructed using the difference value model, based on traditional difference measurement (Krantz et al., 1971), where differences of strength of preference are measured using pairwise comparisons. For instance, given four threats a, b, c, d ∈ A, where A is a set of threats, there exists a dis-value function $V(.) \in \mathbb{R}$ which maps out the relationship between objects in the set A, such that if the strength of dis-preference of a over b (a,b) is at least as large as the strength of dis-preference of c over d (c,d) (Dyer and Sarin, 1979; von Winterfeldt and Edwards, 1986), then:

$$(a, b) \gtrsim (c,d) \Leftrightarrow V(a) - V(b) \geq V(c) - V(d)$$

A set of properties are required for such a dis-value function V(.) function to exist: connectivity, transitivity, summation, cancellation, solvability, and the Archimedian properties (for details see von Winterfeldt & Edwards(1986)).

These partial dis-values scores $V_i(x_i^k)$ can then be aggregated by a function Φ, such that an overall dis-value score for the k-th health threat is assessed (Dyer and Sarin, 1979; Keeney and Raiffa, 1993) by:

$$\mathbf{V^k} = \Phi[V_1(x_1^k), V_2(x_2^k), …, V_N(x_N^k)] \quad \text{(Eq. 1).}$$

We relate our discussion below to these sixteen studies (Table 1), showing to which extent the issues that we found in a single case study are present in other applications.

Gilberto Montibeller, Pratik Patel and Victor J. del Rio Vilas (2019). A Critical Analysis of Multi-Criteria Models for the Prioritisation of Health Threats. European Journal of Operational Research. doi.org/10.1016/j.ejor.2019.08.018

Table 1. Multi-Criteria Health Threat Assessments Reported in the Literature.

Paper	**Purpose of the evaluation**	**Multi-Criteria Method Employed**	**Checks on the properties of attributes** (see Sections 2.2 and 2.4)	**Design of attributes** (see Section 2.2)	**Elicitation of value functions** (see Section 2.3)	**Elicitation of weights** (See Section 2.5)	**Sensitivity/robustness analysis** (See Section 2.6)
Balabanova et al. (2011)	Prioritisation of communicable diseases in Germany	Not mentioned	Not mentioned	Clear definition of levels for qualitative attributes and use of indices for quantitative attributes but with stepwise functions	Three levels with a bivalent scale but with 0 not reflecting a neutral level	Elicited using the notion of direct importance	None
Brookes et al. (2014)	Prioritisation of exotic diseases for the pig industry in Australia	Multi-attribute value theory	Not mentioned	Clear quantitative indices	Assumed linear value functions for all attributes	Use of scenarios representing value trade-offs	Sensitivity analysis of weights on overall scores of threats
Cardoen et al. (2009)	Prioritisation of food born zoonoses in Belgium	Not mentioned	Not mentioned	Not clearly defined	Experts provided directly qualitative labels to assess the threats which were converted into numbers	Points distributed following the notion of direct importance	Analysis of ranges for overall impact of threats from individual expert's assessments
Collineau et al. (2018)	Risk ranking of antimicrobial-resistant hazards in meat in Switzerland	Not mentioned	Not mentioned	Clear definition of levels for qualitative attributes and use of indices for quantitative attributes, but assumed preference independence between incidence and severity	Assumed equal value distance between ordinal levels of the attributes	Weights were not elicited, assuming equal weights	Sensitivity analysis on the weights of the model
Cox et al. (2013)	Prioritising Emerging or Re-Emerging Infectious Diseases Associated with Climate Change in Canada	Multi-attribute value theory	Not mentioned	Clear definition of levels for qualitative attributes and use of indices for quantitative attributes but with stepwise functions	Elicited from experts using the Macbeth software	Elicited from experts using the swing weights procedure	Criteria were weighted using probability distributions representative of expert opinion
Dahl et al. (2015)	Communicable Diseases Prioritised According to Their Public Health Relevance in Sweden	Not mentioned	Not mentioned	Clear definition of levels for qualitative attributes and use of indices for quantitative attributes but with stepwise functions	Three levels with a bivalent scale but with 0 not reflecting a neutral level)	Elicited using the notion of direct importance	None

Gilberto Montibeller, Pratik Patel and Victor J. del Rio Vilas (2019). A Critical Analysis of Multi-Criteria Models for the Prioritisation of Health Threats. European Journal of Operational Research. doi.org/10.1016/j.ejor.2019.08.018

Paper	**Purpose of the evaluation**	**Multi-Criteria Method Employed**	**Checks on the properties of attributes** (see Sections 2.2 and 2.4)	**Design of attributes** (see Section 2.2)	**Elicitation of value functions** (see Section 2.3)	**Elicitation of weights** (See Section 2.5)	**Sensitivity/robustness analysis** (See Section 2.6)
Del Rio Vilas et al. (2013b)	Prioritisation of emerging animal health threats to the UK	Multi-attribute value theory	Not mentioned	Unambiguous descriptions and adequate handling of dependences	Elicited from experts using the Macbeth software	Elicited from experts using the swing weights procedure	None
Domanovici et al. (2017)	Prioritistion of bacterial infections transmitted through substances of human origin in Europe	Not mentioned	Not mentioned	Use of indices for quantitative attributes but with stepwise functions and overlapping thresholds	Four levels with arbitrary conversions from attribute to a value scale	Multiplicative function with weights that do not reflect value trade-offs	None
FAO (2014)	Ranking of food-borne parasites	Not mentioned	Not mentioned	Use of indices for quantitative attributes but with stepwise functions and overlapping thresholds; preferentially dependent attributes (severity and number of cases) were treated as independent	Four levels with arbitrary conversions from attribute to a value scale	Elicited using the notion of direct importance	Sensitivity analysis of performance assessments and weights for different expert groups
Garner et al. (2015)	Assessment of antimicrobial resistant disease threats in Canada	Not mentioned (except for the sensitivity analysis)	Not mentioned	Clear definition of levels for qualitative attributes and use of indices for quantitative attributes but with stepwise functions	Three levels with arbitrary conversions from attribute to a value scale	Elicited using the notion of direct importance	Detailed robustness analysis comparing with outranking results, varying weights and with a different set of weights
Havelaar et al. (2010)	Prioritisation of pathogens in the Netherlands	Not mentioned	Not mentioned	Use of indices for quantitative attributes but with stepwise functions and overlapping thresholds	Convoluted conversion of levels into scores	Elicited with the probabilistic inversion function method	Sensitivity of results to different criteria weights and to distributions on scores
Humblet et al. (2012)	Prioritisation of diseases and zoonoses in Europe	Not mentioned	Not mentioned	Clear definition of levels for qualitative attributes and use of indices for quantitative attributes but with step-wise functions	Discrete set of four levels for quantitative attributes scored between 0 and 4	Unclear procedure for the allocation of points from which weights were calculated	Criteria were weighted using probability distributions

Gilberto Montibeller, Pratik Patel and Victor J. del Rio Vilas (2019). A Critical Analysis of Multi-Criteria Models for the Prioritisation of Health Threats. European Journal of Operational Research. doi.org/10.1016/j.ejor.2019.08.018

Paper	**Purpose of the evaluation**	**Multi-Criteria Method Employed**	**Checks on the properties of attributes** (see Sections 2.2 and 2.4)	**Design of attributes** (see Section 2.2)	**Elicitation of value functions** (see Section 2.3)	**Elicitation of weights** (See Section 2.5)	**Sensitivity/robustness analysis** (See Section 2.6)
Kadohira et al. (2015)	Stakeholder prioritisation of zoonoses in Japan with analytic hierarchy process method	AHP	Not mentioned	Use of indices for quantitative attributes but with stepwise functions and overlapping thresholds; preferentially dependent attributes (severity and number of cases) were treated as independent	Four levels with arbitrary conversions from attribute to a value scale	Elicited using the notion of direct importance (as it is standard in the AHP)	Analysis of impact on ranking from weights defined by different groups of experts
Mehand et al. (2018)	Prioritisation of emerging infectious diseases in need of research and development	AHP	Not mentioned	No attributes were defined and there are preferentially dependent criteria	Value function indirectly assessed via the AHP method	Elicited using the notion of direct importance (as it is standard in the AHP)	Limited sensitivity analysis on the weights of the model
Munyua et al. (2016)	Prioritisation of zoonotic diseases in Kenya	AHP	Not mentioned	Use of indices for quantitative attributes but with stepwise functions; preferentially dependent attributes (severity and prevalence) were treated as independent	Value function indirectly assessed via the AHP method	Elicited using the notion of direct importance (as it is standard in the AHP)	Limited sensitivity analysis on the weights of the model
Pieracci et al. (2016)	Prioritisation of zoonotic diseases in Ethiopia	Not mentioned	Not mentioned	Use of ambiguous levels for indices with stepwise functions; double-counting attributes (severity and burden of disease)	Four levels with arbitrary conversions from attribute to a value scale	Elicited using the notion of direct importance	No

2.2 Definition of Criteria and Attributes

The criteria in a multi-criteria evaluation should reflect the fundamental objectives of concern in the prioritisation (Keeney, 2013, 1992). These fundamental objectives should be identified from an appropriate decision frame (Barcus and Montibeller, 2008; Keeney, 1992), which balances their essentiality in relation to ultimate objectives *versus* their ability to assess impacts of health threats that are only influenced by the threats themselves, but not other events ("controllability"). This property seems to have remained unchecked in certain *ad hoc* models, such as for the 57 attributes employed by Humblet et al. (2012), which might not all be measuring fundamental objectives. In our review we have not found a single application that has mentioned explicitly that such properties have been checked (see fourth column in Table 1 which summarises if checks on the properties of the attributes were conducted in each model).

