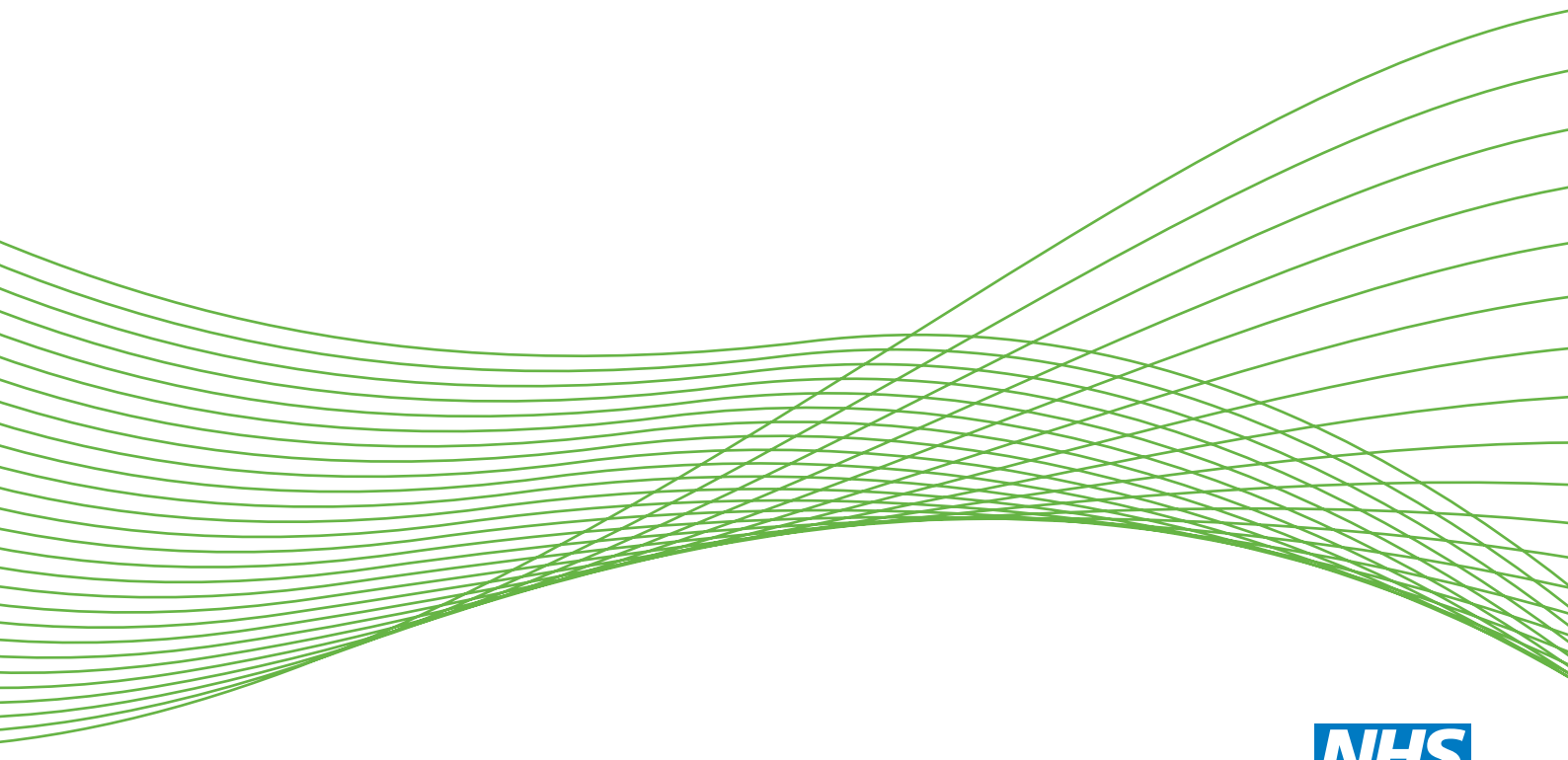
HEALTH TECHNOLOGY ASSESSMENT

VOLUME 17 ISSUE 5 FEBRUARY 2013
ISSN 1366-5278

Development of DEMQOL-U and DEMQOL-PROXY-U: generation of preference-based indices from DEMQOL and DEMQOL-PROXY for use in economic evaluation

B Mulhern, D Rowen, J Brazier, S Smith, R Romeo, R Tait, C Watchurst, K-C Chua, V Loftus, T Young, D Lamping, M Knapp, R Howard and S Banerjee



National Institute for Health Research

DOI 10.3310/HTA17050

Development of DEMQOL-U and DEMQOL-PROXY-U: generation of preference-based indices from DEMQOL and DEMQOL-PROXY for use in economic evaluation

B Mulhern,[1] D Rowen,[1] J Brazier,[1] S Smith,[2] R Romeo,[3] R Tait,[3] C Watchurst,[3] K-C Chua,[3] V Loftus,[3] T Young,[1] D Lamping,[2†] M Knapp,[3,4] R Howard[3] and S Banerjee[5]*

[1]Health Economics and Decision Science, School of Health and Related Research (ScHARR), University of Sheffield, Sheffield, UK
[2]Department of Health Services Research and Policy, London School of Hygiene and Tropical Medicine, London, UK
[3]Institute of Psychiatry, King's College London, London, UK
[4]Personal Social Services Research Unit, London School of Economics, London, UK
[5]Centre for Dementia Studies, Brighton and Sussex Medical School, Brighton, UK

*Corresponding author

[†]In memorium

Declared competing interests of authors: none

Published February 2013
DOI: 10.3310/hta17050

This report should be referenced as follows:

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

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

ISSN 1366-5278 (Print)

ISSN 2046-4924 (Online)

Five-year impact factor: 5.596

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) (http://www.publicationethics.org/).

Editorial contact: nihredit@southampton.ac.uk

The full HTA archive is freely available to view online at http://www.hta.ac.uk/project/htapubs.asp. Print copies can be purchased from the individual report pages.

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

For more information about the HTA programme please visit the website: http://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.

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

Abstract

Development of DEMQOL-U and DEMQOL-PROXY-U: generation of preference-based indices from DEMQOL and DEMQOL-PROXY for use in economic evaluation

B Mulhern,[1] D Rowen,[1] J Brazier,[1] S Smith,[2] R Romeo,[3] R Tait,[3] C Watchurst,[3] K-C Chua,[3] V Loftus,[3] T Young,[1] D Lamping,[2†] M Knapp,[3,4] R Howard[3] and S Banerjee[5*]

[1]Health Economics and Decision Science, School of Health and Related Research (ScHARR), University of Sheffield, Sheffield, UK
[2]Department of Health Services Research and Policy, London School of Hygiene and Tropical Medicine, London, UK
[3]Institute of Psychiatry, King's College London, London, UK
[4]Personal Social Services Research Unit, London School of Economics, London, UK
[5]Centre for Dementia Studies, Brighton and Sussex Medical School, Brighton, UK

*Corresponding author

[†]In memorium

Background: Dementia is one of the most common and serious disorders in later life and the economic and personal cost of caring for people with dementia is immense. There is a need to be able to evaluate interventions in dementia using cost-effectiveness analyses, but the generic preference-based measures typically used to measure effectiveness do not work well in dementia. Existing dementia-specific measures can effectively measure health-related quality of life but in their current form cannot be used directly to inform cost-effectiveness analysis using quality-adjusted life-years as the measure of effectiveness.

Objectives: The aim was to develop two brief health-state classifications, one from DEMQOL and one from DEMQOL-Proxy, to generate health states amenable to valuation. These classification systems consisted of items taken from DEMQOL and DEMQOL-Proxy so they can be derived from any study that has used these instruments.

Data sources: In the first stage of the study we used a large, clinically representative sample aggregated from two sources: a sample of patients and carers attending a memory service in south London and a sample of patients and carers from other community services in south London. This included 644 people with a diagnosis of mild/moderate dementia and 689 carers of those with mild/moderate dementia. For the valuation study, the general population sample of 600 respondents was drawn to be representative of the UK general population. Households were sampled in urban and rural areas in northern England and balanced to the UK population according to geodemographic profiles. In the patient/carer valuation study we interviewed a sample of 71 people with mild dementia and 71 family carers drawn from a memory service in south London. Finally, the instruments derived were applied to data from the HTA-SADD (Study of Antidepressants for Depression in Dementia) trial.

Review methods: This was a complex multiphase study with four linked phases: phase 1 – derivation of the health-state classification system; phase 2 – general population valuation survey and modelling to produce values for every health state; phase 3 – patient/carer valuation survey; and phase 4 – application of measures to trial data.

Results: All four phases were successful and this report details this development process leading to the first condition-specific preference-based measures in dementia, an important new development in this field.

Limitations: The first limitation relates to the lack of an external data set to validate the DEMQOL-U and DEMQOL-Proxy-U classification systems. Throughout the development process we have made decisions about which methodology to use. There are other valid techniques that could be used and it is possible to criticise the choices that we have made. It is also possible that the use of a mild to moderate dementia sample has resulted in classification systems that do not fully reflect the challenges of severe dementia.

Conclusion: The results presented are sufficiently encouraging to recommend that the DEMQOL instruments be used alongside a generic measure such as the European Quality of Life-5 Dimensions (EQ-5D) in future studies of interventions in dementia as there was evidence that they can be more sensitive for patients at the milder end of disease and some limited evidence that the person with dementia measure may be able to reflect deterioration.

Funding: The National Institute for Health Research Health Technology Assessment programme.

Contents

List of abbreviations

AE	absolute error
AEF	adjacent endorsement frequency
ANOVA	analysis of variance
AQL-5D	Asthma Quality of Life Utility Index
BADLS	Bristol Activity of Daily Living Scale
CFA	confirmatory factor analysis
CSDD	Cornell Scale for Depression in Dementia
DCE	discrete choice experiment
df	degrees of freedom
DIF	differential item functioning
DOMINO	DOnepezil and Memantine IN mOderate to severe Alzheimer's disease
EFA	exploratory factor analysis
EORTC-8D	European Organisation for Research and Treatment of Cancer Core Quality of Life Questionnaire
EQ-5D	European Quality of Life-5 Dimensions
GLS	generalised least squares
HRQL	health-related quality of life
HTA	Health Technology Assessment
HTA-SADD	HTA Study of Antidepressants for Depression in Dementia
ICC	intraclass correlation
IRT	item response theory
KMO	Kaiser–Meyer–Olkin
LB	Ljung–Box
MAE	mean absolute error
MCID	minimal clinically important difference
MEF	maximum endorsement frequency
MMSE	Mini Mental State Examination
MVH	Measurement and Valuation of Health study, conducted in York
NICE	National Institute for Health and Clinical Excellence
NIHR	National Institute for Health Research
NPI	Neuropsychiatric Inventory
OAB-5D	Overactive Bladder Questionnaire-5 Dimensions
OLS	ordinary least squares
QALY	quality-adjusted life-year
RCT	randomised controlled trial
RMSE	root-mean-squared error
ScHARR	School of Health and Related Research
SD	standard deviation
SF-6D	Short Form questionnaire-6 Dimensions
SF-36	Short Form questionnaire-36 items

SG	standard gamble	TTO	time trade-off
SRM	standardised response mean	VAS	visual analogue scale

All abbreviations that have been used in this report are listed here unless the abbreviation is well known (e.g. NHS), or it has been used only once, or it is a non-standard abbreviation used only in figures/tables/appendices, in which case the abbreviation is defined in the figure legend or in the notes at the end of the table.

Executive summary

Background

The challenge of dementia

Dementia is one of the most common and serious disorders in later life with a prevalence of 5% and an incidence of 2% per year in the over-65s. In the UK there are currently 750,000 people with dementia and 200,000 new cases every year. It causes irreversible decline in global intellectual, social and physical functioning. Abnormalities in behaviour, insight and judgement are part of the disorder, as are neuropsychiatric symptoms such as psychosis, anxiety and depression. The economic cost of caring for people with dementia is immense. In the UK, the cost of dementia is around £17B per year, greater than the costs of stroke (£3B), heart disease (£4B) and cancer (£2B). More importantly, the negative impacts of dementia on those with the disorder, in terms of deteriorating function, and on carers are profound. Worldwide there are 35 million people with dementia and this costs $600B per year; these numbers are set to double and the costs to at least triple in the next 20 years. The need to improve care for people with dementia is a policy priority.

Evaluation of clinical effectiveness in dementia

Given its importance in public health terms and its devastating effects, it is understandable that there is a large and growing volume of basic, translational and applied research under way investigating the effectiveness of interventions to help people with dementia. This includes evaluations of psychological, educational and social interventions as well as trials of pharmacological treatments. Given the complexity of the syndrome of dementia, there has been discussion about how best to measure the impact of interventions. There is an emerging consensus that we need to measure broad patient-reported outcomes such as health-related quality of life (HRQL) in dementia as well as discrete areas such as cognition or behaviour. A variety of instruments are available to measure discrete areas of function across many of the major domains including cognition, behavioural problems and psychological symptoms, activities of daily living and depression in dementia, often using proxy reports of observable behaviour.

Measuring quality of life in dementia is more challenging, not least because of poor recall, time perception, insight and communication. However, recent studies indicate that meaningful measurements can be made using condition-specific measures, using both subjective and proxy instruments.

Funded by the National Institute for Health Research (NIHR) Health Technology Assessment (HTA) programme, we developed the DEMQOL system, a condition-specific measure of HRQL in dementia. The DEMQOL system consists of two interviewer-administered tools: DEMQOL (28 items), which is completed by the person with dementia (score range 28–112, with a higher score indicating better HRQL); and DEMQOL-Proxy (31 items), a proxy report of the HRQL of the person with dementia, completed by the main carer (score range 31–124). A global quality-of-life item is also included in both instruments but does not contribute to the overall score. The system was designed according to best psychometric practice, and there is some evidence for the validity of the scale. DEMQOL has good psychometric properties for people with mild to moderate dementia [defined as a Mini Mental State Examination (MMSE) score of 10+]. DEMQOL-Proxy can be used across dementia severity, from mild to severe.

Economic evaluation in dementia

The last two decades have seen the increased use of economics to inform the allocation of resources between competing health-care interventions around the world and particularly the use of cost-effectiveness, in which interventions are often assessed in terms of their cost per quality-adjusted life-year (QALY). The QALY provides a way of measuring the benefits of health-care interventions, including improvements in HRQL. Brief generic (i.e. not condition-specific) measures of HRQL are most commonly

used to put the 'Q' into the QALY. Such measures include the European Quality of Life-5 Dimensions (EQ-5D) and Short Form questionnaire-6 Dimensions (SF-6D); it is suggested that these are applicable to all interventions and patient groups. This claim has support across certain conditions, for example rheumatoid arthritis, for which it has passed conventional psychometric tests of reliability and validity, but is more questionable for others, such as visual impairment, hearing loss and schizophrenia.

There is reason to believe that the available brief generic measures of HRQL do not work well in dementia. The inherent impairments in dementia of recall, time perception, insight and expressive and receptive communication mean that it is not possible to assume that what works for a general non-cognitively impaired population will work for those with dementia. This means that instruments to be used in dementia need to be psychometrically tested in populations of people with dementia. When this has been done, the results have suggested that there are major potential difficulties in using such generic measures in dementia, with considerable error likely.

What then is needed to enable cost-effectiveness evaluation in dementia? If the use of the currently available brief generic measures is problematic because of the error inherent in their use, might it be possible to use instruments that can measure HRQL in dementia, such as DEMQOL and DEMQOL-Proxy? These instruments cannot directly be used in economic evaluation in their current form because they do not incorporate preference information. They therefore cannot yet be used to calculate QALYs for use in incremental cost-effectiveness analysis. This is a major limitation in the currently available measurement technology. To meet this need, this study aims to generate a preference-based single index for the two instruments that comprise the DEMQOL system (DEMQOL and DEMQOL-Proxy) for use in economic evaluation using general population values. In addition, we set out to generate patient and carer values for a sample of states to compare with the general population values and to test the new system using a trial data set.

Objectives

1. To derive health-state classification systems that are amenable to valuation from DEMQOL and DEMQOL-Proxy which can be used to categorise all patients with responses to the measures.
2. To generate utility values for every health state defined by the health-state classification systems developed from DEMQOL and DEMQOL-Proxy.
3. To examine whether or not utility values elicited from the general population differ from utility values elicited from patients and carers for dementia health states generated by the classification system.
4. To examine the psychometric performance of the dementia-specific preference-based measures using trial data.

Method and results

The overall aim was to develop two preference-based measures, one from DEMQOL and one from DEMQOL-Proxy. These measures use a subset of items from DEMQOL and DEMQOL-Proxy, respectively, to form classification systems so that utility scores can be produced for any study that has used the existing DEMQOL and/or DEMQOL-Proxy instruments. We have named the new measures DEMQOL-U and DEMQOL-Proxy-U, with the 'U' referring to the utility scores generated in this project. This was a complex multiphase study. The project had four linked phases:

- phase 1 – derivation of the health-state classification system
- phase 2 – general population valuation survey and modelling to produce values for every health state
- phase 3 – patient/carer valuation survey
- phase 4 – application of measures to trial data.

Phase 1a: derivation of the health-state classification system – quantitative evaluation of DEMQOL and DEMQOL-Proxy dimension structure

The analysis used DEMQOL ($n = 1189$) and DEMQOL-Proxy ($n = 1223$) data drawn from two sources: routine data collected from a memory service and data collected from a study assessing HRQL in dementia.

Method

We evaluated rates of missing data, maximum endorsement frequencies, adjacent endorsement frequencies and redundancy to evaluate the factor structure of DEMQOL and DEMQOL-Proxy and to determine the extent to which the conceptual domains are supported.

Results

We identified two separate five-factor models for DEMQOL and DEMQOL-Proxy. Both models reflect aspects of the original conceptual framework but also highlight important differences between self- and proxy reports. The factor structures were robust enough to provide the basis for the development of dementia-specific preference-based measures for patient self-report and proxy report by carers.

Phase 1b: derivation of the health-state classification system – development of a health-state classification system for DEMQOL AND DEMQOL-Proxy

The aim of this stage was to identify one item for each of the dimensions identified in DEMQOL and DEMQOL-Proxy.

Method

To identify the most robust items for use in the health-state classification system, five separate Rasch models were generated for both DEMQOL and DEMQOL-Proxy.

Results: final health-state classification systems

The five items selected to reflect the DEMQOL dimension structure form the basis for the DEMQOL classification system. This was named DEMQOL-U. Each dimension has four response levels that correspond to the options included on the original DEMQOL instrument. Therefore, the DEMQOL-U descriptive system generates a possible 1024 (i.e. 4^5) health states. The four items selected to reflect the DEMQOL-Proxy dimension structure form the basis of the DEMQOL-Proxy classification system. This was named DEMQOL-Proxy-U. It contains four dimensions each with four levels corresponding to those included in the original measure. Therefore, DEMQOL-Proxy-U generates 256 (i.e. 4^4) health states.

Phase 2: general population valuation survey and modelling to produce values for every health state

Preference-based measures have two components: first, a health-state classification system that can be used to categorise all patients with the condition of interest; and second, a means of obtaining a utility score for all states defined by the system. In this phase of the development we generated a preference-based single index for each classification system.

Method

The first stage of generating the preference-based single index involved a valuation study in which a representative sample of the general population valued a sample of health states derived from each classification system. The sample of states that was valued was derived using simulation. The time trade-off (TTO) technique, which asks respondents to trade off years in full health to avoid living in a particular health state, and ranking, in which respondents order health states from best to worst, were used for the valuation study. The analysis used a range of multivariate regression models including ordinary least squares and random-effects generalised least squares to produce a single-index measure from each classification system anchored on a full health–dead 1–0 scale, in which a value of 1 is equal to full health and a value of 0 is equal to being dead.

Results

The data generated were subjected to multiple multivariate regression and preference weights were generated. These enable a health-state utility value to be estimated for every health state defined by each classification system. These preference weights can be used to generate a utility score for a person with dementia each time they complete the DEMQOL questionnaire or their carer completes the DEMQOL-Proxy questionnaire.

Phase 3: patient/carer valuation survey

In the previous stage of the study we estimated a preference-based single index for each classification system using values obtained from the general population. However, such values can be obtained from other sources; here, we investigated patients and carers.

Method

Health states matched with a selection of those valued by the general population were valued using TTO by samples of people with dementia and carers. The elicited values were compared with the general population values.

Results

People with dementia and carers of people with dementia gave systematically lower utility values than members of the general population. These results suggest that the population used to produce dementia health-state utility values may well impact on the results of cost-effectiveness analysis and potentially affect resource allocation decisions, and no systematic adjustment between values is possible.

Phase 4: application of measures to trial data

If the DEMQOL-U and DEMQOL-Proxy-U are to be used alongside or instead of generic preference-based measures it is important to assess their psychometric validity, responsiveness and level of agreement between patient and carer report. This can be assessed by applying psychometric methods to data sets containing responses to the DEMQOL system alongside generic preference-based and non preference-based measures.

Method

We compared the validity, patient/proxy agreement and responsiveness of the EQ-5D and the DEMQOL-U and DEMQOL-Proxy-U utility measures. The data for these analyses were obtained from the HTA Study of Antidepressants for Depression in Dementia (HTA-SADD), a multicentre placebo-controlled pragmatic randomised controlled trial of the clinical effectiveness of sertraline and mirtazapine.

Results

There is some evidence for the acceptability of the DEMQOL system, in particular the DEMQOL-Proxy-U, which displays low missing data rates. There is no clear pattern regarding agreement between patients and carers. In terms of responsiveness, there is evidence that the DEMQOL utility measures and EQ-5D are less sensitive to change than the original DEMQOL and DEMQOL-Proxy. The psychometric performance of the DEMQOL utility measures may be impacted by the sample used, which focused on those with depression in dementia and so may not be representative of all those with dementia. The inconclusive nature of the results means that further testing on a range of samples is required.

Conclusions

We have detailed the development and application of two dementia-specific preference-based measures, one for self-completion (DEMQOL-U) and the other to be completed by carers (DEMQOL-Proxy-U). These measures can be used to generate health-state utility values on the QALY scale for use in economic evaluation of interventions in this group of patients. These are the first condition-specific preference-based measures in dementia. The results of the psychometric analysis are encouraging but the validity and

responsiveness of the instruments require further investigation; therefore, until more evidence is available, we would recommend that the DEMQOL instruments are used alongside a generic measure such as the EQ-5D in future evaluations of interventions for dementia.

Funding

Funding for this study was provided by the Health Technology Assessment programme of the National Institute for Health Research.

Chapter 1 Introduction

Background

The challenge of dementia

Dementia is one of the most common and serious disorders in later life with a prevalence of 5% and an incidence of 2% per year in the over 65s.[1,2] In the UK, there are currently 750,000 people with dementia[3] and 200,000 new cases every year. Dementia causes irreversible decline in global intellectual, social and physical functioning. Abnormalities in behaviour, insight and judgement are part of the disorder, as are neuropsychiatric symptoms such as psychosis, anxiety and depression. The economic cost of caring for people with dementia is immense. In the UK, the cost of dementia is around £17B per year,[3] greater than that of stroke (£3B), heart disease (£4B) and cancer (£2B).[4] More importantly, the negative impacts of dementia on those with the disorder, in terms of deteriorating function, and on carers[5,6] are profound. Worldwide there are 35 million people with dementia, and this costs $600B per year; these numbers are set to double and the costs to at least triple in the next 20 years.[7,8] The need to improve care for people with dementia is a policy priority.[9–12]

Evaluation of clinical effectiveness in dementia

Given dementia's importance in public health terms and its devastating effects, it is understandable that there is a large and growing volume of basic, translational and applied research under way investigating the effectiveness of interventions to help people with dementia. This includes evaluations of psychological, educational and social interventions as well as trials of pharmacological treatments. Given the complexity of the syndrome of dementia, there has been discussion about how best to measure the impact of interventions. There is an emerging consensus that we need to measure broad patient-reported outcomes such as health-related quality of life (HRQL) in dementia as well as discrete areas such as cognition or behaviour.[13] There are a variety of instruments available across all of the major domains such as cognition, behavioural problems and psychological symptoms, activities of daily living and depression in dementia, often using proxy reports of observable behaviour.

Measuring quality of life in dementia is more challenging, not least because of poor recall, time perception, insight and communication.[14] However, recent studies indicate that meaningful measurements can be made using both self- and proxy report condition-specific instruments.[14–17]

Funded by the National Institute for Health Research (NIHR) Health Technology Assessment (HTA) programme we developed the DEMQOL system, a condition-specific measure of HRQL in dementia.[14,17] The DEMQOL system consists of two interviewer-administered tools: DEMQOL (28 items), which is completed by the person with dementia (score range 28–112, with a higher score indicating better HRQL); and DEMQOL-Proxy (31 items), a proxy report of the HRQL of the person with dementia completed by the main carer (score range 31–124). A global quality-of-life item is also included in both instruments but does not contribute to the overall score. DEMQOL has evidence of reliability and validity for people with mild to moderate dementia [defined as a Mini Mental State Examination (MMSE) score of 10+]. DEMQOL-Proxy can be used across dementia severity, from mild to severe.

The development of HRQL instruments has lagged behind the development of measures of discrete function; therefore, they have not yet been widely employed in randomised controlled trials (RCTs) of anti-dementia medication and other treatments, which have instead concentrated on discrete areas of function, most commonly cognition, with the assumption that these are acceptable surrogates for HRQL. We have analysed associations between commonly used measures of specific outcomes in dementia and HRQL.[18] The data generated suggest that HRQL in dementia does not have a simple relationship with cognition or functional limitation. This and other studies[19,20] suggest that cognitive impairment and activity

limitation are inadequate proxies for HRQL in dementia. These data confirm that HRQL in dementia is a complex construct and that simple proxy substitution of cognition or function is likely to miss many important dimensions. They suggest that there may be considerable value in including measures of HRQL along with measures of specific function such as cognition and behaviour in treatment trials in dementia. A failure to include broad outcome measures such as those measuring HRQL, and a reliance on measures of discrete function, could lead to the positive effects of interventions being overlooked or to potential negative effects of interventions being missed.[21]

Economic evaluation in dementia

The situation is not quite so clear for the economic evaluation of treatments in dementia. We may be relatively confident about the data on the clinical effectiveness of treatments for dementia, at least with respect to the cholinesterase inhibitors and memantine for cognition in dementia[22] and carer interventions for psychosocial outcomes,[23] yet there is a lack of directly relevant data with which to ascertain the cost-effectiveness of such treatments.

The last two decades have seen the increased use of economics to inform the allocation of resources between competing health-care interventions around the world and particularly the use of cost-effectiveness, in which interventions are commonly assessed in terms of their cost per quality-adjusted life-year (QALY). The QALY provides a way of measuring the benefits of health-care interventions, including improvements in HRQL. Brief generic (i.e. not condition-specific) preference-based measures of HRQL are most commonly used to put the 'Q' into the QALY. Such measures include the European Quality of Life-5 Dimensions (EQ-5D)[24] and Short Form questionnaire-6 Dimensions (SF-6D),[25] and it is suggested that these are applicable to all interventions and patient groups. This claim has support across certain conditions, for example rheumatoid arthritis, for which it has passed conventional psychometric tests of reliability and validity,[26] but is more questionable for others, such as visual impairment,[27] hearing loss,[28] and schizophrenia.[29]

There is reason to believe that the available brief generic measures of HRQL do not work well in dementia. As noted above the inherent impairments in dementia of recall, time perception, insight and expressive and receptive communication mean that it is not possible to assume that what works for a general non-cognitively impaired population will work for those with dementia. This means that instruments to be used in dementia need to be psychometrically evaluated in populations of people with dementia. In fact, when self-report or proxy report is needed, this generally means that condition-specific measures need to be generated which can measure accurately in dementia whatever attribute is under consideration. When brief generic measures of HRQL have been tested in dementia, the results have been equivocal at best.[30–33]

In practical terms, the unsatisfactory nature of the current evidence base is very clearly illustrated by the major difficulties presented to the National Institute for Health and Clinical Excellence (NICE) in generating its technology appraisal guidance (TA111).[22] There has been a great deal of concern raised (including referral to judicial review) about the assumptions that had to be made with respect to the cost-effectiveness models. The challenges encountered can be attributed to a lack of direct data on cost and quality of life. The conclusions of TAG111 make clear the need for a measure that can be used in dementia trials to generate direct and accurate measurements of cost-effectiveness in quality-of-life terms. This conclusion is echoed in systematic reviews and trials in dementia.[34–36]

What then is needed to enable cost-effectiveness evaluation in dementia? If the use of the currently available generic preference-based measures is problematic, might it be possible to use instruments that are developed to measure HRQL in dementia such as DEMQOL and DEMQOL-Proxy? These instruments cannot directly be used in economic evaluation in their current form because they are too large to incorporate preference information. They therefore cannot yet be used to calculate QALYs for use in incremental cost-effectiveness analysis. This is a major limitation in the currently available measurement technology. However, the methodology is available to allow the benefits of the DEMQOL system to be applied to valuing the benefits of interventions for economic evaluation. This study aims to generate

a preference-based single index for these two instruments (DEMQOL and DEMQOL-Proxy) for use in economic evaluation using general population preference values. In addition, we generated patient and carer values to compare with the general population values and evaluated the preference-based measures developed in comparison with EQ-5D and other condition-specific indicators using a trial data set.

Chapter 2 Plan of investigation

Overview

This is necessarily a complex multiphase study. For ease of understanding this chapter will set out the overall plan of the investigation and development of the preference-based measures from the DEMQOL system. Each of the major elements will then be reported and discussed in its own chapter before bringing the data together to draw conclusions from the programme as a whole.

The project had four linked phases:

- phase 1 – derivation of the health-state classification systems
- phase 2 – general population valuation survey and modelling to produce values for every health state defined by each classification system
- phase 3 – patient and carer valuation survey
- phase 4 – application of measures to trial data.

The methods of deriving a preference-based measure from an existing condition-specific measure are the subject of a recent HTA review.[37] This identified six stages for this process (*Figure 1*). These stages are used as a guide to the key components in the development of a preference-based measure rather than a prescriptive methodology as it is not always practical or possible to follow each stage separately or sequentially. Furthermore, the precise technique used at each of the development stages used may differ depending on the size and structure of the original instrument. Derivatives of this approach have been applied successfully in deriving a preference-based index from the generic Short Form questionnaire-36 items (SF-36)[25] and Short Form questionnaire-12 items (SF-12)[38] and condition-specific measures including the King's Health Questionnaire,[39] the Asthma Quality of Life Questionnaire,[40,41] and the Overactive Bladder Questionnaire.[42]

The problem with deriving preference-based measures from existing measures of HRQL such as DEMQOL and DEMQOL-Proxy is that they are simply too large. The instruments include multiple dimensions and numerous items measuring a range of HRQL constructs and therefore would define many millions of potential health states, which would be too many and complex for preference valuation by respondents. The first challenge in this type of work is therefore to fashion a health-state classification system that is amenable to valuation by sampling items from the original instruments (like the EQ-5D in structure with a small number of dimensions and minimum number of items per dimension). The process of developing a health-state classification system is stages I–IV in *Figure 1*. To carry out stages I–IV, we identified and used existing data sets that included DEMQOL and DEMQOL-Proxy data (however, we were unable to identify data to carry out the validation required for stage IV).

The aim of phase 1 is to develop two brief health-state classification systems, one from DEMQOL and one from DEMQOL-Proxy, that could be used to generate health states for valuation. These classification systems consist of items taken from DEMQOL and DEMQOL-Proxy, respectively, and so can be derived from any study that has used the existing DEMQOL and/or DEMQOL-Proxy instruments. The classification systems and corresponding preference weights have been named DEMQOL-U and DEMQOL-Proxy-U, with the 'U' referring to the utility scores generated for every health state.

Organisations such as NICE and other similar agencies around the world charged with providing judgements on the cost-effectiveness of health-care interventions mostly require HRQL to be valued using a choice-based technique.[43] There are a range of choice-based techniques available including time trade-off (TTO), standard gamble (SG), ranking, and discrete choice experiments (DCEs). In TTO, respondents are

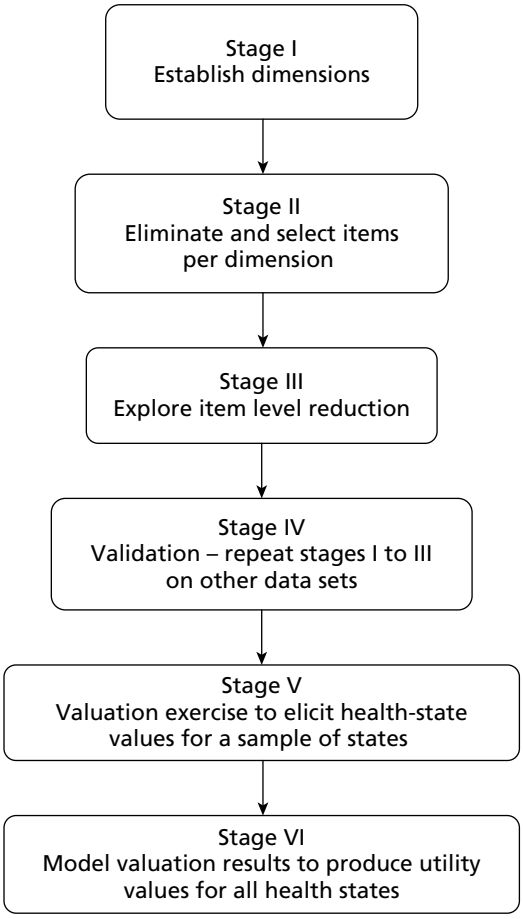


FIGURE 1 The six stages for deriving a preference-based HRQL measure.

asked how many years they would be willing to sacrifice in order to be in full health in comparison with a set number of years (usually 10) in an impaired health state described by the classification system. The time spent in full health is varied until the respondent is indifferent between the options. SG asks respondents how big a risk of death they are willing to take in order to have a chance of ending up in full health. Ranking asks respondents to rank a number of health states from best to worst and DCE asks respondents to choose which is better out of two health states produced by the classification system. For the valuation studies described in this report we used TTO and ranking. TTO was selected for this study over SG to avoid asking respondents how much they would be willing to risk their life. We used the TTO protocol developed in York,[44] which was used to derive the EQ-5D value set currently recommended for use in studies of cost-effectiveness by NICE.[45] This was done to allow for some level of comparability between the utility values produced by the DEMQOL measures and the EQ-5D value set and to hence best meet the reference case of NICE.

The question of whose values should be used to value health is an interesting one, but ultimately political,[39] and so beyond the scope of this project. The main valuation survey in this study used a representative sample of the general public to conform to the requirements of NICE and the majority of other reimbursement authorities around the world. General population values are typically used because they reflect the views of society as a whole, which funds the services, and to enhance comparability between programmes. Data from the general population survey were modelled using regression techniques to estimate preference-based scoring algorithms that can be applied to existing and future DEMQOL and DEMQOL-Proxy data to generate a utility score for each instrument.

However, there are important questions about whether or not the general population is the right valuation group in dementia given the complex nature of the disorder and the error often inherent in public attitudes and understandings of dementia.[11] In preparing the design of this study we engaged in conversation with the Alzheimer's Society through its Quality Research in Dementia forum. It was clear that it wanted the values of people with dementia and their family carers to be investigated directly as part of this research. We therefore took this opportunity to complete a supplementary valuation survey of people with dementia and their carers to explore the size and direction of any deviation from the general public valuations.

In the final phase of the study we applied the general population algorithms to a trial data set that was completed during the project [HTA Study of Antidepressants for Depression in Dementia (HTA-SADD), a placebo-controlled RCT of the treatment of depression in dementia, Chief Investigator SB][46] to investigate the psychometric performance of the DEMQOL utility measures. We had hoped to do the same for the MRC-DOMINO trial data set [DOnepezil and Memantine IN mOderate to severe Alzheimer's Disease, a placebo-controlled RCT of donepezil and memantine (Ebixa®, Lundbeck) alone and in combination for the treatment of patients whose treatment response is questioned, Chief Investigator RH]. However, the data set was not available for analysis by the end of this project and so we focused our analyses on the HTA-SADD data alone. This provided an opportunity to examine the psychometric properties of the indices in comparison with EQ-5D, the original DEMQOL measures and also other dementia-specific indicators.

Phase 1: derivation of health states from DEMQOL and DEMQOL-Proxy

The first task was to construct two health-state classification systems, one from DEMQOL and the other from DEMQOL-Proxy. This corresponds to stages I–IV in *Figure 1*:

- I – confirm the dimensional structure of DEMQOL and DEMQOL-Proxy
- II – select items from each dimension to construct the DEMQOL and DEMQOL-Proxy health-state classification systems
- III – explore item-level reduction
- IV – carry out validation of the classification systems.

Stage I: dimensional structure

The development of the DEMQOL system was based around a five-domain conceptual framework incorporating daily activities, health and well-being, cognitive functioning, social relationships and self-concept. Factor analysis has been used to investigate the dimensional structure of DEMQOL and DEMQOL-Proxy.[47] This technique attempts to identify underlying factors that explain the pattern of correlations within a set of observed variables. This analysis suggested that DEMQOL has four factors and DEMQOL-Proxy two. However, the factor structure proved inconclusive and difficult to define. Therefore, the first stage in this research was to carry out exploratory factor analysis (EFA) on two new data sets in which DEMQOL and DEMQOL-Proxy had been used. These were the Croydon Memory Service (CMS) clinical audit evaluation data set, which has over 1000 cases,[48] and further baseline data of over 100 people with dementia collected in a separate cohort study. These data were used to derive a final set of dimensions for the health-state classification systems using the factor analysis as a guide. Other techniques could have been used to support this stage, such as cluster analysis, but previous experience of the research team found this one to be most helpful. Rasch analysis was also used to confirm whether or not the proposed dimensions did indeed each reflect a unidimensional concept (see the following section for a full explanation).

Stage II: item selection

Each dimension of a health-state classification system is usually represented by one or occasionally two items from the original instrument. The selection of items must be undertaken with great care. This process has been assisted in past research undertaken by the Sheffield team by a combination of conventional psychometric methods and Rasch analysis.[37] Rasch analysis is a mathematical technique that converts qualitative (categorical) responses to a continuous (unmeasured) latent scale using a logit model.

The theory underlying this approach is that the probability of an affirmative response to each item (or each response to each item) depends on the degree of difficulty of the item (or severity in the case of health) and the ability of the respondent. In the development of a health-state classification system, Rasch analysis can then be used to eliminate items that poorly represent the underlying latent scale. Techniques based on item response theory (IRT) could have been used instead of Rasch analysis.

The process of selecting items in a number of studies has been broken down into two components. The first is the elimination of poorly performing items that do not meet key criteria tested for using Rasch analysis. This can leave a number of items in some dimensions and so the second component involves selecting the best item for each dimension.

To choose which items to eliminate in each dimension, separate Rasch models need to be fitted to each of the dimensions established in stage I. This is because Rasch analysis assumes unidimensionality and so it would not have made sense to consider all items together in a single analysis. The assumption of unidimensionality was also tested for each dimension using a Rasch procedure. In deriving the DEMQOL-U and DEMQOL-Proxy-U, Rasch models were fitted and items eliminated using three criteria. First, items unable to display item-level ordering were eliminated from consideration in the classification system as these items demonstrate an inability to distinguish between item response levels. Second, differential item functioning (DIF) was used to establish whether or not respondents with different characteristics respond differently to items. Items that display DIF are of limited value across subgroups of patients defined by the characteristic (as often would be required in an economic evaluation) and are therefore excluded. Third, items that do not fit the underlying Rasch model were eliminated as they do not represent the underlying dimension. These items are identified using Rasch model goodness of fit statistics.

Once items had been eliminated from the selection process, Rasch analysis and traditional psychometric methods were applied to select the 'best' items for the health-state classification system. The item selection criteria included item-level coverage across the latent space using the Rasch model (an indicator of how much of the underlying severity scale an item covers), item goodness of fit using the Rasch model, feasibility (level of missing data) and distribution of responses across response categories.

Stage III: explore item-level reduction

In practice, respondents may not be able to distinguish between item response choices and this is investigated at stage III of the development process. However, items from the DEMQOL instruments had only four response choices and the Rasch analyses for the items used for the health-state classification systems confirmed that respondents were able to distinguish between them. Therefore, there was no need to explore the possibility for further reductions.

Stage IV: validation of the classification system

The application of stages I–III produced health-state classifications for the two DEMQOL instruments. Before proceeding with the valuation process it is recommended that the analysis is repeated on an independent sample (stage IV of the development process) or a subsample of the data. However, no other samples were available during the development of the health-state classification system and the sample size of the data used for the Rasch analysis was not sufficient to allow a subsample to be generated. Therefore, this stage was omitted from the development process.

Phase 2: generation of utility values for all health states

Phase 2 of the project involved the generation of population-based utility values for all of the health states generated by the DEMQOL-U and DEMQOL-Proxy-U classification systems. This forms stages V and VI of the development process detailed in *Figure 1*.

Stage V: general population valuation study

The aim of this phase of the project was to obtain valuations of a selection of health states generated by the DEMQOL-U and DEMQOL-Proxy-U classification systems using the TTO and ranking elicitation

techniques. The key design issues were the sampling of states for valuation, the sampling of respondents and the content of the interviews.

Sample of states

The DEMQOL-U and DEMQOL-Proxy-U health-state classification systems define many health states and so only a sample of health states produced by each system was valued. The selection of health states used a novel approach based on simulation rather than the conventional approaches based on an orthogonal array or a balanced design in order to provide a more efficient basis for selecting states into blocks (i.e. combinations of health states an individual respondent will value) for use in TTO interviews.

Respondents

The main sample of respondents in this survey was drawn to be representative of the UK general population in terms of characteristics such as age and socioeconomic status. Households were sampled in urban and rural areas in northern England and balanced to the UK population according to geodemographic profiles.

The interview

At the interview, respondents self-completed the classification system of the instrument they were valuing (i.e. either the DEMQOL-U or the DEMQOL-Proxy-U) to familiarise themselves with the classification system. They were then asked to rank eight states alongside full health and dead and then value these eight health states using TTO. Respondents were guided through one practice TTO exercise to ensure that they understood the task. They were also asked a number of sociodemographic questions. In the sample size calculation undertaken before phase 1 of the project, we assumed that there would be up to 100 states to value for each classification system and each state was to be valued 30 times, which implies a required sample size for one health-state classification of 375. This sample size or less has been successfully used to value a number of descriptive systems.[25,39,42] We initially assumed the same level of complexity for the items derived from DEMQOL-Proxy-U and so this required a further 375 interviews, suggesting a final sample size of 750. In the event, the final DEMQOL-U and DEMQOL-Proxy-U designs required fewer states to be valued and therefore a smaller total sample size was used.

Stage VI: modelling health-state values

Econometric models were then fitted to the health-state TTO valuations using the TTO value as the dependent variable and each level of each domain, other than the baseline, entered as dummy variables. A range of different specifications were explored, including aggregate models using mean health-state values and ordinary least squares (OLS) and random-effects generalised least squares (GLS) models using individual-level data.[25] The impact of adding interaction terms and various transformations was explored. Rank data were modelled using the rank-ordered logit model and anchored onto the 1–0 full health–dead scale required to produce QALYs using the modelled TTO value of the worst health state. All models were subjected to the standard tests of goodness of fit.[25] The best TTO models were selected and converted into scoring algorithms to be applied to existing and future DEMQOL and DEMQOL-Proxy data sets. The algorithms have been produced in SPSS (SPSS Inc., Chicago, IL, USA) and are publicly available free of charge on the DEMQOL website (www.kcl.ac.uk/iop/depts/hspr/research/ciemh/mha/demqol/index.aspx).

Phase 3: patient and carer valuation survey

The aim of phase 3 was to examine whether or not health-state values elicited from patients and carers for health states defined by DEMQOL-U and DEMQOL-Proxy-U, respectively, differed significantly from those provided by the general public. Respondents from these two groups were recruited from clinical contacts in south London. We recruited people with a clinical diagnosis of mild dementia, with mild severity of dementia defined by a MMSE score of > 18. Their main family carer was also recruited when possible.

Respondents were asked to value a set of eight states using the same methods as the general population valuation survey. These interviews were undertaken in people's homes at a time that was convenient for them by research workers trained in the TTO valuation method.

Assuming a power of 0.8, significance level of 0.05, standard deviation (SD) of 0.3 and expected difference of 0.1, this required a sample of 71 interviews to compare with mean valuations per state from the main general population survey for each of the measures. Mean values obtained from different populations (i.e. general population vs carer and general population vs patients) were compared using simple *t*-tests.

Econometric models were also estimated for each classification system to estimate the impact of population, health-state severity and respondent sociodemographic characteristics on elicited health-state utility values.

Phase 4: application to trial data

If the preference-based single index developed as part of this study is used instead of or alongside generic preference-based measures it is important that it is valid, reliable and responsive in a dementia patient and carer population. This can be assessed by applying psychometric methods to data sets containing both generic and condition-specific preference-based measures. The issue of how condition-specific preference-based measures compare with generic preference-based measures is particularly important as this indicates the likely impact of using condition-specific preference-based measures compared with generic measures to generate QALY values for use in economic evaluation. The recently completed HTA-SADD trial – a multicentre randomised double-blind, placebo-controlled trial of the clinical effectiveness of sertraline and mirtazapine – was used to conduct such an investigation.

