A systematic review of the literature on the development of condition-specific preference-based measures of health

Authors: Elizabeth Goodwin PhD[1], Colin Green PhD[1,2]

[1]Health Economics Group, University of Exeter Medical School, University of Exeter, Exeter, UK

[2]Collaboration for Leadership in Applied Health Research and Care South West Peninsula, University of Exeter Medical School, University of Exeter, Exeter, UK

Contact/Correspondence: Elizabeth Goodwin

Email: e.goodwin@exeter.ac.uk

Telephone: +44 (0) 1392 72 6073

Mailing address: Health Economics Group, University of Exeter Medical School, Room 1.06, South Cloisters, St Luke's Campus, Exeter EX1 2LU

Running title: Developing condition-specific preference-based measures of health

Abstract

Background. Health state utility values (HSUVs) are required to calculate quality-adjusted life-years (QALYs). They are frequently derived from generic preference-based measures of health. However, such generic measures may not capture health attributes of relevance to specific conditions. In such cases, a condition-specific preference-based measure (CSPBM) may be more appropriate.

Objective. This systematic review aimed to identify all published accounts of developing CSPBMs, to describe and appraise the methods used.

Method. A systematic search (of Embase, Medline, PsycINFO, Web of Science, the Cochrane Library, CINAHL, EconLit, ASSIA and the Health Management Information Consortium database) was undertaken to identify published accounts of CSPBM development up to July 2015. Studies were reviewed to investigate the methods used to design classification systems, estimate HSUVs, and validate the measures.

Results. Eighty-six publications were identified, describing 51 CSPBMs. Around two-thirds of these were QALY measures; the remainder were designed for clinical decision-making only. Classification systems for 33 CSPBMs were derived from existing instruments; 18 were developed *de novo*. HSUVs for 34 instruments were estimated using a 'composite' approach, involving statistical modelling; the remainder used a 'decomposed' approach based on multi-attribute utility theory. Half of the papers that described the estimation of HSUVs did not report validating their measures.

Conclusion. Various methods have been used at all stages of CSPBM development. The choice between developing a classification system *de novo* or from an existing instrument may depend on the availability of a suitable existing measure, while the choice between a decomposed or composite approach appears to be determined primarily by the purpose for which the instrument is designed. The validation of CSPBMs remains an area for further development.

Key Points for Decision Makers

There are two main approaches to developing a classification system (derivation from an existing instrument or de novo) and two main approaches to estimating health state utility values (using statistical regression or multi-attribute utility theory) for a condition-specific preference based measure.

There is more evidence available to support the validity of classification systems that were derived from existing measures of health-related quality of life than there is for classification systems developed *de novo*.

The choice of methods for the estimation of health state utility values are primarily determined by the purpose for which the instrument is being developed and the context in which it is to be used.

1. Introduction

The quality-adjusted life-year (QALY) is a measure of health-related quality of life (HRQoL) that is frequently used for evaluating the cost effectiveness of healthcare interventions. QALYs are calculated by weighting each year of life according to its quality, on a numerical scale anchored at 1 (equivalent to full health) and zero (equivalent to being dead), thereby combining length and quality of life into a single measure. These standard "scaling anchors" enable comparisons to be drawn between different health conditions [1]. QALY weights are frequently sourced from preference-based measures (PBMs) of health-related quality of life. PBMs provide a standardised health state classification system and a tariff of quality weights for all health states described by the classification system. Generic PBMs, such as the EuroQol EQ-5D [2], Short-Form 6D [3] or Health Utilities Index [4], are commonly used. In addition to enabling the cost-effectiveness of interventions to be compared across a wide range of conditions, using the common metric of the QALY, PBMs can also be used to inform decision-making at an individual patient-level or between alternative treatments for the same condition. In the latter case, the PBM need not generate QALY weights on the scale from dead (zero) to full health (1.00). Generic PBMs are designed to be applicable for all health conditions; however their broad focus has led to some debate around the extent to which they capture aspects of HRQoL of particular relevance to specific conditions. An alternative approach is to use a PBM with a health state classification system that is specific to a particular condition. By focusing on the aspects of health that are most relevant to the condition of interest, these condition-specific PBMs (CSPBMs) potentially offer greater sensitivity and responsiveness [1].

Here we present a systematic review of the literature describing the development of CSPBMs. The review is structured around the three stages involved in developing a PBM [1].

The first stage is to construct a classification system consisting of a set of dimensions, each of which represents a health attribute. Each dimension has a number of ordinal levels. Health states are constructed by selecting one level from each dimension [5]. The classification system must be sufficiently concise to be amenable to valuation, typically containing no more than nine dimensions [6]. There are two approaches to developing a classification system: constructing a new instrument (*de novo*) or deriving one from an existing HRQoL measure [1]. These existing measures often have large numbers of items with several response levels. In order to produce a classification system that is suitable for valuation, the most appropriate single item is selected to represent each dimension, using statistical analysis of a dataset that contains the original measure [5].

Secondly, in order to use the PBM for the calculation of QALYs, a quality weight, or health state utility value (HSUV), must be assigned to all the health states described by the classification system. Typically, values are obtained directly for a sample of the health states via a valuation survey, in which preferences between different health states are elicited either from a sample of people with the condition or from a sample of the general population. Preference elicitation techniques include the standard gamble (SG), time trade-off (TTO), visual analogue scales (VAS) and discrete choice experiments (DCEs). One of two modelling methods is then used to estimate HSUVs for all health states described by the classification system [6]. The first of these is described as the 'composite' approach because valuations are obtained simultaneously for dimensions and dimension-levels. Regression models are estimated, in which the dependent variable is the observed HSUV

and the independent variables are binary dummy variables representing each level of each dimension. Various criteria can be used to compare the performance of alternative models, enabling a preferred model to be selected. The second method is described as 'decomposed' because dimensions and dimension-levels are valued separately and the results are combined in the modelling phase [5]. In the valuation survey, respondents consider each dimension in isolation and scale the levels of the dimension relative to one another and, in a separate exercise, consider the relative importance, or weighting, of each dimension. HSUVs are estimated by solving a series of equations based on multi-attribute utility theory, using the results of the valuation survey.

Finally, the new PBM should be validated. Brazier et al [1] recommend assessing the convergent validity between the CSPBM and generic PBMs, and comparing them in terms of their discriminative validity and responsiveness. This enables an estimation of the impact on the results of cost effectiveness analyses of using the CSPBM rather than a generic alternative and indicates whether the CSPBM is better able to capture differences and changes in HRQoL and condition severity. Where the descriptive system has been derived from an existing measure, they suggest that the magnitude of the information lost in the process of reducing the number of items should be estimated, by assessing the convergent validity between the CSPBM and its parent measure and by comparing them in terms of their discriminative validity, responsiveness and convergent validity with other relevant measures.

Four previous systematic literature reviews have been undertaken to address research questions that relate to or include the development of CSPBMs. Mortimer and Segal [7] aimed to compare four methods for estimating QALY weights from non-PBMs, including derivation of a PBM from an existing measure, and to determine what effect the choice of method may have on the values generated. Their review included generic and condition-specific measures. Petrillo and Cairns [8] aimed to identify best practice in mapping from condition-specific non-PBMs to generic PBMs and in developing CSPBMs from existing measures. Lin et al [9] compared the content of CSPBMs and mapping studies with the EQ-5D in order to identify health attributes for which additional, 'bolt-on' dimensions to the EQ-5D may be appropriate. These three reviews did not focus solely on the development of CSPBMs, therefore their search terms were not specifically designed for this purpose, and discussion of the CSPBMs and the methods employed in their development were necessarily brief or absent. None of these reviews included studies in which the classification system was developed *de novo*. Another systematic review, undertaken by Brazier at al [1], focused solely on the methods for the development of CSPBMs, both *de novo* and from existing instruments. Their literature search identified 26 papers describing the development of 22 instruments up to December 2010. A number of other CSPBMs have been developed since this review. An informal review of the titles, abstracts and keywords of several papers describing the development of CSPBMs suggested that a search strategy incorporating a wider range of search terms could potentially identify additional papers.

Here we present a systematic review based on a more recent and comprehensive search strategy, with a clear focus on identifying papers that describe the development of a CSPBM. The aim of this review is to identify and appraise the methods used to develop these measures.

2. Methods

A search strategy was developed, based on four groups of terms:

1. to identify papers describing the development of instruments
2. to limit the search to studies in which HSUVs were elicited to estimate a PBM (excluding those in which health states were valued for other purposes)
3. to identify papers discussing the measurement of HRQoL
4. to identify preference based (rather than non-preference based) instruments.

The following databases were searched: Embase, Medline, PsycINFO, Web of Science, the Cochrane Library, CINAHL, EconLit, ASSIA and the Health Management Information Consortium (HMIC) database. In addition, citation searching, a Google search and searches of the Discussion Papers published on the Sheffield School of Health and Related Research (ScHARR) Health Economics and Decision Science website were undertaken. An example search strategy and full details of the databases searched are presented as supplementary material.

The review includes papers that:

- relate to an instrument that is designed for a single named condition or a related group of conditions (eg multiple sclerosis or neurological conditions, glaucoma or visual impairment);
- and relate to an instrument that provides a classification system, based on HRQoL, functioning or symptoms, which is capable of categorising all patients with the condition;
- and relate to an instrument for which an algorithm has been developed to enable preference weights to be calculated for all health states defined by the classification system, or where the intention to develop such an algorithm has explicitly been stated;
- and provide an account of the development or validation of such an instrument, whether all or part of the development process, including revaluations of existing instruments using a different population;
- or review various CSPBMs.

The review excludes papers that:

- use demographic or other factors, either alone or in combination with items drawn from HRQoL questionnaires, to predict utility;
- or are not written in English;
- or are commentaries, editorials, letters, conference abstracts or dissertations;
- or provide an account of health state valuation if this does not form part of the development of a CSPBM;
- or concern the development of a bolt-on dimension to an existing generic measure.

The original search was undertaken in July 2011 and was repeated in July 2013 and July 2015. In total, the searches produced 11,706 results (including some duplicates due to slight overlaps between the dates covered by the searches). Titles and abstracts were assessed independently by two reviewers in order to exclude any papers that were not relevant to the review. Through cross-referencing and discussion, both reviewers agreed on a shortlist of potential studies for inclusion. Conference or dissertations abstracts or project records for which the full text was not available were excluded (the reviewers included full-

text manuscripts corresponding to two conference abstracts, in the form of two papers that had been accepted for publication). The full text of the remaining papers was obtained for assessment against the inclusion and exclusion criteria. In total, across the three literature searches, this resulted in the identification of 86 papers, covering the development of 51 instruments (Figure 1). Assessment of the papers against the inclusion and exclusion criteria was undertaken independently by both authors and no disagreements arose.

Data were extracted using a data extraction form structured around the process for developing a PBM (please see supplementary material).

Fig. 1 Literature search results

3. Results

3.1 Overview of the literature

The literature search yielded 86 publications that met the inclusion criteria for the review, including the four previous reviews outlined above (see Figure 1). The remainder described the development of 51 different CSPBMs, including two classification systems that were yet to be valued. Some authors reported the entire process of developing a measure in a single paper, while others reported the development of the classification system and the estimation of the tariff separately. Table 1 links the papers that discuss the development of each measure, and summarises some of the key features of the CSPBMs, as discussed below.

As outlined above, PBMs can be used in cost effectiveness analyses to inform resource allocation across the healthcare system, by generating the HSUVs required for the calculation of QALYs, and can also be used to inform decision analysis at an individual or clinical level. Only around two-thirds of the identified studies aimed to produce an instrument capable of generating QALY weights. The remainder produced an instrument purely for individual or clinical-level decision-making, and stated that their measures were not designed to generate QALY. These studies aimed to identify the factors that influence patients' experiences of living with disease and that inform patients' and clinicians' decisions with respect to treatment alternatives, to quantify the relative importance of these factors, and to produce instruments capable of informing resource allocation within the condition of interest from a patient perspective.

Common areas for CSPBM development were cancer (eight instruments), urinary, gynaecological and sexual health issues (eight), respiratory conditions (seven), neurological conditions (seven), mental health or cognitive deficits (five), vision (three) and oral health (two). A third of the classification systems focused solely on specific symptoms without considering wider impacts on HRQoL. The majority (62.7%) had between four and six dimensions and the total number of health states ranged from nine to several million. Nearly half of the studies were UK-based and around one fifth were undertaken in the USA. Three different sets of scaling anchors were used: perfect health and being dead, the best health state described by the classification system and being dead, and the best and worst health states described by the classification system. Only two studies [11; 42] produced tariffs that included negative values.

