PharmacoEconomics (2024) 42:767–782
https://doi.org/10.1007/s40273-024-01365-z

ORIGINAL RESEARCH ARTICLE

Quality Appraisal in Systematic Literature Reviews of Studies Eliciting Health State Utility Values: Conceptual Considerations

Muchandifunga Trust Muchadeyi[1,2] · Karla Hernandez-Villafuerte[1] · Gian Luca Di Tanna[3,4] · Rachel D. Eckford[1] · Yan Feng[5] · Michela Meregaglia[6] · Tessa Peasgood[7,8] · Stavros Petrou[9] · Jasper Ubels[1,2] · Michael Schlander[1,10,2]

Accepted: 22 February 2024 / Published online: 29 March 2024

Abstract

Background The increasing number of studies that generate health state utility values (HSUVs) and the impact of HSUVs on cost-utility analyses make a robust tailored quality appraisal (QA) tool for systematic reviews of these studies necessary.
Objective This study aimed to address conceptual issues regarding QA in systematic reviews of studies eliciting HSUVs by establishing a consensus on the definitions, dimensions and scope of a QA tool specific to this context.
Methods A modified Delphi method was used in this study. An international multidisciplinary panel of seven experts was purposively assembled. The experts engaged in two anonymous online survey rounds. After each round, the experts received structured and controlled feedback on the previous phase. Controlled feedback allowed the experts to re-evaluate and adjust their positions based on collective insights. Following these surveys, a virtual face-to-face meeting was held to resolve outstanding issues. Consensus was defined a priori at all stages of the modified Delphi process.
Results The response rates to the first-round and second-round questionnaires and the virtual consensus meeting were 100%, 86% and 71%, respectively. The entire process culminated in a consensus on the definitions of scientific quality, QA, the three QA dimensions—reporting, relevance and methodological quality—and the scope of a QA tool specific to studies that elicit HSUVs.
Conclusions Achieving this consensus marks a pivotal step towards developing a QA tool specific to systematic reviews of studies eliciting HSUVs. Future research will build on this foundation, identify QA items, signalling questions and response options, and develop a QA tool specific to studies eliciting HSUVs.

1 Introduction

Health state utility values (HSUVs), also referred to as "health utilities", "utility weights", "utility values" or "preference-based health-related quality of life measures", are essential quantitative metrics signifying the cardinal strength of an individual's preference for specific health-related outcomes or health states [1–3]. Health state utility values are typically anchored between 0 (representing death) and 1 (representing full health), and are used to adjust the length of life lived in a specific health state based on the quality of life perceived in that state. The quality-adjusted life-year (QALY) is a widely used generic measure of health outcomes in comparative cost-utility analyses. Quality-adjusted life-years are calculated by multiplying the years lived in each health state by its corresponding HSUV [4]. For example, living 1 year with perfect health is regarded as 1 QALY,

Key Points for Decision Makers

Achieving consensus on the definitions of "scientific quality" and "quality appraisal" for systematic reviews of health state utility values is needed for a standardised framework for assessing studies that elicit health state utility values. A standardised framework helps ensure consistent, transparent and reproducible quality appraisal outcomes

A comprehensive quality appraisal in systematic reviews of health state utility values must evaluate three quality appraisal dimensions—reporting, relevance and methodological quality—for a holistic and rigorous assessment

Our research provides a framework and groundwork for developing a quality appraisal tool designed to elevate the assessment process of studies eliciting health state utility values, thus reinforcing evidence-based healthcare decision making

Extended author information available on the last page of the article whereas living the same year with not-so-perfect health with a HSUV of 0.7 will result in 0.7 QALYs (1 × 0.7). Given that the choice of HSUVs has a significant impact on the outcomes of cost-effectiveness analyses, it is essential to have reliable and unbiased HSUV estimates.

Recent years have witnessed an exponential increase in primary studies eliciting HSUVs, with researchers employing direct (standard gamble, time trade-off or visual analogue scale) and indirect methods (i.e. generic preference-based measures such as the European Quality of Life Five Dimension, Short-Form Six Dimensions, Health Utility Index or mapping algorithms) [1, 5, 6]. Consequently, systematic reviews and meta-analyses have become indispensable tools for synthesising these studies across various decision-making contexts. Irrespective of the source of HSUVs used in an economic evaluation, they must be devoid of recognisable sources of bias. The measurement methodology should be validated, aptly suited to the relevant condition and setting, and aligned with the decision-makers' viewpoint [2]. Therefore, quality appraisal (QA) of HSUV elicitation studies, which aims to ensure the credibility and reliability of HSUV estimates, is central when conducting systematic reviews to inform new health technology assessments [7–13].

However, the conduct of QA in systematic reviews of studies that elicit HSUVs needs to be improved. Recent reviews [5, 14] of studies that elicited HSUVs estimated that only 55% appraised the quality of individual studies in the systematic reviews, which is far lower than that in other research fields [15–18]. The low prevalence of conducting a QA in this field could be partly attributed to the lack of a widely accepted and scientifically developed QA tool [5, 14] specific to this context. This gap may arise from the unique features of these studies, which include multiple applicable study designs and elicitation methods [5]. Consequently, identifying an appropriate tool as recommended by other scholars [8], combining multiple existing tools or developing a bespoke tool is a time-consuming endeavour and significant challenge.

Previous reviews have highlighted several QA tools used to appraise the quality of "broader" health economic evaluation studies [6, 19]. Yet, only a few of these QA tools directly apply to evaluating the quality of studies that elicit HSUVs (Table S.1 of the Electronic Supplementary Material [ESM]). Moreover, these tools differ considerably in their QA items, QA dimensions and synthesis of QA results.

Our previous review [5] showed that reviewers typically assess three QA dimensions: reporting, methodological limitations and risk of bias (RoB), and relevance, albeit the extent to which these three dimensions are considered varies. Additionally, the terminology used to describe the QA process varied widely across the systematic reviews analysed [5], and ranged from terms such as QA or assessment [20–37], critical appraisal [38], RoB assessment [39–45], relevancy and quality assessment [46, 47], assessment of quality and data appropriateness [48], methodological quality assessment [49–53], reporting quality [54, 55], credibility checks and methodological review [56] (see Fig. S1 in the ESM). This inconsistency or heterogeneity largely stems from the absence of a standardised conceptual framework deconstructing scientific quality and the QA process, which is the overarching aim of the current study.

Quality is a multidimensional concept that varies according to the context and field of study [57, 58]. The concept of 'scientific quality' has been introduced to differentiate research quality from other forms of quality, such as product or process quality [59]. Nonetheless, to our knowledge, only one study has attempted to find an agreed-upon definition and has been unsuccessful [60]. Existing definitions range widely, from the likelihood of generating unbiased results in comparative clinical effectiveness science [61] to including dimensions such as relevance and applicability [62, 63], generalisability and imprecision in other research fields [13, 58, 64–66].

Stemming from the heterogeneous definitions of scientific quality are the considerable variations and inconsistencies in how QA has been and is currently applied in most systematic reviews (not only in HSUVs) [5, 67]. For example, Viswanathan and colleagues [65, 66], basing their argument on Cochrane's recommendations, differ significantly from the Grading of Recommendations Assessment, Development and Evaluation Working Group (GRADE) framework [68]. On the one hand, Viswanathan and colleagues [65, 66] advocate for the use and evaluation of the RoB rather than quality assessment as the term quality is used variedly across many fields. On the other hand, the GRADE framework considers quality to be more than just RoB, as quality also includes imprecision, inconsistency, indirectness of study results and publication bias. The GRADE framework later uses these dimensions to make overall judgements regarding the strength of the body of evidence [68].

Often, systematic review authors account for reporting quality and consider this as the overall study quality [69], yet the quality of reporting may not reflect the quality of the study [70, 71]. In fact, a focus on reporting quality alone may overestimate the overall quality of a study [69]. We previously posited that a comprehensive QA should encompass all three QA dimensions—all three QA dimensions are *necessary and sufficient components* [5]. Evaluating relevance and methodological limitations is contingent on reporting quality; after all, appraisal can only be based on what is documented.

Given this context, there is a pressing need for available tools and guidelines to offer more explicit directives on the constructs, dimensions and items that are pivotal for rigorous QA. A widely accepted definition of scientific quality—one that identifies the dimensions and constructs pertinent to QA in systematic reviews of studies eliciting HSUVs—might present a viable solution to the existing conundrum [58].

Considering the importance placed on HSUVs in the health economic evaluation of new technologies and interventions [9], and the inconsistencies highlighted above, developing a QA tool specific to HSUVs is pertinent and timely. The present study aims to set the stage for future work on developing an evidence-based QA tool specific to studies eliciting HSUVs by conceptualising scientific quality and QA in systematic reviews of studies eliciting HSUVs. This entails:

1. Establishing a working definition for scientific quality and QA.
2. Establishing and defining the relevant dimensions for the QA of studies eliciting HSUVs.
3. Defining the scope of a QA tool specific to studies eliciting HSUVs.