The aim of this phase was to compare the validity and responsiveness of DEMQOL-U and DEMQOL-Proxy-U with those of the EQ-5D and other dementia-specific indicators. In looking at these measures we will seek to determine (1) if the utility scores derived from the DEMQOL-U and DEMQOL-Proxy-U perform less well than the original measures and (2) the performance of the condition-specific preference-based measures in comparison with a generic preference-based measure.

Source of data

The data for the analyses shown in this chapter were obtained from the HTA-SADD study of the use of antidepressants for depression in dementia. Clinical measures available included the Cornell Scale for Depression in Dementia (CSDD),[49] the MMSE,[50] the Bristol Activity of Daily Living Scale (BADLS),[51] the Neuropsychiatric Inventory (NPI),[52] DEMQOL and DEMQOL-Proxy[14,17] and the EQ-5D.[53]

Statistical analyses

Summary statistics were used to describe the distribution of responses on the self-report EQ-5D, carer report EQ-5D (CEQ-5D), DEMQOL-U and DEMQOL-Proxy-U. Agreement between measures (EQ-5D/DEMQOL-U and CEQ-5D/DEMQOL-Proxy-U) was assessed. The utility values generated from self-report HRQL data were then compared with those generated from carer reports to examine the extent of agreement between the two sets of values. Construct validity was next examined in light of the absence of a gold standard for utility measurement in populations with cognitive impairment. Specifically, the construct validity of the EQ-5D, DEMQOL-U and their proxy equivalents was examined in terms of convergent validity to quantify the association between the utility values and measures of cognitive impairment (MMSE), depression in dementia (CSDD), neurobehavioural problems (NPI) and daily functioning (BADLS). Construct validity was also further examined using known group differences based on recommended thresholds for the MMSE[54] and CSDD.[55] Responsiveness to change was examined using the minimal clinically important difference (MCID) thresholds recommended by the DOMINO trial group.[56]

Outline of chapters in this report

The remainder of this report presents the methods used in more detail, the rationale for the methods used, the findings and the implications of the findings for the conduct of economic evaluation in this area. *Chapter 3* presents a quantitative evaluation of the DEMQOL and DEMQOL-Proxy dimensional structure. This evidence was used to establish the dimensions used for the health-state classification systems. *Chapter 4* presents the Rasch analyses completed to inform the selection of items used to construct the health-state classification systems. The general population valuation survey is presented in *Chapter 5* along with the econometric analyses generating utility values for all states described by the classification systems. The comparison of patient and carer health-state values with those of the general population is described in *Chapter 6*. The DEMQOL-U and DEMQOL-Proxy-U indices are then evaluated in *Chapter 7*. *Chapter 8* provides a brief discussion of the work and how the resultant preference-based measures can be used in economic evaluation.

Chapter 3 Quantitative evaluation of the DEMQOL and DEMQOL-Proxy dimensional structure

The analyses reported in this chapter used classical psychometric techniques to conduct item analyses to inform the development of the DEMQOL and DEMQOL-Proxy health-state classification systems. We evaluated the dimensional structure of DEMQOL and DEMQOL-Proxy to provide a basis for the generation of patient- and carer-reported disease-specific descriptive systems. This represents stage I of the development process described in *Chapter 2*. The analyses reported here replicate a selection of those conducted as part of the original development of DEMQOL and DEMQOL-Proxy[47] but use a substantially bigger and independent sample.

Background to the DEMQOL system

The DEMQOL system[14,17] is a measure of HRQL in dementia. It consists of two interviewer-administered instruments: DEMQOL (self-reported by the patient) and DEMQOL-Proxy (proxy reported by a carer). DEMQOL and DEMQOL-Proxy can be used for all types of dementia and were developed from a five-domain conceptual framework. This was based on in-depth qualitative interviews with people with dementia and their carers[47] and includes health and well-being, cognitive functioning, social relationships, daily activities and self-concept. Items were drafted to represent each of the conceptual framework domains. The items developed were piloted with patients and carers and subsequently evaluated in two-stage field testing (item reduction followed by psychometric evaluation).

DEMQOL contains 28 items reported on a four-point Likert scale (a lot/quite a bit/a little/not at all); all items refer to the last week. A global quality-of-life item is also included but does not contribute to the overall score. Items are scored from 1 to 4, with higher scores indicating better HRQL. In the original psychometric evaluation there was some evidence of content validity (four of the original conceptual domains were represented in the item-reduced version). DEMQOL was also found to have high reliability (internal consistency and test–retest) and moderate validity (convergent and discriminant) in mild and moderate dementia.[14,17] In terms of acceptability, there was some evidence of missing data but floor and ceiling effects were not apparent. Factor analyses established a four-factor solution (defined as daily activities, memory, positive emotion and negative emotion), but the factor model was inconclusive and did not fully support the original conceptual framework.

DEMQOL-Proxy has 31 items reported on a four-point Likert scale (a lot/quite a bit/a little/not at all) and also includes an additional global quality-of-life item. The original psychometric evaluation found that DEMQOL-Proxy has good reliability (internal consistency and test–retest) and acceptable content validity (all five conceptual domains are represented in the item-reduced final version) and that there is some evidence for convergent and discriminant validity across the full range of severity. There is also evidence for acceptability in mild/moderate and severe dementia. Factor analysis suggested a two-factor solution (functioning and emotion), but this did not support the original conceptual framework.

Method

The current analysis used DEMQOL ($n = 1189$) and DEMQOL-Proxy ($n = 1223$) data drawn from two sources: routine data collected from a memory service[48] and data collected from an unpublished study assessing HRQL in dementia. To be consistent with the original DEMQOL development, we excluded those without a definite diagnosis of dementia ($n = 451$) and also those with a MMSE score < 10, indicating severe memory problems ($n = 80$). Although DEMQOL-Proxy can be used for those with severe dementia,

data from carers of patients with severe dementia were also excluded from the analysis so that the patient and proxy samples used for the development of the preference-based measures were consistent in terms of diagnosis and severity. Analyses reported below on non-imputed data are based on this sample of 658 for DEMQOL and 692 for DEMQOL-Proxy.

For the factor analyses we used imputed data. We therefore excluded those for whom imputation (using the standard rule outlined in the DEMQOL scoring) could not be undertaken (DEMQOL: $n = 14$; DEMQOL-Proxy: $n = 10$). The final sample for factor analysis was therefore 644 for DEMQOL and 682 for DEMQOL-Proxy. The demographic characteristics of the sample used for the factor analyses are shown in *Table 1*.

First, we conducted item analysis on non-imputed data. To do this we evaluated rates of missing data, maximum endorsement frequencies (MEFs), adjacent endorsement frequencies (AEFs) and redundancy using the criteria outlined below. Next, we imputed responses for missing data (using the imputation rule specified in the standard scoring instructions for DEMQOL) to evaluate the factor structure of DEMQOL and DEMQOL-Proxy and to determine the extent to which the conceptual domains are supported. The criteria for item analyses are based on well-established techniques and also our own previous psychometric work.[14,57–61]

Item analysis (pre imputation)

Missing data

We selected a more stringent criterion for missing data than the original DEMQOL development study. This was based on consideration of the range of rates of missing data across the sample. In general, the current

TABLE 1 DEMQOL and DEMQOL-Proxy demographics of patient group

Characteristic	DEMQOL	DEMQOL-Proxy
n	644	682
Female (%)	69.9	60.4
Age (years), mean (SD)	78.83 (7.59)	79.23 (7.69)
Age range (years)	44–97	44–106
Ethnicity (%)		
White	85.3	85.7
Asian	4.9	4.9
Black	7.5	6.8
Other	2.3	2.6
MMSE, mean (SD)	20.81 (4.67)	20.59 (4.58)
MMSE severity, *n* (%)		
Mild	272 (42.2)	308 (45.2)
Moderate	372 (57.8)	374 (54.8)
Diagnosis, *n* (%)		
Late-onset AD	286 (44.4)	307 (44.9)
Atypical/mixed	154 (23.9)	166 (24.3)
Other	204 (31.7)	209 (30.8)

AD, Alzheimer's disease.

sample was found to have fewer missing data than in the original development study and so we used the criterion of < 5% for acceptable rates of missing data. This is similar to other work of this type. The high number of missing data in the original development study meant that items were originally selected using more lenient missing data criteria of < 30% for DEMQOL and < 10% for DEMQOL-Proxy.

Maximum endorsement frequencies and adjacent endorsement frequencies

We evaluated the proportion of respondents who endorse each response category, including floor and ceiling effects (i.e. response categories with high endorsement rates at the bottom/top ends of the response scale respectively). The MEF should be < 80%. The sum of any two AEFs should be ≥ 10%.

Redundancy

Redundancy was assessed using inter-item correlations. Items were defined as redundant if the inter-item correlation is > 0.75.

Next, we conducted exploratory factor analyses on imputed data.

Factor analysis (post imputation)

Exploratory factor analysis (principal axis factoring with varimax rotation) was used to provide a preliminary evaluation of the extent to which the a priori conceptual domains were quantitatively supported in DEMQOL and DEMQOL-Proxy. The data were screened by specifying that the Kaiser–Meyer–Olkin (KMO) statistic should be > 0.5 and that Bartlett's test of sphericity should be non-significant. Several factor-analytic models were considered based on eigenvalues > 1 and scree plots and also by considering the conceptual meaningfulness of alternative models obtained by requesting specific numbers of factors. Items were removed from the model if they did not load ≥ 0.40 on any factor, or cross-loaded within 0.20 on more than one factor.[62] These item-removal criteria were also used in the original development of the DEMQOL system. The factor analysis was repeated on randomly generated split-half samples to test the stability of the models. There was no significant difference for either the DEMQOL or the DEMQOL-Proxy total scores across the split halves, which was tested using analysis of variance (ANOVA). Furthermore, there were no conceptual differences between the models produced across the split halves, and therefore the results presented here will focus on the whole sample analysis.

Results

DEMQOL item analysis (pre imputation)

Missing data

No items display missing data rates above the maximum accepted level (5%).

Maximum endorsement frequencies and adjacent endorsement frequencies

All DEMQOL items passed the criterion for MEFs. Ten items had AEFs below the criterion (10%) for the response options 'quite a bit' and 'a lot' (feeling distressed, forgetting who people are, thoughts being muddled, not having enough company, how you get on with other people, having enough affection, people not listening, making yourself understood, getting help when you need it, getting to the toilet on time).

Redundancy

All item pairs met the criterion for redundancy. No pairs were correlated ≥ 0.75.

DEMQOL factor analysis (post imputation)

The five-factor model explained 45.5% of the variance and was supported by the eigenvalue > 1 rule and the scree plots. The four- and six-factor models were also considered but as the factors were conceptually difficult to interpret they were rejected. The five factors were defined as cognition (factor 1; six items),

negative emotion (factor 2; five items), positive emotion (factor 3; five items), social relationships (factor 4; six items) and loneliness (factor 5; two items). There were four non-loading items, and these were excluded from all of the dimensions and also from consideration for inclusion in the health-state classification system. *Table 2* displays the model for both the overall sample and the two split halves (split half 1: five-factor model explained 46.29% of the variance; split half 2: five-factor model explained 51.92% of the variance). There is broad similarity between the overall and split-half models and factor scores were therefore derived from the overall model. All three samples met the data screening criteria (KMO statistic > 0.5; Bartlett's test of sphericity non-significant). We derived dimension scores for each of the factors by adding the (unweighted) items loading on each factor. These were labelled cognitive functioning, negative emotion, positive emotion, social relationships and loneliness.

DEMQOL-Proxy item analysis (pre imputation)

Missing data

Eleven DEMQOL-Proxy items (feeling frustrated, feeling full of energy, feeling sad, feeling content, feeling distressed, feeling lively, feeling fed up, having things to look forward to, getting what you want from the shops, using money to pay for things, looking after your finances) had missing data frequencies above the criterion.

Maximum endorsement frequencies and adjacent endorsement frequencies

One DEMQOL-Proxy item (keeping yourself clean) displayed a MEF above the criterion (80%). For the majority of the items assessing cognition, daily activities and social relationships the most highly endorsed response option was 'not at all'. Eight items displayed AEFs below the criterion for the response options 'quite a bit' and 'a lot' (forgetting things that happened a long time ago, forgetting where you are, keeping yourself clean, keeping yourself looking nice, getting what you want from the shops, using money to pay for things, getting in touch with people, not being able to help other people).

Redundancy

All item pairs met the criterion for redundancy as no pairs were correlated ≥ 0.75.

DEMQOL-Proxy factor analysis (post imputation)

A five-factor model (explaining 49.3% of the variance) also provided the best fit for DEMQOL-Proxy. This was supported by the eigenvalue > 1 rule and the scree plot. The five factors were defined as cognition (factor 1; nine items), negative emotion (factor 2; six items), daily activities (factor 3; three items), positive emotion (factor 4; three items) and appearance (factor 5; two items). The four- and six-factor models were also considered but the factors could not be clearly interpreted and so they were rejected. There were two cross-loading items and five non-loading items, and these were eliminated from the dimensional structure and were not considered further for inclusion in the health-state classification system. *Table 3* shows the five-factor models for the whole sample and also both split-half samples. All of the other factor analysis specifications were met (KMO statistic > 0.5; Bartlett's test of sphericity non- significant).

Discussion

This chapter provides an evaluation of each of the items and also the dimension structure of DEMQOL and DEMQOL-Proxy to determine the extent to which the original conceptual framework is supported by quantitative data. The analysis replicates part of the original development work[14] for DEMQOL and DEMQOL-Proxy using a large, clinically representative sample. This factor analysis provides the basis for the dimensions represented in the dementia-specific preference-based measures developed.

We have identified two separate five-factor models for DEMQOL and DEMQOL-Proxy. Both models reflect aspects of the original conceptual framework but also highlight important differences between

self- and proxy reports. The potential for one dimension (self-concept) to be represented was limited by the particular items that were retained in the final versions of DEMQOL and DEMQOL-Proxy. The four original domains (daily activities, health and well-being, cognitive functioning, social relationships) that were operationalised in the questionnaire were represented in the model for either DEMQOL or DEMQOL-Proxy, but neither model supported all four of these original domains. This supports using DEMQOL and DEMQOL-Proxy as complementary measures in trials and research settings. For example, the domain of health and well-being split into two factors (representing positive and negative emotion) for both DEMQOL and DEMQOL-Proxy. The domain of daily activities was more strongly supported in DEMQOL-Proxy than in DEMQOL, probably reflecting the seven items retained for this domain in the final DEMQOL-Proxy compared with two items in DEMQOL. Similarly, social relationships is evident as a factor in DEMQOL but not in DEMQOL-Proxy, reflecting the five items retained for this domain in DEMQOL compared with two items in DEMQOL-Proxy. However, the emphasis on daily activities and looking after yourself in DEMQOL-Proxy and social relationships in DEMQOL is also consistent with the differences found between self- and proxy reports in previous qualitative work.[14] In addition, the factor analyses reported here suggest that in self-reported data (DEMQOL) social relationships consists of two separate parts (which we have labelled 'social relationships' and 'loneliness') and in proxy-reported data (DEMQOL-Proxy) daily activities and looking after yourself consists of two separate parts (which we have labelled 'daily activities' and 'appearances').

The factor structure reported here better supports the original conceptual domains than the earlier factor analysis conducted during the development of the DEMQOL system. This more positive result is possibly due to the much larger sample used for this study (approximately six times larger than that used for the original validation).

In the development of the classification systems we used EFA rather than confirmatory factor analysis (CFA) to investigate the dimensionality of DEMQOL and DEMQOL-Proxy. We recognise that CFA would also be a valid way to investigate the dimensionality of the DEMQOL system. However, we used EFA because the factor analysis carried out during the development of the DEMQOL system on a much smaller sample was inconclusive and did not clearly match the original conceptual framework, and therefore we did not have an established a priori factor structure to confirm. There are some features of the EFA method used that may impact on the dimension structure established for DEMQOL and DEMQOL-Proxy. First, factors with more items are more likely to be clearly defined using EFA, and this has implications for the item loadings of the smaller factors both within and across factors. Second, factors are based around inter-item correlations and so a minimum of two items is required to generate a factor. However, factors with only two items are not strong, and five or more strongly loading items are recommended.[63] In this study we used EFA as a guide to the dimensionality and after running a range of models selected the five-factor models that fitted the data conceptually and did not contain a high number of non- or cross-loading items. The DEMQOL and DEMQOL-Proxy models both contain a factor with two items but we believe that the items in each of these factors can be clearly defined as a dementia-specific HRQL concept. Also, as the aim is to develop a multiattribute instrument that retains many of the key HRQL concepts from the original instrument, we believe that using the factor model including only strong factors with five or more items would have resulted in a classification system with reduced sensitivity, which has implications for the preference weights derived from the valuation stage.

The factor analysis used orthogonal rotation, which assumes independence between the factors extracted. It could be argued that, as the instruments are measuring constructs relating to dementia, the factors are related, and therefore we should use oblique rotation, which takes into account the relationship between factors. For the DEMQOL system, the factor structures produced using both rotation methods were similar and so orthogonal was used. For both DEMQOL and DEMQOL-Proxy there were items that loaded on more than one factor. However, as we use factor analysis to also exclude items from consideration for the reduced classification system, we followed guidelines used in the development of DEMQOL to exclude items (i.e. items loading within 0.2 of each other across factors or < 0.4 on any factor).

TABLE 2 DEMQOL factor analysis[a]

Item	Original domain	Cognition			Negative emotion			Positive emotion			Social relationships			Loneliness		
		All	Half 1	Half 2	All	Half 1	Half 2	All	Half 1	Half 2	All	Half 1	Half 2	All	Half 1	Half 2
Q1. Felt cheerful	Health and well-being	0.134	0.112	0.176	0.273	0.235	0.242	0.449	0.449	0.465	0.147	0.135	0.214	0.259	0.237	0.234
Q2. Felt worried	Health and well-being	0.309	0.316	0.311	0.418	0.376	0.393	0.256	0.270	0.289	0.258	0.232	0.260	0.083	–0.002	0.249
Q3. Felt that you are enjoying life	Health and well-being	0.108	0.108	0.115	0.243	0.198	0.225	0.579	0.571	0.604	0.033	0.034	0.070	0.289	0.297	0.263
Q4. Felt frustrated	Health and well-being	0.157	0.119	0.282	0.634	0.561	0.581	0.224	0.258	0.237	0.163	0.163	0.060	0.030	0.057	0.103
Q5. Felt confident	Health and well-being	0.188	0.230	0.191	0.190	0.113	0.135	0.568	0.594	0.604	0.096	0.096	0.027	0.009	0.028	0.037
Q6. Felt full of energy	Health and well-being	0.034	0.000	0.092	0.131	0.164	0.093	0.751	0.721	0.754	0.056	0.020	0.058	–0.001	–0.021	0.001
Q7. Felt sad	Health and well-being	0.242	0.197	0.315	0.458	0.449	0.361	0.235	0.229	0.265	0.144	0.102	0.143	0.336	0.342	0.395
Q8. Felt lonely	Health and well-being	0.164	0.130	0.226	0.170	0.181	0.084	0.096	0.111	0.095	0.167	0.134	0.176	0.739	0.742	0.801
Q9. Felt distressed[b]	Health and well-being	0.340	0.316	0.397	0.370	0.304	0.117	0.125	0.190	0.208	0.251	0.289	0.315	0.270	0.272	0.329
Q10. Felt lively	Health and well-being	0.039	0.060	0.024	0.106	0.131	0.043	0.787	0.725	0.836	0.053	–0.075	–0.011	0.054	0.057	0.000
Q11. Felt irritable	Health and well-being	0.232	0.283	0.230	0.536	0.413	0.662	0.109	0.186	0.107	0.138	0.170	0.116	0.072	0.181	–0.041
Q12. Felt fed up	Health and well-being	0.202	0.190	0.274	0.609	0.469	0.673	0.242	0.274	0.282	0.128	0.129	0.126	0.297	0.387	0.257
Q13. Things wanted to do but couldn't[b]	Health and well-being	0.136	0.132	0.107	0.355	0.451	0.321	0.270	0.162	0.368	0.210	0.115	0.291	0.088	0.099	0.128
Q14. Forgetting things that happened recently	Cognitive functioning	0.605	0.630	0.610	0.229	0.191	0.224	0.117	0.130	0.106	0.104	0.085	0.015	0.103	0.103	0.146
Q15. Forgetting who people are	Cognitive functioning	0.539	0.527	0.589	0.107	0.124	0.079	0.037	0.045	0.037	0.230	0.195	0.184	0.095	0.114	0.051

Item	Original domain	Cognition			Negative emotion			Positive emotion			Social relationships			Loneliness		
		All	Half 1	Half 2	All	Half 1	Half 2	All	Half 1	Half 2	All	Half 1	Half 2	All	Half 1	Half 2
Q16. Forgetting what day it is	Cognitive functioning	0.612	0.604	0.648	0.089	0.047	0.115	0.080	0.067	0.089	0.199	0.156	0.194	0.103	0.045	0.125
Q17. Thoughts being muddled	Cognitive functioning	0.682	0.681	0.698	0.220	0.180	0.190	0.052	0.057	0.082	0.194	0.154	0.232	0.130	0.092	0.122
Q18. Difficulty making decisions	Cognitive functioning	0.504	0.536	0.531	0.242	0.290	0.142	0.143	0.181	0.118	0.340	0.310	0.275	0.101	0.093	0.122
Q19. Poor concentration	Cognitive functioning	0.627	0.607	0.693	0.217	0.207	0.184	0.165	0.138	0.193	0.168	0.138	0.076	0.039	0.071	0.047
Q20. Not having enough company	Social relationships	0.114	0.131	0.135	0.087	0.044	0.080	0.116	0.126	0.118	0.260	0.134	0.237	0.656	0.677	0.656
Q21. How you get on with people close to you	Social relationships	0.185	0.168	0.264	0.200	0.276	0.122	0.051	0.015	0.095	0.567	0.549	0.778	0.168	0.101	0.089
Q22. Getting the affection that you want	Social relationships	0.081	0.072	0.163	0.113	0.223	0.023	0.028	–0.017	0.064	0.637	0.692	0.643	0.257	0.198	0.244
Q23. People not listening to you	Social relationships	0.133	0.111	0.262	0.097	0.055	0.136	0.023	0.070	0.009	0.664	0.749	0.427	0.148	0.174	0.131
Q24. Making yourself understood	Cognitive functioning	0.337	0.398	0.370	0.208	0.111	0.260	0.030	–0.002	0.086	0.487	0.537	0.386	0.039	0.078	0.018
Q25. Getting help when you need it	Social relationships	0.313	0.343	0.338	0.110	0.078	0.146	0.092	0.099	0.086	0.527	0.589	0.397	0.159	0.167	0.138
Q26. Getting to the toilet on time	Daily activities	0.213	0.252	0.203	0.121	0.255	0.045	0.047	0.047	0.021	0.450	0.432	0.384	–0.040	–0.087	0.073
Q27. How you feel in yourself[b]	Health and well-being	0.338	0.267	0.457	0.359	0.445	0.280	0.178	0.088	0.233	0.388	0.391	–0.011	0.064	0.073	0.162
Q28. Health overall[b]	Health and well-being	0.272	0.213	0.355	0.397	0.576	0.284	0.267	0.150	0.350	0.264	0.207	0.236	0.012	–0.004	0.089

a Highlighted items load > 0.4.

b These items are non-loaders.

TABLE 3 DEMQOL-Proxy factor analysis[a]

Item	Original domain	Cognition			Negative emotion			Daily activities			Positive emotion			Appearance		
		All	Half 1	Half 2	All	Half 1	Half 2	All	Half 1	Half 2	All	Half 1	Half 2	All	Half 1	Half 2
Q1. Felt cheerful[b]	Health and well-being	0.034	–0.008	0.066	0.499	0.392	0.556	0.072	0.099	0.022	0.471	0.592	0.326	–0.027	0.021	–0.022
Q2. Felt worried	Health and well-being	0.270	0.309	0.266	0.632	0.600	0.649	0.123	0.134	0.127	0.088	0.114	0.044	0.004	–0.052	0.081
Q3. Felt frustrated	Health and well-being	0.237	0.257	0.216	0.618	0.627	0.579	0.079	0.073	0.072	0.143	0.178	0.063	0.050	0.058	0.099
Q4. Felt full of energy	Health and well-being	0.023	0.063	–0.015	0.105	0.073	0.164	0.000	0.029	–0.038	0.810	0.790	0.797	0.021	0.028	0.013
Q5. Felt sad	Health and well-being	0.172	0.169	0.155	0.687	0.741	0.662	0.110	0.123	0.104	0.101	0.142	0.062	0.111	0.026	0.225
Q6. Felt content[b]	Health and well-being	–0.00	–0.064	0.058	0.541	0.460	0.605	0.089	0.072	0.113	0.401	0.463	0.336	0.021	0.062	–0.077
Q7. Felt distressed	Health and well-being	0.269	0.294	0.245	0.681	0.708	0.648	0.156	0.104	0.208	0.014	0.034	–0.029	0.096	0.084	0.134
Q8. Felt lively	Health and well-being	0.037	0.093	–0.003	0.141	0.127	0.174	0.002	0.009	–0.022	0.833	0.817	0.836	0.002	–0.008	0.028
Q9. Felt irritable	Health and well-being	0.121	0.146	0.095	0.531	0.482	0.566	0.045	0.024	0.041	0.103	0.164	0.006	0.113	0.232	0.044
Q10. Felt fed up	Health and well-being	0.205	0.163	0.254	0.666	0.688	0.638	0.091	0.060	0.130	0.197	0.238	0.138	0.087	0.097	0.112

Item	Original domain	Cognition			Negative emotion			Daily activities			Positive emotion			Appearance		
		All	Half 1	Half 2	All	Half 1	Half 2	All	Half 1	Half 2	All	Half 1	Half 2	All	Half 1	Half 2
Q11. Things to look forward to	Health and well-being	0.017	–0.012	0.042	0.314	0.140	0.514	0.073	0.036	0.134	0.454	0.508	0.402	0.010	0.003	0.024
Q12. Memory in general	Cognitive functioning	0.661	0.631	0.699	0.193	0.193	0.174	0.028	0.064	0.022	–0.003	0.027	–0.021	–0.015	–0.032	0.030
Q13. Forget things that happened long ago	Cognitive functioning	0.457	0.519	0.375	0.106	0.086	0.106	0.138	0.127	0.164	–0.024	0.027	–0.092	0.105	0.153	0.116
Q14. Forget things that happened recently	Cognitive functioning	0.755	0.766	0.731	0.182	0.176	0.182	0.046	0.084	0.038	0.018	0.023	–0.006	–0.017	–0.029	0.021
Q15. Forgetting people's names	Cognitive functioning	0.604	0.586	0.602	0.091	0.112	0.066	0.032	0.087	–0.014	–0.049	–0.017	–0.084	0.146	0.155	0.153
Q16. Forgetting where he/she is	Cognitive functioning	0.295	0.343	0.182	0.209	0.149	0.286	0.373	–0.072	0.352	–0.056	0.379	–0.052	0.058	0.145	0.043
Q17. Forgetting what day it is	Cognitive functioning	0.575	0.575	0.550	0.196	0.241	0.166	0.173	0.287	0.084	0.053	0.050	0.059	0.102	0.086	0.119
Q18. Thoughts being muddled	Cognitive functioning	0.695	0.697	0.696	0.263	0.323	0.186	0.163	0.161	0.180	0.083	0.084	0.067	0.084	0.065	0.113
Q19. Difficulty making decisions	Cognitive functioning	0.651	0.644	0.639	0.184	0.187	0.180	0.224	0.212	0.247	0.111	0.101	0.104	0.107	0.154	0.066
Q20. Making him/herself understood	Cognitive functioning	0.471	0.543	0.442	0.206	0.178	0.190	0.204	0.154	0.245	–0.012	0.002	–0.067	0.120	0.230	0.040

continued

TABLE 3 DEMQOL-Proxy factor analysis[a] (*continued*)

Item	Original domain	Cognition			Negative emotion			Daily activities			Positive emotion			Appearance		
		All	Half 1	Half 2	All	Half 1	Half 2	All	Half 1	Half 2	All	Half 1	Half 2	All	Half 1	Half 2
Q21. Keeping him/herself clean	Daily activities	0.174	0.198	0.150	0.121	0.087	0.108	0.092	0.085	0.096	–0.009	–0.016	–0.015	0.730	0.848	0.705
Q22. Keeping him/herself looking nice	Daily activities	0.146	0.143	0.149	0.148	0.105	0.166	0.114	0.122	0.095	0.040	0.026	0.055	0.772	0.671	0.773
Q23. Getting what he/she wants from the shops	Daily activities	0.207	0.311	0.116	0.117	0.101	0.101	0.518	0.516	0.508	0.003	0.054	–0.042	0.279	0.231	0.349
Q24. Using money to pay for things	Daily activities	0.185	0.247	0.086	0.110	0.076	0.136	0.810	0.832	0.778	0.039	0.065	–0.012	0.058	0.061	0.095
Q25 Looking after his/her finances	Daily activities	0.229	0.186	0.247	0.172	0.157	0.203	0.655	0.659	0.688	0.015	0.042	0.005	0.005	0.049	–0.027
Q26. Things taking longer than they used to	Daily activities	0.430	0.398	0.383	0.145	0.085	0.123	0.258	0.290	0.274	0.143	0.147	0.138	0.122	0.313	0.311
Q27. Getting in touch with people[c]	Daily activities	0.097	0.034	0.341	–0.029	–0.05	0.092	0.154	0.188	0.307	0.026	–0.003	0.069	0.231	0.131	0.396
Q28. Not having enough company[c]	Social relationships	0.222	0.175	0.233	0.264	0.339	0.190	0.183	0.266	0.264	0.101	0.071	0.138	0.371	0.342	0.372
Q29. Not being able to help other people[c]	Social relationships	0.318	0.326	0.269	0.090	0.078	0.128	0.084	0.002	0.073	0.052	0.102	0.014	0.235	0.058	0.263
Q30. Not playing a useful part in things[c]	Self	0.349	0.377	0.270	0.161	0.184	0.134	0.081	0.040	0.116	0.081	0.059	0.094	0.190	0.007	0.384
Q31 Physical health[c]	Health and well-being	0.226	0.158	0.279	0.236	0.297	0.209	0.011	0.147	–0.025	0.120	0.101	0.106	0.317	0.300	0.312

a Highlighted items load > 0.4.

b These items are cross-loaders.

c These items are non-loaders.

There are a number of other methods that we could use to investigate the factor structure of the DEMQOL system, including the use of a polychoric matrix or IRT techniques such as Rasch analysis. Using a multiple-factor IRT approach within the six-step guide to developing a condition-specific preference-based measure would mean that both the investigation of dimensionality (step I) and item elimination and selection (steps II–IV) could be based around the same underlying latent model. Further research may assess differences in the factor structures produced using IRT and traditional factor analysis to inform the use of both processes in the development of condition-specific health-state classification systems.

The item analyses revealed that all of the DEMQOL items were robust in terms of missing data and redundancy, although 10 items failed the AEF criterion. There are 17 items in DEMQOL-Proxy that failed at least one of the item analysis criteria. This does not necessarily mean that they are uninformative items, and indeed all items passed all of the criteria in the original development study. These classical item analyses are considered alongside the Rasch analyses in *Chapter 3* to inform the final selection of items for the system.

Conclusion

The factor structures established here provide support for dimensions that are similar but not identical to the domains of the original conceptual framework. The work reported here has not evaluated whether or not the dimensions can be used as scores alongside the overall scores for DEMQOL and DEMQOL-Proxy. This would require a prospective study including carefully chosen validating measures. The factor structures are robust enough to provide the basis for the development of dementia-specific preference-based measures for patient self-report and proxy report by carers.

Chapter 4 Development of a health-state classification system for DEMQOL and DEMQOL-Proxy

Introduction

This chapter presents stages II and III of the development process described previously (see *Figure 1*). The aim of the analysis was to identify one item to represent each dimension in the health-state classification system. This involved the application of Rasch methods to evaluate the performance of each of the items within each dimension.

Rasch analysis is one of a number of IRT techniques that could be used to assess the performance of items included in a range of questionnaires and tests, including measures of HRQL. Rasch converts ordinal or categorical responses to items into a continuous unidimensional latent scale. This is done using logit modelling. When items fitted to the scale assess HRQL, the latent scale represents a continuous measure of quality of life covering the full severity range for the particular HRQL construct being measured. Respondents are also modelled on the logit scale, and an individual's position on the scale represents their level of severity in terms of the construct being measured. Responses to items are assumed to be a function of both the position of the person and the item on the overall latent scale. This means that respondents at the more severe end of the logit scale should be more likely to indicate more severe problems on the items included in the dimension than those at the lower end of the scale. The Rasch model for each dimension allows for an assessment of item performance at the overall dimension and individual response category levels. Therefore, Rasch can be used to inform the selection of items from existing condition-specific measures of HRQL to generate a condition-specific health-state classification system.[40,64] A range of Rasch statistics are used to first eliminate poorly performing items and subsequently select items for each dimension. The tests used are outlined below.

Method

To identify the most robust items to use in the health-state classification system, Rasch models were fitted separately to each of the dimensions established by the factor analysis at stage I (*Tables 4* and *5*). The Rasch model assumes unidimensionality and, as DEMQOL and DEMQOL-Proxy are not unidimensional (see *Chapter 3*), it would not be appropriate to fit a single Rasch model encompassing all DEMQOL/DEMQOL-Proxy items. Five separate Rasch models were therefore generated for both DEMQOL and DEMQOL-Proxy. Stage II of the development process described in this chapter includes two parts. First, we eliminated items across each dimension by assessing the performance of the items included in each model. We assessed item response level ordering, DIF and goodness of fit of items to the Rasch model. Next, we used Rasch criteria to select one item for each dimension. This included the range of item responses on the logit scale and the spread of item responses at logit 0 (the average item difficulty on the severity scale). Each of these processes is described in detail below. Rasch analysis was carried out using Rasch Unidimensional Measurement Models (RUMM2020©, 1997–2004 RUMM Laboratory Pty Ltd, www.rummlab.com.au/).

Stage II: selecting items

Item elimination

Item-level ordering

For each model we evaluated the ordering of the responses to each individual item. Item responses are disordered if respondents cannot differentiate between the response choices and therefore the observed

TABLE 4 DEMQOL: factor analysis revalidation

Factor	Item	Loading
Cognition	Q17. How worried have you been about your thoughts being muddled?	0.682
	Q19. How worried have you been about poor concentration?	0.627
	Q16. How worried have you been about forgetting what day it is?	0.612
	Q14. How worried have you been about forgetting things that happened recently?	0.605
	Q15. How worried have you been about forgetting who people are?	0.539
	Q18. How worried have you been about difficulty making decisions?	0.504
Negative emotion	Q4. Have you felt frustrated?	0.634
	Q12. Have you felt fed up?	0.609
	Q11. Have you felt irritable?	0.536
	Q7. Have you felt sad?	0.458
	Q2. Have you felt worried?	0.418
Positive emotion	Q10. Have you felt lively?	0.787
	Q6. Have you felt full of energy?	0.751
	Q3. Have you felt that you are enjoying life?	0.579
	Q5. Have you felt confident?	0.568
	Q1. Have you felt cheerful?	0.449
Social relationships	Q23. How worried have you been about people not listening to you?	0.664
	Q22. How worried have you been about getting the affection that you want?	0.637
	Q21. How worried have you been about how you get on with people close to you?	0.567
	Q25. How worried have you been about getting help when you need it?	0.527
	Q24. How worried have you been about making yourself understood?	0.487
	Q26. How worried have you been about getting to the toilet on time?	0.450
Loneliness	Q8. Have you felt lonely?	0.739
	Q20. How worried have you been about not having enough company?	0.656
Non- and cross-loaders	Q9. Have you felt distressed?	Non
	Q13. Have you felt that there are things that you wanted to do but couldn't?	Non
	Q27. How worried have you been about how you feel in yourself?	Non
	Q28. How worried have you been about your health overall?	Non

response is not in line with the expected response (*Figure 2*). This can be assessed as item responses are mapped across the logit or severity scale and should be endorsed by respondents who display similar amounts of severity of the underlying construct. If respondents can distinguish between the response choices included on an instrument, responses to the items will be ordered as the observed response will be in line with the expected response at any point of the severity scale. If disordering occurs (i.e. the observed response is different to the expected response at that point of the logit scale), adjacent response levels are collapsed to artificially impose ordering, and the Rasch model is reapplied (*Figure 3*). Items for which disordering occurs are not considered for inclusion in the health-state classification system as it is important that respondents can distinguish between the different levels of the health dimensions used for the preference-based measure. Although items are excluded from further consideration for use as part of

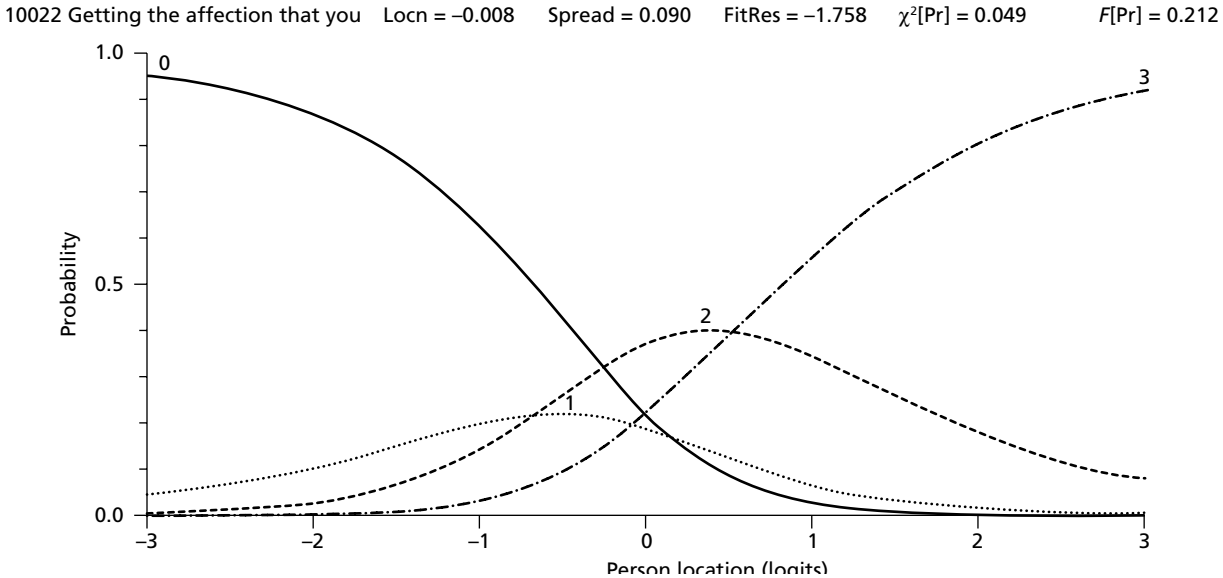


FIGURE 2 Item with disordered response options.

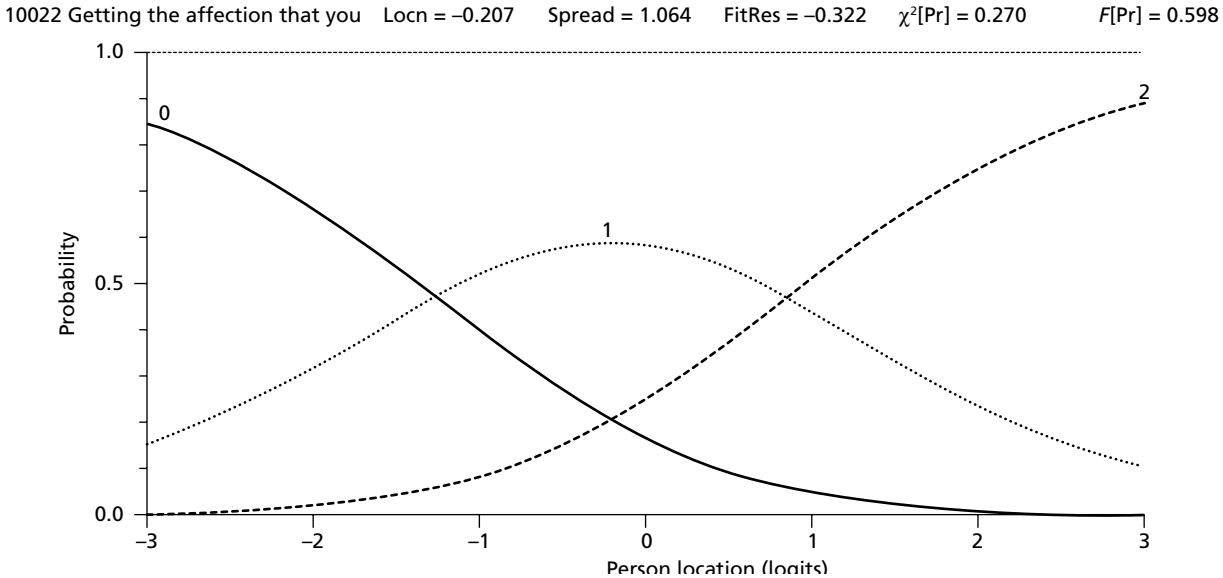


FIGURE 3 Item (from *Figure 2*) with reordered response options.

the health-state classification system they remain in the Rasch model to allow for the whole dimension to be modelled.

Examination of differential item functioning

We investigated each item for DIF. DIF occurs when item responses systematically differ according to a range of patient characteristics when equal amounts of the underlying characteristic are present (at different points of the latent scale). In this study the characteristics analysed were gender and age (split into two groups: < 65 and ≥ 65 years). We considered two categories of DIF – uniform and non-uniform. Uniform DIF occurs when one of the subgroups belonging to each characteristic consistently exhibits a difference in response across the full severity range of the construct being assessed by the item. For example, women may consistently display higher levels of depression as assessed by an individual item than men, and this is the case across the full severity scale of the underlying construct, from mild to

TABLE 5 DEMQOL-Proxy: factor analysis revalidation

Factor	Item		Loading
Cognition	Q14.	How worried would you say [patient] has been about forgetting things that happened recently?	0.755
	Q18.	How worried would you say [patient] has been about his/her thoughts being muddled?	0.695
	Q12.	How worried would you say [patient] has been about his/her memory in general?	0.661
	Q19.	How worried would you say [patient] has been about difficulty making decisions?	0.651
	Q15.	How worried would you say [patient] has been about forgetting people's names?	0.604
	Q17.	How worried would you say [patient] has been about forgetting what day it is?	0.575
	Q20.	How worried would you say [patient] has been about making him/herself understood?	0.471
	Q13.	How worried would you say [patient] has been about forgetting things that happened a long time ago?	0.457
	Q26.	How worried would you say [patient] has been about things taking longer than they used to?	0.430
Negative emotion	Q5.	Would you say that [patient] has felt sad?	0.687
	Q7.	Would you say that [patient] has felt distressed?	0.681
	Q10.	Would you say that [patient] has felt fed up?	0.666
	Q2.	Would you say that [patient] has felt worried?	0.632
	Q3.	Would you say that [patient] has felt frustrated?	0.618
	Q9.	Would you say that [patient] has felt irritable?	0.531
Daily activities	Q24.	How worried would you say [patient] has been about using money?	0.810
	Q25.	How worried would you say [patient] has been about looking after his/her finances?	0.655
	Q23.	How worried would you say [patient] has been about getting what he/she wants from the shops?	0.518
Positive emotion	Q8.	Would you say that [patient] has felt lively?	0.833
	Q4.	Would you say that [patient] has felt full of energy?	0.810
	Q11.	Would you say that [patient] has felt that there are things to look forward to?	0.454
Appearance	Q21.	How worried would you say [patient] has been about keeping him/herself clean?	0.772
	Q22.	How worried would you say [patient] has been about keeping him/herself looking nice?	0.720
Non- and cross-loaders	Q1.	Would you say that [patient] has felt cheerful?	Non
	Q6.	Would you say that [patient] has felt content?	Non
	Q16.	How worried would you say [patient] has been about forgetting where he/she is?	Cross
	Q27.	How worried would you say [patient] has been about getting in touch with people?	Cross
	Q28.	How worried would you say [patient] has been about not having enough company?	Cross
	Q29.	How worried would you say [patient] has been about not being able to help other people?	Cross
	Q30.	How worried would you say [patient] has been about not playing a useful part in things?	Cross
	Q31.	How worried would you say [patient] has been about his/her physical health?	Cross

severe. Non-uniform DIF occurs when responses between characteristic subgroups systematically diverge depending on the level of the attribute present. For example, responses between genders to a depression item may systematically differ depending on the underlying severity. Items for which either form of DIF is present are split into component factors (e.g. male and female subgroups) and the Rasch model is reapplied with each subgroup included separately. Items displaying DIF are not considered for the health-state classification as systematic differences in responses to the items across person characteristics compromises their ability to compare across subgroups.

Examination of goodness of fit

We evaluated the goodness of fit to the Rasch model of individual items included in each factor. Goodness of fit is assessed by analysing fit residuals and item–trait interactions. Fit residuals assess the discrepancy between the expected and observed responses at both the respondent and the item level. Divergence residuals $> |2.5|$ were considered high. Respondents outside these boundaries were also removed from the analysis and the model refitted. When all of the misfitting respondents had been removed, the fit of items was assessed. Items with residuals above the minimum level were excluded from the descriptive system. The overall mean fit residual for each dimension should be approximately 0 and the SD should be approximately 1.

