University of Groningen

A Checklist for Reporting Valuation Studies of Multi-Attribute Utility-Based Instruments (CREATE)

Xie, Feng; Pickard, A. Simon; Krabbe, Paul F. M.; Revicki, Dennis; Viney, Rosalie; Devlin, Nancy; Feeny, David

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Link to publication in University of Groningen/UMCG research database

CrossMark

ORIGINAL RESEARCH ARTICLE

A Checklist for Reporting Valuation Studies of Multi-Attribute Utility-Based Instruments (CREATE)

Feng Xie[1,2,3] · A. Simon Pickard[4] · Paul F. M. Krabbe[5] · Dennis Revicki[6] ·
Rosalie Viney[7] · Nancy Devlin[8] · David Feeny[9]

Abstract Multi-attribute utility-based instruments (MAUIs) assess health status and provide an index score on the full health-dead scale, and are widely used to support reimbursement decisions for new healthcare interventions worldwide. A valuation study is a key part of the development of MAUIs, with the primary goal of developing a scoring algorithm through eliciting societal preferences. We developed the 21-item Checklist for REporting VAluaTion StudiEs (CREATE) by following a modified two-round Delphi panel approach plus an email survey. CREATE is intended to promote good reporting practice as well as guiding developers to thoroughly and carefully think through key methodological elements in designing valuation studies.

Key Points

Good reporting standards assist developers and users in critically appraising studies and improving the reproducibility of published results.

The development of value sets for multi-attribute utility instruments (MAUIs) has proliferated in recent years, with many countries seeking to support local reimbursement decision making with their own societal preference-based valuation studies.

Guidance is lacking on the key components that should be reported in a valuation study.

This study describes the key elements that should be reported for valuation studies of MAUIs—the Checklist for REporting VAluaTion StudiEs (CREATE).

✉ Feng Xie
fengxie@mcmaster.ca

[1] Department of Clinical Epidemiology and Biostatistics, McMaster University, 50 Charlton Ave E, Hamilton, ON L8N 4A6, Canada

[2] Father Sean O'Sullivan Research Center, St. Joseph's Healthcare Hamilton, Hamilton, ON, Canada

[3] Program for Health Economics and Outcome Measures (PHENOM), Hamilton, ON, Canada

[4] Department of Pharmacy Systems, Outcomes and Research, Center for Pharmacoepidemiology and Pharmacoeconomic Research, University of Illinois at Chicago, Chicago, IL, USA

[5] University of Groningen, University Medical Center Groningen, Department of Epidemiology, Groningen, The Netherlands

[6] Outcomes Research, Evidera Inc., Bethesda, MD, USA

[7] Centre for Health Economics Research and Evaluation, University of Technology Sydney, Sydney, NSW, Australia

[8] Office of Health Economics, London, UK

[9] Department of Economics, McMaster University, Hamilton, ON, Canada

1 Introduction

Multi-attribute utility-based instruments (MAUIs) assess health status, typically with a small number of questions that describe health, and facilitate the ability to generate a preference-based summary score (i.e. value or utility, anchored at zero for dead and one for full health, hereafter collectively referred to as 'index value') for each health state the MAUI defines. The development of an MAUI involves two major steps: the development of a descriptive system (i.e. health-state classification), followed by a valuation study in which the preference weights are obtained from a representative sample of the general population. The latter typically involves elicitation of preferences for a subset of health states that are described by the MAUI from each respondent, followed by estimation of a scoring algorithm that allows for the generation of an index value for every health state described by the MAUI (i.e. the value set).

These index values are applied in economic evaluations of healthcare interventions to calculate quality-adjusted life-years (QALYs). A few established MAUIs include the Health Utilities Index [1, 2], the Short-Form (SF)-6D [3], the EQ-5D [4, 5], and the Assessment of Quality of Life (AQoL)-8D [6].

While only a handful of MAUIs are prominent in the literature, many country-specific valuation studies have been conducted for most of these MAUIs as health technology assessment guidelines often suggest local societal preferences be applied to economic evaluations of healthcare interventions to inform resource allocation decision making [7–10]. A valuation study of an MAUI is aimed at developing a scoring algorithm that assigns values to all possible health states for that instrument. Conducting a valuation study is methodologically challenging, time consuming, and resource intensive. However, once a value set is developed, MAUIs are relatively easy to administer and score, and have the advantage of providing comparable valuations across different diseases and interventions. Furthermore, as MAUIs impose modest respondent burden relative to psychometrically-derived measures of health that typically contain multi-scale and multi-item for each scale, MAUIs are increasingly used for a variety of purposes beyond economic evaluations. They provide a profile and single summary score, a succinct indicator of health-related quality of life (HRQL) that can be employed in a clinical context or for monitoring population health [11–14].

