Contents lists available at sciencedirect.com
Journal homepage: www.elsevier.com/locate/jval

Methodology

Developing Criteria for Health Economic Quality Evaluation Tool

David D. Kim, PhD, Lauren A. Do, BS, Patricia G. Synnott, MALD, MS, Tara A. Lavelle, PhD, Lisa A. Prosser, PhD, John B. Wong, MD, Peter J. Neumann, ScD

ABSTRACT

Objectives: Because existing publication guidelines and checklists have limitations when used to assess the quality of cost-effectiveness analysis, we developed a novel quality assessment tool for cost-effectiveness analyses, differentiating methods and reporting quality and incorporating the relative importance of different quality attributes.

Methods: We defined 15 quality domains from a scoping review and identified 72 methods and reporting quality attributes (36 each). After designing a best-worst scaling survey, we fielded an online survey to researchers and practitioners to estimate the relative importance of the attributes in February 2021. We analyzed the survey data using a sequential conditional logit model. The final tool included 48 quality attributes deemed most important for assessing methods and reporting quality (24 each), accompanied by a free and web-based scoring system.

Results: A total of 524 participants completed the methodology section, and 372 completed both methodology and reporting sections. Quality attributes pertaining to the "modeling" and "data inputs and evidence synthesis" domains were deemed most important for methods quality, including "structure of the model reflects the underlying condition and intervention's impact" and "model validation is conducted." Quality attributes pertaining to "modeling" and "Intervention/comparator(s)" domains were considered most important for reporting quality, including "model descriptions are detailed enough for replication." Despite its growing prominence, "equity considerations" were not deemed as important as other quality attributes.

Conclusions: The Criteria for Health Economic Quality Evaluation tool allows users to differentiate methods and reporting as well as quantifies the relative importance of quality attributes. Alongside other considerations, it could help assess and improve the quality of cost-effectiveness evidence to inform value-based decisions.

Keywords: cost-effectiveness, economic evaluations, evidence, priority setting, quality assessment.

VALUE HEALTH. 2023; 26(8):1225–1234

Introduction

In 1996, the Original Panel on Cost-Effectiveness in Health and Medicine proposed a reference case analysis—a set of standard methods guidance for conducting cost-effectiveness analyses (CEAs)—that sought to improve the quality and comparability of CEAs.[1] Since then, the number of published CEAs has grown considerably,[2,3] and CEAs have played an increasingly important role in informing clinical guidelines and coverage/reimbursement decisions.[4] Nevertheless, the proliferation of CEAs has raised questions about their quality, particularly when CEAs on the same topic reach different conclusions. For example, the choice of key data inputs and analytic judgments in CEA models evaluating biological drugs for rheumatoid arthritis were shown to have significant impacts on cost-effectiveness results.[5] The Second Panel's updated report on the future directions for CEA highlighted the need for improvements in quality scoring systems for CEAs to aid decision makers considering economic evidence.[6]

Multiple guidelines and checklists on conducting and reporting CEAs exist,[7-13] but have limitations when assessing the quality of CEAs. For example, the widely cited and recently updated Consolidated Health Economic Evaluation Reporting Standards checklist promotes transparency and consistency in reporting CEAs.[8,14] Still, it does not assess methods quality or delineate the relative importance of different reporting attributes. Although the 2003 Quality of Health Economic Studies (QHES) instrument grades CEA quality, its 16 criteria consist of dichotomous "yes" or "no" responses and primarily focuses primarily on reporting rather than methodological quality.[15] Although the Agency for Healthcare Research and Quality recently published recommendations for conducting modeling and simulation studies[16] and the Second Panel on Cost-Effectiveness in Health and Medicine published detailed guidance on CEA methodology,[7] neither provides specific criteria for quality assessment. An article highlighting the problematic use of quality assessment scale in meta-analysis supports our notion that "many scales include items that are more closely related to reporting quality, ethical issues, or to the interpretation of results rather than to the interval validity of trials (p1058)."[17] Indeed, many existing checklists and tools focus on reporting quality or mixed reporting and method quality. None explicitly differentiates methods from reporting quality and incorporates the relative importance of different quality attributes. Our study aims to fill this gap by developing a quality scoring system—the Criteria for Health Economic Quality Evaluation (CHEQUE) tool—to assess the quality of CEAs, separately for methods and reporting quality, accounting for the relative importance of quality attributes.

Methods

Scoping Review: Determining Initial CEA Quality Domains and Attributes

According to the ISPOR good research practices for conjoint analysis,[18] we conducted a scoping review of textbooks and practice guidelines to define evidence-based quality domains and specific quality attributes based on textbooks and practice guidelines,[7,10,16,19-21] existing checklists and tools,[8,9,12,13,15] and international health technology assessments guidelines.[22-24] For each guideline, an independent reviewer extracted recommended quality domains and relevant attributes, separately for methods and reporting assessment. After the initial data extraction was completed, 3 researchers (D.D.K., L.A.D., and P.G.S.) had a consensus meeting to define 15 broad CEA quality domains (Tables 1 and 2): decision problem and scope; intervention and comparator(s); perspective; population; outcome measures; time horizon; discounting; modeling; data inputs and evidence synthesis; consequences; utilities (preference measures); costs and resource use; analysis; equity considerations; and transparency and reporting (note: "transparency and reporting" domain is only applicable to reporting quality). Within these 15 domains, we identified 72 quality attributes based on relevance to the decision context (ie, determining the quality of CEAs) and whether attributes are related to one another. Half of these attributes (n = 36) assess methods quality, and half (n = 36) assess reporting quality. Appendix Supplement Section A in Supplemental Materials found at https://doi.org/10.1016/j.jval.2023.04.004 provides the full summary of our scoping review, whereas Appendix Tables B1 and B2 in Appendix Supplement Section B in Supplemental Materials found at https://doi.org/10.1016/j.jval.2023.04.004 provide the list of these attributes identified from our initial review.

