



HAL
open science

Early Health Technology Assessment during Nonalcoholic Steatohepatitis Drug Development: A Two-Round, Cross-Country, Multicriteria Decision Analysis

Aris Angelis, Mark Thursz, Vlad Ratziu, Alastair O'brien, Lawrence Serfaty,
Ali Canbay, Ingolf Schiefke, Joao Bana E Costa, Pascal Lecomte, Panos
Kanavos

► **To cite this version:**

Aris Angelis, Mark Thursz, Vlad Ratziu, Alastair O'brien, Lawrence Serfaty, et al.. Early Health Technology Assessment during Nonalcoholic Steatohepatitis Drug Development: A Two-Round, Cross-Country, Multicriteria Decision Analysis. *Medical Decision Making*, 2020, 40 (6), pp.830 - 845. 10.1177/0272989x20940672 . hal-03263405

HAL Id: hal-03263405

<https://hal.sorbonne-universite.fr/hal-03263405>

Submitted on 17 Jun 2021

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

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

Early Health Technology Assessment during Nonalcoholic Steatohepatitis Drug Development: A Two-Round, Cross-Country, Multicriteria Decision Analysis

Aris Angelis¹, Mark Thursz, Vlad Ratziu, Alastair O'Brien, Lawrence Serfaty, Ali Canbay, Ingolf Schiefke, Joao Bana e Costa, Pascal Lecomte, and Panos Kanavos

Medical Decision Making
2020, Vol. 40(6) 830–845
© The Author(s) 2020



Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/0272989X20940672

journals.sagepub.com/home/mdm



Background. The assessment of value along the clinical development of new biopharmaceutical compounds is a challenging task. Complex and uncertain evidence has to be analyzed, considering a multitude of value preferences from different stakeholders. **Objective.** To investigate the use of multicriteria decision analysis (MCDA) to support decision making during drug development while considering payer and health technology assessment (HTA) value concerns, by applying the Advance Value Framework in nonalcoholic steatohepatitis (NASH) and testing for the consistency of the results. **Design.** A multiattribute value theory methodology was applied and 2 rounds of decision conferences (DCs) were organized in 3 countries (England, France, and Germany), with the participation of national key experts and stakeholders using the MACBETH questioning protocol and algorithm. A total of 51 health care professionals, patient advocates, and methodologists, including (ex-) committee members or assessors from national HTA bodies, participated in 6 DCs in the study countries. **Target Population.** NASH patients in fibrosis stages F2 to 3 were considered. **Interventions.** The value of a hypothetical product profile was assessed against 3 compounds under development using their phase 2 results. **Outcome Measures.** DC participants' value preferences were elicited involving criteria selection, options scoring, and criteria weighting. **Results.** Highly consistent valuation rankings were observed in all DCs, always favoring the same compound. Highly consistent rankings of criteria clusters were observed, favoring therapeutic benefit criteria, followed by safety profile and innovation level criteria. **Limitations.** There was a lack of comparative treatment effects, early evidence on surrogate endpoints was used, and stakeholder representativeness was limited in some DCs. **Conclusions.** The use of MCDA is promising in supporting early HTA, illustrating high consistency in results across countries and between study rounds.

Keywords

comparative research, decision conference, drug development, early HTA, MCDA, NASH, value assessment

Date received: September 18, 2019; accepted: May 13, 2020

The assessment of value along the clinical development and regulation of a new medicine is complex and involves different decision problems. It commences with manufacturers' internal decisions regarding which disease areas to invest in as part of research and development (R&D) portfolio prioritization and progresses to choices related to clinical study design, regulatory assessment of the product's

benefit-risk profile, payer evaluation of reimbursement conditions via health technology assessment (HTA),

Corresponding Author

Aris Angelis, Department of Health Policy and LSE Health, London School of Economics and Political Science, Portugal Street, London, WC2A 2AE, UK (a.n.angelis@lse.ac.uk).

and concludes with prescriber behavior based on clinical guidance and individual patient need. Arguably, for better decisions and improved transparency, the preferences of decision makers should more often be quantified and explicitly communicated.¹

The early HTA context, in which payer concerns relating to the expected value of new compounds are explored before the licensing stage, is becoming more important in Europe. This is reflected through the joint work plan of the European Medicines Agency (EMA) and the European Network for Health Technology Assessment (EUnetHTA),^{2,3} an important aspect of which is parallel consultation⁴ and EUnetHTA Early Dialogue.⁵ Early HTA empirical studies are becoming more frequent, and although a number of methods have been used to identify and evaluate technologies, the use of multicriteria decision analysis (MCDA) has been recommended as a tool to support decisions in product development.^{6,7}

Decision analysis methods and, specifically, quantitative modeling approaches such as MCDA can be used to aid medical decision making, to explicitly integrate objective measurement with value judgment while managing subjectivity transparently. This is useful for drug evaluation contexts because, although the clinical evidence concerning different treatments' performance might be objective in nature, the understanding of its value requires subjective interpretation; for example,

relating to the relevance of data for the disease of interest, the meaningfulness of improvement in health benefit, and the value trade-offs with possible risks⁸.

A method of eliciting value preferences from different stakeholder groups is decision conferencing,^{9,10} a form of face-to-face workshops guided by a facilitator while applying decision theory with multiple objectives.¹¹ In medical decision making, MCDA methods in combination with decision conferencing could be used as a tool to elicit and communicate value preferences across a number of evaluation aspects with the view to ranking a set of alternative medical options based on their value.¹²

Since European drug regulators called for more explicit and quantitative methodological approaches in the assessment of drug benefit-risk balance with well-defined evaluation criteria and the valuation of outcomes through numerical weights,¹³ a number of MCDA studies have been conducted in the licensing context.^{14–16} The EMA currently adopts and recommends the use of the “effects table” for the tabulation of the most important favorable and unfavorable effects and their uncertainty,¹⁷ a constituent step of MCDA methodologies¹⁸. Other similar structured frameworks are also implemented by drug regulators for increasing transparency in the communication of benefits and risks,¹⁹ including by the US Food and Drug Administration.²⁰

Due to various limitations of economic evaluation methods, MCDA applications in the context of HTA have intensified in recent years.^{21,22} Good practice guidelines for the use of MCDA and its adaptation to HTA have been developed,^{23–25} and different multicriteria value frameworks have been recommended.^{26–29} A number of empirical studies have been conducted, often simulating different HTA settings^{30–33} or by involving and eliciting the preferences of real decision makers and evaluators across different settings.^{34–37} Still, a number of challenges relating to the appropriate use of such methods and their integration with policy-making have been raised that would be critical for MCDA implementation.^{38,39}

Although MCDA approaches have been explored in the context of product development and treatment selection, fewer studies exist compared with the drug approval context. Relevant case studies include the use of MCDA and decision conferencing to prioritize R&D project portfolios of pharmaceutical companies and budget allocation⁴⁰ and the use of a stochastic MCDA approach for the selection of statins in primary prevention.⁴¹

In the context of early HTA, no quantitative, cross-country MCDA studies exist. More broadly, in the context of health care evaluation, there have been no

Department of Health Policy and LSE Health, London School of Economics and Political Science, London, UK (AA, PK); Imperial College Healthcare NHS Trust and Imperial College London, London, UK (MT); Université Pierre et Marie Curie and the Hôpital Pitié Salpêtrière Medical School, Paris, France (VR); Royal Free London NHS Foundation Trust and University College London, London, UK (AO); Hautepierre Hospital, University of Strasbourg, Strasbourg, France (LS); Department of Gastroenterology, Hepatology and Infectious Diseases, University Magdeburg, Magdeburg, Germany (AC); Department of Internal Medicine, Ruhr-University Bochum, Bochum, Germany (IS); Decision Eyes, Lisbon, Portugal (JB); Novartis Pharma AG, Basel, Switzerland (PL). The authors declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: AA, AC, IS, and PK declare no conflict of interest. MT, VR, AO, and JBC declare honoraria from Novartis as reimbursement for their participation in the decision conference meetings that were part of work leading to this article. LS declares a grant from Gilead Sciences and consulting fees from AbbVie, Gilead Sciences, Novartis, Pfizer, Sanofi, and Theratechnologies Inc. PL is an employee of Novartis. The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Financial support for this study was provided by a research grant from Novartis. The funding agreement ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report. The following author is employed by the sponsor: PL. The study sponsor did not have any influence on the design, evidence collection, or analysis of the study.

validating MCDA studies testing the consistency of their results by repeating the exercise with different groups of participants.