A valuation study involves many considerations, requiring thoughtful planning around its design, implementation, and analysis. Many options for eliciting and modeling valuation data are available, and methods are still under development, for which there is no gold standard that is accepted unanimously by the scientific community.

Rather, it is important for peer reviewers and potential users of these value sets to be able to identify and critically appraise more important aspects of study design and methods that can affect the resulting valuations and, ultimately, whether the values provide a defensible and valid basis for healthcare decision making. Thus, it is important that all relevant details should be disclosed in a manuscript by the authors/developers of value sets. Upon reporting key elements to the informed user, the potential user can then determine whether the value set is of sufficient methodological quality and well-suited to the intended use. A recent review of EQ-5D valuation studies revealed that the reporting quality varied across studies, and highlighted the potential value of a checklist or guidelines to assist developers in reporting valuation studies of MAUIs, and users to assess them [15]. Therefore, our objective was to develop a checklist for valuation studies of MAUIs, the Checklist for REporting VAluaTion StudiEs (CREATE).

2 Methods

2.1 Checklist for REporting VAluaTion StudiEs (CREATE) Expert Panel

The development of CREATE was initiated by Feng Xie and A. Simon Pickard, who then enlisted an expert panel to provide inputs and guidance on the checklist. The panel consisted of five members: Dr. Nancy Devlin, Office of Health Economics, UK; Dr. Paul Krabbe, University of Groningen, The Netherlands; Dr. Rosalie Viney, University of Technology Sydney, Australia; Dr. Dennis Revicki, Evidera Inc., US; and Dr. David Feeny, Department of Economics, McMaster University, Canada. Members of the panel were selected based on their longstanding academic expertise in health utility measurement and HRQL research, and in the estimation of multi-attribute utility functions. This panel consists of a diverse group of scientists who have an interest in health preference-based research and have conducted extensive research on existing MAUIs such as the EQ-5D, SF-6D, Health Utilities Index (HUI), Quality of Well-Being, and AQoL.

2.2 CREATE Development

We followed the international reporting guideline development framework [16], where a review was conducted to identify any published reporting guidelines related to MAUI valuation studies in general. No such guideline was identified in the public domain. A list of 35 items was initially compiled based on a systematic literature review of EQ-5D valuation studies [15]. Although that list was

specifically developed for the EQ-5D valuation studies, many of those items were applicable to other MAUIs, with a simple rephrase. In addition, major themes that represented components of the development of value sets were identified to organize the items.

We adopted a modified Delphi panel approach, asking the expert panel via email to independently assess the content validity, completeness, and wording of these 35 items, and suggest any additional items if needed. Upon receiving input from the expert panel, items were refined and several items dropped, resulting in a checklist of 26 items. Inputs on the 26 candidate items were solicited from members of the EuroQol Research Foundation through an email survey, and a deliberation took place while reviewing all the responses. Members of the EuroQol Research Foundation include developers of the EQ-5D, SF-6D, and 15D. The Group is comprised of approximately 70 researchers from across the globe whose research interests include health utility measurement, many of whom have conducted studies related to value set development. Each participant was asked to assess how important each item was, as included in a reporting checklist for valuation studies, by indicating it as 'required' (defined as essential to understand and evaluate the methodological rigor of the study), 'recommended' (defined as helpful to understand and evaluate the rigor), or 'optional' (defined as not necessary/applicable but may be useful if applicable). If an item was indicated as 'required' by more than 50 % of participants in the survey, the item was included in the final deliberation. Participants were also invited to suggest or drop items and refine item wording. The second round of the Delphi panel focused on reviewing all inputs and finalizing the checklist.

3 Results

After an initial survey invitation and two email reminders, a total of 16 members of the EuroQol Research Foundation responded to the survey; 50 % of respondents were female. Participants came from eight countries in Europe, North America, and Australia, with 87.5 % working in academia and 12.5 % in industry.

A total of 22 items were rated as 'required' by more than 50 % of survey participants. The consensus on these items was high. The number of items with a corresponding percentage of participants who considered these items as 'required' was 19 by more than 75 % of participants, 17 by more than 80 % of participants, 11 by more than 90 % of participants, and 7 by more than 100 % of participants. The item 'the attributes and levels of the health states being valued are described' was rated as 'required' by 80 % of participants. During deliberation we felt that it might not be

necessary or feasible to report this item in the main text of the paper, especially for an MAUI with a large number of possible health states defined. A neighboring item, 'the approach to selecting health states to be valued directly is explained', could provide sufficient detail to allow for replicating the health state selection process and then generating the full list of health states included in the valuation study. Therefore, we decided that the item 'the attributes and levels of the health states being valued are described' be dropped from the final list.