Developing the Best-Worst Scaling Survey

Based on quality domains and relevant attributes extracted, we designed an object case best-worst scaling (BWS) survey—separately for methods and reporting quality assessment—to estimate the relative importance of these attributes.[25] In BWS, respondents are presented with a choice set of 3 or more attributes and asked to select one attribute as "most important" and another attribute as "least important."[26-28] Consistent with adaptation-level theory,[29] BWS assumes that individuals make reliable and valid choices of the 2 most extreme objects/items in a set. The BWS approach has been increasingly used in health services and clinical research to measure various preferences in healthcare settings, such as patients' treatment goals or public preferences for healthcare reform.[30-34]

To determine which quality attributes to include in each choice set, we used a balanced incomplete block design (BIBD), a common type of block design. In the BIBD, every quality attribute (ie, element) appears in the same number of blocks; every block contains the same number of elements; and every pair of elements appears together in the same number of blocks. At the same time, the word "incomplete" indicates that the size of each block is smaller than the number of elements. Our BIBD included a set of design parameters ($v = 36$, $b = 42$, $r = 7$, $k = 6$, $\lambda = 1$) with a block design efficiency of 99.77 involving 42 blocks ($b = 42$), where each block contained 6 elements ($k = 6$) selected from a total of 36 elements ($v = 36$). Each element appeared 7 times ($r = 7$) across the entire design set (see Appendix Table B3 in Appendix Supplement Section B in Supplemental Materials found at https://doi.org/10.1016/j.jval.2023.04.004 for further details). We divided the 42 choice sets into 7 groups and randomized participants to complete 1 group (= 6 choice sets) each from methods and reporting quality assessment to minimize participants' cognitive and response burden.

Survey Design, Administration, and Data Collection

The survey consisted of 4 sections: (1) introduction and informed consent, (2) methods quality assessment (6 choice sets), (3) demographic information (4 questions), and (4) reporting quality assessment (6 choice sets). Respondents selected the most important and least important quality attributes for methods and reporting sections among 6 attributes presented in each choice set with an option to provide any written comments. For demographic information, we asked about their level of CEA knowledge, experiences in health economics and outcomes research, sectors they work in, and their confidence in their responses. Sample questions from the methods quality assessment survey are presented in Figure 1 (see Appendix Fig. B5 in Appendix Supplement Section B in Supplemental Materials found at https://doi.org/10.1016/j.jval.2023.04.004 for additional examples).

We used the Qualtrics survey platform to design and administer an online survey to researchers and practitioners in health economics and outcomes research.[35] Outreach was conducted through personal contacts and mailing lists of the Center for the Evaluation of Value and Risk at Tufts Medical Center, 2500+ ISPOR members in selected Special Interest Groups and Council Groups, and 900+ Society for Medical Decision Making members. We fielded the online survey between February 9, 2021, and March 23, 2021, and provided a $100 Amazon gift card to 15 randomly selected respondents to encourage participation. The institutional review board at Tufts Medical Center/Tufts University Health Sciences determined that this project was exempt in accordance with 45 Code of Federal Regulations 46.104(d), given that this study did not involve the use or collection of protected health information.

Statistical Analyses

We analyzed the object case BWS survey data for methods and reporting assessment separately. First, we used a max-diff model to estimate standardized BWS scores, a unidimensional interval-level scale ranging from -1 (least important) to $+1$ (most important).[28] Within a choice set, we estimated the number of times the attribute was selected as "most important" minus the number of times the attribute was selected as "least important." Then, this number was divided by the number of times the attributes appeared in the survey to estimate BWS scores.

Second, we applied a sequential conditional logit model—which assumes that respondents provide the best choice, followed by the worst choice from remaining attributes—to quantify relative preferences for different quality attributes.[36] We accounted for the correlation among multiple responses from an individual

Table 1. CHEQUE tool: methods quality (final 24-item version).

Domain	Attribute	Rounded importance score	Scoring weight assessment* (yes, somewhat, no, or N/A)	Final score
Decision problem and scope	M1. The analysis answers an important question for decision making.	5		
	M2. The study objective (decision problem) is measurable.	6		
Intervention and comparator(s)	M3. The comparator(s) is/are the best possible option that appropriately measures the opportunity cost of using the new treatment.	4		
Perspective	M4. The analytic perspective(s) is/are appropriate to answer the research question posed.	4		
Population	M5. The scope of the study encompasses all populations affected by the intervention.	1		
Outcome measures	M6. Health outcomes are measured in health metrics that aggregate survival and health-related quality of life or disability (eg, QALY or DALY).	3		
Time horizon	M7. The analytic time horizon is sufficiently long enough to reflect all important differences between intervention(s) and comparator(s).	4		
Discounting	M8. Costs and health effects that occur in the future are discounted to their present value using a recommended discount rate.	2		
Modeling	M9. The chosen model type is appropriate to address study questions.	6		
	M10. The structure of the model reflects the underlying health condition and the impact of the interventions.	8		
	M11. Modeling assumptions are reasonable, given the underlying data.	5		
	M12. The need for extrapolation or integrating multiple data sources is considered.	3		
	M13. Model validation is conducted, including an assessment of the model structure, assumptions, data, and results.	8		
Data inputs and evidence synthesis	M14. A "best available evidence" approach is used to select data sources for model parameters (eg, conducted or references systematic reviews/meta-analyses).	6		
	M15. Data inputs are generated by appropriate statistical and epidemiological techniques.	5		
	M16. The quality of the data, including sources of bias, is assessed appropriately.	6		
Consequences	M17. Major consequences affected by the choice of interventions being compared are identified.	5		
Utilities (preference measures)	M18. Health preferences reflect those of the jurisdiction(s) of interest (as specified in the decision problem).	2		
Costs and resource use	M19. Resource use that is nontrivial in magnitude is included in the reference case analysis.	2		
Analysis	M20. Incremental analyses are conducted (ie, the additional costs generated by one alternative over another are compared with the additional effects generated).	4		
	M21. ICERs are obtained by comparing each intervention with the next most effective option after eliminating dominated options.	3		
	M22. Probabilistic sensitivity analysis is conducted to account for uncertainty in input parameters simultaneously.	3		

continued on next page

Table 1. Continued

Domain	Attribute	Rounded importance score	Scoring weight assessment* (yes, somewhat, no, or N/A)	Final score
	M23. Alternative modeling choices and assumptions (structural uncertainty) are explored through additional sensitivity analysis (ie, scenario analysis).	4		
Equity considerations	M24. Relevant equity or distributional considerations are taken into account.	1		
Total		100		

CHEQUE indicates Criteria for Health Economic Quality Evaluation; DALY, disability-adjusted life-year; ICER, incremental cost-effectiveness ratio; N/A, not applicable; QALY, quality-adjusted life-year.
*Scoring weight: "yes" = 1.0; "somewhat" = 0.5; "no" = 0.0. For "N/A," you will have the option to give full credit (weight=1.0) or exclude it from the total.

survey respondent. The dependent variable indicates whether a specific attribute was chosen as best or worst. The independent variables are effect coding variables indicating each attribute compared with the reference attribute. With effects coding, the model coefficient of each attribute represents its relative importance compared with the mean of all attributes. Attributes more or less important than the mean are expressed with positive or negative values. Using a probability-based rescaling procedure,[37] we transformed the original model coefficients into attribute-specific importance scores so that the sum of all attribute-specific importance scores equals 100. Statistical analyses were conducted using Stata 16 (StataCorp).[38]

Developing the CHEQUE Tool

To develop a practical quality assessment tool, we created a scoring system that reflects the relative importance of each quality attribute based on survey participants' responses. We applied 2 criteria to refine the set of quality attributes deemed most important to evaluate CEA quality. First, we limited attributes to those that exceeded the median importance score (2.7 for methodology and 2.6 for reporting) for each assessment. When examining alternative thresholds in sensitivity analyses, a more generous threshold (ie, lowering the threshold) added 2 to 3 items that would increase user burden. The second criterion was "rescue," where we rescued the highest-scoring attribute for domains in which none of their attributes met the first criterion. For example, we rescued "the scope of the study encompasses all populations affected by the intervention" from the "population" domain and "relevant equity or distributional considerations are taken into account" from the "equity consideration" domain. This approach helped us retain all of the quality domains of CEAs and their most important quality attributes.

