Appendix 2: Assessment of methodological quality

To enable critical assessment of patient preference studies across methodologies we followed the approach recommended by Streiner and Norman,[1] and constructed a practical tool. We performed a rapid review and identified seven relevant methodological reviews:

- An overview of techniques used to elicit public preferences on the provision of healthcare and a discussion of their weaknesses and strengths.[2]
- An appraisal of stated preference valuation techniques and criteria for how studies should be designed from the perspective of competition in the market.[3]
- A summary and discussion of approaches to assess and compare the validity of methods in cost-utility analysis.[4]
- A review of methods for economic evaluation of healthcare and a check -list for assessing methods.[5]
- Critical appraisal principles for cost-effectiveness and cost-utility studies in healthcare.[6]
- Critical appraisal criteria for systematic reviews.[7 8]

We identified an initial list of 75 possible items. Two researchers (OE, KN) piloted the items on four methodologically different studies. Based on our experiences on practicality and usefulness, we reduced the number of items to 50 and then applied the tool to 25 stated preference studies. The agreement of the reviewers on the overall quality of the studies was 80 % (20/25). Formal tests on the consistency and construct validity were not performed. Three studies were removed due to not meeting inclusion criteria. After the initial review we simplified the checklist to 31 items within five categories. The final version of the tool is presented below.

1. What is the external validity of the study? (High/medium/low)

The population included in the study should be representative of patients with the relevant condition.

1. Is the patient population in the study clearly described?
    a. Disease or condition
    b. Age group(s)
    c. Gender distribution
    d. Ethnicity
    e. Socio-economic status
    f. Inclusion and exclusion criteria
    g. Clinical setting
    h. Common comorbidities
2. What was the sample size and is it robust? No guidelines have been established for sample size calculations for stated preference studies. For simple surveys a minimum total sample size of 400 and a size of each cell of interest of at least 75 has been recommended.[5]
3. Did the recruitment procedures ensure generalizability?
4. Was randomisation and/or stratification applied to avoid selection bias?
5. Consider the number of participants enrolled in the study and how many participants were included in the analysis. What was the completion rate? Poor completion may cause non-response error because those who do not respond are different from those who do.
    a. Consider how empty cells, such as questions not replied to in a questionnaire, are treated. Are empty cells given zero value, replaced with the mean or excluded from analysis?
    b. Are the reasons for refusals, attritions, withdrawals, exclusions and re-inclusions reported? If refusals are reported, are they representative of the population?
6. If a decision problem was presented, to what extent did the problem take into account the real decision-making context and involve a constrained choice?

2. What is the quality of construct representation in the study? (High/medium/low)

Construct underrepresentation occurs when "a stimulus presented to a judge fails to fully represent the depth and complexity of information required in actual judgments". [4] In preference studies, construct

underrepresentation threatens validity when options, health states, attributes or health state/attribute levels presented for preference elicitation are inadequate, ambiguous, vague, non-meaningful, unrealistic or incomprehensive.

1. Were all important and relevant options, attributes and attribute levels relevant to the main objective of the study identified? If options were presented to participants, were all relevant alternatives included? Important comparators such as non-pharmaceutical alternatives, do-nothing or the opt-out option should not be excluded.
2. Are the sources used in the construction of options, attributes and attribute levels for the preference elicitation accounted for and are they appropriate? Constructs might be based on systematic literature reviews, patient focus groups and clinician interviews.
3. Were options, attributes and attribute levels presented with sufficient detail and accuracy?
    a. Were attribute levels appropriate and plausible to respondents? Attribute levels should neither be too wide nor too narrow.
    b. If time periods were presented to participants, were the periods clearly stated and appropriate?
    c. If costs were presented to participants, were the measurement units clearly described? Are the sources for the costs clearly referenced?
    d. Were consequences that occur in the future discounted when required?

3. To what extent was the risk of construct-irrelevant variance minimised? (To a high /moderate/small degree)

Construct-irrelevant variation threatens the validity of a study when factors irrelevant to preferences influence measurements of utilities.[4] Construct-irrelevant variance may be caused by a number of factors such as impairments in the cognitive abilities of the participants, numeracy skills, emotions and prejudices, and the elicitation procedure.