The final CREATE consists of a total of 21 items (Table 1), grouped into seven sections: (1) descriptive system; (2) health states valued; (3) sampling; (4) preference data collection; (5) study sample; (6) modeling; and (7) scoring algorithm. For each item, a brief explanation for inclusion of the item is provided, followed by an example. The examples were selected primarily from those published valuation studies with a higher score, using the original version of CREATE [15] or suggestions by the panelists.

Item 1: The Attributes of the Instrument are Described

Explanation

The descriptive system of an MAUI needs to be developed and validated before a valuation study can be carried out. 'Attribute' is used as a general term that is synonymous with health dimension or domain. The number of attributes and the content covered by each attribute should be described.

Example

"The EQ-SD descriptive system consists of 5 dimensions: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression" [17].

Item 2: The Number of Levels in Each Attribute of the Instrument is Described

Explanation

To understand the richness of the descriptive system and issues related to study design, the number of levels in each attribute should be described, in addition to the description of each attribute.

Example

The EQ-5D-5L consists of five dimensions, with each dimension distinguishing five levels of problems, e.g. no, slight, moderate, severe, and extreme problems.

Item 3: The Approach to Selecting Health States to be Valued Directly is Explained

Explanation

Saturation studies, in which all health states described by an MAUI are valued, are generally not feasible. Therefore,

Table 1 CREATE items

Item no.	Section/item	Yes	No
	Descriptive system		
1	The attributes of the instrument are described	☐	☐
2	The number of levels in each attribute of the instrument is described	☐	☐
	Health states valued		
3	The approach to selecting health states to be valued directly is explained	☐	☐
4	The number of health states valued per respondent is stated	☐	☐
5	Method(s) of assigning the health states to respondents are stated	☐	☐
	Sampling		
6	Sample size/power calculations are stated and rationalized	☐	☐
7	Target population is described	☐	☐
8	Sampling method is stated and rationalized	☐	☐
9	Recruitment strategies are described	☐	☐
10	Response rate is reported	☐	☐
	Preference data collection		
11	Mode of data collection is stated	☐	☐
12	Preference elicitation technique(s) are described	☐	☐
	Study sample		
13	Reasons for excluding any respondents or observations are provided	☐	☐
14	Characteristics of respondents included in the analysis are described	☐	☐
	Modeling		
15	The dependent variable for each model is stated	☐	☐
16	Independent variables for each model are explained	☐	☐
17	Model specifications are provided	☐	☐
18	Model estimators are described	☐	☐
19	Goodness-of-fit statistics for each model are reported	☐	☐
	Scoring algorithm		
20	Criteria for selecting the preferred model are stated	☐	☐
21	The scoring algorithm is presented	☐	☐

CREATE Checklist for REporting VAluaTion StudiEs

a subset of health states needs to be selected for valuation. The subset of selected health states should be carefully considered and the selection process described so that the statistical efficiency and appropriateness of those health states and implied functional forms can be evaluated.

Example

"The minimum sample of health states for an additive model was identified using an orthogonal design…which generated 49 health states (out of 18,000). Alternatively, it was desirable to value more states to allow for more complex specifications (allowing for interaction terms). A stratified sampling method (selecting from health states categorized as mild, moderate, severe) was used to supplement the 49 states with a further 200 states, to provide 249 health states for valuation" [3].

Item 4: The Number of Health States Valued per Respondent is Stated

Explanation

When many health states are selected for valuation, the respondent burden could become excessive. It is a common practice that each respondent is required to value a subset of the selected health states. Therefore, the paper should be explicit as to how many health states were assigned to each respondent.

Example

"All stages used a single set of 45 health states, with each health state described on a separate card. Only 15 health states/cards were used with each respondent" [17].

Item 5: Method(s) of Assigning the Health States to Respondents are Stated

Explanation

Further to how many health states were assigned, the methods used to assign the health states should be described so that the concerns regarding framing effects that may bias the estimates can be mitigated (e.g. randomly assigned as blocks of health states to each respondent).

Example

"Each respondent was randomly assigned to one of five groups, which valued 13 health states plus 'immediate death' and 'unconscious', as described in Table 1" [17].

Item 6: Sample Size/Power Calculations are Stated and Rationalized

Explanation

Sample size calculations are useful to understand the basis for the number of respondents and whether or not the study was adequately powered after study completion.

Example

"Sample size calculations were based on the estimated number of respondents needed to perform comparisons among the major racial/ethnic groups in the United States. These indicated that 4000 completed interviews would be needed to detect a between-groups difference in mean TTO valuations of 0.07, with a power of 80 % and probability of type I error of 0.05. Observed differences between groups in previous studies suggested that a 7–10 % difference in valuations was important" [17].

Item 7: Target Population is Described

Explanation

It is important to describe the target population from whom the value set is developed. There are different views on who is the target population. A popular view is that the target population is the population that could potentially be affected by healthcare resource allocation decisions, and thus is the source of societal preference to inform such decision making. It is important that researchers and users of the MAUI find out the view on the target population adopted by the decision makers in their countries.