The Free Online Tool and Final Scoring

For accessibility and usability, we created a free web-based, user-friendly CHEQUE tool (https://uchicago.co1.qualtrics.com/jfe/form/SV_ef9Shen9Ds0JJlk), which let users select their level of assessment for each quality attribute: yes, somewhat, no, or not applicable. Each choice of the level corresponds to a different weight that, cumulatively, sums up to the final score: yes = full credit (ie, the assigned importance score is multiplied by 1.0), somewhat = half credit (ie, the assigned importance score is multiplied by 0.5), no = no credit. Users could also choose 2 scoring options for handling "not applicable" attributes: Option 1: exclude "not applicable" (ie, quality attributes deemed "not applicable" will be excluded from the final score, and the total number of possible points will vary) or Option 2: assign "full credit" to "not applicable" (ie, there will be 100 total possible points for each methods and reporting quality score). After completing the assessment, the online CHEQUE tool automatically calculated 3 final scores: methods quality, reporting quality, and overall quality (the sum of methods and reporting scores). Users could download their assessment results in a PDF format.

Case Study and Inter-Rater Reliability Assessment

To explore the application of the CHEQUE scoring tool, 2 authors (D.D.K. and L.A.D.) independently evaluated the methods and reporting quality of 10 randomly selected CEAs published in 2020 and available in the Tufts Medical Center CEA Registry.[39] We assessed the inter-rater reliability of the CHEQUE scoring tool using kappa statistics to measure our level of agreement.[40]

Results

Survey Results

During the 6-week survey period, 616 respondents participated, 524 (85%) completed the methods quality assessment (the first section of the survey), and 372 (60%) completed both methodology and reporting sections. Among 446 respondents who reported demographic information, 283 were academic/university researchers (63%), 72 individuals worked at consulting/contract research firms (16%), and 54 worked in the biotechnology/pharmaceutical/medical technology industry (12%). More than three-quarters of the participants (n = 345, 77%) expressed a very good understanding of CEA, and approximately 90% indicated they were very confident or somewhat confident in their responses. The characteristics of survey participants are presented in Table 3.

Max-Diff Analysis: Descriptive Analysis of the Initial Data

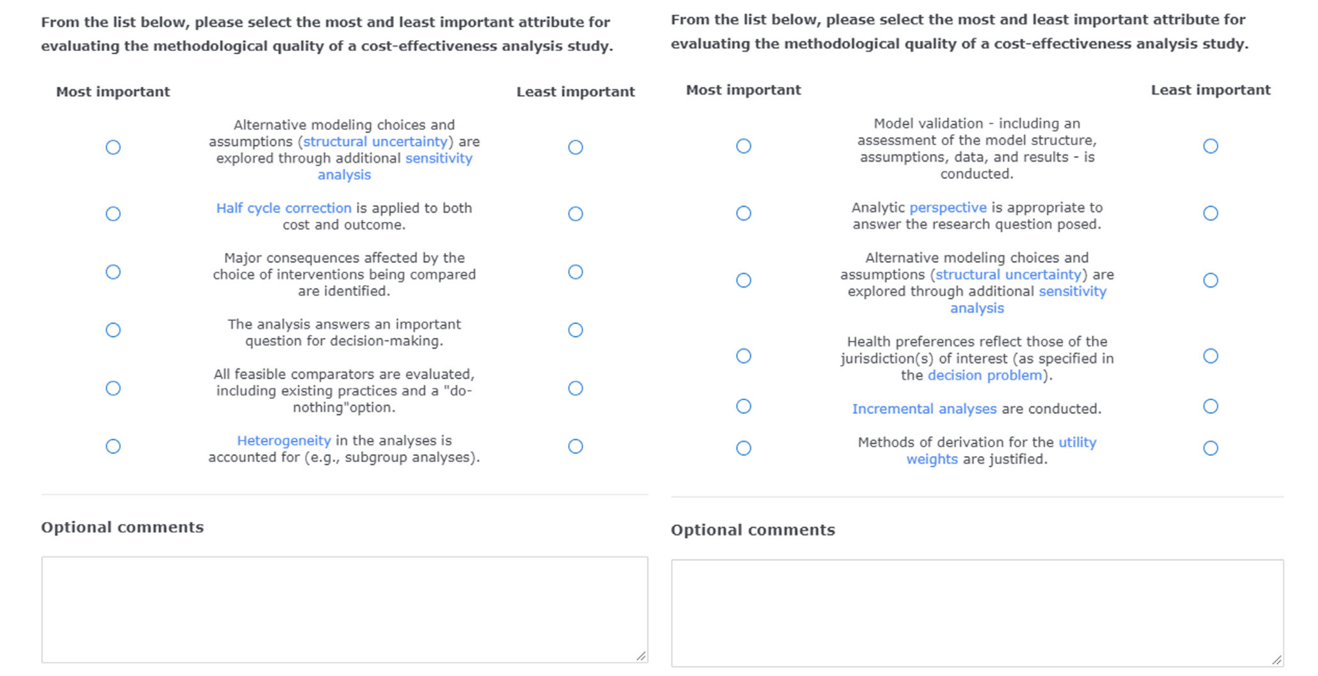
Among methods quality attributes, respondents selected modeling-related attributes as "most important" (ie, 4 of 5 modeling-related attributes were ranked in the top 7), including "structure of the model reflects the underlying health condition and the impact of the intervention" (BWS score: +0.067) and "model validation, including an assessment of the model structure, assumptions, data, and results, is conducted" (BWS score: +0.066). In contrast, the least important attributes were "half-cycle correction is applied to both cost and outcome" (BWS

Table 2. CHEQUE tool: reporting quality (final 24-item version).