In this study, we investigate the use of the Advance Value Framework (AVF), a multicriteria value framework,²⁷ to assess the value of compounds in clinical development for the treatment of nonalcoholic steatohepatitis (NASH). A key objective is to test the consistency of the results, both between and within countries, by conducting 2 rounds of decision conferences (DCs). An early-stage HTA scope is adopted in an attempt to make the best use of available evidence at the time of the study, with the collection of views from key stakeholders across 3 countries (England, France and Germany). Given the compounds' early clinical development stage, a set of hypothetical assumptions are imposed in some cases on their performance for the purpose of enabling their comparative assessment. The output generated aims to engage key experts and decision makers in discussions around compound value at this early stage in their development, to communicate their value prospects while aligning with increasing clinical evidence availability, rather than inform clinical or policy decisions.

Methods

Methodological Framework and Overall Process

The AVF is based on multiattribute value theory (MAVT)^{10,42} and comprises 5 distinct phases: a) problem structuring, b) model building, c) model assessment, d) model appraisal and e) development of action plans.²⁵ The AVF was operationalized using a decision support system (M-MACBETH) enabling the use of graphics to build a model of values, acting as a facilitation tool to inform both the structuring phases ("a," "b") and the evaluation phases ("c," "d")⁴³; the development of action plans phase ("e") was not conducted because of the early development phase of the compounds.

The study consisted of 2 rounds. The first took place between November 2016 and June 2017 and included all 4 phases (a–d) with the involvement of 26 participants in 3 DCs (1 in each country); the second took place between November 2017 and June 2018, acting as a follow-up to validate the results obtained in the first round by conducting the 3 latter phases of the methodological process (b–d) and involving 25 participants, the purpose being to elicit additional preferences through 3 further DCs in the same study countries. Both rounds used an identical methodological approach in relation to value preference elicitation and construction.

Following problem structuring (phase a), as part of model building (phase b), an extensive review of the clinical literature was conducted to understand the clinical endpoints of interest and in consultation with a clinical hepatologist, co-author of the study (M.T.), a preliminary version of a NASH value tree was developed alongside the collection of the appropriate performance data, which was validated at the beginning of each DC. During model assessment (phase c) and model appraisal (phase d), DC participant value preferences were elicited and were used to inform, first, compound scoring against the evaluation criteria and, second, criteria weighting. The process was completed by analyzing the results, including sensitivity analysis.

Phase a. Problem Structuring: Clinical Practice and Scope

Nonalcoholic fatty liver disease (NAFLD) is the most common chronic liver disease worldwide, affecting as many as 30% of adults and 70% to 80% of those who are obese and have type 2 diabetes.⁴⁴ NAFLD represents a histologic spectrum of conditions, ranging from simple steatosis to NASH and cirrhosis. Many with simple steatosis do not develop significant liver disease, however NASH can progress to cirrhosis, hepatocellular carcinoma and end-stage liver disease; it has become the second leading aetiology of liver disease among adults awaiting liver transplantation in the United States.⁴⁵ NASH is projected to become the leading indication for liver transplantation within the next decade (for more clinical information see the Supplementary Appendix).

In the context of this study, a simulation exercise was undertaken focusing on the assessment of the overall value of compounds in clinical development for the treatment of NASH, by adopting an early HTA perspective focusing on market access and coverage decisions with the involvement of key stakeholders, including HTA experts and proxy decision makers.

Phase b. Model Building: The Advance Value Tree Adaptation, Alternative Treatments and Evidence, Attribute Ranges and References

1. The Advance Value Tree (AVT) adaptation for NASH

As part of the AVF, the AVT is a generic value tree providing HTA-related value concerns for new medicines in a hierarchical structure of evaluation criteria.²⁷ Based on a structured process involving systematic review and

expert consultation⁴⁶ and adopting a top-down approach,⁴⁷ the AVT consists of 5 value domains (i.e. criteria clusters), capturing all essential value attributes of new medicines in the HTA context under a prescriptive decision-aid approach: 1) burden of disease (BoD), 2) therapeutic benefit (THE), 3) safety profile (SAF), 4) innovation level (INN), and 5) socioeconomic impact (SOC), such that overall value is a function expressed as follows:

$$Value = f(\mathbf{BoD}, \mathbf{THE}, \mathbf{SAF}, \mathbf{INN}, \mathbf{SOC}) \quad (1)$$

The latest available evidence from clinical studies was used to populate the performance of the alternative options across the respective criteria attributes of the AVT. The AVT was thus adapted for the context of NASH as part of a bottom-up, alternative-focused thinking approach, following the comparison of the alternative compounds, in consultation with a hepatology specialist (M.T., study co-author).^{25,48} This adaptation resulted in the preliminary version of the NASH value tree, which served as the basis of assessing the value of NASH compounds in the first round of DCs, comprising a total of 17 criteria attributes. The burden of disease cluster was removed because all alternative treatments were assessed for the same indication, whereas socioeconomic impact criteria were excluded because of lack of evidence.

Following the completion of the first round of DCs, the final version of the value tree was used as the starting point for the second round of DCs, that is, the preliminary version in the second part of the exercise. In arriving at the NASH-specific attributes and the respective value tree, we strived to adhere to key decision theory properties such as preferential independence and nonredundancy, thus ensuring attribute selection was methodologically correct and theoretically robust.⁴⁹

2. Alternative treatments compared and evidence considered

A total of 4 compound profiles were assessed, 3 with publicly available phase 2 results plus a hypothetical product profile (HPP); the latter was viewed as an “aspirational summary” of a hypothetical product in terms of labeling concepts, leveraging hypothetical information about the compound available at a particular time in development,⁵⁰ effectively providing a prospective summary of the characteristics of a product that could theoretically be achieved.⁵¹ Information on the HPP was communicated by the study sponsor. Because of the commercial in confidence information surrounding the HPP, all compounds are anonymized in the study.

Given the early assessment scope of the exercise with no available drug treatment on the market for NASH,

we used the most relevant clinical evidence for the assessment of the compounds, including a number of assumptions for the performance of the HPP. Expert opinion was used to specify the performance of options across specific criteria in case such information was not available.

The final performance of options used across the different criteria attributes together with the respective range of placebo arms and lower – higher (i.e., least preferred – most preferred) reference levels used in the models is shown in Supplementary Table A1; additional information on evidence considered is discussed in the Supplementary Appendix.

3. Setting attribute ranges and reference levels

For the purposes of scoring and weighting, “higher” (x_h , set at 100) and “lower” (x_l , set at 0) reference levels were defined for each attribute acting as benchmarks of an interval value scale based on which the compounds were scored, that is, $v(x_{\text{higher}}) = 100$ and $v(x_{\text{lower}}) = 0$. The “higher” reference corresponded to the best available performance on that attribute and the “lower” corresponded to the worst available performance across the compounds compared. These reference levels were needed for the construction of criteria partial value functions on interval scales and the elicitation of relative weights. Given the application of an additive aggregation model, a hypothetical overall weighted preference value (WPV) score of 100 for a compound would entail the best possible performance across all criteria, whereas an overall WPV score of 0 would entail the worst possible performance across all criteria (see Appendix Table A1, last 2 columns for worst and best reference levels of all criteria).

Phase c. Model Assessment; and Phase d- Model Appraisal: DCs and MCDA Technique

In each of the 2 DC rounds, the completion of model-building, model assessment and part of the model appraisal phases took place through 3 DC meetings with key stakeholders, mimicking what would occur in the respective settings¹⁰; these took the form of facilitated workshops lasting 1 to 1.5-days each and were conducted in England (London), France (Paris) and Germany (Berlin).