Example

"The target population for the study comprised the roughly 210 million civilian noninstitutionalized English- and Spanish-speaking adults, aged 18 and older, who resided in the United States (50 states plus the District of Columbia) in 2002" [17].

Item 8: Sampling Method is Stated and Rationalized

Explanation

To develop a value set based on population health preference, it is important to recruit a representative sample from the target population. A proper sampling method is the first key step to achieving this goal. It determines to what extent the study sample could represent the target population. Therefore, a clear and explicit description and justification of the sampling method are needed.

Example

"A multistage probability sample was selected from the target population using a sampling frame based on residential mailing lists and Census demographic data. The 2 largest minority groups in the United States, Hispanics and non-Hispanic blacks, were oversampled to ensure adequate numbers of minority respondents" [17].

Item 9: Recruitment Strategies are Described

Explanation

Recruitment strategies are aimed at getting a representative sample, as specified by the sampling method. These strategies must be clearly described in order to understand the population basis on which the study sample is recruited and any potential selection biases which may affect representativeness.

Example

"In Canada, participants were recruited by random cold phone call in 2 multi-ethnic cities: Hamilton and Montreal. English was used as the survey language in Hamilton, whereas French was used in Montreal" [18].

Item 10: Response Rate is Reported

Explanation

Response rate reflects not only feasibility but also generalizability of the recruitment strategy. A poor response rate may indicate poor generalizability or study implementation issues. In contrast, a high response rate suggests that the response generalizes well to the stated target population. If available, a comparison between respondents and non-respondents in terms of demographics could also be helpful in further understanding of generalizability.

Example

"A total of 504 interviews were completed, representing 65 % of eligible respondents who could be contacted (from an initial random sample of 3000 households in the City of Hamilton, ON, Canada)" [19].

Item 11: Mode of Data Collection is Stated

Explanation

There are a few modes of data collection that have been used in MAUI valuation studies. Different modes of data collection may be associated with not just advantages but also potential biases, and no data collection method is superior to others. Therefore, it is important to understand which method was used, and whether quality assurance was implemented to ensure data quality.

Example

"A trained professional interviewer visited the respondent in their home. All interviews were audiotape recorded and a 10 % random sample of interviews were reviewed for quality control" [19]

Item 12: Preference Elicitation Technique(s) are Described

Explanation

Preference elicitation technique refers to a procedure of estimating ordinal or cardinal preference for a health state or multiple health states. There are several commonly employed preference elicitation techniques, for instance, matching versus choice-based indifference search procedure [20]. For the same technique there are often variations; for example, there are several approaches for duration and routing process presented to respondents in time trade-off (TTO) [21] and standard gamble choice-based assessments, and whether there is any imposed boundary in value by design. Similarly, there are several variants of discrete choice experiments (DCEs), including forced choice, best-worst, and inclusion of an immediate death state. These details should be described to allow for replication.

Example

"Each respondent was instructed to put perfect health at the top (100) of the FT. Each respondent was then instructed to decide which state, the most disabled or dead, was the least preferred, and place that state at the bottom (0) of the FT. FT tasks on the left-hand side of the FT board were completed by rating the most disabled state or dead, whichever was most preferred, and then the three marker states. States rated on the left-hand side of the FT board remained in place during subsequent waves of ratings on the right-hand side of FT board" [19].

Item 13: Reasons for Excluding Any Respondents or Observations are Provided

Explanation

Health preference elicitation is a complex, cognitively demanding task. Respondents may not be able to understand fully the task or the task may be designed in a way that is difficult to engage respondents. It is therefore expected that some responses may not meet predefined criteria and thus are deemed inconsistent. There is no standard on defining and handling inconsistent responses. Including or excluding inconsistent responses is also an arbitrary decision and subject to debate, but needs to be explicitly stated. A sensitivity analysis of the resulting models to the excluded responses could also be considered.

Example

"Similar to the MVH and other studies, respondents were excluded from the valuation sample if all health states were given the same TTO value or if all health states were valued worse than death. A number of other criteria were applied to exclude respondents from the Valuation Sample" [17].

Item 14: Characteristics of Respondents Included in the Analysis are Described

Explanation

Describing the characteristics of respondents who were included in the analysis has almost become a standard first step in reporting any study involving human subjects. It is especially so in the context where the societal preference from a representative sample of the general public is a recommended source of health utility measurement. In addition, it is recommended to describe the corresponding characteristics of the general public from which the study sample is drawn for the purpose of assessing 'representativeness'. If any respondents were excluded from the analysis, it was also recommended to describe the characteristics of the excluded sample and how they differ from the included sample.

Example

See the table of characteristics of respondents for the EQ-5D US valuation study [17].