An attribute should be carefully selected to measure the concern expressed by the respective fundamental objective. Whenever a natural quantitative attribute is available (for instance "total number of disease cases") it should be used. If a natural attribute is not available, then the analyst needs to define either a constructed attribute specifically designed for the evaluation, or else, a proxy attribute measuring the achievement of a means objective to the fundamental objective (see Keeney and Gregory (2005) for details).

The definition of attributes in *ad hoc* models provides skewed translations from impact into dis-value. Firstly, quantitative attributes are discretized into categories, with steep jumps in value and, sometimes, overlapping categories. For example, in the model proposed by Havelaar et al. (2010) to prioritise emerging zoonoses, the attribute "transmission in animal reservoirs" C_H (performance metric: X_H = prevalence per 100,000 animals) is discretized into five categories, valued as: $C_H(X_H< 1) = 0$; $C_H(1 \leq X_H \leq 100) = 50$; $C_H(100 \leq X_H \leq 1{,}000) = 500$; $C_H(1{,}000 \leq X_H \leq 10{,}000) = 5{,}000$; $C_H(X_H> 10{,}000) = 50{,}000$. The categories overlap, so it is unclear if $X_H = 100$ has a value of 50 or 500. Furthermore, the model suggests a steep increase in value that the rise of two cases (e.g. from 999 to 1001) would result in (from 500 to 5,000 units of dis-value). Secondly, the discretization of these scales is further compounded by a very limited number of attribute levels, such as in the model proposed by the RKI (Balabanova et al., 2011; Krause, 2008), in which each attribute has only three levels (see Table A1[3]). These attributes with a limited number of levels are present even in models that are rooted on a sound multi-criteria framework (*e.g.* Del Rio Vilas et al. (2013a)).

[3]All tables with simulation data and simulation results are in the Online Appendix II.

Many of these models include attributes that are concerned with intangible issues. They often rely on qualitative labels that are ambiguously defined. For example, the model developed by the FAO (2014) (see Table A2) assesses the attribute C_8"how relevant is the parasite for international trade" using the labels "not at all", "some relevance", and "high relevance". In the context of probability elicitation, evidence has shown that there are wide variations of between-subject interpretation of such labels (Budescu and Wallsten, 1985; Wallsten et al., 1986). These results should be valid also for value assessments, a matter of concern when multiple experts are assessing the threats. In our survey of multi-criteria health threat evaluations, several studies presented ambiguous attributes (see the fifth column in Table 1 which analyses how attributes were designed in each model).

2.3 Definition of Dis-Value Functions

As mentioned previously, to each i-th attribute an associated dis-value function should be defined to assess every k-th threat: $V_i(x_i^k)$. For continuous variables, such as number of disease cases, a continuous function can be defined, anchored at $V_i(x_i^*) = 100$ and $V_i(x_i^0) = 0$. These functions might be linear for some variables (such as number of disease cases) but are often non-linear to indicate decreasing /increasing marginal dis-value. Simulation studies have shown that linear multi-attribute value models are sensitive to the shape of value functions (Stewart, 1996), therefore such non-linearities should be modelled whenever present.

As value functions are measured on interval scales (Dyer and Sarin, 1979), they admit positive linear transformations such as: $V'(.) = \alpha V(.) + \beta$ with $\alpha > 0$. Thus the zero of the scale need not be necessarily set at the lowest level of the attribute (x_i^0) and instead it can be placed at a 'neutral' level (Bana e Costa et al., 2008). This flexibility about the location of the zero needs to be used only for those attributes that are bivalent (i.e. with positive and negative ranges, indicating regions of desirability and undesirability, respectively). This aspect has not been observed in some of the health threat models we analysed. For instance, the model proposed by the RKI (Balabanova et al., 2011; Krause, 2008), features single valence attributes, such as the number of cases (C_1) scored using a bivalent scale: {-1,0,1} (Table A1).

The discretisation of continuous variables into a small set of categories often leads to inconsistencies between impacts and valuations. For instance, in the FAO model (see Table A2), the attribute "number of global food-borne illnesses" (C_1) is scored in the following levels (bins) (where X_1 = number of cases): $C_1(X_1 < 10^4) = 0$; $C_1\ (10^4 \leq X_1 \leq 10^5) = 25$; $C_1(10^5 \leq X_1 \leq 10^6) = 50$; $C_1(10^6 \leq X_1 \leq 10^7) = 75$; $C_1(X_{F1} > 10^7) = 100$. The value of a case in Bin2 is thus $(25 - 0)/(10^6 - 10^5) = 2.77 \times 10^{-4}$, while for Bin3 each case is worth $(75 - 50)/(10^7 - 10^6) = 2.77 \times 10^{-6}$, consequently a case in the former is considered as 100 times more valuable than the latter. It is not clear if these valuations are an

unintended consequence of the way the attribute was built, instead of a deliberate attempt of allocating a non-linear marginal value for life (Dickert et al., 2012).

Another common issue is the implicit assumption of equal marginal value reductions between the categories. Consider the attribute related to the likelihood of increased human burden (C_7) in the FAO model (Table A2). The scoring system assumes that the marginal dis-value in moving from 0-25% (Bin1) to the 25-50% category (Bin2), of 25 value points, is the same as moving from this to the 75-100% category (Bin3). A similar assumption is made in Cox et al. (2013)'s model. As before, this might be an unintended consequence of the way the attribute was defined.

Adequate elicitation procedures should be used to define value functions (von Winterfeldt and Edwards, 1986) either using numerical judgments, indifference judgments, or qualitative judgments that are adequately translated into values (Bana e Costa et al., 2012). Our review shows that value functions are often arbitrarily set, with unrecognized consequences on the evaluation of dis-value of threats (sixth column Table 1 which analyses how value functions were elicited).

2.4 Definition of Properties for the Set of Attributes

As discussed previously, appropriate decision framing should lead to the definition of attributes that are associated with fundamental objectives. Other properties are also necessary for the set of attributes, specifically they should be comprehensive, concise, understandable, decomposable, non-redundant, and if possible, preferentially independent (Belton and Stewart, 2002; Keeney and Raiffa, 1993).

The non-redundancy property is contingent on avoiding double-counting, so the impact is not over-weighted within the aggregation. For example, in the Humblet et al. (2012)'s model for the prioritisation of diseases, it seems that "lower human consumption of animals" and "impact on an [a given] animal industry" could be, at least partially, double-counting the same concern. Attributes in *ad hoc* models might have been chosen because data is readily available for some indices, or alternatively because modellers want to accommodate all the issues that experts raise.

The preferential independence property among attributes must be observed if the multi-criteria model employs a simple weighted sum as the aggregation function Φ (in Eq. 1), i.e.:

$$\mathbf{V^k} = \Phi[V_1(x_1^k), V_2(x_2^k), \ldots, V_N(x_N^k)] = \sum_{i=1}^{N} w_i\, V_i(x_i^k); \qquad \text{with } \sum_{i=1}^{N} w_i = 1 \qquad \text{(Eq. 2).}$$

Where w_i is the weight associated with the i-th attribute. We noticed that this aggregation rule has been commonly used in the *ad hoc* health threat assessments.

In the case of a multi-attribute value function, the required property for performing the aggregation shown in Eq. 2 is weak-difference independence (Dyer and Sarin, 1979; Keeney, 1992). This property can be easily formalized for two attributes x_1 and x_2 with four attribute levels $x_1'' > x_1'$ (with x_1'', $x_1' \in [x_1^*, x_1^0]$) and $x_2'' > x_2'$ (with x_2'', $x_2' \in [x_2^*, x_2^0]$). The attribute x_1 is independent of attribute x_2 if $\mathbf{V}[x_1'', x_2''] - \mathbf{V}[x_1', x_2''] = \mathbf{V}[x_1'', x_2'] - \mathbf{V}[x_1', x_2'] \; \forall x_1'', x_1', x_2''$, and x_2'. This is graphically illustrated in Figure 1. The property is not symmetrical (Keeney, 1992), so the same test must to be done for x_2 against x_1.

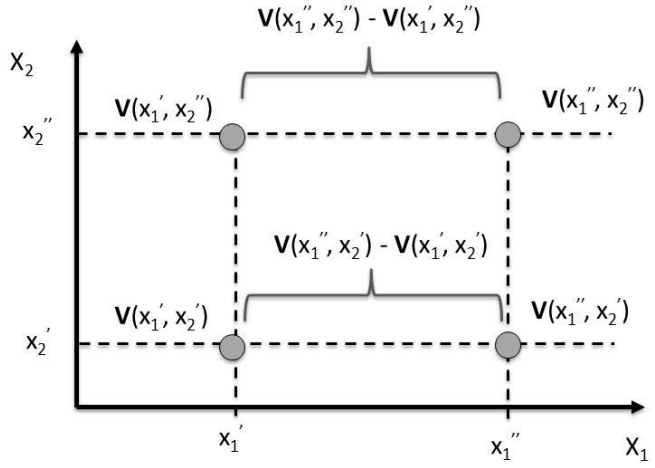

Figure 1: The weak-difference independence condition (adapted from von Winterfeldt and Edwards(1986)).

In the health context, a common case where this preferential independence does not hold is when assuming number of disease cases and disease severity as two preferentially independent attributes. Consider that x_1 is the value for the first attribute, ranging from x_1'= 100 to x_1'' = 1,000 cases; and x_2 is the value for the second attribute, ranging from x_2'= "mild" to x_2'' = "life threatening". In this case the value distance when the disease is life threatening is probably much larger than when it is mild, thus: $\mathbf{V}[x_1'', x_2''] - \mathbf{V}[x_1', x_2''] \gg \mathbf{V}[x_1'', x_2'] - \mathbf{V}[x_1', x_2']$. Some of the *ad hoc* health threat assessments proceed to use an additive value function as if these two attributes were preferentially independent (e.g. Cardoen et al.(2009), Balabanova et al. (2011), Krause et al. (2008)).