Item–trait interactions measure overall differences between observed and expected responses for subgroups of responders (which are grouped dependent on where responders lie on the logit scale). The chi-squared test statistic (which is > 0.01 for a well-fitting model, i.e. non-significant) was used to assess item–trait interactions. Items with the highest level of significance (i.e. the largest divergence between the observed and expected responses) were removed one by one, with the Rasch model reapplied after every item was removed. This process continued until only items that fitted the model remained and the overall goodness of fit statistic was non-significant.

Item selection

To select one item per dimension two main criteria were used. We assessed the range of the item on the logit scale and the spread of the item at logit 0. A large range indicates that an item covers the full severity range of the underlying construct. Items were selected that incorporate both positive and negative values. This indicates that the item is sensitive to responses for both more severe and less severe respondents. In Rasch analysis, the latent scale is centred at zero. This point is the average difficulty on the latent scale relating to a particular item included in the final dimension model. Item levels should be distributed across the latent space depending on the severity of the response level. Therefore, a large spread of levels at logit 0 indicates items for which respondents endorse the full range of possible responses at the average severity or difficulty level. Item goodness of fit statistics and classical psychometric analyses, including MEFs, AEFs and missing data rates (alongside input from clinicians, dementia experts and the original instrument developers), were also used to guide the item selection process.

Stage III: exploration of item-level reduction

Stage III of the six-step guide to developing a condition-specific preference-based measure is to explore item-level reduction using Rasch analysis to examine whether or not respondents can distinguish between the response choices of the items selected for the health-state classification system. This was done by evaluating the ordering of the responses to each individual item selected using the item threshold probability curve.

Stage IV: validation of the classification system

It is recommended that the Rasch analysis described above is repeated on another sample before proceeding with the valuation study. To do this it is possible to use an independent data set or a subsample of the data used for stages I–III (if the sample size is large enough).

Results

Sample

The sample used for the missing data and MEF and AEF analyses was the same as the pre-imputation sample used in *Chapter 3* (DEMQOL: $n = 658$; DEMQOL-Proxy: $n = 692$). The sample used for the Rasch analyses was the post-imputation sample used to establish the dimensional structure of DEMQOL ($n = 644$) and DEMQOL-Proxy ($n = 683$), as described in *Chapter 3*.

DEMQOL

Cognition factor (six candidate items)

The responses to all six items were ordered and none of the items displayed evidence of DIF by either age group or gender (*Table 6*). Item 17 ('worry about thoughts being muddled') did not fit the model and was excluded. Of the five remaining items that could be included in the descriptive system, items 15 ('worry about forgetting who people are'), 16 ('worry about forgetting what day it is') and 18 ('worry about difficulty making decisions') displayed good fit to the Rasch model. However, the severity range covered by the items and the spread at logit 0 were lower than for items 14 ('worry about forgetting things that happened recently') and 19 ('worry about poor concentration') and so items 15, 16 and 18 were not included in the descriptive system. Item 14 was selected for the health-state classification as it displays the largest range and spread at logit 0 of the remaining items. It also measures a key characteristic of dementia and therefore had high face validity suggesting that conceptually it would be a valid item to use for the descriptive system.

Negative emotion factor (five candidate items)

The response categories were ordered on the logit scale and all items displayed good fit to the dimension-level Rasch model (*Table 7*). Item 11 ('felt irritable') showed evidence of uniform DIF by gender, with women scoring at a higher level irrespective of underlying severity. As the stem for most of the DEMQOL items asks, 'how worried have you been about …', we did not consider item 2 for the descriptive system because it asks directly about worry and we wanted to avoid double counting. The remaining items – item 4 ('felt frustrated'), item 7 ('felt sad') and item 12 ('felt fed up') – all cover a large range and spread at logit 0. Item 4 was selected as it was considered to be the most clinically relevant item, has low rates of missing data and displays acceptable MEF and AEF statistics.

Positive emotion factor (five candidate items)

The responses to all five items were ordered and none of the items displayed evidence of DIF. Items 3 ('felt that you are enjoying life'), 6 ('felt full of energy') and 10 ('felt lively') cover a high range of the logit scale and had high spread at logit 0 (*Table 8*). These items were excluded as items 1 ('felt cheerful') and 5 ('felt confident') displayed better statistics and are clearer constructs to allow for a better overall classification of positive emotion. Item 1 was selected on conceptual grounds as item 5 could be misconstrued as a personality trait rather than as a component of positive emotion. Item 1 also had the lowest missing data rates and displayed acceptable MEF and AEF statistics.

Social relationships factor (six candidate items)

The item responses of three items – item 21 ('worry about how you get on with people'), 22 ('worry about getting affection') and 25 ('worry about getting help') – were disordered between the responses 'quite a bit' and 'a lot', and item 26 ('worry about getting to the toilet on time') did not fit the Rasch model (*Table 9*). Of the remaining items – items 23 ('worry about people not listening to you') and 24 ('worry about making yourself understood') – neither displayed evidence of good severity coverage indicated by the item range and both items displayed AEF statistics below the minimum accepted level. Overall, however, item 24 covered more of the severe end of the logit scale and had a better fit to the model and was therefore selected for the classification system.

TABLE 6 Psychometric and Rasch results for DEMQOL cognition dimension

Item	Missing data	MEF	AEF	Responses ordered?	No DIF?	Good fit ($\chi^2 < 0.01$)	Item range	Fit residual	χ^2 *p*-value	Spread at logit
14. Worry about forgetting things that happened recently	✓	✓	✓	✓	✓	✓	–0.769 to 2.037	–0.158	0.740	0.12 to 0.68
15. Worry about forgetting who people are	✓	✓	✗	✓	✓	✓	–2.019 to 0.236	–0.551	0.385	0.44 to 0.88
16. Worry about forgetting what day it is	✓	✓	✓	✓	✓	✓	–0.780 to 1.287	–0.112	0.477	0.22 to 0.69
17. Worry about thoughts being muddled	✓	✓	✗	✓	✓	✗	N/A	N/A	N/A	N/A
18. Worry about difficulty making decisions	✓	✓	✓	✓	✓	✓	–1.196 to 0.800	–0.559	0.354	0.31 to 0.77
19. Worry about poor concentration	✓	✓	✓	✓	✓	✓	–0.641 to 1.905	–0.673	0.279	0.13 to 0.65

N/A, not applicable.

TABLE 7 Psychometric and Rasch results for DEMQOL negative emotion dimension

Item	Missing data	MEF	AEF	Responses ordered?	No DIF?	Good fit ($\chi^2 < 0.01$)	Item range	Fit residual	χ^2 *p*-value	Spread at logit
Q2. Felt worried	✓	✓	✓	✓	✓	✓	–1.262 to 1.564	1.164	0.030	0.17 to 0.78
Q4. Felt frustrated	✓	✓	✓	✓	✓	✓	–0.753 to 1.348	–0.626	0.155	0.21 to 0.68
Q7. Felt sad	✓	✓	✓	✓	✓	✓	–0.916 to 0.980	0.877	0.866	0.27 to 0.71
Q11. Felt irritable	✓	✓	✓	✓	✗	✓	N/A	N/A	N/A	N/A
Q12. Felt fed up	✓	✓	✓	✓	✓	✓	–0.979 to 1.400	–1.558	0.031	0.20 to 0.73

N/A, not applicable.

TABLE 8 Psychometric and Rasch results for DEMQOL positive emotion dimension

Item	Missing data	MEF	AEF	Responses ordered?	No DIF?	Good fit ($\chi^2 < 0.01$)	Item range	Fit residual	χ^2 *p*-value	Spread at logit
Q1. Felt cheerful	✓	✓	✓	✓	✓	✓	–3.104 to 2.427	1.208	0.165	0.08 to 0.96
Q3. Felt that you are enjoying life	✓	✓	✓	✓	✓	✓	–2.757 to 2.104	–0.267	0.348	0.11 to 0.94
Q5. Felt confident	✓	✓	✓	✓	✓	✓	–2.747 to 2.843	0.839	0.915	0.06 to 0.94
Q6. Felt full of energy	✓	✓	✓	✓	✓	✓	–1.387 to 2.474	–0.907	0.326	0.08 to 0.80
Q10. Felt lively	✓	✓	✓	✓	✓	✓	–1.421 to 2.962	–1.575	0.181	0.05 to 0.81

Loneliness factor (two candidate items)

The response categories were ordered and neither item displayed DIF by gender or age group (*Table 10*). Item 20 ('worry about not having enough company') had higher spread at logit 0 but displayed AEF statistics below the minimum accepted level. Item 8 ('felt lonely') covered more of the severe end of the scale, asking directly about loneliness rather than a particular aspect of loneliness, and displayed acceptable MEF and AEF statistics; therefore, item 8 was chosen for the classification system.

DEMQOL-Proxy

Cognition (nine candidate items)

Item 20 ('worry about making self understood') was disordered between the response levels 'quite a bit' and 'a lot' and so was excluded from the descriptive system (*Table 11*). Item 12 ('worry about memory in general') displayed evidence of uniform DIF by age, with those <65 years scoring lower than those aged ≥ 65 years across the full logit scale. Items 14 ('worry about forgetting things that happened recently'), 15 ('worry about forgetting people's names'), 18 ('worry about thoughts being muddled') and 19 ('worry about difficulty making decisions') did not fit the model and so were removed from the descriptive system. Of the remaining items, item 13 ('worry about forgetting things that happened a long time ago') was excluded as it assesses a memory problem that occurs late in the course of dementia; this was supported by the small range and large ceiling effect displayed by the item. Of the two remaining items, item 17 ('worry about forgetting what day it is') was selected for the descriptive system as the context of the item is more relevant to proxy reporting of memory and cognition than item 26 ('worry about things taking longer than they used to'). The chosen item did not display high rates of missing data and had acceptable MEF and AEF statistics.

Negative emotion (six candidate items)

All of the items were ordered and none displayed evidence of DIF. Item 9 ('felt irritable') displayed poor fit to the model and was removed from the descriptive system (*Table 12*). Item 2 ('felt worried') was also excluded to avoid double counting as the stem to most of the DEMQOL-Proxy items uses the phrase, 'how worried have you been about …'. Items 5 ('felt sad') and 10 ('felt fed up') were also excluded as the content of the items may reflect on aspects of comorbid depression, which is prevalent in those with dementia.[65,66] Of the remaining items, item 7 ('felt distressed') had low item fit and both items 7 and 3 ('felt frustrated') displayed missing data rates slightly above the minimum accepted level. Item 3 was chosen for the descriptive system as it displayed strong fit statistics and good range and spread and was also repeated across both the DEMQOL and the DEMQOL-Proxy descriptive systems.

Daily activities (three candidate items)

Item 23 ('worry about getting what he/she wants from the shops') displayed disordering between 'quite a bit' and 'a lot', and items 24 ('worry about using money') and 25 ('worry about looking after his/her finances') did not fit the Rasch model (*Table 13*). Therefore, no items remained for this factor and it could not be included in the classification system for DEMQOL-Proxy.

Positive emotion (three candidate items)

All three items were ordered and there was no evidence of DIF but item 11 ('felt that there are things to look forward to') displayed poor fit to the Rasch model (*Table 14*). Of the remaining items – item 4 ('felt full of energy') and item 8 ('felt lively') – item 8 displayed considerably better range, spread and fit statistics and was therefore chosen for the descriptive system.

Appearance (two candidate items)

Both items were ordered and there was no DIF. The fit of item 21 ('worry about keeping self clean') was approaching significance, but this item displayed lower range and spread statistics and higher MEF and AEF than item 22 ('worry about keeping self looking nice'); therefore, as it performed better on all indicators, item 22 was chosen for the descriptive system (*Table 15*).

TABLE 9 Psychometric and Rasch results for DEMQOL social relationships dimension

Item	Missing data	MEF	AEF	Responses ordered?	No DIF?	Good fit ($\chi^2 < 0.01$)	Item range	Fit residual	χ^2 *p*-value	Spread at logit
Q21. Worry about how you get on with people close to you	✓	✓	✗	✗	✓	✓	N/A	N/A	N/A	N/A
Q22. Worry about getting the affection that you want	✓	✓	✗	✗	✓	✓	N/A	N/A	N/A	N/A
Q23. Worry about people not listening to you	✓	✓	✗	✓	✓	✓	–0.681 to 1.025	–1.754	0.027	0.26 to 0.66
Q24. Worry about making yourself understood	✓	✓	✗	✓	✓	✓	–0.205 to 1.280	–0.087	0.472	0.22 to 0.55
Q25. Worry about getting help when you need it	✓	✓	✗	✗	✓	✓	N/A	N/A	N/A	N/A
Q26. Worry about getting to the toilet on time	✓	✓	✗	✓	✓	✗	N/A	N/A	N/A	N/A

N/A, not applicable.

TABLE 10 Psychometric and Rasch results for DEMQOL loneliness dimension

Item	Missing data	MEF	AEF	Responses ordered?	No DIF?	Good fit ($\chi^2 < 0.01$)	Item range	Fit residual	χ^2 *p*-value	Spread at logit
Q8. Felt lonely	✓	✓	✓	✓	✓	✓	–1.044 to 2.309	0.293	0.256	0.09 to 0.74
Q20. Worry about not having enough company	✓	✓	✗	✓	✓	✓	–2.228 to 1.624	1.154	0.153	0.16 to 0.90

TABLE 11 Psychometric and Rasch results for DEMQOL-Proxy cognition dimension

Item	Missing data	MEF	AEF	Responses ordered?	No DIF?	Good fit ($\chi^2 < 0.01$)	Item range	Fit residual	χ^2 *p*-value	Spread at logit
Q12. Worry about his/her memory in general	✓	✓	✓	✓	✗	✓	N/A	N/A	N/A	N/A
Q13. Worry about forgetting things that happened a long time ago	✓	✗	✓	✓	✓	✓	–2.203 to –0.904	–0.967	0.175	0.71 to 0.90
Q14. Worry about forgetting things that happened recently	✓	✓	✓	✓	✓	✗	N/A	N/A	N/A	N/A
Q15. Worry about forgetting people's names	✓	✓	✓	✓	✓	✗	N/A	N/A	N/A	N/A
Q17. Worry about forgetting what day it is	✓	✓	✓	✓	✓	✓	–0.947 to 0.324	–0.628	0.211	0.42 to 0.72
Q18. Worry about his/her thoughts being muddled	✓	✓	✓	✓	✓	✗	N/A	N/A	N/A	N/A
Q19. Worry about difficulty making decisions	✓	✓	✓	✓	✓	✗	N/A	N/A	N/A	N/A
Q20. Worry about making him/herself understood	✓	✓	✓	✗	✓	✓	N/A	N/A	N/A	N/A
Q26. Worry about things taking longer than they used to	✓	✓	✓	✓	✓	✓	–1.75 to 0.52	0.555	0.393	0.52 to 0.85

N/A, not applicable.

TABLE 12 Psychometric and Rasch results for DEMQOL-Proxy negative emotion dimension

Item	Missing data (%)	MEF	AEF	Responses ordered?	No DIF?	Good fit ($\chi^2 < 0.01$)	Item range	Fit residual	χ^2 *p*-value	Spread at logit
Q2. Felt worried	✓	✓	✓	✓	✓	✓	–1.287 to 2.284	0.823	0.976	0.10 to 0.79
Q3. Felt frustrated	✗	✓	✓	✓	✓	✓	–1.007 to 1.923	1.126	0.695	0.15 to 0.76
Q5. Felt sad	✗	✓	✓	✓	✓	✓	–2.005 to 1.411	–0.902	0.091	0.22 to 0.84
Q7. Felt distressed	✗	✓	✓	✓	✓	✓	–2.081 to 0.655	–1.614	0.014	0.34 to 0.89
Q9. Felt irritable	✓	✓	✓	✓	✓	✗	N/A	N/A	N/A	N/A
Q10. Felt fed up	✗	✓	✓	✓	✓	✓	–1.243 to 2.047	0.624	0.606	0.11 to 0.78

N/A, not applicable.

TABLE 13 Psychometric and Rasch results for DEMQOL-Proxy daily activities dimension

Item	Missing data (%)	MEF	AEF	Responses ordered?	No DIF?	Good fit ($\chi^2 < 0.01$)	Item range	Fit residual	χ^2 *p*-value	Spread at logit
Q23. Worry about getting what he/she wants from the shops	✗	✓	✗	✗	✓	✓	N/A	N/A	N/A	N/A
Q24. Worry about using money	✗	✓	✗	✓	✓	✗	N/A	N/A	N/A	N/A
Q25. Worry about looking after his/her finances	✗	✓	✓	✓	✓	✗	N/A	N/A	N/A	N/A

N/A, not applicable.

Stage III: exploration of item-level reduction

The item threshold probability curves indicated that patients were able to distinguish across the four responses that form the levels of the chosen items of the two descriptive systems. Therefore, it was not necessary to explore item-level reduction further.

Stage IV: validation

It was not possible in this study to validate the classification systems developed either on a subsample of the data or on an external data set. This is because randomly splitting the data used in this chapter into two samples would have resulted in a sample size that was smaller than the optimum recommended for Rasch analysis ($n = 500$). Furthermore, as Rasch excludes extreme scores (in which respondents answer at the most severe or least severe level for the overall dimension), the sample size for the dimensions with two to three items would be reduced further. At the time of development, no external data set including DEMQOL or DEMQOL-Proxy was available to validate the classification systems.

Final health-state classification systems

The 10 DEMQOL and DEMQOL-Proxy dimension Rasch models from which the items were selected all displayed overall goodness of fit (*Table 16*). The final health-state classification systems that were developed following the item selection process outlined in the sections above are displayed in *Table 17* (with health states ordered as they appear in the valuation study). The five items selected to reflect the DEMQOL dimension structure form the basis of the DEMQOL classification system. This was named DEMQOL-U. Each dimension has four response levels that correspond to the options included on the original DEMQOL instrument. Therefore, the DEMQOL-U descriptive system generates a possible 1024 (i.e. 4^5) health states.

The four items selected to reflect the DEMQOL-Proxy dimension structure form the basis of the DEMQOL-Proxy classification system. This was named DEMQOL-Proxy-U. The four-dimension structure each with four response levels means that DEMQOL-Proxy-U generates 256 (i.e. 4^4) health states.

Discussion

This chapter describes the development of condition-specific descriptive systems for dementia for patient self-report (DEMQOL-U) and carer proxy report (DEMQOL-Proxy-U). We selected items to represent all of the five DEMQOL dimensions and four of the five DEMQOL-Proxy dimensions established during stage I of the development process. The methodology used builds on previous work that has applied Rasch analysis to non-preference-based condition-specific instruments to develop condition-specific classification systems that are amenable to valuation.[64,67] This is the first part of the process in developing a condition-specific preference-based measure. The second part is to obtain preference weights so that the measure can be used in the economic evaluation of interventions for dementia. This stage is described in *Chapter 5*.

This chapter also describes the first attempt to derive a condition-specific health-state classification system specifically for proxy report by carers. The descriptive systems reflect both similarities and differences in the factors that are the key focus for patients and carers in terms of evaluating HRQL in dementia. For example, both measures included cognition and emotion dimensions and this reflects the importance of cognitive functioning and mood both for the person with dementia and for those involved in their care. There are discrepancies between patient and proxy report of HRQL in dementia[68] and some evidence suggests that agreement is lower for more subjective domains such as emotional well-being.[69] The development of a proxy-specific measure in which all of the dimensions are meaningful to carers may address some of the concerns about using generic measures for proxy report by carers in dementia.

TABLE 14 Psychometric and Rasch results for DEMQOL-Proxy positive emotion dimension

Item	Missing data (%)	MEF	AEF	Responses ordered?	No DIF?	Good fit ($\chi^2 < 0.01$)	Item range	Fit residual	χ^2 *p*-value	Spread at logit
Q4. Felt full of energy	✗	✓	✓	✓	✓	✓	–2.694 to 2.545	0.221	0.142	0.07 to 0.94
Q8. Felt lively	✗	✓	✓	✓	✓	✓	–3.243 to 3.276	0.629	0.735	0.04 to 0.96
Q11. Felt that there are things to look forward to	✗	✓	✓	✓	✓	✗	N/A	N/A	N/A	N/A

N/A, not applicable.

TABLE 15 Psychometric and Rasch results for DEMQOL-Proxy appearance dimension

Item	Missing data (%)	MEF	AEF	Responses ordered?	No DIF?	Good fit ($\chi^2 < 0.01$)	Item range	Fit residual	χ^2 *p*-value	Spread at logit
Q21. Worry about keeping self clean	✓	✗	✗	✓	✓	✓	–0.782 to 0.703	0.310	0.044	0.33 to 0.69
Q22. Worry about keeping self looking nice	✓	✗	✓	✓	✓	✓	–1.630 to 1.452	0.903	0.938	0.19 to 0.84

TABLE 16 Goodness of fit to the Rasch model for each dimension

Dimension	χ^2 (df)	*p*-value	Item fit (SD)	Person fit (SD)	PSI
DEMQOL					
Cognition	29.54 (30)	0.49	–0.35 (0.36)	–0.30 (0.88)	0.78
Positive emotion	75.44 (51)	0.01	–0.01 (1.01)	–0.23 (0.81)	0.75
Negative emotion	49.85 (45)	0.29	–0.14 (1.17)	–0.45 (1.15)	0.79
Relationships	34.10 (25)	0.11	–0.07 (0.99)	–0.19 (0.77)	0.73
Loneliness	10.74 (7)	0.15	0.72 (0.61)	–0.33 (0.71)	0.73
DEMQOL-Proxy					
Cognition	43.79 (39)	0.28	–0.40 (0.64)	–0.27 (0.75)	0.64
Negative emotion	56.11 (40)	0.05	0.10 (1.27)	–0.28 (0.90)	0.81
Positive emotion	9.66 (9)	0.38	0.43 (0.29)	0.76 (1.14)	0.78
Appearance	10.56 (8)	0.23	0.61 (0.42)	–0.50 (0.98)	0.72

df, degrees of freedom; PSI, person separation index.

TABLE 17 Health-state classification systems: DEMQOL and DEMQOL-Proxy

DEMQOL	DEMQOL-Proxy
Positive emotion	***Positive emotion***
1. I feel cheerful a lot	1. I feel lively a lot
2. I feel cheerful quite a bit	2. I feel lively quite a bit
3. I feel cheerful a little	3. I feel lively a little
4. I do not feel cheerful at all	4. I do not feel lively at all
Cognition	***Cognition***
1. I do not worry at all about forgetting things that happened recently	1. I do not worry at all about forgetting what day it is
2. I worry a little about forgetting things that happened recently	2. I worry a little about forgetting what day it is
3. I worry quite a bit about forgetting things that happened recently	3. I worry quite a bit about forgetting what day it is
4. I worry a lot about forgetting things that happened recently	4. I worry a lot about forgetting what day it is
Relationships	***Appearance***
1. I do not worry at all about making myself understood	1. I do not worry at all about keeping myself looking nice
2. I worry a little about making myself understood	2. I worry a little about keeping myself looking nice
3. I worry quite a bit about making myself understood	3. I worry quite a bit about keeping myself looking nice
4. I worry a lot about making myself understood	4. I worry a lot about keeping myself looking nice
Negative emotion	***Negative emotion***
1. I do not feel frustrated at all	1. I do not feel frustrated at all
2. I feel frustrated a little	2. I feel frustrated a little
3. I feel frustrated quite a bit	3. I feel frustrated quite a bit
4. I feel frustrated a lot	4. I feel frustrated a lot
Loneliness	
1. I do not feel lonely at all	
2. I feel lonely a little	
3. I feel lonely quite a bit	
4. I feel lonely a lot	

The results of the Rasch analysis used to select the items for the descriptive system have not been validated on an external sample and this is a limitation of the development process described in this chapter. It was not possible to carry out validation analyses as the sample size was not sufficient to randomly allocate responses to two subgroups. Furthermore, the data used in *Chapter 7* of this report to assess the psychometric performance of DEMQOL-U and DEMQOL-Proxy-U were not available to validate the descriptive system. The DEMQOL system is being used in a number of studies and so further work validating the classification systems may be possible in the future.

It is also possible to carry out factor analysis using IRT-based techniques and this would enable us to combine stage I of the development process with stages II and III and base the full classification system development on one underlying model. This approach and the differences in factor structure that might

lead from it need further investigation. Indeed, other IRT techniques could be used to select items for the classification system and further research could investigate differences in the items selected using a range of techniques and determine which technique produces the classification system that retains the most information from the original HRQL measure.

There are also concerns around the use of condition-specific preference-based measures, including the extent to which they capture comorbidities.[39] This issue may be addressed by investigating the performance of condition-specific preference-based measures in relation to generic measures such as the EQ-5D in trials and settings where both instruments have been used together. Both DEMQOL-U and DEMQOL-Proxy-U cover a broad range of HRQL issues in dementia. The absence of an activity limitation dimension may be a concern; this was because the Rasch analysis found that the original daily activities items did not meet the minimum threshold for inclusion. Another important limitation is that it has not been possible to consider item responsiveness during the development of the descriptive system. The ability of items to detect change over time is an important psychometric characteristic and so responsiveness needs to be addressed in future research that includes DEMQOL and DEMQOL-Proxy as outcome measures at multiple time points.

It could also be argued that the resultant classification systems reflect a mild description of dementia. This is a result of both the development process (in which the items that performed well reflected mild aspects of HRQL in dementia) and also the sample used, which was patients with a mild or moderate diagnosis. Future work could investigate qualitatively the acceptability and characteristics of the classification system with independent clinical experts and patient groups.

There are a number of strengths of the process described in this chapter. The descriptive systems were based on representative samples from memory and community services where many of the clients have mild to moderate dementia. Although it is widely acknowledged that people with dementia are often able to self-report it is often also necessary to use a proxy report.[31] Administering measures in an interview setting with response cards (as is done with the DEMQOL system) may help to maximise the reliability and validity of such reports. The data also included demographic information that enabled us to investigate DIF characteristics and this helped to strengthen the item selection process and provide a set of descriptive systems that cover as broad a range of HRQL issues in dementia based on the non-preference-based instruments as possible.

Conclusion

In conclusion, item selection procedures have identified robust items to represent all five of the DEMQOL dimensions and four of the five DEMQOL-Proxy dimensions. Using validated measures (DEMQOL and DEMQOL-Proxy) along with a representative sample of patients and carers we have created a classification system that is an appropriate representation of HRQL in dementia. The descriptive systems are amenable to valuation and the next part of the process in developing a condition-specific preference-based measure is to obtain preference weights so that the measure can be used in the economic evaluation of interventions in dementia.

Chapter 5 General population valuation survey and modelling to produce values for every health state: estimating preference-based single-index measures for dementia

Introduction

Cost–utility analysis measures the benefits of treatment using the QALY, which captures changes in both quantity of life and HRQL. The 'Q' quality adjustment weight is typically derived using a preference-based measure. A preference-based measure has two components: first, a health-state classification system that can be used to categorise all patients with the condition of interest; second, a means of obtaining a utility score for all states defined by the system. This chapter reports on work to estimate a preference-based single index for each classification system. First, this chapter reports the valuation study in which a representative sample of the general population valued a sample of health states derived from each classification system using ranking and the TTO elicitation technique. Second, the chapter reports on modelling using the valuation results to produce utility values for all health states described by each classification system. This analysis uses a range of multivariate regression models to produce a single-index measure from each classification system anchored on a full health–dead 1–0 scale, in which a value of 1 is equal to full health and 0 is equal to being dead. Values were obtained from the general population in accordance with recommendations of agencies such as NICE[43,54] and the Washington Panel[70] for use in economic evaluation.

Method

Health-state descriptions for the valuation study were generated using the DEMQOL-U and DEMQOL-Proxy-U classification systems described in *Chapter 4*. *Table 17* presents an overview of the two classification systems. Each health state is generated using one level of each dimension. *Box 1* includes an example health state for each classification system.

The DEMQOL-U health-state classification describes 1024 health states and the DEMQOL-Proxy-U classification describes 256 health states. The large number of health states means that it is impractical to value every state and therefore only a sample of states were selected for valuation. Respondents cannot value a large number of health states because of the nature of the preference elicitation task. In accordance with many previous valuation studies,[37,40,42,67,70–72] each respondent valued eight health states. The study design needs to determine both the total sample of all health states to be included in the valuation study and the combinations of eight health states to be valued by respondents. To get more precise estimates of the worst state and to have a common state valued by all respondents, each combination of health states for valuation included the worst state plus seven other states. Previous valuation surveys have used a variety of methods to select health states for valuation, such as an orthogonal array,[37,71] balanced design[40,42] and the Rasch vignette approach.[66,73] None of these approaches selects the combinations of health states to be valued by respondents, and only the Rasch vignette approach avoids implausible health states. Here, a new approach was used to select health states, which offered the following advantages over existing approaches: (1) it selected the combinations of health states to be valued by respondents; (2) it avoided implausible health states; and (3) it selected the health states that produce the most accurate modelled utility estimates using simulation. The last advantage is gained because the process involves simulation of alternative selections of health states that could be

BOX 1 Example health states

DEMQOL-U state 23424

You feel cheerful *quite a bit*

You worry *quite a bit* about forgetting things that happened recently

You worry *a lot* about making yourself understood

You feel frustrated *a little*

You feel lonely *a lot*

DEMQOL-Proxy-U state 1341

You feel lively *a lot*

You worry *quite a bit* about forgetting what day it is

You worry *a lot* about keeping yourself looking nice

You do *not* feel frustrated *at all*

chosen and hypothetical values for those health states (given some assumptions about how people value health states). The hypothetical values were modelled using similar modelling to the analysis that was applied to the actual utility values in the general population valuation survey. The accuracy of the model predictions for each of the alternative selections of health states was compared using predictions for health states that would not be valued. The best performing selection of health states was chosen. For each selection of health states these are separated into combinations of health states for valuation. Each combination is called a block and each block consisted of seven health states of different severity plus the worst state. The steps used in this process are explained further below; the process was undertaken for each classification system.

The selection of health states can involve a different number of blocks, for example five blocks of seven health states plus the worst state, or six blocks of seven health states plus the worst state, or seven blocks, etc. Furthermore, within each block design, alternative health states could be selected, for example there are many different combinations of health states that could be selected within the five-block design, within the six-block design, etc. Therefore, the first step in the process derived 100 approximate optimal Federov designs[74] for each block design of 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15 blocks (i.e. 100 designs for five blocks, 100 designs for six blocks, etc.). For each block size the optimal design was selected using the D-criterion and the worst state was added to each block. There were 11 designs remaining for inclusion in step 2, one for each block design.

Assuming that the sample size of the valuation study will remain the same, this means that each health state in the 15-block design will be valued fewer times than health states in the 14-block design. It is expected that each block design will involve different health states as well as different numbers of health states in total (e.g. the five-block design will have 35 health states and the six-block design will have 42 health states, assuming that there is no repetition in health states). All of these factors may impact on the reliability and robustness of the modelled utility values. Therefore, the second step in the process involved the simulation of hypothetical utility values for each of the block designs selected in step 1. Utility values were simulated for each block design for 50 sets of 300 individuals using utility functions. The utility function assumed a logarithmic transformation between 'severity' and utility in which severity of a health state was generated using a linear function of the dimensions with equal weights for each dimension and equidistance between levels with random measurement errors. As respondents do not always trade

life-years to avoid mild health states, non-trading behaviour was simulated using a logistic model of severity [such that approximately 10% of the scores would be 1 to represent non-traders (proportion selected using previous valuation studies)].

The third step in the process involved the estimation of a linear regression model for each block design using the appropriate simulated utility data set. The regression model was used to produce predictions of health-state utility values for all health states not selected in the block design. These predictions were compared with the 'true values', the hypothetical values that were generated using the same utility functions explained above, using the sum of squared residuals and average mean sum of squares.

Fourth, the block design was selected that achieved the best predictions of health-state utility values for health states outside the set of health states to be valued (i.e. all health states that were not selected in that block design).

The selected designs involved a sample of 87 unique states across 13 blocks for DEMQOL-U and a sample of 70 unique states across 12 blocks for DEMQOL-Proxy-U. For each measure one block was selected to be valued by a minimum of 71 individuals to enable mean values for these states to be statistically compared with mean values elicited from people with dementia and carers of patients with dementia as reported in *Chapter 6*.

Valuation study

Sample

Each classification system was valued by a representative sample of the general population. This is in accordance with recommendations by agencies such as NICE[43,54] and the Washington Panel,[70] who recommend the use of general population values to produce QALY estimates for economic evaluation. The sample size for each measure was selected on the basis of previous valuation studies for similar measures [e.g. for the cancer-specific EORTC-8D (European Organisation for Research and Treatment of Cancer Core Quality of Life Questionnaire) a sample of 344 respondents valued 85 states; for the asthma-specific AQL-5D (Asthma Quality of Life Utility Index) a sample of 307 respondents valued 99 states; and for the overactive bladder-specific OAB-5D (Overactive Bladder Questionnaire-5 Dimensions) a sample of 312 respondents valued 99 states]. The sample size was selected as 310 for DEMQOL-U and 290 for DEMQOL-Proxy-U. The sample size for DEMQOL-U was larger as the classification system describes more states and had a larger selected study design.

The sample was obtained by sampling 600 households in urban and rural areas in northern England using the AFD Names and Numbers version 3.1.25 database (AFD Software Limited, Ramsey, UK). The sample was balanced to the UK population according to geodemographic profiles. Trained and experienced interviewers conducted home interviews in which respondents valued states from one of the classification systems determined using a card block system. Interviewers worked systematically through blocks; odd and even blocks contained DEMQOL-U and DEMQOL-Proxy-U states respectively. This approach was used to try and ensure that there were no systematic differences across the geodemographic profiles of the samples for each classification system.

Procedure

At the start of the interview respondents first self-completed the EQ-5D to report their own health and then whichever classification system would be used subsequently in the valuation task (DEMQOL-U or DEMQOL-Proxy-U). Respondents were asked to complete the questions for their own health and this was done to help familiarise them with the classification system. Respondents were not informed that the health state was associated with dementia as there are concerns that naming the condition can affect elicited utility values. A recent study found that introducing condition labels into health-state descriptions impacted on health-state utility values elicited from the general population.[37,71] The impact differed by condition and health-state severity but as the study was quantitative the reasoning behind why these

differences occurred was not determined. Further qualitative research is recommended to examine why these differences occurred and until these findings are available it is recommended that condition labels are not included in health-state descriptions to ensure that elicited utility values are not affected by factors that should not affect the quality of life of health states.

Second, respondents ranked eight health states, as well as full health and dead, in order from the best to the worst. This rank task further familiarised respondents with the classification system and the health states to be valued later using TTO.

Third, respondents undertook a practice TTO task using a hypothetical 'practice' state and subsequently valued the (ranked) eight health states using TTO. The intuition underlying the TTO task is that it is designed to determine whether or not respondents are willing to sacrifice time to avoid living in the dementia health state. Essentially this is undertaken by asking respondents whether they would rather live in the dementia health state for 10 years or full health for a shorter time duration than 10 years. If respondents believe that the health state is severe, they should be willing to sacrifice a larger time duration than if they believe that the health state is mild. For some respondents they may not be willing to sacrifice any time at all to avoid the dementia health state. Some respondents may believe that some health states are so severe that they would rather die than live in that health state.

The TTO task for each health state asks first whether respondents prefer to live in a given health state for 10 years after which they will die, or to die immediately. This determines whether the respondent values the health state as better, worse or equal to being dead. For health states considered better than dead, respondents choose between (a) the health state for 10 years, after which they will die, or (b) full health for x years ($x \leq 10$), after which they will die. Years in full health, x, is varied until respondents are indifferent between the two options. For health states considered worse than being dead, respondents choose between (a) the health state for w years followed by full health for x years, after which they will die, or (b) immediate death. Both years in full health, x, and years in the health state, $w = 10 - x$, are varied until respondents are indifferent between the two options. Utility values are generated using the formula $x/10$ for states better than being dead and $-x/10$ for states worse than being dead.[45] A sample TTO scripting for one TTO task is included in *Appendix 1*. The Measurement and Valuation of Health (MVH) study version of TTO was used, including a visual prop designed by the MVH group (University of York),[45] which has been reproduced in *Figure 4* (although note that in the MVH study respondents valued more than eight health states). All TTO questions follow an iterative procedure to determine the point of indifference, with the task completed only when the respondent switches his or her choice between (a) and (b) or reaches the point at which he or she is indifferent between (a) and (b). The TTO elicitation technique was selected in accordance with the UK valuation of the EQ-5D,[45] which also meets the reference case recommended by NICE[54] for use in economic evaluation.

Finally, respondents rated how difficult they found the rank and TTO tasks and answered questions about sociodemographic characteristics and health service use.

Before conducting the full survey a small number of interviews were conducted and the interviewer discussed the findings with one of the research team. As no problems were identified with the survey, the main survey was started without amendment and these interviews were included in the main sample. The survey was approved by the School of Health and Related Research (ScHARR) Research Ethics Committee at the University of Sheffield.

Analysis of time trade-off data

Descriptive statistics of health-state utility values for all health states included in the valuation study are presented and plotted for each measure.

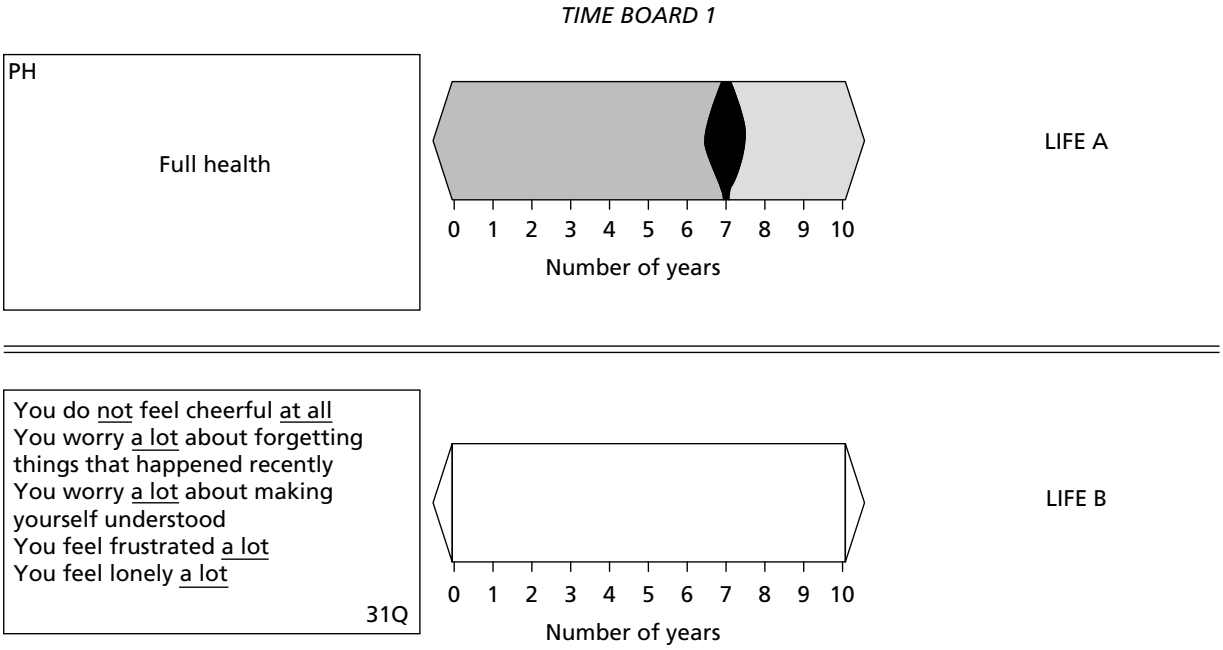


FIGURE 4 Time trade-off visual prop [MVH group (4)]. Progeny of the University of Sheffield – 58 © Copyright MVH Group University of York, 1992.

Modelling to obtain preference weights for the health-state classification

The valuation study elicited health-state utility values for a sample of health states described by each classification system. These values need to be modelled in order to produce estimated health-state utility values for every state defined by each classification system. Regression models were used to estimate a preference weight, known as a 'utility decrement', for every severity level of every dimension of the classification system using level 1 as the baseline. These models assume that full health is assigned a value of 1 and instrument-specific full health is assigned a value of 1 minus the constant term. These models can be used to estimate a utility value for each state, for example DEMQOL-U state 21111 has an estimated utility value of 1 minus the constant term and the utility decrement for level 2 of the positive emotion dimension, and DEMQOL-U state 22111 has an estimated utility value of 1 minus the constant term and the utility decrement for level 2 of the positive emotion dimension and level 2 of the memory dimension.

The regression models used the following general specification:

$$U_{ij} = g(\beta' \mathrm{x}_i) + \varepsilon_{ij} \quad (1)$$

where U represents TTO disvalue (generated as 1 – TTO), $i = 1, 2, \ldots, n$ represents individual health states, $j = 1, 2, \ldots, m$ represents respondents, g is a function specifying the appropriate form, x_i is a vector of binary dummy variables for each level λ of dimension δ of the descriptive system in which the best level of each dimension represents the baseline for that dimension, and ε_{ij} is an error term, whose properties depend on the assumptions of the model (see below).

A variety of mean- and individual-level multivariate regression models were fitted to the valuation data for each classification system. First, models were estimated on observational-level data using OLS. However, OLS does not take into account the structure of the data, as there are repeated observations for each individual as each individual valued eight health states. Random- and fixed-effects models were estimated using GLS maximum likelihood estimation in order to take into account the structure of the data with

repeated observations for each individual.[25] For the random-effects model the error term, ε_{ij}, is subdivided as follows:

$$\varepsilon_{ij} = u_j + e_{ij} \quad (2)$$

where u_j represents the individual random effect and e_{ij} represents the random error term for the *i*th health-state valuation of the *j*th individual. The choice of whether to use the random- or fixed-effects specification depends on the sample design and the purpose of the study. In this valuation study respondents were randomly sampled and the assumptions of the random-effects specification are met when we assume that any differences in valuations across individuals are random. Fixed effects can be interpreted as using dummy variables for each individual, yet here we do not have reason to believe that each individual will be sufficiently distinct. However, the choice will be determined empirically using the Hausman test.

Time trade-off valuation data typically have a large proportion of values at 1, where individuals are not prepared to trade any time to avoid living in the impaired dementia health state. This means that the data can be interpreted as being bounded at 1 and the data are further bounded at –1, although there are typically fewer observations at –1 than at 1. This structure of the data is not taken into account using the OLS and GLS models. A random-effects Tobit model was estimated to take account of the bounded nature of the data; however, the Tobit models had poorer performance and predictive ability and are not reported here (results available from authors on request). Models were also estimated using OLS on mean-level data, with the data consisting of one mean value per state ($n = 87$ for DEMQOL-U, $n = 70$ for DEMQOL-Proxy-U).

Performance of the regression models was assessed using the number of inconsistent coefficients and number of significant coefficients. Predictive ability was assessed using root-mean-squared error (RMSE) of health-state predictions at the health-state level, mean absolute error (MAE) at the state level, number of states with absolute error (AE) $> 5\%$ and $> 10\%$, plots of actual and predicted health-state utility values and the Ljung–Box (LB) test to examine autocorrelation. Autocorrelation is present if there is a correlation between the errors when these are ordered by observed mean health-state utility value. The error indicates the difference between observed and predicted utility values.

Analysis of rank data

Modelling to obtain preference weights for the health-state classification

Regression analysis was also used to estimate preference weights using the rank data. The rank data need to be modelled in order to produce estimated health-state utility values for every state defined by each classification system. The rank data were analysed using the rank-ordered logit model (also referred to as the conditional logit model).[39] The model states that individual *j* has a latent utility function for health state *i*, which is U_{ij}. When the individual is asked to rank two states *j* and *k*, the individual will rank state *i* over state *k* if the utility of state *i* is higher than the utility of state *k*: $U_{ij} > U_{kj}$.

The model specification for rank data has a similar format to the model specification used for the TTO data in *Equation 1*. However, unlike TTO data, rank data are not anchored onto the 1–0 full health–dead utility scale required for the estimation of QALYs. Instead, modelled rank data lie on an unanchored latent utility scale. To anchor the latent scale onto the 1–0 full health–dead utility scale, two approaches are used in the literature. The first approach excludes the data from rankings of the state 'dead' and estimates *Equation 1*.[75] The estimated coefficients are normalised onto the full health–dead scale using the estimated TTO value of the worst state. This means that the value of the worst state in the rank model is anchored at the value of the worst state in the preferred TTO model. The second approach includes the state 'dead' in the regression equation and its coefficient is used to anchor all other coefficients. The general model specification for analysis of the ranking data using the second approach is:[75,76]

$$U_{ij} = f(\beta' x_i + \phi_D) + \varepsilon_{ij} \quad (3)$$

where U represents utility, $i = 1, 2, \ldots, n$ represents individual health states, $j = 1, 2, \ldots, m$ represents respondents, f is a function specifying the appropriate form, x_i is a vector of binary dummy variables for each level λ of dimension δ of the descriptive system in which the best level of each dimension represents the baseline for that dimension, D is a dummy variable for the state 'dead' (equals 1 for 'dead' and 0 for all other states) and ε_{ij} is an error term, whose properties depend on the assumptions of the model (see below). To anchor the values onto the 1–0 full health–dead scale the coefficients for the levels of each dimension are normalised by dividing each level coefficient by the coefficient relating to dead: $\beta_{r\lambda\delta} = \beta_{\lambda\delta} /$, where $\beta_{r\lambda\delta}$ is the rescaled coefficient for level λ of dimension δ, $\beta_{\lambda\delta}$ is the coefficient for level λ of dimension δ and ϕ is the coefficient for dead.