Item 15: The Dependent Variable for Each Model is Stated

Explanation

A core step in the data analysis of a valuation study is to develop a model to predict utilities for all health states defined by an MAUI from observed utilities of the selected states. It needs to be explicitly stated what the dependent variable is in each model. Dependent variables of the models could take different formats, depending on the elicitation techniques. For TTO or SG, the elicited utility is commonly used as the dependent variable in these models. Notably, there are different ways to present the elicited

utility; for example, with or without transformation of raw scores (e.g. for worse than dead states), disutility versus utility, or decrement from full health or best possible health state defined by the MAUI. For DCEs, the dependent variable would be the stated choices.

Example

"Adjusted TTO score *h* of each health state by each respondent was subtracted from 1, and then these were regressed to 11 dummy variables pertaining to the health state evaluated" [22].

Item 16: Independent Variables for Each Model are Explained

Explanation

The choice of independent variables can be viewed as an exercise in search of a reasonable presentation of the descriptive system of an MAUI, in part depending on the choice of functional form (see Item 17). This is one of the most challenging parts in the modeling due to the lack of an agreed theoretical basis on which the descriptive system should be presented. Unfortunately, this has also become a neglected part in many valuation studies. Since the first use of the dummy variables to represent the levels of impairment in the UK EQ-5D valuation study, almost all subsequent countries have simply used similar dummy variables as the core list of independent variables in the modeling. Despite being intuitive and easily interpretable, this approach to presenting the MAUI represents only one of many possibilities and should not be used as the standard. Researchers are encouraged to explore various ways of presenting the MAUI's descriptive system in modeling.

Example

"Where x_{dl} represents ten dummy variables that indicate the presence of either a level 2 or level 3 in a given dimension of the evaluated state. In other words, *d* stands for the dimensions: M for mobility, SC for self-care, UA for usual activities, PD for pain or discomfort, AD for anxiety or depression; and *l* stands for either level 2 or level 3. Since the objective of the exercise is to estimate a function that maps the five-digit description to average TTO, these ten x_{dl} dummy variables form the core of the regression" [22].

Item 17: Model Specifications are Provided

Explanation

A model specification is a process of selecting an appropriate functional form and choosing variables to be included. Three major functional forms are additive (used for the estimation of many EQ-5D scoring algorithms [15] and SF-6D [3]), multiplicative (used for HUI [2] and the AQoL-8D [23]), and multi-linear [24]. Once the dependent and independent variables are defined, there are a number of possible model specifications. Therefore, it is important to describe the specification for each model. Often an equation, supplemented by text description, could be an efficient and straightforward way of presenting this information. If any specification test is used, that should also be stated.

Example

"Thus, the regression equation is as follows: $Y = \alpha + \beta_1 MO + \beta_2 SC + \beta_3 UA + \beta_4 PD + \beta_5 AD + \beta_6 M2 + \beta_7 S2 + \beta_8 U2 + \beta_9 P2 + \beta_{10} A2 + \beta_{11} N3$ i.e. TTO scores were explained by 12 independent variables: two variables for each dimension (one to represent the move from level 1 to level 2 and one to represent the move from level 2 to level 3), a term that picks up whether any dimension is at level 3, and an intercept" [25].

Item 18: Model Estimators are Described

Explanation

Model estimators refer to methods used to estimate the coefficients of interest, often in a regression model. Different estimators (e.g. ordinary least square, fixed effects, and random effects) lead to different coefficient estimates. It is important to explicitly describe the estimators of the model. Often each respondent is asked to complete the valuation task for multiple health states. It is important to explicitly state whether the estimator selection takes account of repeated measures.

Example

"Since each respondent was expected to have a different pattern of response, for example, to offer higher or lower values than the average persistently across all health states, a random effects (RE) estimation or a fixed effects (FE) estimation may be used as estimation methods. Therefore, a series of preliminary analyses was carried out to compare the simple ordinary least-squares (OLS) regressions with RE and FE regressions" [22].

Item 19: Goodness of Fit Statistics for Each Model are Reported

Explanation

Goodness of fit is a measure of how well observed data fit a regression model. Some statistics are commonly used to measure goodness of fit for regression models; for example, the coefficient of determination (commonly denoted R^2), and Akaike Information Criterion. Moreover, it is important to measure the goodness of fit by assessing the discrepancy between observed and predicted utilities by the model since the goal is to predict utilities for all health

states. Mean absolute error (MAE, or sometimes called mean absolute difference or mean absolute deviation) is often used for this purpose. Two approaches aimed at maximizing external validity of the model prediction are commonly used to estimate MAEs in valuation studies. One approach is to split the whole study sample into two subsamples: one for the model development and the other for the estimation of MAEs. Another approach is to calculate the MAE by excluding, in turn, each health state included in the valuation study from the modeling. Since the goal is to predict utilities of all health states (which include those not directly included in the valuation), the second approach is considered a more appropriate indicator of the predictive ability of a model. A variation on the second approach is an assessment of the agreement between directly measured out-of-sample scores (scores not used to estimate the scoring function for the MAUI) and scores generated by the scoring algorithm of the MAUI [19].