2 Methodology

This study builds on our previous rapid review of the current nature of QA in systematic reviews of studies eliciting HSUVs [5]. From this review, we discerned three QA dimensions frequently assessed in such systematic reviews, albeit to varying degrees. Additionally, we collated terminologies (see Fig. S.1 in the ESM) and definitions related to "quality" and "QA" (see Tables S.2 and S.3 in the ESM). For pragmatic reasons, we opted for a modified Delphi technique to facilitate a consensus among experts regarding the definitions of quality and QA, as well as the components that should be integral to QA in our specified context.

2.1 Study Design

The conventional Delphi method, developed by the RAND Corporation in the 1950s, is a structured technique for achieving a formal consensus on specific issues among panel members [72–74]. This method emphasises non-face-to-face interactions, ensuring anonymity or quasi-anonymity among participants throughout the process. Instead, communication within this framework hinges on a series of iterative questionnaires designed to capture insights and opinions from the participants [72, 74–76].

We identify our methodological approach as a modified Delphi technique [72, 75], sometimes referred to as a modified nominal group technique [75]. We adopted this modification for two primary reasons: our panel comprised seven experts, and after two rounds of online questionnaires, we conducted a virtual face-to-face meeting with the experts. Subsequent sections delve into the reasoning and elaborate on critical components of the study, such as participant selection, the preservation of anonymity, the iterative process and group response (or consensus) [72].

2.1.1 Steering Committee

The steering committee comprised four members of the project team: MTM (who served as the project leader), KH, RE and MS. Their responsibilities encompassed conducting literature review(s), developing and piloting the questionnaires, analysing and reporting the responses at each stage, organising the virtual panel discussion and the overall moderation of the process. Importantly, all committee members refrained from expressing their personal opinions during the consensus-building exercises.

2.1.2 Selection of Experts

The steering committee sought to enlist experts seasoned in conducting systematic reviews of studies eliciting HSUVs, and are also recognised experts with knowledge and experience in health technology assessments, mapping studies, health-related quality of life and core health economic evaluations. Moreover, potential contributors should have authored peer-reviewed publications involving one or more of these domains.

Given the stringent inclusion criteria and the limited number of systematic reviews that appraised the quality of their incorporated studies (40 out of 73) [5], the eligible potential participant pool was considerably restricted. To ensure representation from the desired domains in the eligibility criteria, we set a minimum target of five experts, without capping the maximum.

To achieve our recruitment goal, we *purposively* constituted an international multidisciplinary expert panel. Personalised emails were sent to 23 experts between September and December 2021. These individuals were identified through the articles deemed eligible for the rapid review [5], the QA tools referenced in those articles, or reference searches from these articles.

2.2 Modified Delphi Rounds

The appeal of the traditional Delphi method stems from the principles of expert anonymity, controlled feedback and iterative discussions during the process [72–75]. Advocates of the Delphi technique believe that its structured and controlled nature helps counter the drawbacks often seen in face-to-face meetings, such as undue influence from other experts (dominance) and group conformity (defined as groupthink) [72, 73, 75].

2.2.1 Expert Anonymity

True anonymity is ensured when no one (including the researcher) can trace back a response from a respondent [76]. In this study, the project leader managed all communication exclusively through individualised e-mail interactions, instead of sending group e-mails. Furthermore, the analysis was conducted without identifiers that could be linked back to a particular expert. As a result, true anonymity was unattainable. Instead, we maintained quasi-anonymity, where only one researcher (MTM) knew all the respondents and their responses [76].

2.2.2 First-Round Questionnaire

In the first-round questionnaire, Section A was designed to gather information on the characteristics and expertise of contributing experts in conducting systematic reviews of studies eliciting HSUVs. This section aimed to understand the experts' background and prior experience with systematic reviews of HSUVs.

Section B explored the experts' perspectives on various conceptual issues related to quality and QAs in systematic reviews of studies eliciting HSUVs. Specifically, we refrained from providing predefined definitions for "quality" and "quality appraisal." Instead, the experts were requested to offer concise definitions and comments regarding these terms. They were also asked to indicate their agreement on whether they considered QA as an integral part of systematic reviews of studies eliciting HSUVs.

The definitions for reporting, methodological limitations and RoB, and relevance quality dimensions presented in the first-round questionnaire were crafted by the steering committee, drawing upon the prior literature [15–17, 67] and their theoretical understanding of the terms. The experts were then asked to rate their agreement on whether they considered these three dimensions to be fundamental aspects of systematic reviews of studies eliciting HSUVs.

To comprehensively capture the experts' opinions and insights, a combination of a five-point Likert rating scale (with a range from strongly disagree to strongly agree) and open-ended questions were included. The open-ended questions allowed the experts to express and discuss their views and opinions in greater detail. The first-round questionnaire can be found in the ESM.

2.2.3 Controlled Feedback

For controlled feedback, the first-round's responses and comments were descriptively (quantitatively and qualitatively) summarised; any potential identifiers were removed to preserve anonymity. The input from the experts was qualitatively analysed, key concepts and themes were extracted, and then presented in text boxes. Integrating these insights with themes from the existing literature [60–63, 77–79], we formulated working definitions of scientific quality and QA for the second-round questionnaire (see Tables S.2 and S.3 of the ESM). Similarly, based on experts' feedback and previous literature [15–17, 67], we fine-tuned the working definitions of reporting, methodological limitations (RoB) and relevance for the second-round questionnaire (see Table S.4 of the ESM).

In the quantitative analysis, we calculated frequencies for each level on the Likert scale or the responses in "yes" or "no" format, and presented these as bar charts. A comprehensive report of the first-round responses was shared through separate e-mails with all the experts.

2.2.4 Second-Round Questionnaire

The proposed working definitions for quality, QA and three QA dimensions (i.e. reporting, methodological limitations and RoB, and relevance) from the first-round questionnaire were provided as part of the second-round questionnaire. A statement about the purpose and scope of a QA tool for studies eliciting HSUVs was also presented. Consensus was determined a priori when an item reached ≥85% agreement (i.e. six out of seven experts). The consensus level was later changed to 83% (5/6) during the analysis, to align with the observed questionnaire response rate.

Unlike the first-round questionnaire, all responses were coded as binary "yes" or "no" format and analysed descriptively (quantitatively and qualitatively). The second-round questionnaire is presented in the ESM. Similar to the first round, a report of the second-round responses was prepared and shared as controlled feedback with all the experts through individualised e-mails.

2.2.5 Virtual Panel Discussion

Following the second-round questionnaire, it became apparent that agreement on some of the remaining issues would be more effectively facilitated through an open virtual panel discussion instead of a third round of anonymous questionnaire. Thus, in consultation with the experts, the steering committee set aside anonymity and invited all experts to participate in a virtual panel discussion.

All experts were invited to join a 3-hour virtual consensus panel discussion held on 31 May, 2023 using the Zoom virtual conferencing platform. Real-time polling was used during the meeting to facilitate a consensus on each point of discussion. The session was recorded to enable a thorough review and post-discussion transcription.

The video and audio recordings were transcribed by a steering committee member (MTM) and deleted once the transcription had been verified for accuracy. The polling results and discussion content were analysed descriptively and anonymously.

3 Results

3.1 Description of the Expert Panel

Seven international experts, all health economists, except for a biostatistician with a health economics background, consented to participate in the modified Delphi study. The panel members represented a diverse range of countries, including the UK (three experts), Italy (one expert), Australia (two experts who later moved to the UK and another who later moved to Switzerland) and Germany (one expert).

Importantly, the majority (6/7) had significant prior experience with systematic reviews of HSUVs, 5/7 rated themselves as very familiar with systematic reviews, whereas the remaining two considered themselves somewhat familiar. Among them, three had previously co-authored a systematic review of studies eliciting HSUVs, four had participated as contributing health economists and one as a contributing biostatistician.

One expert in the panel had not directly participated in a systematic or rapid review of studies eliciting HSUVs. However, this expert has conducted significant methodological research on HSUVs and contributed to the latest developments in both methodological and applied research in the field.

3.2 Response Rates to the Questionnaires and Virtual Panel Discussion

Response rates to the first-round and second-round questionnaires and the virtual consensus meeting were 100%, 86% and 71%, respectively. The first-round questionnaire was completed on 8 April, 2022, the second round on 23 September, 2022, and the virtual meeting was held on 31 May, 2023. The completion rate was 100% for all questions regarding the characteristics of the experts and conceptual considerations.

3.3 Definition of Scientific Quality and QA

Key constructs extracted from the definitions of quality suggested by the experts, combined with themes from the existing literature, include “using the term scientific quality instead of quality”, “the validity of results and questionnaire”, “attention to methods and methodology”, “accurate and comprehensive reporting”, transparency”, “relevance, “replicability”, “reproducibility” and “bias minimisation”.

Key constructs extracted from the definitions of QA suggested by the experts, combined with themes from the existing literature, include “independence”, “systematic”, “explicit”, “transparency” and “thorough (robust).” Some experts also highlighted the need to consider internal validity, RoB, reporting standards, and methodological or reporting quality in QAs (see Box S.1 of the ESM).

