HEALTH TECHNOLOGY ASSESSMENT

VOLUME 18 ISSUE 9 FEBRUARY 2014
ISSN 1366-5278

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, Yaling Yang, Tracey Young, Brendan Mulhern, Mónica Hernández Alava, Clara Mukuria, Donna Rowen, Jonathan Tosh, Aki Tsuchiya, Pippa Evans, Anju Devianee Keetharuth and John Brazier

NHS
National Institute for
Health Research

DOI 10.3310/hta18090

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

This report should be referenced as follows:

Longworth L, Yang Y, Young T, Mulhern B, Hernández Alava M, Mukuria C, *et al.* Use of generic and condition-specific measures of health-related quality of life in NICE decision-making: a systematic review, statistical modelling and survey. *Health Technol Assess* 2014;**18**(9).

Health Technology Assessment is indexed and abstracted in *Index Medicus*/MEDLINE, *Excerpta Medica*/EMBASE, *Science Citation Index Expanded* (SciSearch®) and *Current Contents*®/ Clinical Medicine.

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).

Editor-in-Chief of *Health Technology Assessment* and NIHR Journals Library

NIHR Journals Library Editors

Abstract

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

Background: The National Institute for Health and Care Excellence recommends the use of generic preference-based measures (GPBMs) of health for its Health Technology Assessments (HTAs). However, these data may not be available or appropriate for all health conditions.

Objectives: To determine whether GPBMs are appropriate for some key conditions and to explore alternative methods of utility estimation when data from GPBMs are unavailable or inappropriate.

Design: The project was conducted in three stages: (1) A systematic review of the psychometric properties of three commonly used GPBMs [EQ-5D, SF-6D and Health Utilities Index Mark 3 (HUI3)] in four broadly defined conditions: visual impairment, hearing impairment, cancer and skin conditions. (2) Potential modelling approaches to 'map' EQ-5D values from condition-specific and clinical measures of health [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)] are compared for predictive ability and goodness of fit using two separate data sets. (3) Three potential extensions to the EQ-5D are developed as 'bolt-on' items relating to hearing, tiredness and vision. They are valued using the time trade-off method. A second valuation study is conducted to fully value the EQ-5D with and without the vision bolt-on item in an additional sample of 300 people.

Setting: The valuation surveys were conducted using face-to-face interviews in the respondents' homes.

Participants: Two representative samples of the UK general population from Yorkshire ($n = 600$).

Interventions: None.

Main outcome measures: Comparisons of EQ-5D, SF-6D and HUI3 in four conditions with various generic and condition-specific measures. Mapping functions were estimated between EORTC QLQ-C30 and FACT-G with EQ-5D. Three bolt-ons to the EQ-5D were developed: EQ + hearing/vision/tiredness. A full valuation study was conducted for the EQ + vision.

Results: (1) EQ-5D was valid and responsive for skin conditions and most cancers; in vision, its performance varied according to aetiology; and performance was poor for hearing impairments. The HUI3 performed well for hearing and vision disorders. It also performed well in cancers although evidence was limited and there was no evidence in skin conditions. There were limited data for SF-6D in all four conditions and limited evidence on reliability of all instruments. (2) Mapping algorithms were estimated to predict EQ-5D values from alternative cancer-specific measures of health. Response mapping using all the domain scores was the best performing model for the EORTC QLQ-C30. In an exploratory analysis, a limited dependent variable mixture model performed better than an equivalent linear model. In the full analysis for the FACT-G, linear regression using ordinary least squares gave the best predictions followed by the tobit model. (3) The exploratory valuation study found that bolt-on items for vision, hearing and tiredness had a significant impact on values of the health states, but the direction and magnitude of differences depended on the severity of the health state. The vision bolt-on item had a statistically significant impact on EQ-5D health state values and a full valuation model was estimated.

Conclusions: EQ-5D performs well in studies of cancer and skin conditions. Mapping techniques provide a solution to predict EQ-5D values where EQ-5D has not been administered. For conditions where EQ-5D was found to be inappropriate, including some vision disorders and for hearing, bolt-ons provide a promising solution. More primary research into the psychometric properties of the generic preference-based measures is required, particularly in cancer and for the assessment of reliability. Further research is needed for the development and valuation of bolt-ons to EQ-5D.

Funding: This project was funded by the UK Medical Research Council (MRC) as part of the MRC-NIHR methodology research programme (reference 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.

Contents

List of tables

List of figures

List of boxes

List of abbreviations

AIC	Akaike information criterion
AMD	age-related macular degeneration
AML	acute myeloid leukaemia
ANOVA	analysis of variance
BDI-SF	Beck Depression Inventory – short form
BIC	Bayesian information criterion
CHQ	child health questionnaire
CINAHL	Cumulative Index to Nursing and Allied Health
CLAD	censored least absolute deviation
DLQI	Dermatology Life Quality Index
ECOG	Eastern Co-operative Oncology Group
EORTC QLQ-C30	European Organisation for Research and Treatment of Cancer Quality-of-life Questionnaire Core 30
EORTC QLQ-C38	European Organisation for Research and Treatment of Cancer Quality-of-life Questionnaire Core 38
EQ-VAS	EuroQol visual analogue scale
ES	effect size
FACT-An	Functional Assessment of Cancer Therapy – Anaemia
FACT-C	Functional Assessment of Cancer Therapy – Colorectal subscale
FACT-F	Functional Assessment of Cancer Therapy – Fatigue Module
FACT-G	Functional Assessment of Cancer Therapy – General Scale
FACT-N	Functional Assessment of Cancer Therapy – Neutropenia
FAI	Frenchay Activities Index
FLIC	Functional Living Index – Cancer
GPBM	Generic preference-based measure
HADS	Hospital Anxiety and Depression Scale
HAQ	Health Assessment Questionnaire
HAQ-DI	Health Assessment Questionnaire Disability Index
HRQL	health-related quality of life
HTA	Health Technology Assessment
HUI1	Health Utilities Index Mark 1
HUI2	Health Utilities Index Mark 2
HUI3	Health Utilities Index Mark 3
LDVMM	Limited Dependent Variable Mixture Model
MAE	mean absolute error
ML	malignant lymphoma
MM	multiple myeloma
MRC	Medical Research Council
MVH	Measurement and Valuation of Health
MYCaW	Measure Yourself Concerns and Well-being questionnaire
NAPSI	Nail Psoriasis Severity Index
NICE	National Institute for Health and Care Excellence
OLS	ordinary least squares
PASI	Psoriasis Area Severity Index
PCQ	Psychological Consequences Questionnaire
PedsQL	Paediatric Quality-of-Life Inventory
PsAQoL	Psoriatic Arthritis Quality-of-life scale
PTA	pure-tone average
QALY	quality-adjusted life-year
QoL	quality of life
QWB	Quality of Well-Being scale
RCT	randomised controlled trial
RE	random effects
RESET	Regression Equation Specification Error Test
RMSE	root-mean-square error

RSCL	Rotterdam Symptom Checklist
SAPASI	self-administered PASI
$S\beta_2M$	serum beta-2-microglobulin
SD	standard deviation
SE	standard error
SF-12	Short Form questionnaire-12 dimensions
SF-36	Short Form questionnaire-36 dimensions
SF-MPQ	Short Form McGill pain questionnaire
SG	standard gamble
STAI	State-Trait Anxiety Inventory
TNM	tumour node metastasis
TPM	two-part model
TTO	time trade-off
VA	visual acuity
VAS	visual analogue scale
VF-14	Visual Function Questionnaire (14 item)
VF-4D	Visual Function Questionnaire (4 dimension)
VFA	Visual Function Assessment
VFQ-20/25	Visual Function Questionnaire-20/25
VISTA	Velcade as Initial Standard Therapy

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.

Chapter 1 Introduction

This report addresses a range of important methodological issues arising from the use of generic and condition specific measures of health-related quality of life (HRQL) in the decision-making of the National Institute for Health and Care Excellence (NICE). It describes a series of studies undertaken to address the key questions of how to determine whether a generic measure of HRQL is valid for use in calculating quality-adjusted life-years (QALYs), what to do when the generic measure is not available (and specifically the use of 'mapping' techniques) and examines a new approach to dealing with situations where the generic measure is found to miss important components of HRQL for specific conditions (i.e. the use of 'bolt-on' dimensions). The rest of this chapter describes the rationale for looking at these questions and presents the key objectives of the research.

Background

Generic preference-based measures (GPBMs) of HRQL are commonly used in the economic evaluation of health interventions. These instruments have many advantages, including that they can incorporate the impact of treatment or ill health on a multidimensional scale and can be combined with data on survival in the form of QALYs. Furthermore, they facilitate comparisons between interventions and across conditions, which is important if there is a need for consistency in decision-making between interventions or if there is a need to compare with a common benchmark or cost-effectiveness threshold. The questionnaires can usually be easily administered to patients for self-completion and the data can incorporate a reflection of the value associated with different levels of health (usually based on values from members of the general population).

In the UK, NICE has specified that Health Technology Assessments (HTAs) submitted to its Technology Appraisal programme should be based on an incremental cost per QALY framework and recommends the use of the EQ-5D as the preferred GPBM.[1] The EQ-5D descriptive classification consists of five dimensions of health: mobility, self-care, usual activities, anxiety/depression, and pain/discomfort.[2] In the older and most commonly used version, each dimension of health has three levels of severity; however, a new five-level version has recently been published.[3] The 3-level version can describe 243 unique health states, to which a preference value can be assigned based on a set of values obtained from a large UK general population survey.[4]

The decision by NICE to recommend the EQ-5D was, in part, a pragmatic decision.[5] It is now widely recognised that the various GPBMs produce different values,[6–8] and this can be problematic for an organisation wanting to make consistent, transparent and predictable decisions. The GPBMs, including EQ-5D, have been criticised for being insensitive or failing to capture important aspects of health.[9,10] While NICE recommends the use of the EQ-5D for its HTAs, in its *Guide to the methods of technology appraisal*,[1] it recognises that the EQ-5D may not be an appropriate measure for all conditions.[1] NICE requests evidence to show that EQ-5D is inappropriate for the condition of interest; however, it does not specify areas where EQ-5D is inappropriate, nor does it provide criteria to determine when a measure is appropriate for a particular condition or treatment.

The first section of this report will describe a systematic assessment of the appropriateness of the EQ-5D and other commonly used GPBMs in four broadly defined health conditions using the criteria of reliability, validity and responsiveness. This assessment uses established psychometric methods but is complicated by the absence of a gold standard measure of HRQL with which to compare the GPBMs. It is not possible to definitively determine whether the generic measures are inappropriate; it still requires an element of judgement. A generic measure may legitimately show no overall change in HRQL in contrast with a disease-specific measure because they are measuring different constructs. For example, a condition-specific measure may show improvements in some symptoms, but the overall impact on HRQL may be weakened as a result of new symptoms or side effects from treatment. However, judgements can be made transparently and systematically based on the totality of the evidence available. The reviews presented here draw on published research and established psychometric methods to establish the performance of the GPBMs.

In addition to acknowledging that the EQ-5D may not always be appropriate, the NICE *Guide to the methods of technology appraisal*[1] also acknowledges that EQ-5D data may not always be available. This may be for a variety of reasons, such as planning the economic evaluation after the trial design, concerns about obtaining data directly from patients and concerns about the views of regulators regarding non-significant differences in HRQL between treatments. In these circumstances NICE suggests incorporating data from other measures of health through the use of 'mapping'. 'Mapping' (sometimes referred to as 'cross-walking') describes a method by which values obtained from GPBMs, such as EQ-5D, can be predicted from other measures or indicators of health.[11,12] No specific guidance is provided on the best methods of mapping other than to state that it must be based on empirical analysis and the methods must be clearly described. In 2013, recommendations on the use of mapping were described;[13] however, these acknowledge that there is limited evidence to provide clear guidelines on many aspects of mapping, in particular the most appropriate model specifications. A recent review of mapping functions showed use of a range of different models including linear models, tobit models, censored least absolute deviation (CLAD), two-part models (TPMs) and response mapping to predict quality of life (QoL).[11] Studies also report a variety of methods to assess model and predictive performance including predicted mean and standard deviation (SD), median, Akaike information criterion (AIC), Bayesian information criterion (BIC), R^2, pseudo-R^2, mean estimates across severity groups, root-mean-square error (RMSE) and mean square error. A further issue in mapping is uncertainty, which is typically ignored. There is uncertainty in utility measure weights, the mapping coefficients, the choice of coefficients and the choice of model and these have not been addressed in the literature.

The second section of this report aims to establish the most appropriate model specifications for mapping based on two separate data sets. The analysis draws on the results of the systematic reviews reported in *Chapter 2* and focuses on conditions where the EQ-5D measure has been found to be appropriate based on the published evidence. An exploratory analysis demonstrates how the uncertainty in the estimates can be better incorporated into analyses.

The third section of the report examines an alternative method for dealing with the situations when the EQ-5D has been demonstrated to be inappropriate for a given condition owing to insensitivity or failing to cover an important dimension of HRQL. One option could be to use alternative GPBMs, but, as discussed above, this leads to a lack of comparability in the estimates compared with the standard EQ-5D approach and also may not cover missing dimension(s). Recently, there has been growing interest to explore an alternative approach by developing preference-based measures from existing and validated condition-specific measures of HRQL (for a full review of this approach, see the HTA monograph by Brazier *et al.*[14] and for recent examples, see papers by Yang *et al.*[15,16]). This approach can offer a useful solution in some situations. There have, however, been concerns raised that these condition-specific preference-based measures also produce very different values to the GPBMs and so may compromise comparability[17] and these differences may continue to arise even when the methods of valuation are designed to be similar with GPBMs.[14]

One possible solution to this problem is to not use comparable methods of valuation only, but also to keep the health state classification systems as similar as possible through the development of 'bolt-on' items to the EQ-5D or the GPBM of interest. Bolt-ons are dimensions that can be appended to another instrument and to which utility values can be attributed to the health states described by the instrument with the bolt-on. Previous research has examined the impact of modifying the EQ-5D descriptive system to include additional dimensions of health.[18,19] Krabbe *et al.*[18] valued EQ-5D health states including a 'cognition' dimension of health and found that it significantly impacted upon health state values.[18] More recently, Yang *et al.*[19] developed a 'sleep' dimension to add to the EQ-5D but found that it did not significantly impact on values.[19] The value of any potential 'bolt-on' dimension to EQ-5D depends crucially on whether its inclusion significantly impacts on the values given to the EQ-5D health states. The design and complexity of 'bolt-on' valuation studies will depend on how the values of the bolt-on levels are affected by the EQ-5D states accompanying it and whether the inclusion of the bolt-on items has a significant impact on the values given to the EQ-5D dimensions. Furthermore, the methods of bolt-on development and valuation are not well developed. Two studies are described in this report to develop potential bolt-ons to the EQ-5D, to quantify the impact they have on EQ-5D values and to assess the implications of this for future bolt-on developments. In undertaking this, a full valuation model is provided for one of the EQ-5D bolt-ons.

Aims and objectives of the report

The overall aim of the study was to develop methods for systematically incorporating information from condition-specific measures into the NICE decision-making framework. Specifically, the project had three related objectives:

1. To examine whether the EQ-5D and other commonly used generic HRQL measures are appropriate for use in calculating QALYs for NICE decision-making in selected specific conditions.
2. To develop mapping functions to predict EQ-5D data from condition-specific or clinical measures, to compare alternative model specifications and to conduct an exploratory analysis around the incorporation of uncertainty in the predicted estimates.
3. To investigate the development and valuation of bolt-ons to expand the EQ-5D descriptive system for those conditions in which the EQ-5D is not appropriate.

The results from the analysis to meet the first objective are used to inform the second and third objectives. Mapping will not be successful if the measure to be predicted does not adequately capture HRQL; therefore, only those conditions where the EQ-5D is found to be appropriate (objective 1) are considered to inform the mapping analyses (objective 2). Conversely, those conditions found to be not adequately captured by EQ-5D (objective 1) are the focus of the analyses of bolt-on measures (objective 3).

Chapter 2 A systematic review of the psychometric properties of generic preference-based measures of health in four conditions

Introduction

The aim of the review reported in this chapter was to assess the reliability, validity and responsiveness of the EQ-5D, Health Utilities Index Mark 3 (HUI3) and SF-6D for measuring HRQL in four broadly defined conditions: visual disorders, hearing disorders, skin conditions and cancer.

The three GPBMs focused on (EQ-5D, HUI3 and SF-6D) were chosen to represent commonly used GPBMs of HRQL in NICE Technology Appraisals.[20] Specifically, as noted previously, the EQ-5D is recommended as the preferred measure by NICE and is the most commonly used measure in its Technology Appraisals.[1,20] The HUI3 was chosen as it is commonly used internationally and is the second most frequently used in NICE Technology Appraisals.[20] The SF-6D was also chosen as it has properties considered important by NICE (as a validated and generic measure of HRQL that also has a set of UK general population values elicited using a choice based method). In addition, the SF-6D questionnaire was derived from the short form questionnaire-36 dimensions (SF-36), which is widely used in clinical trials.

The four conditions were chosen to represent areas where the EQ-5D measure may not be appropriate based on previous published research[21–24] or concerns reported during the development of NICE Technology Appraisals.[25,26] Previous research has reported that the generic instruments, particularly the EQ-5D, do not adequately capture changes in health as a result of visual or hearing loss, but findings are mixed.[21–23,24] In addition, the measurement of HRQL in these conditions has been the subject of debate within NICE Technology Appraisals of treatments for these conditions.[25,26] The appraisals of treatments for skin conditions by NICE have frequently relied upon data from condition-specific measures in analyses rather than directly using generic measures of HRQL. Finally, the condition for which treatments are most frequently appraised by NICE is cancer. There have been suggestions that generic measures, such as the EQ-5D, may not adequately reflect the effects of cancer and related treatments that are considered important to patients (e.g. fatigue); however, a comprehensive review of the evidence has not been previously reported. A similar review has been conducted to examine the appropriateness of the EQ-5D in mental health as part of another Medical Research Council (MRC) funded project.[27,28]

The rest of this chapter discusses the methods used for the systematic literature reviews, the findings and results for the four conditions, each discussed separately, and finally a brief discussion and conclusion is provided.

Methods

The generic preference-based measures

EQ-5D

The EQ-5D describes HRQL in terms of five dimensions: mobility, self-care, usual activities, pain/discomfort and anxiety/depression.[2] Each dimension is usually described in terms of three levels of severity, although a version with five levels has recently been published.[3] The health classification system for the three-level version describes 243 health states and a tariff of values for each health state is available for several countries, including the UK. The UK value set was obtained from valuations provided by 3395 members of the general population using the time trade-off (TTO) valuation method.[4,29]

SF-6D

Derived from the SF-36 and Short Form questionnaire-12 dimensions (SF-12) health questionnaires, the SF-6D has six dimensions (physical functioning, role limitation, social functioning, bodily pain, mental health and vitality) and each dimension has four to six severity levels.[6,30] Any patient who completes the SF-36 or the SF-12 can be uniquely classified according to the SF-6D. The health classification system of SF-6D describes a total of 18,000 health states and a tariff of values for each health state is available for several countries, including the UK. The UK value set was obtained from valuations provided by 611 members of the general population using the standard gamble (SG) valuation method.[30]

Health Utilities Index Mark 3

Health Utilities Index is a group of GPBMs for measuring comprehensive health status and HRQL, including Health Utilities Index Mark 1(HUI1), Health Utilities Index Mark 2 (HUI2) and HUI3. HUI3 has nine dimensions (vision, hearing, speech, ambulation/mobility, pain, dexterity, self-care, emotion and cognition) and each dimension has three to six levels. The health classification system of HUI3 describes almost a million unique health states and a tariff of values for each health state is available for Canada. The Canadian value set was obtained from valuations provided by 504 members of the general population using the visual analogue scale (VAS) and SG valuation methods.[31]

Search strategy and data identification

The search strategy aimed to identify relevant journal papers providing evidence on the reliability, validity and responsiveness of EQ-5D, HUI3 or SF-6D in the following four clinical conditions: vision disorders, hearing impairments, skin disorders and cancer.

Four separate search strategies were developed, one for each of the conditions. The search strategies were developed following consultation with experts in information resources and health economics. An iterative approach to the searches was adopted. The strategies consisted of a broad search to identify studies reporting the use of the GPBMs in patients with each of the four clinical conditions. The search included both free text and controlled terms. Free text words included 'euroqol', 'hui3', 'sf6d' (all with alternative spellings). Condition-specific terms were also included (see *Appendix 2* for the full searches used). 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).

In addition, a database of studies held on the website of the EuroQol Group[32] was searched to check for any missing papers reporting EQ-5D and to check that the search strategies were identifying relevant papers. Comparable databases for the SF-6D and HUI3 are not available. The search strategies are presented in *Appendix 2*.

The inclusion criteria were that (1) the study reported dimensions and/or index values for at least one of the generic instruments EQ-5D, HUI3 or SF-6D and (2) the study reported another measure of QoL [including VAS or EuroQol VAS (EQ-VAS), TTO, SG direct valuation of QoL or another utility measure] or a measure of clinical severity/symptoms that would enable an assessment of validity, responsiveness or reliability.

The condition-specific inclusion criteria were that the studies reported the above data for people with one of the following conditions: vision disorders, hearing disorders, skin disorders or cancer.

There was no restriction relating to the type of study or type of condition within the overall definitions. Owing to resource limitations, only English language studies were reviewed.

Data extraction

Data were extracted from the studies using a standardised set of forms developed for this study after reviewing forms used for similar studies in other disease areas.[27] The data extracted included general characteristics of the study and participants, instruments used in the study, methods and results used in the study for assessment of reliability, construct validity and responsiveness. Data extraction for the different clinical conditions was undertaken by one member of the research team and summarised using items presented in *Table 1*.

Data analysis

Assessment of quality and relevance

For the review, of most importance was the relevance of the study in terms of the patient population and inclusion of evidence to establish the psychometric performance of the generic measures. Studies including a mixed population of patients (i.e. with various conditions) were only included if they reported health-related utility values or dimension responses for subgroups of patients with one of the four specific conditions being evaluated. Nevertheless, a judgment regarding the risk of bias for each study was

TABLE 1 Information extracted from included papers

General	Author name, year
	Country where the study took place
	Type of disease/disorder
	Disease/treatment stage
	Treatment (if any)
	Study design
Participant characteristics	Number of participants
	Age (mean and range)
	Gender (percentage of males)
	Ethnicity
	Missing data, including reasons for non-completion if given
Valuation and descriptive methods	Descriptive systems
	Tariff or source of value sets
	Mean values (SD, range)
	Direct valuations used
	Condition-specific HRQL measures used
	Clinical measures used
	Qualitative questions asked
	Missing data of measures completion
Reliability	Methods
	Results
Validity	Methods
	Results
Responsiveness	Methods
	Results

determined by reviewing the methods of patient recruitment and noting any missing data reported (either study drop-outs or incomplete questionnaires). Studies were not required to be specifically designed to assess validity, responsiveness or reliability, provided that they reported data in sufficient detail to allow an assessment of these traits. The intention of the assessment of quality was not to exclude relevant studies, but to highlight any concerns about quality when findings were interpreted.

Assessment of reliability

The reliability of a measure is defined as its ability to reproduce results when measurements are repeated on an unchanged population,[33] or the comparability of responses across different assessors (for example, patient and proxy report). Reliability can be measured by retesting and reporting either the correlation or difference between estimates. In some circumstances, no change in health status may be expected over time and, subsequently, the values obtained using the measures may be stable. These results were interpreted as evidence of the reliability and stability of instruments. Other assessments of reliability included assessments of inter-rater reliability based on a comparison of responses given by multiple people completing the questionnaire on the patients' behalf. When considering the results of inter-rater comparisons, it is important to note that all of the GPBMs have been designed for self-completion and to report self-assessed HRQL. Therefore, perfect agreement between the intended respondents and their proxies may not be expected. Finally studies reporting internal consistency were also included as assessed through multitrait analysis.

Assessment of construct validity

Validity is defined as how well an instrument measures what it was intended to measure. More specifically, for the GPBMs, whether the dimensions adequately cover the key determinants of health-related utility. Criterion validity is determined by comparing an instrument to an established gold standard; however, a gold standard with which to benchmark HRQL measures against does not exist. Therefore, it is necessary to assess the validity of measures of health-related utility using measures that have evidence of construct validity for that condition, which establishes if patterns in scores confirm constructs or hypotheses about expected patterns.

We assessed the construct validity of the GPBMs using the ‘known-group’ method. The known-group method compares the values obtained from the GPBMs between groups of patients who are expected to differ [qualitatively or statistically using *t*-test or analysis of variance (ANOVA)] in the construct measured by the indicator used to define the groups. The known groups in this context are often defined according to clinical severity using other measures. It should be noted that the usefulness of these comparisons can be limited by sample size, particularly as studies are usually not powered to detect differences according to preference-based measures. In addition, consideration must be given to the appropriateness of the clinical measure and the groups defined by it, and exogenous factors that may influence HRQL. For instance, groups defined solely by the presence of a biomarker may have no impact on HRQL. If patients have a number of comorbidities, then these may have a greater impact on HRQL than the condition of interest. Known groups can also be defined using a case–control analysis in which comparison is between patient and general public population, or defined on the basis of other aspects such as age, gender or countries. However, a more stringent test is to define known groups based on different levels of condition severity (for example, by using a clinical indicator).

We also examined convergent validity, which is a type of construct validity. Convergent validity is defined as the extent to which one measure correlates with another measure of the same or similar concept. In this review, we examined the extent to which the EQ-5D, SF-6D or HUI3 correlate with other measures of QoL or clinical severity. Correlation was defined as ‘low’ if correlation coefficient was < 0.3, ‘moderate’ if between 0.3 and 0.5 and ‘strong’ if > 0.5. Correlations need to be interpreted with caution as it is not always clear how strong the relationship between the generic and condition-specific indicators should be. Furthermore, we interpreted estimation of regression between GPBMs and other measures as another indication of a correlation, focusing on whether some measures were significant predictors of others.

Assessment of responsiveness

Responsiveness assesses the ability of an instrument to measure a change in health-related utility over time. As with construct validity, the measurement of responsiveness is difficult as there is no gold standard measure with which to compare. Nevertheless, we assessed the responsiveness of health-related utility measures by comparing change in health-related utility measured over a period of time in which health status is expected to change (e.g. before and after an intervention) with the change demonstrated by another measure of health. For inclusion in the assessment of responsiveness, the comparator measure must have demonstrated a change in health. We did not review data from studies outside of the review relating to responsiveness of the comparator measures. Good evidence of responsiveness is considered where the GPBM shows statistically significant change in health (e.g. *t*-test) shown by other measures or clinical indicators. Weaker evidence of responsiveness is considered where the same trend of change is shown but the change is not statistically significant. When responsiveness indices for estimates of health-related utility are reported [e.g. effect size (ES) or standard response mean], they were compared with other measures. ES is the mean change in score of a measure between two different time points divided by the SD of the score at baseline. Standardised response mean is the mean change score of a measure between two different time points divided by the SD of the change score. As for the tests of validity, it is important to consider whether the measures of health change that are being used to assess responsiveness are valid. In addition, it is important to consider whether other health changes not directly related to the condition could have impacted upon health-related utility (e.g. side effects of treatment).

Presentation and analysis

Data for each of the four conditions are presented separately. Information on the study design, participant characteristics and the measures included are reported. Within each of the broadly defined conditions, there is a range of underlying aetiologies with different symptoms. The results for visual disorders, skin diseases and cancers are therefore presented for subgroups defined according to type of condition. Subgroups are not presented for hearing impairments as the studies were mainly defined according to the presence or absence of hearing loss and/or extent of hearing loss. For each condition, a summary table is presented which reports an overview of the conclusions drawn from each paper for each of the types of assessment.

Results

Vision

Search results: vision

Bibliographic searching was completed in August 2010 and total of 1025 potentially relevant papers were identified. Abstracts and titles for all papers were screened to identify papers meeting the inclusion criteria; 969 records were excluded and full papers were ordered for the remaining 56 records. After reviewing the full papers, 25 were excluded and a total of 31 papers were included in the review. A flow chart of the study selection process is shown in *Figure 1*.

Quality assessment: vision

A range of recruitment procedures were reported. Some were retrospective analyses of data sets with predetermined inclusion criteria,[34–36] some were case–control analyses,[37–39] and the majority were cross-sectional observational studies.[22,34,36,40–52] The only randomised controlled trial (RCT) had well-defined inclusion criteria.[53] Response rates for questionnaires ranged from 33% to 96%, with completion rates of longitudinal studies > 85% in all but one study[50] (range 52–98%). No study was excluded after the assessment of quality.

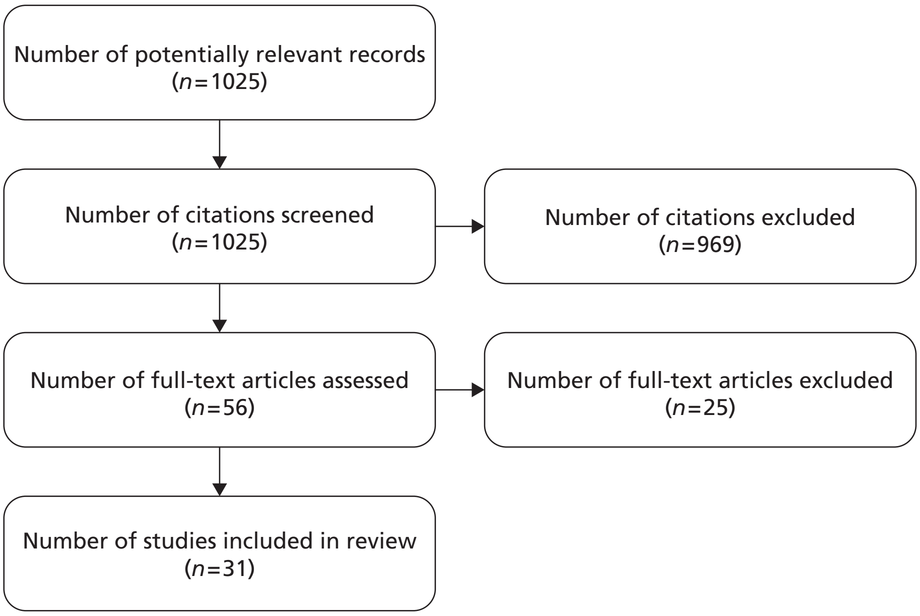
Number of potentially relevant records
(n=1025)
Number of citations screened
(n=1025)
Number of citations excluded
(n=969)
Number of full-text articles assessed
(n=56)
Number of full-text articles excluded
(n=25)
Number of studies included in review
(n=31)

FIGURE 1 Flow diagram showing selection of studies: vision.

Study design and patients' characteristics: vision

Summary characteristics of the 31 studies are presented in *Table 2*. Thirty of the 31 studies were observational studies[22,34,36–39,40–52,54–64] and the remaining study was a RCT.[53] The studies were conducted in different countries including the UK, the USA and Canada and some were multicountry studies. The studies identified included a wide range of visual disorders. Five studies were in patients with glaucoma,[34,44–46,54] seven studies were in patients with age-related macular degeneration (AMD),[22,43,47–49,55,56] eight studies included patients with cataracts,[36–39,53,57–59] two studies were on patients with diabetic retinopathy,[42,50] three were on patients with conjunctivitis[51,60,61] and the remaining studies included people with various other visual conditions.[40,41,52,62,63,64]

TABLE 2 Characteristics of included studies: visual disorders

Study reference grouped by condition (author, year)	Country	Disease/treatment stage	Sample size	Study type
Glaucoma				
Aspinall *et al.*, 2008[44]	UK	Glaucoma and no other ocular comorbidity	72	Cross-sectional
Kobelt *et al.*, 2006[45]	Sweden	Ocular hypertension or open-angle glaucoma	109	Cross-sectional
Mittmann *et al.*, 2001[34]	Canada	Glaucoma – a subset from a study on a range of chronic conditions	137	Cross-sectional
Montemayor *et al.*, 2001[46]	Canada	Chronic open-angle glaucoma, normal-pressure glaucoma or suspected glaucoma with treatment	224	Cross-sectional
Thygesen *et al.*, 2008[54]	Multiple	Late-stage primary open-angle glaucoma	162	Case review

TABLE 2 Characteristics of included studies: visual disorders (*continued*)

Study reference grouped by condition (author, year)	Country	Disease/treatment stage	Sample size	Study type
AMD				
Cruess *et al.*, 2007[47]	Canada	Neovascular AMD	67	Cross-sectional
Espallargues *et al.*, 2005[22]	UK	Wet or dry AMD	209	Cross-sectional
Kim *et al.*, 2010[55]	Korea	–	625	Cohort
Lotery *et al.*, 2007[48]	UK	Bilateral subfoveal neovascular-AMD	75	Cross-sectional
Payakachat *et al.*, 2009[49]	Multiple	Wet AMD	154	Cross-sectional
Ruiz-Moreno *et al.*, 2008[56]	Spain	Bilateral neovascular AMD	89	Prospective case–control
Soubrane *et al.*, 2007[43]	Multiple	Neovascular AMD	401	Cross-sectional
Cataracts				
Asakawa *et al.*, 2008[36]	Canada	With or without other comorbidities	911	Cross-sectional
Black *et al.*, 2009[57]	UK	First or second eye	860	Prospective cohort
Conner-Spady *et al.*, 2005[58]	Canada	–	253	Cohort
Datta *et al.*, 2008[53]	UK	Bilateral cataracts in participants over 70 years of age	289	Secondary analysis of RCT
Jayamanne *et al.*, 1999[59]	UK	First Eye	144	Prospective
Polack *et al.*, 2007[37]	Kenya	–	196	Case–control
Polack *et al.*, 2008[38]	Bangladesh	–	217	Case–control
Polack *et al.*, 2010[39]	Philippines	Participants over 50 years of age	401	Case–control
Diabetic retinopathy				
Lloyd *et al.*, 2008[42]	UK	Diabetic retinopathy due to diabetes	122	Cross-sectional
Smith *et al.*, 2008[50]	USA	Type 2 diabetes	401	Cross-sectional
Conjunctivitis				
Pitt *et al.*, 2004[60]	UK	–	310	Cohort
Rajagopalan *et al.*, 2005[51]	Multiple	Non-Sjögren’s keratoconjunctivitis or Sjögren’s syndrome	210	Cross-sectional
Smith *et al.*, 2005[61]	Spain	–	401	Cohort
Other visual disorders				
Boulton *et al.*, 2006[40]	UK	Vision impairment or blindness in children	100	Cross-sectional
Clark *et al.*, 2008[62]	Australia	Postcataract surgery endophthalmitis	49	Cohort
Kempen *et al.*, 2003[63]	USA	Cytomegalovirus retinitis in patients with acquired immunodeficiency syndrome	961	Prospective cohort
Langelaan *et al.*, 2007[41]	Netherlands	Low-vision patients	120	Cross-sectional
Quinn *et al.*, 2004[64]	USA	Retinopathy of prematurity	244	Cohort
van Nispen *et al.*, 2009[52]	Netherlands	Vision impairment in older people	296	Observational

The inclusion criteria varied across the studies reviewed within each of the specific conditions. Some studies reported that patients were identified through case notes, but no more details are provided. It was noted whether AMD was bilateral or unilateral and wet or dry, whether cataracts were present in the first or second eye and whether glaucoma was primary or multiple. Sample sizes also varied across studies, ranging from 49[62] to 961.[63] One study[40] included children with a mean age of 6 years and used HUI3. The authors reported that the HUI system had been used in a previous study of young children with a range of impairments similar to those included in their study, although it should be noted that this did not refer specifically to the HUI3 at that time. All other studies included adult patients and the AMD studies included patients over 70 years.

Measures used in studies: vision

Table 3 summarises the measures that have been used in the 31 studies included in the review. For the three GPBMs of interest, the EQ-5D was reported in 27 studies[22,37–39,41–63] and therefore was the most commonly utility measure, six studies reported the HUI3[22,34,36,40,42,64] and only one study reported the SF-6D.[22] Ten studies also reported direct valuations of patients' own health states using methods such as the TTO or VAS.[22,44,45,51,58–63] Twenty-three studies reported visual acuity (VA)[22,34,37–39,41–50,52–55,58,61,63,64] to indicate visual severity. In addition, various patient-reported visual-specific QoL measures were used.

Reliability: vision

No tests of reliability were performed on the generic preference-based measures.

Known-group analysis and convergent validity: vision

Known-group analysis was performed in 24 studies:[22,34,36–45,47–51,54,56,60–64] 20 for EQ-5D,[22,37–39,41–45,47–51,54,56,60–63] five for HUI3,[34,36,40,42,64] but no studies for SF-6D. In six of the studies, groups were defined by VA, or by contrast sensitivity, and mean estimates of utility for each defined group were provided.[22,41–43,54,61] The remaining 25 studies had either a case–control design, had different conditions or did not define levels of severity.

Nine of the 31 studies reviewed provide evidence on correlation or regression between GPBMs with either each other or with visual measures.[22,37,38,44,46,50,52–54] Eight studies report evidence of convergent validity in EQ-5D compared with a visual measure,[37,38,44,46,50,52–54] with Espallargues *et al.*[22] also reporting correlations across EQ-5D, SF-6D and HUI3. Details of the data are summarised in *Appendix 3* and below by type of vision disorder.

Glaucoma

Known-group analysis Three studies of people with glaucoma allowed a known-group analysis for EQ-5D where groups were defined by severity of vision problems.[44,45,54] The studies by Aspinall *et al.*[44] and Kobelt *et al.*[45] found that EQ-5D utility values decreased with increasing glaucomatous damage but were not statistically significant. The study by Thygesen *et al.*[54] defined three groups on the basis of the Snellen score and the ordering of mean utility values were consistent and statistically significant. No such data were available for HUI3 or SF-6D by severity groups. However, one paper reported HUI3 in a case–control study, which showed an appropriate and significant difference in HUI3 values between the cases and controls.[34]

Convergent validity Three studies reported correlation statistics for EQ-5D with VA in patients with glaucoma.[44,46,54] Aspinall *et al.*[44] reported moderate and statistically significant correlations for the EQ-5D measure and the mobility, self-care and anxiety dimensions. The study by Thygesen *et al.*[54] also showed a significant correlation between VA and EQ-5D. However, Montemayor *et al.*[46] reported low and non-significant correlations for EQ-5D with VA.

TABLE 3 Instruments used: vision

Study reference grouped by condition (author, year)	GPBM			Direct valuation	Rating scale	Condition specific HRQL instruments					Clinical severity
	EQ-5D	SF-6D	HUI3	TTO	VAS	VFQ-20/25	VF-14/4D	RQLQ	VFA	IDEEL	VA
Glaucoma											
Aspinall *et al.*, 2008[44]	✓			✓							✓
Kobelt *et al.*, 2006[45]	✓				✓						✓
Mittmann *et al.*, 2001[34]			✓								✓
Montemayor *et al.*, 2001[46]	✓								✓		✓
Thygesen *et al.*, 2008[54]	✓										✓
AMD											
Cruess *et al.*, 2007[47]	✓					✓					✓
Espallargues *et al.*, 2005[22]	✓	✓	✓	✓	✓		✓				✓
Kim *et al.*, 2010[55]	✓						✓				✓
Lotery *et al.*, 2007[48]	✓					✓					✓
Payakachat *et al.*, 2009[49]	✓					✓					✓
Ruiz-Moreno *et al.*, 2008[56]	✓					✓					
Soubrane *et al.*, 2007[43]	✓					✓					✓
Cataracts											
Asakawa *et al.*, 2008[36]			✓								
Black *et al.*, 2009[57]	✓						✓				
Conner-Spady *et al.*, 2005[58]	✓				✓				✓		✓
Datta *et al.*, 2008[53]	✓						✓				✓
Jayamanne *et al.*, 1999[59]	✓				✓						✓
Polack *et al.*, 2007[37]	✓					✓					✓
Polack *et al.*, 2008[38]	✓					✓					✓
Polack *et al.*, 2010[39]	✓					✓					✓

continued

TABLE 3 Instruments used: vision (*continued*)

Study reference grouped by condition (author, year)	GPBM			Direct valuation	Rating scale	Condition specific HRQL instruments					Clinical severity
	EQ-5D	SF-6D	HUI3	TTO	VAS	VFQ-20/25	VF-14/4D	RQLQ	VFA	IDEEL	VA
Diabetic retinopathy											
Lloyd *et al.*, 2008[42]	✓		✓			✓					✓
Smith *et al.*, 2008[50]	✓										✓
Conjunctivitis											
Pitt *et al.*, 2004[60]	✓				✓	✓		✓			
Rajagopalan *et al.*, 2005[51]	✓				✓					✓	
Smith *et al.*, 2005[61]	✓				✓	✓		✓			
Other visual disorders											
Boulton *et al.*, 2006[40]			✓								
Clark *et al.*, 2008[62]	✓			✓		✓					
Kempen *et al.*, 2003[63]	✓				✓						✓
Langelaan *et al.*, 2007[41]	✓										✓
Quinn *et al.*, 2004[64]			✓								✓
van Nispen *et al.*, 2009[52]	✓										✓

IDEEL, impact of dry eyes on everyday life questionnaire; RQLQ, rhinoconjunctivitis; VF-14/4D, Visual Function Questionnaire 14/4 dimension; VFA, Visual Function Assessment; VFQ-20/25, Visual Function Questionnaire – 20/25.

Age-related macular degeneration

Known-group analysis In studies of people with AMD, all seven[22,43,47–49,55,56] papers provided evidence to allow an assessment of construct validity of the EQ-5D. Of these, five[22,43,48,49,55] differentiated between groups based on severity of vision disorder and three[43,47,56] included assessments of cases against controls. Three studies defined visual severity groups: two[22,43] in terms of levels of VA and the other[55] based on whether they had unilateral or bilateral AMD. Soubrane *et al.*[43] showed inconsistency with the mean estimates, with normal VA having a worse mean utility when compared with mild, moderate, severe and near blind utility values. The anxiety dimension of the Hospital Anxiety and Depression Scale (HADS) was also inconsistent between the normal and mild VA groups, but this inconsistency was not shown in the Visual Function Questionnaire-25 (VFQ-25). The study did, however, report a significant difference of utility values between those with neovascular AMD and the control group. Kim *et al.*[55] found a statistically significant difference in EQ-5D values between those with unilateral and bilateral AMD. Espallargues *et al.*[22] found a consistent relationship between VA and contrast sensitivity with HUI3, SF-6D, TTO and VAS but not EQ-5D.

Of the three case–control studies, two found that EQ-5D showed an appropriate and statistically significant reduction in HRQL for people with AMD compared with general population controls.[43,56] One reported a difference that was not a statistically significant difference, but the difference was in the appropriate direction.[47]

Convergent validity Three studies provided correlation statistics between generic and visual measures in patients with AMD and all showed poor correlation of EQ-5D with other measures.[22,47,48] Espallergues *et al.*[22] found that the VAS, TTO, HUI3 and SF-6D were all significantly correlated with both VA and contrast sensitivity. However, they did not find significant correlations for EQ-5D with VA or contrast sensitivity.

Cataracts

Known-group analysis Four[36–39] of the seven[37–39,53,57–59] studies in patients with cataracts provided evidence to allow an assessment of the construct validity of the EQ-5D[37–39] and HUI3.[36] Three case–control studies conducted in different countries by Polack *et al.*[37–39] found that there were significant differences in EQ-5D between cases and controls, and found that cases were likely to report a significant difference across all dimensions (except pain dimension in Polack *et al.*[38]). However, Polack *et al.*[39] reported an inconsistent association between EQ-5D and VA.

One study reported HUI3 values for cases and controls and identified a statistically significant and appropriate difference between the two groups.[36]

Convergent validity Four studies provided evidence of the convergent validity of the EQ-5D with VA.[37–39,53] Polack *et al.*[37–39] tested associations between EQ-5D and VA, with one study finding that poorer VA was associated with higher odds of reporting any problem with all EQ-5D dimensions apart from anxiety.[37] The other two studies found no significant associations between VA and EQ-5D dimensions, apart from a borderline association with self-care.[38,39] Datta *et al.*[53] did not find significant correlations for EQ-5D with VA.

Diabetic retinopathy

Known-group analysis Two studies reported EQ-5D identifying a statistically significant difference between the two extreme groups; however, the differences between neighbouring groups were not significant and frequently inconsistent.[42,50] In the study by Lloyd *et al.*[42] the inconsistencies were also shown in VAS ratings of patients' own health and the HUI3. This may be the result of small sample size or, as the authors speculate, it may be the result of a loss of independence of the participants when they reach that level of severity.[42]

Convergent validity Smith *et al.*[50] fitted a linear regression and found visual angle to be a predictor of EQ-5D utility values. They also fitted a non-parametric ordinal logistic regression and this estimated that any degree of visual impairment would result in an increased likelihood of reporting non-perfect utility values.

Conjunctivitis

Known-group analysis All three studies allowed an assessment of construct validity of the EQ-5D in people with conjunctivitis. Two were case–control studies and showed a statistically significant difference between cases and controls.[60,61] One study demonstrated a difference between groups defined according to severity.[51] Within the dimensions of the EQ-5D, the study by Pitt *et al.*[60] found the pain dimension to be the only dimension to show a statistical difference. However, Smith *et al.*[61] reported a significant difference across all dimensions except mobility. No studies provided evidence on the construct validity of the HUI3 or SF-6D.

Convergent validity No papers reported on convergent validity of the measures in patients with conjunctivitis.

Other visual conditions

Known-group analysis The remaining six studies were in unique visual conditions.[40,41,52,62–64] Three of these studies allowed an assessment of the construct validity of the EQ-5D[41,62,63] and two of the HUI3.[40,64] Clark *et al.*[62] and Kempen *et al.*[63] reported an appropriate, but non-significant, difference in the EQ-5D between the control group and those with endophthalmitis and cytomegalovirus, respectively. Langelaan *et al.*[41] undertook a study on visually impaired patients and identified an appropriate, but non-significant, difference in the EQ-5D between low and high visual field groups, but an inconsistent and non-significant difference in the EQ-5D between low- and high-VA groups.

Boulton *et al.*[40] and Quinn *et al.*[64] found the HUI3 identified statistically significant and appropriate differences between groups of patients with unspecified blindness/visual impairment.

Convergent validity A study by van Nispen *et al.*[52] reported a multivariate regression analysis of data from older patients with visual impairment. They found that worsening VA was a significant risk factor for a lower EQ-5D value.

Responsiveness

Only three studies reported responsiveness of the utility measures in visual disorders (*Appendix 4*).[55,57,58] Kim *et al.*[55] reported a statistically significant improvement in both the Visual Function Questionnaire (4 dimension) (VF-4D) and the EQ-5D after photodynamic therapy in patients with AMD. Black *et al.*[57] reported a statistically significant improvement in both the Visual Function Questionnaire (14 item) (VF-14) and the EQ-5D postcataract surgery, although the latter was relatively small. Conner-Spady *et al.*[58] reported a statistically significant improvement in the Visual Function Assessment (VFA) and VA post cataract surgery, but the subsequent mean improvements in EQ-VAS and EQ-5D were small and not statistically significant. This may suggest that the EQ-5D is not responsive in this population; however, it should be recognised that the study was not initially powered to identify statistically significant changes and a mean improvement was identified. In addition, the VAS did not change from pre to post treatment; therefore, the treatment may not significantly impact on HRQL.

Summary of results for visual review

The 31 studies included in this review show a worsening of utility values as visual impairment increased in many though not all studies. The magnitude and statistical significance of the association varied between different GPBMs of HRQL. *Table 4* shows an overview of performance of utility measures in visual impairment.

The largest amount of evidence was found for the EQ-5D compared with the other generic measures and the results were mixed. Nearly all studies showed significant differences between patients with the condition and a control group. Studies comparing EQ-5D scores across severity groups were more mixed, with the majority of studies showing little or no difference between groups defined by clinical measures of visual impairment. No studies allowed an assessment of reliability for any of the measures. There were just three studies on responsiveness. and all were in the form of before-and-after studies of an intervention.[55,57,58] These identified changes consistent with an effective intervention, but differences were statistically significant in only two of three studies.[55,57] The assessment of convergent validity was also concerning, with half of the studies not demonstrating a statistically significant correlation with clinical measures. While there was less evidence for the HUI3, all but one study[42] demonstrated good validity; no studies assessed responsiveness. There was very limited evidence on the SF-6D in patients with visual impairment.

Hearing

Search results: hearing impairment

Bibliographic searching was completed in July 2010. The search strategy identified 119 articles. After reviewing titles and abstracts, 70 papers were excluded. Forty-nine papers were reviewed in full, and a further 31 were excluded and 18 papers were included in the final review. A flow chart of the study selection process is shown in *Figure 2*.

Quality assessment: hearing impairment

A range of study designs was reported in the studies included in the review. Three studies were cross-sectional[65–67] but the majority were prospective or retrospective before-and-after studies.[21,23,68–78] Studies had well-defined inclusion/exclusion criteria for recruitment. For longitudinal studies, no study had extremely high levels of missing data and completion rates for patients in studies ranged from 60%[68] to 100%.[66] The completion rates for the instruments included were usually high, ranging from 71%[67] to 97%.[23] The reporting in these papers was reasonably clear. After quality assessment, no studies were excluded from the review.

Study characteristics: hearing impairment

The main characteristics of the 18 papers included in this review are shown in *Table 5*. The two papers by Joore *et al.*[71,72] and the two papers by Joore[73,74] reported the results of one specific study and, similarly, the two papers by Vuorialho *et al.*[77,78] reported a single study. In total, 14 separate studies were included in the review. The studies were undertaken in a range of countries, including the UK, the Netherlands, the USA, Canada and Finland. Some studies recruited patients with specific hearing problems, e.g. large vestibular aqueduct syndrome,[23] but most were for defined the sample using clinical indicators such as the better ear unaided pure-tone average (PTA). As shown in *Table 5*, the level of hearing loss varied between studies.

The sample sizes of the studies reviewed ranged from 20[68] to 3272.[65] Most studies had approximately 100 participants, but two studies only had approximately 20 participants.[68,79] Five studies included young children with hearing impairments (the mean age of the samples ranged from 7 to 9 years),[66–68,76,80] and the remaining studies included adults, with most focusing on older adults over 60 years of age. The studies involving children used parents or caregivers as proxies to assess HRQL of children.

TABLE 4 Overall performances of EQ-5D, HUI3 and SF-6D in visual disorders

Study reference grouped by measure (author, year)	Conditions	Known group (severity)		Known group (case–control)		Known group (other)		Correlation	Responsiveness		Reliability
		Consistent evidence	Significant	Consistent evidence	Significant	Consistent evidence	Significant		Consistent evidence	Significant	
EQ-5D											
Aspinall *et al.*, 2008[44]	Glaucoma	✓	✗					Moderate			
Kobelt *et al.*, 2006[45]	Glaucoma	✓	✗								
Montemayor *et al.*, 2001[46]	Glaucoma							✓ (low)			
Thygesen *et al.*, 2008[54]	Glaucoma	✓	✓					✓			
Cruess *et al.*, 2007[47]	AMD			✓	✗			✗			
Lotery *et al.*, 2007[48]	AMD	✓	✓					✗			
Payakachat *et al.*, 2009[49]	AMD	✗	✗								
Ruiz-Moreno *et al.*, 2008[56]	AMD			✓	✓						
Soubrane *et al.*, 2007[43]	AMD	✗	✗	✓	✓						
Espallargues *et al.*, 2005[22]	AMD	✗	✗					Low			
Kim *et al.*, 2010[55]	AMD	✓	✓						✓	✓	
Datta *et al.*, 2008[53]	Cataracts							✗			
Polack *et al.*, 2007[37]	Cataracts			✓	✓			✓			
Polack *et al.*, 2008[38]	Cataracts			✓	✓			✗			
Polack *et al.*, 2010[39]	Cataracts			✓	✓			✗			
Conner-Spady *et al.*, 2005[58]	Cataracts								✓	✗	
Black *et al.*, 2009[57]	Cataracts								✓	✓	
Lloyd *et al.*, 2008[42]	Diabetic retinopathy	Mixed evidence	Mixed evidence								
Smith *et al.*, 2008[50]	Diabetic retinopathy	Mixed evidence	Mixed evidence					✓			

Study reference grouped by measure (author, year)	Conditions	Known group (severity)		Known group (case–control)		Known group (other)		Correlation	Responsiveness		Reliability
		Consistent evidence	Significant	Consistent evidence	Significant	Consistent evidence	Significant		Consistent evidence	Significant	
Pitt *et al.*, 2004[60]	Conjunctivitis			✓	✓						
Rajagopalan *et al.*, 2005[51]	Conjunctivitis	✓	✓								
Smith *et al.*, 2005[61]	Conjunctivitis			✓	✓						
Clark *et al.*, 2008[62]	Other			✓	✗						
Kempen *et al.*, 2003[63]	Other	✓	✗								
Langelaan *et al.*, 2007[41]	Other	Mixed evidence	Mixed evidence								
van Nispen *et al.*, 2009[52]	Other							✓			
HUI3											
Mittmann *et al.*, 2001[34]	Glaucoma			✓	✓						
Asakawa *et al.*, 2008[36]	Cataracts			✓	✓						
Lloyd *et al.*, 2008[42]	Diabetic retinopathy	✗	✗								
Boulton *et al.*, 2006[40]	Other	✓	✓								
Quinn *et al.*, 2004[64]	Other	✓	✓								
Espallargues *et al.*, 2005[22]	AMD							Low to moderate			
SF-6D											
Espallargues *et al.*, 2005[22]	AMD							Low to moderate			

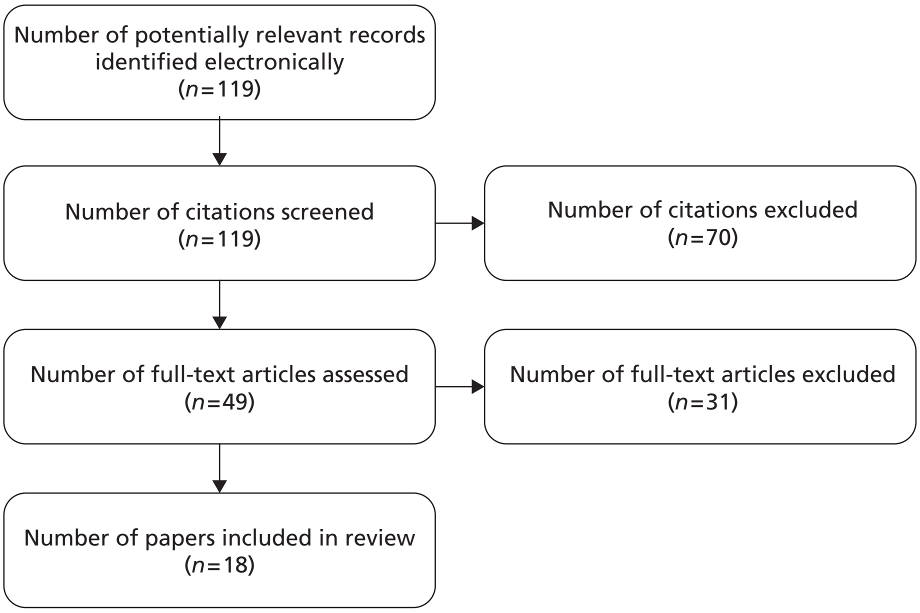
Number of potentially relevant records identified electronically (*n*=119)
Number of citations screened (*n*=119)
Number of citations excluded (*n*=70)
Number of full-text articles assessed (*n*=49)
Number of full-text articles excluded (*n*=31)
Number of papers included in review (*n*=18)

FIGURE 2 Flow diagram showing selection of studies: hearing impairment.

TABLE 5 Characteristics of the studies included in the review: hearing loss

Study reference (author, year)	Country	Hearing disorder	Treatments	Sample size (*n*)	Study design
Barton *et al.*, 2005[21]	UK	Hearing impaired	Hearing aid (analogue and digital signal-processing)	609	Prospective before-and-after
Barton *et al.*, 2006[65]	UK	Hearing impaired	Cochlear implant	3272	Cross-sectional
Damen *et al.*, 2007[69]	Netherlands	Postlingual deafness	Cochlear implant	83	Prospective before-and-after
Grutters *et al.*, 2007[23]	Netherlands	Hearing impaired	Hearing aid	337	Prospective before-and-after
Hol *et al.*, 2004[70]	Netherlands	Conductive or mixed hearing loss	Bone-anchored hearing aid	56	Prospective before-and-after
Joore *et al.*, 2002,[71] 2002,[74] 2003,[72] 2003[73]	Netherlands	First-time hearing-aid users	Hearing aid	126	Prospective before-and-after
Palmer *et al.*, 1999[75]	Canada and USA	Severe to profound hearing impaired	Cochlear implant	62	Prospective before-and-after
Vuorialho *et al.* 2006,[77] 2006[78]	Finland	First-time hearing aid user over 60	Hearing aid	101	Prospective before-and-after
Lee *et al.*, 2006[79]	South Korea	Postlingual deafness	Cochlear implant	26	Retrospective before-and-after
Bichey *et al.*, 2002[68]	USA	Large vestibular aqueduct syndrome	Cochlear implant and hearing aid	20	Retrospective before-and-after
Cheng *et al.*, 2000[80]	USA	Profoundly deaf	Cochlear implant	140	Retrospective
Sach and Barton, 2007[76]	UK	Hearing impaired children	Unilateral cochlear implant	222	Retrospective before-and-after
Lovett *et al.*, 2010[66]	UK	Profoundly deaf	Cochlear implant (bilateral and unilateral)	50	Cross-sectional
Smith-Olinde *et al.*, 2008[67]	USA	Permanent childhood hearing loss	Cochlear implant	146	Cross-sectional

Measures: hearing impairment

Table 6 summarises the measures used in the 18 papers.[21,23,65–80] Eleven papers reported EQ-5D,[21,23,70–74,76–79] 10 reported HUI3[21,23,65–69,75,79,80] and one used the SF-6D[21] (alongside EQ-5D and HUI3). Among those studies that used EQ-5D, most reported the EQ-5D index based on the tariff of UK population values. In two cases, it was unclear which tariff of population values had been used.[71,77] Three papers also reported responses on the EQ-5D dimensions alongside the utility values.[72–74] A total of 11 papers reported patients' rating of own health using VAS[66,70–74,76–80] and two used TTO methods.[79,80] A total of seven studies employed self-reported hearing-specific HRQL measures[66,69–71,74,77,78] and seven studies reported clinical indicators to indicate severity of hearing impairment,[23,65,67–69,75,77] including PTA for the best or worst ear without hearing aid and speech identification tests.

Reliability: hearing impairment

The review found little evidence on the reliability assessments of EQ-5D, HUI3 and SF-6D in hearing impairment. No papers reported test–retest experiments. Although not specifically for test–retest reliability purposes, one study[71] reported EQ-5D responses and VAS indices at baseline and asked respondents to recall them 3 months after a hearing aid fitting. The authors did not find any significant difference between the baseline assessment and the recalled assessment of baseline health for EQ-5D.

Known-group analysis and convergent validity

Out of the 18 papers included in the review, seven papers provided information to enable an assessment of the validity of EQ-5D, HUI3 or SF-6D,[23,65–68,75,76] although most studies were not designed to examine the validity of these measures.[23,65–68,75,76] The results are summarised in *Appendix 5*.

Known-group analysis

Seven studies presented data to allow an assessment of known-group differences of HUI3 and EQ-5D where the groups were defined by the severity of hearing loss.

Assessment for EQ-5D Using ANOVA, the study by Grutters *et al.*[23] demonstrated that EQ-5D failed to detect significant differences by hearing loss severity whereas HUI3 showed a difference. Sach and Barton[76] found that EQ-5D differentiated the group with the most severe hearing loss but not groups defined by milder levels of deafness.

Assessment for Health Utilities Index Mark 3 Barton *et al.*[65] reported that HUI3 mean scores were different between moderate, severe, profound and implanted groups but no statistical test was reported. Palmer *et al.*[75] showed that HUI3 showed significant difference between people with and without hearing aids at two follow-up time points. Similarly, HUI3 discriminated two groups of patients with cochlear implant and with normal hearing aids where the hearing loss of these two groups was different according to their PTA.[68] In a study comparing HUI3 and the quality of well-being scale (QWB) in hearing loss, both scores declined with the degree of hearing loss for children who did not have a cochlear implant with a much greater extent for HUI3 than QWB.[67] Another study found that the HUI3 differentiated between groups defined according to unilateral or bilateral implantation but this was not significant as suggested by the speech measure.[66] However, this finding was also reflected in the VAS measure and might reflect that the additional impact of bilateral implantation in this group and the sample size was small.

Convergent validity

Four studies presented data for an assessment of convergent validity of EQ-5D and HUI3.[21,23,65,69] HUI3 showed moderate correlation with two speech perception tests, which was consistent with a hearing specific QoL measure that also showed similar results.[69] Barton *et al.*[65] reported a regression analysis and showed that for cochlear implant (grouped by age at implantation and duration of use), the average of pure-tone air-conduction thresholds at different frequencies in the better hearing ear and gender were significant predictors of HUI3 in a large cross-sectional study.[65] Grutters *et al.*[23] reported a moderate correlation between EQ-5D and HUI3 and Barton *et al.*[21] reported strong correlations between EQ-5D, HUI3 and SF-6D in their study.

TABLE 6 Measures reported in the papers: hearing loss

Study reference (author, year)	GPBMs			Direct valuation	Rating		
	EQ-5D	HUI3	SF-6D	TTO	VAS	Hearing-specific measures	Clinical indicators
Barton *et al.*, 2005[21]	✓	✓	✓				
Barton *et al.*, 2006[65]		✓					AHL
Grutters *et al.*, 2007[23]	✓	✓					BEPTA
Lee *et al.*, 2006[79]	✓	✓		✓	✓		
Bichey *et al.*, 2002[68]		✓					PTA
[a]Cheng *et al.*, 2000[80]		✓		✓	✓		
Damen *et al.*, 2007[69]		✓				NCIQ	NVA and AN test
Lovett *et al.*, 2010[66]		✓			✓	SSQ	
Palmer *et al.*, 1999[75]		✓					NU-6; audiological mean score for CID sentence recognition
Smith-Olinde *et al.*, 2008[67]		✓					BEPTA
Hol *et al.*, 2004[70]	✓				EQ-VAS	HHDI	
Joore *et al.*, 2002[71]	Index and responses				VAS and EQ-VAS	ADPI	
Joore *et al.*, 2003[72]	Index and responses				VAS and EQ-VAS		
Vuorialho *et al.*, 2006[77]	Index and responses				✓	HHIE-S	BEHL, SRT, WRS
Joore *et al.*, 2003[73]	✓				VAS and EQ-VAS		
Joore *et al.*, 2002[74]	✓				VAS and EQ-VAS	HHIE-S and hearing aid satisfaction/use	
Sach and Barton, 2007[76]	✓				EQ-VAS and QoL VAS		
Vuorialho *et al.*, 2006b[78]	✓				EQ-VAS	HHIE-S, hearing aid satisfaction	

ADPI, audiological disabilities preference index; AHL, average hearing level; AN test, Antwerp–Nijmegen hearing test; BEHL, better ear hearing levels; BEPTA, better ear PTA; CID, central institute for the deaf; HHDI, hearing handicap and disability index; HHIE–S, Hearing Handicap Inventory for the Elderly – Screening; NCIQ, the Nijmegen cochlear implant questionnaire; NU-6, Northwestern University 6-word test; NVA test, Dutch Audiological Society open speech recognition test; SRT, speech reception thresholds; SSQ, speech, spatial and qualities of hearing scale for parents; WRS, word reception scores.

a Parents were used as proxies.

Responsiveness

Twelve papers[21,23,66,69–72,74,77–80] involved a total of nine studies that provided adequate information to allow an assessment of responsiveness of EQ-5D, HUI3 and/or SF-6D (see *Appendix 6*).

Assessment of EQ-5D

Six studies reported evidence to assess the responsiveness of EQ-5D.[21,23,70–72,74,77–79] In most of these studies, no statistically significant changes before and after the hearing intervention were detected[23,70–74,77,78] and the ES where reported were very low. However, for these studies, statistically significant improvements were shown in VAS scores or condition-specific measures or SF-36 social functioning domain.

Assessment of Health Utilities Index Mark 3 Six studies reported the responsiveness of HUI3.[21,23,66,69,79,80] Grutters *et al.*[23] found that HUI2 and HUI3 detected statistically significant change after cochlear implant fitting. The study by Lee *et al.*[77] demonstrated that the increases in EQ-5D, VAS, HUI3 and QWB scores following cochlear implantation were all statistically significant. The results suggest that the EQ-5D was responsive in capturing larger improvements in hearing, as in the study by Lee *et al.*,[79] but was not able to capture the smaller levels of improvement shown in the study by Grutters *et al.*[23]

Cheng *et al.*[80] found that the change in HUI3 overall score was higher than the change in both VAS and TTO scores after cochlear implant fitting, but all changes were statistically significant. Only the change in scores on the hearing and speech dimensions of HUI3 were significant and the change score was greatest for the hearing dimension, while scores on other dimensions were stable over time. Moderate correlations between the change scores of VAS, TTO and HUI3 were found.

Assessment of SF-6D Barton *et al.*[21] detected statistically significant differences ($p < 0.001$) between the changes in HUI3 and EQ-5D values and between the changes in HUI3 and SF-6D values, but not between the changes in EQ-5D and SF-6D values.

Summary and conclusion

Overall, the HUI3 was the most commonly used measure in the studies. In all six cases,[23,65–68,75] the HUI3 detected a difference between groups defined by their severity of hearing impairment and four[23,68,78,79] out of five[23,66,69,79,80] cases detected statistically significant changes as a result of intervention (*Table 7*). Differences picked up by the HUI3 were driven by the hearing dimensions and, in some cases, the speech and emotion dimensions. On the other hand, the findings of the review suggested relatively poor responsiveness of EQ-5D in this condition as, in five[23,70–72,74,77,78] out of six cases,[23,70–72,74,77–79] EQ-5D failed to detect change. The studies that allowed an assessment of known groups using the EQ-5D suggested it had only weak ability to discriminate difference between severity groups. Only one study involved the SF-6D; thus, the information is too limited to conclude on its performance.[21] No studies allowed an assessment of reliability to be made.

Skin conditions

Search results: skin conditions

The bibliographic search was completed in September 2010. The search of electronic databases identified 161 records and two additional records were identified from the EuroQol Group website database. After reviewing titles and abstracts, 122 records were excluded. Forty-one papers were reviewed in full: a further 25 papers were excluded and 16 papers were included in the final review (*Figure 3*).

Quality assessment: skin conditions

Three types of study designs were observed in the review. Eleven studies were RCTs,[81–91] four studies were cross-sectional[92–95] and one was an uncontrolled before-and-after study.[96] The majority of studies provided clear inclusion and exclusion criteria, but two did not.[81,82] Six papers did not report completion rates[82,83,87,88,92,96] and, among the 10 studies reporting this information, completion rates were reasonable or high (ranging from 70%[84] to 97%).[94] The completion rates for specific measures

TABLE 7 The overall performance of EQ-5D, HUI3 and SF-6D in studies of hearing impairment

Study reference grouped by measure (author, year)	Known group (severity)		Known group (case–control)		Known group (other)		Correlation	Responsiveness		Reliability
	Consistent evidence	Significant	Consistent evidence	Significant	Consistent evidence	Significant		Consistent evidence	Significant	
EQ-5D										
Grutters *et al.*, 2007[23]					✓	✓	Moderate	✗	✗	
Sach and Barton, 2007[76]	Mixed evidence	Mixed evidence			✓	✓				
Lee *et al.*, 2006[79]								✓	✓	
Hol *et al.*, 2004[70]								✗	✗	
Joore *et al.*, 2002a, 2002b, 2003a[71,72,74]								Mixed evidence	✗	
Vuorialho *et al.*, 2006a, 2006b[77,78]								✗	✗	
Barton *et al.*, 2005[21]								✓	✗	
HUI 3										
Barton *et al.*, 2006[65]	✓	✓								
Bichey *et al.*, 2002[68]	✓	N/R								
Damen *et al.*, 2007[69]							Moderate	✓	✓	
Grutters *et al.*, 2007[23]	✓	✓					Moderate	✓	✓	
Lovett *et al.*, 2010[66]	✓	✗						✓	✗	
Palmer *et al.*, 1999[75]	✓	✓								
Smith-Olinde *et al.*, 2008[67]	✓	N/R								
Lee *et al.*, 2006[79]								✓	✓	
Cheng *et al.*, 2000[80]								✓	✓	
Barton *et al.*, 2005[21]								✓	✓	
SF-6D										
Barton *et al.*, 2005[21]							Strong	✓	✓	

N/R, not reported.