The classification systems for 33 of the CSPBMs were derived from existing non-PBMs; 18 were developed *de novo*.

Table 1

3.2 Methods for the development of classification systems

3.2.1 Derivation from an existing instrument

Table 2 presents the methods employed by studies that derived a classification system from an existing measure. The main reasons provided for selecting a particular instrument as the basis for a CSPBM were wide usage in clinical and research settings, psychometric properties, suitability for a range of condition subtypes or severity levels, and coverage of dimensions considered important by patients. One study [22] based their choice on a systematic comparison of available HRQoL measures.

Twelve of the 33 classification systems that were derived from an existing instrument adopted the 'health state classification approach' developed by Brazier et al [1], which employs statistical techniques to determine the dimensional structure of an instrument, followed by a combination of traditional psychometric analysis with one of the "new" psychometric techniques (Rasch analysis or item response theory) to select one item to represent each dimension. Rasch analysis is based on the idea that there is a latent scale for the construct being measured (in this case, HRQoL), on which both respondents and item response levels are located. The probability that a respondent will give a particular response to a particular question is calculated as a logistic function of the distance between the position of the respondent and the position of that response level on the latent (logit) scale. Rasch models can be used to analyse the psychometric properties of existing scales and develop new scales [92; 93] and offer a number of advantages compared to traditional psychometric techniques[1] [94].

In the health state classification approach, item selection takes place in two phases. In the item elimination phase, Rasch analysis is typically employed to identify and adjust items where respondents struggle to distinguish between response levels (disordered thresholds) and items that behave differently between different subgroups of respondents (differential item functioning). Items are removed if they exhibit poor fit to the Rasch model, using item chi-squared statistics, item fit residuals and measures of model goodness of fit (the item-trait interaction chi-squared statistic, mean fit residuals and the person separation index). Only the remaining, unadjusted items are considered during the item selection phase. Item selection typically uses both new and traditional psychometric methods to compare items. Rasch analysis is used to assess the extent to which item-levels cover the full range of condition severity (ie their spread across the latent scale). All studies also considered how well items fitted the Rasch model. The psychometric criteria employed to compare items included acceptability (missing data), item difficulty (item response distributions), the extent to which the item measures the intended construct (internal consistency) and sensitivity to change (responsiveness). Several (n = 5) of the studies that took this approach employed expert opinion to enhance content validity and clinical relevance and to avoid redundancy. Brazier et al [1] also recommend exploring whether the number of item response levels can be reduced, using Rasch analysis and response frequencies; only one of these studies [71] did not do so.

[1] Unlike traditional psychometric approaches, Rasch models produce interval scales, which are based on an estimate of the true score rather than relying on observed scores. Item parameters are estimated independently of the sample used for scale construction; similarly, person values are estimated independently of the items used. This avoids two limitations of traditional psychometric methods, in which results for scales are sample dependent and *vice versa*. Rasch models are more tolerant of missing data and do not require imputation of missing values [94].

The remaining studies adopted a variety of methods, primarily psychometric criteria and expert opinion, to determine dimensions and select items. In five studies, individual items were not selected to represent each dimension; instead, item responses were summed to produce overall dimension scores or new response levels were assigned to each dimension.

Table 2

3.2.2 Developing a classification system *de novo*

The methods used for developing a classification system *de novo* are summarised in Table 3. All but four studies used qualitative work with patients at some stage of the process. Six studies took an inductive approach to determining the domains to be included in their classification systems, constructing dimensions by analysing data from interviews with patients and clinicians [10; 21; 27; 75; 76; 77]. Others used a combination of statistical, qualitative and literature-based techniques. A wide variety of methods were employed to construct and select items to represent each dimension, although relatively few studies reported using psychometric techniques [41; 48; 63; 76; 77]. Rather than developing items, eight studies constructed intra-domain statements to describe the severity levels for each dimension from analysis of qualitative interviews with patients [10; 14; 21; 24; 27; 75] or expert opinion [16; 24; 74].

Table 3

3.2.3 Validation

Thirty-six of the 51 classification systems were not validated. Six of the studies that derived their CSPBM from an existing measure repeated the process on a separate dataset and compared the results, as recommended by Brazier et al [1]. Other studies undertook a separate survey to test the psychometric properties of the classification system (including test-retest reliability; convergent validity with PBMs, HRQoL measures and clinical measures; discriminative validity; responsiveness; and unidimensionality). Other methods involved further analysis of the dataset used to develop the classification system, using Rasch tests of unidimensionality and item redundancy [45; 57] or psychometric criteria [17], or constructing a temporary scoring index and assessing the performance of the regression model and the psychometric properties of the values generated [41].

One vision-specific classification system [48] was validated independently by two subsequent publications [50; 52], which described the translation of the instrument into other languages.

3.3 Preference elicitation techniques

Table 4 summarises the methods used to select health states and to elicit preferences. Thirty-two studies elicited preferences from a sample of the general population, 21 elicited preferences from people with the condition and one surveyed clinicians. With the exception of one paper that did not explicitly mention QALYs [17], all studies that obtained valuations from the general public aimed to produce an instrument capable of producing QALY weights.

Eighteen of the 21 studies that obtained valuations from patients developed non-QALY measures.

3.3.1 Composite approach

A composite approach was adopted to estimate the tariffs for 34 classification systems. Two additional studies estimated a new tariff for an existing classification system [31; 71] using a different population [32; 72]. Three papers [16; 59; 64] undertook more than one valuation of the same classification system, aiming to compare the tariffs obtained via different methods.

Five classification systems were small enough for preferences to be elicited directly for all health states from all survey respondents. Among the others, the dominant methodology was to obtain valuations for health states prospectively sampled from the classification system. Samples of health states were selected from most classification systems using a statistically efficient design, such as an orthogonal array [95]. The number of conceptual dimensions represented by three classification systems [23; 45; 87] exceeded the number of statistically distinct dimensions, increasing the likelihood of interdependencies between items, and hence the possibility of a statistical design including implausible combinations of dimension levels. Therefore, these studies employed a 'Rasch vignette approach', which uses Rasch analysis to identify health states typically reported by people experiencing different levels of condition severity. This ensures that the sample contains health states that are likely to be experienced by people with the condition. Two studies [23; 35] involved people with the condition described by the health states to ensure that the selected states were plausible. All but four of these studies aimed to produce a QALY measure, several of which stated that they selected their valuation technique in order to enhance comparability with the EQ-5D or to comply with guidelines [96] issued by the UK National Institute for Health and Care Excellence (NICE) [23; 30; 42; 46; 47; 57; 61; 82; 84; 87; 89]. Three studies selected DCE methods to produce non-QALY instruments to inform decision analysis, stating that DCEs mirror the ways in which patients choose between treatment strategies [25; 28] or are more appropriate than SG or TTO approaches for chronic conditions where death is not a likely outcome [14]. The latter also provided a rationale for using DCEs to elicit preferences for a QALY measure [24].

An alternative methodology, used in three studies that aimed to produce non-QALY instruments for decision analysis, was to ask people with the condition to complete the classification system, thereby reporting the health state they were currently experiencing, and then to complete a preference elicitation exercise for their own current health [18; 44; 62]. The values provided for the respondents' own health states were used in regression analysis to obtain HSUVs for all health states described by the classification system.

3.3.2 Decomposed approach

Tariffs for 15 classification systems were estimated using a decomposed approach. In addition, two studies [66; 69] estimated new tariffs for two existing classification systems [65; 68] in different populations. The sample of health states required to estimate a decomposed model is largely determined by multi-attribute utility theory. Values for dimension-levels are

obtained using ‘single attribute states’: respondents are asked to consider each dimension in isolation and to scale the levels of the dimension relative to one another. The relative weighting of each dimension is assessed using its ‘corner state’, in which the relevant dimension is at its worst level and all others are at their best levels. Finally, respondents value a number of ‘multi-attribute states’ to estimate the power function required to convert VAS values into utility scores [97]. The most common technique was to value all single attribute states, all corner states and a small number of multi-attribute states using VAS and to value the corner states and multi-attribute states using a choice-based method. Two studies omitted a corner state that was considered implausible [15; 63]. Four studies [15; 49; 54; 63] included fewer than the three multi-attribute states required for the multiplicative multi-attribute utility function that they used to predict HSUVs [5]. Six studies provided a rationale for their choice of methods: minimising respondent burden [39; 49], consistency with economic theory and theoretical suitability [38], or replicating the methods of previous studies [17; 66; 69; 81].

An alternative technique did not involve the valuation of health states. Instead, respondents were asked to weight dimensions relative to each other, then to score the levels of each dimension independently of the others, using resource allocation tasks and VAS [10; 21; 27; 75].

Table 4

3.4 Modelling techniques

Table 5 presents details of 32 composite models selected to produce tariffs for 28 classification systems, plus five models reported by one study that did not select a preferred specification [64]. Of the remaining classification systems that were valued using a composite approach, four did not require models because all health states were valued directly [33; 59; 76; 79]. One study that obtained direct valuations for all health states estimated models to predict values for members of the population outside the study sample [16]. One study did not provide details of their model [78]. The majority of models (n = 26) used either an ordinary least squares (OLS) or random effects (RE) specification. The five models based on the results of DCEs were estimated using RE probit or conditional logistic regression. A greater likelihood of interactions between items was anticipated by the developers of three classification systems in which the dimensions were highly correlated[46; 35; 87]. These studies used OLS regression models to examine the relationship between mean observed HSUVs and values on the Rasch logit scale corresponding to the health states that were directly valued.

In 14 studies, some model coefficients were found to be inconsistent with the expected direction of preferences. In three of these studies, the final models included inconsistent coefficients [12; 64; 71]; the remainder merged inconsistent coefficients to produce a consistent model. Twelve studies tested the inclusion of preference interactions. Eight studies employed dummy variables to capture the effect of any dimension being at its highest or lowest level [12; 13], any dimension being at its most severe level [23; 55; 61; 72], any dimension being at least level 3 or at least level 4 [72] or two or more dimensions being at level 4 or 5 [84; 89]. Two studies fitted first-order interaction terms [44; 72] and three did not describe their interaction terms [16; 57; 71]. Only two studies included an interaction term in their preferred model [61; 72].

There are two reasons for evaluating the performance of models: to select the preferred model and to assess the quality of the preferred model. The following summary concerns preferred models only. Given that these models aim to predict values for health states, an important test is the difference between observed and predicted values for the health states that were directly valued. The majority of studies assessed this using the mean absolute error (n = 20); some also reported the number of health states with prediction errors > 0.1 or > 0.05 (n = 11). Most studies reported the number of inconsistent coefficients (n = 27) and the number of significant coefficients (n = 25). These tests can be calculated for any model specification, making them useful for comparisons between alternative models [5]. A number of other statistics were reported, including R-squared or R-squared type statistics, root mean squared error, the Akaike information criterion, the Bayesian information criterion, the Ljung-Box statistic and the Jarque-Bera test of normality of prediction errors. Two papers reported no model performance statistics [25; 28]. Only two models that were selected to produce a tariff included inconsistent coefficients, although 11 of the preferred models included levels that had been merged to remove inconsistencies. In 19 of the 25 models for which this was reported, at least 80% of coefficients were significant ($p < 0.05$). MAEs ranged from 0.008 to 0.065. With the exception of two models [13; 60], the percentage of health states with prediction errors greater than 0.05 ranged from 31% to 44%, while the percentage of errors greater than 0.1 varied from 2% to 16%.

In addition to the models listed in Table 5, two studies produced alternative tariffs for existing instruments. Kharroubi et al [86] estimated new models for two existing measures [84; 89] using non-parametric Bayesian methods in order to address certain limitations of the standard approach, in terms of the size and pattern of prediction errors and the extent to which the effects of covariates are captured. Hernandez-Alava et al [90] rescaled six PBMs, including the AQL-5D [89], onto a common scale, using ranking data.

Table 5

The 13 studies that directly valued individual health states to inform a decomposed model transformed the values obtained from the VAS into SG or TTO utilities prior to fitting the MAU function, using a power curve or function. The multiplicative functional form proved the dominant model due to its ability to allow for some preference interactions between dimensions; one study found that additive models were adequate to estimate tariffs for two of their three population subsamples, but required a multiplicative model for their third subsample and their combined tariff [66]. Only five of these studies reported performance statistics to illustrate the predictive ability of their models [11; 17; 39; 69; 81].