Figure 1 illustrates the consensus from the second questionnaire on the term “scientific quality”, with minor comments regarding word order and spelling (see Box S.2 of the ESM). The agreed working definition of scientific quality following the second round of questionnaires and minor revisions by the steering committee is reported in Table 1.

Nevertheless, disagreement regarding the definition of QA was observed (Fig. 1). Some experts advocated for a more concise definition, while others suggested removing the word “robustness” from the definition (Box S.3 of the ESM). After substantial deliberation and revisions in the virtual meeting (for intermediary results, see Box S.4 of the ESM) and further refinement during manuscript proofreading and revision, the consensus definition is delineated in Table 1.

3.4 Terms That Can be Considered Synonymous with QA

In the first-round questionnaire, the experts presented varied opinions on terms synonymous with QA. Most of the experts (5/7) initially suggested that the terms “quality assessment” and “critical appraisal” could be used synonymously with QA. However, terms like “credibility checks”, “reporting quality assessment” and “data appropriateness” were not viewed as direct synonyms. Instead, most experts regarded these terms as constituents or dimensions of a QA (see Box 1).

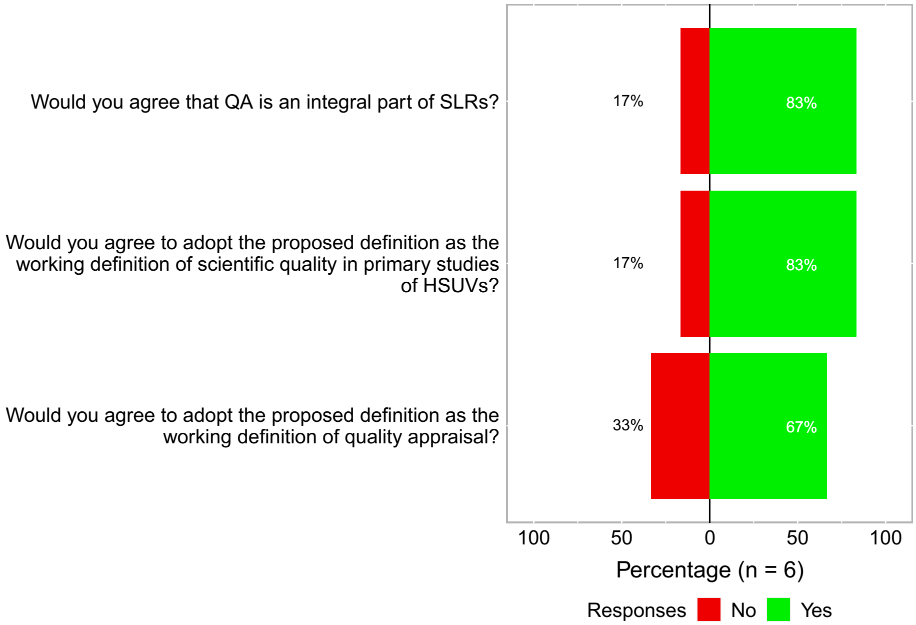


Fig. 1 Second-round questionnaire: experts' opinions on the proposed definitions of "scientific quality" and "quality appraisal (QA)" and how important QA is in systematic literature reviews (SLRs) [systematic reviews]) of studies eliciting health state utility values (HSUVs). > 83% (5/6) = agreement or consensus

Table 1 Key definitions of QA components for systematic literature reviews of studies eliciting health state utility values

Term	Agreed working definition or description
Scientific quality	Scientific quality is the extent[a] to which a study eliciting health state utility values adheres to established standards[b] in its design, conduct[c], analysis and reporting, influencing the internal and external validity of elicited health state utility values
QA	QA is the systematic[d] application of explicit, transparent and reproducible methods to assess key attributes[e] in the design, conduct, analysis and reporting of studies that elicit health state utility values
Reporting quality	Reporting quality is the extent to which a research article explicitly and transparently details its study's design, conduct, analysis and results
RoB[f]	RoB is the likelihood that flaws in the study's design or conduct, analysis or reporting might lead to invalid or misleading results
Relevance[g]	Relevance (or applicability) is the extent to which health state utility values accurately represent the population and context of interest to the review question, decision context policy or decision makers' requirements
Scope of a QA tool	A QA tool specific for systematic literature reviews of primary studies eliciting health state utility values should: 1. Assess both the reporting and methodological quality 2. Assesses the relevance (representativeness, applicability, transferability and or generalisability) 3. Apply to multiple study designs (e.g. randomised controlled trials, cohort case controlled or cross-sectional studies) 4. Apply to different methods of health state utility values elicitation techniques (e.g. direct or indirect methods)

QA quality appraisal, *RoB* risk of bias

[a] For scientific quality to be measurable, it should be expressed as a degree or extent of meeting predefined or generally acceptable standards

[b] Other terms for "established standards" are "norms" or "benchmarks"

[c] The term "conduct", refers to the procedures and protocols followed in the process followed in studies eliciting health state utility values. This includes various aspects such as participant sampling and selection, instrument administration and valuation methods among others

[d] Here, systematic refers to how the QA should be done—and synonyms include orderly, structured, methodological, scientific or organised

[e] The word "attributes" can be used interchangeably with the words "characteristics", "features" or "variables"

[f] The Dictionary of Health Economics" by Antony Culyer (2005) defines bias as any "systematic difference between the empirical results of an analysis and the true facts of the case (e.g. the difference between the distribution of values in a sample and the actual values of the population from which the sample is drawn) [91]. To evaluate the RoB, one must first identify any methodological limitations or flaws in the study's design, conduct, analysis or reporting, then determine the extent to which these flaws make the study susceptible to bias—a two-step process

[g] The purpose of assessing study relevance is to determine the health state utility values' suitability for their intended use. Specifically, evaluation of study relevance entails identifying potential differences between study attributes such as population (participants or patients) characteristics, setting and methods (measurement instrument, health states description and valuation) and the population and context of interest to the review question, decision context, policy or decision makers' requirements

"All of the elements listed here ***are necessary for quality appraisal, but I would not call them synonymous with quality appraisal since, by themselves, they are not sufficient.*** *Still, I crossed critical appraisal and validity check. I crossed critical appraisal since conducting a quality check is inherently a critical appraisal of the methods/results, which covers the risk of bias assessment and the other points. I crossed validity check, because I believe internal and external validity to be the sufficient and necessary conditions of quality appraisal."*

"All the boxes above ***relate to some extent to quality appraisal*** *but the ones that I have not ticked are more focused to specific concepts (data, methods, reporting) whilst I considered quality appraisal concept in a general way."*

"Assessment of reporting quality and data appropriateness ***may be one component of a quality appraisal****. Critical appraisal would include considerations of generalisability but some may separate methodological quality assessment/risk of bias from questions of relevance of applying research findings to a particular context."*

"Many of these terms I think ***refer to components of quality appraisal****. QA is a broader concept."*

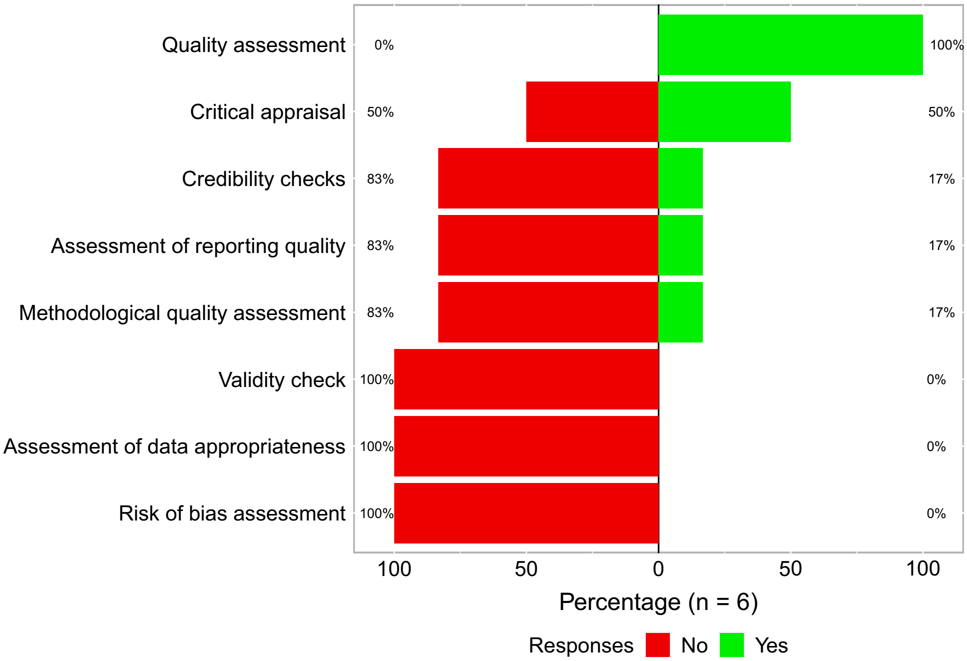
Box 1 First-round questionnaire: an extract of the experts' comments on terms considered synonymous with quality appraisal (QA)

Contrary to the first round, in the subsequent round, there was a unanimous consensus that the term "quality appraisal" can be used synonymously with the term "quality assessment" (Fig. 2). However, the experts indicated that it should not be considered synonymous with a "critical appraisal" (50% disagreed) or any of the other presented terms (100% disagreed).