Results

Valuation data

A total of 600 interviews were successfully conducted across both classification systems, providing a response rate of 41.5% of suitable respondents who answered their door to the interviewers. Respondents who valued the worst state higher than all other states, who valued all states worse than being dead or who valued all states identically but < 1 were excluded from the analysis. Seven respondents were excluded as they valued all states identically and < 1; two of these respondents also valued all states as worse than being dead. The analysis therefore included 306 respondents for DEMQOL-U and 287 for DEMQOL-Proxy-U. *Table 18* compares respondents across each classification system with the general population in South Yorkshire and England. The valuation study sample has a lower proportion of individuals aged 18–40 years, a higher proportion of respondents aged 41–65 years, women, retired individuals and homeowners and a lower mean EQ-5D score in comparison with the general population in South Yorkshire. The DEMQOL-U sample contains a higher proportion of women and a lower proportion of retired individuals than the DEMQOL-Proxy-U sample.

Time trade-off descriptive statistics

Tables 19 and *20* report descriptive statistics of observed TTO values for DEMQOL-U and DEMQOL-Proxy-U respectively. The number of observations per intermediate health state varied from 18 to 78, with 306 and 287, respectively, for the worst DEMQOL-U and DEMQOL-Proxy-U states. The range of mean values was larger for DEMQOL-U (from 0.954 to 0.184) than for DEMQOL-Proxy-U (from 0.961 to 0.331), although only two DEMQOL-U states had mean values outside the DEMQOL-Proxy-U range. Each classification system had one or more states with a mean value lower than that of the worst state defined by the classification system. In addition, DEMQOL-Proxy-U had two states with a mean value higher than that of the best state. These apparent contradictions are most likely observed because of the much smaller number of observations for some states in comparison with worst state and best state. *Figures 5* and *6* present the distribution of observed TTO values for DEMQOL-U and DEMQOL-Proxy-U respectively. There were a large proportion of TTO values at 1 for both measures (26.9% for DEMQOL-U and 28.8% for DEMQOL-Proxy-U) and the distribution of the data was negatively skewed for both measures.

Modelled time trade-off health-state utility values

Regression models estimating preference weights are reported in *Table 21* for DEMQOL-U and *Table 22* for DEMQOL-Proxy-U. Random-effects models are reported as the Hausman test confirmed for both classifications that fixed-effects models would produce similar estimates at reduced efficiency. Models including sociodemographic variables and interaction terms were explored, but the inclusion of these terms did not affect the coefficients of the main effects or improve the predictive performance of the model at the health-state level.

TABLE 18 Characteristics of all respondents[a]

Characteristic	DEMQOL-U (*n* = 306)	DEMQOL-Proxy-U (*n* = 287)	South Yorkshire[b]	England[b]
Age (years), mean (SD)	48.96 (16.60)	49.38 (17.22)	N/A	N/A
Age distribution (%)				
18–40 years	31.0	32.4	41.2	41.6
41–65 years	51.6	47.4	39.1	39.1
> 65 years	17.3	20.2	19.7	19.3
Female (%)	63.1	54.7	51.2	51.3
Married/partner (%)	64.7	66.6	N/A	N/A
Employed or self-employed (%)	50.3	51.2	56.1	60.9
Unemployed (%)	2.6	1.7	4.1	3.4
Long-term sick (%)	6.2	3.5	7.7	5.3
Full-time student (%)	4.9	4.5	7.5	7.3
Retired (%)	23.2	28.9	14.4	13.5
Own home outright or with a mortgage (%)	79.4	77.0	64.0	68.7
Renting property (%)	20.6	23.0	36.0	31.3
Secondary school is highest level of education (%)	31.7	32.4	N/A	N/A
EQ-5D score (SD)	0.83 (0.24)	0.81 (0.26)	N/A	0.86 (0.23)[c]
TTO completion rate (%)	100	96.6		

N/A, not available.

a Seven respondents excluded: all valued every state identically and at < 1, and two of these respondents further valued all states as worse than being dead. Remaining in the analysis are 306 respondents for the DEMQOL-U derived from the DEMQOL and 287 for the DEMQOL-PROXY-U derived from the DEMQOL-Proxy.

b Statistics for South Yorkshire Health Authority and for England in the 2001 Census. Questions used in this study and in the census are not identical. The census includes those aged ≥16 years whereas this study surveys only those aged ≥18 years. Age distribution is here reported as the percentage of all adults aged ≥18 years.

c Interviews conducted in the MVH study in 1993.

The coefficients of the main effects variables indicate the utility decrement for every severity level of every dimension of the classification system using level 1 as the baseline. For coefficients to be consistent they must be positive and increasing in size for each subsequent severity level within each dimension. For the DEMQOL-U models all coefficients were positive as expected. The standard OLS models estimated using pooled observations and mean level data [models (1) and (3)] had inconsistent coefficients for negative emotion at levels 2 and 3, but the random-effects GLS model [model (2)] had consistent coefficients. Model (2) was preferred to standard OLS estimated using pooled observations [model (1)] as it takes into account the structure of the data with repeated observations per respondent. The mean model [model (3)] does not suffer from the theoretical bias occurring when estimating models on data containing a large proportion of values at 1, yet performed similarly to model (2), with a large number of significant coefficients and good predictive ability.

Model (2), the random-effects GLS model, was selected as the preferred model because of its strong relative performance. Model (2) had a high number of significant variables, no inconsistent coefficient estimates and good predictive ability with low RMSE and the lowest percentage of AEs at the state level

TABLE 19 Descriptive statistics of observed TTO values for DEMQOL-U

Health state	Mean	SD	Median	Count
11111	0.954	0.153	1.000	78
14231	0.934	0.093	0.975	19
11231	0.930	0.139	1.000	20
11312	0.918	0.098	0.950	19
13221	0.913	0.105	0.950	18
14311	0.909	0.134	0.950	19
12241	0.903	0.153	1.000	19
12231	0.894	0.160	1.000	78
14431	0.888	0.140	0.950	19
24231	0.860	0.182	0.963	18
13432	0.858	0.187	0.938	18
23411	0.845	0.225	1.000	20
12323	0.845	0.159	0.875	19
21241	0.845	0.253	0.998	19
21143	0.836	0.170	0.875	18
21213	0.835	0.230	0.913	18
41111	0.822	0.228	0.925	19
21233	0.819	0.191	0.875	19
22123	0.818	0.204	0.925	19
12432	0.818	0.278	0.925	18
31331	0.817	0.235	1.000	19
34221	0.812	0.181	0.888	20
12123	0.804	0.118	0.813	20
14233	0.801	0.226	0.850	18
13142	0.800	0.199	0.863	18
32131	0.799	0.251	0.900	19
23413	0.796	0.219	0.813	20
12342	0.790	0.228	0.813	20
32112	0.789	0.217	0.875	19
33431	0.789	0.265	0.925	19
21322	0.786	0.261	0.925	19
23214	0.784	0.271	0.925	37
34123	0.782	0.289	0.925	19
21132	0.773	0.306	0.850	20
14223	0.770	0.221	0.850	19
12243	0.768	0.255	0.800	18

continued

TABLE 19 Descriptive statistics of observed TTO values for DEMQOL-U (*continued*)

Health state	Mean	SD	Median	Count
32441	0.767	0.212	0.813	18
23441	0.754	0.240	0.775	19
21342	0.753	0.250	0.800	19
34142	0.751	0.345	0.925	18
14133	0.742	0.433	0.925	19
13242	0.741	0.264	0.788	20
22414	0.732	0.222	0.688	18
32142	0.730	0.280	0.775	19
44322	0.728	0.251	0.700	19
34132	0.728	0.276	0.813	20
34313	0.721	0.315	0.813	18
23314	0.685	0.353	0.725	39
14323	0.685	0.296	0.650	20
24313	0.680	0.301	0.750	19
31433	0.675	0.340	0.775	18
32143	0.673	0.302	0.663	78
13413	0.670	0.474	0.800	19
32314	0.661	0.352	0.800	19
33421	0.661	0.313	0.625	19
31413	0.655	0.216	0.650	20
24144	0.654	0.335	0.675	19
41233	0.640	0.329	0.575	18
43122	0.629	0.299	0.675	19
43231	0.621	0.257	0.500	19
34423	0.617	0.225	0.600	19
41212	0.611	0.364	0.625	78
23344	0.609	0.362	0.675	19
22334	0.605	0.340	0.625	20
42243	0.592	0.357	0.625	19
43321	0.591	0.387	0.663	56
41241	0.589	0.480	0.725	18
31444	0.587	0.399	0.525	19
42321	0.576	0.300	0.513	20
41231	0.573	0.397	0.675	39
24434	0.568	0.432	0.675	19
23424	0.566	0.340	0.600	78
43123	0.548	0.400	0.588	20
44221	0.544	0.367	0.525	78

TABLE 19 Descriptive statistics of observed TTO values for DEMQOL-U (*continued*)

Health state	Mean	SD	Median	Count
41423	0.534	0.482	0.500	19
44341	0.528	0.356	0.550	20
34234	0.525	0.486	0.600	18
43432	0.462	0.525	0.600	19
43434	0.449	0.429	0.463	18
44242	0.418	0.549	0.500	19
41314	0.417	0.477	0.475	19
43442	0.403	0.412	0.475	78
43344	0.397	0.422	0.500	19
43343	0.364	0.561	0.500	20
41224	0.339	0.542	0.425	20
44444	0.234	0.501	0.200	306
43244	0.184	0.475	0.200	20

TABLE 20 Descriptive statistics of observed TTO values for DEMQOL-Proxy-U

Health state	Mean	SD	Median	Count
1132	0.961	0.062	1.000	19
2221	0.944	0.108	1.000	20
1111	0.918	0.185	1.000	77
2141	0.918	0.172	1.000	20
2421	0.896	0.151	1.000	39
1143	0.893	0.149	0.963	20
1141	0.892	0.153	1.000	19
2312	0.884	0.154	1.000	19
1131	0.869	0.218	1.000	18
1222	0.869	0.174	0.925	77
3121	0.855	0.213	1.000	19
2212	0.846	0.248	0.950	20
3311	0.846	0.181	0.925	19
1214	0.839	0.227	0.925	38
1333	0.833	0.191	0.863	20
1231	0.830	0.267	0.950	19
2131	0.821	0.241	0.888	18
1233	0.821	0.194	0.925	19
1341	0.806	0.236	0.900	77
3112	0.804	0.255	0.925	77

continued

TABLE 20 Descriptive statistics of observed TTO values for DEMQOL-Proxy-U (*continued*)

Health state	Mean	SD	Median	Count
1223	0.799	0.256	0.888	20
3221	0.796	0.224	0.900	19
1422	0.791	0.308	0.900	38
1322	0.789	0.190	0.875	19
2431	0.763	0.396	0.900	19
3331	0.763	0.221	0.775	19
1432	0.751	0.269	0.825	20
2422	0.750	0.349	0.825	19
1313	0.721	0.277	0.800	38
3432	0.719	0.263	0.775	18
2342	0.716	0.304	0.850	18
3341	0.714	0.309	0.788	20
3431	0.705	0.294	0.800	19
1334	0.700	0.256	0.700	19
3342	0.699	0.287	0.750	38
4111	0.697	0.326	0.775	19
2234	0.692	0.279	0.675	19
1324	0.688	0.398	0.775	19
2243	0.684	0.350	0.788	38
3224	0.684	0.305	0.775	58
1423	0.682	0.341	0.800	19
2224	0.676	0.259	0.725	19
3312	0.674	0.256	0.700	19
3133	0.664	0.239	0.575	19
2424	0.662	0.342	0.725	97
4212	0.655	0.412	0.800	40
3234	0.654	0.287	0.675	95
4322	0.645	0.331	0.638	20
3344	0.634	0.370	0.675	20
4123	0.629	0.271	0.625	19
3144	0.625	0.339	0.663	20
4141	0.614	0.310	0.725	19
4342	0.603	0.468	0.788	20
3424	0.596	0.357	0.538	20
2344	0.593	0.410	0.675	19
3443	0.589	0.288	0.588	38

TABLE 20 Descriptive statistics of observed TTO values for DEMQOL-Proxy-U (*continued*)

Health state	Mean	SD	Median	Count
2214	0.588	0.336	0.675	18
4411	0.580	0.395	0.625	77
4242	0.578	0.289	0.500	19
2423	0.552	0.375	0.488	18
2334	0.533	0.352	0.638	18
4133	0.532	0.330	0.400	18
4214	0.505	0.309	0.425	19
3314	0.484	0.356	0.350	19
4311	0.478	0.421	0.500	36
4333	0.476	0.378	0.488	58
4434	0.434	0.477	0.500	19
4443	0.383	0.458	0.425	18
4444	0.357	0.439	0.325	287
4343	0.331	0.492	0.450	18

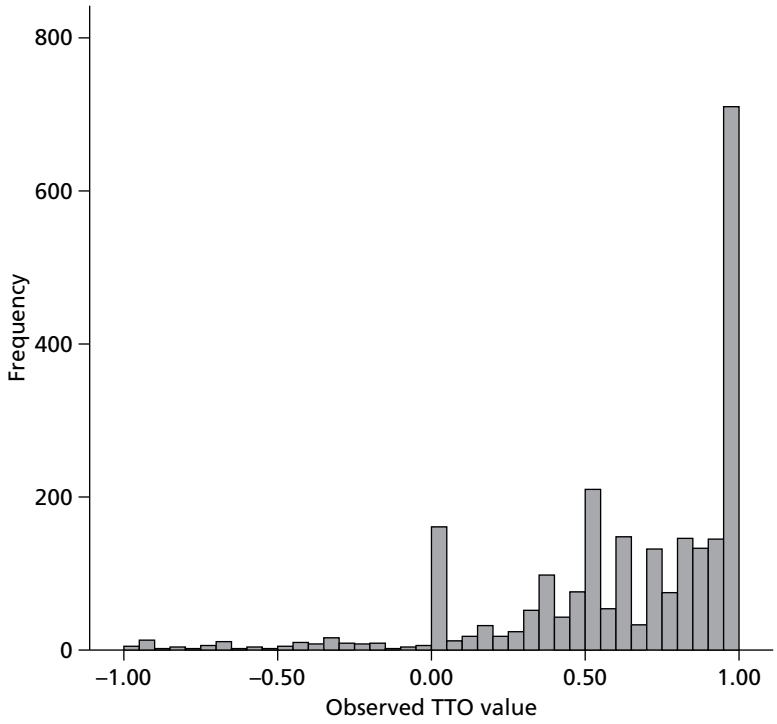


FIGURE 5 Histogram of observed TTO values for DEMQOL-U.

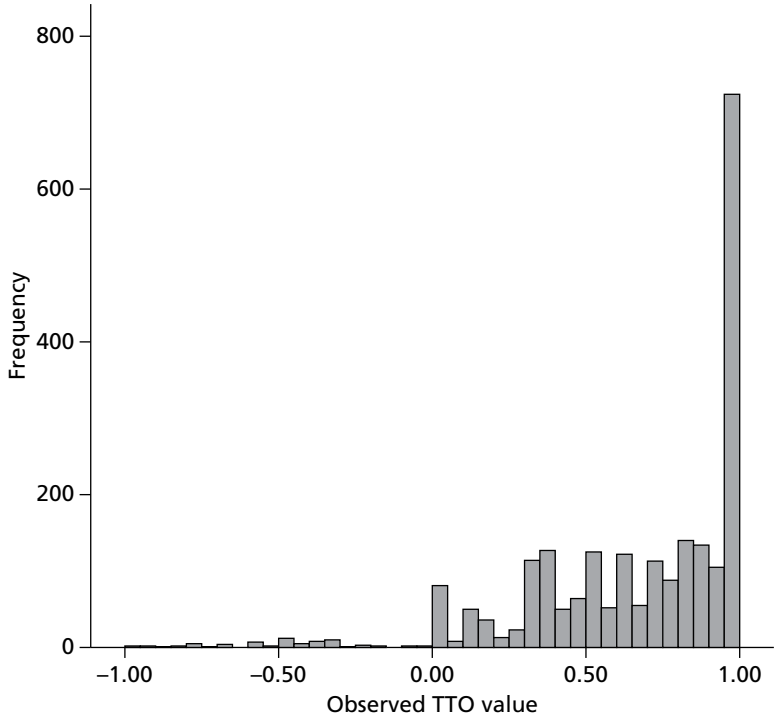


FIGURE 6 Histogram of observed TTO values for DEMQOL-Proxy-U.

>0.05 and >0.10. *Figure 7* plots observed and predicted TTO values at the health-state level for model (2). The figure shows that there was no systematic bias in the predictions by severity, but some health states had large prediction errors. The range of predicted values for this model was from 0.986 (best state 11111) to 0.243 (worst state 44444), which was similar to the observed range for these states (from 0.954 to 0.234).

For DEMQOL-Proxy-U, all models have at least one inconsistent coefficient and at least one negative coefficient for the appearance dimension. These inconsistencies mean that health worsens but predicted utility increases, for example deterioration in appearance from level 1 to level 2 leads to a higher predicted utility value. Other valuation studies in which this has occurred have resolved this issue by merging adjacent inconsistent levels and re-estimating the model to produce a consistent model [see model (10) for example]. This approach was used to estimate models (8) and (10), consistent versions of models (7) and (9), respectively, in which adjacent levels of the appearance dimension have been merged.

For DEMQOL-Proxy-U, the standard OLS model estimated using pooled observations [model (6)], the consistent random-effects GLS model [model (8)] and the consistent mean model [model (10)] performed similarly in terms of predictive ability. As explained above, the consistent version of the random-effects GLS model [model (8)] and the mean-level model [model (10)] are preferred on a theoretical basis to model (6). Both models suffer from autocorrelation between errors to some degree using the LB test, but this is significant only for model (8) (at the 10% level). Despite this finding, *Figure 8* demonstrates little difference in the pattern of the predictions by severity for these models. Therefore, model (8) was selected as the preferred model because of its higher number of significant coefficients. The range of predicted values for this model was from 0.937 (best state 1111) to 0.363 (worst state 4444), which was similar to the observed range for these states (from 0.918 to 0.357).

Modelled rank health-state utility values

Regression models estimating normalised preference weights using the rank data are also reported in *Table 21* for DEMQOL-U and *Table 22* for DEMQOL-Proxy-U. The sample is identical to the sample used in the models estimated using TTO data for comparability. Additional exclusion criteria were explored:

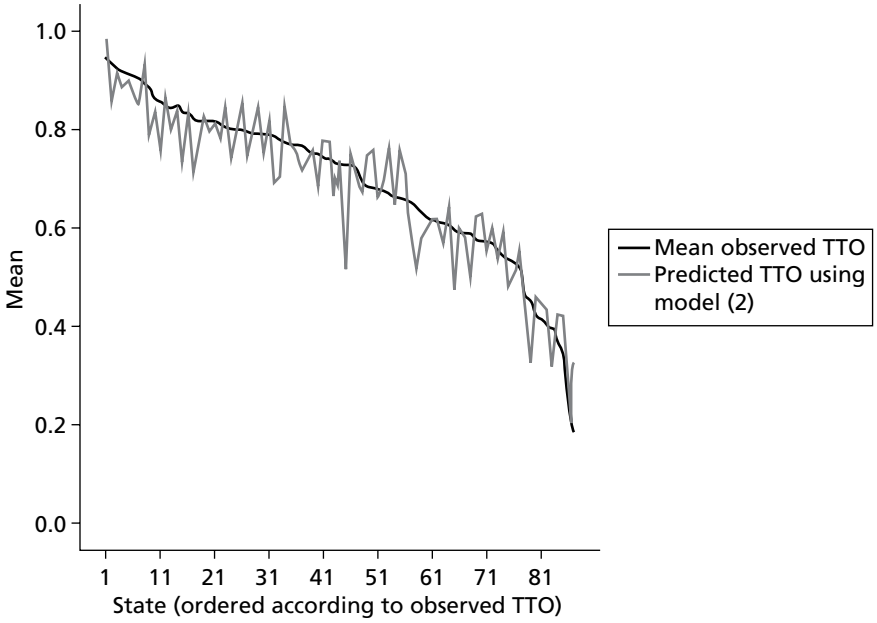


FIGURE 7 Observed and predicted TTO for DEMQOL-U using model (2).

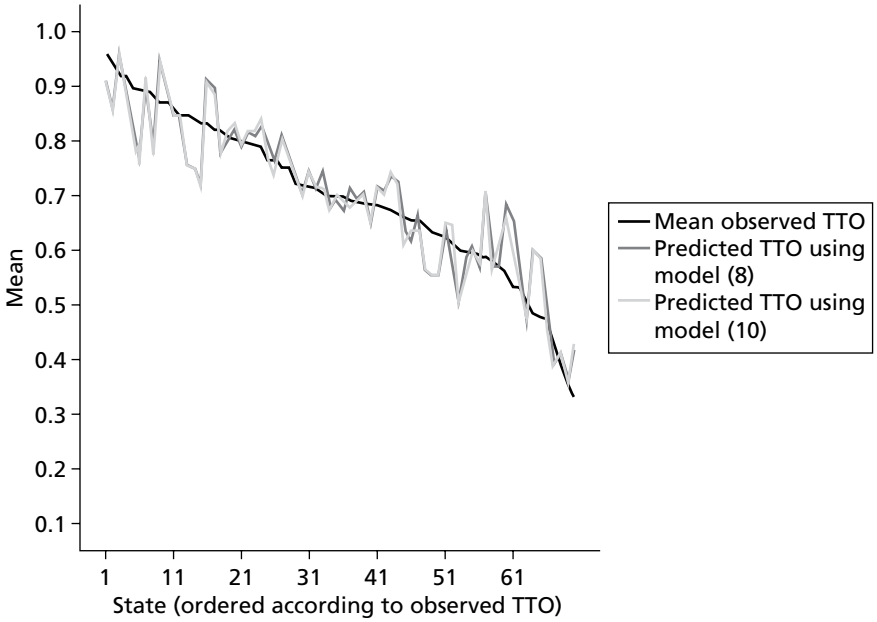


FIGURE 8 Observed and predicted TTO for DEMQOL-PROXY-U using models (8) and (10).

respondents ranking the worst state higher than all other states, respondents ranking all states worse than dead or respondents ranking all states identically. No respondents were excluded using these criteria. These coefficients are interpreted in the same way as the TTO coefficients as explained above. The rank coefficients were normalised onto the 1–0 full health–dead scale using the first approach outlined above, using the modelled TTO values of the worst state generated using the preferred model [model (2) for DEMQOL-U and model (8) for DEMQOL-Proxy-U]. It was not possible to normalise the coefficients using the second approach outlined above, using the coefficient of the dead dummy variable. For DEMQOL-U only 4.9% of individuals ranked one or more states as worse than dead, and for DEMQOL-Proxy-U only 1.0% of individuals ranked one or more states as worse than dead. This meant that the coefficient for the dead dummy variable could not be estimated because of insufficient variance in the ranking of the dead

TABLE 21 Regression models estimating preference weights for DEMQOL-U

	TTO			Rank		
	(1) OLS	(2) RE GLS	(3) Mean model	(4) ROL		(5) ROL consistent
Positive emotion2	0.016	0.025	0.024	0.102[a]	Positive emotion2	0.101[a]
Positive emotion3	0.085[a]	0.087[a]	0.097[a]	0.129[a]	Positive emotion3	0.129[a]
Positive emotion4	0.285[a]	0.284[a]	0.275[a]	0.308[a]	Positive emotion4	0.309[a]
Memory2	0.002	0.000	0.001	0.075[a]	Memory23	0.068[a]
Memory3	0.024	0.027[b]	0.038[b]	0.065[a]		
Memory4	0.046[c]	0.055[a]	0.041[c]	0.116[a]	Memory4	0.117[a]
Relationships2	0.023	0.024	0.033	0.084[a]	Relationships23	0.074[a]
Relationships3	0.033	0.042[c]	0.046[c]	0.063[a]		
Relationships4	0.083[a]	0.083[a]	0.073[a]	0.110[a]	Relationships4	0.107[a]
Negative emotion2	0.047[c]	0.035[c]	0.042[b]	0.031[a]	Negative emotion2	0.031[a]
Negative emotion3	0.039[b]	0.048[a]	0.040[b]	0.046[a]	Negative emotion3	0.052[a]
Negative emotion4	0.102[a]	0.102[a]	0.092[a]	0.072[a]	Negative emotion4	0.075[a]
Loneliness2	0.054[c]	0.055[a]	0.063[a]	0.089[a]	Loneliness2	0.086[a]
Loneliness3	0.095[a]	0.104[a]	0.105[a]	0.112[a]	Loneliness3	0.113[a]
Loneliness4	0.217[a]	0.219[a]	0.216[a]	0.152[a]	Loneliness4	0.150[a]
Constant	0.023	0.014	0.007			
Observations	2448	2448	87	2742		2742
Number of subjects		306		306		306
R^2	0.263		0.875			
No. of inconsistencies	1	0	1	2		0
No. of significant variables	10	12	12	15		13
RMSE	0.060	0.060	0.058			
MAE	0.044	0.045	0.044			
% states with AE > 0.05	37.9	34.5	36.8			
% states with AE > 0.10	9.2	8.0	8.0			
LB test	5.751	5.370	4.184			

RE GLS, random-effects generalised least squares; ROL, rank order list.

a Significant at 1%.

b Significant at 10%.

c Significant at 5%.

TABLE 22 Regression models estimating preference weights for DEMQOL-Proxy-U

	TTO					Rank
	(6) OLS	(7) RE GLS	(8) RE GLS consistent model	(9) Mean model	(10) Mean consistent model	(11) ROL
Positive emotion2	0.043[a]	0.037[a]	0.037[a]	0.058[b]	0.059[b]	0.069[b]
Positive emotion3	0.093[b]	0.091[b]	0.092[b]	0.100[b]	0.101[b]	0.110[b]
Positive emotion4	0.261[b]	0.263[b]	0.265[b]	0.262[b]	0.266[b]	0.289[b]
Memory2	0.027	0.029[c]	0.028[c]	0.031	0.030	0.083[b]
Memory3	0.088[b]	0.085[b]	0.085[b]	0.092[b]	0.093[b]	0.143[b]
Memory4	0.110[b]	0.104[b]	0.102[b]	0.117[b]	0.113[b]	0.145[b]
Appearance2	–0.020	–0.006		–0.019		0.024[b]
Appearance3	0.007	–0.002		0.008	0.019	0.054[b]
Appearance4	0.036[c]	0.045[b]	0.047[b]	0.027	0.037[c]	0.074[b]
Negative emotion2	0.024	0.027[c]	0.026[c]	0.016	0.016	0.054[b]
Negative emotion3	0.118[b]	0.119[b]	0.119[b]	0.123[b]	0.120[b]	0.069[b]
Negative emotion4	0.160[b]	0.161[b]	0.160[b]	0.163[b]	0.162[b]	0.129[b]
Constant	0.067[b]	0.065[b]	0.063[b]	0.062[a]	0.051[a]	
Observations	2295	2295	2295	70	70	2573
Number of subjects		287	287			287
R^2	0.212			0.860	0.858	
No. of inconsistencies	1	2	0	1	0	0
No. of significant variables	9	11	11	8	9	12
RMSE	0.056	0.057	0.057	0.055	0.055	
MAE	0.042	0.042	0.042	0.042	0.042	
% states with AE > 0.05	35.7	35.7	35.7	32.9	32.9	
% states with AE > 0.10	11.4	8.6	8.6	10.0	12.9	
LB test	12.365	14.193	14.301	9.142	9.757	

RE GLS, random-effects generalised least squares; ROL, rank order list.

a Significant at 5%.

b Significant at 1%.

c Significant at 10%.

state across individuals. We expect that this occurred because the severity of the worst state is perceived as being unanimously better than being dead, and this represents the fact that the DEMQOL-U and DEMQOL-Proxy-U have condition-specific classification systems that do not fully capture all dimensions of HRQL. However, the same problem was not encountered for an asthma-specific preference-based measure and an overactive bladder preference-based measure.[77]

For the DEMQOL-U all normalised rank coefficients were positive as expected and significant at the 1% level [models (4) and (5)]. However, in contrast to the models estimated using the TTO data, there were inconsistent coefficients for the memory dimension at levels 2 and 3 and the relationships dimension at levels 2 and 3 in model (4). As explained above, adjacent inconsistent levels were merged and the model was re-estimated to produce a consistent model [model (5)]. The normalised coefficients were larger than the coefficients for the preferred model estimated using the TTO data [model (2)] with the exception of one coefficient.

For the DEMQOL-Proxy-U all normalised rank coefficients were positive and significant at the 1% level [model (11)]. Contrary to the models estimated using the TTO data there were no inconsistent coefficients. As explained above, adjacent inconsistent levels were merged and the model was re-estimated to produce a consistent model [model (5)]. The majority of normalised coefficients were larger than the coefficients for the preferred model estimated using the TTO data [model (8)].

Errors have not been generated for the rank models as they represent the difference in observed and predicted utilities and for rank data there are no directly observed latent utility estimates that the predictions can be compared with.

Discussion

In this chapter we have reported on the estimation of preference weights for the two dementia-specific classification systems derived in *Chapter 4*: DEMQOL-U derived from DEMQOL and DEMQOL-Proxy-U derived from DEMQOL-Proxy. These preference weights enable a health-state utility value to be estimated for every health state defined by each classification system. These preference weights can be used to generate a utility score for a patient with dementia each time the patient completes the DEMQOL questionnaire or their carer completes the DEMQOL-Proxy questionnaire. These utility scores can then be used as the 'Q' quality adjustment weight of the QALY to inform economic evaluation. The use of both the DEMQOL-U and DEMQOL-Proxy-U measures enables utilities to be generated appropriately across the full severity range of dementia. These dementia-specific measures and corresponding utility scores have been designed specifically for use in a cognitively impaired population, offering an advantage over generic measures such as the EQ-5D and SF-36.

The valuation study included rank and TTO tasks. Both rank and TTO data can be used to estimate preference weights that can be used to generate utility values for all health states defined by the classification system. Here, the regression models estimated using the rank data had more significant coefficients than the models estimated using the TTO data, yet had inconsistent coefficients for two DEMQOL-U dimensions. We recommend that the preferred models estimated using TTO data should be used to generate the preference weights. This is in accordance with recommendations from NICE stating that the valuation protocol used to estimate the preference weights should be the protocol used to produce the UK values of the EQ-5D. This promotes comparability of utility values generated using different classification systems from different preference-based measures.

There is currently no accepted standard on the minimum number of observations required for each health state to ensure reliability and robustness of the modelling output. It is possible that more observations per health state may have improved the reliability and robustness of the modelling output and further research into this issue is encouraged. However, the regression models estimated using the TTO data had good

predictive ability (e.g. MAE 0.042–0.045) and performed comparably with regression models estimated for other preference-based measures including the generic EQ-5D (MAE 0.039),[45] generic SF-6D (MAE 0.073–0.079),[38] asthma-specific AQL-5D (0.046–0.057),[40] overactive bladder-specific OAB-5D (0.044–0.076)[42] and cancer-specific EORTC-8D (0.046–0.054).[72]

Each measure taps three common dimensions (although not using all of the same items) in its classification system: positive emotion, negative emotion and memory. DEMQOL-U has two additional dimensions of relationships and loneliness and DEMQOL-Proxy-U has an additional appearance dimension. The appearance dimension had small and inconsistent coefficients across all DEMQOL-Proxy-U regression models, suggesting that deterioration in this dimension did not have a large or stable impact on utility. The coefficients for the dimensions common to both measures were consistently larger for DEMQOL-Proxy-U than for DEMQOL-U (with the exception of negative emotion level 2). This finding is consistent with research which found that adding an extra dimension to a classification system reduced the preference weights for existing dimensions.[78] This suggests that the coefficients of the DEMQOL-U measure with five dimensions may be smaller in the common dimensions than the coefficients of the DEMQOL-Proxy-U measure because of the larger number of dimensions. However, this finding may also be due to differences in the wording of the items selected for each classification system.

Observed and modelled utility scores for DEMQOL-U have a larger range and lower bound than those for DEMQOL-Proxy-U, meaning that DEMQOL-U has the larger severity range. This is likely due to differences in the classification systems: DEMQOL-Proxy-U had fewer dimensions and the appearance dimension had small preference weights. Yet this produces an apparent contradiction as DEMQOL and DEMQOL-U were designed for use in people with mild to moderate dementia whereas DEMQOL-Proxy and DEMQOL-Proxy-U were designed to be appropriate for all levels of dementia. However, the possible range of utility scores for each measure is likely to be different from utility scores of patients because this depends on the distribution of responses to the classification system. This will be evaluated further in *Chapter 7*.

Conclusion

The data presented here suggest that DEMQOL-U and DEMQOL-Proxy-U, dementia-specific preference-based single-index measures derived from DEMQOL and DEMQOL-Proxy, respectively, are suitable to be used to generate utility scores for use in health technology assessment across the full severity range of dementia when both measures are used. Algorithms are presented on the DEMQOL website that generate DEMQOL-U and DEMQOL-Proxy-U utility scores when applied to DEMQOL and DEMQOL-Proxy data (www.kcl.ac.uk/iop/depts/hspr/research/ciemh/mha/demqol/index.aspx). These algorithms use the preference weights generated using modelled TTO data.

Chapter 6 Patient and carer valuation survey

Introduction

In the preceding chapters we have detailed the development of DEMQOL-U and DEMQOL-Proxy-U, dementia-specific preference-based single-index measures from DEMQOL and DEMQOL-Proxy. In *Chapter 5* we reported the estimation of a preference-based single index for each classification system using values obtained from the general population. However, such values can also be obtained from other sources including patients, carers and health-care professionals. Values are typically obtained from the general population in accordance with recommendations from agencies such as NICE and the Washington Panel. General population values are often advocated because public preferences are considered appropriate when health care is publicly funded (partly or fully, depending on a country's health system). In addition, the general population is considered to have no vested interest as they do not have prior knowledge about the future health states they may experience. However, it can also be argued that the general population do not fully understand hypothetical health states because of their lack of experience of similar health states, meaning that patients are more able to provide accurate valuations. This is important because patients often provide different values from those provided by the general population. This means that the source of the values may affect the values obtained and thus have an impact on any economic evaluation based on such data.[79]

A recent review and meta-analysis found that patient values of health states were significantly lower than general population values using the TTO elicitation technique.[80] However, other papers have found that patients provide higher values for health states using a variety of valuation techniques.[79,81] Reasons for these differences can be summarised as (1) valuing different states because of differences in understanding or interpretation of the description; (2) different scales of measurement because of a response shift in different populations; and (3) adaptation to the state by patients. Patients may better understand what it is like to be in the health state than a member of the general population who has to imagine the health state. Furthermore, patients may value the health state differently because they are bringing to the valuation additional information that is not included in the health-state description. However, patients may also not recall experiences of full health or mild health states, and may have lowered their expectations as a result. Knowledge and consideration of adaptation is one important difference between patients and members of the general population. Patients may also have different sociodemographic characteristics from the general population and this may also affect how they value health states.

Much of the literature examining differences between patient and general population preferences has focused on physical health states, and the relationship may differ for conditions affecting mental health and cognition. In two studies people experiencing a range of different health states gave mental health greater weight than physical health, which was the opposite for members of the general public trying to imagine the health states.[82] Public attitudes and understandings of dementia in particular may mean that patient and general population values differ for dementia health states. There may also be differences in values provided by carers of patients with dementia, as carers have experience of how the condition impacts on the patient without the health problems experienced by the patient that may affect their understanding of the valuation techniques. The contrast in weightings of physical and mental health by the general population and patients is important for policy. If general population values are used to inform policy and these are higher than patient values for mental health and cognition states, then mental health and cognition may be given lower priority than people with these conditions feel is warranted. The higher relative weighting of physical health to mental health by the general population potentially prioritises physical health at the expense of mental health.

The literature is also limited as typically studies ask respondents to value only a small number of states, meaning that the data cannot be examined to determine whether or not there are systematic differences across all health states or whether or not the difference varies by health-state severity. Even those studies that value a large sample of health states have not examined whether or not differences vary by severity. For example, one study estimated different value sets for patients and the general population for EQ-5D but when combining data from both populations to estimate a combined value set included only a dummy variable to capture patient effects rather than including interaction effects for the severity levels of each dimension.[83] However, one study did find that differences in visual analogue scale (VAS) values vary by the severity of the health state between members of the general population with no health problems, those with mild health problems and those with moderate health problems.[84] It is also important to control for sociodemographic characteristics of the samples[84] as it has been found that elicited TTO values vary by sociodemographic characteristics,[85] and it is possible that this explains some of the variation in values across populations.

To address these general limitations and to investigate the difference in values that might be obtained between the general population and people with dementia and their carers we undertook a comparative study comparing health-state utility values from samples of the general population, people with dementia and carers of people with dementia for a range of dementia health states of differing severity. The analysis here explores whether population, health-state severity and respondent sociodemographic characteristics impact on elicited utility values.

Methods

Health-state description

Health-state descriptions were generated using the DEMQOL-U and DEMQOL-Proxy-U classification systems as described in *Chapter 3*. The DEMQOL-U classification system is derived from the DEMQOL questionnaire, which is completed by people with dementia, whereas the DEMQOL-Proxy-U classification system is derived from the DEMQOL-Proxy questionnaire, which is completed by carers using proxy report. These classification systems are presented in *Table 14*.

Selection of health states

In *Chapter 5* we described the selection of health states and blocks of combinations of states that were included in the general population valuation study used to estimate a preference-based single index. We selected one of these blocks each for DEMQOL-U and DEMQOL-Proxy-U. For each classification system we chose the block that contained the best health state defined by the classification system (states 11111 and 1111 respectively) and so along with the worst state this ensured that the selected health states covered the full severity range of the classification system.

Samples

To maintain the involvement of each population in the valuation of each classification system, health states for each classification were valued by the population who completed the questionnaire and who were involved in the initial development of the questionnaire. DEMQOL-U was valued by patients and DEMQOL-Proxy-U was valued by carers. In addition, both classification systems were valued by the general population in phase 2 of the project as described in *Chapter 5*. Sample size was chosen to ensure sufficient power for comparison of mean health-state utility values across the different populations for each classification system using simple *t*-tests. This required a total of 71 completed interviews per population per classification system, assuming a power of 0.8, significance level of 0.05, SD of 0.3 and an expected difference of 0.1.

General population

The general population sample and recruitment process is described in detail in *Chapter 5*. The analysis conducted in this chapter uses a subset sample of the 600 general population respondents, consisting of

all respondents valuing our selected card block. To ensure that there were no systematic differences across the geodemographic profiles of the subsamples used here interviewers worked systematically through all card blocks.

People with dementia and their carers

To recruit the sample of people with mild dementia and their carers, two research workers visited clinical teams who were part of the Mental Health for Older Adults and Dementia Clinical Service at the South London and Maudsley NHS Foundation Trust. This included the Croydon Memory Service, the Southwark and Lambeth Memory Service and the community mental health teams in Croydon and Lewisham. The research workers explained the study to the clinical teams and invited team members to refer suitable patients. Letters and information sheets were then sent to all referred patients, after which the research workers contacted them by telephone to arrange appointments. Participants who agreed to take part in the study were visited in their own homes by both research workers. If both the person with dementia and his or her carer were participating, interviews were conducted separately, each in a different room where possible. The research workers received the same training as the interviewers who carried out the general population survey.

Valuation task

For the valuation survey the TTO technique was chosen in accordance with the UK valuation study design of the EQ-5D,[54] which also meets the reference case recommended by NICE (see *Chapter 5* for a detailed overview of the TTO technique and *Appendix 1* for a sample script for one health state). For each classification system all respondents valued the same health states.

At the interview respondents first self-completed the EQ-5D and then whichever classification system would be used subsequently in the valuation task (DEMQOL-U or DEMQOL-Proxy-U). This familiarises respondents with the classification system by using the system to describe their own health. General population respondents then undertook a ranking task of eight health states plus full health and dead. People with dementia and carers did not undertake a rank task because of concerns that this task would be too complex. Ranking requires the respondent to consider all health states simultaneously and, given that dementia is a condition that affects memory, this was considered inappropriate for patients with dementia. Although carers may have been able to successfully undertake the ranking task, the same protocol was used for both people with dementia and carers, for consistency and to ensure that people with dementia felt that they were treated in the same way as their carers.

To familiarise respondents with the TTO task, respondents undertook a practice TTO task using a hypothetical 'practice' state and subsequently valued eight states using TTO. The MVH study version of TTO was used, including a visual prop designed by the MVH group (University of York) (see *Figure 4*). Finally, respondents rated how difficult they found the rank and TTO tasks and answered questions about their sociodemographic characteristics and health service use.

Before conducting the full survey a small number of interviews were conducted and the interviewers discussed their findings with one of the research team. As no problems were identified with the survey, the main survey was started without amendment and these interviews were included in the main sample. The person with dementia and carer survey was approved by the London Research Ethics Committee and the general population survey was approved by the ScHARR Research Ethics Committee at the University of Sheffield.

Analysis of time trade-off data

Significant differences in the observed variation of respondent characteristics across the different populations were analysed for each measure using a factorial ANOVA estimated using a generalised linear model. Descriptive statistics of health-state utility values across the different populations (general population, people with dementia, carers) were presented for each measure. Mean health-state utility

values were compared across the different populations for each measure using simple *t*-tests and Wilcoxon rank-sum tests.

Regression analysis was used to examine the impact of population (general population or people with dementia, general population or carer) on elicited health-state utility values. The analysis further determined the impact of population and health-state severity while controlling for the sociodemographic characteristics of respondents. The standard model specification was:

$$y_{ij} = \alpha + \beta s_j + \gamma q_i + \theta r_{ij} + \delta z_i + \varepsilon_{ij} \quad (4)$$

where $i = 1, 2, \ldots, n$ represents individual respondents and $j = 1, 2, \ldots, m$ represents the eight health states. The dependent variable, y, represents the TTO utility value, s represents the vector of dummy variables for the health states, q represents the dummy variable capturing the population (equals 1 for persons with dementia and carers, 0 for general population), r represents the vector of interaction terms to jointly capture population and severity effects (e.g. in the DEMQOL-U regression analysis a dummy variable equals 1 for state 12231 valued by a person with dementia), z represents the vector of sociodemographic characteristics of respondents and ε_{ij} represents the error term. OLS regressions were estimated but these do not take into account the structure of the data, as there are multiple observations per individual, and these models are not reported here. Random-effects and fixed-effects GLS models were estimated as these take into account the structure of the data when all respondents have multiple observations.[86] Refer to *Chapter 5* for an overview of random-effects and fixed-effects models. The choice of whether to use random- or fixed-effects models was determined empirically using the Hausman test.

Four model specifications were estimated in which each model contained additional explanatory variables: first, regressions were estimated containing only health-state dummies as independent variables; second, the dummy variable capturing population was added to the model specification; third, interaction terms capturing population and severity effects were added; and finally, sociodemographic characteristics of respondents were added. This procedure was undertaken to determine the additive impact of population across all states regardless of severity, and to then determine whether or not model performance was improved by expanding the model specification to include interactions that allowed for impact of population to vary by state severity. This procedure was also used to determine whether or not population effects are important when the differences in sociodemographic characteristics of respondents and across population samples are controlled for, as some differences in values across populations may be due to the difference in sociodemographic composition of the samples rather than the populations per se. The final model specification was also estimated using a sample excluding respondents whose understanding of the TTO task was doubted by the interviewers.

The selection of sociodemographic characteristics for inclusion in the models was informed using the ANOVA analysis (as described above), Spearman rank correlation coefficients, significance of coefficients and performance of the regression models (described below). Spearman rank correlation coefficients were used to indicate the sociodemographic variables that were appropriate for inclusion in the models (which are variables most highly correlated with the TTO utility value) and to indicate sociodemographic variables that should not appear in the models alongside each other (which are variables with high correlations with other variables, for example age and different categories of employment status).

Performance of the regression models was assessed using within *R*-squared, between *R*-squared, overall *R*-squared, RMSE of predictions and the Wald chi-squared test. Stata version 11 was used for all regression analysis (StatCorp LP, College Station, TX, USA) and SPSS version 15 was used for the descriptive statistical analysis.

Results

Samples

Recruitment of the general population sample is summarised in *Chapter 5*. The samples of people with dementia and carers were recruited between 7 November 2010 and 23 May 2011. During this time the teams referred a total of 196 people diagnosed with dementia. Of these, 93 patients and 73 carers agreed to be interviewed and were assessed as suitable by the research workers. On some occasions patients were seen without carers, and a few carers participated alone. Recruitment continued until 71 patients and 71 carers had completed the interview. During the process 21 partial interviews were undertaken, 19 with patients and two with carers, and three patients were not able to participate at the time of the arranged interview. These partial interviews were terminated before the end of the TTO tasks because of respondent fatigue, misunderstanding or distress, and these interviews are therefore excluded from the analysis. Of those that completed the full TTO study, 49 were patient/carer dyads.