All four studies in which dimensions and dimension-levels were valued separately used an additive MAU functional form, which allows for no preference interactions between dimensions. None of these studies assessed the predictive ability of their models.

3.5 Validation methods

Methods for the validation of CSPBMs are illustrated in Table 6 and discussed here in relation to recommendations made by Brazier et al [1]. Twenty-one studies assessed the convergent validity of their measures: four compared the CSPBM with the measure from

which it had been derived; eight compared it with other PBMs. Eighteen studies undertook an evaluation of discriminative validity, although only three compared this with another measure: two with the EQ-5D and one with the EQ-5D, SF-6D and the parent measure. Responsiveness was considered by eight studies, of which only two compared the responsiveness of the CSPBM with another measure. Around half of the papers that described the estimation of a tariff for one or more CSPBMs did not report any validation of their measure(s).

Ten CSPBMs were validated retrospectively in subsequent publications [19; 20; 36; 40; 50-53; 67; 73; 85; 91]. These adopted a range of methods, including comparisons with algorithms that map from the parent measure of the CSPBM to EQ-5D values. Brazier et al [1] conducted retrospective validations of four instruments [46; 71; 84; 89] following their own recommended procedure.

Table 6

4. Discussion

4.1 Methods for developing a classification system

Two methodological frameworks for the development of classification systems have been proposed: one for the derivation of classification systems from existing measures [1] and one for the development of classification systems *de novo* [77].

4.1.1 Derivation from existing instruments

Brazier and colleagues have developed the 'health-state classification approach', which provides a guide to the methods for deriving a classification system from an existing HRQoL measure, while minimising the loss of descriptive information and enabling HSUVs to be estimated from responses to the original measure [1]. The majority of papers describing the derivation of a CSPBM from an existing measure that had been published since 2009 used this approach, although relatively few followed its recommendations for validation. This suggests that the health state classification approach is emerging as the dominant method for deriving a classification system from an existing measure. Differences were apparent between these studies; however, this approach is intended as a guide that should be adapted to each particular application, rather than a rigid methodological process.

Across all studies in which individual items were selected from the original measure, all but one [79] used traditional psychometric criteria, Rasch analysis or factor analysis to do so. Given the strong emphasis that the health-state classification approach places on new psychometric techniques, it is unsurprising that the use of Rasch analysis has increased substantially in recent years. Expert opinion retains a significant role at various points in the process, illustrating that however robust the statistical methods may be, it remains important to check that their results make clinical sense.

Among those studies that used an existing instrument as the basis for a CSPBM, only one [22] reported undertaking a systematic comparison between candidate measures to provide a robust justification for selection of a particular instrument, suggesting that this could be a fruitful area for future research.

Deriving a classification system from an existing measure may increase the scope for the PBM to be adopted in cost-effectiveness studies. The PBM can be applied to existing datasets that contain the original measure, enabling retrospective economic evaluations to be undertaken. Furthermore, selecting an instrument that is well accepted by relevant clinical and research communities may enhance the acceptability of the PBM. However, HRQoL measures were not intended for this purpose and not all will provide a suitable basis for a PBM [5]. Where no suitable measure is available, the classification system will have to be developed *de novo* [1].

4.1.2 Development de novo

Stevens and Palfreyman [77] have recommended a best practice method for developing a classification system *de novo*, in which dimensions and items are derived inductively from analysis of qualitative interviews with patients or the public. One third of the *de novo* studies

employed this approach [10; 21; 27; 75; 76; 77], although the majority of studies involved people with the relevant health condition at some stage.

Notwithstanding their recommendations for a 'bottom-up', qualitative approach, it is notable that Stevens and Palfreyman made extensive use of new and traditional psychometric methods when selecting items for their classification system [78]. Given that classification systems are intended for use as questionnaires to be completed by people with the relevant condition, one may expect developers to use similar methods to those used in the construction of non-preference based HRQoL measures. This typically involves analysis of a dataset containing a draft version of the classification system, using either new or traditional psychometric techniques, in combination with insights derived from patients, clinical experts or the relevant literature [1]. However, only six of the 18 *de novo* studies reported that they had evaluated the psychometric properties of items in the development [41; 48; 63; 76; 77] or validation [41; 63; 74] of their classification systems. This limits the range of evidence available to support the validity of many classification systems that were developed *de novo*, particularly compared to those derived from existing instruments, the majority of which employed psychometric techniques to select items.

4.2 Methods for generating a tariff

Context appears to be an important factor in the selection of methods for estimating HSUVs for a CSPBM. As we have seen, most studies that designed non-QALY instruments for decision analysis in condition-specific settings elicited preferences from patients and used a decomposed approach, which explicitly weights dimensions relative to one another, arguably providing a more appropriate method for identifying the factors of greatest importance to patients. Conversely, the requirements of national decision-making bodies were a key determinant of the methods adopted by studies that aimed to produce a QALY measure; for example, the UK studies tended to follow NICE guidance, which stipulates that tariffs should be statistically modelled from societal preferences, elicited using the Measurement and Valuation of Health variant of the TTO technique [96]. The importance of context is supported by the predominance of QALY measures developed in settings which rely on QALY-based frameworks to inform resource allocation for the healthcare system overall, in contrast to the predominance of non-QALY based decision analysis tools in the US setting, where patients pay for medical treatment either directly or via insurance, leading to a greater focus on patients' perceptions of health status and treatment outcomes.

Dolan [98] has argued that methodological considerations may determine the choice between decomposed and composite approaches. The extent to which this is reflected in the approaches adopted by the included studies is mixed. Firstly, as the size of the classification system increases, so does the size of the sample of health states required to estimate a composite model, whereas the number of states required for a decomposed model remains constant. Dolan suggests, therefore, that the decomposed approach may be better suited to larger classification systems. This appears to be reflected in the included studies: six of the 34 classification systems that were valued using a composite approach described more than 10,000 health states, compared to nine of the 15 that were valued using the decomposed approach (Tables 1 and 4). Secondly, he asserts that the decomposed approach may be less appropriate for classification systems that include correlated dimensions, due to the restrictions it places on preference interactions between

dimensions and its requirement to obtain valuations for potentially implausible corner states [97]. The findings of this review do not indicate that the potential for interactions between dimensions influenced the choice of approach. Most of those that followed a decomposed strategy allowed for limited preference interactions by fitting multiplicative models to their data, whereas only four that followed a composite strategy specifically avoided selecting implausible states and only two found that an interaction term improved the predictive ability of their model (indeed, most did not report testing for preference interactions). A further implication of dimension orthogonality is that dimensions defined according to specific symptoms are more likely to be structurally independent than dimensions that reflect broader impacts on HRQoL. This indicates that a composite strategy may be better suited to classification systems that describe the impact of the condition on HRQoL. There is some evidence that the included studies followed this pattern. As Table 1 shows, 17 classification systems did not include dimensions that describe the impact of the condition on aspects of people's HRQoL, focusing instead on describing symptoms of the condition. Twenty-three of the 34 classification systems that were valued using a composite approach included HRQoL attributes, compared to six of the 15 that were valued using the decomposed approach.

It has been suggested that the decomposed approach provides a stronger theoretical foundation due to its basis in MAU theory. Dolan [98] asserts that this theoretical advantage is largely irrelevant if it does not enhance the ability to predict HSUVs. One of the included studies [61] estimated both a composite and a decomposed algorithm for the same classification system, concluding that the algorithms performed similarly well in terms of responsiveness and discriminative validity. Following their review of the methods for converting condition-specific measures into PBMs, Petrillo and Cairns [8] found no evidence to prefer either a composite or a decomposed approach.

Among the studies that took a composite approach, model selection was generally based on the difference between observed and predicted values, the proportion of coefficients that were consistent and the proportion of coefficients that were significant. The performance of the preferred models varied considerably. Few of the studies that took a decomposed approach reported performance statistics to illustrate the predictive ability of their models (n = 5).

Young et al [87] and Mavranezouli et al [45; 46] developed novel approaches to valuation and modelling based on Rasch analysis. While most studies adopted statistically efficient designs to select health states for inclusion in the valuation survey, these authors noted that such designs may generate implausible health states when applied to classification systems with interdependent dimensions. This led to the development of the Rasch vignette approach, which provides a method for selecting health states based on the combinations of item-levels that are most likely to be experienced by people with the condition. Rather than using individual dimension-levels as the independent variables in the regression analysis, these studies used corresponding Rasch logit values for health states to predict HSUVs. Two subsequent studies have used similar techniques to select a sample of health states [23] or predict HSUVs [35]. Although there remain some issues to resolve, particularly the ability of the technique to predict values for individual health states rather than groups of states based on total dimension-level scores [23], this is a promising new approach that is an important area for further research.

4.3 Validation methods

Brazier et al [1] recommend that the validity and responsiveness of the CSPBM should be assessed in comparison with generic PBMs and (where appropriate) the parent measure. However, relatively few papers reported a thorough validation of their measure. It is worth noting that a number of validations were reported in subsequent papers, therefore validations of some instruments may be reported in future publications. Where validation was undertaken, this seldom adhered to these best practice methods. Convergent and discriminative validity were reasonably well covered, responsiveness less so, possibly due to a lack of access to longitudinal data.

4.4 Tariffs and values

In order for the results of CSPBMs to be used to calculate QALYs, the valuations for health states described by the CSPBM should lie on a scale from full health (one) to being dead (zero) [1]. All instruments with values anchored against the worst and best possible health states were designed for decision analysis only, not for estimating QALYs. Among those studies that sought to develop a QALY measure, all anchored the state of being dead at zero. Eighteen selected an upper anchor of generic full health, while twelve used an upper anchor of condition-specific full health (the best health state described by the classification system). The definition of 'full health' is a subject of debate in the literature. Condition-specific full health is not necessarily equivalent to perfect health. Respondents to a valuation survey may assume that patients in the best possible health state may have decrements on dimensions that are not included in the classification system, due to the condition itself or due to co-morbidities [99]. Brazier et al [1] assert that comparability is problematic between instruments that do not share a common upper anchor; therefore PBMs should be anchored against death and generic full health rather than the best possible health state.

Only two of the condition-specific tariffs included negative HSUVs [11; 42], despite the fact that several instruments were developed for conditions that cause severe decrements in HRQoL. One hypothesis is that the relatively narrow coverage of HRQoL attributes by condition-specific classification systems compared to the EQ-5D allows valuation survey respondents to assume high levels of function on other HRQoL attributes [1]. However, it is notable that another popular generic PBM, the SF-6D, also has no negative values [3]. Further research is required to understand this phenomenon.

4.5 Conclusion

This paper presents a detailed systematic review of the literature describing the development of CSPBMs. The most relevant previous review included 26 papers [1]. The literature search reported here identified 43 papers published since this earlier review and included 17 additional papers, providing an up-to-date and comprehensive review of the literature in this area.

A wide range of methods have been used at all stages of CSPBM development, within two overarching groups of approaches for construction of the classification system (derivation

from an existing instrument and *de novo*) and two for the estimation of the tariff (composite and decomposed methods). The choice between developing a classification system *de novo* or from an existing instrument may depend largely on the availability of a suitable existing measure. The choice between a decomposed or composite approach appears to be determined primarily by the purpose for which the instrument is designed (QALY generation or individual/ clinical level decision analysis), although considerations regarding the size of the classification system and practical constraints on the proportion of health states that can be valued directly may play a part. More comparative studies are required to provide empirical evidence on the relative merits of these competing approaches. Despite the recent publication of recommended methods for validation [1], uptake of these appears to have been relatively low, and this remains an area for further development.

Compliance with ethical standards

Funding: This study formed part of a PhD studentship funded by The Multiple Sclerosis Society of Great Britain and Northern Ireland (grant reference number **928/10**). This research was supported by the National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care South West Peninsula at the Royal Devon and Exeter NHS Foundation Trust. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

Conflicts of interest: Elizabeth Goodwin and Colin Green have no conflicts of interest.

Authors' contributions: Elizabeth Goodwin drafted and revised the manuscript for content, was involved in conceptualisation and design of the work, analysed and interpreted the data, and is guarantor for overall content. Colin Green drafted and revised the manuscript for content, was involved in conceptualisation and design of the work and analysed and interpreted the data.