3.5 QA Dimensions

In the first-round questionnaire, the majority (5/7) of the experts strongly agreed, and 2/7 agreed that reporting and methodological limitations and RoB are essential for QA of systematic reviews of HSUVs. However, opinions on the "relevance" dimension diverged considerably. One expert was neutral (neither agree nor disagree), three agreed and the remaining three strongly agreed on the significance of the relevance dimension to QA in systematic reviews of HSUVs. Some experts suggested that the definitions of these three dimensions required further refinement and elaboration. Notably, although supportive of reporting quality in a QA, one expert pointed out cases in which it may not hold equal weight (see Box 2).

Fig. 2 Second-round questionnaire: experts' opinions on terms considered synonymous with quality appraisal. ≥83% (5/6) = agreement or consensus



"RoB to me does not encompass only methodological quality as it is related to design, conduct and reporting of a study."

"There is a case for not presenting reporting quality as an equal dimension. A poorly conducted study may lead to high reporting quality which could mislead, and a poorly reported study will be reflected in the uncertainty around responses for RoB and Relevance. An initial assessment of reporting quality can be very helpful for the reviewer - it's more a question of how it is presented."

"This relates to my previous answer "Reporting quality" is about transparency of reporting, it's not quite about the quality of the study being included in the review per se. Similar to scores of quality that are applied to RCTs. Whether a QA review then needs to use the information on what is reported in the studies to make judgements (which is what items 2 and 3 relate to), depends on the purpose of the review. Item 3 definition above doesn't make sense grammatically."

"Your three definitions should be improved. For example, what do you mean by "a set of parameters in the design and conduct of a study has been described"? Do you intend to refer to "study reporting" in general? And by "the extent to which the study has been executed"? Do you intend to refer to "methodological soundness"? Moreover, the HSUV study relevance could be also for clinical/epidemiological research".

Box 2 First-round questionnaire: extract of the experts' comments regarding the definitions of the three quality appraisal (QA) dimensions and whether to include these in a QA tool specific for studies eliciting health state utility values (HSUVs). *RCTs* randomised controlled trials, *RoB* risk of bias

After rephrasing the three QA dimensions' definitions (see Table S.4 of the ESM), there was unanimous agreement among the experts to include them in a QA tool specific to HSUVs. Furthermore, as shown in Fig. 3, a consensus (yes ≥83%) was reached on the proposed definition of the three QA dimensions.

A remaining concern was the practical application of the "relevance" dimension. The experts wondered how the issue of study relevance was related to the study perspective (experienced patient utilities vs general population or ex-ante vs post-ante utilities) and the study population for HSUV valuation (see Box 3).

3.6 Scope of a QA Tool for Systematic Reviews of Studies Eliciting HSUVs

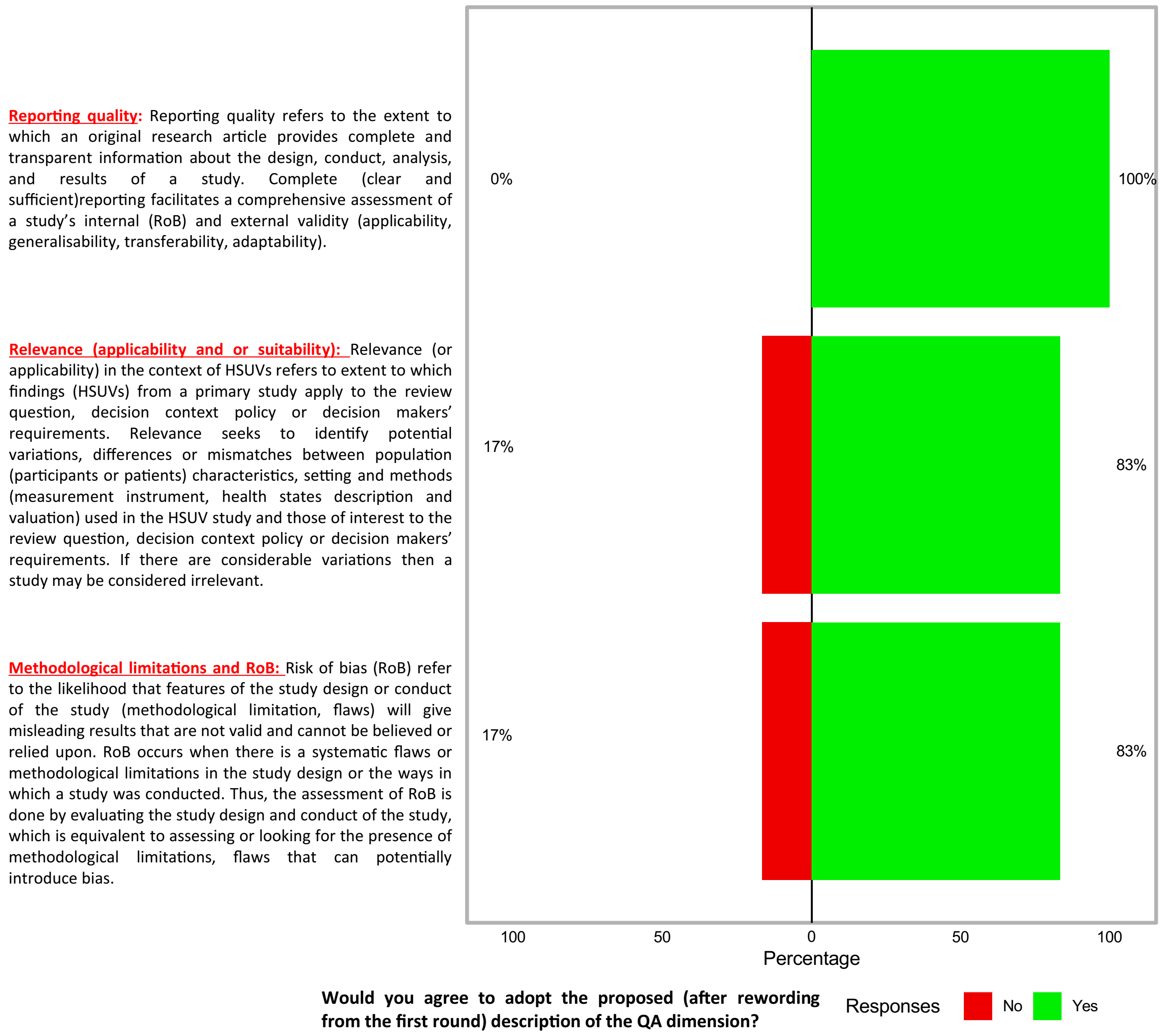
A question was introduced in the second-round questionnaire asking the experts what they would consider a plausible scope of a QA tool in systematic reviews of studies eliciting HSUVs. Figure 4 depicts the experts' views on how broad the scope of the proposed QA tool in development should be.

Notably, the experts disagreed with the suggestion that the questionnaire should be reasonably short and allow a consistent and reliable quality assessment of different backgrounds (Box 4). One expert suggested that a QA tool specific to studies eliciting HSUVs should not exclude mapping studies from the tool's scope a priori because some items evaluated for direct methods, such as sample size, relevance and reporting, also apply to mapping studies. Another concern was whether randomised controlled studies are applicable research design methods for studies eliciting HSUVs.

4 Discussion

We elucidated the essential conceptual considerations for developing a QA tool specific to systematic reviews of studies eliciting HSUVs. The systematic reviews can be for publication, health economic model development or broader health technology assessment purposes .This study defined scientific quality, QA and three QA dimensions by synthesising insights from a comprehensive literature review and opinions from seven international experts via three modified Delphi rounds. Furthermore, this study proposed a preliminary scope (boundaries) for future QA tools.

Scientific quality is a nuanced multidimensional concept highly specific to the context and field of study [13,



Reporting quality: Reporting quality refers to the extent to which an original research article provides complete and transparent information about the design, conduct, analysis, and results of a study. Complete (clear and sufficient)reporting facilitates a comprehensive assessment of a study's internal (RoB) and external validity (applicability, generalisability, transferability, adaptability).

Relevance (applicability and or suitability): Relevance (or applicability) in the context of HSUVs refers to extent to which findings (HSUVs) from a primary study apply to the review question, decision context policy or decision makers' requirements. Relevance seeks to identify potential variations, differences or mismatches between population (participants or patients) characteristics, setting and methods (measurement instrument, health states description and valuation) used in the HSUV study and those of interest to the review question, decision context policy or decision makers' requirements. If there are considerable variations then a study may be considered irrelevant.

Methodological limitations and RoB: Risk of bias (RoB) refer to the likelihood that features of the study design or conduct of the study (methodological limitation, flaws) will give misleading results that are not valid and cannot be believed or relied upon. RoB occurs when there is a systematic flaws or methodological limitations in the study design or the ways in which a study was conducted. Thus, the assessment of RoB is done by evaluating the study design and conduct of the study, which is equivalent to assessing or looking for the presence of methodological limitations, flaws that can potentially introduce bias.