DC participants consisted of small groups of experts, ranging between 5 and 13 participants. We endeavoured to involve all relevant stakeholder groups and perspectives, reflecting actual assessment processes in the study countries; specifically, health care professionals,

Table 1 Decision Conference Participant Numbers, Stakeholder Groups, and Durations

	HCP	METH	PAT	Total	DC Duration, days
Round 1					
England	8	4	1	13	1.5
France	4	2	0	6	1.0
Germany	4	2	1	7	1.0
Round 2					
England	5	4	1	10	1.5
France	6	4		10	1.5
Germany	2	2	1	5	1.5

Note: HCP, health care professional(s); METH, methodologists; PAT, patient(s) and/or patient advocates.

methodologists, and patient representatives were included. These sizes have been shown to be optimal, allowing efficient group processes to emerge while preserving individuality, as they are large enough to represent all major perspectives but small enough to be able to work toward agreement.⁸ A summary of participant numbers and stakeholder groups in each meeting together with each DC duration is shown in Table 1.

An impartial facilitator (J.B.C., study co-author) guided the process interaction while refraining from contributing to content, essentially helping the group in how to think about the issues but not what to think,^{40,52} thus pointing to an interactive model-building process in which debate was encouraged and differences in opinion were actively sought in an iterative manner. Where consensus could not be reached, value judgments were selected based on majority voting, representing a single preference input for the whole group of participants and the relevant parameter was then tested in sensitivity analysis (the Supplementary Appendix contains further information about the decision conferencing process).

Compound overall value was obtained through the application of the additive aggregation model. The AVF was operationalized by adopting a typical simple additive aggregation approach, where the overall value $V(\cdot)$ of an option a would be given by Eq. (2)²⁷:

$$V(a) = \sum_{i=1}^m w_i v_i(a) \quad (2)$$

where $v_i(a)$ is the partial value score of option a obtained by the application of the value function of criterion i to the performance of a in that criterion, w_i is the weight of criterion i , and m is the total number of criteria (attributes). This function $V(\cdot)$ is effectively a multiattribute value function.¹¹

A MACBETH (Measuring Attractiveness by a Categorical-Based Evaluation Technique) protocol was adopted as an approach to elicit value preferences, effectively using qualitative judgments about the difference in value between different pairs of performance levels.^{53,54} MACBETH is based on strong theoretical foundations,⁵⁵ and its usefulness as a decision support tool has been shown through numerous applications for a variety of real-world problems^{56,57} as part of which semantic judgments are converted into a cardinal scale. We used M-MACBETH,⁵⁸ a decision support system based on the MACBETH approach, to elicit value preferences of DC participants and complete the MCDA model. Following the DCs, additional deterministic sensitivity analysis was conducted to address parameter uncertainty on criteria weights, by systematically exploring changes on baseline weights and their impact on the overall value rankings of the compounds (the Supplementary Appendix contains further information on MACBETH, M-MACBETH, and the robustness analysis conducted).

Results

Final Value Trees

The attributes of the final NASH value tree in each country, as emerged following discussions with DC participants, are listed in Table 2. Schematic illustrations of the final value trees emerging from each DC are shown in Supplementary Appendix Figure A1.

Across the 3 study countries over the 2 rounds of DCs, 14 to 19 attributes were included in the value tree, as shown in Table 3. In terms of the different criteria clusters' composition, in round 1 the THE cluster comprised 9 to 10 attributes, followed by SAF (5 attributes) and INN (2–4 attributes) clusters. In round 2, THE comprised 7 to 10 attributes, followed by SAF (5–9 attributes) and INN (1 attribute) clusters. In round 2, the value tree of each country consisted of the same or lower number of attributes compared with round 1. In both rounds of DCs, the largest number of criteria were allocated under the THE cluster, followed by SAF and INN, with the exception of the second round in Germany, where more criteria were included in the SAF cluster, followed by THE and INN.

A common feature of changes in the composition of value trees across all 3 countries over the two rounds was the non-inclusion of spillover effect criteria under the INN cluster because they were considered to be non-relevant for the scope of the exercise. Another common feature as evident from the English and French DCs was the non-inclusion of the SF-36 summary scores (physical

Table 2 Criteria Definitions and Their Consideration in Each Jurisdiction per Round of Decision Conference (Their Presence Denoted by “X”)

Criteria Subcluster	Criteria Names	Attribute Definitions	Country							
			Round 1				Round 2			
			England	France	Germany	Germany	England	France	France	Germany
Criteria cluster 1: therapeutic benefit										
Histologic endpoints	NASH resolution	% of patients experiencing resolution of NASH without worsening of fibrosis	X	X	X	X	X	X	X	X
Histologic endpoints	Fibrosis improvement	% of patients experiencing fibrosis improvement	X	X	X	X	X	X	X	X
Lipids	LDL cholesterol	Mean % or absolute change from baseline in LDL cholesterol, mmol/L	X	X	X	X	X	X	X	X
Lipids	HDL cholesterol	Mean % or absolute change from baseline in HDL cholesterol, mmol/L	X	X	X	X	X	X	X	X
Metabolic factors	HbA1c	Mean % or absolute change from baseline in glycated hemoglobin, A1c, mmol/mol	X	X	X	X	X	X	X	X
Metabolic factors	Body weight	Mean % or absolute change from baseline in body weight	X	X	X	X	X	X	X	X
Metabolic factors	Systolic pressure	Mean % or absolute change from baseline in systolic pressure, mm Hg	X	X	X	X	X	X	X	X
Metabolic factors	Diastolic pressure	Mean % or absolute change from baseline in diastolic pressure, mm Hg	X	X	X	X	X	X	X	X
Metabolic factors	ALT	Mean % or absolute change from baseline in alanine aminotransferase, U/L	X	X	X	X	X	X	X	X
Metabolic factors	Triglycerides	Mean absolute change from baseline in triglycerides, grams/L	X	X	X	X	X	X	X	X
Metabolic factors	Gamma GT	Mean absolute change from baseline in Gamma-glutamyl transferase, U/L	X	X	X	X	X	X	X	X
Metabolic factors	HOMA-IR	Mean absolute change from baseline in HOMA-IR, units	X	X	X	X	X	X	X	X
Quality of life	SF-36 Physical comp	Mean % or absolute change from baseline, score	X	X	X	X	X	X	X	X
Quality of life	SF-36 Mental comp	Mean % or absolute change from baseline, score	X	X	X	X	X	X	X	X
Criteria cluster 2: safety profile										
Adverse events	Treatment related Serious AEs	% of patients experiencing treatment-related serious adverse events	X	X	X	X	X	X	X	X
Adverse events	Overall serious AEs	% of patients experiencing overall serious adverse events	X	X	X	X	X	X	X	X
Adverse events	Nausea	% of patients experiencing nausea	X	X	X	X	X	X	X	X
Adverse events	Pruritus, G1-G2	% of patients experiencing pruritus, Grade 1 and 2	X	X	X	X	X	X	X	X
Adverse events	Pruritus, G2-G3	% of patients experiencing pruritus, Grade 2 and 3	X	X	X	X	X	X	X	X
Adverse events	Pruritus, G3	% of patients experiencing pruritus, Grade 3	X	X	X	X	X	X	X	X
Adverse events	Pruritus, any	% of patients experiencing pruritus, any grade	X	X	X	X	X	X	X	X
Adverse events	Nausea, vomiting, diarrhea	% of patients experiencing nausea, vomit, & diarrhea	X	X	X	X	X	X	X	X
Adverse events	Renal	% of patients experiencing renal adverse events	X	X	X	X	X	X	X	X

(continued)

Table 2 (continued)

Criteria Subcluster	Criteria Names	Attribute Definitions	Country					
			Round 1			Round 2		
			England	France	Germany	England	France	Germany
Adverse events	Fatigue	% of patients experiencing fatigue						X
Adverse events	Cardiovascular	% of patients experiencing cardiovascular adverse events						X
Contraindications & warnings	Contraindications	Existence of any contraindications of use	X	X	X	X	X	X
Contraindications and warnings	Drug-drug interactions	Existence of any drug–drug interactions	X	X	X	X	X	X
Criteria cluster 3: innovation level								
Mechanism of action	ATC L4	The technology's mechanism of action innovation type as reflected through WHO's Anatomical Therapeutic Chemical (ATC) Classification System, level 4						
Spillover effect	New indications in Ph1	Number of other indications investigated as part of Phase 1 clinical trials						
Spillover effect	New indications in Ph2	Number of other indications investigated as part of Phase 2 clinical trials	X					
Spillover effect	New indications in Ph3	Number of other indications investigated as part of Phase 3 clinical trials	X		X			
Spillover effect	New indications in MA	Number of other indications having received a marketing authorization	X	X ^a	X			
Patient convenience	Delivery system and posology	Combination of the delivery system and the posology (i.e., dosing)	X	X	X	X	X	X

HbA1c, hemoglobin A1c; HDL, high-density lipoprotein; HOMA-IR, Homeostatic Model Assessment of Insulin Resistance; LDL, low-density lipoprotein; NASH, nonalcoholic steatohepatitis; U/L, upper limit.