Table 1: Outline of condition-specific preference-based measures

First author(s) date	Aim	Condition	HRQoL	Dimensions	States	Country	Scaling anchors
Bellamy 1996 [10]	NQM	Peridontal disease	Yes	4	320	UK	Best state; pits
Beusterien 2005 [11]	QALY	ALS	No	4	750	USA	Best state; dead
Brazier 2005 [12]	QALY	Menopause	No	7	6,075	UK	Full health; dead
Brazier 2008 [13]	QALY	Urinary tract symptoms	Yes	5	1,024	UK	Full health; dead
Burr 2007 [14]	NQM	Glaucoma	Yes	6	4,096	UK	Best state; pits
Chiou 2005 [15]	QALY	Asthma (paediatric)	No	3	12 (10 valid)	USA	Unclear; dead
Cho 2015 [16]	QALY	COPD	No	3	12 (10 valid)	South Korea	Full health; dead
Cuervo 2014 [17]	UC	Urinary	No	5	243	UK	Best state; dead
Dobrez 2007 [18]; validated by Hess 2013 [19]; Pickard 2012 [20]	NQM	Cancer	Yes	4	48	USA	Best state; dead
Goodey 2000 [21]	NQM	Oral	Yes	5	1,024	UK	Best state; pits
Goodwin 2015a [22]; Goodwin 2015b[a] [23]	QALY	Multiple sclerosis	Yes	8	65,536	UK	Full health; dead
Gu 2013 [24]	QALY	Dupuytren's contracture	No	8	6,561	UK	Best state; dead
Hauber 2010 [25]; re-modelled by Mohamed 2010 [26]	NQM	Obesity	Yes	8	6,561	USA	Best state; pits
Hodder 1997 [27]	NQM	Head and neck cancer	Yes	8	390,625	UK	Best state; pits
Johnson 2006 [28]; Osoba 2006 [29]; re-modelled by Mohamed 2010 [26]	NQM	Cancer	Yes	14	268,435,456	USA	Best state; pits
Kerr 2015 [30]	QALY	Aberrant behaviour in FXS	No	5	2,187	UK	Full health; dead
Kind 2005 [31]; revalued by Lamers 2007 [32]; validated by Pickard 2012 [20]	QALY	Lung cancer	Yes	6 "	64 "	UK Netherlands	Full health; dead Full health; dead
Kok 2002 [33]	QALY	Urinary	No	2	9	Netherlands	Best state; dead
Kowalski 2012 [34]; Rentz 2014 [35] validated by Naik 2013 [36]	QALY	Vision	Yes	6	15,625	Multiple[b]	Full health; dead
Krahn 2000 [37]; Ritvo 2005 [38]; Tomlinson 2012 [39]; translated and validated by Avila 2014 [40]	NQM	Prostate cancer	No	10	6,000,000	Canada	Best state; dead
Kuspinar 2014 [41]	QALY	Multiple sclerosis	Yes	5	243	NA	NA
Lloyd 2014 [42]	QALY	Short bowel syndrome	Yes	6	64	UK	Full health; dead
Martin 1999 [43], 1998 [44]	NQM	Cardiovascular	Yes	4	268,880	Australia	Best state; dead

Table 1 continued

First author(s), date	Aim	Condition	HRQoL	Dimensions	States	Country	Scaling anchors
Mavranezouli 2011 [45]; 2013 [46]	QALY	Mental health	Yes	6	729	UK	Full health; dead
McKenna 2008 [47]	QALY	Pulmonary hypertension	Yes	4	36	UK	Full health; dead
Misajon 2005 [48]; Peacock 2008 [49]; validated by Finger 2013a [50], 2013b [51]; Gothwal 2013a [52], 2013b [53]	QALY	Vision	Yes	6	45,360	Not stated	Best state; dead
Montejo 2011 [54]	NQM	Anti-psychotic side effects	No	8	65,536	Spain	Best state; dead
Mulhern 2012a [55]	QALY	Epilepsy	Yes	6	4,096	UK	Full health; dead
Mulhern 2012b [56]; 2013 [57]; Rowen 2012 [58]	QALY	Dementia (proxy measure)	Yes	5	1,024	UK	Full health; dead
		Dementia (self-complete)	No	4	256	UK	Full health; dead
Palmer 2000 [59]	NQM	Parkinson's disease	No	2	10	USA	Full health; dead
Petrillo 2011 [60]; 2010[c] [61]	QALY	COPD	No	5	UC	UK	Full health; dead
Pickard 2009 [62]	NQM	Cancer	Yes	5	144	USA	Best state; dead
Poissant 2003 [63]	NQM	Stroke	Yes	10	59,049	Canada	Best state; pits
Ratcliffe 2009 [64]	QALY	Sexual quality of life	Yes	3	64	UK	Full health; dead
Revicki 1998a [65]; revalued by Flood 2006 [66]; validated by Bime 2012 [67]	NQM	Asthma	No	5	100,000	USA Multiple[d]	Best state; pits Best state; pits
Revicki 1998b [68]; revalued by Lo 2006 [69]	NQM	Rhinitis	No	5	100,000	USA Hong Kong	Best state; pits Best state; pits
Revicki 2011 [70]	NQM	Acute rhinosinusitis	Yes	5	1,024	USA	Best state; pits
Rowen 2011 [71]; revalued by Kularatna 2015 [72] validated by Rowen 2012 [73]	QALY	Cancer	Yes	8	81,920	UK Sri Lanka	Best state; dead
Scholzel-Dorenbos 2012 [74]	UC	Dementia	Yes	5	243	NA	NA
Shaw 1998 [75]	NQM	Menorrhagia	No	6	4,096	UK	Best state; pits
Stevens 2005 [76]	QALY	Atopic dermatitis	Yes	4	16	UK	Full health; dead
Stevens 2012 [77]; Palfreyman 2011[e] [78]	QALY	Venous ulcers	Yes	5	720	UK	Full health; dead
Stolk 2003 [79]	QALY	Erectile function	No	2	25	Netherlands	Best state; dead
Sundaram 2009 [80]; 2010 [81]	QALY	Diabetes	No	5	768	USA	Best state; dead
Versteegh 2012 [82]	QALY	Arthritis	No	NA	1,024	Netherlands	Best state; dead
	QALY	Cancer	Yes	8	65,536	Netherlands	Best state; dead
	QALY	Multiple sclerosis	Yes	NA	65,536	Netherlands	Best state; dead

Table 1 continued

First author(s), date	Aim	Condition	HRQoL	Dimensions	States	Country	Scaling anchors
Young 2009 [83]; Yang 2009 [84] validated by Desroziers 2013 [85] remodelled by Kharroubi 2014 [86]	QALY	Overactive bladder	Yes	5	3,125	UK	Best state; dead
Young 2010 [87]	QALY	Flushing	No	1	2,500	UK	Full health; dead
Young 2011 [88]; Yang 2011 [89]; validated by McTaggart-Cowan 2008 [91] remodelled by Kharroubi 2014 [86] rescaled by Hernandez-Alava, 2013 [90]	QALY	Asthma	Yes	5	3,125	UK	Best state; dead

[a] data provided by authors; [b] Rentz et al conducted four preference elicitation surveys, in Australia, Canada, the UK and the USA; [c] conference poster; [d] Flood et al conducted three preference elicitation surveys, in the UK, France and Italy; [e] newsletter article

ALS = amyotrophic lateral sclerosis; FXS = fragile X syndrome; COPD = chronic obstructive pulmonary disease; QALY = instrument designed to generate QALY weights; NQM = instrument not designed to generate QALY weights; UC = unclear

HRQoL: "No" indicates a classification system comprised of dimensions that describe specific symptoms rather than wider impacts on HRQoL; "Yes" indicates a classification system that includes one or more dimensions describing the impact of the condition on HRQoL.

Table 2: Methods for deriving classification systems from existing instruments

Method of item selection:	Health state classification approach											Other item selection method										No items selected					Unclear	
	Goodwin 2015a [22]	Kowalski 2011 [34]	Mavranezouli 2011 [45]	Mulhern 2012 [55]	Mulhern 2013[a] [57]	Rowen 2011 [[71]	Versteegh 2012[b] [82]	Versteegh 2012[c] [82]	Young 2009 [83]	Young 2010 [87]	Young 2011 [88]	Brazier 2008 [13]	Cuervo 2014 [17]	Dobrez 2007 [18]	Kerr 2015 [30]	Kind 2005 [31]	Lloyd 2014 [42]	McKenna 2008 [47]	Pickard 2009 [62]	Stolk 2003 [79]	Versteegh 2012[d] [82]	Hauber 2010 [25]	Johnson 2006 [28]	Kok 2002 [33]	Martin 1999 [43]	Sundaram 2009 [80]	Brazier 2005 [12]	Petrillo 2011 [60]
Selecting an instrument																												
Wide usage	X	X	X	X	X	X	X	X				X		X	X	X		X		X	X	X			X			X
Psychometric properties	X	X	X	X		X			X	X	X	X	X	X	X				X			X	X			X	X	X
Wide suitability	X		X		X	X			X													X	X			X		
Importance to patients	X												X	X				X							X	X	X	
Other reason							X	X										X			X					X		X
Determining dimensions																												
Factor analysis	X		X	X	X	X	X	X	X	X	X				X	X					X			X		X		
Psychometric criteria		X	X				X	X		X		X			X						X	X						
Rasch analysis		X				X																				X		
Expert opinion			X	X		X						X				X										X		
Determined by item selection														X			X			X								
Original instrument developers																							X				X	
Conceptual framework	X																											
Not reported							X	X					X					X							X			X
Item selection																												
Rasch criteria																												
Disordered thresholds	X		X	X	X	X			X		X		X						X									
Differential item functioning	X	X	X	X	X	X	X	X	X				X		X													
Item $\chi 2$ and/ or fit residuals	X	X	X	X	X	X	X	X	X	X	X																	
Goodness of fit to Rasch model	X	X	X	X	X	X	X	X	X	X	X		X															
Spread across latent space	X		X	X	X	X		X	X		X		X		X			X	X									

Table 2 continued

	Health state classification approach											Other item selection method										No items selected					Unclear	
	Goodwin 2015a [22]	Kowalski 2011 [34]	Mavranezouli 2011 [45]	Mulhern 2012 [55]	Mulhern 2013[a] [57]	Rowen 2011 [[71]	Versteegh 2012[b] [82]	Versteegh 2012[c] [82]	Young 2009 [83]	Young 2010 [87]	Young 2011 [88]	Brazier 2008 [13]	Cuervo 2014 [17]	Dobrez 2007 [18]	Kerr 2015 [30]	Kind 2005 [31]	Lloyd 2014 [42]	McKenna 2008 [47]	Pickard 2009 [62]	Stolk 2003 [79]	Versteegh 2012[d] [82]	Hauber 2010 [25]	Johnson 2006 [28]	Kok 2002 [33]	Martin 1999 [43]	Sundaram 2009 [80]	Brazier 2005 [12]	Petrillo 2011 [60]
Psychometric criteria																												
Factor analysis				X	X										X	X	X											
Missing data	X	X	X	X	X	X	X	X	X	X	X	X		X	X		X		X		X							
Item response distributions	X	X	X	X	X	X	X	X	X	X	X	X		X	X		X	X			X							
Responsiveness			X	X		X			X		X	X																
Internal consistency	X		X			X	X	X	X		X								X		X							
Construct validity												X		X				X	X									
Other criteria																												
Expert opinion		X	X	X	X			X				X	X		X	X		X		X	X							
Item responses summed																								X	X			
Levels created for dimensions																						X	X			X		
Item selection not reported																											X	X
Item level reduction																												
Disordered thresholds	X	X	X				X	X		X				X							X							
Other Rasch analysis					X				X		X			X														
Item response distributions				X					X		X			X	X												X	
Method not reported													X															
Validation																												
Repeat on different dataset	X		X	X		X			X		X																	
Separate survey																									X	X		
Other methods			X		X								X															
Methods not specified																												X

[a] includes two classification systems (DEMQOL and DEMQOL-proxy); [b] arthritis measure; [c] MS measure; [d] cancer measure

Two studies obtained preference weights for an existing measure without alteration [Montejo et al, 2011; Revicki et al, 2011], and two provided no account of how a classification system was derived from the original measure [Beusterien et al, 2005; Ratcliffe et al, 2009].