Would you agree to adopt the proposed (after rewording from the first round) description of the QA dimension?

Fig. 3 Second-round questionnaire: consensus on defining the three quality appraisal (QA) dimensions for systematic literature reviews of health state utility values (HSUVs). *RoB* risk of bias

58, 64–66]. Contrary to previous unsuccessful attempts to define quality [60], a consensus on the working definition of *scientific quality* was reached during the second-round questionnaire. It took a further virtual panel discussion for the experts to agree on the working definition of QA, underscoring the existing ambiguity and heterogeneity in applying the terms.

In its simplest form, QA involves evaluating a study's *scientific quality*. Our intention is not to dictate precisely how studies that elicit HSUVs should be conducted; however, we recognise that a universal scientific quality standard for these studies is crucial. To facilitate this, we propose a QA tool specific to HSUV eliciting studies that may promote adherence to high-quality standards. This proposition aligns with the sentiments of Wolowacz and colleagues [58], who advocated that HSUVs intended for policy and decision making should be derived from validated methods, aimed at minimising bias and be relevant to the condition, population and decision-maker's perspective. Detailing all elements that constitute scientific quality and QA in systematic reviews of studies eliciting HSUVs is an embarking step toward realising this goal.

Evident from the definition of scientific quality and QA are the primary constituents of QA in systematic reviews

Regarding reporting quality: *"I'm not sure how to interpret the final term 'adaptability'*

Regarding methodological quality: *Rephrase to 'are systematic flaws'*

Regarding relevance*: This is a good definition but I don't think the final sentence adds anything.*

One point I was thinking about is whether the way relevance is being described here might pose a problem for HSUV use in practice. For example, it is well known that the experienced utilities of patients in a certain health state are different from how people from the general public would value those health states. I guess technically one could argue that there is a mismatch here, however, there are cases where this mismatch is actually wanted (i.e., as discussed in the report, some agencies prefer values from the public). I guess this comment is more a question. Would this situation represent a mismatch between population characteristics? And would that mean that according to the tool that is developed in this study, that health utility states elicited from the public are not relevant for a patient population, which would conflict with the use of HSUV in practice?

Box 3 Second-round questionnaire: extract of the experts' comments regarding the definitions of the three quality appraisal dimensions. *HSUVs* health state utility values

of studies eliciting HSUVs—reporting, relevance, methodological limitations and RoB. The intricate relationship between these dimensions warrants special consideration. Methodological limitations or flaws, such as high attrition rates, incorrect use of utility assessment tools or improper statistical methods, can compromise the reliability of the derived HSUV estimates, underscoring the need to inspect all studies contributing to a systematic review for methodological flaws and the likelihood of bias that may result from such.

Unlike the commonly assessed clinical outcomes in clinical effectiveness studies, which can be considered comparable across different settings and not affected by context, HSUVs are context sensitive [5]. Primary studies eliciting HSUVs have limited relevance to a research question, policy or decision framework if they differ significantly in attributes, such as populations included, specific health conditions, measurement instruments, health state descriptions, and valuations or settings. Thus, it is essential to assess the relevance of HSUVs to specific research questions or decision frameworks.

More importantly, a comprehensive study report is essential to assess the previous two dimensions, rarely do systematic review authors undertake additional steps to conduct principal investigators of the primary studies eliciting HSUVs. Thus, all three QA dimensions are *necessary* and *sufficien*t components for a robust QA tool. Achieving a consensus on the definitions of the three QA dimensions will further augment efforts to mitigate ambiguity among these three pillars of scientific quality regarding studies eliciting HSUVs.

Our endeavour to develop a QA tool specific to systematic reviews of studies eliciting HSUVs aligns with the efforts of other research groups [6, 80]. However, our study diverges from these efforts by proposing a unified QA tool designed to distinguish and evaluate reporting, relevance, methodological limitations and potential bias in primary studies that elicit HSUVs.

The three QA dimensions are nothing new to systematic reviews of studies eliciting HSUVs. Many systematic review authors have already considered these dimensions, albeit inconsistently [5]. For instance, 18% of the systematic reviews of studies eliciting HSUVs published between 2015 and 2022 evaluated all three QA dimensions [5]. Furthermore, the need for differentiating these three QA dimensions has been underscored in various studies [62, 63, 69, 81, 82].

The opinions of experts regarding study relevance and perspective require careful consideration and discussion. This study provides a succinct definition of study relevance concerning HSUVs. A generally accepted assumption in conventional normative health economics is that the society is rational and the primary goal of choosing certain health technologies over others is to maximise health [83]. In this regard, the health component is widely measured by using QALYs as a generic outcome. What health economists often disagree on is whose utility values should be used in the QALY computation to maximise health, fueling debates about the perspective of analysis [83, 84].

While the perspective of analysis in health economic evaluations is primarily used to prescribe the breadth of which cost to include, another essential perspective to consider is whose utilities (preferences or utility weights) are considered when valuing health. The choice of which HSUVs are appropriate depends on the viewpoint (perspective) of the decision maker (i.e. whose point of view is the decision on healthcare intervention being made?) [85]. A

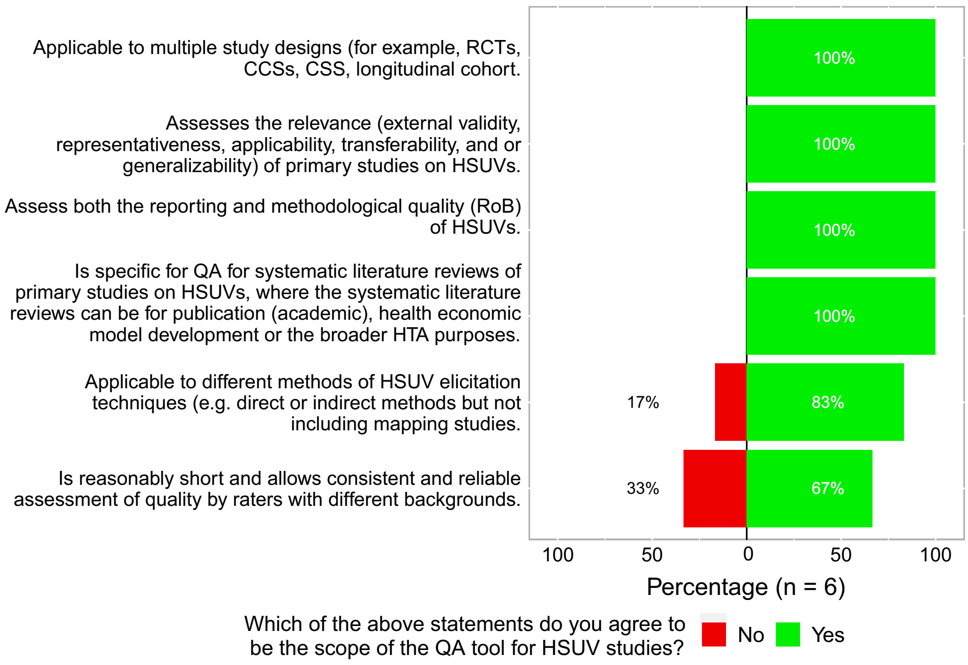


Fig. 4 Second-round questionnaire: experts' opinions on the proposed scope of a quality appraisal (QA) tool specific to studies eliciting health state utility values (HSUVs). *CCSs* case controlled studies, *CSS* cross sectional studies, *HTA* health technology assessment, *RCTs* randomised controlled studies, *RoB* risk if bias, > 83% (5/6) = agreement or consensus

> *I agree to include item 4, but not sure that I will follow the example. How one use a RCT study to develop HSUVs? Do you mean use observational data from a RCT?*
>
> *Despite "mapping" is a particular technique for deriving HSUVs, I would not exclude it a priori. Some considerations (e.g., sample size, relevance, reporting) can also apply to mapping studies.*
>
> *I don't think the length of the tool is relevant here.*

Box 4 Second-round questionnaire: an extract of the experts' comments on the proposed scope of a quality appraisal tool specific to studies eliciting health state utility values (HSUVs). *RCT* randomised controlled trials

related discussion borrowed from mainstream behavioural economics evolves around decision utilities (ex-ante preferences of decision makers or society as a whole for states they have not experienced) and experienced utilities (post-ante preferences of individuals who have experienced the health states) [84]. Different perspectives can include those of the general population, patients, clinicians or decision makers. Notably, utilities derived from patients can differ from other utilities owing to variations in the decisional context and other factors [83]. Each type of utility has its strengths and limitations, and the choice depends on the specific evaluation objectives and requirements. An explicit statement of the purpose of an economic evaluation and a systematic review of studies eliciting HSUVs is crucial, as it determines the analysis perspective, relevant population, setting and health valuation techniques, all of which are integral components of the relevance dimension.

A fundamental first step in any scale or QA tool development is a clearly defined boundary between what the tool can and cannot do [60, 62, 86]. A third of the current panel members disagreed that a QA tool specific to studies eliciting HSUVs should be reasonably short and allow reliable and consistent assessments by raters with different backgrounds. This result can be considered inconsistent with previous tool development exercises [62]. Nevertheless, a pertinent consideration when developing a QA tool is the anticipated burden on the raters. A lengthy questionnaire with too many items and signalling questions may be burdensome and deter the QA process.

