ROBVALU: a tool for assessing risk of bias in studies about people's values, utilities, or importance of health outcomes

Samer G Karam,[1,2] Yuan Zhang,[1,2] Hector Pardo-Hernandez,[3,4] Uwe Siebert,[5,6,7] Laura Koopman,[8] Jane Noyes,[9] Jean-Eric Tarride,[1,10,11] Adrienne L Stevens,[12] Vivian Welch,[13] Zuleika Saz-Parkinson,[14] Brendalynn Ens,[15] Tahira Devji,[16] Feng Xie,[1,10] Glen Hazlewood,[17,18] Lawrence Mbuagbaw,[1,19,20,21,22,23] Pablo Alonso-Coello,[3,4,24] Jan L Brozek,[1,2] Holger J Schünemann[1,25]

For numbered affiliations see end of the article

Correspondence to:
H J Schünemann
schuneh@mcmaster.ca
(ORCID 0000-0003-3211-8479)

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2024;385:e079890
http://dx.doi.org/10.1136/bmj-2024-079890

Accepted: 09 April 2024

People's values are an important driver in healthcare decision making. The certainty of an intervention's effect on benefits and harms relies on two factors: the certainty in the measured effect on an outcome in terms of risk difference and the certainty in its value, also known as utility or importance. The GRADE (Grading of Recommendations, Assessment, Development, and Evaluations) working group has proposed a set of questions to assess the risk of bias in a body of evidence from studies investigating how people value outcomes. However, these questions do not address risk of bias in individual studies that, similar to risk-of-bias tools for other research studies, is required to evaluate such evidence. Thus, the Risk of Bias in studies of Values and Utilities (ROBVALU) tool was developed. ROBVALU has good psychometric properties and will be useful when assessing individual studies in measuring values, utilities, or the importance of outcomes. As such, ROBVALU can be used to assess risk of bias in studies included in systematic reviews and health guidelines. It also can support health research assessments, where the risk of bias of input variables determines the certainty in model outputs. These assessments include, for example, decision analysis and cost utility or cost effectiveness analysis for health technology assessment, health policy, and reimbursement decision making.

SUMMARY POINTS

- Assessing the risk of bias in individual studies is an essential step to determine overall certainty of evidence in a systematic review or health technology assessment and for guideline development
- The Risk of Bias in Values and Utilities (ROBVALU) tool assesses risk of bias in quantitative studies of people's values, utilities, or importance of outcomes
- A sequential mixed methods approach was used to develop ROBVALU, initially based on signalling questions and subdomains developed by the GRADE working group to assess risk of bias; a modified Delphi approach was used for final refinement of the tool
- ROBVALU covers four separate subdomains through which bias might be introduced; individual subdomain judgments inform the overall risk of bias of studies
- ROBVALU has demonstrated high validity and reliability

Healthcare decision making relies on evidence on the relative effectiveness, safety, and cost effectiveness of an intervention evaluated in appropriate studies.[1 2] Choosing between different interventions (such as preventive, diagnostic, or treatment strategies) depends on the importance or value that people place on specific health states or health outcomes.[2] Values have a major role at different levels of decision making, from the individual level to the healthcare system level. In this context, people's values reflect the importance they place on outcomes of interest that result from decisions about using an intervention—for example, taking a certain test or starting a new treatment regimen.[2] We use the term "people" when talking about value because the term is inclusive to patients, healthcare providers, policy makers, and the general public.

Utility instruments are widely used to elicit the absolute value of a health outcome, and provide an index measure anchored on a scale with 1 reflecting perfect health and 0 reflecting being dead.[3 4] Various methods are used to establish values, including direct measures of utility, indirect measurements of utility, or qualitative research.[2 5] The visual analogue scale (VAS) is one of the simplest measures to elicit these values. People are asked to rate a health state on a VAS that is then converted to a utility value.[6 7] While the scale directly measures the importance of an outcome, concerns exist about how accurate and valid it might be.[2] Other direct measures such as the standard gamble and time trade-off require people to choose between their current health state and a treatment option that could result in perfect health or in immediate death.[4 8] Discrete choice experiments ask people to choose between two or more treatment

options where the choices differ in terms of their attributes, that are defined by the investigators.[9] The relative importance of each attribute is then inferred by analysing the responses, assuming that patients choose the option with the highest value.[9] Indirect methods of measuring utility values include validated, health related, quality-of-life instruments, such as the EQ-5D and the Health Utilities Index.[10] The EQ-5D requires respondents to answer questions across five domains that are converted to a utility value using validated scoring systems.[11 12]

General application of utility values in research

These utility values allow researchers to weigh the benefits and harms of an option and, thus, they also are important in health economics and health technology assessments.[3 13] For instance, in decision analysis, they are required to calculate quality adjusted life years. Confidence in studies that report on values needs to be ascertained for decision making in guideline recommendations, health technology assessments, or coverage decision.[14] For example, in a systematic review on people with chronic obstructive pulmonary disease, we found moderate certainty that patients value adverse events as important, but on average valued them as less important than symptom relief.[15] We also found moderate certainty that exacerbation and hospital admission owing to exacerbation are the outcomes that patients with chronic obstructive pulmonary disease rate as most important. In another example, a systematic review on patients' values on venous thromboembolism, we found that people with cancer placed more importance on a reduction in new or recurrent venous thromboembolism than on a decrease in major or minor bleeding events.[16]

The Grading of Recommendations, Assessment, Development, and Evaluation (GRADE) Evidence to Decision frameworks is a widely used approach in guidelines, health technology assessment, and other decisions. The frameworks require judgments about the certainty in how much people value the main outcomes: "Is there important uncertainty about . . . how much people value the main outcomes?"[17 18] A key determinant of certainty is internal validity—that is, how well individual studies were designed and conducted (ie, internal validity, which GRADE and Cochrane label as the risk-of-bias (ROB) domain).