Table 3: Methods for developing a classification system de novo

	Bellamy 1996 [10]	Burr 2007 [14]	Chiou 2005 [15]	Cho 2015 [16]	Goodey 2000 [21]	Gu 2013 [24]	Hodder 1997 [27]	Krahn 2000 [37]	Kupsinar 2014 [41]	Misajon 2005 [48]	Palmer 2000 [59]	Poissant 2003 [63]	Revicki 1998a [65]	Revicki 1998b [68]	Scholzel-Dorenbos 2012 [74]	Shaw 1998 [75]	Stevens 2005 [76]	Stevens 2012 [77]
Identifying dimensions																		
Qualitative work with patients	X	X			X	X	X	X		X			X	X	X	X		X
Expert opinion			X	X		X		X	X		X		X	X	X			
Literature review		X		X				X			X		X	X	X			
Content of existing instruments		X	X					X		X	X				X			
Factor analysis										X								
Other statistical analysis								X	X			X						
Determined by item selection									X								X	
Designing items																		
Qualitative work with patients			X					X		X			X	X			X	X
Expert opinion			X					X					X	X			X	
Literature review/ content of existing instruments			X						X		X	X	X	X				X
Traditional psychometric techniques										X		X					X	X
Rasch analysis or item response theory									X	X							X	X
Other statistical analysis								X	X			X						X
Intra-domain statements	X	X		X	X	X	X								X	X		
Validation																		
Traditional psychometric analysis of validation survey												X			X			
Other statistical analysis of validation survey										X								
Other									X									

Table 4: Summary of valuation and health state selection techniques

First author, date	Method for selecting health states	No of states	Technique	Method	Source	Purpose
Composite studies						
Brazier 2005 [12]	Statistically efficient plus randomly selected states	96	TTO	Interviews	Patients	QALY
Brazier 2008 [13]	Statistically efficient	49	SG	Interviews	Patients	QALY
Burr 2007 [14]	Statistically efficient	32 pairs	DCE	Postal	Patients	NQM
Cho 2015 [16]	All health states valued by all respondents	10	TTO/ VAS	Interviews	Public	QALY
Goodwin 2015b [22]	Rasch vignette plus patient feedback	169	TTO (MVH)	Online	Public	QALY
Gu 2013 [24]	Statistically efficient	26 pairs	DCE	Not stated	Public	QALY
Hauber 2010 [25]	Statistically efficient	48 pairs	DCE	Online	Patients	NQM
Johnson 2006 [28]	Statistically efficient	72 pairs	DCE	Not stated	Patients	NQM
Kerr 2015 [30]	Statistically efficient	20	TTO	Interviews	Public	QALY
Kind 2005 [31]	Statistically efficient	19	VAS	Postal	Public	QALY
Kok 2002 [33]	All health states valued by all respondents		TTO	Interviews	Public	QALY
Kularatna 2015 [72]	Statistically efficient	85	TTO (MVH)	Interviews	Public	QALY
Lamers 2007 [32]	Statistically efficient	19	VAS	Online	Public	QALY
Lloyd 2014 [42]	Statistically efficient	16	LT-TTO	Interviews	Public	QALY
Mavranezouli 2013 [46]	Rasch vignette approach	18	TTO (MVH)	Interviews	Public	QALY
McKenna 2008 [47]	All health states valued by groups of respondents	36	TTO (MVH)	Interviews	Public	QALY
Mulhern 2013 (DEMQOL) [57]	Statistically efficient	87	TTO (MVH)	Interviews	Public	QALY
Mulhern 2013 (proxy) [57]	Statistically efficient	70	TTO (MVH)	Interviews	Public	QALY
Mulhern 2012 [55]	Statistically efficient	50	TTO (MVH)	Interviews	Public	QALY
Palmer 2000 [59]	All health states valued by all respondents		VAS/ SG	Interviews	Patients	NQM
Petrillo 2011 [60]	Statistically efficient plus frequently observed states	41	TTO (MVH)	Interviews	Public	QALY
Ratcliffe 2009 [64]	All health states valued by groups of respondents	64	TTO (MVH)	Interviews	Public	QALY
	All health states valued by groups of respondents	66	Ranking	Interviews	Public	QALY
	Statistically efficient	12 pairs	DCE	Postal	Public	QALY
Rentz 2014 [35]	Unspecified selection method plus patient feedback	8	TTO	Interviews	Public	QALY
Rowen 2011 [71]	Statistically efficient	85	TTO (MVH)	Interviews	Public	QALY
Stevens 2005 [76]	All health states valued by all respondents		SG	Interviews	Public	QALY

Table 4 continued

First author, date	Method for selecting health states	No of states	Technique	Method	Source	Purpose
Palfreyman 2011 [78]	Statistically efficient	26	TTO	Interviews	Public	QALY
Stolk 2003 [79]	All health states valued by all respondents		TTO	Interviews	Public	QALY
Versteegh 2012 (MS) [82]	Statistically efficient plus frequently observed states	100	TTO	CAPI	Public	QALY
Versteegh 2012 (cancer) [82]	Statistically efficient	105	TTO	CAPI	Public	QALY
Versteegh 2012 (arthritis) [82]	Statistically efficient plus frequently observed states	56	TTO	CAPI	Public	QALY
Yang 2009 [84]	Statistically efficient	99	TTO (MVH)	Interviews	Public	QALY
Young 2010 [87]	Rasch vignette approach	16	TTO (MVH)	Interviews	Public	QALY
Yang 2011 [89]	Statistically efficient	99	TTO (MVH)	Interviews	Public	QALY
Dobrez 2007 [18]	Respondent's own current health state		TTO	Interviews	Patients	NQM
Martin 1998 [44]	Respondent's own current health state		TTO, SG	Interviews	Patients	NQM
Pickard 2009 [62]	Respondent's own current health state		TTO	Interviews	Patients	NQM
Decomposed studies						
Bellamy 1996 [10]	No health states valued: dimensions and levels valued separately		RA, VAS	Interviews	Patients	NQM
Beusterien 2005 [11]	VAS & SG: all 4 CS; 5 MA; perfect & normal health; dead. VAS: all SA		VAS, SG	Online	Public	QALY
Chiou 2005[b] [15]	VAS & SG: CS (excluding one implausible); 1 MA; pits. VAS: all SA.		VAS, SG	Interviews	Public	QALY
Cuervo 2014 [17]	VAS: all 5 CS; 3 MA; all SA; best state; pits; dead. TTO: 3 CS; 3 MA		VAS, TTO	Interviews	Public	UC
Flood 2006 [66]	VAS & SG: all 5 CS; 5 MA. VAS: all SA		VAS, SG	Interviews	Patients	NQM
Goodey 2000 [21]	No health states valued: dimensions and levels valued separately		RA, VAS	Interviews	Patients	NQM
Hodder 1997 [27]	No health states valued: dimensions and levels valued separately		VAS	NR	Clinicians	NQM
Lo 2006 [69]	VAS & SG: all 5 CS; 5 MA. VAS: all SA		VAS, SG	CAPI	Patients	NQM
Montejo 2011 [54]	VAS: all SA; all 8 CS; 2 MA; best state; pits. TTO: 2 CS; 2 MA		VAS, TTO	Interviews	Patients	NQM
Peacock 2008 [49]	TTO & RS: all 6 CS; pits. RS: all SA		RS, TTO	Interviews	NS	QALY
Poissant 2003 [63]	VAS: CS (excluding one implausible); best state; pits; dead; unconscious. No other preference elicitation technique used.		VAS	Interviews	Patients, carers	NQM
Revicki 1998a [65]	VAS & SG: all 5 CS; 5 MA. VAS: all SA		VAS, SG	Interviews	Patients	NQM
Revicki 1998b [68]	VAS & SG: all 5 CS; 5 MA. VAS: all SA		VAS, SG	Interviews	Patients	NQM
Revicki 2011 [70]	VAS & SG: all 5 CS; 5 MA. VAS: all SA		VAS, SG	Online	Patients	NQM
Shaw 1998 [75]	No health states valued: dimensions and levels valued separately		RA, VAS	Interviews	Patients	NQM

Table 4 continued

First author, date	Method for selecting health states	Technique	Method	Source	Purpose
Sundaram 2010 [81]	VAS & SG: 3 MA; pits; dead. VAS: 5 CS; 9 SA; best state	VAS, SG	Interviews	Patients	QALY
Tomlinson 2012 [39]	VAS: all 10 CS; 3 MA; dead; all SA. SG: 2 CS; 2 MA	VAS, SG	Interviews	Patients	NQM

[a] conventional TTO for health states considered better than being dead; lead-time TTO for health states considered worse than being dead; [b] All health states valued by all respondents

TTO = time trade-off; LT-TTO = lead-time TTO; MVH = Measurement and Valuation of Health; SG = standard gamble; VAS = visual analogue scale; DCE = discrete choice experiment; RS = rating scale; RA = resource allocation task; CAPI = computer-assisted personal interview; QALY = instrument designed to generate QALY weights; NQM = instrument not designed to generate QALY weights; CS = corner states; MA = multi-attribute states; SA = single attribute states; pits = worst possible health state; NR = not reported; NA = not applicable

Table 5: Preferred models

First author, year	Model type	Inconsistent coefficients	Sig coefficients p< 0.05	Sig coefficients p< 0.1	MAE	Errors > 0.05	Errors > 0.10	R-squared	RMSE
Brazier 2005 [12]	OLS	2	-	33%	0.053	38%	16%	0.178[g]	-
Brazier 2008 [13]	OLS	Merged	27%	36%	0.018	2%	2%	0.32[g]	-
Burr 2007 [14]	CLR	Merged	100%	100%	-	-	-	0.158[h]	-
Cho 2015 (TTO) [15]	Mixed	None	100%	100%	0.008	-	-	0.86[i]	-
Dobrez 2007 [18]	OLS	Merged	57%	86%	0.19	-	-	0.17	-
Goodwin 2015b [23]	RE	None	83%	83%	0.041	31%	7%	-	-
Gu 2013 [24]	CLR	None	100%	100%	-	-	-	-	-
Hauber 2010 [25]	RE probit	-	-	-	-	-	-	-	-
Johnson 2006 [28]	RE probit	-	-	-	-	-	-	-	-
Kerr 2015 [30]	RE	None	93%	-	0.018	-	-	-	-
Kind 2005 A [31]	OLS	None	-	-	-	-	-	0.443	-
Kind 2005 B [31]	OLS	None	-	-	-	-	-	0.413	-
Kularatna 2015 [72]	RE	Merged	100%	100%	-	-	-	-	-
Lamers 2007 A [32]	OLS	None	-	-	0.029	-	-	0.33	-
Lamers 2007 B [32]	OLS	None	-	-	0.043	-	-	0.29	-
Lloyd 2014 [42]	RE	None	100%	100%	0.045	-	-	-	0.063
Martin 1998 [44]	Other	-	-	-	-	-	-	0.42	-
Mavranezouli 2013 [46]	Rasch OLS	-	80%	100%	-	-	-	0.99[g]	0.028
McKenna 2008 [47]	OLS	None	100%	100%	0.041	35%	6%	0.936[g]	-
Mulhern 2012 [55]	RE	Merged	65%	77%	0.065	44%	14%	0.293[g]	0.052
Mulhern 2013[a] [57]	RE	Merged	73%	80%	0.045	35%	8%	-	0.060
Mulhern 2013[b] [57]	RE	Merged	82%	100%	0.042	36%	9%	-	0.057
Petrillo 2011 [60]	FE	Merged	80%	-	0.039	15%	2%	-	0.054
Pickard 2009 [62]	Other	Merged	44%	-	-	-	-	0.082[h]	-
Ratcliffe (TTO) [64]	OLS	2	56%	-	0.040	31%	None	0.517[g]	-
Ratcliffe (TTO) [64]	RE	None	78%	-	0.072	23%	2%	0.207[g]	-
Ratcliffe (DCE) [64]	RE probit	1	56%	-	0.077	38%	5%	0.203[g]	-
Ratcliffe (rank) [64]	Logit	1	78%	-	0.069	28%	6%	0.198[g]	-
Ratcliffe (rank) [64]	Rescaled logit	1	78%	-	0.083	69%	36%	0.198[g]	-
Rentz 2014 [35]	Rasch OLS	None	100%	100%	-	-	-	0.418[g]	-
Rowen 2011 [71]	OLS (ERUM)	2	88%	-	0.046	39%	7%	0.56	-
Versteegh 2012[c] [82]	RE	None	100%	-	0.040	-	-	0.78	-
Versteegh 2012[d] [82]	RE	None	100%	-	0.033	-	-	0.88	-
Versteegh 2012[e] [82]	RE	None	100%	-	0.028	-	-	0.94	-

Table 5 continued

First author, year	Model type	Inconsistent coefficients	Sig coefficients p< 0.05	Sig coefficients p< 0.1	MAE	Errors > 0.05	Errors > 0.10	R-squared	RMSE
Yang 2009 [84]	OLS	Merged	67%	-	0.045	40%	6%	-	-
Yang 2011 [89]	OLS	Merged	65%	-	0.048	32%	9%	0.957	0.046
Young, 2010 [87]	Rasch OLS	-	100%[f]		-	-	-	0.958	0.042

[a] DEMQOL-U; [b] DEMQOL-U-proxy; [c] MS-PBM; [d] QLQ-PBM; [e] HAQ-PBM; [f] p-value not stated; [g] adjusted R-squared; [h] pseudo R-squared; [i] generalised R-squared

No modelling: Kok 2002 [33]; Palmer 2000 [59]; Stevens 2005 [76]; Stolk 2003 [79]. No report of model performance: Palfreyman, 2010 [78].

