Developing and testing methods for deriving preference-based measures of health from condition-specific measures (and other patient-based measures of outcome)

JE Brazier,* D Rowen, I Mavranezouli, A Tsuchiya, T Young, Y Yang, M Barkham and R Ibbotson

Health Economics and Decision Science, School of Health and Related Research (ScHARR), University of Sheffield, Sheffield, UK

*Corresponding author

Executive summary

Health Technology Assessment
NIHR HTA programme
www.hta.ac.uk

Executive summary

Background

Quality-adjusted life-years (QALYs) are increasingly being calculated using health-state values provided by generic preference-based measures of health. However, generic preference-based measures are not used in all studies, may not cover all dimensions of relevance to some medical conditions as their focus is general rather than specific, and may not be appropriate for all conditions. In contrast, condition-specific measures are often used in clinical studies and may be regarded as better able to capture the impact on the health-related quality of life (HRQoL) of patients with that condition, as they are often focused on symptoms or the HRQoL associated with the symptoms of that condition. A limitation with condition-specific measures is that they are not preference based and so cannot be used to estimate QALYs. Recent years have seen the development of methods for deriving preference-based measures from condition-specific measures, including the derivation of health-state classification systems to generate states for valuation. This project sought to review these methods and then to address a range of issues in the development and use of condition-specific preference-based measures (CSPBMs) to estimate QALYs for use in economic evaluation.

Objectives

The specific objectives are as follows:

1. to identify and review the existing literature on current methods for deriving a preference-based measure of health from non-preference-based measures of health
2. to examine and test a new method for generating health states from non-preference-based measures using Rasch modelling
3. to assess the impact of referring to the medical condition (or disease) in the descriptions on health-state values
4. to assess the impact of attempting to capture side effects using CSPBMs on health-state values
5. to assess the impact of comorbidities by testing the additivity assumption and the extent of any violation across two conditions (asthma and common mental health problems)
6. to examine the degree of information loss of moving from the original instrument to the preference-based index
7. to compare CSPBMs with generic preference-based measures (including EQ-5D and SF-6D) in order to examine the degree of agreement and the extent of any gain in psychometric performance
8. to propose a set of conditions that should be satisfied in order to justify the development of CSPBMs for use in economic evaluation
9. to examine whether CSPBMs can be used to inform resource allocation decisions.

Methods and results

Five studies were undertaken to address the objectives of the project.

Study 1: Review of studies developing condition-specific preference-based measures

A six-stage approach to developing CSPBMs was used to structure the review: to establish the dimensionality (I), select items within each dimension (II), test the number of levels (III), validate the health-state classification on an independent sample (for those based on existing condition-specific measures) (IV), valuation survey (V) and modelling of the valuation data (VI). The aim of the review was to identify and appraise existing methods for deriving CSPBMs based on these six stages.

Methods

Current methods for developing CSPBMs were identified from searches of electronic databases. Paper title and abstracts were sifted using agreed exclusion criteria to identify papers for reading in full. Data were extracted on each paper and a narrative review undertaken to examine the methods used to derive health-state classification systems either from existing measures or 'de novo' and the methods of valuation (including modelling the health-state values).

Results

A total of 26 papers revealed a wide range of methods to develop health states from the condition-specific measures and methods of valuation. Around half of the measures were developed from existing condition-specific measures. A substantial proportion did not adequately report on the methods used and many failed to validate the classification system. Some CSPBMs were found to suffer from a narrow scope, focusing on symptoms rather than HRQoL, and this raises problems of unidimensionality addressed in study 2. This narrowness also raises issues about the likely impact of side effects and comorbidities that are explored in study 4.

Study 2: Developing a methodology for deriving measures with a unidimensional component: the Rasch vignette approach

A problem encountered in the development of CSPBMs is a lack of independence between dimensions. This study reports on a new approach that uses Rasch analysis to develop health states from the Clinical Outcomes in Routine Evaluation – Outcome Measure (CORE-OM), a 34-item instrument monitoring clinical outcomes of people with common mental health problems.

Methods

The CORE-OM is characterised by high correlation across its domains. Rasch analysis was used to reduce the number of items and response levels to produce a health-state classification system for valuation. Rasch analysis was used to generate a credible set of health states corresponding to different levels of symptom severity using the Rasch item threshold map. An interview valuation survey was undertaken using the time trade-off (TTO) technique to value the sample of health states. Regression analysis was applied to estimate health-state values for all states.

Results

The CORE-6D was developed – a two-dimensional health-state classification system consisting of a unidimensional five-item emotional component (derived from Rasch analysis) and a physical health dimension. Inspection of the Rasch item threshold map of the emotional component helped identify plausible 'emotional' health states, and these were combined with the response levels of the physical health dimension for valuation. A total of 220 respondents to the valuation survey provided 1496 health-state values. Multivariate regression models were used to predict values for all CORE-6D states using the Rasch logit value of the emotional health-state and the response level of the physical health dimensions as independent variables.

Study 3: The impact of labelling on health-state values

Many descriptions of health used in vignettes and condition-specific measures name the medical condition. This study assessed the impact of referring to the medical condition in the descriptions of health states valued by members of the general population.