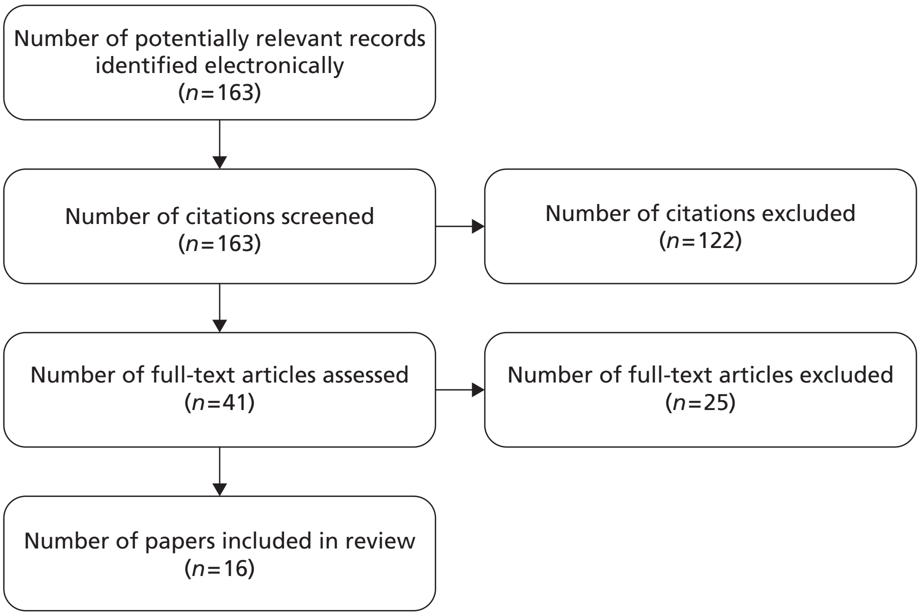
Number of potentially relevant records identified electronically (n=163)
Number of citations screened (n=163)
Number of citations excluded (n=122)
Number of full-text articles assessed (n=41)
Number of full-text articles excluded (n=25)
Number of papers included in review (n=16)

FIGURE 3 Flow diagram showing selection of studies for skin review.

(e.g. item non-response) were generally high (above 90%).[82,91,92,95] No study was excluded after the assessment of quality.

Study design and patients' characteristics: skin conditions

The main characteristics of the 16 papers included in this review are shown in *Table 8*.[81–96] Studies were conducted in various European and American countries, with several multinational studies. All but four studies recruited patients with psoriasis or psoriatic arthritis;[82–88,92–96] the remaining studies recruited patients with acne,[81] eczema,[90] hidradenitis suppurativa[89] or venous leg ulcers.[91] All studies included adults (mean age approximately 43 years), and male respondents accounted for 24–71% of the samples. Sample sizes ranged from 32[91] to 27,994,[95] with most studies including between 100 and 200 participants.

Measures used in studies: skin diseases

Table 9 summarises the measures that have been used in the 16 studies included in the review. Of the three GPBMs of interest, only those studies reporting EQ-5D were identified and included in the review. No studies reported data from SF-6D or HUI3. Fourteen studies also reported patients' valuation of their own health states using VAS.[81,82,84–89,90–92,94–96] Clinical indices were reported in studies to indicate severity of skin problems, including the Psoriasis Area Severity Index (PASI) by eight studies,[85–88,92,94–96] Nail Psoriasis Severity Index (NAPSI) by one study,[96] and the Acne Grade by one study.[81] Various generic measures [e.g. SF-36, Health Assessment Questionnaire – Disability Index (HAQ-DI), Health Assessment Questionnaire (HAQ)], skin-specific HRQL measures [e.g. Dermatology Life Quality Index (DLQI)], or symptom-specific HRQL measures (e.g. HADS, the Depression Inventory) were included in the studies (see *Table 9*).

Reliability: skin conditions

No study reported data on reliability of the three GPBMs.

Known-group analysis and convergent validity: skin conditions

Thirteen studies of patients with skin conditions provided sufficient evidence to allow assessment of known-group analysis and convergent validity of EQ-5D[81–85,88–93,95,96] including: 12 known-group analyses[82–86,88–93,96] and seven convergent validity analyses.[83,87,89–92,95] A summary of the findings is presented below. See *Appendix 7* for details.

TABLE 8 Characteristics of studies included: skin diseases

Study reference grouped by condition (author, year)	Country	Treatment	Sample size	Study type
Plaque psoriasis and psoriatic arthritis				
Bansback *et al.*, 2006[83]	UK	Methotrexate with and without ciclosporin A	72	RCT
Brodszky *et al.*, 2010[92]	Hungary	None	183	Cross-sectional
Christophers *et al.*, 2010[93]	Multiple	None	1660	Cross-sectional
Daudén *et al.*, 2009[84]	Multiple	Continuous vs. paused subcutaneously therapy	720	RCT
Van de Kerkhof 2004[82]	Multiple	Two-compound product (+ ointment vehicle, once daily), Two-compound product (twice daily), calcipotriol (Dovonex®, LEO) (twice daily), ointment vehicle (twice daily)	828	RCT
Luger *et al.*, 2009[96]	Multiple	Continuous and paused etanercept therapy	130	Before-and-after
Reich *et al.*, 2009[85]	Multiple	Etanercept	720	RCT
Revicki *et al.*, 2008[94]	Multiple	Adalimumab (Humira®, AbbVie), methotrexate, placebo	54	Cross-sectional
Shikiar *et al.*, 2006[95]	USA and Canada	Subcutaneously administered adalimumab vs. placebo	27994	Cross-sectional
Shikiar *et al.*, 2007[86]	USA and Canada	Subcutaneously administered adalimumab vs. placebo	142	RCT
Weiss *et al.* 2002[87]	USA	N/R (only baseline data were reported)	271	RCT
Weiss *et al.* 2006[88]	USA	Topical therapy vs. combination clobetasol solution	147	RCT
Acne				
Klassen *et al.* 2000[81]	UK	Isotretinoin or antibiotic, hormonal, physical and topical treatments	148	RCT
Hidradenitis suppurativa				
Matusiak *et al.* 2010[89]	Poland	N/R	233	RCT
Hand eczema				
Moberg *et al.* 2009[90]	Sweden	N/R	35	RCT
Venous leg ulcers				
Walters *et al.* 1999[91]	UK	Compression bandaging in a community clinic setting vs. usual home-based care by district nursing services	32	RCT

N/R, not reported.

Plaque psoriasis and psoriatic arthritis

Known-group analysis Eight studies provided evidence of known-group validity for EQ-5D among people with psoriasis or psoriatic arthritis.[82–85,87,92,93,96] Three studies showed that EQ-5D was able to discriminate between severity groups on the basis of psoriatic arthritis and psoriasis,[93] treatments,[84] pain and nail psoriasis.[96] Three case–control studies confirmed that EQ-5D can differentiate between people

TABLE 9 Measures used in the studies included in the skin review

Study reference grouped by condition (author, year)	GPBMs			Direct valuation	Rating scale	Generic or condition specific HRQL instruments			
	EQ-5D	SF-6D	HUI3	TTO	VAS	SF-36	DLQI	PASI	Others
Plaque psoriasis and psoriatic arthritis									
Bansback *et al.*, 2006[83]	✓								HAQ-DI
Brodszky *et al.*, 2010[92]	✓				✓ (pain, global assessment)			✓	PsAQoL, HAQ, PASI, DAS28, BASDAI, swollen joint count, tender joint count, EQ-VAS, patient pain VAS, patient global assessment VAS
Christophers *et al.*, 2010[93]	✓								BSA, employment disadvantage questionnaires
Daudén *et al.*, 2009[84]	✓				✓	✓	✓		HADS, PSS, BSA, PGA
Van de Kerkhof 2004[82]	✓				✓				Psoriasis Disability Index
Luger *et al.*, 2009[96]	✓				✓	✓	✓	✓	HADS, SGA, PGA, BSA, NAPSI
Reich *et al.*, 2009[85]	✓				✓		✓	✓	FACIT-F, BSA
Revicki *et al.*, 2008[94]	✓				✓		✓	✓	
Shikiar *et al.*, 2006[95]	✓				✓	✓	✓	✓	PGA
Shikiar *et al.*, 2007[86]	✓				✓	✓	✓	✓	PGA
Weiss *et al.*, 2002[87]	✓					✓		✓	SAPASI, SWLS
Weiss *et al.*, 2006[88]	✓				✓		✓	✓	SAPASI, BSA

continued

TABLE 9 Measures used in the studies included in the skin review (*continued*)

Study reference grouped by condition (author, year)	GPBMs			Direct valuation	Rating scale	Generic or condition specific HRQL instruments			
	EQ-5D	SF-6D	HUI3	TTO	VAS	SF-36	DLQI	PASI	Others
Acne									
Klassen *et al.*, 2000[81]	✓				✓		✓		Acne grade
Hidradenitis suppurativa									
Matusiak *et al.*, 2010[89]	✓				✓		✓		BDI-SF, FACIT-F, QLES-Q, GQ 6-item scale, Hurley's classification
Hand eczema									
Moberg *et al.*, 2009[90]	✓				✓				
Venous leg ulcers									
Walters *et al.*, 1999[91]	✓				✓	✓			FAI, SF-MPQ, self-perceived transition question (item 2 of SF-36) with three scales: better, same and worse compared with 3 months earlier

BASDAI, the Bath Ankylosing Spondylitis Disease Activity Index; BDI-SF, Beck Depression Inventory – short form; DAS28, the 28 joint disease activity score; FACIT-F, Functional Assessment of Chronic Illness Therapy – Fatigue; FAI, the Frenchay Activities Index; GQ, Global Question index; PGA, Physician Global Assessment; PsAQoL, Psoriatic Arthritis Quality of Life Scale; PSS, patient satisfaction survey; QLES-Q, Quality of Life Enjoyment and Satisfaction Questionnaire; SAPASI, self-administered PASI; SF-MPQ, Short Form McGill pain questionnaire; SGA, subject global assessment (for joint pain); SWLS, Satisfaction With Life Scale.

with psoriasis and the general population.[82,85,87] Brodszky *et al.*[92] found that the standard mean difference between groups measured by EQ-5D were comparably lower than measured with the Psoriatic Arthritis Quality-of-Life Scale (PsAQoL) or the HAQ; however, the groups were defined not according to severity aspects, but according to possible surrogate markers of severity such as admission to hospital or use of devices.[92]

Convergent validity Good convergent validity of EQ-5D was found among people with psoriasis or psoriatic arthritis in four studies.[83,88,92,95] Three studies showed moderate or strong correlation between EQ-5D and other generic or skin-specific measures.[87,92,95] Bansback *et al.*[83] suggested that the HAQ disability index was a significant predictor of EQ-5D.

Other skin conditions

Four studies had sufficient information to allow assessment of construct and convergent validity in various skin conditions.[81,89–91]

Known-group analysis In a case–control study, Klassen *et al.*[81] found that people with acne reported more problems on most EQ-5D dimensions than the general population. Among those with hidradenitis suppurativa, Matusiak *et al.*[89] found that significant differences according to the severity groups defined by Hurley's classification groups were suggested by EQ-5D, EQ-VAS, DLQI, the Beck Depression Inventory-Short Form (BDI-SF) and other measures. Among patients with hand eczema, Moberg[90] suggested that EQ-5D and EQ-VAS significantly differ between groups defined according to whether they have hand eczema groups, as well as age and gender. For venous leg ulcer patients, Walters *et al.*[91] reported small ESs for the EQ-5D, EQ-VAS, SF-36 and Frenchay Activities Index (FAI) for patients grouped on the basis of their initial leg ulcer size, current ulcer duration, maximum ulcer duration and age; however, the differences were statistically significant only for the EQ-5D, EQ-VAS, FAI and five subscales of the SF-36.

Convergent validity Among those with hidradenitis suppurativa, moderate correlation was reported between EQ-5D with DLQI and EQ-5D with Functional Assessment of Cancer Therapy – Fatigue module (FACT-F). Moberg *et al.*[90] found strong correlation between EQ-5D and EQ-VAS among hand eczema patients, and, similarly, Walters *et al.*[91] found moderate to high correlations with SF-36 subscales.

Responsiveness: skin conditions

A total of 10 studies provided evidence to allow assessment of responsiveness of EQ-5D in skin diseases.[81,82,84–86,88,91,94–96] Among them, eight studies included people with psoriasis or psoriatic arthritis,[82,84–86,88,94–96] one study included people with acne[81] and one study focused on venous leg ulcers.[91] Ten studies examined changes of scores over time or after treatment,[81,82,84–86,88,91,94–96] and two provided details of ES or standard response mean estimation.[81,91] One study checked the correlation between change scores of health measures with changes in clinical measures[95] (see *Appendix 8*).

Plaque psoriasis or psoriatic arthritis

All eight studies among people with psoriasis or psoriatic arthritis confirmed that EQ-5D was responsive to change in health over time in these conditions.[82,84–86,88,94–96] Daudén *et al.*[84] reported that consistent with EQ-VAS, DLQI, HADS-anxiety subscale and the SF-36 vitality dimension, EQ-5D values improved significantly and clinically meaningfully from baseline for both treatment groups. Luger *et al.*[96] demonstrated that EQ-5D values improved significantly (by 29%), as did scores from the EQ-VAS, DLQI, the SF-36 vitality dimension, HADS-depression subscale and HADS-anxiety subscale among patients with joint pain; however, the improvement reported using EQ-5D was not significant for patients with nail psoriasis, whereas improvement using the other measures was significant.[96] Reich *et al.*[85] reported that, at both follow-up time points, the group who received active treatment achieved significant improvement compared with placebo, measured using EQ-5D, EQ-VAS, FACT-F and DLQI (both total and domain scores). Similarly, Revicki *et al.*[94] reported that a statistically significant improvement was detected for treatment groups by EQ-5D, DLQI and PASI and the difference between treatment and placebo groups was significant. Shikiar *et al.*[86,95] also confirmed that the two treatment groups improved significantly more than placebo, measured using EQ-5D, EQ-VAS, DLQI, and most SF-36 domains. Weissi *et al.*[88] reported that, after 2 weeks of therapy, scores of EQ-5D, EQ-VAS, PASI, body surface area (BSA) and self-administered PASI (SAPASI) all improved significantly. Van de Kerkhof[82] showed that a significant improvement was detected by EQ-VAS, Psoriasis Disability Index, and the pain/discomfort and anxiety/depression dimensions of EQ-5D, although no statistical tests were reported.

Acne

Klassen *et al.*[81] reported that EQ-5D detected a significant change after treatment and this was consistent with SF-36 physical component summary score and DLQI. A moderate ES for EQ-5D was reported.

Venous leg ulcers

Walters *et al.*[91] reported mixed results in a study of compression healing of venous leg ulcers in different settings. When patients were grouped according to the status of the leg ulcer healing at 3 months, both EQ-5D and SF-36 showed deterioration in health status, but this conflicted with data from the VAS and the Short Form McGill pain questionnaire (SF-MPQ).

Summary and conclusion: skin conditions

The overall performance of EQ-5D among skin diseases is summarised in *Table 10*. Overall, there was evidence to suggest that EQ-5D is appropriate in terms of construct and convergent validity, as well as responsiveness in some skin conditions. All six studies showed that EQ-5D was able to reflect differences between severity groups[84,89–91,93,96] and only one was not significant.[91] EQ-5D was shown to be able to significantly differentiate between patient and general populations in four case–control studies[85,86–88] (one study did not report statistical tests),[82] as well as groups defined by other aspects rather than severity. Moderate to strong correlations were found between EQ-5D and other measures. Nine[81,82,84–86,88,94–96] out of 10 studies[81,82,84–86,88,91,94–96] demonstrated that EQ-5D was able to detect change appropriately over time. Among these, only one study did not demonstrate a statistically significant difference.[82] 'Skin conditions' were defined in very broad terms for the purpose of the review and incorporate a range of conditions, each of which can affect different aspects of patients' QoL. Most of the studies identified were conducted for patients with psoriasis or psoriatic arthritis. Evidence was limited or unavailable for other skin conditions; however, the limited data available were generally positive. No studies reported evidence for HUI3 and SF-6D and no studies allowed an assessment of reliability for any of the measures.

Cancer

Search results: cancer

Bibliographic searching was completed in August 2010. A total of 5223 potentially relevant papers were identified. Overall, a total of 5000 papers were excluded following screening of title and abstract. Full papers were reviewed for the remaining 223 records which met the inclusion criteria. After reviewing the full papers, 125 were excluded and a total of 98 papers were included in the review. A flow chart of the study selection process is shown in *Figure 4*.

The 98 papers were grouped according to 20 different types of cancers. These included 18 papers on non-specific cancers,[97–114] 11 each for colon cancer[115–125] and cancer survivors,[126–136] 10 for breast cancer,[137–146] eight for gastric cancer[147–154] and seven for prostate cancer,[155–161] and a small number of papers for brain,[162,163] cervical,[164–167] kidney,[168–171] lung[103,172,173] and other cancers[101,124,174–188] (*Table 11* gives details). As different cancers affect HRQL in different ways, the following sections present data according to the different types of cancer.

Quality assessment: cancer

A range of study designs were observed in the review. Some were cross-sectional studies,[13,97–99,102,103,106–108,114,115,121,126,128–130,133,137,138,147,148,151,152,156,161,163,164,173,183,184,189] others were before-and-after studies[110,112,116,117,120,123,139,140,145,190] or cohort studies[100,141,142,155,157,158,160,162,191] and many were RCTs.[118,119,122,125,132,136,143,144,146,149,150,153,154,159,161,165,166,168–171,176,177,180–182,188,192,193] Most RCTs had clear inclusion and exclusion criteria and

TABLE 10 Overall performance of EQ-5D in studies of skin diseases

Study reference grouped by measure (author, year)	Conditions	Known group (severity)		Known group (case–control)		Known group (other)		Correlation	Responsiveness		Reliability
		Consistent evidence	Significant	Consistent evidence	Significant	Consistent evidence	Significant		Consistent evidence	Significant	
EQ-5D											
Bansback *et al.*, 2006[83]	Psoriatic arthritis	✓	✓					✓			
Brodszky *et al.*, 2010[92]	Psoriatic arthritis					✓	✓	Strong			
Christophers *et al.*, 2010[93]	Plaque psoriasis and Psoriatic arthritis	✓	✓								
Daudén *et al.*, 2009[84]	Plaque psoriasis	✓	✓						✓	✓	
Van de Kerkhof, 2004[82]	Plaque psoriasis			✓	N/R				✓	N/R	
Luger *et al.*, 2009[96]	Plaque psoriasis	✓	✓						✓	✓	
Reich *et al.*, 2009[85]	Plaque psoriasis			✓	✓				✓	✓	

continued

TABLE 10 Overall performance of EQ-5D in studies of skin diseases (*continued*)

Study reference grouped by measure (author, year)	Conditions	Known group (severity)		Known group (case–control)		Known group (other)		Correlation	Responsiveness		Reliability
		Consistent evidence	Significant	Consistent evidence	Significant	Consistent evidence	Significant		Consistent evidence	Significant	
Revicki *et al.*, 2008[94]	Plaque psoriasis								✓	✓	
Shikiar *et al.*, 2006[95]	Psoriasis							Moderate to strong	✓	✓	
Shikiar *et al.*, 2007[86]	Psoriasis								✓	✓	
Weiss *et al.*, 2002[87]	Psoriasis			✓	✓			Moderate (significant)	✓	✓	
Weiss *et al.*, 2006[88]	Psoriasis										
Klassen *et al.*, 2000[81]	Acne			✓	✓				✓	✓	
Matusiak *et al.*, 2010[89]	Hidradenitis Suppurativa	✓	✓					Moderate			
Moberg *et al.*, 2009[90]	Hand eczema	✓	✓			✓	✓	Strong			
Walters *et al.*, 1999[91]	Venous leg ulcers	✓	N/R			✓	N/R	Moderate	Mixed evidence	N/R	

N/R, not reported.

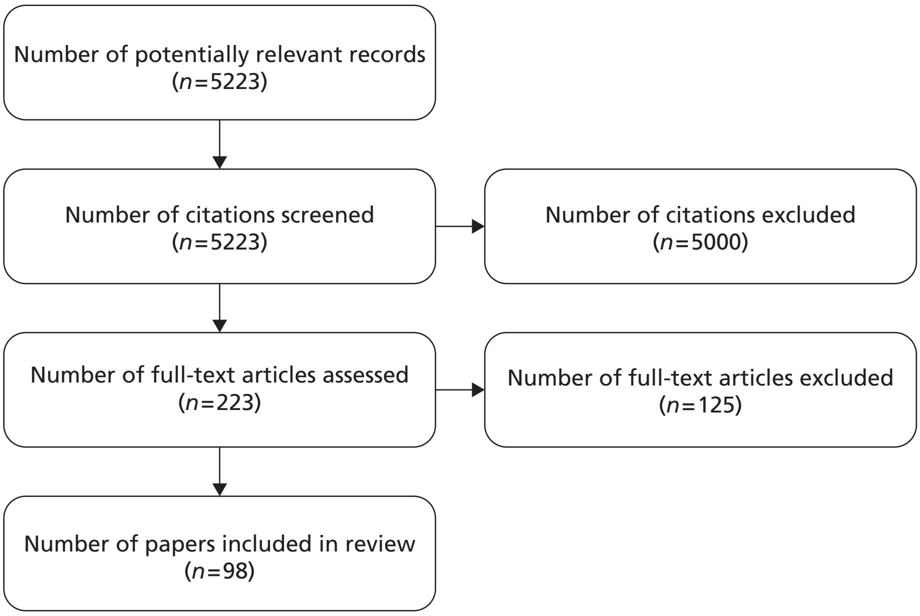
Number of potentially relevant records
(n=5223)
Number of citations screened
(n=5223)
Number of citations excluded
(n=5000)
Number of full-text articles assessed
(n=223)
Number of full-text articles excluded
(n=125)
Number of papers included in review
(n=98)

FIGURE 4 Flow diagram showing selection of studies: cancer.

TABLE 11 Number of papers included in the review by type of cancer

Cancer type	Number of papers
Non-specific	18
Brain	2
Breast	10
Cervical	4
Colon	11
Gastric	8
Hodgkin's lymphoma	2
Kidney	5
Leukaemia and related	3
Liver	3
Lung	2
Lymphoma	3
Lymphoma/leukaemia	2
MM	2
MM/lymphoma	1
Musculoskeletal	1
Pancreatic	1
Prostate	8
Spinal metastases	1
Survivors	11
Total	98

MM, multiple myeloma.

appropriate and explicit methods of randomisation. In some studies, the inclusion criteria were not clearly reported, which occurred mainly for studies of non-specific cancers.[97–100]

Response rates varied between studies. Completion rates for breast cancer studies ranged from 74%[141] to 99%[143] and for colon cancer ranged from 67%[115] to 90%.[120] No study was excluded after the assessment of quality.

Study characteristics: cancer

General characteristics of the 98 studies are presented in *Table 12*. These studies were divided into 20 subgroups according to different types of cancer. Thirty-three studies were cross-sectional analyses,[97–99,102,103,106–108,114,115,121,126–130,133,137,138,147,148,151,152,156,161,163,164,173,181–184,189] 24 were RCTs,[118,119,122,125,143,144,146,149,150,153,154,159,165,166,168–171,176,177,180,188,192,193] 24 were before-and-after or longitudinal studies[99,101,105,109,111–113,116,117,120,123,125,140,145,154,162,167,172,174,178,179,185,187,190] and nine were cohort studies.[100,104,141,142,155,157,158,160,191] Most groups included a mixture of study designs, exceptions were kidney cancer[168,169,170,171,193] and lymphoma[177,188,192] which were all RCTs and both lung cancer studies[103,173] had cross-sectional designs. The selected studies were conducted in different countries across Europe, Asia and North America and eight were multinational studies.[118,146,157,168–171,193] Various treatments were included in the studies including types of surgery,[117,141] radiotherapy and chemotherapy,[100,137,162,175] other medicines and supportive care interventions or referral.[118,143,165,166] Most studies included adults, but some were collected data from children using HUI including studies of brain cancer,[163] Hodgkin's lymphoma,[185,190] and a couple of the studies where recruitment was not limited to a specific type of cancer.[98,99,105,107]

The inclusion criteria for recruiting patients varied across the studies reviewed and within each type of specific cancer. Some studies recruited patients according to specific stages of cancer patients, for example primary tumours,[162] stage II and III breast cancer with poor prognosis,[140] tumour stage I, II and III breast cancer[144] and advanced colorectal cancer.[119] Some studies involved patients after screening, for example studies of screening for cervical cancer.[164–167] For these screening studies, some of the respondents would be asymptomatic and therefore the GPBMs and other measures may not be expected to reflect differences between patients with and without cancer. Sample size varied across studies, ranging from 18[112] to 113,587.[99]

Measures: cancer

Table 13 summarises the measures that have been used in the 98 studies included in the review. For the three GPBMs of interest, EQ-5D was the most commonly used and was reported by 71 studies.[97,98,100,101,103–107,110–123,128,129,137–140,145–154,143,144,156–160,164–173,175–177,179–184,186,188,192–194] Twenty-four studies reported HUI2/HUI3[99,108,109,126,127,130,131–136,141,142,155,161,162,163,174,178,185–187,190] and only three studies reported SF-6D.[98,147,156] Two studies[101,143] used EQ-5D and HUI3 alongside other measures and another three studies[98,147,156] use both EQ-5D and HUI3 alongside other cancer-specific measures. Fifty-eight studies also reported patients' ratings of their own health status using VAS[97,98,100–104,106–109,111,112,114–118,120–123,129,137–140,142–145,148,149,151–155,159,164–166,168–172,175,177,179–181,183–185,190,193,194] and valuations of own health were reported in three studies using TTO[101,138,174] and in one study using the SG method.[161] Five studies also reported generic measures SF-12 or SF-36.[120,141,164,173,184] A wide range of cancer-specific measures of health were used, including the most commonly used European Organization for Research and Treatment of Cancer Quality-of-life Questionnaire (EORTC QLQ) in 26 studies[100,101,111,115–117,121,124,128,129,144,146,147,149–151,153,154,172,179,180,182,183,188,192,193] and the FACT in 13 studies.[102,103,105,106,114,118,120,123,129,143,171,176,190] A range of other measures were reported, including variations of the previously mentioned cancer-specific HRQL measures such as the EORTC QLQ-Core 38 (EORTC QLQ-C38), staging of cancer progression using various staging systems and other measures of symptoms or aspects of health such as the HADS (see *Table 13* for details). Many studies used multiple measures and did not always give consistent results, which make conclusions regarding concordance with results from the GBPMs more difficult to interpret.

TABLE 12 Characteristics of included studies: cancer review

Study reference grouped by condition (author, year)	Country	Sample size	Disease stage	Treatment	Study type
Brain cancer					
Le Gales *et al.*, 1999[163]	France	43	Children with medulloblastoma	Standard treatment protocols	Cross-sectional
McCarter *et al.*, 2006[162]	Canada	93	Primary tumours	Radiotherapy and/or chemotherapy and surgery	Prospective longitudinal
Breast cancer					
Chang *et al.*, 2004[143]	Canada	354	Mildly anaemic (haemoglobin level $\leq$12 g/dl) women with breast cancer	Epoetin alfa (40,000 international units once weekly) vs. standard of care	RCT
Conner-Spady *et al.*, 2001[139]	Canada	52	Stage II and III breast cancer	High-dose chemotherapy with autologous blood stem transplantation	Before-and-after
Conner-Spady *et al.*, 2005[140]	Canada	52	Stage II and III breast cancer with poor prognosis	High-dose chemotherapy with autologous blood stem transplantation	Before-and-after
Crott *et al.*, 2010[146]	5 European countries including UK	220	Locally advanced	Cyclophosphomide, epirubicin and fluorouracil vs. dose-intensified epirubicin and cyclophosphomide-filgrastim	RCT
Freedman *et al.*, 2010[145]	USA	1050	Early stage breast cancer (stage 0, I, II invasive breast cancer)	Breast conserving surgery and radiation	Before-and-after
Jansen *et al.*, 2004[137]	Netherlands	448	Early stage breast cancer	Adjuvant chemotherapy (choice regarding treatment with adjuvant chemotherapy)	Cross-sectional
Kimman *et al.*, 2009[144]	Netherlands	192	Breast cancer (tumour stage I, II, III and unknown)	Curative treatment: surgery and/or radiotherapy and/or chemotherapy	RCT
Lidgren *et al.*, 2007[138]	Sweden	361	Consecutive breast cancer	N/R	Cross-sectional
Lovrics *et al.*, 2008[141]	Canada	85	Breast cancer (tumour grade I, II, III)	Breast-conservation surgery	Cohort
Polsky *et al.*, 2002[142]	USA	1159	Primary T1 or T2, N0 or N1, or NX and M0 invasive breast carcinoma. People aged over 67 years and community dwelling	Mastectomy, breast conservation with radiation, breast conservation only. Choice regarding breast cancer treatment	Cohort study

continued

TABLE 12 Characteristics of included studies: cancer review (*continued*)

Study reference grouped by condition (author, year)	Country	Sample size	Disease stage	Treatment	Study type
Cervical cancer					
Korfage *et al.*, 2010[164]	Netherlands	622	Low grade abnormalities after screening	N/R	Cross-sectional
Maissi *et al.*, 2005[167]	UK	1011	Screening tested for either human papillomavirus (HPV) or abnormal smear or normal smear	N/R	Prospective longitudinal
Whynes *et al.*, 2008[165]	UK	3132	Low-grade abnormalities after screening	Control: cytological surveillance Intervention: immediate referral to colposcopy	RCT
Whynes *et al.*, 2008[166]	UK	191	Low-grade abnormalities after screening	Control: cytological surveillance Intervention: immediate referral to colposcopy	RCT

Study reference grouped by condition (author, year)	Country	Sample size	Disease stage	Treatment	Study type
Colon cancer					
Anderson and Palmer, 1998[119]	UK	545	Advanced colorectal cancer	Raltitrexed (Tomudex®, Hospira) vs. Standard 5-fluorouracil (5-FU) plus leucovorin	RCT
Colwell *et al.*, 2010[125]	USA	391	Metastatic colorectal cancer	Panitumumab (Vectibix®, Amgen) plus best supportive care vs. best supportive care alone	RCT
Doornebosch *et al.*, 2007[115]	Netherlands	62	T1 carcinoma after surgery (TEM), T1 to T3 (35%) (TME)	Total mesorectal excision vs. transanal endoscopic microsurgery	Cross-sectional
Doornebosch *et al.*, 2008[116]	Netherlands	47	People with rectal cancer eligible for TEM	Transanal endoscopic microsurgery	Before-and-after
Gosselink *et al.*, 2006[121]	Netherlands	204	People with rectal cancer in the middle or low third of the rectum after total mesorectal excision	Abdominoperineal resection, transanally double stapled low colorectal anastomosis, coloanal J-pouch anastomosis	Cross-sectional
Hamashima, 2002[117]	Japan	110	Rectal cancer patients who had received surgery as their initial treatment	Surgery	Before-and-after
Janson *et al.*, 2007[122]	Sweden	285	Elective colon cancer patients with potentially curable cancer best treated by right or left hemicolectomy or sigmoid resection	Laparoscopic colon resection vs. open resection	RCT
Ramsey *et al.*, 2000[191]	USA	74 (phase 1), 98 (phase 2)	Colon carcinoma survivors with TNM stage I–IV	Colon cancer related treatment including surgery, chemotherapy, radiation therapy, colostomy appliance	Before-and-after
Sharma *et al.*, 2007[123]	UK	104	Newly diagnosed colorectal cancer scheduled for elective open resection	Elective open resection	Before-and-after
Siena *et al.*, 2007[118]	Multinational	463	Metastatic colorectal cancer patients who had progressed on prior fluoropyrimidine, irinotecan and oxaliplatin	Panitumumab plus best supportive care vs. best supportive care alone	RCT
Wilson *et al.*, 2006[120]	UK	210	Patients undergoing potentially curable open surgery for colorectal cancer	Surgery	Before-and-after

continued

TABLE 12 Characteristics of included studies: cancer review (*continued*)

Study reference grouped by condition (author, year)	Country	Sample size	Disease stage	Treatment	Study type
Gastric (and related) cancer					
Homs *et al.*, 2004[149]	Netherlands	209	Patients with dysphagia from inoperable oesophageal carcinoma	Arm 1: stent placement Arm 2: single-dose brachytherapy	RCT
Kontodimopoulos *et al.*, 2009[147]	Greece	48	N/R	Surgery and 2–4 previous chemotherapy sessions	Cross-sectional (mapping)
McMillan *et al.*, 1999[153]	UK	73	Histologically proven advanced or metastatic gastrointestinal cancer	Arm 1: megestrol acetate (Megace®, Bristol-Myers Squibb) and ibuprofen Arm 2: megestrol acetate and placebo	RCT
O'Gorman *et al.*, 1998[151]	UK	119	Histologically proven advanced or metastatic gastrointestinal cancer	N/R	Cross-sectional
Rogers *et al.*, 2006[148]	UK	224	Oral and oropharyngeal squamous cell carcinoma patients	Primary surgery	Cross-sectional
Shenfine *et al.*, 2009[150]	UK	215	Patients with dysphagia due to oesophageal carcinoma	Arm 1: 18 mm stent Arm 2: 24 mm stent Arm 3: rigid stent Arm 4: non-stent treatments	RCT
Verschuur *et al.*, 2009[154]	Netherlands	109	Patients after surgery for oesophageal or gastric cancer	Arm 1: standard follow-up by surgeons at an outpatient clinic Arm 2: home visits by specialist nurse	RCT
Wildi *et al.*, 2004[152]	USA	50	Newly diagnosed adenocarcinoma/ squamous cell carcinoma of the oesophagus	N/R	Cross-sectional

Study reference grouped by condition (author, year)	Country	Sample size	Disease stage	Treatment	Study type
Hodgkin's lymphoma (children)					
Klaassen *et al.*, 2010[185]	Canada	49	New presentation of Hodgkin lymphoma	First and second course of chemotherapy and radiation	Longitudinal
Klaassen *et al.*, 2010[186]	Canada	49	New presentation of Hodgkin lymphoma	First and second course of chemotherapy and radiation	Longitudinal
Kidney cancer					
Castellano *et al.*, 2009[171]	Multinational	304	mRCC	Sunitab (Sutent®, Pfizer)	RCT
Cella *et al.*, 2008[169]	Multinational	750	mRCC	Sunitab	RCT
Cella *et al.*, 2010[168]	Multinational	750	mRCC	Sunitab	RCT
Sternberg *et al.*, 2010[193]	Multinational	435	aRCC	Oral pazopanib (Votrient®, GSK) Placebo	RCT
Yang *et al.*, 2010[170]	Multinational	270	aRCC	Temsirolimus (Torisel®, Pfizer)	RCT
Leukaemia cancer					
Barr *et al.*, 1997[174]	Canada	18	Acute lymphoblastic leukaemia	Continuing chemotherapy	Prospective longitudinal
Cox *et al.*, 2005[187]	USA	27	Acute lymphoblastic leukaemia	Frontline protocol	Longitudinal
Hahn *et al.*, 2003[176]	USA	865	Chronic myeloid leukaemia	Arm 1: imatinib (Glivec®, Novartis) Arm 2: interferon alfa plus low-dose cytarabine	RCT

continued

TABLE 12 Characteristics of included studies: cancer review (*continued*)

Study reference grouped by condition (author, year)	Country	Sample size	Disease stage	Treatment	Study type
Liver metastases					
Langenhoff *et al.*, 2006[180]	Netherlands	97	Colorectal liver metastases	Arm 1: surgical treatment of liver metastases Arm 2: inoperable disease and underwent exploratory laparotomy only Arm 3: patients with inoperable disease who were not scheduled for operation as were groups 1 and 2	RCT
Mendez Romero *et al.*, 2008[172]	Netherlands	28	Metastatic liver tumour	Stereotactic body radiation therapy	Longitudinal
Krabbe *et al.*, 2004[179]	Netherlands	75	Liver metastases	Liver surgery to eradicate metastatic disease A: resection B: local ablative therapy C: unresectable (so no surgery)	Prospective longitudinal
Lung cancer					
Pickard *et al.*, 2007[103]	USA	50	Advanced lung cancer	At least two cycles of chemotherapy	Cross-sectional
Trippoli *et al.*, 2001[173]	Italy	95	Non-small cell lung cancer	Resection, chemotherapy and/or radiotherapy	Cross-sectional

Study reference grouped by condition (author, year)	Country	Sample size	Disease stage	Treatment	Study type
Lymphoma cancer					
Doorduijn *et al.*, 2005[188]	Netherlands	132	Patients with newly diagnosed aggressive non-Hodgkin's lymphoma	Cyclophosphamide, doxorubicin, vincristine, prednisone chemotherapy	RCT
Van Agthoven *et al.*, 2001[177]	Netherlands	91 PBSCT: 62 ABMT: 29	Intermediate or high-grade Morbus Hodgkin or non-Hodgkin's lymphoma who relapsed after or were refractory to primary chemotherapy	Arm 1: autologous peripheral blood stem cell transplantation Arm 2: autologous bone marrow transplantation	RCT
Witzens-Harig *et al.*, 2009[192]	Germany	91 Treatment: 47 Observation: 44	Patients with CD20+ B cell non-Hodgkin's lymphoma	Arm 1: maintenance therapy with rituximab (MabThera®, Roche) every 3 months for 2 years Arm 2: observation	RCT
ML/AML					
Banks *et al.*, 2008[178]	Canada	29	N/R	One course of chemotherapy	Longitudinal study of patient and proxy report
Slovacek *et al.*, 2007[181]	Czech Republic	Total: 36 ML: 24/AML: 12	N/R	Haematopoietic stem cell transplantation	Retrospective cross-sectional

continued

TABLE 12 Characteristics of included studies: cancer review (*continued*)

Study reference grouped by condition (author, year)	Country	Sample size	Disease stage	Treatment	Study type
MM					
Slovacek *et al.*, 2007[182]	Czech Republic	32	N/R	High-dose chemotherapy followed by autologous peripheral blood progenitor cell transplantation	Retrospective cross-sectional
Uyl-de-Groot *et al.*, 2005[124]	Netherlands	51	Newly diagnosed MM	Tandem transplantation programme	Prospective longitudinal
MM/ML					
Slovacek *et al.*, 2007[181]	Czech Republic	80 recruited. 56 (70%) returned questionnaires	N/R	Progenitor stem cell transplantation	Retrospective cross-sectional
Musculoskeletal cancer					
Lee *et al.*, 2003[184]	South Korea	49	Patients who had been operated on for malignant musculoskeletal tumours and could walk unassisted	Surgery	Cross-sectional
Pancreatic cancer					
Mueller-Nordhorn *et al.*, 2006[183]	Germany	45	First admission to hospital with expected pancreatic cancer	N/R	Cross-sectional

Study reference grouped by condition (author, year)	Country	Sample size	Disease stage	Treatment	Study type
Prostate cancer					
Albertsen *et al.*, 1998[155]	USA	84	Localised	Management	Cohort
Krahn *et al.*, 2003[161]	Canada	235	All prostate cancer stages	Prostatectomy	Cross-sectional
				Radiotherapy	
				Hormonal therapy	
Krahn *et al.*, 2007[160]	Canada	248	1: patients undergoing treatment	Prostatectomy	Cohort
			2: patients with metastatic prostate cancer	Radiation/hormonal therapy (cohort 1)	
			3: all other prostate cancers, majority post treatment		
Sandblom *et al.*, 2004[158]	Sweden	1442	Palliative	Full range of palliative treatments	Cohort
Shimizu *et al.*, 2008[156]	Shimizu	330	Localised and advanced	Prostatectomy	Cross-sectional
				Radiotherapy	
				Brachytherapy	
				Hormonal therapy	
				Watchful waiting	
Sullivan *et al.*, 2007[157]	Multinational	280	Hormone refractory (advanced or metastatic)	Chemotherapy	Cohort
				Laser ablation	
				Radiotherapy	
				Other treatments	
Weinfurt *et al.*, 2005[159]	USA	643	Advanced	Zoledronic acid vs. placebo	RCT

continued

TABLE 12 Characteristics of included studies: cancer review (*continued*)

Study reference grouped by condition (author, year)	Country	Sample size	Disease stage	Treatment	Study type
Spinal metastases					
Falicov *et al.*, 2006[101]	Canada	85	Cancer with bony spinal metastases	Surgery for spinal metastases	Prospective longitudinal
Non-specific cancer					
Barton *et al.*, 2008[98]	UK	2770	N/R	N/R	Cross-sectional
Bowker *et al.*, 2006[99]	Canada	113,587	N/R	N/R	Cross-sectional
Capuano *et al.*, 2008[107]	Italy	164	No previous oncological treatment	N/R	Cross-sectional
Cheung *et al.*, 2009[105]	Singapore	558	Various	Various (54.7% currently on chemotherapy/radiotherapy)	Longitudinal
Chow *et al.*, 2010[106]	Singapore	316	Various	Complementary and alternative medicine	Cross-sectional
Kim *et al.*, 2008[113]	Korea	42	Various. All experiencing nausea or insomnia	Mirtazapine	Longitudinal
Lathia *et al.*, 2008 (abstract only)[102]	Canada	N/R	Various. All experiencing febrile neutropaenia	N/R	Cross-sectional
Mantovani *et al.*, 2004[111]	Italy	28	Advanced disease	Pharmaconutritional support for 16 weeks	Non-randomised
Norum, 1996[100]	Norway	125	N/R	Radiotherapy and/or chemotherapy	Cohorts

Study reference grouped by condition (author, year)	Country	Sample size	Disease stage	Treatment	Study type
Park *et al.*, 2006[191]	Korea	293	Palliative care	Palliative care	Cohort study
Pickard *et al.*, 2007[114]	USA	534	Advanced (stages III and IV)	Various	Cross-sectional
Pickard *et al.*, 2007[190]	USA	424	Various	N/R	Cross-sectional
Ravasco *et al.*, 2003[104]	Portugal	125	Various	Nutritional counselling and radiotherapy	Prospective cohort
Sung *et al.*, 2003[108] (children)	Canada	36	Various (child patient, parent respondent)	Chemotherapy	Cross-sectional
Trudel *et al.*, 1998[109] (children)	Canada	61	Various (child patient)	Assessed during treatment and follow-up	Longitudinal
Vaghela *et al.*, 2007 [112]	UK	18	Various	'Spiritual healing'	Before-and-after (pilot)
Wang *et al.*, 2008[97]	Germany	38	N/R	N/R	Cross-sectional
Weze *et al.*, 2004[110]	UK	35	Various	Healing by 'gentle touch'	Before-and-after

continued

TABLE 12 Characteristics of included studies: cancer review (*continued*)

Study reference grouped by condition (author, year)	Country	Sample size	Disease stage	Treatment	Study type
Cancer survivors					
Barr *et al.*, 1999[127]	Canada	44	Survivors of central nervous system tumours	Operative intervention, radiotherapy and chemotherapy	Cross-sectional
Barr *et al.*, 2000[133]	Canada	78	Survivors of Wilm's tumour and advanced neuroblastoma in childhood	N/R	Cross-sectional
Boman *et al.*, 2009[134]	Sweden	1599	Survivors of CNS tumours	N/R	N/R
Felder-Puig *et al.*, 2000[131]	German	142	Survivors of a range of cancers	N/R	N/R
Fu *et al.*, 2006[130]	Central America	211	Survivors of a range of cancers	Chemotherapy, surgery and radiation	Cross-sectional prospective (patient and proxy report)
Grant *et al.*, 2006[135]	USA	84	Survivors of a range of cancers	N/R	N/R
Korfage *et al.*, 2009[129]	Netherlands	640	Survivors of cervical cancers	N/R	Cross-sectional
Nijdam *et al.*, 2008[128]	Netherlands	119	Survivors of a range of cancers	Brachytherapy or surgery	Cross-sectional
Nixon Speechley *et al.*, 1999[136]	Canada	244	Survivors of a range of cancers	N/R	Retrospective cohort
Pogany *et al.*, 2006[132]	Canada	4584	Survivors of a range of cancers	Operative intervention, radiotherapy and chemotherapy	Retrospective cohort
Shimoda *et al.*, 2005[126]	Brazil	50	Survivors of a range of cancers	Chemotherapy, radiotherapy and surgery	Cross-sectional

AML, acute myeloid leukaemia; aRCC, advanced renal cell carcinoma; ML, malignant lymphoma; MM, multiple myeloma; mRCC, metastatic renal cell carcinoma; N/R, not reported; TEM, transanal endoscopic microsurgery; TME, total mesorectal excision; TNM, tumour node metastasis.

TABLE 13 Measures used: cancer review

Study reference grouped by condition (author, year)	GPBMs			Direct valuation	Rating	Generic measures	Condition specific measures				Clinical index	
	EQ-5D	SF-6D	HUI2/HUI3	TTO	VAS	SF-12 and SF-36	FACT	EORTC	HAD	Others	ECOG	Stage
Brain cancer												
Le Gales *et al.*, 1999[163]			✓							✓		
McCarter *et al.*, 2006[162]			✓							✓		
Breast cancer												
Chang *et al.*, 2004[143]	✓		✓		✓		✓			✓		
Conner-Spady *et al.*, 2001[139]	✓				✓					✓		
Conner-Spady *et al.*, 2005[140]	✓				✓					✓		
Crott *et al.*, 2010[146]	✓							✓				
Freedman *et al.*, 2010[145]	✓				✓							
Jansen *et al.*, 2004[137]	✓				✓				✓			
Kimman *et al.*, 2009[144]	✓				✓			✓				
Lidgren *et al.*, 2007[138]	✓			✓	✓							
Lovrics *et al.*, 2008[141]			✓			✓						
Polsky *et al.*, 2002[142]			✓		✓							
Cervical cancer												
Korfage *et al.*, 2010[164]	✓				✓	✓						
Maissi *et al.*, 2005[167]	✓									✓		
Whynes *et al.*, 2008a[165]	✓				✓				✓	✓		
Whynes *et al.*, 2008b[166]	✓				✓				✓	✓		

continued

TABLE 13 Measures used: cancer review (*continued*)

Study reference grouped by condition (author, year)	GPBMs			Direct valuation	Rating	Generic measures	Condition specific measures				Clinical index	
	EQ-5D	SF-6D	HUI2/HUI3	TTO	VAS	SF-12 and SF-36	FACT	EORTC	HAD	Others	ECOG	Stage
Colon cancer												
Anderson and Palmer, 1998[119]	✓											
Doornebosch *et al.*, 2007[115]	✓				✓			✓				
Doornebosch *et al.*, 2008[116]	✓				✓			✓		✓		
Gosselink *et al.*, 2006[121]	✓				✓			✓		✓		
Hamashima, 2002[117]	✓				✓			✓				
Janson *et al.*, 2007[122]	✓				✓							
Ramsey *et al.*, 1998[190]			✓				✓					
Sharma *et al.*, 2007[123]	✓				✓		✓					
Siena *et al.*, 2007[118]	✓				✓		✓					
Wilson *et al.*, 2006[120]	✓				✓	✓	✓					
Gastric (and related)												
Homs *et al.*, 2004[149]	✓				✓			✓				✓
Kontodimopoulos *et al.*, 2009[147]	✓	✓						✓				
McMillan *et al.*, 1999[153]	✓							✓				
O'Gorman *et al.*, 1998[151]	✓				✓			✓		✓		
Rogers *et al.*, 2006[148]	✓				✓					✓		
Shenfine *et al.*, 2009[150]	✓							✓		✓		
Verschuur *et al.*, 2009[154]	✓				✓			✓				✓
Wildi *et al.*, 2004[152]	✓				✓							✓

Study reference grouped by condition (author, year)	GPBMs			Direct valuation	Rating	Generic measures	Condition specific measures				Clinical index	
	EQ-5D	SF-6D	HUI2/HUI3	TTO	VAS	SF-12 and SF-36	FACT	EORTC	HAD	Others	ECOG	Stage
Hodgkin's lymphoma												
Klaassen *et al.*, 2010[186]			✓		✓					✓		
Klaassen *et al.*, 2010[187]			✓		✓					✓		
Kidney cancer												
Castellano *et al.*, 2009[172]	✓				✓		✓				✓	
Cella *et al.*, 2008[169]	✓				✓						✓	
Cella *et al.*, 2010[168]	✓				✓							
Sternberg *et al.*, 2010[194]	✓				✓			✓				
Yang *et al.*, 2010[170]	✓				✓							
Leukaemia												
Barr *et al.*, 1997[174]			✓	✓						✓		
Cox *et al.*, 2005[187]			✓									
Hahn *et al.*, 2003[176]	✓						✓			✓	✓	
Liver metastases												
Krabbe *et al.*, 2004[179]	✓				✓			✓				
Langenhoff *et al.*, 2006[180]	✓				✓			✓				
Lung cancer												
Mendez Romero *et al.*, 2008[172]	✓				✓			✓				
Pickard *et al.*, 2007[103]	✓						✓				✓	
Trippoli *et al.*, 2001[173]	✓					✓						

continued

TABLE 13 Measures used: cancer review (*continued*)

Study reference grouped by condition (author, year)	GPBMs			Direct valuation	Rating	Generic measures	Condition specific measures				Clinical index	
	EQ-5D	SF-6D	HUI2/HUI3	TTO	VAS	SF-12 and SF-36	FACT	EORTC	HAD	Others	ECOG	Stage
Lymphoma												
Doorduijn *et al.*, 2005[188]	✓							✓		✓		
van Agthoven *et al.*, 2001[177]	✓				✓							
Witzens-Harig *et al.*, 2009[192]	✓							✓				
ML/AML												
Banks *et al.*, 2008[178]			✓							✓		
Slovacek *et al.*, 2007[181]	✓				✓							
MM												
Slovacek *et al.*, 2008[175]	✓				✓							
Uyl-de-Groot *et al.*, 2005[124]	✓							✓				
MM/ML												
Slovacek *et al.*, 2007[182]	✓							✓				
Musculoskeletal cancer												
Lee *et al.*, 2003[184]	✓				✓	✓				✓		

Study reference grouped by condition (author, year)	GPBMs			Direct valuation	Rating	Generic measures	Condition specific measures				Clinical index	
	EQ-5D	SF-6D	HUI2/HUI3	TTO	VAS	SF-12 and SF-36	FACT	EORTC	HAD	Others	ECOG	Stage
Pancreatic cancer												
Mueller-Nordhorn *et al.*, 2006[183]	✓				✓			✓				✓
Prostate cancer												
Albertsen *et al.*, 1998[155]			✓		✓							
Krahn *et al.*, 2003[161]			✓	SG								
Krahn *et al.*, 2007[160]	✓		✓									
Sandblom *et al.*, 2004[158]	✓											
Shimizu *et al.*, 2008[156]	✓	✓										
Sullivan *et al.*, 2007[157]	✓											
Weinfurt *et al.*, 2005[159]	✓				✓							
Spinal metastases												
Falicov *et al.*, 2006[101]	✓		✓	✓				✓			✓	

continued

TABLE 13 Measures used: cancer review (*continued*)

Study reference grouped by condition (author, year)	GPBMs			Direct valuation	Rating	Generic measures	Condition specific measures				Clinical index	
	EQ-5D	SF-6D	HUI2/HUI3	TTO	VAS	SF-12 and SF-36	FACT	EORTC	HAD	Others	ECOG	Stage
Non-specific cancers												
Barton *et al.*, 2008[98]	✓	✓			✓							
Bowker *et al.*, 2006[99]			✓									
Capuano *et al.*, 2008[107]	✓				✓							
Cheung *et al.*, 2009[105]	✓						✓				✓	
Chow *et al.*, 2010[106]	✓				✓							✓
Kim *et al.*, 2008[113]	✓				✓					✓		
Lathia *et al.*, 2008 (abstract only)[102]	✓				✓		✓					
Mantovani *et al.*, 2004[111]	✓				✓			✓		✓	✓	
Norum, 1996[100]	✓				✓			✓				
Park *et al.*, 2006[192]	✓				✓					✓		
Pickard *et al.*, 2007[103]	✓				✓		✓				✓	
Pickard *et al.*, 2007[114]	✓				✓		✓					✓
Ravasco *et al.*, 2003[104]	✓				✓		✓				✓	
Sung *et al.*, 2003[108]			✓		✓					✓		
Trudel *et al.*, 1998[109]			✓		✓					✓		
Vaghela *et al.*, 2007[112]	✓				✓					✓		
Wang *et al.*, 2008[97]	✓				✓					✓		
Weze *et al.*, 2004[110]	✓				✓							

Study reference grouped by condition (author, year)	GPBMs			Direct valuation	Rating	Generic measures	Condition specific measures				Clinical index	
	EQ-5D	SF-6D	HUI2/HUI3	TTO	VAS	SF-12 and SF-36	FACT	EORTC	HAD	Others	ECOG	Stage
Survivors of cancer												
Barr *et al.*, 1999[127]			✓							✓		
Barr *et al.*, 2000[133]			✓									
Boman *et al.*, 2009[134]			✓									
Felder-Puig *et al.*, 2000[131]			✓							✓		
Fu *et al.*, 2006[130]			✓									
Grant *et al.*, 2006[135]			✓							✓		
Korfage *et al.*, 2009[129]	✓				✓		✓	✓		✓		
Nijdam *et al.*, 2008[128]	✓							✓				
Nixon Speechley *et al.*, 1999[136]			✓							✓		
Pogany *et al.*, 2006[132]			✓							✓		
Shimoda *et al.*, 2005[126]			✓							✓		

AML, acute myeloid leukaemia; ECOG, Eastern Co-operative Oncology Group; FACT, Functional Assessment of Cancer Therapy; ML, malignant lymphoma; MM, multiple myeloma.

Reliability: cancer

Fourteen studies[127,130,131,133,134,163,168,174,176–178,184,190,192] reported evidence to allow assessment of reliability of EQ-5D (five studies)[168,176,177,184,192] and HUI3 (nine studies)[127,130,131,133,134,163,174,178,190] in patients with cancer and results are summarised in *Appendix 9*. Cella *et al.*[168] examined EQ-5D in patients with kidney/renal cancer in terms of stability across treatment groups and found that EQ-5D, FACT and VAS scores did not differ between the different country cohorts. This provided some evidence for the reliability of EQ-5D in multinational trials. Similarly, Hahn *et al.*,[176] van Agthoven *et al.*[177] and Witzens-Harig *et al.*[192] reported that no significant differences between the treatment groups were found for EQ-5D, as well as EORTC QLQ-C30, among patients with leukaemia and lymphoma. Two studies examined the internal consistency of EQ-5D and HUI3 for specific questions/dimensions and dimensions/overall scores within measures.[163,184] One study reported that internal consistency was high for EQ-5D (as was the SF-36)[184] and another study reported consistency for most questions for HUI3.[163] Inter-rater reliability of HUI3 was reported in nine studies.[127,130,131,133,134,163,174,178,190] These studies reported completion of HUI3 by multiple respondents and all studies demonstrated high agreement between different raters' assessments of the dimensions of HUI3. Although the instruments are designed for self-completion by adults, the agreement between raters provides some limited evidence of reliability.

Known-group analysis and convergent validity: cancer

Overall, 77 studies[97–109,114,115,117,118,120–123,126–138,141,143–153,156–159,161–173,175,176,178,179–185,187,188,190,193] out of 98 provided evidence to allow for known-group analysis and convergent validity. Known-group analysis was carried out in 54 studies,[97–99,103–106,109,114,115,117–122,126,127,130–135,138,148–152,156–159,162–167,169,170,172,173,175,176,179–183,188,190,193] 41 for EQ-5D,[97,98,103–106,114,115,117–122,138,148–152,156–159,164–167,169,170,172,173,175,176,179–183,188,193] of which two also included the SF-6D[98,156] and 13 included the HUI3.[99,109,126,127,130–135,162,163,190] In most studies, groups were defined by severity of cancer on the basis of a global heath scale,[126,163,179] or disease status[120,127,162,164,195] or by treatment.[18,121,128,148,157] Some studies had case–control design comparing between cancer patients and the general public.[97,117,129,149] Several studies defined groups on the basis of other characteristics such as age and smoking status[175] and country.[130] The differences in the clinical definition of groups, conditions, characteristics of patients and study designs make it difficult to directly compare the utility values, or to conduct meta-analyses across studies.

Convergent validity testing was carried out in 30 studies, 20 for EQ-5D[98,100–103,107,123,137,138,144–146,148,151,156,164,165,173,184,196] and 10 for HUI3[101,108,136,141,143,155,162,178,186,187] and one for SF-6D.[156] In most cases, evidence on the correlation between generic measure of HRQL with either each other or with cancer-specific measures was reported.[138,141,143–145,162] Regressions between scores of different measures were reported by several studies.[102,122,146,147]

Details of the assessments of construct validity of utility measure in different type of cancers are shown in *Appendix 10* and below are briefly summarised by specific types of cancers. For some types of cancer, there were only limited studies (fewer than three) for assessment of validity. The findings of these are summarised under the heading of ‘other cancers’.

Breast cancer

Known-group analysis One study among people with breast cancer allowed a known-group analysis for EQ-5D where groups were defined by severity of breast cancer status.[138] EQ-5D and TTO can distinguish between different groups to some extent but the two measures did not always agree with each other in terms of which groups were different.

Convergent validity Correlation statistics were reported by five studies for EQ-5D[137,138,144–146] (two through regression estimation)[137,146] and two studies for HUI3 with other HRQL measures in patients with breast cancer.[141,143] Moderate to high correlations were found between the EQ-5D index with EQ-VAS or TTO values and the EQ-5D index with EORTC.[138,144,145] Significant regression coefficients were found between EORTC QLQ items and EQ-5D[146] and EQ-5D index, VAS, HADS-depression or anxiety demonstrated similar relationships between treatment choice and chemotherapy.[137] Strong correlations were found between HUI3 index and three subscales with Functional Assessment of Cancer Therapy – Anaemia (FACT-An) and FACT-F,[143] and between HUI3 and SF-36.[141]

Colon cancer

Known-group analysis In studies of patients with colon cancer, six studies for EQ-5D[115,117,118,120–122] and one study for HUI3[190] provided evidence to allow an assessment of construct validity. Of those reporting EQ-5D, five differentiated between groups based on severity of cancer[115,118,120–122] and one included an assessment of case (people with cancer) against controls (general population without cancer).[117] In two studies, EQ-5D scores demonstrated differences between treatment groups.[118,120] In four studies, EQ-5D index revealed no difference between study groups; the results of one study were consistent with no difference on EORTC QLQ-C30,[122] another was consistent with EQ-VAS among patient with or without stoma,[117] and two were consistent with EORTC QLQ-C30 but not EORTC QLQ-C38 among treatment groups.[115,121] The case–control analysis of the Gosselink *et al.*[121] study found that EQ-5D could differentiate between some, but not all, treatment groups with the general population. Ramsey *et al.*[190] found that HUI3 was consistent with the FACT – Colorectal subscale (FACT-C) summary scores and both measures detected significant differences between diagnosis groups.

Convergent validity One study[123] found that EQ-5D and EQ-VAS were not significantly correlated to the cancer tumour node metastasis (TNM) stage and the correlation coefficient was low, whereas other measures (HADS-anxiety subscale, positive and negative affect schedule and the emotional well-being component of the FACT-C module) had moderate correlations.[123]

Kidney cancer

Known-group analysis Three studies found that EQ-5D followed the same pattern across the study follow-up period with VAS, EORTC global health and EORTC global scores.[169,170,193] One study showed that EQ-5D, VAS and FACT scores did not differ between different country cohorts.[168]

Convergent validity The only study that reported convergent validity and illustrated that EQ-5D and EQ-VAS were moderately and significantly correlated with the Functional Assessment of Cancer Therapy – General Scale (FACT-G) and FACT-Kidney Symptom Index (FKSI).[171]

Cancer survivors

Known-group analysis Eight studies allowed known-group analysis for HUI3,[126,127,130–135] which successfully discriminated between cancer severity groups,[131] treatment groups,[132] global health rating[126] and between patients and controls.[132] Some HUI3 dimensions also discriminated between groups.[133,134] The HUI3 values and HUI3 dimensions were not significantly different between diagnosis groups;[130,135] however, it is not clear that any difference HRQL would be expected between these groups.

Two studies reported evidence for known-group assessment for EQ-5D.[128,129] One study found EQ-5D consistent with EORTC QLQ-C30 in that EQ-5D did not differ between treatment groups.[128] Another study found that neither EQ-5D nor the majority of dimensions of SF-36 displayed significant difference between survivors and control groups, but this was not consistent with the finding for the State-Trait Anxiety Inventory (STAI).[129]

Convergent validity Only one study reported moderate to high and significant correlations between HUI3 and the child health questionnaire (CHQ).[136]

Cervical cancer

Known-group analysis Three studies reported evidence to allow an analysis of known group validity for EQ-5D and the results were mixed.[164,166,167] One study[166] found that EQ-5D did not discriminate between the treatment and control group, which was consistent with HADS-anxiety and HADS-depression but not with the Multidimensional Health Locus of Control Scale chance dimension. Korfage *et al.*[164] demonstrated that EQ-5D found non-significant worsening of health for borderline mildly dyskaryotic group, but the increased psychological distress found by the SF-12 mental component summary score, STAI, Psychological Consequences Questionnaire (PCQ) score was significant. In contrast, significantly better physical health was found by the SF-12 physical component summary score. Maissi *et al.*[167] showed that STAI and general health questionnaire were sensitive to health differences at baseline whereas EQ-5D was not.

Convergent Validity One study[165] demonstrated moderate correlation between EQ-5D and EQ-VAS. Through regression, Korfage *et al.*[164] found that perceived risk of being diagnosed with cervical cancer was significantly associated with EQ-5D and PCQ score but not with mental component summary score or STAI.

Gastric cancer

Known-group analysis Five studies provided evidence to allow a known-group analysis for EQ-5D and the findings were generally mixed.[148,149,150–152] Shenfine *et al.*[150] and Rogers *et al.*[148] confirmed that EQ-5D values or the EQ-5D mobility and usual activities dimensions could discriminate between treatment groups. O'Gorman *et al.*[151] showed that, consistent with EORTC, EQ-5D was significantly lower and not significantly different in the weight-losing groups. However Wildi *et al.*[152] reported that the overall difference measured by EQ-5D between groups defined by cancer stage groups was not as expected or significant, although EQ-5D was higher for patients at cancer stage 0 than patients at stage 1–3. Two case–control studies[148,149] confirmed the ability of EQ-5D to discriminate between cancer patients and the general population.

Convergent validity Three studies provided evidence to assess convergent validity for EQ-5D,[147,148,151] and one of them also included SF-6D.[147] Through regression, Kontodimopoulos *et al.*[147] found that three EORTC subscales (physical and emotional function and global health status) were significant predictors of EQ-5D, whereas six EORTC subscales (social and emotional functioning, pain, constipation, dyspnoea and global health status) were significant predictors of SF-6D. Rogers *et al.*[148] showed significant correlation between the EQ-5D mobility, usual activities and anxiety dimensions, and the University of Washington QoL questionnaire overall scores, and between questionnaire subscales scores and specific EQ-5D dimensions.

Prostate cancer

Known-group analysis Four studies with prostate cancer patients allowed a known-group analysis of EQ-5D and the results suggested that EQ-5D discriminated between survival groups,[158] symptom-based severity groups (also shown by SF-6D[156]), and treatment groups.[157,159]

Convergent validity Studies reported low or non-significant correlations between HUI3 and VAS[155] or HUI3 and SG.[160]

Non-specific cancers

Known-group analysis Seven studies in groups not defined according to specific cancers provided evidence to allow a known-group analysis of the EQ-5D.[97,98,103–106,114] Among the six studies, four found that EQ-5D could discriminate groups defined on the basis of cancer severity such as Eastern Co-operative Oncology Group (ECOG) and FACT (statistical significance not reported),[103] high or low risk,[104] ECOG[105] and stage of cancer[106] (statistical significant not reported). For the two case–control studies, one study showed that a significant difference was found by EQ-VAS but not EQ-5D or SF-6D[98] and another study found that cancer patients were more likely to report any problems on the usual activities dimension of EQ-5D than other patients[97] but this was not found by the other dimensions.

Two studies in non-specific cancers provided evidence to allow known-group analysis for HUI3.[99,109] Both studies found that HUI3 scores were statistically different between groups. One study defined groups as cancer, cancer and diabetes, and diabetes only groups compared with no cancer or diabetes group;[99] another study defined groups on the basis of severity.[109]

Convergent validity Five studies examined the relationships of EQ-5D with other measures: two through correlation[100,114] and three through regression.[102,105,107] Pickard *et al.*[114] found statistically significant and moderate correlations between all EQ-5D dimensions, ECOG and subscales of FACT-G.[103] Similarly, Norum[100] found high correlations between EQ-5D, EQ-VAS and EORTC QLQ-C30. Capuano *et al.*[107] found that anaemia and weight loss significantly influenced EQ-5D scores but not inflammation, whereas in study by Lathia *et al.*[102] none of the EQ-5D data were significant predictors of Functional Assessment of Cancer Therapy – Neutropenia (FACT-N).

Two studies provided evidence to examine convergent validity of HUI3.[108,109] One study in children with cancer[108] found a moderate but significant correlation between HUI3 and the CHQ physical scale and between the pain, physical activity and emotion dimensions of HUI3 and the corresponding scale of the CHQ, but not between HUI3 and the psychosocial scale of CHQ. The other study including children reported by Trudel *et al.*[109] found moderate correlations for HUI3 values and the HUI3 dimensions compared with the VAS and a cancer-specific measure.

Liver cancer

Known-group analysis Two studies[179,180] found that EQ-5D could discriminate between treatment groups, which was consistent with the EORTC measure. Another case–control study[172] found that both EQ-5D and EORTC measure were sensitive to differences between a group of patients with liver metastasis and a group of the general population.

Lung cancer

Known-group analysis Two studies demonstrated that EQ-5D is able to distinguish patients groups on the basis of FACT quintiles,[103] and between patients with and without metastasis.[173]

Convergent validity Tripploli *et al.*[173] found that there were significant correlations between the EQ-5D index and VAS, and also between EQ-5D and SF-36.

Malignant lymphoma/acute myeloid leukaemia

Known-group analysis Slovacek *et al.*[181] found significantly higher EQ-5D scores among malignant lymphoma (ML) patients, which indicates that EQ-5D can discriminate between ML patients and acute myeloid leukaemia (AML) patients.

Convergent validity Banks *et al.*[178] demonstrated that there were substantial correlations between proxy HUI2/HUI3 and the CHQ physical score.

Other cancers

Known-group analysis Six studies among various cancer patients provided evidence to allow a known-group analysis for EQ-5D and HUI3. Slovacek[175] found that EQ-5D scores were significantly different depending on age and smoking status among patients with multiple myeloma (MM). Slovacek[182] demonstrated that the EQ-5D could differentiate between patients with MM and ML, with ML patients having significantly higher scores. One case–control study suggested that EQ-5D was consistent with EORTC in that it discriminated well between patients with pancreatic cancer and the general population as well as between gender groups.[183]

Two studies used the HUI3 in patients with brain cancer.[162,163] One study[163] found that the number of impaired HUI3 attributes was lower for children with better health status as reported by physicians, but no significant differences were found according to the level of radiation treatment received. Another study[162] found significant difference of all HUI3 dimensions (except emotion) between patients and the general population group, and between tumour groups although no significance was reported.

Convergent validity Significant correlations were reported between dimensions of the EQ-5D and the Musculoskeletal Tumour Rating Scale (MSTS) in patients with musculoskeletal cancer.[184,185] Klassen *et al.*[81] reported strong correlations between HUI3 and VAS, Pediatric Quality-of-Life Inventory (PedsQL) core and PedsQL-cancer module among patients with Hodgkin's lymphoma.[186] Falicov *et al.*[101] found a low to moderate correlation between EQ-5D and HUI3 among patients with spinal metastases.

Responsiveness: cancer

A total of 39 out of 98 studies among cancer patients provided sufficient evidence to allow assessment of responsiveness for EQ-5D (31 studies),[104,110–113,116,124,139,140,143,144,147,149,153,154,157–160,165,167–169,170,171,176,177,179,180,188,192] for HUI3 (six studies)[101,141,142,174,178,186] and both EQ-5D and HUI3 (two studies).[143,160] Most studies reported mean change of scores over the study period.[116,122,123,143,149,153,154,167] Some studies compared scores or responses at baseline and follow-up.[104,110,119,139] Some studies also reported responsiveness indices including ES or standard response mean,[141,169,171,185] or a correlation between changes of different measures.[143,144] Statistical tests such as the *t*-test, ANOVA and Mann–Whitney *U*-test were conducted by some, but not all, studies. The detailed results are summarised below according to type of cancer. As for validity, cancer types for which only three or fewer studies reporting responsiveness data were available are grouped as ‘other cancers’. See *Appendix 11* for details.