Risk of bias

Similar to other study designs, threats to internal validity arising from the study design, conduct, analysis, and reporting of the study introduce ROB in research on utility values.[2] Poor study quality could result in indirectness which encompasses applicability and external validity, often as a result of PICO (patient/population, intervention, comparison, and outcomes) elements. Another quality issue is low sample size or no sample size calculation, which could result in imprecision. ROB assessment tools are developed to assess biases that result in threats of internal validity and would not measure indirectness and precision. Quality assessment tools and reporting checklists often include all factors of a study's qualities and safeguards, but these tools differ from a ROB assessment tool that aims to present a ROB judgment for a study. A key factor that might introduce bias in values studies is the instrument used to measure utilities of the people in the study. Bias means that a value people place on an outcome in a research study (eg, a value of 0.5 for stroke) would be systematically different from the true value that people would place on that outcome. For example, the true unbiased value might be 0.3 and, thus, use of biased estimates would provide inaccurate answers in the modelling and health decision making context.

ROB assessment tools exist for many study designs, including the Cochrane Risk of Bias 2 (RoB 2) for randomised trials,[19] ROBINS-I for non-randomised studies of the effects of interventions,[20] and ROBINS-E for studies about exposures.[21 22] Critical appraisal tools to assess the quality of a study are also study design specific, such as the Newcastle-Ottawa scale and the Joanna Briggs Institute's critical appraisal tool for cross sectional studies.[23 24] These tools are regularly used by researchers to assess the quality of individual studies or to assess ROB, but they were not developed for studies on utility values. These checklists invariably include questions specific to the study design, which would not always be appropriate to answer in studies about people's values (eg, "Were there deviations from the intended intervention that arose because of the trial context?" or "Was the exposure measured in a valid and reliable way?").[19 21 22] For studies on utility values, a major concern that is not adequately addressed by any commonly used ROB tool is the method used to elicit people's values. The measurement instrument needs to be valid and reliable, be used appropriately, use valid health outcomes, and explore proper understanding of the instrument. No validated tool is available for the nuanced assessment of ROB in individual studies measuring utility values.[9 20 25 26]

Objective

To properly implement evidence based decision making and formulate evidence based recommendations in clinical or public health guidelines, evaluation of ROB is crucial in studies of values, utilities, or importance of outcomes. However, owing to the absence of specialised and validated tools to assess ROB, this evaluation is rarely done. Thus, our goal was to develop, validate, and describe a pragmatic tool for studies measuring the value people place on health outcomes with appropriate guidance to apply it correctly.

Development of the ROBVALU tool and guidance

We used a sequential, mixed methods approach to develop ROBVALU and related guidance document (supplement S1),[27] starting with a qualitative approach and followed by a quantitative phase to assess the psychometric properties of the tool (fig 1). In the qualitative phase, we first considered the ROB signalling questions (appendix table A1) and

subdomains that we had developed for GRADE guidance to assess ROB about values across studies in a body of evidence.[2] For that GRADE guidance, we iteratively developed the subdomains and signalling questions starting with a list of 23 items identified as part of a systematic survey project.[26] The core research group reviewed the 23 items to identify any missing item that might be relevant for the single study ROBVALU tool. After thorough group discussions, a decision was made not to add any new items or subdomains to avoid complexity, thereby improving applicability, feasibility, and adoption of the tool.

We first structured a preliminary version of the tool and added simple considerations to help answer