For example, consider attributes C_1 ("Number of global food-borne illnesses") and C_4 ("Chronic morbidity severity") in the FAO model (Table A2)[4] and assume, without loss of generality, that C_5 ("Fraction of illness that is chronic") is Bin 4 (this just defines a proportion, thus 100% chronic and 0% acute morbidity severity). C_1 and C_4 are normalized between 0 and 100, with the bins equally spaced in value.

[4]The aggregation formulas of the FAO model are detailed in Equations 3 and 4 in Section 3.1 of this paper.

As preferential independence is assumed between these attributes C_1 and C_4, and the baseline weights for both attributes C_1 and C_{345} are identical (see Table A3) this implies iso-value lines as shown in Figure 2, with all combinations over a given line receiving the same score. The aggregate dis-value $C_1 + C_{345}$ for this example is shown in Table A4. Notice the same dis-value gap between Bin 0 and Bin 4 of C_1 for a given Bin of C_4 (Table A4, last column). However, it is likely that this dis-value gap should be higher when the cases are severe (Bin4 on C_4) than when there is no severity (Bin0 on C_4).

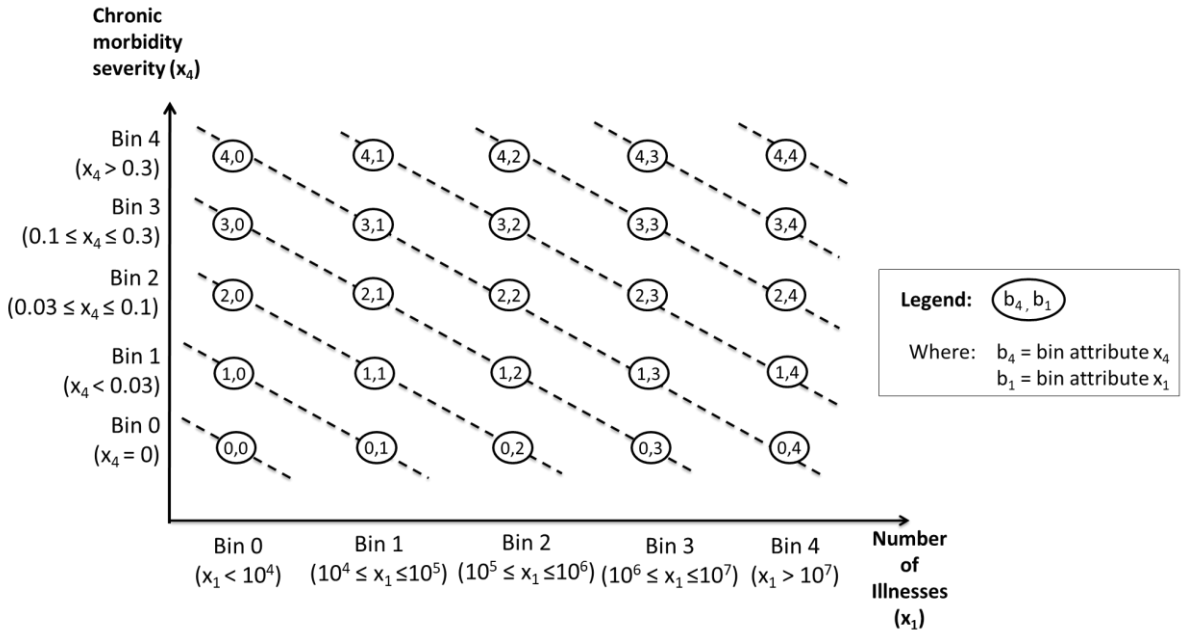

Figure 2: The assumption of preferential independence between attributes C1 and C4 for the FAO model.

A related consequence of considering these types of attributes (number of cases and severity) as preferentially independent is that iso-value curves are generated in a way that may lead to illogical prioritisations. To illustrate, suppose there exist two threats in the FAO model, again setting up C_5 as Bin 0: the first threat leading to more than 10^7 cases but no chronic morbidity, i.e. $C_{14}^1[x_1=\text{Bin4}, x_4=\text{Bin0}]=22$; and a second one with less than 10^4 cases but severe chronic morbidity severity, i.e.$C_{14}^2[x_1=\text{Bin0}, x_4=\text{Bin4}] =22$ (see cells in bold in Table A4). Both have the same dis-value score, and therefore are mathematically considered equally serious, but there is *no* morbidity in the former despite the large number of cases, so we would expect its score to be lower. In our survey, we have not found any explicit checks on the axiomatic or preference properties of the multi-criteria models that were developed (see the fourth column in Table 1 which analyses if these properties were checked in each model).

2.5 Elicitation of Attribute Weights

Weights in a multi-criteria model reflect the trade-offs that must be made among conflicting objectives. In a multi-attribute value function, weights are scaling constants that convert partial values into an overall value (Keeney and Raiffa, 1993), given both the range of each i-th attributes $[x_i^0 , x_i^*]$ associated with attribute $x_i(.)$ and the relative importance of the "swing" from level x_i^0 to level x_i^* in comparison with other swings in the remaining attribute set.

The requirement for weights to be always anchored on the extreme levels of the attributes brings two important consequences. Firstly, any protocol for the elicitation of weights *must* ask for preference information in relation to such anchors. For instance, the trade-off method for elicitation, when applied for two attributes x_1 and x_2, with $\mathbf{V}[x_1^*, x_2^0] < \mathbf{V}[x_1^0, x_2^*]$, will ask the evaluator for a level x_2, such as $\mathbf{V}[x_1^*, x_2^0] = \mathbf{V}[x_1^0, x_2]$. Weights can be then calculated from the two equations: $w_1V_1(x_1^*) + w_2 V_2(x_2^0) = w_1 V_1(x_1^0) + w_2 V_2(x_2)$ (from Eq.2 and the equal value point) and $w_1 + w_2 = 1$ (from Eq. 2). Other methods, such as swing weighting, simplify the elicitation protocol, but again use the anchors to elicit weights (von Winterfeldt and Edwards, 1986). Most decision analysts prefer to elicit such priorities directly from experts, helping them to think explicitly about trade-offs, but conjoint methods can also be employed to derive weights, based on the ranking of dummy threats (Kurowicka et al., 2010). Secondly, any elicitation protocol that does not use anchors, or asks for "direct importance" of attributes should be avoided, as empirical results show that the answers provided are unreliable and incompatible with the way that weights are conceptualized in multi-attribute value functions (Keeney, 2002; von Nitzsch and Weber, 1993).

The direct elicitation of importance, as employed in some health threat prioritisations (see the seventh column in Table 1 which describes how the weights were elicited in the sample we analysed) is known as the "most common critical mistake" in the prioritisation of objectives (Keeney, 2002, 1992) and may have a serious impact on the prioritisation of health threats. It is not surprising, for instance, that the weights elicited asking for direct importance were quite distinct from the ones derived from the application of conjoint analysis, as in the probabilistic inversion method proposed by Kurowicka et al. (2010), with the latter weights being more likely to represent the experts' real value trade-offs than the former.

2.6 Conduct Sensitivity and Robustness Analysis

Despite the uncertainties about impacts that are usually present in health threat assessments, which ideally would require risk modelling (see Morgan and Henrion (1992)), they are typically conceptualised as deterministic evaluations in the models that we searched. In addition, the uncertainty related to the impacts of a threat, coupled with the uncertainty related to the priorities

of policy makers (reflected on the attribute weights) require the analyst to perform a sensitivity analysis on the impact of inputs on the ranking of threats, and there are several tools available to perform it (Rios-Insua, 1990). However, many of the prioritisation models that we analysed portray a rather limited sensitivity analysis (see the eight column of Table 1 which describes how sensitivity analysis was performed in each model).

Furthermore, a full robustness analysis can identify to which extent the ranking of threats is robust to changes in the input parameters. The role of robustness analysis is considered as crucial in decision analysis, given the uncertainties involved in assessments and decision making (Roy, 2010). We have identified only one robustness analysis in the sample of models that we considered, the one developed by Garner et al. (2015).

While this conceptual critique discussed the issues most commonly observed in *ad hoc* multi-criteria health threat assessments, policy makers and experts may question the actual impact of such problems on the ranked order of options. In the following section we assess the possible impacts of specific assumptions made in the rankings of the FAO and RKI models using simulation.

3 SIMULATION OF MODELING ASSUMPTIONS IN HEALTH THREAT PRIORITISATIONS

Simulation has frequently been used in multi-criteria analysis, either for dealing with uncertainties of model parameters (e.g. Butler et al.(1997)) or, instead, for comparing idealized models against non-idealized ones. Within this latter purpose, Stewart (1996) compared the robustness of generic multi-attribute value models to those with missing attributes or preferential dependences, Durbach and Stewart (2009) simulated how simplified models deal with decision under uncertainty against the classical expected utility model, Keisler (2008) evaluated the impact of different weighting rules in portfolio multi-criteria analysis, and Jimenez et al. (2009) considered the assessment of alternatives with some missing performances. To the best of our knowledge, no simulation study has been conducted to assess the impact of modelling assumptions in health threat prioritisations. We suggest a set of guidelines in Table 2 on how these modelling issues might be assessed via simulation studies (the ones indicated with a star (*) in the table will be illustrated in this section).