Furthermore, item and construct overlap is more likely to occur with an increasing number of items and ultimately, the length of the questionnaire. Another reason for revising the QUADAS-1 to QUADAS-2 was that users reported problems rating certain items that seemed to overlap. Their [62] proposed solution was to limit the number of domains and signalling questions. Unfortunately, there are no strict guidelines on how many items to include, how long a QA tool should be, or how long raters should take to complete the QA of a single study eliciting HSUVs.

It is also prudent to use simple language that is easily understandable by raters with different backgrounds. Using complex terms such as "construct-irrelevant variance" and "quality of construct representation", as done by Eiring and colleagues [80], could be misleading and fuel the existing inconsistencies in QA in systematic reviews of studies eliciting HSUVs. In developing the ROBINS-I tool [86], the developers noted the challenges brought about by the variations in terminology used to describe different domains and items. For example, terms such as "selection bias" may be confused with related but different terms such as "confounding". For the same reasons, the recent version of RoB 2 avoided the use of terms such as "selection bias", "performance bias", "attrition bias" and "detection bias" [87]." While most systematic reviews of studies eliciting HSUVs will likely be performed by people with relevant knowledge in health economic evaluations, using easily understood terminology remains essential.

The major strength of this study is its ability to triangulate the findings of a previous literature review with the knowledge of experts in the field. Keeping the experts anonymous during the two rounds of questionnaires ensured that the contributing experts freely expressed their views without being influenced or dominated by others [72–76]. Akin to a focus group discussion, the virtual meeting afforded experts space to ask questions and explain their viewpoints, delving deeper into the subjects at hand—a depth often missing in individual-based questionnaires [88]. It is also widely accepted that a well-facilitated interactive meeting can bolster participants' contributions by providing opportunities to clarify or rephrase questions, thus enhancing both comprehension and quality of response [89]. As a result, we captured the depth and complexity of experts' views, particularly for the more complex topics of interest that were unresolved after the two rounds of questionnaires, and reached a consensus on the definition of QA.

Because of the limited number of researchers with expertise required to contribute to this study, the generalisability of our findings may be limited. Nevertheless, experts were drawn from different countries and had diverse competencies regarding HSUVs (i.e. systematic reviews of HSUVs, health technology assessments, mapping algorithms, health-related quality of life and mainstream health economic evaluations). Furthermore, combining questionnaire-based approaches and face-to-face meetings further reduced the need for a larger sample size. For example, the nominal group technique is typically effective when the group size is small [75, 90].

Another limitation that often affects qualitative research is that the steering committee may impose its views throughout the study. To limit this bias, the steering committee sought to ensure high levels of transparency throughout the study phases, for example, using text boxes to report raw comments from experts.

5 Conclusions

This study defined scientific quality and QA in the context of systematic reviews of studies eliciting HSUVs. In addition, a consensus was reached on the scope and boundaries of a QA tool specific for this context. Based on these, the experts concurred that an effective QA should discern among reporting, relevance and methodological quality while being applicable to multiple design features and health elicitation techniques. This consensus represents a fundamental step towards harmonising or standardising the QA process in this field. Future work should leverage this foundation to identify QA items, signalling questions, and response options and develop a QA tool for systematic reviews of studies eliciting HSUVs.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40273-024-01365-z.

Acknowledgements We extend our appreciation to an expert of health economic evaluations, who participated in the first round of questionnaires for this study. We are keeping the identity anonymous in accordance with the Delphi process. This person's contribution to the study helped inform and enrich our work. We also acknowledge the members of the Division of Health Economics at DKFZ (https://www.dkfz.de/en/gesundheitsoekonomie/index.php) for their questions and constructive feedback when this project was presented in team meetings.

Declarations

Funding Open Access funding enabled and organized by Projekt DEAL. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of interest Muchadeyi M, Hernandez-Villafuerte K, Di Tanna GL, Eckford R, Feng Y, Meregaglia M, Peasgood T, Petrou S, Ubels J, and Schlander M have no conflicts of interest that are directly relevant to the content of this article.

Ethics approval Ethical approval for Zoom conference recording was obtained from all experts.

Consent to participate The conditions to participate were that all analysis and reporting of the findings would be done anonymously and the video and audio recordings would be deleted once the transcriptions had been verified for accuracy.

Consent for publication Not applicable.

Availability of data and material The datasets generated and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Code availability Not applicable.

Author Contributions Members of the steering committee (Muchandifunga Trust Muchadeyi, Karla Hernandez-Villafuerte, Rachel Eckford and Michael Schlander) contributed to the study's conception, initial design, data analysis plan, questionnaire development, overseeing its piloting, refinement and data collection. Muchandifunga Trust Muchadeyi and Karla Hernandez-Villafuerte performed formal analyses. Reporting during the questionnaire rounds was performed by Muchandifunga Trust Muchadeyi, Karla Hernandez-Villafuerte and Rachel Eckford. The first draft of the manuscript was written by Muchandifunga Trust Muchadeyi, and all members of the steering committee (Muchandifunga Trust Muchadeyi, Karla Hernandez-Villafuerte, Rachel Eckford and Michael Schlander) reviewed, edited and commented on previous versions. Michael Schlander provided supervision throughout the project. Expert panellists including (Gian Luca Di Tanna, Yan Feng, Michela Meregaglia, Tessa Peasgood, Stavros Petrou and Jasper Ubels) provided their valuable insights throughout the all phases of the Delphi study as international, multidisciplinary experts. Expert panellists including (Gian Luca Di Tanna, Yan Feng, Michela Meregaglia, Tessa Peasgood, Stavros Petrou and Jasper Ubels) contributed to the revision of the project design and comprehensively revised and edited the manuscript. All authors read and approved the final manuscript.

References

1. Chang EM, Saigal CS, Raldow AC. Explaining health state utility assessment. JAMA. 2020;323(11):1085–6.
2. Wolowacz SE, Briggs A, Belozeroff V, Clarke P, Doward L, Goeree R, et al. Estimating health-state utility for economic models in clinical studies: an ISPOR Good Research Practices Task Force report. Value Health. 2016;19(6):704–19.
3. Payakachat N, Murawski MM, Summers KH. Health utility and economic analysis: theoretical and practical issues. Expert Rev Pharmacoecon Outcomes Res. 2009;9(4):289–92.
4. Whitehead SJ, Ali S. Health outcomes in economic evaluation: the QALY and utilities. Br Med Bull. 2010;96:5–1.
5. Muchadeyi MT, Hernandez-Villafuerte K, Schlander M. Quality appraisal for systematic literature reviews of health state utility values: a descriptive analysis. BMC Med Res Methodol. 2022;22(1):303.
6. Zoratti MJ, Pickard AS, Stalmeier PFM, Ollendorf D, Lloyd A, Chan KKW, et al. Evaluating the conduct and application of health utility studies: a review of critical appraisal tools and reporting checklists. Eur J Health Econ. 2021;22(5):723–33.
7. Longworth L, Rowen D. NICE DSU Technical Support Document 10: The Use of Mapping Methods to Estimate Health State Utility Values [Internet]. London: National Institute for Health and Care Excellence (NICE); 2011. PMID: 28481491.
8. Petrou S, Kwon J, Madan J. A practical guide to conducting a systematic review and meta-analysis of health state utility values. Pharmacoeconomics. 2018;36(9):1043–61.
9. Xie F, Zoratti M, Chan K, Husereau D, Krahn M, Levine O, et al. Toward a centralized, systematic approach to the identification, appraisal, and use of health state utility values for reimbursement decision making: introducing the Health Utility Book (HUB). Med Decis Mak. 2019;39(4):370–8.
10. Ara R, Brazier J, Peasgood T, Paisley S. The identification, review and synthesis of health state utility values from the literature. Pharmacoeconomics. 2017;35(Suppl. 1):43–55.
11. Ara R, Hill H, Lloyd A, Woods HB, Brazier J. Are current reporting standards used to describe health state utilities in cost-effectiveness models satisfactory? Value Health. 2020;23(3):397–405.
12. Ara R, Peasgood T, Mukuria C, Chevrou-Severac H, Rowen D, Azzabi-Zouraq I, et al. Sourcing and using appropriate health state utility values in economic models in health care. Pharmacoeconomics. 2017;35(Suppl. 1):7–9.
13. Sanderson S, Tatt ID, Higgins JP. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. Int J Epidemiol. 2007;36(3):666–76.
14. Yepes-Nuñez JJ, Zhang Y, Xie F, Alonso-Coello P, Selva A, Schünemann H, et al. Forty-two systematic reviews generated 23 items for assessing the risk of bias in values and preferences' studies. J Clin Epidemiol. 2017;85:21–31.
15. Büttner F, Winters M, Delahunt E, Elbers R, Lura CB, Khan KM, et al. Identifying the 'incredible'! Part 1: assessing the risk of bias in outcomes included in systematic reviews. Br J Sports Med. 2020;54(13):798–800.
16. Büttner F, Winters M, Delahunt E, Elbers R, Lura CB, Khan KM, et al. Identifying the 'incredible'! Part 2: spot the difference: a rigorous risk of bias assessment can alter the main findings of a systematic review. Br J Sports Med. 2020;54(13):801–8.
17. Katikireddi SV, Egan M, Petticrew M. How do systematic reviews incorporate risk of bias assessments into the synthesis of evidence? A methodological study. J Epidemiol Commun Health. 2015;69(2):189–95.
18. Marušić MF, Fidahić M, Cepeha CM, Farcaș LG, Tseke A, Puljak L. Methodological tools and sensitivity analysis for assessing quality or risk of bias used in systematic reviews published in the high-impact anesthesiology journals. BMC Med Res Methodol. 2020;20(1):121.
19. Walker DG, Wilson RF, Sharma R, Bridges J, Niessen L, Bass EB, et al. AHRQ methods for effective health care. Best practices for conducting economic evaluations in health care: a systematic review of quality assessment tools. Rockville: Agency for Healthcare Research and Quality (US); 2012.
20. Afshari S, Ameri H, Daroudi RA, Shiravani M, Karami H, Akbari SA. Health related quality of life in adults with asthma: a systematic review to identify the values of EQ-5D-5L instrument. J Asthma. 2022;59(6):1203–12.
21. Brennan VK, Mauskopf J, Colosia AD, Copley-Merriman C, Hass B, Palencia R. Utility estimates for patients with type 2 diabetes mellitus after experiencing a myocardial infarction or stroke: a systematic review. Expert Rev Pharmacoecon Outcomes Res. 2015;15(1):111–23.
22. Buchanan-Hughes AM, Buti M, Hanman K, Langford B, Wright M, Eddowes LA. Health state utility values measured using the EuroQol 5-dimensions questionnaire in adults with chronic