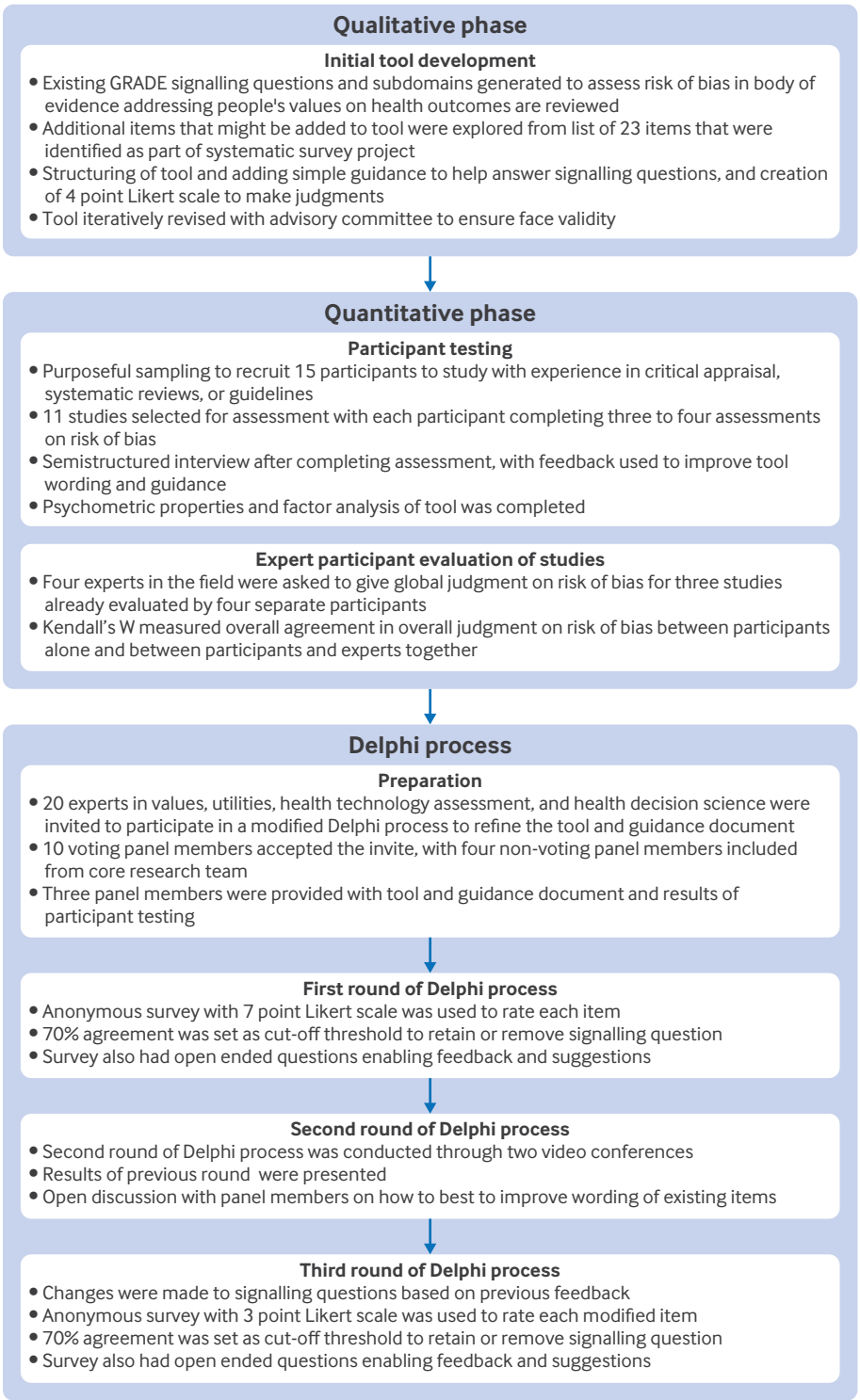

Fig 1 | Tool development process for the Risk of Bias in Values and Utilities (ROBVALU) tool. GRADE=Grading of Recommendations, Assessment, Development, and Evaluation

the signalling questions. These signalling questions were categorised into four subdomains: selection of participants into the study, completeness of data, measurement instrument, and data analysis. We used a 4 point, Likert-type scale (ie, yes, probably yes, probably no, no) to judge the individual items, to avoid a neutral option of a 5 point Likert scale when studies lack sufficient information to make a proper judgment. In each subdomain, the tool asked how important and how serious the ROB issue is. The core research group iteratively revised the tool and accompanying guidance document. An advisory group of experts provided feedback and suggested appropriate changes to establish face and content validity (supplement S2).

Participant testing

We used purposeful sampling to recruit 15 participants with experience in critical appraisal, systematic reviews, or guidelines for user testing and semi-structured interviews (supplement S3). The participants had a broad level of expertise, from masters level students to senior researchers with experience in health research ranging from six months to 30 years (appendix table A2). All users received the ROBVALU tool and the accompanying guidance document (supplement S1). We instructed the participants to complete three to four assessments and every sample study was assessed by four users independently, 11 studies in total were assessed (appendix table A3). Based on feedback received in the semi-structured interview after user testing, we iteratively revised and improved the guidance document throughout the project with a focus on the wordings, spelling, and grammatical structure of the guidance document. The ROBVALU tool demonstrated good psychometric properties with an overall intraclass correlation coefficient of 0.87 and the four subdomains showed good to excellent reliability ranging from 0.80 to 0.91 (table 1 and supplement S4). We also calculated the inter-rater reliability of the global ROB judgment using the ROBVALU tool with Kendall's W, which showed substantial agreement of 0.62 (supplement S4). We invited four expert participants in the field to provide a global judgment for ROB without using the ROBVALU, with each expert rating three to four studies. When we added expert participant responses of the global ROB judgment, the Kendall's W dropped to 0.45, showing moderate agreement (supplement S4). However, only four global judgment responses were more than one level of seriousness higher or lower than the expert participant judgment (appendix table A4).

Table 1 | Reliability of the Risk of Bias in Values and Utilities (ROBVALU) tool. CI=confidence interval

Subdomain	Intraclass correlation (95% CI)
Selection of participants	0.87 (0.79 to 0.93)
Completeness of data	0.90 (0.84 to 0.94)
Measurement instrument	0.80 (0.69 to 0.88)
Data analysis	0.91 (0.86 to 0.95)
Total	0.86 (0.78 to 0.91)

Modified Delphi process

Finally, following our protocol, we used purposeful sampling to invite 20 experts in values, utilities, health technology assessment, and health decision science to participate in a modified Delphi process for final refinement of the tool (supplement S5, fig S8).[28-30] We used our extensive network of global colleagues working in the field of study to identify and invite the expert panel. Ten voting members accepted the invite to participate in the Delphi panel, and four members of the working group participated as non-voting members. We shared the ROBVALU tool draft, guidance document, and results of our participant testing with the panel members.

The first round of the Delphi process involved an anonymous survey to determine the signalling questions to be included. The second round took place via recorded video conferences with the aim of identifying common themes and reaching consensus on simplifying and harmonising language across the tool. The third and final round of the Delphi process included an anonymous survey for final consensus on the wording of the signalling questions and the proposed methods for providing a global ROB judgment. We used Google forms to prepare the surveys; the first survey used a 7 point Likert scale (ie, strongly agree, agree, somewhat agree, neutral, somewhat agree, disagree, and strongly disagree) to rate each item, with 70% agreement set as the cut-off threshold to retain or remove a signalling question. The final survey used a 3 point scale (ie, agree, neutral, and disagree) with a 70% agreement set as the cut-off threshold to retain the signalling question.

We had a 100% response rate in the first round of the Delphi process, with 80-100% consensus to retain all signalling questions. We also collected feedback from open ended questions for suggested edits for the signalling questions (supplement S6). In the second round of the Delphi process, we presented the ROBVALU tool, psychometric properties, exploratory factor analysis, and results of the first round of the Delphi to the panel members. After deliberating on the tool's properties, agreement was reached to edit some signalling questions to simplify the language or to harmonise the language across the tool, which resulted in minor changes only. We also discussed how to make a final judgment for ROB for a study.