Breast cancer

Three studies of breast cancer patients provided evidence to examine responsiveness of EQ-5D, which was shown to perform satisfactorily. Conner-Spady *et al.* (2001)[139] found a significant change in mean scores over time for EQ-5D and three of its dimensions, Functional Living Index – Cancer (FLIC) and three of its subscales and VAS using repeated ANOVA. Large ESs were reported for all measures except for EQ-5D with a moderate ES for severe cancer according to thyroid hormone level (T_3/T_4). Another study by Conner-Spady *et al.*[140] demonstrated that EQ-5D, FLIC and VAS showed a similar pattern of change after high-dose chemotherapy, and a Friedman test showed significant change over time on four of the EQ-5D dimensions; there was no significant change for pain/discomfort. Kimman *et al.*[144] confirmed consistency between EQ-5D and EQ-VAS in terms of showing significant effect in the group that perceived a moderate and large change of global health but found no effect in the group that perceived no or small change of global health. Two studies examined the responsiveness of HUI3 in patients with breast cancer and found that performance was good.[141,142] Both Lovrics *et al.*[141] and Polsky *et al.*[142] found significant decreases in HUI3 score shortly after surgery and improvements in longer term, which was consistent with the VAS and SF-36 subscales.

One study by Chang *et al.*[143] provided evidence for both EQ-5D and HUI3, alongside EQ-VAS, FACT-An and FACT-F. The results of this study were difficult to interpret as it found that both HUI3 and EQ-VAS scores improved in one treatment group but decreased in another, although EQ-5D showed improvement for both groups. In addition, the difference between changes of scores between the treatment groups were statistically significant for HUI3 and EQ-VAS, but not for EQ-5D.

Cervical cancer

Two studies reported evidence for responsiveness assessment of EQ-5D.[165,167] Maissi *et al.*[167] found that mean change on EQ-5D was small but this was consistent with General Health Questionnaire and STAI. Whynes[165] showed that EQ-5D dimensions and HADS were significant predictors of decreasing VAS scores.

Colon cancer

Four studies provided evidence to examine responsiveness of EQ-5D.[116,119,122,123] Anderson and Palmer[119] found similar patterns over time for all EQ-5D dimensions and most subscales of the Rotterdam Symptom Checklist (RSCL) and significant differences were found between the two treatment groups over time using both measures. Doornebosch *et al.*[116] found that 6 months after surgery, significant improvement was detected by both the Faecal Incontinence Severity Index and EQ-VAS, but not EQ-5D. Both Janson *et al.*[122] and Sharma *et al.*[123] found that EQ-5D indicated no significant change over time and that this was not consistent with EORTC QLQ-C30 or the HADS.

Gastric cancer

Three studies provided evidence of responsiveness for EQ-5D.[149,153,154] Two studies[149,153] found consistent results with the EORTC, EQ-5D and EQ-VAS and all showed a change in HRQL, but this change was not significant. McMillan *et al.*[153] demonstrated that EQ-5D detected significant improvement in the intervention arm at follow-up.

Kidney cancer

Five studies included evidence to assess responsiveness of EQ-5D.[168,170,171,193,194] All five studies found that EQ-5D and EQ-VAS could detect differences between treatment groups and two studies[169,171] reported statistically significant differences.

Liver cancer

All three studies with responsiveness evidence suggested that EQ-5D was consistent with EORTC.[173,179,180] In one study,[180] both the EQ-5D and EORTC QLQ showed a response over time following three different surgical procedures. In another study,[172] both measures detected no change and another study[179] found comparable magnitude of change over time in terms of ES.

Prostate cancer

Two studies among the prostate cancer patients reported evidence of responsiveness for EQ-5D[157,159] and another study included both EQ-5D and HUI3.[160] Both Sullivan *et al.*[157] and Weinfurt *et al.*[159] confirmed that EQ-5D was responsive in prostate cancer patients as it detected deterioration in HRQL at follow-up and showed similar ES to other measures. Krahn *et al.*[160] indicated that EQ-5D and HUI3 were less responsive to treatment compared with other measures. Using external responsiveness, EQ-5D and HUI3 were able to discriminate between those whose health had changed and those whose health had not changed.

Non-specific cancer

Five studies among patients with general cancers provided evidence of responsiveness for EQ-5D and all studies found satisfactory performance of EQ-5D. Mantovani *et al.*[111] showed that EQ-5D registered a trend of improvement over time and the improvement at 4 months was statistically significant compared with baseline. Vaghela *et al.*[112] suggested that statistically significant improvement was seen on the anxiety and depression dimension of EQ-5D but was seen by the two first stated concerns of Measure Yourself Concerns and Well-Being Questionnaire (MYCaW), the overall profile and the EQ-VAS but not the well-being measure. Ravasco *et al.*[104] reported that all EQ-5D dimensions (except for pain/discomfort) and EQ-VAS improved following radiotherapy but the difference was statistically significant only for high-risk patients on the EQ-5D. Weze *et al.*[110] demonstrated that only the anxiety/depression and pain dimensions of EQ-5D showed statistically significant improvement whereas the EQ-VAS and stress, fear, sleep, relaxation and coping were significant. Similarly, Kim *et al.*[113] also reported that they found statistically significant differences in the sum of severity levels on pain/discomfort and anxiety/depression after treatment.

Other cancers

Five studies of various cancers provided information to allow assessment of responsiveness of the EQ-5D[124,176,177,188,192] and four studies for HUI3.[101,174,178,186] Hahn *et al.*[176] suggested that EQ-5D was picking up differences in mean change over time between the treatment groups in people with leukaemia, and Uyl-de-Groot *et al.*[124] found a significant mean change for EQ-5D and some EORTC QLQ-C30 dimensions at selected follow-up time points for people with MM. Three studies in patients with lymphoma indicated that EQ-5D changed over the study period, but this change was not always statistically significant.[177,188,192]

For HUI3, Klaasen *et al.*[186] found consistent change in the HUI3 and other measures between two time points with large and clinically relevant ES, but not at two other time points. The remaining three studies indicated good responsiveness of HUI3 across a range of indicators, including similar responsiveness to CHQ, but lower than PedsQL in terms of size of change. The pain dimension of HUI3 was responsive to change with EORTC QLQ-C30.

Summary and conclusion: cancer

The overall performance of EQ-5D, HUI3 and SF-6D are summarised in *Table 14*. Among the 98 studies included in this review, the EQ-5D[97,98,100,101,103–107,110–123,128,129,137–140,143–154,156–160,164–173,175–177,179–184,186,188,192,194] was the most commonly used GPBM, whereas HUI3[99,108,109,126,127,130–136,141,142,155,161–163,174,178,185–187,190] was the second most widely used measure. Few studies reported evidence for SF-6D.[98,147,156]

Overall, the results for EQ-5D compared with the other generic and cancer-specific measures were satisfactory. The majority of studies comparing patients with cancers and a control group of people without cancer showed consistent differences in EQ-5D values.[97,117,121,148,149,152,157,172,173,183] Studies comparing EQ-5D scores across severity groups also showed that, in most cases, EQ-5D differentiated between groups, although this was not always statistically significant.[103–106,114,118–120,122,148,156,158,164,173,176,188] Correlations between EQ-5D and other measures were a mixture of low, moderate and strong. In terms of responsiveness, overall EQ-5D scores or dimensions were able to detect appropriate change-over time points but sometimes the change of scores was small or not statistically significant over all time points. The assessment of reliability of EQ-5D provided some evidence of good reliability with no change being observed in EQ-5D responses when other measures confirmed no reported change in health over time; however, very few of the identified studies were specifically designed to assess test–retest reliability.

Evidence on the performance of EQ-5D varied in different types of cancer. EQ-5D showed good responsiveness and convergent validity in breast cancer[137–140,143–146] but known-group evidence was very limited. For colon cancer studies, the majority of evidence suggested relatively good construct validity,[118–120,122] but the only study available did not support responsiveness of EQ-5D.[116] In prostate cancer studies, EQ-5D appropriately differentiated between groups and detected change over time, but in most cases the differences or changes were not statistically significant.[156–160] In studies of non-specific cancers, EQ-5D was sensitive to change over time and sensitive to differences between severity groups.[103–106,110–114]

There was evidence to support HUI3's ability to differentiate between severity groups and between patients with and without cancers. The ability of HUI3 to detect between groups defined by other non-severity based aspects was more mixed and the responsiveness of HUI3 was also found to be satisfactory. Although HUI3 is essentially designed for self-completion by the patient, several studies examined inter-rater reliability.[133,134,163,174,178,186] These studies generally found that inter-rater reliability for HUI3 was good.

TABLE 14 Overview of performance of EQ-5D, HUI3 and SF-6D in studies of cancer

Study reference grouped by measure (author, year)	Cancer	Known group (severity)		Known group (case–control)		Known group (other)			Responsiveness		Reliability
		Consistent evidence	Significant	Consistent evidence	Significant	Consistent evidence	Significant	Correlation	Consistent evidence	Significant	Consistent evidence
EQ-5D											
Chang *et al.*, 2004[143]	Breast cancer								✓	✓	
Conner-Spady *et al.*, 2001[139]	Breast cancer								✓	✓	
Conner-Spady *et al.*, 2005[140]	Breast cancer								✓	✓	
Crott *et al.*, 2010[146]	Breast cancer							✓			
Freedman *et al.*, 2010[145]	Breast cancer							Strong (with EQ-VAS)			
Jansen *et al.*, 2004[137]	Breast cancer							✓			
Kimman *et al.*, 2009[144]	Breast cancer							Moderate to high	✓	✓	
Lidgren *et al.*, 2007[138]	Breast cancer	Mixed evidence	Mixed evidence					Moderate			
Korfage *et al.*, 2010[164]	Cervical cancer	✓	✗					✓			
Maissi *et al.*, 2005[167]	Cervical cancer			✗	✗				✓	N/R	
Whynes *et al.*, 2008[165]	Cervical cancer					✓	✓	✓ (moderate)	✓	✓	
Whynes *et al.*, 2008[166]	Cervical cancer	Mixed evidence	N/R								

continued

TABLE 14 Overview of performance of EQ-5D, HUI3 and SF-6D in studies of cancer (*continued*)

Study reference grouped by measure (author, year)	Cancer	Known group (severity)		Known group (case–control)		Known group (other)			Responsiveness		Reliability
		Consistent evidence	Significant	Consistent evidence	Significant	Consistent evidence	Significant	Correlation	Consistent evidence	Significant	Consistent evidence
Anderson and Palmer, 1998[119]	Colon cancer								✓	✓	
Doornebosch *et al.*, 2007[115]	Colon cancer	Mixed evidence	Mixed evidence								
Doornebosch *et al.*, 2008[116]	Colon cancer								✗	✗	
Gosselink *et al.*, 2006[121]	Colon cancer	✗	✗	✓	✓						
Hamashima, 2002[117]	Colon cancer			✓	✓						
Janson *et al.*, 2007[122]	Colon cancer	✓	✓								
Sharma *et al.*, 2007[123]	Colon cancer							Low			
Siena *et al.*, 2007[118]	Colon cancer	✓	✓								
Wilson *et al.*, 2006[120]	Colon cancer	✓	✓								
Homs *et al.*, 2004[149]	Gastric cancer			✓	N/R					N/R	N/R
Wildi *et al.*, 2004[152]	Gastric cancer	✗	N/R	✓	✗						
O'Gorman *et al.*, 1998[151]	Gastric cancer					✓	✓	Moderate			
Shenfine *et al.*, 2009[150]	Gastric cancer					✓	✓				
Verschuur *et al.*, 2009[154]	Gastric cancer									✓	N/R

Study reference grouped by measure (author, year)	Cancer	Known group (severity)		Known group (case–control)		Known group (other)			Responsiveness		Reliability
		Consistent evidence	Significant	Consistent evidence	Significant	Consistent evidence	Significant	Correlation	Consistent evidence	Significant	Consistent evidence
Kontodimopoulos *et al.*, 2009[147]	Gastric cancer								✓		
McMillan *et al.*, 1999[153]	Gastric cancer									✓	✓
Rogers *et al.*, 2006[148]	Gastric cancer	✓	✓	✓	N/R			Low to moderate			
Castellano *et al.*, 2009[172]	Kidney cancer							Moderate	✓	✓	
Cella *et al.*, 2008[169]	Kidney cancer					✓	N/R		✓	✓	
Cella *et al.*, 2010[168]	Kidney cancer								✓	N/R	✓
Sternberg *et al.*, 2010[193]	Kidney cancer					✓	N/R				
Yang *et al.*, 2010[170]	Kidney cancer					✓	N/R		✓	N/R	
Hahn *et al.*, 2003[176]	Leukaemia cancer	✓	✓							✓	N/R
Langenhoff *et al.*, 2006[180]	Liver cancer					✓	N/R		✓	N/R	
Mendez Romero *et al.*, 2008[172]	Liver cancer			✓	✓						✓
Krabbe *et al.*, 2004[179]	Liver cancer					✓	N/R		✓	N/R	
Basch *et al.*, 2009[196]	Lung cancer							Low to moderate			
Trippoli *et al.*, 2001[173]	Lung cancer	✓	✓	✓				Moderate to high			

continued

TABLE 14 Overview of performance of EQ-5D, HUI3 and SF-6D in studies of cancer (*continued*)

Study reference grouped by measure (author, year)	Cancer	Known group (severity)		Known group (case–control)		Known group (other)			Responsiveness		Reliability
		Consistent evidence	Significant	Consistent evidence	Significant	Consistent evidence	Significant	Correlation	Consistent evidence	Significant	Consistent evidence
Doorduijn *et al.*, 2005[188]	Lymphoma	✓	✓						✓	✗	
Van Agthoven *et al.*, 2001[177]	Lymphoma								✓	N/R	✓
Witzens-Harig *et al.*, 2009[192]	Lymphoma								✓	✓	✓
Slovacek *et al.*, 2007[181]	ML/acute myeloid leukaemia (AML) cancer					✓	✓				
Slovacek *et al.*, 2008[175]	MM cancer					✓	✓				
Uyl-de Groot *et al.*, 2005[124]	MM cancer					✓	✓		✓	✓	
Slovacek *et al.*, 2007[182]	MM/ML					✓	✓				
Lee *et al.*, 2003[184]	Musculoskeletal cancer							Low to high			High
Mueller-Nordhorn *et al.*, 2006[183]	Pancreatic cancer			✓							
Krahn *et al.*, 2007[160]	Prostate cancer								✓	Mixed evidence	
Sandblom *et al.*, 2004[158]	Prostate cancer	✓	Mixed evidence						✓	Mixed evidence	
Shimizu *et al.*, 2008[156]	Prostate cancer	✓	Mixed evidence					✓			

Study reference grouped by measure (author, year)	Cancer	Known group (severity)		Known group (case–control)		Known group (other)			Responsiveness		Reliability
		Consistent evidence	Significant	Consistent evidence	Significant	Consistent evidence	Significant	Correlation	Consistent evidence	Significant	Consistent evidence
Sullivan *et al.*, 2007[157]	Prostate cancer			✓	Mixed evidence				✓	✓	
Weinfurt *et al.*, 2005[159]	Prostate cancer					✓	Mixed evidence		✓	Mixed evidence	
Falicov *et al.*, 2006[101]	Spinal metastases							Moderate			
Capuano *et al.*, 2008[107]	Non-specific cancer							✓			
Mantovani *et al.*, 2004[111]	Non-specific cancer									✓	✓
Vaghela *et al.*, 2007[112]	Non-specific cancer									✓	✓
Park *et al.*, 2006[192]	Non-specific cancer										
Pickard *et al.*, 2007[103]	Non-specific cancer	✓	N/R								
Pickard *et al.*, 2007[114]	Non-specific cancer	✓	✓					Moderate			
Ravasco *et al.*, 2003[104]	Non-specific cancer	✓	✓							✓	Mixed evidence

continued

TABLE 14 Overview of performance of EQ-5D, HUI3 and SF-6D in studies of cancer (*continued*)

Study reference grouped by measure (author, year)	Cancer	Known group (severity)		Known group (case–control)		Known group (other)			Responsiveness		Reliability
		Consistent evidence	Significant	Consistent evidence	Significant	Consistent evidence	Significant	Correlation	Consistent evidence	Significant	Consistent evidence
Wang *et al.*, 2008[97]	Non-specific cancer			✓	✓						
Weze *et al.*, 2004[110]	Non-specific cancer									✓	✓
Barton *et al.*, 2008[98]	Non-specific cancer			✗	✗						
Cheung *et al.*, 2009[105]	Non-specific cancer	✓	N/R					✓			
Lathia *et al.*, 2008[102]	Non-specific cancer							✗			
Chow *et al.*, 2010[106]	Non-specific cancer	✓	N/R			✓	N/A				
Kim *et al.*, 2008[113]	Non-specific cancer									✓	✓
Norum, 1996 [100]	Non-specific cancer							Moderate to high (significant)			
Korfage *et al.*, 2009[129]	Cancer survivors			✗	✗						
Nijdam *et al.*, 2008[128]	Cancer survivors										✓

Study reference grouped by measure (author, year)	Cancer	Known group (severity)		Known group (case–control)		Known group (other)		Correlation	Responsiveness		Reliability
		Consistent evidence	Significant	Consistent evidence	Significant	Consistent evidence	Significant		Consistent evidence	Significant	Consistent evidence
HUI2/HUI3											
Le Gales *et al.*, 1999[163]	Brain cancer	Mixed evidence	✓			✓	✓				
McCarter *et al.*, 2006[162]	Brain cancer	✓	N/R	✓	✓			Moderate to high			
Chang *et al.*, 2004[143]	Breast cancer							✓ Strong	✓	✓	
Lovrics *et al.*, 2008[141]	Breast cancer							✓ Moderate to strong	✓	✓	
Polsky *et al.*, 2002[142]	Breast cancer								✓	✓	
Ramsey *et al.*, 1998[190]	Colon cancer	✓	N/R						✓	✓	
Klaassen *et al.*, 2010[186]	Hodgkin's lymphoma							Moderate to high	✓	✓	
Klaassen *et al.*, 2010[185]	Hodgkin's lymphoma										Generally substantial agreement
Barr *et al.*, 1997[174]	Leukaemia								✓	Mixed evidence	✓ (inter-rater reliability)
Cox *et al.*, 2005[187]	Leukaemia							Reported accept ability			

continued

TABLE 14 Overview of performance of EQ-5D, HUI3 and SF-6D in studies of cancer (*continued*)

Study reference grouped by measure (author, year)	Cancer	Known group (severity)		Known group (case–control)		Known group (other)			Responsiveness		Reliability
		Consistent evidence	Significant	Consistent evidence	Significant	Consistent evidence	Significant	Correlation	Consistent evidence	Significant	Consistent evidence
Banks *et al.*, 2008[178]	ML/acute myeloid leukaemia (AML)							Low to high (baseline) Low to moderate (change scores)	✓	N/R	Moderate: substantial agreement
Krahn *et al.*, 2007[160]	Prostate cancer								✓	Mixed evidence	
Krahn *et al.*, 2003[161]	Prostate cancer										✓
Albertsen *et al.*, 1998[155]	Prostate cancer							Low			
Falicov *et al.*, 2006[101]	Spinal metastases							Moderate	✓ (pain only)	✓	
Bowker *et al.*, 2006[99]	Non-specific cancer			✓	✓						
Sung *et al.*, 2003[108]	Non-specific cancer							Low to moderate			
Trudel *et al.*, 1998[109]	Non-specific cancer			✓	✓						✓
Barr *et al.*, 2000[133]	Cancer survivors					✓	Mixed evidence				Substantial agreement
Boman *et al.*, 2009[134]	Cancer survivors			✓	✓	✓	Mixed evidence				Moderate: substantial agreement

Study reference grouped by measure (author, year)	Cancer	Known group (severity)		Known group (case–control)		Known group (other)			Responsiveness		Reliability
		Consistent evidence	Significant	Consistent evidence	Significant	Consistent evidence	Significant	Correlation	Consistent evidence	Significant	Consistent evidence
Felder-Puig *et al.*, 2000[131]	Cancer survivors	✓	✗								Moderate: substantial agreement
Fu *et al.*, 2006[130]	Cancer survivors					✓	Mixed evidence				Low: substantial agreement
Barr *et al.*, 1999[127]	Cancer survivors					✓	Mixed evidence				Substantial agreement
Pogany *et al.*, 2006[132]	Cancer survivors			✓	✓						
Grant *et al.*, 2006[135]	Cancer survivors					Mixed evidence	Mixed evidence				
Nixon Speechley *et al.*, 1999[136]	Cancer survivors							Moderate to high			
Shimoda *et al.*, 2005[126]	Cancer survivors	✓	✓								Substantial agreement
SF-6D											
Kontodimopoulos *et al.*, 2009[147]	Gastric cancer								✓		
Shimizu *et al.*, 2008[156]	Prostate cancer	✓	N/R					✓			
Barton *et al.*, 2008[98]	Non-specific cancer			✗	✗						

N/R, not reported.

Chapter 3 Mapping to EQ-5D

Introduction

The review of the performance of GPBMs in the previous chapter showed that EQ-5D is a valid and responsive measure for patients with cancer. Despite these findings, many cancer studies do not include the EQ-5D and are more likely to include one of two cancer-specific questionnaires: the EORTC QLQ-C30 or the FACT-G. Five studies have previously mapped between EORTC QLQ-C30 and EQ-5D.[146,147,197–199] Four of these functions are not necessarily applicable to other samples,[146,147,197,198] Versteegh *et al.*[197] fail to provide the mapping function for other researchers to use and the sample used by Crott and Briggs[146] includes only female patients. Wu *et al.*[198] require data on both the FACT-G and the EORTC QLQ-C30 to produce mapped estimates, although studies may not routinely collect both of these together. Kontodimopoulous *et al.*[147] use a linear regression model to predict EQ-5D scores; however, they state that the model does not produce reliable predictions and is based on a small sample. Potentially the most useful mapping function was published by McKenzie and van der Pol,[199] who produced two mapping functions; the first used linear regression to estimate EQ-5D index scores and gave reasonable predictions and the second used ordered probit models to predict EQ-5D dimension levels and gave poor predictions. Other models such as tobit and TPMs were not explored by any authors but may predict EQ-5D values more accurately, and this needs to be explored further. Only one mapping function has been published using FACT-G data to predict EQ-5D values; it fitted ordinary least squares (OLS) and CLAD models at the domain level and showed that scores were poorly predicted away from the mean.[105]

The aims of this chapter are (1) to estimate mapping functions using two cancer-specific HRQL measures, the EORTC QLQ-C30 and FACT-G, to the EQ-5D for use in future studies and (2) to test the applicability of different mapping approaches that have been used in the literature in order to provide recommendations for future mapping studies. In particular, the analysis was aimed at providing comprehensive information on how to select the mapping function and information on uncertainties around the predictions. We assessed different modelling techniques that have been applied in the literature and used standard criteria to identify the most appropriate mapping functions. We also provide information on uncertainty.

Methods

Measures

Target measure: EQ-5D

Our target measure for mapping was the EQ-5D.

Source measures

The cancer data sets included two widely used cancer-specific measures and these were selected as the source measures: EORTC QLQ-C30 and the FACT-G.

The EORTC QLQ-C30 is a cancer-specific HRQL measure that has been found to be valid for many cancer conditions and has been widely used in cancer clinical trials across Europe and Canada.[200] The EORTC QLQ-C30 has 30 items, 28 with four levels (not at all, a little, quite a bit and very much) and two items (overall health and overall QoL) with seven levels (ranging from very poor to excellent). The items cover five functioning scales (physical, role, social, emotional and cognitive functioning), plus a global QoL scale and nine symptoms scales (fatigue, nausea and vomiting, pain, dyspnoea, sleep disturbance, appetite loss, constipation, diarrhoea, financial impact). Each summary scale ranges from 0 to 100. Higher scores for the functioning and global QoL scales indicate higher functioning levels, whereas higher scores for the symptoms scales indicate higher symptom levels. Mapping functions were developed using the dimensions scores and items.

The FACT-G has been shown to be a reliable and validated HRQL measure.[201] The questionnaire consists of 27 items in four subscales (physical well-being, social/family well-being, emotional well-being and functional well-being). Each item has a range of five options ranging from not at all (score 0) to very much (score 4) and item scores are added to form a subscale score and subscale scores are added to form a global score. Global scores can range from 0 to 108. Mapping functions were developed using the total score, dimension scores and items.

Data sets

Four data sets were used for the mapping study; three contained the EORTC QLQ-C30 and EQ-5D while one contained the FACT-G and EQ-5D. The three data sets containing EORTC QLQ-C30 were pooled into a single data set.

European Organization for Research and Treatment Quality-of-life Questionnaire Core 30

One EORTC QLQ-C30 data set came from a randomised trial [Velcade as Initial Standard Therapy (VISTA)][202] while the other two data sets came from a cancer clinic. The VISTA data were collected in a Phase III randomised open-label trial for patients newly diagnosed with MM. Patients were requested to complete both the EQ-5D and EORTC QLQ-C30 at their screening visit, day 1 of each of the nine cycles of treatment, at the end of each treatment visit and during the post-treatment phase (every 6 or 8 weeks) until disease progression. For the mapping analysis, only responses at screening visit were used. The mean age of the screening sample was 72 years (SD 5.5 years) and 50% were male. Severity was measured using the International Staging System for Multiple Myeloma, according to which patients are classed as having stage I disease if serum beta-2-microglobulin ($S\beta_2M$) is < 3.5 mg/l and serum albumin ≥ 3.5 g/dl (median survival 62 months). Patients are classed as having stage II disease if they do not meet the criteria for stages I or III and as having stage III disease if $S\beta_2M \geq 5.5$ mg/l.[203]

The other data were collected at the Vancouver Cancer Clinic. Women diagnosed with breast cancer and attending an outpatient clinic were asked to complete EQ-5D and EORTC QLQ-C30. The mean age of the full sample was 68 years (SD 18.2 years). Severity was measured using the stage of disease, with stage I indicating that the cancer is localised and stage IV indicating that cancer has metastasised or spread to other areas of the body. Patients diagnosed with lung cancer attending an outpatient clinic were also asked to complete EQ-5D and EORTC QLQ-C30. The mean age of the full sample was 62 years (SD 21.1 years) and 48% were male. As with the data set from patients with breast cancer, severity was measured using the stage of disease.

Functional Assessment of Cancer Therapy – General Scale

The FACT-G data set contained 538 cases from USA of which 530 provided self-reported data on HRQL. Participants were from a validation survey of different cancer scales and had one of 11 cancers at stage 3 or 4 and had undergone at least two cycles of chemotherapy, for non-cyclical treatments, and had received treatment for more than 1 month.[189] Participants completed the EQ-5D (both the three- and five-level versions), FACT-G and ECOG performance measures, cancer and treatment distress scale, FACT-G cancer disease-specific add-on questions, the renal cell carcinoma symptom index and the symptom checklist for depression and anxiety. For the mapping study, we focus on mapping between EQ-5D and FACT-G and use the ECOG performance status measure as a measure of cancer severity. The sample consisted of 273 (52%) male patients and 255 (48%) female patients with an average age of 59 years (SD 11.9 years, range 24–88 years).

Data analysis

Preliminary analysis

Spearman's rank correlations of the independent variables were used to determine whether any variables were highly correlated and therefore not recommended for inclusion in the same regression model. A high correlation was defined as a correlation coefficient > |0.7|.[204] Spearman's rank correlations were also used to determine correlations between the dependent and independent variables to inform model specification and this was undertaken for the EQ-5D utility values and dimension levels and the total scores, dimensions scores and items of the EORTC QLQ-C30 and the FACT-G. The distribution of the EQ-5D was also examined to determine the distribution of the scores and whether this differed by data set. This was used to determine the appropriate model specifications for the regression equations mapping the two cancer measures onto EQ-5D.

Specification

The mapping analysis involves using regression techniques to estimate the relationship between the EQ-5D and the cancer-specific measures. The relationship can be specified in different ways. The simplest additive model regresses the EQ-5D onto the global score of the starting measure, for example, the FACT-G global score. This specification assumes that all the items/dimensions contributing to the global score have equal weight and response choices to each item lie on a similar interval scale (e.g. the intervals between 'all of the time', 'most of the time' and 'some of the time', etc., are equal). These assumptions can be relaxed by including dimension scores and item responses as independent variables. We assessed global scores, dimension scores and item responses for each cancer measure, where appropriate. Global and dimension scores were treated as continuous variables and item responses were modelled as discrete dummy variables.

We included squared terms for dimensions that displayed non–linear relationships. We also tested the inclusion of interaction terms where there was evidence of correlations between dimensions. We tested for the inclusion of interaction terms for the dimension scores based on high correlations (> |0.7|). Squared and interaction terms were not included for item models.

Modelling techniques

Models were fitted to the overall EQ-5D score using linear regressions estimated by OLS, tobit models, TPMs and splining. Further models were fitted to the individual dimensions of the EQ-5D using response mapping. A limited dependent variable mixture model (LDVMM) was also used in an illustrative analysis.

Ordinary least squares

The most common model used in the literature for mapping between QoL instruments is OLS, which assumes that the relationship between the dependent variable (EQ-5D index values) and the independent variable(s) (EORTC QLQ-C30 or FACT-G) can be expressed as a linear function of the parameters. OLS models are typically able to predict the mean scores but are poor at predicting those in poor health and full health.

Tobit model

Ordinary least squares does not allow for the fact that the EQ-5D is bounded at –0.594 at the bottom and 1 at the top of the scale and thus predictions could be greater than 1 or less than –0.594. The tobit model can be used to take into account the upper and lower limits of EQ-5D so predictions are limited to the credible range.

Two-part model

The TPM uses a combination of two different model types to predict different parts of the distribution of the data. These have been used in cost analysis to predict whether resource use is incurred (see Lipscomb *et al.*[205] for example) and in mapping, where logistic regression is applied to model the probability of whether responders are in full health or not and OLS or another suitable model used to model scores less than full health. The results from the two parts of the model are combined to obtain an overall score. We fitted a logistic regression model to estimate the probability of being in full health (yes/no) and a truncated OLS model to predict EQ-5D score if not in full health, where for the truncated OLS model scores cannot exceed a value of 1.[206] Predicted EQ-5D scores were calculated as follows, where FH is full health:

$$\text{Expected(EQ-5D)} = \text{probability(FH)} + \{\text{predicted EQ-5D score if not FH} * [1-\text{probability(FH)}]\} \tag{1}$$

Splining

One of the issues in mapping to EQ-5D scores is that they rarely follow or approximate to the normal distribution. Transformations can be used to account for this but another option is to use splining to identify changes (cut points) in the distribution of the data and to model these changes using different mathematical functions; this approach is also known as fractional polynomials. The first stage of the process is to identify possible cut-off values, which was done using the multivariable fractional polynomials function in Stata version 12, StataCorp LP, College Station, TX, USA,[207] which fits all possible polynomial functions to the data and identifies the best-fitting model. We applied splining functions to the best-fitting dimension-based models to test whether splines offered an improvement over including squared terms in our models.

Response mapping

An alternative to modelling the EQ-5D index is to fit models to the dimensions of the EQ-5D using ordinal or multinomial logistic regression models known in the literature as response mapping.[208,209] We fitted multinomial logistic regression models to each of the five dimensions of the EQ-5D. Using an approach previously reported in the response mapping literature,[210] the expected value of the EQ-5D was then calculated by multiplying the probability of being in each response level by the standard UK tariff.[4]

$$\begin{aligned}\text{Expected(EQ-5D)} = {} & 1-(\text{Prmob2}\times 0.069)-(\text{Prmob3}\times 0.314)-(\text{Prcare2}\times 0.104)-(\text{Prcare3}\times 0.214)\\ & -(\text{Pruact2}\times 0.036)-(\text{Pruact3}\times 0.094)-(\text{Prpain2}\times 0.123)-(\text{Prpain3}\times 0.386)\\ & -(\text{Pranx2}\times 0.071)-(\text{Pranx3}\times 0.236)-(1-\text{PrPerfect})\times 0.081-\text{PrN3}\times 0.269\end{aligned} \tag{2}$$

where Prmob2 is the probability of being in mobility level 2 on EQ-5D, Prmob3 is the probability of being in mobility level 3 on EQ-5D, Prcare2 is the probability of being in self-care level 2 on EQ-5D, Prcare3 is the probability of being in self-care level 3 on EQ-5D, Pruact2 is the probability of being in usual activities level 2 on EQ-5D, Pruact3 is the probability of being in usual activities level 3 on EQ-5D, Prpain2 is the probability of being in pain or discomfort level 2 on EQ-5D, Prpain3 is the probability of being in pain or discomfort level 3 on EQ-5D, Pranx2 is the probability of being in anxiety or depression level 2 on EQ-5D and Pranx3 is the probability of being in anxiety or depression level 3 on EQ-5D. PrN3 is the probability of any of EQ-5D dimensions being at level 3.

$$\begin{aligned}&\text{PrPerfect is the probability of being in perfect health}\\ &\quad = \text{Prmob1}\times\text{Prcare1}\times\text{Pruact1}\times\text{Pr pain1}\times\text{Pranx1 and PrN3 is the probability of being}\\ &\quad \text{in level } 3 = 1-(1-\text{Prmob3})\times(1-\text{Prcare3})\times(1-\text{Pruact3})\times(1-\text{Prpain3})\times(1-\text{Pranx3})\end{aligned} \tag{3}$$

where Prmob1 is the probability of being in mobility level 1 on EQ-5D, Prcare1 is the probability of being in self-care level 1 on EQ-5D, Pruact1 is the probability of being in usual activities level 1 on EQ-5D, Prpain1 is the probability of being in pain or discomfort level 1 on EQ-5D and Pranx1 is the probability of being in anxiety or depression level 1 on EQ-5D.

Limited dependent variable mixture model

A further model was fitted, the LDVMM. Although the models described in the preceding section for modelling the index of EQ-5D are widely used in the literature, they have been shown to be inappropriate in several studies as they are unable to take into account the characteristics of EQ-5D data and their distribution across individuals.[211–213] These characteristics include the bounded nature of the EQ-5D data, a large proportion of respondents at 1 (full health), a large gap between this top value and the next allowable EQ-5D value and the multimodality of the distribution. These are the features that the standard models are unable to generate and has led to the development of new, more advanced models, one of which is the LDVMM of Hernández Alava *et al.*[211,212] Finite mixture models provide a very flexible semiparametric framework in which to model complex nonstandard distributions in cases where standard models are unable to provide a satisfactory model for all the data. By combining several distributions (also referred to as components) using probability weights, mixture models can approximate any distribution arbitrarily well and are able to generate characteristics such as skewness and multimodality. These probability weights can be functions of any relevant variables. Thus, the covariates in these models can determine EQ-5D directly by inclusion in the individual components but also indirectly through their effect in the probability of component membership. This flexibility generates a rich and complex pattern of relationships between the explanatory variables and EQ-5D where the same variable can be highly significant in certain components but not in others and can also have an independent indirect effect through its significance in the probability of component membership. Insignificance of variables in standard models (i.e. models with only one component) may be the result of differing patterns of significance across components and might lead to the erroneous exclusion of variables under usual practice. The LDVMM combines the flexibility of the mixture model approach with specially designed components that are limited at 1 (full health) and at –0.594 and have an adjustment to generate the gap in feasible values of the UK EQ-5D tariff between 1 and 0.883. For a more technical description of this model, see Hernández Alava *et al.*[211] The LDVMM was fitted to the FACT-G data set using domain level covariates to illustrate new model developments in this area which take into account the idiosyncrasies of EQ-5D data.

Model specification

Models were fitted using backwards regression where all possible variables are included in the model and the least significant removed until only significant variables remain ($p < 0.1$), except in the implementation of the LDVMM. To avoid overfitting models, we use the rule of 10 participants per variable for continuous models and 10 events for the smallest category for response mapping models. When variables were highly correlated, the variable that was most likely to map to the EQ-5D was selected, based on the analyst's judgement. Standard errors (SEs) of regression co-efficents were calculated from bootstrap estimates and 5000 bootstrap samples were run for each model.

Insignificant variables were not automatically dropped in the LDVMM analysis. This process of data mining increases the risk of fitting a model to the specific sample data set being used but that lacks generalisability. It leads to an estimated model with an improved in-sample fit but tends to perform poorly out of sample. This is particularly important when the number of observations is relatively small, as in the present case, since often these data sets present many idiosyncrasies not seen in larger samples. The aim of the LDVMM analysis was to fit a model that predicted well in sample and that captured the general characteristics of EQ-5D data sets but at the same time avoided 'fitting the model to the data' in excess.

Model goodness of fit

Model goodness of fit was measured using AIC and BIC, where the smaller the value, the better the model fit. For each model, we also reported the model RMSE and mean absolute error (MAE). For OLS models, we reported the R^2 and adjusted R^2 and used the Ramsey Regression Equation Specification Error Test (RESET) to test non-linear combinations of variables in the model. For tobit, logistic regression and Response mappings, we used the pseudo-R^2. Sigma was reported for the tobit and truncated regression models and the link test was used to check model specification. The Hosmer–Lemeshow test was used to assess goodness of fit for logistic regression models.

Model performance and discrimination

Summary statistics including mean and range were examined to assess overall model predictions. However, a more stringent test was applied using a severity measure to assess the discriminative performance of the predicted EQ-5D score. For FACT-G, respondents were asked a variation of the ECOG performance status. ECOG has five categories ranging in severity from 0 to 4 (worst)[214] and five response categories: normal activity without symptoms, some symptoms but do not require bed rest during the waking day, require bed rest for less than 50% of the waking day, require bed rest for over 50% of the waking day and unable to get out of bed. No patients were in the most severe level (unable to get out of bed) and few patients [*n* = 21 (4%)] required bed rest for more than 50% of the waking day; therefore, these two categories are merged with the level do not require bed rest less than 50% of the waking day. The ECOG responses are included in mapping models as a measure of disease severity and to test the predictive ability of the mapping models across different severity groups. There was no common severity measure in the EORTC QLQ-C30 data sets and the item reporting health status was used instead. Response options ranged from poor (1) to excellent (7). Discriminative ability across severity groups using these measures was tested using ANOVA. MAEs were reported for each subgroup.

Model performance was also assessed visually by plotting observed and predicted EQ-5D values by health state. As a further comparison for the LDVMM, EQ-5D data sets were simulated using each model in turn as the data generating process based on 100 replications per individual in the sample for a total of 53,000 simulated EQ-5D data points. Only one data set per model was generated and, therefore, small variations for different generated errors can be expected for the individual simulated data points; however, enough simulations have been generated to ensure an accurate overall distribution. Plots of the observed EQ-5D distribution in the data were compared with distributional plots of the simulated data sets. A model that correctly fits the data should generate a distribution of simulated values which displays similar characteristics to the observed EQ-5D distribution in the data.

Model validation

Internal model validation was carried out using bootstrapping to estimate a shrinkage factor. We used the bootstrapping techniques reported by Steyerberg *et al.*[215] to assess all models (except in the implementation of LDVMM) and shrinkage coefficients are reported in order to counter overoptimism of estimates.[215] Five thousand bootstrap estimates were run to calculate shrinkage factors. A shrinkage coefficient of less than 1 (typical value expected for a shrinkage coefficient) reflects an 'overfitting' of the data.

Model selection

When producing a mapping model, the factors that are important in selecting a model are accuracy of the predicted mean and SE, MAE, shrinkage and the reproducibility of the model across different severity states. Mapping and model fitting literature does not suggest a single criteria for use in selecting the best-fitting model and the criteria that we might focus on when selecting a model may depend on what we want the mapping function to achieve. For each type of model (OLS, tobit, etc.) we gave equal weighting to all model selection and performance statistics and ranked across models based on these statistics, a mean rank per model was then estimated. The model with the best mean ranking was selected. The best-performing models per model type were then compared and ranked to select the best overall model. In the event of there being no clear difference between models, we gave priority to models that best estimated the mean and were able to discriminate across disease severity.

Table 15 presents an overview of the analysis that was carried out. For each modelling technique (with the exception of LDVMM) we assessed the performance of a series of model specifications based on overall cancer instrument score, dimensions scores, dimensions scores plus squared and square root terms, dimensions scores plus squared, square root terms and interations, item level models and the best fitting of these models plus patient characteristics.

Mapping models that we fitted between EORTC QLQ-C30 or FACT-G and EQ-5D were:

Model 1 EQ-5D Index = Global Index Score (FACT-G only).

Model 2 EQ-5D Index = All dimensions.

TABLE 15 Overview of analysis

Dependent variables	Independent variables	Model selection and specification	Model type	Performance	Validation	Uncertainty
EQ-5D index	EORTC QLQ-C30	Used standard statistical techniques to examine the data prior to mapping estimation (including frequency tables and correlations)	Linear OLS	*Goodness of fit*	Application and assessment of mapping algorithm was validated using bootstrapping	Estimate for best performing model using probabilistic sensitivity analysis (based on regression estimates and correlation matrix)
EQ-5D dimension levels	Dimension summary scores		Tobit	Statistical significance, sign and size of coefficients		
	+ interaction terms		TPM			
	+ squared terms	Fully described the data set used to estimate the regression model including both range of EQ-5D and plots showing EQ-5D distribution, to determine whether unimodal/bimodal/ trimodal or skewed	SPL			
	+ square root terms		Response mapping	R^2, adjusted R^2 and pseudo-R^2		
	Item level models			AIC and BIC		
	Sociodemographic variables		LDVMM (FACT-G dimension summary scores only)	Further tests of model fit such as Ramsey RESET		
	FACT-G					
	Total score			*Predictive ability*		
	Dimension summary scores			MAE, MAE by subset of severity range of EQ-5D and/or predictive measure(s)		
	+ interaction terms					
	+ squared terms					
	+ square root terms			Plots of observed and predicted EQ-5D scores		
	Item level models					
	Sociodemographic variables					

SPL, splining.

Model 3 EQ-5D Index = Significant dimensions only.

Model 4 EQ-5D Index = Significant dimensions, squared and square root terms.

Model 5 EQ-5D index = Significant dimensions, squared, square root and interaction terms.

Model 6 EQ-5D index = Significant items.

Model 7 EQ-5D index = Significant items collapsed item levels.

Model 8 best performing mode selected from Models 1 to 7 above plus significant patient and disease characteristics.

Models 6 and 7 were not fitted for splining as this is performed on continuous variables. Response mapping fitted models to each of the EQ-5D domains rather than the EQ-5D index. We assessed model performance by assessing models across these specifications for each modelling technique to select the best-fitting model specification. We then used the same criteria to compare the best-fitting models across the modelling techniques. LDVMM were fitted only for the FACT-G dimension scores.

Representing for uncertainty in mapping methods

Probabilistic sensitivity analysis was used to allow for uncertainty in mapping coefficients for the best performing FACT-G model. Regression coefficients were assumed to follow a normal distribution and the covariance matrix for the model was used to allow for variability and correlations between variables. It was necessary to run 100,000 simulations to obtain convergence to a mean across simulations. For each simulation mean, the EQ-5D score was calculated and percentiles were used to summarise the variability around the mean estimate.

Results

European Organization for Research and Treatment Quality-of-life Questionnaire Core 30 Preliminary analysis

Table 16 shows the characteristics of the full sample and for each data set for those with complete data. Mean age and proportion of males varied by data set. The breast cancer data set had the lowest mean age and contained only females, and the MM data set had the highest mean age and the highest proportion of males. The mean EQ-5D score also varied by data set, the MM data set had a mean EQ-5D value of 0.519 whereas the breast and lung cancer data sets had higher mean EQ-5D values of 0.765 and 0.742, respectively. Only the MM data set covered the entire range of the EQ-5D and had fewer ceiling effects than the other data sets, with 8% of responses at full health on EQ-5D, compared with 24% and 17% for the breast and lung cancer data sets, respectively. *Figure 5* shows the histograms of the EQ-5D index for each data set and the combined data set showing that the distributions differ by data set but without further information we cannot conclude whether this is differences in the severity of the patients in each data set or differences in the pattern of EQ-5D by condition. Separate assessment of the scores for the EORTC QLQ-C30 scales most noticeably varied across the three data sets for physical functioning, role functioning, pain, dyspnoea, constipation and global QoL (see *Table 16*).

Assessment of the correlations between the independent variables indicated that the highest correlations were between role functioning, physical functioning and fatigue variables (see *Appendix 12*). Assessment of the correlations between the EORTC QLQ-C30 summary scales and EQ-5D dimensions and utility score indicated that overall physical functioning, role functioning, social functioning, fatigue, pain and global QoL were most highly correlated with EQ-5D dimensions and score. However, as global QoL is likely to encompass the other conceptual domains, it is theoretically preferable to exclude this from consideration in the mapping function alongside the other summary scale variables. Correlations between the EORTC

TABLE 16 Characteristics of the patient samples

	All cancers (*n* = 771)		Breast cancer (*n* = 100)		Lung cancer (*n* = 99)		MM (*n* = 572)	
Variables	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Age (years)	68	9.6	54	10.9	63	11	72	5.4
Male (%)	44		0		48		50	
EQ-5D								
EQ-5D utility score	0.579	0.342	0.765	0.202	0.742	0.199	0.519	0.360
Proportion reporting EQ-5D = 1 (%)	11		24		17		8	
Range of EQ-5D	–0.594 to 1		–0.144 to 1		0.088 to 1		–0.594 to 1	
EORTC QLQ-C30 dimensions								
Physical functioning	65	25.6	78	19.9	70	19.6	62	26.5
Role functioning	59	33.2	73	27.7	68	27.0	55	34.2
Emotional functioning	70	24.9	73	22.7	76	21.5	68	25.6
Cognitive functioning	76	22.7	77	22.8	77	20.5	76	23.1
Social functioning	69	29.8	72	26.2	74	23.8	68	31.3
Fatigue[a]	45	26.2	39	20.9	43	23.1	47	27.3
Nausea[a]	9	17.9	11	19.9	10	16.8	8	17.7
Pain[a]	40	33.0	23	24.3	23	23.5	47	33.5
Dyspnoea[a]	25	29.0	17	22.5	37	30.7	24	29.1
Sleep disturbance[a]	33	32.6	34	31.1	31	28.3	33	33.6
Appetite loss[a]	27	32.5	20	28.5	29	32.3	29	33.1
Constipation[a]	23	30.7	12	23.4	23	30.0	25	31.6
Diarrhoea[a]	10	19.9	16	27.0	11	20.3	8	18.4
Financial impact[a]	20	28.8	24	30.5	23	28.8	19	28.4
Global QoL	53	23.2	68	18.2	62	21.0	48	22.8

a Higher scores for symptom scales indicate worse symptoms. EORTC QLQ-C30 dimension score range 0–100, higher scores indicate better functioning and QoL.

QLQ-C30 item levels by domains indicated that items within physical, role, emotional and social functioning, QoL, fatigue and pain were highly correlated, suggesting that not all items within these domains need to be selected for item level models.

European Organization for Research and Treatment Quality-of-life Questionnaire Core 30 Mapping Analysis Results

Selecting models

We illustrate how the best performing model was selected using the OLS results for the EORTC QLQ-C30. *Table 17* summarises the predicted EQ-5D scores and model performance of the six models that were undertaken. Physical, role and emotional functioning dimensions were statistically significant and positive as expected. Pain and sleep disturbance were statistically significant and negative as expected but dyspnoea was positive. Inclusion of squared terms improved the model (model 4) but interactions (model 5) had no impact and results from these were therefore not reported. Items related to dimensions

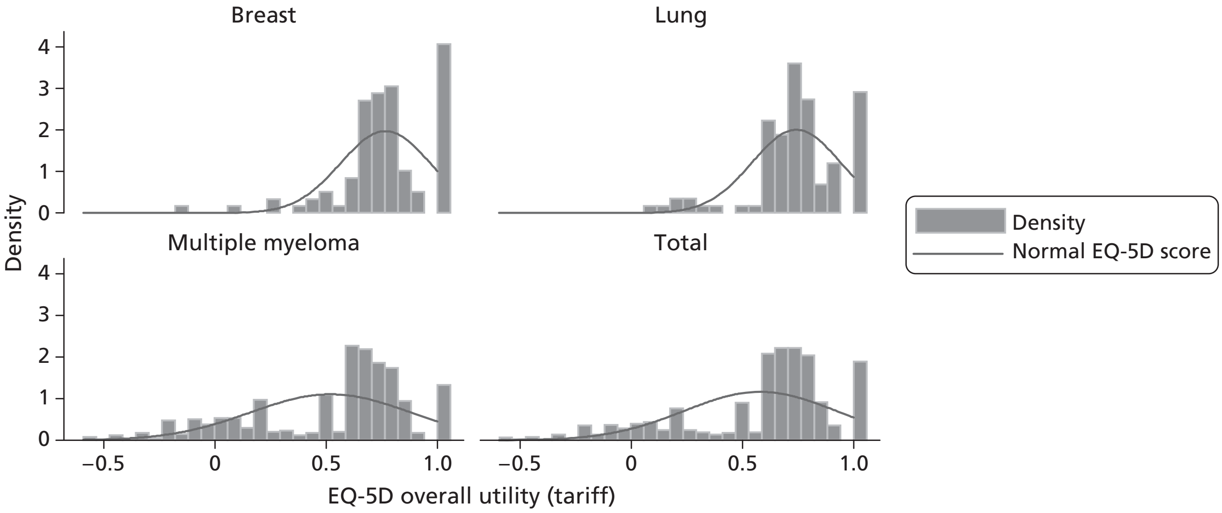

FIGURE 5 Histogram of EQ-5D utilities: All data sets.

of physical, role, cognitive, emotional and social functioning and fatigue, pain, sleep disturbance, appetite loss and constipation symptoms were statistically significant. Collapsing unordered levels (model 7) did not improve the results; however, including age improved the results (model 8).

Model performance statistics indicate that item models consistently performed better than the domain-level models. All the models tended to underpredict EQ-5D scores for those in near perfect or full health and overpredicted those in poorer health (*Figure 6*). Dimension level models predicted individuals in full health, with values above 1, but item-level models did not. At the severe end of health, item-level models performed better (see *Figure 6*). The models were able to discriminate across severity (see *Table 17*). There was some evidence that the error was associated with severity, with higher MAE for poor health compared with excellent health.

Table 18 presents the ranking for each of the models performance statistics. Model 8, the item model with age, was the best-fitting model (mean rank = 2.08) although it did not predict any EQ-5D scores in full health. Domain-level models without squared terms and interactions gave the poorest estimates.

Best-fitting models – European Organization for Research and Treatment Quality-of-Life Questionnaire Core 30

The process described above for EORTC QLQ-C30 was repeated for the tobit, two-part, splining and response mapping models, the results are presented in *Appendix 12* and are summarised here (*Table 19* and *Figure 7*).

The results for tobit models were similar to OLS models in terms of the dimensions that were significant and model performance statistics with item-level models performing better than the domain-level models. The best performing model was model 8, which was the item-level model with age included. OLS and tobit models were best at predicting the mean EQ-5D value. All the TPMs overpredicted mean scores and median values were lower than the observed values (see *Appendix 12*). Model 8, i.e. the item-level model (model 6) with age, was the best-performing model and was best at predicting the median EQ-5D values.

Only one splining mode was fitted to EORTC QLQ-C30 data for significant domain scores as identified in OLS model 3 a single spline was included for physical functioning at a score of 47. This model did not perform better than the best OLS (model 8), but had the least deviation from the shrinkage coefficient of 1.

For response mapping, it was necessary to collapse EORTC QLQ-C30 items into two levels (no problem and any problem). The mean and median EQ-5D predicted values were lower than the observed values (see *Appendix 12*). Response mapping models were able to discriminate between different severity groups and predicted scores were associated with level of severity. The best-fitting model was model 8 (the domain

TABLE 17 European Organization for Research and Treatment Quality-of-life Questionnaire Core 30 mean observed and predicted EQ-5D values per model and summary model performance: OLS

Summary statistics and model performance tests	*n*	Observed values	OLS model 2	OLS model 3	OLS model 4	OLS model 6	OLS model 7	OLS model 8
			All dimensions	Significant dimensions	Significant + squared terms	Significant items	Significant items + collapsed	Significant items + age
Mean (SD)	771	0.5793 (0.3423)	0.5793 (0.2797)	0.5793 (0.2792)	0.5793 (0.2830)	0.5793 (0.2863)	0.5793 (0.2844)	0.5793 (0.2866)
Median		0.6910	0.6281	0.6244	0.6451	0.6498	0.6557	0.6502
Range		–0.5940 to 1	–0.1846 to 1.02	–0.1915 to 1.031	–0.3712 to 0.9419	–0.4078 to 0.9713	–0.3670 to 0.9430	–0.4046 to 0.9714
R^2			0.668	0.665	0.684	0.700	0.691	0.701
Adjusted R^2			0.662	0.662	0.681	0.689	0.683	0.690
AIC			–286	–294	–340	–338	–330	–339
BIC			–216	–257	–307	–207	–237	–205
Ramsey RESET			$F_{3,753} = 12.57$, $p < 0.001$	$F_{3,761} = 13.09$, $p < 0.001$	$F_{3,761} = 1.56$, $p = 0.198$	$F_{3,737} = 1.00$ $p = 0.3945$	$F_{3,736} = 0.58$, $p = 0.6310$	$F_{3,736} = 0.88$, $p = 0.449$
MAE			0.149	0.151	0.143	0.139	0.142	0.139
Shrinkage			0.836	0.996	0.997	1.060	1.072	1.042

continued

TABLE 17 European Organization for Research and Treatment Quality-of-life Questionnaire Core 30 mean observed and predicted EQ-5D values per model and summary model performance: OLS (*continued*)

Summary statistics and model performance tests	*n*	Observed values	OLS model 2		OLS model 3		OLS model 4		OLS model 6		OLS model 7		OLS model 8	
			All dimensions		Significant dimensions		Significant + squared terms		Significant items		Significant items + collapsed		Significant items + age	
Health status	**n**	***Mean***	***Mean***	***MAE***	***Mean***	***MAE***	***Mean***	***MAE***	***Mean***	***MAE***	***Mean***	***MAE***	***Mean***	***MAE***
1 (very poor)	42	–0.0057	0.1213	0.221	0.1212	0.224	0.0638	0.201	0.0636	0.206	0.0685	0.210	0.0642	0.205
2	53	0.1763	0.2571	0.193	0.2631	0.194	0.2569	0.191	0.2471	0.181	0.2590	0.194	0.2470	0.179
3	144	0.4286	0.4403	0.193	0.4410	0.195	0.4577	0.189	0.4650	0.181	0.4684	0.183	0.4629	0.182
4	226	0.6220	0.5685	0.154	0.5661	0.155	0.5839	0.145	0.5829	0.137	0.5794	0.141	0.5823	0.138
5	186	0.7180	0.7145	0.112	0.7158	0.113	0.7179	0.108	0.7170	0.109	0.7147	0.109	0.7176	0.109
6	94	0.8321	0.8494	0.110	0.8470	0.109	0.8165	0.102	0.8145	0.100	0.8148	0.103	0.8181	0.098
7 (excellent)	26	0.9029	0.8958	0.073	0.9008	0.075	0.8538	0.086	0.8553	0.081	0.8504	0.080	0.8546	0.080
ANOVA		$F_6 = 97$, $p < 0.001$	$F_6 = 126$, $p < 0.001$		$F_6 = 126$, $p < 0.001$		$F_6 = 118$, $p < 0.001$		$F_6 = 112$, $p < 0.001$		$F_6 = 109$, $p < 0.001$		$F_6 = 114$, $p < 0.001$	

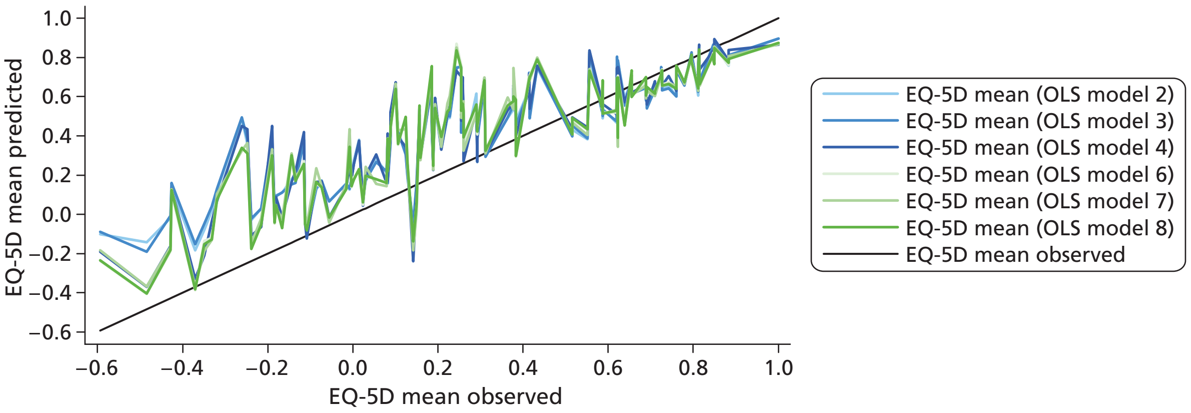

FIGURE 6 European Organization for Research and Treatment Quality-of-life Questionnaire Core 30 Plots of observed and predicted EQ-5D scores for OLS models.

TABLE 18 European Organization for Research and Treatment Quality-of-life Questionnaire Core 30 mean ranking of summary statistics and model performance tests for OLS models

	OLS model 2		OLS model 3		OLS model 4		OLS model 6		OLS model 7		OLS model 8	
Ranking components	All dimensions		Significant dimensions		Significant + squared terms		Significant items		Significant collapsed items		Significant items + age	
Mean (SD)	1 (6)		1 (5)		1 (4)		1 (2)		1 (3)		1 (1)	
Median	5		6		4		3		1		2	
Range	6		5		3		1		4		2	
R^2	5		6		4		2		3		1	
Adjusted R^2	5		5		4		2		3		1	
AIC	6		5		1		3		4		2	
BIC	6		2		1		4		3		5	
MAE	5		6		4		1		3		1	
Shrinkage	6		2		1		4		5		3	
Health status	***Mean***	***MAE***	***Mean***	***MAE***	***Mean***	***MAE***	***Mean***	***MAE***	***Mean***	***MAE***	***Mean***	***MAE***
1 (very poor)	4	4	6	5	3	3	2	2	5	5	1	1
2	6	5	5	6	2	1	1	3	4	4	3	2
3	1	5	2	6	3	4	5	1	6	3	4	2
4	5	5	6	6	1	4	2	1	4	3	3	2
5	6	5	4	6	1	1	3	2	5	2	2	2
6	4	6	2	5	3	3	6	2	4	4	1	1
7 (excellent)	2	1	1	2	5	6	3	5	6	3	4	3
Mean ranking	4.58		4.38		2.79		2.54		3.67		2.08	

model including all the items, age and gender). In terms of predictive ability, the response mapping models had the lowest MAEs on average. The best-fitting response-mapping model differs from other model techniques where the best-fitting models were item models. This was a result of collapsing item levels in order to estimate these models.

TABLE 19 European Organization for Research and Treatment Quality-of-life Questionnaire Core 30 mean observed and predicted EQ-5D values per model and summary model performance: best-fitting model across modelling techniques

Summary statistics and model performance tests	*n*	Observed values	OLS model 8		Tobit model 8		TPM model 8		SPL model 3		Response mapping model 8	
			Significant items + age		Significant items + age		Significant items + age (P1)		Significant dimensions		All dimensions + age/gender	
Mean (SD)	771	0.5793 (0.3423)	0.5793 (0.2866)		0.5792 (0.2891)		0.6066 (0.2997)		0.5793 (0.2833)		0.5726 (0.2914)	
Median		0.6910	0.6502		0.6517		0.6892		0.6457		0.6569	
Range		–0.594 to 1	–0.4046 to 0.9714		–0.3937 to 0.9463		–0.3936 to 0.9898		–0.3718 to 0.9438		–0.3376 to 0.9416	
MAE			0.139		0.139		0.140		0.143		0.134	
Shrinkage			1.042		1.020		0.940		0.997		1.179	
Health status			***Mean***	***MAE***	***Mean***	***MAE***	***Mean***	***MAE***	***Mean***	***MAE***	***Mean***	***MAE***
1 (very poor)	42	–0.0057	0.0642	0.205	0.0638	0.203	0.0649	0.195	0.0660	0.245	0.0473	0.181
2	53	0.1763	0.2470	0.179	0.2433	0.177	0.2670	0.185	0.3345	0.236	0.2262	0.159
3	144	0.4286	0.4629	0.182	0.4602	0.183	0.4808	0.184	0.5166	0.142	0.4515	0.182
4	226	0.6220	0.5823	0.138	0.5816	0.139	0.6091	0.141	0.5694	0.143	0.5827	0.139
5	186	0.7180	0.7176	0.109	0.7205	0.109	0.7566	0.107	0.7353	0.084	0.7094	0.097
6	94	0.8321	0.8181	0.098	0.8195	0.099	0.8511	0.104	0.8151	0.072	0.8137	0.100
7 (excellent)	26	0.9029	0.8546	0.080	0.8546	0.081	0.8925	0.060	0.8660	0.134	0.8596	0.075
ANOVA		F_6=97, p<0.001	F_6=114, p<0.001		F_6=113, p<0.001		F_6=114, p<0.001		F_6=117, p<0.001		F_6=116, p<0.001	

SPL, splining.

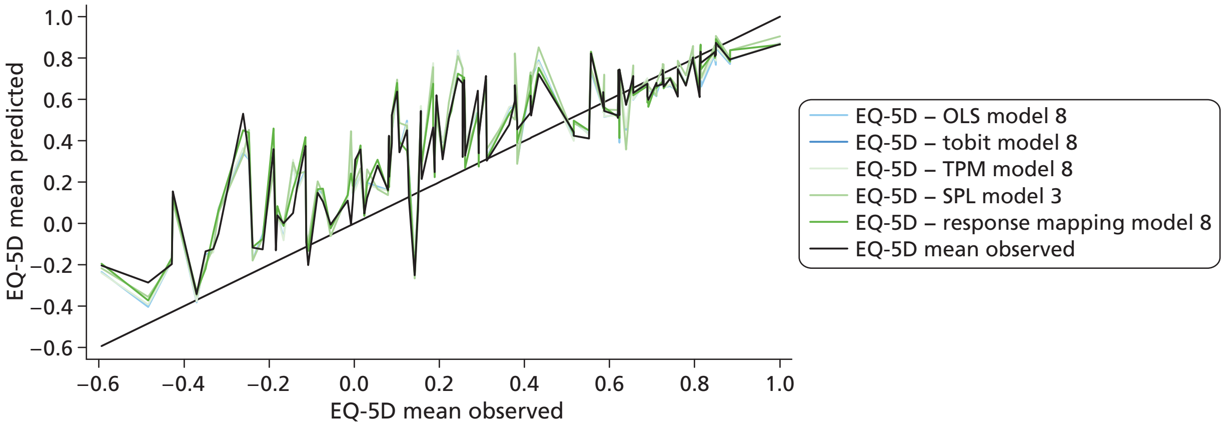

FIGURE 7 Observed and predicted EQ-5D scores for best performing models for EORTC QLQ-C30. SPL, splining.

Table 20 presents ranking of model performance statistics and the mean ranking across the common criteria for the best-fitting models from the different techniques used. Response mapping was the best performing model across all model performance statistics (mean ranking = 2.4) followed by OLS (mean = 2.7) and the tobit model (mean = 2.75).

Table 21 presents the model coefficients for the response-mapping model.

TABLE 20 European Organization for Research and Treatment Quality-of-life Questionnaire Core 30 mean ranking of summary statistics and model performance tests: best performing model across techniques

	OLS model 8		Tobit model 8		TPM model 8		SPL model 3		Response mapping model 8	
Ranking components	Significant items + age		Significant items + age		Significant items + age (P1)		Significant dimensions		All dimensions + age/gender	
Mean (SD)	1 (4)		3 (3)		5 (1)		1 (5)		4 (2)	
Median	3		4		1		5		2	
Range	2		3		1		4		5	
MAE	2		2		4		5		1	
Shrinkage	3		2		4		1		5	
Health status	***Mean***	***MAE***	***Mean***	***MAE***	***Mean***	***MAE***	***Mean***	***MAE***	***Mean***	***MAE***
1 (very poor)	3	4	2	3	4	2	5	5	1	1
2	3	3	2	2	4	4	5	5	1	1
3	3	2	2	4	4	5	5	1	1	2
4	3	1	4	2	1	4	5	5	2	2
5	1	4	2	4	5	3	4	1	3	2
6	2	3	1	2	5	5	3	1	4	4
7 (excellent)	4	3	4	4	1	1	2	5	3	2
Mean rank	2.7		2.75		3.2		3.65		2.4	

SPL, splining.

TABLE 21 European Organization for Research and Treatment Quality-of-life Questionnaire Core 30 Best-fitting response-mapping model

Variables	Mobility 2	Mobility 3	Self-care 2	Self-care 3	Usual acts 2	Usual acts 3	Pain 2	Pain 3	Anxiety/ depression 2	Anxiety/ depression 3
Physical functioning (SE)	–0.072*** (0.010)	–0.167*** (0.037)	–0.049*** (0.008)	–0.099 (0.119)	–0.036*** (0.010)	–0.085*** (0.014)	–0.001 (0.009)	–0.013 (0.013)	–0.014** (0.006)	–0.044*** (0.016)
Role functioning (SE)	–0.011* (0.006)	–0.007 (0.017)	–0.017*** (0.006)	–0.030 (0.023)	–0.032*** (0.007)	–0.055*** (0.010)	0.001 (0.007)	–0.001 (0.011)	0.005 (0.005)	0.019 (0.013)
Emotional functioning (SE)	0.010 (0.006)	0.024 (0.019)	0.008 (0.006)	0.008 (0.015)	0.021*** (0.007)	0.028*** (0.010)	0.009 (0.008)	0.011 (0.011)	–0.078*** (0.008)	–0.148*** (0.017)
Cognitive functioning (SE)	–0.011* (0.006)	–0.006 (0.015)	–0.010* (0.006)	–0.009 (0.029)	0.004 (0.007)	–0.001 (0.010)	0.003 (0.008)	0.015 (0.011)	–0.007 (0.006)	0.006 (0.014)
Social functioning (SE)	0.003 (0.006)	0.011 (0.016)	–0.009* (0.006)	–0.005 (0.017)	–0.021*** (0.007)	–0.034*** (0.009)	0.005 (0.008)	–0.001 (0.010)	0.006 (0.006)	0.008 (0.011)
Fatigue (SE)	0.006 (0.008)	0.002 (0.019)	–0.022*** (0.008)	–0.025 (0.027)	0.028*** (0.009)	0.033** (0.013)	0.007 (0.008)	0.006 (0.013)	–0.006 (0.007)	0.007 (0.019)
Nausea and vomiting (SE)	0.001 (0.007)	0.016 (0.018)	0.007 (0.008)	0.019 (0.029)	0.022** (0.010)	0.022* (0.013)	0.005 (0.017)	–0.004 (0.019)	–0.007 (0.008)	–0.009 (0.015)
Pain (SE)	0.023*** (0.005)	0.043*** (0.017)	0.016*** (0.005)	0.024* (0.015)	0.020*** (0.006)	0.023*** (0.008)	0.100*** (0.012)	0.164*** (0.016)	0.002 (0.004)	–0.012 (0.012)
Dyspnoea (SE)	0.002 (0.005)	0.004 (0.012)	–0.005 (0.005)	–0.015 (0.014)	–0.005 (0.006)	–0.015** (0.008)	0.010* (0.006)	0.008 (0.008)	0.000 (0.004)	–0.018 (0.011)
Sleep disturbance (SE)	0.002 (0.004)	0.010 (0.012)	0.002 (0.004)	–0.000 (0.010)	–0.001 (0.005)	–0.002 (0.007)	0.013*** (0.005)	0.021*** (0.008)	–0.003 (0.004)	0.012 (0.008)
Appetite loss (SE)	–0.009* (0.005)	0.004 (0.012)	–0.000 (0.004)	0.010 (0.013)	–0.010* (0.006)	–0.011 (0.008)	–0.013* (0.006)	–0.008 (0.009)	0.006 (0.004)	0.016* (0.009)

Variables	Mobility 2	Mobility 3	Self-care 2	Self-care 3	Usual acts 2	Usual acts 3	Pain 2	Pain 3	Anxiety/ depression 2	Anxiety/ depression 3
Constipation (SE)	–0.004 (0.005)	–0.012 (0.013)	–0.004 (0.005)	–0.009 (0.014)	–0.000 (0.005)	0.004 (0.007)	0.006 (0.006)	0.010 (0.008)	0.004 (0.004)	0.001 (0.009)
Diarrhoea (SE)	–0.005 (0.006)	0.010 (0.016)	0.003 (0.006)	0.005 (0.024)	–0.009 (0.006)	–0.011 (0.009)	–0.004 (0.008)	–0.008 (0.012)	0.002 (0.006)	0.002 (0.013)
Financial Impact (SE)	–0.001 (0.005)	–0.003 (0.010)	0.005 (0.004)	0.015 (0.010)	0.008 (0.006)	0.006 (0.008)	0.010* (0.005)	0.012 (0.008)	0.012*** (0.004)	0.015* (0.009)
Age (SE)	0.028** (0.013)	–0.021 (0.056)	0.048*** (0.015)	0.131* (0.069)					0.026** (0.011)	0.008 (0.028)
Female (SE)	–0.349 (0.251)	–1.397* (0.831)								
Constant (SE)	3.169** (1.598)	3.542 (5.250)	0.498 (1.467)	–6.619 (5.120)	3.494** (1.436)	5.675*** (1.835)	–3.255** (1.410)	–9.819*** (2.086)	4.562*** (1.316)	6.024* (3.123)
Observations	771	771	771	771	771	771	771	771	771	771
Pseudo R squared	0.449	0.449	0.392	0.392	0.461	0.461	0.455	0.455	0.364	0.364

* Statistically significant at the 10% level.
** Statistically significant at the 5% level.
*** Statistically significant at the 1% level.

