# Guidance for the assessment of treatment moderation and patients' preferences

4

*EDITORS: Ralph van Hoorn, Marcia Tummers, Wietske Kievit, Gert Jan van der Wilt*

INTEGRATE-HTA

This project is co-funded by the European Union under the Seventh Framework Programme (Grant Agreement No. 306141)

***PLEASE CITE THIS PUBLICATION AS:***

*VAN HOORN, R., TUMMERS, M., KIEVIT, W., VAN DER WILT, G.J. (eds.) (2016) Guidance for the assessment of treatment moderation and patients' preferences [Online]. Available from: http://www.integrate-hta.eu/downloads/*

***CONTACT:***

*For questions regarding this document, contact INTEGRATE-HTA (info@integrate-hta.eu)*

***DATE:***

*Version of 01/02/2016*

***PROJECT:***

*Integrated Health Technology Assessment for Evaluating Complex Technologies (INTEGRATE-HTA)*

***COORDINATOR:***

Universität Bremen

***PARTNER:***

*The research leading to these results has received funding from the European Union Seventh Framework Programme ([FP7/2007-2013] [FP7/2007-2011]) under Grant Agreement No. 306141.*

***DISCLAIMER:***

*The sole responsibility for the content of this publication lies with the authors. It does not necessarily reflect the opinion of the European Union. The European Commission is not responsible for any use that may be made of the information contained therein.*

# About this guidance

## Who would find this guidance useful?

This guidance is primarily intended for use by health technology assessment (HTA) researchers and guideline developers who wish to take account of heterogeneity among patients in their capacity to respond (favourably or unfavourably) to treatments or their preferences towards the outcomes of these treatments.

---

## Purpose and scope of this guidance

This guidance can be used to retrieve and appraise current knowledge of moderators and predictors for treatment effects, and patient preferences for treatment outcomes. It can also be used to determine whether such knowledge should affect recommendations regarding the funding and use of the relevant treatments.

---

## Added value for an integrated assessment of complex technologies

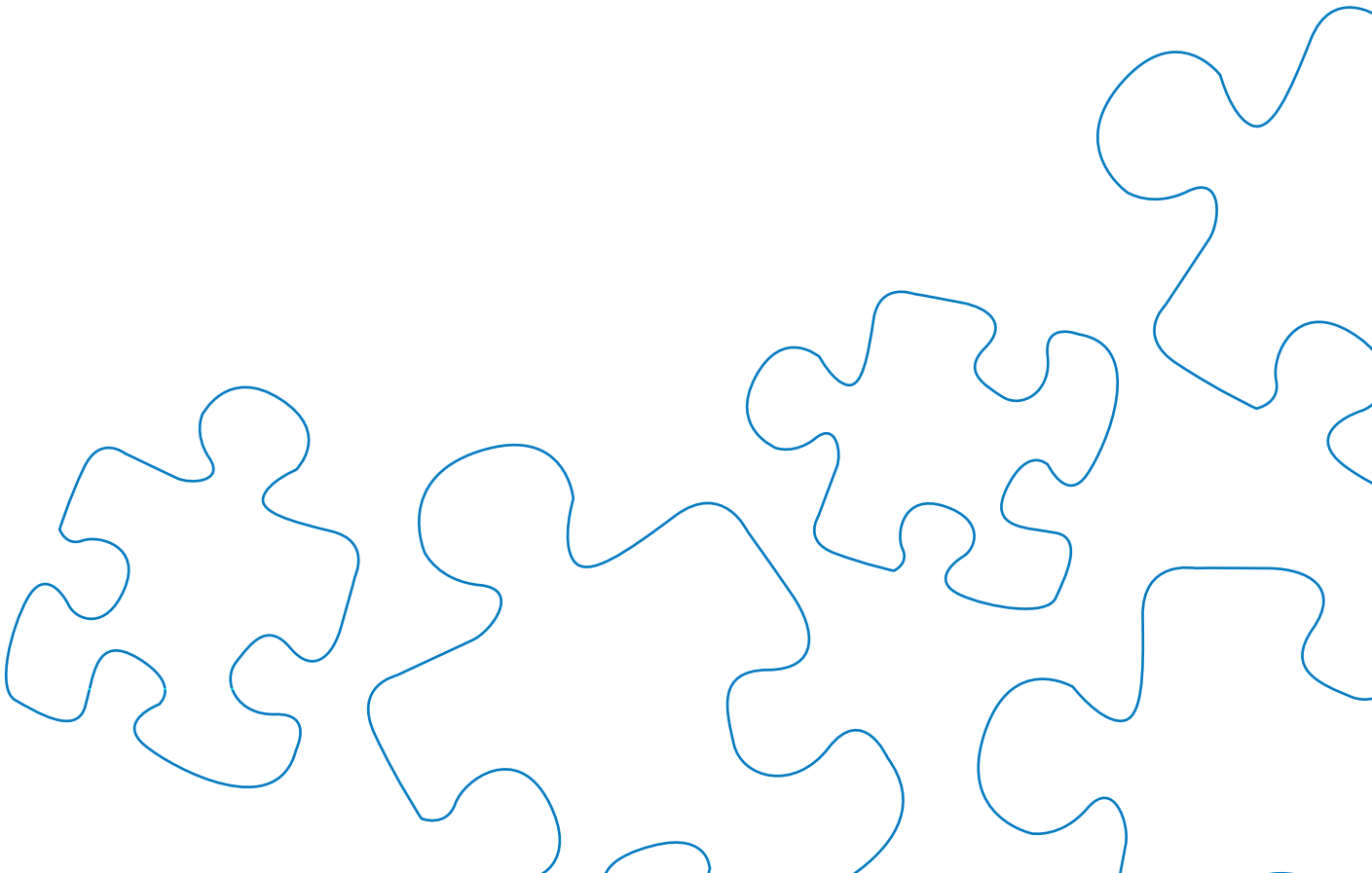
Apparently similar patients may respond quite differently to treatments. It is one of the sources of complexity in the context of healthcare. Another source is the heterogeneity in the valuation of treatment outcomes among patients. This guidance should help users to judge whether, and if so how, this heterogeneity should be taken into account when assessing the value of healthcare technologies.

---

## INTEGRATE-HTA

INTEGRATE-HTA is an innovative project that has been co-funded by the European Union under the Seventh Framework Programme from 2013 till 2015. Using palliative care as a case study, this project has developed concepts and methods that enable a patient-centred, comprehensive, and integrated assessment of complex health technologies.

*Guidance for the assessment of treatment moderation and patients' preferences*



# Executive Summary

## Background

In recent years there have been major advances in the development of health technology assessment (HTA). However, HTA, still has certain limitations when assessing technologies which

- are complex, i.e. consist of several interacting components, target different groups or organizational levels, have multiple and variable outcomes, and/or permit a certain degree of flexibility or tailoring (Craig et al., 2008),
- are context-dependent - current HTA usually focuses on the technology, not on the system within it is used,
- perform differently depending on the way they are implemented,
- have different effects on different individuals.

This guidance deals with heterogeneity in how patients who share a particular disease respond to specific treatments. The nature and magnitude of a beneficial effect may vary, as may the time of onset and its sustainability over time. The same holds for adverse outcomes. Although this heterogeneity is as old as medicine itself, interest in this phenomenon appears to be fairly recent. Arguably, taking such heterogeneity among patients into account in the development, assessment or use of health technologies could considerably enhance their appropriateness. To make this happen, patient-related factors that influence the effects of such technologies (moderators and predictors) need to be explored and findings need to be shared through published reports. Concomitantly, to enable incorporation of the resulting insights in treatment guidelines, it is important to develop search strategies to retrieve this information efficiently and to develop guidance to appraise the information.

In addition to heterogeneity of effect of medical interventions, patients can and often do have differing views on the relative importance of certain treatment outcomes. Recognising this heterogeneity could improve the appropriateness of technologies by assessing them on outcomes that are important to patients. If multiple outcomes of a technology are evaluated, Patients Preferences for Treatment Outcomes (PPTOs) may also help in trading off between outcomes such as adverse effects and beneficial effects of a treatment. This should help informing decision makers on weighing factors to determine which treatment is preferred for a (sub)group of patients. Finally, knowledge about PPTO may also help prioritising new research.

These two sources of complexity, (moderators and predictors on one hand, and PPTOs on the other) should be acknowledged in HTA, otherwise we may fail to realize the potential value of healthcare technologies. An HTA should, therefore, reveal whether there is evidence for clinically important moderators or predictors of treatment effects. It should also reveal whether there is evidence to suggest whether and how patients differ in their appreciation of various treatment outcomes. The results should help guideline developers to decide whether such factors need to be taken into account.

## Purpose and scope of the guidance

The first part of this guidance provides methods for performing efficient searches of the literature on moderators and predictors of treatment outcome. As the quality of this literature is important to consider before implementing results in guidelines, an appraisal checklist was developed to enable for critical appraisal. The checklist covers not only study quality, but also relevance of the findings to the research question.

The second part of this guidance offers methods for finding and appraising literature on PPTOs. This part also contains an overview of methods that are commonly used for such purpose.

When there is evidence of moderation of treatment effect and/or differential preferences for treatment outcomes among patients, HTA researchers or guideline developers need to decide whether and in what way this should be used in order to inform or select patients for treatment. This can be facilitated by determining the effect of adding such factors to a clinical decision and comparing the expected additional effect against the costs, for instance of performing the diagnostic tests required to classify patients in subgroups. Therefore, the last section of this guidance proposes a framework to test variables individually or in combination by modelling their effects on treatment outcomes. The results of such a modeling exercise should reveal whether there is added value of using moderators, predictors and/or preferences to make choices for different treatments in specific subgroups.

## Development of the guidance

To guide the identification of literature on moderators and predictors of treatment effects, a set of search filters for PubMed was developed. It was created by firstly collecting relevant articles on moderators and predictors of treatment effects. This was done by hand-searching a large volume of articles and selecting papers that reported on moderators or predictors of treatment effects. Subsequently, search terms were retrieved from these papers and algorithmically combined to find the optimal combination of search terms for finding articles on moderators and predictors of treatment effects.

For the development of the appraisal checklist, a literature review of the reporting and critical appraisal of moderators and predictors of treatment effects was conducted. The review was based on searches in PubMed and Google Scholar, citation chasing, author searches, related articles and consultation with experts. Subsequently, a Delphi procedure was used following the Research ANd Development Appropriateness Method guidelines to value a set of eligible appraisal criteria retrieved from the literature. Based on these results, a final selection of criteria was made and included in a test version of the appraisal checklist. After testing this version, adjustments were made based on feedback from testers and external reviewers to create a final version.

To guide the identification of literature on PPTOs, the methods to create search filters for finding moderators and predictors in PubMed were repeated for studies describing PPTOs.

The development of the appraisal checklist for PPTOs required a different approach due to the diversity of studies and methods used to elicit preferences, as well as the wide range of literature and appraisal checklists already available. The construction of the appraisal checklist started with a literature search for the identification of methods to elicit PPTOs and literature searches for appraisal criteria for each of these methods. For the various methods that were found, additional searches were performed to find evidence on how these methods could be used for eliciting patient preferences. These results served as basis for the guidance on generating new evidence. It was decided to guide the user towards existing tools as much as possible, while breaking down the appraisal task into a set of key items. Each of the items would direct the user towards individual criteria or appraisal tools described elsewhere.

To estimate the effects of moderators, predictors and PPTOs, a framework was developed based on Value Of Information-analysis. The framework was developed in such a way that it guides the user through building

up a model to estimate effects with and without a factor of interest. To determine which model should make up the core of this framework, several possible models and their usage considerations were identified through literature and summarised. Lastly, the use and possibilities of the framework as a whole was demonstrated using an example from the field of palliative care.

## *Application of this guidance*

For a comprehensive integrated assessment of a complex technology we have developed a five step process, the INTEGRATE-HTA model. In Step 1 the HTA objective and the technology are defined with the support from a panel of stakeholders. A system-based logic model is developed in Step 2. It provides a structured overview of technology, the context, implementation issues, and relevant patient groups. It then frames the assessment of the effectiveness, as well as economic, ethical, legal, and socio-cultural aspects in Step 3. In Step 4 a graphical overview of the assessment results, structured by the logic model, is provided. Step 5 is a structured decision-making process informed by the HTA (and is thus not formally part of the HTA but follows it).

This guidance is used in Step 2 to identify factors predicting treatment outcomes and patients' preferences towards treatment outcomes which feed into the logic model. The first sections for the identification of moderators and predictors of treatment outcomes, and the second section for the identification of patient preferences for treatment outcomes, can be used in parallel. Each of the sections describe the evidence retrieval and subsequent critical appraisal of the found literature.

The final section of this guidance on the integration of patient preferences and moderators for treatment outcomes, should be used after the other two sections have been completed. It will help creating an overview of the found evidence and uncover gaps in evidence. Hence, during this section it may be possible to go back to the other sections to supplement the evidence using more sensitive searches directed towards these gaps.

## *Conclusions*

Search term combinations which help to retrieve an optimal set of relevant papers concerning moderators and predictors of treatment outcome were established and tested. Combinations optimised for sensitivity, specificity, number needed to read, and accuracy are presented. The guidance describes how to use and choose between these search filters.

A checklist for the appraisal of studies on moderators and predictors was developed and tested as well. It should help those that seek guidance in the appraisal of moderator or predictor effects reported in the literature. The benefit of using the filters is likely to depend on the disease area (available literature) and research question, but they are easy to use and can help in increasing the process of identifying factors influencing treatment outcomes.

Using the methodology described in this guidance will likely aid the uptake of evidence on moderators and predictors for treatment outcome in for instance treatment guidelines. Even though a more systematic inquiry of moderators, predictors and PPTOs may increase the workload for HTA researchers, it is expected to improve the relevance of HTA results. More accurate estimations of effectiveness of a technology can be made and different conclusions for groups of patients may be drawn to optimise the usage of a technology. Moderators and predictors may also have ethical, legal, and socio-cultural implications. Taking the effects of different factors into account when making treatment decisions could considerably enhance the appropriateness of medical care.

This guidance can be used to systematically retrieve and appraise evidence of patient preferences for treatment outcomes. PPTOs can be used to support effectiveness valuation in case of positive as well as negative effects, or identify subgroups where a particular technology or focus is required. They can directly inform

effectiveness assessments to weigh multiple outcomes, or can be used in a decision process to judge different aspects of technologies. As such, it can be considered as a step in the direction of a more stratified healthcare as well as a more stratified assessment of technologies.

Hence, a framework for integrating evidence on moderation and prediction of treatment effect and patients' preferences was built. The added value of doing this compared to a 'one-size-fits-all' approach was demonstrated with a hypothetical example. However, its evidential requirements are substantial: reliable data are needed on [1] factors that modify treatment response and [2] whether substantial differences exist across patients in their valuation of outcomes. Furthermore, it should be noted that implementing personalized health care in daily practice also incurs certain costs. These costs are associated with the generation and use of further evidence, and with the introduction of false (positive or negative) treatment decisions. The modelling exercise can be used to test if a stratified approach could have added value based on what is described in the medical literature. The framework described in this guidance can be used to explore these benefits and relate them to the costs associated with stratified healthcare.

# List of abbreviations

| | |
|---|---|
| Ac | Accuracy |
| AMSTAR | A MeaSurement Tool to Assess systematic Reviews |
| CASP | Critical Appraisal Skills Programme |
| CERQual | Confidence in the Evidence from Reviews of Qualitative Research |
| CONSORT | CONsoldiated Standards Of Reporting Trials |
| CPR | Clinical Prediction Rules |
| DA | Decision Aid |
| DCE/DCM | Discrete Choice experiment/modelling |
| EBP | Evidence Based practice |
| EVPPI | Expected Value of Partial (or Parameter) Perfect Information |
| GRADE | Grading of Recommendations Assessment, Development, and Evaluation |
| HTA | Health Technology Assessment |
| ICC | Intra-Class Correlation |
| MCDA | Multi-criteria Decision Analysis |
| ME | Magnitude Estimation |
| MeSH | Medical SubHeading |
| NICE | National Institute for Health and Care Excellence |
| NNR | Number Needed to Read |
| PPTO | Patient Preferences for Treatment/Technology Outcome |
| PRISMA | Preferred Reporting Items for Systematic Reviews and Meta-Analyses |
| PTO | Person Trade-off |
| QALY | Quality Adjusted Life Year |
| QoL | Quality of Life |
| RAND | Research ANd Development |
| RCT | Randomised Controlled Trial |
| Se | Sensitivity |
| SG | Standard Gamble |
| Sp | Specificity |
| STROBE | STrengthening the Reporting of Observational studies in Epidemiology |
| TTO | Time-TradeOff |
| VAS | Visual Analogue Scale |
| VOI | Value of Information |
| WTP/WTT | Willingness To Pay/Travel |

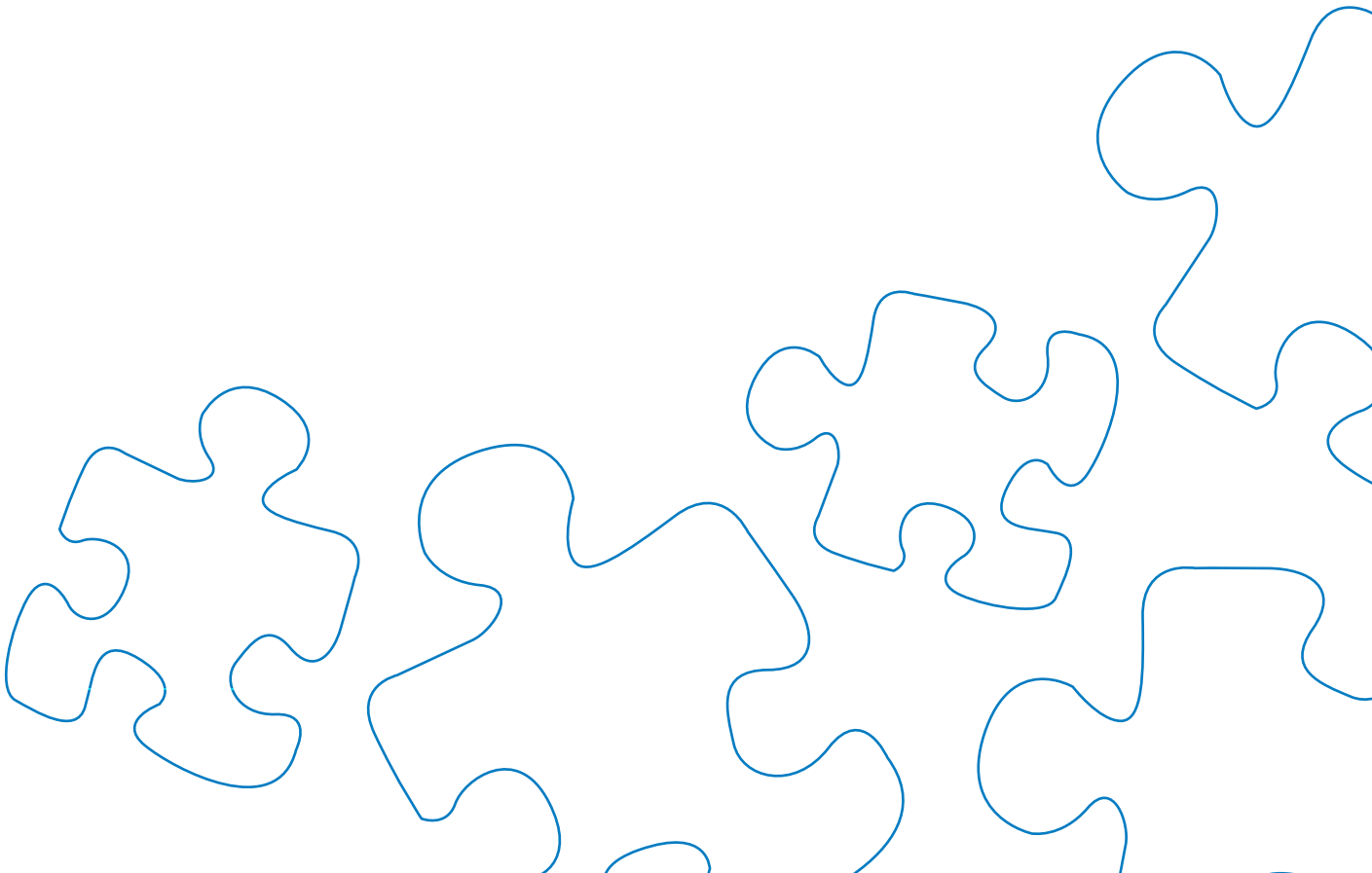
# Table of contents

# List of Tables



# List of Figures

*Guidance for the assessment of treatment moderation and patients' preferences*



# 1 INTRODUCTION

## 1.1 PURPOSE AND SCOPE OF THE GUIDANCE

Complexity in the context of healthcare partly derives from the fact that patients with a particular disease condition may respond quite differently to a specific treatment. This is true of both beneficial and adverse effects. Moreover, patients may differ in how they value particular treatment outcomes. For instance, in the context of treatment of patients with epilepsy, a specific drug may be known to lead, on average, to slightly superior control of seizures, improved mood, but also weight gain. Even though this may be true on average, seizure control will be better achieved in some patients using other drugs. Also, weight gain, if it occurs, may be less of a problem to some patients as compared to others. Given the vast number of possible treatments and combinations thereof, it can be quite challenging to find the optimal treatment for an individual patient. Arguably, to this end, it would be of great value to know in advance whether particular patient characteristics are predictive of the onset of specific treatment outcomes and whether differences exist in how these treatment outcomes are valued by patients. Taking into account such knowledge when discussing treatment strategies with patients could greatly enhance the appropriateness of healthcare.

This guidance deals with these issues. It takes the perspective of HTA researchers who wish to make use of the best available evidence in order to develop recommendations as to how and for whom healthcare technologies may be optimally used. This guidance consists of three sections:

1. The retrieval and critical appraisal of literature on moderators and predictors of treatment effects. This section of the guidance draws attention of HTA researchers to heterogeneity in treatment response: how widely do patients differ in their response to certain treatments, both beneficially and adversely, and what is known about patient characteristics that seem to be associated with this variability? For HTA researchers it is important to know how such knowledge can be found efficiently and how it can be critically appraised for its validity and relevance. For this part of the guidance, specific search filters and an appraisal checklist were developed and tested.
2. The retrieval and critical appraisal of literature on patient preferences for treatment outcomes. This section of the guidance draws attention of HTA researchers to differences across patients in how they value specific outcomes of treatment: what is important to them and how do patients vary in this respect? Here, too, it is important for HTA researchers to know how relevant information on this subject can be found and how it can be critically appraised. Also for this part of the guidance, specific search filters and an appraisal checklist were developed and tested.
3. Guidance on the integration of moderators of and patient preferences for treatment outcomes. The third and last section of this guidance aims to support HTA researchers in using information on moderators or predictors and preferences when developing recommendations regarding the use of healthcare technologies. Given what is known about differences in treatment response between patients, about associated patients' characteristics and about differences in valuation of these outcomes, can a case be made for a stratified approach? It is important for HTA researchers to be critical on the fact that a more stratified approach may not always be the most effective. The methods described in this guidance describe how to synthesize this evidence in a model in order to determine the possible effects, but also the costs, of making treatment decisions better tailored to groups of patients.

Although the three sections of this guidance can be used separately, the integration step (last section) is an important step in the estimation of relevancy of moderators, predictors or patient preferences. This requires an integrated look at the evidence as described in the third section of this guidance.

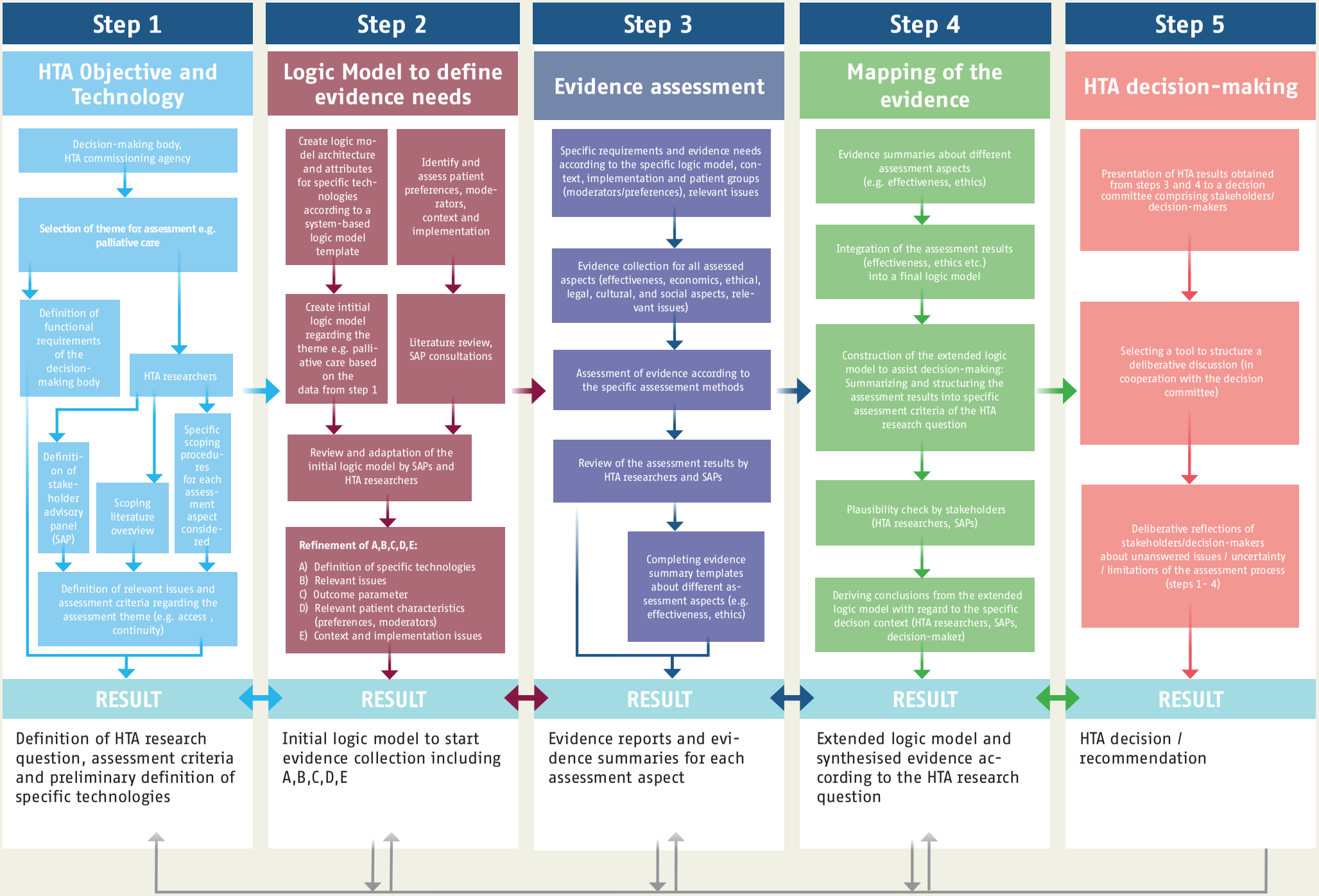
### 1.1.1 Target audience for this guidance

Health professionals such as HTA-researchers and guideline developers, who wish to take account of heterogeneity in treatment response and outcome preferences among patients, can make use of the methods described in this guidance to assess the value of acknowledging this information within technology assessments or guidelines.

### 1.1.2 How this guidance relates to an integrated assessment process

In order to achieve an integrated HTA, the application of the methodological guidances is structured into a systematic assessment process to strive for integration from the very beginning of the HTA. The INTEGRATE-HTA Model (see Figure 1) consists of five steps (Wahlster et al., 2016). After an initial definition of the HTA objective and the technology in accordance with the support of the stakeholders in step 1, a specific logic model in step 2 provides a structured overview of the factors and aspects surrounding the technology. Patient characteristics, context and implementation issues inform the assessment of effectiveness, and economic, ethical, legal, and socio-cultural aspects in step 3. In step 4, a graphical overview of the assessment results is structured according to the HTA objective and the logic model is created. Finally, the presentation of the results in step 5 forms the basis for a structured decision-making process.

*Figure 1: INTEGRATE-HTA process model for integrated assessment of complex technologies.*



The methods described in this guidance can be used to retrieve information on moderators and predictors for treatment effects from the literature, as well as the preferences associated with these outcomes. It focuses on heterogeneity of patients and the influence of this heterogeneity on treatment outcomes. The results of the methods described in this guidance can, for instance, be used to identify subgroups of patients where a particular technology is expected to better than in de total group of patients. This relates to an effectiveness assessment within an HTA (Burns et al., 2016). However, applied into the INTEGRATE-HTA Model it provides a more comprehensive, iterative and integrated process. How the guidance can feed into the INTEGRATE-HTA Model is indicated in Figure 1. Moderators, predictors and preferences feed into the logic model in step 2 as indicated by the yellow marked boxes. On the one hand, information on preferences for treatment outcomes can guide the focus of the effectiveness assessment. On the other hand, socio-cultural, legal, or ethical issues may determine preferences for treatment outcomes. Furthermore, specific socio-cultural, legal, or ethical issues may only apply to specific groups of patients.

## 1.2 BACKGROUND

A treatment is usually defined as a specific medical intervention. Moderators, predictors, and patient preferences, the subject of this guidance, are usually viewed from a medical or epidemiological viewpoint where the term treatment is a well established and comprehensive term. In HTA the term (health) technology is used more often as it is considered a more comprehensive term and including a broader area of interventions that may be assessed. As the contents of this guidance are based on many principles originating from epidemiology and clinical research, we will be using the term treatment, instead of technology, to stay true to the source of most of our information. The general concepts presented in this guidance, however, are valid in the wider context of (health) technologies.

So far, there has been relatively little attention within for moderators, predictors and patient preferences for treatment outcomes. We argue that, especially for complex technologies, the (combined) effects of these factors may be larger than anticipated and ignoring these factors can then lead to inappropriate HTAs or guidelines. When moderators, predictors and patient preferences are used to stratify or adjust recommendations, better assessments and results may be achieved.

### 1.2.1 Complexity

The UK Medical Research Council (MRC) defines complex interventions as being characterised by the number of interacting components within the experimental and control interventions, the number and difficulty of behaviours required by those delivering or receiving the intervention, the number of groups or organisational levels targeted by the intervention, the number and variability of outcomes, and the degree of flexibility or tailoring of the intervention permitted (Medical Research Council, 2008). Shiell et al. (2008) highlight

*Table 1: Synthesis of potentially relevant characteristics of complexity in HTA.*

| Characteristic | Short explanation |
|---|---|
| 1 Multiple and changing perspectives | The variety of perspectives is caused by the many components (social, material, theoretical, and procedural), actors, stakeholders and organisational levels that are involved in the intervention. These are interconnected and interacting, and accordingly exposed to changes. |
| 2 Indeterminate phenomena | The intervention or condition cannot be strictly defined or delimited due to characteristics like flexibility, tailoring, self-organization, adaptivity and evolution over time. |
| 3 Uncertain causality | Factors like synergy between components, feedback loops, moderators and mediators of effect, context and symbolic value of the intervention lead to uncertain causal pathways between intervention and outcome. |
| 4 Unpredictable outcomes | The outcomes of the intervention may be many, variable, new, emerging and unexpected. |
| 5 Historicity, time and path dependence | Complex systems evolve through series of irreversible and unpredictable events. The time, place and context of an intervention therefore impact on the effect, generalizability and repeatability of an intervention. |

that complexity is a characteristic of the system within which an intervention acts as well as being an inherent characteristic of an intervention itself. They describe complex systems as being adaptive to their local environment, as behaving non-linearly and as being part of hierarchies of other complex systems.

Many of the traditional methods of analysis in HTA rely upon specific assumptions about the structure, content and objectives of an intervention, its implementation, the system within which it is intended to act and the potential interplay and co-evolution of the system and the intervention. However, to avoid misleading conclusions, HTA should take the complexity of a technology and/or the complexity of its environment into account. For example, when assessing a technology such as an educational program to prevent the transmission of the human immunodeficiency virus (HIV) the success or failure might depend on the message itself (e.g. abstention or condoms or both), the messenger (a young celebrity or a respected religious leader), the target group (sexually active adolescents or elderly religious persons), the medium transmitting the message (internet spots or lectures), the perceived prevalence of the disease (omnipresent threat or small chance), and so on. Simply to focus on the content of the program without considering these other factors is not sufficient.