In the two studies that are presented here, we have used Monte-Carlo simulation to test the parametric stability of the rankings, running 1,000 iterations and ensuring the outputs showed convergence (for the mean values, with a tolerance of 3% and a confidence level of 95%). For the FAO model we simulated four modelling assumptions that we discussed previously: the limited number of levels for attributes in the scoring functions, the presence of preferential dependencies among attributes, the use of inadequate elicitation protocols for weighting, and the ambiguity of

attribute definitions. For the RKI model we simulated only the limited number of levels for attributes and the inadequate elicitation of weights.

The analyses focus was on changes of rankings, instead of scores of threats, as the former is typically more relevant as input in health decision processes. For the two studies we assume that the evidence about the impacts of each health threat is adequate, and any changes of ranking are due only to modelling assumptions. Therefore, we are not questioning the quality of the evidence about the threats employed in those two assessments.

Table 2. Guidelines for simulation studies modelling issues in health threat prioritisations.

Modelling Step Issue	**Specific Modelling Issue**	**Possible Simulation**
Misspecification of fundamental objectives	Absence of a fundamental objective and its respective attribute	Include a new attribute x_i (i = N +1) and assess the impact on the ranking by varying its weight in Eq. 2.
	Presence of redundant objectives and respective attributes	Vary weight of the double counted attribute in Eq. 2, from the current weight down to zero, and assess its impact on the ranking.
Misspecification of attributes	Ambiguity in the assessment of, or classification of, performances	Vary up and downwards the impacts of each threat x_i^k within reasonable bounds, and assess consequent impact on the ranking (*).
	Inadequate use of a proxy attribute instead of using a more direct one	Vary weight of the proxy attribute from current weight downwards to zero, as it tends to receive excessive weight (Fischer et al., 1987), and assess its impact on the ranking.
Inadequate elicitation of value/utility functions	Inadequate elicitation of continuous value/utility functions	Consider different (monotonic) shapes of value function $V_i(x_i)$ by using a parametric function and varying its parameter(s), then assess their impact on the ranking.
	Inadequate elicitation of discrete value/utility functions	Range score associated with each attribute level of a value function $V_i(x_i)$, within reasonable bounds, then assess their impact on the ranking (*).
Inadequate elicitation of attribute weights	Incorrect assumptions of preference structure	Compare the assumption of a simple weighted sum in Eq. 2 against the aggregation required by the preference structure and assess its impact on the ranking (*).
	Incorrect elicitation of attribute weights	Vary weights w_i in Eq. 2 considering only ordinal information, or alternatively, no prior information and assess the impact on the ranking (*).

Obs.: (*) Simulation employed in the two case studies.

3.1 Simulation 1: The FAO Ranking of Parasites

The first multi-criteria model we analyse is the one by the FAO to rank food-borne parasites (FAO, 2014). This model was developed with 21 international experts supported by a team of risk analysts, who identified the attributes to be assessed, evaluated and scored the impact of each parasite on these attributes, and weighted these attributes in terms of their "importance". The published report presents details about the project and the results of this analysis with the overall ranking of parasites.

In this model, let C_{i} be the dis-value function associated with each i-th attribute shown in Table A2 (except the last row), w_{i}^{j} being the i-th attribute weight for the j-th group of experts (see Table A3) and b_{i}^{k} the bin of the i-th attribute for the k-th parasite. Then the overall scores for a given k-th parasite $\boldsymbol{C^k}$ is calculated by:

$$\boldsymbol{C^k} = w_1^j\left(b_1^k\right) + w_2^j C_2\left(b_2^k\right) + w_{345}^j C_{345}\left(b_3^k, b_4^k, b_5^k\right) + w_6^j C_6\left(b_6^k\right) + \cdots + w_9^j C_9\left(b_9^k\right) \quad \text{(Eq. 3)}$$

With:

$$C_{345}\left(b_3^k, b_4^k, b_5^k\right) = C_3\left(b_3^k\right)\left[1 - \frac{C_5\left(b_5^k\right)}{100}\right] + C_4\left(b_4^k\right)C_5\left(b_5^k\right)/100 \quad \text{(Eq. 4)}$$

The bins are scored between 0 (Bin 0) and 100 (highest Bin) for each attribute and are equally spaced in value.

We created replica parasites to conduct the simulation using goal programming (Jones and Tamiz, 2010), as the individual expert judgments were not presented in the report (see Online Appendix I). A replica parasite is one that would be classified in the bins of each attribute in a way that it would provide overall values similar to the ones found in the report (see Table A5). We are thus replicating an individual expert reaching the actual assessments and the resulting ranking provided in the FAO report (see Table A6).

The first four simulations described below (Sections 3.1.1 to 3.1.4) assume a single expert making an individual assessment of the threats. The fifth simulation to be presented (Section 3.1.5) assumes a group of experts, each one making an individual assessment of the threats, with the individual results then aggregated. This latter analysis tries to assess the impact of modelling issues on the individual rankings, as well as the impact of the ambiguity of attributes on the group's results. Next, we detail the simulation model and the results for the assumptions analysed.

3.1.1 Scoring Functions with Limited Number of Attribute Levels

The scoring functions for the attributes in FAO's model feature stepwise increases from one category to the next (for example a jump in dis-value from 25 to 50 when the disability weighting reaches 0.03, as depicted in Figure 3 – attribute C_3). What would be the impact on the ranking of the threats if we considered a less steep scoring function? To answer this question we assumed in our simulation model a continuous and monotonic function, which is linear between two consecutive levels (e.g. V_3 in the same figure) and uniformly distributed: V_3(Bin0)=0, V_3(Bin1) ~U(0.001,25), V_3(Bin2) ~U(25.001,50), V_3(Bin3) ~U(50.001,75), and V_3(Bin4) ~U(75.001,100). We are hence using a conservative assumption in the simulation model, which favours the existing model as much as possible. We have created similar scoring functions for the other attributes. The only exception was for attribute C_2 where the index is an integer and so we kept the original deterministic scores for all bins except Bin3, where we defined a distribution V_2(Bin3) ~U(75,99.999).

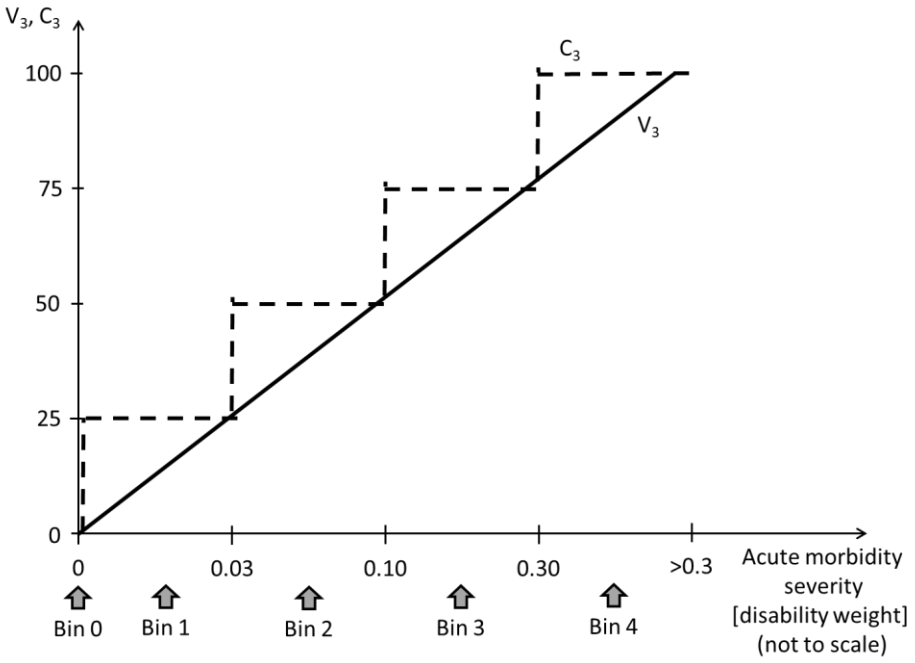

Figure 3: Score function for Attribute 3 (C3 = FAO model; V3 = simulation model).

We then compared the actual scores from the FAO model (baseline results in Table A5) with those of our simulation (Table A7, column R1), which are aggregated into an overall simulated score $\boldsymbol{V}^k$ (using the same Eqs. 3 and 4 above, replacing each C_i by the respective V_i). We calculated the Spearman's rank order correlation (ROC) between the FAO ranking and the simulated ranking (based on the overall simulated score $\boldsymbol{V}^k$for each k-th threat) for every iteration of the simulation. The mean ROC between the two rankings was 0.9688 with a range of [0.9226, 0.9956].

We also compared the percentage of iterations where the simulated rankings were the same as those of the FAO model (Table A7, column R1). For instance, the parasite ranked 1st in the FAO model was also ranked 1st in 79.1% of the iterations of the simulation, however dropping down to the 3rd position in some instances (highlighted in bold in the table). Notice that the positions in the simulated ranks vary rather drastically after the 5th parasite, and even top ranked parasites have a widespread position in the simulated results (e.g. Parasite 4, ranging from the 1st to 6th position).

3.1.2 Neglecting Preferential Dependences

As we discussed previously, the FAO model assumes that attribute C_1 is preferentially independent from C_3 and C_4 (Table A2). What is the impact of this assumption on the overall ranking of parasites? To analyse this issue, we compare the FAO ranking with an alternative model, where the ordinal ranking of levels considers simultaneously the number of cases (C_1) and morbidity (C_4) and follows the strict preferences displayed by the arrows in Figure 4 (for example, the combination (x_1=0, x_3=0) is always less serious than (x_1=1, x_3=0) and thus always receives a lower score). This suggested ordering, in effect displaying dependencies between attributes, may represent a more logical sequence than the original one (Figure 2).