Functional Assessment of Cancer Therapy – General Scale preliminary analysis

The mean EQ-5D index score for FACT-G data set was 0.721 (SD = 0.22) with a median of 0.735, scores ranged from –0.135 to 1 and 18% of responders are in full health and 0.9% scored less than 0. *Figure 8* presents the distribution of the EQ-5D index which displays the usual characteristics: there is a mass of observations at 1 (full health), there is a large gap between these observations and the next allowable value according to the EQ-5D tariff score with two additional peaks in the distribution. Patients with very poor HRQL were not included in the sample and, therefore, the data set does not span the full range of EQ-5D. *Table 22* shows that no responder had extreme problems for mobility and few responders had extreme problems for self-care, usual activities, pain/discomfort or anxiety/depression.

Average FACT-G scores were 20, 23, 18 and 18 for the physical, social, emotional and functional dimensions respectively (*Table 23*). The average overall score was 78 and ranged from 33 to 108, with no responders at the worse end of the FACT-G score (0–32). This is similar to the EQ-5D, where there are no respondents at the worst levels. The correlation between EQ-5D domains is presented in *Appendix 13*.

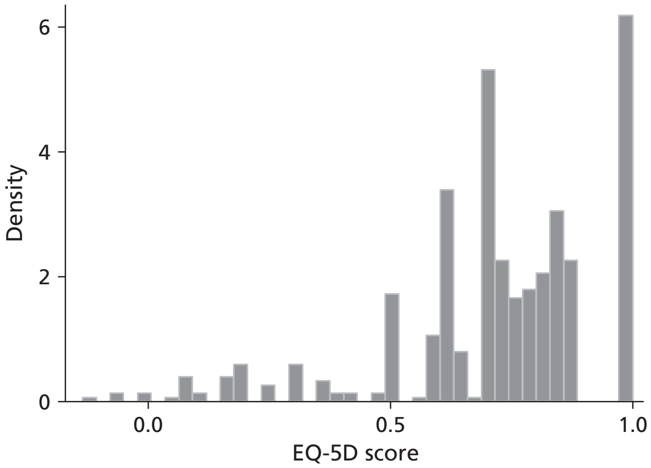

FIGURE 8 Distribution of EQ-5D scores for FACT-G data set.

TABLE 22 Responses to EQ-5D dimensions

EQ-5D item levels	Mobility	Self-care	Usual activities	Pain/ discomfort	Anxiety/ depression
No problems	316 (59.6%)	456 (86.0%)	206 (38.9%)	235 (44.3%)	260 (49.1%)
Some problems	214 (40.4%)	72 (13.6%)	292 (55.1%)	278 (52.5%)	260 (49.1%)
Unable/extreme problems	0 (0%)	2 (0.4%)	32 (6.0%)	17 (3.2%)	10 (1.9%)

TABLE 23 Summary of the FACT-G overall and domain scores

Summary statistics	Physical	Social	Emotional	Functional	Total score
n	530	530	530	530	530
Mean (SD)	20 (5.7)	23 (4.8)	18 (4.5)	18 (5.9)	78 (15.2)
Median	21	24	18	18	79
IQR	17–25	20–26	15–21	13–22	68–89
Range	1–28	1–28	4–24	0–28	33–108

IQR, interquartile range.

The only correlation of note is that between the physical domain and functional domain, which can be regarded as a moderate correlation ($p = 0.570$); all other correlations were below 0.4.

There was a modest relationship between FACT-G overall score and EQ-5D (Spearman's rank-order correlation = 0.575) (see *Appendix 13*). The EQ-5D also had a reasonable correlation with the physical and functional domains of the FACT-G, EQ-5D usual activities correlate modestly with FACT-G physical and functional scales and EQ-5D anxiety/depression correlates modestly with FACT-G emotion.

Best-fitting model Functional Assessment of Cancer Therapy – General Scale

The model selection process described above for EORTC QLQ-C30 was repeated for FACT-G and the best-fitting OLS, tobit, two-part, splining and response mapping models are summarised in *Table 24* and *Figure 9*. *Appendix 13* summarises individual OLS, tobit, two-part, splining and response mapping model results.

The best OLS and tobit models included significant items (model 6). For OLS, these were 'lack of energy', 'trouble meeting the needs of family' and 'pain' from the physical domain, 'feeling sad' and 'losing hope' from the emotional domain and 'able to work' from the functional domain. Level 0 (very much) of 'I feel sad' had fewer than 20 observations; therefore, these item levels were merged with level 1 and model 6 was then refitted. Collapsing item levels did not improve the overall model fit. Model 6 predicted the overall mean, underestimated those in near perfect or full health and overestimated those in poorer health states (see *Figure 9*). OLS gave the best mean estimates overall and by severity group, and had one of the two largest ranges of predicted scores (the TPM covered the widest range). OLS was the poorest at predicting the median and had the lowest shrinkage factor, suggesting it would be the most likely to overpredict results in studies applying the mapping algorithm.

The tobit model included two items from the physical domain (lack of energy and pain) and two items from the functional domain (able to work and enjoy life).

The best performing TPM included significant domain and squared terms (model 4) and domain, squared, interaction and gender and education (model 8). Females were less likely to report full health, whereas those with college degrees or professional degrees were more likely to report full health. Level of education was classified using an American system[201] and not all studies collect educational information in this way, meaning that this model may have limited generalisability to other studies. We therefore recommend model 4 as the best-fitting TPM as the estimates from models 4 and 8 were similar. Generally TPMs resulted in poorer mean predictions than tobit and OLS models but did have a slightly wider coverage of EQ-5D predicted scores.

Splining model 3 included significant domain scores and produced better estimates than model 1 (global FACT-G score model). Fractional polynomials identified cut-offs at a score of 25 for the physical domain and a score of 15 for the emotional well-being domain – no cut-off was necessary for the functional domain.

The best response-mapping models for predicting EQ-5D were the simplest models using significant domain scores (model 3); this model was unable to predict the full range of EQ-5D scores owing to the small proportion of responses at level 3, meaning that there were not enough data to obtain reliable estimates at the lower level. The response-mapping model gave reasonable estimates of the mean and median but the poorest MAE across severity groups.

A mean ranking of models across the different model performance statistics showed that OLS gave the best predictions (mean = 2.08), followed by the tobit model (mean = 2.42), with response mapping (mean = 3.5) and TPMs (mean = 3.58) giving the poorest predictions (*Table 25*). *Table 26* presents model coefficients for the best-fitting model. All models failed to predict anyone in full health, underpredicting at the top of the EQ-5D scale and overpredicting at the bottom end of the scale. However, the underprediction at the lower end of the scale is perhaps unsurprising given that few responders in the FACT-G data set reported severe problems with QoL.

TABLE 24 Summary of observed and predicted values for best performing models: FACT-G data set

			OLS model 6		Tobit model 6		TPM model 4		SPL model 3		Response mapping model 3	
Summary statistics and model performance tests		Observed values	Significant item levels		Significant item levels		Significant domain scores, squared and square root terms		Significant domain scores		Significant domain scores	
Mean (SD)		0.721 (0.223)	0.721 (0.163)		0.723 (0.161)		0.739 (0.154)		0.723 (0.144)		0.720 (0.133)	
Median		0.735	0.755		0.738		0.753		0.736		0.737	
Range		–0.135 to 1	0.115 to 0.962		0.132 to 0.957		0.119 to 0.993		0.312 to 0.974		0.268 to 0.934	
MAE			0.111		0.181		0.120		0.198		0.125	
Shrinkage			0.850		0.962		0.944		0.982		1.019	
Health states	***n***	**Mean**	**Mean**	**MAE**	**Mean**	**MAE**	**Mean**	**MAE**	**Mean**	**MAE**	**Mean**	**MAE**
1 (best)	122	0.8645	0.8464	0.088	0.8498	0.088	0.8302	0.090	0.8460	0.097	0.7933	0.101
2	256	0.7219	0.7318	0.108	0.7320	0.111	0.7359	0.121	0.7277	0.121	0.7201	0.122
3 (worst)	152	0.6055	0.6033	0.135	0.6074	0.137	0.6713	0.141	0.6152	0.148	0.6601	0.149
ANOVA		$F_{2527} = 55$, $p < 0.001$	$F_{2527} = 107$, $p < 0.001$		$F_{2527} = 109$, $p < 0.001$		$F_{2527} = 122$, $p < 0.001$		$F_{2527} = 130$, $p < 0.001$		$F_{2527} = 120$, $p < 0.001$	

SPL, splining.

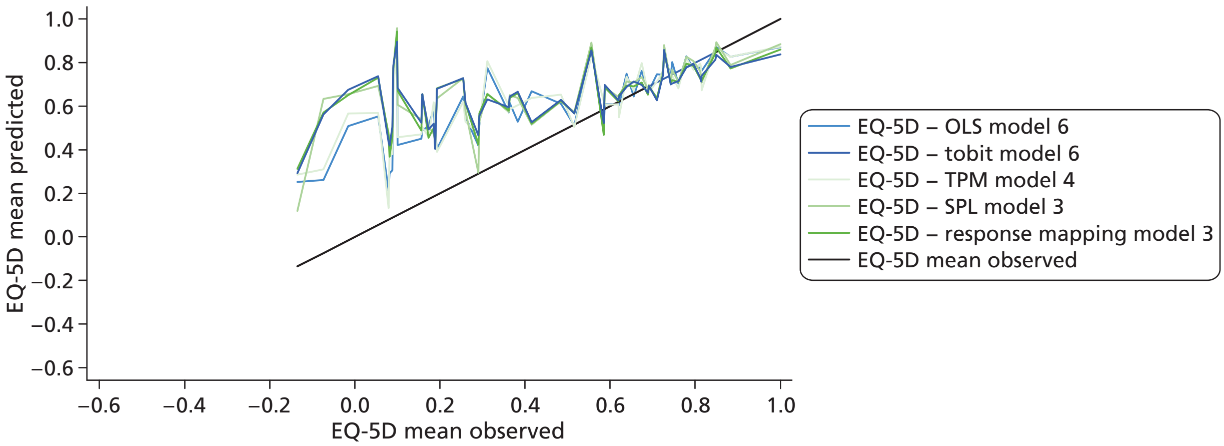

FIGURE 9 Summary of best FACT-G model predictions. SPL, splining.

TABLE 25 FACT-G mean ranking of summary statistics and model performance tests: best performing model across techniques

	OLS model 6		Tobit model 6		TPM model 4		SPL model 3		Response mapping model 3	
Ranking components	Significant item levels		Significant item levels		Significant domain scores, squared and square root terms		Significant domain scores		Significant domain scores	
Mean (SD)	1 (1)		3 (2)		5 (3)		3 (4)		2 (5)	
Median	5		3		4		1		2	
Range	2		3		1		5		4	
MAE	1		4		2		5		3	
Shrinkage	5		3		4		2		1	
Health states	Mean	MAE	Mean	MAE	Mean	MAE	Mean	MAE	Mean	MAE
1 (best)	2	1	1	1	4	3	3	4	5	5
2	3	1	4	2	5	4	2	3	1	5
3 (worst)	2	1	1	2	5	3	3	4	4	5
Mean rank	2.08		2.42		3.58		3.25		3.5	

SPL, splining.

Limited dependent variable mixture model: Functional Assessment of Cancer Therapy – General Scale

A model with all four FACT-G domains was selected as a possible candidate for estimation. Gender and age have been consistently shown to be important when estimating mapping functions. In addition, these two variables are typically used as explanatory variables for a host of parameter values used to populate decision analytic cost-effectiveness models and thus are also included. All four FACT-G domains are allowed to determine the mean EQ-5D of each latent component directly as well as indirectly through the probability of component membership. Gender and age determine the mean of EQ-5D in each class but are excluded from the probabilities. Models with up to five different components were fitted. Given the difference in variable and model selection procedures in this section to those included in *Best-fitting model Functional Assessment of Cancer Therapy-General*, a linear model with the same six covariates was also fitted for direct comparisons.

TABLE 26 Coefficients for best-fitting mapping model from FACT-G: item level OLS

Domain	Item	Item level	OLS model 6
Physical	Lack of energy	Very much (baseline level)	$F_{4505} = 3.62$, $p = 0.007$
		Quite a bit	0.045 (0.032)
		Somewhat	0.036 (0.030)
		A little bit	0.071 (0.033)*
		Not at all	0.118 (0.033)***
	Trouble meeting need of family	Very much (baseline level)	$F_{4505} = 2.75$, $p = 0.028$
		Quite a bit	0.028 (0.056)
		Somewhat	0.049 (0.050)
		A little bit	0.088 (0.050)*
		Not at all	0.098 (0.050)*
	Pain	Very much (baseline level)	$F_{4505} = 29.09$, $p < 0.001$
		Quite a bit	0.125 (0.073)*
		Somewhat	0.219 (0.069)**
		A little bit	0.240 (0.071)**
		Not at all	0.342 (0.070)***
Emotional	I feel sad	Very much (baseline level)	$F_{4505} = 2.45$, $p = 0.045$
		Quite a bit	–0.085 (0.105)
		Somewhat	–0.019 (0.101)
		A little bit	0.006 (0.099)
		Not at all	0.004 (0.099)
	Losing hope	Very much (baseline level)	$F_{4505} = 3.68$, $p = 0.006$
		Quite a bit	–0.081 (0.122)
		Somewhat	–0.007 (0.079)
		A little bit	0.013 (0.076)
		Not at all	0.060 (0.075)
Functional	Able to work	Not at all (baseline level)	$F_{4505} = 10.22$, $p < 0.001$
		A little bit	0.113 (0.031)***
		Somewhat	0.130 (0.028)***
		Quite a bit	0.150 (0.028)***
		Very much	0.152 (0.030)***
Constant			0.186 (0.0141)***

* Statistically significant at the level of 10%.
** Statistically significant at the level of 5%.
*** Statistically significant at the level of 1%.
The values in brackets refer to SEs.

Table 27 presents summary measures of overall fit and prediction for the linear model as well as the LDVMM with three to five components. The AIC decreases steadily from the linear model to the five-component LDVMM as more classes are added and it is lowest for the five-component LDVMM. However, the BIC is lowest for the linear model, reflecting the much higher penalty for model complexity

TABLE 27 Summary of overall model fit and prediction measures

Model performance tests	Linear model	LDVMM, three classes	LDVMM, four classes	LDVMM, five classes
Log-likelihood	175.36	202.54	223.79	272.35
AIC	–336.71	–337.07	–353.58	–424.70
BIC	–306.80	–191.79	–152.76	–168.33
MAE	0.126	0.123	0.121	0.119
RMSE	0.174	0.171	0.170	0.168

of this information criterion given the size of the data set. Out of the three LDVMMs, BIC selects the simplest model, a model with three components. Measures of in-sample predictions such as MAE and RMSE are lowest for the LDVMM with five components.

Table 28(a) displays comparisons of the observed and predicted EQ-5D means by the ECOG performance status, which measures the progression of the disease and its effect on the individual daily living activities. There are no individuals in the last ECOG group corresponding to 'completely disabled' in this data set and only 21 patients in the 'capable of only limited self-care' category, leaving only three groups of severity with enough patients to make any kind of comparison and even these have relatively small sample sizes across a broad range of severity. This prevents a more thorough analysis of systematic differences. *Table 28(b)* and *(c)* present the MAE and RMSE for each ECOG category. The differences in absolute value between the different LDVMMs are small as are their differences with the linear model. This is the typical

TABLE 28 Comparisons of observed vs. predicted means and in sample predictions split by the ECOG

Statistic	*n*	Observed EQ-5D	Linear model	LDVMM, thee classes	LDVMM, four classes	LDVMM, five classes
(a) Mean						
Health status (ECOG)						
1 (best)	122	0.8645	0.8342	0.8375	0.8410	0.8423
2	256	0.7219	0.7334	0.7328	0.7339	0.7308
3	131	0.6301	0.6182	0.6194	0.6194	0.6262
4[a] (worst)	21	0.4517	0.5610	0.5537	0.5555	0.5527
(b) MAE						
Health status (ECOG)						
1 (best)	122		0.096	0.094	0.092	0.091
2	256		0.123	0.122	0.118	0.118
3	131		0.145	0.140	0.143	0.135
4[a] (worst)	21		0.205	0.207	0.199	0.195
(c) RMSE						
Health status (ECOG)						
1 (best)	122		0.124	0.121	0.122	0.122
2	256		0.177	0.174	0.172	0.171
3	131		0.186	0.181	0.183	0.180
4[a] (worst)	21		0.275	0.276	0.270	0.265

[a] Owing to the small sample size in this group, numbers are only reported for completeness and should be taken with caution.

finding given the insensitivity of these measures when applied to individual level data sets and the small range covered by the EQ-5D scale. In terms of the mean, the LDVMM with five components is closer to the observed mean. Both the MAE and the RMSE are smallest for the LDVMM with five components with the exception of the RMSE of the first category of ECOG in which the smallest corresponds to the LDVMM model with three components.

Figure 10 depicts the percentage distribution of EQ-5D in the data set on the top left corner as well as the distributions of the simulated data from each model (100 replications per individual in the sample). The accompanying *Table 29* presents some descriptive statistics of the same simulated data sets. It is clear from *Figure 10* that the linear model is not capable of reproducing any of the characteristic features seen in EQ-5D data. It generates points well above one in considerable numbers: 10.5% of the simulated data set (see *Table 29*). The lack of observations at the bottom of the EQ-5D range allows a smaller estimated variance of the error term in the linear model than it would have been otherwise without penalising the likelihood excessively.

The key characteristics of EQ-5D are reflected in all three of the mixture models (see *Figure 10*). The mass of observations at one, the gap to the next feasible values and the multimodal distribution are all clearly generated by the use of this modelling method. There is a clear, separate peak in the observed data around 0.8, which is replicated in the five-class model and to a lesser extent in the three-class model.

Of these models, and based on these various aspects of model suitability, the five-class model is the optimal approach for estimating the index of EQ-5D from FACT-G domain scores, although that is based on fit for this particular data set which has features that may not be typical of the true relationship owing to the small sample size such as the separate peak at around 0.8. If this is the case, the four-class model, which offers similar performance, will be a better alternative. *Table 30* presents the parameter estimates as well as robust SE for these two LDVMMs.

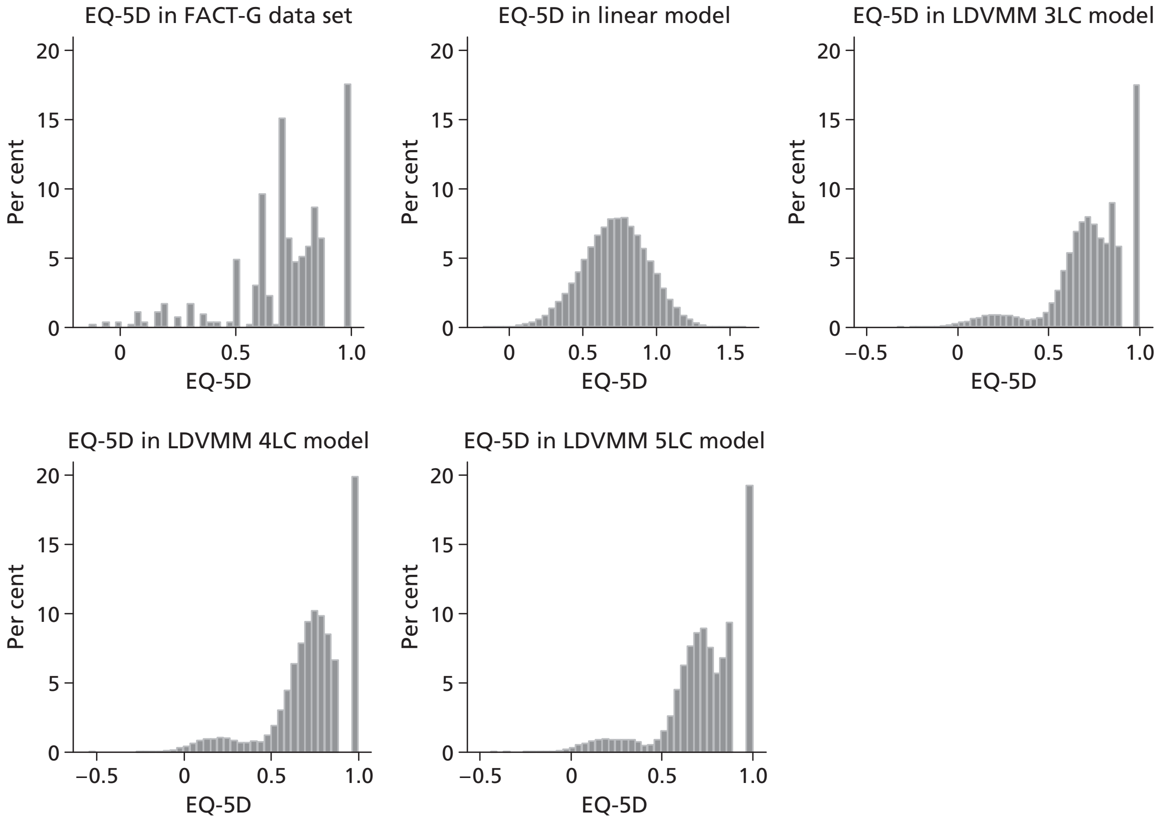

FIGURE 10 Observed EQ-5D distribution vs. simulated distributions from the models.

TABLE 29 Summary statistics of the observed EQ-5D distribution and simulated distributions from the models

Summary statistics	Observed EQ-5D	Linear model	LDVMM, three classes	LDVMM four, classes	LDVMM, five classes
Mean (SD)	0.7213 (0.2226)	0.7224 (0.2230)	0.7216 (0.2214)	0.7235 (0.2261)	0.7251 (0.2235)
Median	0.735	0.7280	0.7436	0.7494	0.7422
Range	–0.135–1	–0.176–1.605	–0.330–1	–0.544–1	–0.442–1
Percentage of values equal to 1 (%)	17.55	0	17.51	19.91	19.28
Percentage of values bigger than 1 (%)	0	10.48	0	0	0

TABLE 30 Parameter estimates and robust SEs of the LDVMM models

		LDVMM, four classes		LDVMM, five classes	
Individual components	Variables	Parameter	Robust SE	Parameter	Robust SEs
Component 1	Intercept	0.4404	0.0398	–0.1326	0.1690
	Physical/10	0.0823	0.0125	0.0894	0.0488
	Social/10	0.0079	0.0136	–0.0370	0.0410
	Emotional/10	0.0534	0.0135	0.0741	0.0380
	Functional/10	0.0531	0.0133	0.0834	0.0984
	Female	0.0065	0.0107	–0.0700	0.0582
	Age/10	–0.0106	0.0043	0.0262	0.0232
	Variance	–0.0084	0.0008	0.0096	0.0034
Component 2	Intercept	0.0475	0.0854	0.5255	0.0426
	Physical/10	0.0277	0.0449	0.0694	0.0144
	Social/10	0.0033	0.0172	0.0051	0.0128
	Emotional/10	–0.3663	0.0290	0.0242	0.0140
	Functional/10	0.3657	0.0420	0.0304	0.0142
	Female	–0.3006	0.0203	–0.0021	0.0109
	Age/10	0.1210	0.0146	–0.0095	0.0046
	Variance	0.0011	0.0009	0.0070	0.0008
Component 3	Intercept	0.2362	0.1543	0.9048	0.0500
	Physical/10	0.0649	0.0459	–0.0284	0.0078
	Social/10	0.0356	0.0415	–0.0022	0.0082
	Emotional/10	0.0403	0.0379	0.0255	0.0152
	Functional/10	–0.1298	0.0417	–0.0112	0.0121
	Female	–0.0814	0.0376	–0.0054	0.0065
	Age/10	–0.0131	0.0179	0.0025	0.0020
	Variance	–0.0112	0.0036	0.0001	0.0001

continued

TABLE 30 Parameter estimates and robust SEs of the LDVMM models (*continued*)

		LDVMM, four classes		LDVMM, five classes	
Individual components	Variables	Parameter	Robust SE	Parameter	Robust SEs
Component 4	Intercept	0.5016	0.0622	1.0421	0.0915
	Physical/10	1.4330	0.0049	0.1591	0.0144
	Social/10	0.2381	0.0063	–0.1523	0.0315
	Emotional/10	–1.0314	0.0338	–0.0642	0.0160
	Functional/10	0.3333	0.0129	–0.2793	0.0149
	Female	–0.3035	0.0048	–0.0645	0.0148
	Age/10	–0.2678	0.0045	–0.0267	0.0069
	Variance	0.0000	0.0000	–0.0008	0.0003
Component 5	Intercept			0.4723	0.0161
	Physical/10			1.5171	0.0237
	Social/10			0.2630	0.0056
	Emotional/10			–1.0861	0.0168
	Functional/10			0.3506	0.0060
	Female			–0.3354	0.0073
	Age/10			–0.2845	0.0051
	Variance			0.0000	0.0000
Probability of component membership					
Component 1	Intercept	26.3220	9.5991	28.7320	4.8843
	Physical/10	–6.4481	2.0328	–7.7485	1.3572
	Social/10	0.6283	0.5400	0.2088	0.7066
	Emotional/10	–5.9957	3.1645	–5.3613	1.2930
	Functional/10	1.0201	0.8806	–2.0401	0.6475
Component 2	Intercept	24.0409	9.7623	24.5811	4.5782
	Physical/10	–6.9441	2.1127	–6.9610	1.3098
	Social/10	–0.1360	0.9505	0.4083	0.4989
	Emotional/10	–4.8344	3.2401	–3.7666	1.0400
	Functional/10	1.3550	1.2704	–0.1056	0.4365
Component 3	Intercept	28.8196	9.6574	17.1879	4.3081
	Physical/10	–7.4868	2.0489	–5.6688	1.3442
	Social/10	0.8007	0.7136	0.6917	0.5661
	Emotional/10	–6.8488	3.2067	–3.3006	1.0858
	Functional/10	–0.3731	0.9861	0.8492	0.5185
Component 4	Intercept			20.2004	4.9627
	Physical/10			–7.8089	1.3512
	Social/10			1.4606	0.8702
	Emotional/10			–2.3890	1.4754
	Functional/10			–1.1400	0.6987

Even though the models in this section are not directly comparable to those in the preceding section owing to differing selection procedures, compared with the equivalent dimension models of the FACT G (*Appendix 13*), the overall MAE of LDVMM is better than the three models with (significant) domain scores. Dropping insignificant terms from the components and from the probabilities of component membership would increase both AIC and BIC for the LDVMM model and would tend to improve other measures of fit, making the LDVMM model appear to fit better. However, in doing this, there is a risk of fitting the model to this particular data set in excess and has not been pursued here.

Uncertainty

After allowing for uncertainty, the mean EQ-5D estimates ranged from 0.541 to 0.944 (mean = 0.721).

Discussion

Different regression techniques were explored to develop mapping functions for two widely used cancer measures, the EORTC QLQ-C30 and the FACT-G to the EQ-5D. In addition to methods such as OLS, which are widely used in the literature, a newer method that takes into account the characteristics typically seen in the distribution of the EQ-5D the LDVMMs were also applied for the FACT-G.

Response mapping gave the best predictions for the combined EORTC QLQ-C30 data sets. This model used all dimension scores, age and gender to estimate the EQ-5D index. Compared with other models fitted to this data set, this was best at predicting the overall MAE and mean and MAE per health status group. The mapping function is based on pooled data from three data sets, which was necessary in order to give a large enough sample to produce more reliable and representative mapping estimates. The data were from three different types of cancer and, therefore, could be argued to be more representative for use in other populations of mixed cancer types than other published mapping models. We also explored a range of models not previously examined in other studies.[146,147,197–199,216]

Only one previous study had mapped from FACT-G to EQ-5D and the mapping estimates were not reliable.[105] At this stage we do not know whether our findings are generalisable to other studies. Given the small amount of patients in the severest levels of HRQL in the FACT-G data set, the generalisability of those estimates are likely to be limited when compared with other populations containing patients at the severe end of the HRQL scale. Of the models fitted to the FACT-G data set, OLS and the tobit model using significant items gave the best estimates according to the mean predictions for the overall sample and the subgroups defined according to severity, and these models also performed well in the EORTC QLQ-C30 data sets. The model based on splining gave better median predictions and the response-mapping model performed best in terms of shrinkage. Only one LDVMM specification was fitted for the FACT-G data set, which included only dimension level information, gender and age. This model performed better than the equivalent linear model for the FACT-G and was shown to generate the main characteristics of the original distribution of EQ-5D in the data set. Even though the response mapping model results did not fit the data as well as other techniques, it is the only one, with the exception of the LDVMM, which can generate the features observed in the distribution of EQ-5D data. It does, however, ignore the ordinality of the data, and it is possible that more flexible models for response mapping, such as those presented in Hernández Alava *et al.*,[213] or further developments will increase the predictive ability of this modelling approach.

When considering the development of mapping functions, we could consider the size of sample needed to produce reliable functions. However, there are no rules for sample sizes in predictive modelling like prognostic modelling and mapping modelling but a rule of thumb is to have at least 20 individuals per independent variable.[217] For simple models like OLS, this would mean that a model including four dimensions would require a minimum of 80 individuals and a model including 27 items, each with five levels, would require 2160 observations (4 × 27 × 20). For response mapping models, the number of variables would relate to the smallest response category (usually level 3 for EQ-5D dimensions) and to work out sample size requirements you could work backwards from the expected number of respondents in level 3.

For example, if this was 3%, for a model including four dimensions you would need 2667 (80/0.03) observations or 27 items with five levels 72,000 (2160/0.03) observations.

To our knowledge, this is the first time that uncertainty has been accounted for in parameter (coefficient) estimates from mapping functions. At this stage we do not know what potential allowing for this uncertainty will have on NICE decisions. Future research needs to build on this and allow for uncertainty in the original EQ-5D estimates as well as the selection of appropriate models.

Generally, both OLS and tobit models using item level EORTC QLQ-C30 and FACT-G models gave some of the best model estimates and for FACT-G produced the best models, while for TPMs, domain level models gave better predictions. Other studies have fitted CLAD and generalised linear models as mapping functions. Like the tobit model, the CLAD model also deals with the limited nature of the data and produces consistent estimates in the presence of heteroscedasticity and non-normality. Median based models are not usually used for economic evaluation as, particularly when applied to costs, when aggregated, may not accurately reflect the total cost or benefit for the population.[218] Therefore, this model was not fitted here. Generalised linear models were not fitted either as they did not improve model fit over OLS models.

In terms of model selection, mapping studies in the literature report different model fit and model selection criteria, some focusing on model goodness of fit, others on the predictive ability of the model. Models should be selected mainly on their predictive ability, but other considerations may also be taken into account. Even still, there are still a number of criteria from which a model can be selected and different choices can result in alternative models being selection. In this chapter, rather than choosing one performance statistic to select the best model, we have given equal weighting to the overall mean, median, MAE, shrinkage and the mean and MAE per health status group. Further work should be undertaken to examine whether the criteria we have included are the optimal criteria to be used when judging mapping functions. For example, measures such as MAE and RMSE are not often used in other analyses of individual level data because heterogeneity across individuals is considerable, making these measures very insensitive to model improvements. This is an even greater problem when using dependent variables that span an extremely small range such as EQ-5D. The ranking method used here does not account for the magnitude of the predictions and how close they are to the observed data; further work should be undertaken to incorporate this into selecting the best models.

One of the other methodological factors that should be taken into consideration when carrying out mapping is the sample size used when producing the mapping functions. Response mapping produced poor predictions for FACT-G, although it was the best-fitting model for EORTC QLQ-C30. This was a result of the sample not covering the poorer health states but is also a function of sample size. With a larger sample, it would be possible to obtain more accurate predictions of the 3% of the sample being in level 3 for an EQ-5D dimension, for example. Further work is needed on sample size recommendations for the more complex models such as response mapping and LDVMMs. However, given the typically small size of cancer studies, it may be difficult to find studies with large enough samples to carry out the analysis. Combining data sets, as carried out for the EORTC QLQ-C30, offers an alternative when available and using mapping functions based on simpler techniques, such as OLS, may be the only option when these are not available.

Chapter 4 Developing 'bolt-on' items to EQ-5D

Background and aim

This chapter details the methods and results of the studies to develop 'bolt-on' items to the EQ-5D. Bolt-ons are dimension(s) that can be appended to another instrument in order to overcome perceived inadequacies of the parent instrument in a particular population. Utility values can then be obtained for the health states described by the instrument with the bolt-on. The bolt-ons reported here have been developed and tested with reference to the EQ-5D as the parent instrument. The EQ-5D was chosen for this purpose as it is the most widely used GPBM for economic evaluation internationally, and is recommended as the preferred GPBM by NICE in England and Wales. The systematic reviews of the published literature reported in *Chapter 2* found that the EQ-5D performed poorly in conditions affecting hearing and in some vision impairments. Therefore, these two clinical areas were selected for development of bolt-ons. In addition, energy was also selected as a potential bolt-on. Although the review of the measures in cancer did not find any particular problems related to cancer or cancer-related fatigue, it is an area where concern has been raised by NICE and its stakeholders (as summarised in Wailoo *et al.*[219]).

This chapter describes the bolt-on items and two valuation studies. The first study was an exploratory study to test the impact of the three bolt-on items on EQ-5D health states chosen to reflect mild, moderate and severe health states. Following from this study, the bolt-on having the largest impact was chosen for further evaluation and the second study was designed to allow a full valuation of that bolt-on with the EQ-5D.

Methods

Development of bolt-on items

The labels for the three bolt-on items were developed to be consistent with the labels of the three-level version of the EQ-5D so that they include categories of 'no problems', 'some problems' and 'extreme problems'. In addition, the measures identified in the systematic reviews reported in *Chapter 2* were considered in the development of the descriptions of the condition-specific labels. The review highlighted that some measures of vision and hearing give explicit reference to the use of supportive equipment, such as hearing aids and glasses. The use and provision of equipment such as these are commonplace in developed countries and, in most cases, easily address vision and hearing problems. A decision was taken to include reference to the use of supportive equipment so that the bolt-on instrument captures vision severity after taking into account the use of equipment. If the use of equipment was not explicitly addressed, the bolt-on item would fail to distinguish more severe problems that cannot be corrected using standard equipment. The references to equipment were developed to follow a similar format to that for the 'usual activities' dimension of EQ-5D, which includes a clarification in parentheses in the heading of the item. This referred to glasses or contact lenses in the vision bolt-on: 'Vision (using glasses or contact lenses if needed)', and to hearing aids as an example in the hearing bolt-on: 'Hearing (using equipment if needed, e.g. hearing aids)'. The wording of the three bolt-on dimensions is shown in *Box 1*.

Methods of the exploratory study

Health state selection

The aim of the exploratory study was to test the impact of the three bolt-on items on EQ-5D health state values. In brief, each possible level of severity of the bolt-ons was added to a selection of EQ-5D health states, each of which also represented a different level of severity, the health states were valued and

BOX 1 The three bolt-on items used in the exploratory study

Hearing (using equipment if needed, e.g. hearing aids)

I have no problems hearing □

I have some problems hearing □

I have extreme problems hearing □

Vision (using glasses or contact lenses if needed)

I have no problems seeing □

I have some problems seeing □

I have extreme problems seeing □

Tiredness

I am not tired □

I am moderately tired □

I am extremely tired □

compared with values obtained for corresponding EQ-5D states without the bolt-ons. It was hypothesised that:

- adding a mild level (no problems) of the bolt-on to a mild-state would have little impact compared with a mild EQ-5D state without a bolt-on
- adding a moderate level (some problems) of the bolt-on to a moderate-state would have little impact compared with a moderate EQ-5D state without a bolt-on
- adding a severe level (extreme problems) of the bolt-on to a severe-state would have little impact compared with a severe EQ-5D state without a bolt-on.

It was, however, also recognised that other effects could logically occur. For example, it could be that people assume no problems on the impairment or symptom reflected in the bolt-on if it is not presented. If this were the case, adding on a 'level 1' (no problems) of the bolt-on would be expected to have no impact on the EQ-5D health state regardless of the severity of that state. Therefore, we chose the study design for the exploratory study to reflect our weak priors regarding the impact of the bolt-ons and to explore it further.

Three EQ-5D health states were chosen as 'core' states for valuation. The health states were selected following consideration of three criteria: (1) to cover a range of severity levels, (2) to select from the set of 43 states that have previously been valued in a large UK general population study used to develop the UK EQ-5D tariff,[4,29] (3) to include combinations of problems that are not implausible or rare. This third criterion was assessed by examining health states that occur with relative high frequency in the Health Survey for England.[220] The final selection included three with a logically determined ordering of severity: a mild EQ-5D state, a moderate state and a severe state. The notation used to describe the health states in this report reflects the severity (level 1, 2, 3) on each of the five dimensions in the EQ-5D classification in the order presented in the questionnaire. The chosen mild state consists of no problems on the first three and last dimensions (mobility, self-care, usual activities and anxiety/depression) and moderate problems on the fourth dimension (pain/discomfort); therefore, is represented by the classifier 11121. The moderate state included some or moderate problems on all five dimensions (22222). The severe state included moderate problems on the first three dimensions (mobility, self-care and usual activities) and severe problems on the last two dimensions (pain/discomfort and anxiety/depression) and is represented by the classifier 22233.

All three levels of the bolt-on item (with severity levels of 1, 2 or 3) were added to each EQ-5D state resulting in nine states for valuation for each bolt-on dimension. The three core EQ-5D states without the bolt-on items were also valued. In order to ensure consistency in the number of states valued between groups and to allow a comparison of EQ-5D states with previous studies, six further EQ-5D states were selected for valuation from the previous large UK valuation study. The final selection of health states valued is shown in *Table 31*.

Data collection

Respondents to the survey were allocated to one of the four questionnaire variants: EQ-5D with each of the bolt-ons and EQ-5D alone. Five trained interviewers undertook the interviews. Interviewers were instructed to use each questionnaire variant in turn, so their first respondents completed questionnaire one, the next questionnaire two and so on and then back to questionnaire one. This ensured an even distribution of the variants between interviewers and minimised the risk of an interviewer effect biasing the results. The interviews followed a similar format to the UK EQ-5D valuation study.[4] After agreeing to participate in the study, respondents were asked to describe their own health using the EQ-5D and the bolt-on dimension they were about to value. Then the respondents rated their own health using the EQ-VAS, which is bounded by 0 ('worst imaginable health state') and 100 ('best imaginable health state'). Respondents then ranked six hypothetical states described on separate cards as a 'warm-up' task to familiarise respondents with the health state cards and with the process of stating their preferences towards the health states. The six states consisted of four states randomly chosen by interviewers or respondents from the nine states for each instrument, plus the best state described by the instrument and 'immediate death'.

Respondents then completed the main valuation exercise using the TTO method.[221] The best health state (11111 or 11111 + 1) described by the EQ-5D with/without bolt-on was used as the upper anchor. The respondent was asked to imagine 10 years of life in the health state under valuation, relative to a shorter duration in the best state, followed by 'immediate death'. A 'TTO board' was used as a visual aid to assist respondents with one side for valuing health states better than dead and the other side for those health states worse than dead. A conventional approach was taken to valuing states considered to be worse than dead. If respondents indicated that they would rather die immediately than live in the imperfect health state for any number of years, the TTO board was reversed as they were asked to state their preferred option between the imperfect health state for *t* years followed by (10 – *t*) years in full health or immediate death and the value of *t* was varied until the respondent was indifferent between the two options. Respondents valued a practice health state and then each of the nine health states as described in *Table 31*. Finally, respondents were asked to complete sociodemographic questions and their health status described using the remaining bolt-on items.

TABLE 31 Health states selected for valuation in the exploratory study

EQ-5D	EQ-5D + hearing	EQ + vision	EQ-5D + tiredness
11121	11121 + 1	11121 + 1	11121 + 1
22222	11121 + 2	11121 + 2	11121 + 2
22233	11121 + 3	11121 + 3	11121 + 3
11112	22222 + 1	22222 + 1	22222 + 1
11122	22222 + 2	22222 + 2	22222 + 2
21232	22222 + 3	22222 + 3	22222 + 3
22323	22233 + 1	22233 + 1	22233 + 1
33232	22233 + 2	22233 + 2	22233 + 2
33333	22233 + 3	22233 + 3	22233 + 3

Analysis

The sample size was estimated to detect a difference of 0.1 in mean values for health states with and without the bolt-on item using independent *t*-tests. Based on an assumed power of 0.8, significance level of 0.05 and SD of 0.3, based on a previously conducted study including a bolt-on item,[19] 73 respondents were required in each bolt-on group. Therefore, study recruitment aimed to survey 300 people randomly allocated to four groups of 75 people. Recruitment aimed to achieve a good spread across age, gender, ethnicity and social class. The sample was selected on the basis of postal address within South Yorkshire using the Names and Numbers software, AFD software (Ramsey, Isle of Man).

Time trade-off valuations were transformed using the transformation reported for the UK EQ-5D tariff to ensure all health state values are bound between –1 and 1:[4]

for states valued as better than dead TTO = $t/10$ and

for states valued as worse than dead TTO = $-t/10$.

The number of observations, mean transformed TTO values, SDs, maximum and minimum values are reported for all health states. Tests for differences in the sociodemographic characteristics between the four groups were compared using a chi-square test for categorical variables, a chi-square gamma statistic for ordered variables and ANOVA for continuous variables.

Paired *t*-tests were used to compare each health state with the bolt-on to the core EQ-5D state without the bolt-on. Regression analyses were used to 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 and allows for the fact that the error term may not be independent of the respondent.

The general model is:

$$y_{ij} = (\alpha + \beta \mathrm{x}_{ij} + \delta q_j + \theta r_j + \gamma z_i) + \varepsilon_{ij} \qquad (4)$$

where:

y_{ij} = TTO utility values for health state *j* valued by respondent *i*

i = 1, 2, . . ., m represents individual respondents

j = 1, 2, . . ., *n* represents health states valued

x = vector of dummy variables for the three EQ-5D core health state

q = vector of dummy variables for each variant (including EQ-5D and three bolt-ons)

r = vector of dummy variables for the three severity levels of the bolt-ons

z = vector of sociodemographic characteristics, including respondent's gender, age, experience of the bolt-on condition

ε_{ij} = an error term whose autocorrelation structure and distributional properties depend on the assumptions underlying the particular regression model used.

Stata version 10 (StataCorp LP, College Station, TX, USA) was used for all regression analysis, and SPSS v. 18 (SPSS Inc., Chicago, IL, USA) was used for the descriptive statistical analysis. A level of statistical significance was assumed where $p < 0.05$.

Methods of the full valuation study

The primary aim of the full valuation study was to develop a model for valuing all possible health states described by EQ-5D with one of the bolt-ons. Secondary aims included assessing the impact of the bolt-on to the coefficients representing the five EQ-5D dimensions.

In order to choose a bolt-on for this study, the results of the exploratory study were examined to identify the bolt-on with the most frequently statistically significant and consistent impact on health state values. Based on this assessment, the bolt-on for vision (EQ + vision) was selected for inclusion in the full valuation study. No change was made to the labelling or format of the vision bolt-on.

Health state selection

Health states were selected based on an orthogonal design of EQ + vision states. This required values for 18 health states assuming a main effects additive model. As the orthogonal design included mainly severe health states, two additional mild states were added to the orthogonal set. The set of EQ-5D only health states was selected from dropping the sixth dimension of the 20 EQ + vision states. Both sets of 20 health states were split in two in order to produce four groups of 10 states. The health states valued within the survey are shown in *Table 32*.

Data collection

A further sample of 300 members of the general public in South Yorkshire was recruited to participate in face-to-face interviews. The methods of sampling were the same as described for the exploratory study but people who had previously participated in the exploratory study were excluded. Survey respondents were allocated to one of four questionnaire variants and, as in the previous survey, each interviewer undertook valuations of each questionnaire variant in turn. The interviews followed a similar format to the exploratory study following amendments to take account of updated valuation methods for EQ-5D as recommended by the EuroQol Group. Specifically these included referring to 'dead' rather than 'immediate death' and 'full health' rather than the description of state 11111. In summary, after agreeing to participate in the study, respondents completed the EQ-5D for their own health (with the vision bolt-on if valuing EQ + vision states), then completed a 'warm up' task of ranking four health states plus the state 'dead' and valuing a

TABLE 32 Health states selected to value in the full valuation survey

EQ-5D states		EQ + vision states	
Group 1	Group 2	Group 3	Group 4
23133	32231	23133 + 3	32231 + 1
13122	21221	13122 + 1	21221 + 2
23212	22323	23212 + 2	22323 + 1
21332	13331	21332 + 1	13331 + 2
31133	31312	31133 + 2	31312 + 3
12232	12313	12232 + 3	12313 + 2
22111	33321	22111 + 3	33321 + 3
32122	33213	32122 + 2	33213 + 1
11121	11223	11121 + 1	11223 + 3
33333	11112	33333 + 3	11112 + 2

practice health state using the TTO method (22222/+ 2), and then valuing the 10 health states using the TTO method using the same approach as described for the exploratory study. The final task was for respondents to complete sociodemographic questions and describe their health status using the vision bolt-on (for those respondents valuing the EQ-5D only).

Analysis

The transformation of TTO valuations, statistical summaries of values, statistical software tests for differences in the sociodemographic characteristics are the same as those reported above for the exploratory study. Models were developed for both instruments separately using EQ-5D and EQ + vision. RE models were used in analyses to account for repeated observations. The dependent variable in each model was '1 – TTO value' and dummy variables were used to represent the levels on each dimension. The variables considered for inclusion in the analysis are shown in *Table 33*. The impact of the vision dimension was assessed by its statistical significance in the model after accounting for the EQ-5D dimensions. Alternative model specifications were explored including models published for other large EQ-5D data sets for the standard UK and USA EQ-5D value sets.[4,222]

The coefficients of the EQ-5D dimension dummy variables in the final model were compared using the *z*-test in order to make an assessment of the impact of the vision bolt-on to the values given to the EQ-5D dimensions.

TABLE 33 Variables considered for inclusion in the multivariate analysis of EQ-5D and EQ+vision

Variable	Description
Mobility	EQ-5D mobility dimension: level 1 (ref), level 2, level 3
Self-care	EQ-5D self-care dimension: level 1 (ref), level 2, level 3
Activities	EQ-5D usual activities dimension: level 1 (ref), level 2, level 3
Pain	EQ-5D pain/discomfort dimension: level 1 (ref), level 2, level 3
Anxiety	EQ-5D anxiety/depression dimension: level 1 (ref), level 2, level 3
Vision	EQ + vision dimension: level 1 (ref), level 2, level 3
Gender	Male (ref) or female
Age	Age categories: (1) 18–24 years (ref), (2) 25–34 years, (3) 35–44 years, (4) 45–54 years, (5) 55–64 years, (6) 65 + years
Marital	Marital status: (1) single (ref), (2) married, (3) separated, (4) divorced, (5) widowed
Yourself	Reporting experience serious of illness in yourself (0 reporting experience, 1 otherwise)
Family	Reporting experience serious of illness in your family (0 reporting experience, 1 otherwise)
Carer	Reporting experience serious of illness in caring for others (0 reporting experience, 1 otherwise)
Activity	Main activity: (1) employed or self-employed (ref), (2) retired, (3) homemaker, (4) student, (5) seeking work, (6) other
Education	Educated beyond school leaving age (0 yes, 1 no)
Home	Housing status: (1) own home (ref), (2) rent in public sector, (3) rent privately
SRVision	Self-reported level vision problems: level 1 (ref), level 2, level 3
N3	1 if any level 3 problems included in the health state, 0 otherwise
I2	Number of dimensions at level 2 beyond the first
I3	Number of dimensions at level 3 beyond the first
D1	Number of dimensions not at level 1 beyond the first

ref, reference level for dummy variables.

Results of the exploratory study

Three hundred face-to-face interviews were successfully completed, evenly split ($n = 75$) across four groups valuing each of the three bolt-ons and a group valuing EQ-5D alone. The characteristics of the respondents are shown in *Table 34*. Overall the characteristics of the groups were well balanced with very few statistically significant differences between the groups. Statistically significant differences were found

TABLE 34 Characteristics of respondents to the exploratory bolt-on valuation study

Characteristic	EQ-5D ($n = 75$)	EQ + hearing ($n = 75$)	EQ + vision ($n = 75$)	EQ + tiredness ($n = 75$)	χ^2 or *t*-statistic (*p*-value)
Age group (%)					
18–24	5	17	9	11	24.0 (0.065)
25–34	21	7	11	17	
35–44	20	16	24	8	
45–54	16	19	27	23	
55–64	20	19	12	23	
65 +	17	23	17	19	
Male (%)	32	40	49	39	4.78 (0.189)
Relationship status (%)					
Single	21	32	23	28	12.5 (0.408)
Married	53	40	60	48	
Separated	3	7	6	5	
Divorced	12	15	5	9	
Widowed	11	5	5	9	
Experience of serious illness (%)					
In yourself	29	33	23	37	4.34 (0.223)
In your family	68	68	71	79	2.71 (0.439)
In caring for others	55	36	40	52	12.2 (0.007)
Main activity (%)					
Employment	52	36	45	39	13.5 (0.563)
Retired	24	29	27	35	
Housework	6	12	9	6	
Student	3	5	5	6	
Seeking work	6	12	6	3	
Other	8	5	6	11	
Educated after minimum school leaving age (%)	64	60	56	55	1.48 (0.688)
Degree (%)	29	27	29	25	0.45 (0.930)
Home ownership (%)					
Own home	71	65	75	69	5.01 (0.543)
Rent (local authority)	17	16	19	17	
Rent (private sector)	12	19	7	12	

between the groups in terms of experience in caring for others, with more people in the EQ-5D and EQ + tiredness groups reporting experience of this than those in the EQ + vision and EQ + hearing groups.

Self-reported health status is shown in *Table 35*. Few people reported severe problems on any of the dimensions of health. The only differences in self-reported health between the groups were in the number of respondents reporting problems with vision; fewer people in the group allocated to valuing the EQ + vision reported current problems with vision. EQ-VAS scores and EQ-5D index values were similar between the groups.

TABLE 35 Self-reported health of respondents in the exploratory study

EQ-5D dimension and level		EQ-5D (%) (*n* = 75)	EQ + hearing (%) (*n* = 75)	EQ + vision (%) (*n* = 75)	EQ + tiredness (%) (*n* = 75)	χ^2 or *F*-statistic (*p*-value)
Mobility	1	62	59	58	48	11.4 (0.077)
	2	13	15	17	27	
	3	0	1	0	0	
Self-care	1	70	70	66	67	4.23 (0.646)
	2	5	4	8	8	
	3	0	1	1	0	
Usual activities	1	62	61	61	52	7.63 (0.266)
	2	11	11	11	21	
	3	1	2	3	2	
Pain/discomfort	1	46	48	53	41	4.36 (0.628)
	2	24	22	18	27	
	3	5	5	4	7	
Anxiety/depression	1	57	58	63	57	4.82 (0.567)
	2	15	13	11	12	
	3	3	4	1	6	
Hearing	1	64	63	61	63	2.65 (0.851)
	2	11	11	14	11	
	3	0	1	0	1	
Vision	1	44	43	61	45	13.3 (0.038)
	2	30	30	13	29	
	3	1	2	1	1	
Tiredness	1	39	40	42	40	6.17 (0.405)
	2	28	29	31	25	
	3	8	6	2	10	
Mean self-reported VAS (SD)		77.1 (21.2)	80.9 (17.2)	78.9 (17.8)	74.7 (21.5)	1.38 (0.250)
Mean self-reported EQ-5D index (SD)		0.83 (0.26)	0.80 (0.28)	0.84 (0.28)	0.75 (0.32)	13.36 (< 0.01)

Comparison of health state values

A total of 2697 TTO values were elicited from the 300 respondents. On average, each state was valued around 75 times. Summary statistics for the TTO values given to the EQ-5D health states (without bolt-on) are shown in *Table 36*. Mean values from the Measurement and Valuation of Health (MVH) study[29] used to generate the social tariff of EQ-5D values for the UK are also presented for comparison.[223,224]

In general, the values given in the NICEQoL study were higher than those obtained through the MVH study used to generate the social tariff of EQ-5D values for the UK. This is consistent with some international valuation studies of EQ-5D health states conducted since the MVH study, which have also reported higher mean TTO values compared with the original MVH study.[223,224]

The mean values for each of the bolt-on health states are presented in *Table 37* alongside the values for the 'core' EQ-5D states. The results of *t*-test comparing TTO values between the three core EQ-5D states and the corresponding nine states with specific bolt-ons are also reported in *Table 37*.

TABLE 36 Mean TTO values for all EQ-5D heath states (no bolt-on)

	Values from valuation study						Values from MVH study
State	*n*	Mean	SD	Median	Minimum	Maximum	Mean
11121	76	0.94	0.11	1.00	0.50	1	0.85
22222	74	0.71	0.30	0.80	–0.30	1	0.50
22233	74	0.41	0.40	0.43	–0.80	1	–0.14
11112	75	0.93	0.14	1.00	0.40	1	0.83
11122	75	0.87	0.19	1.00	0.20	1	0.72
21232	76	0.52	0.40	0.50	–0.80	1	0.06
22323	75	0.46	0.43	0.50	–0.93	1	0.04
33232	74	0.11	0.40	0.01	–0.93	1	–0.33
33333	75	–0.02	0.40	0.00	–0.93	1	–0.54

TABLE 37 Comparison between mean TTO values for EQ-5D and EQ-5D with bolt-ons

	EQ-5D		EQ-5D + hearing		EQ + vision		EQ-5D + tiredness	
EQ-5D state	Mean	Bolt-on state	Mean	*p*-value	Mean	*p*-value	Mean	*p*-value
11121	0.94	111211	0.94	0.89	0.94	0.82	0.94	0.71
		111212	0.90	0.07	0.90	0.01	0.90	0.06
		111213	0.85	0.001	0.69	< 0.001	0.82	< 0.001
22222	0.71	222221	0.80	0.04	0.74	0.54	0.79	0.09
		222222	0.77	0.18	0.76	0.25	0.74	0.54
		222223	0.70	0.82	0.59	0.02	0.72	0.85
22233	0.41	222331	0.40	0.92	0.41	0.99	0.45	0.51
		222332	0.45	0.56	0.41	0.99	0.45	0.52
		222333	0.36	0.43	0.32	0.16	0.34	0.33

The ordering of the mean values of the three core EQ-5D states was consistent with the logical ordering of these health states. Within each questionnaire variant, the TTO values were consistent with the domain levels with the exception of levels 1 and 2 added to the severe health state, where level 2 was higher than level 1 for the hearing bolt-on and there was no difference in values between bolt-on levels 1 and 2 in the corresponding state for vision and tiredness.

For the mild state (11121), there were no differences between the mean value for the 'core' EQ-5D state and for the states with the level 1 (no problems) bolt-on included. The states with a bolt-on level 2 added to the mild state (111212) resulted in lower values for all three bolt-ons. This difference was statistically significant for vision and approached significance for hearing ($p = 0.07$) and tiredness ($p = 0.06$). The inclusion of the level 3 bolt-on to form state 111213 resulted in significantly lower mean health state values across all bolt-ons. Among the three bolt-ons, adding on a level 3 (severe problems) for vision showed the greatest impact on the TTO value as the mean value decreased from 0.94 to 0.69, compared with 0.85 for hearing and 0.82 for tiredness.

The pattern of values for the bolt-on items to the moderate (22222) and severe (22233) states was more complex. For the moderate EQ-5D state (22222), including levels 1 or 2 of the bolt-ons increased the health state values, although only the level 1 hearing bolt-on showed a statistically significant difference. There was little impact of adding a level 3 for hearing and tiredness, but there were significantly lower values for the level 3 vision bolt-on. SDs were consistently higher for the more severe health states (with or without the bolt-ons).

For the severe state (22233), none of the bolt-on items had a statistically significant impact on the TTO value; however, the variance was also greater for these states. After adding level 1 and level 2 of the bolt-ons, the mean TTO values showed no difference for vision, small increases for tiredness and a slight increase for level 2 hearing, but none of the differences were statistically significant. Although not statistically significant, the addition of level 3 led to a reduction in mean TTO values for all bolt-ons (although it approached significance at the 0.1 level).

Table 38 shows the results of the multivariate analysis using a RE model. The primary aim of this analysis was to assess whether any of the differences in background characteristics between the groups had an impact on the values given to the health states. A secondary aim was to assess the impact of background characteristics on values more generally. The coefficients representing the severity of the core EQ-5D states were logically ordered and highly statistically significant. Similarly, the coefficients for the level of the bolt-on were consistently ordered; however, only the most severe level was statistically different from level 1. There were no significant differences in the coefficients for the type of bolt-on. Overall, the coefficients are difficult to interpret as the impact of the bolt-on depends on the severity of the state to which it is added.

The results of the multivariate analysis show that those background characteristics that differed between the groups (experience in caring for others and self-reported vision problems) had no significant impact upon the valuations given to the health states described by the instruments. Of the other background characteristics, marital status significantly impacted upon the values with single people giving lower values to health states than some of the other groups. In addition, those seeking work and people who had no further education after minimum school leaving age gave higher values to the health states.

Conclusions from the exploratory study

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 level of severity of the bolt-on and the severity of 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 one or two bolt-on to the moderate state led to higher values, but this was only statistically significant for the level 1 hearing bolt-on. Adding a level 3

TABLE 38 Analysis of the impact of background characteristics on the health state values in the exploratory study

Explanatory variables	Coefficient	SE	*p*-value
Core states			
11121	Ref		
22222	–0.151***	0.012	< 0.001
22233	–0.487***	0.012	< 0.001
Bolt-ons			
No bolt-on	Ref		
Hearing	0.055	0.036	0.132
Vision	0.005	0.12	0.902
Tiredness	0.037	1.05	0.292
Bolt-on levels			
Level 1	Ref		
Level 2	–0.015	0.013	0.235
Level 3	–0.114***	0.013	< 0.01
Female	–0.019	0.027	0.486
Age (years)			
18–24	Ref		
25–34	0.052	0.053	0.331
35–44	0.038	0.056	0.501
45–54	0.004	0.058	0.952
55–64	0.040	0.063	0.526
65 +	0.067	0.076	0.382
Marriage status			
Single	Ref		
Married	0.091**	0.037	0.014
Separated	0.052	0.061	0.394
Divorced	0.094*	0.051	0.067
Widowed	0.153**	0.060	0.011
No experience of serious illness			
In yourself	0.000	0.032	0.995
In your family	–0.028	0.030	0.349
In caring for others	–0.047	0.029	0.104
Main activities			
Employed	Ref		
Retired	–0.036	0.049	0.456
House work	0.013	0.045	0.769
Student	0.020	0.065	0.757
Seeking work	0.103**	0.052	0.048
Others	0.060	0.051	0.241

continued

TABLE 38 Analysis of the impact of background characteristics on the health state values in the exploratory study (*continued*)

Explanatory variables	Coefficient	SE	*p*-value
Education (to minimum school leaving age only)	0.057**	0.026	0.031
House ownership			
Rent from local authority	0.027	0.035	0.436
Rent from private sector	0.032	0.041	0.433
Self-reported health			
Hearing1	Ref		
Hearing2	–0.034	0.035	0.327
Hearing3	0.114	0.151	0.449
Vision1	Ref		
Vision2	0.012	0.029	0.689
Vision3	–0.137	0.094	0.148
Tiredness1	Ref		
Tiredness2	0.015	0.027	0.581
Tiredness3	–0.065	0.049	0.183
Constant	0.804***	0.068	< 0.001
Observations	2219		

Ref, reference value for dummy variables.
* *p* < 0.1.
** *p* < 0.05.
*** *p* < 0.01.

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. It should be noted that the severe states had the highest SDs associated with the mean values and so the comparisons had the lowest power.

Although there were a couple of statistically significant differences in the sociodemographic composition of the subgroups (specifically for experience in caring for others and vision problems), the regression analysis confirmed that these characteristics did not have a significant impact upon valuations. There did not appear to be substantial differences between the three bolt-ons, but overall, the impact appeared to be stronger for the vision bolt-on, therefore, this was selected for the full valuation study.

Results of the full valuation study of EQ + vision

In total, 302 people completed the interviews: 155 for EQ-5D alone and 157 for EQ + vision. The sociodemographic characteristics of respondents are presented in *Table 39*. There was a similar age and gender balance between the two groups. A summary of the self-reported health status of the respondents is shown in *Table 40*. There were no statistically significant differences in the sociodemographic characteristics or the self-reported health between the two groups.

TABLE 39 Sociodemographic characteristics of respondents of the EQ+vision valuation study

Characteristic	EQ-5D (*n* = 155)	EQ + vision (*n* = 157)	χ^2 (*p*-value)
Age group (years) (%)			
18–24	9.7	10.2	2.72 (0.742)
25–34	14.2	14.6	
35–44	18.7	22.9	
45–54	21.3	14.6	
55–64	16.8	16.6	
65+	19.4	21.0	
Male (%)	45.8	38.9	1.55 (0.214)
Relationship status (%)			
Single	16.8	25.5	7.05 (0.133)
Married	57.4	59.2	
Separated	3.2	1.9	
Divorced	11.6	7.0	
Widowed	11.0	6.4	
Experience of serious illness (%)			
In yourself	24.7	30.1	1.16 (0.282)
In your family	74.7	71.6	0.369 (0.544)
In caring for others	50.6	42.2	2.21 (0.137)
Main activity (%)			
Employment	52.9	45.9	6.70 (0.244)
Retired	25.2	25.5	
Housework	12.3	14.0	
Student	0	1.9	
Seeking work	3.9	8.3	
Other	5.8	4.5	
Educated after minimum school leaving age	56.1	57.3	0.05 (0.831)
Degree	32.9	36.5	0.45 (0.501)
Home ownership (%)			
Own home	72.9	74.5	1.28 (0.527)
Rent (local authority)	18.7	14.6	
Rent (private sector)	8.4	10.8	

The distribution of TTO values for each state is shown in *Figure 11* and a summary is provided in *Table 41*. Between 76 and 80 valuations were obtained for health state. Mean values ranged from 0.05 (state 33333) to 0.96 (state 11121) for the EQ-5D and from –0.04 (state 33333 + 3) to 0.95 (state 11112 + 2) for the EQ + vision. The rank ordering of several of the EQ-5D 'core' states differed to the rank ordering of the EQ + vision states. For example, for the EQ-5D, state 11121 was valued most highly followed by 11112; however, the rankings of these two states were reversed when the vision bolt-on was included (the mean value for 11112 + 2 was higher than that for 11121 + 1). SDs were generally higher for states considered to be most severe and the range of values given for most states was large.

TABLE 40 Self-reported health status of respondents in the EQ+vision study

EQ-5D dimension and level, VAS and index		EQ-5D (n = 155)	EQ + vision (n = 157)	χ^2 or t-test (p-value)
Mobility (%)	Level 1	83.9	75.2	4.25 (0.119)
	Level 2	16.1	24.2	
	Level 3	0.0	0.6	
Self-care (%)	Level 1	93.5	90.4	1.69 (0.431)
	Level 2	6.5	8.9	
	Level 3	0.0	0.6	
Usual activities (%)	Level 1	81.9	76.4	1.55 (0.462)
	Level 2	16.1	20.4	
	Level 3	1.9	3.2	
Pain/discomfort (%)	Level 1	66.5	61.1	1.63 (0.443)
	Level 2	30.3	33.1	
	Level 3	3.2	5.7	
Anxiety/depression (%)	Level 1	80.0	77.7	3.73 (0.154)
	Level 2	16.8	21.7	
	Level 3	3.2	0.6	
Vision (%)	Level 1	61.3	69.4	2.45 (0.294)
	Level 2	35.5	28.7	
	Level 3	3.2	1.9	
Mean self-reported VAS (SD)		79.2 (17.6)	75.8 (20.0)	–1.58 (0.114)
Mean EQ-5D index (SD)		0.85 (0.24)	0.82 (0.26)	–1.17 (0.244)

Multivariate analysis

The regression models for the EQ-5D and EQ + vision (both excluding sociodemographic variables) are presented in *Tables 42* and *43*, respectively.

The model specifications were estimated as below:

Model 1: including main effects only.

Model 2: including the N3 term to account for interactions as per the standard UK tariff (where N3 is a dummy variable for any dimension at level 3).

Model 3: including the D1 terms to account for interactions as considered in the US tariff (where the D1 terms are a set of interaction terms representing moves away from full health and the number of dimensions at level 3 beyond the first).

Model 4: the preferred model including all sociodemographic characteristics.

The models specified for the regression analysis are reported in *Tables 42* and *43* for the EQ-5D and EQ + vision data, respectively. The terms representing interactions did not have statistically significant coefficients in either of the models and so have been excluded from the final models presented here.