Complexity is not a binary property, and exists rather along a spectrum. All interventions could, therefore, be considered complex to a certain extent. This guidance, however, focuses on those health technologies where the presence of complexity has strong implications for the planning, conduct and interpretation of the HTA. Table 1 lists potentially relevant characteristics of complexity.

Consequently, when starting an assessment of (any) health technology these factors should be carefully reviewed with the purpose of

1. describing the complexity of an intervention and the system within which it acts,
2. understanding whether this complexity matters for decision making and therefore needs to be addressed in an HTA,
3. understanding the implications of complexity for the methods of HTA analysis in assessing the ethical, legal, effectiveness, economic and socio-cultural aspects of an intervention, and
4. exposing important factors that decision makers need to consider in interpreting the HTA.

### *1.2.2 Definitions*

There are multiple ways in which patient characteristics may modify the effect of a treatment. The definitions used here are based on papers by Baron and Kenny (1986) and Kraemer et al. (2002). They distinguish three different types of treatment effect modification based on the exact causal models: moderators, predictors and mediators. Two of these, moderators and predictors, are relevant to this guidance.

First of all, it is important to realize that moderator and predictor are names for theoretical variables in a causal model, not correlational variables as are commonly described in the medical literature (e.g. statistical models) (MacCorquodale & Meehl, 1948). Moderators and predictors can be investigated using subgroup or regression analysis in epidemiological studies, randomised controlled trials (RCTs) or between-study comparisons in meta-analysis, but require an underlying theoretical model (March & Curry, 1998; Kraemer et al., 2008; Viechtbauer, 2008).

Moderators (Figure 2, left) are variables which influence the strength of a relation between two other variables, for instance that of a treatment and an effect. Moderators do not change due to a treatment, and thus they should be measured prior to randomization (Kraemer et al., 2002; Nicholson et al., 2005). Age and gender are common moderators. The term is similar to the possibly more widely known epidemiological and statistical terms 'effect modifier' and 'interaction'. Equation 1 shows a generic statistical function where the Outcome is defined as a treatment effect ( ß0 × treatment), a predictor effect as ß1, and a moderator as ß2 × treatment × variable. In the formula below, ß2 signifies the moderator effect of the variable on the outcome.

$$Outcome = \beta_0 \times treatment + \beta_1 \times variable + \beta_2 \times treatment \times variable \quad (1)$$

The strength of moderation can range from partial to complete. In extreme cases, moderators can turn the positive effect in one subgroup to a negative effect in another subgroup. In such a case, this effect may be the cause of an RCT finding no overall treatment effect (Hunter & Schmidt, 2004). In many cases, however, individual moderator effects are small, especially when the causal models are more complicated (Kraemer, 2013). This is why moderator analysis is often challenging: high statistical power is needed not only to compensate for the smaller groups on which the analysis is performed, but also to find the smaller effects of moderators (Aguinis, 2004; Gabler et al., 2009).

Predictors (Figure 2, right) are characteristics or variables that influence the outcome of a treatment similarly to moderators. However, predictors are not related to the specifically used treatment. The effect would be the same if a different treatment was applied. For example, a patients' age may influence survival regardless of whether a patient is receiving a treatment. Statistically, predictors are tested by a main effect ( $\beta_1$ in equation 1) without interaction with the treatment. Prognostic factors act on outcomes when there is no intervention taking place. For instance, age may be a prognostic factor on survival in cancer patients. Prognostic factors

can be used to determine which (group of) patients should be treated, while predictors (and moderators) can be used to determine which treatment to give based on expected results. Please be aware that some define prognostic factors as we define predictors, and define predictive factors as we define moderators (e.g. Adolfsson & Steineck, 2000). Throughout this guidance the focus will be on treatment-related factors, not prognostic factors. To avoid confusion the moderator-predictor distinction is preferred. The term prognostic factor will be used for those effects which are not related to treating patients (e.g. prognosis).

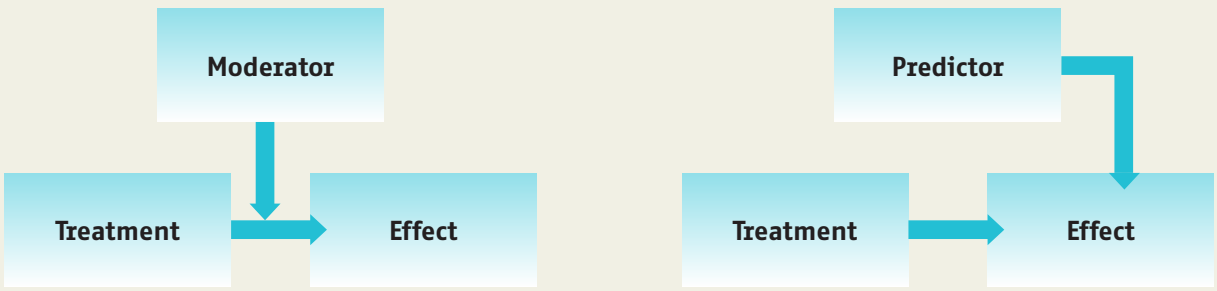
A preference is defined by the Oxford English Dictionary as 'A greater liking for one alternative over another or others' (http://www.oxforddictionaries.com; accessed Nov. 2014). In other words, a preference is a latent value of an individual, with which choices are made between two or more options. In healthcare research, individual subjects may weigh or define risks and benefits differently due to different personal beliefs, risk perception, knowledge and so on (Hanley et al., 2001). Furthermore, preferences are not static, and may change over time within the same individual (Fried et al., 2006). It is therefore important to acknowledge that these preferences, contrary to treatment outcomes, are difficult to predict, and that they should be actively elicited from individual patients or groups of patients to determine them for any given situation.

The choice for one treatment or another (as made by DAs) is made, at least partly, by preferences for [specific] treatment outcomes: distinct desired states of health preferred over other states of health. Such preferences may concern the health state as a whole or specific aspects of it (such as pain or mobility), but may also concern the occurrence (odds) of certain (adverse) events, without looking at the specific effect on health as a whole or the probability of its occurrence.

The distinction between treatments and their outcomes is not always clear: A wound following invasive surgery may be seen either as outcome (undesirable effect) or as part of the treatment (surgery without making wounds is often impossible). These terms may best be regarded as opposite ends of a spectrum (Street et al., 2012), illustrating the relationship between treatment preferences and treatment outcome preferences. A treatment preference usually comprises the combination of multiple treatment outcomes, their preferences, the odds of getting these outcomes as well as treatment-specific properties (e.g. where it is performed or how long it takes). An outcome preference only involves one (or a set of) effect(s). It does not consider the probability of obtaining the outcome. Ideally, standardized descriptions of outcomes are constructed and applied across different studies and treatments, but the construction of such universal descriptions has proven to be challenging (MacLean et al., 2012).

Throughout this guidance we will refer to stratified medicine when it comes to addressing patient preferences for treatment outcomes and moderators and predictors for treatment outcomes. This term relates to the ability to identify subgroups of patients where a particular intervention is more (or less) applicable than others. It is comparable to the term personalised medicine (matching patients with their best treatment using predictive knowledge and better diagnostics (i.e. by genomic assays or patient characteristics) (Redekop & Mladsi, 2013)), but focuses on recommendations for subgroups of patients instead of individual patients.

*Figure 2: Schematic overview of two causal models containing a simple relationship between an input variable (e.g. treatment), an output variable (e.g. effect) and a moderator (left) or predictor (right).*



# 2 THE RETRIEVAL AND CRITICAL APPRAISAL OF LITERATURE ON MODERATORS AND PREDICTORS OF TREATMENT EFFECTS

*By: Ralph van Hoorn, Marcia Tummers, Andrew Booth, Ansgar Gerhardus, Bjørn Hofmann, Eva Rehfuess, Wietske Kievit, Gert Jan van der Wilt*

## 2.1 INTRODUCTION

### 2.1.1 Purpose and scope of the guidance

#### *Aim of this guidance*

Patient heterogeneity is one source of complexity in evaluating interventions. If this complexity is not acknowledged sufficiently, we may fail to realize the full potential of a technology or produce results which do not reflect optimal outcomes for certain subgroups. This document will help by guiding the user to identify heterogeneity in the form of moderators and predictors of treatment effects from the medical literature.

#### *The added value of this guidance in relation to existing guidances*

Identifying and using information from the literature concerning moderators or predictors of treatment effect requires a search strategy and a subsequent appraisal of the literature that was found. Although various search strategies do exist to retrieve medical literature on a specific subjects (e.g. PubMed's Clinical Queries), none of the existing queries are directed towards moderators or predictors of treatment effects. Hence, those interested in incorporating moderators or predictors in their clinical protocols or decisions need to hand-search the literature in the area of interest. A search more focused on moderator or predictor-research may improve efficiency in this task.

Most guidance on literature appraisal is aimed at valuing the primary outcome of a study: the overall validity, methods and other study properties that determine the relevance and credibility of the presented outcomes. Moderator and predictor analysis are sometimes mentioned but generally not expanded into explicit items. Guidance on subgroup analysis exists but is not explicitly aimed at appraisal of evidence for use in HTAs. Although core principles from reporting guidelines can be used for appraisal purposes, this translation step requires more in depth knowledge of methodologies. Moderator or predictor analyses are different from analyses related to the main outcome in multiple ways. They are often more complicated, and since they are based on subpopulations of the main study population, they are associated with greater uncertainties. Also, subgroup analyses are often reported with strong claims, while the credibility of these claims is low (Sun et al., 2012). The added value of this guidance in view of appraisal lies in the fact that it is the first to address specifically moderators and predictors of treatment effects from the viewpoint of their credibility, transferability and relevance for use in HTAs. An appraisal checklist based on existing sources of good practice in (the reporting or methods of) moderator, predictor, or subgroup analysis in general, can help professionals to appraise this body of knowledge.

### 2.1.2 Description of theoretical background and available approaches

To facilitate retrieval of relevant papers from the literature, search filters have been developed across many topic areas including aetiology, diagnosis, prognosis and therapy (McKibbon et al., 2009). Moderators and predictors for treatment effects may be identified in the literature as well. One option to do so could be to use search filters for the retrieval of papers for clinical queries on prediction ('Clinical Prediction Rules'(CPRs)) (Wong et al., 2003; Haynes et al., 2005). CPRs combine multiple variables to quantify their individual contribution to the diagnosis, prognosis, or likely response to treatment in a patient (McGinn et al., 2008). Typically, however, CPRs focus on prediction of prognosis or diagnosis (which we define as prognostic effects); only a minority focuses on moderators or predictors of treatment effects (Cook, 2008; Haskins et al., 2012). Furthermore, many moderators or predictors of treatment effect may exist that have not been integrated into CPRs. Therefore, we aimed to develop search filters for one of the most popular medical literature search interfaces, PubMed, to guide the retrieval of papers on moderators and predictors of treatment effects in a broader sense.

Once relevant literature has been identified, appraisal of this literature could be a next step. First, existing methods to do so were investigated. The Cochrane Handbook for Systematic Reviews of Interventions (Higgins & Green, 2011) mentions moderators (subgroup or additional analysis) and explains how analysis should be performed in a descriptive way. Arguably, important predictor analyses could be performed in non-randomized studies. Therefore the Cochrane critical appraisal tool, which is prominently geared towards randomized studies is not sufficient for our purpose. The GRADE (Grading of Recommendations Assessment, Development, and Evaluation) system (Nasser & Fedorowicz, 2011) is a tool for appraisal of bodies of evidence. GRADE, however, does not mention subgroup analysis as a specific topic. There are resources informing on good practice in moderator/predictor (or subgroup) analysis, but these are not specifically geared towards moderator and predictor appraisal for use in HTAs (e.g. Pincus et al., 2011; Sun et al., 2012; Gagnier et al., 2013).

We conclude that there is a need for the further development and refinement of appraisal of individual articles on moderator or predictors of treatment outcome based on the lack of suitable existing methods. As often the best evidence on moderators and predictors is found by combining (pooling) multiple studies, the appraisal of bodies of evidence will also be explicitly addressed. This guidance aims to combine relevant evidence on appraisal into an appraisal checklist to help assess the credibility, transferability and relevance of findings for use in HTAs or similar research.

## 2.2 GUIDANCE DEVELOPMENT

This chapter provides an overview of the guidance development. For further details, please see the appendix (chapter 7.1,) and a related paper by Tummers et al. (2016, paper in preparation).

### *Development of the PubMed search filters*

For the development and validation of the search filters for PubMed two steps were taken: [1] a comprehensive set of search terms and combinations of terms was constructed and [2] the results of these combinations of terms were tested in a set of relevant papers. To this end all articles published in 2011 in a selection of six journals were hand searched to identify articles on moderators and predictors of treatment outcome (Annals of the rheumatic diseases, Arthritis care & research, Arthritis research & therapy, Arthritis and rheumatism, The Lancet, and The New England Journal of Medicine). We excluded articles where no intervention was described. The entire set of articles was randomly (1:1) divided into a development set (in which the search strategies were developed) and a validation set (in which the search strategies were tested). The articles deemed relevant in the development set were submitted to PubReMiner, an online resource to retrieve all search terms related to the submitted articles. Using a computer algorithm, search terms were tested and combined to create search filters with optimal sensitivity, specificity, number needed to read (NNR) and accuracy. The applied methods follow accepted good practice in search filter creation (White et al., 2001; Jenkins, 2004).

### *Development of the appraisal checklist: design*

First, a literature search for existing appraisal instruments (i.e. development or testing studies or reviews) was performed to determine what the checklist should look like. The design was based on four main ideas:

- The aim for this guidance is not to define how moderator and/or predictor analysis should be done and when they are performed sufficiently, but to help users to identify possible problems in used methodology. This covers problems with a study (e.g. analysis) but also problems relating to the relevance of the study findings for the HTA. A study may be perfectly executed but still produce irrelevant results due to problems with transferability.
- Closed (i.e. multiple choice) questions allow for the best, fastest and easiest comparisons between studies and/or different appraisers. Furthermore, they enable an easier way to come to an overall judgment of a paper.
- Creating an overall score, with or without specific thresholds and weights, seems inappropriate because of the diverse circumstances in which moderator or predictor analysis can be appraised and the associated problems and insecurities on determining the thresholds and weights (Juni et al., 1999). The lack of a scientific basis for an overall score outweighs the benefits that may possibly come from arbitrarily defined scores and weights (Katrak et al., 2004).
- Based on experiences and reviews of existing tools, comprehensiveness is a key factor.

*Development of the appraisal checklist: contents and testing*

The selection of quality criteria in the appraisal checklist was based on a literature review, a modified Delphi consensus process, a testing phase, and user and expert feedback.

Literature was systematically searched to identify methodology, examples and other general information concerning moderator, predictor and subgroup analysis. The search was extended with citation chasing, authors searching and looking at (development studies of) existing appraisal tools. The results from the complete search was also used to identify those working in the field who could possibly take part in the Delphi consensus process.

The search resulted in information on how the checklist could be designed and a list of criteria eligible for use in an appraisal instrument. To prepare for the modified Delphi procedure the items were grouped on a conceptual basis. During two rounds, the Delphi panel condensed the number of appraisal items by excluding those deemed less relevant. The Research ANd Development Appropriateness Method (Fitch et al., 2001) was used to rate the appropriateness of the list of potential criteria to be included in the checklist. Additionally, rewordings, additions and other considerations were shared and implemented. The goal of these two rounds was to create a preliminary checklist which could be tested.

During the testing phase, the appraisal checklist was applied to a number of papers by multiple users. Based on feedback several items were merged or reworded, and background information for the interpretation of individual items was added.

A revised version of the checklist was presented to two experts who were part of the Delphi panel. They were asked to provide in-depth feedback on the revised version of the checklist, after which further improvements were made.

The next step of the development of the checklist consists of presenting the revised checklist to the entire Delphi panel for final feedback and endorsement. This step is currently taking place, and results of this step are expected in 2016. Future research should also include a second test round where reliability and agreement are assessed.

## 2.3 APPLICATION OF THE GUIDANCE

The application of the search filters and the appraisal checklist to find and subsequently appraise literature on moderators and predictors of treatment effect is described below. The latest version of the appraisal checklist can be found in the appendix (chapter 7.2).

### 2.3.1 Finding literature on moderators and predictors of treatment effect

The user starts with a search query that captures articles relevant to the field of investigation, for instance the disease or intervention being researched. One usually starts by making an inventory of relevant keywords of the field using expert knowledge, and combining these terms using the OR-operator (though other operators may be required to restrict to a certain area only identifiable by combinations of terms). PubMed's Clinical Queries (http://www.ncbi.nlm.nih.gov/pubmed/clinical) can be used as a starting point.

This field specific search query can then be combined with one of the search filters to retrieve moderators and predictors of treatment effects (Table 2) using the AND-operator. In Table 2 four sets of three search filters are listed: each set of three is the top three optimized for the respective performance measure. The filter optimized for sensitivity will return a relatively large share of irrelevant papers but is least likely to miss papers containing the required information. Search filters with high accuracy, specificity or low number of papers needed to screen (NNR) will return less irrelevant papers at the cost of possibly missing important information. Which of the four strategies is the best ultimately depends on the amount of usable retrieved papers and the amount of time the user is willing and able to invest.

Once a number of articles is found, it should be determined whether they truly contain the information the researcher requires by examining them by abstract and full-text. Moderators of treatment can be used directly to stratify treatment recommendations for groups of patients, and are therefore of more value for decision making then predictors. However, moderators are less commonly found in the medical literature compared to predictors. Predictors are also of interest because they can give information about for which patient the chances on good response or adverse effects are higher. This might also be of great value for decision making.

### 2.3.2 Appraisal of studies on moderators or predictors of treatment effects

Before the moderator or predictor analysis can be appraised it is necessary to identify the strengths and weaknesses of an article in order to assess the useful-

ness and validity of research findings. The overall validity of targeted papers (i.e. the overall study design concerning the analysis of its main outcome(s)) needs to be appraised before the moderator or predictor analysis can be appraised. Depending on the type of study, different appraisal methods may be most suitable. The appraisal guidance (chapter 7.2) proposes a number of existing critical appraisal tools which can be used for this purpose. Many other appraisal tools may be applicable too. The user should choose a tool that is applicable and he or she is most familiar with. Only after the overall validity of a paper is deemed adequate, the ana-

*Table 2: Search filters for articles on moderators and predictors of treatment effects. From: (Tummers et al., 2016, paper in preparation).*

| Search term | Se (%) | Sp (%) | Ac (%) | NNR |
|---|---|---|---|---|
| **Optimal sensitivity** | | | | |
| ("Epidemiologic Methods"[mesh] OR assign* OR control*[tiab] OR trial*[tiab]) AND therapy*[sh] | 100.0 | 79.4 | 80.3 | 5.5 |
| ("Epidemiologic Methods"[mesh] OR assign* OR control*[tiab]) AND (therapy*[sh] OR primary*[tiab]) | 100.0 | 79.1 | 80.0 | 5.6 |
| ("Epidemiologic Methods"[mesh] OR analys* OR predict* OR trial*[tiab]) AND therapy*[sh] | 100.0 | 78.5 | 79.5 | 5.7 |
| **Optimal specificity** | | | | |
| group*[tw] AND therapy* | 75.3 | 94.8 | 94.0 | 2.5 |
| randomi* AND treat* | 78.4 | 94.6 | 93.9 | 2.5 |
| group*[tw] AND treat*[tw] | 77.3 | 94.5 | 93.8 | 2.6 |
| **Optimal accuracy** | | | | |
| group*[tw] AND therapy* | 75.3 | 94.8 | 94.0 | 2.5 |
| (randomi* OR hazard*) AND treat* | 79.4 | 94.6 | 93.9 | 2.5 |
| randomi* AND treat* | 78.4 | 94.6 | 93.9 | 2.5 |
| **Optimal NNR** | | | | |
| (randomi* OR hazard*) AND treat* | 79.4 | 94.6 | 93.9 | 2.5 |
| (randomi* OR multivariate) AND treat* | 79.4 | 94.5 | 93.9 | 2.5 |
| randomi* AND (treat* OR death*) | 79.4 | 94.5 | 93.9 | 2.5 |

*Se: Sensitivity, Sp: Specificity, Ac: Accuracy, NNR: Number Needed to Read*

lysis on moderators and predictors can be performed. A study with low overall quality is expected to have a low quality moderator/predictor analysis at best.

If the overall validity of a study is considered to be good, one can look more in detail towards the moderator/predictor or subgroup analysis. Towards this purpose, the appraisal checklist contains twelve items which can help the user to identify possible problems in a study's design, population and measurement, analysis, results or transferability with regard to moderator and/or predictor effects. At the end of the checklist, an overall judgement can be made based on the answers of the twelve questions.

The criteria listed in the checklist are meant to provide insight in the quality of the moderator or predictor analysis presented in papers as well as usability of the evidence towards an HTA (e.g. addressing transferability). No claims are made that all possible problems with moderator or predictor analysis are captured, but the credibility, transferability and relevance of claims will increase if fewer problems are found using this guidance. It is up to the user to make an overall judgement of a moderator or predictor, by weighing the importance of the claims against the problems encountered during the appraisal. One may find that a study is not applicable for their use or that it has not been rigorously executed, or they may conclude they need extra expertise to properly appraise the studies.

If multiple studies cover the same moderator or predictor, it is possible to appraise these studies together. That is, after the individual papers have been appraised, a set of additional items can inform on the credibility, transferability and relevance of the findings in view of the body of literature. A set of five extra items are listed in the appraisal checklist for this purpose. As the evidence of moderators or predictors is much stronger when found in multiple papers, this is an important last step in the appraisal of single moderators or predictors.

### 2.3.3 Application of the search strategy and appraisal checklist in a case study

In order to test and demonstrate the methods described in INTEGRATE-HTA, a case study was set up. In the case study the following research question was developed from stakeholder input: Are reinforced models of home based palliative care acceptable, feasible, appropriate, meaningful, effective, cost-effective model for providing patient-centred palliative care [compared to non-reinforced (i.e. 'usual') models of home based palliative care] in adults (defined as those aged 18 years old and over) and their families? A detailed description of the case study can be found in the INTEGRATE-HTA case study report (Brereton et al., 2016).

This guidance was applied in a case study to identify moderators and predictors for treatment effects in the literature for models of home based palliative care. The results of applying the search strategy and checklist are described below.

#### *Exploring evidence relating to moderators and predictors of treatment effect*

In the case study the search filters as presented in Table 2 were used to explore what is known on moderators and predictors for treatment effects concerning home-based palliative care. We started with a search strategy developed by Gomes et al. (2013) to find relevant papers on home-based palliative care and translated their search query so it could be used in PubMed. Since we expected that relevant information could also be found in non-randomised trials, we removed the study type restrictions in Gomes' search strategy. The resulting search filter was combined with the AND-operator with the best four filters (optimal sensitivity, specificity, NNR and accuracy, combined using the OR-operator) to find relevant papers while maximizing the number and diversity of returned papers containing information on moderators or predictors related to the case study.

By applying the search filters we identified a total of 6928 papers that potentially contained relevant information on moderators or predictors for the case study. As expected, the combination of our four search filters with the filter based on Gomes et al. resulted in a very sensitive, but not highly specific result. After evaluating 3464 (random selection of 50%) of these on title and abstract, 266 were selected for full text screening, of which 67 were deemed relevant. The number of papers needed to screen, 3464/67=51.7 is higher than the NNR calculated in the study where the filters were developed. Three factors may explain the high NNR in this search. Firstly, using a combination of four of our filters instead of just one. For instance, had we used only the NNR-optimised filter, the NNR would be an estimated 26 papers based on the papers we reviewed full text. Secondly, a high false-positive rate of Gomes' filter may have existed. In their review only 53 out of 7594 identified papers were relevant, although they did use multiple searches and databases. Thirdly, the translation of Gomes' filter to PubMed was complicated by the fact that the [adj]-operator lacked in Pub-

Med and was replaced by the AND-operator, resulting in more hits). Although a broad search (thus expecting more false-positive findings) was intended, more careful selection of the disease-specific filter could have reduced at least some of the work in this regard.

Of the 67 papers deemed relevant, 34 were appraised using the appraisal checklist. During this process, an additional 12 papers were excluded due to not covering the analysis or subject well enough. We identified several different outcomes under which we grouped the moderators and predictors described in the remaining 24 papers: factors influencing cares' feelings of comprehensibility, manageability and grief, such as various social and relational aspects; factors influencing location of death, such as living situation and availability of help; and lastly, factors influencing survival in general, but also for very disease and intervention dependent (e.g. hemodialysis in end-stage renal disease).

Overall, the use of the appraisal checklist was complicated due to the nature of some of these papers (descriptive, non-intervention studies). There were many studies that did perform moderator/predictor research but did not report effect sizes (e.g. only reporting significance). The evidence quality was diverse. We were unable to identify studies describing the same moderator or predictor, which is perhaps a sign of the broadness of the research area. It proved challenging to link the found effects to the output of the rest of the case study (e.g. outcomes) due to the fact that much of the evidence was qualitative or related specific interventions or outcomes. A better disease-specific search filter, with an explicit intervention and/or outcome contained within, may have solved this issue. On the other hand, the methods did identify several moderators/predictors that would not have come up when looking at obvious outcomes, underlining the importance of this guidance in HTAs.

## 2.4 CONCLUSIONS

The search filters presented in this section help to retrieve moderators and predictors of treatment effects in the medical literature. The checklist presented in this section will support the appraisal of moderator or predictor claims in medical literature. It will help in assessing the value of these claims and whether they should be taken into account in an HTA.

To our knowledge, the search filters presented here are the only ones that have been specifically developed to guide the retrieval of articles on moderators and predictors of treatment effects without pre-specification of these moderators or predictors. It is therefore complementary to the PubMed Clinical Query for clinical prediction rules (Auston, 2005). The systematic creation and selection of search filter keywords is a great strength, although external validity (i.e. outside the scope in which the filters were created) has not yet been tested. One remaining challenge in using the search filters, especially when dealing with complex technologies, is the creation of the disease/setting specific search filter. The identification of relevant literature is more likely to be of practical significance if the topic of interest and any relevant interventions can be clearly defined. Without such a definition, relevant studies may be missed or more irrelevant studies may be picked up.

The appraisal checklist was created by combining knowledge from the literature as well as knowledge from experts working in the methodological area. The testing phase ensured usability in the area it was going to be applied as well. The fact that we did not develop an overall quality score could be a reason for critique. However this was explicitly decided, since creating such an overall score brings along many limitations of its own. We did add an overall judgement to structure a summary of problems encountered in a study. The strength of this appraisal checklist lies in the fact that it enables the user to appraise a study without requiring in-depth knowledge. In addition it will direct discussion about the influence of specific risks of biases in specific cases. It was not possible to create a critical appraisal tool that could assess the validity of a study on a detailed level. We do think the level of detail presented in the current guidance is sufficient for policy research of HTA.

There is an increasing demand for more stratified medicine. This guidance intends to contribute to this development by aiding in the identifying the factors that are claimed to influence treatment, the quality of these claims and their consequences. The number of factors that are identified to influence treatment effects is growing. As more factors are identified, we need to be able to make better decisions on which factors are important enough to serve as a basis for patient care. This relates also to the efforts (in kind or resources) that need to be taken to identify a specific moderator or predictor in clinical practice in relation to harms that can be prevented by identifying that moderator or predictor. For more information on this process, see the section on integration in this guidance (chapter 4).

# 3 THE RETRIEVAL AND CRITICAL APPRAISAL OF LITERATURE ON PATIENT PREFERENCES FOR TREATMENT OUTCOMES

*By: Ralph van Hoorn, Wietske Kievit, Andrew Booth, Kristin Bakke Lysdahl, Pietro Refolo, Dario Sacchini, Kati Mozygemba, Bjørn Hofmann, Ansgar Gerhardus, Lisa Pfadenhauer, Marcia Tummers, Gert Jan van der Wilt*

## 3.1 INTRODUCTION

### 3.1.1 Purpose and scope of the guidance

#### *Aim of this guidance*

The aim of this guidance is to provide a systematic approach to identify patient preferences for treatment outcome (PPTO). These preferences can either be retrieved from the literature or elicited from patients. Information on PPTOs can be used in the evaluation of health technologies such as treatments or protocols, to determine research agendas or to evaluate the patient-perceived value of treatment outcomes used in trials for systematic reviews or Health Technology Assessment (HTA). This guidance describes the steps needed to retrieve PPTOs, provides an overview of the most commonly used methods to elicit them, and describes the appraisal of studies where such methods are used.

#### *The added value of this guidance in relation to existing guidances*

It is acknowledged that it is important to take individual patient preferences for treatment outcome into account. In the literature on guidelines it is often mentioned that patients should take part in their development (for instance, NICE has a patient and public involvement policy). However specific guidelines on literature retrieval, appraisal and elicitation of PPTOs are lacking.

### 3.1.2 Problem definition

Measuring preferences is not as straightforward as measuring blood pressure or temperature, as there is no single tool or gold standard to do so. One way of measuring preferences is by offering subjects choices in real or hypothetical situations and monitoring the results. For individual cases Decision Aids (DAs) may be used: these are forms which inform patients on the treatment options and ask questions to come to a decision (Stacey et al., 2011). Decision aids are valuable to determine patient preferences on a personal level. However, it is difficult to extrapolate the choices made with DAs to groups and not possible to use the results for other settings or diseases than the one(s) for which the DA was designed.

This is partly because preferences are sometimes regarded as 'constructed' (Slovic, 1995; Hoeffler & Ariely, 1999): A preference is made explicit only after a subject is asked to make a choice. The preference can be influenced not only by the environment in which the choice is made, such as experiences or socio-demographic characteristics of a person, but also by the method by which it was elicited (Jansen et al., 2000; Merlino et al., 2001; Jung et al., 2003; Lloyd, 2003; Anderson & Mellor, 2009). Under similar circumstances, the same subject or method may yield different preferences. The framing of options, as well as the alternatives that are offered can also greatly influence the results. The inability to directly measure, model, or validate preferences provides one explanation for the heterogeneity of preference elicitation methods and the lack of a gold standard. This impedes the generation of comparable, group-based preference data on which to base HTAs or clinical guidelines.

#### *Importance of patient preferences for treatment outcomes*

The importance of patient preferences can be illustrated using the example of an HTA of the paediatric cochlear implant. Whereas in the literature, mainly outcomes on hearing and speech were reported, the deaf community was at least equally interested in social and emotional development as an outcome (Reuzel et al., 2001). One explanation for this disparity is that trials, as well as research in general, are often performed to test outcomes deemed important by researchers or physicians (Hanley et al., 2001). However, even physicians, who closely work with their patients, may have an imperfect understanding of which treatment outcomes actually

matter to their patients (Mulley et al., 2012; Muhlbacher & Juhnke, 2013). Thus, interventions may be considered superior in aspects deemed important to medical professionals but not to patients. The value of interventions should, therefore, also be established from the viewpoint of the target population, i.e. the patients. Including patient preferences in HTA can increase the (public) acceptance of health policy, increase transparency and legitimacy by involving stakeholders, and is therefore an essential part of good HTA practice (Bridges & Jones, 2007).

### *When are patient preferences for treatment outcomes important?*

PPTOs play an important role in every aspect of healthcare, but the elicitation of patient preferences for treatment outcomes may not always be appropriate. Examples of preference-sensitive decisions are medical treatments that improve one condition (e.g. pain) while worsening another (e.g. nausea), or treatments where benefits and harms need to be weighed (Boyd et al., 2012). Examples of decisions where preferences may be less important or practical to elicit are the acute setting (i.e. first aid) or when treatments seem to have overwhelming beneficial effects. However, even such treatments may be subject to preferences (e.g. life-saving blood transfusions for Jehovah's Witnesses (Lin et al., 2012)). Not acknowledging these preferences in clinical practice could result in over- or underestimation of the value of treatments. Decisions that involve risks, side-effects, quality and length of life, and financial considerations are mostly preference sensitive (van der Weijden et al., 2010). There is a Decisional Conflict Scale that can help to indentify preference-sensitive situations (0'Connor, 1995).