MAE = mean absolute error; RMSE = root mean squared error; OLS = ordinary least squares; CLR = conditional logistic regression; mixed = model incorporating fixed and random effects; RE = random effects; FE = fixed effects; ERUM = episodic random utility model

Table 6: Validation of condition-specific preference-based measures

Test / Comparator measure	Convergent validity: Parent	GPBM	Other	Map	Values	Discriminative validity: Parent	GPBM	Other	Map	NC	Responsiveness: Parent	GPBM	Other	Map	NC
Beusterien 2005 [11]					X										
Brazier 2008 [13]											X				
Burr 2007 [14]		X	X												
Chiou 2005 [15]		X			X					X					
Cuervo 2014 [17]					X										
Dobrez 2007 [18]					X					X					
Flood 2006 [66]			X							X					
Goodwin 2015b [23]	X	X				X	X				X	X			
Kharroubi 2014 [86]					X										
Kok 2002 [33]					X										
Lamers 2007 [32]										X					X
Lo 2006 [69]			X							X					
McKenna 2008 [47]	X						X								
Montejo 2011 [54]		X	X												
Mulhern 2013 [57]	X	X	X							X					X
Palmer 2000 [59]			X							X					
Petrillo 2011 [60]					X					X					X
Poissant 2003 [63]			X							X		X			
Rentz 2014 [35]	X									X					
Revicki 1998a [65]		X	X							X					
Revicki 1998b [68]		X	X							X					
Revicki 2011 [70]			X							X					X
Sundaram 2010 [81]			X		X					X					
Tomlinson 2012 [39]					X										
Versteegh 2012 [82]		X					X								X
Retrospective validations {original papers}															
Avila 2014 [40] {39}						X									
Bime 2012 [67] {65}			X												X
Brazier 2012 [1] {46; 71; 84; 89}	X	X				X	X				X	X			
Desroziers 2013 [85] {84}		X										X			
Gothwal 2013b [53] {49}					X			X							
Hess 2013 [19] {16}			X	X				X	X				X	X	
Finger 2013b [51] {49}							X								

Table 6 continued

Test / Comparator measure	Convergent validity					Discriminative validity					Responsiveness				
	Parent	**GPBM**	**Other**	**Map**	**Values**	**Parent**	**GPBM**	**Other**	**Map**	**NC**	**Parent**	**GPBM**	**Other**	**Map**	**NC**
McTaggart-Cowan 2008 [91] {89}	X	X	X			X	X	X							
Naik 2013 [36] {35}		X				X					X				
Pickard 2012 [20] {18; 31}		X	X	X			X		X						
Rowen 2012b [73] {71}	X	X		X		X	X		X		X	X		X	

Parent = measure from which classification system was derived; GPBM = generic preference-based measure; values = directly elicited HSUVs; map = mapping algorithm from parent measure to EQ-5D; NC = no comparator

References

1. Brazier JE, Rowen D, Mavranezouli I, Tsuchiya A, Young T, Yang Y, Barkham M, Ibbotson R. Developing and testing methods for deriving preference-based measures of health from condition-specific measures (and other patient-based measures of outcome). Health Technol Assess. 2012;16(32):1-114

2. Dolan P. Modeling valuations for EuroQol health states. Med Care. 1997;35:1095-1108.

3. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. J Health Econ. 2002;21:271-292.

4. Horsman J, Furlong W, Feeny D and Torrance G. The Health Utilities Index (HUI): concepts, measurement properties and applications. Health Qual Life Out. 2003;1:54.

5. Brazier J, Ratcliffe J, Salomon J, Tsuchiya A. Measuring and Valuing Health Benefits for Economic Evaluation. Oxford: Oxford University Press; 2007.

6. Feeny D. The multi-attribute utility approach to assessing health-related quality of life. In: Jones A, editor. The Elgar Companion to Health Economics. Cheltenham: Edward Elgar; 2006. pp.359-370.

7. Mortimer D, Segal L. Comparing the incomparable? A systematic review of competing techniques for converting descriptive measures of health status into QALY-weights. Med Decis Making 2008;28 (1):66-89.

8. Petrillo J, Cairns J. Converting condition-specific measures into preference-based outcomes for use in economic evaluation. Expert Rev Pharmacoecon Outcomes Res. 2008;8(5):453-61.

9. Lin F-J, Longworth L, Pickard A. Evaluation of content on EQ-5D as compared to disease-specific utility measures. Qual Life Res. 2013;22(4):853-74.

10. Bellamy CA, Brickley MR, McAndrew R. Measurement of patient-derived utility values for periodontal health using a multi-attribute scale. J Clin Periodonto. 1996;23(9):805-9.

11. Beusterien K, Leigh N, Jackson C, Miller R, Mayo K, Revicki D. Integrating preferences into health status assessment for amyotrophic lateral sclerosis: The ALS Utility Index. Amyotroph Lateral Scler Other Motor Neuron Disord. 2005;6(3):169-76.

12. Brazier JE, Roberts J, Platts M, Zoellner YF. Estimating a preference-based index for a menopause specific health quality of life questionnaire. Health Qual Life Out. 2005;3:13.

13. Brazier J, Czoski-Murray C, Roberts J, Brown M, Symonds T, Kelleher C. Estimation of a preference-based index from a condition-specific measure: The King's Health Questionnaire. Med Dec Making. 2008;28(1):113-26.

14. Burr JM, Kilonzo M, Vale L, Ryan M. Developing a preference-based glaucoma utility index using a discrete choice experiment. Optom Vis Sci. 2007;84(8):797-808.

15. Chiou CF, Weaver MR, Bell MA, Lee TA, Krieger JW. Development of the multi-attribute pediatric asthma health outcome measure (PAHOM). Int J Qual Health Care. 2005;17(1):23-30.

16. Cho S, Kim H, Kim SH, Ock M, Oh YM, Jo MW. Utility estimation of hypothetical chronic obstructive pulmonary disease health states by the general population and health professionals. Health Qual Life Out. 2015;13:34.

17. Cuervo J, Castejón N, Khalaf KM, Waweru C, Globe D, Patrick DL. Development of the Incontinence Utility Index: estimating population-based utilities associated with urinary problems from the Incontinence Quality of Life Questionnaire and Neurogenic Module. Health Qual Life Out. 2014;12:147.

18. Dobrez D, Cella D, Pickard AS, Lai JS, Nickolov A. Estimation of patient preference-based utility weights from the functional assessment of cancer therapy - General. Value Health. 2007;10(4):266-72.

19. Hess LM, Brady WE, Havrilesky LJ, Cohn DE, Monk BJ, Wenzel L, et al. Comparison of methods to estimate health state utilities for ovarian cancer using quality of life data: a Gynecologic Oncology Group study. Gynecol Oncol. 2013;128(2):175-80.

20. Pickard AS, Ray S, Ganguli A, Cella D. Comparison of FACT- and EQ-5D-based utility scores in cancer. Value Health. 2012;15(2):305-11.

21. Goodey RD, Brickley MR, Armstrong RA, Shepherd JP. The minor oral surgery outcome scale: a multi-attribute patient-derived outcome measure. J Oral Maxillofac Surg. 2000;58(10):1096-101.

22. Goodwin E, Green C. A Quality-Adjusted Life-Year Measure for Multiple Sclerosis: Developing a Patient-Reported Health State Classification System for a Multiple Sclerosis-Specific Preference-Based Measure. Value Health. 2015; in press.

23. Goodwin E, Green C, Spencer A. Estimating a preference-based index for an eight dimensional health state classification system derived from the Multiple Sclerosis Impact Scale (MSIS-29). Value Health. 2015; in press.

24. Gu NY, Botteman MF, Gerber RA, Ji X, Postema R, Wan Y, Sianos G, Anthony I, Cappelleri JC, Szczypa P, van Hout B. Eliciting health state utilities for Dupuytren's contracture using a discrete choice experiment. Acta Orthop. 2013;84(6):571-578.

25. Hauber AB, Mohamed AF, Johnson FR, Oyelowo O, Curtis BH, Coon C. Estimating importance weights for the IWQOL-Lite using conjoint analysis. Qual Life Res. 2010;19(5):701-9.

26. Mohamed AF, Hauber AB, Johnson FR, Coon CD. Patient preferences and linear scoring rules for patient-reported outcomes. Patient. 2010;3(4):217-27.

27. Hodder SC, Edwards MJ, Brickley MR, Shepherd JP. Multiattribute utility assessment of outcomes of treatment for head and neck cancer. Br J Cancer. 1997;75(6):898-902.

28. Johnson FR, Hauber AB, Osoba D, Hsu MA, Coombs J, Copley-Merriman C. Are chemotherapy patients' HRQoL importance weights consistent with linear scoring rules? A stated-choice approach. Qual Life Res. 2006;15(2):285-98.

29. Osoba D, Hsu M-A, Copley-Merriman C, Coombs J, Johnson FR, Hauber B, Manjunath R, Pyles A. Stated preferences of patients with cancer for health-related quality-of-life (HRQOL) domains during treatment. Qual Life Res. 2006;15:273–283.

30. Kerr C, Breheny K, Lloyd A, Brazier J, Bailey DB, Berry-Kravis E, Cohen J, Petrillo J. Developing a utility index for the Aberrant Behavior Checklist (ABC-C) for fragile X syndrome. Qual Life Res. 2015;24(2): 305-314.

31. Kind P, Macran S. Eliciting social preference weights for functional assessment of cancer therapy-lung health states. PharmacoEconomics. 2005;23(11):1143-53.

32. Lamers LM, Uyl-de Groot CA, Buijt I. The use of disease-specific outcome measures in cost-utility analysis: The development of Dutch societal preference weights for the FACT-L scale. PharmacoEconomics. 2007;25(7):591-603.

33. Kok ET, McDonnell J, Stolk EA, Stoevelaar HJ, Busschbach JJV. The valuation of the international prostate symptom score (IPSS) for use in economic evaluations. Eur Urol. 2002;42(5):491-7.

34. Kowalski JW, Rentz AM, Walt JG, Lloyd A, Lee J, Young TA, et al. Rasch analysis in the development of a simplified version of the National Eye Institute Visual-Function Questionnaire-25 for utility estimation. Qual Life Res. 2012;21(2):323-34.

35. Rentz AM, Kowalski JW, Walt JG, Hays RD, Brazier JE, Yu R, Lee P, Bressler N, Revicki DA. Development of a preference-based index from the National Eye Institute Visual Function Questionnaire-25. JAMA Ophthalmology. 2014;132(3):310-318.

36. Naik RK, Gries KS, Rentz AM, Kowalski JW, Revicki DA. Psychometric evaluation of the National Eye Institute Visual Function Questionnaire and Visual Function Questionnaire Utility Index in patients with non-infectious intermediate and posterior uveitis. Qual Life Res. 2013;22(10):2801-2808.

37. Krahn M, Ritvo P, Irvine J, Tomlinson G, Bezjak A, Trachtenberg J, et al. Construction of the Patient-Oriented Prostate Utility Scale (PORPUS): A multiattribute health state classification system for prostate cancer. J Clin Epidemiol. 2000;53 (9):920-30.