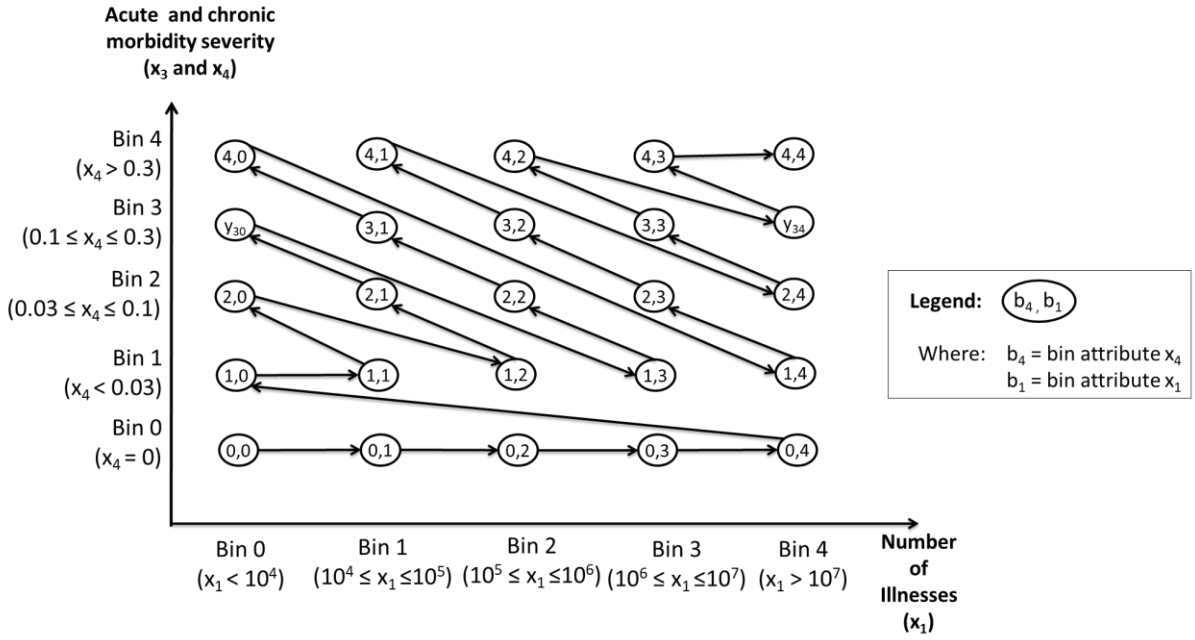

Figure 4: Strict preference relations assumed for the preferential dependence simulation in the FAO model.

The bounds of the dis-value scales are $V_1(x_1^*) + V_{345}(x_3^*, x_4^*) = 200$ and $V_1(x_1^0) + V_{345}(x_3^0, x_4^0) = 0$. The scale was thus built up, given the 24 possible ordered combinations, such as $V_{13}(x_1 = 4, x_3 = 4) \sim U(200\times23/24 + \Delta, 200)$; $V_{13}(x_1 = 4, x_3 = 3) \sim U(200*22/24+ \Delta, 200*23/24)$, …, $V_{00}(x_1 = 0, x_3 = 0) \sim U(0,$

200*1/24), with Δ = 0.0001. The same type of uniform distributions were defined for V_{14}, and V_{1345} was given by a convex combination of them: $V_{1345} = (1 - C_5)V_{13} + C_5 V_{14}$.

The inclusion of preferential dependences had some impact on the ROC between the FAO ranking and the simulated ranking. The mean ROC was 0.9388, with a range of [0.9061,0.9661]. Table A7 (column R2) presents the simulation results. Notice that the assumption of preferential independence has a major impact on the parasites ranked in the bottom half of the ranking – as they were placed in lower bins of attributes 3 and 4 and thus over-valued in the FAO scoring system.

3.1.3 Modelling Attribute Weights as Measuring Direct Importance

The elicitation of weights in the FAO model adopted the concept of direct importance instead of value trade-offs (which would require an elicitation protocol that considered the range of the attributes). What would be the impact on the ranking of threats, if the FAO elicitation protocol of weights had indeed affected the values that the experts provided for such parameters? We again make a conservative assumption, which favours the FAO model as much as possible, that the order of the baseline weights in Table A3 is always preserved, i.e.:

$$w_1 = w_{345} > w_6 > w_2 > w_8 = w_9 > w_7$$

The weights in the simulation are randomly generated for every iteration using the procedure suggested by Butler et al. (1997) for rank order weights, but following the order described above. We kept the scores as the original baselines, to isolate the effect of weighting in the simulation, obtaining a mean ROC of 0.9485 and a range of [0.8696,0.9974]. As Table A7 shows (column R3), the impacts on the ranking are also significant in terms of differences of allocation.

3.1.4 Considering Modelling Assumptions Simultaneously

We now simulate the scales, preferential dependences, and weights, simultaneously, and compare resultant rankings against the FAO rankings. The ROC mean in this case falls to 0.9059, with a range of [0.7713, 0.9696]. Table A7 (Column R4) shows the impact of these three modelling assumptions on the parasite rankings. Once again there are wide variations in the simulated rankings, for instance for Parasite 4, which may fall to the 9th ranked position, and for Parasite 1, which fell to the 4th position in some iterations.

We examined the lowest ROC for the FAO rankings, as well as the lowest ROC for the top 15, 10 and 5 parasites[5], as shown in Table A8. Notice that there are some instances where the highest scored

[5]While calculating ROC is not advisable below 15 data points, this analysis allows us to identify the most dramatic changes in the top tiers of the ranking, positions that are particularly important for policy making.

parasite in the FAO model takes only the third place in the simulation (row *Top 5*). There are also other differences, for instance Parasite 15 in the FAO ranking is now placed in 6th position in the simulation and Parasite 4 is now in the 10th position (row *All 24*).

We proceeded to perform dominance analysis between each pair of t-th and s-th parasites, considering their simulated overall score (V^t and V^s) for each iteration, where the dominance index d is defined as:

$$\begin{cases} d = 1, if\ V^t > V^S\ ; with\ t\ \neq s\ and\ t, s = 1, 2, \ldots, 24. \\ d = 0, if\ V^t \leq V^S\ ; with\ t\ \neq s\ and\ t, s = 1, 2, \ldots, 24. \end{cases}$$

The mean of the dominance index, $\bar{d}$, is shown in Table A9. The table assess whether a higher ranked threat might be dominated by a lower ranked threat, which happens whenever $\bar{d}$< 100%. Note that such cases are prevalent in the table. For instance, there were iterations where Parasite 1 in the FAO ranking was dominated by Parasite 5, and others where Parasite 6 was dominated by Parasite 24 (highlighted in bold in the same table). In general, the local ranking is unstable among parasites in 3 or 4 subsequent positions below (e.g. Parasite 7, against Parasites 8, 9, 10, and 11 – see highlighted range in the table). In addition, the ranking of parasites in adjacent positions are also unstable, as identified by values in the diagonal portion of the table (e.g. Parasite 2 dominates Parasite 3 in only 86.6% of the iterations).

3.1.5 Developing Ambiguous Attributes

Up to now we have considered a scenario where a single evaluator assessed the parasites. However, in the FAO evaluation 21 evaluators independently assessed the health threats. What impact can ambiguous attributes have on the ranking of parasites when assessed by multiple experts?

To simulate the ambiguity of attributes we assumed that different experts, despite having the same underlying evidence, might choose distinctive bins when assessing each parasite. Notwithstanding evidence that qualitative labels for probability elicitation have very wide quantitative ranges among subjects (Budescu and Wallsten, 1985; Wallsten et al., 1986) we again take a conservative position. We assume in the simulation model that there is a 20% chance that an assessor will chose a category above or below the one originally selected for qualitative attributes (e.g. Bin1 in C_9 in Table A2), and that there is a 1% chance that s/he will select a category above or below the original for quantitative attributes where the bounds overlap (e.g. C_1 between Bin2 versus Bins 1 or 3 in the same table).

To emulate the single-expert evaluation process, in this simulation the bins are initially selected according to the ambiguity of the labels: for those bins without ambiguity (e.g. Bin0 in C_1) the original bins have been used (Table A2), for those with ambiguity the bins are determined following

the discrete distributions described in the previous paragraph. Then 1000 iterations of the model described in the previous section are performed. This is repeated 21 times, each one representing an expert (the same seed for the random number generator is used for every simulation).

Aggregate results from these 21 simulations are shown in Table A7 (column R5). It shows the range of the proportion of iterations in which the threat was positioned in that ranking and the mean of these proportions across the 21 simulations. For instance, considering the first threat, one simulated expert positioned this parasite as the first one in only 31.9% of its iterations, while for another simulated expert this figure was 99.1%. The mean proportion among the 21 simulated experts for this parasite was 79.0%. In this analysis, while the average mean ROC has decreased only slightly from the previous model to 0.8984, as well as the lower bound of the range of ROCs [0.7609,0.9696], individual simulated results for the threats are significantly divergent from the FAO ranking. In particular, both the percentage agreement between the FAO model and the simulated model has decreased considerably for most parasites, and the spreads from the highest to the lowest rank for each threat have all heavily increased. These increased spreads are result of the ambiguity of levels in combination with the other modelling assumptions described in Section 3.1.4.

3.2 Simulation 2: the RKI Disease Prioritisation Ranking

The second multi-criteria model analysed is the RKI disease prioritisation system, which assesses 127 diseases, the rankings of which is fed into a set of priority blocks: Highest, High, Medium and Low priority (Balabanova et al., 2011; Krause, 2008). This study was conducted for the German public health institute and involved a group of 10 external experts to score the diseases according to ten attributes. An additional group of 86 experts were employed to weight the attributes. Details about the project and the analysis of results were published in PloS ONE.

