Appraisal checklist for patient preferences for treatment outcomes

This appraisal checklist for patient preferences for treatment outcomes (PPTOs) is intended to be used to appraise reported findings concerning PPTOs. The document is set up into two sections: [1] the questions of the appraisal checklist and [2] a background section on how to interpret the items in the checklist in detail.

The six criteria listed in this checklist can help to evaluate methods used to elicit patients’ preferences. The criteria are not meant to be used as a checklist to determine the quality of research on the basis of particular cut-off levels. They should rather be seen as a set of key quality indicators for research: if more criteria are met, the greater the likelihood that a study was adequately performed. In-depth knowledge on the specific methods used is often required to appraise specific aspects or appropriateness of that method, thus additional considerations not captured in this checklist remain important.

Questions 1 and 2 can be used as filter questions: studies not qualifying for those criteria do not need to be appraised any further. The list of appraisal criteria has no overall score nor any weighting system attached to the individual criteria. There is one summarising question to provide a single, overall, evaluation of a study. This checklist is aimed at the appraisal of individual studies; for the appraisal of body of evidence please see the GRADE/CERQual guidelines (http://www.gradeworkinggroup.org/).

|  |  |
| --- | --- |
|  | YesNoDon't knowNot applicable |
| 1. Does the study address relevant patient preferences for treatment outcome? |  |
| 2. Is the description of population, methods, and analysis clear and complete? |  |
| 3. Are the data collection methods appropriate and appropriately used? |  |
| 1. Is the format of included questions appropriate?
 |  |
| 1. Is the chosen mode of application for included questions appropriate?
 |  |
| 4. Are any theories, assumptions or models on which the research is based adequately described? |  |
| 5. Were the methods properly executed and the results reliable and valid? |  |
| 6. Are the results transferable to the target population? |  |

|  |  |
| --- | --- |
| **Overall judgment** | YesNoDon't know |
| Considering your conclusions regarding items 1-6, would you say that claims regarding patient preferences for treatment outcome are sufficiently substantiated and sufficiently relevant to take into account when making recommendations for treatment decisions? Please clarify your main arguments to support your conclusion: |  |
|  |  |

**Background on the appraisal items**

Below a short explanation on each of the items listed above is displayed. This information should help interpret the appraisal items either by providing the criteria that need to apply to the study that is being appraised or by linking to relevant (external) guidance.

1. **Does the study address relevant patient preferences for treatment outcome?**

Screening of any paper starts with the relevancy of the data; if the data is not of relevance for the research, stop here.

1. **Is the description of population, methods, and analysis clear and complete?**

It is argued that no proper appraisal is possible without adequate reporting (Carroll et al., 2012), hence it must be clear to the appraiser what the research entailed in order to appraise the methods used. Transparency of methods is a commonly mentioned appraisal criterion (Ryan et al., 2001; Clark, 2003; Guyatt et al., 2006; Kuper et al., 2008b). This question should be answered positively if the research can be repeated on the basis of the descriptions of the study. A lack of a clear description on its own is not a sign that a study is biased. However an inadequately described study is very difficult to appraise; and without a clear description or the use of non-standardized methods the study may be more prone to bias. The inventory of possible bias and their consequences may inform the usefulness of the data presented in a study. To do so, at least some of the study’s methods should be clear. If the methods are (completely) unclear, stop here.

For qualitative research, appraisal criteria address specific phases of research, that is: data generation, analysis and interpretation (see further down for details on specific methods)(Facey et al., 2010). For all stages of the research it should be described who performed the action and when (Nicholas et al., 2008). Furthermore, an appropriate presentation of findings, detailed description of sample and context are required (Nicholas et al., 2008). The use of standardized methods (e.g. specific computer programs) may make this step of the appraisal easier, as full transparency in qualitative research is often not possible due to various reasons (e.g. size restrictions, confidentiality)(McLellan et al., 2003)**.**

1. **Are the data collection methods appropriate and appropriately used?**

The choice between data collection methods depends on the information required. Though it is rather difficult (especially in the appraisal phase) to determine appropriateness of a certain method, the choice for one should clearly follow from the paper at the very least. Value, reliability and validity of data may be compromised if the wrong method is used. Phellas et al. (2011) described some of the considerations usable to chose between various methods.