38. Ritvo P, Irvine J, Naglie G, Tomlinson G, Bezjak A, Matthew A, et al. Reliability and validity of the PORPUS, a combined psychometric and utility-based quality-of-life instrument for prostate cancer. J Clin Epidemiol. 2005;58 (5):466-74.

39. Tomlinson G, Bremner KE, Ritvo P, Naglie G, Krahn MD. Development and validation of a utility weighting function for the patient-oriented prostate utility scale (PORPUS). Med Decis Making. 2012;32(1):11-30.

40. Avila M, Pardo Y, Castells M, Ferrer F, et al on behalf of The Multicentric Spanish Group of Clinically Localized Prostate Cancer. Adaptation and validation of the Spanish version of the Patient-Oriented Prostate Utility Scale (PORPUS). Qual Life Res. 2014;23(9): 2481-2487.

41. Kuspinar A, Finch L, Pickard S, Mayo NE. Using existing data to identify candidate items for a health state classification system in multiple sclerosis. Qual Life Res. 2014;23(5): 1445-1457.

42. Lloyd A, Kerr C, Breheny K, Brazier J, Ortiz A, Borg E. Economic evaluation in short bowel syndrome (SBS): an algorithm to estimate utility scores for a patient-reported SBS-specific quality of life scale (SBS-QoLTM). Qual Life Res. 2014;23(2):449-458.

43. Martin AJ, Glasziou PP, Simes RJ. A cardiovascular extension of the Health Measurement Questionnaire. J Epidemiol Community Health. 1999;53(9):548-57.

44. Martin AJ, Glasziou PP, Simes RJ, Lumley T. Predicting patients' utilities from quality of life items: An improved scoring system for the UBQ-H. Qual Life Res. 1998;7(8):703-11.

45. Mavranezouli I, Brazier JE, Young TA, Barkham M. Using Rasch analysis to form plausible health states amenable to valuation: the development of CORE-6D from a measure of common mental health problems (CORE-OM). Qual Life Res. 2011;20(3):321-33.

46. Mavranezouli I, Brazier JE, Rowen D, Barkham M. Estimating a Preference-Based Index from the Clinical Outcomes in Routine Evaluation-Outcome Measure (CORE-OM): Valuation of CORE-6D. Med Decis Making. 2013;33(3):381-95.

47. McKenna SP, Ratcliffe J, Meads DM, Brazier JE. Development and validation of a preference based measure derived from the Cambridge Pulmonary Hypertension Outcome Review (CAMPHOR) for use in cost utility analyses. Health Qual Life Out. 2008;6:65.

48. Misajon R, Hawthorne G, Richardson J, Barton J, Peacock S, Iezzi A, et al. Vision and quality of life: The development of a utility measure. Invest Ophthalmol Vis Sci. 2005;46(11):4007-15.

49. Peacock S, Misajon R, Iezzi A, Richardson J, Hawthorne G, Keeffe J. Vision and quality of life: Development of methods for the VisQoL vision-related utility instrument. Ophthalmic Epidemiol. 2008;15(4):218-23.

50. Finger RP, Kortuem K, Fenwick E, von Livonius B, Keeffe JE, Hirneiss CW. Evaluation of a vision-related utility instrument: the German vision and quality of life index. Invest Ophthalmol Vis Sci. 2013a;54(2):1289-94.

51. Finger RP, Fenwick E, Hirneiss CW, Hsueh A, Guymer RH, Lamoureux EL, Jill E. Keeffe JE. Visual Impairment as a Function of Visual Acuity in Both Eyes and Its Impact on Patient Reported Preferences. PLoS ONE. 2013b;8:12.

52. Gothwal VK, Bagga DK. Vision and Quality of Life Index: validation of the Indian version using Rasch analysis. Invest Ophth Vis Sci. 2013a;54(7):4871-4881.

53. Gothwal VK, Bagga DK. Utility values in the visually impaired: comparing time-trade off and VisQoL. Optometry Vision Sci 2013b;90(8):843-854.

54. Montejo AL, Correas-Lauffer J, Maurino J, Villa G, Rebollo P, Diez T, et al. Estimation of a Multiattribute Utility Function for the Spanish Version of the TooL Questionnaire. Value Health. 2011;14(4):564-70.

55. Mulhern B, Rowen D, Jacoby A, Marson T, Snape D, Hughes D, et al. The development of a QALY measure for epilepsy: NEWQOL-6D. Epilepsy Behav. 2012a;24(1):36-43.

56. Mulhern B, Smith SC, Rowen D, Brazier JE, Knapp M, Lamping DL, et al. Improving the measurement of QALYs in dementia: developing patient- and carer-reported health state classification systems using Rasch analysis. Value Health. 2012b;15(2):323-33.

57. Mulhern B, Rowen D, Brazier J, Smith S, Romeo R, Tait R, et al. Development of DEMQOL-U and DEMQOL-PROXY-U: Generation of preference-based indices from DEMQOL and DEMQOL-PROXY for use in economic evaluation. Health Technol Assess. 2013;17(5):1-140.

58. Rowen D, Mulhern B, Banerjee S, Hout Bv, Young TA, Knapp M, et al. Estimating preference-based single index measures for dementia using DEMQOL and DEMQOL-Proxy. Value Health. 2012a;15(2):346-56.

59. Palmer CS, Schmier JK, Snyder E, Scott B. Patient preferences and utilities for 'off-time' outcomes in the treatment of Parkinson's disease. Qual Life Res. 2000;9 (7):819-27.

60. Petrillo J, Cairns J. Development of the EXACT-U: a preference-based measure to report COPD exacerbation utilities. Value Health. 2011;14(4):546-54.

61. Petrillo J, Cairns J. Evaluation of COPD exacerbations using the EXACT-U. Annual European Respiratory Society Congress; Barcelona. 2010.

62. Pickard AS, Shaw JW, Lin HW, Trask PC, Aaronson N, Lee TA, et al. A patient-based utility measure of health for clinical trials of cancer therapy based on the European Organization for the Research and Treatment of Cancer Quality of Life Questionnaire. Value Health. 2009;12(6):977-88.

63. Poissant L, Mayo NE, Wood-Dauphinee S, Clarke AE. The development and preliminary validation of a Preference-Based Stroke Index (PBSI). Health Qual Life Out. 2003;1:43.

64. Ratcliffe J, Brazier J, Tsuchiya A, Symonds T, Brown M. Using DCE and ranking data to estimate cardinal values for health states for deriving a preference-based single index from the sexual quality of life questionnaire. Health Econ. 2009;18(11):1261-76.

65. Revicki DA, Leidy NK, Brennan-Diemer F, Sorensen S, Togias A. Integrating patient preferences into health outcomes assessment: The multiattribute asthma symptom utility index. Chest. 1998a;114(4):998-1007.

66. Flood EM, De Cock E, Mork A-C, Revicki DA. Evaluating preference weights for the Asthma Symptom Utility Index (ASUI) across countries. Health Qual Life Out. 2006;4:51.

67. Bime C, Wei CY, Holbrook JT, Sockrider MM, Revicki DA, Wise RA. Asthma symptom utility index: reliability, validity, responsiveness, and the minimal important difference in adult asthmatic patients. J Allergy Clin Immunol. 2012;130(5):1078-84.

68. Revicki DA, Leidy NK, Brennan-Diemer F, Thompson C, Togias A. Development and preliminary validation of the multiattribute Rhinitis Symptom Utility Index. Qual Life Res. 1998b;7(8):693-702.

69. Lo PSY, Tong MCF, Revicki DA, Lee CC, Woo JKS, Lam HCK, et al. Rhinitis symptom utility index (RSUI) in Chinese subjects: A multiattribute patient-preference approach. Qual Life Res. 2006;15(5):877-87.

70. Revicki DA, Margolis MK, Thompson CL, Meltzer EO, Sandor DW, Shaw JW. Major Symptom Score Utility Index for patients with acute rhinosinusitis. Am J Rhinol. 2011;25(3):E99-E106.

71. Rowen D, Brazier J, Young T, Gaugris S, Craig BM, King MT, et al. Deriving a preference-based measure for cancer using the EORTC QLQ-C30. Value Health. 2011;14(5):721-31.

72. Kularatna S, Whitty JA, Johnson NW, Jayasinghe R, Scuffham PA. Development of an EORTC-8D utility algorithm for Sri Lanka. Med Decis Making. 2015;35(3):361-370.

73. Rowen D, Young T, Brazier J, Gaugris S. Comparison of generic, condition-specific, and mapped health state utility values for multiple myeloma cancer. Value Health. 2012b;15(8):1059-68.

74. Scholzel-Dorenbos CJM, Arons AMM, Wammes JJG, Rikkert MGMO, Krabbe PFM. Validation study of the prototype of a disease-specific index measure for health-related quality of life in dementia. Health Qual Life Out. 2012;10:118.

75. Shaw RW, Brickley MR, Evans L, Edwards MJ. Perceptions of women on the impact of menorrhagia on their health using multi-attribute utility assessment. Br J Obstet Gynaecol. 1998;105(11):1155-9.

76. Stevens KJ, Brazier JE, McKenna SP, Doward LC, Cork MJ. The development of a preference-based measure of health in children with atopic dermatitis. Br J Dermatol. 2005;153(2):372-7.

77. Stevens K, Palfreyman S. The use of qualitative methods in developing the descriptive systems of preference-based measures of health-related quality of life for use in economic evaluation. Value Health. 2012;15(8):991-8.

78. Palfreyman S. The SPVU-5D: A preference-based measure of health related quality of life for use with venous leg ulceration. Patient Reported Outcomes Newsletter. 2011;45:7-9. http://www.pro-newsletter.com/content/view/350/74/

79. Stolk EA, Busschbach JJV. Validity and feasibility of the use of condition-specific outcome measures in economic evaluation. Qual Life Res. 2003;12(4):363-71.

80. Sundaram M, Smith MJ, Revicki DA, Elswick B, Miller LA. Rasch analysis informed the development of a classification system for a diabetes-specific preference-based measure of health. J Clin Epidemiol. 2009;62(8):845-56.

81. Sundaram M, Smith MJ, Revicki DA, Miller LA, Madhavan S, Hobbs G. Estimation of a valuation function for a diabetes mellitus-specific preference-based measure of health: The diabetes utility index. PharmacoEconomics. 2010;28(3):201-16.

82. Versteegh MM, Leunis A, Uyl-de Groot CA, Stolk EA. Condition-specific preference-based measures: benefit or burden? Value Health. 2012;15(3):504-13.

83. Young T, Yang Y, Brazier JE, Tsuchiya A, Coyne K. The first stage of developing preference-based measures: Constructing a health-state classification using Rasch analysis. Qual Life Res. 2009;18(2):253-65.

84. Yang Y, Brazier J, Tsuchiya A, Coyne K. Estimating a preference-based single index from the overactive bladder questionnaire. Value Health. 2009;12(1):159-66.

85. Desroziers K, Aballéa S, Maman K, Nazir J, Odeyemi I, Hakimi Z. Estimating EQ-5D and OAB-5D health state utilities for patients with overactive bladder. Health Qual Life Out. 2013;11:200.

86. Kharroubi SA, Brazier J, Yang Y. Modeling a preference-based index for two condition-specific measures (asthma and overactive bladder) using a nonparametric Bayesian method. Value Health. 2014;17(4): 406-415.

87. Young TA, Rowen D, Norquist J, Brazier JE. Developing preference-based health measures: using Rasch analysis to generate health state values. Qual Life Res. 2010;19(6):907-17.

88. Young TA, Yang Y, Brazier JE, Tsuchiya A. The use of rasch analysis in reducing a large condition-specific instrument for preference valuation: the case of moving from AQLQ to AQL-5D. Med Decis Making. 2011;31(1):195-210.

89. Yang YL, Brazier JE, Tsuchiya A, Young TA. Estimating a Preference-Based Index for a 5-Dimensional Health State Classification for Asthma Derived from the Asthma Quality of Life Questionnaire. Med Decis Making. 2011;31(2):281-91.

90. Hernández Alava M, Brazier J, Rowen D, Tsuchiya A. 2013. Common Scale Valuations across Different Preference-Based Measures: Estimation Using Rank Data. Med Decis Making. 33(6):839-852.

