Use of generic and condition-specific measures of health-related quality of life in NICE decision-making: a systematic review, statistical modelling and survey

Louise Longworth,[1]* Yaling Yang,[1] Tracey Young,[2] Brendan Mulhern,[2] Mónica Hernández Alava,[2] Clara Mukuria,[2] Donna Rowen,[2] Jonathan Tosh,[2] Aki Tsuchiya,[2] Pippa Evans,[2] Anju Devianee Keetharuth[2] and John Brazier[2]

[1]Health Economics Research Group, Brunel University, Uxbridge, Middlesex, UK
[2]School of Health and Related Research, University of Sheffield, Sheffield, UK

*Corresponding author

Declared competing interests of authors: LL, YY, AT and JB are members of the EuroQol Group (a non-profit making organisation responsible for the development of EQ-5D). LL is a member of the Board of the EuroQol Group and Brunel University receives reimbursement for her time. JB developed the SF-6D and the University of Sheffield receives reimbursement for its commercial use. LL is a member of the National Institute for Health and Care Excellence (NICE) Technology Appraisal Committee.

Published February 2014
DOI: 10.3310/hta18090

Scientific summary

Measures of health-related quality of life in NICE decision-making

Health Technology Assessment 2014; Vol. 18: No. 9
DOI: 10.3310/hta18090

NIHR Journals Library www.journalslibrary.nihr.ac.uk

Scientific summary

Background

Generic preference-based measures (GPBMs) of health-related quality of life (HRQL) are commonly used in the economic evaluation of health interventions. They provide a multidimensional description of health that is combined with survival to generate quality-adjusted life-years (QALYs). To enhance comparability, the National Institute for Health and Care Excellence (NICE) prefers the use of one of the GPBMs, EQ-5D, for measuring HRQL. This report addresses a number of important methodological issues arising from the use of GPBMs in NICE decision-making. It describes a series of studies undertaken to address the key questions of how to determine whether a GPBM is valid for use in calculating QALYs, what to do when the GPBM is not available (and specifically the use of 'mapping' or 'cross-walking' techniques to predict EQ-5D values) and what to do when the GPBM is found to miss important components of HRQL for a specific condition through the use of a new approach using 'bolt-on' dimensions.

Objectives

- To examine the appropriateness of three GPBMs of HRQL [EQ-5D, Health Utilities Index Mark 3 (HUI3) and SF-6D] for vision loss, hearing loss, skin disorders and cancer.
- To compare alternative methods for mapping from condition-specific or clinical measures onto EQ-5D, and to conduct exploratory analysis of the incorporation of uncertainty in the predicted estimates.
- To estimate mapping functions for use by researchers and policy-makers in conditions in which the EQ-5D has been found to be appropriate.
- To explore a new method for measuring HRQL in patient groups in which a generic measure has been shown to miss important dimensions ('bolt-ons').
- To estimate the impact of three 'bolt-on' dimensions on the value of EQ-5D health states.
- To estimate a new value set containing one of the EQ-5D bolt-ons and compare it with a value set without the EQ-5D bolt-ons.

Methods and results

Study 1: a systematic review of the performance of generic preference-based measures of health in four disease areas – visual disorders, hearing impairments, skin conditions and cancer

Methods

A systematic review of the literature was conducted for three GPBMs of HRQL: EQ-5D, HUI3 and SF-6D. Search strategies included free text and controlled terms. The following electronic databases were searched: BIOSIS (1969 to 2010), Cumulative Index to Nursing and Allied Health (CINAHL) (1982 to 2010), Cochrane Library comprising the Cochrane Database of Systematic Reviews (CDSR), Cochrane Central Register of Controlled Trials (CENTRAL), Cochrane Methodology Register, NHS Economic Evaluations Database (NHS EED) (1991 to 2010), EMBASE (1980 to 2010), MEDLINE (in process and non-indexed to 2010), PsycINFO (1806 to 2010) and Web of Science (1900 to 2010). Relevant websites were also searched. For inclusion, the studies had to report dimensions and/or index values and another measure of HRQL or clinical severity to allow an assessment of validity. Searching was completed in August 2010.