Domain	Attribute	Rounded importance score	Scoring weight assessment* (yes, somewhat, no, or N/A)	Final score
Decision problem and scope	R1. The study objectives (or decision problems) are clearly stated.	7		
Intervention and comparator(s)	R2. All aspects of the interventions that may affect their cost-effectiveness are clearly defined (eg, frequency of delivery, setting of delivery, and specific technologies used).	7		
	R3. The comparator(s) is/are clearly stated.	6		
Perspective	R4. The analytic perspective(s) is/are clearly stated.	4		
Population	R5. The target population is clearly stated.	4		
Outcome measures	R6. Primary outcome measures are clearly stated.	5		
	R7. ICERs are reported.	4		
Time Horizon	R8. The analytic time horizon is clearly stated.	3		
Discounting	R9. The discount rate is clearly stated.	2		
Modeling	R10. The type of model used is clearly stated.	4		
	R11. Justification of modeling choices and assumptions is provided.	8		
	R12. Model descriptions are detailed enough to allow for replication.	9		
	R13. The description of how the model was validated is provided.	3		
	R14. The software used to develop the model is clearly stated.	5		
Data inputs and evidence synthesis	R15. All data sources are clearly referenced.	5		
Consequences	R16. Comprehensive identification of potential consequences is summarized (eg, using an Impact Inventory table in Second Panel's report).	2		
Utilities (preference measures)	R17. Sources for the utility weights are clearly stated.	3		
Costs and resource use	R18. Quantities of resources are reported separately from the prices (unit costs) of those resources.	2		
Analysis	R19. The approach to secondary analyses (eg, sensitivity, scenario, or subgroup analysis) is sufficiently described.	3		
Equity considerations	R20. Discussion section includes a description of any significant ethical implications of the CEA results.	1		
Transparency and reporting	R21. Results are presented in a disaggregated format for transparency.	3		
	R22. The relevance of study results to specific decision problems is discussed.	3		
	R23. Implications of uncertainty for decision making, including the need for future research, are explored.	3		
	R24. Potential bias and limitations are discussed.	4		
Total		100		

CEA indicates cost-effectiveness analysis; CHEQUE, Criteria for Health Economic Quality Evaluation; ICER, incremental cost-effectiveness ratio; N/A, not applicable.
*Scoring weight: "yes" = 1.0; "somewhat" = 0.5; "no" = 0.0. For "N/A," you will have the option to give full credit (weight = 1.0) or exclude it from the total.

Figure 1. Sample questions from the online survey administered: methods quality assessment.



score: −0.094) and "CEAs are conducted in constant dollars that remove general price inflation" (BWS score: −0.065).

Among reporting quality attributes, modeling-related attributes were also deemed "most important (ie, 4 of 5 modeling-related attributes were ranked in the top 14)," such as "model descriptions are detailed enough to allow for replication" (BWS score: +0.078) and "justification of modeling choices and assumptions are provided" (BWS score: +0.067). In contrast, the least important attributes were "software used to develop the model is clearly stated" (BWS score: −0.092) and "source code for the simulation model(s) used in the analysis is available" (BWS score: −0.053). See Appendix Figs. B7 and B8 in Appendix Supplement Section B in Supplemental Materials found at https://doi.org/10.1016/j.jval.2023.04.004 summarizes the top 5 (most important) and the bottom 5 (least important) quality attributes for methods and reporting assessment (See Appendix Figs. B6 and B7 in Appendix Supplement Section B in Supplemental Materials found at https://doi.org/10.1016/j.jval.2023.04.004 for additional results).

CHEQUE Tool

Of the 36 initial quality attributes for each methods or reporting assessment, 24 met the inclusion criteria for the final quality scoring tool. After re-estimating the importance score with the final 24 items from the conditional logit model, the final attribute-specific importance scores ranged from 1 to 8 for methods attributes (mean 4.21, SD 1.96) and 1 to 9 for reporting attributes (mean 4.17, SD 1.95). A complete list of the CHEQUE tool criteria and attribute-specific importance scores is presented in Tables 1 (methods quality assessment) and 2 (reporting quality assessment).

Case Study

After the independent assessment, the 2 reviewers' level of agreement was "substantial" (Cohen's kappa statistic = 0.638, percent agreement = 83.8%) for methods quality and "near-perfect" (kappa statistic = 0.868, percent agreement = 94.2%) for reporting quality (Appendix Table B9 in Appendix Supplement Section B in Supplemental Materials found at https://doi.org/10.1016/j.jval.2023.04.004). The 3 lowest-scored articles in methods quality assessment (score < 80) exhibited the greatest disagreement.[41-43] Across 10 randomly selected CEAs published in 2020,[41-50] the mean scores between 2 reviewers were 86.6 (range 74.5-97.5) for methods quality and 90.9 (range 83.5-97.5) for reporting quality (Appendix Table B10 in Appendix Supplement Section B in Supplemental Materials found at https://doi.org/10.1016/j.jval.2023.04.004). We observed a significant positive correlation between methods and reporting quality (Pearson's correlation coefficient 0.837, P = .0026) (Fig. 2).

Discussion

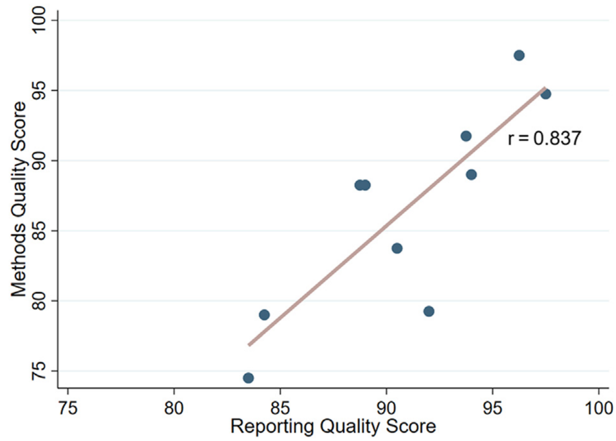
When creating a new instrument, such as the CHEQUE tool, that seems to duplicate the function of other existing tools (eg, the Consolidated Health Economic Evaluation Reporting Standards or the QHES), many often wonder why the tool was developed and what practical purpose it may serve over and above existing tools. Building upon existing checklists and practice guidelines for economic evaluations, we developed a CHEQUE tool that offers 3 major enhancements. First, it can help differentiate methods and reporting quality. Second, it quantifies the relative importance of different quality attributes, estimated from a BWS survey of > 500 participants. Unlike traditional rating scale approaches (eg, the 9-point Likert scale), which assumes all respondents use the same numerical scale and do not trade-off one attribute against another, BWS provides more statistical information about a respondent's preference.[51] Third, quality scoring systems incorporating both methods and reporting attributes can help decision makers, researchers, and practitioners assess the relative quality of different CEAs. The accompanying free, web-based, and user-friendly

Table 3. Characteristics of survey participants.