## 3.2 GUIDANCE DEVELOPMENT

This section of the guidance can be used to:

- retrieve patient preferences for treatment outcome from the literature;
- to appraise relevant literature;
- and to perform primary research on patient preferences for treatment outcome.

### *Development of the search strategy*

In order to find relevant literature in PubMed on PPTOs a set of search filters was created. The process is similar to the development of filters for finding moderators and predictors. For the development and validation of the search filters two steps were taken: [1] a comprehensive set of search terms and combinations of terms was constructed and [2] the results of these combinations of terms were tested in a set of relevant papers.

A selection of journals was hand searched to identify relevant articles on patient preferences for treatment outcome. All articles, published in the year 2011, were manually screened to determine whether they contained information on PPTOs. The entire set of articles was randomly (1:1) divided into a development set (in which the search strategies were developed) and a validation set (in which the search strategies were tested). The articles deemed relevant in the development set were submitted to PubReMiner, an online resource to retrieve all search terms related to the submitted articles. Using a computer algorithm, search terms were tested and combined to create search filters. The applied methods follow accepted good practice in search filter creation (White et al., 2001; Jenkins, 2004).

Four sets of PPTO search filters were generated. A sensitive set was created for use when relevant literature is expected to be scarce or when the other filters do not return enough relevant literature. A set of specificity-optimized filters was constructed to minimise retrieval of irrelevant articles, at the cost of excluding some relevant studies. Though these filters may miss a few relevant articles, they are a good starting point if the likely effect of missing relevant literature is not considered critical (e.g. given a large amount of relevant literature available). The Number Needed to Read (NNR) optimized filters aim to return a set of articles which contain low numbers of irrelevant papers. Finally, a set of accuracy-optimized filters was created to mitigate the effects of incorrectly included and incorrectly excluded papers. The choice of filters may depend on the broadness of the problem under investigation. It should be fairly straightforward to test the sensitivity-optimised filters first and then to fall back on the other filters if too few papers are excluded due to the filter. Where a topic is narrow or where too many papers are excluded the order is reversed i.e. from specific to sensitive. Clearly, the choice will depend on the time-constraints and needs of the user. The PPTO search filters in this guidance were designed for use in combination with one or more subject-specific search strategies to identify relevant literature for the disease, population or intervention under investigation.

For more detailed information on the development of these search filters, please see van Hoorn et al. (2016, paper in preparation).

*Development of the appraisal checklist*

The aim of the appraisal checklist is to determine whether a study reporting on PPTOs has been executed rigorously and whether the findings are relevant to the HTA research question. In order to create such a checklist, we first explored which methods are used to elicit patient preferences for treatment outcomes. For each method, we tried to identify existing guidance or tools to appraise these methods.

To explore the most common methodologies used to elicit patient preferences for treatment outcome, we analysed the papers identified to develop the search strategy, as well as expert opinion, and additional PubMed and Google Scholar searches. A separate search was performed, for each method found, to identify appraisal criteria specifically for that method. These searches combined method-related search terms with appraisal related search terms, such as 'appraisal' or 'quality'. The search resulted in various studies detailing quality criteria of potential value when appraising studies on PPTOs.

Despite the large variety of methods available to elicit patient preferences (see e.g. Janssen et al., 2014), there is considerable overlap in how data are collected or interpreted between methods. Grouping of appraisal criteria was performed not by methodology but primarily by conceptual background. After the creation of a test version, the tool was tested in a case study and revised based on user feedback.

*Development of primary research on PPTOs*

Despite adequate search and appraisal strategies, research may still fail to supply reliable evidence on PPTO. If this is the case, primary research is the only option to gain insights into PPTOs. The lack of standardization (Opmeer et al., 2010), the diversity of methods and applications, and the lack of evidence on which method is best for which situation (Brett Hauber et al., 2013; Muhlbacher & Juhnke, 2013), makes it impossible to develop a concise guidance on preference elicitation covering all possible methods. However, general considerations on which method to use can be based on method descriptions, critical appraisal criteria and further logical considerations. The final part of this guidance elucidates core decisional criteria that can help users to decide between the different methods.

## 3.3 APPLICATION OF THE GUIDANCE

This section explains how to perform the entire process of retrieving PPTOs for any HTA, including how and when to use the three methods described in this guidance. The three methods comprise: applying a search filter, appraising the identified literature and performing primary research for eliciting PPTOs. Figure 3 outlines the process of identifying and appraising literature.

I. At the onset of the HTA, the researcher should determine in which parts patient preferences for treatment outcome may play a role.

II. If there are indications that different preferences might result in choosing different treatments, and therefore different treatment outcomes, it is advisable to explore the literature. The search guidance can help to reduce irrelevant articles in a body of evidence by filtering out papers unrelated to PPTOs. See chapter 3.3.1.

III. The appraisal checklist can be used to determine the study quality of the evidence on PPTOs. Based on the results the overall evidence quality and quantity can be determined. See chapter 3.3.2.

IV. Based on the amount and quality of the evidence, the user should make a judgement as to whether the body of evidence is sufficient for his or her purpose. Does the evidence help to choose the right treatment for a specific patient? If this is not the case, the user can return to the second step to search for more literature, or if no additional literature can be found, move on to step V to create additional evidence.

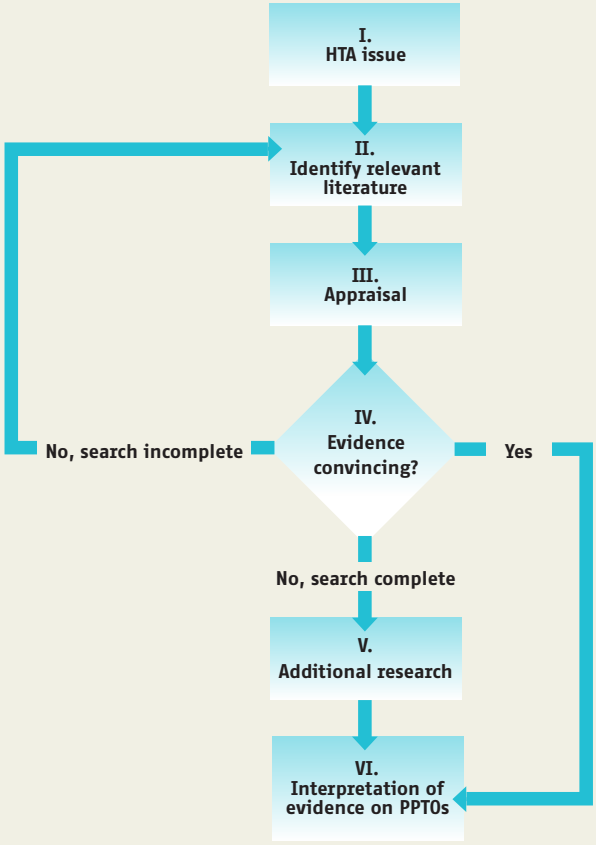
V. New primary research can be used to elicit PPTOs directly. To this end, considerations for which method to use and descriptions and references of the most common methods are listed. See chapter 3.3.3.

VI. The final step is to interpret the available evidence. Depending on the needs and available evidence, this step may consist of summarising and comparing evidence, reflecting on the evidence and translation into recommendations or conclusions. See chapter 3.3.4.

### 3.3.1 Search guidance

The user starts with a search query that captures articles relevant to the field of investigation, for instance the disease or intervention being researched.

*Figure 3: Outline of process to identify patient preferences for treatment outcomes.*




One usually starts by making an inventory of relevant keywords of the field using expert knowledge, and combining these terms using the OR-operator (though other operators may be required to restrict to a certain area only identifiable by combinations of terms). PubMed's Clinical Queries (http://www.ncbi.nlm.nih.gov/pubmed/clinical) can be used as a starting point.

This field specific search query can then be combined with one of the search filters presented in Table 3 to retrieve evidence on PPTOs using the AND-operator. In Table 3 four sets of three search filters are listed: each set of three is the top three optimized for the respective performance measure. The filter optimized for sensitivity will return a relatively large share of irrelevant papers but is least likely to miss papers containing the required information. Search filters with high accuracy, specificity or low number of papers needed to read (NNR) will return less irrelevant papers at the cost of possibly missing important information. Which of the four strategies is the best ultimately depends on the amount of usable retrieved papers and the amount of time the user is willing and able to invest.

Once a number of articles is found, it should be determined whether they truly contain the information the user requires by examining them by abstract and full-text.

### *3.3.2 Appraisal of the literature*

To appraise literature describing PPTOs an appraisal checklist was created. The checklist can be found in the appendix (chapter 7.4). It consists of six questions and an in-depth explanation of the rationale behind each question to assist in appraising studies concerning PPTOs. By answering the individual items, users should be able to identify relevant quality issues. It is up to the user how to deal with such problems. The items in the checklist should be regarded as a set of key quality indicators: the more these criteria are met, the greater the likelihood a study was adequately performed. They should not be used as strict criteria, as many variations in research concerning PPTOs are possible and valid. In-depth knowledge on the specific methods used is often required for the appraisal of specific aspects or to determine the appropriateness of the method. The checklist was not designed to summarize study quality findings or provide overviews on study quality.

If multiple studies on the same subject exist and need to be appraised (i.e. a body of evidence) we refer to the GRADE (Grading of Recommendations Assessment, Development and Evaluation) criteria and CERQual (Confidence in the Evidence from Reviews of Qualitative Research), both of which can be found on http://www.gradeworkinggroup.org/ [accessed Oct 2015]. GRADE is a tool that can be used to appraise a body of evidence for use to support guideline development or HTAs. It is not specific for PPTOs, but will help to appraise a body of evidence when combined with the checklist described in this guidance for the appraisal of individual papers. CERQual is a tool that is currently still being operationalized. It aims to aid in the appraisal a body of evidence from qualitative research on four criteria: methodology, relevance, adequacy of data and coherence. A combination of our checklist, GRADE and/or CERQual, should allow the assessment of any study on PPTOs towards HTAs or guideline development.

*Table 3: Search filters for articles on patient preferences for treatment outcomes. From: van Hoorn et al. (2016a, paper in preparation).*

| Search term | Se (%) | Sp (%) | Ac (%) | NNR |
|---|---|---|---|---|
| **Optimal sensitivity** | | | | |
| "Patient Satisfaction"[mesh] OR prescrib*[tiab] OR experiment*[tiab] OR preferen*[ti] | 100.0 | 93.8 | 91.5 | 30.7 |
| choice* OR practitioner*[tiab] OR preferen*[ti] OR prescrib*[tiab] | 100.0 | 93.8 | 87.6 | 44.6 |
| "Patient Satisfaction"[mesh] OR ask*[tiab] OR preferen*[ti] OR prescrib*[tiab] | 100.0 | 93.6 | 86.9 | 46.9 |
| **Optimal specificity** | | | | |
| "Patient Preference"[mesh] OR adheren*[tw] | 78.6 | 96.3 | 96.3 | 17.5 |
| choice* OR practitioner*[tiab] | 78.6 | 96.0 | 96.0 | 18.9 |
| "Patient Satisfaction"[mesh] OR preferen*[tiab] | 78.6 | 95.9 | 95.9 | 19.2 |
| **Optimal accuracy** | | | | |
| "Patient Preference"[mesh] OR adheren*[tw] | 78.6 | 96.3 | 96.3 | 17.5 |
| choice* OR practitioner*[tiab] | 78.6 | 96.0 | 96.0 | 18.9 |
| "Patient Satisfaction"[mesh] OR preferen*[tiab] | 78.6 | 95.9 | 95.9 | 19.2 |
| **Optimal NNR** | | | | |
| "Patient Preference"[mesh] OR adheren*[tw] | 78.6 | 96.3 | 96.3 | 17.5 |
| "Patient Satisfaction"[mesh] OR prescrib*[tiab] OR preferen*[ti] | 92.9 | 95.3 | 95.3 | 18.8 |
| choice* OR practitioner*[tiab] OR preferen*[ti] | 85.7 | 95.7 | 95.6 | 18.8 |

*Se: Sensitivity; Sp: Specificity; Ac: Accuracy; NNR: Number Needed to Read*

### 3.3.3 Primary research

New research on PPTOs may need to be conducted if no relevant or usable evidence is found, for instance if the evidence is not specific or transferrable. For HTA, preferences are used to support guideline development or assessment of existing methods. Integration of patient preferences for treatment outcome in medical care involves active preference elicitation from individual patients. Below a step-by-step approach to determine patient preferences for treatment outcome is described.

#### Step 1: Qualitative and/or quantitative data requirements

The first step consists of determining the goal of the research to determine whether qualitative, quantitative, or mixed methods are required.

Qualitative methods are generally used to explain findings, create new theories, making sense or interpreting phenomena in terms of how people perceive them. In the context of PPTOs, they can be used to compile inventories of what patients consider important or to make sense of why certain outcomes are considered important or not (to explore/predict heterogeneity or relations with other phenomena). They can be used to identify gaps in research or generate new hypotheses.

Quantitative methods can help determine the strength of preferences, relate them to each other (e.g. rank or prioritize PPTOs), or measure and statistically test relationships between patient characteristics and certain preferences. Using models, preferences for certain situations and groups of patients can be predicted (to a certain extent). The choice for a specific method should be based on the required product (i.e. summarizing or more exhaustive). The choice between qualitative, quantitative methods or a combination is generally determined by what research is already available, what type of data is required and the resources available to the researcher. As the appraisal criteria indicate, experience of the researcher with the chosen method is of special importance in qualitative methods. Mixed methods require extra resources and knowledge of both quantitative and qualitative methods, while researchers are often only well-versed in one type (Creswell & Plano, 2007). Researchers interested in conducting mixed method research are encouraged to consult additional resources (see, for example, Creswell & Plano, 2007; Dellinger & Leech, 2007; Barnett-Page & Thomas, 2009; Onwuegbuzie et al., 2009; Creswell & Clark, 2010; Facey et al., 2010; Cameron, 2011).

#### Step 2: Choosing a specific method

The choice of a method includes the choice of a model or theory on which to base the analysis of the data (e.g. grounded theory for qualitative methods, a specific statistical model or empirical design for quantitative research). Chapter 7.4 provides an overview of frequently used preference elicitation methods. This overview can help in choosing a method.

#### Step 3: Determining the target population

When choosing a target group from which to elicit PPTOs, the transferability of any findings must be considered. The research population should be similar to the population in which the results will be used. The criteria listed under **item 6** ('Are the results transferrable to my patient(s)?') in the appraisal checklist (chapter 7.4) may help in this regard. The number of participants that is sufficient will depend on the method and research question (see, for examples for DCEs and interviews, Francis et al., 2010; de Bekker-Grob et al., 2015). A researcher should make sure that the group is sufficiently large to represent all possible subgroups that may be applicable according to the hypothesis under investigation.

#### Step 4: Choosing a mode of administration

As most quantitative methods depend on larger number of respondents, electronic administration modes (e.g. email, online questionnaires) or paper-based modes are often chosen. They offer more standardized input and analysis, which helps in quantifying results. Interviews and focus groups benefit much more from face-to-face contact, especially when the method relies on inter-personal communication. Not only will the feedback loop be much shorter, it will also benefit from non-verbal communication. Ultimately the researcher is free to choose a mode of administration, but some care must be given towards how well the mode of administration, and the exact execution (e.g. wording of questions) fit the research design. The background information of the appraisal criteria in the appraisal checklist (chapter 7.4) can help in this aspect.

### 3.3.4 Interpretation of evidence

The final phase in the process of to identify PPTOs (see Figure 3) is possibly the most difficult. It involves the assessment of the evidence accumulated through the previous phases in view of the possible applications. This phase consists of addressing a number of questions:

1. Do the available data represent clear and comprehensive evidence of the outcomes that may be expected to result from the use of the technology under study?
2. Do the available data improve our understanding of how the various outcomes are valued by patients?
3. Is there any evidence to suggest that patients may substantially differ from each other in this respect?

First, it must be determined what the outcomes described in the evidence are and how they relate to the expected outcomes. The quality and integrity of the study findings as well as estimations on its validity in the target context also need to be determined. The appraisal step should result in an answer to this question. Once answered, the user must decide whether the results have additional value (over what is already known) on the knowledge about the preferences that are under investigation. Relevancy and (clinical) importance of the results as well as the relevancy of the research question and methods used are important considerations here. If so, the evidence can be used to weigh treatment outcomes (e.g. to assess effectiveness on multiple outcomes) or prioritize research (by focusing on the most prudent problems).

Once it has been determined that the evidence contains relevant PPTOs, it must also be determined whether these preferences differ between patients. This can lead to a judgment whether a patient-centred approach to the use of the healthcare technology should be recommended. In order to answer this question, one should look for any correlation between patient characteristics (including that of their environment, social context, etc) and the preference weights found in the study. Note that the question is formulated as 'evidence to suggest' instead of 'evidence that proves'. Differences in preferences may be quite subtle, hence not likely to result in significant or convincing (statistical) evidence. It would be wrong to dismiss such findings solely based on statistical results. There may be additional factors that make it likely that patients think differently on outcomes, which should be taken into account when determining the evidence can be used to recommend a patient-centred approach.

If heterogeneity in PPTOs is found and the difference is quantifiable and accurately estimated, the heterogeneity could be used to estimate different weights for different groups of patients. In cases where the heterogeneity is only assumed to exist, or found but without accurate quantifiable estimations, the implication may be that HTAs should incorporate the heterogeneity as stratification-factor, or at the very least include it as limitation for its conclusion. Larger heterogeneity may also be considered a sign that the PPTOs may not likely result in accurate estimation of patient preferences and thus should be used only with extreme caution in HTAs, or advising personalised medicine to base decisions on decision aids instead of reported literature on PPTOs.

### 3.3.5 Application of the search strategy, appraisal checklist and primary research in a case study

The methods described in this guidance were applied to a case study concerning reinforced models of palliative home care. We applied the search filters to find information on PPTOs relating to (non)reinforced models of palliative care and appraised the identified papers. Although we did not perform new research to elicit PPTOs, we will describe possible applications within the setting of home-based palliative care.

#### *Exploring evidence relating to PPTOs*

The search filters, as presented in Table 3, were used to explore what is known on PPTOs concerning home-based palliative care. The starting point was a disease-specific search strategy developed by Gomes et al. (2013) to find papers relating to models of home-based palliative care. We translated their search query so it could be used in PubMed and removed the study type restrictions present in the query as these were too strict for finding studies on PPTOs. The resulting query was combined with the AND-operator to a combination of the best four filters. To this end the best four filters (the ones with the most optimal Se, Sp, Ac and NNR) were combined using the OR-operator. It was decided to use all four filters in order to maximize the number and diversity of returned papers.

By applying the search filters we indentified a total of 1606 potentially relevant papers. After evaluating a random selection of 50% of these articles (n=803) on title and abstract, 37 were selected for full text screening, 24 of which contained information on patient preferences

for treatment outcomes. The list of selected papers on PPTOs can be found in the case study report (Brereton et al., 2016). It could be concluded that use of the search filter helped to retrieve papers on PPTOs relevant in the context of reinforced models of home-based palliative care. We needed to screen on average 803/24=33.5 papers in order to find one paper that was considered relevant to the question of PPTOs in this specific case. This number is in the order of magnitude one would expect considering the use of the best filter of each of the four categories and their respective NNRs in Table 3 (ranging from 17.5 to 30.7).

### *Appraisal of evidence*

The appraisal checklist was applied to 24 papers that were retrieved using the search strategies and that were considered relevant after full-text examination. We found a number of preferences that were applicable to home-based palliative care. There was a mix of qualitative research (mostly telephone-based interviews) and quantitative research (for instance variants of discrete choice experiments).

We identified three groups of preferences concerning home-based palliative care: quality of life versus quantity (length) of life, location of death and treatment-specific care goals. The literature showed that most patients valued quality of life over quantity of life. However, depending on cultural background differences do exist. The same applies to preferences on the location of death; if patients expressed a preference, this was usually a preference to die at home. In some papers patients reported that other factors of the treatment, such as the reduction of emotional or physical burden on family and symptom relief), were of higher importance than location of death. These findings suggest that an individual assessment of a patient's preferences would be appropriate. A more detailed description of the results can be found in the INTEGRATE-HTA case study report (Brereton et al., 2016).

### *Example of primary research relating to home-based palliative care*

Using the case study as example, two types of PPTO data could be retrieved: exploratory findings and confirmatory findings. The first one concerns the exploration of relevant treatment outcomes. For instance, patients, but also caregivers, professionals and others who are active in the field of home-based palliative care, could be interviewed to find out which problems they encounter. An example of how this could be done is described in the INTEGRATE-HTA case study report (Brereton et al., 2016).

A more quantitative type of PPTO data concerns the relative importance of these outcomes. When interventions are assessed on multiple outcomes, the importance of individual outcomes can help determine an overall best option. A discrete choice experiment can be set up, where health states are compared with each other. These health states should be described using various levels of all outcomes deemed relevant in the earlier exercise. By asking patients or relatives to compare pairs of health states described with random outcome levels, individual weights related to the importance of these outcomes can be determined.

## *3.4 Conclusions*

Even though patient preferences are widely regarded as an important part of HTA, there is no consensus on how to take this into account when producing evidence-based guidelines. The methods described in this section of the guidance can be used to retrieve patient preferences for treatment outcome from the literature; to appraise relevant literature; and to perform primary research on patient preferences for treatment outcome.

Preference elicitation methods, whether concerning treatment outcomes or in general, are diverse and continuously evolving. There is no one-size-fits-all solution to elicit patient preferences. Methodological reviews identified different problems and proposed solutions but no definite answers (Murphy et al., 1998; Ryan et al., 2001; Facey & Hansen, 2011). This guidance makes no claim of being exhaustive, but it does cover relevant aspects of the methods currently in use. We believe this document will help those unfamiliar in this field of research to master the basics, and direct those who require further information to existing resources. As we demonstrated in the case study, the number of articles that are retrieved from the literature when searching for PPTOs may be quite large. This step, then, does add to the workload of conducting an HTA. In view of the importance of taking a more personalized approach to healthcare, we think that the extra effort can be justified.

Acknowledging patient heterogeneity coincides with the personalised medicine movement that advocates for patient-centred research. Improving the use of patient preferences for intervention outcomes in research and assessments enables a stratified approach in healthcare and a more patient-focussed comparison of technologies incorporating patients' views. As a result, it may help to reduce wasteful research (Chalmers et al., 2014).

# 4 GUIDANCE ON THE INTEGRATION OF MODERATORS OF AND PATIENT PREFERENCES FOR TREATMENT OUTCOMES

*By: Ralph van Hoorn, Gert Jan van der Wilt, Marcia Tummers, Wietske Kievit*

## 4.1 INTRODUCTION

### 4.1.1 Purpose and scope of the guidance

#### *Aim of this guidance*

This guidance describes a framework towards the effect estimation of moderators or predictors on treatment outcomes, as well as the effects of eliciting patient preferences on overall treatment valuation. This guidance can be used to assess the additional value of including these factors in decisions instead of having all patients receive the treatment that is considered the best for the entire population. This is important as treatments may only be effective for a specific group of patients, or effectiveness of treatments may be different if moderators, predictors and preferences are taken into account. At the moment this is insufficiently acknowledged in HTAs. With this guidance we aim to offer a framework that ultimately should help focusing these problems in HTA.

The assessment of interventions should incorporate moderators and predictors of treatment effects as well as preferences for treatment outcomes to allow evaluation of technologies on their fullest potential. However, the amount of factors which could potentially influence clinical decision making is large, especially for complex technologies. Each additional factor included in assessments or decision making complicates the evaluation further and will incur additional costs. Ultimately, including such factors in decision making may not in all cases outweigh the costs associated with determining the factors. Hence, there is a need to determine the added value of using additional diagnostics, patient characteristics and elicitation of patient preferences in such situations. A first step in doing so is to collect available evidence described in the literature. Information on moderators and predictors that are found to (possibly) influence treatment outcomes and their value can be retrieved from the medical literature (see previous sections). The assumption that these factors influence a treatments' outcome significantly needs to be tested by comparing the value of adding factors in the decision process to a one-size-fits-all decision. The framework described in this guidance allows users to do so in a systematic way.

#### *The added value of this guidance in relation to existing guidances*

The importance of including information on patients' preferences and moderators of treatment effect when evaluating healthcare technology has been widely acknowledged. For instance, both issues are covered by the National Institute for Health and Care Excellence (NICE) guide to the methods of technology appraisal (2013). Other guidances that aim to support the implementation, assessment or decision making on complex issues, such as the EVIDEM framework (http://www.evidem.org) or MRC guidance (Moore et al., 2015) also acknowledge patient-related factors that may complicate evaluations. These guidances describe how evidence relating to these factors may be evaluated and used in the assessment, and how to ensure these factors are part of the evaluation process. However, the direct comparison of personalized versus a 'one-size-fits-all' approach has not been explicitly addressed. This is, however, a necessary consideration in both clinical practice and medical research related to the implementation of personalised health care. This type of research is related to value of information (VOI)-analysis where the value of information (e.g. diagnostics) is assessed in view of possible gains for patients. For instance, to prioritise research or support decision making (Claxton & Sculpher, 2006; Eckermann & Willan, 2007). Conceptually, the methods described in this guidance are similar to VOI, but the analyses are different. The difference being the fact that VOI analyses is based on analyses on the uncertainty in the input parameters, while we propose a different approach to combinations of scenario analyses.

### 4.1.2 Description of theoretical background and available approaches

Presumably, two developments have contributed to an increased interest in a more personalised approach in healthcare. Firstly, there seems to

have been a trend towards a more prominent role of the individual in Western societies (Beck, 2002). This relates to the trend of participatory medicine and one way to implementing this is to incorporate PPTOs in clinical decision making. Secondly, technological developments enable for the collection of a vast amount of data from patients through the use of technologies such as whole exome sequencing and technologies that can be used to measure a wide variety of biomarkers (Murdoch & Detsky, 2013).

A more personalised approach to medical decision making is often regarded as improving quality of care, efficiency of care, and saving costs (Isaacs & Ferraccioli, 2011; Jakka & Rossbach, 2013). However, before implementing a more personalised approach, it is important to know what the added value could be of such an approach compared to a 'one-size-fits-all' approach. Furthermore, moderators or predictors may be scientifically valid, significant and relevant, but using them in clinical decision making will incur costs. For instance, pharmacogenetics is based on the idea that identifying the genetic profile of a patient can help decide which pharmaceutical agent is likely to work best for a patient (Kalow, 2002). Determining the genetic profile of large numbers of patients may further add to the costs of healthcare. It stands to reason, therefore, that the costs of procuring this type of information must be weighed against the benefits in the assessment of health technologies.

The same holds for patient preferences. Eliciting preferences requires that patients are informed of the various treatment strategies and possible outcomes, and are assisted in making up their mind about how they value these outcomes. On the other hand implementing shared decision making may result in a better compliance, a sense of control or a better disease control but the associated costs need to be compared with the benefits. In the end, the development of such a tool may be costly and use of such a tool requires extra time from a clinician. It may be questioned whether the benefits outweigh these costs.

The addition of moderators, predictors and patient preferences complicates HTA significantly. However, especially for complex interventions, the influence of these factors may severely impact the outcome of HTAs and therefore important to assess. This applies even for apparently less complex interventions where such effects may easily be underestimated.

## 4.2 GUIDANCE DEVELOPMENT

This guidance describes how to assess the value of individual moderators, predictors, and PPTOs by comparing their effect on modelled outcomes. The assessment of the value of knowledge in (health) economics is known as Value Of Information (VOI)-analysis. Such analyses can inform on the expected (health) benefits when obtaining more information on a decision to be made. A requirement of VOI-analysis is knowledge on the effects of all considered decision options; VOI-analysis are often based on models that predict decision outcomes. By extrapolating the effects of a more personalised treatment towards the entire population, HTAs can be informed on the effects of adding the factor, and simultaneously better inform effectiveness estimations.

One example of VOI-analysis is determining the cost-effectiveness of a diagnostic tool. As this information is often only part of the uncertainty concerning a decision, this information is often called Expected Value of Perfect Partial/Parameter Information (EVPPI). With EVPPI it is assumed that the model structure is correct, but that the knowledge of the value of a specific model parameter or set of model parameters is of interest. Most commonly, EVPPI is determined in a model where outcomes for decisions are predicted by comparing the most optimal decision option for every patient including a specific parameter of interest, with the most optimal decision option for the entire modelled cohort. This way an estimation can be made of how much more optimal the decision could be made using that parameter. EVPPI may be useful to determine the value of moderators and predictors, however focus in this guidance lies on the added value of the parameter in decision making, not in the added value in view of reducing uncertainty.

### *Guidance design*

The design of the guidance started with formulating a number of desiderata. The goal was to develop a framework for a model that integrates moderators of treatment effect with patients' preferences for treatment outcomes, allowing for identification of optimal treatment strategies and comparison with 'one-size-fits-all' strategies. Furthermore, the model used for integrating the factors in this guidance should be accessible, versatile, able to handle many different types of data (e.g. different distributions, interval or nominal data), incorporate parameter

uncertainty and able to handle complex interactions. At the same time it should be valid for the targeted use and not too complicated to build using standard spreadsheet software.

An exploratory literature search for 'medical decision modelling' and 'prediction model' in PubMed and Google Scholar was used to identify methods appropriate for modelling exercises. The systematic review performed by Brailsford et al. (2009) was used as a starting point. This review identified several methods for modelling and simulating in health care. Most of the methods were related to estimating outcomes and included Markov modelling, Monte-Carlo simulations, and various advanced modelling and simulation methods. We investigated these models to see what the weak points and strong points of each model were in relation to each other, and up to what extent they could possibly be used as part of the framework presented in this guidance. This knowledge was expanded by collecting experiences from building such a model. Using that knowledge, we describe a list of prerequisites that any model should adhere to in order to be usable for our purpose.

## 4.3 APPLICATION OF THE GUIDANCE

### 4.3.1 General description of the framework

The framework to integrate evidence on moderators, predictors and PPTOs can be broken down into a number of steps. These steps deal with the collection of data (step 1); the construction of a model to estimate treatment effects (step 2); the sequence of moderators, predictors and PPTOs to evaluate in the framework (step 3); the evaluation of the model (step 4); and the optional step 5 to determine the robustness of the model input in scenario or sensitivity analysis. The entire process is outlined in Figure 4. Throughout this section we will refer to moderators, predictors and PPTOs as 'factors'.

Before the individual steps of the framework are described, the model to estimate the effects of considered treatments will be shortly introduced. This model (which will be constructed in step 2 as model M1 (base case) and adapted in step 4 as model M2) is the core of the framework.

### 4.3.2 General description of the model for effect estimation

The setup of the model for effect estimation is outlined in Figure 5. In this model, a population of patients can receive treatment 1 (e.g. usual care) or treatment 2 (e.g. a new intervention). The effectiveness of both strategies is compared by estimating the relevant treatment outcomes of each treatment using the treatment main effects. Later, these estimations can be made more accurate with the effects of moderators, predictors and patient-preferences interacting with patient characteristics. Interactions between the different components can be added to the model as well (none are displayed in Figure 5 to conserve tractability). The steps in the next chapter describe how to construct the model, how to expand the model with factors, and how to compare the model(s) to assess the value of each factor.

Although the model can be made in any (statistical or spreadsheet) program, the preferred option is a program that can be used to perform automated sets of analysis in case one needs to deal with uncertainty of parameters (i.e. scenario or sensitivity analysis).

### 4.3.3 Step-by-step setup of the framework

Below the steps of constructing and analysing the framework are described (steps 1-5 outlined in Figure 4). The chapter concludes by presenting an example from the field of palliative care.

**Step 1: Exploration of the problem**

The first step in generating a simulation model on moderators and PPTOs is the retrieval of evidence and subsequent critical appraisal. Guidance to do this is described in the previous two sections of this guidance. The exploration of important outcomes should be based on all those outcomes that are considered important to any stakeholder (e.g. patients, policy makers). It is useful during the exploration step to make an overview of the outcomes that need to be modelled and their corresponding moderators, predictors, preferences and/or other interactions. For each of the identified outcomes, relevant patient preferences, moderators and effect sizes should be collected (see below). In economic evaluations it is common methodology to use the

Quality Adjusted Life Year (QALY) in all cases (if not directly available, try to find methods to calculate it using other outcomes); however in many cases it will not suffice because no conversion to QALYs is possible or valid. It may be necessary to list an outcome twice or more if the outcome or concept is measured by different instruments or characteristics that cannot be translated into a common metric. In this exercise, costs can be considered an effect as well.