The RKI model provides us with the opportunity to assess the robustness of a model which exhibits attributes with less levels than the FAO model, coupled with similar issues in weight elicitation practices. The overall score for disease k, $\boldsymbol{C^k}$is calculated by:

$$\boldsymbol{C}^k = w_1 C_1\left(b_1^k\right) + w_2 C_2\left(b_2^k\right) + \cdots + w_{10} C_{10}\left(b_{10}^k\right) \qquad \text{(Eq. 5)}$$

where C_i denotes the dis-value function associated with the i-th attribute, w_i is the weight assigned to each attribute and b_i^k represents the bin of the i-th attribute for disease k. A comprehensive breakdown of each disease score and weights was available for this model online[6].

3.2.1 Scoring Functions with Limited Number of Attribute Levels

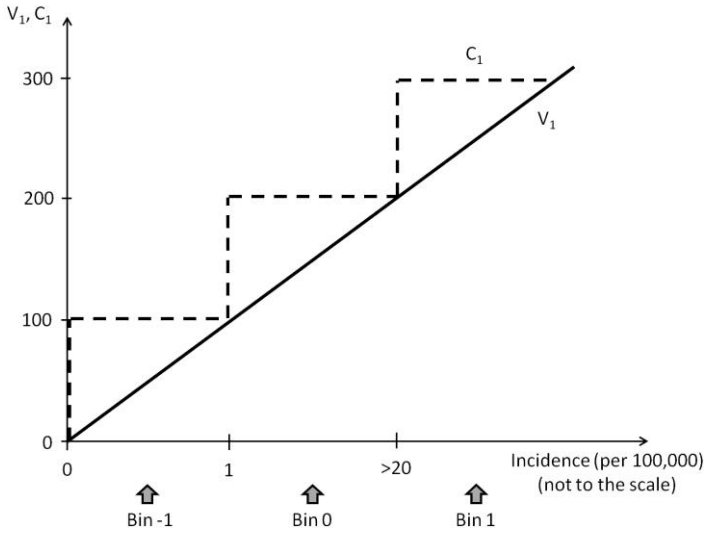

Figure 5: Dis-value function for Attribute 1 (C1 = RKI model; V1 = simulation model).

The attribute scoring functions of the RKI model exhibit similar step increases to the FAO model, where for each attribute there were three possible bin values (-1,0,1) (see Table A1). An example of this is depicted in Figure 5 (dis-value function C_1), where disease incidence rates score a bin value of 1 after the threshold of 20/100,000 is crossed. Similar to the FAO simulations, we compare the step-increasing RKI model against a value function that takes a continuous linear distribution between consecutive score bins. The -1, 0, 1 scales were converted into [0,100], [100,200], and [200,300] to avoid using bivalent scales for single valence attributes. This is performed for all attributes, as scoring labels for the model do not indicate a discrete value and are all of continuous ranges. As shown in Figure 5, the simulated values are uniformly distributed, with: V_1(Bin-1) ~U(0.0001,100), V_1(Bin0) ~U(100.0001,200), V_1(Bin1) ~U(200.0001,300). Thus, the simulation, similarly to the FAO one, adopts a conservative assumption, favouring the existing model as much as possible.

For all iterations, we calculated the ROC between the RKI ranking and the simulated ranking based on the less steep scoring functions, calculated from the overall simulated score $\boldsymbol{C^k}$. The mean ROC was 0.9643 with a range of [0.9431,0.9787] (Table A10, column R1). Due to the large number of entries in the RKI model, we adopt a different approach to summarize results of our simulations,

[6]Data available at: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0025691

opting to split each disease priority block into four clusters of roughly equal size. While each priority block is not of equal length (for example, the highest block contains 26 diseases, whereas the lowest contains 17), quartile divided blocks allow us to compare diseases of perceivably differing priority levels rather than just simply by rank obtained. There is little consistency between the ranks yielded by the simulation and the original RKI ranks, with no disease cluster, bar the lowest exhibiting a case of over 10% of simulated events providing the same rank.

3.2.2 Varying the Attribute Weights

As the weighting methodology in the RKI model also used the concept of direct importance, we used the rank order weighting method similar to the FAO model analysis to simulate this issue, ensuring the following relation from the model's baseline weights is maintained:

$$w_5 > w_4 = w_6 = w_9 > w_{10} = w_1 > w_2 = w_3 > w_7 = w_8$$

By multiplying the simulated weights with the baseline scores, we obtained the results in shown in Table A10 (column R2). The resulting mean ROC was much lower than the scoring simulation at 0.8892, with an increased range of [0.6075,0.9992]. Notice the increase in discrepancy between the highest and lowest simulated rank, particularly for the Highest to Medium disease clusters, with in 11/16 instances the range being wider than 70 places, compared to six cases in the score simulation. This may be partially reflected in the mean percentage of instances where the simulated rank equated the reported rank; in all but three cases the results were lower than obtained in the score simulation.

3.2.3 Considering Modelling Assumptions Simultaneously

Finally, we consider simultaneously the impact of both conceptual assumptions in the RKI system on the overall model. In the last column of Table A10 (column R3) we report the results of the combined simulation, where the mean ROC is lower than previously at 0.8472 and the range has widened to [0.4388,0.9727]. In addition, the differences between the highest and lowest possible rank obtained in each disease cluster have increased further, with the range exceeding 100 places in 11/16 instances.

Similar to our analysis of the FAO case, we compared the existing rankings with those obtained from the lowest ROCs in the model, firstly with the full list of diseases, secondly with those allocated to the highest priority group, and finally with the top 15, 10 and 5 ranked diseases. The comparisons are displayed in Table A11, where we observe that the consistency between simulated and actual rankings decrease substantially when fewer diseases are analysed (for the top 15, 10 and 5 lowest

ROC simulated rankings, the ROC falls below zero). This highlights the extent to which even rankings in the highest tail of the RKI model can change.

To analyse the results further, Table A12 provides simulation results exclusively of the 26 diseases in the highest priority block. The simulations rarely yield the same rank for each disease, even in the instances where ordinal ranks are tested solely among the top 26 where no rank consistency exceeds 30%. For the simulation simultaneously assessing the impact of both modelling assumptions (column R3), the mean ROC falls further than the separate assessments, at 0.5832, with the largest range at [0.2150,0.8762].

Furthermore, we notice that scoring range (Table A12, column R1) or weighting elicitation issues (Table A12, column R2) do not impact rankings uniformly throughout the model, with scoring adjustments substantially reducing mean observed ROC values more in the Highest priority block compared to weighting throughout the whole model. When comparing the ROC of the entire model to just the highest priority group, the drop from 0.9643 (Table A10, column R1) to 0.6811 (Table A12, column R1) for score simulations is substantially larger than the equivalent for weights, which drops from 0.8892 (Table A10, column R2) to 0.7417 (Table A12, column R2).

To identify where the main discrepancies are occurring throughout, a similar dominance analysis to the one done for the FAO model was performed across priority blocks. The dominance index in this case is defined as 1 in events where all scores obtained in disease cluster t exceed all scores in cluster s, where t>s, and a 0 value otherwise. The mean values of this dominance index are shown in Table A13. The table portrays considerable numbers of instances of diseases in higher blocks being frequently dominated by lower ranked entries.

4 Conclusions

Multi-criteria evaluations have been increasingly employed to prioritise health threats. A number of models reported in the literature, however, lack a clear conceptual framework rooted on Multi-Criteria Decision Analysis – they are *ad hoc* models from this perspective. In this paper we critically analysed the main conceptual assumptions made in these ad hoc models, illustrating the issues with several examples from published papers and technical reports, in comparison with the best practices of multi-attribute value analysis. Our critique pointed out serious conceptual issues, which may lead to ranking of threats that do not reflect either the relative significance of the aggregate impacts, or the values and priorities of policy makers.

We also considered in depth two health threat prioritisations: the ranking of foodborne parasites developed by FAO (FAO, 2014) and the ranking of diseases developed by the RKI (Balabanova et al., 2011). For each one of them, we developed a simulation model to assess the impact of their modelling choices on the ranking of threats. While there was earlier evidence that multi-attribute linear models are sensitive to their parameters (Stewart, 1996), here we were interested instead in simulating the impact of specific modelling assumptions on the health threat rankings. In these two simulation analyses, we found significant changes in their ranking of threats. It should be a concern for both policy makers and analysts that, for instance, a threat ranked in the first position in the original ranking could have been placed in the 6th position, as well as the large spreads in ranking that the evaluation system may generate, both issues caused by weak conceptual assumptions and poor design choices made in these evaluation models[7]. In addition, it is a matter of concern that recent attempts to promote standard assessments in health threat prioritisations have promoted modelling practices that are suboptimal, such as defining weights as measurement of direct importance (ECDC, 2017; O'Brien et al., 2016) and overlapping attribute levels for quantitative indices (ECDC, 2017).

As in any type of conceptual analyses and simulation study, ours has also some important limitations. Firstly, we have compared *ad hoc* models with the best practices in multi-attribute value analysis. While there are good reasons for our choice, given the resemblance of *ad hoc* models with this type of framework, a comparison with other types of well-conceptualized multi-criteria methodologies (see Belton and Stewart(2002)) could lead to different conclusions. Secondly, in trying to simulate the modelling issues in *ad hoc* models, we made several assumptions, e.g. regarding the shape of value functions, the ranking of weights, and the clustering of threats. While they were conservative, our results rely on their choice. Thirdly, the results from the two in-depth case studies cannot be directly generalized to other *ad hoc* models (Yin, 2008), albeit, we believe, they are indicative of trends in models that have similar issues.