Performance was assessed in terms of (1) *construct validity*, the extent to which the measure differentiated between groups defined according to severity (*known group*) or a weaker test of differences between

people with and without the condition (*case–control*); (2) *convergent validity*, the strength of association between the EQ-5D and other measures of HRQL or clinical severity assessed using correlation coefficients or statistical significance and regression methods; (3) *responsiveness*, the extent (size and statistical significance) to which EQ-5D shows change where change has been observed using other HRQL or clinical measures; and (4) *reliability*, the extent to which the EQ-5D shows no change where no change in health has been observed using other measures.

Results

Visual disorders

Most of the 31 studies considered in this review found a worsening of utility values as visual impairment increases. Most evidence was found for the EQ-5D. Nearly all studies found significant differences between patients with the condition and a control group without it. Studies comparing EQ-5D scores across severity groups were more mixed, with most finding little or no difference between groups defined by clinical measures of visual impairment. No studies reported evidence on reliability for any of the measures. Three studies only allowed assessment of responsiveness and these identified changes consistent with an effective intervention, but differences were statistically significant in only two of three studies. The assessment of convergent validity was more concerning, with several studies not demonstrating a statistically significant correlation with clinical measures. While there was less evidence for the HUI3, all but one study demonstrated good validity and no studies assessed responsiveness. There was very limited evidence on the SF-6D.

Hearing impairment

Of the 18 studies found in the review, the HUI3 was the most commonly used measure. In all six cases that used the HUI3, this measure detected differences between groups defined by their severity and statistically significant changes were detected in five out of six cases as a result of intervention. Differences picked up by the HUI3 were driven by the hearing dimensions, and, in some cases, the speech and emotion dimensions. The findings suggested relatively poor responsiveness of EQ-5D in this condition as in four out of five cases it failed to detect change. A study suggested it only had weak ability to discriminate differences between severity groups. Only one study involved the SF-6D; thus, the information is too limited to conclude on its performance. No studies reported evidence to allow an assessment of reliability for any of the measures.

Skin diseases

Out of the 16 papers found, there was evidence to suggest the EQ-5D has good construct and convergent validity and responsiveness in skin disorders. All six studies reporting data for groups defined according to severity showed EQ-5D was able to reflect differences between groups and only one was not significant. EQ-5D was able to significantly differentiate patient and general populations in four case–control studies (one study did not report statistical tests), as well as groups defined by non-severity. Moderate to strong correlations were found between EQ-5D and other measures. Nine out of ten studies demonstrated that the EQ-5D measure was able to detect change appropriately over time, and, among them, only one study was not statistically significant. Most of the studies included patients with psoriasis or psoriatic arthritis. No studies reported evidence for HUI3 and SF-6D, and no studies reported evidence on reliability for any of the measures.

Cancer

Ninety-eight studies were found across 20 different types of cancer. Most evidence was found for the EQ-5D and the results were, overall, satisfactory. The majority of studies found significant differences in EQ-5D values between patients with various cancers and a control group. In most cases, the EQ-5D differentiated between severity groups, although the differences were not always statistically significant. Correlations between EQ-5D and other measures were mixed. In terms of responsiveness, overall EQ-5D

scores or dimensions were able to detect appropriate change over time points, but sometimes the change in scores was small or not statistically significant. Evidence on the performance of EQ-5D varied in different types of cancer. There was some limited evidence of reliability for the EQ-5D, but most studies had not been specifically designed to assess reliability. There was evidence to support the ability of the HUI3 to differentiate between severity groups and between patients with or without cancer. The responsiveness of the HUI3 was also found to be satisfactory but evidence of reliability was mainly limited to assessments of inter-rater reliability. Few studies reported evidence to allow a judgement to be made on the validity, reliability or responsiveness of the SF-6D.