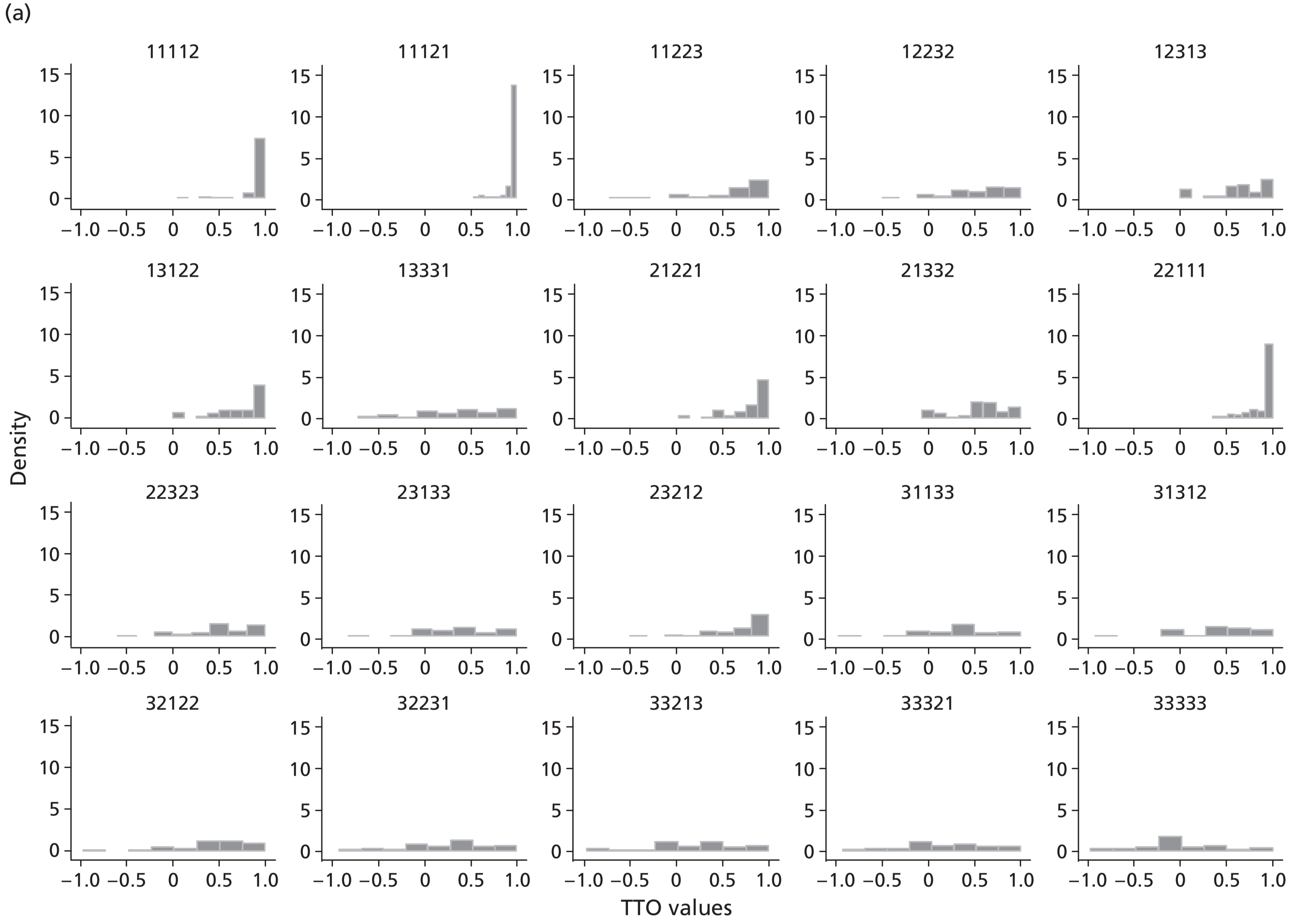

FIGURE 11 Distribution of TTO values for EQ-5D and EQ+vision health states. (a) EQ-5D health states; (b) EQ+vision health states. (*continued*)

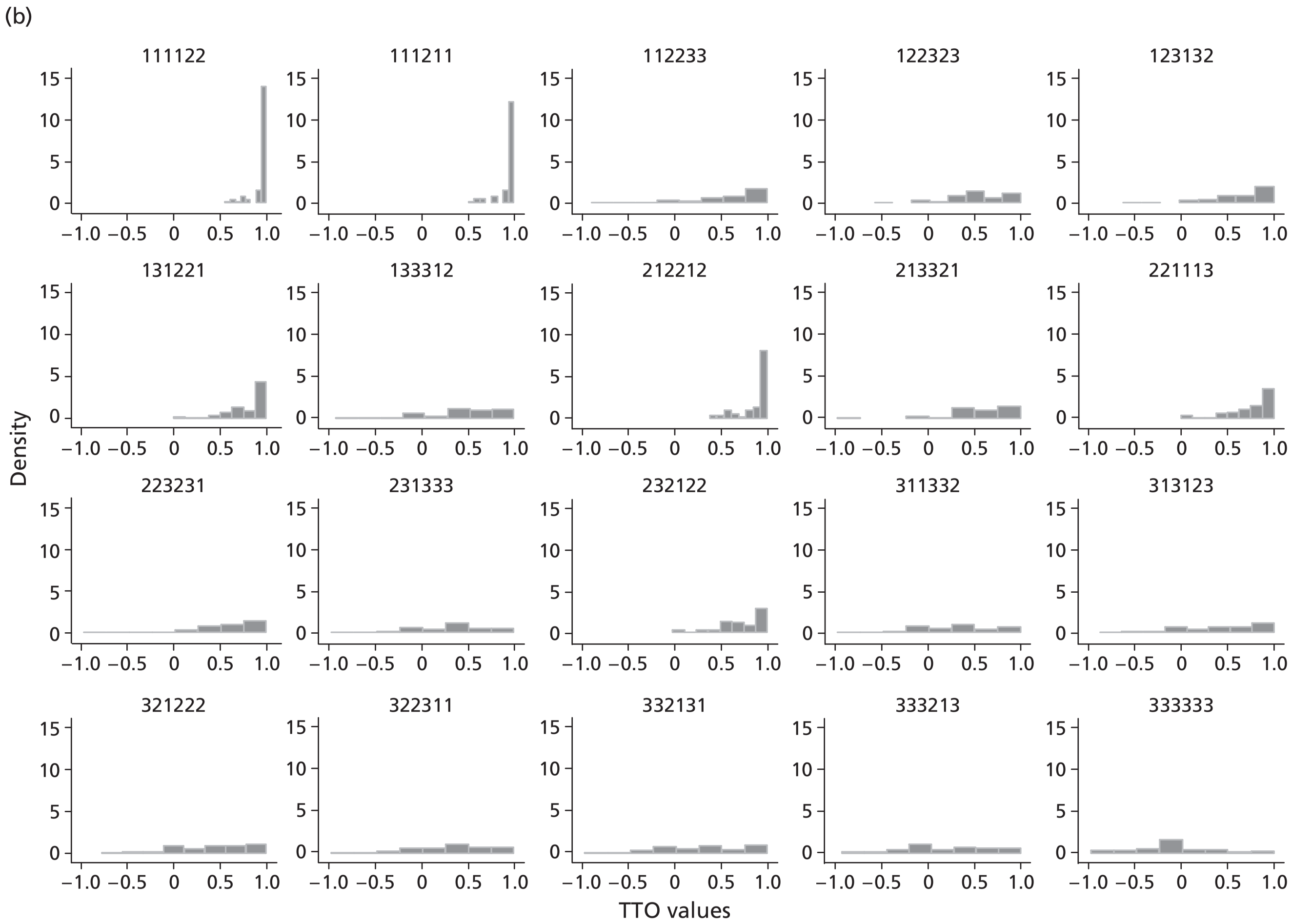

FIGURE 11 Distribution of TTO values for EQ-5D and EQ + vision health states. (a) EQ-5D health states; (b) EQ + vision health states. (*continued*)

TABLE 41 Mean TTO values for EQ-5D and EQ+vision

Health state	Mean	*n*	SD	Minimum	Maximum	Median
EQ-5D states						
11112	0.93	77	0.17	0.05	1	1.00
11121	0.96	78	0.10	0.53	1	1.00
11223	0.67	77	0.37	–0.73	1	0.78
12232	0.57	79	0.32	–0.5	1	0.63
12313	0.62	76	0.32	0	1	0.68
13122	0.75	78	0.30	0	1	0.85
13331	0.39	77	0.45	–0.73	1	0.40
21221	0.82	76	0.23	0.03	1	0.90
21332	0.55	77	0.30	–0.08	1	0.60
22111	0.90	77	0.15	0.35	1	1.00
22323	0.55	78	0.36	–0.6	1	0.51
23133	0.42	78	0.38	–0.83	1	0.43
23212	0.72	79	0.32	–0.5	1	0.80
31133	0.35	78	0.40	–0.98	1	0.38
31312	0.46	77	0.40	–0.93	1	0.50
32122	0.50	77	0.38	–0.98	1	0.53
32231	0.28	76	0.45	–0.93	1	0.38
33213	0.26	78	0.47	–0.98	1	0.35
33321	0.21	78	0.47	–0.93	1	0.20
33333	0.05	79	0.42	–0.98	1	0.00
EQ-5D + vision states						
111122	0.95	79	0.10	0.55	1	1.00
111211	0.94	79	0.12	0.5	1	1.00
112233	0.59	77	0.40	–0.9	1	0.70
122323	0.53	77	0.34	–0.57	1	0.53
123132	0.63	76	0.36	–0.63	1	0.70
131221	0.80	79	0.23	0	1	0.90
133312	0.48	79	0.41	–0.93	1	0.50
212212	0.89	77	0.17	0.38	1	1.00
213321	0.60	79	0.35	–0.98	1	0.63
221113	0.77	79	0.25	0	1	0.83
223231	0.58	78	0.39	–0.98	1	0.70
231333	0.30	79	0.45	–0.98	1	0.33
232122	0.71	80	0.27	–0.03	1	0.75
311332	0.31	80	0.44	–0.98	1	0.30
313123	0.42	77	0.45	–0.88	1	0.50

continued

TABLE 41 Mean TTO values for EQ-5D and EQ+vision (*continued*)

Health state	Mean	*n*	SD	Minimum	Maximum	Median
321222	0.43	79	0.41	–0.78	1	0.50
322311	0.35	79	0.43	–0.98	1	0.40
332131	0.34	80	0.46	–0.98	1	0.35
333213	0.24	78	0.48	–0.93	1	0.21
333333	–0.04	79	0.45	–0.98	1	0

TABLE 42 Models estimated for EQ-5D

	Model 1: main effects only			Model 2: including N3 term			Model 3: including D1 term		
Variable	Coefficient	SE	*p*-value	Coefficient	SE	*p*-value	Coefficient	SE	*p*-value
Mobility 2	0.019	0.018	0.309	0.021	0.018	0.253	0.017	0.049	0.726
Mobility 3	0.315	0.017	< 0.001	0.308	0.018	< 0.001	0.293	0.083	< 0.001
Self-care 2	0.079	0.018	< 0.001	0.067	0.020	0.001	0.083	0.033	0.012
Self-care 3	0.185	0.018	< 0.001	0.170	0.022	< 0.001	0.166	0.069	0.016
Activities 2	0.076	0.020	< 0.001	0.066	0.022	0.002	0.091	0.035	0.011
Activities 3	0.150	0.021	< 0.001	0.136	0.024	< 0.001	0.136	0.071	0.056
Pain 2	0.071	0.018	< 0.001	0.060	0.020	0.003	0.082	0.030	0.006
Pain 3	0.236	0.020	< 0.001	0.221	0.024	< 0.001	0.220	0.078	0.005
Anxiety 2	0.036	0.020	0.070	0.014	0.027	0.610	0.039	0.031	0.206
Anxiety 3	0.120	0.018	< 0.001	0.100	0.025	< 0.001	0.101	0.062	0.103
Vision 2									
Vision 3									
N3				0.043	0.038	0.259			
D1							0.020	0.041	0.635
I^2							–0.029	0.069	0.675
$I2^2$							–0.001	0.022	0.970
I^3							0.023	0.102	0.821
$I3^2$							–0.007	0.011	0.509
Constant	0.009	0.031	0.768	0.017	0.032	0.598			
Number of observations	1550			1550			1550		
Number of groups	155			155			155		
Log-likelihood	–340			–339			–338		
p-value from the chi-squared test	< 0.001			< 0.001			< 0.001		

TABLE 43 Models estimated for EQ+vision

	Model 1: main effects only			Model 2: including N3 term			Model 3: including D1 term		
Variable	Coefficient	SE	*p*-value	Coefficient	SE	*p*-value	Coefficient	SE	*p*-value
Mobility 2	0.034	0.018	0.062	0.032	0.018	0.079	0.025	0.039	0.533
Mobility 3	0.320	0.017	< 0.001	0.314	0.018	< 0.001	–0.069	0.354	0.846
Self-care 2	0.091	0.018	< 0.001	0.077	0.022	< 0.001	0.104	0.051	0.042
Self-care 3	0.158	0.018	< 0.001	0.147	0.021	< 0.001	–0.255	0.379	0.501
Activities 2	0.032	0.020	0.118	0.029	0.021	0.165	0.090	0.071	0.204
Activities 3	0.104	0.021	< 0.001	0.097	0.022	< 0.001	–0.209	0.306	0.495
Pain 2	0.062	0.019	0.001	0.062	0.019	0.001	0.071	0.038	0.062
Pain 3	0.219	0.020	< 0.001	0.216	0.020	< 0.001	–0.100	0.324	0.756
Anxiety 2	0.038	0.020	0.056	0.029	0.021	0.170	0.070	0.053	0.193
Anxiety 3	0.159	0.018	< 0.001	0.150	0.020	< 0.001	–0.161	0.319	0.612
Vision 2	0.037	0.018	0.040	0.039	0.018	0.031	0.030	0.034	0.389
Vision 3	0.130	0.018	< 0.001	0.127	0.018	< 0.001	–0.246	0.361	0.495
N3				0.035	0.033	0.293			
D1							0.444	0.381	0.244
I^2							–0.555	0.429	0.196
$I2^2$							0.026	0.017	0.128
I^3							–0.236	0.168	0.160
$I3^2$							0.042	0.036	0.237
Constant	–0.018	0.035	0.608	–0.026	0.036	0.477			
Number of observations	1570			1570			1570		
Number of groups	157			157			157		
Log-likelihood	–361			–361			–361		
p-value from the chi-squared test	< 0.001			< 0.001			< 0.001		

In the model for EQ-5D, all the coefficients followed a logical order, the decrement in utility attributed to level 3 problems was greater than that for level 2 problems. The coefficients for all dimensions were statistically significant, except for the dummy variables representing some mobility problems and moderate anxiety/depression. The largest impact on EQ-5D values was level 3 mobility problems (being confined to bed), followed by level 3 problems with pain/discomfort and self-care.

There were some similarities to the existing main UK data set for EQ-5D (level 3)[4] that was based on a large UK general population study. That study also found that level 3 mobility and pain/discomfort had the largest impact on EQ-5D values; however, level 3 self-care problems had the fifth largest impact. Overall, the size of the utility decrements were smaller in this this study compared with the previous UK population study and this reflects the higher TTO values reported by respondents in this study.

The model for EQ + vision demonstrated a similar pattern to that for EQ-5D. The coefficients followed a logical ordering, including the coefficients for the vision bolt-on. The vision coefficients were statistically significant, which indicates that vision has a significant impact on EQ-5D values after taking into account the 5 standard EQ-5D dimensions. As with the model for EQ-5D, the coefficients representing some mobility problems and moderate anxiety/depression were not statistically significant, which also applied to the coefficient representing some problems carrying out usual activities. The coefficients with the largest impacts were still level 3 mobility problems (being confined to bed) and level 3 problems with pain/discomfort; the coefficient for level 3 vision problems was the fifth largest, ahead of level 3 problems performing usual activities.

Table 44 shows the results of the analysis including background characteristics. The values of people who reported that they had some or extreme vision problems did not value the health states significantly differently from people who reported no vision problems. There were some statistically significant differences in the health state values according to age with the youngest age group giving lower values than the other age groups. Some differences were also seen in the valuation of the EQ + vision health states according to experience of caring for others and those seeking work compared with employed respondents.

TABLE 44 Final models (with background characteristics)

	EQ-5D			EQ + vision		
Variable	Coefficient	SE	*p*-value	Coefficient	SE	*p*-value
Mobility 2	0.020	0.019	0.270	0.036	0.019	0.051
Mobility 3	0.318	0.017	< 0.001	0.320	0.017	< 0.001
Self-care 2	0.079	0.018	< 0.001	0.091	0.018	< 0.001
Self-care 3	0.185	0.018	< 0.001	0.163	0.018	< 0.001
Activities 2	0.076	0.020	< 0.001	0.033	0.021	0.105
Activities 3	0.149	0.021	< 0.001	0.108	0.021	< 0.001
Pain 2	0.072	0.018	< 0.001	0.060	0.019	0.002
Pain 3	0.238	0.020	< 0.001	0.216	0.020	< 0.001
Anxiety 2	0.039	0.020	0.051	0.037	0.020	0.062
Anxiety 3	0.122	0.018	< 0.001	0.158	0.018	< 0.001
Vision 2				0.033	0.018	0.068
Vision 3				0.127	0.018	< 0.001
Gender	–0.040	0.043	0.352	–0.071	0.045	0.117
Age 1						
Age 2	–0.104	0.090	0.248	–0.212	0.087	0.014
Age 3	–0.202	0.098	0.040	–0.316	0.088	< 0.001
Age 4	–0.188	0.103	0.069	–0.226	0.102	0.027
Age 5	–0.160	0.111	0.149	–0.208	0.105	0.048
Age 6	–0.147	0.132	0.264	–0.122	0.129	0.344
M_single						
M_married	0.010	0.069	0.890	0.057	0.062	0.356
M_sep	–0.001	0.124	0.991	0.449	0.148	0.002
M_div	0.078	0.091	0.390	0.058	0.097	0.549
M_widow	0.089	0.097	0.359	0.044	0.103	0.666
Yourself	0.022	0.052	0.681	–0.031	0.050	0.54

TABLE 44 Final models (with background characteristics) (*continued*)

	EQ-5D			EQ + vision		
Variable	Coefficient	SE	*p*-value	Coefficient	SE	*p*-value
Family	0.031	0.048	0.522	0.057	0.045	0.207
Carer	–0.046	0.043	0.290	0.101	0.044	0.022
Activity_Emp						
Activity_Retired	0.055	0.085	0.518	–0.123	0.087	0.159
Activity_home	–0.076	0.066	0.247	0.013	0.063	0.839
Activity_student				–0.027	0.149	0.856
Activity_seeking	–0.046	0.103	0.653	–0.191	0.084	0.024
Activity_Other	0.058	0.087	0.505	0.032	0.105	0.759
Education	–0.017	0.043	0.697	–0.045	0.046	0.331
Home_own						
Home_rentLA	–0.021	0.056	0.705	–0.031	0.064	0.626
Home_rentp	–0.077	0.080	0.339	–0.142	0.072	0.047
Self-reported Vision 1						
Self-reported Vision 2	0.008	0.047	0.860	–0.031	0.044	0.478
Self-reported Vision 3	–0.036	0.113	0.748	0.037	0.195	0.848
Constant	0.159	0.097	0.101	0.217	0.101	0.032

The coefficients for the five EQ-5D dimensions were compared between the EQ-5D model and the EQ + vision model. A difference in the coefficients would suggest that including the additional vision dimension leads to different valuations of the five EQ-5D dimensions; for example, if having some problems with self-care is valued differently depending on whether vision problems are present in the health state. The results of the *z*-test are presented in *Table 45*. There were no statistical differences in the coefficients at the predefined level for statistical significance; however, some of the coefficients appeared to be qualitatively different and approached the level for significance. In particular, the coefficients for the usual activities dimension differed by 0.045 and 0.046 for levels 2 and 3, respectively, each with *p*-values of less than 0.1. The difference in the coefficients for level 3 anxiety and depression was also of a similar magnitude (0.039) that also had a *p*-value of less than 0.1.

Discussion

The results from the exploratory study and the main valuation study demonstrate that bolt-on items can potentially have a significant impact upon EQ-5D valuations. In these studies, bolt-ons representing vision impairment, hearing impairment and tiredness all significantly impacted on at least some health states.

The findings from both of the empirical studies presented here demonstrate that the relationship of the bolt-ons to the EQ-5D state valuations is complex. The exploratory study shows that the impact of the bolt-ons depends on the severity of the bolt-on item and the severity of the state to which they are added. The inclusion of bolt-ons representing ‘no problems’ is not always of no consequence. When included alongside severe health states, it can lead to higher valuations than not mentioning the absence of problems. This has significant implications for the valuation of bolt-ons as it suggests that including the bolt-on valuation as a simple decrement in, for example, an additive model, is inadequate. This confirms findings in another bolt-on study looking at the addition of pain to a condition specific instrument.[225]

TABLE 45 Comparison of the model coefficients

Dimension and level	EQ-5D model	EQ + vision model	*p*-value (*z*-test)
Mobility 2	0.019	0.0344	0.271
Mobility 3	0.315	0.320	0.428
Self-care 2	0.079	0.091	0.313
Self-care 3	0.185	0.158	0.140
Activities 2	0.076	0.032	0.059[a]
Activities 3	0.150	0.104	0.062[a]
Pain 2	0.071	0.062	0.367
Pain 3	0.236	0.219	0.285
Anxiety 2	0.036	0.038	0.468
Anxiety 3	0.120	0.159	0.062[a]
Vision 2		0.0378	
Vision 3		0.130	

a $p < 0.1$.

The hearing and vision bolt-on items referred explicitly to the use of equipment and were designed to detect more serious problems that cannot be corrected by the use of standard equipment such as glasses. As a result, it is possible that the bolt-on items may not be responsive for some interventions that remove the need for the use of that equipment; for example, laser eye surgery to remove the need for wearing glasses. While accepting this limitation, this was considered preferable to the alternative of excluding the use of equipment, as this could drive differences between levels of severity and would not pick up the most severe levels of vision and hearing problems which are not readily correctable using standard equipment.

Our results differ to an earlier study that investigated the impact of including 'tiredness' as a dimension within the EQ-5D (i.e. a potential EQ-6D) using VAS.[226] We found that the inclusion of a level of 'no tiredness' on the bolt-on led to higher values compared with no bolt-on, as well as lower values reflecting 'extreme tiredness'. One could hypothesise that the differences between the two studies could be the result of the combinations of levels each has chosen to investigate. However, this appears not to be the case as both studies included a common health state (11121). The study by Gudex[226] found that the inclusion of level 2 tiredness problems did not significantly affect the valuations, whereas our study found that it was associated with near significantly lower values. There are notable differences between the two studies that could perhaps explain the discrepancy, including the valuation methods and the number of levels/labelling of the tiredness dimensions. Gudex[226] used visual analogue ratings whereas this study used the TTO method. In addition, the tiredness bolt-on consisted of two possible levels in the study by Gudex,[226] whereas the bolt-on in this study included three levels. On the other hand, similar results were reported in a previous studying adding on a sleep dimension to EQ-5D.[19] A significant difference was found after adding on level 1 to a moderate EQ-5D state (11233) but no statistically significant differences were found where various severity levels of the sleep dimension were added to five other relatively moderate or severe EQ-5D states.

The complexities in the valuations were also found in the main valuation of the vision bolt-on, in which a full valuation model was reported. One of the aims of the study was to establish whether the inclusion of the bolt-on with the EQ-5D health state description had a significant impact on the valuation of the five EQ-5D dimensions. This is an important question as it affects whether future bolt-ons need to be valued alongside the EQ-5D descriptions each time, which leads to substantial resources being required for the valuation of each bolt-on. Unfortunately, the results from the analysis were not conclusive. Although not significant at the predefined level for statistical significance, the project team were unable to conclude that the impact was not qualitatively different. In particular, the vision bolt-on appeared to affect the coefficients for the usual activities dimension and the most severe level of anxiety and depression.

The EQ-5D was selected as the base measure for which bolt-ons were developed in this study. The EQ-5D was chosen as the reference measure as it is the most commonly used GPBM in economic evaluation and is recommended as the preferred GPBM by NICE in the UK.[1] A similar approach could be employed for other GPBMs if evidence were to suggest concerns regarding their responsiveness or validity. Indeed, one of the early studies in this area included a generic bolt item with a condition-specific measure of HRQL.[225] Developing bolt-ons to the other GPBMs considered in the review would require additional considerations to those identified for EQ-5D. For example, the valuation methods for bolt-ons to the HUI systems would need to be carefully considered. For the SF-6D, consideration would need to be given as to how the bolt-on would be presented to respondents given that the SF-6D values are usually derived by applying the SF-6D algorithm to responses from the SF-12 or SF-36 instruments.

The development of bolt-ons to EQ-5D could have significant implications for researchers and policy-makers who use QALY-based evaluations to inform their decision-making. Bolt-ons are likely to be particularly useful where there has been concern about the psychometric properties of EQ-5D in specific conditions, such as for hearing and some vision impairments as identified in the review presented here. Inclusion of the bolt-on items could improve the performance of generic measures, such as EQ-5D, for specific conditions, for example by increasing their responsiveness. This could be very attractive for policy-makers who require a degree of consistency in decision making, for example if they want to compare results with a common threshold value or to studies using the same outcome measure. The degree of consistency with the 'standard' EQ-5D value set is essentially an empirical issue, and needs to be considered relative to the alternative approaches or instruments. The results presented for the valuation of the EQ + vision bolt-on suggest that there are likely to be differences in EQ-5D values depending on the bolt-on included. While acknowledging this, one would expect the use of a common valuation methodology and a very similar descriptive system to produce more consistent valuations than an entirely different descriptive system and/or valuation method, although this needs to be confirmed empirically. The development of bolt-on items should not be viewed as an 'easy option' for those wishing to improve on the responsiveness or validity of EQ-5D or other GPBMs. A substantial amount of research has been conducted to develop and validate the EQ-5D and the other GPBMs included in the review presented here. If this approach is taken forward, it will be important to ensure that appropriate high-quality research underpinning each individual bolt-on is conducted and for EQ-5D, the valuation methods to be comparable to other EQ-5D valuation studies. This is a substantial and resource-intensive exercise. It should also include validation of the bolt-on measure, which is an area of further research for the bolt-on items developed for this study.

The main weakness of the studies presented is that the sample sizes were limited by the constraints of the costs of conducting face-to-face interviews with respondents. Many of the differences appeared potentially important and approached the 0.05 level of significance. Further research needs to be conducted using larger sample sizes.

Chapter 5 Discussion

The project had three main related objectives: (1) to establish where EQ-5D and other commonly used GPBMs are appropriate for measuring HRQL for economic evaluation, (2) to develop mapping functions to predict EQ-5D outcomes from condition-specific or clinical measures and to compare the performance of alternative model specifications, and (3) to investigate the development and valuation of bolt-ons to the EQ-5D descriptive system for those conditions in which EQ-5D is not sufficient. We have systematically reviewed the evidence and provided a narrative analysis of the performance of EQ-5D and two other widely used GPBMs (SF-6D and HUI3) in four broadly defined conditions: cancers, hearing impairments, skin conditions and vision impairment. We have tested alternative model specifications to map from cancer-specific measures of HRQL to EQ-5D. Finally, we have developed three potential bolt-ons to EQ-5D and estimated a full value set for a bolt-on for vision (EQ + vision). While the framing for this research has been to inform the methods of assessment used by NICE in its decision-making, the results are generalisable to other jurisdictions and/or uses of GPBMs.

Psychometric properties of the generic preference-based measures

Overall, the number of studies that use the EQ-5D is much larger than for HUI3 or SF-6D, with the exception of hearing-related conditions. The systematic reviews indicated that EQ-5D performs well in most cancers but performs poorly in hearing-related conditions. The evidence from studies of skin conditions suggested that EQ-5D performs well; however, most of the data relate to psoriasis and psoriatic arthritis. The results were mixed for conditions affecting vision and the performance depends on the nature and aetiology of the condition; specifically, the evidence showed good performance in cataracts and conjunctivitis but poor ability to assess severity in AMD and diabetic retinopathy, and mixed evidence for glaucoma. The evidence suggested that HUI3 is able to assess severity of HRQL for hearing impairments and some cancers and there was some evidence, albeit limited, that HUI3 captures HRQL for vision impairment with the exception of diabetic retinopathy. However, there was no evidence from HUI3 for skin conditions. There was very little evidence available to make an assessment of the performance of SF-6D in these conditions. There was also a complete lack of evidence on the reliability of all of the measures in vision, hearing and skin conditions.

A limitation of the assessment of validity and responsiveness of the GPBMs is that there is no gold standard of HRQL with which to compare the measures and, therefore, there will always be an element of subjectivity in this type of analysis. This limitation is not unique to the measures chosen here or to the focus on preference-based instruments. The literature review presented here utilised psychometric tests to evaluate the ability the GPBMs to reflect the impact of the conditions assessed on HRQL.[227] While many of the studies were not specifically designed to test the psychometric properties of the instruments, most reported data in sufficient detail to allow an assessment to be made of how well an instrument seems to capture the impact of a condition or treatment on HRQL. This approach to assessment relies on the measures used as comparators to adequately reflect HRQL. Some of the measures, particularly clinical indicators (such as VA) do not measure HRQL specifically and may only give a narrow representation of the disease and even where broader condition-specific measures are used, these do not reflect preferences or the relative values placed on different symptoms or health effects. Unfortunately, it was possible to extract only very limited information on the reliability and acceptability of the instruments from the studies identified.

With these limitations in mind, we were able to form an overall assessment on the performance of the GPBMs by considering the totality of the data reported by other instruments within the same studies. We have taken a systematic approach to reviewing and summarising the data. The conclusions from these assessments have been tabulated in terms of consistency in the direction of measures and statistical significance. It is not clear from the findings whether the poor performance of EQ-5D in hearing and some vision impairments are due to the inadequacies in the description of the five dimensions or in the number of levels for each dimension. Currently, the evidence points to the EQ-5D not properly capturing the impact of sensory impairments generally; however, there was some evidence that EQ-5D could distinguish between the most extreme differences in these conditions (for example, from the case–control studies). If this is the case, it is possible that increasing the number of levels of the instrument could improve performance; however, whether increasing the number of levels to five, as in the new version of EQ-5D, will be sufficient to overcome the problems of EQ-5D in sensory impairments remains to be seen. Research to establish whether the new descriptive system and the forthcoming new valuation set improve the performance of EQ-5D in these conditions would be helpful to understand the full impact of the additional levels.

The review of the performance of EQ-5D has focused on the three-level version of the instrument, as this is the most widely used version. A new version of the EQ-5D has been developed with the number of levels increased to five;[3] however, reported data from the new version are currently limited. This could improve the ability of EQ-5D to differentiate between levels of disease severity or assess responsiveness in hearing and vision impairments and further research in this area would be helpful.

Mapping to predict EQ-5D outcomes when data are unavailable

The results of the systematic review found that EQ-5D performed well in relation to cancers and skin conditions. Of these, cancer was chosen to be the focus of the mapping analyses and data sets containing EQ-5D and one or more condition-specific or clinical indicator were obtained. The data sets obtained included one of two commonly used cancer-specific QoL questionnaires: FACT-G or EORTC QLQ-C30.

There is little consensus in the published literature about how to select the best model from a mapping exercise, different criteria (AIC/BIC) will select different models as they give importance to different issues. Different measures of accuracy of predictions (e.g. MAE, RMSE) typically used in mapping studies were not developed for use in this situation where we are faced with individual level data. They are very insensitive given the high level of individual heterogeneity and the small range of the utility scale. Even after this, different measures will lead to different models being selected since they weight errors in different ways. There is no test for what model is best so a range of criteria need to be considered and a judgement made. A range of alternative model specifications were explored using both measures and data sets, and included different standard modelling approaches, explanatory variables and different representations of the dependent variable (EQ-5D index or dimensions). A variety of statistics were reported with some focusing on model goodness of fit and others on the predictive ability of the model. As the purpose of mapping is to predict values, it could be argued that we should give more weight to predictive ability; however, there are still a number of criteria that can be used to assessed predictive performance such as mean predictions, MAE and shrinkage, which we have reported here. Where mapping is used in practice, the aim is usually to estimate mean values for a set of health states, often defined in terms of severity, included within an economic model. Reviews of published mapping functions have found that they frequently do not give accurate predictions at the lower and upper end of the utility scale. In the analyses presented here, we have examined the accuracy of predictions for subgroups of responses defined according to different levels of severity using an external reference measure of health. Ideally, we would assess the define severity according to the measure(s) of severity included in the economic model for which the mapping has been conducted and then assess the predicted values relative to an external sample representative of the population of interest. The response mapping models to predict responses at the dimension level for EQ-5D performed best for the EORTC QLQ-C30 data. This was not the case for the FACT-G data set, which included a more limited range of EQ-5D data and few patients reporting very severe levels of health. Therefore, it was not possible to reliably map to all of the EQ-5D dimension levels. In that analysis, the linear regression models using OLS performed the best of the standard models according to the mean predictions and MAE for the overall sample and the subgroups defined according to severity; however, the model based on splining gave better median predictions and the response mapping model performed best in terms of shrinkage.

It is now widely observed, and has been further demonstrated in the analyses presented here, that the distribution of EQ-5D values observed in patient data are typically not normally distributed when the UK tariff is applied. The distribution is usually bimodal or multimodal and usually exhibits a large peak at 1 (full health) for all but the most severe of health conditions. In addition, there is a sizable gap in the values between the largest value (1) and the next largest (0.88). This is likely to cause problems for some of the standard statistical models. Response mapping has the capability of reflecting these features. Similarly, the limited dependent mixture model, reported here in illustrative analyses, is designed taking such features into account. For individual patient sampling models, these features are critical and also ensure that values outside the feasible range are not predicted. For cohort models, where the interest is in estimating the mean (and its uncertainty) for subgroups of patients, these models also offer the advantage that neither mean estimates nor their sampled values taking into account uncertainty lie outside the feasible range. When the number of subgroups is large and/or lie at the extremes of the EQ-5D range, then these features are of particular importance given how they are to be used in economic evaluation. When compared against an equivalent linear model, the LDVMM performs better on almost all relevant measures both for the sample as a whole and for severity defined subgroups. 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. Unfortunately, such data were not available for us to conduct this analysis but it would be a useful piece of further research if such data sets exist. The generalisability of the mapping algorithm predicting from the FACT-G study to populations including patients in the severest levels of health is limited as the data set only included few observations at the lower end of the HRQL scale.

The bolt-on studies

Mapping is not an effective solution to the problem of measuring and valuing HRQL where EQ-5D has been found to be inappropriate. There may be a preference for using the EQ-5D to maintain consistency between analyses and, therefore, adaptations to the questionnaire may be a potential solution. We examined a new approach of bolt-ons and developed and tested three bolt-on items in an exploratory study valuing nine health states from each descriptive system. We focused on the two areas where the EQ-5D was identified to have some problems in the literature review: hearing and vision. Furthermore, although the review found that EQ-5D performs well in cancer, there have been concerns about the face validity regarding the lack of an energy dimension in EQ-5D and this was also included in the exploratory analysis.

All three of the bolt-on items had an impact on TTO values for the EQ + bolt-on states, but the results suggested that the relationship may not be straightforward. The extent and direction of the impact of the bolt-on varied according to the level of severity of the bolt-on and the severity of the core EQ-5D state to which it was added. In most cases, including a level 1 bolt-on resulted in no difference or higher values, the addition of level 2 was mixed and the addition of level 3 led to lower values.

The results for the tiredness bolt-on differed to those from a previous study assessing the inclusion of a similar dimension within the EQ-5D.[226] However, this disagreement could be attributed to any number of differences in study design including the method of valuation (VAS compared with TTO) and the number and labelling of the bolt-on levels. All three of the bolt-ons in the exploratory study showed some impact. There did not appear to substantial differences between the three bolt-ons, although the impact appeared to be marginally strongest for the vision bolt-on and this was selected for full valuation. However, we believe that based on the results of the exploratory analysis, the tiredness and hearing bolt-on items warrant further investigation and development.

Given the results of the exploratory study, a full valuation of the vision bolt-on was conducted using face-to-face TTO interviews with members of the general public. The results of this study show that the vision bolt-on had a significant impact on EQ-5D state valuations. 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). This could be due to people focusing on the positive aspect of the health state or considering the absence of vision problems to be 'ray of light' in an otherwise severe health state. Some qualitative exploration of what people consider when responding would be informative.

It would be easier and less resource-intensive if future bolt-on items could be valued separately rather than conducting a valuation of the full bolt-on classification including the EQ-5D. In addition, it could potentially be advantageous for decision-makers if the values of the bolt-on items could be related back to a standard tariff. However, based on the results presented here, a model with a simple decrement for each of the bolt-on levels is not appropriate. A more sophisticated analysis that takes into account both the severity of the bolt-on and the severity of the core EQ-5D state to which it is added may be feasible. Whether a full valuation of the EQ + bolt-on instrument is required for each new bolt-on item is not clear. Unfortunately, the analysis comparing the coefficients of the models with and without the bolt-on was not conclusive. It showed that there were no statistically significant differences between the coefficients at the 5% level. However, the size of some differences in coefficients was not trivial and the lack of significant differences could have been due to the sample size. There is also the possibility that the impact is specific to the condition to which the bolt-on relates.

The limitations of the study include that the interviews were based in a specific region of the UK and may not be generalisable to other countries or indeed regions in the UK, although there is no clear reason to suppose that the pattern of results would be different elsewhere. Some differences in reported problems with vision were found between the groups in the exploratory study; however, the regression analysis showed that these characteristics did not significantly impact on values and the same finding was observed in the full valuation of EQ + vision. Another limitation is the lack of qualitative research to investigate acceptability and alternative phrasing of the bolt-ons; however, the labelling builds on the framework of the EQ-5D and the qualitative research that has been used to develop it. Finally, this study has focused on the three-level version for the EQ-5D and it is not clear if similar results would be seen with the five-level version.

A key feature of the EQ-5D is that it can be used across a range of conditions or diseases. This has a substantial advantage for economic evaluation and healthcare decision-making as it means decisions can be based on a common measure and applied consistently across evaluations. For specific conditions, where EQ-5D has been demonstrated to lack validity, the development of bolt-on instruments can offer a solution by improving the sensitivity of the instrument. While this may be at the expense of a level of consistency in the measurement and valuation of HRQL between conditions, retaining the EQ-5D as the basis for measurement may be beneficial. By retaining the EQ-5D as the core basis for measurement and by using a common valuation methodology, the degree of inconsistency in the estimates of HRQL is likely to be less than if alternative GPBMs or condition-specific PBMs are used instead.

Conclusion

This report has presented three substantial pieces of research. We have considered when specific GPBMs are appropriate for the measurement of HRQL, alternative methods for predicting outcomes when GPBMs have been found to be appropriate but data are unavailable and a method for developing bolt-ons to EQ-5D to improve its sensitivity. We have systematically reviewed the evidence on the performance of EQ-5D and two other commonly used GPBMs in four, very broadly defined, clinical areas. We found that EQ-5D performs well in most cancers and skin conditions, although evidence on reliability was lacking for the latter. We also found that EQ-5D shows mixed results in vision impairment and performs poorly in hearing-related conditions. Even where EQ-5D appears to be an appropriate measure of HRQL, data are not always collected within clinical studies. We have developed algorithms to predict EQ-5D outcomes from two commonly used cancer-specific measures of QoL and explored a range of alternative model specifications. Models predicting EQ-5D dimension-level responses performed best for one of the measures (EORTC QLQ-C30); however, this approach did not work well in an alternative data set including the FACT-G as it included patients with a narrower range of disease severity. In this latter data set, when considering standard models, the OLS regression performed best in terms of the accuracy of mean predictions for the whole sample and the subgroups defined according to severity. The LDVMM outperformed the linear model in illustrative analysis of a selected model. Three bolt-on items to EQ-5D were developed and tested in an exploratory study and a bolt-on for vision was tested further and a full set of valuations for EQ + vision obtained. The results of these studies show that the inclusion of a bolt-on item has a complex impact on EQ-5D values and the results have important implications for that valuation of future bolt-ons.

Recommendations for further research

Generic preference-based measures are widely used in the economic evaluation of health interventions and are used to inform the decision-making of bodies such as NICE in the UK. The research presented here has consolidated some of the existing research in this area and presented new areas of methodology. In order to ensure the most appropriate use of generic and condition-specific measures in HTA and health-care decision-making, further research is required. We have highlighted the areas that we consider to be priorities for further research below.

Psychometric properties of the generic preference-based measures in different conditions

The reviews of the psychometric properties of the GPBMs focused on four broadly defined conditions: hearing impairment, vision impairment, skin conditions and cancers. We recommend extending these reviews of the psychometric literature to more conditions. This would provide useful information and lead to recommendations on the use of the GPBMs for researchers conducting HTAs of interventions in other conditions.

Given the widespread use of the measures in HTA, the amount of evidence on psychometric properties of the instruments was limited and, in most cases, the studies had not been specifically designed to examine these issues. We recommend that more primary research or analyses of primary data sets into the psychometric properties of GPBMs is undertaken, particularly in cancer, and particularly of the reliability of the measures in the other conditions.

Mapping

It was not possible to validate the mapping functions estimated in this project using an external data set, but this is recommended to assess the external validity of the functions.

In addition, we recommend comparing alternative statistical models in larger data sets, including those for EORTC QLQ-C30 and FACT-G.

The development and use of bolt-ons to EQ-5D

The development and use of bolt-ons to EQ-5D is still a new but growing area of methodological research. The research presented in this report offers insights that can be used when developing future bolt-ons. Further research to validate the EQ + vision measure presented here would be useful. The results of the exploratory study of the hearing and tiredness bolt-ons suggest that these measures would also benefit from validation and further valuation. There are still methodological issues relating to bolt-on development that require further investigation. We recommend that the best way to undertake this is to develop a systematic programme of research into bolt-ons for EQ-5D.

Acknowledgements

This project was funded by the UK MRC as part of the MRC-NIHR methodology research programme (ref: G0901486). The authors would like to thank Rachel Ibbotson from the Centre for Health and Social Care Research Faculty of Health and Wellbeing at Sheffield Hallam University and the interviewers for conducting the valuation surveys. We would particularly like to thank all respondents for their participation in the studies. We are grateful to advice given by Diana Papaionnou, two cancer data sets kindly provided by Stuart Peacock and a further data set provided by Simon Pickard. Research ethics approval has been given by the ethics committee at Brunel University in line with Brunel University research governance and ethics requirement.

Contributions of authors

Louise Longworth led the project and contributed to the methodology and interpretation of results at each stage.

Yaling Yang led the literature reviews of the GPBMs in hearing impairment and skin conditions, and contributed to the literature review of the GPBMs in cancers reported in *Chapter 2*. She also contributed to the data collection, analysis and interpretation of results for the bolt-on studies reported in *Chapter 4*.

Tracey Young led the mapping analyses reported in *Chapter 3* and contributed to the methodology and interpretation of results at each stage.

Brendan Mulhern co-ordinated the literature review of the GPBMs in cancers reported in *Chapter 2*.

Mónica Hernández Alava contributed to the mapping analyses reported in *Chapter 3*.

Clara Mukuria contributed to the mapping analyses reported in *Chapter 3*.

Donna Rowen contributed to the methodology and interpretation of results at each stage.

Jonathan Tosh led the literature review of the GPBMs in vision impairments reported in *Chapter 2*.

Aki Tsuchiya contributed to the methodology and interpretation of results at each stage.

Pippa Evans conducted the search strategies for the literature reviews of GPBMs reported in *Chapter 2*.

Anju Devianee Keetharuth contributed to the literature review of the GPBMs in cancers reported in *Chapter 2*.

John Brazier contributed to the methodology and interpretation of results at each stage.

LL, **TY**, **MH**, **DR**, **AT** and **JB** contributed to the conceptualisation and overall design of the project. All authors contributed to the drafting of the report.

References

1. National Institute of Health and Care Excellence (NICE) (formerly the National Institute of Health and Clinical Excellence). *NICE Guide to the Methods of Technology Appraisal*. London: NICE; 2008.
2. Brooks R. EuroQol: the current state of play. *Health Policy* 1996;**37**:53–72. http://dx.doi.org/10.1016/0168-8510(96)00822-6
3. Herdman M, Gudex C, Lloyd A, Janssen M, Kind P, Parkin D, *et al.* Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Qual Life Res* 2011;**20**:1727–36. http://dx.doi.org/10.1007/s11136-011-9903-x
4. Dolan P. Modeling Valuations for EuroQol health states. *Med Care* 1997;**35**:1095–108. http://dx.doi.org/10.1097/00005650-199711000-00002
5. Longworth L, Longson C. NICE methodology for technology appraisals: cutting edge or tried and trusted? *Pharmacoeconomics* 2008;**26**:729–32. http://dx.doi.org/10.2165/00019053-200826090-00003
6. Brazier J, Roberts J, Tsuchiya A, Busschbach J. A comparison of the EQ-5D and SF-6D across seven patient groups. *Health Econ* 2004;**13**:873–84. http://dx.doi.org/10.1002/hec.866
7. Longworth L, Bryan S. An empirical comparison of EQ-5D and SF-6D in liver transplant patients. *Health Econ* 2003;**12**:1061–7. http://dx.doi.org/10.1002/hec.787
8. Whitehurst DG, Bryan S, Lewis M. Systematic review and empirical comparison of contemporaneous EQ-5D and SF-6D group mean scores. *Med Decis Making* 2011;**31**:E34–44. http://dx.doi.org/10.1177/0272989X11421529
9. Guyatt G. Generic and specific measurement of health-related quality of life in a clinical trial of respiratory rehabilitation. *J Clin Epidemiol* 1999;**52**:187–92. http://dx.doi.org/10.1016/S0895-4356(98)00157-7
10. Jenkinson C, Gray A, Doll H, Lawrence K, Keoghane S, Layte R. Evaluation of index and profile measures of health status in a randomized controlled trial. Comparison of the medical outcomes study 36-item short form health survey, EuroQol, and disease specific measures. *Med Care* 1997;**35**:1109–18. http://dx.doi.org/10.1097/00005650-199711000-00003
11. Brazier JE, Yang Y, Tsuchiya A, Rowen DL. A review of studies mapping (or cross walking) non-preference based measures of health to generic preference-based measures. *Eur J Health Econ* 2010;**11**:215–25. http://dx.doi.org/10.1007/s10198-009-0168-z
12. Longworth L, Buxton MJ, Sculpher M, Smith DH. Estimating utility data from clinical indicators for patients with stable angina. *Eur J Health Econ* 2005;**6**:347–53. http://dx.doi.org/10.1007/s10198-005-0309-y
13. Longworth L, Rowen D. Mapping to obtain EQ-5D utility values for use in NICE health technology assessments. *Value Health* 2013;**16**:202–10. http://dx.doi.org/10.1016/j.jval.2012.10.010
14. 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**:1–14.
15. Yang Y, Brazier J, Tsuchiya A, Coyne K. Estimating a preference-based single index from the overactive bladder questionnaire. *Value Health* 2009;**12**:159–66. http://dx.doi.org/10.1111/j.1524-4733.2008.00413.x

16. Yang Y, 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**:281–91. http://dx.doi.org/10.1177/0272989X10379646

17. Brazier J, Tsuchiya A. Preference-based condition-specific measures of health: what happens to cross programme comparability? *Health Econ* 2010;**19**:125–9. http://dx.doi.org/10.1002/hec.1580

18. Krabbe PFM, Stouthard MEA, Essink-Bot ML, Bonsel GJ. The effect of adding a cognitive dimension to the EuroQol multiattribute health-status classification system. *J Clin Epidemiol* 1999;**52**:293–301. http://dx.doi.org/10.1016/S0895-4356(98)00163-2

19. Yang Y, Brazier JE, Tsuchiya A. The effect of adding a 'sleep' dimension to the EQ-5D descriptive system [published online ahead of print March 22 2013]. *Med Decis Making* 2013. http://dx.doi.org/10.1177/0272989X13480428

20. Tosh JC, Longworth LJ, George E. Utility values in National Institute for Health and Clinical Excellence (NICE) technology appraisals. *Value Health* 2011;**14**:102–9. http://dx.doi.org/10.1016/j.jval.2010.10.015

21. Barton GR, Bankart J, Davis AC. A comparison of the quality of life of hearing-impaired people as estimated by three different utility measures. *Int J Audiol* 2005;**44**:157–63. http://dx.doi.org/10.1080/14992020500057566

22. Espallargues M, Czoski-Murray CJ, Bansback NJ, Carlton J, Lewis GM, Hughes LA, *et al.* The impact of age-related macular degeneration on health status utility values. *Invest Ophthalmol Vis Sci* 2005;**46**:4016–23. http://dx.doi.org/10.1167/iovs.05-0072

23. Grutters JPC, Joore MA, van der Horst F, Verschuure H, Dreschler WA, Anteunis LJC. Choosing between measures: comparison of EQ-5D, HUI2 and HUI3 in persons with hearing complaints. *Qual Life Res* 2007;**16**:1439–49. http://dx.doi.org/10.1007/s11136-007-9237-x

24. Oostenbrink R, Moll A, Essink-Bot ML. The EQ-5D and the Health Utilities Index for permanent sequelae after meningitis: a head-to-head comparison. *J Clin Epidemiol* 2002;**55**:791–9.

25. National Institute of Health and Care Excellence (NICE). *Cochlear implants for children and adults with severe to profound deafness*. NICE Technical Appraisal TA166. London, UK; 2009.

26. National Institute of Health and Care Excellence (NICE). *Ranibizumab and pegaptanib for the treatment of age-related macular degeneration*. NICE Technical Appraisal TA 155. London, UK; 2008.

27. Papaioannou D, Brazier J, Parry G. How valid and responsive are generic health status measures, such as EQ-5D and SF-36, in schizophrenia? A systematic review. *Value Health* 2011;**14**:907–20. http://dx.doi.org/10.1016/j.jval.2011.04.006

28. Papaioannou D, Brazier J, Parry G. How to measure quality of life for cost effectiveness analyses in personality disorders? A systematic review. *J Pers Disor* 2013;**27**:383–401. http://dx.doi.org/10.1521/pedi_2013_27_075

29. Dolan P, Gudex C, Kind P, Williams A. The time trade-off method: results from a general population study. *Health Econ* 1996;**5**:141–54. http://dx.doi.org/10.1002/(SICI)1099-1050(199603)5:2<141::AID-HEC189>3.0.CO;2-N

30. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. *J Health Econ* 2002;**21**:271–92. http://dx.doi.org/10.1016/S0167-6296(01)00130-8

31. Feeny D, Furlong W, Torrance GW, Goldsmith CH, Zhu Z, DePauw S, *et al.* Multiattribute and single-attribute utility functions for the health utilities index mark 3 system. *Med Care* 2002;**40**:113–28. http://dx.doi.org/10.1097/00005650-200202000-00006

32. The EuroQol Group. *EuroQol*. URL: www.euroqol.org (accessed 1 August 2010).

33. Brazier J, Deverill M. A checklist for judging preference based measures of health related quality of life: learning from psychometrics. *Health Econ* 1999; **8**:41–51. http://dx.doi.org/10.1002/(SICI)1099-1050(199902)8:1<41::AID-HEC395>3.3.CO;2-R

34. Mittmann N, Chan D, Trakas K, Risebrough N. Health utility attributes for chronic conditions. *Dis Manag Health Out* 2001;**9**:11–21. http://dx.doi.org/10.2165/00115677-200109010-00002

35. Cheung MC, Imrie KR, Friedlich J, Buckstein R, Hicks LK, Zaretsky Y, *et al.* The potential for lost productivity and daily activity impairment in patients with follicular (FL) and other indolent non-Hodgkin's lymphoma (NHL). *Blood* 2006;**108**:344.

36. Asakawa K, Rolfson D, Senthilselvan A, Feeny D, Johnson JA. Health Utilities Index Mark 3 showed valid in Alzheimer disease, arthritis, and cataracts. *J Clin Epidemiol* 2008;**61**:733–9. http://dx.doi.org/10.1016/j.jclinepi.2007.09.007

37. Polack S, Kuper H, Mathenge W, Fletcher A, Foster A. Cataract visual impairment and quality of life in a Kenyan population. *Br J Ophthalmol* 2007;**91**:927–32. http://dx.doi.org/10.1136/bjo.2006.110973

38. Polack S, Kuper H, Wadud Z, Fletcher A, Foster A. Quality of life and visual impairment from cataract in Satkhira district, Bangladesh. *Br J Ophthalmol* 2008;**92**:1026–30. http://dx.doi.org/10.1136/bjo.2007.134791

39. Polack S, Eusebio C, Fletcher A, Foster A, Kuper H. Visual impairment from cataract and health related quality of life: results from a case–control study in the Philippines. *Ophthal Epidemiol* 2010;**17**:152–9. http://dx.doi.org/10.3109/09286581003731536

40. Boulton M, Haines L, Smyth D, Fielder A. Health-related quality of life of children with vision impairment or blindness. *Dev Med Child Neurol* 2006;**48**:656–61. http://dx.doi.org/10.1111/j.1469-8749.2006.tb01335.x

41. Langelaan M, de Boer MR, van Nispen RM, Wouters B, Moll AC, van Rens GH, *et al.* Impact of visual impairment on quality of life: a comparison with quality of life in the general population and with other chronic conditions. *Ophthal Epidemiol* 2007;**14**:119–26. http://dx.doi.org/10.1080/09286580601139212

42. Lloyd A, Nafees B, Gavriel S, Rousculp MD, Boye KS, Ahmad A. Health utility values associated with diabetic retinopathy. *Diabet Med* 2008;**25**:618–24. http://dx.doi.org/10.1111/j.1464-5491.2008.02430.x

43. Soubrane G, Cruess A, Lotery A, Pauleikhoff D, Mones J, Xu X, *et al.* Burden and health care resource utilization in neovascular age-related macular degeneration: findings of a multicountry study. *Arch Ophthalmol* 2007;**125**:1249–54. http://dx.doi.org/10.1001/archopht.125.9.1249

44. Aspinall PA, Johnson ZK, Azuara-Blanco A, Montarzino A, Brice R, Vickers A. Evaluation of quality of life and priorities of patients with glaucoma. *Invest Ophth Visual* 2008;**49**:1907–15. http://dx.doi.org/10.1167/iovs.07-0559

45. Kobelt G, Jonsson B, Bergstrom A, Chen E, Linden C, Alm A. Cost-effectiveness analysis in glaucoma: what drives utility? Results from a pilot study in Sweden. *Acta Ophthalmologica Scand* 2006;**84**:363–71. http://dx.doi.org/10.1111/j.1600-0420.2005.00621.x

46. Montemayor F, Sibley LM, Courtright P, Mikelberg FS. Contribution of multiple glaucoma medications to visual function and quality of life in patients with glaucoma. *Can J Ophthalmol* 2001;**36**:385–90.

47. Cruess A, Zlateva G, Xu X, Rochon S. Burden of illness of neovascular age-related macular degeneration in Canada. *Can J Ophthalmol* 2007;**42**:836–43. http://dx.doi.org/10.3129/i07-153

48. Lotery A, Xu X, Zlatava G, Loftus J. Burden of illness, visual impairment and health resource utilisation of patients with neovascular age-related macular degeneration: results from the UK cohort of a five–country cross-sectional study. *Br J Ophthalmol* 2007;**91**:1303–7. http://dx.doi.org/10.1136/bjo.2007.116939

49. Payakachat N, Summers KH, Pleil AM, Murawski MM, Thomas J 3rd, Jennings K, *et al.* Predicting EQ-5D utility scores from the 25-item National Eye Institute Vision Function Questionnaire (NEI-VFQ 25) in patients with age-related macular degeneration. *Qual Life Res* 2009;**18**:801–13. http://dx.doi.org/10.1007/s11136-009-9499-6

50. Smith DH, Johnson ES, Russell A, Hazlehurst B, Muraki C, Nichols GA, *et al.* Lower visual acuity predicts worse utility values among patients with type 2 diabetes. *Qual Life Res* 2008;**17**:1277–84. http://dx.doi.org/10.1007/s11136-008-9399-1

51. Rajagopalan K, Abetz L, Mertzanis P, Espindle D, Begley C, Chalmers R, *et al.* Comparing the discriminative validity of two generic and one disease-specific health-related quality of life measures in a sample of patients with dry eye. *Value Health* 2005;**8**:168–74. http://dx.doi.org/10.1111/j.1524-4733.2005.03074.x

52. van Nispen RM, de Boer MR, Hoeijmakers JG, Ringens PJ, van Rens GH. Co-morbidity and visual acuity are risk factors for health-related quality of life decline: five-month follow-up EQ-5D data of visually impaired older patients. *Health Qual Life Outcomes* 2009;**7**:18. http://dx.doi.org/10.1186/1477-7525-7-18

53. Datta S, Foss AJE, Grainge MJ, Gregson RM, Zaman A, Masud T, *et al.* The importance of acuity, stereopsis, and contrast sensitivity for health-related quality of life in elderly women with cataracts. *Invest Ophth Visual* 2008;**49**:1–6. http://dx.doi.org/10.1167/iovs.06-1073

54. Thygesen J, Aagren M, Arnavielle S, Bron A, Frohlich SJ, Baggesen K, *et al.* Late-stage, primary open-angle glaucoma in Europe: social and health care maintenance costs and quality of life of patients from 4 countries. *Curr Med Res Opin* 2008;**24**:1763–70. http://dx.doi.org/10.1185/03007990802111068

55. Kim J, Kwak HW, Lee WK, Kim HK. Impact of photodynamic therapy on quality of life of patients with age-related macular degeneration in Korea. *Jpn J Ophthalmol* 2010;**54**:325–30. http://dx.doi.org/10.1007/s10384-010-0825-x

56. Ruiz-Moreno JM, Coco RM, Garcia-Arumi J, Xu X, Zlateva G. Burden of illness of bilateral neovascular age-related macular degeneration in Spain. *Curr Med Res Opin* 2008;**24**:2103–11.

57. Black N, Browne J, van der Meulen J, Jamieson L, Copley L, Lewsey J. Is there overutilisation of cataract surgery in England? *Br J Ophthalmol* 2009;**93**:13–17. http://dx.doi.org/10.1136/bjo.2007.136150

58. Conner-Spady BL, Sanmugasunderam S, Courtright P, Mildon D, McGurran JJ, Noseworthy TW, *et al.* The prioritization of patients on waiting lists for cataract surgery: validation of the Western Canada waiting list project cataract priority criteria tool. *Ophthal Epidemiol* 2005;**12**:81–90. http://dx.doi.org/10.1080/09286580590932770

59. Jayamanne DG, Allen ED, Wood CM, Currie S. Correlation between early, measurable improvement in quality of life and speed of visual rehabilitation after phacoemulsification. *J Cataract Refr Surg* 1999;**25**:1135–9. http://dx.doi.org/10.1016/S0886-3350(99)00138-8

60. Pitt AD, Smith AF, Lindsell L, Voon LW, Rose PW, Bron AJ. Economic and quality-of-life impact of seasonal allergic conjunctivitis in Oxfordshire. *Ophthal Epidemiol* 2004;**11**:17–33. http://dx.doi.org/10.1076/opep.11.1.17.26437

61. Smith AF, Pitt AD, Rodruiguez AE, Alio JL, Marti N, Teus M, *et al.* The economic and quality of life impact of seasonal allergic conjunctivitis in a Spanish setting. *Ophthal Epidemiol* 2005;**12**:233–42. http://dx.doi.org/10.1080/09286580590967781

62. Clark A, Ng JQ, Morlet N, Tropiano E, Mahendran P, Spilsbury K, *et al.* Quality of life after postoperative endophthalmitis. *Clin Exp Ophthalmol* 2008;**36**:526–31. http://dx.doi.org/10.1111/j.1442-9071.2008.01827.x

63. Kempen JH, Martin BK, Wu AW, Barron B, Thorne JE, Jabs DA, *et al.* The effect of cytomegalovirus retinitis on the quality of life of patients with AIDS in the era of highly active antiretroviral therapy. *Ophthalmology* 2003;**110**:987–95. http://dx.doi.org/10.1016/S0161-6420(03)00089-7

64. Quinn GE, Dobson V, Saigal S, Phelps DL, Hardy RJ, Tung B, *et al.* Health-related quality of life at age 10 years in very low-birth-weight children with and without threshold retinopathy of prematurity. *Arch Ophthalmol* 2004;**122**:1659–66.

65. Barton GR, Stacey PC, Fortnum HM, Summerfield AQ. Hearing-impaired children in the United Kingdom. IV: Cost-effectiveness of pediatric cochlear implantation. *Ear Hear* 2006;**27**:575–88. http://dx.doi.org/10.1097/01.aud.0000233967.11072.24

66. Lovett RES, Kitterick PT, Hewitt CE, Summerfield AQ. Bilateral or unilateral cochlear implantation for deaf children: an observational study. *Arch Dis Child* 2010;**95**:107–12. http://dx.doi.org/10.1136/adc.2009.160325

67. Smith-Olinde L, Grosse SD, Olinde F, Martin PF, Tilford JM. Health state preference scores for children with permanent childhood hearing loss: a comparative analysis of the QWB and HUI3. *Qual Life Res* 2008;**17**:943–53. http://dx.doi.org/10.1007/s11136-008-9358-x

68. Bichey BG, Hoversland JM, Wynne MK, Miyamoto RT. Changes in quality of life and the cost-utility associated with cochlear implantation in patients with large vestibular aqueduct syndrome. *Otol Neurotol* 2002;**23**:323–7. http://dx.doi.org/10.1097/00129492-200205000-00016

69. Damen GWJA, Beynon AJ, Krabbe PFM, Mulder JJS, Mylanus EAM. Cochlear implantation and quality of life in postlingually deaf adults: long-term follow-up. *Otolaryng Head Neck* 2007;**136**:597–604. http://dx.doi.org/10.1016/j.otohns.2006.11.044

70. Hol MKS, Spath MA, Krabbe PFM, van der Pouw CTM, Snik AFM, Cremers CWRJ, *et al.* The bone-anchored hearing aid – quality-of-life assessment. *Arch Otolaryng Head Neck* 2004;**130**:394–9. http://dx.doi.org/10.1001/archotol.130.4.394

71. Joore M, Brunenberg D, Zank H, van der Stel H, Anteunis L, Boas G, *et al.* Development of a questionnaire to measure hearing-related health state preferences framed in an overall health perspective. *Int J Technol Assess* 2002;**18**:528–39.

72. Joore MA, van der Stel H, Peters HJM, Boas GM, Anteunis JC. The cost-effectiveness of hearing-aid fitting in the Netherlands. *Arch Otolaryng Head Neck* 2003;**129**:297–304. http://dx.doi.org/10.1001/archotol.129.3.297

73. Joore MAB, Brunenberg DEM, Chenault MN and Anteunis LJC. Societal effects of hearing aid fitting among the moderately hearing impaired. *Int J Audiol* 2003;**42**;152–60. http://dx.doi.org/10.3109/14992020309090424

74. Joore MA, Potjewijd J, Timmerman AA, Anteunis LJ. Response shift in the measurement of quality of life in hearing impaired adults after hearing aid fitting. *Qual Life Res* 2002;**11**:299–307.

75. Palmer CS, Niparko JK, Wyatt JR, Rothman M, de Lissovoy G. A prospective study of the cost-utility of the multichannel cochlear implant. *Arch Otolaryng Head Neck* 1999;**125**:1221–8. http://dx.doi.org/10.1001/archotol.125.11.1221

76. Sach TH, Barton GR. Interpreting parental proxy reports of (health-related) quality of life for children with unilateral cochlear implants. *Int J Pediatr Otorhi* 2007;**71**:435–45. http://dx.doi.org/10.1016/j.ijporl.2006.11.011

77. Vuorialho A, Karinen P, Sorri M. Counselling of hearing aid users is highly cost-effective. *European Arch Oto-Rhino-Lary* 2006;**263**:988–95. http://dx.doi.org/10.1007/s00405-006-0104-0

78. Vuorialho A, Karinen P, Sorri M. Effect of hearing aids on hearing disability and quality of life in the elderly. *Int J Audiol* 2006;**45**:400–5. http://dx.doi.org/10.1080/14992020600625007

79. Lee HY, Park EC, Joong Kim H, Choi JY, Kim HN. Cost-utility analysis of cochlear implants in Korea using different measures of utility. *Acta Otolaryngol* 2006;**126**:817–23. http://dx.doi.org/10.1080/00016480500525213

80. Cheng AK, Rubin HR, Powe NR, Mellon NK, Francis HW, Niparko JK. Cost-utility analysis of the cochlear implant in children. *JAMA* 2000;**284**:850–6. http://dx.doi.org/10.1001/jama.284.7.850

81. Klassen AF, Newton JN, Mallon E. Measuring quality of life in people referred for specialist care of acne: Comparing generic and disease-specific measures. *J Am Acad Dermatol* 2000;**43**:229–33. http://dx.doi.org/10.1067/mjd.2000.105507

82. Van De Kerkhof PCM. The impact of a two-compound product containing calcipotriol and betamethasone dipropionate (Daivobet/Dovobet) on the quality of life in patients with psoriasis vulgaris: A randomized controlled trial. *Br J Dermatol* 2004;**151**:663–8. http://dx.doi.org/10.1111/j.1365-2133.2004.06134.x

83. Bansback NJ, Ara R, Barkham N, Brennan A, Fraser AD, Conway P, *et al.* Estimating the cost and health status consequences of treatment with TNF antagonists in patients with psoriatic arthritis (Structured abstract). *Rheumatology* 2006;**45**:1029–38.

84. Daudén E, Griffiths CE, Ortonne JP, Kragballe K, Molta CT, Robertson D, *et al.* Improvements in patient-reported outcomes in moderate-to-severe psoriasis patients receiving continuous or paused etanercept treatment over 54 weeks: the CRYSTEL study. *J Eur Acad Dermatol* 2009;**23**:1374–82. http://dx.doi.org/10.1111/j.1468-3083.2009.03321.x

85. Reich K, Segaert S, Van de Kerkhof P, Durian C, Boussuge MP, Paolozzi L, *et al.* Once-weekly administration of etanercept 50 mg improves patient-reported outcomes in patients with moderate-to-severe plaque psoriasis. *Dermatology* 2009;**219**:239–49. http://dx.doi.org/10.1159/000237871

86. Shikiar R, Heffernan M, Langley RG, Willian MK, Okun MM, Revicki DA. Adalimumab treatment is associated with improvement in health-related quality of life in psoriasis: Patient-reported outcomes from a Phase II randomized controlled trial. *J Dermatol Treat* 2007;**18**:25–31. http://dx.doi.org/10.1080/09546630601121060

87. Weiss SC, Kimball AB, Liewehr DJ, Blauvelt A, Turner ML, Emanuel EJ. Quantifying the harmful effect of psoriasis on health-related quality of life. *J Am Acad Dermatol* 2002;**47**:512–18. http://dx.doi.org/10.1067/mjd.2002.122755

88. Weiss SC, Rehmus W, Kimball AB. An assessment of the cost-utility of therapy for psoriasis. *Therapeutics Clin Risk Manag* 2006;**2**:325–8. http://dx.doi.org/10.2147/tcrm.2006.2.3.325

89. Matusiak L, Bieniek A, Szepietowski JC. Psychophysical aspects of hidradenitis suppurativa. *Acta Dermato-Venereologica* 2010;**90**:264–8.

90. Moberg C, Alderling M and Meding B. Hand eczema and quality of life: a population-based study. *Br J Dermatol* 2009;**161**:397–403. http://dx.doi.org/10.1111/j.1365-2133.2009.09099.x

91. Walters SJ, Morrell CJ, Dixon S. Measuring health-related quality of life in patients with venous leg ulcers. *Qual Life Res* 1999;**8**:327–36.

92. Brodszky V, Péntek M, Bálint PV, Géher P, Hajdu O, Hodinka L, *et al.* Comparison of the psoriatic arthritis quality of life (PsAQoL) questionnaire, the functional status (HAQ) and utility (EQ-5D) measures in psoriatic arthritis: results from a cross-sectional survey. *Scand J Rheumatol* 2010;**39**:303–9. http://dx.doi.org/10.3109/03009740903468982

93. Christophers E, Barker JN, Griffiths CE, Daudén E, Milligan G, Molta C, *et al.* The risk of psoriatic arthritis remains constant following initial diagnosis of psoriasis among patients seen in European dermatology clinics. *J Eur Acad Dermatol* 2010;**24**:548–54. http://dx.doi.org/10.1111/j.1468-3083.2009.03463.x

94. Revicki D, Willian MK, Saurat JH, Papp KA, Ortonne JP, Sexton C, *et al.* Impact of adalimumab treatment on health-related quality of life and other patient-reported outcomes: Results from a 16-week randomized controlled trial in patients with moderate to severe plaque psoriasis. *Br J Dermatol* 2008;**158**:549–57. http://dx.doi.org/10.1111/j.1365-2133.2007.08236.x

95. Shikiar R, Willian MK, Okun MM, Thompson CS, Revicki DA. The validity and responsiveness of three quality of life measures in the assessment of psoriasis patients: results of a phase II study. *Health Qual Life Outcomes* 2006;**4**:71.

96. Luger TA, Barker J, Lambert J, Yang S, Robertson D, Foehl J, *et al.* Sustained improvement in joint pain and nail symptoms with etanercept therapy in patients with moderate-to-severe psoriasis. *J Eur Acad Dermatol* 2009;**23**:896–904. http://dx.doi.org/10.1111/j.1468-3083.2009.03211.x

97. Wang HM, Beyer M, Gensichen J, Gerlach FM. Health-related quality of life among general practice patients with differing chronic diseases in Germany: cross sectional survey. *BMC Public Health* 2008;**8**:246. http://dx.doi.org/10.1186/1471-2458-8-246

98. Barton GR, Sach TH, Doherty M, Avery AJ, Jenkinson C, Muir KR. An assessment of the discriminative ability of the EQ-5Dindex, SF-6D, and EQ VAS, using sociodemographic factors and clinical conditions. *Eur J Health Econ* 2008;**9**:237–49. http://dx.doi.org/10.1007/s10198-007-0068-z

99. Bowker SL, Pohar SL, Johnson JA. A cross-sectional study of health-related quality of life deficits in individuals with comorbid diabetes and cancer. *Health Qual Life Outcomes* 2006;**4**:17.

100. Norum J. Quality of life (QoL) measurement in economical analysis in cancer: a comparison of the EuroQol questionnaire, a simple QoL-scale and the global QoL measure of the EORTC QLQ-C30. *Oncol Rep* 1996;**3**:787–91.

101. Falicov A, Fisher CG, Sparkes J, Boyd MC, Wing PC, Dvorak MF. Impact of surgical intervention on quality of life in patients with spinal metastases. *Spine* 2006;**31**:2849–56. http://dx.doi.org/10.1097/01.brs.0000245838.37817.40

102. Lathia N, Isogai P, Mittmann N, DeAngelis C, Cheung M, Sandra K, *et al.* Comprehensiveness of quality of life instruments in capturing concerns related to chemotherapy-induced neutropenia. *Blood* 2008;**112**:1311.

103. Pickard AS, Neary MP, Cella D. Estimation of minimally important differences in EQ-5D utility and VAS scores in cancer. *Health Qual Life Outcomes* 2007;**5**:70. http://dx.doi.org/10.1186/1477-7525-5-70

104. Ravasco P, Monteiro-Grillo I, Camilo ME. Does nutrition influence quality of life in cancer patients undergoing radiotherapy? *Radiother Oncol* 2003;**67**:213–20. http://dx.doi.org/10.1016/S0167-8140(03)00040-9

105. Cheung YB, Thumboo J, Gao F, Ng GY, Pang G, Koo WH, *et al.* Mapping the English and Chinese versions of the functional assessment of cancer therapy-general to the EQ-5D utility index. *Value Health* 2009;**12**:371–6. http://dx.doi.org/10.1111/j.1524-4733.2008.00448.x

106. Chow WH, Chang P, Lee SC, Wong A, Shen HM, Verkooijen HM, *et al.* Complementary and alternative medicine among Singapore cancer patients. *Ann Acad Med Singapore* 2010;**39**:129–7.