We had a 100% response rate in the third and final Delphi round, with 80-100% consensus on the tool's signalling questions, including those with minor adjustments to the wording. We also established a consensus of >70% that the overall ROB judgment should match the most severe ROB judgment on an item, unless appraisers can provide justifications to rate the overall ROB lower (eg, many concerns on many items) or higher (eg, concern seems not to have an important influence on overall ROB). For example, if multiple subdomains were rated as very serious, the final judgment could be rated as extremely serious (supplement S7).

Risk-of-bias subdomains

ROBVALU includes seven key signalling questions across four subdomains: selection of participants into the study, completeness of data, measurement instrument, and data analysis (table 2).

Selection of participants into the study

Precise research questions include a clear definition of the target population. The study population of any empirical study must be representative for this target population, and is therefore, a critical component

Table 2 | Subdomains and considerations in the Risk of Bias in Values and Utilities (ROBVALU) tool

Subdomain and signalling question	Rationale and examples
Selection of participants	
Question: Was an appropriate study sample selected from the study's sampling frame?	Reviewers should determine whether the sampling strategy was conducted in a manner to minimise the risk of selection bias. In a comparison study, selection bias refers to systematic differences between baseline characteristics of the groups that are compared. Here, for risk of bias, we only refer to bias internal to the study, rather than inadequate generalisability (applicability or directness); that is, selection bias that could happen when the achieved sample is deviated from the intended sample (as described in the protocol or the methods section of the study), rather than from the population we intend to extrapolate the conclusion to (ie, the target population of the research question). We need to assess to what extent the achieved sample is similar to the intended sample. The sampling strategy is a critical component because it will influence the results through the population the researcher's had studied. For example, for a cross sectional study, a stratified random sampling strategy would minimise the risk, while a convenience sample would probably be a biased sample for the study population.
Consider what is the sampling strategy (eg, random sample or consecutive sample); what subsets of the population are more or less likely to be reached with this sampling strategy.	
• Yes • Probably yes • Probably no • No	
Completeness of data	
Question: Was the attrition rate sufficiently low to minimise the risk of bias?	In addition to sampling strategy, in surveys, response rate also influences the representativeness of the achieved sample. The higher the response rate, the less likely that risk of bias is a concern. Response could be influenced by various factors, including study design, study purposes, sampling strategy, and survey administration. There is no single rule for an inadequate response rate, however; if the judgment is not an acceptable response rate, provide justification. For longitudinal studies with follow-up periods planned and used, the attrition rate such as dropouts, loss to follow-up, and exclusions could be another source of concern.
Consider the response rate; if follow-up periods were planned and used, what was the attrition rate during the follow-up period; were the participants who responded systematically different from those who have not responded.	
• Yes • Probably yes • Probably no • No	
Measurement instrument	
Question: Was the instrument used to measure patient values and preferences in a valid and reliable manner?	Measurement instrument refers to direct measures of utility (eg, standard gamble and time trade-off, conjoint analysis with discrete choice experiments) and indirect measurement instruments of utility such as EQ-5D. A variety of measurement instruments could be chosen, including those providing utility measurements (eg, standard gamble, time trade-off, visual analogue scale), willingness to pay, discrete choice, or other structured scales. For a specific study, the validity and reliability of the instrument might not always have been determined. In these cases, to be considered a reliable and valid instrument, either the researchers provide the validity and reliability information in the study being evaluated, or the measurement instruments are widely accepted as both reliable and valid.
Consider what was the measurement instrument selected; does the instrument have validity and reliability that is well constructed; or is this instrument widely accepted in this area to have adequate reliability and validity (translation and culturally adapted in guidance).	
• Yes • Probably yes • Probably no • No	
Question: Was the instrument used in the intended way?	Faulty measurements could be a source of bias, either due to inherent shortcomings in a measurement tool or via administration error. For a specific study, the researchers should demonstrate the measurement tools were used correctly or in a manner conforming to their rationale to minimise the risk of introducing bias. If applicable, tools should be used in a consistent manner across different subpopulations.
• Yes • Probably yes • Probably no • No	
Question: Was a valid representation of the outcome (health state) used?	The description of health states is another possible source of bias. High quality description provides participants with best available evidence, while wrong or insufficient information based on low quality evidence might mislead participants and bias the measurement. High quality description consists of the experience, probability, duration, and consequences of a health state and should be presented in an understandable format.
• Yes • Probably yes • Probably no • No	
Question: Did the researchers check for understanding of the instrument?	If the participants have problems to understanding the techniques, the results they provide are likely to be misleading. There is a gradient in the understanding of measurement techniques, depending on whether the understanding is checked formally, and whether the understanding is adequate.
• Investigators tested the understanding, and understanding was adequate • Investigators did not formally test the understanding, but evidence suggested adequate understanding • Investigators did not formally test the understanding, but evidence suggested inadequate understanding • Investigators tested the understanding, but understanding was inadequate	
Data analysis	
Question: Were the results analysed appropriately to avoid influence of bias and confounding?	The appropriateness of data analysis would include the strategy to deal with missing data or excluded cases from analysis. If confounding factors or other influential factors exist, statistical techniques such as stratification or regression analyses for adjustment of measured confounding factors might be taken when appropriate. Often, in an outcome valuation study, no adjustment is made, and the results are reported in different subgroups. Furthermore, the appropriateness of model selection (if any) or analysis strategy should be checked.
Consider whether the adjustment, stratification, strategy to deal with missing data, and model selection, if any, was appropriate.	
• Yes • Probably yes • Probably no • No	

because bias in the selection will lead to biased estimates of the values people place on outcomes in the target population.[2] When assessing selection bias, users should consider the study's sampling strategy, in particular if the achieved sample population deviates from the intended sample population,[2] because this might lead to biased estimates for the study's population of interest owing to threats to internal validity. If the achieved sample population does not deviate from the intended sample population but differs from the population researchers intend to extrapolate the results to, this difference will result in a lack of generalisability. We refer to this lack of generalisability as indirectness, which encompasses applicability and external validity. The ROBVALU tool is not intended to deal with indirectness, a different domain in assessing the certainty of a body of evidence according to GRADE, but we are developing a tool that is specific to indirectness separately.