Knowledge about factors that influence an outcome is needed to identify all relevant factors that may influence which treatment is the best for which patient. Quite often one will find literature on such factors lacking in the way the outcome or the factors themselves are measured (using different scales or instruments), or in the way the literature assumes they interact (e.g. the type of regression model that was used, the interactions that were included in the model). Additionally, something that applies specifically for complex interventions, are the interactions or relations between different factors. Sometimes these interactions are known, but cannot be quantified. Hence, a procedure where such data is combined in a single model to estimate the value of its individual components is very data hungry.

### *Moderators and predictors for treatment effects*

Once a list of relevant outcomes has been compiled, it is necessary to identify which predictors and moderators interact with each of these outcomes (as well as with each other). Relevant moderators/predictors are all moderators/predictors that influence

*Figure 4: Process of evaluation.*

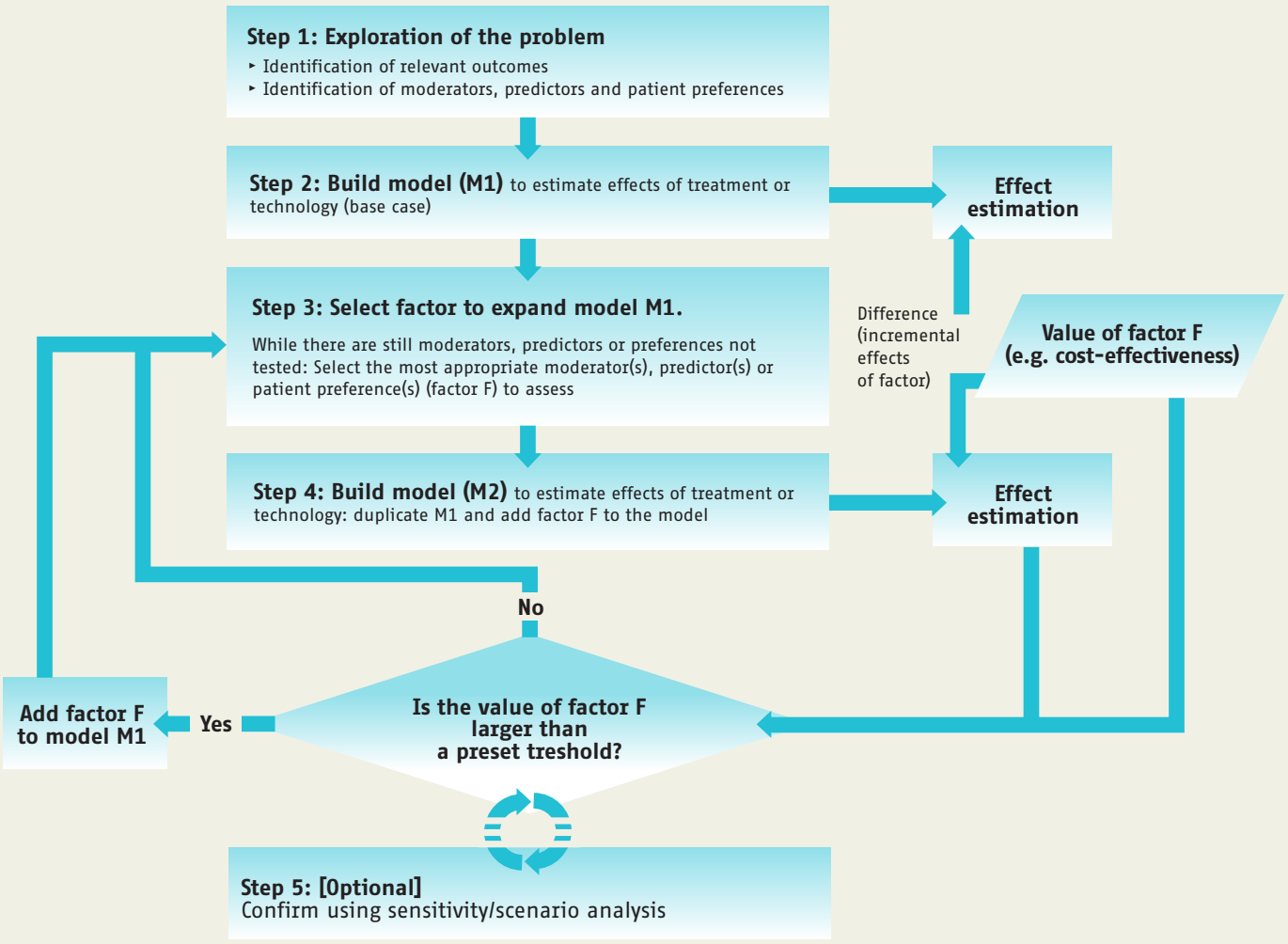
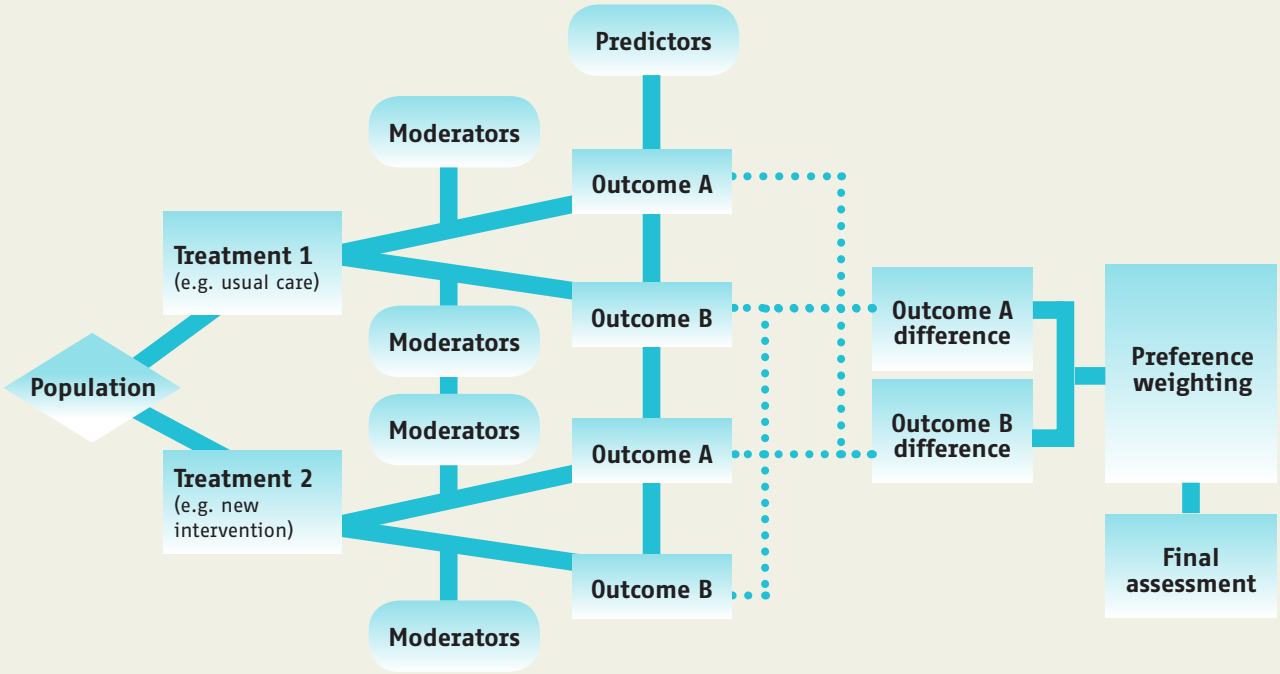


*Figure 5: Setup of the model.*




the occurrence (chance or effect size) of a certain outcome. Often one will find moderators/predictors investigated in multivariable regression analysis. In this case all variables of the final multivariable regression model should be used (not only the significant ones); without the non-significant variable the accuracy of the model will be different.

### *Patient preferences for treatment outcomes*

PPTOs can be used to weigh individual treatment outcomes simulated in the model, in order to compare treatments on multiple outcomes. In order to do the latter, preference data is required that compares every treatment outcome with one of the others in the list. The type of evidence required for the model are weights attributed to specific outcomes such as usually retrieved from discrete choice experiments (DCE) or other types of multi-criteria decision analysis (MCDA). It is also possible to use evidence describing the percentage of patients preferring outcome A over outcome B, but then assumptions have to be made on the strength of preference for these patients.

Model input should include as much information as available. This means that input based on distributions is preferred over point estimates. In order to properly model such data, the following information is required: [1] which distribution describes the variable best (e.g. normal-, binomial-, or Weibull-distribution); and [2] a number (usually two) parameters used to describe the distribution (e.g. a mean and standard deviation, n and p, or alpha and beta). The problem of not using such stochastic input for the model is the inability to account for variations of the parameter, making the outcomes lest robust and therefore less informative. Parameters that can vary widely may be found to have large effects on the overall outcome, or parameters with small variability may be found to have small effects on the outcome. The opposite may also be true. Generally, being able to identify and quantify these influences helps to make accurate estimations of the influence of parameters on their outcome by providing confidence intervals or similar statistics. This can be considered as main outcome (probabilistic sensitivity analysis, where outcomes are generated as distributions instead of single measure) or as secondary analysis (step 6).

**Step 2: Build model (M1) to estimate effects of treatment or technology (base case)**

There are many approaches to decision modelling, there is no single model design that is able to support all applications. A report of the ISPOR-SMDM Modeling Good Research Practices Task Force (Caro et al., 2012) provides an overview of different types of models that may be used in these cases. The report describes for instance that the design of a model needs to be described to the user (decision maker) and it should be clear and transparent why the model was designed the way it is, what the origin of the data is, which uncertainties may play a role and how they are dealt with (e.g. scenario or sensitivity analysis). Another consideration they describe concerns the fact that the design of a model should not depend on the data found, but on the (disease) process that is being modelled (for which stakeholders and subject experts need to be involved). Although it is straightforward to apply this to the general structure of the model, many important factors in a model (i.e. interactions or correlations between input and output, the in- or exclusion of moderators and predictors, the weighting of outcomes) most certainly depend on what is known from the literature.

Concerning which type of model to use, the following recommendations were reported:

- Static cohort models, such as decision trees, may be most appropriate for relatively simple and static decision problems.
- Dynamic cohort models, such as state-transition models, may be most appropriate if the disease or process can be expressed as a series of health states (dynamic decision problems, i.e. decisions based on a specific state). Especially with specialized software they can be easy to create, test, analyse and report. Although they have a presupposed independence of transition probabilities on past history, this can easily be addressed by increasing the number of states in the model.
- Dynamic models that simulate individuals including time, such as discrete event simulation models, are preferred if there are stochastic time-to-event intervals.
- Dynamic cohort models simulating individuals, such as micro simulations, are an alternative if the number of health states is very large (e.g. because of a history-dependence of transition probabilities), or the model is very complex or large. They are extremely versatile, but require a little more knowledge to build as there is no preset model structure.
- In specific cases more specialised models may be required. For instance, in the case of interactions between individuals (e.g. infections) disease transmission models can be advised.

Most of these models require specialist knowledge of software, but offer relatively straightforward effects estimations. We therefore recommend to use these only if such knowledge and software is available. The availability of more powerful software and hardware has made micro simulations for larger cohorts much more accessible and popular. They can be simulated in spreadsheet programs as well, making them the preferred choice for this exercise. Micro simulation models are very flexible and have the added benefit of being able to include (properties of) all the aforementioned models. Often hybrids or combinations of model types are used, which can be appropriate for specific decisions. If multiple outcomes are tested on moderators, predictors or preferences, each outcome may have a different model that is most appropriate. In these cases it may be especially beneficial to use a single (micro) simulation model that incorporates all individual types of models over estimating different models for different outcomes. It is recommended to use micro-simulation models as these allow insights into individual patients' effects, allowing, for instance, identification of subgroups and easy integration of preferences. However, the type of decision that is being modelled may still enforce any other type of model as part of this simulation model.

Ultimately, the results of the model need to be presented to different stakeholders and therefore transparency and simplicity are important criteria. That is, the simplest model that can accommodate all outcomes, moderators, predictors and patient preferences for treatment outcomes should be preferred to allow easier communication of model uncertainties and relate the results to assumptions on which the model is based.

All model design considerations need to be related to the validity of the model. Various types of validity of the model need to be checked throughout the design and use of the model. Another report by the ISPOR Modeling task force details about the validation as well as transparency of decision mo-

dels (Eddy et al., 2012). One type of validation that is extra important here is the comparison of the decision part of the model (i.e. which moderators, predictors or preferences are dealt with) in comparison with standing guidelines or protocols. Al other types of validations apply equally well for decision models in the current context. The model described in this guidance is used to explore aspects of interventions that make the evaluation thereof complex. This makes model validation itself a complex issue. Often there is no complete overview, i.e. there is no single person or group of persons who can oversee all relevant issues that could possibly play a role in that context. The consequence of which is, that it is difficult to determine if the model is complete and based on valid assumptions when it comes to its interacting parts. This guidance will address some methods for assessing model validity in the next chapter. For a more extensive model validation, we recommend to validate the model on the four aspects (conceptual model, input data, code, outcomes) mentioned in the AdVISHE tool (Vemer et al., 2014); however at the time of writing this tool is not made public yet.

Once a model design is chosen, there are three separate sections of the model that need to be constructed: main effects, moderator/predictor effects and patient preferences. The base case model consists of only the main effects (i.e. one-size-fits-all: all patients would receive the treatment that is considered the most optimal over the entire cohort). In later steps, this model can be expanded with moderator/predictor effects and patient preferences. Below is described how to construct the base case model (M1) using main effects only. The addition of moderators, predictors and PPTOs is described in the following steps.

### *Main effects*

The choice for which outcomes to model depends on patient preferences (important outcomes are those that are considered important by patients), relevance for decision makers (e.g. costs, number of deaths), but most of all which moderators and predictors are to be investigated. If a single outcome is reported in different studies using different scales which cannot be directly converted then the outcome should be included multiple times. If logical interactions can be made (e.g. quality of life and survival can be combined into a QALY), this should be an additional outcome, not one that replaces the individual outcomes. It is important for the model to be complete and transparent. Lack of data will likely limit the outcomes by which moderators, predictors and preferences are valued. This is especially true for data that helps quantify interactions between different components of a model, such as correlations between patient characteristics or risks on specific outcomes.

### *Model iterations / size*

For some models (including micro simulations), a required number of patients or iterations needs to be determined. The required number of patients (or iterations of a model) is greatly dependant on the number and spread of parameter distributions entered in the model. Setting this number too low and it will cause fluctuations in the results due to the relatively large impact of random sampling. Setting the number too high may overestimate small effects. A starting point could be the number of patients in one of the studies from which data was derived. It is preferred to run the model a number of times while increasing the number of patients stepwise, until the outcomes of the evaluation are more or less stable. Ending up with a simulation of thousands of patients is no exception.

## Step 3: Selection of factor to expand model M1

The previous step describes the creation of a base-case model (i.e. a one-size-fits-all model or alternatively one that includes the variables used in current practice). In step 3 a moderator, predictor or PPTO is selected to expand the base case model with. The sequence in which moderators, predictors (and later PPTOs) are added to the model greatly influences the results. For instance, it may be so that expensive genetic tests result in significant moderation effects, but if patients are first stratified on age and gender, that this only applies in a specific subgroup of patients. The selection of which of these factors is added to the model first, can be made by:

- feasibility of the assessment in patients. For instance, age or gender are easy to determine and may have great influence;
- if specific factors are generally determined together (e.g. systolic and diastolic blood pressure) these should be added simultaneously to the model;
- likelihood of producing large effects (start with factors that are most likely to produce large effects).

In any case, start with moderators and predictors first. Do not include preferences until the end as they are often associated with larger uncertainties and many assumptions underlying their values.

It is also possible to use a more algorithmic approach to test the factors in the model. Analogous to stepwise regression, there are three ways in which individual factors in outcome prediction can be investigated in a systematic way. Forward selection, where one starts with a 'empty model' (the outcome is determined by a treatments' main effects – a one-size-fits-all model) and adds factors (moderators, predictors) to the model as long as the new factor has a significant increase of the models output. Backward selection starts with the full model (i.e. the model equal to the 'outcome prediction' part) and eliminating factors one by one. A third approach uses a mix of these two, where (re)adding and (re)removing factors can occur at any step. Such methods are only recommended if many factor are considered and thus require automation. If any information is available on which factors are most suited for adding to the model at every iteration, this should be used for the selection of the factor.

**Step 4: Build model M2 to estimate effects of treatment or technology**

In this step, the factor selected in step 3 is added to the model created in step 2. As the aim is to compare the model created in step 2 with the model created here, it is preferable to have the models (or at least their output) side-by-side. This step can be performed by copying the code for the base-case model and adding the factor to the code (thus creating an extra outcome prediction for every outcome in the base-case model). The differences between predicted incremental effects (i.e. the value of treatment A minus treatment B) between models M1 (base case) and M2 (i.e. without versus with the moderator/predictor/PPTO) thus will be a reflection on the value of adding this factor to clinical decision making. The effect of adding the moderator (or predictor or preference) is, the difference of the incremental effects for each patient calculated separately.

### *Including moderator or predictor effects*

Although some models (e.g. decision trees) do not require specific population characteristics to be inserted, doing so will increase interpretability of results especially in the context of a specific population. As moderators and predictors can be regarded as interactions between patient characteristics and outcomes, patient characteristics are needed to quantify the effect in specific populations. In order to do this groups of (simulated) patients are needed. One pitfall in simulating individual patient characteristics is that patients with unlikely combinations may be generated (e.g. young, healthy and low Karnofski performance scale scores). Independent sampling of patient characteristics may result in an overrepresentation of these cases. Although such unlikely combinations may actually exist, in reality patient characteristics often correlate with each other. One should try to preserve these correlations. Ideally, an existing (validated) model structure or dataset is used for the population; an existing dataset can be sampled from, or used to generate a regression model which in turn is used to simulate patients. If only correlation information is available, copula-based simulation methods may be used instead (see for example Kumar and Shoukri (2007)).

Often one may find that moderators and predictors interact with each other, or that outcomes are related to each other (e.g. quality of life and length of life). These interactions can be inserted in the model if required.

### *Including patient preferences*

Each of the outcomes modelled can be weighed by patient preferences to yield a patient-preference weighed outcome, given the availability of relevant information. Instead of weighing the absolute outcomes, it is more useful to weigh outcome differences (i.e. outcome of treatment 1 minus the outcome of treatment 2). This ensures the preferences reflect improvement of one outcome versus another. If different outcomes are measured using different instrument and thus scales, it is difficult to interpret the absolute value of overall added effect. To improve interpretability, it may be an option to transform the incremental effects into a Cohen's standardized mean difference. This statistic, also known as Cohen's' d, is calculated by subtracting the outcome of one treatment from the outcome of another treatment and dividing this difference by the combined standard deviation (i.e. over all simulated patients). It is thereby translated into an effect that can more easily be compared between different outcomes which are

on different scales. This statistic can subsequently be used to weigh treatment outcomes over the entire study population.

To estimate the effect of patient preferences, a reasonable estimate of preference weights is necessary. MCDA methods such as discrete choice modelling could generate such weights empirically, however these are still subject to many assumptions so should be used carefully. Preferences are difficult to model, hence many interactions between patient characteristics and preferences may exist that are not reported, and vice versa. Also, often the reported preferences may not be directly usable in the model. For instance because the preferences are described qualitatively, or because the reported weights exclude a specific outcome. This means that the patient-preference weighted outcome estimation cannot be performed, or only by making a number of assumptions or approximations. One solution would be to estimate preferences weights and determine how this influences the model (i.e. threshold analysis, see step 5). Another solution in such cases would be to discuss the results of the moderators simulation model in light of the preferences, i.e. describe the preferences in a summary accompanying the results of the moderators simulation model. Then, the integration of both sources of evidence takes place at the decision makers' level (either feeding into a final judgment, or use the output for conducting any form of MCDA (see also Wahlster et al. (2016)) to generate weights to add to the model).

### *Effect estimation*

The (weighted) sum, or average, over the entire population can be compared with costs or effects to determine the value of adding the moderator (or predictor or preference). For instance, if the outcome is in QALYs and another outcome is the cost of a moderator test, then dividing these outcomes will result in a cost-effectiveness measure (as we are dealing with effect and cost differences, this is a ICER – incremental cost effectiveness). Additionally, the results per patient (if available) may be used to identify specific subgroups or decision algorithms which may be even more beneficial than the one investigated.

If distributions were used as model parameters, the outcome may be expressed as a distribution itself (mean effect with confidence interval or error) or visually (e.g. as histograms). The value of such a probabilistic sensitivity analysis can be regarded as a more valuable outcome than just a single outcome measure. Step 6 describes how such analysis can be used for specific scenarios.

### *Effect interpretation*

Ultimately, the effects of every evaluated moderator or predictor needs to be assessed by decision makers or by comparing the effects with a certain threshold (e.g. a cost/effectiveness ratio). These effects may be related to other effects (e.g. number of adverse events per gained year of life, or costs per QALY). The interpretation of the effect differences (whether it is clinically significant, for instance) cannot be performed based on model data alone. Hence, the report of the model should include not only all considerations on model design and parameter input (Bennett & Manuel, 2012) but also all evidence that may help in relating effects and preferences to each other but was not suitable for inclusion in the model. This information includes qualitative data, considerations on overall study quality (from which the model input originates) and/or transferability of the findings to the population under investigation. If preferences are evaluated, outcomes should be evaluated without preference weights as well (i.e. do not evaluate a moderator only on preference-weighted outcome). For instance, to assess cost-effectiveness or societal-based benefits.

### *Model adaption*

If the evaluation of a single moderator, predictor or PPTO is considered relevant for the model, it may be permanently included in the base-case model (after which it is no longer a base-case model, but the comparator for the addition of more factors in the model) In order to do so, the code that builds the second model should replace the code for the base-case model. After doing so, steps 3 and 4 can be repeated until no more moderators, predictors or PPTOs need to be assessed.

## Step 5. Perform sensitivity / scenario analysis

As an optional step, one may be interested in the effect of varying the value of certain parameters (such as moderators, preference, population characteristics or preferences) in the model. For instance, if confidence intervals were unfounded or founded on low grade evidence, one may be interested in the impact of these assumptions on any conclusions based on the models' output to get an idea on the overall model validity. Once the base model has been run, the model calculations can be repeated with different inputs. There are three objectives of a sensitivity or scenario analysis:

1. Identifying subgroups where different treatments outcomes or valuations prompt different treatment decisions. Subgroups where different treatment outcomes or valuations can be identified by performing a simulation for various (separate) groups of patients. For instance, set the gender for all patients to 'male', and then simulate the model to see the effects. Determine the outcomes of the model (per outcome, incremental differences) and repeat with gender set to 'female'. All patient characteristics in the model influence the outcome through moderators, so all these variables are eligible for possibly creating subgroups this way.
2. Determining the effect of uncertainty in individual parameters to test assumptions and strength of the model. The effect of individual parameters may be tested by varying them (e.g. increase in steps between a certain range) while recalculating the model in between. The dispersion of the output of the model is then a measure of how much the uncertainty influences the found effects: the more effect one finds, the larger the uncertainty is on the output of the model. If conclusions are stable across different values of the input parameter, the model can be considered as robust. It is advisable to perform sensitivity analyses on input parameter that are of uncertain origin. For instance, if the input is based on sparse data or expert opinion or when effects have large confidence intervals.
3. Determine a threshold for specific input parameters (such as population characteristics or moderator/predictor/preference interaction effect with a stochastic definition). While varying a specific input parameter, the threshold by which a certain approach is not beneficial or cost-effective anymore can be determined. Or the other way around: how large should the effect of a certain moderator or preference weight be in the personalised approach in order to become cost-effective.

It is important to determine a priori which sensitivity analysis or scenarios are going to be tested, why, and what the possible consequence could be. It is recommended to involve different stakeholders in this selection, but especially those that are going to use the results the analysis in decision making. These considerations should be presented as part of the results of the simulation model, to allow decision makers to correctly interpret the findings.

A limited number of sensitivity and scenario analyses are possible by hand: adjusting the formulas in the table, updating the calculations and viewing the results. Often though, one wants to perform a range of sensitivity analyses, which would be time-consuming if not conducted by computer. To this end, more advanced scripting knowledge is needed, or more specialised software packages should be used.

### *4.3.4 Example*

We will demonstrate the practical application of the framework using an example. A general case study was set up to test the methods described in the entire INTEGRATE-HTA project. In the case study the following research question was developed from stakeholder input: Are reinforced models of home based palliative care acceptable, feasible, appropriate, meaningful, effective, cost-effective model for providing patient-centred palliative care [compared to non-reinforced (i.e. 'usual') models of home based palliative care] in adults (defined as those aged 18 years old and over) and their families? A detailed description of the case study can be found in the INTEGRATE-HTA case study report (Brereton et al., 2016).

For the case study, we applied the sections on the retrieval of moderators, predictors and PPTOs. Furthermore, the outcomes explored in the guidance on effectiveness (Burns et al., 2016) was used to collect relevant outcomes. Although multiple outcomes, moderators, predictors and PPTOs were identified, there was a lack of coherence: moderators that would only apply to specific interventions, or only affect a particular type of outcome; PPTOs relating to a mostly different set of outcomes; and a general mix of specific interventions which were difficult to group or even find relations between. Hence, there was insufficient data to build any model on. As the intention of the example is to demonstrate each of the steps described in this guidance, we have selected a specific case within palliative care concerning advanced cancer care. It is a notorious example in palliative care where treatment decisions require to make tradeoffs between quality of life (QoL), length of life and the risks and harms caused by treatments. In this example, we will evaluate palliative chemotherapy with irinotecan for use as secondline therapy in colorectal cancer. We have decided for a micro simulation in this example because it allows a stepwise build up, without using specialised software and furthermore is the most versatile option in decision analysis.

### Step 1: exploring the problem

The data used to illustrate the example is freely interpreted from findings from two studies. The first one, by Cunningham et al. (1998), is a phase III randomised trial where the effectiveness of irinotecan including supportive care (SC) was compared to SC only. In the study 279 patients were randomly allocated between the two treatment options. Investigated outcomes included overall survival (up to 12 months) and quality of life (using the EORTC-QLQ30). Although Cunningham et al. investigated multiple moderators, we will use only one in this example: the presence of liver metastases.

The second study was performed by Voogt et al. (2005). In this study, patient preferences for the outcomes quality of life and quantity of life were investigated with the Quality Quantity Questionnaire (QQQ). The QQQ consists of two sets of four questions, resulting in a score representing a preference for quality of life (range 4-28) and a score representing a preferences for quantity of life (range 4-28). Outcomes of the QQQ will be used to weigh quantity and quality of life effects. A higher score means that patients have increased preferences for either quality or quantity of life.

In summary, the simulation model will consist of two outcomes, survival time in months and quality of life, one moderator, the presence of liver metastases, and preferences for either of the outcomes. The retrieved values are displayed in Table 4.

### Step 2: Build model (M1) to estimate effects of treatment or technology (base case)

The data shown in Table 4 was used to build a model. For illustrational purposes, a table-based microsimulation model was chosen. It was constructed in Microsoft Excel 2007 (Microsoft, Redmond, USA).
The model structure will be demonstrated by building the simulation in the form of a table (see Table 5). In the table each row represents a simulated patient with outcomes for each of the treatments that are included in the comparison. Below, the individual columns are described.
As we do not have data on a specific target population, we based the simulations on population characteristics reported by Cunningham et al.. The study population of Cunningham et al. consisted in 57% of cases of patients with a liver metastasis. As a base case analysis, the simulation model (Table 5) calculated survival and QoL for both treatment options using a 'one-size-fits-all' calculation (i.e. not including moderators or preferences). These calculations were based on the simulation input in Table 4 (Columns 2 and 3). QoL was assumed to be normally distributed. Survival was also assumed to be normally distributed. Standard deviations were estimated based on the range of the mean survival. For easier interpretation, survival was converted from months to years. The calculations for the data are shown in the grey row.

#### *Running the model*

Once the simulation model is built, it needs to be tested. Furthermore, it needs to be determined how many hypothetical patients are to be simulated. Starting with 250 patients (rows), the model was run 5 times. The standard deviations of the simulated outcomes (two outcomes and two interventions) as proportion of the mean outcomes served as measure of the models' stability. We found that between 500 and 1000 patients, this proportion approached zero and stabilising as the number of patients was increased. Based on these findings, it was decided to go for 1000 patients to simulate.

### Step 3: Selection of factor to expand model M1

The base-case model can be expanded with two factors that were retrieved from the literature: preference weights (for survival versus QoL) and the effect of the existence of liver metastases. We first expanded the simulation model with the liver metastases, and afterwards added preference weight. These two iterations of the framework will be described in the next step.

### Step 4: Build model M2 to estimate effects of treatment or technology

The simulation model as described in Table 5 will be expanded by adding the moderator 'liver metastases' to the calculation. The hazard ratio reported by Cunningham et al. shows that patients with liver metastases have a 1.64 times as large a chance of death in their 10-year follow-up period compared to patients without liver metastases. It does not inform on how much sooner patients die, but for the sake of this example we will assume that patients with liver metastases have a 61% (1/1.64) survival compared to those who do not. We will also assume this is a moderator in that is it only applies

*Table 4: Simulation model input.*

| Outcomes | Irinotecan + supportive care | Supportive care | Moderators / Predictors | Preferences* |
|---|---|---|---|---|
| Survival (months) | 9.2 (range 0–18.9) | 6.5 (range 0.7–19.3) | Hazard ratio (of death) 1.64 (if liver metastases present) | 13.3 (SD 3.3) |
| Quality of life (EORTC-QLQ global health scale** | 47.47 (SE 1.97 ) | 38.47 (SE 2.80) | (none used) | 15.5 (SD 3.9) |

to patients treated with SC only. To this end, the formula for calculating the survival of patients on SC (column I in Table 6) was updated by a formula that multiplies its outcome by 0.61 if the patient has liver metastases (column F). The presence of liver metastases could be expressed as a random draw from a binomial distribution, but due to lack of data we opted for a uniform distribution instead.

With this new model side-by-side with the base case model, we can calculate the incremental effects per patient per model (e.g. for survival: the average of all values n columns B minus D and columns I minus G). Furthermore, we can standardise these outputs by dividing this value by the standard deviations of these columns.

Now that model M2 has been built, it needs to be run. In the case of micro simulation models, M1 and M2 may be run together. After running the model(s), the differences in incremental effects of both models, (M1 and M2) with and without the factor, are now a measure of the added value of this moderator. In Table 7 the calculated outcomes

*Table 5: Calculation of treatment effects in simulation table (model M1, base case).*

| | A | B | C | D | E |
|---|---|---|---|---|---|
| | | Irinotecan + supportive care | | Supportive care | |
| | | survival* | QoL** | survival* | QoL** |
| Calculation | | N(0.77, 0.39) | N(47.47, 24.76) | N(0.54,0.39) | N(38.47, 23.92) |
| Patient 1 | | 0.55 | 13.05 | 0.44 | 50.33 |
| Patient 2 | | 0.57 | 61.24 | 0.46 | 11.79 |
| Patient 3 | | 1.40 | 24.68 | 0.31 | 33.85 |

*N(A,B) signifies a random draw from a normal distribution with mean A and standard deviation B. *Survival is expressed in years. **Quality of Life (QoL) is expressed as the EORTC-QLQ30 global health scale, with a range of 0-100.*

*Table 6: Calculation of treatment effects in simulation table (M2) including liver metastases as moderator.*

– Table 5 –

| A-E | F | G | H | I | J |
|---|---|---|---|---|---|
| | Liver metastases present? | Irinotecan + supportive care | | Supportive care | |
| | | survival* | QoL** | survival* | QoL** |
| | (57% yes) | N(0.77, 0.39) | N(47.47, 24.76) | N(0.54,0.39); if (F='Yes') * 0.61 | N(38.47, 23.92) |
| | Yes | 0.55 | 13.05 | 0.27 | 50.33 |
| | Yes | 0.57 | 61.24 | 0.28 | 11.79 |
| | No | 1.40 | 24.68 | 0.31 | 33.85 |

*N(A,B) signifies a random draw from a normal distribution with mean A and standard deviation B. *Survival is expressed in years. **Quality of Life (QoL) is expressed as the EORTC-QLQ30 global health scale, with a range of 0-100.*

*Table 7: Comparison output of simulation with and without effect of liver metastases on survival.*

| | Irinotecan + supportive care | | Supportive care | | Incremental effects (standardised) | |
|---|---|---|---|---|---|---|
| Model | Survival | QoL | Survival | QoL | Survival | QoL |
| M1 | 0.78 (0.39) | 46.28 (24.71) | 0.53 (0.39) | 37.98 (23.60) | 0.61 (1.32) | 0.34 (1.39) |
| M2 | 0.78 (0.37) | 47.20 (23.93) | 0.43 (0.33) | 39.48 (22.91) | 0.90 (1.23) | 0.33 (1.39) |

*Shown values are means of each of the columns (standard deviation). QoL: Quality of life. M1: simulation model with default preference weights and no moderators (base case). M2: simulation model with default preference weights and moderator 'liver metastases'. *Higher value equals preference towards irinotecan+SC. **Where the patient-preference weighed value > 0;*** Where the patient-preference weighed value < 0.*

with and without the factor are displayed. For the two outcomes, survival and QoL, the table shows the average effects and average incremental effects for equal preference weights (averages of the values in columns B, C, D, E, B minus D and C minus E of Table 5 and the corresponding columns of M2) for these outcomes for both treatment options. The last column shows the average patient-preference weighted value. This value is negative when SC results in more optimal outcomes, and positive if SC+irinotecan results in the most optimal outcomes. The higher the absolute value, the stronger the difference is (either because the difference between the treatments is larger or more cases show that an effect in the same direction). On average, adding the factor resulted in roughly 0.10 years of less survival for the SC option. Hence, we were able to improve the prediction of survival by 0.10 years by adding the factor to the model of the more personalized approach.