Methods

An interview valuation study was conducted using TTO. All respondents valued essentially the same health states, but for each respondent descriptions featured either no label, an irritable bowel syndrome (IBS) label or a cancer label. Random effects generalised-least-squares regressions were used to estimate the impact of each label and experience of the condition on health-state values.

Results

No significant difference was found between health-state values when the description contains no label or an IBS label. The inclusion of a cancer label in health-state descriptions affected health-state values and the impact was dependent on the severity of the state, with a significant reduction in values for more severe health states but no significant difference for mild states. Without qualitative research the reason why values differed for states with the cancer label cannot be determined.

Study 4: Adaptation of condition-specific measures to examine the impact of side effects and comorbidities on condition-specific preference-based measures

Condition-specific preference-based measures are often criticised for their inability to capture comorbidities and side effects. Excluded dimensions may impact on health-state values directly via their own decrement or indirectly by interacting with other dimensions. This study examined these potential effects by adding an extra dimension to two CSPBMs.

Methods

First, using the results of study 2, a physical health dimension was added to the emotional component of the CORE-6D. Values of 18 CORE-6D states with a physical dimension were compared with four states containing only the five emotional domains. Second, a pain/discomfort dimension was added to the AQL-5D (asthma-specific CSPBMs) to create the AQL-6D. States for valuation were sampled using an orthogonal array designed to estimate an additive model using regression methods to estimate the coefficients of the dimensions. Out of these states, four were matched states that differed only in the additional dimension. Interview valuation studies were conducted using TTO on general population samples in which respondents valued a selection of health states defined by one CSPBM.

Results

The addition of the extra generic dimension at the worst level reduced health-state values for both CSPBMs. However, the addition of the generic dimension at intermediate or lowest levels increased health-state values. Modelling of the AQL-6D values to produce utilities for all states found the additional pain dimension had a significant and relatively large coefficient and impacted significantly on the coefficients of the other dimensions, but the degree of impact differed by dimension (largest changes for shortness of breath and activities) and severity level. These results suggest that preference weights for extra dimensions added to existing preference-based measures cannot necessarily be treated as simply additive to the existing preference weights for the original dimensions.

Study 5: Performance of condition-specific preference-based measures in comparison with the original measure and generic preference-based measures

This study addressed two questions: (1) How do the CSPBMs compare with the original non-preference-based measure used to derive them in terms of psychometric performance of validity and responsiveness to change?; and (2) Do CSPBMs offer an improvement over existing generic preference-based measures in terms of these psychometric properties?

Methods

The study compared EQ-5D and SF-6D with the condition-specific AQL-5D (asthma), CORE-6D (common mental health problems), EORTC-8D (cancer) and the OAB-5D (overactive bladder) across nine data sets. The analyses focused on validity, measured in terms of the extent to which measures were able to reflect known group differences, and responsiveness to change before and after treatment. These were assessed in terms of statistical significance and effect sizes (mean differences or changes divided by the standard deviation for baseline of change, respectively). For economic evaluation it is the agreement in absolute values that matters most and these were compared across the generic preference-based measures and CSPBMs in terms of mean values and intraclass correlation.

Results

There was little evidence of information loss from moving from the original condition-specific measure to the CSPBMs derived from them across the four conditions (asthma, common mental health problems, cancer and overactive bladder). The performance of the CSPBMs compared with generic preference-based measures was similar as regards responsiveness in capturing change following treatment, but CSPBMs were better at discriminating between groups with different severity. Although the benefits of CSPBMs over generic preference-based measures may not be as marked as expected, effect sizes were larger, which is important for trials and for the uncertainty in the values they generate. The larger effect sizes were due to smaller standard deviations, as mean change and differences were larger for the EQ-5D than for the CSPBMs. The large mean change and standard deviation of EQ-5D may be due to the UK value set used here. Ceiling effects were lower for the CSPBMs than for the EQ-5D, suggesting greater responsiveness for respondents at the upper end of HRQoL.

Conclusions

This project has outlined the six stages of developing CSPBMs and reviewed the range of methods used. It also built on this literature by offering a new approach to developing preference-based measures from existing instruments with high correlations across domains.

There are now more than 20 CSPBMs, but there remain some fundamental concerns about using them in economic evaluations comparing interventions in different conditions and programmes of care. It has been argued that the only way to achieve cross-programme comparability is to use the same generic preference-based measures in all studies. Comparability is important to policy-makers such as the National Institute for Health and Clinical Excellence (NICE) and is one reason why NICE has expressed a preference for the EQ-5D. The argument against relying on one measure is that EQ-5D (or whatever instrument is chosen) may not be available in the relevant studies (e.g. pivotal trials or other studies used to populate economic models) or may not be appropriate for the condition or patient group.

An argument in favour of using CSPBMs is that comparability can be achieved by using a common numéraire, such as a year in full health, provided that the values are obtained using the same valuation technique, with the same tightly controlled protocol, common anchors and the same type of respondents (e.g. general population). This would imply that there is no need to have a common classification system in order to achieve consistency in decision-making. However, there are a number of obstacles to achieving comparability, even if these requirements are met, arising from using different classification systems, including the problem of naming the condition, the exclusion of side effects and comorbidities, focusing effects and the lack of a common anchor.