hepatitis C: a systematic literature review and meta-analysis. Qual Life Res. 2019;28(2):297–319.

23. Carrello J, Hayes A, Killedar A, Von Huben A, Baur LA, Petrou S, et al. Utility decrements associated with adult overweight and obesity in Australia: a systematic review and meta-analysis. Pharmacoeconomics. 2021;39(5):503–19.
24. Carter GC, King DT, Hess LM, Mitchell SA, Taipale KL, Kiiskinen U, et al. Health state utility values associated with advanced gastric, oesophageal, or gastro-oesophageal junction adenocarcinoma: a systematic review. J Med Econ. 2015;18(11):954–66.
25. Di Tanna GL, Urbich M, Wirtz HS, Potrata B, Heisen M, Bennison C, et al. Health state utilities of patients with heart failure: a systematic literature review. Pharmacoeconomics. 2021;39(2):211–29.
26. Gheorghe A, Moran G, Duffy H, Roberts T, Pinkney T, Calvert M. Health utility values associated with surgical site infection: a systematic review. Value Health. 2015;18(8):1126–37.
27. Han R, François C, Toumi M. Systematic review of health state utility values used in European pharmacoeconomic evaluations for chronic hepatitis C: impact on cost-effectiveness results. Appl Health Econ Health Policy. 2021;19(1):29–44.
28. Kua WS, Davis S. PRS49: systematic review of health state utilities in children with asthma. Value Health. 2016;19(7):A557.
29. Li L, Severens JLH, Mandrik O. Disutility associated with cancer screening programs: a systematic review. PLoS ONE. 2019;14(7): e0220148.
30. Paracha N, Abdulla A, MacGilchrist KS. Systematic review of health state utility values in metastatic non-small cell lung cancer with a focus on previously treated patients. Health Qual Life Outcomes. 2018;16(1):179.
31. Park HY, Cheon HB, Choi SH, Kwon JW. Health-related quality of life based on EQ-5D utility score in patients with tuberculosis: a systematic review. Front Pharmacol. 2021;12: 659675.
32. Petrou S, Krabuanrat N, Khan K. Preference-based health-related quality of life outcomes associated with preterm birth: a systematic review and meta-analysis. Pharmacoeconomics. 2020;38(4):357–73.
33. Saeed YA, Phoon A, Bielecki JM, Mitsakakis N, Bremner KE, Abrahamyan L, et al. A systematic review and meta-analysis of health utilities in patients with chronic hepatitis C. Value Health. 2020;23(1):127–37.
34. Szabo SM, Audhya IF, Malone DC, Feeny D, Gooch KL. Characterizing health state utilities associated with Duchenne muscular dystrophy: a systematic review. Qual Life Res. 2020;29(3):593–605.
35. Van Wilder L, Rammant E, Clays E, Devleesschauwer B, Pauwels N, De Smedt D. A comprehensive catalogue of EQ-5D scores in chronic disease: results of a systematic review. Qual Life Res. 2019;28(12):3153–61.
36. Ward Fuller G, Hernandez M, Pallot D, Lecky F, Stevenson M, Gabbe B. Health state preference weights for the Glasgow Outcome Scale following traumatic brain injury: a systematic review and mapping study. Value Health. 2017;20(1):141–51.
37. Yuan Y, Xiao Y, Chen X, Li J, Shen M. A systematic review and meta-analysis of health utility estimates in chronic spontaneous urticaria. Front Med (Lausanne). 2020;7: 543290.
38. Blom EF, Haaf KT, de Koning HJ. Systematic review and meta-analysis of community- and choice-based health state utility values for lung cancer. Pharmacoeconomics. 2020;38(11):1187–200.
39. Foster E, Chen Z, Ofori-Asenso R, Norman R, Carney P, O’Brien TJ, et al. Comparisons of direct and indirect utilities in adult epilepsy populations: a systematic review. Epilepsia. 2019;60(12):2466–76.
40. Haridoss M, Bagepally BS, Natarajan M. Health-related quality of life in rheumatoid arthritis: systematic review and meta-analysis of EuroQoL (EQ-5D) utility scores from Asia. Int J Rheum Dis. 2021;24(3):314–26.
41. Jiang M, Ma Y, Li M, Meng R, Ma A, Chen P. A comparison of self-reported and proxy-reported health utilities in children: a systematic review and meta-analysis. Health Qual Life Outcomes. 2021;19(1):45.
42. Landeiro F, Mughal S, Walsh K, Nye E, Morton J, Williams H, et al. Health-related quality of life in people with predementia Alzheimer’s disease, mild cognitive impairment or dementia measured with preference-based instruments: a systematic literature review. Alzheimers Res Ther. 2020;12(1):154.
43. Rebchuk AD, O’Neill ZR, Szefer EK, Hill MD, Field TS. Health utility weighting of the Modified Rankin Scale: a systematic review and meta-analysis. JAMA Netw Open. 2020;3(4): e203767.
44. Tran AD, Fogarty G, Nowak AK, Espinoza D, Rowbotham N, Stockler MR, et al. A systematic review and meta-analysis of utility estimates in melanoma. Br J Dermatol. 2018;178(2):384–93.
45. Yang Z, Li S, Wang X, Chen G. Health state utility values derived from EQ-5D in psoriatic patients: a systematic review and meta-analysis. J Dermatol Treat. 2020;9:1–8.
46. Golicki D, Jaśkowiak K, Wójcik A, Młyńczak K, Dobrowolska I, Gawrońska A, et al. EQ-5D-derived health state utility values in hematologic malignancies: a catalog of 796 utilities based on a systematic review. Value Health. 2020;23(7):953–68.
47. Paracha N, Thuresson PO, Moreno SG, MacGilchrist KS. Health state utility values in locally advanced and metastatic breast cancer by treatment line: a systematic review. Expert Rev Pharmacoecon Outcomes Res. 2016;16(5):549–59.
48. Cooper JT, Lloyd A, Sanchez JJG, Sörstadius E, Briggs A, McFarlane P. Health related quality of life utility weights for economic evaluation through different stages of chronic kidney disease: a systematic literature review. Health Qual Life Outcomes. 2020;18(1):310.
49. Aceituno D, Pennington M, Iruretagoyena B, Prina AM, McCrone P. Health state utility values in schizophrenia: a systematic review and meta-analysis. Value Health. 2020;23(9):1256–67.
50. Li YK, Alolabi N, Kaur MN, Thoma A. A systematic review of utilities in hand surgery literature. J Hand Surg Am. 2015;40(5):997–1005.
51. Magnus A, Isaranuwatchai W, Mihalopoulos C, Brown V, Carter R. A systematic review and meta-analysis of prostate cancer utility values of patients and partners between 2007 and 2016. MDM Policy Pract. 2019;4(1):2381468319852332.
52. Meregaglia M, Cairns J. A systematic literature review of health state utility values in head and neck cancer. Health Qual Life Outcomes. 2017;15(1):174.
53. Ó Céilleachair A, O'Mahony JF, O'Connor M, O'Leary J, Normand C, Martin C, et al. Health-related quality of life as measured by the EQ-5D in the prevention, screening and management of cervical disease: a systematic review. Qual Life Res. 2017;26(11):2885–97.
54. Khadka J, Kwon J, Petrou S, Lancsar E, Ratcliffe J. Mind the (inter-rater) gap. An investigation of self-reported versus proxy-reported assessments in the derivation of childhood utility values for economic evaluation: a systematic review. Soc Sci Med (1982). 2019;240:112543.
55. Kwon J, Kim SW, Ungar WJ, Tsiplova K, Madan J, Petrou S. A systematic review and meta-analysis of childhood health utilities. Med Decis Mak. 2018;38(3):277–305.
56. Hatswell AJ, Burns D, Baio G, Wadelin F. Frequentist and Bayesian meta-regression of health state utilities for multiple myeloma incorporating systematic review and analysis of individual patient data. Health Econ. 2019;28(5):653–65.
57. Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S. Assessing the quality of randomized controlled trials: an

annotated bibliography of scales and checklists. Control Clin Trials. 1995;16(1):62–73.