These limitations suggest some directions for further research. Firstly, *ad hoc* multi-criteria models could be compared with other well-known frameworks in multi-criteria analysis, for instance multi-

[7]Another conclusion from this analysis would be to stress that the mean ROCs are indeed stable, which means that "on average" the simulated scoring system produced similar rankings to those of the original prioritisation. Although this is true, ROCs are relatively insensitive to small variations and do not distinguish whether the changes are occurring in the first tier of the ranking, which typically is more relevant for policy makers, or in its bottom tiers. Furthermore, contrary to standard investment problems, in which expected values matter (as it is assumed that such decisions are made repeatedly) these health threat rankings are typically employed in one-off studies with the ranking recommendations provided to policy makers.

attribute utility theory (Keeney and Raiffa, 1993), in which risk attitude is formally modelled, or outranking methods (Roy, 1996), which enable the representation of more sophisticated preference relations. Secondly, within a MAVA framework, it would be interesting to assess the extent to which different modelling issues, other than the ones we examined here, would impact on the rankings of the simulated models. In addition, the issue of ambiguity in the definition of attributes is very relevant as most of these assessments are made by multiple experts and many attributes are highly ambiguous. Future studies could explore other ways of simulating the impact of ambiguity in more sophisticated ways. In the same vein, the elicitation of weights as measurements of direct importance is prevalent in these models and could be simulated in more detail. Thirdly, extending this type of analysis to other fields, applications and types of multi-criteria prioritisation may highlight similar issues or shed light on new relevant ones. Furthermore, comparisons of these models with standard risk analysis models could provide fruitful insights.

We conclude the paper on a positive note. While the issues we analysed here are frequently observed in multi-criteria health threat prioritisation, there are some examples of sound practice in this field, such as the prioritisation of diseases for the pig industry in Australia (Brookes et al., 2014) and of emerging animal health threats in the UK (Del Rio Vilas et al., 2013b), among others. This has been accomplished either by assembling interdisciplinary teams of health experts and decision analysts, or by health experts trained in multi-criteria decision analysis. Recent attempts to establish best practices in health prioritisations are also welcome (Brookes et al., 2015; Marsh et al., 2016; O'Brien et al., 2016) as well as the use of participative processes to engage with the experts and share their knowledge and expertise (Dias et al., 2018; Franco and Montibeller, 2010). We hope that the paper also helps to sensitize more experts involved in these assessments to the importance of a careful design of multi-criteria evaluation models. It is crucial that they are robust from a methodological perspective if they are being used to inform decision processes in health risk management.

References

Angelis, A., Kanavos, P., Montibeller, G., 2017. Resource Allocation and Priority Setting in Health Care: A Multi-criteria Decision Analysis Problem of Value? Global Policy 8, 76–83. https://doi.org/10.1111/1758-5899.12387

Balabanova, Y., Gilsdorf, A., Buda, S., Burger, R., Eckmanns, T., Gärtner, B., Groß, U., Haas, W., Hamouda, O., Hübner, J., Jänisch, T., Kist, M., Kramer, M.H., Ledig, T., Mielke, M., Pulz, M., Stark, K., Suttorp, N., Ulbrich, U., Wichmann, O., Krause, G., 2011. Communicable Diseases Prioritized for Surveillance and Epidemiological Research: Results of a Standardized Prioritization Procedure in Germany, 2011. PLoS ONE 6, e25691. https://doi.org/10.1371/journal.pone.0025691

Baltussen, R., Youngkong, S., Paolucci, F., Niessen, L., 2010. Multi-criteria decision analysis to prioritize health interventions: Capitalizing on first experiences. Health Policy 96, 262–264. https://doi.org/10.1016/j.healthpol.2010.01.009

Bana e Costa, C.A., De Corte, J.-M., Vansnick, J.-C., 2012. MACBETH. International Journal of Information Technology & Decision Making 11, 359–387. https://doi.org/10.1142/S0219622012400068

Bana e Costa, C.A., Lourenço, J.C., Chagas, M.P., Costa, J.C.B. e, 2008. Development of Reusable Bid Evaluation Models for the Portuguese Electric Transmission Company. Decision Analysis 5, 22–42. https://doi.org/10.1287/deca.1080.0104

Barcus, A., Montibeller, G., 2008. Supporting the allocation of software development work in distributed teams with multi-criteria decision analysis. Omega 36, 464–475.

Batz, M.B., Doyle, M.P., Morris, J.G., Painter, J., Singh, R., Tauxe, R.V., Taylor, M.R., Wong, D.M.A.L.F., for the Food Attribution Working Group, 2005. Attributing Illness to Food. Emerging Infectious Diseases 11, 993–999. https://doi.org/10.3201/eid1107.040634

Belton, V., Stewart, T.J., 2002. Multiple Criteria Decision Analysis: An Integrated Approach. Springer, Norwell, MA.

Brookes, V.J., Hernández-Jover, M., Cowled, B., Holyoake, P.K., Ward, M.P., 2014. Building a picture: Prioritisation of exotic diseases for the pig industry in Australia using multi-criteria decision analysis. Preventive Veterinary Medicine 113, 103–117. https://doi.org/10.1016/j.prevetmed.2013.10.014

Brookes, V.J., Vilas, V.J.D.R., Ward, M.P., 2015. Disease prioritization: what is the state of the art? Epidemiology & Infection 143, 2911–2922. https://doi.org/10.1017/S0950268815000801

Budescu, D.V., Wallsten, T.S., 1985. Consistency in interpretation of probabilistic phrases. Organizational Behavior and Human Decision Processes 36, 391–405. https://doi.org/10.1016/0749-5978(85)90007-X

Butler, J., Jia, J., Dyer, J., 1997. Simulation techniques for the sensitivity analysis of multi-criteria decision models. European Journal of Operational Research 103, 531–546. https://doi.org/10.1016/S0377-2217(96)00307-4

Cardoen, S., Van Huffel, X., Berkvens, D., Quoilin, S., Ducoffre, G., Saegerman, C., Speybroeck, N., Imberechts, H., Herman, L., Ducatelle, R., Dierick, K., 2009. Evidence-Based Semiquantitative Methodology for Prioritization of Foodborne Zoonoses. Foodborne Pathogens and Disease 6, 1083–1096. https://doi.org/10.1089/fpd.2009.0291

Collineau, L., Carmo, L.P., Endimiani, A., Magouras, I., Müntener, C., Schüpbach-Regula, G., Stärk, K.D.C., 2018. Risk Ranking of Antimicrobial-Resistant Hazards Found in Meat in Switzerland. Risk Analysis 38, 1070–1084. https://doi.org/10.1111/risa.12901

Cox, R., Sanchez, J., Revie, C.W., 2013. Multi-Criteria Decision Analysis Tools for Prioritising Emerging or Re-Emerging Infectious Diseases Associated with Climate Change in Canada. PLoS ONE 8, e68338. https://doi.org/10.1371/journal.pone.0068338

Dahl, V., Tegnell, A., Wallensten, A., 2015. Communicable Diseases Prioritized According to Their Public Health Relevance, Sweden, 2013. PLOS ONE 10, e0136353. https://doi.org/10.1371/journal.pone.0136353

Del Rio Vilas, V.J., Burgeno, A., Montibeller, G., Clavijo, A., Vigilato, M.A., Cosivi, O., 2013a. Prioritization of capacities for the elimination of dog-mediated human rabies in the Americas: building the framework. Pathogens and Global Health 107, 340–345. https://doi.org/10.1179/2047773213Y.0000000122

Del Rio Vilas, V.J., Montibeller, G., Franco, L.A., 2011. Letter to the editor: Prioritisation of infectious diseases in public health: feedback on the prioritisation methodology. Eurosurveillance 16, 1–2. https://doi.org/10.1179/2047773213Y.0000000122

Del Rio Vilas, V.J., Voller, F., Montibeller, G., Franco, L.A., Sribhashyam, S., Watson, E., Hartley, M., Gibbens, J.C., 2013b. An integrated process and management tools for ranking multiple emerging threats to animal health. Preventive Veterinary Medicine 108, 94–102. https://doi.org/10.1016/j.prevetmed.2012.08.007

Dias, L.C., Morton, A., Quigley, J., 2018. Elicitation - The Science and Art of Structuring Judgement. Springer.

Dickert, S., Västfjäll, D., Kleber, J., Slovic, P., 2012. Valuations of human lives: normative expectations and psychological mechanisms of (ir)rationality. Synthese 189, 95–105. https://doi.org/10.1007/s11229-012-0137-4

Dillon-Merrill, R.L., Parnell, G.S., Buckshaw, D.L., Hensley, W.R., Caswell, D.J., 2008. Avoiding Common Pitfalls in Decision Support Frameworks for Department of Defense Analyses. Military Operations Research 13, 19–31. https://doi.org/10.5711/morj.13.2.19

Domanović, D., Cassini, A., Bekeredjian-Ding, I., Bokhorst, A., Bouwknegt, M., Facco, G., Galea, G., Grossi, P., Jashari, R., Jungbauer, C., Marcelis, J., Raluca-Siska, I., Andersson-Vonrosen, I., Suk, J.E., 2017. Prioritizing of bacterial infections transmitted through substances of human origin in Europe. Transfusion 57, 1311–1317. https://doi.org/10.1111/trf.14036

Durbach, I.N., Stewart, T.J., 2009. Using expected values to simplify decision making under uncertainty. Omega 37, 312–330.