1. Was the study piloted?
2. Was a pre-test procedure performed to ensure that participants understood all the questions, and/or were post-test diagnostic questions included to explore the extent to which the respondents understood their tasks?
3. Was a "cheap talk" script included? A cheap talk script is a script that explicitly highlights the hypothetical bias problem before participants make a decision.
4. Is there evidence of starting point bias, e.g. resulting from an anchoring on initial stated values?[4]
5. Were the questions neutral in tone? A negative or positive tone in the questions can result in framing effects.
6. If an interviewer was present - what was the possible influence on the respondent´s answers?
7. Is there evidence of high cognitive load, resulting in fatigue and frustration bias?
8. Were tests or parts of tests repeated for different subjects? With repetition, a subject's ability to express his or her preferences can improve, and result in changes in the responses.[4]
9. Is there evidence that emotions, hidden prejudices or reduced cognitive abilities and skills impaired the judgement of the participants?
10. For analogue scaling methods, consider the risk of sequencing effects. For rating scales, consider whether the values obtained influenced the appearance of the scale. For standard gamble and time trade-off, consider how the gamble was framed, the bottom anchor of the gamble, and the procedure used to find a subject´s indifference point. For standard gamble methods, consider the specific probabilities used. For time trade-off, consider the duration of survival in the base case.[4]

4. What is the quality of the reporting and analysis? (High/medium/low)

1. How complete are the outcome data? Are all pre-specified measures reported? Selective reporting such as extensive use of sub-group analysis, use of data from participants with consistent results only, and the deletion of outliers or extreme values can lead to reporting bias.
2. Is there a tendency for the assessments to take a few discrete values? Are the data skewed or normally distributed?

3. What statistical techniques were used and are they appropriate? This aspect of quality should be assessed by a health economist or statistician with experience in the field.
4. Is allowance made for uncertainty in the estimates?
5. Is heterogeneity and patient subgroups analysed when relevant?

5. Do other aspects strengthen or weaken the study? (Strengthen/no difference/weaken)

1. Were formal tests of internal validity performed and what were the results?
2. To what extent are well-defined, answerable research questions stated and answered?
3. Was more than one assessment method used? If so, is the rank ordering of preferences for health states consistent? [4]
4. Do the authors outline how their piece of work compares and adds to the current evidence base? If so, are allowances made for potential differences in study methodology?
5. Is the protocol or supplementary information about the study available? If so, examine the assessment protocol.
6. Which limitations and weaknesses of the study are cited by the authors themselves?
7. Are there other aspects of the study that strengthen or weaken the quality?

6. Based on the above, what is the overall quality of the study? (High/medium/low)

References

1. Streiner DL NG. Health measurement scales: A practical guide to their development and use. 3rd ed. Oxford: Oxford University Press, 1995.
2. Ryan M, Scott DA, Reeves C, et al. Eliciting public preferences for healthcare: a systematic review of techniques. Health technology assessment (Winchester, England) 2001;5(5):1-186.
3. Review of stated preference and willingness to pay methods. London: Accent and RAND Europe, April 2010. URL: http://webarchive.nationalarchives.gov.uk/+/http://www.competition-commission.org.uk/our_role/analysis/summary_and_report_combined.pdf
4. Lenert L, Kaplan RM. Validity and interpretation of preference-based measures of health-related quality of life. *Medical care* 2000;38(9 Suppl):Ii138-50.
5. Drummond MF SM, Torrance GW, O´Brien BJ, Stoddart GL. Methods for the Economic Evaluation of Health Care Programmes. 3rd Ed. ed. New York: Oxford University Press, 2005.
6. Soares M, Dumville JC. Critical appraisal of cost-effectiveness and cost-utility studies in health care. *Evidence-based nursing* 2008;**11**(4):99-102.
7. Higgins JPT, Altman DG. Assessing Risk of Bias in Included Studies. Cochrane Handbook for Systematic Reviews of Interventions: John Wiley & Sons, Ltd, 2008:187-241.
8. CRD AJ. Systematic Reviews: CRD´s guidance for undertaking reviews in health care: Centre for Reviews and Dissemination, University of York, 2009.