* + 1. Is the format of included questions appropriate?

Interviews or questionnaires questions can be either closed (e.g. multiple choice) or open. Closed questions are generally easier to analyse and report (simpler answers) but require pre-planning (e.g. validated questionnaire or questions based on previous research). Open questions are more suitable for exploratory research.

The total number of questions in an interview or questionnaire should not be too high (Passmore et al., 2002; Burns et al., 2008; Streiner, 2008). There is no agreed-upon maximum for the number of questions, as it depends greatly on the target population, type of questions and medium. Longer questionnaires seem to decrease response rates and quality of answers towards the end of the questionnaire (Iglesias & Torgerson, 2000). Longer duration of interviews or discussions may both increase the opportunity to retrieve data as well as tire out respondents.

The contents/format of questionnaires or questions asked in interviews may not always be reported. In these cases, the quality of questions cannot be directly appraised. An indication of the quality of questions may be inferred from the retrieved data (e.g. face validity of results, consistency), although one should be careful doing so – unexpected results may be caused by real effects.

* + 1. Is the chosen mode of application for included questions appropriate?

**For face-to-face interviews**, the interview location should be described and appropriate for the interview (e.g. not too noisy, location which enables a secure feeling for the respondent (such as at their home), enough privacy, etc.)(MacLean et al., 2004; Cook, 2008).The use of transcriptions, tapes and multiple researchers to interpret the data also improve quality (Giacomini et al., 2000; Nicholas et al., 2008). Some consider interview duration (>=1 hour or multiple interviews) a quality criterion (Nicholas et al., 2008). Indications that the interviewer was capable (e.g. shown experience, required language and/or communication skills) may also help to appraise the quality of the interviewer (Ryan et al., 2001; Szolnoki & Hoffmann, 2013).

**Phone-based interviews** are best performed by interviewers experienced in such methods (Szolnoki & Hoffmann, 2013). No specific appraisal criteria were found for phone-based interviews, except a number of considerations: 1) complex issues or questions with many options are easier to answer in face-to-face or paper (or webbased) media; 2) the length of the phone-based interview cannot be as long as for face-to-face interviews; and 3) phone-based interviews make it less likely that the interviewer is affected by characteristics of the interviewee (e.g. clothes). Phone-based interviews tend to produce less in-depth responses, making face-to-face interviews better suited if that information is needed (Irvine, 2011).

**Focus groups** rely on the group dynamics (interactions, alliances, differences and similarities in views, dominant and silent responders or views, agreements and disagreements, consensus, emotions and conflicts) for information, and thus should clearly be described (Kitzinger, 1994; Stevens, 1996; Webb & Kevern, 2001; Rabiee, 2004). Group members should be selected on appropriateness (i.e. have something to say and are willing to share it within a group) rather than a random sample (Rabiee, 2004). Depending on the issues, homogeneity in the focus group may be needed, but this is not required beforehand (Corfman, 1995; Greenwood et al., 2014). In their focus group guidelines, (Onwuegbuzie et al., 2009) state that it may be beneficial for focus groups to have more than one moderator, if the session consists of multiple sessions lasting 1-2 hours, and consists of 6-12 persons.

**If the questions are posed on paper or are web-based**

The quality of the data relies heavily on the answering possibilities (e.g. the possible multiple-choice options, possibility to add comments, or even the size of the boxes in which the write the answer). For multiple choice questions, quality may be influenced by the number (and content) of available options per question: they should not be too few or too many (±7 items is considered to be the maximum (Streiner, 2008; Marsden & Wright, 2010)). Furthermore, the range of options should allow respondents to answer according to their views (Passmore et al., 2002; Burns et al., 2008; Streiner, 2008).

If surveys or questionnaires are used, they should preferably be standardized (pre-existing)(Rattray & Jones, 2007; Streiner, 2008; Edwards, 2010). Otherwise, the reason for not using an existing tool should be clear (McColl et al., 2001; Passmore et al., 2002), and any development of the newly created tool described (Was it pilot-tested, were redundant items removed, was it reviewed by experts (Passmore et al., 2002; Burns et al., 2008)).