Completeness of data

When judging completeness of data, reviewers need to consider the response rate of the study population, the attrition rate if follow-up was involved, and the differential responders compared with non-responders.[2] High response rates and low proportion of loss to follow-up are clearly preferable, and a high proportion of non-response or dropout rates could be problematic.[2] Participants providing responses could plausibly differ from those who do not, and researchers should consider that results coming only from those participants who responded or completed follow-up might be misleading.[2]

Measurement instrument

Reliable and valid instruments should be used to measure the relative importance of outcomes in values, preferences, and utility studies.[2] Using unreliable or poorly validated instruments can result in biased measurements of the outcome. Similarly, utility values for specific health states based on instruments not sufficiently validated that are used as input parameters for decision analytical models can result in biased estimates, such as quality adjusted life years derived from state transition models.[31 32] Researchers conducting primary empirical studies should provide information regarding the measurement properties of their chosen instrument.[2]

Researchers should also demonstrate that the instrument has been used correctly and in a consistent manner across all participants in a study. For example, if the standard gamble is to be administered by an interviewer, buta subset of participants used self-administration, this could result in biased utility estimates that could be due to systematic differences between the two groups. In addition, an optimal representation of the outcome or health state should be presented or described in a way that accurately reflects the attribute the researchers intended to measure. This information could include a detailed explanation of how the outcome defines the experience, the probability of the outcome, durations, and possible consequences. Finally, researchers should evaluate whether participants had a proper understanding of the instrument to complete the tasks.

Data analysis

Studies should explore heterogeneity in values when appropriate and present results for the different subgroups. The data analysis plan and exploration of heterogeneity should be outlined a priori before collection of data. A causal framework that helps delineate health state and outcome interactions with possible confounding factors will help make assumptions explicit. If heterogeneity is found, the evaluator needs to consider whether the adjustment, stratification, or model selection used in the study reporting on values was appropriate.[2] Adjusting for important confounding factors (such as age if it is associated with the intervention and influences the estimated values) or reporting values in a stratified manner reduces biased estimates of the value placed on an outcome. In addition, self-inflicted biases, including selection bias or immortal time bias should be controlled for appropriately using modern causal inference methods (eg, target trial emulation or g methods for time varying confounding).[33]

ROBVALU tool application

The assessment of ROB in studies evaluating the value people place on outcomes follows seven steps:

1. Specify the research or review question.
2. Specify the outcome being assessed.
3. Identify the sampling frame, the response rate and/or attrition rate, the measurement instrument used, and the data analysis plan.
4. Answer the signalling questions of the four subdomains.
5. Make a judgment if the four subdomains have important ROB concerns.
6. Formulate a ROB judgment for the four subdomains.
7. Formulate an overall ROB judgment for the study outcome being assessed.

The ROBVALU tool (supplement S8) provides users with space to record vital information of the study being assessed, and signalling questions to all four subdomains that must be answered. We validated a 4 point Likert-type scale (yes, probably yes, probably no, no) to respond to the individual signalling questions (items). When rating individual signalling questions, we suggest following the flowchart in figure 2 for consistent answers between raters. In each subdomain, the tool asks to specify how important the ROB issue is on a 4 point Likert-type scale (yes, probably yes, probably no, no), and how serious the overall ROB issue is on a 4 point Likert-type scale (not serious, serious, very serious, extremely serious). Responses to the signalling questions should provide the basis for the subdomain level judgment, of how important

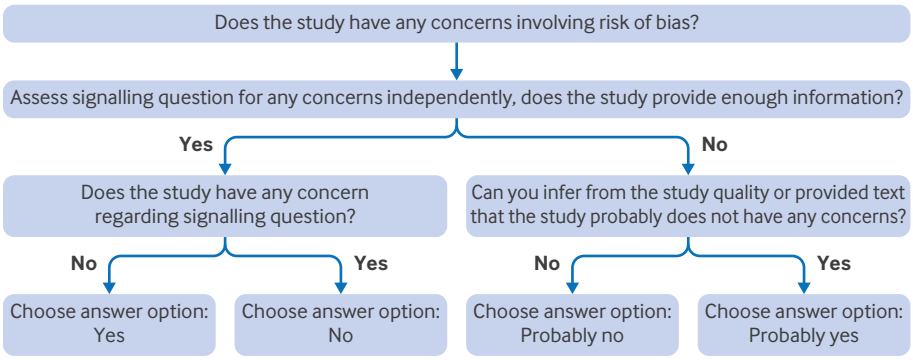

Fig 2 | Rating individual signalling questions in the Risk of Bias in Values and Utilities (ROBVALU) tool

and how serious the ROB issues are in the study. Raters should provide a rationale for the response as free text, to justify their judgments. We suggest that the final judgment for each subdomain inversely correlates with the signalling question judgment. For example, in the measurement instrument subdomain, if the answer to "Was the instrument administered in the intended way?" was "No," then the answer to "Are there important risk of bias issues concerning the measurement instruments?" should be "Yes." If raters believe that the lowest signalling question judgment does not reflect the overall subdomain judgment, they might choose not to deem the results of the study at ROB for that subdomain, but they are asked to provide explanations for why they would not do this.