Characteristics	n (%)
Participated	616 (100)
Completed the method section only	524 (85)
Completed both method and reporting sections	372 (60)
Did not complete at least one section	88 (14)
Declined to take the survey	4 (0.6)
Completed demographic information	446 (100)
Rate your understanding of CEA	
Very good understanding	345 (77)
Limited understanding	80 (18)
No understanding	3 (1)
Prefer not to answer	18 (4)
How long have you been working in the HEOR field?	
30+ years	22 (5)
21-30 years	75 (17)
11-20 years	117 (26)
5-10 years	104 (23)
< 5 years	112 (25)
No experience	7 (2)
Prefer not to answer	9 (2)
Which of the following best describes the sector you work in? *	
University/academic	283 (63)
Consulting/contract research	72 (16)
Biotechnology/pharmaceutical/medical technology	54 (12)
Government/nonprofit organization	41 (9)
Hospital/medical care delivery	39 (9)
No experience	17 (4)
Prefer not to answer	3(1)
How confident do you feel about your responses to the survey questions?	
Very confident	131 (29)
Somewhat confident	261 (59)
Not confident	37 (8)
Prefer not to answer	17 (4)

CEA indicates cost-effectiveness analysis; HEOR, health economics and outcomes research.
*Survey participants could select multiple categories, so the sum of all categories was > 100%.

Figure 2. Correlation between methods and reporting quality scores among 10 CEAs.



CEA indicates cost-effectiveness analysis.

CHEQUE tool (https://uchicago.co1.qualtrics.com/jfe/form/SV_ef9Shen9Ds0JJlk) facilitates accessibility and usability.

Our case study evaluating the quality of 10 CEAs showed substantial inter-rater reliability, especially agreement for assessing reporting quality. Overall, lower-quality studies were more likely to result in greater disagreement because of some methods quality attributes that require subjective judgment, such as "M17: major consequences affected by choice of interventions being compared are identified." We also found that, on average, reporting quality was higher than methods quality. For example, a CEA evaluating the cost-effectiveness of annual lung cancer screening in Italy clearly stated their use of a 5-year analytic time horizon (ie, Yes, for the time horizon reporting quality criterion),[41] but a 5-year time horizon would be insufficient to capture all of the important differences in consequences among screened versus unscreened individuals (ie, No, for the time horizon methods quality criterion).[52]

Our survey elicited the relative importance of different attributes. Respondents rated quality attributes in "modeling" and "data inputs and evidence synthesis" as most important for assessing methods quality. Similarly, "modeling"-related attributes were deemed most important for reporting quality, for example, whether model descriptions are detailed enough to allow for replication and clear justification of modeling choices and assumptions. These findings highlight the growing use of advanced evidence synthesis (eg, network meta-analysis) and complex modeling (eg, microsimulation and agent-based modeling) in economic evaluation and their essential role in ensuring high-quality evidence and modeling methodology underlying CEA. Nevertheless, as we found in our case study, some modeling-related attributes—for example, M10: structure of the model reflects the underlying health condition and the impact of interventions; M11: modeling assumptions are reasonable given the underlying data requires technical challenges; R12: model descriptions are detailed enough to allow for replication—should ideally be assessed by reviewers with technical backgrounds because they rely on subjective judgment, combined with subject matter expertise, which could lead to greater disagreement among reviewers. When resolving disagreement around these attributes, we recommend having an independent adjudicator with more technical modeling expertise in addition to subject matter reviewers. Although we did not provide explicit recommendations evaluating methods attributes, we recommend that users of this tool refer to best practice guidelines in modeling approaches specific to their assessment.[16,53,54]

Despite growing interest, quality attributes for "equity consideration" were not deemed as important as other attributes in both methods and reporting quality assessments. Methods advancement in the field, such as the development of extended CEA and distributional CEA,[55,56] allows analysts to incorporate financial risk protection and health equity into conventional economic evaluation. A recent review found some growth but limited cost-effectiveness evidence on equity effects.[57] Still, unresolved conceptual debates on which types of individuals should be prioritized and the lack of empirical evidence on how much weight to give to particular subgroups remain a key challenge in addressing ethical and distributive issues in CEA.[58-60] Thus, despite its societal importance, respondents might judge "equity consideration" as outside the scope for quality assessment of cost-effectiveness evidence and should instead be part of deliberative processes in health technology assessments for resource allocation at the population level.[61]

A few notable original quality attributes were excluded from the final scoring tool because of relatively low importance scores. For example, "any potential conflicts of interest relevant to the analysis, including funding sources, are disclosed." Studies have shown that industry-funded CEAs are more likely to report favorable cost-effectiveness ratios,[62,63] and sponsorship and conflicts of interest have been one of the commonly cited factors related to potential bias in cost-effectiveness evidence.[64] Nevertheless, it begs the question of whether such evidence could tell us about the "quality" of the cost-effectiveness evidence. The existing studies did not attempt to ask why the study sponsorship indicates low study quality or whether it is an "important" quality attribute. Sponsorship biases may be discerned through evaluating assumptions or data inputs deemed not justifiable or reasonable. These quality attributes could be captured in the "modeling" or "data inputs and evidence synthesis" domains. Thus, a study could be deemed low-quality not because a specific interest party is the study sponsor, but because that such study is more likely to use unreasonable assumptions to generate favorable ratios, which should be captured by other quality attributes included in the CHEQUE tool.

Other omitted attributes include "written protocol for the design and conduct of the CEA are developed and included" and "source code for the simulation model(s) used in the analysis is available." These attributes have recently gained more attention as the need for transparent reporting and conduct of CEAs to ensure reproducibility continues to grow.[6,65,66] It is important to note that the BWS approach is designed to measure the relative importance of quality attributes, not absolute importance. Although excluded attributes were not deemed as important as those included in the current version of the CHEQUE tool, we believe they may still be important quality attributes that should require the user's attention. For those wishing to use the complete list of the original attributes (36 items each, instead of 24 items for methods and reporting quality), Appendix Tables B1 and B2 in Appendix Supplement Section B in Supplemental Materials found at https://doi.org/10.1016/j.jval.2023.04.004 provide the expanded version of the final CHEQUE tool with 36-attribute-specific importance scores that sum up to 100.