^aFor related disease.

Table 3 Number of Criteria Attributes and Their Relative Weights per Criteria Cluster across the 3 Countries

Criteria Clusters/ Countries	England		France		Germany	
	Criteria (n)	Criteria Weights (%)	Criteria (n)	Criteria Weights (%)	Criteria (n)	Criteria Weights (%)
Round 1 (2017)						
Therapeutic benefit	10	64.7	9	66.31	10	58.48
Safety profile	5	19.79	5	27.71	5	31.57
Innovation level	4	15.51	2	5.98	3	9.95
Total	19	100	16	100	18	100.0
Round 2 (2018)						
Therapeutic benefit	8	59.84	10	61.01	7	35.96
Safety profile	5	28.37	5	37.6	9	62.28
Innovation level	1	11.79	1	1.39	1	1.74
Total	14	100	16	100	17	100.0

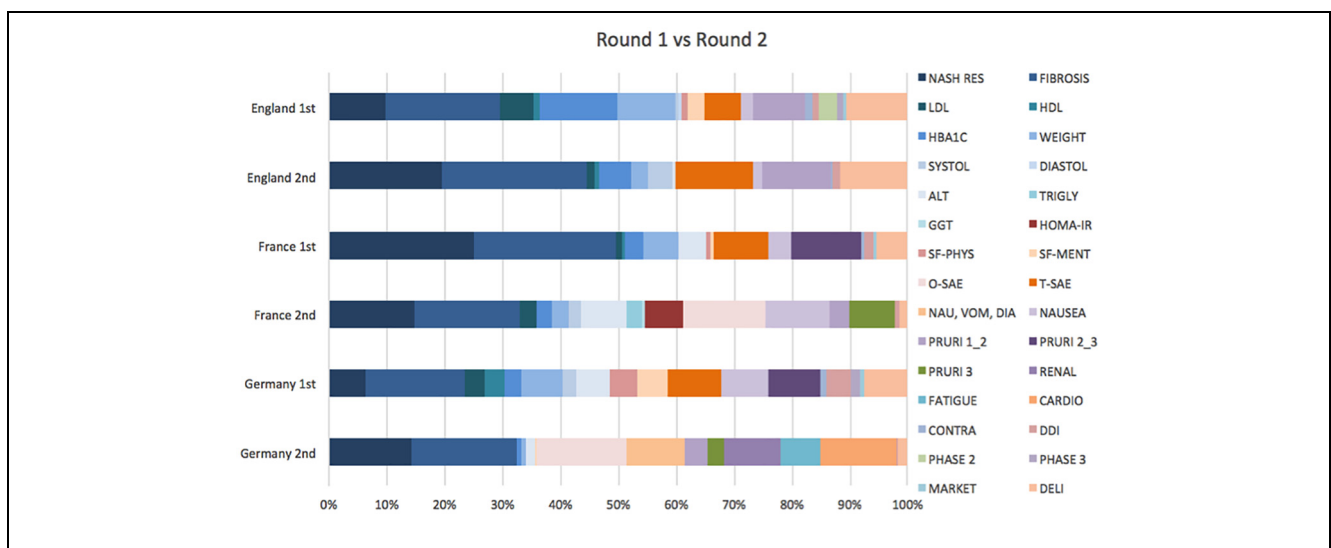


Figure 1 Relative weights of criteria across the 2 rounds of decision conferences. NASH RES, NASH Resolution; FIBROSIS, fibrosis improvement; LDL, low-density lipoprotein cholesterol; HDL, high-density lipoprotein cholesterol; HbA1c, hemoglobin A1c; WEIGHT, body weight; SYSTOL, systolic blood pressure; DIASTOL, diastolic blood pressure; ALT, alanine aminotransferase; TRIGLY, triglycerides; GGT, gamma-glutamyl transferase; HOMA-IR, Homeostatic Model Assessment of Insulin Resistance; SF-PHYS, SF-36 physical component score; SF-MENT, SF-36 mental component score; T-SAE, treatment-related serious adverse events; O-SAE, overall serious adverse events; NAUSEA, nausea; NAU, VOM, DIA, nausea, vomiting, and diarrhea; PRURI 1_2, pruritus grades 1 and 2; PRURI 2_3, pruritus grades 2 and 3; PRURI 3, pruritus grade 3; RENAL, renal events; FATIGUE, fatigue; CARDIO, cardiovascular events; CONTRA, contraindications; DDI, drug-drug interactions; PHASE 2, phase 2 indications; PHASE 3, phase 3 indications; MARKET, market authorized indications; DELI, delivery system and posology.

and mental component scores) as they were judged to be non-clinically meaningful.

Criteria Weights

The relative weights of individual criteria across both rounds are illustrated in Figure 1. In terms of the total

weights assigned across the different criteria clusters, the THE cluster always obtained the highest rank, followed by SAF and INN, with the exception of round 2 of the German DC, in which the SAF cluster outranked THE. The fact that the THE cluster always contributed to more than 50% of the model’s total weight means that an improvement from the worst to the best reference

Table 4 Overall Weighted Preference Value (WPV) Scores of the Compounds across the 3 Countries over the 2 Rounds of Decision Conferences^a

	England		France		Germany	
	2017	2018	2017	2018	2017	2018
Compound A	42.0 ³	40.9 ³	37.7 ³	37.0 ³	46.3 ³	42.5 ²
Compound B	31.4 ⁴	27.8 ⁴	24.7 ⁴	31.1 ⁴	38.2 ⁴	30.3 ⁴
Compound C	61.4 ²	63.3 ²	62.2 ²	60.8 ²	57.9 ²	33.8 ³
Compound D	67.4 ¹	78.5 ¹	74.2 ¹	70.0 ¹	70.0 ¹	90.6 ¹

^aSuperscript numerals indicate the compound ranking: 1 = first ranked; 2 = second ranked; 3 = third ranked; 4 = fourth ranked.

level in this cluster's criteria would always be considered more valuable than an improvement in the other criteria clusters combined. Across countries, the relative weight of the SAF cluster increased in round 2 compared with round 1 to the detriment of the THE and INN clusters, whose relative weights declined. In terms of individual criteria, the number of highest-ranked criteria assigned a relative weight of 10% or more was always observed to be higher in the second compared with the first round of DCs across all 3 countries. The relative weights of criteria and their differences across countries reflect the actual value preferences of the participants but are also influenced by the number of criteria being considered in each value tree.

In England, the THE cluster accounted for the highest proportion of the model's relative weight, followed by SAF and INN; this result was consistent in both rounds. Four criteria were assigned a relative weight of 10% or more in round 1 (fibrosis improvement [19.8%], HbA1c [13.4%], delivery system and posology [10.7%], body weight [10.2%]), in contrast to 5 criteria in round 2 (fibrosis improvement [25.3%], NASH resolution [19.4%], treatment-related serious adverse events [13.5%], pruritus G1,2 [12.1%], delivery system and posology [11.8%]).

In France, across both rounds, the THE cluster accounted for the highest proportion of the model's relative weight, followed by SAF and INN. Three criteria were assigned a relative weight of 10% or more in the first round (NASH resolution [25.0%], fibrosis improvement [24.5%], pruritus G2,3 [12%]) compared with 4 criteria in the second round (fibrosis improvement [18.1%], NASH resolution [14.9%], overall serious adverse events [14.4%], nausea, vomiting, and diarrhea [11.1%]).

In Germany, there were some differences in the overall weight of clusters between rounds, such that the THE cluster was outranked by the SAF cluster in the second round in terms of their total weights, followed by INN in third place. Although only a single criterion was associated with a relative weight of 10% or more in round 1

(fibrosis improvement [17.0%]), a total of 5 criteria were assigned a relative weight of 10% or more in round 2 (fibrosis improvement [18.3%], overall serious adverse events [15.3%], NASH resolution [14.2%], cardiovascular adverse events [13.1%], nausea, vomiting, and diarrhea [10.2%]).