Respondents who valued the worst state higher than all other states, who valued all states worse than being dead or who valued all states identically but < 1 were excluded from the analysis. No patients with dementia or carers were excluded but one general population respondent was excluded for valuing all states identically but < 1.

Table 23 shows a comparison of respondents by population and classification system. The comparison of respondents valuing the DEMQOL-U classification system shows that the patient population sample is significantly older (at the 10% level) using the chi-squared *p*-value, with a mean (SD) age of 78.4 (7.7) years, than the general population sample, with a mean (SD) age of 49.6 (17.0) years. The samples have significant differences in employment status and education, with the patient sample having a much higher proportion of retired individuals and individuals for whom secondary school was their highest level of education. EQ-5D scores were similar between the general population and patient DEMQOL-U samples.

The comparison of respondents valuing the DEMQOL-Proxy-U classification system shows that the carer population sample was significantly older, with a mean (SD) age of 69.8 (12.7) years in comparison to a mean (SD) age of 50.7 (16.2) years for the general population sample. The samples have significant differences in employment status (using chi-squared *p*-values), marital status and homeownership (using ANOVA), with the carer sample including a significantly lower proportion of employed and unemployed individuals and students, but a significantly higher proportion of retired individuals, married individuals and individuals owning their own home. EQ-5D scores were similar between the general population and carer DEMQOL-Proxy-U samples.

The differences in sociodemographics may affect elicited values and these differences can be taken into account when modelling the data using the inclusion of sociodemographic characteristics in the model specification. Overall, the person with dementia and carer samples had a significantly higher proportion of respondents for whom the interviewers reported that it was doubtful that they understood the TTO tasks. This can be taken into account by estimating the preferred regression model excluding respondents for whom it was doubtful that they understood the TTO tasks.

Descriptive statistics of health-state utility values

Table 24 presents descriptive statistics of observed TTO values for DEMQOL-U elicited from the general population (as also reported in *Chapter 5*) and from people with dementia. For every health state, the mean and median TTO values are higher for the general population sample than for the patient sample. The range of mean utility values is larger for the patient sample (from 0.816 to –0.023) than for the general population sample (from 0.955 to 0.190). The ordering of states differs by population. The ordering is logically consistent for each population as for many states it is not possible to determine a logical ordering as one state is not consistently better across all dimensions.

TABLE 23 Characteristics of respondents by population and classification system

Characteristic	DEMQOL-U, general population (*n* = 78)	DEMQOL-U, patient population (*n* = 71)	ANOVA *p*-value	DEMQOL-Proxy-U, general population (*n* = 77)	DEMQOL-Proxy-U, carer population (*n* = 71)	ANOVA *p*-value
Age (years), mean (SD)	49.6 (17.0)	78.4 (7.7)	< 0.001	50.7 (16.2)	69.8 (12.7)	< 0.001
Age distribution (%)[a]			< 0.001			< 0.001
18–40 years	30.8	0.0		32.5	0.0	
41–65 years	48.7	8.5		45.4	33.8	
> 65 years	20.5	91.5		22.1	66.2	
Female (%)	52.6%	54.9%	0.77	51.9%	63.4	0.16
Married/partner (%)	66.7	62.0	0.55	74.0	85.9	0.07
Employment status (%)[a]			< 0.001			0.002
Employed or self-employed	53.8	2.8		45.5	29.6	
Unemployed	2.6	0.0		2.6	0.0	
Long-term sick	3.8	0.0		1.3	0.0	
Full-time student	1.3	0.0		5.2	0.0	
Retired	24.4	90.1		33.8	62.0	
Own home outright or with a mortgage (%)	79.5	87.3	0.20	72.7	94.4	< 0.001
Secondary school is highest level of education (%)	34.6	50.7	0.05	35.1	38.0	0.71
Interviewer reported that was doubtful respondent understood TTO tasks (%)	1.3	12.7	0.001	0.0	7.0	0.02
EQ-5D score (SD)	0.87 (0.19)	0.85 (0.14)		0.79 (0.25)	0.78 (0.19)	

a To adhere to statistical assumptions categories are merged if there are zero observations or fewer than five expected observations and chi-squared *p*-values are reported.

TABLE 24 Descriptive statistics of observed TTO values for DEMQOL-U

	General population (*n* = 78)			Patient population (*n* = 71)		
Health state	Mean	SD	Median	Mean	SD	Median
11111	0.955	0.153	1.000	0.816	0.241	0.900
12231	0.894	0.160	1.000	0.633	0.297	0.700
32143	0.673	0.302	0.663	0.399	0.398	0.500
41212	0.611	0.364	0.625	0.436	0.347	0.500
23424	0.566	0.340	0.600	0.435	0.327	0.500
44221	0.544	0.367	0.525	0.327	0.439	0.475
43442	0.403	0.412	0.475	0.161	0.455	0.225
44444	0.190	0.484	0.250	–0.023	0.456	0.000

Table 25 reports descriptive statistics of observed TTO values for DEMQOL-Proxy-U elicited from the general population (as also reported in *Chapter 5*) and from carers of people with dementia. For every health-state mean and median TTO values are higher for the general population sample than for the carer sample. The range of mean utility values is larger for the carer sample (from 0.857 to 0.049) than for the general population sample (from 0.918 to 0.370). The ordering of states differs by population, but the ordering is logically consistent for each population. *Figure 9* shows differences in the distribution of TTO values by population for each measure. General population TTO values are negatively skewed with a large proportion of responses at 1, whereas patient and carer TTO values are also negatively skewed but peak at 0.5.

Simple *t*-tests revealed that health-state utility values for the general population and for people with dementia are significantly different for DEMQOL-U [general population mean (SD) 0.604 (0.410), person with dementia mean (SD) 0.398 (0.447), $t = 8.279$, degrees of freedom (df) = 1153, $p < 0.001$] and health-state utility values for the general population and for carers are significantly different for DEMQOL-Proxy-U [general population mean (SD) 0.707 (0.350), carer mean (SD) 0.531 (0.410), $t = 7.946$, df = 1120, $p < 0.001$]. Wilcoxon rank-sum tests reached the same conclusions ($p < 0.001$ and $p < 0.001$ respectively).

Regression analysis

Regression analysis examining the relationship between elicited health-state utility values, population, health-state severity and sociodemographic characteristics of respondents is presented in *Table 26* for DEMQOL-U and *Table 27* for DEMQOL-Proxy-U. Random-effects models are reported as the Hausman test confirmed for both classification systems that fixed-effects models would produce similar estimates at reduced efficiency. For the DEMQOL-U classification system, TTO values were regressed on state-level dummy variables in model (1); state-level dummy variables and a population dummy variable in model (2); state-level dummy variables and interaction terms to reflect the interaction between the specific health state and population in model (3); and explanatory variables used in model (3) while controlling for sociodemographic characteristics in model (4). Model (5) used the same specification as model (4) and was estimated on a sample excluding respondents whose understanding of the TTO task was doubted by the interviewers. Models (6), (7), (8), (9) and (10) estimated for the DEMQOL-Proxy-U classification system have the same specification and criteria for sample selection as models (1), (2), (3), (4) and (5) respectively. Models using a range of sociodemographic variables as explanatory variables were estimated and the best models (using significance of coefficients, within *R*-squared, between *R*-squared, overall *R*-squared, RMSE of predictions and Wald chi-squared) are presented here.

TABLE 25 Descriptive statistics of observed TTO values for DEMQOL-Proxy-U

	General population ($n = 77$)			Carer population ($n = 71$)		
Health state	Mean	SD	Median	Mean	SD	Median
1111	0.918	0.185	1.000	0.857	0.211	1.000
1222	0.869	0.174	0.925	0.731	0.251	0.700
3112	0.804	0.255	0.925	0.666	0.313	0.725
1341	0.807	0.236	0.900	0.640	0.250	0.700
3234	0.672	0.272	0.700	0.458	0.361	0.500
2424	0.638	0.341	0.700	0.464	0.364	0.500
4411	0.580	0.395	0.625	0.380	0.406	0.500
4444	0.370	0.482	0.400	0.049	0.465	0.000

TABLE 26 Regression analysis of DEMQOL-U health-state values across general population and patient respondents

	(1)	(2)	(3)	(4)	(5)
States					
12231	–0.119[a]	–0.119[a]	–0.061	–0.061	–0.061
32143	–0.346[a]	–0.346[a]	–0.281[a]	–0.281[a]	–0.281[a]
41212	–0.361[a]	–0.361[a]	–0.344[a]	–0.344[a]	–0.344[a]
23424	–0.385[a]	–0.385[a]	–0.388[a]	–0.388[a]	–0.388[a]
44221	–0.448[a]	–0.448[a]	–0.411[a]	–0.411[a]	–0.411[a]
43442	–0.601[a]	–0.601[a]	–0.552[a]	–0.552[a]	–0.552[a]
44444	–0.800[a]	–0.800[a]	–0.765[a]	–0.765[a]	–0.765[a]
Patient		–0.206[a]			
Patient interaction terms					
11111 × patient			–0.139[b]	–0.134[c]	–0.106
12231 × patient			–0.261[a]	–0.257[a]	–0.235[a]
32143 × patient			–0.274[a]	–0.270[a]	–0.236[a]
41212 × patient			–0.175[a]	–0.171[b]	–0.164[b]
23424 × patient			–0.131[b]	–0.127[c]	–0.125[c]
44221 × patient			–0.217[a]	–0.213[a]	–0.181[a]
43442 × patient			–0.241[a]	–0.237[a]	–0.211[a]
44444 × patient			–0.213[a]	–0.208[a]	–0.194[a]
Sociodemographics					
Female				–0.073	–0.123[a]
Unemployed				–0.116	–0.137
Long-term sick				–0.077	–0.111
Retired				–0.036	–0.030
Student				–0.290	–0.263
Homemaker				–0.147	–0.138
Secondary school is highest level of education				0.008	–0.025
Renting accommodation				0.021	0.096[c]
Constant	0.888[a]	0.987[a]	0.954[a]	1.023[a]	1.046[a]
Observations	1192	1192	1192	1192	1120
Number of subjects	149	149	149	149	140
Within R^2	0.000	0.478	0.483	0.483	0.489
Between R^2	0.000	0.137	0.137	0.181	0.234
Overall R^2	0.287	0.341	0.385	0.362	0.394
Root MSE	0.264	0.264	0.264	0.264	0.261
Wald χ^2	948.43	971.76	986.02	993.39	962.79

a Significant at 10% level.
b Significant at 5% level.
c Significant at 1% level.
Reference state is 11111 valued by the general population.

TABLE 27 Regression analysis of DEMQOL-Proxy-U health-state values across general population and carer respondents

	(6)	(7)	(8)	(9)	(10)
States					
1222	–0.087[a]	–0.087[a]	–0.050	–0.050	–0.050
3112	–0.151[a]	–0.151[a]	–0.114[a]	–0.114[a]	–0.114[a]
1341	–0.163[a]	–0.163[a]	–0.112[a]	–0.112[a]	–0.112[a]
3234	–0.320[a]	–0.320[a]	–0.246[a]	–0.246[a]	–0.246[a]
2424	–0.335[a]	–0.335[a]	–0.280[a]	–0.280[a]	–0.280[a]
4411	–0.405[a]	–0.405[a]	–0.338[a]	–0.338[a]	–0.338[a]
4444	–0.673[a]	–0.673[a]	–0.548[a]	–0.548[a]	–0.548[a]
Carer		–0.177[a]			
Carer interaction terms					
1111×carer			–0.061	–0.053	–0.044
1222×carer			–0.138[a]	–0.131[b]	–0.117[b]
3112×carer			–0.138[a]	–0.131[b]	–0.111[c]
1341×carer			–0.167[a]	–0.159[a]	–0.138[b]
3234×carer			–0.215[a]	–0.207[a]	–0.203[a]
2424×carer			–0.174[a]	–0.167[a]	–0.156[a]
4411×carer			–0.200[a]	–0.192[a]	–0.180[a]
4444×carer			–0.321[a]	–0.314[a]	–0.313[a]
Sociodemographics					
Female				–0.022	–0.020
Unemployed				0.025	0.027
Long-term sick				–0.447[c]	–0.449[c]
Retired				–0.062	–0.066
Student				–0.075	–0.067
Homemaker				0.080	0.126
Secondary school is highest level of education				0.031	0.046
Renting accommodation				–0.023	–0.029
Constant	0.889[a]	0.974[a]	0.918[a]	0.950[a]	0.943[a]
Observations	1184	1184	1184	1184	1144
Number of subjects	148	148	148	148	143
Within R^2	0.000	0.431	0.445	0.445	0.450
Between R^2	0.000	0.131	0.131	0.180	0.181
Overall R^2	0.263	0.314	0.322	0.341	0.344
Root MSE	0.246	0.246	0.244	0.244	0.243
Wald χ^2	780.38	802.48	840.11	848.25	837.676

a Significant at 10% level.
b Significant at 5% level.
c Significant at 1% level.
Reference state is 1111 valued by general population.

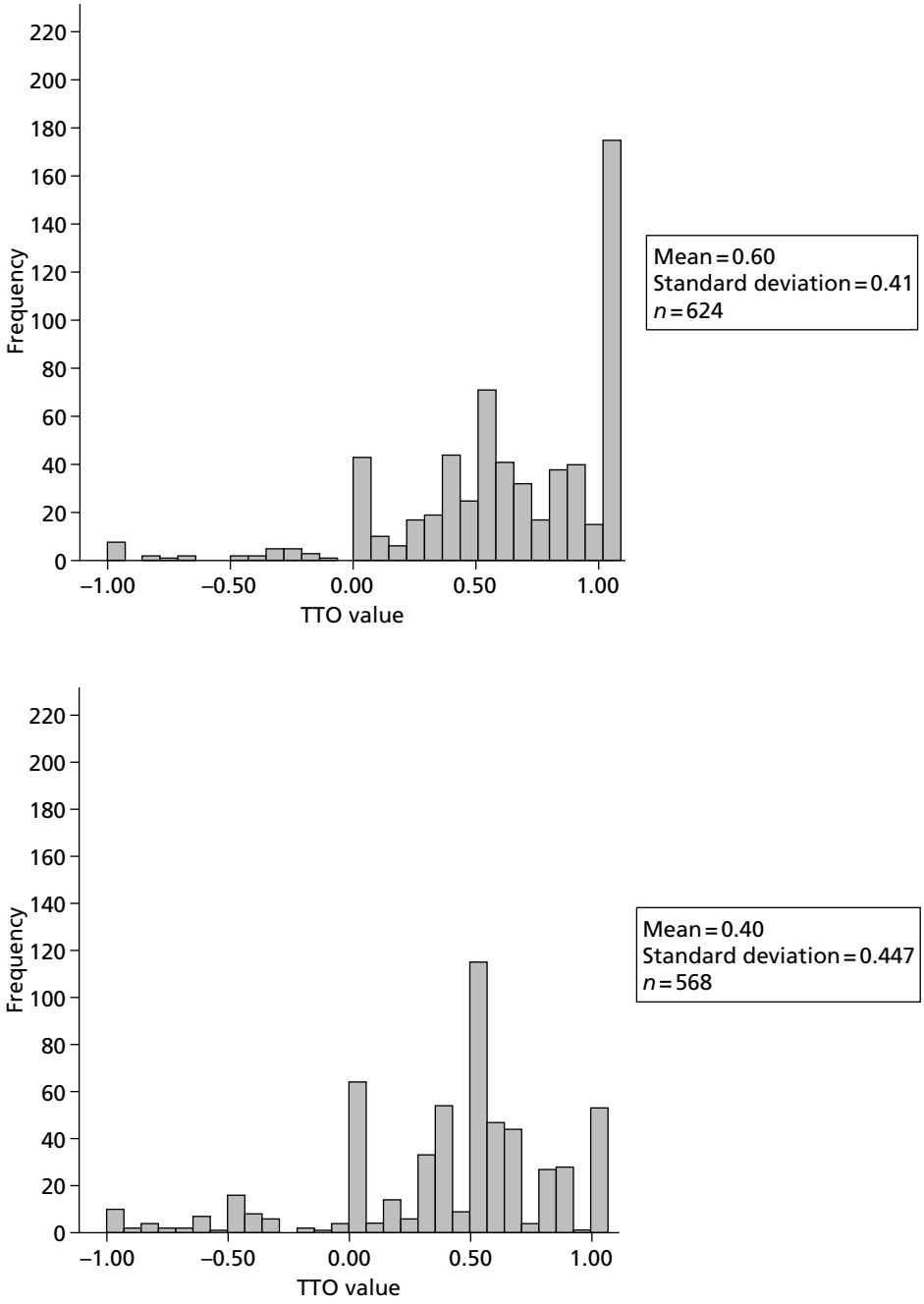


FIGURE 9 Histogram of observed TTO values: (a) DEMQOL-U valued by the general population sample; (b) DEMQOL-U valued by the patient sample; (c) DEMQOL-Proxy-U valued by the general population sample; (d) DEMQOL-Proxy-U valued by the carer sample.

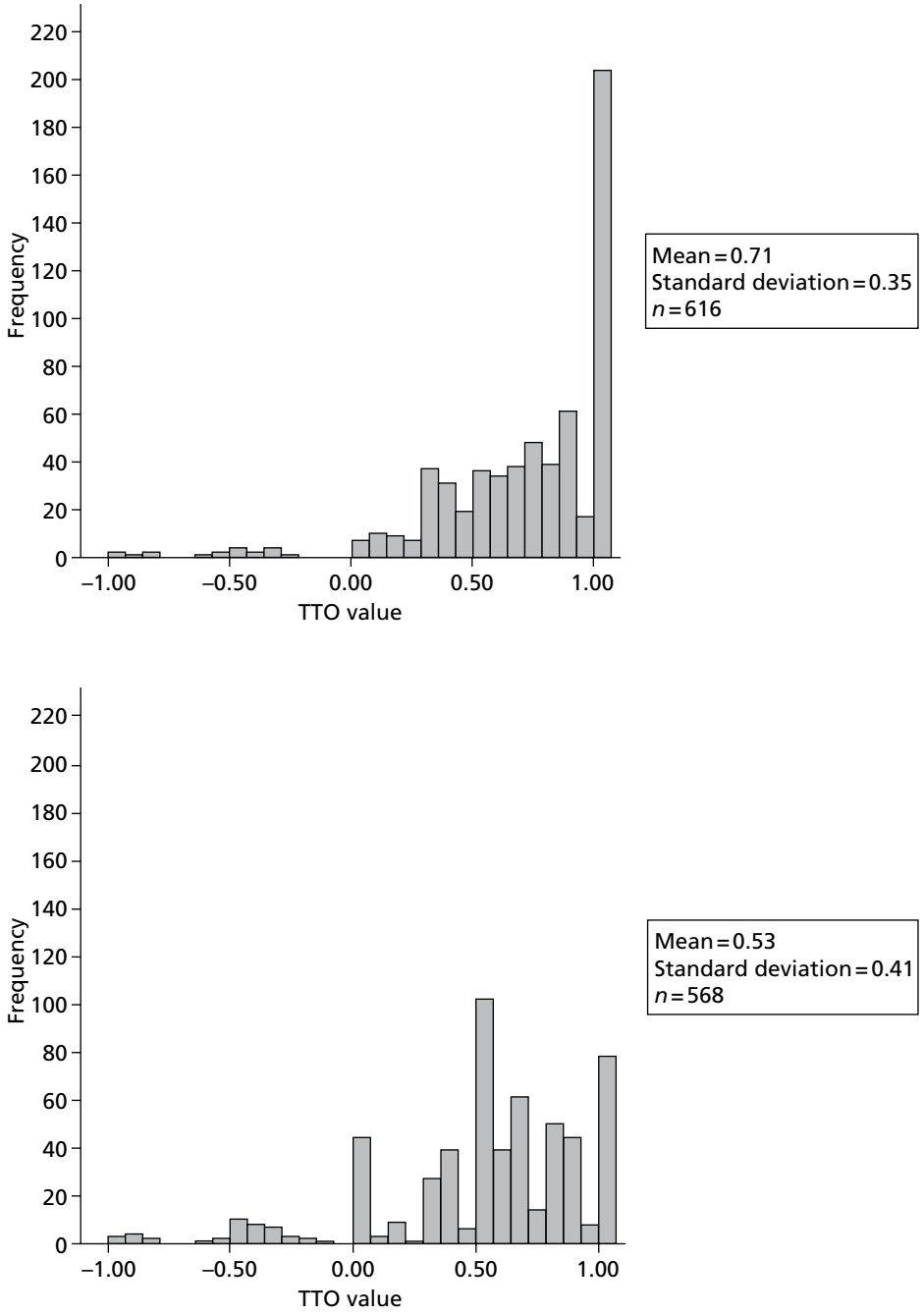


FIGURE 9 Histogram of observed TTO values: (a) DEMQOL-U valued by the general population sample; (b) DEMQOL-U valued by the patient sample; (c) DEMQOL-Proxy-U valued by the general population sample; (d) DEMQOL-Proxy-U valued by the carer sample. (*continued*)

Spearman correlation coefficients were used to indicate correlations among the sociodemographic characteristics that can be used as potential explanatory variables. All sociodemographic variables included in the regression had poor intercorrelation ($< |0.3|$) with the exception of moderate correlation ($< |0.7|$) between retired and patient variables for DEMQOL-U (correlation 0.67) and between renting and unemployed variables for DEMQOL-Proxy-U (correlation 0.32). This means that there was no problem of multicollinearity in the explanatory variables included in the regressions. Age was not included as an explanatory variable as it was correlated with the employment status predictors, which had a greater improvement in model performance than age and age-squared. Health status using EQ-5D was not included as an explanatory variable because of a concern that it was not an accurate measure of health status for the patient population. Furthermore, inclusion of health status was not significant and did not improve model performance for the DEMQOL-Proxy-U models. The population variable was not included in models (3), (4), (5), (8), (9) and (10) because of perfect collinearity between the population label and the eight population×state interaction terms.

The results show that health-state dummy variables were significant at the 1% level in all models except for state 12231 in models (3), (4) and (5) and state 1222 in models (8), (9) and (10). The size of the coefficients was consistent as the decrement in the elicited utility value was larger for more severe health states (indicated using the modelled utility values from the original valuation study). The only exception is an inconsistency between health states 3112 and 1341 in models (8), (9) and (10) for DEMQOL-Proxy-U.

The inclusion of a dummy variable to represent population in models (2) and (7) was significant at the 1% level and improved model performance across all goodness of fit statistics. The coefficient was negative, demonstrating that respondents with dementia and carers value states significantly lower than the general population.

Interaction effects reflecting the interaction between the specific health state and the person with dementia or carer population had negative coefficients, meaning that there is a reduction in utility value for respondents from the patient or carer populations. The inclusion of the interaction terms reduced the absolute size of the coefficient for the health-state dummy variables with the exception of DEMQOL-U state 23424 in model (3). The size of the interaction coefficients varied by health state, indicating that the impact of population differs by health-state severity, but there was not a clear pattern. Interaction effects for DEMQOL-U health states valued by people with dementia with more severe health problems in the (fourth) negative emotion dimension (health states 12231, 32143, 43442, 44444) had larger coefficients than the other health states, but this could be due to coincidence. Interaction effects for DEMQOL-Proxy-U health states valued by carers with the most severe level in one or more dimensions (health states 1341, 2424, 3234, 4411, 4444) had larger coefficients than other states, but again this could be due to coincidence. The coefficient for the interaction effect for the worst health state for DEMQOL-Proxy-U valued by carers is noticeably larger than those for the other interaction effects. The inclusion of interaction effects improved model performance as measured using within and overall *R*-squared and Wald chi-squared.

The addition of the sociodemographic characteristics in models (4) and (9) had only a minor impact on the coefficients of the other variables and led to only a small improvement in model performance. Few sociodemographic variables were significant. For the DEMQOL-U models, employment status variables of 'homemaker' and 'retired' were significant in model (1) but not in models (2) and (3). No sociodemographic variables were significant in model (4) while the variables 'female' and 'renting accommodation' were significant in model (5). For the DEMQOL-Proxy-U models the employment status variable 'long-term sick' was significant in models (9) and (10) (although note that the number of respondents in this category was small). This suggests that, although the sociodemographics are different for the different samples, as found in *Table 1*, these differences do not impact on elicited health-state utility values.

Excluding respondents whose understanding of the TTO task was doubted by the interviewers did not have a noticeable impact on coefficients for health-state severity and interaction effects. However, the 'female' and 'renting accommodation' sociodemographic variables became significant in model (5) for DEMQOL-U.

Discussion

The results demonstrate that the population used to value dementia health states affects health-state utility values elicited using TTO. This is in accordance with the literature.[80] Health-state utility values elicited from people with dementia and carers of people with dementia were lower than health-state utility values elicited from the general population. This is different from many studies which have found that patients provide higher values than the general population for other health states using a variety of valuation techniques, but is consistent with a recent review and meta-analysis which found that patient values were significantly lower using the TTO elicitation technique in particular.[80] Furthermore, the majority of the literature focuses on physical health states whereas this study focused on mental health and cognition, with previous research suggesting that patients and the general population weight problems in physical and mental health differently.[82]

Modelling of the health-state utility values indicated that the differences in values were due to the population per se and not to differences in the sociodemographic composition of the populations. This was demonstrated because controlling for sociodemographic characteristics had a minimal impact on the coefficients of the other variables and resulted in only a minor improvement in model performance and few of the sociodemographic coefficients were significant. Regression analysis and descriptive statistics of observed values indicated that the difference in values by population differed for each health state, sometimes indicating a different ordering of the health states. This different ordering of health states is not an inconsistency across populations; it represents differences in the valuation of and trade-off between different dimensions. This suggests that the choice of whose values are used to produce utility values for use in cost-effectiveness analysis will have an impact on the results and could potentially affect resource allocation decisions. This suggests the need to produce a full value set for all DEMQOL-U and DEMQOL-Proxy-U health states using values elicited from people with dementia and from carers of people with dementia respectively.

The finding that general population values are higher than patient values for dementia states suggests that members of the general population may be systematically undervaluing the impact on quality of life arising from problems with cognition and mental health. This may reflect public perception in general of health problems such as dementia. However, it is important to note that it may be due to the wording and labelling of the classification system. General population respondents were not informed at any point that the study was about dementia, whereas people with dementia and carers knew that the study was about dementia. This contextualised health states for the people with dementia and carers, meaning that they had a greater understanding of the health states, but it also meant that they knew the underlying cause of the health state. A recent study examining labelling effects found that introducing condition labels into health-state descriptions impacted on health-state utility values elicited from the general population, and that the impact differed by condition and health-state severity.[37,71] It can be argued that the inclusion of a condition label can better inform the general population about the health state because respondents may value differently, for example, feeling frustration as a result of epilepsy and feeling frustration as a result of dementia. However, it can also be argued that the inclusion of a condition label may mean that respondents take into account preconceptions of the condition or mortality of the condition, and these are factors that should not be taken into account in the elicitation of health-state utility values. As the aim of the general population valuation survey was to elicit values to inform resource allocation decisions across all conditions and patient groups, there must be comparability between the health-state descriptions used here and in other valuation surveys, regardless of the underlying condition. Otherwise, this could imply that, for example, a given generic EQ-5D state has a lower utility value for a patient with epilepsy than for a patient with diabetes. The exclusion of a dementia health-state label in the general population study may

therefore have impacted on elicited utility values, and there is a possibility that if general population values were obtained with the inclusion of a dementia label these values may be closer to the values elicited from people with dementia and carers. However, an important difference is likely to remain: the difference in the valuation of and trade-off between the different dimensions in the classification system. Qualitative research or a full-scale valuation study similar to the study outlined in *Chapter 5* for people with dementia and carers would indicate differences in the trade-off between and relative importance of the dimensions in comparison with the general population. Further research in this area is encouraged but is likely to be constrained by the difficulties of conducting this type of research in these populations.

One potential limitation in the comparisons made between the general population values and the values from people with dementia and carers is that the general population study involved ranking as a warm-up task prior to the TTO task whereas this was excluded from the survey for people with dementia and carers. Starting the TTO task without the rank task, which helps respondents think about how they would value different health states, may have affected the understanding of the TTO task and may also have affected the values provided by people with dementia and carers. However, it could be argued that people with dementia and carers knew that the states were dementia states and may already have an idea before the interview about how they value different dementia health states.

Among people with dementia and carers there were significantly higher proportions of respondents who were reported by the interviewers as doubtful that they understood the TTO task, but at 12.7% and 7.0% of respondents, respectively, these proportions are not especially high. The preferred regression models were also estimated on samples excluding respondents whose understanding of the TTO task was doubted by the interviewers. The exclusion of these respondents did not have a large impact on the coefficients for health-state severity and interaction effects but some sociodemographic characteristic variables became significant.

Another potential limitation of this study is the involvement of people with dementia with neuropsychological and cognitive problems to value health states using a cognitively demanding elicitation technique. The study was designed at all stages to take into consideration the health and competencies of the patient population: the team was engaged in conversation with the Alzheimer's Society through its Quality Research in Dementia panel prior to designing the study; patients were referred to the study by clinicians specialising in dementia; and the interviews were immediately terminated if the person with dementia suffered from fatigue, misunderstanding or distress. For these reasons the samples are made up of patients with mild dementia and carers of people with mild dementia. This means that the samples may not be representative of people with more severe dementia and their carers but this is a constraint of conducting a valuation study for this patient group. The MVH TTO protocol has been widely used in valuation studies of the general population but may be more challenging for respondents with cognitive problems. The TTO is no more challenging than many other elicitation techniques, such as SG or person trade-off, and arguably is less cognitively demanding than ranking, which requires the simultaneous consideration of multiple health states. Other options that may be less cognitively demanding include ordinal techniques such as discrete choice techniques and best–worst scaling. However, these require many more data and either increase the burden on each respondent and/or require a much larger sample. There are also challenges regarding the anchoring of these values onto the 1–0 full health–dead scale (see reference 19 for an overview). A VAS may be cognitively easier for respondents, but to obtain values anchored onto the 1–0 full health–dead scale would require the valuation of dead alongside each state.

The study used the same valuation protocol as the EQ-5D as recommended by NICE to ensure that elicited utilities are comparable with the UK tariff of the EQ-5D. However, the choice of TTO and the valuation protocol may have impacted on results. A recent meta-analysis of patient and general population values found that results differed by valuation technique[80] and other research has found that valuation results may in general differ by protocol.[87] Further research in this area is needed.

Conclusion

Dementia health-state utility values elicited using TTO differ by the population used to elicit the utility values, with people with dementia and carers of people with dementia giving systematically lower utility values than members of the general population. These differences in values were due to the population per se and not to differences in the sociodemographic characteristics of the populations. The general population underestimated the impact of dementia compared with people with dementia and their carers. The ordering of health states also differed by population for some health states, indicating a difference in the valuation of and trade-off between different dimensions. These results suggest that the population used to produce dementia health-state utility values could impact on the results of cost-effectiveness analysis and potentially affect resource allocation decisions.

Chapter 7 Application of the preference-based index to the Health Technology Assessment Study of Antidepressants for Depression in Dementia trial data

Introduction

If the DEMQOL-U and DEMQOL-Proxy-U are to be used alongside or instead of generic preference-based measures it is important to assess their psychometric validity, responsiveness and level of agreement between patient and carer report. This can be assessed by applying psychometric methods to data sets containing responses to the DEMQOL system alongside generic preference-based and non-preference-based measures. The issue of how condition-specific preference-based measures compare with generic preference-based measures is particularly important as this indicates the likely impact of using DEMQOL-U and DEMQOL-Proxy-U instead of generic measures to generate QALY values for use in economic evaluation. The psychometric testing of DEMQOL-U and DEMQOL-Proxy-U is reported in this chapter using data from the HTA-SADD trial, a multicentre randomised double-blind placebo-controlled trial of the clinical effectiveness of sertraline and mirtazapine.[46] This study included an assessment of HRQL using the generic preference-based measure EQ-5D and the DEMQOL system.

In this chapter we compared the validity, patient/proxy agreement and responsiveness of the EQ-5D and the DEMQOL-U and DEMQOL-Proxy-U utility measures. Validity was examined in terms of the ability to discriminate between different levels of severity (i.e. known group validity), the convergence between the DEMQOL utility measures and the other measures of dementia-related constructs, and the level of agreement between DEMQOL-U, DEMQOL-Proxy-U and patient- and carer-reported EQ-5D. Agreement between patient- and carer-reported utilities was assessed. Responsiveness was assessed by investigating sensitivity to change in quality of life over time.

Method

Data source

The data used for the psychometric analyses were from the HTA-SADD study of the use of antidepressants for depression in dementia.[46] HTA-SADD is a multicentre parallel-group double-blind placebo-controlled pragmatic RCT of the clinical effectiveness of sertraline and mirtazapine. Participants were eligible to participate in the study if they had probable or possible Alzheimer's disease and depression (4+ weeks' duration), and a CSDD score of 8+ (indicating significant depressive symptoms). They were drawn from nine English old age psychiatry services. Patients were excluded only if they were clinically too critical (e.g. were at risk of suicide), had a contraindication to medication, were taking antidepressants, were in another trial or had no carer. Participants were randomly allocated to one of three treatment arms, sertraline, mirtazapine or placebo, all with usual care. Target doses were 150 mg of sertraline or 45 mg of mirtazapine daily. The objective of the study was to determine the clinical effectiveness of sertraline and mirtazapine compared with placebo in reducing depression at 13 weeks post randomisation. In total, 326 participants were randomised. The study found no statistically significant differences in depression score between groups at 13 or 39 weeks. The placebo group had fewer adverse reactions (26%) than the sertraline (43%) or mirtazapine (41%) groups and fewer serious adverse events rated as severe.

Measures

DEMQOL-U and DEMQOL–Proxy-U

The DEMQOL-U and DEMQOL-Proxy-U utility scores were generated using the general population algorithms reported in *Chapter 5*. The utility scores are anchored on the 1–0 full health–dead scale, with scores < 0 equivalent to states worse than dead. The range of scores for DEMQOL-U is from 0.243 to 0.986 and that for DEMQOL-Proxy-U is from 0.363 to 0.937.

European Quality of Life-5 Dimensions

The EQ-5D[53] is a standardised instrument used as a measure of health outcome and in the assessment of cost–utility. It is designed for self-completion by respondents and can also be completed by proxy. The EQ-5D consists of a descriptive health-state classification system with five domains (mobility, self-care, usual activity, pain/discomfort and anxiety/depression) and a VAS 'health thermometer'. Each attribute in the classification system has three levels ('no problem', 'some problems' and 'major problems'), thus defining a total of 243 possible health states, to which 'unconscious' and 'dead' have been added for a total of 245 health states. Preferences for the scoring function used in the HTA-SADD trial were measured using the MVH UK value set produced using modelled utility values obtained using the TTO technique on a random sample of approximately 3000 adults in the UK.[45] EQ-5D utility scores range from –0.594 to 1. The 'health thermometer' is a subjective evaluation of the respondent's health status on a scale between 0 and 100, with 0 representing the worst imaginable health state and 100 representing the best imaginable health. In the HTA-SADD trial the EQ-5D was reported by both the patient and the carer (CEQ-5D). The patient completion data were compared with DEMQOL-U data and the carer completion data were compared with DEMQOL-Proxy-U data.

Cornell Scale for Depression in Dementia

The CSDD was specifically developed to assess signs and symptoms of depression in patients with dementia using both patient and carer report.[49] The final ratings of the 19 CSDD items represent the research worker's impression rather than the responses of the informant or the patient. Each item is rated on symptom severity ('absent', 'mild or intermittent' and 'severe'). A score of ≥ 11 indicates probable depression and a score of ≥ 18 indicates definite depression. The scale has been described as the best available to assess mood in the presence of cognitive impairment.[88]

Mini Mental State Examination

The MMSE is widely used in screening for dementia. It generates scores between 0 and 30, with scores between 0 and 9 indicating severe impairment, scores between 10 and 20 indicating moderate impairment and scores between 21 and 30 indicating mild impairment.[50] The person with dementia is assessed on orientation, memory and attention, and ability to follow commands on tasks, including copying a diagram of overlapping hexagons, writing a sentence and reading and following a printed instruction.

Bristol Activity of Daily Living Scale

The BADLS was specifically designed for use with people with dementia living in the community and participating in clinical trials.[51] The levels of disability between which the scale aims to discriminate were generated by carers. The BADLS is sensitive to change and has good test–retest reliability. Scores range from 0 to 60, with 0 representing unimpaired daily activities and 60 representing impaired daily activities.[51] The BADLS is completed using proxy report by carers.

Neuropsychiatric Inventory

The NPI assesses 12 behavioural disturbances in dementia using a screening strategy. The nature, frequency and severity of behaviour and psychiatric symptoms in dementia are assessed. NPI scores range from 0 (no disturbance) to 144 (maximum disturbance). In the original validation study, the NPI demonstrated content and concurrent validity as well as inter-rater, test–retest and internal consistency reliability.[52] NPI is completed using proxy report by carers.

Analysis

Descriptive analysis

Summary statistics were generated to describe the distribution of responses on the self-report EQ-5D, CEQ-5D, DEMQOL-U and DEMQOL-Proxy-U. The proportion of responses endorsing the best (i.e. ceiling effect) and worst (i.e. floor effect) ratings was explored. This is because large numbers at the ceiling or floor imply that the measure cannot capture an improvement or deterioration in health status respectively. The extent of missing data was also calculated to provide an indication of the acceptability of each measure for respondents with cognitive impairment.

Agreement

Bland–Altman plots[89] were employed to assess agreement between responses to EQ-5D and responses to DEMQOL-U and DEMQOL-Proxy-U to see if the DEMQOL measures displayed a similar pattern of response. As the true tariff for each participant is not known, the horizontal (x) axis uses the average of the two utility values for every individual [e.g. $x = (\text{utility 1} + \text{utility 2})/2$]. The vertical ($y$) axis is the difference between the utility scores. To assess agreement at the less severe end of the scale (where a ceiling effect is commonly reported for EQ-5D[90]), we also assessed respondent scores on the DEMQOL-U and DEMQOL-Proxy-U for those who reported full health on EQ-5D and CEQ-5D respectively.

Patient/proxy agreement

The utility values generated from self-report were compared with those generated from carer report to examine the extent of agreement between the two sets of values. Good agreement would suggest that self- and carer-based utility scores could possibly be used interchangeably for evaluation studies or alongside each other as complementary measures. The mean difference between the rater scores was assessed. Intraclass correlation coefficients, or ICCs,[91] were employed to quantify the extent of agreement, while Bland–Altman plots provided a graphical representation of the discrepancy between self- and carer-based utility values. The y-axis of these plots represents the discrepancy between self- and carer-based utility values (e.g. $y = \text{DEMQOL-U} - \text{DEMQOL-Proxy-U}$) while the x-axis represents the participants' utility values. However, as the true utility for each participant is not known, the horizontal axis uses the average of self- and carer-based utility values for every individual [e.g. $x = (\text{DEMQOL-U} + \text{DEMQOL-Proxy-U})/2$]. When there is good agreement (i.e. $\text{DEMQOL-U} - \text{DEMQOL-Proxy-U} \approx 0$), a scatter of points would cluster horizontally around the line $y = 0$ (which indicates no discrepancy) for all possible utility values on the x-axis. All the plots were generated using the Stata module provided by Mander.[92]

Convergent validity

Convergent validity assesses the association between the DEMQOL utility measures and existing validated measures of similar constructs. Evidence of convergent validity would require a reliable association between the utility scores, which represent the value placed on health, and the scores of the other indicators, which measure perceptions of health across a number of related constructs. The convergent validity of the EQ-5D, DEMQOL and DEMQOL-U, and the CEQ-5D, DEMQOL-Proxy and DEMQOL-Proxy-U was examined in relation to measures of cognitive impairment (MMSE), depression (CSDD), neurobehavioural problems (NPI) and daily functioning (BADLS). A range of measures were assessed as there is no gold standard for the measurement of populations with cognitive impairment. Correlations were assessed as very strong (≥ 0.6), strong (< 0.6 to ≥ 0.5), moderate (< 0.5 to ≥ 0.3) and weak (< 0.3).

Known-group validity

Known-group validity was assessed using severity thresholds recommended for the MMSE and CSDD. Specifically, we hypothesised that the group with mild cognitive impairment (MMSE score > 20) would have the highest utility values relative to those with moderate (MMSE score 10–20) and severe (MMSE score < 10) impairment. Similarly, patients with less severe depressive symptoms (CSDD score ≤ 10) should have utility values that are higher than those of the groups with moderate (CSDD score 11–17) and severe (CSDD score ≥ 18) depression. Effect size indices were also calculated to provide an indication of the average group difference taking into account the variability observed in the group with least impairment

(MMSE score > 20; CSDD score ≤ 10). Based on commonly cited guidelines,[93] values of 0.2, 0.5 and 0.8 denote small, medium and large effect sizes respectively.

Responsiveness to change

Responsiveness to change was examined using the MCID thresholds recommended by the DOMINO trial group.[56] Specifically, a score change of 1.4 on the MMSE, 3.5 on the BADLS and 8.0 on the NPI was deemed a clinically significant change in health status. Based on these thresholds, we classified the participants into three groups: MCID improvement, no MCID change and MCID deterioration. We hypothesised that the group with MCID improvement (in MMSE, BADLS or NPI) should demonstrate the largest increment in utility or HRQL score relative to the other groups at follow-up. On the other hand, the group with MCID deterioration should demonstrate a decline in utility or HRQL score at follow-up. In the absence of MCID guidelines for the CSDD, we adopted a common practice in antidepressant trial investigations of defining clinically significant changes as a percentage of the baseline score for each patient. Specifically, a 30% reduction from baseline CSDD score was classified as clinical improvement. Conversely, a 30% increase from baseline CSDD score would be classified as clinical deterioration. As before, patients with clinical improvement should demonstrate the largest increment in utility or HRQL score. Those with clinical deterioration, on the other hand, should demonstrate a decline in utility or HRQL score. Standardised response means (SRMs) were also used to compare the responsiveness of the respective utility and non-preference-based measures. Based on commonly cited guidelines,[93] SRM values of 0.2, 0.5 and 0.8 denote fair, moderate and strong responsiveness respectively.

Results

Sample characteristics

The study sample comprised participants recruited from the HTA-SADD trial. The current analyses are based on data from 326 patients (*Table 28*). All but one had data for at least one of the preference-based measures. The mean age of the sample was 79 years, and 68% were female. The majority of the sample reported a mild or moderate level of cognitive impairment (88%), and 69% self-reported depression for a duration of > 6 months.

Descriptive statistics

Considering patient responses, the mean utility score at baseline for the EQ-5D was 0.68 and for DEMQOL-U was 0.80 (*Table 29*). The median value for the EQ-5D is 0.75, and as the median value exceeds the mean the EQ-5D scores are positively skewed. The median for DEMQOL-U is 0.82. For both measures the scores improved at 13 and 39 weeks' follow-up, although the mean EQ-5D score increased by more than the mean DEMQOL-U score (0.10 vs 0.02). The mean carer-reported EQ-5D score at baseline was 0.47 and the mean carer-reported DEMQOL-Proxy-U score was 0.79. Again, scores on both measures improved at 13 weeks' follow-up but CEQ-5D scores decreased slightly at 39 weeks. However, the overall increase in CEQ-5D scores was larger than the increase in DEMQOL-Proxy-U scores (0.06 vs 0.02).

Ceiling effects were higher for the patient-reported EQ-5D utilities than for DEMQOL-U utilities (16.0 vs 3.3) at baseline (*Table 30*). There was no evidence of a floor effect. At week 13, the proportion of the sample at the ceiling on the EQ-5D increased to 27.7% in comparison with 5.8% on DEMQOL-U. At week 39, the EQ-5D ceiling effect is maintained (28.7%) but the proportion at the ceiling on DEMQOL-U drops to 1.9%. In terms of the dimension scores, both measures display evidence of a ceiling effect at baseline, which increases at follow-up (data not shown). However, the dimensions cannot directly be compared because of differences in wording and number of dimension levels. The DEMQOL-U also displays some evidence of a floor effect for some dimensions. The high ceiling effects observed for the EQ-5D and DEMQOL-U for some dimensions means that it may be difficult for some dimensions to be sensitive to improvements in health over the course of a study.

TABLE 28 Summary of baseline characteristics

Characteristic	*n*[a]	Mean (SD) or %
Participant profile		
Age (years)	325	79.4 (8.5)
Female	220	67.7
Residence (care home)	50	15.4
Dementia vascularity	240	2.1 (1.3)
Cognitive impairment		
MMSE total score	250	18.1 (6.6)
Mild (MMSE score > 20)	102	40.8
Moderate (MMSE score 10–20)	118	47.2
Severe (MMSE score < 10)	30	12.0
Depression duration		
Duration < 1 month	10	3.1
Duration 1–2 months	20	6.3
Duration 2–6 months	68	21.3
Duration > 6 months	221	69.3
Severity of depression		
CSDD total score	295	12.9 (4.2)
Non-case (CSDD score < 6)	96	32.5
Probable (CSDD score 6–10)	165	55.9
Definite (CSDD score > 10)	34	11.5
Activities of daily living		
BADLS total score	179	17.6 (11.9)
Neurobehavioral problems		
NPI total score	325	28.9 (18.6)
Carer profile		
Age (years)	260	63.0 (14.5)
Female	213	65.7
Living with participant	185	66.6

a Subgroup sizes do not add up because of missing data.