The CHEQUE tool has some limitations. Above all, a quality scoring system is just one way to evaluate the potential usefulness of the study findings. CEAs with low quality scores may still provide valuable insights, so the CHEQUE tool users should carefully judge each study using their own context-specific set of criteria in evaluating the quality and relevance of the CEA to their decision. We acknowledge that the hazards of scoring the quality of clinical trials for meta-analysis have been previously described.[17] Nevertheless, unlike the meta-analyses—which primarily aim to summarize well-defined (and hence, often less susceptible to bias) clinical outcomes from randomized controlled trials or observational studies, CEA involves a range of choices and layers of uncertainty to produce the value measures (ie, incremental cost-effectiveness ratios) through the use of multiple data sources, modeling, and assumptions. Thus, a well-developed quality assessment tool might be more needed than the case for the meta-analysis. The CHEQUE tool should complement decision making and not necessarily be the primary decision criterion or be a substitute for a deliberative process when making resource allocation decisions.

Similarly, we do not define a threshold separating high- from low-quality cost-effectiveness evidence. Although a threshold-based "absolute" ranking can be used to differentiate high- versus low-quality CEAs, the CHEQUE tool is designed to provide a "relative ranking" of CEA studies. Two primary reasons for not providing an absolute threshold include (1) because it requires comprehensive quality assessment using the CHEQUE tool to estimate "mean" quality scores among published CEAs and such efforts are deemed outside of the scope of this study focusing on "development" of the tool (note: we provided the average quality score among the 10 randomly selected CEAs for the case study, but this should not be considered an absolute threshold" because of the small sample size) and (2), more importantly, a single "threshold" for quality assessment might not be as helpful given that the CHEQUE tool users are likely focused on certain disease areas or particular interventions or specific regions/countries. Although a quality threshold may emerge over time, users should currently judge their confidence in the findings of a CEA without an arbitrary threshold.

Depending on the topic and decision-maker perspective, the importance of the different attributes of a well-conducted and well-reported CEA may vary. For example, extrapolation might not be necessary for evaluating short-term outcomes, such as antibiotics for strep throat. In promoting users' judgment in quality assessment, our CHEQUE tool allows users to choose alternative scoring methods for handling "not applicable" attributes. In addition, methods quality scores often rely on good reporting. It may not be feasible to assess specific methods and approaches if they are not well described in the CEAs, including the technical appendix. Finally, we acknowledge that selecting the final quality attributes in the CHEQUE tool reflects the survey respondent's current understanding of the field. As the field evolves, the perceived importance of different attributes may change. Consequently, the CHEQUE tool may need to update the criteria and the scoring mechanism to reflect the state-of-the-art recommendations. Other future directions include assessing the reproducibility and comparability of the results with other checklists/tools and testing for reliability and cross-validity for CEA assessment.

Conclusions

The CHEQUE tool's quality scoring system differentiates the methods and reporting quality as well as quantifies the relative importance of various quality attributes in published CEAs. In addition, the CHEQUE tool can help identify sources of variation across studies and interpret individual study results. Alongside other considerations, a systematic quality assessment of cost-effectiveness evidence would be valuable for researchers, practitioners, payers, and policy makers to determine the best cost-effectiveness evidence to inform value-based decisions in their settings.

Supplemental Material

Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/j.jval.2023.04.004.

Article and Author Information

Accepted for Publication: April 4, 2023

Published Online: May 5, 2023

doi: https://doi.org/10.1016/j.jval.2023.04.004

Author Affiliations: Section of Hospital Medicine, Department of Medicine, The University of Chicago, Chicago, IL, USA (Kim); Center for the Evaluation of Value and Risk in Health (CEVR), Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, Boston, MA, USA (Do, Synnott, Lavelle, Neumann); Department of Medicine, Tufts University School of Medicine, Boston, MA, USA (Lavelle, Wong, Neumann); The Susan B. Meister Child Health Evaluation and Research

(CHEAR) Center, Michigan Medicine, Ann Arbor, MI, USA (Prosser); Department of Pediatrics, University of Michigan Medical School, Ann Arbor, MI, USA (Prosser); Department of Health Management and Policy, University of Michigan School of Public Health, Ann Arbor, MI, USA (Prosser); Division of Clinical Decision Making, Tufts Medical Center, Boston, MA, USA (Wong).

Correspondence: David D. Kim, PhD, Biological Sciences Division and the College, The University of Chicago, 5841 S Maryland Ave, MC 5000, Chicago, IL 60637, USA. Email: ddk@uchicago.edu

Author Contributions: *Concept and design*: Kim, Do, Synnott, Lavelle, Prosser, Neumann
Acquisition of data: Kim, Do, Synnott
Analysis and interpretation of data: Kim, Do, Synnott, Lavelle, Prosser, Wong
Drafting of the article: Kim, Wong
Critical revision of the article for important intellectual content: Kim, Synnott, Lavelle, Prosser, Wong, Neumann
Statistical analysis: Kim
Provision of study materials or patients: Do
Obtaining funding: Kim
Administrative, technical, or logistic support: Do, Synnott, Neumann
Supervision: Kim

Conflict of Interest Disclosures: All authors reported receiving grants from PhRMA Foundation (2020 Value Assessment Research Award), during the conduct of the study. Dr Kim reported receiving research grants from National Institutes of Health/National Heart, Lung, and Blood Institute and Arnold Ventures and personal fees from the American College of Physicians outside the submitted work. Dr Lavelle reported receiving grants from Genentech and personal fees from the Institute for Clinical and Economic Review outside the submitted work. Dr Prosser is an editor for *Value in Health* and had no role in the peer-review process of this article. Dr Wong reported receiving personal fees from the American College of Physicians, UpToDate, and Cambridge Press outside the submitted work. Dr Neumann reported being on the Panel of Health Advisors for the Congressional Budget Office and receiving funds from the Cost-Effectiveness Analysis Registry sponsors (the Cost-Effectiveness Analysis Registry has been funded by the National Science Foundation, National Library of Medicine, Agency for Healthcare Research and Quality, the Centers for Disease Control and Prevention, and a variety of pharmaceutical and device companies who subscribe to the data). He also reported receiving grants from the National Institutes of Health, Arnold Ventures, No Patient Left Behind, Bristol-Myers of Squibb, the National Pharmaceutical Council, and the Alzheimer's Association; being on advisory boards for Bayer, Merck, Novartis Gene Therapy, Otsuka, ArgenX, and Panalgo; and consulting with Sarepta, G1 Therapeutics, and the Analysis Group. No other disclosures were reported.

Funding/Support: This work was supported by the 2020 Value Assessment Research Award from the sponsor id="gs19" sponsor-id="https://doi.org/10.13039/100001797">PhRMA Foundation.

Role of the Funder/Sponsor: The funder had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the article; and the decision to submit the article for publication.

Acknowledgment: The authors are grateful for the editor and reviewers' helpful comments on an earlier draft and the support from ISPOR and Society for Medical Decision Making in distributing our survey to their members.