107. Capuano G, Pavese I, Satta F, Tosti M, Del Grosso A, Di Palma M. Correlation between anemia, unintentional weight loss and inflammatory status on cancer-related fatigue and quality of life before chemo and radiotherapy. *e-SPEN* 2008;**3**:e147–51. http://dx.doi.org/10.1016/j.eclnm.2008.04.008

108. Sung L, Greenberg ML, Doyle JJ, Young NL, Ingber S, Rubenstein J, *et al.* Construct validation of the Health Utilities Index and the Child Health Questionnaire in children undergoing cancer chemotherapy. *Br J Cancer* 2003;**88**:1185–90. http://dx.doi.org/10.1038/sj.bjc.6600895

109. Trudel JG, Rivard M, Dobkin RL, Leclerc JM, Robaey R. Psychometric properties of the Health Utilities Index Mark 2 system in paediatric oncology patients. *Qual Life Res* 1998;**7**:421–32. http://dx.doi.org/10.1023/A:1008857920624

110. Weze C, Leathard HL, Grange J, Tiplady P, Stevens G. Evaluation of healing by gentle touch in 35 clients with cancer. *Eur J Oncol Nurs* 2004;**8**:40–9. http://dx.doi.org/10.1016/j.ejon.2003.10.004

111. Mantovani G, Madeddu C, Macciò A, Gramignano G, Lusso MR, Massa E, *et al.* Cancer-related anorexia/cachexia syndrome and oxidative stress: An innovative approach beyond current treatment. *Cancer Epidemiology Biomarkers Prev* 2004;**13**:1651–9.

112. Vaghela C, Robinson N, Gore J, Peace B, Lorenc A. Evaluating healing for cancer in a community setting from the perspective of clients and healers: a pilot study. *Complement Ther Clin Pract* 2007;**13**:240–9. http://dx.doi.org/10.1016/j.ctcp.2007.03.004

113. Kim SW, Shin IS, Kim JM, Kim YC, Kim KS, Kim KM, *et al.* Effectiveness of mirtazapine for nausea and insomnia in cancer patients with depression. *Psychiatry Clin Neurosci* 2008;**62**:75–83. http://dx.doi.org/10.1111/j.1440-1819.2007.01778.x

114. Pickard AS, Kohlmann T, Janssen MF, Bonsel G, Rosenbloom S, Cella D. *et al.* Evaluating equivalency between response systems: application of the Rasch model to a 3-level and 5-level EQ-5D. *Med Care* 2007;**45**:812–9. http://dx.doi.org/10.1097/MLR.0b013e31805371aa

115. Doornebosch PG, Tollenaar RA, Gosselink MP, Stassen LP, Dijkhuis CM, Schouten WR, *et al.* Quality of life after transanal endoscopic microsurgery and total mesorectal excision in early rectal cancer. *Colorectal Dis* 2007;**9**:553–8. http://dx.doi.org/10.1111/j.1463-1318.2006.01186.x

116. Doornebosch PG, Gosselink MP, Neijenhuis PA, Schouten WR, Tollenaar RAEM, de Graaf EJR. Impact of transanal endoscopic microsurgery on functional outcome and quality of life. *Int J Colorectal Dis* 2008;**23**:709–13. http://dx.doi.org/10.1007/s00384-008-0442-z

117. Hamashima C. Long-term quality of life of postoperative rectal cancer patients. *J Gastroenterol Hepatol* 2002;**17**:571–6. http://dx.doi.org/10.1046/j.1440-1746.2002.02712.x

118. Siena S, Peeters M, Van Cutsem E, Humblet Y, Conte P, Bajetta E, *et al.* Association of progression-free survival with patient-reported outcomes and survival: Results from a randomised phase 3 trial of panitumumab. *Br J Cancer* 2007;**97**:1469–74. http://dx.doi.org/10.1038/sj.bjc.6604053

119. Anderson H, Palmer MK. Measuring quality of life: impact of chemotherapy for advanced colorectal cancer. Experience from two recent large phase III trials. *Br J Cancer* 1998;**77**(Suppl. 2):9–14. http://dx.doi.org/10.1038/bjc.1998.420

120. Wilson TR, Alexander DJ, Kind P. Measurement of health-related quality of life in the early follow-up of colon and rectal cancer. *Dis Colon Rectum* 2006;**49**:1692–702. http://dx.doi.org/10.1007/s10350-006-0709-9

121. Gosselink MP, Busschbach JJ, Dijkhuis CM, Stassen LP, Hop WC, Schouten WR, *et al.* Quality of life after total mesorectal excision for rectal cancer. *Colorectal Dis* 2006;**8**:15–22. http://dx.doi.org/10.1111/j.1463-1318.2005.00836.x

122. Janson M, Lindholm E, Anderberg B, Haglind E. Randomized trial of health-related quality of life after open and laparoscopic surgery for colon cancer. *Surg Endosc* 2007;**21**:747–53. http://dx.doi.org/10.1007/s00464-007-9217-9

123. Sharma A, Sharp DM, Walker LG, Monson JR, Sharma A. Predictors of early postoperative quality of life after elective resection for colorectal cancer. *Ann Surg Oncol* 2007;**14**:3435–42. http://dx.doi.org/10.1245/s10434-007-9554-x

124. Uyl-de-Groot CA, Buijt I, Gloudemans IJM, Ossenkoppele GJ, van den Berg HP, Huijgens PC. Health related quality of life in patients with multiple myeloma undergoing a double transplantation. *Eur J Haematol* 2005;**74**:136–43. http://dx.doi.org/10.1111/j.1600-0609.2004.00346.x

125. Colwell HH, Mathias SD, Turner MP, Lu J, Wright N, Peeters M, *et al.* Psychometric evaluation of the FACT Colorectal Cancer Symptom Index (FCSI-9): reliability, validity, responsiveness, and clinical meaningfulness. *Oncologist* 2010;**15**:308–16. http://dx.doi.org/10.1634/theoncologist.2009-0034

126. Shimoda S, de Camargo B, Horsman J, Furlong W, Lopes LF, Seber A, *et al.* Translation and cultural adaptation of Health Utilities Index (HUI) Mark 2 (HUI2) and Mark 3 (HUI3) with application to survivors of childhood cancer in Brazil. *Qual Life Res* 2005;**14**:1407–12. http://dx.doi.org/10.1007/s11136-004-6127-3

127. Barr RD, Simpson T, Whitton A, Rush B, Furlong W, Feeny DH. Health-related quality of life in survivors of tumours of the central nervous system in childhood – A preference-based approach to measurement in a cross-sectional study. *Eur J Cancer* 1999;**35**:248–55. http://dx.doi.org/10.1016/S0959-8049(98)00366-9

128. Nijdam WM, Levendag PC, Noever I, Schmitz PI, Uyl-de Groot CA. Longitudinal changes in quality of life and costs in long-term survivors of tumors of the oropharynx treated with brachytherapy or surgery. *Brachytherapy* 2008;**7**:343–50. http://dx.doi.org/10.1016/j.brachy.2008.05.001

129. Korfage IJ, Essink-Bot ML, Mols F, Poll-Franse L, Kruitwagen R, van BM, *et al.* Health-related quality of life in cervical cancer survivors: a population-based survey. *Int J Radiat Oncol Biol Phys* 2009;**73**:1501–9. http://dx.doi.org/10.1016/j.ijrobp.2008.06.1905

130. Fu L, Talsma D, Baez F, Bonilla M, Moreno B, Ah-Chu M, *et al.* Measurement of health-related quality of life in survivors of cancer in childhood in Central America: feasibility, reliability, and validity. *J Pediatr Hematol Oncol* 2006;**28**:331–41. http://dx.doi.org/10.1097/00043426-200606000-00003

131. Felder-Puig R, Frey E, Sonnleithner G, Feeny D, Gadner H, Barr R, *et al.* German cross-cultural adaptation of the Health Utilities Index and its application to a sample of childhood cancer survivors. *Eur J Pediatr* 2000;**159**:283–8. http://dx.doi.org/10.1007/s004310050071

132. Pogany L, Barr RD, Shaw A, Speechley KN, Barrera M, Maunsell E. Health status in survivors of cancer in childhood and adolescence. *Qual Life Res* 2006;**15**:143–57. http://dx.doi.org/10.1007/s11136-005-0198-7

133. Barr RD, Chalmers D, De Pauw S, Furlong W, Weitzman S, Feeny D. Health-related quality of life in survivors of Wilms' tumor and advanced neuroblastoma: a cross-sectional study. *J Clin Oncol* 2000;**18**:3280–7.

134. Boman KK, Hoven E, Andclair M, Lannering B, Gustafsson G. Health and persistent functional late effects in adult survivors of childhood CNS tumours: a population-based cohort study. *Eur J Cancer* 2009;**45**:2552–61. http://dx.doi.org/10.1016/j.ejca.2009.06.008

135. Grant J, Cranston A, Horsman J, Furlong W, Barr N, Findlay S, *et al.* Health status and health-related quality of life in adolescent survivors of cancer in childhood. *J Adolesc Health* 2006;**38**:504–10. http://dx.doi.org/10.1016/j.jadohealth.2005.08.002

136. Nixon Speechley K, Maunsell E, Desmeules M, Schanzer D, Landgraf JM, Feeny DH, *et al.* Mutual concurrent validity of the child health questionnaire and the health utilities index: an exploratory analysis using survivors of childhood cancer. *Int J Cancer Suppl* 1999;**12**:95–105.

137. Jansen SJ, Otten W, van de Velde CJ, Nortier JW, Stiggelbout AM. The impact of the perception of treatment choice on satisfaction with treatment, experienced chemotherapy burden and current quality of life. *Br J Cancer* 2004;**91**:56–61. http://dx.doi.org/10.1038/sj.bjc.6601903

138. Lidgren M, Wilking N, Jonsson B, Rehnberg C. Health related quality of life in different states of breast cancer. *Qual Life Res* 2007;**16**:1073–81. http://dx.doi.org/10.1007/s11136-007-9202-8

139. Conner-Spady B, Cumming C, Nabholtz JM, Jacobs P, Stewart D. Responsiveness of the EuroQol in breast cancer patients undergoing high dose chemotherapy. *Qual Life Res* 2001;**10**:479–86.

140. Conner-Spady BL, Cumming C, Nabholtz JM, Jacobs P, Stewart D. A longitudinal prospective study of health-related quality of life in breast cancer patients following high-dose chemotherapy with autologous blood stem cell transplantation. *Bone Marrow Transpl* 2005;**36**:251–9. http://dx.doi.org/10.1038/sj.bmt.1705032

141. Lovrics PJ, Cornacchi SD, Barnabi F, Whelan T, Goldsmith CH. The feasibility and responsiveness of the health utilities index in patients with early-stage breast cancer: a prospective longitudinal study. *Qual Life Res* 2008;**17**:333–45. http://dx.doi.org/10.1007/s11136-007-9305-2

142. Polsky D, Keating NL, Weeks JC, Schulman KA. Patient choice of breast cancer treatment: impact on health state preferences. *Med Care* 2002;**40**:1068–79. http://dx.doi.org/10.1097/00005650-200211000-00008

143. Chang J, Couture FA, Young SD, Lau CY, Lee MK. Weekly administration of epoetin alfa improves cognition and quality of life in patients with breast cancer receiving chemotherapy. *Support Cancer Ther* 2004;**2**:52–8. http://dx.doi.org/10.3816/SCT.2004.n.023

144. Kimman ML, Dirksen CD, Lambin P, Boersma LJ. Responsiveness of the EQ-5D in breast cancer patients in their first year after treatment. *Health Qual Life Outcomes* 2009;**7**:11. http://dx.doi.org/10.1186/1477-7525-7-11

145. Freedman GM, Li T, Anderson PR, Nicolaou N, Konski A. Health states of women after conservative surgery and radiation for breast cancer. *Breast Cancer Res Treat* 2010;**121**:519–26. http://dx.doi.org/10.1007/s10549-009-0552-5

146. Crott R, Briggs A. Mapping the QLQ-C30 quality of life cancer questionnaire to EQ-5D patient preferences. *Eur J Health Econ* 2010;**11**:427–34. http://dx.doi.org/10.1007/s10198-010-0233-7

147. Kontodimopoulos N, Vassilis HA, Dimitris P, Dimitris N. Mapping the cancer-specific EORTC QLQ-C30 to the preference-based EQ-5D, SF-6D, and 15D instruments. *Value Health* 2009;**12**:1151–7. http://dx.doi.org/10.1111/j.1524-4733.2009.00569.x

148. Rogers SN, Miller RD, Ali K, Minhas AB, Williams HF, Lowe D, *et al.* Patients' perceived health status following primary surgery for oral and oropharyngeal cancer. *Int J Oral Maxillofac Surg* 2006;**35**:913–19. http://dx.doi.org/10.1016/j.ijom.2006.07.017

149. Homs MY, Essink-Bot ML, Borsboom GJ, Steyerberg EW, Siersema PD, Dutch SIREC Study Group. Quality of life after palliative treatment for oesophageal carcinoma – a prospective comparison between stent placement and single dose brachytherapy. *Eur J Cancer* 2004;**40**:1862–71. http://dx.doi.org/10.1016/j.ejca.2004.04.021

150. Shenfine J, McNamee P, Steen N, Bond J, Griffin SM, Shenfine J, *et al.* A randomized controlled clinical trial of palliative therapies for patients with inoperable esophageal cancer. *Am J Gastroenterol* 2009;**104**:1674–85. http://dx.doi.org/10.1038/ajg.2009.155

151. O'Gorman P, McMillan DC, McArdle CS. Impact of weight loss, appetite, and the inflammatory response on quality of life in gastrointestinal cancer patients. *Nutr Cancer* 1998;**32**:76–80. http://dx.doi.org/10.1080/01635589809514722

152. Wildi SM, Cox MH, Clark LL, Turner R, Hawes RH, Hoffman BJ, *et al.* Assessment of health state utilities and quality of life in patients with malignant esophageal Dysphagia. *Am J Gastroenterol* 2004;**99**:1044–9. http://dx.doi.org/10.1111/j.1572-0241.2004.30166.x

153. McMillan DC, Wigmore SJ, Fearon KC, O'Gorman P, Wright CE, McArdle CS, *et al.* A prospective randomized study of megestrol acetate and ibuprofen in gastrointestinal cancer patients with weight loss. *Br J Cancer* 1999;**79**:495–500. http://dx.doi.org/10.1038/sj.bjc.6690077

154. Verschuur E, Steyerberg E, Tilanus H, Polinder S, Essink-Bot ML, Tran K, *et al.* Nurse-led follow-up of patients after oesophageal or gastric cardia cancer surgery: a randomised trial. *Br J Cancer* 2009;**100**:70–6. http://dx.doi.org/10.1038/sj.bjc.6604811

155. Albertsen PC, Nease RF Jr, Potosky AL. Assessment of patient preferences among men with prostate cancer. *J Urol* 1998;**159**:158–63. http://dx.doi.org/10.1016/S0022-5347(01)64043-6

156. Shimizu F, Fujino K, Ito YM, Fukuda T, Kawachi Y, Minowada S, *et al.* Factors associated with variation in utility scores among patients with prostate cancer. *Value Health* 2008;**11**:1190–3. http://dx.doi.org/10.1111/j.1524-4733.2008.00336.x

157. Sullivan PW, Mulani PM, Fishman M, Sleep D. Quality of life findings from a multicenter, multinational, observational study of patients with metastatic hormone-refractory prostate cancer. *Qual Life Res* 2007;**16**:571–5. http://dx.doi.org/10.1007/s11136-006-9156-2

158. Sandblom G, Carlsson P, Sennfalt K, Varenhorst E. A population-based study of pain and quality of life during the year before death in men with prostate cancer. *Br J Cancer* 2004;**90**:1163–8. http://dx.doi.org/10.1038/sj.bjc.6601654

159. Weinfurt KP, Castel LD, Li Y, Saad F, Timbie JW, Glendenning A, *et al.* The significance of skeletal-related events for the health-related quality of life of patients with metastatic prostate cancer. *Ann Oncol* 2005;**16**:579–84. http://dx.doi.org/10.1093/annonc/mdi122

160. Krahn M, Bremner KE, Tomlinson G, Ritvo P, Irvine J, Naglie G, *et al.* Responsiveness of disease-specific and generic utility instruments in prostate cancer patients. *Qual Life Res* 2007;**16**:509–22. http://dx.doi.org/10.1007/s11136-006-9132-x

161. Krahn M, Ritvo P, Irvine J, Tomlinson G, Bremner KE, Bezjak A, *et al.* Patient and community preferences for outcomes in prostate cancer: implications for clinical policy. *Med Care* 2003;**41**:153–64. http://dx.doi.org/10.1097/00005650-200301000-00017

162. McCarter H, Furlong W, Whitton AC, Feeny D, DePauw S, Willan AR, *et al.* Health status measurements at diagnosis as predictors of survival among adults with brain tumors. *J Clin Oncol* 2006;**24**:3636–43. http://dx.doi.org/10.1200/JCO.2006.06.0137

163. Le Gales C, Costet N, Gentet JC, Kalifa C, Frappaz D, Edan C, *et al.* Cross-cultural adaptation of a health status classification system in children with cancer. First results of the French adaptation of the Health Utilities Index Marks 2 and 3. *Int J Cancer Suppl* 1999;**12**:112–18.

164. Korfage IJ, van Ballegooijen M, Huveneers H, Essink-Bot ML. Anxiety and borderline PAP smear results. *Eur J Cancer* 2010;**46**:134–41. http://dx.doi.org/10.1016/j.ejca.2009.07.003

165. Whynes DK, TOMBOLA Group. Correspondence between EQ-5D health state classifications and EQ VAS scores. *Health Qual Life Outcomes* 2008;**6**:94. http://dx.doi.org/10.1186/1477-7525-6-94

166. Whynes DK, Woolley C, Philips Z. Trial of management of borderline and other low-grade abnormal smears group. Management of low-grade cervical abnormalities detected at screening: which method do women prefer? *Cytopathology* 2008;**19**:355–62.

167. Maissi E, Marteau T, Hankins M, Moss S, Legood R, Gray A. The psychological impact of human papillomavirus testing in women with borderline or mildly dyskaryotic cervical smear test results: 6-month follow-up. *Br J Cancer* 2005;**92**:990–4. http://dx.doi.org/10.1038/sj.bjc.6602411

168. Cella D, Michaelson MD, Bushmakin AG, Cappelleri JC, Charbonneau C, Kim ST, *et al.* Health-related quality of life in patients with metastatic renal cell carcinoma treated with sunitinib vs interferon-alpha in a phase III trial: final results and geographical analysis. *Br J Cancer* 2010;**102**:658–64. http://dx.doi.org/10.1038/sj.bjc.6605552

169. Cella D, Li JZ, Cappelleri JC, Bushmakin A, Charbonneau C, Kim ST, *et al.* Quality of life in patients with metastatic renal cell carcinoma treated with sunitinib or interferon alfa: results from a phase III randomized trial. *J Clin Oncol* 2008;**26**:3763–9. http://dx.doi.org/10.1200/JCO.2007.13.5145

170. Yang S, de Souza P, Alemao E, Purvis J. Quality of life in patients with advanced renal cell carcinoma treated with temsirolimus or interferon-alpha. *Br J Cancer* 2010;**102**:1456–60. http://dx.doi.org/10.1038/sj.bjc.6605647

171. Castellano D, del Muro XG, Perez-Gracia JL, Gonzalez-Larriba JL, Abrio MV, Ruiz MA, *et al.* Patient-reported outcomes in a phase III, randomized study of sunitinib versus interferon-α as first-line systemic therapy for patients with metastatic renal cell carcinoma in a European population. *Ann Oncol* 2009;**20**:1803–12. http://dx.doi.org/10.1093/annonc/mdp067

172. Mendez Romero A, Wunderink W, van Os RM, Nowak PJ, Helimen BJ, Nuyttens JJ, *et al.* Quality of life after stereotactic body radiation therapy for primary and metastatic liver tumors. *Int J Radiat Oncol* 2008;**70**:1447–52. http://dx.doi.org/10.1016/j.ijrobp.2007.08.058

173. Trippoli S, Vaiani M, Lucioni C, Messori A. Quality of life and utility in patients with non-small cell lung cancer. *Pharmacoeconomics* 2001;**19**:855–63. http://dx.doi.org/10.2165/00019053-200119080-00007

174. Barr R, Petrie C, Furlong W, Rothney M, Feeny D. Health-related quality of life during post-induction chemotherapy in children with acute lymphoblastic leukemia in remission: an influence of corticosteroid therapy. *Int J Oncol* 1997;**11**:333–9.

175. Slovacek L, Slovackova B, Hrstka Z, Mcingova Z, Jebavy L, Horacek JM. Health-related quality of life in multiple myeloma survivors treated with high dose chemotherapy followed by autologous peripheral blood progenitor cell transplantation: a retrospective analysis. *Neoplasma* 2008;**55**:350–5.

176. Hahn EA, Glendenning GA, Sorensen MV, Hudgens SA, Druker BJ, Guilhot F, *et al.* Quality of life in patients with newly diagnosed chronic phase chronic myeloid leukemia on imatinib versus interferon alfa plus low-dose cytarabine: results from the IRIS Study. *J Clin Oncol* 2003;**21**:2138–46. http://dx.doi.org/10.1200/JCO.2003.12.154

177. van Agthoven M, Vellenga E, Fibbe WE, Kingma T, Uyl-de Groot CA. Cost analysis and quality of life assessment comparing patients undergoing autologous peripheral blood stem cell transplantation or autologous bone marrow transplantation for refractory or relapsed non-Hodgkin's lymphoma or Hodgkin's disease: a prospective randomised trial. *Eur J Cancer* 2001;**37**:1781–9.

178. Banks BA, Barrowman NJ, Klaassen R. Health-related quality of life: changes in children undergoing chemotherapy. *J Pediatr Hematol Oncol* 2008;**30**:292–7. http://dx.doi.org/10.1097/MPH.0b013e3181647bda

179. Krabbe PF, Peerenboom L, Langenhoff BS, Ruers TJ. Responsiveness of the generic EQ-5D summary measure compared to the disease-specific EORTC QLQ C-30. *Qual Life Res* 2004;**13**:1247–53. http://dx.doi.org/10.1023/B:QURE.0000037498.00754.b8

180. Langenhoff BS, Krabbe PF, Peerenboom L, Wobbes T, Ruers TJ. Quality of life after surgical treatment of colorectal liver metastases. *Br J Surg* 2006;**93**:1007–14. http://dx.doi.org/10.1002/bjs.5387

181. Slovacek L, Slovackova B, Jebavy L, Mcingova Z. Psychosocial, health and demographic characteristics of quality of life among patients with acute myeloid leukemia and malignant lymphoma who underwent autologous hematopoietic stem cell transplantation. *Sao Paulo Med J* 2007;**125**:359–61. http://dx.doi.org/10.1590/S1516-31802007000600012

182. Slovacek L, Slovackova B, Blazek M, Jebavy L. Quality of life in patients with multiple myeloma and malignant lymphoma undergoing autologous progenitor stem cell transplantation: the effect of selected psychosocial and health aspects on quality of life: a retrospective analysis. *Reports Pract Oncol Radiother* 2007;**12**:101–8. http://dx.doi.org/10.1016/S1507-1367(10)60046-6

183. Mueller-Nordhorn J, Roll S, Boehmig M, Nocon M, Reich A, Braun C, *et al.* Health-related quality of life in patients with pancreatic cancer. *Digestion* 2006;**74**:118–25.

184. Lee SH, Kim DJ, Oh JH, Han HS, Yoo KY, Kim HS. Validation of functional evaluation system in patients with musculoskeletal tumors. *Clin Orthop Relat Res* 2003;**411**:217–26. http://dx.doi.org/10.1097/01.blo.0000069896.31220.33

185. Klaassen RJ, Krahn M, Gaboury I, Hughes J, Anderson R, Grundy P, *et al.* Evaluating the ability to detect change of health-related quality of life in children with Hodgkin disease. *Cancer* 2010;**116**:1608–14. http://dx.doi.org/10.1002/cncr.24883

186. Klaassen RJ, Barr RD, Hughes J, Rogers P, Anderson R, Grundy P, *et al.* Nurses provide valuable proxy assessment of the health-related quality of life of children with Hodgkin disease. *Cancer* 2010;**116**:1602–7. http://dx.doi.org/10.1002/cncr.24888

187. Cox CL, Lensing S, Rai SN, Hinds P, Burghen E, Pui CH, *et al.* Proxy assessment of quality of life in pediatric clinical trials: application of the Health Utilities Index 3. *Qual Life Res* 2005;**14**:1045–56. http://dx.doi.org/10.1007/s11136-004-4714-y

188. Doorduijn J, Buijt I, Holt B, Steijaert M, Uyl-de Groot C, Sonneveld P. Self-reported quality of life in elderly patients with aggressive non-Hodgkin's lymphoma treated with CHOP chemotherapy. *Eur J Haematol* 2005;**75**:116–23. http://dx.doi.org/10.1111/j.1600-0609.2005.00438.x

189. Pickard AS, De Leon MC, Kohlmann T, Cella D, Rosenbloom S. Psychometric comparison of the standard EQ-5D to a 5 level version in cancer patients. *Med Care* 2007;**45**:259–63. http://dx.doi.org/10.1097/01.mlr.0000254515.63841.81

190. Ramsey SD, Andersen M, Etzioni R, Moinpour C, Peacock S, Potosky A, *et al.* Quality of life in survivors of colorectal carcinoma. *Cancer* 2000;**88**:1294–303. http://dx.doi.org/10.1002/(SICI)1097-0142(20000315)88:6<1294::AID-CNCR4>3.3.CO;2-D

191. Park SM, Park MH, Won JH, Lee KO, Choe WS, Heo DS, *et al.* EuroQol and survival prediction in terminal cancer patients: a multicenter prospective study in hospice-palliative care units. *Support Care Cancer* 2006;**14**:329–33. http://dx.doi.org/10.1007/s00520-005-0889-1

192. Witzens-Harig M, Reiz M, Heib C, Benner A, Hensel M, Neben K, *et al.* Quality of life during maintenance therapy with the anti-CD20 antibody rituximab in patients with B cell non-Hodgkin's lymphoma: sesults of a prospective randomized controlled trial. *Ann Hematol* 2009;**88**:51–7.

193. Sternberg CN, Davis ID, Mardiak J, Szczylik C, Lee E, Wagstaff J, *et al.* Pazopanib in locally advanced or metastatic renal cell carcinoma: results of a randomized phase III trial. *J Clin Oncol* 2010;**28**:1061–8.

194. Cella D, Cappelleri JC, Bushmakin A, Charbonneau C, Li JZ, Kim ST, *et al.* Quality of life predicts progression-free survival in patients with metastatic renal cell carcinoma treated with sunitinib versus interferon alfa. *J Oncol Pract* 2009;**5**:66–70. http://dx.doi.org/10.1200/JOP.0922004

195. Lundy JJ. Assessing psychometric equivalence of paper-and-pencil and interactive voice response (ivr) modes of administration for the eq-5d and the qlq-c30. *Dissertation Abstracts International: Section B: The Sciences and Engineering* 6045;**69**.

196. Basch E, Jia X, Heller G, Barz A, Sit L, Fruscione M, *et al.* Adverse symptom event reporting by patients vs. clinicians: relationships with clinical outcomes. *J Natl Cancer Inst* 2009;**101**:1624–32. http://dx.doi.org/10.1093/jnci/djp386

197. Versteegh M, Rowen D, Brazier J, Stolk E. Mapping onto EQ-5D for patients in poor health. *Health Qual Life Outcomes* 2010;**8**:141.

198. Wu EQ, Mulani P, Farrell MH, Sleep D. Mapping FACT-P and EORTC QLQ-C30 to patient health status measured by EQ-5D in metastatic hormone-refractory prostate cancer patients. *Value Health* 2007;**10**:408–14. http://dx.doi.org/10.1111/j.1524-4733.2007.00195.x

199. McKenzie L, van der Pol M. Mapping the EORTC QLQ C-30 onto the EQ-5D instrument: the potential to estimate QALYs without generic preference data. *Value Health* 2009;**12**:167–71. http://dx.doi.org/10.1111/j.1524-4733.2008.00405.x

200. Aaronson NK, Ahmedzai S, Bergman B, Bullinger M, Cull A, Duez NJ, *et al.* The European Organization for Research and Treatment of Cancer QLQ-C30: a Quality-of-Life instrument for use in International Clinical Trials in oncology. *J Natl Cancer Inst* 1993;**85**:365–76. http://dx.doi.org/10.1093/jnci/85.5.365

201. Cella DF, Tulsky DS, Gray G, Sarafian B, Linn E, Bonomi A, *et al.* The Functional Assessment of Cancer Therapy scale: development and validation of the general measure. *J Clin Oncol* 1993;**11**:570–9.

202. San Miguel JF, Schlag R, Khuageva NK, Dimopoulos MA, Shpilberg O, Kropff M, *et al.* Bortezomib plus melphalan and prednisone for initial treatment of multiple myeloma. *N Engl J Med* 2008;**359**:906–17.

203. Greipp PR, Miguel JS, Durie BGM, Crowley JJ, Barlogie B, Bladé J *et al.* International staging system for multiple myeloma. *J Clin Oncol* 2005;**23**:3412–20. http://dx.doi.org/10.1200/JCO.2005.04.242

204. Tabachnick BG, Fidell LS, Osterlind SJ. *Using Multivariate Statistics*. 4th edn. Boston, MA: Allyn and Bacon; 2001.

205. Lipscomb J, Ancukiewicz M, Parmigiani G, Hasselblad V, Samsa G, Matchar DB. Predicting the cost of illness: a comparison of alternative models applied to stroke. *Med Decis Making* 1998;**18**(Suppl. 2):S39–56. http://dx.doi.org/10.1177/0272989X9801800207

206. Huang IC, Frangakis C, Atkinson MJ, Willke RJ, Leite WL, Vogel WB, *et al.* Addressing ceiling effects in health status measures: a comparison of techniques applied to measures for people with HIV disease. *Health Serv Res* 2008;**43**:327–39. http://dx.doi.org/10.1111/j.1475-6773.2007.00745.x

207. Royston P, Sauerbrei W. Multivariable modeling with cubic regression splines: a principled approach. *Stata Journal* 2007;**7**:45–70.

208. Gray AM, Rivero-Arias O, Clarke PM. Estimating the association between SF-12 responses and EQ-5D utility values by response mapping. *Med Decis Making* 2006;**26**:18–29. http://dx.doi.org/10.1177/0272989X05284108

209. Rivero-Arias O, Ouellet M, Gray A, Wolstenholme J, Rothwell PM, Luengo-Fernandez R. Mapping the Modified Rankin Scale (mRS) measurement into the generic EuroQol (EQ-5D) health outcome. *Med Decis Making* 2010;**30**:341–54. http://dx.doi.org/10.1177/0272989X09349961

210. Le QA, Doctor JN. Probabilistic mapping of descriptive health status responses onto health state utilities using Bayesian networks: an empirical analysis converting SF-12 into EQ-5D utility index in a national US sample. *Med Care* 2011;**49**:451–60. http://dx.doi.org/10.1097/MLR.0b013e318207e9a8

211. Hernandez Alava M, Wailoo AJ, Ara R. Tails from the Peak District: adjusted limited dependent variable mixture models of EQ-5D questionnaire health state utility values. *Value Health* 2012;**15**:550–61. http://dx.doi.org/10.1016/j.jval.2011.12.014

212. Hernandez Alava M, Wailoo AJ, Wolfe F, Michaud K. The relationship between EQ-5D, HAQ and pain in patients with rheumatoid arthritis. *Rheumatology* 2013;**52**:944–50. http://dx.doi.org/10.1093/rheumatology/kes400

213. Hernandez Alava M, Wailoo AJ, Wolfe F, Michaud K. *A comparison of direct and indirect methods for the estimation of health utilities from clinical outcomes*. HEDS Discussion Paper No 12.12. 2012. URL: www.sheffield.ac.uk/polopoly_fs/1.283053!/file/12.12.pdf (accessed 11 January 2013).

214. Oken MM, Creech RH, Tormey DC, Horton J, Davis TE, McFadden ET, *et al.* Toxicity and response criteria of the Eastern Cooperative Oncology Group. *Am J Clin Oncol* 1982;**5**:649–55. http://dx.doi.org/10.1097/00000421-198212000-00014

215. Steyerberg EW, Eijkemans MJC, Harrell FE, Habbema JD. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. *Stat Med* 2000;**19**:1059–79. http://dx.doi.org/10.1002/(SICI)1097-0258(20000430)19:8<1059::AID-SIM412>3.3.CO;2-S

216. 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**:977–88. http://dx.doi.org/10.1111/j.1524-4733.2009.00545.x

217. Harrell FE Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996;**15**:361–87. http://dx.doi.org/10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4

218. Powell JL. Least absolute deviations estimation for the censored regression model. *J Econometrics* 1984;**25**:303–25. http://dx.doi.org/10.1016/0304-4076(84)90004-6

219. Wailoo A, Davis S, Tosh J. *The incorporation of health benefits in cost-utility analysis using the EQ-5D: A report by the NICE DSU*. 2010. URL: www.nicedsu.org.uk/PDFs%20of%20reports/DSU%20EQ5D%20final%20report%20-%20submitted.pdf (accessed 11 January 2013).

220. NHS. Health Survey for England. *Health and Social Care information centre, 2006*. URL: www.hscic.gov.uk/pubs/hse06cvdandriskfactors (accessed 6 January 2013).

221. Gudex C. *Time trade-off user manual: props and self-completion methods*. University of York, York: Centre for Health Economics; 1994.

222. Shaw JW, Johnson JA, Coons SJ. US valuation of the EQ-5D health states: development and testing of the D1 valuation model. *Med Care* 2005;**43**:203–20.

223. Szende A, Oppe M, Devlin N. EQ-5D Value Sets: inventory, comparative review and user guide. AA Dordrecht: Springer; 2007.

224. Busschbach J. A comparison of EQ-5D time trade-off values obtained in Germany, the United Kingdom and Spain. In: Brooks R, Rabin R, de Charro F, editors. *The Measurement and Valuation of Health Status using EQ-5D: a European Perspective*. Dordrecht: Kluwer Academic Publishers; 2003. pp. 143–66.

225. Brazier J, Rowen D, Tsuchiya A, Yang Y, Young TA. The impact of adding an extra dimension to a preference-based measure. *Soc Sci Med* 2011;**73**:245–53. http://dx.doi.org/10.1016/j.socscimed.2011.05.026

226. Gudex C. *Are we lacking a dimension of energy in the EuroQol instrument?* Paper presented at the 8th Plenary Meeting of the EuroQol Group. Lund, Sweden: EuroQol Conference Proceedings; 1991.

227. Streiner DL, Norman GR. *Health Measurement Scales: A Practical Guide to their Development and Use.* 4th edn. Oxford: Oxford University Press; 2008.

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

229. Ara R, Brazier J. Deriving an algorithm to convert the eight mean SF-36 dimension scores into a mean EQ-5D preference-based score from published studies (where patient level data are not available). *Value Health* 2008;**11**:1131–43. http://dx.doi.org/10.1111/j.1524-4733.2008.00352.x

230. Rowen D, Brazier J, Roberts J. Mapping SF-36 onto the EQ-5D index: how reliable is the relationship. *Health Qual Life Outcomes* 2009;**7**:27. http://dx.doi.org/10.1186/1477-7525-7-27

231. Buxton MJ, Lacey LA, Feagan BG, Niecko T, Miller DW, Townsend RJ. Mapping from disease-specific measures to utility: an analysis of the relationships between the Inflammatory Bowel Disease Questionnaire and Crohn's Disease Activity Index in Crohn's disease and measures of utility. *Value Health* 2007;**10**:214–20. http://dx.doi.org/10.1111/j.1524-4733.2007.00171.x

Appendix 1 Project protocol

Aims and objectives

The National Institute for Health and Care Excellence (NICE) Technology Appraisals (TA) Methods Guide recommends the use of GPBMs of HRQL (specifically naming the EQ-5D) for its economic evaluations. Generic measures have been criticised for being insensitive or failing to capture important aspects of health. The NICE TA Guide recognises that EQ-5D data may not always be available and/or appropriate and offers some advice for these circumstances. However, it does not identify those areas where EQ-5D is inappropriate nor does it provide criteria to determine when a measure is appropriate for a particular condition or treatment. Information from condition-specific measures can be incorporated into the standard framework of analysis adopted by NICE using techniques such as mapping from condition-specific or clinical measures to the generic measure, modifying the generic HRQL instrument (e.g. creating 'add-on' dimensions) and valuing condition-specific measures directly (i.e. creating preference-based condition-specific measures).

The overall aim of the proposal will be to develop methods for systematically incorporating condition-specific and other non-reference case measures into the NICE decision-making framework.

Research objectives:

1. To examine where commonly used generic HRQL measures are not appropriate for use in calculating QALYs for NICE decision-making by undertaking a review of the published literature on the use of generic measures (EQ-5D, SF-6D and HUI) for different conditions and treatments.
2. To consider the use of condition-specific measures when data from generic instruments are not available by mapping from condition-specific and clinical measures to generic measures. Specifically, to generate functions to map from three key condition-specific or clinical measures to EQ-5D and conduct exploratory analysis around the incorporation of uncertainty in the predicted estimates.
3. To explore new methods for developing new preference-based measures. Specifically, by investigating the use of new 'add-on's to expand the EQ-5D descriptive system for those conditions in which EQ-5D is not appropriate (as determined by part 1).

Description of the project methodology

Stage 1: Review of the appropriateness of generic measures of health-related quality of life

A review of the published literature will identify studies in which generic instruments (EQ-5D, HUI 3 and SF-6D) have been used to obtain health-state utility values (HSUVs) in four key areas: visual impairment, aural impairment, cancer and skin conditions. The review will be conducted using MEDLINE, EMBASE, NHS Economic Evaluation Database (EED) and HTA and OHE Health Economic Evaluations Database (HEED). The records from these databases will be supplemented by a review of a database held by the EuroQol Group containing more than 1500 references of studies relating to the use of the EQ-5D. The papers identified from the search will be sifted to identify papers that report data on the use of generic instruments or systematic reviews related to the use of generics. Systematic reviews will be assessed for quality and relevance to the study. Where good quality systematic reviews exist, these will be used to guide the subsequent review for that particular condition/treatment. Studies will then be grouped into condition groups based on ICD-9 codes. Papers that report the use of EQ-5D from another study will be reviewed, and the original articles considered for inclusion. Papers reporting primary data collection will be included in the review if they report data on the use of EQ-5D, SF-6DF or HUI 3 in sufficient detail to allow an assessment of their validity. Therefore empirical papers must include data on the resulting health classification systems or utility values and include data on other measures of health outcomes (e.g. visual analogue scale data, disease specific-measures and/or clinical measures of severity). Papers reporting qualitative studies on the use of the generic instruments will also be included.

The assessment of the validity of a preference-based measure of health such as the EQ-5D is fraught with conceptual and empirical problems owing to the lack of a gold standard.[228] A common mistake is to assume that because a condition specific measure finds a difference, then a generic measure should reflect that difference, when the general population may not regard the difference as sufficiently important in the valuation task. The approach adopted here follows that suggested in Brazier and Deverill,[33] that distinguishes the validity of the descriptive system from that of the preference-based index. It will examine the descriptive validity of the EQ-5D as a descriptive system in terms of its content, face and construct validity. Contact and face validity will be examined using evidence from qualitative studies. Construct validity of the descriptive system will be assessed in terms of whether the distribution of responses by dimension level agrees with other measures of those dimensions and other relevant indicators. The empirical validity of the index will be based on the convergence with other measures of stated preferences and hypothetical preferences (other indicators of likely preferences). The former will use GPBMs and directly elicited preferences (e.g. time trade-off or standard gamble). Hypothetical preferences will be assessed by looking at convergence with other measures and clinical indicators, but care will be taken in the interpretation to ensure that these are likely to reflect genuine differences in preferences. These tests will be applied to cross sectional data to examine validity and the longitudinal data to examine the responsiveness of the measure.

Stage 2: Mapping from condition-specific and clinical measures to EQ-5D

The results of the literature review and existing research conducted in ScHARR[11,229,230] will form the basis for this section of the current project. This comprehensive literature review identified some methodological issues in the use mapping to predict health-related utility values. Most published papers have focused upon mapping between alternative generic instruments (e.g. SF-12 to EQ-5D[209]) or from existing condition-specific instruments to generic instruments (e.g. Asthma quality of life questionnaire to EQ-5D[16]). Relatively few published studies have focused on mapping from clinical measures of disease activity or severity.[12,231] However, these severity indices may form the basis of health outcomes estimation included in submissions to NICE (e.g. Psoriasis Area and Severity Index and Crohn's Disease Activity Index). A key issue is arising from the literature review is that the uncertainty around the resulting predictions is usually ignored when the mapping algorithms are applied, and thus the estimates do not reflect that the health-state utility values are estimated and not observed.

At least three mapping functions will be developed during this stage of the project, including at least one using clinical scales rather than patient-reported outcomes. This stage will also include exploratory work around some methodological issues, specifically the incorporation of uncertainty into the predicted estimates and an assessment of whether methods differ for mapping from clinical outcome measures. Data sets held within ScHARR will be considered for use to generate the mapping functions. In addition members of the EuroQol Group will be approached for access to data sets that include EQ-5D data and responses to a condition-specific measure and to clinical measures of severity. (One of the terms and conditions for the use of the EQ-5D is that data should be made available to other EuroQol Group members if requested.) Data sets will include those that include data from the EQ-5D and other condition-specific measures and/or clinical measures of severity.

The mapping functions will be made publicly available at no charge via the ScHARR website, where the predictive ability is considered adequate for use by others, along with guidance on incorporating the uncertainty in the estimates.

Phase 1: A potential list of mapping functions will be drawn up based on the availability of datasets of sufficient size to undertake the mapping exercise. This 'long list' will be reduced to a recommended list of possible mapping functions based on how widespread the use of a condition-specific or clinical measure is and whether good quality mapping functions have already been published for that measure. We will consult with representatives from NICE before making the final decision about which condition-specific and clinical measures to create mapping functions from, however at least one will be derived from a condition-specific measure and at least one will be developed from clinical measure/s of severity. It is anticipated that at least 3 mapping functions will be developed.

Phase 2: The mapping function will be estimated. The datasets will be randomly split into two subsets in order to provide a subset for model estimation and a subset for assessing the predictive ability of the models. Alternative models will be considered to estimate the mapping function. Simple OLS models will be explored as these have been most frequently used in published studies.[11] However simple OLS models ignore the bounded nature of health-state utility data (i.e. the maximum value is 1) will result in biased and inconsistent estimates. Tobit and censored least absolute deviations (CLAD) models will also be explored as appropriate alternatives.[230] Both the tobit and the CLAD models use the same structure to generate both the continuous and the censored observations. Rejection of this assumption would render the asymptotic properties of the CLAD model invalid. Therefore, we will first explore the validity of this assumption by estimating different two-part models. In addition, there is also an issue of efficiency loss of the CLAD estimator when compared to maximum likelihood if the assumed distribution of the errors is correct. Consideration will also be given to Generalised Linear Models with RE, Adjusted Least Square Regression Model (ALS), and Weighted Least Squares models. Most published mapping models predict the single index utility value from the generic instrument. However, there are advantages to predicting the responses to the health state classification system (e.g. the descriptor 12112 on the EQ-5D) as this better reflects the data that would have collected had the relevant generic instrument been included in the study and enables alternative sets of utility data or 'tariffs' to be applied to the health state descriptions. Models that map to the single index utility value and those which map to descriptive classification will be considered.

Phase 3: The goodness of fit and predictive ability of the alternative models will be assessed in order to recommend a preferred model for each condition-specific or clinical measure.

The goodness of fit of the models obtained will be assessed using standard statistics (variance explained, range, mean and SE). The predictive abilities will be compared by charting the observed and predicted preference-based scores together with the residuals. The mean error, mean absolute error, RMSE and the proportion of predicted values within the minimum clinically important difference for the preference-based index will also be reported.

Phase 4: The literature describing results of mapping exercises rarely report the full range of statistics required to independently assess the functions. Analysts who wish to use the results of the mapping functions are not provided with the data required to estimate uncertainties in the predicted values. In addition, there are no clear recommendations of methods to incorporate the uncertainty arising from the predicted values into their application when estimating QALYs. Exploratory work into the appropriate methods for incorporating uncertainty in the predicted values into practical analyses will be conducted. This will include using probabilistic simulation; however this requires the underlying distribution of the values to be appropriately specified. For example, if the tobit model is found to be the most appropriate model specification it will be necessary to ensure that the distribution takes into account the censoring of the dependent variable is properly taken into account when incorporating the uncertainty into practical analyses.

Stage 3: Developing new measures by extending existing generic measures: 'add-on's to the EQ-5D

The review conducted in Stage 1 will identify those conditions in which generic instruments, and specifically the EQ-5D, are not appropriate. In these cases it is not meaningful to map from a condition-specific measure to the generic because the generic measure does not adequately capture the important aspects of health for that condition. Previous work has been conducted on taking existing condition-specific measures and deriving preference-based measures[15,16] and a further study is currently investigating some of the methods around this (COSMeQ study).[14] Problems associated with this approach include a loss of information when condensing the original measure into a new measure for which preferences can be obtained, the introduction of labelling effects and the failure to reflect side effects and co-morbidities. An alternative approach is 'add-on' additional dimensions to existing generic measures.

The focus of this part of the study will be to the EQ-5D due to its prominence in the NICE Methods Guide. A new 5 level version of the EQ-5D has recently become available however empirical data, including data on value set of corresponding utility estimates, aren't currently available to allow its routine use in decision-making. Therefore evidence from this review will relate to the 3 level version of the EQ-5D. The focus for this element of the project will focus on adding additional dimensions to the 3 level version of the EQ-5D. The proposed approach will be similar to that adopted in a recent study to investigate the addition of a sleep dimension to the EQ-5D.[19]

Phase 1: Approximately 3 conditions will be selected where the EQ-5D has been shown to be insufficient for capturing changes on HRQL from those identified in the review of the literature described in stage 1 of the proposal. Six EQ-5D health states will be selected covering a range of severity (2 mild, 2 moderate and 2 severe), plus full health (11111) and the worst possible health state (33333). The same states with the addition of an extra 'add-on' dimension relating to the condition of interest will be described. The description of the add-on will be based on the results of review described in section 1. The description of levels will follow the approach used for the EQ-5D (no problems, some problems and extreme problems). There will be four variants of the questionnaire to avoid contamination between them (original EQ-5D, plus 3 versions with add-ons).

The effect of including the additional dimension will be assessed by comparing mean valuations with and without the additional dimension using an independent *t*-test. Assuming a power of 0.8, significance level of 0.05, SD of 0.3 and a difference of 0.1, then this requires a sample of 73 interviews in each group for each instrument. In order to obtain 75 valuations per variant of the questionnaire, this will require a sample of 300 people (4 × 75 people). A sample of 300 members of the general public in South Yorkshire will be selected randomly from the electoral register. Three groups will each be allocated to one of the add-on instruments and one will be allocated to the original EQ-5D questionnaire. The methods of valuation will be compatible with the original EQ-5D valuation study[29] and will use the time trade-off method using full health as the top anchor and using a time board for visual props. The recruitment of patients and conduct of the interviews will be commissioned from Sheffield Hallam University who have extensive experience of conducting this kind of study.[14]

Phase 2: Based on the results of phase 1, the new 'add-on' instrument that is found to add the most additional information will be selected for further study in order to develop a valuation system and to generate a set of methods that can be used by others when considering expanding the descriptive system of a generic instrument. Based on an orthogonal design for an instrument with six dimensions, values for 18 health states are required for an additive model. For five dimensions, 16 states are required. Therefore it will be necessary to obtain valuations for 34 states. A sample of 300 members of the general public in South Yorkshire will be selected randomly from the electoral register and recruited to the study (to get 75 valuations per state; 4 groups each valuing 8 or 9 states). They will be split into four groups: two groups will each value 9 health states from the new add-on instrument and two groups will each value 8 health states from the standard EQ-5D. The recruitment of people and methods of valuation will be the same as described in Phase 1.

A model will be developed to estimate a tariff of values for all health states. The impact of the inclusion of the add-on will be assessed in two ways: 1) examining the significance of the co-efficient of the extra dimension and 2) examining the impact on the other dimensions from having a new dimension added to the descriptive system. The model developed for the add-on instrument and the descriptive system will be made publicly available at no charge.

Appendix 2 Search strategies for literature review

MEDLINE Search strategy for vision review (first search)

1. (vision disorder$ or micropsia$ or metamorphopsia$ or hemeralopia$ or day blindness or macropsia$).mp.
2. [mp=title, original title, abstract, name of substance word, subject heading word, unique identifier] (18729)
3. vision disorders/ or visually impaired persons/ (18807)
4. 1 or 2 (19258)
5. (euroqol or euro qol or eq5d or eq 5d or eq-5d or (euro adj qol) or (eur adj qual) or (eq adj 5d)).mp. [mp=title, original title, abstract, name of substance word, subject heading word, unique identifier] (1868)
6. (hui3 or hui 3 or health utilities index mark 3 or health utilities mark three or hui III or huiIII).mp. [mp=title, original title, abstract, name of substance word, subject heading word, unique identifier] (207)
7. (sf6D or sf 6D or short form 6D or shortform 6D or sf six D or sfsixD or shortform six D or short form sixD or sf-6d or 6d or 6-d or 6 dimension).mp. [mp=title, original title, abstract, name of substance word, subject heading word, unique identifier] (4204)
8. 4 or 5 or 6 (6114)
9. 3 and 7 (14)
10. from 8 keep 1-14 (14)

Search strategy including specific visual disorders (second search)

amblyopia

acuity

age related macular degeneration

anisometropia

astigmatism

blurred vision

cataracts

conjunctivitis

corneal opacities

cytomegalovirus

cytomegalovirus retinitis

day blindness

diabetic retinopathy

diplopia

double vision

dry eye

dystrophy

edema

far sightedness

glaucoma

hemeralopia*

hemianopia

hypermetropia

lazy eye

macropsia*

macular degeneration

metamorphopsia*

micropsia*

near sightedness

night blindness

nystagmus

ocular hypertension

Oedema

onchocerciasis

phaco

phacoemulisification

quandrantanopia

retinitis pigmentosa

retinopathy

river blindness

strabismus

trachoma

vision

vision disorder*

visual*

visually impaired persons

disorder adj (eyelid or lacrimal system or orbit or conjunctiva or sclera or cornea or iris or ciliary body or lens or choriod or retina or vitreous body or globe or optic nerve or visual pathways or ocular muscles or binocular movement or accomadation or refraction or eye or adnexa)

(euroqol or euro qol or eq5d or eq 5d or eq-5d or (euro adj qol) or (eur adj qual) or (eq adj 5d)).mp.

(hui3 or hui 3 or health utilities index mark 3 or health utilities mark three or hui III or huiIII).mp.

(sf6D or sf 6D or short form 6D or shortform 6D or sf six D or sfsixD or shortform six D or short form sixD or sf-6d or 6d or 6-d or 6 dimension).mp.

MEDLINE search strategy used for hearing review

1. (euroqol or euro qol or eq5d or eq 5d or eq-5d or (euro adj qol)Or eur adj qual) or (eq adj 5d).mp.
2. (hui3 or hui 3 or health utilities index mark 3 or health utilities mark three or hui III or huiIII).mp.
3. (sf6D or sf 6D or short form 6D or shortform 6D or sf six D or sfsixD or shortform six D or short form sixD or sf-6d or 6d or 6-d or 6 dimension).mp.
4. (hearing disorder or dysacusis or paracousis or paracusis or Distorted hearing).mp.
5. (hearing loss or hearing complaints or hearing aids or cochlearimplants).mp.
   [mp=title, original title, abstract, name of substance word, subject heading word, unique identifier]
6. hearing disorders/
7. 1 or 2 or 3
8. 4 or 5 or 6
9. 7 and 8

MEDLINE search strategy used for skin review

1. (euroqol or euro qol or eq5d or eq 5d or eq-5d or (euro adj qol) or (eur adj qual) or (eq adj 5d)).mp. (2151)
2. (hui3 or hui 3 or health utilities index mark 3 or health utilities mark three or hui III or huiIII).mp. (231)
3. (sf6D or sf 6D or short form 6D or shortform 6D or sf six D or sfsixD or shortform six D or short form sixD or sf-6d or 6d or 6-d or 6 dimension).mp. (4538)
4. 1 or 2 or 3 (6722)
5. Staphylococcal scalded skin syndrome.mp. or Staphylococcal Scalded Skin Syndrome/ (414)
6. Impetigo.mp. or Impetigo/ (1457)
7. boil.mp. or Furunculosis/ (1278)
8. furunculosis.mp. (1165)
9. Cutaneous abscess.mp. (66)
10. Cellulitis/ or Cellulitis.mp. (8369)
11. Acute lymphadenitis.mp. (30)
12. Pilonidal cyst.mp. (116)

13. Pyoderma/ or Pyoderma.mp. (3928)
14. Erythrasma.mp. or Erythrasma/ (175)
15. Pemphigus/ or Pemphigus.mp. (7253)
16. Pemphigoid.mp. or Pemphigoid, Bullous/ (4942)
17. Dermatosis.mp. or Skin Diseases/ (46511)
18. Acantholysis/ or Acantholytic disorder.mp. (660)
19. Dermatitis/ or Dermatitis.mp. (59308)
20. Eczema/ or eczema.mp. (13886)
21. prurigo.mp. or Prurigo/ (1207)
22. Pruritus.mp. or Pruritus/ (13152)
23. Lichen simplex chronicus.mp. or Neurodermatitis/ (1396)
24. Dyshidrosis.mp. (104)
25. Erythema intertrigo.mp. (2)
26. Pityriasis alba.mp. (79)
27. Papulosquamous.mp. (861)
28. Psoriasis.mp. or Psoriasis/ (27853)
29. Acrodermatitis/ or Acrodermatitis continua.mp. (1813)
30. Pustulosis.mp. (1302)
31. Urticaria/ or Urticaria.mp. (12733)
32. erythema.mp. or Erythema/ (25199)
33. Sunburn.mp. or Sunburn/ (2693)
34. Dermatitis, Phototoxic/ (528)
35. Dermatitis, Photoallergic/ or Photoallergic.mp. (700)
36. Solar urticaria.mp. (228)
37. Actinic keratosis.mp. or Keratosis, Actinic/ (944)
38. Actinic reticuloid.mp. (139)
39. Cutis rhomboidalis nuchae.mp. (12)
40. Poikiloderma of Civatte.mp. (36)
41. Cutis laxa senilis.mp. (0)
42. Actinic granuloma.mp. (49)
43. Acne.mp. (11465)
44. Rosacea.mp. or Rosacea/ (2084)
45. Vitiligo.mp. or Vitiligo/ (4053)
46. 5 or 6 or 7 or 8 or 9 or 10 or 11 or 12 or 13 or 14 or 15 or 16 or 17 or 18 or 19 or 20 or 21 or 22 or 23 or 24 or 25 or 26 or 27 or 28 or 29 or 30 or 31 or 32 or 33 or 34 or 35 or 36 or 37 or 38 or 39 or 40 or 41 or 42 or 43 or 44 or 45 (212215)
47. 4 and 46 (60)

MEDLINE search strategy used for cancer review

1. (euroqol or euro qol or eq5d or eq 5d or eq-5d or (euro adj qol) or (eur adj qual) or (eq adj 5d)).mp.
2. (hui3 or hui 3 or health utilities index mark 3 or health utilities mark three or hui III or huiIII).mp.
3. (sf6D or sf 6D or short form 6D or shortform 6D or sf six D or sfsixD or shortform six D or short form sixD or sf-6d or 6d or 6-d or 6 dimension).mp.
4. 1 or 2 or 3

Adenocarcinoma

Astrocytoma

Blastoma

Cancer

carcinoma

Cytoma

Cytosis

Ependymoma

Glioblastoma

heavy chain disease

hepatoma

hogkin's disease

Kahler's disease

Leiomyoma

Leukaemia

lymphoma

Lymphosarcoma

Malignant glioma

Malignant neoplasm

melanoma

mesothelioma

Multiple myeloma

myeloma

Myelomatosis

Myelosis

neoplasms

Neuroblastoma

Non-Hodgkin lymphoma

Non-Hodgkin's Lymphoma

Oligodendroglioma

Osteosarcoma

Retinoblastoma

sarcoma

Thymoma

Tumour/Tumor

Waldenström's macroglobulinaemia

5. 4 and 5 (270)
6. from 6 keep 1-270 (270)

Appendix 3 Summary of validity for utility measures: visual disorders

Study reference grouped by condition (author, year)	Utility measures	Methods	Results
Glaucoma			
Aspinall *et al.*, 2008[44]	EQ-5D	Known groups (severity) Convergence	EQ-index stratified by mild, moderate and severe visual field loss. EQ-index, mobility, self-care and anxiety statistically significantly correlated with VA. Mobility and self-care correlated with severity of visual field loss
Kobelt *et al.*, 2006[45]	EQ-5D	Known groups (severity)	EQ-5D utility decreased with increased severity, but difference between groups only statistically significant for severe disease after controlling for co-morbidity
Mittmann *et al.*, 2001[34]	HUI3	Known groups (case–control)	Mean HUI3 values (SD): glaucoma patients 0.924 (0.086); no condition patients 0.953 (0.068)
Montemayor *et al.*, 2001[46]	EQ-5D	Convergence	EQ-5D correlated with age (health status only) and VFA score. Not correlated with diagnosis, VA, mean deviation in the better or worse eye, corrected pattern standard deviation in the better or worse eye. VFA was the best predictor of EQ-5D
Thygesen *et al.*, 2008[54]	EQ-5D	Convergence Known groups (severity)	Better VA is correlated with higher EQ-5D ($p = 0.005$). EQ-5D was consistent with the severity groups defined by Snellen scores
AMD			
Cruess *et al.*, 2007[47]	EQ-5D	Known groups (case–control) Convergence	EQ-5D not significantly lower in subjects compared with control (14% relative difference, $p = 0.064$). Moves in the right direction. No association between EQ-5D and VA stratification found
Espallargues *et al.*, 2005[22]	EQ-5D, SF-6D and HUI3	Convergence Known groups (severity)	All preference-based measures were correlated and significant to 1% level with VF-14. EQ-5D was not correlated to a significant level with CS or VA. VAS was correlated with 5% significant level with CS and VA. SF-6D was correlated with CS (1% level) and VA (5% level). HUI3 and TTO were correlated with both CS and VA to 1% level. VA and CS were consistent with HUI3, SF-6D, TTO and VAS but not with the EQ-5D
Lotery *et al.*, 2007[48]	EQ-5D	Known groups (severity) Convergence	EQ-5D and VFQ-25 differentiated between groups (statistically significant). No apparent relationship was found between EQ-5D and severity of vision loss. This was found for the NEI-VFQ-25 (no *p*-value reported)
Payakachat *et al.*, 2009[49]	EQ-5D	Known groups (severity)	Subjects reported full health in EQ-5D but had visual problems, as elicited by the NEI-VFQ-25
Ruiz-Moreno *et al.*, 2008[56]	EQ-5D	Known groups (case–control)	Adjusted mean scores 0.68 vs. 0.79 $p < 0.05$ for neovascular-AMD vs. control
Soubrane *et al.*, 2007[43]	EQ-5D	Known groups (case–control and severity)	Adjusted mean scores of EQ-5D 0.65 vs. 0.75 $p < 0.001$ for neovascular-AMD vs. control. No significant difference across VA levels of neovascular-AMD (and does not follow degree of severity)
Kim *et al.*, 2010[55]	EQ-5D	Known groups (severity)	Significant differences were found in EQ-5D scores for people with unilateral and bilateral AMD

Study reference grouped by condition (author, year)	Utility measures	Methods	Results
Cataracts			
Asakawa *et al.*, 2008[36]	HUI3	Known groups (case–control, gender)	Adjusted mean differences in single-attribute vision utility scores for cataracts were negative, quantitatively important (difference > 0.05) and statistically significant
Datta *et al.*, 2008[53]	EQ-5D	Convergence	No visual variables were significantly associated with EQ-5D. VF-14 was strongly associated with acuity, stereopsis and contrast sensitivity. Acuity was less important than either stereopsis or contrast sensitivity for EQ-5D, which may suggest that acuity is required for function tasks, but stereopsis and contrast sensitivity were more important determinant of generic QoL
Polack *et al.*, 2007[37]	EQ-5D	Known groups (case–control)	Cases were significantly more likely to report problems with mobility, self-care, usual activities and anxiety than controls
		Convergence	No significant association between VA and EQ-5D across all dimensions, except for self-care which has a borderline ($p = 0.05$) association
Polack *et al.*, 2008[38]	EQ-5D	Known groups (case–control) Convergence	Significant difference ($p < 0.001$) across all EQ-5D dimensions between cases and controls. Poorer VA was associated with higher odds or reporting any problem with mobility, self-care, usual activities and pain. There was no significant association for depression
Polack *et al.*, 2010[39]	EQ-5D	Known groups (case–control) Convergence	Significant difference between cases and controls using VF20 and self-rated health scale. Cases were significantly more likely to report problems with all five EQ-5D domains compared with controls after adjustment for age, gender and socioeconomic status. Inconsistent association between EQ-5D and VA level. Borderline trend with VA shown with self-care ($p = 0.05$), driven by the higher prevalence of reported problems among cases with perception of light compared with those with moderate visual impairment. The lack of association with the remaining domains may reflect the fact that relatively few cases (< 25%) reported no problem, resulting in a lack of variation in the data
Diabetic retinopathy			
Lloyd *et al.*, 2008[42]	EQ-5D and HUI3	Known groups (severity) Convergence	EQ-5D index, EQ-VAS and HUI3 all show some inconsistency when compared with degree of severity. Pattern on VFQ-25 consistent. Between each level of VA, not every difference in utility was significant or consistent. Results show a significant trend with EQ-5D and HUI3 worsening as VA worsens. A regression was undertaken and VFQ-25 and LogMAR were identified as independent significant predictors of utility. The data from the EQ-5D, HUI and VFQ-25 suggest that relatively mild vision loss (6/12 to 6/18) can be associated with very substantial declines in utility, with lower scores than people with worse vision
Smith *et al.*, 2008[50]		Convergence (through regressions) Known groups (severity)	No clear pattern from mean values. OLS model used to estimate the impact on utility of a doubling of the visual angle. Utility values dropped by approximately seven points for each doubling (assuming linear relationship between acuity and utility). Doubling visual angle results in utility loss of about 0.03. A non-parametric ordinal logistic model was fitted and this estimated that anyone who suffered any degree of visual impairment were more likely to report non-perfect utility values (OR 1.44, 95% confidence interval 1.08 to 1.91)

Study reference grouped by condition (author, year)	Utility measures	Methods	Results
Conjunctivitis			
Pitt *et al.*, 2004[60]	EQ-5D	Known groups (case–control)	Inconsistent results comparing SAC to controls. Only the pain domain and the EQ-5D were significantly worse in the SAC group compared with the control. In some cases, the remaining domains were worse in the control (but non-significant). RQLQ was statistically significant across all domains. VFQ-25 was statistically significant across the mean vision score and the general health score
Rajagopalan *et al.*, 2005[51]	EQ-5D	Known groups (severity)	EQ-5D showed significant differences in scale scores across the varying severity levels (EQ-5D, $p < 0.05$, and VAS, $p < 0.0001$). Significant differences were seen across all IDEEL scales except treatment satisfaction. EQ-5D and IDEEL were consistent in their ranking of severity. Strength of difference analysis was provided and the IDEEL outperformed EQ-5D and SF-36 across all severity levels. Mean (SD) EQ-5D scores: control 0.87 (0.03), non-SS KCS 0.82 (0.02) and SS 0.74 (0.03). Mean (SD) EQ-5D VAS score: 88.93 (2.06), non-SS KCS 82.45 (1.19) and SS 66.94 (2.43)
Smith *et al.*, 2005[61]	EQ-5D	Known groups (case–control)	EQ-VAS and all EQ-5D dimensions, except mobility, are statistically significant ($p < 0.02$) between SAC and control groups. Interestingly, VFQ-25 showed significantly lower scores in all domains in the SAC group, except for the general health domain, which returned a lower (non-significant) value for the control group
Other visual disorders			
Boulton *et al.*, 2006[40]	HUI3	Known groups (severity)	Statistically different (unknown to what level) mean HUI3 scores between groups
Clark *et al.*, 2008[62]	EQ-5D	Known groups (case–control)	Significant differences between cases and controls using NEI VFQ-25, but not with EQ-5D or TTO. Only the mobility domain had a significant difference. Patients had a significant difference using the VFQ-25; however, no difference was significant when stratified by visual impairment. Postoperation VA was statistically significantly different
Kempen *et al.*, 2003[63]	EQ-5D	Known groups (severity)	Does not distinguish between groups (non-significant) and direction of trend is counter-intuitive. VAS distinguished newly-diagnosed group. No statistically significant difference in EQ-5D and borderline between VAS
Langelaan *et al.*, 2007[41]	EQ-5D	Known groups (severity)	None were statistically significant at the 5% level. VA saw an appropriate movement in EQ-5D; however, VF moved in the wrong direction
Quinn *et al.*, 2004[64]	HUI3	Known groups (severity)	HUI3 mean (SD) scores: All 0.59 (0.39). Blind or low vision in better eye 0.25 (0.37). Sighted in better eye 0.78 (0.25). No-ROP subjects 0.90 (0.16). Statistical significance of VA not given but appears to be statistically significant and appropriate. HUI3 showed a significantly lower score ($p < 0.001$) for the blind group compared with the sighted group and the non-ROP group compared with the sighted group ($p < 0.001$)
van Nispen *et al.*, 2009[52]	EQ-5D	Convergence (through regression)	LogMAR VA is a significant risk factor for lower QoL

IDEEL, impact of dry eyes on everyday life questionnaire; KCS, Keratoconjunctivitis sicca; LogMAR, logarithm of the minimum angle of resolution; NEI-VFQ-25, National Eye Institute Visual Functioning Questionnaire – 25; ROP, retinopathy of prematurity; RQLQ, rhinoconjunctivitis QoL questionnaire; SS, Sjögren's syndrome.

Appendix 4 Summary of responsiveness for utility measures: visual disorders

Study (author, year)	Utility measures	Method	Results
Cataracts			
Conner-Spady *et al.*, 2005[58]	EQ-5D	Pre–post treatment comparison of VFA, EQ-VAS and EQ-5D	EQ-VAS and EQ-5D show a non-significant improvement. Mean difference (SD): EQ-VAS 1.93 (13.27) and EQ-5D 0.03 (0.17). Per cent better/worse: EQ-VAS 49/33, EQ-5D 38/23
Black *et al.*, 2009[57]	EQ-5D	Pre–post treatment comparison of VF-14 and EQ-5D	Statistically significant improvement in both EQ-5D and VF-14 ($p = 0.003$). Mean (SD) VF-14 scores: preoperation 82.7 (17.3), postoperation 93.7 (13.2); mean EQ-5D scores: preoperation 0.82, postoperation 0.79
AMD			
Kim *et al.*, 2010[55]	EQ-5D	Pre–post treatment comparison of VF-4D and EQ-5D	Statistically significant improvement in both EQ-5D and VF-14 ($p < 0.001$). Mean VF-4D scores: before treatment 0.411, after treatment 0.353. Mean EQ-5D scores: before treatment 0.729, after treatment 0.793

Appendix 5 Summary of validity for utility measures: hearing impairments

Study (author, year)	Instrument	Assessment	Methods	Summary of results
Barton *et al.*, 2005[21]	HUI3/EQ-5D/SF-6D	Convergence	Correlations between measures	Moderate to strong correlations were found between HUI3, EQ-5D and SF-6D
Barton *et al.*, 2006[65]	HUI3	Known groups (severity) Convergence	HUI3 scores and severity groups defined by AHL level	HUI3 mean scores were different between moderate, severe, profound1, profound2 and implanted groups (significance not reported). Cochlear implant (grouped by age at implantation and duration of use), AHL and gender were significant predictors of HUI3 ($p < 0.01$)
Bichey *et al.*, 2002[68]	HUI3	Known groups (severity)	HUI3 scores and PTA (presented by cochlear implant and hearing aid group)	HUI3 mean scores: 0.82 (cochlear implant) vs. 0.62 (hearing aid), consistent with PTA. No statistical test reported
Damen *et al.*, 2007[69]	HUI3	Convergence	Spearman's rank correlations between mean score of different measures at the follow-up	Correlation coefficients: 0.33 (HUI3 and AN test, $p < 0.05$), 0.39 (HUI3 and NVA test, $p <0.05$), 0.48 (NCIQ and AN test, $p < 0.05$), 0.32 (NCIQ and NVA test, $p < 0.05$)
Lovett *et al.*, 2010[66]	HUI3	Known groups (severity)	HUI3 index scores and SSQ, VAS scores presented by unilateral and bilateral implantation groups	A significant difference ($p < 0.05$) was detected in favour of the bilateral group using the SSQ; no significant ($p = 0.2$) differences detected (HUI3 and VAS)
Palmer *et al.*, 1999[75]	HUI3	Known groups (severity)	HUI3 index scores presented by Cochlear implant and non-Cochlear implant groups at enrolment, 6 and 12 months after cochlear implant	Difference between cochlear implant and non-cochlear implant groups by HUI3: not significant (baseline), significant ($p < 0.1$) difference (0.76 for cochlear implant and 0.58 for non-cochlear implant) at both 6 and 12 months after intervention
Smith-Olinde *et al.*, 2008[67]	HUI3	Known groups (severity)	HUI3 utility index presented by four groups defined by degree of hearing loss	Both HUI3 and QWB scores declined with the degree of hearing loss, the decline was greater for HUI3 than QWB. No statistical significance were presented
Grutters *et al.*, 2007[23]	EQ-5D (UK and Dutch tariff), HUI3	Known groups (age, gender and severity) Convergence	Utility scores compared between age, gender (EQ-5D) and clinically distinctive groups (HUI3) Agreements between utility scores by Kendall's tau correlation and ICC	Significant differences detected: age and gender (by EQ-5D) and clinically distinctive groups (by HUI3). Kendall's Tau correlations: 0.36 to 0.41 (between EQ-5D with UK or Dutch tariff and HUI2, HUI3) ICC: 0.44 to 0.51 (between utility measures)
Sach and Barton, 2007[76]	EQ-5D	Known groups (through regressions)	Multiple linear regression were estimated between the child's EQ-5D scores and CAP, as well as other variables	Statistically significant coefficients ($p < 0.05$) for children with or without additional disabilities, gender, a more severe deaf condition (measured by CAP); non statistical significant coefficients ($p > 0.05$) for children with a mild deaf condition (in the top three levels of the CAP) and other socioeconomic factors

AHL, average hearing level; AN test, Antwerp-Nijmegen hearing test battery; CAP, categories of auditory perception; ICC, intraclass correlation; NCIQ, the Nijmegen Cochlear Implant questionnaire; NVA test, Dutch Audiological Society open speech recognition test; SSQ, speech, spatial and qualities of hearing scale for parents.