The ceiling effect is not as high for the CEQ-5D and DEMQOL-Proxy-U utility scores at baseline (2.5% and 6.0% respectively), but the EQ-5D still displays a higher proportion at the ceiling. This is maintained at week 13 (6.0% and 1.2%) and week 39 (8.5% and 1.5%). CEQ-5D has less of a ceiling effect than EQ-5D at the dimension level, but direct comparisons are not possible. The only dimension that can be directly compared across the DEMQOL-U and DEMQOL-Proxy-U is negative emotion, and the patient-reported measure displays more evidence of a ceiling effect on this dimension.

TABLE 29 Summary statistics for the EQ-5D, CEQ-5D, DEMQOL-U and DEMQOL-Proxy-U across the study period

Measure	*n*	Mean	Median
EQ-5D			
Baseline	293	0.68	0.75
Week 13	220	0.75	0.80
Week 39	174	0.78	0.81
DEMQOL-U			
Baseline	277	0.80	0.82
Week 13	207	0.81	0.84
Week 39	161	0.82	0.85
CEQ-5D			
Baseline	321	0.47	0.59
Week 13	251	0.56	0.66
Week 39	211	0.53	0.64
DEMQOL-Proxy-U			
Baseline	317	0.79	0.79
Week 13	244	0.81	0.82
Week 39	207	0.81	0.82

TABLE 30 Baseline missing data and floor and ceiling effects for the EQ-5D, CEQ-5D, DEMQOL-U and DEMQOL-Proxy-U across the study period

Measure	*n*	% missing	% floor	% ceiling
EQ-5D				
Tariff	293	10.1	0	16.0
Mobility	305	6.4	0.7	58.0
Self-care	303	7.1	3.3	77.6
Usual activities	296	9.2	7.8	59.5
Pain/discomfort	304	6.7	5.6	59.5
DEMQOL-U				
Tariff	277	15.0	0	3.3
Positive emotion	296	9.2	18.2	7.8
Memory	285	12.6	12.3	37.9
Relationship	283	13.2	7.4	62.9
Negative emotion	292	10.4	20.2	35.6
Loneliness	293	10.1	11.6	51.5
CEQ-5D				
Tariff	321	1.5	3.0	2.5
Mobility	325	0.3	2.2	45.8

TABLE 30 Baseline missing data and floor and ceiling effects for the EQ-5D, CEQ-5D, DEMQOL-U and DEMQOL-Proxy-U across the study period (*continued*)

Measure	*n*	% missing	% floor	% ceiling
Self-care	325	0.3	15.1	44.6
Usual activities	323	0.9	22.3	29.1
Pain/discomfort	324	0.6	11.1	41.7
Anxiety/depression	324	0.6	17.3	14.5
DEMQOL-Proxy-U				
Tariff	317	2.8	3.0	6.0
Positive emotion	322	1.2	55.9	2.5
Memory	320	1.8	21.6	43.1
Appearance	322	1.2	5.6	67.7
Negative emotion	322	1.2	18.6	18.0

Acceptability

DEMQOL-U displays higher missing data rates at baseline than EQ-5D both for the overall utility score and at the dimension level (see *Table 30*). The carer-reported measures display lower missing data rates than the patient-reported measures.

Agreement

Figure 10 presents Bland–Altman plots depicting the extent of the agreement between the patient measures (EQ-5D and DEMQOL-U) and between the carer measures (CEQ-5D and DEMQOL-Proxy-U) across the range of possible utility values. For the patient measures at baseline, 4.78% of the responses fall outside the 95% limits of agreement, and there is a small increase at follow-up. The same pattern is apparent for the CEQ-5D and DEMQOL-Proxy-U, for which 3.81% of responses fall outside the 95% limits of agreement at baseline. The charts indicate that the majority of the disagreement occurs at the more severe end of the scale, where a small number of respondents are providing a higher utility score for DEMQOL-U than for the EQ-5D. This may be related to the difference in the ranges of the utility scales, which means that respondents in poor health may generate a substantially lower utility score on EQ-5D than on DEMQOL-U. Agreement between the measures is increased at the higher end of the utility scale, and the mean difference decreases at follow-up. This indicates that measurement convergence between the generic and the condition-specific indicators is better among those in better health at baseline, and also as health status improves over the course of a trial. The mean difference for the CEQ-5D and DEMQOL-Proxy-U is larger, indicating more disagreement between the proxy-reported measures.

Table 31 presents respondent scores for DEMQOL-U and DEMQOL-Proxy-U when the patient is in full health according to the EQ-5D (state 11111 is equivalent to a utility score of 1). There is wide variation in the DEMQOL-U and DEMQOL-Proxy-U utility values when the same respondents report optimal health on the EQ-5D (or CEQ-5D). At baseline the DEMQOL-U scores that are found when the EQ-5D utility is 1 range from 0.63 to 0.96. The range increases at the 39-week follow-up (0.48–0.96). The same pattern is apparent for DEMQOL-Proxy-U, where the range at baseline is 0.60–0.81, increasing to 0.65–0.94 at 39 weeks' follow-up.

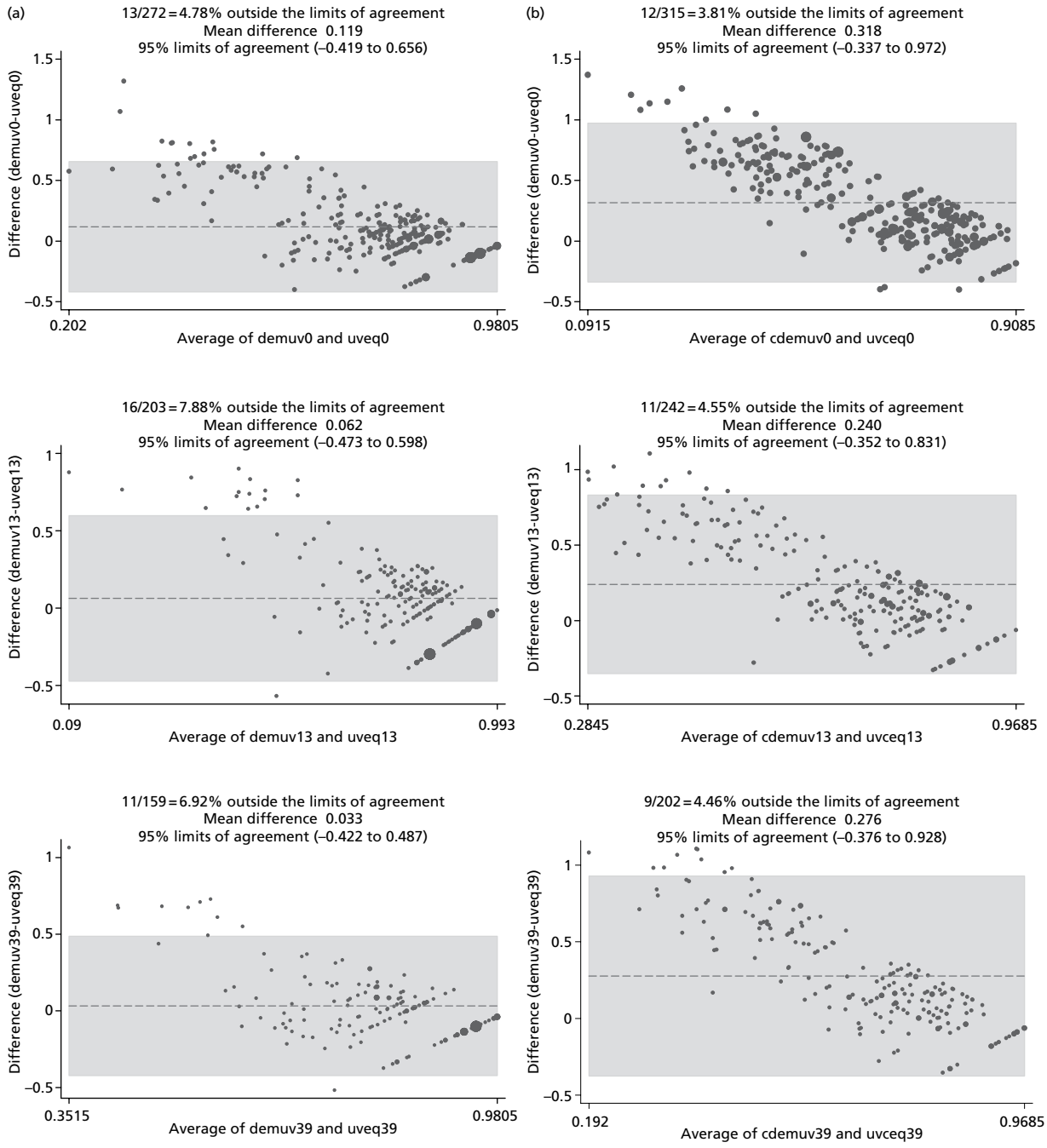


FIGURE 10 Agreement between DEMQOL-U and EQ-5D, and CEQ-5D and DEMQOL-Proxy-U utility values at baseline, week 13 and week 39. (a) DEMQOL-U vs EQ-5D; (b) DEMQOL-Proxy-U vs CEQ-5D.

Patient/proxy agreement

In light of the common reliance on carer reports for people with cognitive impairment, the agreement between carer- and patient-reported utility values was examined (*Table 32*). In terms of mean differences, DEMQOL-Proxy-U utilities differed very little from DEMQOL-U utilities. The difference between the EQ-5D and CEQ-5D scores was relatively larger (0.18–0.20). However, the ICC values indicated that there was better agreement between the EQ-5D and the CEQ-5D than between DEMQOL-U and DEMQOL-Proxy-U. This means that the EQ-5D may have better agreement between patients and carers. However, the patient- and carer-reported EQ-5D utilities did have greater variability than was observed for DEMQOL-U and DEMQOL-Proxy-U, and this may be related to the substantially larger utility range of EQ-5D.

TABLE 31 Distribution of DEMQOL-U AND DEMQOL-Proxy-U utility values when EQ-5D is in full health

Measure	*n*	Mean	SD	Median	Min.	Max.
EQ-5D = 11111						
DEMQOL-U week 0	43	0.85	0.09	0.87	0.63	0.96
DEMQOL-U week 13	57	0.82	0.10	0.84	0.61	0.99
DEMQOL-U week 39	45	0.85	0.10	0.89	0.48	0.96
CEQ-5D = 11111						
DEMQOL-Proxy-U week 0	8	0.74	0.07	0.76	0.60	0.81
DEMQOL-Proxy-U week 13	15	0.79	0.09	0.77	0.67	0.94
DEMQOL-Proxy-U week 39	18	0.83	0.10	0.86	0.65	0.94

TABLE 32 Agreement between utility values derived from self- and carer-rated HRQL data

Measure	*n*	Mean difference[a]	95% LoA[b]	ICC[c]	95% CI
DEMQOL-U and DEMQOL-Proxy-U					
Baseline	273	0.02	–0.24 to 0.28	0.15	0.04 to 0.26
Week 13	202	0.01	–0.01 to 0.03	0.25	0.12 to 0.38
Week 39	154	0	–0.26 to 0.26	0.17	0.01 to 0.32
EQ-5D and CEQ-5D					
Baseline	290	0.20	–0.46 to 0.85	0.32	0.12 to 0.47
Week 13	216	0.18	–0.42 to 0.79	0.36	0.14 to 0.53
Week 39	170	0.19	–0.43 to 0.81	0.23	0.04 to 0.39

CI, confidence interval; LoA, limits of agreement.
a Mean difference = self-rating – proxy rating.
b Bland–Altman plots.
c ICC model: two-way random absolute agreement.

Agreement between patients and carers on DEMQOL-U and DEMQOL-Proxy-U and on the EQ-5D and CEQ-5D was also examined using Bland–Altman plots (*Figure 11*). Although the mean difference between DEMQOL-U and DEMQOL-Proxy-U is less than that between the EQ-5D and the CEQ-5D, a higher percentage of responses are outside the 95% agreement limits at baseline (6.23% vs 5.86%). However, the magnitude of the difference is not large and this indicates that agreement is similar across the DEMQOL and EQ-5D measure pairs. The difference in the magnitude of agreement remains low across all time points, with the percentage outside the 95% agreement limits lower for the DEMQOL-U and DEMQOL-Proxy-U response pairs. The plots indicate that most of the difference occurs at the more severe end of the scale, where the mean utility score is lower. This means that patients are rating their utility as substantially better than their carer perceives it, or vice versa. The mean difference between ratings is higher for the EQ-5D, but this is explained by the larger range of utility values generated by this instrument.

Convergent validity

Table 33 reports the associations between the EQ-5D, DEMQOL-U and DEMQOL and the non-preference indicators of dementia-related health status. All three measures are significantly associated with depression scores (with a better EQ-5D, DEMQOL-U or DEMQOL score associated with lower levels of depression, as measured by CSDD, with DEMQOL-U displaying the highest correlation); however, the correlations

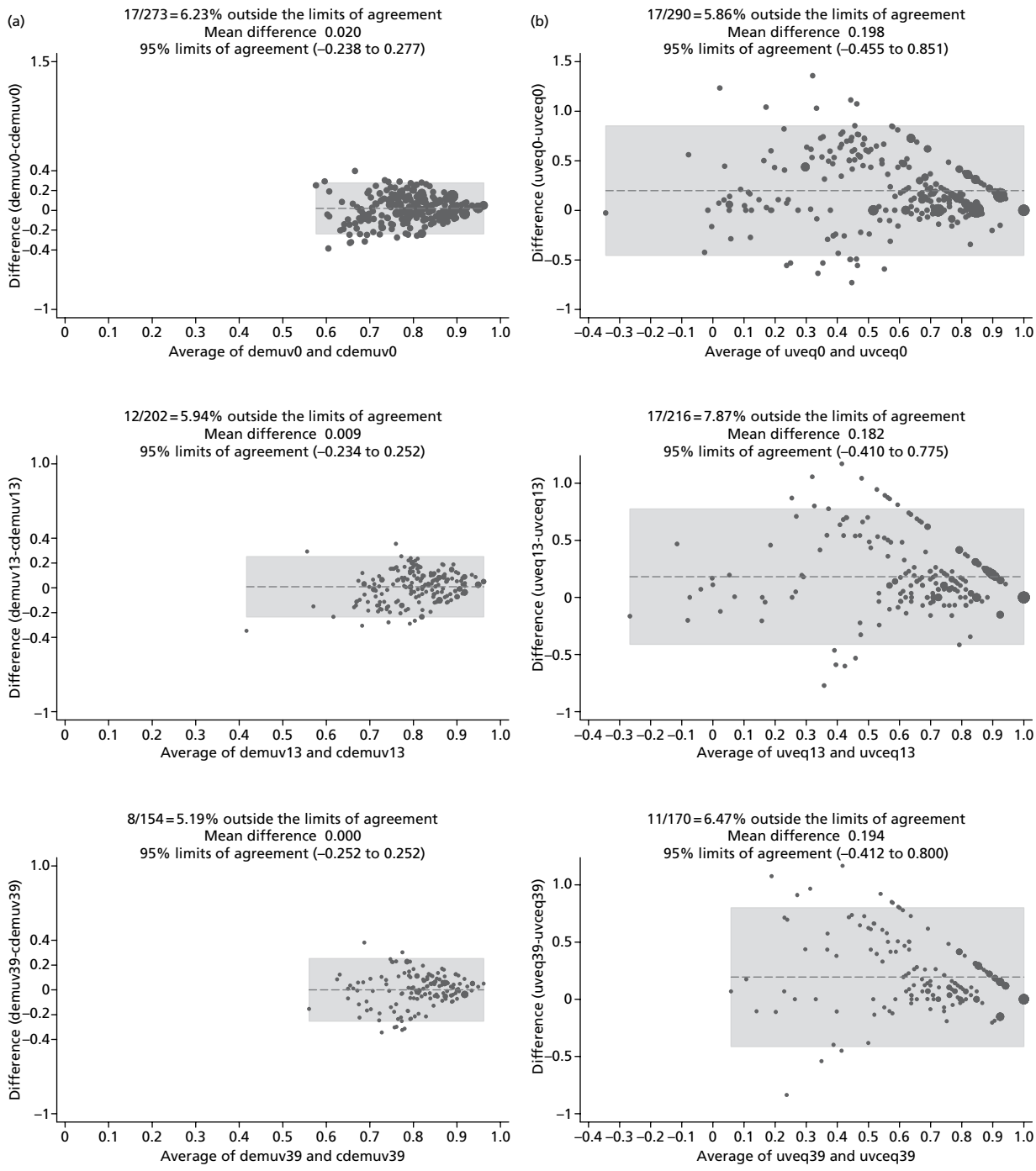


FIGURE 11 Agreement between self- and carer-based utility values at baseline, week 13 and week 39. (a) DEMQOL-U vs DEMQOL-Proxy-U; (b) EQ-5D vs CEQ-5D.

are in the range defined as low, indicating little convergence. There is no significant association between DEMQOL-U and EQ-5D and the external indicators of cognition (MMSE), daily activities (BADLS) and behavioural disturbances (NPI). This indicates that there is little or no association between the patient-report utility scores and the dementia-specific indicators.

The CEQ-5D was significantly associated with depression, cognition, daily activities and behavioural disturbances in the expected direction. The correlations with the BADLS and NPI were in the moderate

TABLE 33 Spearman correlation between HRQL utilities/scores and health status at baseline

Measure	*n*	EQ-5D	*n*	DEMQOL-U	*n*	DEMQOL
MMSE	240	0.01	235	0.03	223	–0.05
CSDD	213	–0.16[a]	198	–0.30[b]	187	–0.16[a]
BADLS	168	–0.08	159	0.03	150	0.08
NPI	293	–0.06	277	–0.07	260	0.02
Measure	***n***	**CEQ-5D**	***n***	**DEMQOL-Proxy-U**	***n***	**DEMQOL-Proxy**
MMSE	250	0.22[c]	247	–0.01	247	–0.01
CSDD	230	–0.17[d]	226	–0.14[a]	200	–0.08
BADLS	179	–0.50[b]	174	–0.05	157	–0.18[a]
NPI	325	–0.39[b]	317	–0.20[c]	279	–0.39[b]

a $p < 0.05$.
b $p < 0.0001$.
c $p < 0.001$.
d $p < 0.01$.

range, indicating some convergence between generic health status as measured by CEQ-5D and daily activities/behavioural disturbances. DEMQOL-Proxy-U was mildly associated with depression and behavioural disturbances and DEMQOL-Proxy with daily activities and behavioural disturbances. This indicates little association between DEMQOL-Proxy-U and the condition-specific measures.

Known-group validity

Table 34 indicates that EQ-5D and DEMQOL-U scores decrease slightly across cognition severity groups in the expected direction, but the effect sizes are in the range defined as small to moderate. The magnitude of the change on the EQ-5D and DEMQOL-U between the severity groups is similar, but the DEMQOL-U effect sizes are larger because of the bigger SD of the EQ-5D scores, which is related to the larger range in utility score of this instrument. Both measures have a small level of known-group validity across cognitive impairment groups as measured by the MMSE. The original DEMQOL measure does not discriminate as well as the utility measures as there are slight increases in HRQL as measured by DEMQOL as cognitive functioning decreases, but the magnitude of the effect is smaller than for the utility measures. The *n* of the severe impairment group is small and so the results need to be interpreted with caution. For depression the utility scores decrease as depression worsens. The effect size for the probable depression sample is small across the patient utility measures indicating a low level of discriminant validity, with the DEMQOL measure displaying a slightly larger effect size indicating increased discriminative ability. There is a large effect size for both the EQ-5D and DEMQOL-U for those with a definite case of depression; however, the *n* is small and so again this result needs to be interpreted with caution.

In terms of carer report, CEQ-5D scores decrease across cognition severity groups, with effect sizes in the moderate range. This is in contrast to DEMQOL-Proxy-U, for which scores remain relatively stable, indicating that the CEQ-5D displays a higher level of known-group validity. The original DEMQOL-Proxy scores display a low effect size, and the direction of change is not stable across the severity groups. In terms of depression severity, both CEQ-5D and DEMQOL-Proxy-U effect sizes are in the range defined as small for the probable group. The effect sizes are larger for the definite depression group, but again the *n* is small and so the results need to be interpreted with caution. Overall, DEMQOL-U displays a higher level of known-group validity than DEMQOL-Proxy-U. In contrast, CEQ-5D discriminates better between severity groups than EQ-5D.

TABLE 34 HRQL utilities and scores (mean and SD) by health status (MMSE, CSDD) at baseline

Measure	*n*	EQ-5D	Effect size	*n*	DEMQOL-U	Effect size	*n*	DEMQOL	Effect size
Cognitive impairment									
Mild (MMSE score > 20)	101	0.71 (0.26)	–	100	0.82 (0.09)	–	93	84.5 (14.0)	–
Moderate (MMSE score 20–10)	114	0.69 (0.27)	0.08	113	0.80 (0.11)	0.21	109	84.7 (13.5)	0.01[a]
Severe (MMSE score < 10)	25	0.67 (0.33)	0.13	22	0.79 (0.12)	0.29	21	85.5 (17.7)	0.06[a]
Measure	***n***	**CEQ-5D**	**Effect size**	***n***	**DEMQOL-Proxy-U**	**Effect size**	***n***	**DEMQOL-Proxy**	**Effect size**
Mild (MMSE score > 20)	101	0.57 (0.28)	–	99	0.79 (0.09)	–	89	87.6 (13.2)	–
Moderate (MMSE score 20–10)	118	0.47 (0.33)	0.33	118	0.78 (0.10)	0.11	104	85.8 (16.3)	0.12
Severe (MMSE score < 10)	30	0.43 (0.31)	0.47	30	0.79 (0.11)	0.00	26	90.0 (15.1)	0.17[a]
	n	**EQ-5D**	**Effect size**	***n***	**DEMQOL-U**	**Effect size**	***n***	**DEMQOL**	**Effect size**
Depression									
Non-case (CSDD score < 11)	163	0.72 (0.26)	–	153	0.81 (0.11)	–	147	85.8 (13.1)	–
Probable (CSDD score 11–17)	41	0.70 (0.29)	0.07	36	0.79 (0.09)	0.20	35	81.3 (17.5)	0.29
Definite (CSDD score > 17)	9	0.46 (0.35)	0.84	9	0.70 (0.14)	0.87	5	86.6 (20.3)	0.05[a]
	n	**CEQ-5D**	**Effect size**	***n***	**DEMQOL-Proxy-U**	**Effect size**	***n***	**DEMQOL-Proxy**	**Effect size**
Non-case (CSDD score < 11)	170	0.52 (0.30)	–	169	0.79 (0.09)	–	152	88.7 (13.3)	–
Probable (CSDD score 11–17)	46	0.46 (0.33)	0.19	47	0.78 (0.09)	0.11	39	88.7 (15.6)	0
Definite (CSDD score > 17)	11	0.16 (0.33)	1.14	10	0.74 (0.08)	0.59	9	79.0 (15.6)	0.60

a Difference from reference group not in hypothesised direction.

Effect size indices refer to the magnitude of the difference from the reference group (MMSE mild, CSDD non-case).

Responsiveness to change

Table 35 reports the observed change in HRQL utilities and scores for three groups of clinical progression using the EQ-5D, DEMQOL-U and DEMQOL: (1) MCID/clinical improvement, (2) no clinical change and (3) MCID/clinical deterioration. We expected HRQL utilities and scores to decline in general for the groups with MCID/clinical deterioration. At week 13 there is no evidence for responsiveness in the clinical deterioration groups for the EQ-5D and DEMQOL-U. DEMQOL responded to a decrease in depression but as the *n* is small this needs to be interpreted with caution. At week 39 the DEMQOL-U responds to deterioration in cognition, daily activities and behavioural disturbances with SRMs in the small range (*Table 36*). The SRM for depression is higher but again the *n* is small and so this must be interpreted carefully. In terms of

TABLE 35 Mean change in HRQL utilities and scores by clinical change in MMSE, BADLS, NPI and CSDD (baseline to week 13)

Measure	*n*	Δ EQ-5D, mean (SD)	SRM	*n*	Δ DEMQOL-U, mean (SD)	SRM	*n*	Δ DEMQOL, mean (SD)	SRM
MCID: MMSE									
Improved (+1.4)	45	0.07 (0.21)	0.35	41	0.02 (0.09)	0.24	36	5.42 (9.08)	0.60
No MCID Δ	63	0.09 (0.29)	0.32	61	0.02 (0.12)	0.17	58	2.88 (11.18)	0.26
Deteriorated	64	0.02 (0.21)	0.08	60	0.00 (0.12)	0.01	55	2.27 (12.78)	0.18
MCID: BADLS									
Improved (–3.5)	15	0.14 (0.40)	0.35	15	0.02 (0.11)	0.17	13	7.46 (8.91)	0.84
No MCID Δ	59	0.02 (0.22)	0.10	50	0.01 (0.10)	0.08	44	4.55 (11.80)	0.39
Deteriorated	30	0.08 (0.24)	0.32	26	0.02 (0.10)	0.19	25	2.76 (14.30)	0.19
MCID: NPI									
Improved (–8.0)	98	0.08 (0.28)	0.27	90	0.01 (0.13)	0.04	79	6.44 (11.68)	0.55
No MCID Δ	81	0.01 (0.25)	0.03	76	0.02 (0.10)	0.23	72	1.33 (10.70)	0.12
Deteriorated	34	0.06 (0.21)	0.30	29	0.02 (0.09)	0.26	28	2.79 (13.66)	0.20
Clinical Δ CSDD									
Improved (–30%)	120	0.07 (0.27)	0.26	111	0.02 (0.12)	0.16	102	5.57 (10.81)	0.52
No clinical Δ	55	–0.02 (0.27)	0.09[a]	50	0.02 (0.11)	0.16	45	1.82 (12.06)	0.15
Deteriorated	12	0.06 (0.22)	0.29	9	0.00 (0.08)	0.00	7	–9.14 (10.56)	0.87[a]

Measure	*n*	Δ CEQ-5D, mean (SD)	SRM	*n*	Δ DEMQOL-U-Proxy, mean (SD)	SRM	*n*	Δ DEMQOL-Proxy, mean (SD)	SRM
MCID: MMSE									
Improved (+1.4)	48	0.14 (0.29)	0.49	48	0.03 (0.08)	0.34	37	9.54 (11.37)	0.84
No MCID Δ	66	0.01 (0.29)	0.03	64	0.03 (0.11)	0.26	54	7.85 (14.62)	0.54
Deteriorated	68	0.08 (0.27)	0.31	66	0.02 (0.10)	0.25	53	7.36 (13.61)	0.54
MCID: BADLS									
Improved (–3.5)	16	0.25 (0.48)	0.52	14	0.04 (0.10)	0.41	12	12.83 (9.57)	1.34
No MCID Δ	63	0.08 (0.30)	0.28	59	0.01 (0.09)	0.14	50	5.98 (14.22)	0.42
Deteriorated	35	–0.04 (0.29)	0.13[a]	35	0.02 (0.10)	0.23	28	6.32 (14.14)	0.45
MCID: NPI									
Improved (–8.0)	114	0.13 (0.32)	0.41	112	0.03 (0.10)	0.35	90	11.31 (12.55)	0.90
No MCID Δ	96	0.06 (0.26)	0.23	92	0.02 (0.09)	0.16	73	3.40 (13.13)	0.26
Deteriorated	38	–0.03 (0.30)	0.10[a]	36	0.02 (0.09)	0.19	30	2.00 (14.85)	0.13
Clinical Δ CSDD									
Improved (–30%)	132	0.11 (0.30)	0.36	128	0.02 (0.10)	0.25	106	8.49 (13.53)	0.63
No clinical Δ	65	0.00 (0.25)	0.02	65	0.01 (0.08)	0.16	54	2.63 (13.55)	0.19
Deteriorated	11	–0.02 (0.40)	0.05[a]	10	0.03 (0.04)	0.84	7	5.43 (6.08)	0.89

a SRM value for decline in HRQL utility/score.

TABLE 36 Mean change in HRQL utilities and scores by clinical change in MMSE, BADLS, NPI and CSDD (baseline to week 39)

Measure	*n*	Δ EQ-5D, mean (SD)	SRM	*n*	Δ DEMQOL-U, mean (SD)	SRM	*n*	Δ DEMQOL, mean (SD)	SRM
MCID: MMSE									
Improved (+1.4)	23	0.02 (0.31)	0.05	23	0.06 (0.12)	0.51	22	5.27 (8.69)	0.61
No MCID Δ	49	0.07 (0.22)	0.30	46	0.00 (0.12)	0.01	42	4.81 (9.68)	0.50
Deteriorated	71	0.07 (0.24)	0.27	67	–0.02 (0.14)	0.14[a]	63	4.30 (10.64)	0.40
MCID: BADLS									
Improved (–3.5)	9	0.13 (0.41)	0.31	10	0.03 (0.05)	0.65	10	13.00 (8.89)	1.46
No MCID Δ	34	0.07 (0.20)	0.35	32	–0.01 (0.14)	0.07[a]	26	5.23 (10.08)	0.52
Deteriorated	39	0.03 (0.22)	0.13	31	–0.02 (0.19)	0.09[a]	28	6.29 (11.01)	0.57
MCID: NPI									
Improved (–8.0)	82	0.11 (0.23)	0.46	75	0.02 (0.13)	0.11	70	7.26 (11.45)	0.63
No MCID Δ	62	0.01 (,18)	0.06	58	0.02 (0.14)	0.15	54	5.59 (8.60)	0.65
Deteriorated	27	0.02 (0.37)	0.06	25	–0.04 (0.12)	0.30[a]	23	0.09 (12.49)	0.01
Clinical Δ CSDD									
Improved (–30%)	94	0.07 (0.22)	0.32	85	0.02 (0.14)	0.18	81	7.43 (10.48)	0.71
No clinical Δ	48	0.05 (0.28)	0.19	45	0.00 (0.14)	0.01[a]	40	0.90 (11.14)	0.08
Deteriorated	7	–0.12 (0.30)	0.41[a]	5	–0.07 (0.09)	0.85[a]	5	–4.00 (11.22)	0.36[a]
Measure	***n***	**Δ CEQ-5D, mean (SD)**	**SRM**	***n***	**Δ DEMQOL-U-Proxy, mean (SD)**	**SRM**	***n***	**Δ DEMQOL-Proxy, mean (SD)**	**SRM**
MCID: MMSE									
Improved (+1.4)	26	0.05 (0.43)	0.11	25	0.07 (0.07)	0.92	23	6.30 (8.92)	0.71
No MCID Δ	48	0.07 (0.27)	0.25	47	0.04 (0.11)	0.36	44	7.34 (17.73)	0.41
Deteriorated	73	0.05 (0.24)	0.21	68	0.02 (0.10)	0.24	54	7.65 (13.07)	0.58
MCID: BADLS									
Improved (–3.5)	10	0.35 (0.40)	0.87	11	0.09 (0.08)	1.15	7	15.86 (13.75)	1.15
No MCID Δ	38	0.08 (0.31)	0.27	36	0.02 (0.10)	0.19	32	7.22 (11.43)	0.63
Deteriorated	47	–0.07 (0.25)	0.27[a]	43	0.03 (0.10)	0.32	36	3.78 (16.05)	0.24
MCID: NPI									
Improved (–8.0)	108	0.09 (0.33)	0.28	105	0.03 (0.11)	0.31	84	12.08 (14.43)	0.84
No MCID Δ	66	–0.01 (0.24)	0.03[a]	67	0.01 (0.11)	0.06	59	4.36 (14.14)	0.31
Deteriorated	33	–0.01 (0.40)	0.02[a]	29	0.08 (0.12)	0.67	23	2.39 (13.07)	0.18
Clinical Δ CSDD									
Improved (–30%)	107	0.09 (0.27)	0.33	107	0.03 (0.11)	0.25	91	9.23 (13.22)	0.70
No clinical Δ	56	–0.03 (0.34)	0.10[a]	53	0.02 (0.09)	0.23	45	7.18 (14.02)	0.51
Deteriorated	11	–0.08 (0.35)	0.23[a]	11	0.08 (0.11)	0.72	10	–2.10 (17.89)	0.12[a]

a SRM value for decline in HRQL utility/score.

sensitivity to improvement, both the EQ-5D and DEMQOL-U display evidence of responsiveness, but at a lower level than the original DEMQOL measure.

For proxy report, the CEQ-5D responds to deterioration in daily activities, behavioural disturbances and depression in the small effect size range, but DEMQOL-Proxy-U does not. This effect is maintained at 39 weeks when, again, the DEMQOL-Proxy-U does not respond to health deterioration. Both measures respond to health improvements, with SRMs in the small to moderate range. The original DEMQOL-Proxy proved to be more responsive than CEQ-5D and DEMQOL-Proxy-U.

Discussion

In this chapter we have documented initial evidence regarding the acceptability, validity, patient/carer agreement and responsiveness of the DEMQOL-U and DEMQOL-Proxy-U dementia-specific preference-based HRQL measurement system (DEMQOL-U and DEMQOL-Proxy-U) and the EQ-5D generic preference-based measure in comparison with external indicators of dementia-related health status and the original DEMQOL and DEMQOL-Proxy. We also assessed agreement between the preference-based instruments and found that overall there is a good level of agreement and that the majority of the disagreement between the measures occurs at the more severe end of the utility scale. There is some evidence for the acceptability of the DEMQOL system, in particular the DEMQOL-Proxy-U, which displays low missing data rates. However, missing data rates are lower for the EQ-5D. There is less evidence regarding the convergent validity of DEMQOL-U and DEMQOL-Proxy-U in comparison with other measures of cognition, depression and activity level in dementia. DEMQOL-U displays a higher level of known-group validity between cognition and depression severity groups than DEMQOL-Proxy-U, but the effect sizes are in the low range and there are differences from the discriminative ability of the patient- and carer-reported EQ-5D. There is no clear pattern regarding agreement between patients and carers. In terms of responsiveness, there is evidence that the DEMQOL utility measures and EQ-5D are less sensitive to change than the original DEMQOL and DEMQOL-Proxy.

The patient- and carer-reported EQ-5D have lower missing data rates than DEMQOL-U and DEMQOL-Proxy-U respectively. This may indicate that the EQ-5D is more acceptable to respondents. However, the differences in missing data rates are not large and may be expected because of the size of the instruments (EQ-5D includes five items whereas DEMQOL and DEMQOL-Proxy, which are used to generate the utility values, include 29 and 32 items respectively). Therefore, it is more likely that data will be missing from the DEMQOL utility measures.

DEMQOL-U displays a lower level of agreement with the patient-reported EQ-5D than DEMQOL-Proxy-U does with the CEQ-5D, and this may be because carers are more likely to report stable responses than people with dementia in terms of the general level of severity reported. However, for both pairs the majority of the scores fall within the 95% agreement range, indicating a good level of agreement. The majority of the disagreement occurs when the mean utility value is at the more severe end of the scale (where DEMQOL utility measure scores are substantially higher than EQ-5D/CEQ-5D scores). This may be due to differences in the classification systems, which may be sensitive to different HRQL issues, and also to large differences in the possible utility scale range.

Exploring the distribution of the utility values of each measure revealed that DEMQOL-U and DEMQOL-Proxy-U utility values tend to be higher than EQ-5D values for poorer health states. This may be expected as the lowest possible EQ-5D utility score is –0.594 whereas the lowest possible utility scores for DEMQOL-U and DEMQOL-Proxy-U are 0.243 and 0.363 respectively. Similarly, EQ-5D utility values tend to be higher than DEMQOL-U and DEMQOL-Proxy-U values for better health states. The EQ-5D had high ceiling effects and therefore may not be sensitive to mild HRQL impairment in patients with better health states. This is also consistent with the considerable variation observed in DEMQOL-U and DEMQOL-Proxy-U utility values, which suggests that patients with full health on EQ-5D may be experiencing varying levels

of impaired HRQL. A likely reason for the ceiling effects observed in the EQ-5D is that there are three levels of responses for each dimension, which may result in insensitivity for detecting small differences in health states.[94]

As carer report may be a necessary substitute for self-report among people with dementia, a utility measure that allows for good agreement between the carer and the patient perspective would be preferred. Results regarding the agreement between patients and carers across the DEMQOL utility measures and patient- and carer-reported EQ-5D are mixed. There is a higher level of patient/carer agreement between the EQ-5D and the CEQ-5D at each time point, and more agreement between the ratings at baseline; however, DEMQOL-U and DEMQOL-Proxy-U ratings display more agreement at follow-up. However, overall, the agreement as measured using ICC values is low for both measures. As the dimensions included in the EQ-5D are matched across the patient and carer versions, it might be expected that the EQ-5D displays a higher level of agreement between ratings than the DEMQOL measures, for which the dimensions and questions used vary. However, it is well established that dementia patients and proxies give different reports,[14,68] and therefore agreement level cannot be seen as a clear indicator of the performance of the measures. Further work in this area could investigate the pattern of agreement further, and this could be done by investigating agreement at different dementia severity levels, both for the DEMQOL utility measures and the EQ-5D.

We psychometrically assessed the convergent and known-group validity of DEMQOL-U and DEMQOL-Proxy-U. DEMQOL-U displayed a moderate correlation with the CSDD (at a higher level than the EQ-5D and DEMQOL), suggesting some convergence in terms of measurement of depression related to dementia. However, correlations between DEMQOL-U and the EQ-5D and the indicators of cognition, daily activities and behavioural disturbances are low. The correlations between the CEQ-5D and the external condition-specific indicators (which are low to moderate) are higher than the DEMQOL-Proxy-U correlations (which are low).

There is no clear pattern to the convergent validity results, which makes interpretation difficult. It could be argued that the lack of convergence for the patient measures is due to differences in the focus of the classification systems in terms of both the dimensions included and the dimensions omitted. However, the EQ-5D has dimensions directly assessing depression and daily activities and DEMQOL-U includes a cognition dimension and so better correlations may be expected. Further analysis should investigate relationships between the measures at the dimension level and also investigate convergence using a range of other indicators. The CEQ-5D correlations are higher, particularly with the other measures completed by carers, and this may mean that patient report in dementia is unreliable and unstable across assessments.

There is also no clear pattern to the known-group validity of the DEMQOL utility measures or the EQ-5D across cognitive impairment and depression severity groups. Overall, DEMQOL-U performs better than DEMQOL-Proxy-U, but the CEQ-5D performs better than the EQ-5D. This may be because the DEMQOL-U descriptive system generates more health states than DEMQOL-Proxy-U and so therefore has more discriminative ability. The better performance of the CEQ-5D may be due to the more stable nature of proxy report in dementia.

For the patient utility measures across the cognitive impairment and depression severity groups (excluding the definite depression group for which the sample size is low), the effect sizes are in the small to moderate range. The magnitude of the change is similar, but the DEMQOL-U effect sizes are larger because the SD of the EQ-5D scores is larger because of differences in utility score range. This means that, in trials, the DEMQOL utility values can be estimated with more precision. This has also been found for other condition-specific preference-based measures in comparison with the EQ-5D.[37]

In terms of carer report, the CEQ-5D is more sensitive to differences across cognition severity groups than DEMQOL-Proxy-U. However, both measures display effect sizes in a similar range for depression severity. Again, there is no clear pattern and further research could investigate the discriminative ability of the DEMQOL utility measures using a range of further indicators. Furthermore, more work needs to be carried out on the discriminative ability of the measures among those with severe problems, as the sample size used in this study is too small to draw any conclusions.

Although the DEMQOL utility instruments and the EQ-5D respond to improvements in health status over time, this is not at the same level as with the original DEMQOL and DEMQOL-Proxy. DEMQOL-U was sensitive to deterioration at a higher level than DEMQOL-Proxy-U. The reasons for these differences in results are not clear and more research is required to assess responsiveness in more detail using a range of indicators and different data sets.

For the psychometric analysis reported above we used the utility values derived from the general population, and this allows for a level of comparability with the EQ-5D value set, which was also derived from the general population. However, the difference between the general population and patient/carer valuations of DEMQOL-U and DEMQOL-Proxy-U health states reported in *Chapter 6* suggests that a full valuation study with patients and carers would produce a substantially different utility scale. If a full valuation study was to be carried out it would be important to investigate the psychometric performance of the patient- and carer-derived utility scales in comparison with both the general population values and external dementia-related indicators. It was not possible to compare the valuations produced in *Chapter 5* as only mean TTO values were produced.

It is important to note that the psychometric performance of the DEMQOL utility measures may be impacted by the HTA-SADD sample used (in which the focus was on depression). DEMQOL-U and DEMQOL-Proxy-U are designed specifically for dementia and therefore further testing on other dementia samples is recommended. Furthermore, this analysis has focused on the psychometric properties of each measure and has not combined the utility values with duration to produce QALY estimates. Further comparative research using QALY estimates is also recommended.

Conclusions

Health outcomes in economic evaluation are often measured using a composite measure of both quality and length of life – the QALY. This composite measure can be used in health policy decision-making to compare the efficiency of different treatment strategies. For utilities (such as those derived here) to be of any value for decision-makers, they must be included in a QALY measure that also considers length of life. The analysis described above presents some early evidence of the level of validity, patient/carer agreement level and responsiveness of the DEMQOL utility measures in comparison with the widely used generic preference-based measure EQ-5D. The results suggest that both the EQ-5D and the DEMQOL utility instruments have advantages and disadvantages over each over, but the pattern is unclear. Further research investigating the complex psychometric performance of the measures is required using different data sets incorporating a range of clinical indicators and dementia severity levels. We originally planned to carry out the analysis described above on a second data set but these data did not become available during the study period. Until further results are available we would recommend using both the EQ-5D and the DEMQOL utility measures alongside each other in dementia studies to allow for both generic and condition-specific utilities to be generated.

Chapter 8 Conclusions

This report has detailed the development and application of two dementia-specific preference-based measures, one for self-completion (DEMQOL-U) and the other to be completed by carers (DEMQOL-Proxy-U). These measures can be used to generate health-state utility values on the QALY scale for use in economic evaluation of interventions in this group of patients. These are the first condition-specific preference-based measures in dementia. The results of the psychometric analysis presented in this report are encouraging but there are a number of concerns regarding the validity and responsiveness of the instruments that require further investigation. Therefore, before more evidence is available we would recommend that the DEMQOL instruments are used alongside a generic measure such as the EQ-5D in future evaluations of interventions for dementia.

A detailed discussion of each part of the study is provided in each chapter, which covers specific concerns over the methods used and the main findings and their implications. In this final chapter we review and summarise our findings and check these against the study aims to determine the extent to which these aims have been met. This chapter also provides an overview of the implications of the work for economic evaluation and for future research.

The following sections show that we have been able to deliver on all four major aims.

To derive health-state classification systems from DEMQOL and DEMQOL-Proxy that can be used to categorise all patients with responses to the measures

This project has used an established multistage process to guide the derivation of preference-based measures from the DEMQOL instruments. The first stage was to identify the dimensions to be used in the health-state classification using factor analysis as a guide to the dimensionality. We chose to use exploratory analysis because of the inconclusive nature of the earlier factor analysis of the DEMQOL system (in comparison with the original conceptual framework), which means that we did not have an a priori structure to confirm. However, the factor structures established here provide some support for the original conceptual framework. The factor structures are robust enough to provide the basis for the development of dementia-specific health-state classification systems: a five-dimensional structure was found for both DEMQOL and DEMQOL-Proxy.

We then used Rasch analyses to select an item to represent each factor. Using stringent criteria to investigate item performance we were able to select one item for each of the five DEMQOL domains and one item for four of the DEMQOL-Proxy domains (one domain was dropped as none of the items performed at an acceptable level in the Rasch analysis). Each dimension has four-level responses corresponding to the response choices for the items in the original DEMQOL instruments. These new instruments were named DEMQOL-U (which produces a possible 1024 health states) and DEMQOL-Proxy-U (which produces 256 possible states). These health-state classifications generate health states that are amenable to valuation and the subsequent generation of dementia-specific QALYs.

To generate utility values for every health state defined by the health-state classification systems derived from DEMQOL and DEMQOL-Proxy

Interviews were successfully conducted with a representative sample of 593 members of the general population who were able to complete the rank and TTO tasks and provide valuations of a sample of

health states generated by DEMQOL-U and DEMQOL-Proxy-U. The TTO valuations were modelled to estimate two preference-based algorithms, one each for DEMQOL-U and DEMQOL-Proxy-U. This means that utility values can be estimated for all 1024 DEMQOL-U health states and all 256 DEMQOL-Proxy-U health states. We applied a range of models to the TTO and rank data and selected the TTO model that fitted best overall, with the highest number of significant variables and best-fit statistics. The models fitted to the data performed favourably compared with similar models estimated for other preference-based measures in terms of fit, prediction and the consistency of the coefficients with the descriptive system. The DEMQOL-U model did not have any inconsistent variables (in which a decrease in health status leads to an increase in utility) and DEMQOL-Proxy-U had just one. The modelling process means that it is now possible to generate health-state utility values from any data set in which one or both of DEMQOL and DEMQOL-Proxy have been completed.