The global ROB judgment for a study corresponds to the lowest subdomain judgment (table 3), because any domain level bias will lower our confidence in the study results. If users do not believe that the lowest subdomain judgment reflects the global ROB judgment, they should provide a justification. For example, if a study has a low response rate resulting in very serious ROB domain judgment and the study results are comparable to better quality studies, a reviewer might consider that the subdomain judgment does not reflect the global ROB judgment. Box 1 presents an illustrative example of a completed assessment (supplement S9).

Discussion

We have developed and validated the ROBVALU tool, a new instrument to assess ROB in studies measuring the value, utility, or relative importance that people place on health outcome. We followed a sequential mixed methods approach, by first adapting the signalling questions from the GRADE guidance for judging ROB across studies. ROBVALU differs from existing GRADE guidance by specifically assessing ROB in individual studies as opposed to across studies.[2] We iteratively revised the tool with our core group and an advisory group. The final draft tool contains 15 items in four subdomains: selection of participants, completeness of data, measurement instrument, and data analysis. We conducted a validation exercise with 15 participants that showed good reliability. Additional refinement using a modified Delphi process established construct validity on the final content of the tool.

Strengths and limitations

Assessing ROB is an essential step to assess the overall certainty of the evidence in a systematic review or health technology assessment and to develop a guideline. This assessment has often relied on adapting ROB tools not specifically designed for this type of research.[26] However, the lack of validation could lead to unreliable certainty of the evidence assessments, both for single

Table 3 | Response options for judgments on risk of bias at an overall study level, according to the Risk of Bias in Values and Utilities (ROBVALU) tool

Response option	Criteria
Not serious risk of bias	Study is judged to have no serious risk of bias for all subdomains
Serious risk of bias	Study is judged to be at serious risk of bias in at least one subdomain, but not very or extremely serious risk of bias in any subdomain
Very serious risk of bias (study has some important problems)	Study is judged to be at very serious risk of bias in at least one subdomain, but not at extremely serious risk of bias in any subdomain
Extremely serious risk of bias	Study is judged to be at extremely serious risk of bias in at least one subdomain

Box 1: Example application of the Risk of Bias in Values and Utilities (ROBVALU) tool to assess risk of bias in values assigned to exacerbation of chronic obstructive pulmonary disease[34]

In a study of 65 men and women with chronic obstructive pulmonary disease, researchers assessed the utility value that participants placed on an exacerbation, at seven study sites in the US when they visited an outpatient clinic within 48 hours of symptom onset.[34] Eligible participants were at least 40 years old and were current or former smokers with a history of at least 10 pack years. Of 65 participants, 59 completed the study, three were lost to follow-up, and three were ineligible. Utility values were measured using the EQ-5D.

An assessment using the ROBVALU tool revealed the following (supplement 9):

- Selection of participants into the study would likely lead to risk of bias.
  - Exacerbations that required hospital admission were considered severe and were excluded from this study and might importantly bias the estimates. Thus, the population was deemed to be probably not representative of the intended population.
- Completeness of data was present:
  - Only three patients were lost to follow-up, which did not cause risk of bias.
- Measurement instrument caused some concern about risk of bias:
  - It was not clear whether the instrument was used in a valid and reliable manner, but it was applied in the intended way using a valid representation of the outcome. Patients also appeared to show an understanding of the instrument that was used and did not encounter difficulties, but this was not reported.
- Data analysis did not cause concern for risk of bias:
  - Adjustment, stratification, and model selection was appropriate based on a plan created a priori.

ROBVALU assessment

Overall risk of bias was deemed serious because of issues related to the selection of participants into the study and the way the measurement instrument was used.

studies and for a body of evidence. By using ROBVALU, evaluators can incorporate the ROB assessment into their meta-analysis, such as performing a sensitivity analysis to evaluate how studies with higher ROB might affect the study's conclusion or primary outcomes. An advantage of the ROBVALU tool is the use of standardised GRADE terminology and judgments to facilitate assessment when establishing the certainty of the evidence. The ROBVALU tool can also be used to assess ROB in all elicitation studies of values, utilities, and importance of outcomes that use discrete choice, ranking, indifference, and rating methods.[35] Finally, the tool can be used in individual studies that use indirect methods to elicit people's preferences, such as quality of life and EQ-5D scores.

This study and the derived tool also has several limitations. The new tool focuses on assessing values quantitively. For any given intervention, there is usually qualitative literature exploring what patients want to achieve and what they value (or not) from interventions; this information could be important for decision making. While some of the signalling questions might be used for qualitative studies, other signalling questions will not apply. Further exploration with qualitative studies should be performed to assess how ROBVALU can be adapted for that particular use, or whether a different tool is required. Furthermore, an exploratory factor analysis showed that one item in the tool had relatively poor fit (Was a valid representation of the outcome (health state) used?), but this poor fit could be due to the relatively small sample size. However, we retained this item because of feedback from the Delphi panel, who deemed it important. External validation of ROBVALU's reliability by different users and on different studies will help refine the guidance and the tool.

Future implications

ROBVALU allows researchers to appraise individual studies reporting utilities, values, or the importance of outcomes for risk of bias. For example, in health technology assessments, the certainty of input variables from an individual study determines the certainty of outputs from decision analytical models (eg, cost utility and cost effectiveness analyses).[32 36] ROBVALU should also help with evaluating ROB as part of a systematic review, health technology assessment, or formal health guideline, to develop recommendations and make judgments across the overall body of this type of evidence (eg, assessing overall certainty of the evidence when following the GRADE approach).