Of the 6 rounds, fibrosis improvement ranked first 5 times and second once, with a relative weight ranging from 17.0% to 25.3%. NASH resolution appeared in the top-3 for a total of 4 times, and serious adverse events (treatment related or overall, depending on the country) ranked second or third 4 times.

Overall Compound Rankings and Value Composition

In terms of the compounds' overall rankings, across the 3 countries and the 2 rounds of DCs, compound D was always ranked first (overall weighted preference value (WPV) score range: 67.4–90.6). With the exception of the second round of the German DC, in which compound A ranked second and compound C third (overall scores of 42.5 and 33.8, respectively), in all remaining rounds across countries compound C ranked second and compound A ranked third (overall score ranges: 33.8–63.3 and 37.0–46.3, respectively). Compound B always ranked last (overall score range: 24.7–38.2). The largest difference between the first and second ranked compound was 48.1 points (D over A in the second DC in Germany), whereas the smallest difference was 6 points (D over C in the first DC in England).

The overall WPV scores for all compounds across the study countries and the 2 DC rounds are shown in Table 4. Stacked bar plots of the compounds' overall WPV scores across the 3 countries over the 2 rounds of DCs are shown in Figures 2 to 4, with absolute value contributions of each criterion. It should be clear that the overall WPV score of each option depends on the criteria included in the value model, the shape of the value

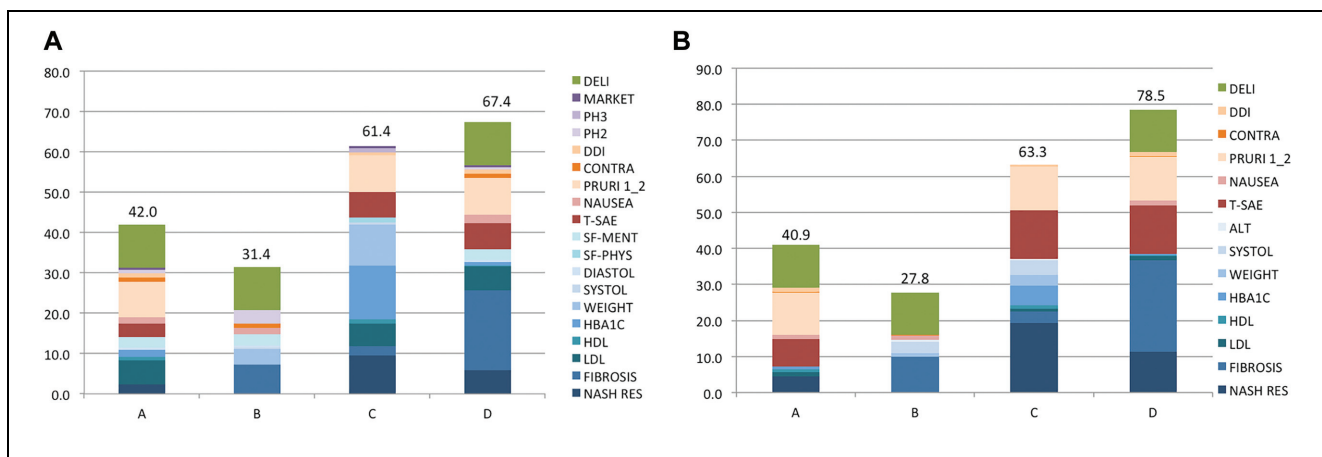


Figure 2 Stacked bar plots of compounds' overall weighted preference value scores over the 2 rounds of decision conferences in England. (a) Round 1. (b) round 2.

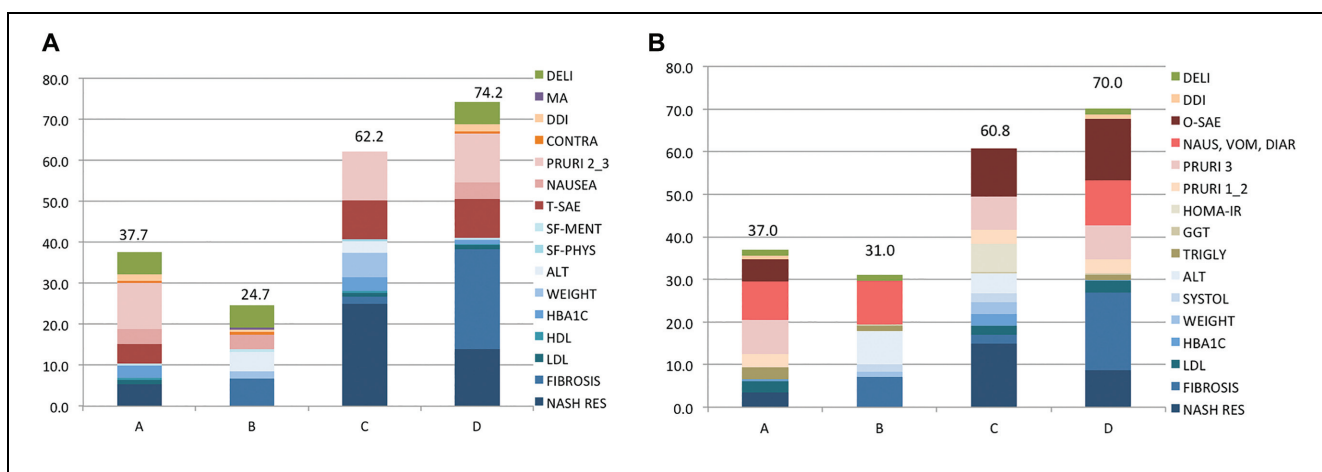


Figure 3 Stacked bar plots of compounds' overall weighted preference value scores over the 2 rounds of decision conferences in France. (a) Round 1. (b) Round 2.

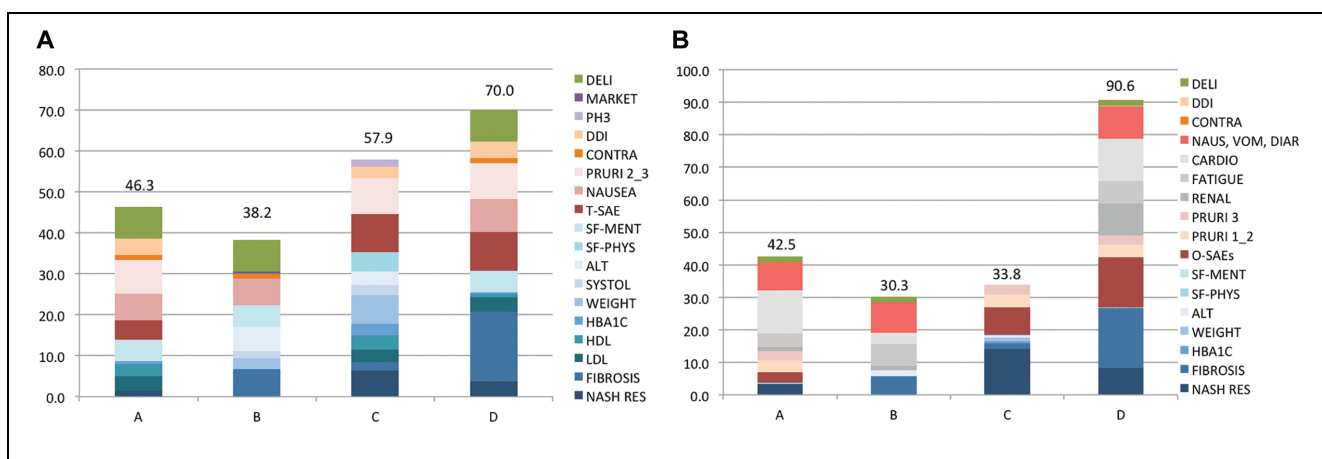


Figure 4 Stacked bar plots of compounds' overall weighted preference value scores over the 2 rounds of decision conferences in Germany. (a) Round 1. (b) Round 2.

functions that influence the value scores, and the criteria weights that are elicited using the 2 reference levels.

In the English and French DCs, the largest proportion of the leading compound's value contribution (compound D) was always due to its performance on the THE cluster, followed by the SAF and INN clusters. By contrast, in the German DC, the largest proportion of compound's D value was accounted for by the SAF cluster in both rounds, followed by the THE and INN clusters. In terms of individual criteria, across all 3 countries and over both rounds, fibrosis improvement was always the most influential value dimension on compound's D overall value composition, contributing the greatest value score.