The potential benefit of a personalised approach where the presence of liver metastases is included in the decision process, can be determined using the incremental effects. The absolute difference in incremental effects of the one-size-fits-all model and the incremental effects of the model including the moderator shows how much better the model predicts outcomes. If the values in Table 7 are used (i.e. for QoL: (0.78-0.43) - (0.78-0.53) = 0.10) this effect would be underestimated as on a case-by-case level, the difference may be inversed for some patients. Therefore, this calculation is performed for every simulated patient separately. In this example, the predicted average difference in survival is increased by 0.58 (sd 0.45) years and the predicted average QoL is increased by 37 (sd 28) points on the QoL scale. These differences can be related to the costs of determining whether liver metastases are present, to determine, for instance, a cost-effectiveness estimate. If, for example, determining the presence of liver metastases costs €5,000, then we can relate this cost to the survival (€5,000 / 0.58 = €8,621 per gained life year) or the more often used QALY (€5,000 / (0.58*0.37) = €23,300/QALY). These figures illustrate the value of such a diagnostic in a very straightforward way. If a threshold is available (e.g. €80,000 per QALY) the value of this diagnostic may be considered cost-effective and should be added to the model. If the cost-effectiveness (or other measure) did not reach the threshold, it may still be possible to see if the threshold would be reached in specific scenarios. Otherwise, the factor on whether liver metastases are present may be better left out of the model.

### *Adding patient preferences for treatment outcome*

The relevancy of the change in survival may best be regarded in view of patient-preference weighed values: does information on patient preferences show that the effect is important, and if so, how important? In order to analyse this step, we will add preferences to the simulation. First, a default set of preferences (with all weights equal) will be added to the base case in order to provide a comparator. See Table 8 (expansion for Table 6) for the calculation of the preferences weights and overall weighted value of the effects. The incremental effects were standardised to allow weighting of outcomes between the measures of survival and QoL; this was necessary because both outcomes were measured on a different scale. It was performed by dividing the incremental effects (i.e. columns B minus D and C minus E) by their respective standard deviations. A similar calculation for M2 should be added as well. The output of these new columns can then be compared between the base-case and M2.

Table 9 shows the output of the simulation including the default preference weights for both base-case and M2 models. The last rows of the table shows for how many patients that treatment would be the optimal one, that is whether the sum of incremental outcomes, weighted by patient preferences, is more optimal for one treatment or the other. The simulation shows that for two-thirds of the simulated patients the irinotecan option would be the best option. It will also result in the highest QoL and the best survival.

One could see that, when looking at the preference-weighed preferred treatment, this would increase the number of cases where irinotecan would be considered best (based on incremental outcomes and preference weights) by 6.4% (in Table 9, 756-682 = 64 patients).

It was decided not to include the moderator to assess the effect of preferences at this point, so the next evaluation would be based only on the addition of the preference weights (and not preference weights plus the moderator). Had we decided to keep the moderator in the model to iteratively

*Table 8: The inclusion of patient preference weights for the different outcomes (base case).*

– Table 6 –

| A-J | K | L | M | N | O |
|---|---|---|---|---|---|
| | Incremental effects | | Preference weights | | Patient-preference weighted value |
| | Survival* | QoL* | Survival | QoL | |
| | B-D/stdev(B,D) | C-E/stdev(C,E) | (Default=1) | (Default=1) | (KxM/2)+(LxN/2) |
| | 0.26 | -1.52 | 1 | 1 | -0.63 |
| | 0.25 | 2.02 | 1 | 1 | 1.13 |
| | 2.65 | -0.37 | 1 | 1 | 1.14 |

*The letters in the calculation row refer to column names. Columns A-J are described in table 3. The second (grey) row shows the calculations. The values in columns K and L (incremental effects) are positive if the irinotecan options outperforms the SC option.*Incremental effects were standardised*

expand the model to compare, we had to copy the code for the M2 model into the columns for M1 and continue below.

### *Adding patient preference for treatment outcome weights*

In a similar way to the moderator, preference weights can be added to the model. To this end, the preference weights in model M2 were replaced with formulas that sampled from two normal distributions (one for each outcome specific preference) based on the values in the last column of Table 4. The effects on the output of the simulation are displayed in Table 10.

Inclusion of the preference weights in the model increased the weight of the QoL outcome compared to the survival effect resulting in an overall higher patient-preference weighted value (stronger preference towards irinotecan). However, over the entire simulated population only a small effect can be seen when these preferences are taken into account: about 2% more cases show that irinotecan is the most optimal treatment (704 versus 682). This can be explained by the fact that irinotecan is more effective on both survival AND QoL compared to SC only. Therefore inclusion of a higher preference for the QoL outcome will increase the patient-preference weighted value but not the number of patients for which irinotecan would be the most optimal treatment (because in most cases it is already the most optimal treatment). It probably would have changed the number of patients for which irinotecan is the most optimal when there would have been a positive effect on survival, but a negative effect on QoL. This is demonstrated in the next step.

## Step 6: Sensitivity and scenario analyses

For this step we assumed that irinotecan results in higher survival but lower QoL compared to SC. We will perform a sensitivity analysis where the mean preference weight for QoL compared to survival is varied from 0 to 3 to determine its effect on patient-preference weighed treatment effects.

Figure 6 shows the results of the sensitivity analysis. In dark grey the results of the model where all

*Table 9: Comparison output of simulation with and without effect of liver metastases on survival including preference weights.*

| | Irinotecan + supportive care | | Supportive care | | Incremental effects (standardised) | | Patient-preference weighted value |
|---|---|---|---|---|---|---|---|
| Model | Survival | QoL | Survival | QoL | Survival | QoL | |
| M1 | 0.78 (0.39) | 46.28 (24.71) | 0.53 (0.39) | 37.98 (23.60) | 0.61 (1.32) | 0.34 (1.39) | 0.47 (0.98) |
| M2 | 0.78 (0.37) | 47.20 (23.93) | 0.43 (0.33) | 39.48 (22.91) | 0.90 (1.23) | 0.33 (1.39) | 0.61 (0.92) |

| | M1 | M2 |
|---|---|---|
| No. of cases where irinotecan + supportive care has best outcome** | 682 | 756 |
| No. of cases where supportive care has best outcome*** | 318 | 244 |

*Shown values are means of each of the columns (standard deviation). QoL: Quality of life. M1: simulation model with default preference weights and no moderators (base case). M2: simulation model with default preference weights and moderator 'liver metastases'. *Higher value equals preference towards irinotecan+SC. **Where the patient-preference weighed value > 0;*** Where the patient-preference weighed value < 0.*

preference weights are equal, and in light gray the results of the model where the preference weights are varied. The first (dark grey) model shows that if the default preference weights are used, roughly 550 cases show irinotecan to be the best treatment option. As the preference weights are not varied in this model, a change is not expected here. In the second (light grey) model, the number of cases where irinotecan is the most optimal treatment varies depending on the QoL preference weights: from roughly 650 to 450 patients. As to be expected, an increasing preference weight for QoL results in less patients where irinotecan is the most optimal choice.

The sensitivity analysis shows us not which patients benefit from individually elicited preferences, but shows us the influence these preferences may have on which treatment can be considered to give the best results. Assuming a preference weight of QoL over survival between 0-3 is credible, then the maximal effect of adding the individual preferences according to our simulation is about 100 patients (the maximal difference between the light grey and dark grey lines is approximately 100 patients) or 10% of all patients. Of course, if preference weights are less extreme (e.g. between 0.5 and 1.5) the difference between the two models is small compared to the random fluctuations of the simulated numbers.

In order to determine the added value of using patient preferences in medical decision making concerning these two interventions, one should relate this number of patients to the cost (in terms of mo-

*Table 10: Comparison of output of simulation with and without different preferences for QoL and survival.*

| | Irinotecan + supportive care | | Supportive care | | Incremental effects (standardised) | | Patient-preference weighted value* |
|---|---|---|---|---|---|---|---|
| Model | Survival | QoL | Survival | QoL | Survival | QoL | |
| M1 | 0.78 (0.39) | 46.28 (24.71) | 0.53 (0.39) | 37.98 (23.60) | 0.61 (1.32) | 0.34 (1.39) | 0.47 (0.98) |
| M2 | 0.78 (0.37) | 47.20 (23.93) | 0.54 (0.38) | 39.48 (22.91) | 0.62 (1.31) | 0.33 (1.39) | 0.52 (0.97) |

| | M1 | M2 |
|---|---|---|
| No. of cases where irinotecan + supportive care has best outcome** | 682 | 704 |
| No. of cases where supportive care has best outcome*** | 318 | 296 |

*Shown values are means of each of the columns (standard deviation). ). QoL: Quality of life. M1: simulation model with default preference weights and no moderators (base case). M2: simulation model with different preference for the two outcomes. *Higher value equals preference towards irinotecan+SC. **Where the patient-preference weighed value > 0;*** Where the patient-preference weighed value < 0.*

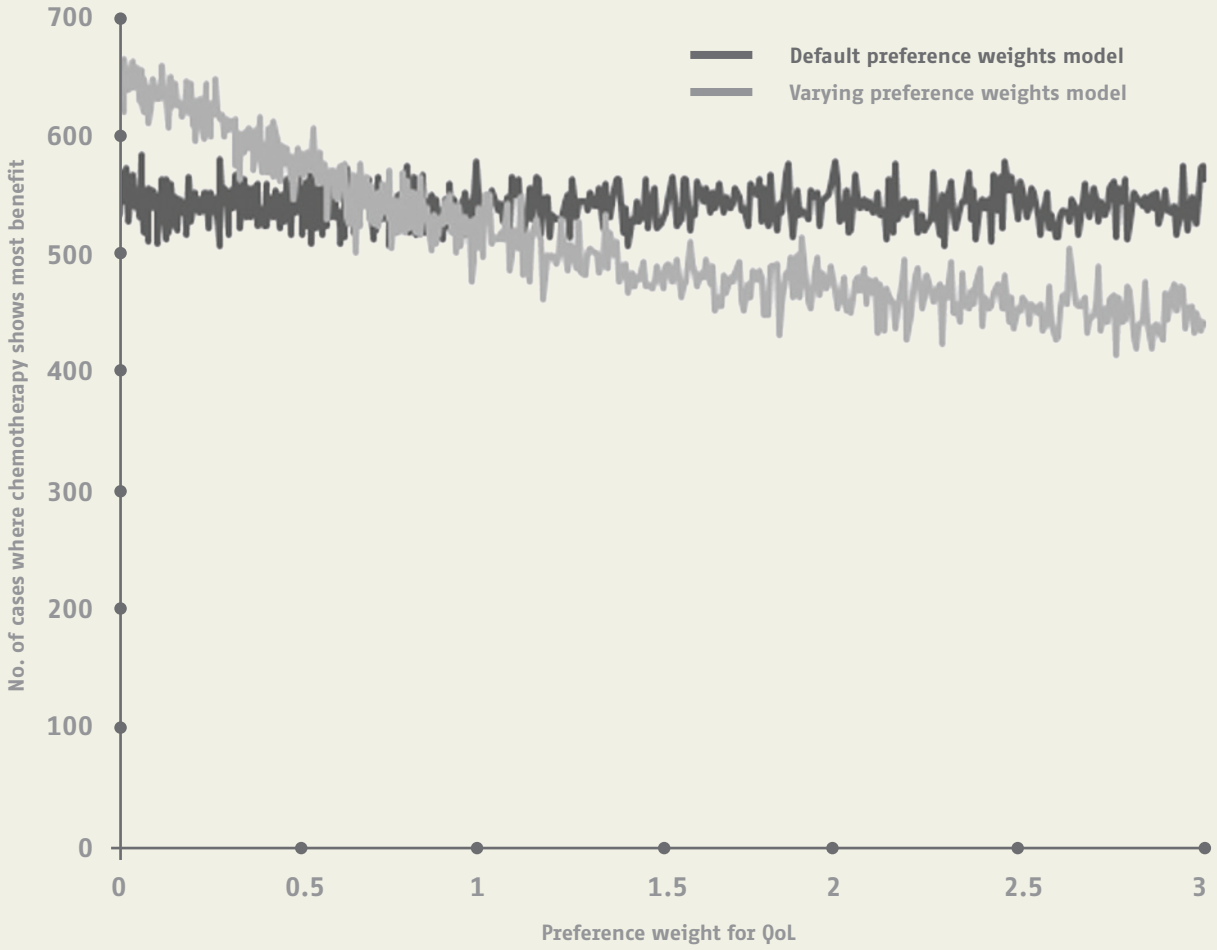
ney, time or tools required) of eliciting these preferences.

Above we have looked at the number of patients receiving one treatment or the other, but in reality we should also look at the effects. It may very well be possible that adding patient preferences results in more patients receiving a treatment that will actually result in less beneficial outcomes on the group as a whole. In the base case scenario irinotecan + SC results (on average) in the best outcome, so the more patients receiving only SC due to their preferences, the more patients will (on average) have a worse outcome. This apparent contradiction is typical for most decisions in medicine, and is a clear example of why the integration of moderators, predictors and PPTOs may result in different outcomes than when looking at overall outcomes. On the other hand, such findings may result in ethical issues on a personal level. All such considerations need to be taken into account when interpreting the results of this exercise.

## 4.4 CONCLUSIONS

The methods described in this section of this guidance offer a way to include multiple factors that increases complexity in an HTA. We provide a framework for assessing the overall effect of a treatment strategy in which moderators and predictors

*Figure 6: Sensitivity analysis of varying strengths of preference for survival.*




of and patient preferences for treatment outcome are included. Furthermore we give guidance on how to assess the added value of such a personalised approach compared to a 'one-size-fits-all' approach. The results of such a comparison may provide insight in how much effort or money should be or can be invested in a personalized approach in order to break even. This information could be helpful for the policy making process.

The model presented here can estimate the effects of adding or removing factors in decisions, and in turn, help decision makers determine when certain considerations, parameters or even preferences matter more or less and or may be ignored to simplify the overview of all complex interactions that exist when a technology is applied. These findings can inform decisions for larger or specific groups of patients.

One of the limitations of a modelling exercise in general is the feeling of a black box. Therefore giving transparency in reporting is a key issue: a clear description of all assumptions made, the sources of the models' parameters (and quality thereof) should be given. In part, many of the assumptions can be tested using sensitivity or threshold analysis, which should be utilized as much as possible to inform decision makers on the robustness of the conclusions. Furthermore, we recommend to use a reporting guideline for modelling studies (e.g. Bennett and Manuel (2012)) while drafting your report.

Another limitation in this example in particular is the assumption that patients can be given one treatment or another. It does not consider the fact that in daily practice, patients may switch treatments if the one they are given is found to be ineffective. This means the model overestimates the negative consequences of receiving the wrong treatment. So the modelling could improve when a sequential treatment regimen in corporate in the 'one-size-fits-all' approach. From a simulation viewpoint, this requires the incorporation of a specific decision rule on when a treatment is re-evaluated, and on what conditions it is decided that a switch should be made. This information has to be retrieved from guidelines or professionals used in daily practice (medical decision making). Such a simulation is more in line with daily practice (and its results, consequently, more relevant), however it requires very detailed information which may not be available in all settings.

Usability of patient preferences for treatment outcomes beyond identifying relevant outcomes, requires a comparison of all different outcomes of an intervention. Furthermore, to improve interpretability of the patient preference-weighted effects, a rescaling to for instance a utility score would help improve the interpretability of certain improvements. Although it is of great additional value if such weighing would be possible, in practice this can only rarely be adequately performed.

The main limitations of the presented methods is the fact that the framework does not generate new evidence, it only synthesizes existing evidence. Thus, the quality of the output is entirely dependent on the existence of good evidence that can be synthesised. This became very clear when the case study on home based models of palliative care was used to build an example: we were not able to retrieve the required information to build a sufficiently satisfactory model to estimate relevant outcome, let alone demonstrate the different steps of this framework. The construction of the simulation model can help identifying these problems. Ideally, one should go back to the literature and try to identify additional evidence on moderators, predictors or preferences. If none are found, additional research into these factors may be needed. This may then either prompt the research, or form part of a recommendation to the stakeholders. The model can also be used to perform a value of information analysis. This specific analysis can be used to estimate what the monetary value is of having perfect information i.e. no uncertainty in the input parameters. This information can help to decide if investment in addition research is justified.

As knowledge on complex interventions increases, and more complex interventions are identified because of the many factors, simple prediction models will no longer suffice. There is a need for versatile, expandable and understandable models. A variety of models exist to do so, each suitable for a specific situation. However, the biggest challenge is the same for each model: the translation from practice to the model. Although this is a considerable challenge, computer (simulation) models are already finding their way in clinical decision making: offering clinicians the tools to predict the outcomes of treatments for individual patients before they are given helps supporting their decision making (Sadiq et al., 2008; Souza et al., 2011). They can also be used as training environments (Flores et al., 2012; Moss & Berner, 2015). In our view these types of models will likely become increasingly used in HTA.

# 5 REFERENCES


ADOLFSSON, J., STEINECK, G. (2000) Prognostic and treatment-predictive factors-is there a difference? Prostate Cancer and Prostatic Diseases, 3, 265-268.

AGUINIS, H. (2004) Regression analysis for categorical moderators. Guilford Press.

ANDERSON, L., MELLOR, J. (2009) Are risk preferences stable? Comparing an experimental measure with a validated survey-based measure. Journal of Risk and Uncertainty, 39, 137-160.

AUSTON, I. (2005) Clinical Prediction Guides Added to PubMed® Clinical Queries. NLM Tech Bull, May-Jun.

BARNETT-PAGE, E., THOMAS, J. (2009) Methods for the synthesis of qualitative research: a critical review. BMC Medical Research Methodology, 9, 59.

BARON, R.M., KENNY, D.A. (1986) The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. Journal of Personality and Social Psychology, 51, 1173-1182.

BECK, U. (2002) Individualization: Institutionalized individualism and its social and political consequences. Sage.

BENNETT, C., MANUEL, D.G. (2012) Reporting guidelines for modelling studies. Medical Research Methodology, 12, 168.

BOYD, C.M., MCNABNEY, M.K., BRANDT, N., CORREA-DE-ARAUJUO, R., DANIEL, M., EPPLIN, J., FRIED, T.R., GOLDSTEIN, M.K., HOLMES, H.M., RITCHIE, C.S., SHEGA, J.W. (2012) Guiding principles for the care of older adults with multimorbidity: an approach for clinicians: American Geriatrics Society Expert Panel on the Care of Older Adults with Multimorbidity. Journal of the American Geriatrics Society, 60, E1-E25.

BRAILSFORD, S.C., HARPER, P.R., PATEL, B., PITT, M. (2009) An analysis of the academic literature on simulation and modelling in health care. Journal of Simulation, 3, 130-140.

BRERETON, L., WAHLSTER, P., LYSDAHL, K.B., MOZYGEMBA, K., BURNS, J., CHILCOTT, J.B., WARD, S., BRÖNNEKE, J.B., TUMMERS, M., VAN HOORN, R., VAN DER WILT, G.J., PFADENHAUER, L., GERHARDUS, A., ROHWER, A., REHFUESS, E., OORTWIJN, W., REFOLO, P., SACCHINI, D., LEPPERT, W., BLAZEVICIENE, A., PRESTON, L., CLARK, J., GOYDER, E., ON BEHALF OF THE INTEGRATE-HTA TEAM (2016) Integrated assessment of home based palliative care with and without reinforced caregiver support: 'A Demonstration of INTEGRATE-HTA methodological guidances'.

BRETT HAUBER, A., FAIRCHILD, A.O., REED JOHNSON, F. (2013) Quantifying benefit-risk preferences for medical interventions: an overview of a growing empirical literature. Applied Health Economics and Health Policy, 11, 319-329.

BRIDGES, J.F., JONES, C. (2007) Patient-based health technology assessment: a vision of the future. International Journal of Technology Assessment in Health Care, 23, 30-35.

BURNS, J., CHILCOTT, J.B., VAN HOORN, R., KIEVIT, W., REHFUESS, E. (2016) Guidance to assess effectiveness aspects. In: LYSDAHL, K.B., MOZYGEMBA, K., BURNS, J., CHILCOTT, J.B., BRÖNNEKE, J.B., HOFMANN, B. (eds.). Guidance for assessing effectiveness, economic aspects, ethical aspects, socio-cultural aspects and legal aspects in complex technologies [Online]. Available from: http://www.integrate-hta.eu/downloads/.

CAMERON, R. (2011) Mixed methods research: the Five Ps framework. The Electronic Journal of Business Research Methods, 9, 96–108.

CHALMERS, I., BRACKEN, M.B., DJULBEGOVIC, B., GARATTINI, S., GRANT, J., GÜLMEZOGLU, A.M., HOWELLS, D.W., IOANNIDIS, J.P.A., OLIVER, S. (2014) How to increase value and reduce waste when research priorities are set. The Lancet, 383, 156-165.

COOK, C.E. (2008) Potential pitfalls of clinical prediction rules. J Man Manip Ther, 16, 69-71.

CRAIG, P., DIEPPE, P., MACINTYRE, S., MICHIE, S., NAZARETH, I., PETTICREW, M., MEDICAL RESEARCH COUNCIL, G. (2008) Developing and evaluating complex interventions: the new Medical Research Council guidance. BMJ, 337, a1655.

CRESWELL, J.W., CLARK, V.L.P. (2010) Designing and Conducting Mixed Methods Research. (Second Edition edition ed.).Los Angeles: SAGE Publications, Inc.

CRESWELL, J.W., PLANO, L. (2007) Designing and conducting mixed methods research. Thousand Oaks, CA, US: Sage Publications, Inc.

CUNNINGHAM, D., PYRHONEN, S., JAMES, R.D., PUNT, C.J., HICKISH, T.F., HEIKKILA, R., JOHANNESEN, T.B., STARKHAMMAR, H., TOPHAM, C.A., AWAD, L., JACQUES, C., HERAIT, P. (1998) Randomised trial of irinotecan plus supportive care versus supportive care alone after fluorouracil failure for patients with metastatic colorectal cancer. The Lancet, 352, 1413-1418.

DE BEKKER-GROB, E.W., DONKERS, B., JONKER, M.F., STOLK, E.A. (2015) Sample Size Requirements for Discrete-Choice Experiments in Healthcare: a Practical Guide. Patient, 8, 373-384.

DELLINGER, A.B., LEECH, N.L. (2007) Toward a Unified Validation Framework in Mixed Methods Research. Journal of Mixed Methods Research, 1, 309-332.

FACEY, K., BOIVIN, A., GRACIA, J., HANSEN, H.P., LO SCALZO, A., MOSSMAN, J., SINGLE, A. (2010) Patients' perspectives in health technology assessment: a route to robust evidence and fair deliberation. International Journal of Technology Assessment in Health Care, 26, 334-340.

FACEY, K.M., HANSEN, H.P. (2011) Patient-focused HTAs. International Journal of Technology Assessment in Health Care, 27, 273-274.

FITCH, K., BERNSTEIN, S.J., AGUILAR, M.D., BURNAND, B., LACALLE, J.R. (2001) The RAND/UCLA appropriateness method user's manual. DTIC Document.

FLORES, C.D., BEZ, M.R., RESPÍCIO, A., FONSECA, J.M. (2012) Training Clinical Decision-Making through Simulation. Decision Support Systems–Collaborative Models and Approaches in Real Environments. Springer.

FRANCIS, J.J., JOHNSTON, M., ROBERTSON, C., GLIDEWELL, L., ENTWISTLE, V., ECCLES, M.P., GRIMSHAW, J.M. (2010) What is an adequate sample size? Operationalising data saturation for theory-based interview studies. Psychol Health, 25, 1229-1245.

FRIED, T.R., BYERS, A.L., GALLO, W.T., AL, E. (2006) Prospective study of health status preferences and changes in preferences over time in older adults. Archives of Internal Medicine, 166, 890-895.

GABLER, N.B., DUAN, N., LIAO, D., ELMORE, J.G., GANIATS, T.G., KRAVITZ, R.L. (2009) Dealing with heterogeneity of treatment effects: is the literature up to the challenge. Trials, 10, 43.

GAGNIER, J., MORGENSTERN, H., ALTMAN, D., BERLIN, J., CHANG, S., MCCULLOCH, P., SUN, X., MOHER, D., GROUP, F.T.A.A.C.H.C. (2013) Consensus-based recommendations for investigating clinical heterogeneity in systematic reviews. BMC Medical Research Methodology, 13, 106.

GOMES, B., CALANZANI, N., CURIALE, V., MCCRONE, P., HIGGINSON, I.J. (2013) Effectiveness and cost-effectiveness of home palliative care services for adults with advanced illness and their caregivers. The Cochrane Library.

HANLEY, B., TRUESDALE, A., KING, A., ELBOURNE, D., CHALMERS, I. (2001) Involving consumers in designing, conducting, and interpreting randomised controlled trials: questionnaire survey. BMJ, 322, 519-523.

HASKINS, R., RIVETT, D.A., OSMOTHERLY, P.G. (2012) Clinical prediction rules in the physiotherapy management of low back pain: a systematic review. Man Ther, 17, 9-21.

HAYNES, R.B., MCKIBBON, K.A., WILCZYNSKI, N.L., WALTER, S.D., WERRE, S.R., HEDGES, T. (2005) Optimal search strategies for retrieving scientifically strong studies of treatment from Medline: analytical survey. BMJ, 330, 1179.

HIGGINS, J.P.T., GREEN, S. (2011) Cochrane handbook for systematic reviews of interventions version 5.1. 0 [updated March 2011]. The Cochrane Collaboration [Online]. Available from: www.cochrane-handbook.org.

HOEFFLER, S., ARIELY, D. (1999) Constructing Stable Preferences: A Look Into Dimensions of Experience and Their Impact on Preference Stability. Journal of Consumer Psychology, 8, 113-139.

HUNTER, J.E., SCHMIDT, F.L. (2004) Treatment Effects: Experimental Artifacts and Their Impact. Methods of Meta-Analysis. (2 ed). Thousand Oaks, CA: SAGE Publications, Inc.

ISAACS, J.D., FERRACCIOLI, G. (2011) The need for personalised medicine for rheumatoid arthritis. Annals of the Rheumatic Diseases, 70, 4-7.

JAKKA, S., ROSSBACH, M. (2013) An economic perspective on personalized medicine. The HUGO Journal, 7, 1-6.

JANSEN, S.J., STIGGELBOUT, A.M., WAKKER, P.P., NOOIJ, M.A., NOORDIJK, E.M., KIEVIT, J. (2000) Unstable preferences: a shift in valuation or an effect of the elicitation procedure? Medical Decision Making, 20, 62-71.

JANSSEN, I.M., GERHARDUS, A., SCHRÖER-GÜNTHER, M.A., SCHEIBLER, F. (2014) A descriptive review on methods to prioritize outcomes in a health care context. Health Expectations.

JENKINS, M. (2004) Evaluation of methodological search filters--a review. Health Info Libr J, 21, 148-163.

JUNG, H.P., BAERVELDT, C., OLESEN, F., GROL, R., WENSING, M. (2003) Patient characteristics as predictors of primary health care preferences: a systematic literature analysis. Health Expectations, 6, 160-181.

JUNI, P., WITSCHI, A., BLOCH, R., EGGER, M. (1999) The hazards of scoring the quality of clinical trials for meta-analysis. JAMA, 282, 1054-1060.

KALOW, W. (2002) Pharmacogenetics and personalised medicine. Fundam Clin Pharmacol, 16, 337-342.

KATRAK, P., BIALOCERKOWSKI, A.E., MASSY-WESTROPP, N., KUMAR, S., GRIMMER, K.A. (2004) A systematic review of the content of critical appraisal tools. BMC Med Res Methodol, 4, 22.

KRAEMER, H.C. (2013) Discovering, comparing, and combining moderators of treatment on outcome after randomized clinical trials: a parametric approach. Stat Med, 32, 1964-1973.

KRAEMER, H.C., KIERNAN, M., ESSEX, M., KUPFER, D.J. (2008) How and why criteria defining moderators and mediators differ between the Baron & Kenny and MacArthur approaches. Health Psychol, 27, S101-108.

KRAEMER, H.C., WILSON, G.T., FAIRBURN, C.G., AGRAS, W.S. (2002) Mediators and moderators of treatment effects in randomized clinical trials. Archives of General Psychiatry, 59, 877-883.

KUMAR, P., SHOUKRI, M.M. (2007) Copula based prediction models: an application to an aortic regurgitation study. Medical Research Methodology, 7, 21.

LIN, E.S., KAYE, A.D., BALUCH, A.R. (2012) Preanesthetic Assessment of the Jehovah's Witness Patient. The Ochsner Journal, 12, 61-69.

LLOYD, A.J. (2003) Threats to the estimation of benefit: are preference elicitation methods accurate? Health Economics, 12, 393-402.

MACCORQUODALE, K., MEEHL, P.E. (1948) On a distinction between hypothetical constructs and intervening variables. Psychological review, 55, 95.

MACLEAN, S., MULLA, S., AKL, E.A., JANKOWSKI, M., VANDVIK, P.O., EBRAHIM, S., MCLEOD, S., BHATNAGAR, N., GUYATT, G.H., AMERICAN COLLEGE OF CHEST, P. (2012) Patient values and preferences in decision making for antithrombotic therapy: a systematic review: Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. Chest, 141, e1S-23S.

MARCH, J.S., CURRY, J.F. (1998) Predicting the outcome of treatment. Journal of abnormal child psychology, 26, 39-51.

MCGINN, T., O'CONNOR-MOORE, N., ALFANDRE, D., GARDENIER, D., WISNIVESKY, J. (2008) Validation of a hepatitis C screening tool in primary care. Arch Intern Med, 168, 2009-2013.

MCKIBBON, K.A., WILCZYNSKI, N.L., HAYNES, R.B., HEDGES, T. (2009) Retrieving randomized controlled trials from medline: a comparison of 38 published search filters. Health Info Libr J, 26, 187-202.

MEDICAL RESEARCH COUNCIL, U.M. (2008) Developing and evaluating complex interventions: new guidance.

MERLINO, L.A., BAGCHI, I., TAYLOR, T.N., UTRIE, P., CHRISCHILLES, E., SUMNER, W., MUDANO, A., SAAG, K.G. (2001) Preference for fractures and other glucocorticoid-associated adverse effects among rheumatoid arthritis patients. Medical Decision Making, 21, 122-132.

MOSS, J., BERNER, E.S. (2015) Evaluating clinical decision support tools for medication administration safety in a simulated environment. International Journal of Medical Informatics, 84, 308-318.