REFERENCES

1. Gold MR, Siegel JR, Russell LB, et al., eds. *Cost-Effectiveness in Health and Medicine*. New York, NY: Oxford University Press; 1996.
2. Neumann PJ, Thorat T, Shi J, Saret CJ, Cohen JT. The changing face of the cost-utility literature, 1990-2012. *Value Health*. 2015;18(2):271–277.
3. Kim DD, Silver MC, Kunst N, Cohen JT, Ollendorf DA, Neumann PJ. Perspective and costing in cost-effectiveness analysis, 1974-2018. *Pharmacoeconomics*. 2020;38(10):1135–1145 [published correction appears in *Pharmacoeconomics*. 2020;38(12):1377].
4. Kim DD, Basu A. How does cost-effectiveness analysis inform health care decisions? *AMA J Ethics*. 2021;23(8):E639–E647.
5. Drummond MF, Barbieri M, Wong JB. Analytic choices in economic models of treatments for rheumatoid arthritis: what makes a difference? *Med Decis Making*. 2005;25(5):520–533.
6. Neumann PJ, Kim DD, Trikalinos TA, et al. Future directions for cost-effectiveness analyses in health and medicine. *Med Decis Making*. 2018;38(7):767–777.
7. Sanders GD, Neumann PJ, Basu A, et al. Recommendations for conduct, methodological practices, and reporting of cost-effectiveness analyses: second panel on cost-effectiveness in health and medicine. *JAMA*. 2016;316(10):1093–1103 [published correction appears in *JAMA*. 2016;316(18):1924].
8. Husereau D, Drummond M, Petrou S, et al. Consolidated Health Economic Evaluation Reporting Standards (CHEERS) statement. *BMJ*. 2013;346:f1049.
9. Drummond MF, Jefferson TO. Guidelines for authors and peer reviewers of economic submissions to the BMJ. The BMJ economic evaluation working party. *BMJ*. 1996;313(7052):275–283.
10. Drummond MF, Sculpher MJ, Claxton K, et al. *Methods for the Economic Evaluation of Health Care Programmes*. Oxford, United Kingdom: Oxford university press; 2015.
11. Wilkinson T, Sculpher MJ, Claxton K, et al. The international decision support initiative reference case for economic evaluation: an aid to thought. *Value Health*. 2016;19(8):921–928.
12. Williams A. The cost-benefit approach. *Br Med Bull*. 1974;30(3):252–256.
13. Philips Z, Bojke L, Sculpher M, Claxton K, Golder S. Good practice guidelines for decision-analytic modelling in health technology assessment: a review and consolidation of quality assessment. *Pharmacoeconomics*. 2006;24(4):355–371.
14. Husereau D, Drummond M, Augustovski F, et al. Consolidated Health Economic Evaluation Reporting Standards 2022 (CHEERS 2022) statement: updated reporting guidance for health economic evaluations. *Value Health*. 2022;25(1):3–9.
15. Chiou CF, Hay JW, Wallace JF, et al. Development and validation of a grading system for the quality of cost-effectiveness studies. *Med Care*. 2003;41(1):32–44 [published correction appears in *Med Care*. 2003;41(3):446].
16. Dahabreh IJ, Trikalinos TA, Balk EM, Wong JB. Recommendations for the conduct and reporting of modeling and simulation studies in health technology assessment. *Ann Intern Med*. 2016;165(8):575–581.
17. Jüni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA*. 1999;282(11):1054–1060.
18. Bridges JF, Hauber AB, Marshall D, et al. Conjoint analysis applications in health–a checklist: a report of the ISPOR Good Research Practices for Conjoint Analysis Task Force. *Value Health*. 2011;14(4):403–413.
19. Neumann PJ, Sanders GD, Russell LB, et al. *Cost-Effectiveness in Health and Medicine*. Oxford, United Kingdom: Oxford University Press; 2016.
20. Garrison Jr LP, Neumann PJ, Willke RJ, et al. A health economics approach to US value assessment frameworks-summary and recommendations of the ISPOR Special Task Force Report [7]. *Value Health*. 2018;21(2):161–165.
21. Neumann PJ, Willke RJ, Garrison Jr LP. A health economics approach to US value assessment frameworks-introduction: an ISPOR Special Task Force Report [1]. *Value Health*. 2018;21(2):119–123.
22. Developing NICE guidelines: the manual. The National Institute for Health and Care Excellence. https://www.nice.org.uk/process/pmg20/chapter/incorporating-economic-evaluation#the-reference-case. Accessed February 14, 2021.
23. Value assessment framework. Institute for Clinical and Economic Review. https://icer.org/wp-content/uploads/2022/01/ICER_2020_2023_VAF_120821.pdf. Accessed February 14, 2021.
24. Guidelines for the economic evaluation of health technologies: Canada. Canadian Agency for Drugs and Technologies in Health (CADTH). https://www.cadth.ca/sites/default/files/pdf/guidelines_for_the_economic_evaluation_of_health_technologies_canada_4th_ed.pdf. Accessed February 14, 2021.
25. Louviere J, Lings I, Islam T, Gudergan S, Flynn T. An introduction to the application of (case 1) best–worst scaling in marketing research. *Int J Res Mark*. 2013;30(3):292–303.
26. Auger P, Devinney TM, Louviere JJ. Using best–worst scaling methodology to investigate consumer ethical beliefs across countries. *J Bus Ethics*. 2007;70(3):299–326.
27. Cohen S, Neira L. Measuring preference for product benefits across countries. ESOMAR. https://www.warc.com/fulltext/ESOMAR/79133.htm. Accessed February 14, 2021.
28. Lee JA, Soutar G, Louviere J. Measuring values using best-worst scaling: the LOV example. *Psychol Mark*. 2007;24(12):1043–1058.
29. Helson H. *Adaptation-Level Theory: An Experimental and Systematic Approach to Behavior*. New York, NY: Harper and Row; 1964.
30. Flynn TN, Louviere JJ, Peters TJ, Coast J. Best–worst scaling: what it can do for health care research and how to do it. *J Health Econ*. 2007;26(1):171–189.
31. Cheung KL, Wijnen BF, Hollin IL, et al. Using best–worst scaling to investigate preferences in health care. *Pharmacoeconomics*. 2016;34(12):1195–1209.
32. Mühlbacher AC, Kaczynski A, Zweifel P, Johnson FR. Experimental measurement of preferences in health and healthcare using best-worst scaling: an overview. *Health Econ Rev*. 2016;6(1):2.
33. Bridges JF, Beusterien K, Heres S, et al. Quantifying the treatment goals of people recently diagnosed with schizophrenia using best–worst scaling. *Patient Prefer Adherence*. 2018;12:63.