91. McTaggart-Cowan HM, Marra CA, Yang Y, Brazier JE, Kopec JA, FitzGerald JM, et al. The validity of generic and condition-specific preference-based instruments: The ability to discriminate asthma control status. Qual Life Res. 2008;17(3):453-62.

92. Hedayat A, Sloane N, Stufken J. Orthogonal Arrays: Theory and Applications. New York: Springer; 1999.

93. Hagquist C, Bruce M, Gustavsson J. Using the Rasch model in nursing research: An introduction and illustrative example. Int J Nurs Stud. 2009;46:380–93.

94. Tennant A, Conaghan P. The Rasch Measurement Model in Rheumatology: What Is It and Why Use It? When Should It Be Applied, and What Should One Look for in a Rasch Paper? Arthritis Rheum. 2007;57(8):1358-62.

95. O'Connor R. Measuring Quality of Life in Health. Edinburgh: Churchill Livingstone; 2004.

96. NICE. Guide to the methods of technology appraisal 2013. National Institute for Health and Care Excellence, 2013.

97. Feeny D. The utility approach to assessing population health. In: Murray C, Salomon J, Mathers C, Lopez A, editors. Summary Measures of Population Health: Concepts, Ethics, Measurement and Applications. Geneva: World Health Organisation; 2002. pp.515-528.

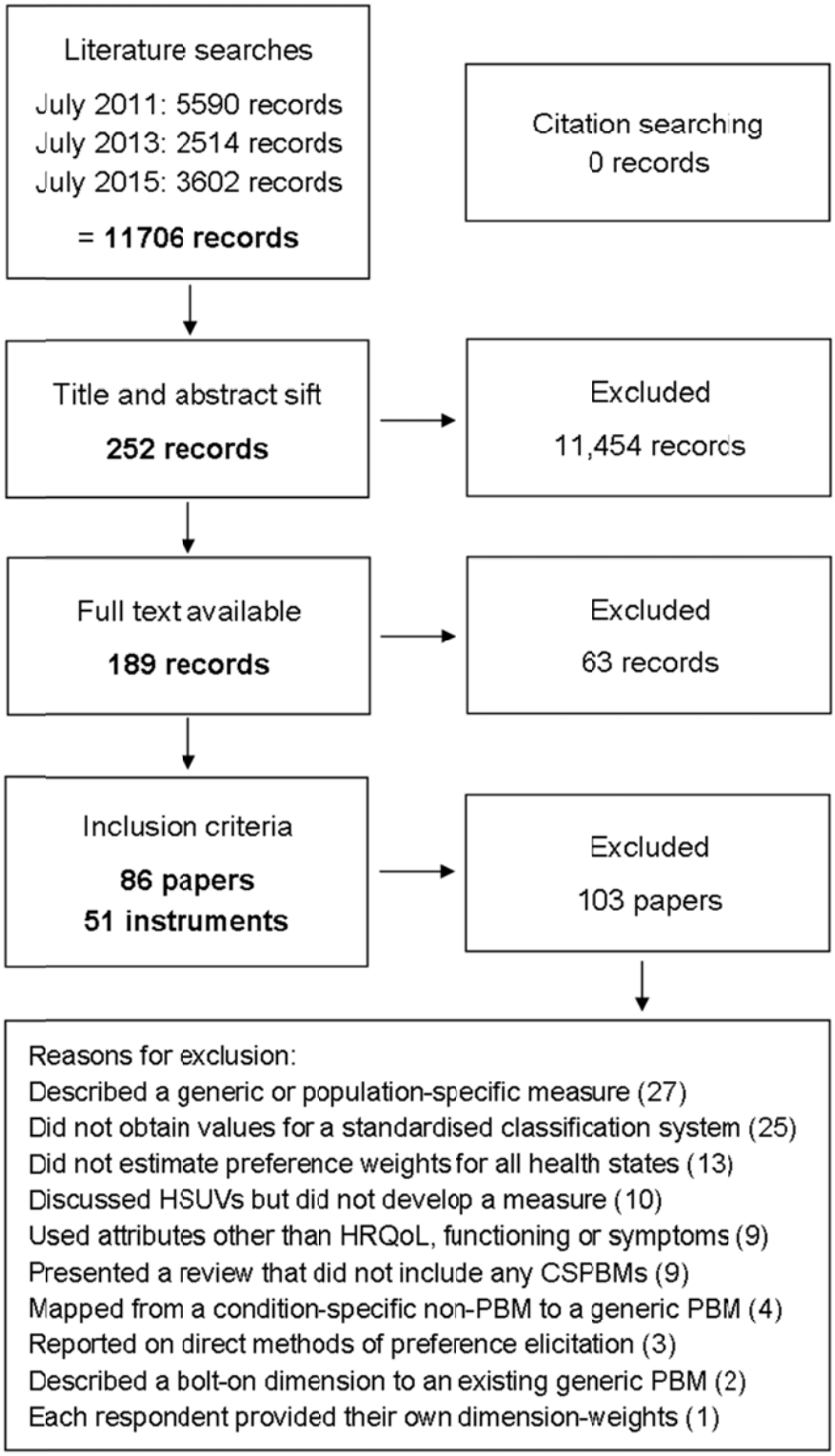
98. Dolan P. Modelling the relationship between the description and valuation of health states. In: Murray C, Salomon J, Mathers C, Lopez A, editors. Summary Measures of

Population Health: Concepts, Ethics, Measurement and Applications. Geneva: World Health Organisation; 2002. pp.501-514.

99. Drummond M, Sculpher M, Torrance G, O'Brien B, Stoddart G. Methods for the Economic Evaluation of Health Care Programmes. 3rd edition. Oxford: Oxford University Press; 2005.

Figure legend:

Fig 1 Results of systematic literature search



Supplementary material

Literature search strategy

Search terms	Hits 2011	Hits 2013	Hits 2015
1. “quality of life”/	166328	223308	289794
2. quality of life.mp	203988	276051	354960
3. qol.mp	18459	28351	40268
4. hrqol.mp	5882	9378	13671
5. hrql.mp	2173	3004	3812
6. health status/	63583	78316	93296
7. health status.mp	77473	96043	115196
8. quality adjusted life year/	7309	10733	14342
9. quality adjusted life year$.mp	8691	12696	16536
10. qaly$.mp	4432	7587	10705
11. develop$.mp	2850491	3523944	4249388
12. deriv$.mp	1984064	2331151	2600597
13. estimat$.mp	597063	757811	948382
14. creat$.mp	386614	496471	624253
15. generat$.mp	638314	806968	984354
16. construct$.mp	295607	362776	435422
17. valu$.mp	1415469	1727519	2120858
18. transform$.mp	356439	431408	517037
19. transfer$.mp	486773	586677	691188
20. translat$.mp	157242	208766	263142
21. conver$.mp	369549	453000	545448
22. map$.mp	328352	408076	473423
23. preference base$.mp	556	824	1117
24. utilit$.mp	109102	144102	182591
25. preference weight$.mp	147	239	319
26. valuation weight$.mp	2	3	4
27. preference valu$.mp	113	147	181
28. qaly weight$.mp	28	40	53
29. health state$	3505	5013	6681
30. valu$	1415469	1727519	2120858
31. preference$	86019	109509	136584
32. rank$	94243	131160	179437
33. 30 OR 31 OR 32	1566231	1927111	2380666
34. (health state$ ADJ5 (valu$ OR preference$ OR rank$)).mp	796	1134	1409
35. instrument$.mp	409212	487262	526455
36. measure$.mp	2072299	2548309	3096367
37. classification system$.mp	12404	16542	21106
38. health state classification$.mp	52	72	91
39. descriptive system$.mp	135	195	279
40. index.mp	414191	553790	730951
41. indices.mp	95349	115862	141360
42. indexes.mp	25599	31604	38698
43. “quality of life index”/	799	1361	1931
44. tool$.mp	336360	457131	597143
45. questionnaire/ OR structured questionnaire/	274878	360022	454615
46. questionnaire$.mp	367011	480246	610647
47. scale$.mp	441790	588311	769671
48. rating scale/	67425	78984	91093

Table continues overleaf

Search terms	Hits 2011	Hits 2013	Hits 2015
49. 1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10	269913	357153	452503
50. 11 OR 12 OR 13 OR 14 OR 15 OR 16 OR 17 OR 18 OR 19 OR 20 OR 21 OR 22	7423215	8986145	10649837
51. 23 OR 24 OR 25 OR 26 OR 27 OR 28 OR 34	109791	145010	183761
52. 35 OR 36 OR 37 OR 38 OR 39 OR 40 OR 41 OR 42 OR 43 OR 44 OR 45 OR 46 OR 47 OR 48	3441043	4294449	5251866
53. 49 AND 50 AND 51 AND 52	**3431**	6888	9283
54. LIMIT 53 to English language		6672	9042
55. LIMIT 54 to yr="2011 –Current"		**2100**	
55. LIMIT 54 to yr="2013 –Current"			**2761**

Please note: The search strategy also includes terms to identify mapping studies, to contribute to other research projects.

Databases searched: Embase; Medline (R); Medline (R) In-Process and Other Non-Indexed Citations; PsycINFO; Web of Science (Science Citation Index, Social Sciences Citation Index, Conference Proceedings Citation Index – Science, Conference Proceedings Citation Index – Social Science and Humanities); the Cochrane Library (Cochrane reviews, other reviews, clinical trials, methods studies, technology assessments and economic evaluations); CINAHL; EconLit; ASSIA; and the Health Management Information Consortium (HMIC) database. In addition, citation searching, a Google search and searches of the Discussion Papers published on the Sheffield School of Health and Related Research (ScHARR) Health Economics and Decision Science website were undertaken.

Data extraction form

Basic details				
1st author:		**Year:**		
Condition:				
Non-PBM / *de novo*:				
New PBM name:				
Citation:				
Article coverage:		**Brazier review?**	Y/N	
Classification system				
Original instrument:	**Dimensions:**	**Items:**	**Levels:**	
New instrument:	**Dimensions:**	**Items:**	**Levels:**	**States:**
Rationale for choosing instrument:				
Dimensions of new instrument:				
Dimensionality design methods:				
Item reduction/ selection methods:				
Level reduction/ selection methods:				
Dataset used:	**Sample size:**	**Population:**		
Classification system validation:				
Additional notes:				
Valuation methods				
Health state selection method:				
States per respondent:				
Total n^{o} states valued:		**N^{o} per respondent:**		
Valuation source:	*Patient/public*	**Country:**		
Sample size:		**N^{o} of respondents:**		
Respondents excluded:				
Valuation technique:	*eg TTO*			
Valuation method:	*eg interview*	**Condition label:**		
Rationale for choice of technique:				
States worse than dead:				
Upper anchor:		**Lower anchor:**		
Additional notes:				

Modelling			
Modelling approach:	*Decomposed / statistical*	**Chosen model type:**	*Additive/ multiplicative etc*
Chosen model level:	*Individual/ aggregate*	**Constant:**	
Preference interactions:			
Transformations:	*For statistical approach, did they adjust for skewness, truncation and non-continuity of data?*		
VAS to SG conversion:	*If VAS was used, how was this converted to SG?*		
Models tested:	*Particularly if individual OLS model selected, were RE and mean level models considered?*		
Rationale for selecting preferred model:			
Preferred model specification:			
Respondent characteristics:	*Were any significant differences found between groups of respondents?*		
Additional notes:			
Model performance			
Mean absolute error		**MAE 95% CIs**	
N^{o} of errors > 0.05		**N^{o} of errors > 0.10**	
Root mean sq. error		**Mean error**	
R-squared		**Adj R-squared**	
Max predicted score compared to obs		**Min predicted score compared to obs**	
Proportion of coefficients p<0.05		**Proportion of coefficients p<0.1**	
Coefficients with unexpected sign		**Inconsistent coefficients**	
Other goodness of fit measures	*Eg t-test and the normality of prediction errors (eg the Jarque-Bera test) to assess bias in predictions; testing of MAUT models*		
Additional notes:			
Validation of overall instrument			
Acceptability:			
Reliability:			
Validity:			
Responsiveness:			
Other notes on quality:			
Any other notes			

Title: A systematic review of the literature on the development of condition-specific preference-based measures of health

Journal: Applied Health Economics and Health Policy

Authors: Elizabeth Goodwin PhD, Colin Green PhD

Corresponding author: Elizabeth Goodwin, Health Economics Group, University of Exeter Medical School, University of Exeter, Exeter, UK

Email: E.Goodwin@exeter.ac.uk