MUHLBACHER, A.C., JUHNKE, C. (2013) Patient preferences versus physicians' judgement: does it make a difference in healthcare decision making? Applied Health Economics and Health Policy, 11, 163-180.

MULLEY, A.G., TRIMBLE, C., ELWYN, G. (2012) Stop the silent misdiagnosis: patients' preferences matter. BMJ, 345, e6572.

MURDOCH, T.B., DETSKY, A.S. (2013) The inevitable application of big data to health care. JAMA, 309, 1351-1352.

MURPHY, E., DINGWALL, R., GREATBATCH, D., PARKER, S., WATSON, P. (1998) Qualitative research methods in health technology assessment: a review of the literature. Health Technology Assessment (Winchester, England), 2, iii-ix, 1-274.

NASSER, M., FEDOROWICZ, Z. (2011) Grading the quality of evidence and strength of recommendations: the GRADE approach to improving dental clinical guidelines. J Appl Oral Sci, 19.

NICHOLSON, R.A., HURSEY, K.G., NASH, J.M. (2005) Moderators and Mediators of Behavioral Treatment for Headache. Headache: The Journal of Head and Face Pain, 45, 513-519.

O'CONNOR, A.M. (1995) Validation of a decisional conflict scale. Medical Decision Making, 15, 25-30.

ONWUEGBUZIE, A.J., DICKINSON, W.B., LEECH, N.L., ZORAN, A.G. (2009) Toward More Rigor in Focus Group Research: A New Framework for Collecting and Analyzing Focus Group Data. International Journal of Qualitative Methods, 8, 1-21.

OPMEER, B.C., DE BORGIE, C.A., MOL, B.W., BOSSUYT, P.M. (2010) Assessing Preferences Regarding Healthcare Interventions that Involve Non-Health Outcomes: An Overview of Clinical Studies. Patient, 3, 1-10.

PINCUS, T., MILES, C., FROUD, R., UNDERWOOD, M., CARNES, D., TAYLOR, S. (2011) Methodological criteria for the assessment of moderators in systematic reviews of randomised controlled trials: a consensus study. BMC Medical Research Methodology, 11, 14.

REDEKOP, W.K., MLADSI, D. (2013) The faces of personalized medicine: a framework for understanding its meaning and scope. Value In Health, 16, S4-9.

REUZEL, R.P., VAN DER WILT, G.J., TEN HAVE, H.A., DE VRIES ROBBE, P.F. (2001) Interactive technology assessment and wide reflective equilibrium. Journal of Medicine and Philosophy, 26, 245-261.

RYAN, M., SCOTT, D.A., REEVES, C., BATE, A., VAN TEIJLINGEN, E.R., RUSSELL, E.M., NAPPER, M., ROBB, C.M. (2001) Eliciting public preferences for healthcare: a systematic review of techniques. Health Technology Assessment (Winchester, England), 5, 1-186.

SADIQ, S.K., MAZZEO, M.D., ZASADA, S.J., MANOS, S., STOICA, I., GALE, C.V., WATSON, S.J., KELLAM, P., BREW, S., COVENEY, P.V. (2008) Patient-specific simulation as a basis for clinical decision-making. Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, 366, 3199-3219.

SHIELL, A., HAWE, P., GOLD, L. (2008) Complex interventions or complex systems? Implications for health economic evaluation. BMJ, 336, 1281-1283.

SLOVIC, P. (1995) The construction of preference. American Psychologist, 50, 364-371.

SOUZA, N.M., SEBALDT, R.J., MACKAY, J.A., PROROK, J.C., WEISE-KELLY, L., NAVARRO, T., WILCZYNSKI, N.L., HAYNES, R.B. (2011) Computerized clinical decision support systems for primary preventive care: a decision-maker-researcher partnership systematic review of effects on process of care and patient outcomes. Implementation Science, 6, 87-99.

STACEY, D., BENNETT, C.L., BARRY, M.J., COL, N.F., EDEN, K.B., HOLMES-ROVNER, M., LLEWELLYN-THOMAS, H., LYDDIATT, A., LÉGARÉ, F., THOMSON, R. (2011) Decision aids for people facing health treatment or screening decisions. The Cochrane database of systematic reviews, 10.

STREET, R.L., ELWYN, G., EPSTEIN, R.M. (2012) Patient preferences and healthcare outcomes: an ecological perspective. Expert Review of Pharmacoeconomics & Outcomes Research, 12, 167-180.

SUN, X., BRIEL, M., BUSSE, J.W., YOU, J.J., AKL, E.A., MEJZA, F., BALA, M.M., BASSLER, D., MERTZ, D., DIAZ-GRANADOS, N., VANDVIK, P.O., MALAGA, G., SRINATHAN, S.K., DAHM, P., JOHNSTON, B.C., ALONSO-COELLO, P., HASSOUNEH, B., WALTER, S.D., HEELS-ANSDELL, D., BHATNAGAR, N., ALTMAN, D.G., GUYATT, G.H. (2012) Credibility of claims of subgroup effects in randomised controlled trials: systematic review. BMJ (Clinical research ed.), 344, e1553.

TUMMERS, M., VAN HOORN, R., LEVERING, C., BOOTH, A., VAN DER WILT, G.J., KIEVIT, W. (2016) Optimal search strategies for identifying moderators of treatment outcome in PubMed. (In preparation).

VAN DER WEIJDEN, T., LEGARE, F., BOIVIN, A., BURGERS, J.S., VAN VEENENDAAL, H., STIGGELBOUT, A.M., FABER, M., ELWYN, G. (2010) How to integrate individual patient values and preferences in clinical practice guidelines? A research protocol. Implementation Science, 5, 10.

VAN HOORN, R., KIEVIT, W., BOOTH, A., MOZYGEMBA, K., LYSDAHL, K.B., REFOLO, P., SACCHINI, D., GERHARDUS, A., VAN DER WILT, G.J., TUMMERS, M. (2016) The development of PubMed search strategies for patient preferences for treatment outcomes. (In preparation).

VEMER, P., CORRO RAMOS, I., VAN VOORN, G., AL, M.J., FEENSTRA, T.L. (2014) PRM79 - Advishe: a New Tool to Report Validation of Health-Economic Decision Models. Value in Health, 17, A556-A557.

VIECHTBAUER, W. (2008) Analysis of Moderator Effects in MetaAnalysis. In: OSBORNE, J. (ed.). Best practices in quantitative methods. Thousand oaks, CA: SAGE Publications, Inc.

VOOGT, E., VAN DER HEIDE, A., RIETJENS, J.A., VAN LEEUWEN, A.F., VISSER, A.P., VAN DER RIJT, C.C., VAN DER MAAS, P.J. (2005) Attitudes of patients with incurable cancer toward medical treatment in the last phase of life. Journal of Clinical Oncology, 23, 2012-2019.

WAHLSTER, P., BRERETON, L., BURNS, J., HOFMANN, B., MOZYGEMBA, K., OORTWIJN, W., PFADENHAUER, L., POLUS, S., REHFUESS, E., SCHILLING, I., VAN HOORN, R., VAN DER WILT, G.J., BALTUSSEN, R., GERHARDUS, A. (2016) Integrated assessment of complex health technologies – The INTEGRATE-HTA Model [Online]. Available from: http://www.integrate-hta.eu/downloads/.

WHITE, V.J., GLANVILLE, J.M., LEFEBVRE, C., SHELDON, T.A. (2001) A statistical approach to designing search filters to find systematic reviews: objectivity enhances accuracy. Journal of Information Science, 27, 357-370.

WONG, S.S., WILCZYNSKI, N.L., HAYNES, R.B., RAMKISSOONSINGH, R., HEDGES, T. (2003) Developing optimal search strategies for detecting sound clinical prediction studies in MEDLINE. AMIA Annual Symposium Proceedings, 728-732.

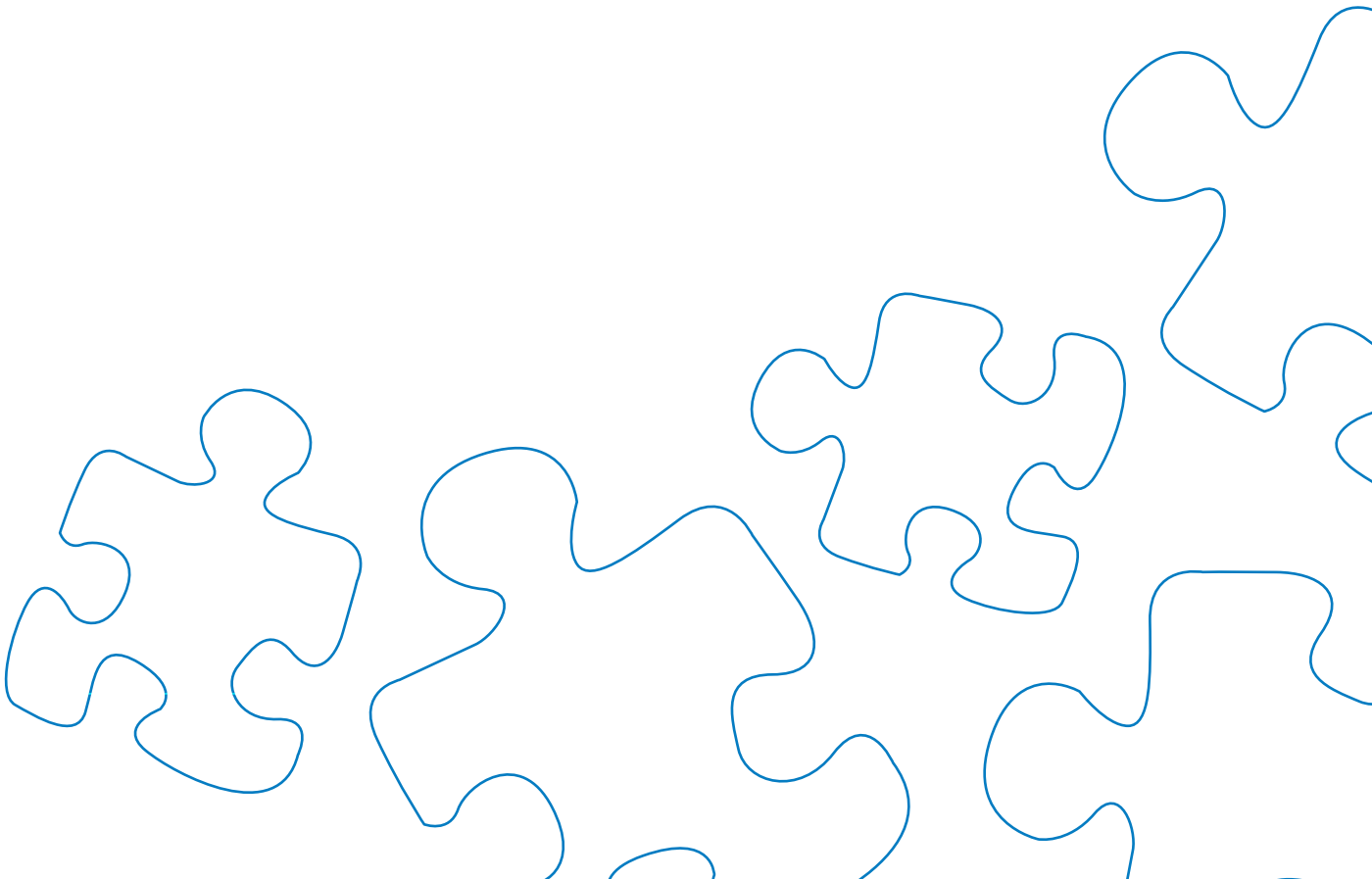
# 6 ACKNOWLEDGEMENT

We would like to thank all those who contributed to the establishment of this guidance: external experts, stakeholders who participated in the case study and members of the INTEGRATE-HTA project. We also thank the European Union for funding this project.

We would also like to thank those that took part in the Delphi panel to review the items for the appraisal checklist and gave valuable input for improving the appraisal checklist: Patrick Bossuyt; Pim Cuijpers; Thomas Debray; Rogier Donders; Atle Fretheim; Jelle Goeman; Julian Higgins; Daniel Hind; Helena Kraemer; Peter Langhorne; Tamar Pincus; Maroeska Rovers; Peter Tugwell; Martin Underwood; Vivian Welch.

We also wish to thank Eddy Adang for reviewing the guidance on integration and giving valuable feedback.

The guidance was reviewed by the following external experts: Albert Brühl, Pascale Lehoux, Nora Ibargoyen Roteta, Claudia Wild and Stefan Sauerland.



# 7 APPENDIX

## 7.1 DEVELOPMENT OF THE APPRAISAL CHECKLIST

The development of the appraisal checklist is based on a literature search, a modified Delphi process, a testing phase, and expert and user feedback. Below the individual processes are described in detail.

*Literature search towards appraisal criteria*

Searching for methodological criteria is not straightforward, and the methods we employed were aimed to increase sensitivity of the search in favour of being systematic or repeatable. First, a general search of PubMed was performed to find (methodological) information about moderators of treatment effects and their appraisal. This includes both methodological articles and existing reviews that included appraisal. Google Scholar was additionally used because of its wider range of journals and more in-depth text search as well as its more advanced relevancy sorting. Search results were scanned for possible relevant content based on title and abstract (PubMed) or title and shown snippets in which the search words were found (Google Scholar).

An exploratory search towards criteria in PubMed using related MeSH-terms (e.g. "Effect modifier, Epidemiologic" and "Randomized Controlled Trials as Topic", "moderators of treatment effects", "moderators", "subgroup", "heterogeneity" and combinations thereof) with and without keywords related to critical appraisal was conducted (e.g. "critical appraisal", "appraisal", "guidance", "methodology"). A similar search was performed in Google Scholar from which approximately 100 articles were selected for full-text reading by relevancy (title/abstract or Google Scholar text snippets) from these search results (including related articles and citations).

Five articles from these results (Gabler et al., 2009; Sun et al., 2009; Sun et al., 2010; Pincus et al., 2011; Gagnier et al., 2013) were considered to be key publications (most relevant based on their contents) and were used to inform a more systematic search method. Four different search strategies were used in total:

- Searching specifically for the combinations (OR'ed, as the AND-operator yielded only a limited amount of papers) of MeSH-terms under which these five most relevant articles were listed in PubMed. This yielded large amounts of articles, but (percent wise) few hits deemed to be relevant.
- Starting with an index article that defines moderators as used in this study, Kraemer et al. (2002), articles citing the index article were reviewed for relevant information. 1074 articles cited Kraemer et al. (citation count according to Google Scholar). Of the first 1000 articles, 945 articles were excluded based on title and abstract (articles clearly concerning application instead of methodology of moderation analysis). Most of the remaining articles contained criteria useful for this study.
- The aforementioned five most relevant articles all shared certain keywords in their title ("heterogeneity", "subgroup*" or "moderator*"). A generic search towards these keywords in titles yielded nearly 25000 articles, so the search query was expanded with the string ("Randomized Controlled Trials as topic/methods"[MeSH] or "Meta-Analysis as topic "[MeSH]). This search resulted in 179 articles.
- Additional relevant articles were identified through citation chasing of the articles found in the search strategy.

The search process identified key authors (through the number and type of articles in the search results). Additional articles were identified by searching for articles written by these authors. This list of authors was also used to identify potential contributors to a Delphi panel.

Lastly, existing appraisal and quality reporting guidances (such as those listed on the EQUATOR network, Cochrane reviews, Risk of bias analysis) were investigated. Most of these had been previously identified by the authors of this guidance or from the existing literature, e.g. Katrak et al. (2004). Specifically, we looked at the existence of information on moderator analysis or subgroup analysis in general, both to inspire the design of our own checklist and to determine possible overlap with our checklist and these pre-existing guidances.

*Inclusion/exclusion criteria*

Articles from which to extract appraisal criteria were not considered if they were not methodological (i.e. only applied analysis), if they were systematic reviews or meta-analysis which, or if after reading them full text revealed a lack of in-depth discussion on appraisal criteria. Articles that only repeated

*Table A-1: Inclusive selection of appraisal criteria found in the literature.*

**Group A. Is the variable a genuine moderator?**

1. Was an interaction test used to test for a subgroup effect, and P-values for tests of interaction or heterogeneity reported?[1,7]
2. Is the subgroup variable a characteristic measured before randomization?[1,10,15]
3. Continuous moderators without meaningful zero points are centered.[4,9]

**Group B. Is the result possibly spurious?**

4. Are all variables examined reported?[3,5,6]
5. Are subgroup (-analyse)s defined a priori?[12]
6. Are subgroup hypotheses defined a priori?[5,6,10,11,15,16]
7. Was the subgroup effect one of a small number of hypothesized effects tested?[6,7,15,16]

**Group C. Theoretical basis of moderator variables**

8. The selection of moderators to consider was pre-specified and based on expert knowledge and theoretical considerations.[1]
9. Report defines which moderators are defined a priori and which post hoc.[3]
10. Is there indirect evidence that supports the hypothesized interaction (biological rationale)?[6,15,16]
11. Was measurement of baseline and process factors reliable and valid (from published information) in target population?[1]

**Group D. Analysis**

12. Was there an equal distribution of moderators between groups at baseline?[1]
13. Was the regression significant at $P < 0.05$, or (if more than three comparisons) corrected or significance adjusted to $P < 0.01$ or using Bonferroni or similar corrections?[1,2,7]
14. Did the authors explore residual variances of interactions if carrying out multiple two-way interactions?[1]
15. All moderators are analysed in a single model.[2]

**Group E. Power analysis**

16. Power analysis has been performed specifically to detect moderation effects. [3,4]
17. Do authors report a power analysis for moderator effect (a-priori or posthoc, but using an a-priori effect size, not the observed one?[1]
18. Was sample size adequate for the moderator analysis (at least 4 fold the required sample size for main treatment effect in the lowest sub-group for the moderator factor)?[1]
19. If not, were there at least 20 people in the smallest sub-group of the moderator?[1]
20. Have authors employed analysis to compensate for insufficient power (i.e. boot-strapping techniques?)[1]

**Group F. Effect size**

21. Effect sizes, confidence intervals and statistical significance are reported for all subgroups.[3]
22. Was the direction of the subgroup effect specified a priori?[15,16]
23. Were the magnitude of the differences in reported subgroup analysis large enough to support different recommendations for the subgroups?[1]
24. Is there indirect evidence that supports the hypothesized interaction?[15,16]
25. Does the interaction test suggest a low likelihood that chance explains the apparent subgroup effect?[10,15,16]
26. Is the significant interaction effect independent of other potential subgroup effects?[10,15]

**27.** Is the interaction consistent across closed related outcomes within the study?[10,15]

**28.** Is there evidence that the measurement error of the instrument is likely to be sufficiently small to detect the differences between subgroups that are likely to be important?[1]

**29.** Was there adjustment for other baseline factors?[1]

**30.** Is there an explicit presentation of the differences in outcome between baseline sub-groups (e.g. standardized mean difference between groups, Cohen's d).[1]

### Group G. Applicability and relevance

**31.** Are the patients in the subgroup comparable to my patients?[12]

**32.** Would the outcome influence my decision making?[3]

**33.** Is the subgroup effect or interaction clinically important?[3,12]

**34.** Is the between-subgroup treatment effect clinically important?[1,3,12]

### Group H. Items for meta-analyses and meta-regressions

**35.** Is the effect suggested by comparisons within rather than between studies?[1,11,15,16]

**36.** The review/investigative team should include clinical experts or state a plan for consulting clinical experts during the review protocol development and implementation (e.g., when choosing clinical covariates and when interpreting the findings).[17]

**37.** Reviewers should think through all potentially relevant variables to explore and not rely on statistical measures of heterogeneity to justify such investigations.[17]

**38.** Is the interaction consistent across studies?[12,10]

**39.** A logical hierarchy of clinical covariates should be formed and investigated only if there is sufficient rationale and a sufficient number of trials available (10 trials per covariate).[1]

**40.** Studies that did not include the moderator of interest were excluded from the analysis.[2]

**41.** A mixed model was used to model the effect of moderators [over studies].[2]

**42.** The reviewers stated any plans to include additional covariates after looking at the data (post hoc) from included studies (e.g., forest plots, radial plots) and how they plan to do this.[17]

**43.** Was described a priori how the results of any findings were going to be interpreted and used in the overall synthesis of evidence?[17]

**44.** If participant level covariates are investigated, individual patient data was used.[13,14,17]

### Group I: Additional items for a body of evidence

**45.** Is the subgroup difference consistent across studies?[12,10]

**46.** Is the (combined) subgroup effect or interaction clinically important?[1,3,12]

**47.** Are the patients in the subgroup comparable to the target population?[12]

**48.** Is the between-subgroup treatment effect clinically important?[1,3,12]

**49.** Are there ethical issues involved in using the moderator to selecting patients for treatments?

*Sources of the criteria are: 1 (Pincus et al., 2011); 2 (Viechtbauer, 2008); 3 (Gabler et al., 2009); 4 (Wu & Zumbo, 2008); 5(Thompson & Higgins, 2002); 6 (Yusuf et al., 1991); 7 (Assmann et al., 2000); 9 (Kraemer et al., 2002); 10 (Sun et al., 2009); 11 (Guyatt et al., 2011); 12 (Dijkman et al., 2009); 13 (Groenwold et al., 2009); 14 (Bigger, 2003); 15 (Sun et al., 2010); 16 (Oxman & Guyatt, 1992); 17 (Gagnier et al., 2012)*

considerations reported in other articles (i.e. quotations or citations) were also excluded after ensuring the referenced article was included in our list of references. For all other articles we tried to be as inclusive as possible.

*Data extraction and study quality criteria*

Possible criteria for studies concerning moderators or predictors of treatment effects were extracted and listed in a long list of all possible criteria (duplicates were eliminated unless distinctly differently worded). No specific quality criteria were employed to exclude papers or critical appraisal criteria. Criteria were grouped conceptually to facilitate item selection in the following development phase.
Table A-1 lists the criteria retrieved from the literature and references to the articles from which they were retrieved (note that some criteria were reworded or constructed from statements made in the referenced literature). A total of 49 criteria were identified and grouped into nine categories.

*Modified Delphi process*

A modified Delphi consensus process was used to assess the appropriateness of the list of items in Table A-1. The participants (selected professionals working in the field and identified through the literature searches) were asked to rate the individual appraisal criteria retrieved from the literature on their appropriateness for inclusion in our checklist. The RAND/UCLA Appropriateness Method is a modification of the Delphi consensus method that allows the selection of appropriate items of a appraisal checklist in a very efficient way (Fitch et al., 2001). This method was conducted through an online questionnaire. Individual criteria were rated on a scale from 1(not relevant) to 9(very relevant). An opt-out option was also provided. Additionally, participants were asked for ideas on additional criteria, rewording of existing criteria or other remarks concerning the criteria or the list of items as a whole. While asking for the appropriateness of individual items, participants were asked specifically to take the other items in account (e.g. if items overlapped, rate the best of the two high and the other one low). An overview of all items was given at the end of the questionnaire to aid participants in suggesting additional appraisal criteria.

Agreement on the appropriateness of items was determined using the IPRAS (Interpercentile Range Adjusted for Symmetry). It was calculated as described in the RAND/UCLA appropriateness method manual, leaving the Correction Factor for Asymmetry at 1.5 and the Required InterPrecentile Range at 2.35, as described in the RAND manual.

After a first round, items for which agreement was reached and with a mean appropriateness score >=5 were selected to be included in a test version of the appraisal checklist; items scoring <3 and that had agreement were excluded. Proposed rewordings of, and additions to the items were carefully considered by the research team.

In a second Delphi round, results from the first round were reported: per criterion a histogram and the mean score were presented. Any changes made to the list of criteria were displayed with changes in wording marked in the text, if these were adopted by the researchers. Items that were selected to be included in the test version of the checklist due to the results of the first Delphi round were marked as such, but still displayed to allow participants to compare new items to already included items. The Delphi panel was asked to re-rate the remaining items. The need to exclude redundant items to shorten the appraisal checklist was emphasised in the introduction text to try to reduce the list of items agreed to be included. At the end of this round criteria that were only reworded mildly and for which agreement was reached were accepted (agreed to be included in the test-version of the instrument).

*Modified Delphi process results*

Of the in total 37 experts from the fields of HTA, evidence synthesis, statistics, epidemiology and psychology identified as eligible and invited, 20 agreed to participate in the modified Delphi procedure. Of these 20, 15 experts (75%) provided answers in the first round. Twelve participants provided answers in the second Delphi round. In these two rounds all items listed in table A-1 were rated for appropriateness. Often participants of the Delphi procedure motivated their answer using the comment sections. We critically reviewed these comments to see if certain items needed to be reworded, added, merged, or moved to another section.

In the first Delphi round, none of the criteria reached a score low enough for excluding the item. Twenty-five criteria were considered to be relevant (score >=5 and agreement reached according to IPRAS method). One additional item was suggested by a participant (item 28 in Table A-2). These items were to be evaluated in the next Delphi round. After the second round, agreement on exclusion was reached on 3 items (items 3, 20, and 15, Table A-1). and agreement on inclusion was

reached on 32 items. We also merged several items due to overlapping contents. The remaining 15 items either reached no agreement or resulted in an indecisive appropriateness. Two of these items, 11 and 14 in Table A-1, were selected based on distribution of appropriateness valuation and comments of the participants. Furthermore, item 49 in Table A-1 was added as expansion on item 20 in the checklist. The remaining 12 items were discarded. Based on the comments, specific adjustments to criteria were made. Furthermore, the items were rearranged under headings matching different sections as often found in papers (background, analyses, results) to improve usability. The test-version of the appraisal checklist (see Table A-2) ultimately included 31 items. The following answering possibilities were included: 'Yes', 'Partially', 'No', 'Don't know', and 'Not (clearly) described'.

*Testing phase*

The test version of the appraisal checklist (see Table A-2) was pilot tested on a sample of papers to detect problems with the items and practical use of the checklist. To this end a set of 23 papers concerning moderators or predictors of treatment outcome was appraised. Each study was appraised twice by different testers: five INTEGRATE-HTA researchers with backgrounds in HTA, medicine, evidence based information practice and one Master's student Biomedical sciences. Each of the testers was familiar with the case study and the subject of home-based palliative care. They were asked to critically appraise a random selection of the 23 studies relevant to the subject of home-based palliative care. Along with the appraisal checklist, a set of appraisal tools were handed out to allow quick access to a small set of general appraisal tools for various study types. Users could also choose to use a different appraisal checklist to determine overall study quality. Testers were asked to first appraise the study on overall quality. If the overall study quality was inadequate, they were asked not to further appraise the moderator analysis.

One papers was not appraised by one tester due to inadequate overall quality and one due to the tester considering the appraisal checklist inappropriate for that type of study due to a lack of a specific intervention and control group. The remaining papers were appraised by two users independently. Feedback and usage statistics (frequency of 'don't know' and 'not applicable' answers and inter-rater agreement) were collected.

The testing revealed an overall poor inter-rater agreement. Some items were more often than anticipated answered with 'don't know' or 'not applicable'. Users indicated problems with determining whether a study was hypothesis-generating or hypothesis-testing (which was a condition for some of the items to apply). Furthermore the testing phase showed that some item need more explanations on how to apply the item. It was also found that the test version mentions only moderators, giving the appearance that the items do not apply to predictors. However, most items do apply equally well to predictor analyses. Lastly, some of the items seem to apply only to specific study types, which, after rephrasing, could be applicable to more study designs.

Further development based on expert opinion and finalisation of the appraisal checklist

Based on the feedback and subsequent discussions between a subset of testers and two experts that also were part of the Delphi panel, final adjustments were made to the checklist. The major revisions (excluding rewordings of individual items) were as follows:

- The addition of a background section that would improve the interpretation and agreement between different users.
- The number of answering categories was reduced. From the initial set ('Yes', 'No', 'Partially', 'Not (clearly) described', 'Don't know' and 'Not applicable'), 'Not (clearly) described' and 'Partially' were dropped. After reviewing the background information it was found that the answering categories partially overlapped, and furthermore, did not contribute to the overall valuation.
- An initial section specifically for meta-analyses was removed. The items described under this section overlapped significantly with the items on body of evidence, and some items concerned more general validity of systematic reviews (e.g. the use of random-effects models in meta-analysis).
- One item was added to assess whether the statistical analyses were performed sufficiently.
- One question was added specifically to make the overall judgement of a moderator or predictor per paper more explicit.
- The process of first assessing overall validity, and then moderator/predictor appraisal was further emphasised by creating a single document to improve overview on the use of the checklist and emphasise the two-step appraisal (overall validity first, then moderator/predictor effects).

*Table A-2: Items in test version of appraisal checklist.*

**Study design**

1. Was the selection of moderators to consider based on expert knowledge and/or theoretical considerations?
2. Is there (in)direct evidence that supports the hypothesized interaction (e.g. biological rationale)?
3. In the case of hypothesis-testing, are moderator hypotheses and their analyses defined a priori?
4. In the case of hypothesis-testing, was the selection of moderators to consider pre-specified?

**Population & measurements**

5. Was the moderator variable measured before randomisation?
6. Was measurement of baseline and process moderators/predictors reliable and valid (from published information) in target population?
7. Is there evidence that the measurement error of the instrument is likely to be sufficiently small to detect the differences between subgroups that are likely to be important?
8. Is the sampled population comparable to my patients?

**Analysis**

9. Was an interaction test used to test for a moderator effect?
10. In the case of hypothesis-testing, are the moderator effects one of a small number of hypothesized effects tested?
11. In the case of hypothesis-testing, has power-analysis been performed specifically to detect moderation effects?

**Results**

12. Are all moderator and predictor variables examined reported?
13. Was distinction made between which moderators were defined a priori and which post hoc?
14. Was sample size adequate for the moderator analysis?
15. Were effect sizes, confidence intervals and statistical significance reported?
16. Was the moderator significant?
17. Does the interaction test suggest a low likelihood that chance explains the apparent moderator effect?
18. Is there an explicit presentation of the differences in outcome between baseline subgroups (e.g. standardized mean difference between groups, Cohen's d)?
19. Is the interaction consistent across closely related outcome measures within the study?
20. Is the moderator effect clinically important, i.e., does it support making different clinical decisions for different patients in this population?

### Meta-analyses

**21.** Do within and between study comparisons agree with each other?

**22.** Did the review/investigative team include clinical experts or have they stated a plan for consulting clinical experts during the review protocol development and implementation?

**23.** Were studies that did not include the moderator of interest excluded from the analysis?

**24.** Did the model for evaluating the effects of moderators over studies account for between-study heterogeneity in (moderator) effects?

**25.** Did the reviewers state any plans to include additional covariates after looking at the data (pots hoc) from included studies (e.g., forest plots, radial plots) and how they plan to do this?

**26.** Was described a priori how the results of any findings were going to be interpreted and used in the overall synthesis of evidence?

**27.** If participant level covariates are investigated, was individual patient data used?

**28.** Is the moderator effect size consistent across studies?

### Body of evidence

**29.** Is the moderator or predictor effect reasonably homogenous across studies ?

**30.** Is the synthesized moderator or predictor effect clinically important (i.e. would the result influence clinical decision making)?

**31.** Are the patients in the sampled population comparable to the population for which the information will be used?

The revised appraisal checklist can be found in chapter 7.2. The next step in the finalisation of the appraisal checklist is the review by the Delphi panel members and additional validation and testing of reliability in a new test round. These final steps of the development of the checklist are expected to take place and be published in 2016.

## 7.2 APPRAISAL CHECKLIST FOR MODERATORS AND PREDICTORS OF TREATMENT EFFECTS

*About the checklist*

In this document, a checklist is presented for the critical appraisal of claims concerning moderators and predictors of treatment effects. The checklist aims to appraise evidence and assess relevance of the evidence for HTAs or guidelines. The document has three sections: an introduction, the appraisal checklist itself, and a background section on how the checklist may be used. Further information on how the checklist was developed can be found in the previous chapter 7.1.

The checklist is designed to provide users with a structured way of looking at a set of key quality and relevancy indicators. The checklist is intended to be applicable for various study types such as (randomised) intervention trials, observational studies and systematic reviews. The background information provided at the end of this document is intended to help users answering the items and interpreting the consequences of negative answers for the overall judgment of credibility and relevance of treatment moderation or prediction.