The sensitivity and robustness analyses demonstrated that treatment rankings are robust to the relative criteria weights across the different settings. The only case in which a change of less than 100% of a criterion's weight could affect the compound ranking was in the first DC round in England and France, in which a 29% reduction in fibrosis improvement (from 19.8% to 14.0%) and a 77% increase in NASH resolution (from 25% to 44.2%) would be adequate for compound C to become better ranked than compound D.

Discussion

This two-round, cross-country exercise demonstrated the application of a recently developed MCDA methodological process and value framework^{25,27} for a set of alternative compounds under clinical development versus a HPP in the context of early HTA. In terms of design, implementation and review of the analysis, the process adopted was in alignment with good practice guidelines on the use of MCDA in health care decisions.^{23,24}

Compound Ranking and Value Preferences

Based on the MCDA model used, the set of evaluation criteria considered, and the respective performance of the compounds used, we found that the ranking of the compounds was virtually identical across the study countries and both DC rounds. Compound D always ranked first based on DC participants' overall WPV scores, followed by compounds C, A, and B, with the exception of the second DC in Germany, where compound A ranked higher than compound C.

In terms of the various value reflections, with the exception of the second German DC, the THE cluster had the highest relative importance for participants across all 3 countries (59%–66% of model weight), always followed by the SAF (20%–38%) and the INN

clusters (1%–16%). With the exception of the second German DC, the SAF cluster outranked the THE cluster with 62% versus 36%.

Fibrosis was the most important criterion across countries and DC rounds, having the highest relative weight in most cases, with the exception of the French DC (round 1 only), where it was assigned the second highest weight but still very close to the top-ranked criterion. In France, the 2 histologic endpoints played the most important role to stakeholders in both rounds; this partly explains why the THE cluster produced the highest relative weight across all 3 countries.

Some differences were recorded between rounds and the criteria accounting for at least 50% of the model weight in each case. In the English DCs, for at least 50% of the model's weight to be accounted for, weights of the 3 leading criteria had to be included, adding up to 58% in the second round (the 2 histologic endpoints and treatment-related serious adverse events). By contrast, the 4 highest ranking criteria in the first round (fibrosis improvement, HbA1c, delivery system and posology, and body weight) accounted for 54% of the model's weight. In the French DCs, for at least half the model's weight to be captured, the weights of the four highest ranking criteria in the second round had to be included, adding up to 59% (the 2 histologic endpoints, overall serious adverse events and nausea, vomiting, and diarrhea); by contrast, the top 3 criteria in the first round accounted for 61% of total model weight (the 2 histologic endpoints and pruritus G2,3). In the German DCs, for at least half the importance of the model's weight to be captured, in the second round the weights of the 4 highest ranking criteria had to be included, adding up to 61% (the 2 histologic endpoints, overall serious adverse events and cardiovascular adverse events). In comparison, the 5 highest ranking criteria in the first round added up to 51% (fibrosis improvement, treatment-related serious adverse events, pruritus G2,3, nausea and delivery system and posology).

Overall, nearly identical results were observed across countries and rounds in terms of compound rankings. Understandably, the direct comparison of overall value scores between options requires identical value models, comprising the same criteria sets, weights, and value functions. The ranking comparisons made in this study using ordinal scales reflect these restrictions. Nevertheless, when trying to interpret differences in results between countries, it could be argued that any disparities might be due to either "real" country differences relating to the consideration of different fundamental objectives, priorities, and preferences (as reflected through differences in

criteria inclusion, relative weights assignment, and scores elicitation), or differences relating to individual participants. This was indirectly addressed through the extensive sensitivity analysis conducted, aiming to reduce parameter and, by extension, intracountry uncertainty, but also through the further insights generated in the second round of DCs, which practically validated the results of the first round.

Limitations

One of the study limitations is the lack of relative effect estimates as part of clinical evidence and the use of their absolute effects from different clinical trials with the assumption that they can be compared directly. Given the absence of head-to-head clinical trials directly comparing the compounds of interest, the small number of clinical trials available and their early phase, absolute effects from the respective single randomized clinical trials of the alternative compounds were used. If more clinical studies become available, an indirect treatment comparison could be conducted first using a common comparator to estimate the relative effects of 2 treatments versus the comparator, or a network meta-analysis combining both direct and indirect evidence available through a mixed treatment comparison. Although the placebo range was disclosed to participants during the workshops, the incremental performance difference from placebo would be needed to better understand the associated value. In real-world evaluations aiming to inform decision making, evidence synthesis would be required to take place together with evidence collection as part of the model-building phase.

A second limitation relates to the clinical trials used as the source of evidence. Only data from a single, phase 2, randomized clinical trial per compound was used with relatively small sample sizes (number of subjects randomly assigned in either arm ranging from 26–142 patients).

Third, the studies used had different populations (disease stage), different trial durations, and one did not meet its primary endpoint, all of which have implications in terms of their comparability and perceived efficacy. Differences in the definitions of histologic endpoints across studies were assumed to be comparable, considering that a consensus across NASH primary outcome definitions is lacking. Given the uncertainty around treatment effects, participants were instructed to assume that they corresponded to a specific sample size of patients, chosen based on the largest number of patients in a single active arm evident across the studies ($N =$

142), which helped participants to provide a value judgment across the clinical attributes.

Finally, in relation to the participatory approach adopted, not all DCs were of the same size or had the same composition of experts, particularly the second DC in Germany (5 participants), which may have contributed to a different overall compound ranking order.

Lessons Learned

Despite the limitations, this study demonstrated consistency of MCDA results following 2 rounds of preference elicitation via DCs with different participants in 3 settings, suggesting that value preferences and their differences between countries are fairly reliable.

One contribution of this study relates to the value attribution of NASH compounds in different countries as reflected through the key drivers of their overall scores. The 2 histologic endpoints of fibrosis improvement and NASH resolution had a clear influence in this context. Regarding differences in the acceptability of the evaluation criteria considered, the German experience (and, to a lesser degree, the French experience) suggests that it is paramount to submit evidence on clinically meaningful outcomes relating to mortality, morbidity, and quality of life. Alternatively, if surrogate endpoints are used, they should be accompanied with a clear impact on relevant clinically meaningful outcomes or demonstrate minimum clinically important differences. This could be informed by randomized prospective interventional studies or by focusing on the magnitude of the studies rather than follow-up duration.

A number of insights relating to the usefulness of MCDA in early HTA have been generated. In terms of the phase of clinical development, it became obvious that there will be a trade-off between data availability and prospects of influencing drug development. Regarding the preference for data on clinical versus surrogate endpoints, this seems to be universally true, but surrogates can probably be accepted if they are validated and can predict the clinical endpoint. It is recommended to provide correlations of surrogates with outcomes and/or clinically meaningful thresholds.

A critical issue observed across the different country settings was the current clinical debate given the lack of a clear disease definition, including the challenges around disease diagnosis and identification of patients in greater need. The debate around clinically meaningful outcomes was indicative in that context, with most clinical studies using histologic endpoints as surrogate markers. This suggests that a better understanding of the disease together with the development of improved clinical

guidelines could potentially benefit the elevation of NASH in the national policy agendas of decision and policy makers.

In terms of methodological contribution, the analytical objectives in early HTA contexts should always be clearly defined (e.g., evaluation of current product profiles vs. design of future phase 3 trials) to facilitate preference elicitation as part of DCs or other participatory processes. With regard to improvements around evidence synthesis and summary, assuming the use of nonsynthesized clinical evidence and the existence of a common comparator (such as placebo arms), the use of performance levels compared with the comparator (e.g., incremental placebo differences) should be explored. Using performance levels, it could be easier for participants to comprehend the significance of any relevant treatment effects. However, it should be considered whether this might have an influence on the definition and interpretation of the reference levels, as an “incremental” treatment effect might also have to be present across other non-clinical criteria for a more homogeneous and easier interpretation of the results.

Importantly, for the use of non disease-specific clinically meaningful outcomes, it might be useful to derive clinically meaningful thresholds or correlations that could inform the elicitation of value judgments and preferences, for example, minimally clinically important difference for PRO instruments.