Important factors of questionnaires include the (cognitive) burden (length, complicatedness of questions), general questionnaire quality (e.g. can the respondent read and understand the questions, is it clear what is expected), contents (type of questions, open/closed questions, number of options), layout, and the application itself (how and where were participants recruited)? Most of these issues can be best examined if a copy or link to (a part of) the used questionnaire was provided (Kelley et al., 2003). Other factors to look at are statistical measures such as inter-rater agreement or Cronbachs alpha. Note also that online application is more susceptible to bias (Wright, 2005; Cook, 2010).

If the user is not familiar with appraising questionnaires, the following resources may help identifying problems in used questionnaires:

* For questionnaires in general: the overview and checklist for designing and developing questionnaires by Rattray and Jones (2007).
* For health status and quality-of-life instruments: the attributes and review criteria by Aaronson et al. (2002).
* For studies on measurement properties of health status measurement instruments: the COSMIN checklist (http://www.cosmin.nl)

*Items 4 concerns qualitative research. For further information on specific quantitative research, please see item 5.*

1. **Are any theories, assumptions or models on which the research is based adequately described?**

Qualitative data always arises from certain methodological, theoretical and analytical positions: the chosen methods and hypotheses and even execution of methods are based on pre-existing knowledge and can greatly influence the outcome of the research (Malterud, 2001).

The following resources may help identifying possible problems and critical properties of qualitative research and findings:

* CASP (Critical Appraisal Skills Programme) checklist entitled "10 questions to help you make sense of qualitative research" (available at http://www.casp-uk.net/) and other generic checklists for qualitative research.
* The RATS (Relevance of study question, Appropriateness of qualitative method, Transparency of procedures, and Soundness of interpretive approach) (Clark, 2003). Available from: http://www.biomedcentral.com/authors/rats
* Anderson (2010) provides a list of (generic) criteria for appraising qualitative research.
* Carroll et al. (2012) argued appraisal of qualitative research should only target reporting quality of studies and generated a short criteria list.
1. **Were the methods properly executed and were results reliable and valid?**

For some methods to elicit patient preferences, more detailed appraisal criteria are described. These can be found below.

Brazier et al. (1999) reviewed the **Time Tradeoff (TTO)**, **Standard Gamble (SG)**, **Person Trade-off (PTO)**, **Magnitude Estimation (ME)** and **Visual Analogue Scale (VAS)** for use in economic evaluations. They list several points for attention which can be used to appraise studies using one of these (or similar) methods: practicality (completion rate and time, response rate); reliability (inter-rater and test-retest reliability; sample size); content, face and construct validity; choice for valuation technique (see below); quality of data (background characteristics of the population; degree of variation; evidence on understanding of the task; and finally, empirical validity and whether revealed, stated or hypothesized preferences are discussed or shown.

Other considerations specifically for some of these methods are:

* The used method should be appropriate for the subject (i.e. **SG** for choices where risks are involved, **TTO** for choices that influence the chronically, **Willingness to pay (WTP)** for subjects that can ethically and logically be expressed as a monetary value, etc.) (Wakker & Stiggelbout, 1995; Ryan & San Miguel, 2000).
* For **SG** and **TTO** the shown alternative should be appropriately chosen (Patrick et al., 1994; Robinson & Spencer, 2006; Lamers, 2007).
* For **SG**, the gamble (odds, chance) should be explained and conveyed appropriately and clearly (Garcia-Retamero et al., 2012).
* For **TTO**, the time span should be chosen appropriately (also in view of possible non-linearity or valuations worse than dead)(Patrick et al., 1994; Robinson & Spencer, 2006; Lamers, 2007).
* For **VAS**, the scale endpoints and interval should be chosen appropriately (McCormack et al., 1988).
* For **WTP**, if results are to be applicable to a situation outside the context of the payer, financial context (of the respondent or the environment) should be taken into account (Damschroder et al., 2007).

A more complete checklist of issues appropriate for TTO, SG, Discrete Choice Experiments (DCE) and VAS is described by Attema et al. (2013).