34. Louviere JJ, Flynn TN. Using best-worst scaling choice experiments to measure public perceptions and preferences for healthcare reform in Australia. *Patient*. 2010;3(4):275–283.
35. *Qualtrics. Qualtrics Software*. Utah: Version XM. Provo; 2020.
36. Louviere JJ, Flynn TN, Marley AAJ. *Best-Worst Scaling: Theory, Methods and Applications*. Cambridge, United Kingdom: Cambridge University Press; 2015.
37. The MaxDiff system technical paper. Sawtooth Software. https://sawtoothsoftware.com/resources/technical-papers/maxdiff-technical-paper. Accessed July 23, 2023.
38. Stata Statistical Software: Release 16 [program]. College Station, TX: StataCorp LLC; 2019.
39. CEA registry. Center for the Evaluation of Value and Risk in Health, Tufts Medical Center. https://cevr.tuftsmedicalcenter.org/databases/cea-registry. Accessed July 23, 2023.
40. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. *Fam Med*. 2005;37(5):360–363.
41. Veronesi G, Navone N, Novellis P, et al. Favorable incremental cost-effectiveness ratio for lung cancer screening in Italy. *Lung Cancer*. 2020;143:73–79.
42. Wilkoff BL, Boriani G, Mittal S, et al. Cost-effectiveness of an antibacterial envelope for cardiac implantable electronic device infection prevention in the US healthcare system from the WRAP-IT trial. *Circ Arrhythm Electrophysiol*. 2020;13(10):e008503.
43. Fuller SD, Hu J, Liu JC, et al. Five-year cost-effectiveness modeling of primary care-based, nonmydriatic automated retinal image analysis screening among low-income patients with diabetes. *J Diabetes Sci Technol*. 2022;16(2):415–427.
44. Sørensen AR, Hammeken LH, Qvist AH, Jensen SL, Ehlers LH. Operative treatment of displaced midshaft clavicular fractures is not cost-effective. *J Shoulder Elbow Surg*. 2020;29(1):27–35.
45. Paulsen RT, Sørensen J, Carreon LY, Andersen MØ. Cost-effectiveness of postoperative rehabilitation after surgery for lumbar disc herniation: an analysis based on a randomized controlled trial [published online January 17, 2020]. *J Neurosurg Spine*. https://doi.org/10.3171/2019.11.SPINE191003.
46. Wolff HB, Alberts L, van der Linden N, et al. Cost-effectiveness of stereotactic body radiation therapy versus video assisted thoracic surgery in medically operable stage I non-small cell lung cancer: a modeling study. *Lung Cancer*. 2020;141:89–96.
47. Spahillari A, Zhu J, Ferket BS, et al. Cost-effectiveness of contemporary statin use guidelines with or without coronary artery calcium assessment in African American individuals. *JAMA Cardiol*. 2020;5(8):871–880.
48. Johnston K, Danchenko N, Hansen R, et al. Cost effectiveness and impact on quality of life of abobotulinumtoxinA and onabotulinumtoxinA in the treatment of children with lower limb spasticity in Canada. *J Med Econ*. 2020;23(6):631–640.
49. Powell J, Williams V, Atherton H, et al. Effectiveness and cost-effectiveness of a self-guided Internet intervention for social anxiety symptoms in a general population sample: randomized controlled trial. *J Med Internet Res*. 2020;22(1):e16804.
50. Peultier AC, Pandya A, Sharma R, Severens JL, Redekop WK. Cost-effectiveness of mechanical thrombectomy more than 6 hours after symptom onset among patients with acute ischemic stroke. *JAMA Netw Open*. 2020;3(8):e2012476.
51. Flynn TN, Marley AA. *Best-Worst Scaling: Theory and Methods. Handbook of Choice Modelling*. Cheltenham, United Kingdom: Edward Elgar Publishing; 2014.
52. Kim DD, Wilkinson CL, Pope EF, Chambers JD, Cohen JT, Neumann PJ. The influence of time horizon on results of cost-effectiveness analyses. *Expert Rev Pharmacoecon Outcomes Res*. 2017;17(6):615–623.
53. Caro JJ, Briggs AH, Siebert U, Kuntz KM. ISPOR-SMDM Modeling Good Research Practices Task Force. Modeling good research practices–overview: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-1. *Med Decis Making*. 2012;32(5):667–677.
54. Kuntz KM, Russell LB, Owens DK, et al. Decision models in cost-effectiveness analysis. In: Neumann PJ, Sanders GD, Russell LB, Siegel JE, Ganiats TG, eds. *Cost-Effectiveness in Health and Medicine*. New York, NY: Oxford University Press; 2016:105–136.
55. Verguet S, Kim JJ, Jamison DT. Extended cost-effectiveness analysis for health policy assessment: a tutorial. *Pharmacoeconomics*. 2016;34(9):913–923.
56. Cookson R, Mirelman AJ, Griffin S, et al. Using cost-effectiveness analysis to address health equity concerns. *Value Health*. 2017;20(2):206–212.
57. Avanceña ALV, Prosser LA. Examining equity effects of health interventions in cost-effectiveness analysis: a systematic review. *Value Health*. 2021;24(1):136–143.
58. Brock DW, Daniels N, Neumann PJ, et al. Ethical and distributive considerations. In: Neumann PJ, Sanders GD, Russell LB, Siegel JE, Ganiats TG, eds. *Cost-Effectiveness in Health and Medicine*. New York, NY: Oxford University Press; 2017:319–327.
59. Nord E, Pinto JL, Richardson J, Menzel P, Ubel P. Incorporating societal concerns for fairness in numerical valuations of health programmes. *Health Econ*. 1999;8(1):25–39.
60. Williams A. Cost-effectiveness analysis: is it ethical? *J Med Ethics*. 1992;18(1):7–11.
61. Bond K, Stiffell R, Ollendorf DA. Principles for deliberative processes in health technology assessment [published online August 4, 2020]. *Int J Technol Assess Health Care*. https://doi.org/10.1017/S0266462320000550.
62. Bell CM, Urbach DR, Ray JG, et al. Bias in published cost effectiveness studies: systematic review. *BMJ*. 2006;332(7543):699–703.
63. Garattini L, Koleva D, Casadei G. Modeling in pharmacoeconomic studies: funding sources and outcomes. *Int J Technol Assess Health Care*. 2010;26(3):330–333.
64. Xie F, Zhou T. Industry sponsorship bias in cost effectiveness analysis: registry based analysis. *BMJ*. 2022;377:e069573.
65. Catalá-López F, Caulley L, Ridao M, et al. Reproducible research practices, openness and transparency in health economic evaluations: study protocol for a cross-sectional comparative analysis. *BMJ Open*. 2020;10(2):e034463.
66. Cohen JT, Neumann PJ, Wong JB. A call for open-source cost-effectiveness analysis. *Ann Intern Med*. 2017;167(6):432–433.