To examine whether or not utility values elicited from the general population differ from utility values elicited from patients and carers for dementia health states generated by the classification systems

We interviewed a sample of 71 people with mild dementia and 71 family carers. They were able to complete the TTO task and assign values to a limited number of the health states generated from the new instruments. This is the first time that the values of dementia health states from the general population have been compared with patient and carer values in this way. The main finding was that the values obtained from people with dementia and carers were systematically different from those obtained from the general public. The general public tended to undervalue the impact of dementia compared with people with dementia and carers. This finding is in contrast to the general finding in the literature – that patients tend to give higher values than the general public, at least for physical health conditions.[79] Patient values are thought to be higher because of processes related to adaptating to the condition. In mental health the opposite has been observed, and this may reflect an inability of the general population to understand the consequences for quality of life of problems associated with mental disorders compared with physical health.[82] This finding may reflect a similar inability to understand the consequences of the syndrome of dementia, even though they are described in everyday terms.

It is not possible to model the pattern of the difference between patients and carers and the general population. Too few states were valued to estimate specific algorithms for patients and carers, as the aim of this part of the project was only to test the hypothesis that there are differences. However, it was also observed that there were differences in the ordering of states, which suggests that there may well be differences in the relative weight being given to different dimensions. This requires further investigation.

These results suggest that the population used to produce dementia health-state utility values could impact on the results of cost-effectiveness analysis and potentially affect resource allocation decisions.

To examine the psychometric performance of the dementia-specific preference-based measures using trial data

Finally, we tested the new measures using a data set from a recently completed trial. In comparing the psychometric performance of DEMQOL-U and DEMQOL-U-Proxy with that of their non-preference-based as well as generic preference-based counterparts, this report documents the first early evidence of validity and responsiveness of a condition-specific preference-based measure for people with dementia. We also assessed the utility measures in comparison with clinically based measures. However, these measures may be indirectly related to many of the dimensions covered by the DEMQOL utility instruments and the EQ-5D and further research should test the measures using a range of clinical indicators.

The DEMQOL utility instruments and the EQ-5D have advantages and disadvantages over each other in terms of psychometric performance; however, the pattern is unclear. There is evidence for the acceptability of the DEMQOL system but missing data rates are higher than for EQ-5D. EQ-5D displays high ceiling effects and therefore may not be as sensitive as the DEMQOL utility measures to mild HRQL impairment in patients with better health states.

There is a good level of agreement between the preference-based instruments. The majority of the disagreement between the measures occurs at the more severe end of the utility scale. This may be linked to differences in the dimensions described by the classification system, which may be sensitive to different HRQL issues, and also to large differences in the possible utility scale range. In terms of agreement between patients and carers we found that carer report was more stable. In dementia, carer report is often essential and therefore a utility measure that allows for good agreement between the carer and the patient perspective is important.

There is no clear pattern to the convergent validity results. Correlations between DEMQOL-U and the EQ-5D and the external indicators are low. However, DEMQOL-U displays some convergence in terms of measurement of depression related to dementia. The correlations between the CEQ-5D and the external indicators are higher than the DEMQOL-Proxy-U correlations. It could be argued that the lack of convergence for the patient measures is due to differences in the focus of the classification systems in comparison with the external indicators. Assessment against a wider range of indicators is required.

DEMQOL-U discriminates better than DEMQOL-Proxy-U but the CEQ-5D discriminates better than the EQ-5D. This may be linked to classification system differences in that DEMQOL-U generates more health states than DEMQOL-Proxy-U and so has increased discriminative ability. The better performance of the CEQ-5D may be linked to the more stable nature of proxy report in dementia across matched classification systems.

In terms of responsiveness, there is evidence that the DEMQOL utility measures and the EQ-5D are less sensitive to change than the original DEMQOL and DEMQOL-Proxy. The reasons for these differences are not clear and responsiveness needs to be assessed in more detail using a range of indicators and different dementia-specific data sets.

The psychometric performance of the DEMQOL utility measures may be impacted by the sample used, which focused on depression. The inconclusive nature of the results means that further testing on a range of samples is required.

Limitations of the study

There are a number of limitations in the development process and testing of the DEMQOL utility measures that currently limit the use of the instruments.

The first limitation relates to the lack of an external data set to validate the DEMQOL-U and DEMQOL-Proxy-U classification systems. This is an important step in the process of deriving a condition-specific preference based measure from an existing measure but is reliant on the data available to developers. In this study, no external DEMQOL or DEMQOL-Proxy data were available to validate the system, and the HTA-SADD data were available only towards the end of the study when the psychometric validation stage of the project had been completed and the valuation phase had commenced. It is also possible to use a split-half approach to validation in which a subsample of the development data are used. However, the optimum sample size for Rasch analysis is 500 and splitting the sample used in this study into two would result in data sets that are too small for the task. This problem is made worse for those factors with a small number of items, as the Rasch programme excludes extreme scores within the dimension (i.e. respondents at the floor or ceiling), resulting in a smaller sample size.

Throughout the development process we have made decisions about which methodology to use. There are other valid techniques that could have been used and it is possible to criticise the choices that we have made. First, it is possible to criticise the use of EFA to investigate the dimensional structure of the instruments, and this study may be limited by not investigating a wider range of possible methods for deriving a dimensional structure. We used EFA as the earlier factor analysis of the DEMQOL system was inconclusive and so we were not confirming an existing structure. However, EFA may not be as robust as CFA and it may have benefited the development process to also study some CFA models.

The health-state classification systems that we have derived are a result of the original measure and the sample and analysis techniques used. It is possible that the use of a mild to moderate dementia sample has resulted in classification systems describing a comparatively mild conceptualisation of dementia. This is supported by the results of the valuation study. However, we used the mild to moderate sample as DEMQOL has been validated for use in this group, and it was decided to use a matched sample in terms of severity for DEMQOL-Proxy. It is also possible that the resultant classification systems are missing important dimensions, in particular daily activity limitations (this did not appear as a dimension in DEMQOL-U and was excluded from DEMQOL-Proxy-U because of poorly performing items). However, it should also be noted that the lack of association between dementia-related HRQL and activity limitations has been found elsewhere.[20]

Another concern might be the absence of the direct involvement of patients and their carers in the development of the health-state classifications. The process benefited from the results of in-depth interviews with patients and their carers, a detailed psychometric analysis of patient- and carer-reported outcomes and the expertise of clinicians working with patients on a daily basis. However, the process of selection might have been further improved with more direct patient involvement in the decisions and this is something to consider in any further qualitative work that might be undertaken with the measure (see *Recommendations for future research*).

In the original proposal it was suggested that health-state utility values could also be derived using ranking and a further DCE study. The field has developed since the time that this study was designed and the proposed methods of estimating values on the 0–1 scale required for calculating QALYs have been superseded.[95] Although we have conducted the analysis of the rank data collected in the valuation study, it was decided not to send respondents an additional postal survey given that the results would not be used to inform policy as the TTO results provide greater comparability with the EQ-5D and other measures.

Conclusions for evaluation in dementia

We would recommend that those designing trials and studies on dementia consider the use of DEMQOL-U and DEMQOL-Proxy-U. However, to clarify the nature and extent of the advantages and disadvantages of the new preference-based measures developed here we need further research into the psychometric performance and acceptability of the instruments. Ongoing and future trials and studies using the DEMQOL system will allow for further analyses to be carried out comparing the output of the generic measures with the general population utility values derived from DEMQOL-U and DEMQOL-Proxy-U.

More generally there are important issues surrounding whether or not QALYs estimated from condition-specific preference-based measures can be used to inform resource allocation across programmes. This issue has been examined in detail in a recent HTA report reviewing the development of preference-based measures from existing condition-specific measures.[37] Policy-makers must weigh the potential gains in terms of greater relevance and sensitivity to the condition from using a particular condition-specific preference-based measure with the reduction in comparability. Condition-specific preference-based measures have a number of potential disadvantages including missing side effects of treatment and the impact of comorbidities (where their effect is not additive). There are also concerns that the values for the

condition-specific states are subject to focusing effects from respondents failing to take into account their overall health.

Currently there is not enough evidence to be able to determine the likely benefits of using the DEMQOL-U and DEMQOL-Proxy-U measures compared with using the EQ-5D or other generic measures. At this stage it is not possible to recommend that they are used in place of the EQ-5D. We recommend that they are used together with the EQ-5D (or other generic measures depending on jurisdiction) in future studies to ensure that the needs of policy-makers can be met. This will also provide further evidence for testing these new instruments and examining their impact on the cost-effectiveness of interventions.

Recommendations for future research

1. The selection of dimensions and items in DEMQOL-U and DEMQOL-PROXY-U from the original instruments needs to be validated in another data set. The lack of validation is a limitation of the DEMQOL utility measures and we would aim to carry out this process when relevant data become available. This would complete stage IV of the six-step process and provide more evidence regarding the stability and robustness of the classification systems.
2. Further evaluation of DEMQOL-U and DEMQOL-Proxy-U compared with generic preference-based measures such as the EQ-5D, other relevant dementia-specific indicators and the original DEMQOL system should be carried out. The initial psychometric validation reported here is inconclusive, with evidence both for and against the validity and responsiveness of the DEMQOL utility measures, and agreement between preference-based measures and self- and carer report. Therefore, there is a need to expand the analyses completed for the HTA-SADD trial to other trials and evaluations that have used the DEMQOL system using a range of other relevant indicators. This will provide further evidence regarding the psychometric performance of the utility measures, and the samples in which the measures perform best. Further psychometric analysis will allow policy-makers to understand the implications of using these new dementia-specific preference-based measures to estimate QALYs compared with EQ-5D in cost-effectiveness analyses of new interventions.
3. We have shown that it is feasible to obtain health-state values from people with dementia and carers and that the valuations differ from those provided by the general population. A future study is needed to value a sufficient number of health states to estimate patient and carer preference weights for all states defined by the classification system so that economic evaluations can be conducted using both patient and carer and general population utility weights. This will inform an important policy debate about the consequences of using patient or carer values rather than those of the general population.
4. When there are data sets that share the use of the EQ-5D and DEMQOL and the proxy versions, these should be pooled to allow for further analysis of the system. Studies are ongoing that include the DEMQOL system and the EQ-5D and these data may become available to the authors for further analysis.
5. One potential area for future research is to investigate the acceptability and validity of the classification systems qualitatively with dementia patients and their carers.

Acknowledgements

We would like to thank the participants in the north of England and south London who agreed to participate in the valuation studies included in this programme of research. This report presents independent research commissioned by the NIHR. This project was funded by the NIHR Health Technology Assessment programme (project number 07/73/01). See the HTA programme website for further project information. The views and opinions expressed therein are those of the authors and do not necessarily reflect those of the NHS, the NIHR, the NIHR Evaluation, Trials and Studies Coordinating Centre (NETSCC), the HTA programme or the Department of Health.

Contribution of authors

Sube Banerjee (Professor of Mental Health and Ageing, King's College London, Institute of Psychiatry) was the principal investigator for this project overall. He was the principal investigator for the development of DEMQOL and initiated the proposal that led to this work. He led the team at the Institute of Psychiatry, participating in the design and conduct of all stages of the research and report preparation.

John Brazier (Professor of Health Economics, Health Economics and Decision Science, ScHARR, University of Sheffield) was the co-principal investigator of the study, leading the economics team at the University of Sheffield. He participated in the design and conduct of all stages of the research and report preparation.

Brendan Mulhern (Research Associate, Health Economics and Decision Science, ScHARR, University of Sheffield) acted as study co-ordinator, participated in the design and conduct of all stages of the research and report preparation other than the grant preparation and co-ordinated the input of the Sheffield elements of the research. He completed the psychometric and economic work informing the factor structure and item selection.

Donna Rowen (Research Fellow, Health Economics and Decision Science, ScHARR, University of Sheffield) participated in the design and conduct of all stages of the research and report preparation other than the grant preparation. She completed the valuation elements of the study.

Sarah Smith (Research Fellow, Department of Health Services Research and Policy, London School of Hygiene and Tropical Medicine) completed the psychometric work informing the factor structure and item selection. She participated in the design and conduct of all stages of the research and report preparation.

Renee Romeo (Lecturer, Centre for the Economics of Mental Health, Institute of Psychiatry, King's College London) completed the psychometric and economic evaluation of the system using the HTA-SADD data set. She participated in the design and conduct of all stages of the research and report preparation.

Rhian Tait (Research Worker, Institute of Psychiatry, King's College London) recruited and interviewed the south London valuation sample. She participated in all stages of report preparation.

Caroline Watchurst (Research Worker, Institute of Psychiatry, King's College London) recruited and interviewed the south London valuation sample. She participated in all stages of report preparation.

Kia-Chong Chua (PhD student, Institute of Psychiatry, King's College London) assisted RR in the analyses of the HTA-SADD data set. He participated in all stages of report preparation.

Vanessa Loftus (Specialist Registrar, South London and Maudsley NHS Foundation Trust) helped co-ordinate the King's College London team and facilitated recruitment from clinical settings. She participated in all stages of the study conduct and report preparation.

Tracey Young (Senior Research Fellow, Health Economics and Decision Science, ScHARR, University of Sheffield) participated in the design and conduct of all stages of the research and report preparation.

Donna Lamping (Professor of Psychology, Department of Health Services Research and Policy, London School of Hygiene and Tropical Medicine) was the presiding genius behind the development of the DEMQOL system. She participated in the design and conduct of all stages of the research up to her very sad and untimely death. This report is dedicated to her.

Martin Knapp (Professor of Health Economics, Centre for the Economics of Mental Health, Institute of Psychiatry, King's College London) participated in the design and conduct of all stages of the research and report preparation.

Robert Howard (Professor of Old Age Psychiatry and Psychopathology, Department of Old Age Psychiatry, Institute of Psychiatry, King's College London) participated in the design and conduct of all stages of the research and report preparation.

SB and JB are the guarantors for the work reported here.

Publications

1. Mulhern B, Smith S, Rowen D, Brazier J, Knapp M, Lamping D, *et al.* Improving the measurement of QALYs in dementia: developing patient- and carer-health-state classification systems using Rasch analysis. *Value Health* 2012;**15**:323–33.
2. Rowen D, Mulhern B, Banerjee S, van Hout B, Young T, Knapp M, *et al.* Estimating preference based single index measures for dementia using DEMQOL and DEMQOL-Proxy. *Value Health* 2012;**15**:346–56.

References

1. Hofman A, Rocca WA, Brayne C, Breteler MM, Clarke M, Cooper B, *et al.* The prevalence of dementia in Europe. *Int J Epidemiol* 1991;**20**:736–48.
2. Launer LJ, Brayne C, Dartigues J-F, Hofman A. Epilogue. *Neuroepidemiology* 1992;**11**:119–21.
3. Knapp M, Prince, M, Albanese E, Banerjee S, Dhanasiri S, Fernandez J-L, *et al. Dementia UK: the full report*. London: Alzheimer's Society; 2007.
4. Lowin A, Knapp M, McCrone P. Alzheimer's disease in the UK: comparative evidence on cost of illness and volume of research funding. *Int J Geriatr Psychiatry* 2001;**16**:1143–8.
5. Murray J, Schneider J, Banerjee S, Mann A. EUROCARE: a cross-national study of co-resident spouse carers for people with Alzheimer's dementia. II: a qualitative analysis of the experience of caregiving. *Int J Geriatr Psych* 1999;**14**:665–61.
6. Schneider J, Murray J, Banerjee S, Mann A. EUROCARE: a cross national study of co-resident spouse carers for people with Alzheimer's disease. I: factors associated with carer burden. *Int J Geriatr Psych* 1999;**14**:651–61.
7. Prince M, Jackson J. *World Alzheimer's report 2009*. London: Alzheimer's Disease International; 2009.
8. Wimo A, Prince M. *World Alzheimer's Report 2010: the global economic impact of dementia*. London: Alzheimer's Disease International; 2010.
9. Department of Health. *National service framework for older people.* London: Department of Health; 2001.
10. Department of Health. *Everybody's business*. London: Care Services Improvement Partnership; 2005.
11. Department of Health. *Living well with dementia, a national dementia strategy*. London: The Stationery Office; 2008.
12. National Audit Office. *Improving services and support for people with dementia*. Report by the Comptroller and Auditor General, HC 604, Session 2006–2007. London: The Stationery Office; 2007.
13. Whitehouse PJ. Harmonization of dementia drug guidelines: a report of the International Working Group for the Harmonization for Dementia Drug Guidelines. *Alzheimer Dis Assoc Disord* 2000;**14**:S119–22.
14. Smith SC, Lamping DL, Banerjee S, Harwood R, Foley B, Smith P, *et al.* Measurement of health-related quality of life for people with dementia: development of a new instrument (DEMQOL) and an evaluation of current methodology. *Health Technol Assess* 2005;**9**(10).
15. Brod M, Stewart AL, Sands L, Walton P. Conceptualization and measurement of quality of life in dementia. *Gerontologist* 1999;**39**:25–35.
16. Logsdon RG, Gibbons LE, McCurry SM, Teri L. Assessing quality of life in older adults with cognitive impairments. *Psychosomat Med* 2002;**64**:510–19.
17. Smith SC, Lamping DL, Banerjee S, Harwood RH, Foley B, Smith P, *et al.* Development of a new measure of health-related quality of life for people with dementia: DEMQOL. *Psychol Med* 2007;**37**:737–46.
18. Banerjee S, Smith SC, Lamping DL, Harwood RH, Foley B, Smith P, *et al.* Quality of life in dementia: more than just cognition. *J Neurol Neurosurg Psychiatry* 2006;**77**:146–8.

19. Woods RT, Thorgrimsen L, Spector A, Royan L, Orrell M. Improved quality of life and cognitive stimulation in dementia. *Aging Ment Health* 2006;**10**:219–26.

20. Banerjee S, Samsi K, Petrie CD, Alvir J, Treglia M, Schwam EM, *et al.* What do we know about quality of life in dementia? A review of the emerging evidence on the predictive and explanatory value of disease specific measures of health related quality of life in people with dementia. *Int J Geriatr Psychiatry* 2009;**24**:15–25.

21. Banerjee S. The development of a new measure of health related quality of life for people with dementia – DEMQOL: use in research and clinical practice. *J Alzheimers Dis Dement* 2007;**3**:116–71.

22. National Institute for Health and Clinical Excellence. *Donepezil, galantamine, rivastigmine (review) and memantine for the treatment of Alzheimer's disease*. NICE: London; 2006.

23. Brodaty H, Green A, Koschera A. Meta-analysis of psychosocial interventions for caregivers of people with dementia. *J Am Geriatr Soc* 2003;**51**:657–64.

24. Brooks R, EuroQol Group. EuroQol: the current state of play. *Health Policy* 1996;**37**:53–72.

25. Brazier JE, Roberts J, Deverill M. The estimation of a preference based measure of health from the SF-36. *J Health Econ* 2002;**21**:271–92.

26. Marra CA, Woolcott JC, Kopec JA, Shojania K, Offer R, Brazier JE, *et al.* A comparison of generic, indirect utility measures (the HUI2, HUI3, SF-6D and the EQ-5D) and disease-specific instruments (the RAQoL and the HAQ) in rheumatoid arthritis. *Soc Sci Med* 2005;**60**:1571–82.

27. Tosh J, Brazier J, Evans P, Longworth L. A review of generic preference-based measures of health-related quality of life in visual disorders. *Value Health* 2012;**15**:118–27.

28. Longworth L, Mulhern B, Yang Y, Tosh J, Keetharuth A, Rowen D, *et al. A systematic review of the performance of EQ-5D in four disease areas*. EuroQol Group Scientific Meeting, Oxford, 2011.

29. Papaioannou D, Brazier J, Parry G. How valid and responsive are generic health status measures, such as the EQ-5D and SF-36, in schizophrenia? A systematic review. *Value Health* 2011;**14**:907–20.

30. Kavirajan H, Hays RD, Vassar S, Vickrey BG. Responsiveness and construct validity of the health utilities index in patients with dementia. *Med Care* 2009;**47**:651–61.

31. Coucill W, Bryan S, Bentham P, Buckley A, Laight A. ED-5D in patients with dementia. *Med Care* 2001;**39**:760–71.

32. Hounsome N, Orrell M, Edwards RT. EQ-5D as a quality of life measure in people with dementia and their carers: evidence and key issues. *Value Health* 2011;**14**:390–9.

33. Karlawish JH, Zbrozek A, Kinosian B, Gregory A, Ferguson A, Glick HA. Preference-based quality of life in patients with Alzheimer's disease. *Alzheimers Dement* 2008;**4**:193–202.

34. D2000 study group. Long-term donepezil treatment in 565 patients with Alzheimer's disease (AD2000). *Lancet* 2004;**363**:2105–15.

35. Loveman E, Green C, Kirby J, Takeda A, Picot J, Payne E, *et al.* The clinical and cost-effectiveness of donepezil, rivastigmine, galantamine and memantine for Alzheimer's disease. *Health Technol Assess* 2006;**10**(1).

36. Takeda A, Loveman E, Clegg A, Kirby J, Picot J, Payne E, *et al.* A systematic review of the clinical effectiveness of donepezil, rivastigmine and galantamine on cognition, quality of life and adverse events in Alzheimer's disease. *Int J Ger Psych* 2006;**21**:17–28.

37. Brazier JE, Rowen D, Mavranezouli I, Tsuchiya T, Yang Y, Barkham M, *et al.* Developing and testing methods for deriving preference-based measure of health from condition-specific measures (and other patient based measures of outcome). *Health Technol Assess* 2012;**16**(32).

38. Brazier J, Roberts J. The estimation of a preference-based measure of health from the SF-12. *Med Care* 2004;**42**:851–9.

39. Brazier JE, Ratcliffe J, Tsuchiya A, Salomon J. *Measuring and valuing health for economic evaluation*. Oxford: Oxford University Press; 2007.

40. Yang Y, Brazier J, 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.

41. Young T, Yang Y, Brazier J, Tsuchiya A. The use of Rasch analysis in reducing a large condition-specific instrument for preference valuation: the case of moving from AQLQ to AQL-5D. *Med Decis Making* 2011;**31**:195–210.

42. Yang Y, Brazier J, Tsuchiya A, Coyne K. Estimating a preference-based single index from the Overactive Bladder Questionnaire. *Value Health* 2006;**12**:159–66.

43. National Institute for Clinical Excellence. *Guide to the methods of technology appraisal*. London: NICE; 2004.

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

45. Dolan P. Modeling valuations for EuroQol health states. *Med Care* 1997;**35**:1095–108.

46. Banerjee S, Hellier J, Dewey M, Romeo R, Ballard C, Baldwin R, *et al.* Sertraline or mirtazapine for depression in dementia (HTA-SADD): a randomised, multicentre, double-blind, placebo-controlled trial. *Lancet* 2011;**378**:403–11.

47. Smith SC, Murray J, Banerjee S, Foley B, Cook JC, Lamping DL, *et al.* What constitutes health-related quality of life in dementia? Development of a conceptual framework for people with dementia and their carers. *Int J Geriatr Psychiatry* 2005;**20**:889–95.

48. Banerjee S, Willis R, Matthews D, Contell F, Chan J, Murray J. Improving the quality of care for mild to moderate dementia: an evaluation of the Croydon Memory Service Model. *Int J Geriatr Psychiatry* 2007;**22**:782–8.

49. Alexopoulos GS, Abrams RC, Young RC, Shamoian CA. Cornell Scale for Depression in Dementia. *Biol Psychiatry* 1988;**23**:271–84.

50. Folstein MF, Folstein SE, McHugh PR. Mini mental state. *J Psychiatric Res* 1975;**12**:189–98.

51. Bucks RS, Ashworth DL, Wilcock GK, Siegfried K. Assessment of activities of daily living in dementia: development of the Bristol Activities of Daily Living Scale. *Age Ageing* 1996;**25**:113–30.

52. Cummings JL, Mega M, Gray K, Rosenberg-Thompson S, Carusi DA, Gornbein J. The Neuropsychiatric Inventory: comprehensive assessment psychopathology in dementia. *Neurology* 1994;**44**:2308–14.

53. EuroQol Group. EuroQol – a new facility for the measurement of health-related quality of life. *Health Policy* 1990;**16**:199–208.

54. National Institute for Health and Clinical Excellence. *Guide to the methods of technology appraisal*. London: NICE; 2008. URL: www.nice.org.uk/media/B52/A7/TAMethodsGuideUpdatedJune2008.pdf (accessed 2 September 2011).

55. Alexopoulos GS. *The Cornell Scale for Depression in Dementia: administration and scoring guidelines 2002*. URL: www.scalesandmeasures.net/files/files/The%20Cornell%20Scale%20for%20Depression%20in%20Dementia.pdf (accessed 31 August 2011).

56. Howard R, Phillips P, Johnson T, O'Brien J, Sheehan B, Lindesay J, *et al.* Determining the minimum clinically important differences for outcomes in the DOMINO trial. *Int J Geriatr Psychiatry* 2011;**26**:812–17.

57. Scientific Advisory Committee of the Medical Outcomes Trust. Assessing health states and quality of life instruments: attributes and review criteria. *Qual Life Res* 2002;**11**:193–205.

58. Streiner DL, Norman GR. *Health measurement scales: a practical guide to their development and use*. Oxford: Oxford Medical Publications; 1995.

59. WHOQOL Group. The World Health Organization quality of life assessment (WHOQOL): development and general psychometric properties. *Soc Sci Med* 1998;**46**:1569–85.

60. Lamping DL, Hobart JC, Schroter S, Smith SC. Developing short-form outcome measures: methodological approaches to item reduction. In D Lamping (Chair), *State of the art on item reduction: identifying the good, bad and ugly items: analysis and evaluation* (Symposium). International Society for Quality of Life Research Annual Conference, Orlando, FL, 30 October–2 November 2002.

61. Lamping DL, Schroter S, Kurz X, Kahn SR, Abenhaim L. Evaluating outcomes in chronic venous disorders of the leg: development of a scientifically rigorous, patient-reported measure of symptoms and quality of life. *J Vasc Surg* 2003;**37**:410–19.

62. Ferguson E, Cox T. Exploratory factor analysis: a user's guide. *Int J Select Assess* 1993;**1**:84–94.

63. Costello AB, Osborne JW. Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. *PARE* 2005;**10**:1–9.

64. Young T, Yang YL, Brazier JE, Tsuchiya A, Coyne K. The first stage of developing preference-based measures: constructing a health-state classification using Rasch analysis. *Qual Life Res* 2009;**18**:253–65.

65. Prado-Jean A, Couratier P, Druet-Cabanac M, Nubukpo P, Bernard-Bourzeix L, Thomas P, *et al.* Specific psychological and behavioural symptoms of depression in patients with dementia. *Int J Geriatr Psychiatry* 2010;**25**:1065–72.

66. Steinberg M, Shao H, Zandi P, Lyketsos CG, Welsh-Bohmer KA, Norton MC, *et al.* Point and 5-year period prevalence of neuropsychiatric symptoms in dementia: the Cache County Study. *Int J Geriatr Psychiatry* 2008;**23**:170–7.

67. Young T, Rowen D, Norquist J, Brazier JE. Developing preference-based health measures: using Rasch analysis to generate health state values. *Qual Life Res* 2010;**19**:907–17.

68. Novella JL, Jochum C, Jolly D, Morrone I, Ankri J, Bureau F, *et al.* Agreement between patients' and proxies' reports of quality of life in Alzheimer's disease. *Qual Life Res* 2001;**10**:443–52.

69. Boyer F, Novella JL, Morrone I, Jolly D, Blanchard F. Agreement between dementia patient report and proxy reports using the Nottingham Health Profile. *Int J Geriatr Psychiatry* 2004;**19**:1026–34.

70. Gold MR, Siegel JE, Russell LB, Weinstein MC. *Cost-effectiveness in health and medicine*. Oxford: Oxford University Press; 1996.

71. Rowen D, Brazier J, Tsuchiya A, Young T, Ibbotson R. It's all in the name, or is it? The impact of labelling on health state values. *Med Decis Making* 2012;**32**:31–40.

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

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

74. Federov VV. *Theory of optimal experiments*. New York, NY: Academic Press; 1972.

75. Salomon JA. Reconsidering the use of rankings in the valuation of health states: a model for estimating cardinal values from ordinal data. *Popul Health Metr* 2003;**1**:12.

76. McCabe C, Brazier J, Gilks P, Tsuchiya A, Roberts J, O'Hagan A, *et al.* Using rank data to estimate health state utility models. *J Health Econ* 2006;**25**:418–31.

77. Brazier J, Rowen D, Yang Y, Tsuchiya A. Comparison of health state utility values derived using time trade-off, rank and discrete choice data anchored on the full health–dead scale. *Eur J Health Econ* 2012;**13**:575–87.

78. 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.

79. Brazier J, Akehurst R, Brennan A, Dolan P, Claxton K, McCabe C, *et al.* Should patients have a greater role in valuing health states? *Appl Health Econ Health Policy* 2005;**4**:201–8.

80. Dolders MGT, Zeegers MPA, Groot W, Ament A. A meta-analysis demonstrates no significant differences between patient and population preferences. *J Clin Epidemiol* 2006;**59**:653–64.

81. Boyd NF, Sutherland HJ, Heasman ZJ, Cummings BJ. Whose values for decision making? *Med Decis Making* 1990;**10**:58–67.

82. Brazier J. Measuring and valuing mental health for use in economic evaluation. *J Health Serv Res Policy* 2008;**13**:70–5.

83. Krabbe PFM, Tromp N, Ruers TJM, van Riel PLCM. Are patients' judgments of health status really different from the general population? *Health Qual Life Outcomes* 2011;**9**:31.

84. Insinga RP, Fryback DG. Understanding differences between self-ratings and population ratings for health in the EuroQOL. *Qual Life Res* 2003;**12**:611–19.

85. Dolan P, Roberts J. To what extent can we explain time trade-off values from other information about respondents? *Soc Sci Med* 2002;**54**:919–29.

86. Goldstein H. *Multilevel statistical methods*. New York, NY: Halstead Press; 1995.

87. Dolan P, Gudex C, Kind P, Williams A. Valuing health states: a comparison of methods. *J Health Econ* 1996;**15**:209–31.

88. Burns A, Lawlor B, Criag S. Rating scaes in old age psychiatry. *Br J Psychiatry* 2002;**180**:161–7.

89. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;**1**:307–10.

90. Bharmal M, Thomas J. Comparing the EQ-5D and the SF-6D descriptive systems to assess their ceiling effects in the US general population. *Value Health* 2006;**9**:262–71.

91. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979;**86**:420–8.

92. Mander A. *BATPLOT: Stata module to produce Bland–Altman plots accounting for trend*. Statistical Software Components, Boston College Department of Economics, 2005. URL: http://ideas.repec.org/c/boc/bocode/s448703.html (accessed 30 March 2011).

93. Cohen J. *Statistical power analysis for the behavioral sciences*. New York, NY: Academic Press; 1969.

94. 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.

95. Flynn TN, Louviere JJ, Marley AA, Coast J, Peters TJ. Rescaling quality of life values from discrete choice experiments for use as QALYs: a cautionary tale. *Popul Health Metr* 2008;**6**:6.

Appendix 1 Time trade-off process

INTERVIEWER CHECK:

PICK UP PACK OF 8 GREEN HEALTH STATE CARDS (SHUFFLED).
TAKE OUT FIRST CARD TO BE VALUED. ENTER LETTERS OF THE CARD: _____

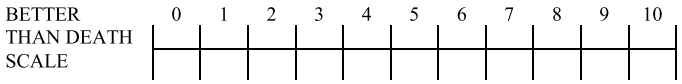
PASS CARD TO THE RESPONDENT.
Please read this card through carefully.

b. HAVE TTO BOARD WITH SIDE '1' FACING UPWARDS.
PLACE GREEN CARD IN POCKET FOR LIFE B.
MOVE BOARD MARKER FOR LIFE A TO 0 YEARS.
Now you would either die immediately, or you would live in Life B for 10 years and then die. Would you prefer to die immediately or to have Life B, or are they the same?

Life A	1. GO TO h. (STATE WORSE THAN DEATH)
Life B	2. GO TO c. (STATE BETTER THAN DEATH)
The same	3. GO TO C5

ASK IF 'LIFE B' (code 2) AT b.

c. STATE BETTER THAN DEATH
MARK 'X' UNDER 0 ON THE SCALE BELOW.



CONTINUE TO USE TIME BOARD WITH SIDE '1' UPWARDS
SET BOARD MARKER FOR LIFE A TO 5 YEARS (t=5).

d. Now you would either live in Life A for 't' years and then die, or you would live in Life B for 10 years and then die. Would you prefer Life A or Life B, or are they the same?
CONTINUE TO WRITE ON SCALE ABOVE ON THIS PAGE.

IF A:	✓ UNDER 't'	MOVE MARKER 1 YEAR TO THE LEFT. REPEAT d. WITH 't' 1 LESS THAN LAST TIME.
IF B:	× UNDER 't'	MOVE MARKER 1 YEAR TO THE RIGHT. REPEAT d. WITH 't' 1 MORE THAN LAST TIME.
IF SAME:	= UNDER 't'	GO TO C5

REPEAT d. UNTIL:	
A) YOU ENTER '='	GO TO C5 OR
B) '×' AND '✓' APPEAR NEXT TO EACH OTHER	GO TO e.

ASK IF d. ENDED WITH '×' AND '✓' NEXT TO EACH OTHER

e. LET 't' NOW BE HALFWAY BETWEEN THE ADJACENT CROSS AND TICK, I.E. 'SOMETHING AND 6 MONTHS'

What if you would either live in Life A for 't' and then die, or you would live in Life B for 10 years and then die. Would you prefer Life A or Life B, or are they the same?

Life A	1. GO TO C5
Life B	2. GO TO f.
The same	3. GO TO C5

IF 'LIFE B' (code 2) AT e.

f. INTERVIEWER CHECK:

IF THERE IS A × UNDER 9	1. GO TO g.
IF THERE IS NOT A × UNDER 9	2. GO TO C5

ASK IF THERE IS '×' UNDER 9 AND '✓' UNDER 10

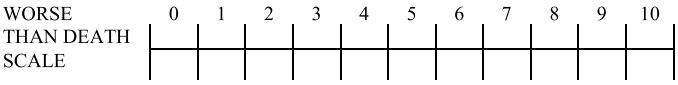
g. Would you be prepared to sacrifice any time in order to avoid Life B?
IF YES: How many weeks?
ENTER WEEKS: _____

Yes	1.
No	2. GO TO C5

ASK IF 'LIFE A' (code 1) AT b.

h. STATE WORSE THAN DEATH

MARK '✓' UNDER 0 ON SCALE BELOW.



TURN TTO BOARD SIDE '2' UPWARDS.
MOVE GREEN CARD TO TOP LEFT POCKET ON SIDE '2'.
SET BOARD MARKER FOR LIFE A TO 5 YEARS (t = 5).
Now here is a different choice.

i. Life A is now 't' years of this state (POINT TO THE GREEN CARD) followed by '10-t' years in this other state (POINT TO THE PINK CARD). Or instead of that you could choose to die immediately (POINT TO LIFE B). Would you prefer Life A, or to die immediately, or are they the same?
WRITE ON SCALE ABOVE ON THIS PAGE.

IF A:	✓ UNDER 't'	MOVE MARKER 1 YEAR TO THE RIGHT. REPEAT i. WITH 't' 1 MORE THAN LAST TIME.
IF B:	× UNDER 't'	MOVE MARKER 1 YEAR TO THE LEFT. REPEAT i. WITH 't' 1 LESS THAN LAST TIME.
IF SAME:	= UNDER 't'	GO TO C5

REPEAT i. UNTIL:	
A) YOU ENTER '='	GO TO C5 OR
B) '✓' AND '×' APPEAR NEXT TO EACH OTHER	GO TO j.

ASK IF i. ENDED WITH '✓' AND '×' NEXT TO EACH OTHER

j. LET 't' NOW BE HALFWAY BETWEEN THE ADJACENT TICK AND CROSS, I.E. 'SOMETHING AND 6 MONTHS'.
What if Life A was 't' of this state (POINT TO THE GREEN CARD) followed by '10-t' in this other state (POINT TO THE PINK CARD). Or instead of that you could choose to die immediately (POINT TO LIFE B). Would you prefer Life A, or to die immediately, or are they the same?

Life A	1.
Life B	2. GO TO C5
The same	3.

Appendix 2 Protocol

GENERATION OF PREFERENCE-BASED INDICES FROM DEMQOL AND DEMQOL-PROXY FOR USE IN ECONOMIC EVALUATION

DETAILED PROJECT DESCRIPTION

1.0 Background

1.1 The challenge of dementia

Dementia is one of the most common and serious disorders in later life with a prevalence of 5% and an incidence of 2% per year in the over 65s (Hofman *et al* 1991; Launer *et al* 1992) equating to 700,000 people with dementia at any one time (Knapp *et al* 2007) and 200,000 new cases every year in the UK. It causes irreversible decline in global intellectual, social and physical functioning. Abnormalities in behaviour, insight and judgement are part of the disorder, as are neuropsychiatric symptoms such as psychosis, anxiety and depression. The economic cost of caring for people with dementia is immense. In the UK the costs of dementia are around £17 billion (Knapp *et al* 2007), greater than stroke (£3 billion), heart disease (£4 billion) and cancer (£2 billion) (Lowin *et al* 2001). We can also predict that the challenges posed by dementia will only grow in the next decades. In the next 30 years the number of people with dementia in the UK will double to 1.4 million (Knapp *et al* 2007) and the costs will treble, with the growth in costs faster than the costs of caring for an ageing population more generally (Comas-Herrera *et al* 2007). More importantly, the negative impacts of dementia on those with the disorder, in terms of deteriorating function, and on carers (Murray *et al* 1999; Schneider *et al* 1999) are profound. The need to improve care for people with dementia is a policy priority (DH 2001, 2005; NAO 2007).

1.2 Evaluation of clinical effectiveness in dementia

Given its importance in public health terms and its devastating effects, it is understandable that there is a large and growing volume of basic, translational, and applied research underway. This includes evaluations of psychological and social interventions as well as trials of pharmacological treatments. Given the complexity of dementia, there has been discussion about how best to measure the effects of interventions in dementia. There is an emerging consensus that we need to measure broad patient-reported outcomes such as health-related quality of life (HRQL) in dementia as well as discrete areas like cognition or behaviour (Whitehouse 2000). The technologies for measuring discrete areas of function are well developed with a variety of psychometric robust instruments available for evaluating all major domains (eg cognition, behaviour, activities of daily living, depression) often using proxy reports of observable behaviour.

Measuring HRQL in dementia is more challenging, not least because of poor recall, time perception, insight and communication. However, recent studies indicate that meaningful measurements can be made using disease-specific measures, based on self- and proxy-report (Brod *et al* 1999, Logsdon *et al* 2002, Smith *et al* 2005, 2007).

Funded by the HTA we have developed the DEMQOL system, a disease-specific measure of HRQL in dementia (Smith *et al* 2005; 2007). The DEMQOL system consists of two interviewer-administered tools: DEMQOL (28 items) is completed by the person with dementia (score range 28 to 112; higher scores indicate better HRQL); and DEMQOL-Proxy (31 items) a proxy report of the person with dementia's HRQL completed by the main carer (score range 31 to 124; higher scores indicate better HRQL). The system was designed according to best psychometric practice and has undergone extensive validation in the UK.

DEMQOL has good psychometric properties for people with mild to moderate dementia (defined as a MMSE score of 10+). DEMQOL-Proxy can be used across disease severity from mild to severe.

The development of such instruments has lagged behind the measures of discrete function. So they have not yet been widely employed in randomised controlled trials (RCTs) of anti-dementia medication (ADM) and other treatments as have measures of discrete areas of function, most commonly cognition. We have analysed associations between commonly used measures of specific outcomes in dementia and HRQL (Banerjee *et al* 2006). The data suggest that HRQL in dementia does not have a simple relationship with cognition or functional limitation. This and other studies (Woods *et al* 2006) suggest that cognitive impairment is an inadequate proxy for HRQL improvement in dementia. They suggest that there may be considerable value in including measures of HRQL along with measures of specific function such as cognition and behaviour in treatment trials in dementia. A failure to include broad outcome measures such as HRQL, and a reliance on measures of discrete functions (eg cognition), could lead to the positive effects of interventions being overlooked or to potential negative effects of intervention being missed (Banerjee 2007).

1.3 Economic evaluation in dementia

Things are not quite so clear for the economic evaluation of treatments in dementia. Even though we can be relatively confident on data on the clinical effectiveness of treatments for dementia, there is a real lack of directly relevant data with which to ascertain the cost-effectiveness of treatments. The last decade has seen the increased use of economics to inform the allocation of resources between competing health care interventions around the world and particularly the use of cost-effectiveness analysis, where interventions are assessed in terms of their cost per Quality-Adjusted Life Year (QALY) gained. The QALY provides a way of measuring the benefits of health care interventions, including improvements in HRQL. Brief generic (ie not disease-specific) measures of HRQL are most commonly used to put the 'Q' into the QALY. Such measures include the EQ-5D (Brook, 1996) and other generic preference-based measures such as SF-6D (Brazier *et al*, 2002). It is suggested that these generic measures are applicable to all interventions and patient groups, a claim that has some support in some conditions, such as rheumatoid arthritis where it has managed to pass conventional psychometric tests of reliability and validity (eg Marra *et al*, 2006), but is more questionable in others, such as visual impairment (Espallargues *et al*, 2006) and hearing loss (Barton *et al*, 2004).

Such brief generic measures of HRQL do not work well in dementia. As noted above the inherent impairments in dementia of recall, time perception, insight and expressive and receptive communication, mean that it is not possible to assume that what works for a general non-cognitively impaired population will work in those with dementia. This means that instruments to be used in dementia need to be psychometrically tested in populations of people with dementia. In fact, where self or proxy report is needed, this generally means that disease-specific measures need to be generated which can measure accurately in dementia whatever attribute is under consideration. Brief generic measures of HRQL have been tested in dementia and there is data that these can be completed by people with dementia (Jonsson *et al*, 2006; Coucill *et al*, 2001; Naglie *et al*, 2006), the questions revolve around the validity of the data generated. There are questions about how patient insight into cognitive impairments and activity limitations affect preference ratings (Vogel *et al*, 2006). Equally the instruments used have limits in the validation available data especially with respect to the complex co-morbidities found in dementia. The EQ5D has been reported to have more correlations than the HUI3 but limitations were observed for both instruments, with for example 43% to 57% of people with dementia rating themselves as having perfect health (Naglie *et al*, 2006). The likelihood that such instruments may well not capture the impact of dementia is raised by their not specifically having been developed for use in this population and the fact that the HUI2 includes a single item on cognition and the EQ-5D none.

In practical terms, the unsatisfactory nature of the current evidence base is very clearly illustrated by the major difficulties presented to the National Institute for Health and Clinical Excellence (NICE) in generating their recent (2006) Technology Appraisal Guidance (TAG111). There has been a great deal of concern

raised (including referral to Judicial Review) about the assumptions they had to make with respect to their cost-effectiveness models. The challenges encountered can be attributed to a lack of direct data on cost and quality of life. The conclusions of TAG111 make clear the need for a technology that can be used in dementia trials to generate direct and accurate measurements of cost and effectiveness in quality-of-life terms. This conclusion is echoed in systematic reviews and trials in dementia (AD2000 2004, Loveman *et al* 2006; Takeda *et al* 2006).

What then is needed to enable cost-effectiveness evaluation in dementia? If the use of the brief generic measures is problematic due to the error inherent in their use, might it be possible to use instruments that can measure HRQL in dementia such as the QOL-AD or DEMQOL and DEMQOL-proxy? These instruments essentially cannot be used in cost–utility evaluations directly in their current form because they are too large to incorporate preference information. They therefore cannot be used to calculate QALYs for use in incremental cost-effectiveness analysis. This is a major limitation in the current available measurement technology. However the methodology is available to allow the benefits of the DEMQOL system in being able to measure HRQL in dementia accurately to be applied to valuing the benefits of interventions in this area in economic evaluation. This study therefore aims to generate a preference-based single index for these two instruments (DEMQOL and DEMQOL-Proxy) for use in economic evaluation using general population and patient/carer values.

2.0 Method

2.1 Overview of the plan of investigation

To derive preference-based single index measures from the two DEMQOL instruments, we propose to apply the methods that have been developed by one of the applicants (JB) in Sheffield. These methods have been applied with success to the SF-36 (Brazier 1998, 2002), the King's Health Questionnaire (Brazier *et al* 2007), the Asthma Quality of Life Questionnaire (AQLQ – Yang *et al* 2006b), and the Over Active Bladder Questionnaire (OABq – Yang *et al* 2006a). The problem with deriving preference-based measures from such measures of HRQL is that they are simply too large. With multiple dimensions and numerous items they would define many millions of potential health states and produce health states too large for valuation by respondents.

The first stage in this type of work is therefore to fashion a health-state classification (like the EQ-5D in structure) by sampling items from the original instruments using conventional and advanced psychometric methods. We will use existing data sets in which DEMQOL and DEMQOL-Proxy have been used. The overall aim will be to develop two brief health-state classifications, one from DEMQOL and one from DEMQOL-Proxy. These will be used to generate health states for valuation. They will consist of existing items from DEMQOL and DEMQOL-Proxy respectively, and so can be derived from any study using the existing DEMQOL and DEMQOL-Proxy instruments. They will be known as DEMQOL-nD and DEMQOL-Proxy-nD (where n is the number of items in the shortened measure). Consensus from organisations such as NICE around the world require HRQL to be valued using a choice-based technique. We have selected a time trade-off technique (TTO) for this study, where respondents are asked how many years they would be willing to sacrifice in order to be in full health. TTO was selected over standard gamble, since the latter asks respondents to consider how much they are willing to risk their life; given that dementia is predominantly a disorder of later life TTO seemed a more appropriate line of questioning. We propose to use the TTO version developed in York for the EQ-5D since this will allow comparison with the EQ-5D population value set used by NICE. In addition we propose to use two ordinal methods for valuation, ranking and a discrete choice experiment (DCE).