Study 2: mapping from cancer-specific measures to EQ-5D – a comparison of methods

Methods

The aims of this study were to estimate mapping functions from two cancer-specific HRQL measures, the European Organisation for Research and Treatment of Cancer Quality-of-life Questionnaire Core 30 (EORTC QLQ-C30) and Functional Assessment of Cancer Therapy – General Scale (FACT-G), for estimating EQ-5D and to test the applicability of different mapping approaches that have been used in the literature. In particular, the analysis aimed to provide comprehensive information on how to select the mapping function and incorporate information on uncertainties around the predictions. Ordinary least squares (OLS), tobit model, two-part models (TPMs), splining models and response mapping models were used and an illustrative analysis using a limited dependent mixture model for a selected FACT-G model was also conducted. We used a range of criteria to identify the most appropriate mapping functions including mean absolute error (MAE), severity groups and shrinkage. Analysis for the FACT-G instrument was based on 530 patients with various cancers and the EORTC QLQ-C30 was based on 771 patients with multiple myeloma (MM), breast cancer and lung cancer.

Results

The mean observed EQ-5D value for the FACT-G data set was 0.722 [standard deviation (SD) = 0.224], ranging from –0.135 to 1, with 17% of participants reporting full health. For the sample with EORTC QLQ-C30 data, the mean, range and per cent in full health was 0.57 (SD = 0.35), –0.594 to 1 and 11% respectively.

Based on the range of criteria used, response mapping using all the domain scores was the best-performing model for the EORTC QLQ-C30. This was followed by OLS and tobit model, both of which were based on significant item-level models. Results for the FACT-G showed OLS gave the best predictions, followed by tobit model, with both based on item-level models. Response mapping and TPMs gave the poorest predictions. The limited dependent variable mixture model (LDVMM) performed better than an equivalent linear model in an exploratory analysis.

Generally, both OLS and tobit models using item levels gave some of the best estimates for EORTC QLQ-C30 and, for FACT-G, produced the most reliable models. Response mapping worked best for the EORTC QLQ-C30 functions but did not perform well for the FACT-G. This is because the FACT-G data set did not cover the full range of severity on both the EQ-5D scale and FACT-G; therefore, the mapping functions for this measure should be used only in non-severe populations.

Different selection methods for choosing the best model are currently used in mapping studies and can result in selecting different models therefore a range of criteria should be considered. We used criteria that were common across the different modelling techniques to select the best models. Further work is required on the most appropriate criteria to use in model selection.

Study 3: a new approach to dealing with inappropriateness – developing 'bolt-on' items to EQ-5D

Study 3a: testing the impact of three 'bolt-ons' to the EQ-5D methods

Three 'bolt-on' dimensions were developed following the systematic review of the performance of the EQ-5D. Two were developed in conditions in which EQ-5D was shown to be problematic: hearing and vision. A third was developed in fatigue, since this has been raised as a problem area in cancer (although, overall, EQ-5D was found to be satisfactory for cases of cancer). The description of levels follows the approach used for EQ-5D ('no problems' as level 1, 'some problems' as level 2 and 'extreme problems' as level 3). Three core EQ-5D health states were selected for valuation covering a range of severity: a mild state, a moderate state and a severe state. To each of these states, three levels of the extra dimension (with severity levels of 1, 2 or 3) were added, resulting in nine EQ-5D states for each bolt-on. The three core EQ-5D states without the bolt-ons were also valued, plus another six EQ-5D states. A valuation survey was undertaken using a sample of the general public in South Yorkshire, UK, using face-to-face interviews and the time trade-off (TTO) method. Individuals were allocated into four groups – three groups each valued one of the bolt-on variants and one group valued EQ-5D with no bolt-ons.

Mean values for each bolt-on health state were compared with the corresponding core EQ-5D state using paired *t*-tests. Regression analyses were used to further examine whether any differences between the groups could explain any potential differences between the values for the bolt-on states. Random effects (RE) models were used to take account of the clustering of data by respondents.

Results

Three hundred interviews were successfully completed, evenly split ($n = 75$) across three groups valuing each of the three bolt-ons and a group valuing EQ-5D alone. The characteristics of the groups were well balanced with the exception of fewer people in the group allocated to valuing the EQ + vision reporting current problems with vision.