AUTHOR AFFILIATIONS

[1]Department of Health Research Methods, Evidence and Impact, McMaster University, Hamilton, ON, Canada

[2]Michael G DeGroote Cochrane Canada and McMaster GRADE Centres, McMaster University, Hamilton, ON, Canada

[3]Iberoamerican Cochrane Centre, Sant Antoni Maria Claret, Barcelona, Spain

[4]Centro de Investigación Biomédica en Red de Epidemiología y Salud Pública (CIBERESP), Madrid, Spain

[5]Department of Public Health, Health Services Research and Health Technology Assessment, Institute of Public Health, Medical Decision Making and Health Technology Assessment, UMIT TIROL-University for Health Sciences and Technology, Hall in Tirol, Austria

[6]Center for Health Decision Science and Departments of Epidemiology and Health Policy and Management, Harvard T H Chan School of Public Health, Boston, MA, USA

[7]Institute for Technology Assessment and Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

[8]Department of Specialist Medical Care, National Health Care Institute, Diemen, Netherlands

[9]School of Medical and Health Sciences, Bangor University, Wales, UK

[10]Centre for Health Economics and Policy Analysis, McMaster University Faculty of Health Sciences, Hamilton, ON, Canada

[11]Programs for Assessment of Technologies in Health, St Joseph's Healthcare Hamilton, Hamilton, ON, Canada
[12]Centre for Immunisation Programmes, Public Health Agency of Canada, ON, Canada
[13]Bruyère Research Institute and, School of Epidemiology and Public Health, University of Ottawa, Ottawa, ON, Canada
[14]European Commission, Joint Research Centre, Ispra, Italy
[15]Implementation Support and Knowledge Mobilisation, Canadian Agency for Drugs and Technologies in Health, Ottawa, ON, Canada
[16]Temerty Faculty of Medicine, University of Toronto, Toronto, ON, Canada.
[17]Department of Medicine, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada
[18]Department of Community Health Sciences, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada
[19]Department of Anaesthesia, McMaster University, Hamilton, ON, Canada
[20]Department of Paediatrics, McMaster University, Hamilton, ON, Canada
[21]Biostatistics Unit, Father Sean O'Sullivan Research Centre, St Joseph's Healthcare, Hamilton, ON, Canada
[22]Centre for Development of Best Practices in Health, Yaoundé Central Hospital, Yaoundé, Cameroon
[23]Division of Epidemiology and Biostatistics, Department of Global Health, Stellenbosch University, Cape Town, South Africa
[24]Institut de Recerca Sant Pau (IR SANT PAU), Sant Quintí, Barcelona, Spain
[25]Clinical Epidemiology and Research Centre (CERC), Humanitas Universityand Humanitas Research Hospital, Via Rita Levi Montalcini 4, 20090 Pieve Emanuele, Milan, Italy

Contributors: The authors are epidemiologists, statisticians, systematic reviewers, and health services researchers, many of whom are involved with methods research and GRADE. Development of ROBVALU was informed by GRADE guideline 19, previously published tools for assessing risk of bias in intervention studies, systematic reviews of available tools to assess risk of bias in values and preferences, and the authors' experience of developing similar tools to assess risk of bias. All authors contributed to development of the ROBVALU tool and to writing associated guidance. SGK, YZ, JLB, and HJS designed the study and formed the core group. YZ, JLB, and HJS conceived of the project. HJS oversaw the project and is guarantor. SGK, YZ, TD, JLB, and HJS drafted the ROBVALU tool. JN, PAC, FX, and US formed the advisory group. SGK led working groups and conducted the semi-structured interviews. SGK and LM analysed the data. HP-H, GH, YZ, and PAC assessed studies. PAC, FX, BE, ZSP, VW, ALS, J-ET, JN, LK, and US participated in the Delphi process as voting members, and HJS, YZ, SGK, and JLB were non-voting members. SGK and HJS drafted the manuscript. YZ, JLB, and HJS obtained funding for the study. All authors reviewed and commented on drafts of the manuscript. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Funding: The study was funded by the Canadian Institutes of Health Research (grant 401310 to HJS and JLB). The funders had no role in considering the study design or in the collection, analysis, interpretation of data, writing of the report, or decision to submit the article for publication.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/disclosure-of-interest/ and declare: support from the Canadian Institutes of Health Research for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Ethical approval: This international study was designed and coordinated at McMaster University after approval by the Hamilton Integrated Research Ethics Board (project ID 5634), and interviews and meetings were conducted in person or over video conference. All participants provided informed consent.

Provenance and peer review: Not commissioned; externally peer reviewed.