Example

"The R^2 of 0.46 (in both cases) was very high given the type of (cross-sectional) data analyzed here,…". "It can be seen that the predictive power of the model remained high; only five states had a predicted value that was more than 0.1 different from the actual value, and the mean absolute difference was again below 0.05. Table 5 shows the predicted value for each state when direct values for that state are excluded from the modeling. Only three states had a predicted value that was more than 0.1 different from its actual value: the biggest difference was for the most extreme state (ie, 33333), which has a predicted value that is 0.173 below its actual value. The mean absolute difference was once again below 0.05" [25].

Item 20: Criteria for Selecting the Preferred Model are Stated

Explanation

For any valuation study, an extensive modeling exercise with different specifications and estimators is always necessary before arriving at a best-performed model for the value set development. Given multiple indicators could be used to compare the performance of alternative models, it is extremely important to describe clearly and explicitly the criteria used for selecting the best model. Consistency in predicted utilities between health states (sometimes referred to as 'face validity') is intuitive and commonly used as the first criterion. In addition, multiple goodness-of-fit statistics might be available for each model. It is necessary to state the order in which these criteria are compared. This description makes the model selection process transparent and justified.

Example

"To make a choice between different ways of representing the relationship between the valuations of EuroQol health states and the different dimensions and levels, the model that ultimately was chosen had to predict a higher score for one state, A, than for another, B, if A was logically better than B on at least one dimension and no worse on any other dimension. In choosing between the many models that satisfy this consistency condition, the one that best explained the differences in the valuations given to those states on which there was direct data was chosen. For models with comparable goodness-of-fit statistics, the ultimate choice was made according to parsimony, ie, the simplest model (both in terms of the number of independent variables and the ability to explain them) was chosen. The results presented below are from the "best" model according to these criteria" [25].

Item 21: The Scoring Algorithm is Presented

Explanation

After a comparison between alternative models using pre-specified criteria, it is very likely that a best-performed model can be selected. For the model selected as the scoring algorithm, full details, including coefficient estimates and corresponding standard errors (or 95 % confidence intervals), should be provided. Ideally, the full value set (i.e. mean prediction and corresponding standard error or 95 % confidence intervals) derived from this scoring algorithm is attached to the paper; however, given the space constraint, especially in peer-reviewed journals, this may not be feasible. Therefore, it is recommended to provide an example of how the utility for a health state can be calculated using the scoring algorithm, accompanied by, for instance, a website link where the full value set can be accessed and via different formats (e.g. in SAS, R, or STATA syntax codes).

Example

"Hence, the predicted value for state 11223 is 1.000-0.000-0.000-0.140-0.173-0.450-(−0.280)-0.011-0.000-0.000 = 0.506" [17].

4 Discussion

MAUIs have gained widespread popularity in clinical and economic research [26]. Given their potential impact on reimbursement decision for new healthcare interventions, such as innovative but expensive pharmaceuticals, MAUI-based valuation studies should be reported in a transparent and adequate manner so that the methodological rigor and

judgment of the developers can be appraised. The CRE-ATE is intended to promote good reporting practice and to guide developers through key elements in study design and methodology.

Preference-based instruments differ from non-preference-based in how they are scored and who the potential users are. A scoring algorithm for preference-based instruments needs to be based on health preferences directly elicited from the target population, which is much more complex than the psychometrically-derived scoring for non-preference-based instruments which typically assume equal weights across scales and items. Expertise in health economics, health preference measurement, and econometric modeling is essential to ensure the quality of valuation studies. In contrast, healthcare professionals and policy makers who are users of MAUIs may have little or no expertise in these fields. The role of MAUIs in healthcare resource allocation decision making implies that the potential impact of using MAUIs is substantial, albeit implicit, and beyond individual research findings. For example, if a value set overestimated average distance between health states, the incremental cost-effectiveness ratio would be smaller (thus favoring the intervention under evaluation), everything else being equal. This could be translated into an inefficient use of scarce healthcare resource at the societal level. Furthermore, there is potential for gaming of the choice of MAUIs and valuation algorithms by sponsors seeking reimbursement of new interventions, and it is important for those involved in evaluating such decisions to be able to identify any possible sources of bias. These differences build up barriers to communication, appraisal, and use of valuation studies among developers and users and, on the other hand, highlight the need to enhance the reporting quality for these types of studies.