Each of the bolt-on items had a significant impact on at least one EQ-5D health state. The extent and direction of the impact of the bolt-on varied according to the severity of the bolt-on and the state to which it was added. Adding a level 1 bolt-on to a mild state had no impact, but adding more severe levels led to lower values. Adding a level 1 or 2 bolt-on to the moderate state led to higher values, but was only statistically significant for the level 1 hearing bolt-on. Adding a level 3 bolt-on to the moderate state led to statistically significant lower values for the vision bolt-on. Adding a level 1 or 2 to the severe state has little impact or increased the health state values, though not significantly. Adding level 3 to the severe state reduced the value, but not significantly. The severe state had the highest SDs associated with the mean values and so the comparisons had the lowest power. The regression analysis confirmed that the differences in characteristics did not have a significant impact upon the valuations.

Study 3b: estimating the impact of a vision bolt-on to EQ-5D valuation model

Methods

The aim was to examine the impact of the vision bolt-on on EQ-5D health state values and the overall model parameters. A valuation study was undertaken using face-to-face interviews to obtain TTO values from members of the general public in South Yorkshire, UK. Half of respondents valued health states described using the EQ-5D plus vision bolt-on (EQ-5D + vision), and for comparative purposes, half of respondents valued EQ-5D states without the bolt-on. An orthogonal design of a six-dimension three-level instrument included 18 states, most of which were severe. Starting from these, 20 health states each for EQ-5D + vision and EQ-5D were selected for valuation, including two mild states. The set of EQ-5D states consisted of the same EQ-5D + vision states but without the vision bolt-on item. Two RE models were estimated for both instruments separately. TTO values were regressed on dimension or level models and coefficients for each of the five EQ-5D dimensions were compared for the two models using *z*-values.

Results

Three hundred people completed the interviews and 3120 TTO values were obtained. The two groups valuing EQ-5D and EQ-5D + vision were comparable in terms of age, gender, education, and health status. The results indicate that the inclusion of a vision bolt-on has a statistically significant impact on the valuation of EQ-5D health states. As with the exploratory analysis, the results suggest a somewhat complex relationship between the bolt-on and EQ-5D. Health states with a level 3 (extreme) vision problems included are unsurprisingly lower than the corresponding EQ-5D health state; however, the values given to severe EQ-5D states are higher if 'no problems' on vision are explicitly mentioned (EQ + vision) compared with if vision is not mentioned at all (EQ-5D only). There was also a suggestion that the coefficients on usual activity and anxiety and depression dimensions were lower with the introduction of the vision bolt-on; however, this difference did not quite reach the 5% level of significance.

Conclusion

This report has presented three substantial pieces of research.

The reviews of performance of the GPBMs were limited by the amount of evidence available, particularly for HUI3 and SF-6D. It is also difficult to prove the validity or otherwise of EQ-5D given the absence of a gold standard. However, the systematic review established that EQ-5D was a valid and responsive method for cases of cancer and some skin conditions, performance varied according to aetiology for vision, and performance was poor for hearing disorders. The HUI3 performed well for hearing and vision disorders and it also performed well in cases of cancer, although evidence was limited and there was no evidence for skin-related conditions. There were limited data for the SF-6D in all four conditions. There was very little evidence on reliability of all the instruments in all four conditions.

Mapping algorithms were estimated to predict EQ-5D values from alternative cancer-specific measures of health (FACT-G and EORTC QLQ-C30). While some differences were found in performance between models examined and some models did perform noticeably better across most criteria, conclusions about the best method are hard to draw owing to small sample sizes and the limited coverage of the patient groups. Further work is needed to determine the most important criteria for model selection. Ideally, all the mapping functions would be estimated in bigger data sets spanning the full spectrum of disease and then validated against an external, but similar, sample. Such data sets were not available for us to conduct this analysis but would be useful for further research.

The exploratory valuation study found that bolt-on items for vision, hearing and tiredness significantly impacted on values of the health states. The direction and magnitude of differences depended on the severity of the health state. A full model to obtain values for all EQ-5D + vision health states was estimated. The vision bolt-on item had a statistically significant impact on EQ-5D health state values, but the impact was not simply additive. The results from the vision study suggest that it may be necessary to estimate new models for some bolt-ons where there is an impact on the coefficients of the five core dimensions. The development of bolt-ons is a significant development for researchers and policy-makers using GPBMs in their evaluations. A proliferation of bolt-ons could be problematic if they reduce lead to many different value sets and the research to develop them is not conducted appropriately. However, bolt-ons could be very useful by improving on the performance of EQ-5D in specific conditions where there may be specific concerns.