Appendix 6 Summary of responsiveness for utility measures: hearing impairments

Study (author, year)	Instruments	Assessment methods	Results summary
Barton *et al.*, 2005[21]	EQ-5D, SF-6D and HUI3	Correlation between change scores of different measures	Statistically significant difference ($p < 0.001$) between score changes of the HUI compared with SF-6D or EQ-5D, but not between the EQ-5D and SF-6D. Pearson correlation coefficients between score changes were small (around or below 0.2)
Grutters *et al.*, 2007[23]	EQ-5D (UK and Dutch tariff), HUI2 and HUI3	Mean change of scores after hearing aid fitting, ES and SRM	Mean change score of HUI2 and HUI3 were significantly different from E-5D (UK or Dutch tariff); ES and SRM of EQ-5D were small (0.02–0.05), ES and SRM of HUI2 and HUI3 were large (around 0.6)
Lee *et al.*, 2006[79]	EQ-5D, QWB, VAS, HUI3	Paired *t*-test for change of scores after cochlear implant for EQ-5D, QWB, VAS, HUI and its dimensions	Mean change scores were statistically significant ($p < 0.05$) for EQ-5D,VAS, QWB, HUI3, HUI hearing and emotion dimensions
Hol *et al.*, 2004[70]	EQ-5D, EQ-5D responses, VAS, HHDI and SF-36	Change and ES of EQ-5D, EQ-5D responses, VAS, HHDI domains and SF-36 domains after bone-anchored hearing aid fitting	For both air-conduction hearing aid and : conventional bone-conduction hearing aid groups, mean change scores of EQ-5D and EQ-5D index and its five dimensions, VAS, SF-36 and subdomains were small and not significant. ESs were also small at 0.05 for EQ-5D and 0.1 for VAS. ES for mobility, self-care and pain dimension of EQ-5D and role limitation (emotional), mental health and pain domains were larger at around 0.3. Mean change ES for HHDI disability and handicap dimensions were large at 1.42 and 0.79
Joore *et al.*, 2002a,[71] 2002b,[74] 2003a[72]	EQ-5D responses, EQ-VAS, ADPI, hearing VAS, SF-36 social domain, Amsterdam Inventory	Change of scores of different measures after hearing aid fitting	After a hearing aid fitting, the mean scores on the first five questions of ADPI, Amsterdam Inventory and hearing VAS showed a significantly significant reduction. The largest improvements were found in 'detection of sounds' and 'intelligibility in quiet' and the smallest improvement was in 'intelligibility in noise'. This change maintained to the second follow-up. Change in ADPI from baseline to T2 and hearing loss (BEPTA) were not correlated (r = –0.066); the correlation between gain in ADPI and reported degree of satisfaction with the hearing aid at the second follow-up measurement was higher (r = 0.389, $p < 0.01$). EQ-5D VAS showed slight improvement after the hearing aid fitting (paired differences = 0.02, non-significant). The correlation between change in EQ-5D VAS and ADPI scores was low at –0.039. Response to EQ-5D dimensions showed little change over time with only the feeling dimension improving significantly from baseline to T1
Vuorialho *et al.*, 2006a,[77] 2006b[78]	EQ-5D, VAS, HHIE, SRT and WRS	Mean change and statistical test (paired *t*-test or Wilcoxon signed ranks tests) for different measures after hearing aid	The hearing aid improved the mean SRT and also slightly improved the mean WRS. The mean HHIE-S scores changed from 28.7 to 12.7 6 months after fitting the hearing aid for the first time. The EQ-VAS score changed significantly 6 months after the hearing aid fitting. No change was detected for the EQ-5D index

Study (author, year)	Instruments	Assessment methods	Results summary
Cheng *et al.*, 2000[80]	HUI3, VAS, TTO	Perceived change scores and correlations between change scores	VAS: 92% perceived improvement of QoL, 4% no change, 4% decrease (one required reimplantation; one encountered difficulty during rehabilitation). HUI: 95% improved and 5% decreased. TTO: 78% improved and 22% no change. Pearson correlations: VAS/TTO: 0.57 (*n* = 49); VAS/HUI: 0.44 (*n* = 22); TTO/HUI: 0.48 (*n* = 15)
Damen *et al.*, 2007[69]	HUI3, NCIQ	Statistically significant difference between scores of different instruments and their subdomains pre and post cochlear implant	Where significant changes in five of the NCIQ domains (speech perception advanced, speech perception basic, speech production, self-esteem, activities) were found, HUI3 index also showed significant improvement
Lovett *et al.*, 2010[66]	HUI3, VAS, SSQ	Gain in scores of different measures	SSQ demonstrated significant difference between gains of unilateral and bilateral groups but HUI3 and VAS did not show this

ADPI, Audiological Disabilities Preference Index; BEPTA, better ear PTA; HHDI, hearing handicap and disability index; HHIE, Hearing Handicap Inventory for the Elderly; HHIE-S, Hearing Handicap Inventory for the Elderly – Screening; NCIQ, Nijmegen Cochlear Implant questionnaire; SRM, standardised response mean; SRT, Speech reception thresholds; SSQ, Speech, Spatial and Qualities of hearing scale for parents; WRS, Word Reception Scores.

Appendix 7 Summary of validity for utility measures: skin conditions

Study reference grouped by condition (author, year)	Assessment methods	Results
Psoriasis and psoriatic arthritis		
Bansback *et al.*, 2006[83]	Known groups (regression model predicts EQ-5D from HAQ-DI) Convergent validity	Coefficient: –0.31 (*p* = 0.03)
Brodszky *et al.*, 2010[92]	Known groups (other) Convergent validity	All groups: standard mean differences of EQ-5D were comparably lower than PsAQoL and HAQ. Significant differences were found for two groups for EQ-5D, three groups for PsAQoL and four groups for HAQ. Strong Spearman's rank-order correlation (> 0.5) between EQ-5D and HAQ, PsAQoL, the patient pain VAS, the patient global VAS and the BASDAI
Christophers *et al.*, 2010[93]	Known groups (severity) Psoriasis and psoriatic arthritis	EQ-5D of psoriatic arthritis is lower than psoriasis patients (0.56 vs. 0.82, *p* < 0.0005). Psoriasis effects on every day tasks [lower than psoriatic arthritis patients (2.34 vs. 2.85 *p* < 0.001)]
Daudén *et al.*, 2009[84]	Known groups (severity) Continuous vs. paused therapy	After treatment, difference (*p* < 0.05) found for EQ-5D, EQ-VAS and DLQI, but not for HAD-D, HAD-A or SF-36 vitality and satisfaction survey
Van de Kerkhof, 2004[82]	Known groups (case–control) Psoriasis patients vs. general population	Psoriasis patients reported greatest problems on EQ-5D pain and anxiety than general population (no significant data reported)
Luger *et al.*, 2009[96]	Known groups (Severity) Patients with/without joint pain; with/without nail psoriasis	Joint pain groups: differences (*p* < 0.1) for EQ-5D, EQ-VAS, PASI, DLQI, SF-36 vitality and HADS but no significant difference (*p* > 0.10) for BSA. Nail psoriasis group: differences (*p* < 0.1) for EQ-5D, EQ-VAS, BSA, PASI and HADS-depression but no significant difference (*p* > 0.1) for SF-36 vitality scores and HADS-anxiety subscale
Reich *et al.*, 2009[85]	Known groups (case–control) Psoriasis patients and the UK general population	The EQ-5D, EQ-VAS, FACIT-F and DLQI scores of people with psoriasis were lower than those of UK population
Shikiar *et al.*, 2006[95]	Convergent validity	EQ-5D showed moderate to strong correlations with DLQI, PASI, PGA EQ-VAS and SF-36 domains. EQ-5D and DLQI more highly correlated with the PASI and PGA than any of the SF-36 domains
Weiss *et al.*, 2002[87]	Known groups (case–control) Psoriasis patients vs. population with no chronic conditions	EQ-5D and SF-36 of psoriasis patients were lower (*p* < 0.01) than a population of healthy subjects
	Convergent validity	Patient's SWLS scores were correlated with EQ-5D (0.46, *p* = 0.006) and VAS (0.48, *p* = 0.004) and all eight dimensions of SF-36 (0.34–0.65, *p* < 0.05). EQ-5D (0.62–0.78, *p* < 0.001) and EQ-VAS (0.48–0.76, *p* ≤ 0.003) correlated with the eight dimensions of SF-36
Acne		
Klassen *et al.*, 2000[81]	Known groups (case–control) Acne patients vs. population sample (20–39 years)	Acne patients reported higher proportions of moderate or severe problems for most EQ-5D dimensions (especially pain and anxiety) than population sample

Study reference grouped by condition (author, year)	Assessment methods	Results
Hidradenitis suppurativa		
Matusiak *et al*., 2010[89]	Known group (severity) Hurley classification I, II and III	Differences ($p < 0.01$) between Hurley's classification groups were found for EQ-5D, EQ-VAS, DLQI, BDI-SF, FACIT-F, QLES-Q and the GQ 6-item scale
	Convergent validity	Moderate correlations were found between the number of sites affected and the EQ-5D, DLQI and FACIT-F (correlations ranged from 0.28 to 0.39, $p < 0.05$)
Hand eczema		
Moberg *et al*., 2009[90]	Known groups (severity): with/without hand eczema Known groups (others): age, gender	Hand eczema: EQ-5D and EQ-VAS differ ($p < 0.05$) between groups, as well as between age and gender subgroups. The proportions of people reporting problems in EQ-5D dimensions were significantly larger in the group with hand eczema compared with patients without hand eczema
	Convergent validity	Strong correlations were found between EQ-5D and EQ-VAS
Venous leg ulcers		
Walters *et al*., 1999[91]	Known groups (non severity and severity): age, Mobility, Initial ulcer size Current and maximum ulcer duration	Age group: ES of EQ-5D and SF-MPQ were less than 0.2. Difference ($p < 0.05$) detected by SF-36 (PF, GHP and MH) and the FAI. Mobility group: differences ($p < 0.05$) detected by five dimensions of the SF-36 (PF, RL, Pain, VT and SF), EQ-5D, FAI and EQ-VAS. Initial leg ulcer size group: small ES for four measures. Current ulcer duration and maximum ulcer duration: small ESs observed for four measures
	Convergent validity	EQ-5D achieved moderate to high correlations with SF-36 dimensions, FAI and SF-MPQ (larger than between other measures)

BASDAI, Bath Ankylosing Spondylitis Disease Activity Index; FACIT-F, Functional Assessment of Chronic Illness Therapy – Fatigue; GHP, general health profile; GQ, Global question index; MH, mental health; PF, physical functioning; PGA, Physician Global Assessment of psoriasis; QLES-Q, Quality of Life Enjoyment and Satisfaction Questionnaire; RL, role limitations; SF, social functioning; SWLS, Satisfaction With Life Scale; VT, vitality.

Appendix 8 Summary of responsiveness for utility measures: skin conditions

Study reference grouped by condition (author, year)	Method of assessment	Responsiveness results
Psoriasis and psoriatic arthritis		
Daudén *et al.*, 2009[84]	Examine changes between baseline and 54 weeks' follow-up for two treatment groups	EQ-5D, EQ-VAS, DLQI, HADS-anxiety and SF-36 vitality improved statistically ($p < 0.05$) and clinically meaningfully from baseline for both treatment groups
Van de Kerkhof, 2004[82]	Examine change between baseline and end of treatment at 4 weeks	Significant improvement detected by psoriasis disability index, EQ-VAS, VAS ($p < 0.01$) and EQ-5D pain/discomfort and anxiety/depression (no significant information)
Luger *et al.*, 2009[96]	Examine change before and after treatments in 54 weeks	Joint pain patients: DLQI improved by 8.86 (61%), EQ-5D by 0.17 (29%), EQ-VAS by 12.87 (23%), SF-36 vitality by 5.55 (11%), HADS-depression scores by 1.9 (29%) and HADS-anxiety sores by 2.27 (28%) (all $p < 0.001$). Patients with joint pain had greater improvement than patients without joint pain in DLQI, EQ-5D utility index and HADS-depression and HADS-anxiety after treatment. Nail psoriasis patients: NAPSI improved by 2.38 (51%). Significant improvement observed for DLQI and EQ-VAS but not EQ-5D
Reich *et al.*, 2009[85]	Test improvement after treatments	At week 12, treatment group achieved significant improvement than placebo in total DLQI (–7.4 vs. –1.2, $p < 0.0001$), six DLQI domains ($p < 0.01$), EQ-5D (17% vs. 3%, $p < 0.05$), EQ-VAS (11% vs. 8%, $p < 0.01$) and moderate improvement in FACT-F (1.3 vs. 0.3, no significant difference between groups). A total of 37.5% of treatment group and 2.2% placebo group achieved a PASI 75 response ($p < 0.0001$). At week 24, both treatment and placebo groups DLQI total and domain scores improved (–9.6 vs. –7.1), EQ-5D (23% vs. 19%), VAS (29% vs. 3.9%) and FACT-F (3.7 vs. 2.9). A total of 71.1% of the treatment group and 44% of the placebo achieved PASI 75 response ($p < 0.05$)
Revicki *et al.*, 2008[94]	Examine change of scores over time	At week 16, DLQI improved by 9.1 (adalimumab), 3.4 (methotrexate), 5.7 (placebo) and differences between improvements in groups was statistically significant. Statistically significant improvement for the adalimumab group detected by EQ-5D, DLQI, PASI and significantly different with placebo ($p < 0.001$)
Shikiar *et al.*, 2006[95]	Examine correlations between changes of patient-reported outcomes with changes in clinical measures (PASI and PGA); Compare improvements between two groups (defined as PASI responder and non responder)	Correlations 0.69 ($p < 0.001$) for changes of DLQI with PASI, 0.71 ($p < 0.001$) for DLQI with PGA, 0.57 ($p < 0.001$) for EQ-5D PASI and –0.44 ($p < 0.001$) for EQ-5D and PGA. EQ-5D, DLQI, PASI, PGA, EQ-VAS and most SF-36 domains detected significant differences between responders and non-responders. DLQI was the most responsive (ES 0.4) and EQ-5D and EQ-VAS were similar with several SF-36 domain scores (ES 0.12)
Shikiar *et al.*, 2007[86]	Examine changes of measures between baseline and 12 weeks follow-up by treatment groups	Two treatment groups improved greater than placebo in DLQI (10 vs. 1.3), EQ-5D ($p < 0.01$), EQ-VAS ($p < 0.01$) and most SF-36 domains ($p < 0.05$, except physical functioning)
Weiss *et al.*, 2006[88]	Examine change of scores after treatment	At the end of 2 weeks of therapy, PASI achieved 35% improvement ($p < 0.001$), EQ-5D 11.5% improvement ($p = 0.007$), BSA improved 20.4%($p < 0.001$), DLQI improved 40.2% ($p < 0.001$) and EQ-VAS improved 8.2% ($p < 0.001$). The patient's perception of disease severity by SAPASI improved 26.2% ($p = 0.04$)

Study reference grouped by condition (author, year)	Method of assessment	Responsiveness results
Acne		
Klassen *et al.*, 2000[81]	Examine change after treatment (4 and 12 months) and ES of change	After treatment, the proportion of subjects to report a moderate problem on EQ-5D dimensions dropped greatly. EQ-5D, SF-36 PCS, DLQI, and acne grade changed significantly at 4 months. Change was small for EQ-VAS. ESs were 1.57 (the acne grades), 0.98 (DLQI), 0.3–0.45 (SF-36 summary score) and 0.44 and 0.53 (EQ-5D)
Venous leg ulcers		
Walters *et al.*, 1999[91]	Assess change over time and SRM against patient's group by leg ulcer healed status or by response to the self-perceived question (item two of SF-36)	By leg ulcer healed status: at 3 months, EQ-5D detected deterioration of health status for both groups, which was agreed by SF-36 but conflicted with SF-MPQ and VAS. There were small and insignificant SRMs for EQ-5D, SF-36 and FAI but moderate to large SRMs for SF-MPQ. There was no different health change between the healed and no healed groups except for Pain Rating Index (sensory) of SF-MPQ and VAS. After 12 months, changes in EQ-5D and most SF-36 domains were detected over time and the differences were significant between groups. By the transition question: at 3 months, a significant pattern (ANOVA) found for all instruments, except PF and RL dimensions of SF-36, with a worse response of the transition question associated with negative scores but a better response not associated with positive changes

PF, physical functioning; PCS, physical component score; PGA, Physician Global Assessment of psoriasis; RL, role limitations; SRM, standardised response mean.

Appendix 9 Summary of reliability for utility measures: cancers

Study reference grouped by condition (author, year)	Instrument	Assessment methods	Results
Brain cancer			
Le Gales *et al*., 1999[163]	HUI3	Internal consistency Inter-rater reliability	Multitrait analysis was used to assess internal reliability. Correlations of questions within the attribute to which they contribute were examined to check that they were higher than correlations with other attributes. The authors confirmed that this was the case, except for questions 11 and 12 and the cognition attribute when completed by the parent; however, the authors also noted that numerous unexpected correlations were found to be statistically significant There were significant correlations between patient, parent and physician assessments. The hearing dimension demonstrated the greatest amount of agreement between raters. This was followed by speech, ambulation and dexterity. The weakest agreement was between raters of the emotion, cognition and pain dimensions
Hodgkin's lymphoma			
Klaassen *et al*., 2010[186]	HUI3	Inter-rater reliability	Fair to substantial agreement
Kidney/renal cancer			
Cella *et al*., 2010[168]	EQ-5D	Stability across treatment groups	EQ-5D, VAS and FACT scores do not differ between the country cohorts, which provides some evidence for the reliability of the instruments in multinational trials
Leukaemia			
Barr *et al*., 1997[174]	HUI3	Inter-rater reliability	No differences were found on other measures. No significant effect of assessor on HUI3 score was found. This was also apparent at the dimension level (for the mobility, emotion and pain dimensions)
Hahn *et al*., 2003[176]	EQ-5D	Stability across treatment groups at baseline	No differences were found on other measures. As expected, no significant differences between the treatment groups as baseline for EQ-5D
Lymphoma			
van Agthoven *et al*., 2001[177]	EQ-5D	Stability across treatment groups at baseline	As expected, no significant differences between the treatment groups as baseline for EQ-5D and EORTC QLQ-C30
Witzens-Harig *et al*., 2009[192]	EQ-5D	Stability across treatment groups at follow-up	As expected, no significant difference in QoL scores between the groups at follow-up for EQ-5D, EORTC QLQ-C30
ML/AML			
Banks *et al*., 2008[178]	HUI2/HUI3	Inter-rater reliability	The HUI2 showed substantial accordance between the child and parent report, whereas, for the HUI3, the concordance was moderate. The concordance for the PedsQL was lower. Indicates reliability of HUI assessments across raters

Study reference grouped by condition (author, year)	Instrument	Assessment methods	Results
Musculoskeletal cancer			
Lee *et al.*, 2003[184]	EQ-5D	Internal consistency	The authors examined the validity and reliability of a condition-specific system (Musculoskeletal Tumor Society functional evaluation system) relative to EQ-5D and SF-36. They examined the internal consistency of EQ-5D dimensions relative to the overall score and of individual SF-36 questions to summary scores. The authors concluded that internal consistency was in the range defined as high for both measures
Cancer survivors			
Barr *et al.*, 2000[133]	HUI3	Inter-rater reliability	At least 81% agreement across the HUI2/3 domains for both Wilm's tumour and neuroblastoma
Boman *et al.*, 2009[134]	HUI3	Inter-rater reliability	Agreement range across the dimensions 60% (pain) to 95.5% (hearing) for survivors/parents. ICC's in the range of 0.40 (pain) to 0.96 (self-care)
Felder-Puig *et al.*, 2000[131]	HUI3	Inter-rater reliability	Percentage agreement between the three raters ranged from 56% to 100%. Kappa estimates ranged from 0.14 to 1, exhibiting a broader range
Fu *et al.*, 2006[130]	HUI3	Inter-rater reliability	Substantial agreement across the raters for the vision, hearing and ambulation domains and low agreement across the raters for the emotion domain. Patients are more likely to report moderate or severe emotion ($p < 0.001$) and cognition ($p < 0.003$) than parents/physicians, and patients and parents are more likely to report moderate or severe pain than physicians ($p < 0.001$)
Barr *et al.*, 1999[127]	HUI2	Inter-rater reliability	Inter-rater reliability was higher for the more observable attributes of mobility and self-care. Pain also displayed reasonable agreement at a higher level than emotion. ICCs indicate that there is a strong agreement between raters for HUI2 utility scores

ICC, intraclass correlation.

Appendix 10 Summary of validity for utility measures: cancers

Study reference grouped by condition (author, year)	Measure	Assessment methods	Results
Brain cancer			
Le Gales *et al.*, 1999[163]	HUI3	Known-group validity (severity) Face validity	Difference in the number of impaired HUI attributes is significantly different between levels of health status as assessed by physicians. No significant differences across groups defined according to levels of radiation therapy received were found. Approximately 70% of children and 80% of parents concluded that all of the important aspects of health status were covered. Physicians were more ambivalent
McCarter *et al.*, 2006[162]	HUI3	Known-group validity (case–control and severity) Convergent validity	All of the HUI3 dimensions were significantly different except emotion with the patient sample reporting lower utility scores. The utility scores differ between the tumour groups but no statistical tests of significance were reported. The majority of correlations between the KPS, MMSE and HUI dimensions were moderate or strong (> 0.35)
Breast cancer			
Chang *et al.*, 2004[143]	HUI3	Convergent validity	A strong and significant correlation was observed for HUI3 and FACT-An and FACT-F. A less strong correlation was observed between three subscales of HUI3 (ambulation, emotion, cognition) and FACT-An and FACT-F
Crott *et al.*, 2010[146]	EQ-5D	Convergent validity (through regression)	A statistical relationship was estimated between EORTC QLQ-C30 and EQ-5D. Individual EORTC QLQ items better explained EQ-5D values than the total EORTC QLQ-C30 score. The preferred model showed good fit (adjusted R^2 of 0.801 and RMSE of 0.096). The statistically significant items were physical, emotional and social functioning, pain, constipation and diarrhoea. Items that were not included (not significant) were role and cognitive function, fatigue, nausea-vomiting, dyspnoea, appetite and financial problems
Freedman *et al.*, 2010[145]	EQ-5D	Convergent validity	Strong correlations were observed between EQ-5D index and EQ-VAS
Jansen *et al.*, 2004[137]	EQ-5D	Convergent validity (through regression)	The pattern of results for EQ-5D was consistent with other measures. None of the measures, including EQ-5D index, VAS, HADS-anxiety and HADS-depression, had a significant relationship with perceived choice or chemotherapy ($p > 0.05$) but did for interactions of choice and chemotherapy ($p < 0.05$) and age ($p < 0.05$)
Kimman *et al.*, 2009[144]	EQ-5D	Convergent validity	Correlation coefficients: 0.423 (EQ-5D index vs. EROTC) and 0.634 (EQ-VAS vs. EORTC). EQ-VAS and EQ-5D index both moved in the expected direction with EORTC
Lidgren *et al.*, 2007[138]	EQ-5D	Known-group validity (severity) Convergent validity	The EQ-5D index differentiated between groups categorised according to those in their first year after primary breast cancer (state P) and those in the metastatic disease (state M) compared with patients in their second or more years after primary cancer or recurrence (state S), but did not differentiate patients in their first year after recurrence (state R) compared with state S. The TTO differentiated group M with S, but not groups P and R with S. For all breast cancer states except 'state R', TTO values were significantly higher than EQ-5D indices with the correlation being 0.44

Study reference grouped by condition (author, year)	Measure	Assessment methods	Results
Breast cancer			
Lovrics *et al*., 2008[141]	HUI3	Convergent validity	Most Pearson correlations between HUI3 and SF-36 were statistically significant ($p < 0.01$). HUI3 showed moderate to very strong positive correlations to SF-36 PCS scores and moderate to substantial positive correlations to SF-36 MCS scores
Cervical cancer			
Korfage *et al*., 2010[164]	EQ-5D	Known-group validity (severity) Convergent validity	The EQ-5D scores indicate that the borderline/mild dyskaryosis group has worse HRQL than the healthy population but the difference is not statistically significant. In contrast, the differences in the SF-12 PCS and MCS, and STAI are all significant. Mixed patterns were observed the EQ-5D, which did not find the group differences that were found using other generic (SF-12) and condition specific (STAI/PCQ) measures. Perceived risk of being diagnosed with cervical cancer was associated with EQ-5D ($p = 0.004$) and PCQ ($p < 0.005$) score, but not with MCS ($p = 0.12$) or STAI ($p = 0.18$)
Maissi *et al*., 2005[167]	EQ-5D	Known-group validity (severity)	EQ-5D is as sensitive to HRQL issues in cervical cancer (around anxiety and distress) as measured by the STAI and General Health Questionnaire
Whynes *et al*., 2008a[165]	EQ-5D	Convergent validity	A range of significant validity results demonstrating that the EQ-5D is correlated with both the EQ-5D VAS and the HADS, which is a widely used measure of anxiety and depression
Whynes *et al*., 2008b[166]	EQ-5D	Known-group validity (severity)	At post-study follow-up, the EQ-5D, HADS-anxiety and HADS-depression do not discriminate between the control and intervention groups, but the chance dimension of the MHLCS does
Colon cancer			
Doornebosch *et al*., 2007[115]	EQ-5D	Known-group validity (severity)	Mean EQ-VAS scores were similar after treatments (TEM, TME and controls), EQ-5D indices did not differ between the three groups, sores of EORTC QLQ-C30 subscales showed no differences across between groups and EORTC QLQ-CR38 showed a significant difference between TEM and TME groups regarding defecation problems with TEM patients having less defection problems than TME patients ($p < 0.05$)
Gosselink *et al*., 2006[121]	EQ-5D	Known-group validity (case–control, severity)	Mean EQ-5D index of CPA was significantly higher than the gender-age matched general population whereas LRA and APR groups were similar with general population. EQ-5D indices did not differ between the three treatment groups. EQ-VAS scores of CPA were significantly higher than the gender matched general population whereas LRA and APR groups were similar to the general population. Significant differences were found between the groups who had CPA and LRA, and between the CPA and LRA groups. Significant differences between the three groups were found on five subscales of the EORTC measures
Hamashima, 2002[117]	EQ-5D	Known-group validity (case–control)	No significant differences were revealed between with and without stoma groups on the basis of EQ-5D index, EQ-VAS and stoma-specific QoL questions relating to outing and travel question
Janson *et al*., 2007[122]	EQ-5D	Known-group validity (severity)	EQ-5D index, EQ-VAS and EORTC QLQ-C30 revealed no differences between study groups at baseline

Study reference grouped by condition (author, year)	Measure	Assessment methods	Results
Colon cancer			
Ramsey *et al.*, 1998[190]	HUI3	Known-group validity	FACT-C summary scores showed little variation over time by tumour stage at diagnosis. The smoothed curves of HUI3 values suggested that the pattern of scores over time differs depending on the initial stages at diagnosis. HUI3 values did not different significantly by tumour stage at diagnosis. FACT-C scores showed a non-significant trend toward declining health for more advanced stages of disease and showed little variation over time by tumour stage at diagnosis. The smoothed curves of HUI3 values suggested that the pattern of scores over time differs depending on the initial stages at diagnosis. HUI3 values did not differ significantly by tumour stage at diagnosis. FACT-C scores showed a non-significant trend toward declining health for more advanced stages of colorectal carcinoma
Sharma *et al.*, 2007[123]	EQ-5D	Convergent validity	Only HADS-anxiety scores, positive and negative affect schedule score and FACT emotional well-being subscale score were moderately significantly correlated with TNM stage. Other measures, including EQ-5D index and EQ-VAS, were not significant and had low correlation to the TNM stages
Siena *et al.*, 2007[118]	EQ-5D	Known-group validity (severity)	Results for the FACT colorectal symptom index and EQ-5D for all treatment groups were similar regardless of imputation method. Similar results for panitumumab and best supportive care patients stratified by tumour progress status were observed for EQ-VAS and EORTC global scale
Wilson *et al.*, 2006[120]	EQ-5D	Known-group validity (severity)	Except the SF-12 MCS score, EQ-5D index, EQ-VAS, SF-12 general health, SF-12 PCS, QLQ general health, FACT-C total scores declined with advancing preoperative ECOG performance status. Multivariate analysis demonstrated that EQ-5D, EQ-VAS, SF-12 GH, SF-12 PCS and QLQ-GH scores were significantly different between ECOG performance status groups
Gastric (and related) cancer			
Homs *et al.*, 2004[149]	EQ-5D	Known-group validity (case–control)	Large difference in EQ-5D scores between the general and study population groups, with those in the study population reporting lower EQ-5D utility scores at baseline
Kontodimopoulos *et al.*, 2009[147]	EQ-5D/ SF-6D	Convergent validity (through regression)	EORTC physical and emotional functioning and global health status significantly predicted EQ-5D utility scores. Indicates relationship between some EORTC dimensions and EQ-5D
			EORTC social and emotional functioning, pain, constipation, dyspnoea and global health status predicted SF-6D utility score (significant predictor). Indicates relationship between some EORTC dimensions and SF-6D
O'Gorman *et al.*, 1998[151]	EQ-5D	Known-group validity (severity) Convergent validity	EQ-5D scores and most of the EORTC subscales are significantly lower in the weight-losing group. Within the weight-losing group, no significant difference in EQ-5D or EORTC values but KPS significantly lower in the inflammatory response group. Significant correlations between appetite scores and EQ-5D (0.43)/EORTC (0.61)/KPS (0.55) scores. Overall, EQ-5D is demonstrating validity in comparison to condition specific measures

Study reference grouped by condition (author, year)	Measure	Assessment methods	Results
Gastric (and related) cancer			
Rogers *et al.*, 2006[148]	EQ-5D (dimensions)	Known-group validity (case–control, severity) Convergent validity	Higher percentage of patients reporting problems in the EQ-5D dimensions than a general population reference group but significance not reported. Patients having radiotherapy report significantly lower VAS and higher EQ-5D mobility/usual activity dimension scores than those not having radiotherapy. University of Washington QoL questionnaire overall QoL score significantly correlated with EQ-5D mobility/usual activity and anxiety/depression dimensions. University of Washington QoL questionnaire activity/recreation and EQ-5D usual activity/mobility/self-care dimensions are significantly correlated. University of Washington QoL questionnaire is correlated with anxiety/depression, pain and usual activities dimensions
Shenfine *et al.*, 2009[150]	EQ-5D	Known-group validity (severity)	EQ-5D significantly discriminates between the treatment groups at follow-up
Wildi *et al.*, 2004[152]	EQ-5D	Known-group validity (severity)	Those at stage 0 (low severity) display higher EQ-5D utility scores than stages 1–3. However, the overall difference between the stages is not significant and the EQ-5D scores do not decrease as expected between stages 1–3. This provides limited evidence for the known-group validity of EQ-5D
Hodgkin's lymphoma			
Klaassen *et al.*, 2010[185]	HUI3	Convergent validity	Strong correlation between HUI3 and other measures
Kidney/renal cancer			
Castellano *et al.*, 2009[171]	EQ-5D	Convergent validity	EQ-5D index scores are significantly correlated with the FACT-G and FACT kidney symptom index at 0.6 or above. The EQ-5D and EQ-5D VAS are more highly correlated with the condition specific instruments than with each other
Cella *et al.*, 2008[169]	EQ-5D	No formal tests but pattern was observed	EQ-5D, VAS and FACT scores follow a similar pattern across the study follow-up period
Cella *et al.*, 2010[168]	EQ-5D	No formal tests but pattern was observed	EQ-5D, VAS and FACT scores do not differ between the country cohorts, which provides some evidence for the validity of the instruments in multinational trials
Sternberg *et al.*, 2010[193]	EQ-5D	No formal tests but pattern was observed	EQ-5D, VAS and EORTC global health follow the same pattern across the study period
Yang *et al.*, 2010[170]	EQ-5D	No formal tests but pattern was observed	EQ-5D and VAS scores follow the same pattern which indicates agreement between the measures
Leukaemia			
Cox *et al.*, 2005[187]	HUI3	Acceptability Missing data/ceiling effects/proxy completer comments	A significant quantity of data were missing, despite the fact that proxies had undergone extensive orientation to the HUI. Speech was the only category with no missing data. There is a high ceiling effect across all HUI attributes, with vision, hearing and dexterity displaying the highest levels. Comments that there is missing data because attribute was not able to be observed. Comments that measures functional performance, not QoL. Limited evidence for the acceptability of HUI3
Hahn *et al.*, 2003[176]	EQ-5D	Known-group validity (severity)	EQ-5D demonstrated treatment differences at all follow-up time points

Study reference grouped by condition (author, year)	Measure	Assessment methods	Results
Liver metastases			
Langenhoff *et al.*, 2006[180]	EQ-5D	Known-group validity (severity)	Both the EQ-5D and EORTC QLQ-C30 are sensitive to differences between patient treatment groups
Mendez Romero *et al.*, 2008[172]	EQ-5D	Known-group validity (case–control)	Both the EQ-5D and EORTC are sensitive to differences between a metastatic liver tumour patient group and a general population group similar in terms of age in the expected direction
Krabbe *et al.*, 2004[179]	EQ-5D	Known-group validity (severity)	The EQ-5D/VAS and EORTC global health scale discriminated well between the three treatment groups, and followed a similar pattern across the study period
Lung cancer			
Pickard *et al.*, 2007[103]	EQ-5D	Known-group validity (severity)	Minimally important differences for the EQ-5D index by FACT quintile subgroups reveal that the EQ-5D is able to distinguish between the patients at the various FACT quintiles. However, the results should be interpreted with caution owing to the small sample size
Trippoli *et al.*, 2001[173]	EQ-5D	Known-group validity (severity)	The EQ-5D significantly distinguishes between patients with metastasis and those without
		Convergent validity	There are significant correlations between the EQ-5D index score and VAS and also between the EQ-5D and SF-36
Lymphoma			
Doorduijn *et al.*, 2005[188]	EQ-5D	Known-group validity (severity)	EQ-5D significantly discriminates between clinical indicator severity levels, with those at a more severe level reporting lower EQ-5D index scores
ML/AML			
Banks *et al.*, 2008[178]	HUI2/HUI3	Convergent validity	There were correlations of at least 0.2 between all pairs of measures used at baseline. The proxy HUI2/3 was substantially correlated with the PedsQL generic scores. The proxy HUI2/3 and the PedsQL generic showed substantial correlations with the CHQ physical score. Indicates concurrent validity of HUI
Slovacek *et al.*, 2007[181]	EQ-5D	Known-group validity (severity)	Difference between ML and AML EQ-5D scores index and dimension scores, with ML indicating significantly higher scores. Indicates that EQ-5D can discriminate between the level of HRQL associated with different types of cancer. However, sample size was small
MM			
Slovacek *et al.*, 2008[175]	EQ-5D	Known-group validity (others)	EQ-5D significantly decreases as age increases and non-smokers have significantly higher EQ-5D scores. Indicates known-group validity of EQ-5D across demographic variables
MM/ML			
Slovacek *et al.*, 2007[182]	EQ-5D	Known-group validity (severity)	Difference between MM and ML EQ-5D scores, with ML indicating significantly higher scores. Indicates that EQ-5D can discriminate between the level of HRQL associated with different types of cancer
Musculoskeletal cancer			
Lee *et al.*, 2003[184]	EQ-5D	Convergent validity	EQ-5D dimensions were significantly correlated with all dimensions of MSTS. Results discussed in terms of MSTS. Limited evidence for convergent validity of EQ-5D dimensions

Study reference grouped by condition (author, year)	Measure	Assessment methods	Results
Pancreatic cancer			
Muller-Nordhorn *et al.*, 2006[183]	EQ-5D	Known-group validity (case–control, other)	Male cancer patients were significantly more likely to report any problems on all five EQ-5D dimensions than the general population reference sample. However, female patients were only significantly more likely to report problems on the anxiety/depression domain EQ-5D VAS significantly discriminates between the cancer patient and general population samples for both males and females. There were no significant differences in EQ-5D and EORTC scores between males and females or patients with or without metastases
Prostate cancer			
Albertsen *et al.*, 1998[155]	HUI3	Convergent validity	The association between HUI3 and the self-administered questionnaire was not significant
Krahn *et al.*, 2007[160]	HUI3	Convergent validity	Low ICC between HUI3 and SG utilities
Sandblom *et al.*, 2004[158]	EQ-5D	Known-group validity (severity)	EQ-5D scores discriminate between survival groups
Shimizu *et al.*, 2008[156]	EQ-5D/SF-6D	Known-group validity (severity)	EQ-5D and SF-6D discriminate between severity groups as indicated by the number of symptoms. A higher number of symptoms resulted in lower utility scores, as expected
Sullivan *et al.*, 2007[157]	EQ-5D	Known-group validity (severity)	The change in HRQL seemed worse for patients undergoing chemotherapy and TURP than those who did not. Some evidence that generic and condition specific instruments discriminate between different treatment groups
Weinfurt *et al.*, 2005[159]	EQ-5D	Known-group validity	The generic and condition specific instruments are able to pick up effects by patients groups experiencing the different types of SRE
Spinal metastases			
Falicov *et al.*, 2006[101]	EQ-5D/HUI3	Convergent validity	Low/moderate correlation between the utility measures
Non-specific cancer			
Capuano *et al.*, 2008[107]	EQ-5D	Convergent validity	Anaemia (p = 0.031) and weight loss (p = 0.002) were significantly influenced EQ-5D scores. Inflammation was not statistically significant and relationship with fatigue was not directly tested, but both anaemia and weight loss significantly impacted on fatigue
Pickard *et al.*, 2007[103]	EQ-5D	Known groups (severity)	A trend was seen in line with expectations according to severity. Statistical significance not presented. EQ-5D scores decrease as ECOG increases (i.e. as performance status worsens) and as functional assessment (FACT) increases. This applies to both US and UK tariffs, although is more pronounced with the UK tariff
Ravasco *et al.*, 2003[104]	EQ-5D	Known groups (severity)	High-risk patients had statistically significantly worse scores than low risk patients on all dimensions at baseline (p = 0.001) and at the end of the study (p = 0.01)

Study reference grouped by condition (author, year)	Measure	Assessment methods	Results
Non-specific cancer			
Wang *et al.*, 2008[97]	EQ-5D	Known groups (case–control)	The likelihood of reporting any problem was statistically significantly higher for cancer patients compared with other patients for the usual activities dimension ($p < 0.01$) but not for the other dimensions. On the SF-36, there were statistically significant differences on the physical functioning and general health domains but not any of the others. Cancer was a significant explanatory variable for EQ-VAS scores, but not for SF-36 summary scores
Pickard *et al.*, 2007[114]	EQ-5D	Correlations Known groups (severity)	All dimensions statistically significant at varying strengths. Crude summary score decreases as ECOG scores increase as expected. EQ-5D summary scores: ECOG 0 = 89.7 (n = 98), ECOG 1 = 76.0 (n = 205); ECOG 2 = 68.6 (n = 100) and ECOG 3 = 57.0 (n = 20)
Barton *et al.*, 2008[98]	EQ-5D/ SF-6D	Known groups (case–control)	Significant differences between cancer and non-cancer groups were found for EQ-VAS ($p < 0.05$) but not EQ-5D or SF-6D
Bowker *et al.*, 2006[99]	HUI3	Known-group validity (case–control)	Mean difference in scores, adjusted for sociodemographics, were statistically significantly different ($p < 0.001$) for cancer, cancer and diabetes, and diabetes only compared with no cancer or diabetes. Unadjusted mean (SD) scores were statistically significantly different (ANOVA $p < 0.001$) for cancer, cancer and diabetes, diabetes only and no cancer
Cheung *et al.*, 2009[105]	EQ-5D	Known group (severity) Convergent (through regression)	At baseline/follow-up, ECOG 0 = 0.899/0.921, ECOG 1 = 0.791/0.773, ECOG 2 = 0.718/0.737 and ECOG 3 = 0.596/0.530 Social domain of FACT-G was not statistically significant in any of the models, but all other dimensions and total score were. R^2 ranged from 0.345 to 0.451
Lathia *et al.*, 2008 (abstract only)[102]	EQ-5D	Convergent (through regression)	Strongest relationship with FACT-N was with pain/discomfort ($p = 0.18$). Model fit was poor $R^2 = -0.04$
Chow *et al.*, 2010[106]	EQ-5D	Known group (severity) (stage and treatment group)	Appropriate trend found in EQ-5D scores by stage (statistical significance between stages not reported). Similar pattern was found for VAS scores. Mean (SE) HUI3 scores for CAM users: cancer stages 0, I and complete responders 0.82 (0.03); stages II/III: 0.80 (0.02); and stage IV: 0.77 (0.02). Mean (SE) HUI3 scores for non-CAM users: Cancer stages 0, I and complete responders 0.86 (0.04); stages II/III: 0.80 (0.03) and; stage IV: 0.56 (0.06). Multivariate regression analysis found that there was no statistically significant difference in EQ-5D or VAS scores between treatment groups after adjusting for covariates
Norum, 1996[100]	EQ-5D	Convergent	All three measures were highly correlated with each other (all $p < 0.0001$) based on Persons correlation and Mantel–Haenszel test
Sung *et al.*, 2003[108]	HUI3	Convergent validity Acceptability	Significant correlations between CHQ pain and HUI pain, CHQ physical and HUI mobility, CHQ mental health and HUI2 emotion. HUI utility significantly correlated with the CHQ physical scale but not the psychosocial scale. A total of 89% reported that the CHQ and HUI were easy to complete
Trudel *et al.*, 1998[109]	HUI3	Convergent validity Known-group validity Content validity	The correlations between the HUI3 utility and dimension scores and the other measures included are in the moderate range. The difference between the groups is statistically significant for the HUI3 emotion, pain, self-care and overall utility score. HUI3 was adequate as a descriptive health system but does not include neuropsychological or psychosocial aspects

Study reference grouped by condition (author, year)	Measure	Assessment methods	Results
Cancer survivors			
Barr *et al.*, 2000[133]	HUI3	Known-group validity	The hearing ($p = 0.01$) and speech ($p = 0.02$) dimensions significantly discriminate between the samples but no other dimension reaches significance
Boman *et al.*, 2009[134]	HUI3	Known-group validity (case–control)	All HUI3 attributes display significant difference between survivors and controls (survivors better health) except emotion and pain. Range of significant differences between the tumour diagnoses and controls
Felder-Puig *et al.*, 2000[131]	HUI3	Known-group validity (severity)	Significant relationship between degree of severity and HUI2 scores for the majority of groups ($p < 0.05$). For attributes, difference significant for pain and emotion
Fu *et al.*, 2006[130]	HUI3	Group differences (other, severity)	The HUI3 score for the vision dimension was higher in the Hodgkin's group compared with acute lymphoblastic leukaemia ($p < 0.01$). The difference between the emotion ($p < 0.01$) and HRQL ($p < 0.05$) scores are significantly different with the Canadian group displaying higher mean scores. As expected, the differences in mean single attribute scores between acute lymphoblastic leukaemia and Hodgkin's disease patients were not statistically significant
Grant *et al.*, 2006[135]	HUI3	Group differences (severity)	As expected, the attribute and overall utility scores were not statistically different between the two diagnosis groups
Korfage *et al.*, 2009[129]	EQ-5D	Known-group validity (case–control)	When controlling for differences in background variables, neither the EQ-5D nor the majority of the SF-36 dimensions display significant group differences between the survivors and control group (only the mental health domain of the SF-36 is significant). The STAI score is significantly different between the groups
Nijdam *et al.*, 2008[128]	EQ-5D	Known-group validity (severity)	The EQ-5D and QLQ-C30 do not differ between the treatment groups, providing evidence that the measures are performing in the same way
Nixon Speechley *et al.*, 1999[136]	HUI3	Convergent validity	Significant correlations between the HUI and CHQ across a range of similar dimensions
Barr *et al.*, 1999[127]	HUI2	Known-group validity (severity)	HUI2 can discriminate between radiotherapy treatment and disease status groups
Pogany *et al.*, 2006[132]	HUI3	Known-group validity (case–control, severity)	HUI3 utility scores discriminate between survivors and controls. There are significant differences between survivors and controls across the HUI3 dimensions and some significant discrimination by treatment groups
Shimoda *et al.*, 2005[126]	HUI3	Known-group validity (severity) Acceptability	Mean HUI scores significantly decreased in line with global health ratings for nurse and physician assessors ($p < 0.02$). For patients, the HUI3 significantly decreased ($p = 0.05$) but the HUI2 did not ($p = 0.117$). No assessor reported problems understanding and answering the questions

APR, abdominoperineal resection; CAM, complementary and alternative medicine; CPA, coloanal J-pouch anastomosis; ICC, intraclass correlation; KPS, Karnofsky performance score; LRA, low colorectal anastomosis; MCS, mental component score; MHLCS, multidimensional health locus of control scale; MMSE, mini mental state examination; PCS, physical component score; SRE, skeletal-related events; TEM, transanal endoscopic microsurgery; TME, total mesorectal excision.

Appendix 11 Summary of responsiveness for utility measures – cancers

Study reference grouped by condition (author, year)	Instrument	Assessment methods	Results
Breast cancer			
Chang *et al.*, 2004[143]	EQ-5D and HUI3	Mean change over time between groups. Correlations between change scores of HUI3 and condition-specific measures over time	Over time, HUI3 overall scores improved in the epoetin alfa group (mean 0.018, SD 0.024) but decreased in the standard of care group (mean –0.041, SD 0.254, $p = 0.036$). The difference of change score between the two groups was significant ($p = 0.036$). Emotion, ambulation and cognition of HUI3 also detected significant improvement in the epoetion alfa group compared with the standard of care group. A strong and significant correlation was observed between change scores of HUI3 and FACT-An and FACT-F scores. Less strong but significant correlations were observed for emotion, ambulation and cognation subscales of HUI3 with FACT-An and FACT-F. Over time, EQ-5D detected improvement in the epoetin alfa group: base line mean 0.71 (SD 0.22) to follow-up mean 0.78 (SD 0.15); standard of care group: baseline mean 0.72 (SD 0.23) to follow-up mean 0.76 (SD 0.19). The difference of change scores between the two groups was not significant ($p = 0.639$). Over time, for the epoetin alfa group, EQ-VAS improved from 62.13 at baseline to 70.05 at follow-up; for the standard of care group, EQ-VAS decreased from 62.88 to 60.83. The difference between two groups was significant ($p = 0.018$)
Conner-Spady *et al.*, 2001[139]	EQ-5D	ES, paired groups *t*-test; ANOVA; Friedman test	All ES EQ-5D over time was large, except EQ-5D index (T3–T4) what was 0.66. There was no significant differences in ES between EQ-5D and FLIC at T1–T3 and T3–T4. EQ-5D was consistent with other measures: significant changes in mean scores over time for EQ-5D, VAS, FLIC and FLIC subscales (physical well-being, social well-being, hardship, and nausea subscales). EQ-5D dimensions of mobility, self-care and usual activities showed significant change over time
Conner-Spady *et al.*, 2005[140]	EQ-5D	Friedman test, one-way ANOVA to assess differences in HRQL over time	EQ-5D, FLIC and QoL VAS showed a similar pattern of change. They all decreased following high-dose chemotherapy and returned to baseline level after high-dose chemotherapy. There was a significant decrease in HRQL from T1 to T3 and a return to baseline level by T8. From T4 to T7, FLIC showed a significant improvement and EQ-5D and QoL VAS showed a non-significant improvement. The Friedman test showed significant changes over time for EQ-5D mobility, self-care, usual activity and anxiety but not for pain
Kimman *et al.*, 2009[144]	EQ-5D	Correlations between anchor scores and measures of interests, SRM for subgroups, Games–Howell post hoc procedure to compare mean change scores between 'no change' subgroup and other subgroups	In the subgroup of patients with no changed global health, neither SRM of EQ-5D index nor EQ-VAS indicated an effect. For subgroups with a small deterioration or improvement, SRMs of EQ-5D index were too small to be considered as an effect, SRMs of EQ-VAS indicated a small effect. For subgroups with moderate and large improvements or deteriorations, SRMs indicated a moderate effect (> 0.5) on EQ-5D index and a large effect (> 0.8) on EQ-VAS. For the EQ-5D index, mean change scores of subgroups reporting moderate and large improvement that differed significantly from 'no change' group, the subgroups reporting small improvements or a small or moderate and large deterioration could not be differentiated from the 'no change' group. EQ-VAS differed significant between 'no change' and 'moderate and large improvement' and 'moderate and large deterioration'

Study reference grouped by condition (author, year)	Instrument	Assessment methods	Results
Breast cancer			
Polsky *et al.*, 2002[142]	HUI3	Test for change over time	Significant differences were found in VAS and HUI3 5 months after surgery. Emotion attribute of HUI3 was the only one of significance. Differences were non-significant 1 and 2 years after surgery. Choice has a short-term impact on health state preferences but no long-term benefits
Lovrics *et al.*, 2008[141]	HUI3	ANOVA and paired comparisons ES	Significant changes over time were demonstrated ($p < 0.01$) for both measures. Both scores decrease after surgery and improve over time but remain below normative values at all postoperative time points ($p < 0.01$). The HUI3 multiattribute, pain and ambulation scores and the SF-36 PCS, BP, PF, RP, VT and social functioning scores all showed a large downward ES from intensive care to the postoperative time. By 24 months, the ES for these physical variables were small or trivial
Cervical cancer			
Maissi *et al.*, 2005[167]	EQ-5D	Mean change across the study period	Mean change on EQ-5D, General Health Questionnaire and STAI is small but no significance testing is reported
Whynes *et al.*, 2008a[165]	EQ-5D	Regression predicting decrease in VAS scores between baseline and follow-up	VAS score decreases were significantly predicted ($p < 0.01$) by EQ-5D dimension increases (worsening health), decreases (improving health) and HADS increases (worsening health). Regression demonstrates that change over the study period for EQ-5D is apparent
Colon cancer			
Anderson and Palmer, 1998[119]	EQ-5D	OR for responses of EQ-5D dimensions between baseline and weeks 5 and 15 over the two groups ANOVA was used to assess RSCL in weeks 2, 5, 10 and 15	At week 2, there were significant differences between Raltitrexed and 5-FU + LV in changes from baseline for all dimensions and subdimensions of the RSCL, with the exception of the psychological symptoms and disease categories, which fell just outside the significant range. At week 2, there was a highly significant difference in favour of Raltitrexed in four EQ-5D dimensions and general health question. Patients (Raltitrexed) were three times less likely to have problems with mobility and usual activities than patients in the 5-FU + LV group (OR 2.9 and $p < 0.02$). They were also at least twice as likely to have a better general health (OR 2.3, $p < 0.001$) and they were two to three times as capable of self-care as patients in the 5-FU + LV group, but not significantly. Subsequently, the differences between the two treatment groups diminished but there were still some statistically non-significant trends in favour of reltitrexed on the EQ-5D scale and in total symptom advantages that were maintained to week 10
Doornebosch *et al.*, 2008[116]	EQ-5D	Wilcoxon signed-rank test and Mann-Witney *U*-test for change scores within or between groups. Spearman's rank-order correlation coefficient between change scores	Six months after surgery, mean Faecal Incontinence Severity Index scores decreased significantly, depicting an improvement in faecal continence. Reduction of Faecal Incontinence Severity Index was significantly greater in patients with a tumour location within 7 cm from the denatate line ($p = 0.01$) (significant correlations). EQ-VAS was significantly higher 6 months after TEM ($p < 0.02$). The observed change in EQ-VAS showed no correlation with the postoperative alterations in Faecal Incontinence Severity Index scores or tumour characteristics. Both pre and postoperative EQ-5D index scores were similar to those of the gender-age matched general population. The EQ-5D index was not affected by age and gender of the patients, surgical aspects and tumour characteristics. FIQL showed a significant improvement in two of the four domains (embarrassment and lifestyle). The domains of lifestyle, coping and behaviour and embarrassment were correlated with the Faecal Incontinence Severity Index. FIQL scores were not affected by age and gender of the patients and surgical aspects and tumour characteristics

Study reference grouped by condition (author, year)	Instrument	Assessment methods	Results
Colon cancer			
Janson *et al.*, 2007[122]	EQ-5D	Mean changes of scores between groups	ANOVA analysis of change over time of the EQ-5D index indicated no significant differences. For EORTC QLQ-C30, there was a significant benefit of LCR at the 2- and 4-week assessments. At the 12-week assessment, a borderline significance was found. In role function, there was a significant benefit of LCR at the 2-week assessment
Sharma *et al.*, 2007[123]	EQ-5D	Mean changes of scores before and after surgery	Depression measured by the HADS scale was significantly higher in the 6-week postdischarge measure (3.6 vs. 4.8, $p < 0.05$). There was no statistically significant difference in the other scores
Gastric (and related) cancer			
Homs *et al.*, 2004[149]	EQ-5D	Mean change	Stent group shown to have significantly reduced QoL on EORTC role/emotional/cognitive/social scales ($p < 0.05$). The EQ-5D and VAS show a decrease but the scores for each group are not significantly different. Limited evidence for the responsiveness of EQ-5D at a lower level than selected dimensions of the condition specific EORTC
McMillan *et al.*, 1999[153]	EQ-5D	Mean change	EQ-5D index demonstrating significant improvement in the intervention arm at follow-up. EQ-5D is responding to change in the intervention group
Verschuur *et al.*, 2009[154]	EQ-5D	Mean change	Both EQ-5D and EORTC display mean change in the expected direction over time, with both measures displaying improvement at follow-up. This provides some evidence for the responsiveness of EQ-5D in gastric cancer
Hodgkin's lymphoma			
Klaassen *et al.*, 2010[185]	HUI3	*t*-Tests, ES and area under receiver operating characteristic curve	All measures showed a significant change in summary scores between Time 1 and Time 4. All of the ESs were large and clinically relevant. The HUI had negligible to small ESs between Time 2–3 and Time 3–4, whereas the PedsQl, Lanksy Play-Performance scale and VAS had moderate to large ESs
Kidney/renal cancer			
Castellano *et al.*, 2009[171]	EQ-5D	ES/significance level	The difference between the treatment groups is statistically significant overall, demonstrating that the EQ-5D index and VAS respond to treatment effects across the study period. However, the ESs are in the range defined as small
Cella *et al.*, 2008[169]	EQ-5D	ES	The difference between the treatment groups is statistically significant overall, demonstrating that the EQ-5D index and VAS respond to treatment effects across the study period. However, the ESs are in the range defined as small
Cella *et al.*, 2010[168]	EQ-5D	No formal statistical tests	The EQ-5D and VAS results indicate that the measures respond to change in treatment groups but no formal tests have been conducted
Sternberg *et al.*, 2010[193]	EQ-5D	No formal statistical tests	It is not clear whether QoL differences between the treatment groups were not picked up by the instruments because they were not present or because of the lack of responsiveness of the questionnaires
Yang *et al.*, 2010[170]	EQ-5D	No formal statistical tests	There is some evidence that EQ-5D and VAS are able to distinguish between treatments over time but no formal tests have been conducted

Study reference grouped by condition (author, year)	Instrument	Assessment methods	Results
Leukaemia			
Barr *et al.*, 1997[174]	HUI2	Change over the study period	The HUI2 proves to be responsive across a range of indicators and in comparison to four temporary health states for which utility scores are available
Hahn *et al.*, 2003[176]	EQ-5D	Mean change over the study period	EQ-5D is picking up differences in mean change over time between the treatment groups. At three of four follow-up time points, the reduction in the EQ-5D score reflects the proportion of the sample that is showing a clinically relevant decline on the trial outcome index. This provides evidence for the responsiveness of EQ-5D
Liver metastases			
Langenhoff *et al.*, 2006[180]	EQ-5D	ES of change over the study period	Both the EORTC and EQ-5D are responding over time and demonstrating sensitivity to change in HRQL following different surgical procedures across three groups. The EORTC is responding to improvement following surgery and also a subsequent change in two groups who receive chemotherapy. However, the EQ-5D is not picking up this change as clearly
Mendez Romero *et al.*, 2008[172]	EQ-5D	Statistical significance between baseline scores and follow-up scores	The EQ-5D and EORTC findings are consistent as, overall, neither measure demonstrates significant differences in responsiveness apart from one EORTC dimension at one of the three follow-up points
Krabbe *et al.*, 2004[179]	EQ-5D	ES of change over the study period	ESs of comparable magnitude across the EQ-5D index, dimensions, and EORTC scores. Evidence for responsiveness of the EQ-5D/EORTC in comparison to each other
Lymphoma			
Doorduijn *et al.*, 2005[188]	EQ-5D	Mean change over study period	Most EQ-5D mean change scores are not significant. Some EORTC dimensions are significant. EORTC may be more responsive than EQ-5D
Van Agthoven *et al.*, 2001[177]	EQ-5D	Mean change over study period	EQ-5D index scores decrease after treatment and then improve after discharge but significance not reported. Limited evidence for EQ-5D responsiveness
Witzens-Harig *et al.*, 2009[192]	EQ-5D	Mean change over study period	Change within the intervention group over the study period is being captured by both the EQ-5D and EORTC. Evidence for the responsiveness of both measures
ML/AML			
Banks *et al.*, 2008[178]	HUI2/HUI3	Mean change in proxy report	The HUI displays a low level of change over the study period in comparison to the PedsQL but change is at a similar level to the CHQ
MM			
Uyl-de-Groot *et al.*, 2005[124]	EQ-5D	Mean change over study period	Significant mean change for EQ-5D index score and a range of EORTC QLQ-C30 dimensions at selected follow-up time points. There is some evidence for the responsiveness of EQ-5D in comparison to the condition specific EORTC QLQ-C30, but this is not consistent across time points

Study reference grouped by condition (author, year)	Instrument	Assessment methods	Results
Prostate cancer			
Krahn *et al.*, 2007[160]	EQ-5D/HUI3	Standardized ES, standardized mean response Mean change in utility Area under receiver operator curve Differential responsiveness	Internal responsiveness: generic instruments were less responsive to treatments as shown by smaller effects compared with disease specific instruments External responsiveness: utility measures – generic and disease specific – were able to discriminate between those whose health changed and those whose health did not. EQ-5D most consistently reported a high area under receiver operator curve
Sullivan *et al.*, 2007[157]	EQ-5D	Mean change across study period	Patients underwent rapid deterioration in FACT-P, EQ-5D and 10 out of 14 EORTC domains over the 9-month follow-up. This provides some evidence of responsiveness of the instruments
Weinfurt *et al.*, 2005[159]	EQ-5D	ES	The ESs for radiation to bone are larger in comparison. There is evidence to suggest that for radiation to bone SRE, ESs are significant for the total FACT-G score and the EQ-5D utility score. For pathological fracture type SRE, the ES is significant for the EQ-5D utility score
Spinal metastases			
Falicov *et al.*, 2006[101]	HUI3 (pain dimension)	Mean change over study period	The HUI3 pain dimension and EORTC QLQ-C30 significantly respond to changes in QoL/pain over the study period. Responsiveness of one dimension of the HUI3 is good
Non-specific cancer			
Mantovani *et al.*, 2004[111]	EQ-5D	Change over time compared with external measure (up to 4 months)	EQ-5D shows a trend of improvement over time, with slight reduction in utility between months 2 and 4. The improvement at 4 months was statistically significant compared with baseline (p = 0.029). EQ-5D mean (SD): baseline: 0.33 (0.4), 1 month: 0.45 (0.3), 2 month: 0.59 (0.3) and 4 month: 0.54 (0.3). The EORTC QLQ-C30, EQ-VAS and MFSI-SF fatigue showed similar trends in scores over time and were all statistically significant at months 1 and 2, but not at month 4. MFSI-SF vigour showed a small non-statistically significant improvement at all time points
Vaghela *et al.*, 2007[112]	EQ-5D	Change over time compared with external measure (up to 6 weeks)	Statistically significant improvements were found on the first two stated concerns of MYCaW, the overall MYCAW profile and the EQ-VAS but not on the well-being measure. A statistically significant improvement was only seen on the anxiety and depression dimension of EQ-5D
Ravasco *et al.*, 2003[104]	EQ-5D	EQ-5D domains and VAS scores presented before and after radiotherapy	All dimensions improved following radiotherapy (except for pain and discomfort) but this was only statistically significant for high-risk patients (p = 0.004). Pain worsened; however, severe symptoms also worsened (anorexia, diarrhoea, dysphagia, odynophagia). Mobility, usual activities and anxiety/depression were associated with presence of malnutrition and reduced energy intake. The VAS scores showed an increase following radiotherapy in all groups, but this was only statistically significant for high-risk patients (p = 0.001)

Study reference grouped by condition (author, year)	Instrument	Assessment methods	Results
Non-specific cancer			
Weze *et al.*, 2004[110]	EQ-5D	EQ-5D and VAS data presented before and after therapy	Statistically significant improvement on the anxiety/depression dimension (*p* = 0.005) and borderline on the pain dimension (*p* = 0.058). No changes on the other dimensions Mean EQ-VAS score increased by 12.5 (*p* = 0.008). Other improvements in VAS scores that were statistically significant (*p* < 0.05) were: stress, fear, pain, sleep, relaxation and coping. Non-statistically significant VAS scores included: panic, anger, disability and immobility
Kim *et al.*, 2008[113]	EQ-5D	EQ-5D domains (summed) and VAS scores presented before and after mirtazapine	Statistically significant differences found in sum of levels found on pain/discomfort and anxiety/depression dimensions after treatment. No differences were found for mobility or self-care. Usual activities: mean 2.1/2.0, pain/discomfort: mean 2.1/1.9, anxiety/depression: mean 2.3/1.8. Statistically significant differences found on all other outcome measures

BP, bodily pain; FACT-P, Functional Assessment of Cancer Therapy – Prostate Scale; FIQL, Faecal Incontinence QoL; LCR, laparoscopic colon resection; MFSI-SF, Multidimensional Fatigue Symptom Inventory-Short Form; OR, odds ratio; PCS, physical component score; PF, physical functioning; RP, role physical; SRE, skeletal-related events; SRM, standardised response mean; TEM, transanal endoscopic microsurgery; VT, vitality.

Appendix 12 Results from mapping from European Organization for Research and Treatment Quality-of-Life Questionnaire Core 30 to EQ-5D

TABLE 46 Spearman's rank-order correlation coefficients among EORTC QLQ-C30 summary scales for all data

EORTC QLQ-C30 scale	pf	rf	ef	cf	sf	fa	nv	pa	ql	dy	sl	ap	co	di	fi
pf	1														
rf	0.7470**	1													
ef	0.3412**	0.3938**	1												
cf	0.3861**	0.3819**	0.3835**	1											
sf	0.5824**	0.6450**	0.4251**	0.4023**	1										
fa	–0.6839**	–0.7023**	–0.5018**	–0.4971**	–0.5813**	1									
nv	–0.2811**	–0.2758**	–0.2867**	–0.2739**	–0.2683**	0.3319**	1								
pa	–0.6127**	–0.6454**	–0.4177**	–0.3487**	–0.4871**	0.5747**	0.2698**	1							
ql	0.6266**	0.6387**	0.4268**	0.3199**	0.5147**	–0.6237**	–0.2731**	–0.6021**	1						
dy	–0.3396**	–0.2875**	–0.2370**	–0.2550**	–0.2601**	0.4104**	0.2634**	0.1976**	–0.2590**	1					
sl	–0.2082**	–0.2907**	–0.4069**	–0.2848**	–0.2550**	0.3817**	0.1876**	0.2954**	–0.2880**	0.2412**	1				
ap	–0.4350**	–0.4474**	–0.3992**	–0.3526**	–0.4321**	0.5442**	0.4294**	0.4017**	–0.4769**	0.2583**	0.2730**	1			
co	–0.3091**	–0.3042**	–0.2821**	–0.2707**	–0.2780**	0.3218**	0.2353**	0.3091**	–0.3224**	0.1519**	0.2181**	0.3265**	1		
di	–0.0731	–0.0825	–0.1332**	–0.1435**	–0.1044**	0.1041**	0.2191**	0.1017**	–0.0669	0.1570**	0.1109**	0.1665**	0.0255	1	
fi	–0.2785**	–0.2588**	–0.2365**	–0.2459**	–0.3745**	0.2260**	0.0975**	0.2088**	–0.2134**	0.1335**	0.1760**	0.2004**	0.1410**	0.0643	1

ap, appetite loss; cf, cognitive functioning; co, constipation; di, diarrhoea; dy, dyspnoea; ef, emotional functioning; fa, fatigue; fi, financial difficulties; nv, nausea and vomiting; pa, pain; pf, physical functioning; ql, QoL; rf, role functioning; sf, social functioning; sl, sleep disturbance.

** $p < 0.05$.

Correlations > l0.5l are highlighted.

TABLE 47 Spearman's rank-order correlation coefficients among EORTC QLQ-C30 summary scales

Data set	EQ-5D index, dimension or term	pf	rf	ef	cf	sf	fa
All	EQ-5D	0.7001**	0.6875**	0.4862**	0.3935**	0.5649**	–0.6245**
	eq1	–0.6923**	–0.5845**	–0.2344**	–0.3116**	–0.4272**	0.5121**
	eq2	–0.5806**	–0.5403**	–0.2311**	–0.2822**	–0.4281**	0.3795**
	eq3	–0.7086**	–0.7218**	–0.2803**	–0.3369**	–0.5932**	0.6011**
	eq4	–0.4708**	–0.4845**	–0.3066**	–0.2527**	–0.3704**	0.4397**
	eq5	–0.3111**	–0.2963**	–0.6674**	–0.3159**	–0.3213**	0.3747**
	N3	–0.5324**	–0.5391**	–0.3821**	–0.3037**	–0.4581**	0.4729**
Breast	EQ-5D	0.4980**	0.3450**	0.4236**	0.3547**	0.3270**	–0.4447**
	eq1	–0.6431**	–0.4769**	–0.0342	–0.1702	–0.253	0.4321**
	eq2	–0.2564	–0.1135	–0.0847	–0.0992	–0.2008	0.1279
	eq3	–0.5227**	–0.6129**	–0.0626	–0.2992**	–0.5064**	0.5070**
	eq4	–0.3105**	–0.1578	–0.1608	–0.2075	–0.086	0.2031
	eq5	–0.1705	–0.1129	–0.6216**	–0.2497	–0.2789**	0.3022**
	N3	–0.1076	–0.1486	–0.3386**	–0.3262**	–0.22	0.2721**
Lung	EQ-5D	0.5790**	0.6098**	0.3810**	0.3116**	0.5209**	–0.6029**
	eq1	–0.5609**	–0.3125**	–0.0925	–0.1678	–0.3241**	0.3835**
	eq2	–0.4686**	–0.2664**	0.0243	–0.1864	–0.2855**	0.3173**
	eq3	–0.5586**	–0.5831**	–0.2364	–0.1939	–0.5133**	0.5753**
	eq4	–0.2386	–0.4029**	–0.2255	–0.2113	–0.2979**	0.3649**
	eq5	–0.2810**	–0.3519**	–0.5505**	–0.1972	–0.2812**	0.2655**
	N3	–0.2269	–0.3667**	–0.2227	–0.0684	–0.226	0.3057**
Multiple Myeloma	EQ-5D	0.7287**	0.7206**	0.4979**	0.4304**	0.6214**	–0.6472**
	eq1	–0.6890**	–0.6006**	–0.2592**	–0.3648**	–0.4745**	0.5275**
	eq2	–0.6050**	–0.5817**	–0.2490**	–0.3223**	–0.4738**	0.3974**
	eq3	–0.7310**	–0.7344**	–0.2994**	–0.3691**	–0.6232**	0.6028**
	eq4	–0.4946**	–0.5097**	–0.3247**	–0.2723**	–0.4242**	0.4674**
	eq5	–0.3191**	–0.2932**	–0.6776**	–0.3441**	–0.3315**	0.3933**
	N3	–0.5907**	–0.5796**	–0.3932**	–0.3381**	–0.5090**	0.5062**

ap, appetite loss; cf, cognitive functioning; co, constipation; di, diarrhoea; dy, dyspnoea; ef, emotional functioning; eq1, EQ-5D mobility; eq2, EQ-5D self-care; eq3, EQ-5D usual activities; eq4, EQ-5D pain/discomfort; eq5, EQ-5D anxiety/depression; fa, fatigue; fi, financial difficulties; N3, EQ-5D N3 term; nv, nausea and vomiting; pa, pain; pf, physical functioning; ql, QoL; rf, role functioning; sf, social functioning; sl, sleep disturbance.