Finally, DCs should be conducted over a 2-day period to allow sufficient time for discussion and evidence synthesis tasks, the greatest trade-off being the challenge in recruiting adequate numbers of participants. Venues should be in alignment with DC good practice guidelines with U-shaped seating for direct eye contact and 2 projector screens (one with the value tree, the other with the options performance data). The pool of potential experts to contact to secure adequate DC participants should not be underestimated, particularly in follow-on rounds, as more than 125 experts were contacted to secure 25 in the second round.

Conclusion


In this study, we explored the application of the Advance Value Framework in the context of early HTA for NASH compounds in 3 EU countries and tested the consistency of results by conducting 2 rounds of preference elicitation via decision conferencing. The use of MCDA proved to be promising for early HTA while illustrating high consistency in results across countries and between study rounds. The complexity of NASH management,

given the existence of patients with multiple comorbidities and clinical endpoints, enables MCDA to act as a transparent and potentially consistent approach for evaluating compounds in clinical development before market entry. The results can be used to ensure that relevant and important endpoints are included in upcoming clinical studies and to identify potential discrepancies across HTA bodies in terms of value assessment requirements. Ongoing and future research could validate and possibly supplement insights provided regarding differences in value preferences in different settings, whether for NASH or HTA appraisals more generally.

Acknowledgments

We would like to thank all the participants of the decision conferences for their valuable input and feedback. Katie McClain provided excellent administrative and research support throughout the project. Huseyin Naci provided valuable advice on evidence synthesis. MT acknowledges the support of the NIHR Imperial Biomedical Research Centre. Finally, we are grateful to 3 anonymous referees for providing valuable comments and suggestions in an earlier draft of the paper. All outstanding errors remain our own.

ORCID iD

Aris Angelis  <https://orcid.org/0000-0002-0261-4634>

Supplemental Material

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

References

1. Tervonen T. Quantifying preferences in drug benefit-risk decisions. *Clin Pharmacol Thera*. 2019;106(5):955–9.
2. European Medicines Agency. EMA—EUnetHTA three-year work plan. Available from: https://www.ema.europa.eu/en/documents/other/ema-eunetha-three-year-work-plan-2017-2020_en.pdf 2017
3. European Network for Health Technology Assessment. EMA. Available from: <https://www.eunetha.eu/ema/> 2019
4. European Medicines Agency. Parallel consultation with regulators and health technology assessment bodies. Available from: <https://www.ema.europa.eu/en/human-regulatory/research-development/scientific-advice-protocol-assistance/parallel-consultation-regulators-health-technology-assessment-bodies> 2019
5. European Network for Health Technology Assessment. Early dialogues. Available from: <https://www.eunetha.eu/services/early-dialogues/> 2019
6. Ijzerman MJ, Steuten LMG. Early assessment of medical technologies to inform product development and market

- access: a review of methods and applications. *Appl Health Econ Health Policy*. 2011;9(5):331–47.
7. Ijzerman M, Koffijberg H, Fenwick E, Krahn M. Emerging Use of early health technology assessment in medical product development: a scoping review of the literature. *Pharmacoeconomics*. 2017;35(7):727–40.
 8. Angelis A, Phillips LD. Advancing structured decision making in drug regulation at the FDA and EMA. *Br J Clin Pharmacol*. 2020;1–11. <https://doi.org/10.1111/bcp.14425>.
 9. Phillips L, Phillips M. Facilitated work groups: theory and practice. *J Oper Res Soc*. 1993;44(6):533–49.
 10. Phillips L. Decision conferencing. In: Edwards W, Miles R, von Winterfeldt D, eds. *Advances in Decision Analysis: From Foundations to Applications*. Cambridge (UK): Cambridge University Press; 2007.
 11. Keeney R, Raiffa H. *Decisions with Multiple Objectives: Preferences and Value Trade-offs*. Cambridge (UK): Cambridge University Press; 1993.
 12. Angelis A, Kanavos P, Phillips L. ICER Value Framework 2020 Update: Recommendations on the Aggregation of Benefits and Contextual Considerations. *Value in Health*. 2020 in press
 13. Eichler HG, Abadie E, Raine JM, Salmonson T. Safe drugs and the cost of good intentions. *N Engl J Med*. 2009;360(14):1378–80.
 14. Phillips L, Fasolo B, Zafiroopoulos N, Beyer A. Is quantitative benefit-risk modelling of drugs desirable or possible? *Drug Discov Today Technol*. 2011;8(1):e1–e42.
 15. Hughes D, Waddingham E, Mt-Isa S, et al. Recommendations for benefit–risk assessment methodologies and visual representations. *Pharmacoepidemiol Drug Saf*. 2016;25(3):251–62.
 16. Felli JC, Noel RA, Cavazzoni PA. A multiattribute model for evaluating the benefit-risk profiles of treatment alternatives. *Med Decis Making*. 2009;29(1):104–15.
 17. European Medicines Agency. Guidance document on the content of the <Co-> Rapporteur day <60*> <80> critical assessment report. Available from: https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/day-80-assessment-report-quality-guidance-rev1217_en.pdf 2018
 18. Hammond J, Keeney R, Raiffa H. *Smart Choices: A Practical Guide to Making Better Decisions*. Cambridge (MA): Harvard University Press; 1999.
 19. Pignatti F, Ashby D, Brass EP, et al. Structured frameworks to increase the transparency of the assessment of benefits and risks of medicines: current status and possible future directions. *Clin Pharmacol Ther*. 2015;98(5):522–33.
 20. Food and Drug Administration. *Benefit-Risk Assessment in Drug Regulatory Decision-Making*. Draft PDUFA VI implementation plan (FY 2018 -2022). Rockville (MD): Food and Drug Administration; 2018.
 21. Oliveira M, Mataloto I, Kanavos P. Multi-criteria decision analysis for health technology assessment: addressing methodological challenges to improve the state of the art. *Eur J Health Econ*. 2019;20(6):891–918.
 22. Marsh K, Lanitis T, Neasham D, Orfanos P, Caro J. Assessing the value of healthcare interventions using multi-criteria decision analysis: a review of the literature. *Pharmacoeconomics*. 2014;32(4):345–65.
 23. Thokala P, Devlin N, Marsh K, et al. Multiple criteria decision analysis for health care decision making—an introduction: report 1 of the ISPOR MCDA Emerging Good Practices Task Force. *Value Health*. 2016;19(1):1–13.
 24. Marsh K, Ijzerman M, Thokala P, et al. Multiple criteria decision analysis for health care decision making-emerging good practices: report 2 of the ISPOR MCDA Emerging Good Practices Task Force. *Value Health*. 2016;19(2):125–37.
 25. Angelis A, Kanavos P. Value-based assessment of new medical technologies: towards a robust methodological framework for the application of multiple criteria decision analysis in the context of health technology assessment. *Pharmacoeconomics*. 2016;34(5):435–46.
 26. Goetghebeur MM, Wagner M, Khoury H, Levitt RJ, Erickson LJ, Rindress D. Evidence and value: impact on decision making—the EVIDEM framework and potential applications. *BMC Health Serv Res*. 2008;8:270.
 27. Angelis A, Kanavos P. Multiple criteria decision analysis (MCDA) for evaluating new medicines in health technology assessment and beyond: the advance value framework. *Soc Sci Med*. 2017;188:137–56.
 28. Working Group on Mechanism of Coordinated Access to Orphan Medicinal Products. *Process on Corporate Social Responsibility in the Field of Pharmaceuticals Platform on Access to Medicines in Europe*. Final report. Brussels: European Commission; 2013.
 29. Lakdawalla DN, Doshi JA, Garrison LP, Phelps CE, Basu A, Danzon PM. Defining elements of value in health care—a health economics approach: an ISPOR Special Task Force report [3]. *Value Health*. 2018;21(2):131–9.
 30. Goetghebeur MM, Wagner M, Khoury H, Rindress D, Grégoire J, Deal C. Combining multicriteria decision analysis, ethics and health technology assessment: applying the EVIDEM decision making framework to growth hormone for Turner syndrome patients. *Cost Eff Resour Alloc*. 2010;8:4.
 31. Sussex J, Rollet P, Garau M, Schmitt C, Kent A, Hutchings A. A pilot study of multicriteria decision analysis for valuing orphan medicines. *Value Health*. 2013;16(8):1163–9.
 32. Angelis A, Montibeller G, Hochhauser D, Kanavos P. Multiple criteria decision analysis in the context of health technology assessment: a simulation exercise on metastatic colorectal cancer with multiple stakeholders in the English setting. *BMC Med Inform Decis Making*. 2017;17(1):149.
 33. Wagner M, Khoury H, Bennetts L, et al. Appraising the holistic value of Lenvatinib for radio-iodine refractory differentiated thyroid cancer: a multi-country study applying pragmatic MCDA. *BMC Cancer*. 2017;17(1):272.
 34. Tony M, Wagner M, Khoury H, et al. Bridging health technology assessment (HTA) with multicriteria decision analyses (MCDA): field testing of the EVIDEM framework for coverage decisions by a public payer in Canada. *BMC Health Serv Res*. 2011;11(1):329.