A condition label can affect health-state values, but this is dependent on the specific condition and severity. We recommend avoiding condition labels in health-state descriptions or CSPBMs (where possible) to ensure that values are not affected by prior knowledge or preconception of the condition that may distort the health-state being valued.

Comparability between measures requires that the impact of different dimensions on preferences is additive, whether or not they are included in the classification system. For example, the impact of breathlessness on health-state values should be the same whether or not the patient has other problems not covered by the classification system, such as joint pain. In this way an intervention for asthma on health-state values can be estimated without regard to comorbidities. Likewise, the impact of side effects can be estimated separately from the CSPBMs and simply added or subtracted in the cost-effectiveness model as required. Our results cast doubt on this assumption, implying that the selected measure in a trial, for example, should contain all important and relevant dimensions in its classification system. This poses a considerable challenge for all measures, as both known and unknown comorbidities impact on health. Our research suggests that respondents to valuation surveys make assumptions about the excluded dimensions and so, when intermediate or mild levels of an additional dimension are added to severe health states, the value increases. The assumptions being made by respondents may not be appropriate for the population to which the values are going to be applied.

Whether or not a reduction in comparability should be accepted depends on the extent of any gain in validity and responsiveness arising from the use of CSPBMs. The performance of CSPBMs is better than or similar to that of generic preference-based measures in terms of discriminative validity across severity groups and responsiveness to change following treatment in four conditions. The performance of CSPBMs is similar to that of the measure from which they are derived, suggesting that CSPBMs based on existing condition-specific measures are likely to offer an improvement over generic preference-based measures only if the original condition-specific measure offers an improvement on the generic preference-based measures. The development of CSPBMs from existing measures for use in economic evaluation should be limited to measures that have been shown to offer an improved performance compared with generic preference-based measures, typically where the generic measure is inappropriate. There might also be a case for developing CSPBMs de novo and so avoiding the limitations that come from existing measures.

Condition-specific preference-based measures have an important role when generic measures are inappropriate for a given condition. Inappropriateness is difficult to prove in this area in the absence of a gold standard, but recent reviews would suggest there are some conditions for which generic measures are not sensitive to potentially important differences. In this case, CSPBMs have an important role to play in order to ensure that the benefits of health-care interventions are properly reflected in the QALY estimates for economic evaluation for all patient groups.

Future work recommendations

To meet the demand for CSPBMs, the following research is recommended.

To examine the appropriateness of generic preference-based measures in more conditions.

Further quantitative and qualitative work is required into the impact of, and reasons for, labelling effects.

The use of add-ons should be explored further for condition-specific measures (for side effects and comorbidities) and as a solution to the limitation of generic measures.

Finally CSPBMs should be systematically compared with generic measures in order to establish any advantages they may have the consequences of using them.

Funding

Publication

Brazier JE, Rowen D, Mavranezouli I, Tsuchiya A, Young T, Yang Y, *et al*. 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).

NIHR Health Technology Assessment programme

The Health Technology Assessment (HTA) programme, part of the National Institute for Health Research (NIHR), was set up in 1993. It produces high-quality research information on the effectiveness, costs and broader impact of health technologies for those who use, manage and provide care in the NHS. 'Health technologies' are broadly defined as all interventions used to promote health, prevent and treat disease, and improve rehabilitation and long-term care.

The research findings from the HTA programme directly influence decision-making bodies such as the National Institute for Health and Clinical Excellence (NICE) and the National Screening Committee (NSC). HTA findings also help to improve the quality of clinical practice in the NHS indirectly in that they form a key component of the 'National Knowledge Service'.

The HTA programme is needs led in that it fills gaps in the evidence needed by the NHS. There are three routes to the start of projects.

First is the commissioned route. Suggestions for research are actively sought from people working in the NHS, from the public and consumer groups and from professional bodies such as royal colleges and NHS trusts. These suggestions are carefully prioritised by panels of independent experts (including NHS service users). The HTA programme then commissions the research by competitive tender.

Second, the HTA programme provides grants for clinical trials for researchers who identify research questions. These are assessed for importance to patients and the NHS, and scientific rigour.

Third, through its Technology Assessment Report (TAR) call-off contract, the HTA programme commissions bespoke reports, principally for NICE, but also for other policy-makers. TARs bring together evidence on the value of specific technologies.

Some HTA research projects, including TARs, may take only months, others need several years. They can cost from as little as £40,000 to over £1 million, and may involve synthesising existing evidence, undertaking a trial, or other research collecting new data to answer a research problem.

The final reports from HTA projects are peer reviewed by a number of independent expert referees before publication in the widely read journal series *Health Technology Assessment*.

Criteria for inclusion in the HTA journal series

Reports are published in the HTA journal series if (1) they have resulted from work for the HTA programme, and (2) they are of a sufficiently high scientific quality as assessed by the referees and editors.

Reviews in *Health Technology Assessment* are termed 'systematic' when the account of the search, appraisal and synthesis methods (to minimise biases and random errors) would, in theory, permit the replication of the review by others.