The checklist requires that overall study quality (unrelated to treatment moderation or prediction) is assessed, using appropriate appraisal tools. The user of this checklist is asked to report the conclusion from this appraisal before proceeding with the appraisal of claims regarding moderation or prediction. The checklist itself consists of fourteen items grouped under the headings of design, analysis, results, and transferability of results. An additional set of five items can be used to appraise claims regarding moderation or prediction of treatment effects based on a body of evidence.

*Moderators versus predictors: definition*

For moderators and predictors of treatment effects we used the definitions by (Baron & Kenny, 1986). According to their definitions, moderators and predictors are variables such as patient characteristics that influence the effect of a treatment. The difference between moderators and predictors is that a moderator influences the effect of a specific treatment (e.g. the protective effect of aspirin is moderated by gender), while a predictor influences outcomes regardless of any treatment (e.g. old age predicts higher probability of infections). A predictor will show the same effect over all treatment arms within subgroups, while a moderator will show a different effect in each arm. Because of this difference, moderators should be tested using statistics such as interaction tests (e.g. a treatment by characteristic-interaction in a regression model), while predictors are not tested for interaction. The effects of moderators or predictors can be additive (e.g. a linear regression coefficient) or multiplicative (e.g. an odds ratio).With the exception of the interaction term, the analysis of moderator and predictor analyses adhere largely to the same quality criteria. Moderators can only be retrieved from (preferably randomised) intervention studies or systematic reviews as the investigation of moderators should include the evaluation of an interaction term between moderator variable and intervention (intervention vs. control group). Predictors on the other hand may be retrieved from many different kinds of studies, including observational studies.

## Appraisal of overall study quality

A large number of critical appraisal tools is available to assess overall study quality. Widely used appraisal tools for the various study designs include:

- Systematic review / meta-analysis: AMSTAR checklist (http://amstar.ca); CASP Systematic Review Checklist[1]; Cochrane's Risk Of Bias[2] tool (used on the individual studies included in the review).If the systematic review is based on individual patient data (IPD), the guidance by Tierney et al. may be useful (see http://www.ncbi.nlm.nih.gov/pubmed/26196287).
- **Randomised controlled trial:** Cochrane's Risk of Bias tool[2]; CASP Randomised Controlled Trial1 checklist.
- **Cohort study:** CASP Cohort Study Checklist[1] ; Newcastle-Ottawa scale (http://www.ohri.ca/programs/clinical_epidemiology/nosgen.pdf).
- **Cross-sectional / descriptive study:** Cross sectional appraisal tool (https://reache.files.wordpress.com/2010/03/cross-sectional-appraisal-tool.pdf).
- **Case-control study:** CASP checklist for case-control studies[1]; Newcastle-Ottawa scale (http://www.ohri.ca/programs/clinical_epidemiology/nosgen.pdf).
- **Prognostic study:** Quality In Prognosis Studies (QUIPS) (http://www.biomedcentral.com/content/supplementary/1546-0096-12-19-S1.pdf)

Further appraisal tools have been identified by (Katrak et al., 2004). Please indicate the tool that was used to critically appraise the overall study quality, and the major findings below.

---

***Critical appraisal tool used for assessing overall study validity*** *(if any)*

---

---

***Overall study validity*** *– Outcome of the used appraisal tool or argumentation on the quality of the study*

---

---

---

***Is the overall study quality of a level that findings related to moderators and predictors are likely to be of sufficient quality?***

☐ Yes ▸ please continue appraisal

☐ No ▸ do not continue appraisal

☐ Don`t know ▸ please continue the appraisal, but mind possible bias

---

[1] CASP appraisal tools are available from http://www.casp-uk.net/#!checklists/cb36

[2] http://ohg.cochrane.org/sites/ohg.cochrane.org/files/uploads/ Risk%20of%20bias%20assessment%20tool.pdf

## Appraisal of moderators and predictors for treatment effects

Note: multiple candidate moderators or predictors of treatment outcome may have been explored within a single study. In such cases, credibility or relevance may differ across these factors, depending on how they were measured, strength of association, etc. It is up to the user to decide whether conclusions apply to all candidate factors that were examined or to subsets only.

---

| | *Yes* | *No* | *Don't know* | *Not applicable* |
|---|---|---|---|---|
| **Design** | | | | |
| **1.** A priori plausibility: was there sufficient empirical or theoretical support for the moderator or predictor that was examined? | ☐ | ☐ | ☐ | ☐ |
| **2.** Was the moderator or predictor specified a priori? | ☐ | ☐ | ☐ | ☐ |
| **3.** Was the moderator or predictor variable measured before the allocation or start of the intervention? | ☐ | ☐ | ☐ | ☐ |
| **4.** Was measurement of the moderator or predictor reliable and valid in the target population? | ☐ | ☐ | ☐ | ☐ |
| **Analysis** | | | | |
| **5.** In case of a moderator, was an interaction test used? | ☐ | ☐ | ☐ | ☐ |
| **6.** Was a limited number of moderators and predictors tested? | ☐ | ☐ | ☐ | ☐ |
| **7.** Was sample size adequate for the moderator or predictor analysis? | ☐ | ☐ | ☐ | ☐ |
| **Results** | | | | |
| **8.** Were results presented for all candidate moderators or predictors that were examined? | ☐ | ☐ | ☐ | ☐ |
| **9.** Did statistical tests or confidence intervals indicate that observed moderator or predictor effects were unlikely to be merely due to chance variation? | ☐ | ☐ | ☐ | ☐ |
| **10.** Was the moderator or predictor effect consistent with related moderators or predictors, or across related outcomes measured within the study? | ☐ | ☐ | ☐ | ☐ |
| **Transferability** | | | | |
| **11.** Were the setting and study population comparable to the setting and population in which the information would be used? | ☐ | ☐ | ☐ | ☐ |
| **12.** Is the moderator or predictor effect clinically important? | ☐ | ☐ | ☐ | ☐ |

### Overall judgement

| | Yes | No | Don't know |
|---|---|---|---|
| Considering your conclusions regarding overall quality and items 1-12, would you say that claims regarding moderation or prediction of treatment outcomes 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:***

## Appraisal of moderators and predictors for treatment effects in a body of evidence

Specific candidate moderators or predictors of treatment effect may have been explored in multiple studies. Appraisal of such a body of evidence is important as some aspects of appraisal, such as comparison of effects between studies or relevance of a moderator or predictor effect, become apparent only after moderator or predicting findings have been collected from multiple studies

For such cases, items 10-12 are repeated here as they apply to the summarised or pooled effect and may be answered differently as compared to individual studies. For example, smaller studies may find apparently large and clinically important effects, but when pooled with other, larger studies, the effect may no longer be statistically or clinically significant.

### Body of evidence

| | Yes | No | Don't know | Not applicable |
|---|---|---|---|---|
| **10.** Was the moderator or predictor effect consistent with related moderators or predictors, or across related outcomes measured between the studies? | ☐ | ☐ | ☐ | ☐ |
| **11.** Were the setting and study population comparable to the setting and population in which the information would be used? | ☐ | ☐ | ☐ | ☐ |
| **12.** Is the moderator or predictor effect clinically important? | ☐ | ☐ | ☐ | ☐ |
| **13.** Was the moderator or predictor effect reasonably homogenous across studies? | ☐ | ☐ | ☐ | ☐ |
| **14.** Was the moderator or predictor measured similarly across the included studies, or was an adequate conversion performed? | ☐ | ☐ | ☐ | ☐ |

## Overall judgment

| | Yes | No | Don't know |
|---|---|---|---|
| Considering the answers for individually appraised studies as well as items 10-14, would you say that claims regarding moderation or prediction of treatment outcomes 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, for each item a brief rationale is presented, considerations that you may want to take into account, and a brief discussion of possible implications when a study does not seem to meet the relevant criterion.

### Item 1. *A priori plausibility: was there sufficient empirical or theoretical support for the moderator or predictor that was examined?*

Basically, this item asks about independent empirical or theoretical support for the candidate moderator(s) or predictor(s) of treatment effect. In case of such support, it is less likely that the moderator or predictor effect was a spurious result.

***Consider:*** Whether the authors provided a plausible (biological) working mechanism. Preferably, this is based on experimental studies and generally accepted as a possible biological pathway.

***Implications:*** A moderator or predictor effect is more likely to be a false-positive finding if there is no underlying theory on how the effect could influence the outcome. However, the observed effect may have been caused by other mechanisms, yet unknown.

### Item 2. *Was the moderator or predictor specified a priori?*

A moderator or predictor effect should preferably be specified a priori. A finding is less likely to be a chance finding if the moderator or predictor effect (direction and/or size) was hypothesised before the start of the study.

***Consider:*** Whether the hypothesised effects and analyses were specified in a previously published study protocol, or whether they were explained by studies referenced in the paper, or whether authors explicitly stated that candidate factors were pre-specified. Any analysis to estimate the statistical power of the study for specified moderator/predictor effects also indicates that the moderator or predictor was pre-specified.

***Implications:*** A moderator or predictor effect is more likely to be false-positive if its analysis was not pre-specified. In such cases, findings should be considered exploratory, in need of further verification

### Item 3. *Was the moderator or predictor variable measured before the allocation or start of the intervention?*

The earlier variables are measured in a study, the less prone their measurements are to bias (e.g. measurements errors correlated to the treatment arm the patients were allocated to).

***Consider:*** Whether the variable was measured before allocation or start of the intervention, for instance because the variable was used for stratified allocation or because it is explicitly stated. This does not apply to variables that are unlikely to be affected by treatment.

***Implications:*** If the variable is straightforward to measure without bias and insensitive to treatment (e.g. age, gender) there is little reason for concern. Otherwise, two problems may arise: [1] the measured effect may be a mediator: the variable explains part of the outcome because it is part of the causal relationship between treatment and outcome. In this case the variable cannot be used to stratify treatments. [2] different types of measurement bias could have occurred. If the study was double-blinded this is less likely.

### **Item 4.** *Was measurement of the moderator or predictor reliable and valid in the target population?*

Unreliable or invalid measurements of predictors or moderators can result in either under- or overestimation of the moderator or predictor effect. If moderator or predictor is not measured using a reliable and valid method, the subgroup effect may be underestimated and the main treatment effect may be overestimated, or vice versa.

***Consider:*** Whether the measurement method is reliable and valid in the target population as evidenced by pilot testing and/or existing publications on the measurement method.

***Implications:*** Credibility of the effect (size) is compromised, in proportion to doubts regarding reliability or validity of the measurement.

### **Item 5.** *In case of a moderator, was an interaction test used?*

In the literature, the terms 'moderator' and 'predictor' are occasionally used interchangeably (or other terms are used to describe their effects, such as effect modifier, determinant or interaction effect). Hence, it is important that the user first identifies whether the effect that is being appraised is actually a moderator, a predictor or other effect (e.g. mediating effect or main effect). Please be aware that if the study which is being appraised is not a study with an intervention group and a control group or a systematic review of such studies, a subgroup difference should not be considered a moderator effect; a moderator effect can only be shown through interaction effects between a moderator and a treatment group.

***Consider:*** Whether an interaction-test was used (a treatment by moderator interaction). Whether the effect can be considered a main effect instead (i.e. direct theoretical or statistical association with treatment).

***Implications:*** If no interaction-test was used, or the results of such test were not statistically significant, the observed effect (if any) should be considered a predictor, not a moderator. If the moderator is associated/correlated with the treatment, it is a mediator or an effect, respectively.

### **Item 6.** *Was a limited number of moderators and predictors tested?*

There are two reasons for keeping the number of moderators and predictors tested at a minimum: The probability of finding false-positive effects due to chance (related to alpha-level), and the risk of overfitting of (regression) models.

***Consider:*** Consider the total number of moderators and predictors tested in a study. There are no firm criteria to determine what number of tests can still be considered adequate. These problems are (at least partially) related to the amount of tests performed in relation to study size. There are some rules of thumb relating to multivariate analysis: Some sources state that 20 subjects per moderator is the bare minimum (Pincus et al., 2011), other rules of thumb range between 2-20 cases per regression parameter, and up to 50 per parameter in the case of stepwise regression (Voorhis & Morgan, 2007; Austin & Steyerberg, 2015). The minimal number of cases per parameter increases when effect sizes are expected to be small, when there are substantial measurement errors or when data are skewed (Voorhis & Morgan, 2007). For univariate analysis, specific statistical methods may be used to perform or correct for multiple testing (e.g. Bonferroni correction); statistical expertise may be required to appraise such cases.

***Implications:*** Depending on the p-value that is still considered significant (often set at 0.05) or the size of the confidence interval (often set at 95%), a certain number of hypotheses are expected to be significant based on chance alone (in the case of a p-value of 0.05, this is 5% of all tested hypotheses). Hence, if more tests are performed, the more likely that a finding is false-positive.

**Item 7.** *Was sample size adequate for the moderator or predictor analysis?*

Without adequate sample size, the odds of false-negative moderator or predictor effects is increased (i.e. moderators or predictors may be missed). This, in turn, may result in underestimation of other effects if such findings are used as a basis for multivariate analysis. The user may wish to perform post-hoc power analysis or look at similar studies that do provide a power analysis if none is provided in the study.

***Consider:*** Consider whether a power analysis (ad hoc or post hoc) was performed or other consideration on study size were described; did the researchers succeed in including and following up the requisite number of patients? Also, consider measures such as model fit and error sizes. Please keep in mind that any significant finding can indicate that a large enough sample was used.

***Implications:*** If the sample size of a study was inadequate, then any effects that did not reach significance may have actually been significant effects if a larger study was performed. Hence, observed effects to be not statistically significant cannot be dismissed. If interaction terms are not significant and study size was too small, there may yet be a moderator effect.

**Item 8.** *Were results presented for all candidate moderators or predictors that were examined?*

Conceivably, more candidate moderators or predictors were investigated, but only those for which significant associations were found are being reported (selective reporting or reporting bias). This may be established by examining published study protocols.

***Consider:*** Any clues that there were more moderators or predictors tested, such as 'all other variables were not significant' or 'results were corrected for baseline characteristics' without explicit statement of the results. Consider variables that were mentioned in the protocol (if any) but not reported, or moderators or predictors that were much more likely to be researched than those reported in the study. On the other hand, if a study reports insignificant moderators/predictors, this may be considered an indication that the researchers were comprehensive in reporting moderation or prediction effects.

***Implications:*** If it is likely that more moderators or predictors were investigated than reported, it is possible that results were selectively reported. This increases the likelihood that reported effects were chance findings.

**Item 9.** *Did statistical tests or confidence intervals indicate that observed moderator or predictor effects were unlikely to be merely due to chance variation?*

Statistical tests help distinguish chance findings from real findings. Although the results of these tests do not prove that a moderator or predictor effect is real or not, they do add to the evidence that an observed effect is likely to be true.

***Consider:*** Whether the (pooled) moderator or predictor effect shows significance (confidence interval not including 1, or regression parameter range not including 0, or p-value lower than the value considered significant; usually 0.05 or lower in the case of many tests or the utilisation of a Bonferroni or other correction). If a correction for multiple testing was used, consider the validity of this method as well as its assumptions (see also items 5 and 6).

***Implications:*** Smaller intervals and better significance add to the credibility of the observed effect. Insignificance of moderator (interaction-term) effects may indicate that there is another effect (e.g. predictor-effect) or that study size was inadequate. Insignificance of a predictor-effect may indicate the predictor-effect does not exist or that the study was too small.

**Item 10.** *Was the moderator or predictor effect consistent with related moderators or predictors, or across related outcomes measured within the study [or between studies]?*

Consistence between moderator or predictor effects adds to the credibility of the results. Inconsistence between findings may suggest chance findings or incorrect assumptions in theories or analysis (e.g. correlations between two regression models parameters may result in two oppositely directed effects). 'Related' means sharing a (pathophysiological) pathway or related characteristics (e.g. employment and income). The stronger such measures are related, the more consistent the results should be. For systematic reviews or a body of evidence, this item does not involve the comparison of the same moderator or predictor effect across different studies, but the comparison of different moderator or predictor effects.

***Consider:*** Whether multiple related moderators or predictors show the same effects (variation in effect size may apply) for the same outcome; whether a single moderator or predictor shows similar effect across different related outcomes (e.g. survival and incidence of infection).

***Implications:*** if related moderators or outcomes show similar effects, this contributes to the credibility of individual findings. If contradictory effects are found, other effects may underlie the observed moderator or predictor effects, or observed effects apply only under specific conditions.

**Item 11.** *Were the setting and study population comparable to the setting and population in which the information would be used?*

The identification of moderators and predictors helps increasing the transferability of findings, but it remains important to determine whether findings from a study can be validly applied to the target setting.

***Consider:*** Whether target and sampled population are similar for moderators and predictors that have been tested (e.g. same range of age), as well as similar on moderators and predictors that have not been tested (geographical location, socio-economical-status, support from others, etc.).

***Implications:*** Credibility of findings is compromised when they have to be extrapolated to different settings or populations. Moderator or predictor effects may behave differently in other settings (being moderated themselves), for instance because of underlying or related moderator or predictor effects, or practical differences in the treatment applied.

**Item 12.** *Is the moderator or predictor effect clinically important?*

A clinically important moderator or predictor is one that has a considerable effect (i.e. larger than any measurement error as well as sufficient size) and one that is likely to be able to be implemented in practice. For example, ethical, legal or practical issues may prohibit the use of moderators or predictors in clinical practice or assessment, thus reducing its usefulness. Unless authors make explicit statements on the clinical importance of a moderator or predictor effect, specific expertise may be needed to address this item.

Differences between subgroups (predictor or moderator effects) should always be interpreted with caution, even if they are based on formal tests. Explicit presentation of findings in different subgroups may help in the correct interpretation of the relevance of the results. The difference between subgroups can be presented as a mean difference (with standard deviation), standardized mean differences or Cohen's d.

***Consider:*** consider whether the reported group moderator or predictor effect is clinically relevant. Also consider whether confidence intervals or statistical significance were reported; as without these measures clinical importance is more difficult to estimate. Confidence intervals add to the interpretability of effects and add to the validity of the estimation of effects when used in models.

Furthermore, whether the effects that were found, in relation to their practical implementation, may be considered of benefit. This needs to be related to any difficulties one may encounter when implementing the moderator or predictor. For instance, ethical problems may be overcome if the effect is very large. If the authors made statements on clinical importance, be critical and try to assess whether their arguments stand firm.

***Implications:*** for individual studies the implications may be small. It is better to reconsider clinical importance in view of all the evidence available on the moderator or predictor. If the finding is from a large study or body of evidence, the implication may be that the observed effect, although credible, may not be practicable in daily care.

## Items for body of evidence

### Item 13. *Was the moderator or predictor effect reasonably homogenous across studies?*

Moderator and predictor effects should be homogenous across different studies. Effect sizes are likely to vary to some degree, but the greater the differences the less likely the moderator or predictor effect is a true effect. This is especially true if the effects over studies are contradictory (e.g. one study showing a protective effect, while another showing a harmful effect of a moderator). If low agreement between studies is found, it is important to determine possible causes: differences in study populations, study designs, methods of analysis, or actually different moderator or predictor effects?

***Consider:*** Whether the moderator or predictor effects are approximately equally sized (i.e. the conclusion of the moderator or predictor effect would not change if studies are excluded) across studies and point in the same direction.

***Implications:*** Differences in the size or direction of effects impact the credibility of the findings. This impact can be quantified by performing a meta-analysis. If larger differences are found, or even worse, differences in direction of an effect, the credibility of findings is clearly compromised.

It may also be possible that differences between studies can be explained. If this is the case, one could say that there are moderated moderators or predictors. Establishing the effects of such an extra level of moderation should, however, be regarded as an additional analysis. That is, all considerations in this checklist apply to that moderator as well. It may be quite challenging to estimate the effects of these factors to a useful degree without performing further research.

### Item 14. *Was the moderator or predictor measured similarly across the included studies, or was there an adequate conversion performed?*

One problem that moderators and predictor share with many outcomes is that they can be measured differently between studies. Another problem with moderators and predictors is that if moderators or predictors are investigated in a multivariate method, they can greatly influence the effect of the other moderators or predictors in the same model.

***Consider:*** Whether the moderator or predictor effect is measured using a similar instrument (e.g. the same questionnaire, scale or other tool). If this is not the case, consider whether conversions are possible. Also consider, if moderators or predictors are assessed in multivariate analysis, whether a similar set of moderators/predictors/other factors is taken into account.

***Implications:*** If measurements of moderators or predictors are not comparable, it is difficult to determine a pooled effect, to determine usefulness of results or credibility of a body of evidence. If that is the case, one should adhere to the tools that are used in the target setting.

If the moderator or predictor was included in multivariate models with different sets of other included factors, those other factors may greatly influence the observed effect of a moderator, and thus the comparability of the moderator or predictor across studies. Furthermore, it may be more difficult to determine the effect size if it cannot be established which other moderator or predictors need to be taken into account.

## 7.3 CREATION OF APPRAISAL CHECKLIST FOR PPTOS

For the development of the appraisal checklist for PPTOs different types of methods used to elicit preferences were explored The search for methodologies was complicated by the fact that literature in which these can be found is diverse and of an unknown quantity. An extensive literature search was conducted to find methods for eliciting patient preferences for treatment outcome. Aim of the search strategy was to perform searches with low precision but high sensitivity. To this end, Google Scholar was used alongside PubMed.

Not all results of every search were reviewed, as this would result in large numbers of low-yield results. Instead, the first 300 results were scanned on title and displayed data (text snippet for Google Scholar, authors and publication type/location for PubMed). For both Google Scholar and PubMed the 'relevance' sorting method was used, no restrictions on publication year or language were applied. No preselection of journals was applied either. Only promising results were scanned further on abstract, then full text. Searches were continued until no new results came up and/or theoretical saturation occurred (Black et al., 1998). Additional literature was found through citation chasing, i.e. following references or searching for authors of interesting papers found through the initial search or incremental searches.

After drawing up an initial inventory of methods using expert opinion and literature (including methodological reviews and results from the hand-search for articles on patient preferences for treatment outcome) further searches were performed for each method individually to expand further details on that method. In PubMed, MeSH-terms were used as much as possible to include these methods in the search query, but equivalent non-MeSH keywords were used as well.

The following keywords were used in the search strategy (as MeSH-terms or other):

- Preference*
- Elicitation*
- Treatment Outcome*
- Outcome*
- Patient Preference
- Patient
- Preference
- Method*
- Methodological Review
- Health Status Indicator
- Randomized Controlled Trials as topic/methods
- Meta-Analysis as topic
- Research guidelines as topic
- Decision making
- Appraisal
- Critical Appraisal
- Guideline

Extremely generic terms, such as 'appraisal' and 'critical appraisal', were used in combination with other terms to keep the result set focused enough.

Selected articles were first scanned on abstract then full-text, to determine whether they contained possible information on the methodology, guidelines or appraisal on (treatment outcome) preference elicitation methods. Though it was not our purpose to generate reporting guidelines for preference elicitation methods, we did include reporting guidelines in our search results as they can give valuable information on important aspects of performed research, which can be used to appraise and execute the methods. Methods that were used to assess health-states were included as well, as they can be used to assess treatment outcome preferences indirectly.

Identified data was summarized. A list of items was created by identifying a number of key factors (recurring appraisal themes) and describing the appraisal criteria as background for these items. To improve the ability of the checklist (e.g. to compare multiple studies or raters) it was initially decided to provide a fixed set of answering categories for each item (i.e. Yes, Partially, Not (clearly) described, No, Don't know, Not applicable). The contents of the appraisal checklist as it was tested can be found in the box below.

**Box 1: Test version of the appraisal checklist**

1. Does the study address the patient preferences for treatment outcome I am interested in?
2. Is it clear what the researchers did (i.e. was the description of population, methods, and analysis clear and complete)?
3. Are the data collection methods appropriate and appropriately used?
   *a.* Is the format of included questions appropriate?
   *b.* 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. Is the quality of the researcher or research team adequate?
6. Were the methods properly chosen, executed (reliable and valid)?
7. Are the qualitative research results (if applicable) reliable and valid?
8. Are the results transferable?

## Testing the appraisal checklist: The palliative care test case

Following the application of the search strategy, 24 articles concerning patient preferences that related to home-based palliative care, were eligible to test the appraisal checklist on. For the testing procedure, several project partners were asked to apply the appraisal checklist. The test participants came from HTA-fields spanning medical (bio)ethics, public health, epidemiology, palliative care research, and philosophy.

The 24 papers were randomly distributed among the testers. Each tester was assigned 3-4 papers. The testers were asked to extract relevant PPTO information and apply the appraisal checklist on their assigned papers and feed back any comments on using the checklist. The case study aimed to evaluate the added benefit of caregiver support in home-based palliative care, however we included all papers that concerned patient preferences relevant to home-based palliative care in general (i.e. not specifically to caregiver support).

*User feedback and suggestions for adjustments*

The testing phase showed that the critical appraisal checklist was able to show differences in study quality. We did not measure inter-rater agreement, so we cannot present any results in that regard.

Most feedback concerned the background information (i.e. the provided information to help answer the items). A large number of minor textual changes and clarifications were applied to the explanatory text for each of the items to resolve several issues reported when interpreting this information. A few larger changes to the checklist are discussed below.

**Item 1:** *Does the study address the patient preferences for treatment outcome I am interested in?*

**Reworded to** 'Does the study address relevant patient preferences for treatment outcome'.

It was pointed out that the word 'interesting' is a value-laden item. Users might interpret it as relevant (concerning the subject) or as sizable (i.e. is the finding sufficiently large), usable (can I do something with this finding), etc.. In order to remove the possible ambiguity, the item was reworded.

**Item 5:** *Is the quality of the researcher or research team adequate?*

**Removed:** Even though the quality of the researcher was mentioned often to be of great influence on the quality of research, and in qualitative research in particular, it was decided to remove this item from

the appraisal checklist. It was stated that the item was difficult to answer, as it is challenging to judge the quality of a researcher or research group. An inexperienced researcher may produce good research and the opposite may also be true. Hence, it was decided to drop this item for the definite appraisal checklist.

**Item 6:** *Were the methods properly chosen, executed (reliable and valid)?*

**Merged with item 7:** Users often used the option 'don't know' with this item. This is the item that has most method-specific appraisal requirements. Although no reasons were stated for these answers, we found that all studies where users used this option were interview-based (without using specific quantitative methods). The background information of item 6 in the test version did not include information on determining the reliability or validity of such interviews – which might explain this problem.

**Item 7:** *Are the qualitative research results (if applicable) reliable and valid?*

**Merged with item 6:** Item 7 was considered not applicable in almost 40% of the cases. As the item specifically concerned qualitative findings this result was in line with the findings in regard to item 6. Based on these results and the thematic overlap between items 6 and 7 (reliability and validity) it was decided to merge these items to simplify the appraisal checklist.

The initial set of answering categories ('Yes', 'No', 'Partially', 'Not (clearly) described', 'don't know' and 'Not applicable') proved to be less beneficial than anticipated. As the appraisal was aimed towards informing HTAs and similar research, there is little benefit from this amount of answering possibilities. In order to simplify the checklist, the number of answering categories was reduced to 'Yes', 'No', and 'Don't know'. Furthermore, we added an extra item asking for an overall judgement to force the user to think explicitly about the results of the checklist and its effect on study quality, instead of just the checked answers.

The revised appraisal checklist can be found in chapter 7.4.

## 7.4 APPRAISAL CHECKLIST FOR PATIENT PREFERENCES FOR TREATMENT OUTCOMES

This appraisal checklist for studies on patient preferences for treatment outcomes (PPTOs) is intended to be used to appraise reported findings concerning PPTOs. It addresses evidence quality but also relevancy of the findings. The checklist 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. This checklist can be used for many different study types, including qualitative and quantitative methods. The criteria are not meant to be used as a list 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. Furthermore, it aims to help the user to determine if the results of the study are sufficiently relevant to take into account when making recommendation for treatment decisions (e.g. creating protocols or assessing technologies). 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/).

| | Yes | No | Don't know | Not 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? | ☐ | ☐ | ☐ | ☐ |
| ***a.*** Is the format of included questions appropriate? | ☐ | ☐ | ☐ | ☐ |
| ***b.*** 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

| | Yes | No | Don'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.

### **Item 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.

### **Item 2.** *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).

### **Item 3.** *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.

***a. 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.

***b. 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.*

---

### **Item 4.** *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.

### **Item 5.** *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.

### **Item 6.** *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.

## 7.5 TYPES OF PREFERENCE ELICITATION METHODS

In this chapter an overview of preference elicitation methods is presented. The list is not exhaustive, but should provide enough information to choose which method can be used in which cases to determine PPTOs. This information can be used to select preference elicitation methods to perform new research (see chapter 3.3.3) but may also help to better understand key aspects of specific types of research when using the appraisal checklist.

### *Qualitative methods*

#### *Interview*

Interviews are among the most popular qualitative methods (Facey et al., 2010). Usually one interviewer poses questions to an interviewee (or a group thereof) and the responses are recorded and analyzed later. If the interview is face-to-face and the interviewer needs to go more in-depth on a certain answer, or if it was not clear enough, he or she can ask directly. The interview can be structured (i.e. using a list of pre-defined interview-subjects or specific questions), open (no specific questions or subjects), or anywhere in between. The success or failure of interviews depends entirely on the quality of the interviewers, questions asked and the respondent selection. There are various interview techniques and analysis techniques available. For further reading, see McLellan et al. (2003); DiCicco-Bloom and Crabtree (2006); or Gill et al. (2008).

#### *Focus group*

Focus groups are similar to group interviews in many ways. The key difference lies in the importance of interviewee-interviewee communication. Focus groups rely on participant interaction as primary mode of data generation; the description of the interaction of participants is as important as the answers to the questions posed by the interviewer (Kitzinger, 1994). This makes focus groups more suitable for complicated matters, where interaction is needed to get to the depth of the matters required. See additionally Kitzinger (1995); Powell and Single (1996); Stevens (1996); Hyde et al. (2005); Gill et al. (2008); or Onwuegbuzie et al. (2009).

### *Quantitative methods*

#### *Ranking-based methods*

Ranking methods ask respondents to rank two or more items. For instance, by numbering items or reordering items in a list. Ranking methods are fairly straightforward for respondents to perform, however the analysis requires some thought (e.g. whether or not to include anchoring items). Furthermore, the method has some limitations in the number of items, and respondents will not be able to convey the strength of their preferences. See also Ben-Akiva et al. (1992). Due to their low cognitive burden, ranking exercises are suitable for clinical decisions, but due to their simplicity less suitable for HTA (Weernink et al., 2014).

#### *Rating-based methods*

Rating methods allow respondents to place items on a certain scale, anchored between a preset minimum and maximum (e.g. on a bar), or assign a score (e.g. 1-5) to individual items. Though this exercise has only a low cognitive burden, comparisons between multiple items scored are not: Individual ratings may not be comparable (an item with a score twice as high may not be twice as (un) important), or be on the same scale (is the underlying score linear?). Furthermore, respondents may give multiple items the same score, making preferences between these items indistinguishable. Lastly, the sequence of items to rate can influence the value respondents give. Similar to the ranking methods, rating-based methods are most suitable for clinical decision making (Weernink et al., 2014).

#### *Visual Analogue Scale*

The VAS is a special kind of rating method, usually used to rate health states or pain on a scale, usually by asking respondents to place the item on a line, usually anchored on a minimum value (e.g. no pain) to a maximum (e.g. unbearable pain). It may even be considered a ranking exercise when respondents are asked to place multiple items on the same line. However, this may be a rather difficult exercise for respondents. See also McCormack et al. (1988); Brazier et al. (1999); Ryan et al. (2001); or Boer et al. (2004).

*DCE/DCM*

One of the more common methods in HTA to determine patient preferences, discrete choice experiments (or models) are procedures where participants are asked to make a choice between two scenarios with different treatment outcomes. By repeating this question multiple times with varying outcomes in the scenarios, a relative importance of each of these outcomes can be determined using a mathematical model. This method is ideal for rating the importance of various outcomes, however it can be a rather difficult task for respondents and analysis of the answers is not as straightforward as other methods. See for additional information: de Bekker-Grob et al. (2012); World Health Organization (2012); or de Bekker-Grob et al. (2015).