**$p < 0.05$.

Correlations > |0.5| are highlighted.

nv	pa	ql	dy	sl	ap	co	di	fi
–0.2709**	–0.7348**	0.6687**	–0.2340**	–0.3518**	–0.4326**	–0.3302**	–0.0726	–0.2713**
0.1935**	0.5602**	–0.5436**	0.2419**	0.1920**	0.3037**	0.2402**	0.0471	0.1653**
0.1949**	0.5044**	–0.4716**	0.1296**	0.1560**	0.3006**	0.2274**	0.0598	0.2225**
0.2447**	0.5869**	–0.5883**	0.2170**	0.2081**	0.3723**	0.2956**	0.0423	0.2582**
0.1865**	0.7244**	–0.4790**	0.1815**	0.2983**	0.2762**	0.2664**	0.0597	0.2016**
0.1887**	0.3227**	–0.3991**	0.1758**	0.2831**	0.3408**	0.2388**	0.0876	0.2222**
0.2225**	0.5172**	–0.5310**	0.1326**	0.2628**	0.3844**	0.2920**	0.0705	0.2060**
–0.2945**	–0.6974**	0.4318**	–0.3227**	–0.3391**	–0.3290**	–0.1919	–0.2701**	–0.0658
0.3358**	0.5192**	–0.3258**	0.3859**	0.1322	0.2971**	0.1464	0.3179**	–0.0213
0.2534	0.2285	–0.1047	0.1692	0.0278	0.2273	0.0559	0.0394	0.0515
0.3145**	0.5480**	–0.4500**	0.3725**	0.0818	0.3299**	0.2669**	0.2571**	0.2187
0.2555	0.6696**	–0.2656**	0.2933**	0.2975**	0.0842	0.0176	0.1387	0.0123
0.0677	0.2375	–0.3017**	0.0633	0.1518	0.2932**	0.2301	0.1487	0.0695
0.0597	0.15	–0.1187	0.0538	0.3189**	0.2735**	0.2163	0.1999	0.0621
–0.2733**	–0.6294**	0.5297**	–0.2453	–0.3190**	–0.4577**	–0.2479	0.048	–0.2735**
0.0848	0.3159**	–0.3370**	0.3334**	0.2034	0.2425	0.2191	–0.0549	0.1297
0.1449	0.1538	–0.2484	0.0325	–0.1316	0.2624**	0.1356	–0.0253	–0.0501
0.1884	0.1954	–0.4579**	0.2860**	0.1463	0.4168**	0.2404	0.0927	0.1667
0.1963	0.7295**	–0.2994**	0.1229	0.2849**	0.2675**	0.1911	–0.0071	0.2718**
0.2069	0.2691**	–0.3799**	0.0342	0.2557	0.3276**	0.1185	–0.0257	0.2792**
0.0349	0.2397	–0.1911	–0.0021	0.2256	0.1989	–0.082	–0.0266	0.01
–0.3320**	–0.7282**	0.6833**	–0.2627**	–0.3819**	–0.4598**	–0.3398**	–0.1187**	–0.3630**
0.2420**	0.5389**	–0.5436**	0.2388**	0.2196**	0.3099**	0.2223**	0.0663	0.2485**
0.2454**	0.5139**	–0.4756**	0.1639**	0.2074**	0.3165**	0.2264**	0.1197**	0.3179**
0.2824**	0.6032**	–0.5897**	0.1951**	0.2435**	0.3646**	0.2819**	0.039	0.3171**
0.2122**	0.7095**	–0.5071**	0.1944**	0.3164**	0.3030**	0.2887**	0.0959	0.2574**
0.2331**	0.3157**	–0.3849**	0.2397**	0.3158**	0.3522**	0.2470**	0.1223**	0.2687**
0.2995**	0.5509**	–0.5752**	0.1752**	0.2817**	0.4238**	0.3199**	0.1096**	0.2820**

TABLE 48 European Organization for Research and Treatment Quality-of-Life Questionnaire Core 30 mean observed and predicted EQ-5D values per model and summary model performance: OLS

Summary statistics and model performance tests	*n*	Observed values	OLS 2		OLS 3		OLS 4		OLS 6		OLS 7		OLS 8	
Mean (SD)	771	0.5793 (0.3423)	0.5793 (0.2797)		0.5793 (0.2792)		0.5793 (0.2830)		0.5793 (0.2863)		0.5793 (0.2844)		0.5793 (0.2866)	
Median		0.6910	0.6281		0.6244		0.6451		0.6498		0.6557		0.6502	
Range		–0.5940–1	–0.1846–1.02		–0.1915–1.031		–0.3712–0.9419		–0.4078–0.9713		–0.3670–0.9430		–0.4046–0.9714	
R^2			0.668		0.665		0.684		0.700		0.691		0.701	
Adjusted R^2			0.662		0.662		0.681		0.689		0.683		0.690	
AIC			–286		–294		–340		–338		–330		–339	
BIC			–216		–257		–307		–207		–237		–205	
Ramsey RESET			$F_{3,753}$ = 12.57, p = 0.000		$F_{3,761}$ = 13.09, p = 0.000		$F_{3,761}$ = 1.56, p = 0.198		$F_{3,737}$ = 1.00, p = 0.3945		$F_{3,736}$ = 0.58, p = 0.6310		$F_{3,736}$ = 0.88, p = 0.449	
MAE			0.149		0.151		0.143		0.139		0.142		0.139	
Shrinkage			0.836		0.996		0.997		1.060		1.072		1.042	
Health status														
(EORTC QLQ-C30 item 29)	**n**	***Mean***	***Mean***	***MAE***	***Mean***	***MAE***	***Mean***	***MAE***	***Mean***	***MAE***	***Mean***	***MAE***	***Mean***	***MAE***
1 (very poor)	42	–0.0057	0.1213	0.221	0.1212	0.224	0.0638	0.201	0.0636	0.206	0.0685	0.210	0.0642	0.205
2	53	0.1763	0.2571	0.193	0.2631	0.194	0.2569	0.191	0.2471	0.181	0.2590	0.194	0.2470	0.179
3	144	0.4286	0.4403	0.193	0.4410	0.195	0.4577	0.189	0.4650	0.181	0.4684	0.183	0.4629	0.182
4	226	0.6220	0.5685	0.154	0.5661	0.155	0.5839	0.145	0.5829	0.137	0.5794	0.141	0.5823	0.138
5	186	0.7180	0.7145	0.112	0.7158	0.113	0.7179	0.108	0.7170	0.109	0.7147	0.109	0.7176	0.109
6	94	0.8321	0.8494	0.110	0.8470	0.109	0.8165	0.102	0.8145	0.100	0.8148	0.103	0.8181	0.098
7 (excellent)	26	0.9029	0.8958	0.073	0.9008	0.075	0.8538	0.086	0.8553	0.081	0.8504	0.080	0.8546	0.080
ANOVA		F_6 = 97, p = 0.000	F_6 = 126, p = 0.000		F_6 = 126, p = 0.000		F_6 = 118, p = 0.000		F_6 = 112, p = 0.000		F_6 = 109, p = 0.000		F_6 = 114, p = 0.000	

TABLE 49 Best-fitting EORTC QLQ-C30 OLS model

			OLS model 8
Domain	Item	Item level	Regression coefficient (SE)
Physical functioning	Trouble strenuous activities	Not at all (base)	$\chi^2_3 = 6.77$, $p = 0.080$
		A little	–0.0460** (0.018)
		Quite a bit	–0.0375* (0.021)
		Very much	–0.0326 (0.029)
	Short walk	Not at all (base)	$\chi^2_3 = 22.92$, $p = 0.000$
		A little	–0.0551*** (0.021)
		Quite a bit	–0.0975*** (0.033)
		Very much	–0.2160*** (0.047)
	Need help eating/dressing	Not at all (base)	$\chi^2_3 = 42.39$, $p = 0.000$
		A little	–0.1199*** (0.027)
		Quite a bit	–0.2516*** (0.051)
		Very much	–0.3118*** (0.069)
Role functioning	Limited work/housework	Not at all (base)	$\chi^2_3 = 21.34$, $p = 0.000$
		A little	–0.0245 (0.017)
		Quite a bit	–0.0938*** (0.027)
		Very much	–0.1546*** (0.037)
Emotional functioning	Irritable	Not at all (base)	$\chi^2_3 = 9.47$, $p = 0.024$
		A little	–0.0442*** (0.016)
		Quite a bit	–0.0416 (0.030)
		Very much	–0.1086* (0.062)
	Depressed	Not at all (base)	$\chi^2_3 = 22.03$, $p = 0.000$
		A little	–0.0517*** (0.016)
		Quite a bit	–0.0839*** (0.029)
		Very much	–0.1601*** (0.046)
Social functioning		Not at all (base)	$\chi^2_3 = 7.74$, $p = 0.052$
		A little	–0.0317* (0.017)
		Quite a bit	–0.0140 (0.025)
		Very much	–0.0765** (0.034)
Pain	Pain	Not at all (base)	$\chi^2_3 = 86.11$, $p = 0.000$
		A little	–0.0574*** (0.016)
		Quite a bit	–0.1473*** (0.022)
		Very much	–0.2958*** (0.035)
Constipation	Constipation	Not at all (base)	$\chi^2_3 = 8.85$, $p = 0.031$
		A little	–0.0150 (0.016)
		Quite a bit	–0.0753*** (0.028)
		Very much	0.0244 (0.038)

continued

TABLE 49 Best-fitting EORTC QLQ-C30 OLS model (*continued*)

Domain	Item	Item level	OLS model 8
			Regression coefficient (SE)
Age (years)			$\chi^2_1 = 3.98$, $p = 0.046$
			–0.0014** (0.001)
Constant			1.0458*** (0.048)
Observations			771
R^2			0.701
Adjusted R^2			0.690
MAE			0.139
AIC			–339
BIC			–205
Ramsey RESET			$F_{3,736} = 0.88$, $p = 0.449$

* Statistically significant at the 10% level.
** Statistically significant at the 5% level.
*** Statistically significant at the 1% level.

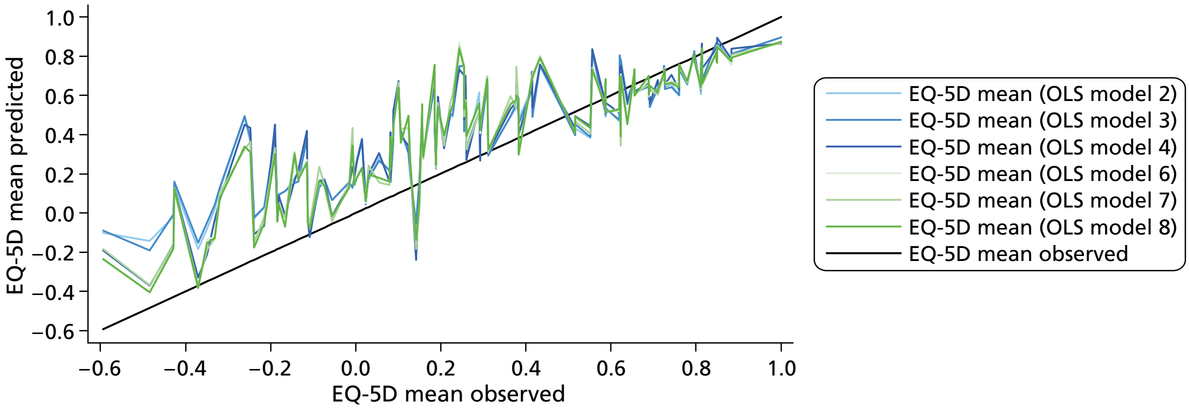

FIGURE 12 Summary of performance of all OLS models.

TABLE 50 European Organization for Research and Treatment Quality-of-Life Questionnaire Core 30 mean observed and predicted EQ-5D values per model and summary model performance: tobit

Summary statistics and model performance tests	*n*	Observed	Tobit model 2		Tobit model 2		Tobit model 2		Tobit model 2		Tobit model 2		Tobit model 2	
			All dimensions		Significant dimensions		Significant and squared terms		Significant items		Significant items collapsed		Significant items + age	
Mean (SD)	771	0.5793 (0.3423)	0.5769 (0.2804)		0.5768 (0.2797)		0.5786 (0.2850)		0.5792 (0.2887)		0.5790 (0.2866)		0.5792 (0.2891)	
Median		0.6910	0.6419		0.6393		0.6570		0.6551		0.6581		0.6517	
Range		–0.5940–1	–0.2201–0.94		–0.2297–0.948		–0.3574–0.9143		–0.3971–0.9408		–0.3622–0.9286		–0.3937–0.9463	
Pseudo R^2			1.024		1.019		1.044		1.094		1.066		1.101	
Log-likelihood			9.83		7.53		17.92		63.57		26.53		40.66	
AIC			12		3		–19		–18		–11		–21	
BIC			87		45		17		117		87		118	
MAE			0.147		0.148		0.143		0.139		0.142		0.139	
Sigma			0.214		0.215		0.210		0.197		0.208		0.204	
Shrinkage			0.921		0.999		0.989		1.04		1.06		1.02	
Health status (EORTC QLQ-C30 item 29)	**n**	***Mean***	***Mean***	***MAE***	***Mean***	***MAE***	***Mean***	***MAE***	***Mean***	***MAE***	***Mean***	***MAE***	***Mean***	***MAE***
1 (very poor)	42	–0.0057	0.1021	0.211	0.1014	0.213	0.0578	0.198	0.0631	0.204	0.0676	0.209	0.0638	0.203
2	53	0.1763	0.2476	0.188	0.2534	0.189	0.2498	0.188	0.2433	0.179	0.2551	0.193	0.2433	0.177
3	144	0.4286	0.4418	0.191	0.4423	0.193	0.4555	0.189	0.4630	0.181	0.4664	0.183	0.4602	0.183
4	226	0.6220	0.5737	0.152	0.5715	0.153	0.5851	0.146	0.5821	0.138	0.5784	0.141	0.5816	0.139
5	186	0.7180	0.7163	0.108	0.7172	0.110	0.7201	0.107	0.7198	0.108	0.7174	0.108	0.7205	0.109
6	94	0.8321	0.8327	0.110	0.8304	0.110	0.8154	0.105	0.8158	0.101	0.8154	0.103	0.8195	0.099
7 (excellent)	26	0.9029	0.8693	0.083	0.8718	0.085	0.8491	0.091	0.8548	0.082	0.8498	0.083	0.8546	0.081
ANOVA		$F_6 = 97$, $p = 0.000$	$F_6 = 124$, $p = 0.000$		$F_6 = 123$, $p = 0.000$		$F_6 = 119$, $p = 0.000$		$F_6 = 112$, $p = 0.000$		$F_6 = 109$, $p = 0.000$		$F_6 = 113$, $p = 0.000$	

TABLE 51 European Organization for Research and Treatment Quality-of-Life Questionnaire Core 30 best-fitting tobit model

			Tobit model 8	
Domain	Item	Item level	Regression coefficient (SE)	
Physical	Trouble strenuous activities	Not at all (base)	$\chi^2_3 = 8.22$	$p = 0.042$
		A little	–0.0675***	(0.024)
		Quite a bit	–0.0581**	(0.026)
		Very much	–0.0549*	(0.033)
	Short walk	Not at all (base)	$\chi^2_3 = 23.65$	$p = 0.000$
		A little	–0.0607***	(0.021)
		Quite a bit	–0.0974***	(0.034)
		Very much	–0.2215***	(0.048)
	Need help eating/dressing	Not at all (base)	$\chi^2_3 = 42.40$	$p = 0.000$
		A little	–0.1158***	(0.027)
		Quite a bit	–0.2477***	(0.051)
		Very much	–0.3096***	(0.069)
Role	Limited work/housework	Not at all (base)	$\chi^2_3 = 22.72$	$p = 0.000$
		A little	–0.0344*	(0.021)
		Quite a bit	–0.1048***	(0.028)
		Very much	–0.1649***	(0.038)
Emotional	Irritable	Not at all (base)	$\chi^2_3 = 10.46$	$p = 0.012$
		A little	–0.0519***	(0.018)
		Quite a bit	–0.0481	(0.032)
		Very much	–0.1212*	(0.067)
	Depressed	Not at all (base)	$\chi^2_3 = 23.28$	$p = 0.000$
		A little	–0.0629***	(0.017)
		Quite a bit	–0.0921***	(0.030)
		Very much	–0.1683***	(0.047)
Social functioning	Interfered social activities	Not at all (base)	$\chi^2_3 = 9.71$	$p = 0.021$
		A little	–0.0435**	(0.018)
		Quite a bit	–0.0211	(0.026)
		Very much	–0.0849**	(0.035)
Pain	Pain	Not at all (base)	$\chi^2_3 = 101.35$	$p = 0.000$
		A little	–0.0896***	(0.020)
		Quite a bit	–0.1788***	(0.025)
		Very much	–0.3252***	(0.035)

TABLE 51 European Organization for Research and Treatment Quality-of-Life Questionnaire Core 30 best-fitting tobit model (*continued*)

			Tobit model 8
Domain	Item	Item level	Regression coefficient (SE)
Constipation	Been constipated	Not at all (base)	$\chi^2_3 = 8.59$, $p = 0.035$
		A little	–0.0165 (0.018)
		Quite a bit	–0.0785*** (0.029)
		Very much	0.0209 (0.039)
	Age (years)		$\chi^2_1 = 5.50$, $p = 0.019$
			–0.0020** (0.001)
Constant			1.1677*** (0.061)
Observations			771
Sigma			0.204
Pseudo R^2			1.101
MAE			0.139
AIC			–21
BIC			118

* Statistically significant at the 10% level.
** Statistically significant at the 5% level.
*** Statistically significant at the 1% level.
Linear predictions using the above predictions need to be adjusted to take into account upper and lower limits.

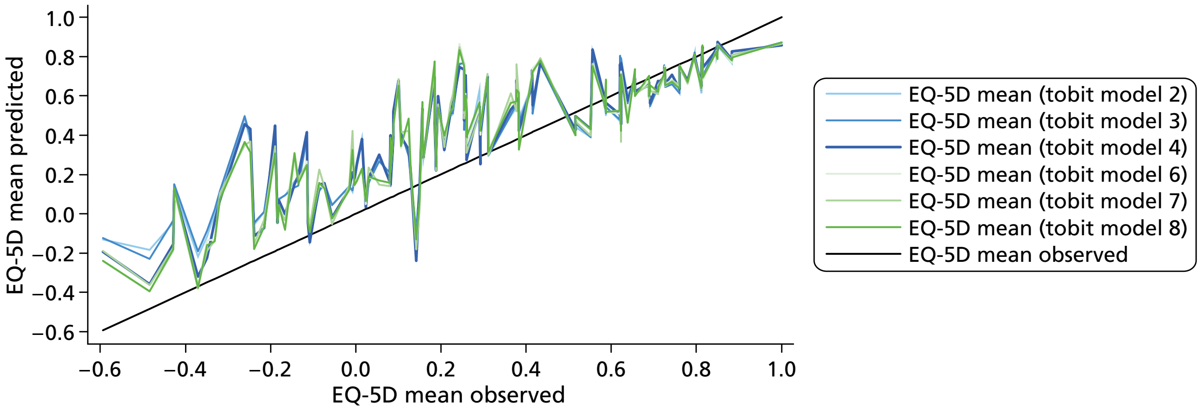

FIGURE 13 Summary of performance of all tobit models.

TABLE 52 European Organization for Research and Treatment Quality-of-Life Questionnaire Core 30 mean observed and predicted EQ-5D values per model and summary model performance: TPM (*continued*)

			TPM model 2		TPM model 3		TPM model 4		TPM model 6		TPM model 7		TPM model 8	
Summary statistics and model performance tests	*n*	Observed values	All dimensions		Significant dimensions		Significant and squared terms		Collapsed items		Significant collapsed items		Significant collapsed items + age	
Mean (SD)	771	0.5793 (0.3423)	0.6102 (0.3025)		0.6108 (0.3021)		0.6066 (0.2994)		0.6073 (0.2987)		0.6074 (0.2985)		0.6066 (0.2997)	
Median		0.6910	0.6788		0.6851		0.6797		0.6897		0.6892		0.6892	
Range		–0.5940–1	–0.2463–0.9896		–0.2636–0.9889		–0.3997–0.9731		–0.3915–0.9747		–0.4109–0.9744		–0.3936–0.9898	
			Part 1	***Part 2***	***Part 1***	***Part 2***	***Part 1***	***Part 2***	***Part 1***	***Part 2***	***Part 1***	***Part 2***	***Part 1***	***Part 2***
Pseudo R^2			0.437		0.411		0.411		0.376		0.376		0.406	
AIC			334	–369	326	–379	326	–384	351	–390	351	–390	336	
BIC			403	–294	344	–341	344	–356	383	–283	383	–287	373	
Model goodness of fit			$\chi^2_{753} = 421$, $p = 1.000$		$\chi^2_{422} = 237$, $p = 1.000$		$\chi^2_{422} = 237$, $p = 1.000$		$\chi^2_{51} = 79$, $p = 0.008$		$\chi^2_{51} = 79$, $p = 0.008$		$\chi^2_{429} = 519$, $p = 0.002$	
Log likelihood			–151.85	200.25	–158.76	197.27	–158.76	198.17	–168.38	217.86	–168.38	216.83	–160.14	217.86
Sigma				0.232		0.234		0.226		0.216		0.217		0.216
MAE			0.147		0.150		0.146		0.140		0.140		0.140	
Shrinkage			0.920		0.923		0.936		0.942		0.943		0.940	

Summary statistics and model performance tests	*n*	Observed values	TPM model 2 All dimensions		TPM model 3 Significant dimensions		TPM model 4 Significant and squared terms		TPM model 6 Collapsed items		TPM model 7 Significant collapsed items		TPM model 8 Significant collapsed items + age	
Health status (EORTC QLQ-C30 item 29)	**n**	***Mean***	***Mean***	***MAE***	***Mean***	***MAE***	***Mean***	***MAE***	***Mean***	***MAE***	***Mean***	***MAE***	***Mean***	***MAE***
1 (very poor)	42	–0.0057	0.0996	0.203	0.0992	0.207	0.0603	0.207	0.0675	0.196	0.0663	0.200	0.0649	0.195
2	53	0.1763	0.2562	0.186	0.2652	0.196	0.2666	0.202	0.2689	0.186	0.2649	0.185	0.2670	0.185
3	144	0.4286	0.4637	0.192	0.4655	0.195	0.4824	0.193	0.4835	0.186	0.4838	0.186	0.4808	0.184
4	226	0.6220	0.6071	0.152	0.6069	0.153	0.6103	0.145	0.6102	0.141	0.6111	0.140	0.6091	0.141
5	186	0.7180	0.7605	0.112	0.7607	0.114	0.7514	0.113	0.7565	0.107	0.7567	0.107	0.7566	0.107
6	94	0.8321	0.8855	0.116	0.8842	0.115	0.8577	0.102	0.8473	0.106	0.8475	0.106	0.8511	0.104
7 (excellent)	26	0.9029	0.9234	0.065	0.9203	0.067	0.8928	0.072	0.8934	0.061	0.8939	0.061	0.8925	0.060
ANOVA		$F_6 = 97$, $p = 0.000$	$F_6 = 123$, $p = 0.000$		$F_6 = 120$, $p = 0.000$		$F_6 = 116$, $p = 0.000$		$F_6 = 112$, $p = 0.000$		$F_6 = 114$, $p = 0.000$		$F_6 = 114$, $p = 0.000$	

TABLE 53 European Organization for Research and Treatment Quality-of-Life Questionnaire Core 30 TPM best-fitting model

			TPM model 8 part 1	TPM model 8 part 2
Domain	Item	Item level	Regression coefficient (SE)	Regression coefficient (SE)
			Logistic	Truncated regression
Physical	Strenuous activity	Not at all (base)	–0.9099*** (0.326)	
		A little		
		Quite a bit		
		Very much		
	Short walk	Not at all (base)		χ^2_3 = 31.53, *p* = 0.000
		A little		–0.0739*** (0.025)
		Quite a bit		–0.1299*** (0.036)
		Very much		–0.2585*** (0.048)
	Stay in bed/chair	Not at all (base)	–0.6943*** (0.392)	
		A little		
		Quite a bit		
		Very much		
	Need help eating/dressing	Not at all (base)		χ^2_3 = 44.11, *p* = 0.000
		A little		–0.0315 (0.026)
		Quite a bit		–0.1194*** (0.033)
		Very much		–0.1833*** (0.043)
Role	Limited work	Not at all (base)	–0.7200* (0.380)	χ^2_3 = 22.28, *p* = 0.000
		A little		–0.0315 (0.026)
		Quite a bit		–0.1194*** (0.033)
		Very much		–0.1833*** (0.043)
	Depressed	Not at all (base)	–1.5256*** (0.374)	χ^2_3 = 19.04, *p* = 0.000
		A little		–0.0671*** (0.022)
		Quite a bit		–0.0928*** (0.032)
		Very much		–0.1750*** (0.050)
	Interfered social activities	Not at all (base)	–1.0215*** (0.380)	
		A little		
		Quite a bit		
		Very much		
Pain	Pain	Not at all (base)	–1.9394*** (0.333)	χ^2_3 = 63.60, *p* = 0.000
		A little		–0.0429 (0.029)
		Quite a bit		–0.1405*** (0.034)
		Very much		–0.2933*** (0.042)

TABLE 53 European Organization for Research and Treatment Quality-of-Life Questionnaire Core 30 TPM best-fitting model (*continued*)

			TPM model 8 part 1	TPM model 8 part 2
Domain	Item	Item level	Regression coefficient (SE)	Regression coefficient (SE)
Sleep disturbance	Trouble sleeping	Not at all (base)		χ^2_3 = 10.98, *p* = 0.0012
		A little		–0.0545** (0.023)
		Quite a bit		–0.0628** (0.029)
		Very much		–0.1082*** (0.038)
Appetite loss	Lacked appetite	Not at all (base)		χ^2_3 = 9.36, *p* = 0.025
		A little		0.0089 (0.023)
		Quite a bit		–0.0812*** (0.031)
		Very much		–0.0285 (0.043)
Age (years)			–0.0549*** (0.014)	
Constant			4.8139*** (0.992)	0.9625*** (0.031)
Observations			771	685
Log-likelihood			–160.14	217.86
Pseudo R^2			0.406	
Sigma				0.217
MAE			0.140	
AIC			336	–389
BIC			373	–283

* Statistically significant at the 10% level.
** Statistically significant at the 5% level.
*** Statistically significant at the 1% level.

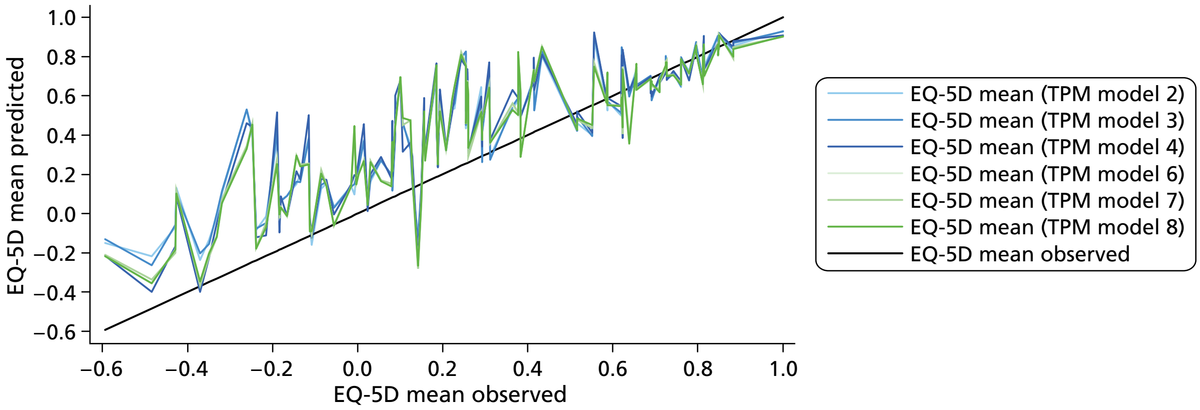

FIGURE 14 Summary of performance of all TPM models.

TABLE 54 European Organization for Research and Treatment Quality-of-Life Questionnaire Core 30 mean observed and predicted EQ-5D values per model and summary model performance: splining

			SPL model 3	
Summary statistics and model performance tests	***n***	**Observed values**	**Significant dimensions**	
Mean (SD)	771	0.5793 (0.3423)	0.5793 (0.2833)	
Median		0.6910	0.6457	
Range		–0.5940 to 1	–0.3718 to 0.9438	
R^2			0.685	
AIC			–343	
BIC			–310	
MAE			0.143	
Shrinkage			0.997	
Ramsey RESET			$F_{3,761}$ = 1.17, p = 0.321	
Health status (EORTC QLQ-C30 item 29)	**n**	***Mean***	***Mean***	***MAE***
1 (very poor)	42	–0.0057	0.0660	0.245
2	53	0.1763	0.3345	0.236
3	144	0.4286	0.5166	0.142
4	226	0.6220	0.5694	0.143
5	186	0.7180	0.7353	0.084
6	94	0.8321	0.8151	0.072
7 (excellent)	26	0.9029	0.8660	0.134
ANOVA			F_6 = 117, p = 0.000	

SPL, splining.

TABLE 55 European Organization for Research and Treatment Quality-of-Life Questionnaire Core 30 best-fitting OLS dimension model with splines

	SPL model 3
Domain	Regression coefficient (SE)
Physical functioning 1	0.1197*** (0.013)
Physical functioning 2	0.0528*** (0.007)
Role functioning	0.0012*** (0.000)
Emotional functioning	0.0020*** (0.000)
Pain	–0.0035*** (0.000)
Sleep disturbance	–0.0007** (0.000)
Constant	0.5339*** (0.044)
Observations	771
Pseudo R^2	0.685

SPL, splining.

** Statistically significant at the 5% level.

*** Statistically significant at the 1% level.

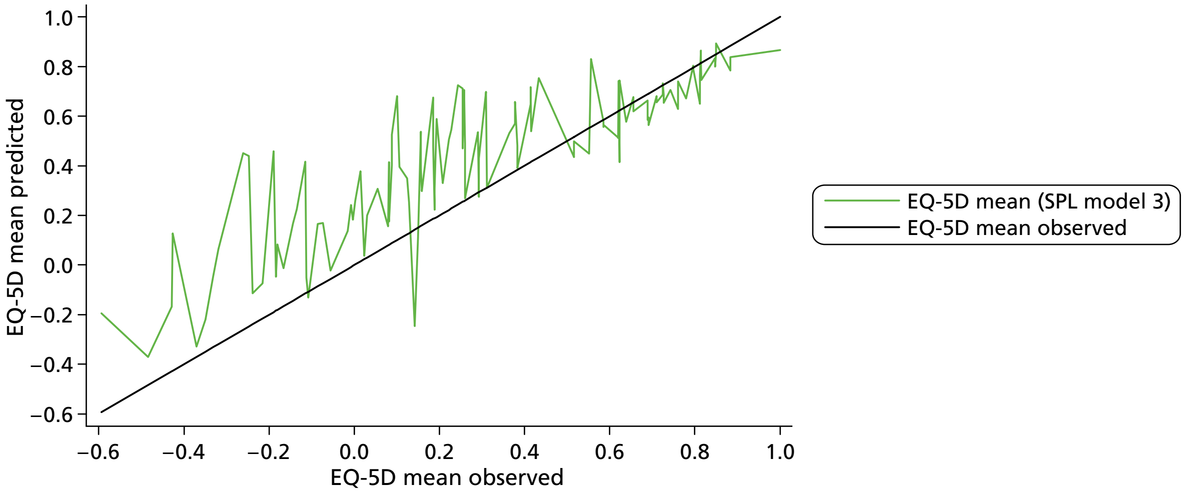

FIGURE 15 Summary of performance of splining model. SPL, splining.

TABLE 56 European Organization for Research and Treatment Quality-of-life Questionnaire Core 30 mean observed and predicted EQ-5D values per model and summary model performance: response mapping

Summary statistics and model performance tests	*n*	Observed values	Response mapping 2		Response mapping 3		Response mapping 4		Response mapping 6		Response mapping 8	
			All dimensions		Significant dimensions		Significant and squared terms		Significant collapsed items		Significant collapsed items + age/gender	
Mean (SD)		0.5793 (0.3423)	0.5726 (0.2913)		0.5715 (0.2891)		0.5724 (0.2883)		0.5363 (0.2734)		0.5726 (0.2914)	
Median		0.6910	0.6605		0.6552		0.6518		0.6066		0.6569	
Range		–0.5940–1	–0.3420–0.9405		–0.3371–0.9406		–0.3846–0.9516		–0.1170–0.9332		–0.3376–0.9416	
MAE			0.134		0.138		0.138		0.192		0.134	
Shrinkage			1.065		0.997		0.998		0.962		1.179	
Health status (EORTC QLQ-C30 item 29)	**n**	***Mean***	***Mean***	***MAE***	***Mean***	***MAE***	***Mean***	***MAE***	***Mean***	***MAE***	***Mean***	***MAE***
1 (very poor)	42	–0.0057	0.0473	0.183	0.0408	0.196	0.0370	0.201	0.1833	0.291	0.0473	0.181
2	53	0.1763	0.2314	0.162	0.2295	0.177	0.2346	0.176	0.3026	0.241	0.2262	0.159
3	144	0.4286	0.4506	0.181	0.4513	0.183	0.4548	0.182	0.4296	0.239	0.4515	0.182
4	226	0.6220	0.5836	0.139	0.5801	0.144	0.5785	0.144	0.5193	0.187	0.5827	0.139
5	186	0.7180	0.7088	0.097	0.7106	0.102	0.7106	0.104	0.6488	0.162	0.7094	0.097
6	94	0.8321	0.8120	0.102	0.8094	0.099	0.8121	0.097	0.7313	0.149	0.8137	0.100
7 (excellent)	26	0.9029	0.8585	0.077	0.8621	0.075	0.8674	0.068	0.8122	0.097	0.8596	0.075
ANOVA		$F_6 = 97, p = 0.000$	$F_6 = 114, p = 0.000$		$F_6 = 120, p = 0.000$		$F_6 = 122, p = 0.000$		$F_6 = 57, p = 0.000$		$F_6 = 116, p = 0.000$	

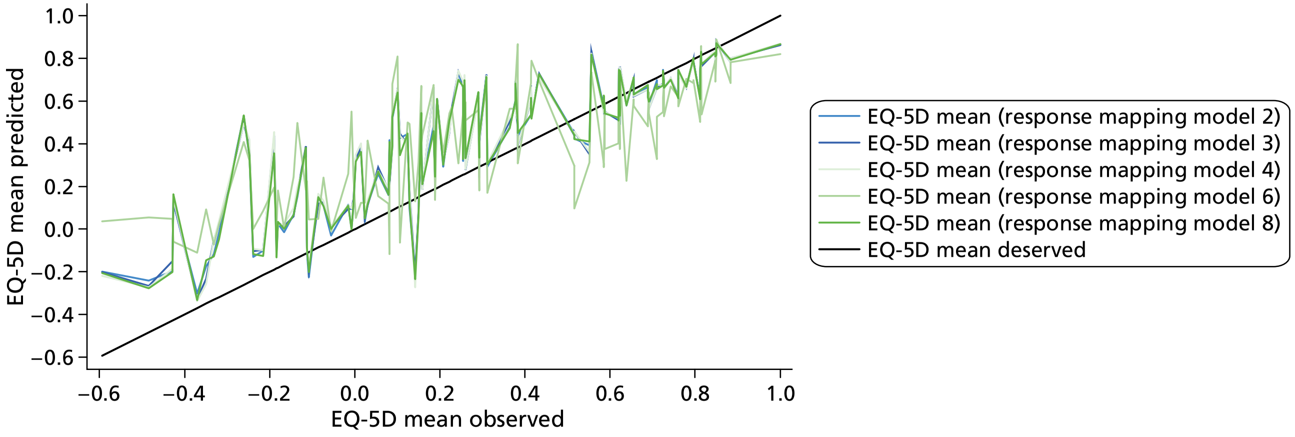

FIGURE 16 Summary of performance of all response mapping models.

Appendix 13 Results from mapping from Functional Assessment of Cancer Therapy – General Scale to EQ-5D

TABLE 57 Spearman's rank correlation coefficients among the FACT-G summary scales

FACT-G summary scale	Physical	Social/family	Emotional	Functional
Physical	1			
Social/family	0.185	1		
Emotional	0.378	0.321	1	
Functional	0.570	0.290	0.442	1

Correlations > l0.5l are highlighted.

TABLE 58 Spearman's rank correlation coefficients between EQ-5D and FACT-G summary scales and total score

EQ-5D index and dimensions	Physical	Social/family	Emotional	Functional	Total
EQ-5D	0.566	0.178	0.382	0.501	0.575
eq1	–0.383	–0.083	–0.172	–0.341	–0.353
eq2	–0.323	–0.085	–0.118	–0.303	–0.300
eq3	–0.504	–0.128	–0.214	–0.504	–0.487
eq4	–0.460	–0.116	–0.227	–0.304	–0.396
eq5	–0.309	–0.245	–0.560	–0.349	–0.493
n3	–0.310	–0.067	–0.198	–0.297	–0.310

Correlations > l0.5l are highlighted.

TABLE 59 Summary of observed and predicted values per model: OLS

Summary statistics and model performance tests	Observed values	OLS model 1	OLS model 2		OLS model 3		OLS model 4		OLS model 5		OLS model 6		OLS model 7	
		Total score	Domain scores		OLS significant domains		OLS significant domains and squared terms		OLS significant domains, squared and interaction terms		OLS item levels: significant levels only		OLS item levels: significant levels only, collapse unordered items	
Mean (SD)	0.721 (0.223)	0.721 (0.128)	0.721 (0.138)		0.721 (0.138)		0.721 (0.144)		0.721 (0.146)		0.721 (0.163)		0.721 (0.161)	
Median	0.735	0.730	0.735		0.735		0.738		0.744		0.755		0.750	
Range	–0.135–1	0.319–0.975	0.357–0.971		0.357–0.972		0.198–0.981		0.161–0.946		0.115–0.962		0.169–0.961	
R^2		0.331	0.383		0.383		0.417		0.432		0.535		0.524	
Adjusted R^2		0.330	0.378		0.379		0.413		0.425		0.513		0.507	
AIC		–298.40	–335.20		–337.12		–365.34		–374.98		–445.43		–443.38	
BIC		–289.86	–313.84		–320.11		–343.97		–345.07		–338.60		–357.92	
Ramsey RESET		$F_{3,525} = 3.19$, $p = 0.024$	$F_{3,522} = 0.83$, $p = 0.477$		$F_{3,525}$=0.84, $p = 0.471$		$F_{3,524} = 2.96$, $p = 0.032$		$F_{3,521} = 2.06$, $p = 0.104$		$F_{3,502} = 0.72$, $p = 0.539$		$F_{3,507} = 1.17$, $p = 0.320$	
MAE		0.129	0.126		0.126		0.124		0.122		0.111		0.112	
Shrinkage		1.005	0.992		0.996		0.995		0.991		0.850		0.909	
	n	Mean	Mean	MAE	Mean	MAE	Mean	MAE	Mean	MAE	Mean	MAE	Mean	MAE
ECOG														
Normal, no symptoms	122	0.8645	0.8156	0.1113	0.8339	0.0958	0.8339	0.0958	0.8429	0.0966	0.8404	0.0973	0.8464	0.0868	0.8464	0.0870
Some symptoms	256	0.7219	0.7280	0.1220	0.7325	0.1237	0.7325	0.1236	0.7263	0.1227	0.7281	0.1214	0.7318	0.1080	0.7319	0.1087
Require some bed	152	0.6055	0.6344	0.1568	0.6121	0.1540	0.6121	0.1540	0.6154	0.1465	0.6143	0.1451	0.6033	0.1353	0.6032	0.1343
ANOVA		$F_{2,527} = 55$, $p < 0.001$	$F_{2,527} = 92$, $p < 0.001$		$F_{2,527} = 134$, $p < 0.001$		$F_{2,527} = 135$, $p < 0.001$		$F_{2,527} = 125$, $p < 0.001$		$F_{2,527} = 117$, $p < 0.001$		$F_{2,527} = 107$, $p < 0.001$		$F_{2,527} = 108$, $p < 0.001$	

TABLE 60 Model coefficients for best performing OLS model (model 6)

			OLS model 6
Domain	Item	Item level	Regression coefficient (SE)
Physical	Lack of energy	Very much (baseline level)	$F_{4,505}$ = 3.62, p = 0.007
		Quite a bit	0.045 (0.032)
		Somewhat	0.036 (0.030)
		A little bit	0.071 (0.033)*
		Not at all	0.118 (0.033)***
	Trouble meeting need of family	Very much (baseline level)	$F_{4,505}$ = 2.75, p = 0.028
		Quite a bit	–0.028 (0.056)
		Somewhat	0.049 (0.050)
		A little bit	0.088 (0.050)*
		Not at all	0.098 (0.050)*
	Pain	Very much (baseline level)	$F_{4,505}$ = 29.09, $p < 0.001$
		Quite a bit	0.125 (0.073)*
		Somewhat	0.219 (0.069)**
		A little bit	0.240 (0.071)**
		Not at all	0.342 (0.070)***
Emotional	I feel sad	Very much (baseline level)	$F_{4,505}$ = 2.45, p = 0.045
		Quite a bit	–0.085 (0.105)
		Somewhat	–0.019 (0.101)
		A little bit	–0.006 (0.099)
		Not at all	–0.004 (0.099)
	Losing hope	Very much (baseline level)	$F_{4,505}$ = 3.68, p = 0.006
		Quite a bit	–0.081 (0.122)
		Somewhat	–0.007 (0.079)
		A little bit	0.013 (0.076)
		Not at all	0.060 (0.075)
Functional	Able to work	Not at all (baseline level)	$F_{4,505}$ = 10.22, $p < 0.001$
		A little bit	0.113 (0.031)***
		Somewhat	0.130 (0.028)***
		Quite a bit	0.150 (0.028)***
		Very much	0.152 (0.030)***
Constant			–0.597 (0.0141)***

* Statistically significant at the 10% level.
** Statistically significant at the 5% level.
*** Statistically significant at the 1% level.

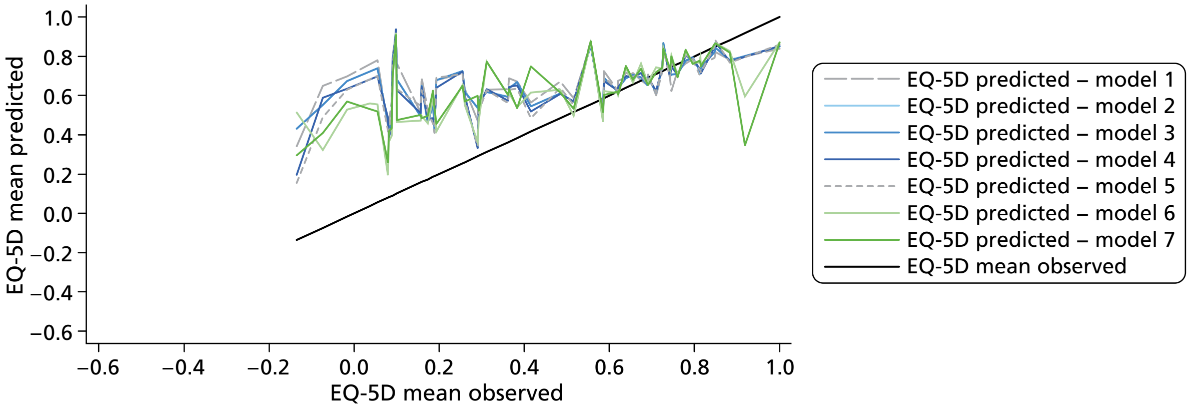

FIGURE 17 Summary of performance of all OLS models.

TABLE 61 Summary of observed and predicted values per model: tobit models

		Tobit model 1	Tobit model 2	Tobit model 3	Tobit model 4	Tobit model 5	Tobit model 6	Tobit model 8
Summary statistics and model performance tests	Observed values	Total score	Domain scores	Significant domains	Significant domains and squared terms	Significant domains, squared and interaction terms	Item levels: significant items only	Item levels: significant items only and significant patient characteristics
Mean (SD)	0.721 (0.223)	0.723 (0.133)	0.724 (0.143)	0.724 (0.143)	0.723 (0.147)	0.723 (0.151)	0.723 (0.161)	0.723 (0.159)
Median	0.735	0.743	0.750	0.750	0.736	0.739	0.738	0.735
Range	–0.135–1	0.264–0.939	0.322–0.939	0.322–0.939	0.201–0.953	0.191–0.992	0.132–0.957	0.188–0.963
Pseudo R^2		0.826	0.976	0.976	1.093	1.178	1.367	1.338
Log-likelihood		–23.06	–3.18	–3.18	12.28	23.57	48.71	44.97
AIC		52.12	18.35	16.35	–12.82	–27.14	–61.42	–55.94
BIC		64.93	43.99	37.72	17.09	15.58	15.49	16.70
MAE		0.130	0.127	0.127	0.124	0.122	0.113	0.116
Sigma		0.211	0.202	0.202	0.196	0.192	0.181	0.182
Shrinkage		0.965	0.948	0.952	0.967	0.958	0.962	0.953

	n	Mean	Mean	MAE	Mean	MAE	Mean	MAE	Mean	MAE	Mean	MAE	Mean	MAE	Mean	MAE
ECOG																
Normal, no symptoms	122	0.8645	0.8167	0.119	0.8345	0.0976	0.8345	0.0976	0.8466	0.0953	0.8494	0.0930	0.8498	0.0878	0.8713	0.0890
Some symptoms	256	0.7219	0.7331	0.1210	0.7385	0.1233	0.7385	0.1233	0.7218	0.1226	0.7271	0.1229	0.7320	0.1108	0.7224	0.1137
Require some bed	152	0.6055	0.6325	0.1598	0.6100	0.1566	0.6100	0.1567	0.6155	0.1479	0.6155	0.1454	0.6074	0.1365	0.6057	0.1409
ANOVA		$F_{2,527} = 55$, $p < 0.001$	$F_{2,527} = 87$, $p < 0.001$		$F_{2,527} = 126$, $p < 0.001$		$F_{2,527} = 126$, $p < 0.001$		$F_{2,527} = 122$, $p < 0.001$		$F_{2,527} = 116$, $p < 0.001$		$F_{2,527} = 109$, $p < 0.001$		$F_{2,527} = 146$, $p < 0.001$	

TABLE 62 Coefficients for best performing tobit model (model 6)[a]

			Tobit model 6
Domain	Item	Item level	Regression coefficient (SE)
Physical	Lack of energy	Very much (baseline level)	
		Quite a bit	0.055 (0.034)
		Somewhat	0.053 (0.033)
		A little bit	0.113 (0.037)**
		Not at all	0.200 (0.044)***
	Pain	Very much (baseline level)	
		Quite a bit	0.164 (0.075)*
		Somewhat	0.255 (0.070)***
		A little bit	0.293 (0.071)***
		Not at all	0.431 (0.072)***
Functional	Able to work	Not at all (baseline level)	
		A little bit	0.097 (0.033)**
		Somewhat	0.110 (0.031)***
		Quite a bit	0.149 (0.032)***
		Very much	0.151 (0.036)***
	Enjoy life	Not at all (baseline level)	
		A little bit	–0.098 (0.092)**
		Somewhat	–0.012 (0.088)*
		Quite a bit	–0.010 (0.087)
		Very much	–0.057 (0.088)
Constant			0.231 (0.115)*
Sigma			0.181 (0.009)

[a] A Stata programme (do) file is available from the authors on request.
* Statistically significant at the 10% level.
** Statistically significant at the 5% level.
*** Statistically significant at the 1% level.

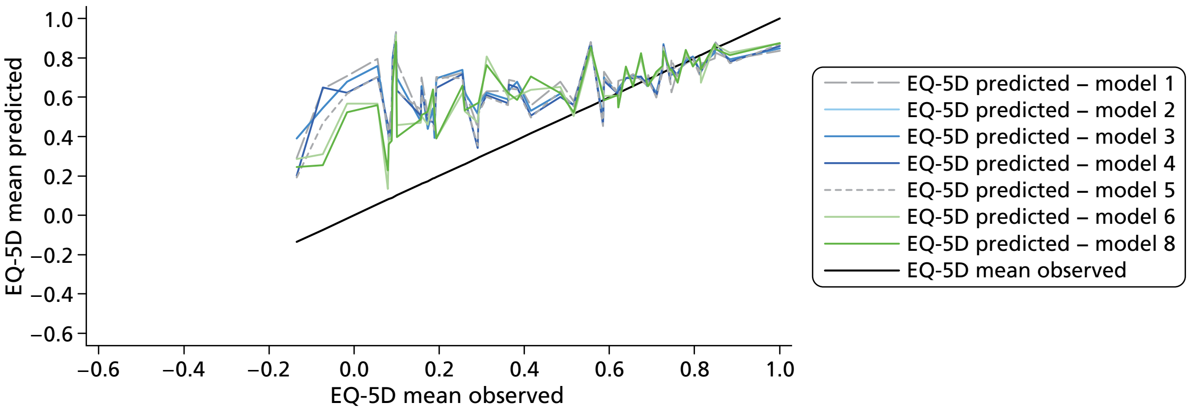

FIGURE 18 Summary of performance of all tobit models.

TABLE 63 Summary of observed and predicted values per model: TPMs

			Model 1		Model 2		Model 3		
Summary statistics and model performance tests		Observed values	Total score		Domain scores		Two-part significant domains		
Mean (SD)		0.721 (0.223)	0.744 (0.139)		0.741 (0.150)		0.743 (0.148)		
Median		0.735	0.755		0.758		0.760		
Range		–0.135–1	0.314–0.977		0.350–0.980		0.336–0.975		
MAE			0.129		0.125		0.125		
Shrinkage			0.922		0.911		0.930		
			Part 1	***Part 2***	***Part 1***	***Part 2***	***Part 1***		***Part 2***
Model goodness of fit			$\chi^2_{153} = 131$, $p = 0.896$		$\chi^2_{518} = 826$, $p < 0.001$		$\chi^2_{256} = 662$, $p < 0.001$		
Log-likelihood			–189	170	–177	184	–179		182
Sigma			N/A	0.203	N/A	0.194	N/A		0.195
Pseudo R^2			0.234	N/A	0.280	N/A	0.272		N/A
AIC			381	–333	364	–356	364		–356
BIC			390	–321	386	–330	377		–339
	n	**Mean**	**Mean**	**MAE**	**Mean**	**MAE**	**Mean**		**MAE**
ECOG									
Normal, no symptoms	122	0.8645	0.8144	0.1035		0.8265	0.0906	0.8279	0.0895
Some symptoms	256	0.7219	0.7444	0.1219		0.7420	0.1249	0.7437	0.1252
Require some bed	152	0.6055	0.6857	0.1593		0.6721	0.1534	0.6735	0.1531
ANOVA		$F_{2,527} = 55$, $p < 0.001$	$F_{2,527} = 91$, $p < 0.001$			$F_{2,527} = 135$, $p < 0.001$		$F_{2,527} = 145$, $p < 0.001$	

N/A, not applicable.

a Model 6 would not converge for logistic regression (some levels were dropped owing to having no observations reducing the sample size $n = 404$). This model is not compared with the other models as it is based on a different sample.

Model 4		Model 5		Model 6a		Model 7		Model 8	
Two-part significant domains and squared terms		Two-part significant domains, squared and interaction terms		Two-part item levels: significant levels only		Two-part item levels: significant levels only, collapse unordered items		Two-part significant domains, squared terms, interaction terms and significant patient characteristics	
0.739 (0.154)		0.739 (0.153)		0.791 (0.132)		0.744 (0.149)		0.739 (0.154)	
0.753		0.763		0.809		0.735		0.759	
0.119–0.993		0.106–0.971		0.321–0.992		0.476–0.987		0.106–0.988	
0.120		0.120		0.093		0.122		0.118	
0.944		0.946		0.589		0.917		0.953	
Part 1	***Part 2***	***Part 1***	***Part 2***	***Part 1***	***Part 2***	***Part 1***	***Part 2***	***Part 1***	***Part 2***
$\chi^2_{479} = 451$, $p < 0.001$		$\chi^2_{255} = 215$, $p = 0.967$		$\chi^2_{348} = 517$, $p < 0.001$		$\chi^2_{41} = 56$, $p = 0.058$		$\chi^2_{463} = 444$, $p = 0.733$	
–165	200	–170	203	–131	260	–175	188	–160	203
N/A	0.184	N/A	0.182	N/A	0.154	N/A	0.195	N/A	0.182
0.328	N/A	0.307	N/A	0.399	N/A	0.288	N/A	0.350	N/A
343	–389	350	–390	332	–444	364	–360	336	–390
369	–363	367	–355	482	–281	394	–326	370	–310
Mean	**MAE**	**Mean**	**MAE**	**Mean**	**MAE**	**Mean**	**MAE**	**Mean**	**MAE**
0.8302	0.0896	0.8269	0.0907	0.8470	0.0764	0.8313	0.0871	0.8278	0.0876
0.7359	0.1211	0.7374	0.1205	0.7665	0.0919	0.7401	0.1204	0.7375	0.1185
0.6713	0.1410	0.6716	0.1435	0.7028	0.1161	0.6797	0.1513	0.6706	0.1412
$F_{2,527} = 122$, $p < 0.001$		$F_{2,527} = 117$, $p < 0.001$		$F_{2,401} = 62$, $p < 0.001$		$F_{2,527} = 112$, $p < 0.001$		$F_{2,527} = 112$, $p < 0.001$	

TABLE 64 Coefficients for modelling to FACT-G domain scores: TPMs

	TPM model 4	
	Significant summary scores and squared terms (SE)	
Domains	Part 1	Part 2
Physical	–0.458 (0.161)**	
Social		
Emotional		–0.105 (0.022)***
Functional	0.420 (0.178)*	
Physical^2	0.016 (0.004)***	0.0005 (0.00007)***
±Emotion	1.540 (0.455)**	0.825 (0.163)***
±Functional	–2.76 (1.431)*	0.075 (0.015)***
Constant	–2.574 (3.482)	–1.369 (0.308)***
Number of observations	530	437

* Statistically significant at the 10% level.
** Statistically significant at the 5% level.
*** Statistically significant at the 1% level.

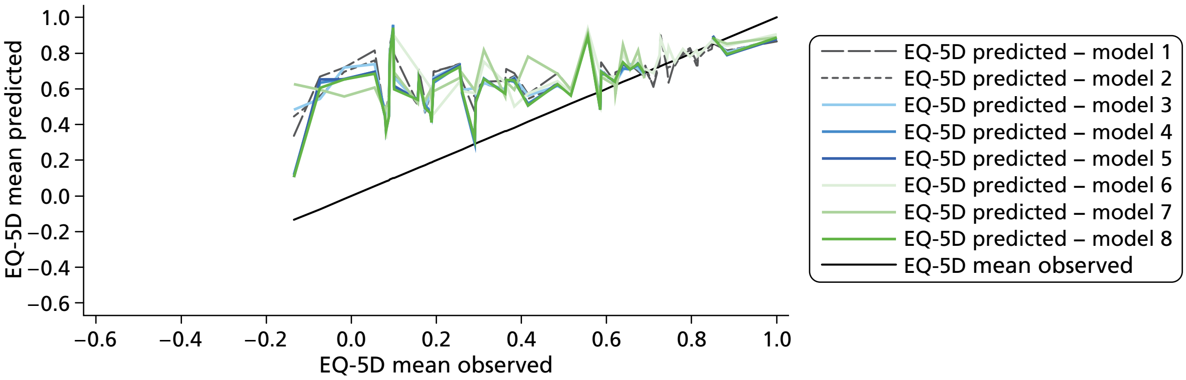

FIGURE 19 Summary of performance of all TPMs.

TABLE 65 Summary of observed and predicted values per model: splining

Summary statistics and model performance tests		Observed values	SPL model 1		SPL model 3	
			Total score		Significant domains	
Mean (SD)		0.721 (0.223)	0.724 (0.134)		0.723 (0.144)	
Median		0.735	0.745		0.736	
Range		–0.135–1	0.250–0.937		0.312–0.974	
Pseudo R^2			0.827		1.079	
Log-likelihood			–23.02		10.45	
AIC			54.04		–6.91	
BIC			71.13		23.00	
MAE			0.130		0.123	
Sigma			0.210		0.198	
Shrinkage			0.961		0.982	
	n	Mean	Mean	MAE	Mean	MAE
ECOG						
Normal, no symptoms	122	0.8645	0.8163	0.112	0.8460	0.097
Some symptoms	256	0.7219	0.7334	0.121	0.7277	0.121
Require some bed	152	0.6055	0.6325	0.160	0.6152	0.148
ANOVA		$F_{2,527} = 55, p < 0.001$	$F_{6,527} = 87, p < 0.001$		$F_{6,527} = 130, p < 0.001$	

SPL, splining.

TABLE 66 Coefficients for modelling to FACT-G significant domain scores

Summary statistics and model performance tests	SPL: model 2 (SE)
Physical (0–25)	0.013 (0.002)***
Physical score (> 25)	0.079 (0.016)***
Emotional (0–15)	0.020 (0.005)***
Emotional (> 15)	0.001 (0.004)
Functional	0.010 (0.002)***
Constant	–0.006 (0.075)
Number of observations	530

SPL, splining.

*** Statistically significant at the 1% level.

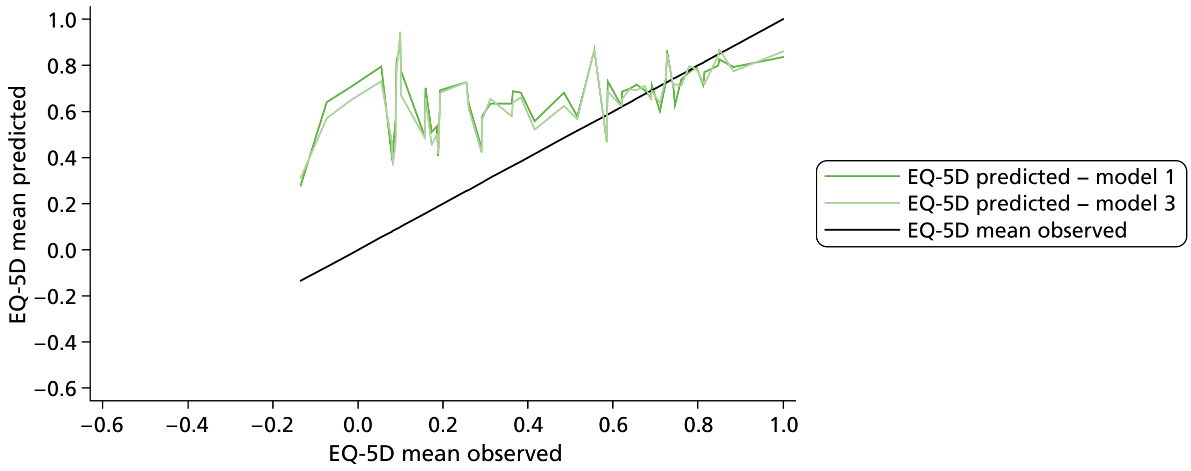

FIGURE 20 Summary of performance of all splining models.

TABLE 67 Summary of observed and predicted values per model: response mapping

			Response mapping 1		Response mapping 2		Response mapping 3		Response mapping 4		Response mapping 5	
Summary statistics and model performance tests		Observed values	Total score		Significant domain scores		Significant domain scores, squared and square root terms		Significant domain scores, squared, square root and interaction terms		Significant domain scores, squared, square root, interaction terms and characteristics	
n		530	530		530		530		530		530	
Mean (SD)		0.721 (0.223)	0.715 (0.126)		0.720 (0.133)		0.677 (0.122)		0.732 (0.175)		0.762 (0.178)	
Median		0.735	0.728		0.737		0.695		0.780		0.817	
Range		–0.135 to 1	0.223–0.930		0.268–0.934		0.194–0.848		0.214–0.960		0.157–0.975	
MAE			0.130		0.125		0.136		0.131		0.138	
Shrinkage			1.027		1.019		1.968		1.146		0.981	
	n	Mean	Mean	MAE	Mean	MAE	Mean	MAE	Mean	MAE	Mean	MAE
ECOG												
Normal, no symptoms	122	0.8645	0.7784	0.1161	0.7933	0.1009	0.7439	0.1273	0.8265	0.0882	0.8509	0.0913
Some symptoms	256	0.7219	0.7160	0.1216	0.7201	0.1219	0.6761	0.1289	0.7336	0.1288	0.7694	0.1355
Require some bed rest	152	0.6055	0.6635	0.1565	0.6601	0.1485	0.6241	0.1539	0.6537	0.1700	0.6798	0.1786
ANOVA	$F_{2,527} = 55, p < 0.001$		$F_{2,527} = 87, p < 0.001$		$F_{2,527} = 120, p < 0.001$		$F_{2,527} = 112, p < 0.001$		$F_{2,527} = 127, p < 0.001$		$\mathrm{F}_{2,527} = 251, p < 0.001$	

TABLE 68 Model 3: coefficients for FACT-G significant domain scores

Summary statistics and model performance tests	Mobility		Self-care		Usual activities		Pain		Anxiety/depression	
	Some problems	Extreme problems	Some problems	Extreme problems	Some problems	Extreme problems	Some problems	Extreme problems	Some problems	Extreme problems
Physical	–0.111 (0.023)***	N/A	–0.100 (0.024)***	–0.244 (2.191)	–0.237 (0.044)***	–0.285 (0.056)***	–0.206 (0.030)***	–0.319 (0.051)***		
Emotional		N/A							–0.331 (0.036)***	–0.607 (5.147)
Functional	–0.074 (0.020)***	N/A	–0.104 (0.027)***	–0.307 (6.663)	–0.124 (0.030)***	–0.266 (0.053)***	–0.057 (0.023)*	0.010 (0.053)	–0.047 (0.021)*	–0.197 (1.465)
Constant	3.089 (0.418)***	N/A	1.633 (0.427)***	2.017 (60.731)	7.737 (0.895)***	8.239 (1.210)***	5.499 (0.574)***	3.510 (1.045)**	6.773 (0.660)***	8.839 (47.729)
Log-likelihood	–310.22		–189.70		–338.3		–346.92		–302.08	
Pseudo R^2	0.132		0.151		0.263		0.191		0.263	
AIC	626.44		391.39		688.27		705.84		616.16	
BIC	639.26		417.03		713.91		731.48		641.80	

* Statistically significant at the 10% level.

** Statistically significant at the 5% level.

*** Statistically significant at the 1% level.

N/A, not applicable as there is no one with extreme problems for mobility.

Values in brackets are the standard errors of regression coefficients.

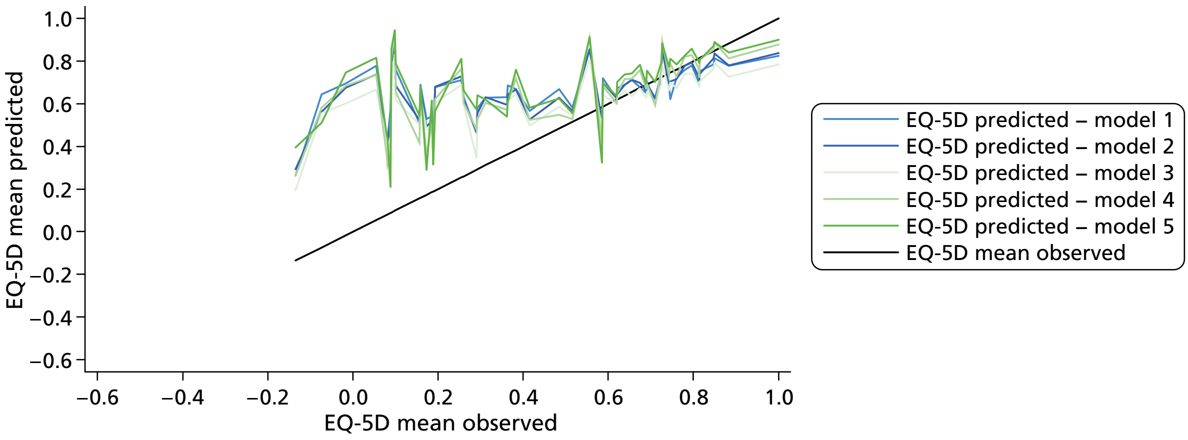

FIGURE 21 Mean predicted EQ-5D scores and observed scores.

Appendix 14 Summary of time trade-off values for all health states included in the exploratory bolt-on study

	Count	Mean	SD	Median	Minimum	Maximum
EQ-5D						
11121 (mild)	76	0.94	0.11	1.00	0.50	1
22222 (moderate)	74	0.71	0.30	0.80	–0.30	1
22233 (severe)	74	0.41	0.40	0.43	–0.80	1
11112	75	0.93	0.14	1.00	0.40	1
11122	75	0.87	0.19	1.00	0.20	1
21232	76	0.52	0.40	0.50	–0.80	1
22323	75	0.46	0.43	0.50	–0.93	1
33232	74	0.11	0.40	0.01	–0.93	1
33333	75	–0.02	0.40	0.00	–0.93	1
EQ-5D + hearing						
111211	76	0.94	0.13	1.00	0.40	1
111212	75	0.90	0.18	1.00	0.10	1
111213	75	0.85	0.24	0.98	0.00	1
222221	74	0.80	0.25	0.90	0.00	1
222222	75	0.77	0.27	0.90	–0.30	1
222223	75	0.70	0.30	0.75	–0.05	1
222331	75	0.40	0.44	0.47	–0.98	1
222332	74	0.45	0.44	0.50	–0.98	1
222333	76	0.36	0.41	0.45	–0.98	1
EQ + vision						
111211	74	0.94	0.11	1.00	0.45	1
111212	74	0.90	0.13	0.93	0.47	1
111213	75	0.69	0.28	0.75	0.00	1
222221	75	0.74	0.23	0.75	0.20	1
222222	75	0.76	0.21	0.75	0.20	1
222223	75	0.59	0.29	0.60	0.00	1
222331	75	0.41	0.35	0.46	–0.63	1
222332	76	0.41	0.34	0.43	–0.50	1
222333	75	0.32	0.33	0.35	–0.50	1

	Count	Mean	SD	Median	Minimum	Maximum
EQ-5D + tiredness						
111211	74	0.94	0.14	1.00	0.35	1
111212	73	0.90	0.15	1.00	0.38	1
111213	77	0.82	0.26	0.93	–0.38	1
222221	75	0.79	0.26	0.93	–0.17	1
222222	75	0.74	0.30	0.80	–0.38	1
222223	75	0.72	0.27	0.80	–0.43	1
222331	75	0.45	0.43	0.50	–0.90	1
222332	77	0.45	0.42	0.50	–0.80	1
222333	75	0.34	0.45	0.40	–0.90	1

EME
HS&DR
HTA
PGfAR
PHR

Part of the NIHR Journals Library
www.journalslibrary.nihr.ac.uk

Published by the NIHR Journals Library