Dyer, J.S., Sarin, R.K., 1979. Measurable Multiattribute Value Functions. Operations Research 27, 810–822. https://doi.org/10.1287/opre.27.4.810

ECDC, 2017. ECDC Tool for the Prioritization of Infectious Disease Threats. European Centre for Disease Prevention and Control, Stockholm.

ECDC, 2015. Literature review: Best practices in ranking emerging infectious disease threats. European Centre for Disease Prevention and Control.

FAO, 2014. Multicriteria-Based Ranking for Risk Management of Foodborne Parasites, Microbiological Risk Assessment Series. Food and Agriculture Organization of the United Nations, Rome, Italy.

FAO/WHO, 2014. Ranking of low moisture foods in support of microbiological risk management (Preliminary Report). Food and Agriculture Organization of The United Nations & World Health Organization, Rome, Italy.

Fischer, G.W., Damodaran, N., Laskey, K.B., Lincoln, D., 1987. Preferences for proxy attributes. Management Science 33, 198–214.

Franco, L.A., Montibeller, G., 2011. Problem structuring for Multicriteria Decision Analysis interventions, in: Wiley Encyclopedia of Operations Research and Management Science. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Franco, L.A., Montibeller, G., 2010. Facilitated modelling in operational research. European Journal of Operational Research 205, 489–500. https://doi.org/10.1016/j.ejor.2009.09.030

Garner, M.J., Carson, C., Lingohr, E.J., Fazil, A., Edge, V.L., Waddell, J.T., 2015. An Assessment of Antimicrobial Resistant Disease Threats in Canada. PLOS ONE 10, e0125155. https://doi.org/10.1371/journal.pone.0125155

Greenberg, M., Haas, C., Cox, A., Lowrie, K., McComas, K., North, W., 2012. Ten Most Important Accomplishments in Risk Analysis, 1980–2010. Risk Analysis 32, 771–781. https://doi.org/10.1111/j.1539-6924.2012.01817.x

Hammond, J.S., Keeney, R.L., Raïffa, H., 1999. Smart Choices: A Practical Guide to Making Better Decisions. Harvard Business Press.

Havelaar, A.H., van Rosse, F., Bucura, C., Toetenel, M.A., Haagsma, J.A., Kurowicka, D., Heesterbeek, J. (Hans) A.P., Speybroeck, N., Langelaar, M.F.M., van der Giessen, J.W.B., Cooke, R.M., Braks, M.A.H., 2010. Prioritizing Emerging Zoonoses in The Netherlands. PLoS One 5. https://doi.org/10.1371/journal.pone.0013965

Humblet, M.-F., Vandeputte, S., Albert, A., Gosset, C., Kirschvink, N., Haubruge, E., Fecher-Bourgeois, F., Pastoret, P.-P., Saegerman, C., 2012. Multidisciplinary and Evidence-based Method for Prioritizing Diseases of Food-producing Animals and Zoonoses. Emerg Infect Dis 18, e1. https://doi.org/10.3201/eid1804.111151

Jiménez, A., Mateos, A., Ríos-Insua, S., 2009. Missing consequences in multiattribute utility theory. Omega 37, 395–410. https://doi.org/10.1016/j.omega.2007.04.003

Jones, D., Tamiz, M., 2010. Practical goal programming. Springer, New York; London.

Kadohira, M., Hill, G., Yoshizaki, R., Ota, S., Yoshikawa, Y., 2015. Stakeholder prioritization of zoonoses in Japan with analytic hierarchy process method. Epidemiol. Infect. 143, 1477–1485. https://doi.org/10.1017/S0950268814002246

Keeney, R.L., 2013. Identifying, prioritizing, and using multiple objectives. EURO J Decis Process 1, 45–67. https://doi.org/10.1007/s40070-013-0002-9

Keeney, R.L., 2002. Common mistakes in making value trade-offs. Operations Research 50, 935–945.

Keeney, R.L., 1992. Value-Focused Thinking: A Path to Creative Decisionmaking. Harvard University Press, Cambridge, MA.

Keeney, R.L., Gregory, R.S., 2005. Selecting attributes to measure the achievement of objectives. Operations Research 53, 1–11.

Keeney, R.L., Raiffa, H., 1993. Decisions with Multiple Objectives: Preferences and Value Trade-Offs, 2nd ed. Cambridge University Press, Cambridge.

Keisler, J.M., 2008. The value of assessing weights in multi-criteria portfolio decision analysis. J. Multi-Crit. Decis. Anal. 15, 111–123. https://doi.org/10.1002/mcda.427

Krantz, D.H., Luce, R.D., Suppes, P., Tversky, A., 1971. Foundations of Measurement. Academic Press, New York, NY.

Krause, G., 2008. Prioritisation of infectious diseases in public health - call for comments. Eurosurveillance 13, 1–6.

Kujawski, E., Triantaphyllou, E., Yanase, J., 2019. Additive Multicriteria Decision Analysis Models: Misleading Aids for Life-Critical Shared Decision Making. Med Decis Making 0272989X19844740. https://doi.org/10.1177/0272989X19844740

Kurowicka, D., Bucura, C., Cooke, R., Havelaar, A., 2010. Probabilistic Inversion in Priority Setting of Emerging Zoonoses. Risk Analysis 30, 715–723. https://doi.org/10.1111/j.1539-6924.2010.01378.x

Marsh, K., IJzerman, M., Thokala, P., Baltussen, R., Boysen, M., Kaló, Z., Lönngren, T., Mussen, F., Peacock, S., Watkins, J., Devlin, N., 2016. Multiple Criteria Decision Analysis for Health Care Decision Making—Emerging Good Practices: Report 2 of the ISPOR MCDA Emerging Good Practices Task Force. Value in Health 19, 125–137. https://doi.org/10.1016/j.jval.2015.12.016

Mehand, M.S., Millett, P., Al-Shorbaji, F., Roth, C., Kieny, M.P., Murgue, B., 2018. World Health Organization Methodology to Prioritize Emerging Infectious Diseases in Need of Research and Development. Emerg Infect Dis 24. https://doi.org/10.3201/eid2409.171427

Montibeller, G., von Winterfeldt, D., 2015. Cognitive and motivational biases in Decision and Risk Analysis. Risk Analysis 35, 1230–1251. https://doi.org/10.1111/risa.12360

Montibeller, G., Winterfeldt, D. von, 2018. Individual and group biases in value and uncertainty judgments, in: Dias, L.C., Morton, A., Quigley, J. (Eds.), Elicitation, International Series in Operations Research & Management Science. Springer, Cham, pp. 377–392. https://doi.org/10.1007/978-3-319-65052-4_15

Morgan, M.G., 2014. Use (and abuse) of expert elicitation in support of decision making for public policy. PNAS 111, 7176–7184. https://doi.org/10.1073/pnas.1319946111

Morgan, M.G., Henrion, M., 1992. Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis. Cambridge University Press, Cambridge.

Morton, A., 2017. Treacle and Smallpox: Two Tests for Multicriteria Decision Analysis Models in Health Technology Assessment. Value in Health 20, 512–515. https://doi.org/10.1016/j.jval.2016.10.005

Mourits, M., Oude Lansink, A., 2007. Multi-criteria decision making to evaluate quarantine disease control strategies. New approaches to the economics of plant health. Heidelberg, Germany: Springer 131–144.

Munyua, P., Bitek, A., Osoro, E., Pieracci, E.G., Muema, J., Mwatondo, A., Kungu, M., Nanyingi, M., Gharpure, R., Njenga, K., Thumbi, S.M., 2016. Prioritization of Zoonotic Diseases in Kenya, 2015. PLOS ONE 11, e0161576. https://doi.org/10.1371/journal.pone.0161576

O'Brien, E.C., Taft, R., Geary, K., Ciotti, M., Suk, J.E., 2016. Best practices in ranking communicable disease threats: a literature review, 2015. Eurosurveillance 21. https://doi.org/10.2807/1560-7917.ES.2016.21.17.30212

Pieracci, E.G., Hall, A.J., Gharpure, R., Haile, A., Walelign, E., Deressa, A., Bahiru, G., Kibebe, M., Walke, H., Ermias Belay, 2016. Prioritizing zoonotic diseases in Ethiopia using a one health approach. One Health 2, 131–135. https://doi.org/10.1016/j.onehlt.2016.09.001

Rios-Insua, D., 1990. Sensitivity Analysis in Multi-objective Decision Making, Lecture Notes in Economics and Mathematical Systems. Springer Science & Business Media, Berlin.

Roy, B., 2010. Robustness in operational research and decision aiding: A multi-faceted issue. European Journal of Operational Research 200, 629–638. https://doi.org/10.1016/j.ejor.2008.12.036

Roy, B., 1996. Multicriteria Methodology for Decision Aiding. Springer.

Stewart, T.J., 1996. Robustness of Additive Value Function Methods in MCDM. Journal of Multi-Criteria Decision Analysis 5, 301–309. https://doi.org/10.1002/(SICI)1099-1360(199612)5:4<301::AID-MCDA120>3.0.CO;2-Q

von Nitzsch, R., Weber, M., 1993. The effect of attribute ranges on weights in multiattribute utility measurements. Management Science 39, 937–943.

von Winterfeldt, D., Edwards, W., 1986. Decision Analysis and Behavioral Research. Cambridge University Press, New York, NY.

Wallsten, T.S., Budescu, D.V., Rapoport, A., Zwick, R., Forsyth, B., 1986. Measuring the vague meanings of probability terms. Journal of Experimental Psychology: General 115, 348–365. https://doi.org/10.1037/0096-3445.115.4.348

Yin, R.K., 2008. Case Study Research: Design and Methods. SAGE Publications.