The CREATE is a methodology-oriented reporting checklist, due to the characteristics of MAUIs and the nature of valuation studies. The components of the checklist are presented in a manner slightly different from what is commonly seen in other reporting checklists (i.e. title, introduction, methods, results, and discussion) [27, 28]. Those items typically seen have been either implicitly embedded in the valuation study itself or covered by other items in the checklist. For example, the title and objectives must be clearly presented in order to be considered as an MAUI valuation study (i.e. to develop a value set for an instrument for a specific target population). In addition, it is important to provide sufficient information about the MAUI itself to allow for a clear understanding of the instrument under valuation. Two items under 'descriptive system' cover what and how the introduction section should be reported. We have elaborated the justification for this presentation in the discussion. We feel this way of presentation is more consistent with the order in which

components of MAUI valuation studies are typically reported.

More recently, there is interest in developing disease-specific utility-based HRQL instruments [29, 30] or applying utility-based approaches to existing disease-specific instruments [31, 32]. This research endeavor probably arises in response to the criticism of MAUIs lacking sufficient sensitivity in measuring the condition-specific impact on HRQL [33–35]. The CREATE should be conducive to assessment of the reporting quality of valuation studies for disease-specific as well as generic HRQL instruments.

There are a few limitations with the development of CREATE. First, there were a relatively small number of participants in the survey, and all were members of the EuroQol Group. Nevertheless, research interest, expertise, and experience of these participants are beyond just the EQ-5D. Development of CREATE is an ongoing process and, in future, we plan to involve more researchers outside the EuroQol Group. Second, we used categorical responses and arbitrarily chose the majority rule instead of commonly used ranking and mean scores. The main reason for this consideration was to minimize the impact of potentially large discrepancies in responses among participants which cannot be revealed through a mean score. Given the high consensus on the final items, the results would remain the same should we assign a score to each category and compare the mean scores. Regardless of which method is used, any threshold may be inevitable and often arbitrary. Lastly, it is important to note that the CREATE is developed based on the current theory and methodology in health preference measures and its application in developing a scoring algorithm for MAUIs. Future advances in measurement concept and theory may indicate that an ongoing update on the CREATE is necessary.

5 Conclusion

The CREATE is aimed at facilitating and promoting transparent reporting for valuation studies of MAUIs. This checklist is methodology-oriented and can assist users in their critical appraisal of value sets and help guide research related to the design, execution, and reporting of health valuation studies.

Rivero-Arias, Luciana Scalone, Elly Stolk, Caitlyn Solem, Aki Tsuchiya, Matthijs Versteegh, and Rosalie Viney.