*Standard Gamble*

The Standard Gamble (SG) method offers respondents two choices: A comparator scenario (e.g. full health or a status quo with a certain treatment outcome) and a scenario where the respondents is offered a gamble (risk) between resolving the problem (treatment outcome) and death. This way, the standard gamble can be used to determine the severity of a treatment outcome compared to death. Some of the properties, advantages and disadvantages of the SG-method are described in Brazier et al. (1999). Risks are difficult to convey to lay people, especially when the risks are very large or small or involve serious events, such as death. People may then become risk-averse (do not want to take any chances). This may make this method less suitable for when such risks or events are concerned (Garcia-Retamero et al., 2012).

*Time-TradeOff*

In the Time-TradeOff (TTO), respondents are asked to choose between two scenarios, such as health states or treatment outcomes, at various durations (e.g. number of years it lasts). Variants exist as well, such as the Person TradeOff, where respondents are asked to make tradeoffs in number of deaths, for instance. By repeating the question for various numbers, a value can be determined at which point respondents are indifferent between the two offered scenarios. The inverse ratio between the two durations is equal to the ratio of the values of the offered outcomes. By repeating the question with varying items, the relative preference value of individual outcomes can be determined.

Time-tradeoff methods are often used to value health states to derive weights for health statues questionnaires. The TTO method results in good repeatability but higher rates of internal consistencies (Ryan et al., 2001). Its results are comparable to that of the SG. One of the criticisms of the TTO method is the assumption of time-independence, or 'constant proportionality'. That is, a health state lasting two year is valued exactly twice as much as the same health state lasting one year. Various studies indicate this assumption does not (always) hold (Craig, 2009). As with the SG, there is the problem of non-traders: people who are unwilling to trade any length of life for quality of life. The TTO method is mostly advocated for chronic or longer lasting conditions (Dolan & Gudex, 1995). Some of the properties, advantages and disadvantages of the TTO-method are described in Brazier et al. (1999).

*Willingness To Pay*

The Willingness To Pay (WTP) method asks respondents how much they are willing to pay to reach a certain status, gain a certain good, etc. There are similar measures, such as willingness to travel, for things that cannot easily be expressed in monetary value. The WTP method is often used in cost-effectiveness research, but it has some limitations where items cannot easily be expressed in monetary values, which is often the case in health-related matters. See also Ryan and San Miguel (2000). Some of the properties, advantages and disadvantages of the SG-method are described in Brazier et al. (1999).

***Other methods***

*Delphi panel*

The Delphi procedure is a method for gaining consensus on specific issues. It starts by offering one participant a question, questionnaire, interview or similar question. The answer(s) of the expert are processed by the researcher: Summarized, reworded, transcribed, etc. The next participant is offered the same questions as the first, but the product of the previous step is displayed as possible answer. The participant is asked to reword, add, or correct the answers provided and again this adjusted answer is processed and shown to the next participant. The process continues until the adjustments made by the different participants are no

longer significant: agreement has been reached. Variants exist where the answers are purely numeric, for instance the Research ANd Development appropriateness method (Fitch et al., 2001): each participant answers multiple choice questions, and the next round tries to answer these again, but this time with the results from the previous rounds given (e.g. how many chose A, how many chose B, etc). The iteration steps in the Delphi procedure should ultimately result in a consensus, however the researchers responsible for the processing of answers play a large role in shaping the consensus (and determining when the process stops). Delphi procedures do not require the participants to be in the same room (ideally the answers are blinded even) which makes them very suitable for gathering opinions from people from all over the world, relatively easy. This is even more so the case with the RAND-appropriateness method, as all participants can enter a round in parallel. If the participants differ too much in their opinion, the process may go on forever; in these cases it may be best to group experts by their opinion/view/background. See also Hasson et al. (2000); Okoli and Pawlowski (2004); Baker et al. (2006); or Hsu and Sandford (2007).

*Citizen jury*

In the citizen jury, a relatively large group of people convenes and presentations or lectures are given to educate the participants on the problem at hand, background information required to make certain decisions, and views from patients, experts, policy makers and other stakeholders. Eventually, participants have to make a decision, rating, ranking or voting on the issues at hand. One way in which patient preferences can be determined is by first making an inventory of issues and subsequently have participants vote on the most important ones. Citizen juries offer a lot of possibilities, allow even lay participants to make well-informed decisions. However, they cost a lot of resources and time to organize. See also Lenaghan (1999); Iredale and Longley (2007); Menon and Stafinski (2008); Street et al. (2014); or Whitty et al. (2014).

*Nominal group technique / Concept mapping*

In the nominal group technique and concept mapping methods, several stakeholders convene and try to identify possible issues (e.g. specific treatment outcomes) and through a process of discussions, selections and prioritizations come up with a list of important items. Both selection and prioritization can be done visually, by mapping items on a grid and grouping them on various properties, or purely by discussions, voting or rating. The contents of these methods are not set in stone, and can be different depending on the preferences of the moderator/organizer, the problems being discussed, the participants or other properties. The resource costs for these methods are much lower than for citizen juries, making them quite attractive for structured, discussion-based preference elicitation. For further reading see Deip et al. (1977); Trochim (1989); Trochim and Kane (2005); Novak and Cañas (2006); Novak and Cañas (2008); or Hiligsmann et al. (2013).

## 7.6 REFERENCES


AARONSON, N., ALONSO, J., BURNAM, A., LOHR, K.N., PATRICK, D.L., PERRIN, E., STEIN, R.E. (2002) Assessing health status and quality-of-life instruments: attributes and review criteria. Quality of Life Research, 11, 193-205.

ANDERSON, C. (2010) Presenting and evaluating qualitative research. American Journal of Pharmaceutical Education, 74, 141.

ASSMANN, S.F., POCOCK, S.J., ENOS, L.E., KASTEN, L.E. (2000) Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet, 355, 1064-1069.

ATTEMA, A.E., EDELAAR-PEETERS, Y., VERSTEEGH, M.M., STOLK, E.A. (2013) Time trade-off: one methodology, different methods. The European Journal of Health Economics, 14, 53-64.

AUSTIN, P.C., STEYERBERG, E.W. (2015) The number of subjects per variable required in linear regression analyses. J Clin Epidemiol, 68, 627-636.

BAKER, J., LOVELL, K., HARRIS, N. (2006) How expert are the experts? An exploration of the concept of 'expert' within Delphi panel techniques. Nurse Researcher, 14, 59-70.

BARON, R.M., KENNY, D.A. (1986) The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. Journal of Personality and Social Psychology, 51, 1173-1182.

BEN-AKIVA, M., MORIKAWA, T., SHIROISHI, F. (1992) Analysis of the reliability of preference ranking data. Journal of Business Research, 24, 149-164.

BIGGER, J.T. (2003) Issues in Subgroup Analyses and Meta-Analyses of Clinical Trials. Journal of cardiovascular electrophysiology, 14, S6-S8.

BLACK, N., REEVES, B., BRAZIER, J., FITZPATRICK, R. (1998) Health Services Research Methods: A Guide to Best Practice. (1 edition ed.).London: BMJ Books.

BOER, A.G.E.M.D., LANSCHOT, J.J.B.V., STALMEIER, P.F.M., SANDICK, J.W.V., HULSCHER, J.B.F., HAES, J.C.J.M.D., SPRANGERS, M.A.G. (2004) Is a single-item visual analogue scale as valid, reliable and responsive as multi-item scales in measuring quality of life? Quality of Life Research, 13, 311-320.

BOOTH, A., CARROLL, C., ILOTT, I., LOW, L.L., COOPER, K. (2013) Desperately Seeking Dissonance Identifying the Disconfirming Case in Qualitative Evidence Synthesis. Qualitative Health Research, 23, 126-141.

BOULKEDID, R., ABDOUL, H., LOUSTAU, M., SIBONY, O., ALBERTI, C. (2011) Using and Reporting the Delphi Method for Selecting Healthcare Quality Indicators: A Systematic Review. PLoS ONE, 6, e20476.

BRAZIER, J., DEVERILL, M., GREEN, C., HARPER, R., BOOTH, A. (1999) A review of the use of health status measures in economic evaluation. Health Technology Assessment (Winchester, England), 3, i-iv, 1-164.

BRERETON, L., WAHLSTER, P., LYSDAHL, K.B., MOZYGEMBA, K., BURNS, J., CHILCOTT, J.B., WARD, S., BRÖNNEKE, J.B., TUMMERS, M., VAN HOORN, R., PFADENHAUER, L., POLUS, S., INGLETON, C., GARDINER, C., VAN DER WILT, G.J., GERHARDUS, A., ROHWER, A., REHFUESS, E., OORTWIJN, W., REFOLO, P., SACCHINI, D., LEPPERT, W., BLAZEVICIENE, A., SPAGNOLO A.G., PRESTON, L., CLARK, J., GOYDER, E., ON BEHALF OF THE INTEGRATE-HTA TEAM (2016) Integrated assessment of home based palliative care with and without reinforced caregiver support: 'A demonstration of INTEGRATE-HTA methodological guidances' [Online]. Available from: http://www.integrate-hta.eu/downloads/

BRIDGES, J.F.P., HAUBER, A.B., MARSHALL, D., LLOYD, A., PROSSER, L.A., REGIER, D.A., JOHNSON, F.R., MAUSKOPF, J. (2011) Conjoint analysis applications in health--a checklist: a report of the ISPOR Good Research Practices for Conjoint Analysis Task Force. Value in Health, 14, 403-413.

BURNS, K.E.A., DUFFETT, M., KHO, M.E., MEADE, M.O., ADHIKARI, N.K.J., SINUFF, T., COOK, D.J. (2008) A guide for the design and conduct of self-administered surveys of clinicians. Canadian Medical Association Journal, 179, 245-252.

CARROLL, C., BOOTH, A., LLOYD-JONES, M. (2012) Should We Exclude Inadequately Reported Studies From Qualitative Systematic Reviews? An Evaluation of Sensitivity Analyses in Two Case Study Reviews. Qualitative Health Research, 22, 1425-1434.

CLARK, J.P. (2003) How to peer review a qualitative manuscript. Peer review in health sciences, 2, 219–235.

CLARK, M.D., DETERMANN, D., PETROU, S., MORO, D., DE BEKKER-GROB, E.W. (2014) Discrete choice experiments in health economics: a review of the literature. PharmacoEconomics, 32, 883-902.

CLIBBENS, N., WALTERS, S., BAIRD, W. (2012) Delphi research: issues raised by a pilot study. Nurse Researcher, 19, 37-44.

COAST, J., AL-JANABI, H., SUTTON, E.J., HORROCKS, S.A., VOSPER, A.J., SWANCUTT, D.R., FLYNN, T.N. (2012) Using qualitative methods for attribute development for discrete choice experiments: issues and recommendations. Health Economics, 21, 730-741.

COOK, C. (2008) Potential pitfalls of clinical prediction rules. The Journal of Manual & Manipulative Therapy, 16, 69-71.

COOK, C. (2010) Mode of administration bias. The Journal of Manual & Manipulative Therapy, 18, 61-63.

CORFMAN, K. (1995) The importance of member homogeneity to focus group quality. Advances in Consumer Research, 22, 354–354.

CRAIG, B.M. (2009) The duration effect: a link between TTO and VAS values. Health Economics, 18, 217-225.

DAMSCHRODER, L.J., UBEL, P.A., RIIS, J., SMITH, D.M. (2007) An Alternative Approach for Eliciting Willingness-to-Pay: A Randomized Internet Trial.

DE BEKKER-GROB, E.W., DONKERS, B., JONKER, M.F., STOLK, E.A. (2015) Sample Size Requirements for Discrete-Choice Experiments in Healthcare: a Practical Guide. Patient, 8, 373-384.

DE BEKKER-GROB, E.W., RYAN, M., GERARD, K. (2012) Discrete choice experiments in health economics: a review of the literature. Health Economics, 21, 145-172.

DEIP, P., THESEN, A., MOTIWALLA, J., SESHARDI, N. (1977) Systems tools for project planning: Nominal Group Technique. Bloomington, Indiana: International Development Institute.

DICICCO-BLOOM, B., CRABTREE, B.F. (2006) The qualitative research interview. Medical Education, 40, 314-321.

DIJKMAN, B., KOOISTRA, B., BHANDARI, M., EVIDENCE-BASED SURGERY WORKING, G. (2009) How to work with a subgroup analysis. Can J Surg, 52, 515-522.

DOLAN, P., GUDEX, C. (1995) Time preference, duration and health state valuations. Health Economics, 4, 289-299.

EDWARDS, P. (2010) Questionnaires in clinical trials: guidelines for optimal design and administration. Trials, 11, 2.

FACEY, K., BOIVIN, A., GRACIA, J., HANSEN, H.P., LO SCALZO, A., MOSSMAN, J., SINGLE, A. (2010) Patients' perspectives in health technology assessment: a route to robust evidence and fair deliberation. International Journal of Technology Assessment in Health Care, 26, 334-340.

FITCH, K., BERNSTEIN, S.J., AGUILAR, M.S., BURNAND, B., LACALLE, J.R., LAZARO, P., VAN HET LOO, M., MCDONNELL, J., VADER, J., KAHAN, J.P. (2001) The RAND/UCLA Appropriateness Method User's Manual [Online].

FLYNN, T.N., LOUVIERE, J.J., PETERS, T.J., COAST, J. (2007) Best–worst scaling: What it can do for health care research and how to do it. Journal of Health Economics, 26, 171-189.

FORTUNE, A.E., REID, W.J., MILLER, R. (2013) Qualitative Research in Social Work. Columbia University Press.

GABLER, N.B., DUAN, N., LIAO, D., ELMORE, J.G., GANIATS, T.G., KRAVITZ, R.L. (2009) Dealing with heterogeneity of treatment effects: is the literature up to the challenge. Trials, 10, 43.

GAGNIER, J., MORGENSTERN, H., ALTMAN, D., BERLIN, J., CHANG, S., MCCULLOCH, P., SUN, X., MOHER, D., GROUP, F.T.A.A.C.H.C. (2013) Consensus-based recommendations for investigating clinical heterogeneity in systematic reviews. BMC Medical Research Methodology, 13, 106.

GAGNIER, J.J., MOHER, D., BOON, H., BEYENE, J., BOMBARDIER, C. (2012) Investigating clinical heterogeneity in systematic reviews: a methodologic review of guidance in the literature. BMC Med Res Methodol, 12, 111.

GARCIA-RETAMERO, R., OKAN, Y., COKELY, E.T. (2012) Using Visual Aids to Improve Communication of Risks about Health: A Review. The Scientific World Journal, 2012, e562637.

GIACOMINI, M.K., COOK, D.J., FOR THE EVIDENCE-BASED MEDICINE WORKING GROUP (2000) Users' guides to the medical literature: Xxiii. qualitative research in health care a. are the results of the study valid? JAMA, 284, 357-362.

GILL, P., STEWART, K., TREASURE, E., CHADWICK, B. (2008) Methods of data collection in qualitative research: interviews and focus groups. British Dental Journal, 204, 291-295.

GREENWOOD, N., ELLMERS, T., HOLLEY, J. (2014) The influence of ethnic group composition on focus group discussions. BMC Medical Research Methodology, 14, 107.

GROENWOLD, R.H., DONDERS, A.T., VAN DER HEIJDEN, G.G., HOES, A.W., ROVERS, M.M. (2009) COnfounding of subgroup analyses in randomized data. Archives of Internal Medicine, 169, 1532-1534.

GUEST, G., BUNCE, A., JOHNSON, L. (2006) How Many Interviews Are Enough? An Experiment with Data Saturation and Variability. Field Methods, 18, 59-82.

GUYATT, G., GUTTERMAN, D., BAUMANN, M.H., ADDRIZZO-HARRIS, D., HYLEK, E.M., PHILLIPS, B., RASKOB, G., LEWIS, S.Z., SCHÜNEMANN, H. (2006) Grading strength of recommendations and quality of evidence in clinical guidelines: report from an american college of chest physicians task force. Chest, 129, 174-181.

GUYATT, G.H., OXMAN, A.D., KUNZ, R., WOODCOCK, J., BROZEK, J., HELFAND, M., ALONSO-COELLO, P., GLASZIOU, P., JAESCHKE, R., AKL, E.A., NORRIS, S., VIST, G., DAHM, P., SHUKLA, V.K., HIGGINS, J., FALCK-YTTER, Y., SCHUNEMANN, H.J., GROUP, G.W. (2011) GRADE guidelines: 7. Rating the quality of evidence--inconsistency. J Clin Epidemiol, 64, 1294-1302.

HASSON, F., KEENEY, S., MCKENNA, H. (2000) Research guidelines for the Delphi survey technique. Journal of Advanced Nursing, 32, 1008-1015.

HEYVAERT, M., HANNES, K., MAES, B., ONGHENA, P. (2013a) Critical Appraisal of Mixed Methods Studies. Journal of Mixed Methods Research, 7, 302-327.

HEYVAERT, M., MAES, B., ONGHENA, P. (2013b) Mixed methods research synthesis: definition, framework, and potential. Quality & Quantity, 47, 659-676.

HILIGSMANN, M., VAN DURME, C., GEUSENS, P., DELLAERT, B.G., DIRKSEN, C.D., VAN DER WEIJDEN, T., REGINSTER, J.-Y., BOONEN, A. (2013) Nominal group technique to select attributes for discrete choice experiments: an example for drug treatment choice in osteoporosis. Patient Preference and Adherence, 7, 133-139.

HSU, C.-C., SANDFORD, B.A. (2007) The Delphi technique: making sense of consensus. Practical Assessment, Research & Evaluation, 12, 1–8.

HYDE, A., HOWLETT, E., BRADY, D., DRENNAN, J. (2005) The focus group method: Insights from focus group interviews on sexual health with adolescents. Social Science & Medicine, 61, 2588-2599.

IGLESIAS, C., TORGERSON, D. (2000) Does Length of Questionnaire Matter? A Randomised Trial of Response Rates to a Mailed Questionnaire. Journal of Health Services Research & Policy, 5, 219-221.

IREDALE, R., LONGLEY, M. (2007) From passive subject to active agent: the potential of Citizens' Juries for nursing research. Nurse Education Today, 27, 788-795.

IRVINE, A. (2011) Duration, Dominance and Depth in Telephone and Face-to-Face Interviews: A Comparative Exploration. International Journal of Qualitative Methods, 10, 202-220.

JONES, J., HUNTER, D. (1995) Qualitative Research: Consensus methods for medical and health services research. BMJ, 311, 376-380.

KATRAK, P., BIALOCERKOWSKI, A.E., MASSY-WESTROPP, N., KUMAR, S., GRIMMER, K.A. (2004) A systematic review of the content of critical appraisal tools. BMC Medical Research Methodology, 4, 22.

KELLEY, K., CLARK, B., BROWN, V., SITZIA, J. (2003) Good practice in the conduct and reporting of survey research. International Journal for Quality in Health Care, 15, 261-266.

KITZINGER, J. (1994) The methodology of Focus Groups: the importance of interaction between research participants. Sociology of Health & Illness, 16, 103-121.

KITZINGER, J. (1995) Qualitative Research: Introducing focus groups. BMJ, 311, 299-302.

KLØJGAARD, M.E., BECH, M., SØGAARD, R. (2012) Designing a Stated Choice Experiment: The Value of a Qualitative Process. Journal of Choice Modelling, 5, 1-18.

KRAEMER, H.C., WILSON, G.T., FAIRBURN, C.G., AGRAS, W.S. (2002) Mediators and moderators of treatment effects in randomized clinical trials. Archives of General Psychiatry, 59, 877.

KUPER, A., LINGARD, L., LEVINSON, W. (2008a) Critically appraising qualitative research. BMJ, 337, a1035.

KUPER, A., REEVES, S., LEVINSON, W. (2008b) An introduction to reading and appraising qualitative research. BMJ, 337, a288-a288.

LAMERS, L.M. (2007) The transformation of utilities for health states worse than death: consequences for the estimation of EQ-5D value sets. Medical Care, 45, 238-244.

LANCSAR, E., LOUVIERE, J. (2008) Conducting discrete choice experiments to inform healthcare decision making: a user's guide. PharmacoEconomics, 26, 661-677.

LENAGHAN, J. (1999) Involving the public in rationing decisions. The experience of citizens juries. Health Policy, 49, 45-61.

LOUVIERE, J.J., PIHLENS, D., CARSON, R. (2011) Design of Discrete Choice Experiments: A Discussion of Issues That Matter in Future Applied Research. Journal of Choice Modelling, 4, 1-8.

MACLEAN, L.M., MEYER, M., ESTABLE, A. (2004) Improving Accuracy of Transcripts in Qualitative Research. Qualitative Health Research, 14, 113-123.

MALTERUD, K. (2001) Qualitative research: standards, challenges, and guidelines. The Lancet, 358, 483-488.

MANGHAM, L.J., HANSON, K., MCPAKE, B. (2009) How to do (or not to do) ... Designing a discrete choice experiment for application in a low-income country. Health Policy Plan, 24, 151-158.

MARSDEN, P.V., WRIGHT, J.D. (2010) Handbook of Survey Research. Emerald Group Publishing.

MCCOLL, E., JACOBY, A., THOMAS, L., SOUTTER, J., BAMFORD, C., STEEN, N., THOMAS, R., HARVEY, E., GARRATT, A., BOND, J. (2001) Design and use of questionnaires: a review of best practice applicable to surveys of health service staff and patients. Health Technology Assessment (Winchester, England), 5, 1-256.

MCCORMACK, H.M., DE L. HORNE, D.J., SHEATHER, S. (1988) Clinical applications of visual analogue scales: a critical review. Psychological Medicine, 18, 1007–1019.

MCLELLAN, E., MACQUEEN, K.M., NEIDIG, J.L. (2003) Beyond the Qualitative Interview: Data Preparation and Transcription. Field Methods, 15, 63-84.

MENON, D., STAFINSKI, T. (2008) Engaging the public in priority-setting for health technology assessment: findings from a citizens' jury. Health Expectations, 11, 282-293.

MULHERN, B., BANSBACK, N., BRAZIER, J., BUCKINGHAM, K., CAIRNS, J., DEVLIN, N., DOLAN, P., HOLE, A.R., KAVETSOS, G., LONGWORTH, L., ROWEN, D., TSUCHIYA, A. (2014) Preparatory study for the revaluation of the EQ-5D tariff: methodology report. Health Technology Assessment (Winchester, England), 18, vii-xxvi, 1-191.

NICHOLAS, D.B., GLOBERMAN, J., ANTLE, B.J., MCNEILL, T., LACH, L.M. (2008) Processes of Metastudy: A Study of Psychosocial Adaptation to Childhood Chronic Health Conditions. International Journal of Qualitative Methods, 5, 55-66.

NOVAK, J.D., CAÑAS, A.J. (2006) The Origins of the Concept Mapping Tool and the Continuing Evolution of the Tool. Information Visualization, 5, 175-184.

NOVAK, J.D., CAÑAS, A.J. (2008) The theory underlying concept maps and how to construct and use them. Florida Institute for Human and Machine Cognition.

O'REILLY, M., PARKER, N. (2013) 'Unsatisfactory Saturation': a critical exploration of the notion of saturated sample sizes in. Qualitative Research, 13, 190-197.

OKOLI, C., PAWLOWSKI, S.D. (2004) The Delphi method as a research tool: an example, design considerations and applications. Information & Management, 42, 15-29.

ONWUEGBUZIE, A.J., DICKINSON, W.B., LEECH, N.L., ZORAN, A.G. (2009) Toward More Rigor in Focus Group Research: A New Framework for Collecting and Analyzing Focus Group Data. International Journal of Qualitative Methods, 8, 1-21.

OXMAN, A.D., GUYATT, G.H. (1992) A consumer's guide to subgroup analyses. Ann Intern Med, 116, 78-84.

PASSMORE, C., DOBBIE, A.E., PARCHMAN, M., TYSINGER, J. (2002) Guidelines for constructing a survey. Family Medicine, 34, 281-286.

PATRICK, D.L., STARKS, H.E., CAIN, K.C., UHLMANN, R.F., PEARLMAN, R.A. (1994) Measuring preferences for health states worse than death. Medical Decision Making, 14, 9-18.

PHELLAS, C.N., BLOCH, A., SEALE, C. (2011) Structured methods: interviews, questionnaires and observation. Researching Society and Culture. London: SAGE Publications Ltd, 181–205.

PINCUS, T., MILES, C., FROUD, R., UNDERWOOD, M., CARNES, D., TAYLOR, S. (2011) Methodological criteria for the assessment of moderators in systematic reviews of randomised controlled trials: a consensus study. BMC Medical Research Methodology, 11, 14.

POWELL, R.A., SINGLE, H.M. (1996) Focus Groups. International Journal for Quality in Health Care, 8, 499-504.

RABIEE, F. (2004) Focus-group interview and data analysis. The Proceedings of the Nutrition Society, 63, 655-660.

RATTRAY, J., JONES, M.C. (2007) Essential elements of questionnaire design and development. Journal of Clinical Nursing, 16, 234-243.

REED JOHNSON, F., LANCSAR, E., MARSHALL, D., KILAMBI, V., MÜHLBACHER, A., REGIER, D.A., BRESNAHAN, B.W., KANNINEN, B., BRIDGES, J.F.P. (2013) Constructing experimental designs for discrete-choice experiments: report of the ISPOR Conjoint Analysis Experimental Design Good Research Practices Task Force. Value in Health, 16, 3-13.

ROBINSON, A., SPENCER, A. (2006) Exploring challenges to TTO utilities: valuing states worse than dead. Health Economics, 15, 393-402.

RYAN, M., SAN MIGUEL, F. (2000) Testing for consistency in willingness to pay experiments. Journal of Economic Psychology, 21, 305-317.

RYAN, M., SCOTT, D.A., REEVES, C., BATE, A., VAN TEIJLINGEN, E.R., RUSSELL, E.M., NAPPER, M., ROBB, C.M. (2001) Eliciting public preferences for healthcare: a systematic review of techniques. Health Technology Assessment (Winchester, England), 5, 1-186.

SAGE PUBLICATIONS (2008) The SAGE Encyclopedia of Qualitative Research Methods. 2455 Teller Road, Thousand Oaks California 91320 United States: SAGE Publications, Inc.

SCHMIDT, R.C. (1997) Managing Delphi Surveys Using Nonparametric Statistical Techniques*. Decision Sciences, 28, 763-774.

SHAW, J.W., JOHNSON, J.A., COONS, S.J. (2005) US valuation of the EQ-5D health states: development and testing of the D1 valuation model. Medical Care, 43, 203-220.

STEVENS, P.E. (1996) Focus groups: collecting aggregate-level data to understand community health phenomena. Public Health Nursing (Boston, Massachusetts), 13, 170-176.

STREET, J., DUSZYNSKI, K., KRAWCZYK, S., BRAUNACK-MAYER, A. (2014) The use of citizens' juries in health policy decision-making: a systematic review. Social Science & Medicine, 109, 1-9.

STREET, J.M., BRAUNACK-MAYER, A.J., FACEY, K., ASHCROFT, R.E., HILLER, J.E. (2008) Virtual community consultation? Using the literature and weblogs to link community perspectives and health technology assessment. Health Expectations, 11, 189-200.

STREINER, D.L. (2008) Health Measurement Scales: A practical guide to their development and use. (4 edition ed.).Oxford ; New York: OUP Oxford.

SUN, X., BRIEL, M., BUSSE, J., AKL, E., YOU, J., MEJZA, F., BALA, M., DIAZ-GRANADOS, N., BASSLER, D., MERTZ, D., SRINATHAN, S., VANDVIK, P., MALAGA, G., ALSHURAFA, M., DAHM, P., ALONSO-COELLO, P., HEELS-ANSDELL, D., BHATNAGAR, N., JOHNSTON, B., WANG, L., WALTER, S., ALTMAN, D., GUYATT, G. (2009) Subgroup Analysis of Trials Is Rarely Easy (SATIRE): a study protocol for a systematic review to characterize the analysis, reporting, and claim of subgroup effects in randomized trials. Trials, 10, 101.

SUN, X., BRIEL, M., WALTER, S.D., GUYATT, G.H. (2010) Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. BMJ, 340.

SZOLNOKI, G., HOFFMANN, D. (2013) Online, face-to-face and telephone surveys—Comparing different sampling methods in wine consumer research. Wine Economics and Policy, 2, 57-66.

TERWEE, C.B., BOT, S.D.M., DE BOER, M.R., VAN DER WINDT, D.A.W.M., KNOL, D.L., DEKKER, J., BOUTER, L.M., DE VET, H.C.W. (2007) Quality criteria were proposed for measurement properties of health status questionnaires. Journal of Clinical Epidemiology, 60, 34-42.

THOMPSON, S.G., HIGGINS, J.P. (2002) How should meta-regression analyses be undertaken and interpreted? Stat Med, 21, 1559-1573.

TONG, A., SAINSBURY, P., CRAIG, J. (2007) Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. International Journal for Quality in Health Care, 19, 349-357.

TORRANCE, G.W., FURLONG, W., FEENY, D., BOYLE, M. (1995) Multi-attribute preference functions. Health Utilities Index. PharmacoEconomics, 7, 503-520.

TROCHIM, W., KANE, M. (2005) Concept mapping: an introduction to structured conceptualization in health care. International Journal for Quality in Health Care, 17, 187-191.

TROCHIM, W.M.K. (1989) An introduction to concept mapping for planning and evaluation. Evaluation and Program Planning, 12, 1-16.

VIECHTBAUER, W. (2008) Analysis of Moderator Effects in MetaAnalysis. In: OSBORNE, J. (ed.). Best practices in quantitative methods. Thousand oaks, CA: SAGE Publications, Inc.

VOORHIS, C.R.W.V., MORGAN, B.L. (2007) Understanding Power and Rules of Thumb for Determining Sample Sizes. Tutorials in Quantitative Methods for Psychology.

WAKKER, P., STIGGELBOUT, A. (1995) Explaining distortions in utility elicitation through the rank-dependent model for risky choices. Medical Decision Making, 15, 180-186.

WEBB, C., KEVERN, J. (2001) Focus groups as a research method: a critique of some aspects of their use in nursing research. Journal of Advanced Nursing, 33, 798-805.

WEERNINK, M.G.M., JANUS, S.I.M., VAN TIL, J.A., RAISCH, D.W., VAN MANEN, J.G., IJZERMAN, M.J. (2014) A Systematic Review to Identify the Use of Preference Elicitation Methods in Healthcare Decision Making. Pharmaceutical Medicine, 28, 175-185.

WHITTY, J.A., BURTON, P., KENDALL, E., RATCLIFFE, J., WILSON, A., LITTLEJOHNS, P., SCUFFHAM, P.A. (2014) Harnessing the potential to quantify public preferences for healthcare priorities through citizens' juries. International Journal of Health Policy and Management, 3, 57-62.

WORLD HEALTH ORGANIZATION (2012) WHO | How to Conduct a Discrete Choice Experiment for Health Workforce Recruitment and Retention in Remote and Rural Areas: A User Guide with Case Studies [Online].

WRIGHT, K.B. (2005) Researching Internet-Based Populations: Advantages and Disadvantages of Online Survey Research, Online Questionnaire Authoring Software Packages, and Web Survey Services. Journal of Computer-Mediated Communication, 10, 00-00.

WU, A.D., ZUMBO, B.D. (2008) Understanding and using mediators and moderators. Social Indicators Research, 87, 367-392.

YUSUF, S., WITTES, J., PROBSTFIELD, J., TYROLER, H.A. (1991) Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. JAMA, 266, 93-98.

1 Integrated health technology assessment for evaluating complex technologies (INTEGRATE-HTA): An introduction to the guidances

2 Guidance on the integrated assessment of complex health technologies – The INTEGRATE-HTA Model

3 Guidance for assessing effectiveness, economic aspects, ethical aspects, socio-cultural aspects and legal aspects in complex technologies

5 Guidance for the Assessment of Context and Implementation in Health Technology Assessments (HTA) and Systematic Reviews of Complex Interventions: The Context and Implementation of Complex Interventions (CICI) Framework

6 Guidance on the use of logic models in health technology assessments of complex interventions

7 Guidance on choosing qualitative evidence synthesis methods for use in health technology assessments of complex intervention

8 Integrated assessment of home based palliative care with and without reinforced caregiver support: A demonstration of INTEGRATE-HTA methodological guidances – Executive Summary

INTEGRATE-HTA

This project is co-funded by the European Union under the Seventh Framework Programme (Grant Agreement No. 306141)