1 Boyd CM, Singh S, Varadhan R, et al. Methods for benefit and harm assessment in systematic reviews. 2012. https://www.ncbi.nlm.nih.gov/books/NBK115750/pdf/Bookshelf_NBK115750.pdf.
2 Zhang Y, Alonso-Coello P, Guyatt GH, et al. GRADE Guidelines: 19. Assessing the certainty of evidence in the importance of outcomes or values and preferences-Risk of bias and indirectness. *J Clin Epidemiol* 2019;111:94-104. doi:10.1016/j.jclinepi.2018.01.013
3 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. *Handbook of health decision science*. Springer, 2016:3-13doi:10.1007/978-1-4939-3486-7_1
4 Dolan P, Gudex C, Kind P, Williams A. Valuing health states: a comparison of methods. *J Health Econ* 1996;15:209-31. doi:10.1016/0167-6296(95)00038-0
5 McDonough CM, Tosteson AN. Measuring preferences for cost-utility analysis: how choice of method may influence decision-making. *Pharmacoeconomics* 2007;25:93-106. doi:10.2165/00019053-200725020-00003
6 Torrance GW, Feeny D, Furlong W. Visual analog scales: do they have a role in the measurement of preferences for health states?*Med Decis Making* 2001;21:329-34. doi:10.1177/02729890122062622
7 Rashidi AA, Anis AH, Marra CA. Do visual analogue scale (VAS) derived standard gamble (SG) utilities agree with Health Utilities Index utilities? A comparison of patient and community preferences for health status in rheumatoid arthritis patients. *Health Qual Life Outcomes* 2006;4:25. doi:10.1186/1477-7525-4-25
8 Bleichrodt H, Johannesson M. Standard gamble, time trade-off and rating scale: experimental results on the ranking properties of QALYs. *J Health Econ* 1997;16:155-75. doi:10.1016/S0167-6296(96)00509-7
9 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;14:403-13. doi:10.1016/j.jval.2010.11.013
10 Horsman J, Furlong W, Feeny D, Torrance G. The Health Utilities Index (HUI): concepts, measurement properties and applications. *Health Qual Life Outcomes* 2003;1:54. doi:10.1186/1477-7525-1-54
11 Devlin N, Parkin D, Janssen B. *Methods for analysing and reporting EQ-5D data*. Springer Nature, 2020. doi:10.1007/978-3-030-47622-9
12 Devlin N, Parkin D, Janssen B. An introduction to EQ-5D instruments and their applications. *Methods for analysing and reporting EQ-5D data*. Springer Nature, 2020:1-22. doi:10.1007/978-3-030-47622-9_1
13 Slaughter KB, Meyer EG, Bambhroliya AB, et al. Direct assessment of health utilities using the standard gamble among patients with primary intracerebral hemorrhage. *Circ Cardiovasc Qual Outcomes* 2019;12:e005606. doi:10.1161/CIRCOUTCOMES.119.005606
14 Schünemann HJ, Reinap M, Piggott T, et al. The ecosystem of health decision making: from fragmentation to synergy. *Lancet Public Health* 2022;7:e378-90. doi:10.1016/S2468-2667(22)00057-3
15 Zhang Y, Morgan RL, Alonso-Coello P, et al. A systematic review of how patients value COPD outcomes. *Eur Respir J* 2018;52:1800222. doi:10.1183/13993003.00222-2018
16 Etxeandia-Ikobaltzeta I, Zhang Y, Brundisini F, et al. Patient values and preferences regarding VTE disease: a systematic review to inform American Society of Hematology guidelines. *Blood Adv* 2020;4:953-68. doi:10.1182/bloodadvances.2019000462
17 Alonso-Coello P, Schünemann HJ, Moberg J, et al, GRADE Working Group. GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 1: Introduction. *BMJ* 2016;353:i2016. doi:10.1136/bmj.i2016
18 Conrad S, Kaiser L, Kallenbach M, Meerpohl J, Morche J. [GRADE: Evidence to Decision (EtD) frameworks - a systematic and transparent approach to making well informed healthcare choices. 2: Clinical guidelines]. *Z Evid Fortbild Qual Gesundhwes* 2019;140:63-73. doi:10.1016/j.zefq.2019.02.006
19 Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ* 2019;366:l4898. doi:10.1136/bmj.l4898
20 Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ* 2016;355:i4919. doi:10.1136/bmj.i4919
21 Morgan RL, Thayer KA, Santesso N, et al, GRADE Working Group. A risk of bias instrument for non-randomized studies of exposures: A users' guide to its application in the context of GRADE. *Environ Int* 2019;122:168-84. doi:10.1016/j.envint.2018.11.004

22 Higgins JPT, Morgan RL, Rooney AA, et al. A tool to assess risk of bias in non-randomized follow-up studies of exposure effects (ROBINS-E). *Environ Int* 2024;186:108602. doi:10.1016/j.envint.2024.108602
23 Wells GA, Shea B, O'Connell D, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. 2000.
24 Joanna Briggs Institute. The Joanna Briggs Institute critical appraisal tools for use in JBI systematic reviews checklist for analytical cross sectional studies. 2017.
25 Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Qual Life Res* 2010;19:539-49. doi:10.1007/s11136-010-9606-8
26 Yepes-Nuñez JJ, Zhang Y, Xie F, 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. doi:10.1016/j.jclinepi.2017.04.019
27 Creswell JW, Clark VLP. *Designing and conducting mixed methods research*. Sage publications, 2017.
28 Helmer -Hirschberg O. Analysis of the future: The Delphi method. Rand, 1967.
29 Nasa P, Jain R, Juneja D. Delphi methodology in healthcare research: How to decide its appropriateness. *World J Methodol* 2021;11:116-29. doi:10.5662/wjm.v11.i4.116
30 Murphy MK, Black NA, Lamping DL, et al. Consensus development methods, and their use in clinical guideline development. *Health Technol Assess* 1998;2:i-iv, 1-88. doi:10.3310/hta2030
31 Siebert U, Alagoz O, Bayoumi AM, et al, ISPOR-SMDM Modeling Good Research Practices Task Force. State-transition modeling: a report of the ISPOR-SMDM modeling good research practices task force-3. *Value Health* 2012;15:812-20. doi:10.1016/j.jval.2012.06.014
32 Siebert U. *When should decision-analytic modeling be used in the economic evaluation of health care?* Springer, 2003:143-50.
33 Kuehne F, Arvandi M, Hess LM, et al. Causal analyses with target trial emulation for real-world evidence removed large self-inflicted biases: systematic bias assessment of ovarian cancer treatment effectiveness. *J Clin Epidemiol* 2022;152:269-80. doi:10.1016/j.jclinepi.2022.10.005
34 Goossens LM, Nivens MC, Sachs P, Monz BU, Rutten-van Mölken MP. Is the EQ-5D responsive to recovery from a moderate COPD exacerbation?*Respir Med* 2011;105:1195-202. doi:10.1016/j.rmed.2011.02.018
35 Soekhai V, Whichello C, Levitan B, et al. Methods for exploring and eliciting patient preferences in the medical product lifecycle: a literature review. *Drug Discov Today* 2019;24:1324-31. doi:10.1016/j.drudis.2019.05.001
36 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;32:667-77. doi:10.1177/0272989X12454577

Web appendix 1: Supplementary materials
Web appendix 2: Appendix tables