35. Jaramillo HEC, Goetghebeur M, Moreno-Mattar O. Testing multi-criteria decision analysis for more transparent resource-allocation decision making in Colombia. *Int J Technol Assess Health Care*. 2016;32(4):307–14.
36. Angelis A. Evaluating the benefits of new drugs in health technology assessment using multiple criteria decision analysis: a case study on metastatic prostate cancer with the dental and pharmaceuticals benefits agency (TLV) in Sweden. *MDM Policy Pract*. 2018;3(2).
37. Angelis, A. et al., Multiple Criteria Decision Analysis for HTA across four EU Member States: Piloting the Advance Value Framework. *Social Science & Medicine*. 2020; 246: 112595.
38. Morton A. Treacle and smallpox: two tests for multicriteria decision analysis models in health technology assessment. *Value Health*. 2017;20(3):512–5.
39. Marsh K, Sculpher M, Caro J, Tervonen T. The use of MCDA in HTA: great potential but more effort is needed. *Value Health*. 2018;21(4):394–7.
40. Phillips LD, Bana e Costa CA. Transparent prioritisation, budgeting and resource allocation with multi-criteria decision analysis and decision conferencing. *Ann Oper Res*. 2007;154(1):51–68.
41. Tervonen T, Naci H, van Valkenhoef G, et al. Applying multiple criteria decision analysis to comparative benefit-risk assessment: choosing among statins in primary prevention. *Med Decis Making*. 2015;35(7):859–71.
42. von Winterfeldt D, Edwards W. *Decision Analysis and Behavioral Research*. Cambridge (UK): Cambridge University Press; 1986.
43. Bana e Costa CA, Ensslin L, Corrêa ÉC, Vansnick J-C. Decision support systems in action: integrated application in a multicriteria decision aid process. *Eur J Oper Res*. 1999;113(2):315–35.
44. Chalasani N, Younossi Z, Lavine JE, et al. The diagnosis and management of non-alcoholic fatty liver disease: practice Guideline by the American Association for the Study of Liver Diseases, American College of Gastroenterology, and the American Gastroenterological Association. *Hepatology*. 2012;55(6):2005–23.
45. Wong R, Aguilar M, Cheung R, et al. Nonalcoholic steatohepatitis is the second leading etiology of liver disease among adults awaiting liver transplantation in the United States. *Gastroenterology*. 2015;148(3):547–55.
46. Angelis A, Lange A, Kanavos P. Using health technology assessment to assess the value of new medicines: results of a systematic review and expert consultation across eight European countries. *Eur J Health Econ*. 2018;19(1):123–52.
47. Keeney R. *Value Focused Thinking: A Path to Creative Decision Making*. Cambridge (MA): Harvard University Press; 1992.
48. Franco L, Montibeller G. Problem structuring for multicriteria decision analysis interventions analysis interventions. In: Cochran JJ, Cox LA Jr., Keskinocak P, Kharoufeh JP, Smith JC, eds. *Wiley Encyclopedia of Operations Research and Management Science*. New York: Wiley; 2010.
49. Keeney RL, Gregory RS. Selecting attributes to measure the achievement of objectives. *Oper Res*. 2005;53(1):1–11.
50. Food and Drug Administration. Guidance for industry and review staff target product profile—a strategic development process tool. Available from: <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/target-product-profile-strategic-development-process-tool> 2007
51. European Medicines Agency. ICH topic Q 8 (R2) pharmaceutical development, step 5, note for guidance on pharmaceutical development (EMA/CHMP/167068/2004). Available from: https://www.ema.europa.eu/en/documents/scientific-guideline/note-guidance-pharmaceutical-development_en.pdf 2009
52. Schein E. *Process Consultation Revisited: Building the Helping Relationship*. Reading (MA): Addison-Wesley; 1999.
53. Bana e Costa CA, Vansnick J-C. MACBETH—an interactive path towards the construction of cardinal value functions. *Int Transact Oper Res*. 1994;1(4):489.
54. Bana E Costa CA, De Corte J-M, Vansnick J-C. MACBETH. *Int J Inform Technol Decis Making*. 2012;11(2): 359–87.
55. Bana e Costa C, De Corte J, Vansnick J. On the mathematical foundations of MACBETH. In: Greco S, Ehrgott M, Figueira J, eds. *Multiple Criteria Decision Analysis: State of the Art Surveys*. New York: Springer; 2016.
56. Bana e Costa CA, Vansnick J-C. Applications of the MACBETH approach in the framework of an additive aggregation model. *J Multicriteria Decis Anal*. 1997;6(2):107–14.
57. Bana e Costa CA, Corrêa ÉC, De Corte J-M, Vansnick J-C. Facilitating bid evaluation in Public call for tenders: a socio-technical approach. *Omega*. 2002;30(3):227.
58. Bana e Costa C, De Corte J, Vansnick J. M-MACBETH web site. 2016. Available from: http://www.m-macbeth.com/help/pdf/M-MACBETH%203.0.0%20Users%20Guide_BETA.pdf
59. Angulo P. Liver fibrosis, but no other histologic features, is associated with long-term outcomes of patients with nonalcoholic fatty liver disease. *Gastroenterology*. 2015;149(2): 389–97.e10.
60. Ekstedt M. Fibrosis stage is the strongest predictor for disease-specific mortality in NAFLD after up to 33 years of follow-up. *Hepatology*. 2015;61(5):1547.
61. Francque SM, van der Graaff D, Kwanten WJ. Non-alcoholic fatty liver disease and cardiovascular risk: pathophysiological mechanisms and implications. *J Hepatol*. 2016; 65(2):425–43.
62. Wong VW. High prevalence of colorectal neoplasm in patients with non-alcoholic steatohepatitis. *Gut*. 2011;60(6): 829.
63. Wong VW. Coronary artery disease and cardiovascular outcomes in patients with non-alcoholic fatty liver disease. *Gut*. 2011;60(12):1721.
64. Wong VW. Noninvasive biomarkers in NAFLD and NASH—current progress and future promise. *Nat Rev Gastroenterol Hepatol*. 2018;15(8):461.

65. Musso G, Cassader M, Gambino R. Non-alcoholic steatohepatitis: emerging molecular targets and therapeutic strategies. *Nat Rev Drug Discov.* 2016;15(4):249–74.
66. Wree A, Schlattjan M, Bechmann LP, et al. Adipocyte cell size, free fatty acids and apolipoproteins are associated with non-alcoholic liver injury progression in severely obese patients. *Metabolism.* 2014;63(12):1542–52.
67. Canbay A. Non-invasive assessment of NAFLD as systemic disease: a machine learning perspective. *Plos One.* 2019;14(3):e0214436.
68. World Health Organization Collaborating Centre. ATC/DDD index 2016. July 25, 2016. Available from:http://www.whocc.no/atc_ddd_index/
69. National Institutes of Health. ClinicalTrials.gov. Bethesda (MD): National Institutes of Health; 2016.
70. Franco LA, Montibeller G. Facilitated modelling in operational research. *Eur J Oper Res.* 2010;205(3):489–500.
71. Phillips L. A theory of requisite decision models. *Acta Psychol.* 1984;56:29–48.
72. Bana e Costa C, Vansnick J. The MACBETH approach: basic ideas, software, and an application. In: Meskens N, Roubens M, eds. *Advances in Decision Analysis*: Amsterdam: Springer Netherlands; 1999.
73. Fasolo B, Bana e Costa CA. Tailoring value elicitation to decision makers' numeracy and fluency: expressing value judgments in numbers or words. *Omega.* 2014;44:83–90.