**Set (e.g. health-state) based methods**

If attribute sets such as health-states or other combinations of specific treatment outcomes are used in the qualitative methods, attention must be given to how these are stated (Torrance et al., 1995). Descriptions can be conveyed narratively or using predefined attributes with distinct levels. Generally, these should be based on theory or some form of qualitative research in the same research area and validated. This to make sure that the resulting preference weights are not biased by the absence of important attributes or clearness in the descriptions provided (Lancsar & Louviere, 2008; Mangham et al., 2009; Louviere et al., 2011; Coast et al., 2012; Kløjgaard et al., 2012; Reed Johnson et al., 2013; Clark et al., 2014). In any case, the research should clearly state which descriptions were used and how they were devised (including tests for (face) validity of the descriptions).

**Rank-based and rating-based methods**

Items should be described and predefined or generated appropriately (Ryan et al., 2001). The total number of items in a ranking or rating exercise should be appropriate as the cognitive burden depends on the number of items and the complexity of the descriptions of individual items (Ben-Akiva et al., 1992; Flynn et al., 2007). Furthermore, the influence of the starting sequence of the items should be taken into account (e.g. by randomizing the starting sequence)(Attema et al., 2013). Lastly, it should be explained what model or anchoring points were used to map the resulting data (weights, sequences of rating results) on a preference scale (Attema et al., 2013).

**Decision modelling (conjoint analysis, DCE/DCM)**

Items should be described and generated appropriately, with an appropriate number of choices (Terwee et al., 2007; Lancsar & Louviere, 2008; Mangham et al., 2009; Louviere et al., 2011; Coast et al., 2012; Kløjgaard et al., 2012; Reed Johnson et al., 2013). It should be described how the items presented to respondents were selected (e.g. use of experimental designs) (Louviere et al., 2011; Hiligsmann et al., 2013; Reed Johnson et al., 2013). Furthermore, appraisal should take into account the used model to determine weights of individual attributes/levels and the consistency of the results (Shaw et al., 2005; Terwee et al., 2007; Louviere et al., 2011; Mulhern et al., 2014).

For users not familiar with these methods and requiring more guidance, the following checklists can help identify key issues for the appraisal:

* The 'Conjoint Analysis Applications in Health - a Checklist: A Report of the ISPOR Good Research Practices for Conjoint Analysis Task Force' checklist (Bridges et al., 2011) lists a best-practice list of criteria which could be used to appraise the entire process of discrete choice experiments. (available from http://www.ispor.org/taskforces/documents/ISPOR-CA-in-Health-TF-Report-Checklist.pdf)
* A slightly different checklist was published by Lancsar and Louviere (2008)

**Methods containing both quantitative and qualitative components**

Sometimes referred to as **mixed methods**, these methods combine quantitative and qualitative methods. Key aspect of these methods is how both method types are integrated in terms of sequence, interaction and duration (Heyvaert et al., 2013a; Heyvaert et al., 2013b). See Heyvaert et al. for an overview on the critical appraisal of mixed methods.

**Delphi procedure**

The Delphi procedure is mainly a qualitative method used to generate consensus among policy makers and/or professionals and/or patients, but can also be used to generate consensus in a quantitative way using modifications such as the Research ANd Development appropriateness method (separately or in the same study) (Fitch et al., 2001). To determine whether the procedure was performed correctly, the following items may be helpful:

* Were the participants chosen appropriate and substantiated? Was there no sign of selection bias on this end (Jones & Hunter, 1995; Baker et al., 2006; Hsu & Sandford, 2007)?
* Group members should be blinded from each other (Okoli & Pawlowski, 2004).
* The methods for consolidating responses should be clearly described and fed back to the participants (Schmidt, 1997; Hasson et al., 2000).
* Was pilot testing performed to test measurement methods, consensus thresholds and controlled feedback (Clibbens et al., 2012)?
* Were interim results of each round described (Boulkedid et al., 2011)?
* Was there a low amount of response rate reduction after each round (response rate reduction gives rise to response bias) (Ryan et al., 2001)?