References

1. Torrance GW, Furlong W, Feeny D, Boyle M. Multi-attribute preference functions. Health Utilities Index. Pharmacoeconomics. 1995;7:503–20.
2. Torrance GW, Feeny DH, Furlong WJ, Barr RD, Zhang Y, Wang Q. Multiattribute utility function for a comprehensive health status classification system. Health Utilities Index Mark 2. Med Care. 1996;34:702–22.
3. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. J Health Econ. 2002;21:271–92.
4. Brooks R. EuroQol: the current state of play. Health Policy. 1996;37:53–72.
5. Herdman M, Gudex C, Lloyd A, Janssen M, Kind P, Parkin D, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Qual Life Res. 2011;20:1727–36.
6. Richardson J, Iezzi A, Khan MA, Maxwell A. Validity and reliability of the Assessment of Quality of Life (AQoL)-8D multi-attribute utility instrument. Patient. 2014;7:85–96.
7. Committee Pharmaceutical Benefits Advisory. Guidelines for preparing submissions to the Pharmaceutical Benefits Advisory Committee. Canberra: Australia Commonwealth Department of Health and Ageing; 2006.
8. Ontario Ministry of Health and Long-Term Care. Ontario guidelines for economic analysis of pharmaceutical products. Toronto: Ministry of Health; 1994.
9. National Institute for Clinical Excellence. Guide to the methods of technology appraisal. London: National Institute for Clinical Excellence; 2013.
10. Canadian Agency for Drugs and Technologies in Health. Guidelines for the economic evaluation of health technologies: Canada. 3rd ed. Ottawa: Canadian Agency for Drugs and Technologies in Health; 2006.
11. Standfield L, Norris S, Harvey C, Elliot L, Riordan J, Hall S, et al. Relationship between rheumatoid arthritis disease severity, health-related utility, and resource use in Australian patients: a cross-sectional, multicenter study. Clin Ther. 2010;32:1329–42.
12. Stavem K, Froland SS, Hellum KB. Comparison of preference-based utilities of the 15D, EQ-5D and SF-6D in patients with HIV/AIDS. Qual Life Res. 2005;14:971–80.
13. Stavem K, Bjornaes H, Lossius MI. Properties of the 15D and EQ-5D utility measures in a community sample of people with epilepsy. Epilepsy Res. 2001;44:179–89.
14. Clemens S, Begum N, Harper C, Whitty JA, Scuffham PA. A comparison of EQ-5D-3L population norms in Queensland, Australia, estimated using utility value sets from Australia, the UK and USA. Qual Life Res. 2014;23:2375–81.
15. Xie F, Gaebel K, Perampaladas K, Doble B, Pullenayegum E. Comparing EQ-5D valuation studies: a systematic review and methodological reporting checklist. Med Decis Mak. 2014;34:8–20.
16. Moher D, Schulz KF, Simera I, Altman DG. Guidance for developers of health research reporting guidelines. PLoS Med. 2010;7:e1000217.
17. Shaw JW, Johnson JA, Coons SJ. US valuation of the EQ-5D health states: development and testing of the D1 valuation model. Med Care. 2005;43:203–20.
18. Krabbe PF, Devlin NJ, Stolk EA, Shah KK, Oppe M, van Hout B, et al. Multinational evidence of the applicability and robustness of discrete choice modeling for deriving EQ-5D-5L health-state values. Med Care. 2014;52:935–43.
19. Feeny D, Furlong W, Torrance GW, Goldsmith CH, Zhu Z, DePauw S, et al. Multiattribute and single-attribute utility functions for the health utilities index mark 3 system. Med Care. 2002;40:113–28.
20. Stalmeier PF, Goldstein MK, Holmes AM, Lenert L, Miyamoto J, Stiggelbout AM, et al. What should be reported in a methods section on utility assessment? Med Decis Mak. 2001;21:200–7.
21. Attema AE, Edelaar-Peeters Y, Versteegh MM, Stolk EA. Time trade-off: one methodology, different methods. Eur J Health Econ. 2013;14(Suppl 1):S53–64.
22. Tsuchiya A, Ikeda S, Ikegami N, Nishimura S, Sakai I, Fukuda T, et al. Estimating an EQ-5D population value set: the case of Japan. Health Econ. 2002;11:341–53.
23. Richardson J, Sinha K, Iezzi A, Khan MA. Modelling utility weights for the Assessment of Quality of Life (AQoL)-8D. Qual Life Res. 2014;23:2395–404.
24. Keeney RL. Building models of values. Eur J Oper Res. 1988;37:149–57.
25. Dolan P. Modeling valuations for EuroQol health states. Med Care. 1997;35:1095–108.
26. Brauer CA, Rosen AB, Greenberg D, Neumann PJ. Trends in the measurement of health utilities in published cost-utility analyses. Value Health. 2006;9:213–8.
27. Husereau D, Drummond M, Petrou S, Carswell C, Moher D, Greenberg D, et al. Consolidated health economic evaluation reporting standards (CHEERS) statement. BMJ. 2013;346:f1049.
28. Calvert M, Blazeby J, Altman DG, Revicki DA, Moher D, Brundage MD. Reporting of patient-reported outcomes in randomized trials: the CONSORT PRO extension. JAMA. 2013;309:814–22.
29. Casey R, Tarride JE, Keresteci MA, Torrance GW. The Erectile Function Visual Analog Scale (EF-VAS): a disease-specific

utility instrument for the assessment of erectile function. Can J Urol. 2006;13:3016–25.

30. Wasserman J, Aday LA, Begley CE, Ahn C, Lairson DR. Measuring health state preferences for hemophilia: development of a disease-specific utility instrument. Haemophilia. 2005;11:49–57.

31. Kularatna S, Whitty JA, Johnson NW, Jayasinghe R, Scuffham PA. Development of an EORTC-8D utility algorithm for Sri Lanka. Med Decis Mak. 2015;35(3):361–70.

32. Lamers LM, Uyl-de Groot CA, Buijt I. The use of disease-specific outcome measures in cost-utility analysis: the development of Dutch societal preference weights for the FACT-L scale. Pharmacoeconomics. 2007;25:591–603.

33. Kvam AK, Fayers PM, Wisloff F. Responsiveness and minimal important score differences in quality-of-life questionnaires: a comparison of the EORTC QLQ-C30 cancer-specific questionnaire to the generic utility questionnaires EQ-5D and 15D in patients with multiple myeloma. Eur J Haematol. 2011;87:330–7.

34. Krahn M, Bremner KE, Tomlinson G, Ritvo P, Irvine J, Naglie G. Responsiveness of disease-specific and generic utility instruments in prostate cancer patients. Qual Life Res. 2007;16:509–22.

35. Marra CA, Woolcott JC, Kopec JA, Shojania K, Offer R, Brazier JE, et al. A comparison of generic, indirect utility measures (the HUI2, HUI3, SF-6D, and the EQ-5D) and disease-specific instruments (the RAQoL and the HAQ) in rheumatoid arthritis. Soc Sci Med. 2005;60:1571–82.