Recommendations for further research

- Extend the reviews of the psychometric literature to more conditions.
- Undertake more primary research or analyses of primary data sets into the psychometric properties of GPBM particularly in cancer.
- Compare alternative statistical models in larger data sets, including those for EORTC QLQ-C30 and FACT-G.
- Develop a systematic programme of research into bolt-ons for EQ-5D.

Funding

This project was funded by the UK Medical Research Council (MRC) as part of the MRC-NIHR methodology research programme (ref: G0901486) and will be published in full in *Health Technology Assessment*; Vol. 18, No. 9. See the NIHR Journals Library website for further project information.

Health Technology Assessment

HTA MRP–MRC

ISSN 1366-5278 (Print)

ISSN 2046-4924 (Online)

Five-year impact factor: 5.804

Health Technology Assessment is indexed in MEDLINE, CINAHL, EMBASE, The Cochrane Library and the ISI Science Citation Index and is assessed for inclusion in the Database of Abstracts of Reviews of Effects.

This journal is a member of and subscribes to the principles of the Committee on Publication Ethics (COPE) (www.publicationethics.org/).

Editorial contact: nihredit@southampton.ac.uk

The full HTA archive is freely available to view online at www.journalslibrary.nihr.ac.uk/hta. Print-on-demand copies can be purchased from the report pages of the NIHR Journals Library website: www.journalslibrary.nihr.ac.uk

Criteria for inclusion in the *Health Technology Assessment* journal

Reports are published in *Health Technology Assessment* (HTA) if (1) they have resulted from work for the HTA programme or, commissioned/managed through the Methodology research programme (MRP), and (2) they are of a sufficiently high scientific quality as assessed by the reviewers 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.

HTA programme

The 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 journal is indexed in NHS Evidence via its abstracts included in MEDLINE and its Technology Assessment Reports inform National Institute for Health and Care Excellence (NICE) guidance. HTA research is also an important source of evidence for National Screening Committee (NSC) policy decisions.

For more information about the HTA programme please visit the website: www.hta.ac.uk/

This report

This issue of the Health Technology Assessment journal series contains a project commissioned/managed by the Methodology research programme (MRP). The Medical Research Council (MRC) is working with NIHR to deliver the single joint health strategy and the MRP was launched in 2008 as part of the delivery model. MRC is lead funding partner for MRP and part of this programme is the joint MRC–NIHR funding panel 'The Methodology Research Programme Panel'.

To strengthen the evidence base for health research, the MRP oversees and implements the evolving strategy for high quality methodological research. In addition to the MRC and NIHR funding partners, the MRP takes into account the needs of other stakeholders including the devolved administrations, industry R&D, and regulatory/advisory agencies and other public bodies. The MRP funds investigator-led and needs-led research proposals from across the UK. In addition to the standard MRC and RCUK terms and conditions, projects commissioned/managed by the MRP are expected to provide a detailed report on the research findings and may publish the findings in the HTA journal, if supported by NIHR funds.

The authors have been wholly responsible for all data collection, analysis and interpretation, and for writing up their work. The HTA editors and publisher have tried to ensure the accuracy of the authors' report and would like to thank the reviewers for their constructive comments on the draft document. However, they do not accept liability for damages or losses arising from material published in this report.

This report presents independent research funded by the National Institute for Health Research (NIHR). The views and opinions expressed by authors in this publication are those of the authors and do not necessarily reflect those of the NHS, the NIHR, the MRC, NETSCC, the HTA programme or the Department of Health. If there are verbatim quotations included in this publication the views and opinions expressed by the interviewees are those of the interviewees and do not necessarily reflect those of the authors, those of the NHS, the NIHR, NETSCC, the HTA programme or the Department of Health.

Published by the NIHR Journals Library (www.journalslibrary.nihr.ac.uk), produced by Prepress Projects Ltd, Perth, Scotland (www.prepress-projects.co.uk).