**Citizen juries**

Citizen juries share similarities with focus groups, hence appraisal should cover similar aspects. Additionally, the following criteria may help in the critical appraisal:

* Sampling of participants should be stratified, have substantial honorarium and follow systematic methods (Street et al., 2008).
* The jury should last 4-5 days (longer duration reduces bias from expert input into the jury)(Street et al., 2014).
* The moderators' role should be predefined and objective (Street et al., 2014).
* The study should include reflection of the researchers and feedback to the participants (Malterud, 2001; Street et al., 2014).

**Concept mapping and nominal group technique**

Similar to focus groups and citizen juries, the influence of the moderator is large, and therefore key in appraising these methods (see appraisal criteria 'focus groups'). Though no evidence-based guidance or appraisal was found, the concepts of these methods are described and demonstrated (Deip et al., 1977; Trochim, 1989; Trochim & Kane, 2005; Novak & Cañas, 2006; Novak & Cañas, 2008). In terms of transparency, though, research should clearly state the exact methods used and describe the (group) process itself as well. There are some parallels with Delphi methods as well in terms of participant selection (Jones & Hunter, 1995).

**Qualitative research methods**

There is a diversity of methods to be used for qualitative research, however some aspects apply to all types of qualitative research: Respondent answers should be validated using feedback (Malterud, 2001; Ryan et al., 2001; Nicholas et al., 2008); Contradictory evidence should be actively sought (Booth et al., 2013; Fortune et al., 2013). Data analysis should be appropriate, standardized and well described (e.g. by using software packages, employing certain framework methods such as grounded theory)(Ryan et al., 2001; Anderson, 2010). Comparisons with other sources, methods or triangulation should be used to determine validity of findings (Malterud, 2001; Nicholas et al., 2008). It should also be clear who speaks for whom, all stakeholders involved should be described (Nicholas et al., 2008). There should be a consistent and logical flow of arguments (Nicholas et al., 2008).

If data saturation is used to determine the end-point of data collection, methods and progress towards data saturation should be systematic and appropriate. There are many methods described to determine the point of data saturation, and the speed with which this is reached depends on the heterogeneity of the population and the topic (SAGE Publications, 2008). As an example, Guest et al. (2006) demonstrate the progress of data saturation in in-depth interviews. Some contest the idea of reaching data saturation being a measure of quality (O’Reilly & Parker, 2013).

The reporting guidelines found in the COnsolidated criteria for REporting Qualitative Research (COREQ) (Tong et al., 2007) may give some extra information on which properties of qualitative research one can look to appraise study design and analysis.

Ryan et al. (2001) identified the following items regarding validity and consistency in qualitative research:

* Results should be consistent with existing research / (a priori) expectations;
* The method should be able to measure all things deemed important in the design of the construct (e.g. considering all domains of quality of life);
* The method should allow respondents to give the answers reflecting their (true) preferences (e.g. allow for completeness of answering choices);
* Mind possible framing effects (bias due to the circumstances under which the study is performed);
* Mind strategic bias (respondents may be giving answers to steer the outcome of a study) (for instance, giving answers that quickly terminates a questionnaire);
* reproducibility of methods and internal consistency (though this may be hard to define in qualitative research).

To determine the ‘completeness’ of the answers respondents can give, in-depth knowledge of the subject that is required. Identifying clues for strategic bias starts with identifying possible benefits for the respondent and trying to find clues that answering strategies have been employed that could lead to such outcomes.

1. **Are the results transferable to the target population?**

Results are transferable if the population, setting and values are comparable (Kuper et al., 2008a; Kuper et al., 2008b; Facey et al., 2010). The extent of which these barriers affect transferability is greatly dependant on many contextual factors. It is therefore not possible to indicate how much transferability is affected if these barriers are crossed. Generally, the greater the differences in culture or geography, the more transferability is impacted.

For qualitative research, additional or alternative criteria may apply concerning transferability: For example whether saturation was addressed (Nicholas et al., 2008) and whether the study adequately addresses potential ethical issues, including reflexivity (Kuper et al., 2008a; Kuper et al., 2008b; Nicholas et al., 2008). An appropriate sample (well explained and justified), is important as well.