58. Belcher BM, Rasmussen KE, Kemshaw MR, Zornes D. Defining and assessing research quality in a transdisciplinary context. Res Eval. 2016;25:1–17.
59. Oxman AD, Guyatt GH. Validation of an index of the quality of review articles. J Clin Epidemiol. 1991;44(11):1271–8.
60. Verhagen AP, de Vet HC, de Bie RA, Kessels AG, Boers M, Bouter LM, et al. The Delphi list: a criteria list for quality assessment of randomized clinical trials for conducting systematic reviews developed by Delphi consensus. J Clin Epidemiol. 1998;51(12):1235–41.
61. Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? Control Clin Trials. 1996;17(1):1–12.
62. Whiting P, Rutjes AWS, Reitsma JB, Bossuyt PMM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol. 2003;3(1):25.
63. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529–36.
64. Centre for Reviews and Dissemination (CRD). Systematic reviews: CRD's guidance for undertaking reviews in health care. 3rd ed. York: Centre for Reviews and Dissemination, University of York Centre for Reviews and Dissemination, University of York; 2013.
65. Viswanathan M, Patnode CD, Berkman ND, Bass EB, Chang S, Hartling L, et al. AHRQ methods for effective health care assessing the risk of bias in systematic reviews of health care interventions. Methods guide for effectiveness and comparative effectiveness reviews. Rockville: Agency for Healthcare Research and Quality (US); 2008.
66. Viswanathan M, Patnode CD, Berkman ND, Bass EB, Chang S, Hartling L, et al. Recommendations for assessing the risk of bias in systematic reviews of health-care interventions. J Clin Epidemiol. 2018;97:26–34.
67. Armijo-Olivo S, Fuentes J, Ospina M, Saltaji H, Hartling L. Inconsistency in the items included in tools used in general health research and physical therapy to evaluate the methodological quality of randomized controlled trials: a descriptive analysis. BMC Med Res Methodol. 2013;17(13):116.
68. Balshem H, Helfand M, Schünemann HJ, Oxman AD, Kunz R, Brozek J, et al. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol. 2011;64(4):401–6.
69. Kim DD, Do LA, Synnott PG, Lavelle TA, Prosser LA, Wong JB, et al. Developing criteria for health economic quality evaluation tool. Value Health. 2023;26(8):1225–34.
70. Littlewood C, Ashton J, Chance-Larsen K, May S, Sturrock B. The quality of reporting might not reflect the quality of the study: implications for undertaking and appraising a systematic review. J Man Manip Ther. 2012;20(3):130–4.
71. Jüni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. JAMA. 1999;282(11):1054–60.
72. Nasa P, Jain R, Juneja D. Delphi methodology in healthcare research: how to decide its appropriateness. World J Methodol. 2021;11(4):116–29.
73. Hsu C-C, Sandford BA. The Delphi technique: making sense of consensus. Pract Assess Res Eval. 2007;12:10.
74. Jones J, Hunter D. Consensus methods for medical and health services research. BMJ (Clin Res Ed). 1995;311(7001):376–80.
75. Murphy MK, Black NA, Lamping DL, McKee CM, Sanderson CF, Askham J, et al. Consensus development methods, and their use in clinical guideline development. Health Technol Assess. 1998;2(3):i–iv, 1–88.
76. Keeney S, Hasson F, McKenna H. Consulting the oracle: ten lessons from using the Delphi technique in nursing research. J Adv Nurs. 2006;53(2):205–12.
77. Al-Jundi A, Sakka S. Critical appraisal of clinical research. J Clin Diagn Res. 2017;11(5):JE01–5.
78. Burls A. What is critical appraisal? 2014. http://www.bandolier.org.uk/painres/download/whatis/What_is_critical_appraisal.pdf. Accessed 5 Nov 2021.
79. Higgins JPT, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ (Clin Res Ed). 2011;343: d5928.
80. Eiring Ø, Landmark BF, Aas E, Salkeld G, Nylenna M, Nytrøen K. What matters to patients? A systematic review of preferences for medication-associated outcomes in mental disorders. BMJ Open. 2015;5(4): e007848.
81. Brazier J, Ara R, Azzabi I, Busschbach J, Chevrou-Séverac H, Crawford B, et al. Identification, review, and use of health state utilities in cost-effectiveness models: an ISPOR Good Practices for Outcomes Research Task Force report. Value Health. 2019;22(3):267–75.
82. Shamliyan T, Kane RL, Dickinson S. A systematic review of tools used to assess the quality of observational studies that examine incidence or prevalence and risk factors for diseases. J Clin Epidemiol. 2010;63(10):1061–70.
83. Drummond MF, Sculpher MJ, Claxton K, Stoddart GL, Torrance GW, Drummond D. Methods for the economic evaluation of health care programmes. 4th ed. Oxford: Oxford University Press; 2015.
84. Oliver A. Distinguishing between experienced utility and remembered utility. Public Health Ethics. 2016;10(2):122–8.
85. Pieterse AH, Stiggelbout AM. What are values, utilities, and preferences? A clarification in the context of decision making in health care, and an exploration of measurement issues. In: Diefenbach MA, Miller-Halegoua S, Bowen DJ, editors. Handbook of health decision science. New York: Springer New York; 2016. p. 3–13.
86. Sterne JAC, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ (Clin Res Ed). 2016;355: i4919.
87. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ (Clin Res Ed). 2019;366: l4898.
88. Morgan DL, Krueger RA. When to use focus groups and why. Successful focus groups: advancing the state of the art. Thousand Oaks: Sage Publications, Inc.; 1993: p. 3–19.
89. Massey OT. A proposed model for the analysis and interpretation of focus groups in evaluation research. Eval Program Plann. 2011;34(1):21–8.
90. Bernstein SJ, Laouri M, Hilborne LH, Leape LL, Kahan JP, Park RE, et al. Coronary angiography: a literature review and ratings of appropriateness and necessity. Santa Monica: RAND Corporation; 1992.
91. Culyer AJ. The dictionary of health economics. Cheltenham: Edward Elgar Publishing; 2005.

Authors and Affiliations

Muchandifunga Trust Muchadeyi[1,2] · **Karla Hernandez-Villafuerte**[1] · **Gian Luca Di Tanna**[3,4] · **Rachel D. Eckford**[1] · **Yan Feng**[5] · **Michela Meregaglia**[6] · **Tessa Peasgood**[7,8] · **Stavros Petrou**[9] · **Jasper Ubels**[1,2] · **Michael Schlander**[1,10,2]

✉ Michael Schlander
m.schlander@dkfz-heidelberg.de

Muchandifunga Trust Muchadeyi
m.muchadeyi@dkfz-heidelberg.de

Karla Hernandez-Villafuerte
karla.hernandez-villafuerte@wifor.com

Gian Luca Di Tanna
gianluca.ditanna@supsi.ch

Rachel D. Eckford
r.eckford@dkfz-heidelberg.de

Yan Feng
yan.feng@qmul.ac.uk

Michela Meregaglia
michela.meregaglia@unibocconi.it

Tessa Peasgood
Tessa.Peasgood@unimelb.edu.au

Stavros Petrou
stavros.petrou@phc.ox.ac.uk

Jasper Ubels
j.ubels@dkfz-heidelberg.de

[1] Division of Health Economics, German Cancer Research Centre (DKFZ) Foundation under Public Law, Im Neuenheimer Feld 280, 69120 Heidelberg, Germany

[2] Medical Faculty Mannheim, University of Heidelberg, Mannheim, Germany

[3] Department of Business Economics, Health and Social Care (DEASS), University of Applied Sciences and Arts of Southern Switzerland, Manno, Lugano, Switzerland

[4] The George Institute for Global Health, University of New South Wales (UNSW Sydney), Sydney, NSW, Australia

[5] Wolfson Institute of Population Health, Queen Mary University of London, London, UK

[6] Centre for Research on Health and Social Care Management (CERGAS), SDA Bocconi School of Management, Milan, Italy

[7] Melbourne School of Population and Global Health, University of Melbourne, Melbourne, VIC, Australia

[8] Sheffield Centre for Health and Related Research, University of Sheffield, Sheffield, UK

[9] Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, UK

[10] Alfred Weber Institute for Economics (AWI), University of Heidelberg, Heidelberg, Germany