The question of whose values should be used to value health is ultimately political and beyond the scope of this proposal (Brazier *et al* 2007). We propose that the main valuation survey in this study should use a representative sample of the general public to conform to the requirements of NICE and other reimbursement authorities. However there are active issues in whether this is the right group given the

nature of the disorder and the state of public attitudes and understandings of dementia. We therefore also propose to complete a supplementary valuation survey of people with dementia and their carers in order to explore the size and direction of any deviation from the general public's values. In preparing this proposal we have engaged in conversation with the Alzheimer's Society through its Quality Research in Dementia (QRD). They are clear that they would wish the values of people with dementia and their family carers to be directly investigated as part of this research.

Data from the general population survey will be modelled to estimate preference-based scoring algorithms that can be applied to existing and future DEMQOL and DEMQOL-Proxy data. The final phase of the study will be to apply these algorithms to two trial data sets that are underway (HTA-SADD, a placebo-controlled RCT of the treatment of depression in dementia, Chief Investigator (CI) SB; and MRC-DOMINO, a placebo-controlled RCT of donepezil and memantine alone and in combination for the treatment of those where treatment response is questioned, CI RH) to estimate QALYs directly. In each case, the trial proposal agreed by the funding body explicitly noted our intention to explore the possibility of generating QALY measures to run alongside the disease-specific outcome measures. MK is responsible for the economic evaluation in both HTA-SADD and MRC-DOMINO. We will also test the psychometric properties of the indices; and compare the results of using the two DEMQOL indices, the EQ-5D, and population and patient/carer valuations.

The project divides itself into five linked phases. These are:

- derivation of the health-state classification
- main population valuation survey
- patient/carer valuation survey
- modelling, and
- application to trial data

2.2 Phase 1: Derivation of health states from DEMQOL and DEMQOL-Proxy

The first task is to derive two health-state classifications, one from DEMQOL and the other from the DEMQOL-Proxy. To accomplish this, extensive psychometric analyses will be undertaken to: i) confirm the dimensional structure of the measures, and (ii) select items from each domain to construct the health-state classifications.

2.2.1 Dimensional structure

The conceptual framework which guided the development of the DEMQOL instruments covers five domains (daily activities, health and well-being, cognitive functioning, social relationships, self-concept). In our initial validation study (Smith *et al* 2005, 2007) factor analyses identified a 4-factor model for DEMQOL and a 2-factor model for DEMQOL-Proxy. The first stage in this proposed research will be to investigate further the factor structure of DEMQOL and DEMQOL-Proxy in data obtained in our subsequent work including: clinical evaluation of the Croydon Memory Service (CMS) – over 1,000 DEMQOL and DEMQOL-Proxy scores (Banerjee *at al* 2007); and a longitudinal study of the natural history of HRQL in dementia – 100 baseline and six month follow-up DEMQOL scores. The results of these further factor analyses will be compared with those from the initial validation work, and new Rasch analyses undertaken on all three data sets to derive a final set of dimensions for the health-state classification.

2.2.2 Item selection

The purpose of this stage of the psychometric analyses is to reduce the 28-item DEMQOL and the 31-item DEMQOL-Proxy to the minimum number of items needed to derive a classification system for a preference-based utility index.

We will undertake item reduction using a clearly defined strategy developed in our previous work (Lamping *et al* 2002a, 2002b, 2003; Hilari *et al* 2003; Hobart *et al*, 2004; Smith *et al* 2005, 2007) and based on state-of-the-art psychometric methods (Nunnally & Bernstein, 1994; Scientific Advisory Committee,

2002; Streiner & Norman, 2003), including Rasch analyses (refs). Our strategy defines *a priori* the psychometric tests to be conducted and the specific criteria for item elimination/retention (see *Table 1*). During item reduction, we will use both classical psychometric tests and Rasch analyses to consider items for elimination, based on examination of missing data, maximum endorsement frequencies, aggregate adjacent endorsement frequencies, redundancy, item–total correlations, factor analysis, item convergent/discriminant analysis, responsiveness (where available) and Rasch item threshold probability curves.

Table 1 shows the psychometric tests and criteria used in the development and validation of DEMQOL. Tests and criteria for the proposed Rasch analyses are described below. Given that the aim of this study is to produce a preference-based measure from an existing validated, item-reduced questionnaire, more conservative psychometric criteria for item reduction will be applied than those used in our previous work in order to carry out more extensive item elimination. For example, a more stringent criterion than the standard $< 5\%$ for missing data will be used to determine whether an item should be retained. Criteria for all other psychometric tests applied during the item reduction phase (eg internal consistency, item–total correlations, item convergent and discriminant validity, etc.) will be reviewed by DL, JB, SS and TY to reach a consensus agreement about the more conservative criteria to be applied in this stage of the analyses.

Rasch analysis

Rasch analysis is a mathematical technique that converts qualitative (categorical) responses to a continuous (unmeasured) latent scale using a logit model. This technique has been used successfully to assist in the selection of items from the AQLQ and OABq in the development of preference-based indices (Young *et al* 2005, 2007). RUMM2020 will be used for the analysis.

Prior to item selection and having fitted the data to the Rasch model, the first step is to see whether items from the DEMQOL instruments fit the Rasch model. The assumption from Rasch analysis being that each set of items measure an underlying trait, eg self esteem related HRQOL and that this HRQOL can be measured on a unidimensional scale. The first step in this process is to see whether the levels of each item are ordered correctly on the logit scale. If levels of a given item are unordered or even disordered, it indicates that responders are unable to distinguish between these levels so these should be merged. Rasch analysis is then repeated until ordering is achieved. To achieve uniformity across all items, the same levels will be merged across items. This could potentially leave some items with disordered responses, and these will be considered for removal from the final item selection of the preference based indices.

The second step in the initial Rasch model fitting is to examine items for the presence of differential item function (DIF). DIF indicates that the way items are answered varies between responders by background characteristics such as age (eg less than or greater than 75), gender, and symptom severity (mild, moderate and severe). For example if women consistently answer items differently to males and those consistently have a worse/better QOL score for a particular item. If items are consistently answered better than others with a different characteristic, this suggests such items must be interpreted separately by these characteristics, or 'split'. In normal applications of Rasch analysis such items can be modified or treated separately, but given the aims of this phase of the work, such items will be considered for removal.

The third step in producing a well-fitting Rasch model is to remove individual items which do not fit the model – these are identified by studying the item goodness of fit statistics. Poor fitting items will be removed one by one until the overall Rasch model goodness of fit becomes insignificant (χ^2 p-value > 0.01). These items will not be considered further in the development of the preference based indices. Rasch models will be fitted for each dimension of the DEMQOL instrument as determined from the factor analysis.

After achieving a good fitting Rasch model the final step is to check each of the remaining items, by domain, in terms of the Rasch statistics that could be used in the selection process of items from the DEMQOL instruments. The spread across the latent variable (spread at logit 0) will be used as the main

criterion to choose the item to represent a given domain. The goodness-of-fit statistics (χ^2 and *p*-value) and results from the psychometric analysis will also be taken into account.

The final selection of items will be made by a panel of the study investigators. They will take into account the results of all the analyses described above, combined with their own understanding of the instrument and the need to derive a health-state classification that will be amenable to valuation.

2.2.3 Final derivation of health states

The aim will be to derive two multi-dimensional health-state classifications. We have achieved this with success in a number of instruments, including the AQLQ and OABq. In a study being undertaken by a PhD student supervised by JB it was found that it was only possible to derive two dimensions from a mental health outcome measure, the CORE-OM. This was because the items were very highly correlated and there were only two factors possible. Given the nature of the subject, this is possible in the work proposed here. Therefore, rather than limit the descriptive systems to two items in such circumstances which would risk the measure being unreliable and loosing a lot of information, the approach we will use is to use the Rasch model to define typical respondents (and hence health states) at different points along the latent variable as defined by 5 or 6 items. The health states valued in this way can then be mapped onto the latent variable in order to generate values for other states generated by the selected items. Clearly, until the full psychometric analyses have been completed it is not possible to predict the best way to generate states (whether by a health-state classification or not) or the precise number of health states. The remainder of the proposal assumes it will be possible to generate a health-state classification for DEMQOL and DEMQOL-Proxy.

2.3 Phase 2: Main valuation survey

The aim of this phase of the project will be to obtain valuations of states defined by items from the DEMQOL and DEMQOL-Proxy questionnaires.

2.3.1 Sample of states

Health-state classifications are likely to define many thousands of states, so a sample of states will be valued and modelling techniques will extrapolate values for all other states. The selection will be based on a balanced design, which ensures that any dimension-level (level λ of dimension δ) has an equal chance of being combined with all levels of the other dimensions. Deciding the number of states to value is rather more difficult to do formally, but recent studies of condition specific measures have successfully used 100 states to estimate additive functions and explore simple summary terms to reflect interactions (Yang *et al* 2006). The design of the DCE will be developed using the library of orthogonal arrays at http://www.research.att.com/~njas/oadir/ to enable all states to be valued.

2.3.2 Respondents

The main sample of respondents in this survey will be representative of the general population and reflect the variability of the population in terms of characteristics such as age, socio-economic status and level of education. The sample will be drawn using a two-stage cluster random selection design. The primary units will be postcode sectors stratified by percentage of households with a non-manual occupation. Postcode sectors will be selected, and addresses randomly selected from each of these. Where more than one adult (ie 16 or over) is found in household, one will be selected at random by the interviewer using a standard Kish selection grid.

2.3.3 The interview

A study of 20 respondents, comprising members of the general population will be undertaken prior to the main study to check respondents' understanding of the tasks and to check that they are completing each task as expected. At the interview, respondents will be taken through the descriptive system and asked to confirm that they understand it. They will then be asked to rank and value 8 states using TTO. Respondents will be taken through one TTO to ensure they understand the task. They will also be asked a number of

background questions. Assuming there are 100 states to value for each descriptive system and each state is valued 30 times, the required sample size for one health-state classification is 375. This sample size or less has been successfully used to value a number of descriptive systems (Brazier *et al* 2002; Brazier *et al* 2007; Yang *et al* 2006). A further 25 interviews will be conducted to provide the 64 required for the comparison with user values in phase 4. Assuming the same level of complexity for the items derives from DEMQOL-Proxy this will require a further 375 interviews, suggesting a final sample size of 775.

After the interview, general public respondents will be asked to consent to being sent a postal questionnaire containing the DCE exercise. Consenting respondents will be sent the questionnaire four weeks after their interview. This approach has been used in two previous valuation surveys undertaken at the University of Sheffield, achieving response rates of over 50% (Brazier *et al* 2006). To increase the sample size, the questionnaire will also be sent to a sample of 600 individuals who have not been interviewed in order to supplement this sample and to ascertain whether the interview itself altered people's valuations.

2.4 Phase 3: Patient and carer valuation survey

The aim will be to examine whether health-state values elicited for health states defined by the DEMQOL and DEMQOL-Proxy by patients and cares respectively, differ significantly from the general public. Respondents from these two groups will be recruited from clinical contacts in South London. We have been successful in recruiting large cohorts of people with dementia and their family carers for study in this way. The people with dementia to be interviewed will be assessed as having mild dementia (defined by MMSE > 18) and their main family carers will also be recruited where possible. In addition the family carer group will be enhanced by a balanced sample of those caring for people with moderate and severe dementia.

Respondents will be asked to value a set of 8 states using the same methods as the general population valuation survey. These interviews will be undertaken in people's own homes at times that are convenient for them with great care by researchers from the Institute of Psychiatry experienced in interviewing people with dementia patients and their carers. Training in the valuation methods will be provided by the Sheffield team.

Assuming a power of 0.8, significance level of 0.05, standard deviation of 0.3 and an expected difference of 0.1, then this requires a sample of 71 interviews to compare to mean valuations per state from the main survey for each of the instruments. Given the separate work needed on DEMQOL-nD and DEMQOL-Proxy-nD and the need to minimise respondent burden to enable the completion of the cognitively complex tasks, we will therefore need to recruit 142 people with dementia and 142 family carers who can complete the assessments. Mean values will be compared using simple *t*-tests.

The participants in the patient and carer valuation survey will also be used to carry out a direct evaluation of the psychometric properties of the item-reduced, preference-based DEMQOL-nD and DEMQOL-Proxy-nD instruments embedded within DEMQOL in an independent sample. This will use standard psychometric methods (as described above and in *Table 1*) to evaluate acceptability, reliability (internal consistency and test–retest), validity (content, convergent, discriminant and known group differences).

2.5 Phase 4: Modelling health-state values

The econometric models will have an additive specification, with the TTO value as the dependent variable and each level of each domain, other than the baseline, entered as dummy variables. A range of different specifications will be explored, including aggregate models using mean health-state values and random effects models using individual level data (Brazier *et al* 2002). The DCE data will be analysed using a random effects probit model that also assumes an additive relationship between the dimensions. The impact of adding interaction terms and various transformations will be explored. All models will be subjected to the standard tests of goodness of fit, *t*-tests of the coefficients, heteroskedasticity, normality

of errors, and robustness (Brazier *et al* 2002). A new Bayesian statistical model will also be estimated that is more theoretically appropriate than conventional models and has been found to perform better in terms of predicting out of sample values (Kharroubi *et al* 2007). This will also be used to examine the impact of covariates, including the respondent's mental health.

The best models will be selected and converted into scoring algorithms that can be applied to existing and future DEMQOL and DEMQOL-Proxy data sets. The algorithms will be produced in SPSS, SAS and Excel and will be made publicly available free of charge.

2.6 Phase 5: application to trial data

2.6.1 HTA-SADD

HTA-SADD is a multi-centre double-blind placebo-controlled RCT of the clinical and cost effectiveness of two classes of antidepressants, and more specifically, mirtazapine and sertraline, from baseline to 3 months (13 weeks) and 9 months (39 weeks), enabling estimation of short and long-term impacts of these antidepressants on depression in dementia. The primary objective of the study is to determine the clinical and cost-effectiveness of the two classes of antidepressants for depression in dementia (compared with placebo) as measured by the Cornell Scale for Depression in Dementia and (on the cost side) the societal resource impacts. Secondary objectives include an investigation of: differences in the clinical and cost-effectiveness, and, in terms of adverse events, withdrawals from treatment and adherence to treatment; differences in the clinical and cost-effectiveness of mirtazapine or sertraline compared to placebo on patient (eg quality of life, cognition) and family carer (eg carer burden, carer quality of life) outcomes; and the influence on clinical and cost-effectiveness of clinical characteristics including: dementia severity, dementia type, depression type, depression severity, care arrangements, neuropsychiatric symptoms, and physical illness.

Setting and selection

The trial is set in secondary care, using referrals to old age psychiatric services and memory clinics in 9 regional sites each covering a catchment area of 100,000 older people (Birmingham, Cambridge, Leicester, Liverpool, Manchester, Newcastle, North London, Southampton and South London) aided by the Department of Health Mental Health Research Network (MHRN) and DeNDRoN. This is a pragmatic trial. The criteria for inclusion are as close to clinical practice as possible. We will recruit those where a secondary care doctor makes a clinical diagnosis of mild to moderate probable or possible Alzheimer's Disease and a co-existing depressive illness of at least four weeks duration, likely to need treatment with antidepressants. The local research worker (RW) will then assess the patient's depression severity and those with a Cornell Scale for Depression in Dementia (CSDD) of 8+ will be eligible for entry into the trial. The other trial exclusions will be: currently taking antidepressants, the case being too critical to be randomised; absolute contra-indications to trial medications, being on another trial, treatment with antidepressants in the past four weeks, and no family or professional carer to give collateral information.

Randomisation and assessment

Patients will be allocated to placebo, sertraline or mirtazapine (ratio 1 : 1 : 1) by the Mental Health & Neurology Clinical Trials Unit based at the Institute of Psychiatry. Allocation will be stratified by centre by stratified block randomisation with randomly varying block sizes. Cases identified will be assessed by RW who will collect baseline and follow-up data (0 m, 3 m, and 9 m). The primary outcomes will be depression score – CSDD and cost – Client Service Receipt Inventory (CSRI). Secondary outcomes will include: adverse events, compliance, patient quality of life (disease-specific DEMQOL, generic EQ5D), cognition (MMSE), behavioural and psychological symptoms (NPI), carer burden (Zarit), carer stress (GHQ12), and carer quality of life (SF12 v2). The analysis of the economic impact of the interventions is a central, fully integrated element of the proposed study. The comprehensive costs of care for all participants will be calculated (including the costs of formal care such as that provided by health and social services and also the costs of informal care) using data gathered using the CSRI completed by key workers or family

carers at baseline, 13 w and 39 w. Unit costs will be best national estimates of the long-run marginal opportunity costs. Informal care will be costed. An overall sample size of 507 patients will provide 90% power to detect a 2 point difference in CSDD (SD 5; SES 0.4) for the primary comparisons of mirtazapine vs. placebo and sertraline vs. placebo at 13 weeks and 86% power for the secondary analysis of these comparisons at 39 weeks. This allows for 10% loss to follow-up at 13 weeks and 20% loss to follow-up at 39 weeks, correlation between baseline and outcome CSDD > 0.6, and up to 12.5% of those randomized (per comparison) to be either drop-outs or drop-ins using an analysis of covariance with 2-sided 5% significance levels. Allowing for the same levels of loss to follow-up, an overall sample of 507 patients would also enable us to calculate 2-sided 95% confidence intervals for the difference in the proportion of pre-specified adverse events between the antidepressant arms of (a clinically significant) 10% (i.e. 5% vs. 15%) $\pm$ 6% at 13 weeks and $\pm$ 7% at 39 weeks.

Analyses

CSDD score at 13 weeks will be analysed by ANCOVA adjusted for baseline CSDD and centre with contrasts for (a) sertraline vs. placebo and (b) mirtazipine vs. placebo. Secondary Analyses – The ANCOVA of CSDD score at 13 weeks will further include a contrast for mirtazapine vs. sertraline. CSDD score at 39 weeks will be analysed by ANCOVA adjusted for baseline CSDD and centre with contrasts for (a) sertraline vs. placebo; (b) mirtazipine vs. placebo, and (c) mirtazapine vs. sertraline. Secondary outcomes will be compared using the same contrasts as above within a [longitudinal] generalised linear model framework adjusting for the respective baseline scores and centre. The significance level will be 5% (2-sided) for all specified analyses of the primary outcome variable and 1% (2-sided) for all specified analyses of secondary outcome variables. From the cost and the outcome data, we will compare total and component (by service or agency) costs, incremental cost-effectiveness ratios and net benefits (using the primary outcome measure CSDD), cost–utility ratios (using utility scores computed from the EQ-5D and societal weights) and cost–consequences results (using all non-cost outcomes measures). The primary evaluation will be the cost-effectiveness analyses with CSDD change as the outcome. The evaluation will include the plotting of cost-effectiveness acceptability curves generated from bootstrap analyses. Sensitivity analyses will explore the impact of differences in key costs and outcome assumptions. Modelling will be conducted to predict costs and outcomes beyond the duration of the trial. The evaluation will be conducted from (a) societal, (b) public sector and (c) NHS perspectives. The projected date for a full data set for evaluation is September 2008, fitting well with the time frame of this proposal.

Application the new system to the HTA-SADD data set

We will work with the existing trial statisticians and economists to generate a further analysis strategy that will be applied to the trial data set. This will include the derivation of DEMQOL-nD and DEMQOL-Proxy-nD scales from the trial data, application of both the population and the patient/carer valuations and leading to the generation of cost effectiveness analyses using the derived dementia QALY. This will allow us to go beyond the originally (and funded) plan of analysis which was to generate utility scores from the EQ-5D for the purposes of the cost-effectiveness analyses. This further work will be led by MK.

2.6.2 MRC-DOMINO

MRC-DOMINO is a pragmatic, multi-centre, double-blind, randomised, placebo-controlled (double dummy), parallel group, 2×2 factorial clinical trial. The aim of the DOMINO study is to determine, in a factorial (2×2) design whether there is worthwhile benefit for patients for whom there is uncertainty on whether or not to continue cholinesterase inhibitors from: 1) adding memantine to cholinestease inhibitors, 2) switching to memantine, or 3) continuing cholinesterase inhibitors, as compared to 4) placebo.

Memantine	Donepezil	
	Continue	Discontinue
Add	*Group 1*	*Group 2*
	Donepezil	Donepezil placebo
	Memantine	Memantine
No	*Group 3*	*Group 4*
	Donepezil	Donepezil placebo
	Memantine placebo	Memantine placebo

Setting and selection

There will be 15 clinical recruiting centres: 1. The Institute of Psychiatry and South London and Maudsley NHS Trust, (Professors Howard, Ballard, Banerjee), 2. Bristol (Professor Wilcock), 3. Bath (Professor Jones), 4. Birmingham (Dr Bentham), 5. Manchester (Professor Burns), 6. Leicester (Professor Lindsay), 7. Newcastle (Professors McKeith, O'Brien), 8. Warwick (Dr Sheehan), 9. Perth and Tayside (Dr Findlay), 10. Imperial College (Dr Ritchie), 11. Paisley (Dr Hughes), 12. Oxford (Professor Jacoby), 13. Cambridge (Dr Dening), 14. Southampton (Professor Holmes), and 15. Belfast (Dr Passmore). Inclusion Criteria – participants will be patients who meet NINCDS-ADRDA criteria for probable or possible Alzheimer's disease (McKhann *et al*, 1984). In addition they will meet all of the following criteria: (1) Continuously prescribed donepezil for at least 3 months; (2) No change in dosage of donepezil in previous 6 weeks; (3) No changes in prescription of any psychotropic (antipsychotic, antidepressant, benzodiazepine) medication in previous 4 weeks; (4) Prescribing clinician considers (based on NICE guidance, discussions with patient and carer and clinical judgement) that change of drug treatment (i.e. stop donepezil or introduce memantine) may be appropriate and MMSE = 5 to 13 (13 chosen as NICE threshold of 10 plus 1 SD on MMSE score); (5) Patient is community resident and has family or professional carer or is visited on at least a daily basis by carer; (6) Patient agrees to participate; (7) Main carer (informal or institutional) consents to their own involvement. Exclusion Criteria – To maximise the generalisability of the study data, exclusions will be kept to a minimum. These will include: (1) Patient has severe, unstable or poorly controlled medical conditions apparent from physical examination or clinical history; (2) Patient is already prescribed memantine; (3) Patient is unable to take trial medications; (4) Patient is involved in another clinical trial; (5) Patient has absolute contraindication to either donepezil or memantine.

Randomisation and assessment

Randomisation will be done centrally by telephone to the MRC Clinical Trials Unit (CTU) in London, using a dedicated hotline. Proposed duration of treatment period. 52 weeks. All study measures will be assessed at randomisation, at 5 weeks to address the acute effects of withdrawal of donepezil, at 26 weeks and at 52 weeks. Finally, participants will be followed up every 26 weeks for 208 weeks by telephone interview to establish whether and on what date they have entered a care institution. There will be three Primary outcome measures. (1) Cognitive Function measured with the Severe Impairment Battery (SIB) (Panisset *et al* 1994). The SIB is a 51-item scale with scores ranging from 0 to 100 which has shown greater sensitivity to change in cognitive function than the Standardized MMSE in similarly affected populations of patients to the participants in DOMINO (Feldman *et al* 2001, Tariot *et al* 2004). (2) Activities of Daily Living measured with the Bristol Activities of Daily Living scale (BADLS) (Bucks *et al* 1996). The BADLS is well validated psychometrically and as a surrogate for estimating costs and scores deteriorate at a steady rate of 5 points per year in AD across a wide range of functional disability. (3) Cost-effectiveness measured as the combination of costs generated from the CSRI (Beecham *et al* 1992) and the SIB, BADLS, DEMQOL or utility measure generated from the EQ-5D. Secondary outcome measures. (1) Non-cognitive dementia symptoms measured with the NPI (Cummings *et al* 1994) and the Cohen-Mansfield Agitation Inventory (Cohen-Mansfield *et al* 1992). (2) Cognition with the MMSE (Folstein *et al* 1975). Although this is less sensitive to change than the SIB within the dementia severity range under study, the MMSE has been used in so many studies that its inclusion is important to allow comparisons with earlier trial data and

to increase generalisability of DOMINO's outcome data. (3) HRQL measured with the EQ-5D (Euroqol Group 1990) and the DEMQOL-Proxy. (4) Institutionalisation defined as permanent transition from living in an independent household to a care home, NHS continuing care unit or hospital and measured with questions taken from the CSRI over 4-year follow up. (5) Caregiver burden measured with the GHQ-12 (Goldberg *et al* 1988).

Analyses

The primary comparisons will be performed on an intention-to-treat basis. The results from the trial will be presented as comparative summary statistics (difference in response rates or means) with 95% confidence intervals. Primary outcomes SIB/BADLS: Depending on the distribution of the change in the SIB/BADLS from baseline over the study period (52 weeks), if appropriate, an analysis using linear mixed effects models with repeated measures will be performed, adjusting for baseline value and stratification covariates, plus other variables that the physicians consider of prognostic importance. We will formally assess the distribution of the change from baseline for evidence of departure from normality. If necessary, data will either be transformed or analysed using a non-parametric equivalent. Cost comparisons will be made between interventions to match the cognitive and ADL comparisons, with adjustments probably needed to adjust for non-normality of data (transformation or non-parametric test). For each hypothesis, relevant perspective and outcome, an incremental cost-effectiveness ratio will be computed using SIB, BADLS, DEMQOL and utility (from EQ5D) measures and compared with results from other studies where appropriate. Cost-effectiveness acceptability curves will be plotted for appropriate pairwise comparisons. For secondary outcomes we will formally assess the distributions of the changes in the continuous secondary outcome measures for evidence of departure from normality. In instances where such changes in outcome are not normally distributed, data will be transformed and analysed as detailed above, or tested using non-parametric equivalents. Service utilisation patterns, carer inputs and all associated costs will be calculated for each patient, based on data collected using a modified version of the CSRI, completed by a family carer or care professional. Unit costs to reflect long-run marginal opportunity costs will be attached using national figures where available. Each cost-effectiveness analysis will be conducted from the perspective of (a) the NHS and social services, and (b) society. The SIB, BADLS, DEMQOL and a utility measure generated from the EQ-5D will be used in turn in a series of cost-effectiveness analyses, the last of these to generate QALY measures (with societal weights). We will also examine the associations between EQ-5D, DEMQOL and SIB scores and changes therein, given uncertainty about the validity of EQ-5D measures as QALY generators within this population. Parallel work is needed to explore the utility generating properties of the DEMQOL but is beyond this study's scope. Cost-effectiveness acceptability curves will be plotted using bootstrap analyses to locate the findings of the economic evaluation in their wider decision-making context. Sensitivity analyses will also examine the consequences of key assumptions in the cost-effectiveness analysis. In addition, we will use a mathematical model, developed from the AD2000 database and using NPI and BADLS data, to estimate risks of institutionalisation in treatment groups over four years.

Application the new system to the MRC-DOMINO data set

We will work with the existing trial statisticians and economists to generate a further Analysis Strategy that will be applied to the trial data set. This will include the derivation of DEMQOL-nD and DEMQOL-Proxy-nD scales from the trial data, application of both the population and the patient/carer valuations and leading to the generation of cost effective analyses using the derived dementia QALY. This will allow us to go beyond the originally (and funded) plan of analysis which was to generate utility scores from the EQ-5D for the purposes of the cost-effectiveness analyses. This further work will be led by MK.

2.6.3 Further application and development

Testing the preference-based index

Using the participants in the patient and carer valuation survey, the psychometric properties of the item-reduced, preference-based DEMQOL-nD and DEMQOL-Proxy-nD will be evaluated using standard

psychometric methods (as described above and in *Table 1*) to evaluate acceptability, reliability (internal consistency and test–retest), validity (content, convergent, discriminant and known group differences)

Comparison of indices with the original sub-scale scores

The aim will be to test whether moving from the full DEMQOL sub-scale scores to the indices results in a significant loss of psychometric performance in terms of missing data, reliability, validity and responsiveness.

Comparison with EQ-5D

An important issue is whether the measures derived from DEMQOL and DEMQOL-Proxy perform any differently to the EQ-5D in terms of psychometric properties. The disease-specific measure must be demonstrated to be psychometrically superior in order to justify its further use rather than the generic EQ-5D. The size of any differences in health-state utility values found in the trials and the implications of this in terms of the incremental cost-effectiveness ratios will also be examined using the data generated by the economic evaluations in the HTA-SADD and MRC-DOMINO trials.

References

1. AD2000 study group (2004). Long-term donepezil treatment in 565 patients with Alzheimer's disease (AD2000). *Lancet*;**363**:2105–115.
2. Banerjee S (2007). Development and use of a disease-specific measure of HRQL. *Alzheimer's and Dementia* (2007) (epub ahead of print).
3. Banerjee S, Smith SC, Lamping DL, *et al.* Quality of life in dementia: more than just cognition (2006). *JNNP* **77**;146–48.
4. Banerjee S, Willis R, Matthews D, *et al* (2007). Improving the quality of dementia care – an evaluation of the Croydon Memory Service Model. *Int J Geriatr Psych* **22**(8):782–8.
5. Barton GR, Bankart J, Davis AC, Summerfield QA. (2004) Comparing utility scores before and after hearing-aid provision. *Appl Health Econ Health Policy* **3**(2):103–105.
6. Beecham JK, Knapp MRJ. Collecting and estimating costs (1992), *Measuring Mental Health Needs*. London: Gaskell.
7. Brazier JE, Harper R, Thomas K, Jones N, Underwood T (1998). Deriving a preference based single index measure from the SF-36. *J Clin Epidemiol* **51**(11):1115–129.
8. Brazier JE, Roberts J, Deverill M (2002). The estimation of a preference based measure of health from the SF-36, *J Health Econ* **21**(2):271–92.
9. Brazier JE, Yang Y, Tsuchiya A (2006). *Using rnak and discrete choice data to estimate health state utlity values: the case of the AQL-5D*. 69th Health Economists Study Group Meeting, University of York.
10. Brazier JE, Czoski-Murray C, Roberts J, Brown M, Symonds T, Kelleher C (2007). Estimation of a preference-based index from a condition specific measure: the King's Health Questionnaire. *Med Decis Making* (in press).
11. Brazier JE, Ratcliffe J, Tsuchiya A, Solomon J (2007). *Measuring and valuing health for economic evaluation*. Oxford: Oxford University Press.
12. Brazier JE, Yang Y, Tsuchiya A, Young T, Coyne K, Pretrillo J. (2006). *Estimating a preference-based index from the Over Active Bladder Questionnaire*, Poster presentation at the Annual ISPOR Meeting, May 20–24 2006, Washington, USA.

13. Brod M, Stewart AL, Sands L, *et al* (1999). Conceptualization and measurement of quality of life in dementia, *Gerontologist*;**39**(1):25–35.

14. Brooks, R and EuroQol Group (1996). EuroQol: the current state of play. *Health Policy;* **37**:53–72.

15. Bucks RS, Ashworth DL, Wilcock GK, Siegfried K (1996). Assessment of activities of daily living in dementia: development of the Bristol activities of daily living scale, *Age and Ageing*;**25**:113–20.

16. Cohen-Mansfield J, Marx MS, Werner P (1992). Agitation in elderly persons: an integrative report of findings in a nursing home. *Int Psychogeriatr* **4**(suppl 2):221–240.

17. Comas-Herrera A, Wittenberg R, Pickard L, Knapp M and MRC CFAS (2007). Cognitive impairment in older people: the implications for future demand for long-term care services and their costs. *Int J Geriatr Psychiatry* (in press).

18. Coucill W, Bryan S, Bentham P, *et al* (2001). ED-5D in patients with dementia. *Med Care* **39**:760–71.

19. Cummings J, Mega M, Gray K, *et al* (1994). The Neuropsychiatric Inventory: comprehensive assessment of psychopathology in dementia. *Neurology*;**44**:2308–14.

20. Department of Health (2001). *National Service Framework for Older People.* London: DH.

21. DH (2005). *Everybody's Business*. London: CSIP.

22. Espallargues M, Czoski-Murray C, Bansback N, Carlton J, Lewis G, Hughes L, *et al.* (2005) The impact of Age Related Macular Degeneration on health state utility values. *Investigative Ophthalmology and Visual Science*;**46**:4016–4023.

23. EuroQoL Group (1990). EuroQoL – a new facility for the measurement of health-related quality of life. *Health Policy*;**16**:199–208.

24. Feldman H, Gauthier S, Hecker J, Vellas B, Subbiah P, Whalen E and the Donepezil MSAD Study Investigators Group (2001). A 24-week randomized double-blind study of donepezil in moderate to severe Alzheimer's disease. *Neurology* **57**:613–20.

25. Folstein MF, Folstein SE, McHugh PR (1975). Mini Mental State. *J Psychiatric Res* **12**:189–198.

26. Goldberg D, Williams P (1988). *A user's guide to the General Health Questionnaire.* Windsor: NFER-NELSON.

27. Hilari K, Byng S, Lamping DL, Smith SC (2003). Stroke and Aphasia Quality of Life scale-39 (SAQOL-39): Evaluation of acceptability, reliability and validity. *Stroke*;**34**:1944–50.

28. Hobart JC, Riazi A, Lamping DL, Fitzpatrick R, Thompson AJ (2004). Improving the evaluation of therapeutic interventions in multiple sclerosis: development of a patient-based measure of outcome. *Health Technol Assess* **8**(9).

29. Hofman A, Rocca WA, Brayne C, *et al* (1991). The prevalence of dementia in Europe. *Int J Epid* **20**:736–48.

30. 3Jonsson L, Andreasen N, Kilander L, *et al* (2006). Patient- and proxy-reported utility in Alzheimer's Disease using the Euroqol. *Alzheimer's Disease and Associated Disorders* **20**:49–55.

31. Kharroubi SA, Brazier JE, Roberts J, O'Hagan A (2007). Modelling SF-6D health state preference data using a nonparametric Bayesian method. *J Health Econ*;**26**:597–612.

32. Knapp M, *et al* (2007). *Dementia UK: the full report*, London: Alzheimer's Society.

33. Liang, *et al.* (1990). Comparisons of five health status instruments for orthopaedic evaluation. *Med Care* **28**:632–42

34. Lamping DL, Hobart JC, Schroter S and Smith SC (2002b). Developing short-form outcome measures: Methodological approaches to item reduction. In D Lamping (Chair), *State of the art on item reduction: Identifying the good, bad and ugly items: Analysis and evaluation* (Symposium). International Society for Quality of Life Research, Orlando.

35. Lamping DL, Schroter S, Kurz X, Kahn SR, Abenhaim L (2003). Evaluating outcomes in chronic venous disorders of the leg: Development of a scientifically rigorous, patient-reported measure of symptoms and quality of life. *J Vasc Surg* **37**:410–19.

36. Lamping DL, Schroter S, Marquis P, Marrel A, Duprat-Lomon I, Sagnier PP (2002a). The Community-Acquired Pneumonia Symptom questionnaire: A new patient-based outcome measure to evaluate symptoms in patients with community-acquired pneumonia. *Chest*;**122**:920–929.

37. Launer LJ, Brayne C, Dartigues J-F, *et al.* Epilogue. (1992) *Neuroepidemiology*;**11**(suppl 1):119–121.

38. Logsdon RG, Gibbons LE, McCurry SM, *et al* (2002). Assessing quality of life in older adults with cognitive impairments. *Psychosomat Med*;**64**:510–19.

39. Loveman E, Green C, Kirby J, *et al* (2006). The clinical and cost-effectiveness of donepezil, rivastigmine, galantamine and memantine for AD. *Health Technol Assess*;**10**:1–160.

40. Lowin A, Knapp M, McCrone P (2001). Alzheimer's disease in the UK: comparative evidence on cost of illness and volume of research funding. *Int J Geriatr Psychiatry* **16**:1143–8.

41. Marra CA, Woolcott JC, Kopec JA, Shojania K, Offer R, Brazier JE, *et al* (2005). A comparison of generic, indirect utility measures (the HUI2, HUI3, SF-6D and the EQ-5D) and disease-specific instruments (the RAQoL and the HAQ) in rheumatoid arthritis. *Social Science and Medicine* **60**:1571–82.

42. McKhann G, Drachman D, Folstein M, *et al* (1984). Clinical diagnosis of Alzheimer's Disease: report of the NINCDS-ADRDA work group. *Neurology*;**34**:939–44.

43. Murray J, Schneider J, Banerjee S, *et al* (1999). EUROCARE a cross-national study of co-resident spouse carers for people with Alzheimer's dementia II: a qualitative analysis of the experience of caregiving. *Int J Geriatr Psych*;**14**:665–661.

44. Naglie G, Tomlinson G, Tansey C, *et al* (2006). Utility-based quality of life measures in Alzheimer's Disease. *Qual Life Res* **15**:631–43.

45. NAO (2007). *National Audit Office Report. Improving Services and Support for People with Dementia*. HC 604, Report by the Comptroller and Auditor General, Session 2006–2007, TSO: London.

46. NICE (2006). *Donepezil, galantamine, rivastigmine (review) and memantine for the treatment of Alzheimer's disease*. NICE: London.

47. Nunnally JC and Bernstein IH (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw Hill.

48. Panisset M, Roudier M, Saxton J (1994). Severe Impairment Battery, a neurological test for severely demented patients. *Arch Neurol* **51**:41–45.

49. Schneider J, Murray J, Banerjee S, *et al* (1999). EUROCARE: a cross national study of co-resident spouse carers for people with Alzheimer's Disease. I – factors associated with carer burden. *Int J Geriatr* Psych;**14**:665–661.

50. Scientific Advisory Committee of the Medical Outcomes Trust (2002). Assessing health status and quality-of-life instruments: Attributes and review criteria. *Qual Life Res*;**11**:193–205.

51. Smith SC, Lamping DL, Banerjee S, Harwood RH, Foley B, Smith P, *et al* (2007). Development of a new measure of health-related quality of life for people with dementia: DEMQOL. *Psychol Med*;**37**:737–46.

52. Smith SC, Lamping DL, Banerjee S, Harwood R, Foley B, Smith P, *et al* (2005). Measurement of health-related quality of life for people with dementia: development of a new instrument (DEMQOL) and an evaluation of current methodology. *Health Technol Assess*;**9**(10).

53. Streiner DL and Norman GR (2003). *Health measurement scales: A practical guide to their development and use* (3rd ed.). Oxford: Oxford University Press.

54. Takeda A, Loveman E, Clegg A, *et al* (2006). A systematic review of the clinical effectiveness of donepezil, rivastigmine and galantamine on cognition, quality of life and adverse events in Alzheimer's disease. *Int J Ger Psych*;**21**:17–28.

55. Tariot PN, Farlow MR, Grossberg GT, Graham SM, McDonald S, Gergel (2004). Memantine treatment in patients with moderate to severe Alzheimer disease already receiving donepezil. *JAMA*;**291**:317–24

56. Vogel A, Mortensen EL, Hasselbalch CG (2006). Patient versus informant reported quality of life in the earliest phases of Alzheimer's Disease, *Int J Geriatr Psychiatr* **21**:1132–38.

57. Whitehouse PJ (2000). Harmonization of Dementia Drug Guidelines: a report of the International Working Group for the Harmonization for Dementia Drug Guidelines. *Alzheimer Dis Assoc Disord*;**14** Suppl 1:S119–2.

58. Woods RT, Thorgrimsen L, Spector A, Royan L, Orrell M (2006). Improved quality of life and cognitive stimulation in dementia. *Aging Ment* Health;**10**:219–26.

59. Yang Y, Tsuchiya A, Brazier J, Young T (2006). *Deriving a preference-based measure for health from the AQLQ*. January HESG, City University,

60. Young T, Yang Y, Brazier J, Tsuchiya A (2005). *Using Rasch analysis to aid the construction of preference based measures from existing quality of life instruments*. Paper presented at the Health Economics Study Group Meeting, University of Newcastle.

61. Young T, Yang Y, Brazier J, Tsuchiya A (2007). *The Use of Rasch Analysis as a Tool in the Construction of a Preference Based Measure: The Case of AQLQ*. HEDS Discussion Paper (07/01) URL: www.shef.ac.uk/content /1/c6/01/87/47/ HEDS%20DP%200701.pdf.

TABLE 1 DEMQOL and DEMQOL-Proxy: Psychometric Tests and Criteria

Psychometric Property	Definition/Test	Criteria for Acceptability
1. Item analysis/reduction	Identify items for possible elimination due to weak psychometric performance; assessed on the basis of: i) unrotated principal component factor analysis to determine whether all items are measuring a single factor; and ii) item analyses for all items	*Principal component factor analysis* • All items should load on the first unrotated factor > 0.30 *Item analyses (applied to all items)* • Missing data < 5% • No item redundancy (inter-item correlations ≤0.75) • Item–total correlations >0.25 • Maximum endorsement frequencies ≤80% (i.e. the proportion of respondents who endorse each response category), including floor/ceiling effects < 80% (i.e. response categories with high endorsement rates at the bottom/top ends of the scale, respectively) • Aggregate adjacent endorsement frequencies ≥10% (Liang *et al* 1990)
2. Acceptability	The quality of data; assessed by completeness of data and score distributions	• Missing data for summary scores < 5% • Even distribution of endorsement frequencies across response categories • Floor/ceiling effects for summary scores < 10%
3. Reliability		
3.1 Internal consistency	The extent to which items comprising a scale measure the same construct (e.g. homogeneity of the scale); assessed by Cronbach's alpha (160) and item–total correlations	• Cronbach's alphas for summary scores ≥0.70 (Streiner and Norman 2003) • Item–total correlations ≥0.20 (Streiner and Norman 2003)
3.2 Test–retest reliability	The stability of a measuring instrument; assessed by administering the instrument to respondents on two different occasions and examining the correlation between test and retest scores	• Test–retest reliability correlations for summary scores ≥0.70 (Scientific Advisory Committee of the Medical Outcomes Trust 2002)
3.3 Inter-rater reliability	Agreement between independent raters/observers; assessed by intra-class correlations (ICC)	• ICC ≥0.70 (Scientific Advisory Committee of the Medical Outcomes Trust 2002)
3.4 Parallel (alternate) forms reliability	Agreement between two or more parallel/alternate forms or different versions of the same measure (e.g. form A/B, short/long form, etc.) that indicates that they can be used interchangeably; assessed on the basis of correlations between parallel/alternate forms of a measure	• High correlation between parallel/alternate forms of the measure (e.g. between long and short form)
4. Validity		
4.1 Content validity	The extent to which the content of a scale is representative of the conceptual domain it is intended to cover; assessed qualitatively during the questionnaire development stage through pre-testing with patients, expert opinion and literature review	• Qualitative evidence from pre-testing with patients, expert opinion and literature review that items in the scale are representative of the construct being measured

Psychometric Property	Definition/Test	Criteria for Acceptability
4.2 Criterion-related validity		
4.2.1 Concurrent validity	Evidence that the scale predicts a gold standard criterion that is measured at the same time; assessed on the basis of correlations between the scale and the criterion measure	• High correlation between the scale and the criterion measure
4.2.1 Predictive validity	Evidence that the scale predicts a gold standard criterion that is measured in the future; assessed on the basis of correlations between the scale and the criterion measure	• High correlation between the scale and the criterion measure
4.3 Construct validity		
4.3.1 Within-scale analyses	Evidence that a single entity (construct) is being measured and that items can be combined to form a summary score; assessed on the basis of evidence of good internal consistency and correlations between scale scores (which purport to measure related aspects of the construct)	• Internal consistency (Cronbach's alpha) ≥0.70 • Moderate to high correlations between scale scores
4.3.2 Analyses against external criteria		
4.3.2.1 Convergent validity	Evidence that the scale is correlated with other measures of the same or similar constructs; assessed on the basis of correlations between the measure and other similar measures	• Correlations are expected to vary according to the degree of similarity between the constructs that are being measured by each instrument. Specific hypotheses are formulated and predictions tested on the basis of correlations
4.3.2.2 Discriminant validity	Evidence that the scale is not correlated with measures of different constructs; assessed on the basis of correlations with measures of different constructs	• Low correlations between the instrument and measures of different constructs
4.3.2.3 Known groups differences	The ability of a scale to differentiate known groups; assessed by comparing scores for subgroups who are expected to differ on the construct being measured	• Significant differences between known groups or difference of expected magnitude
5. Responsiveness	Tthe ability of a scale to detect clinically important change over time; assessed by comparing scores before and after an intervention of known efficacy (on the basis of various methods including *t*-tests (161), effect sizes (162, 163), standardised response means (164), or responsiveness statistics (165)	• Significant differences between known groups or difference of expected magnitude

Smith *et al*. 2005; adapted from Lamping *et al*., 2002a, 2002b, 2003; Hilari *et al*., 2003.

Appendix 3 DEMQOL and DEMQOL-Proxy

DEMQOL

Instructions: Read each of the following questions (in bold) verbatim and show the respondent the response card.

I would like to ask you about your life. There are no right or wrong answers. Just give the answer that best describes how you have felt in the last week. Don't worry if some questions appear not to apply to you. We have to ask the same questions of everybody.

Before we start we'll do a practise question; that's one that doesn't count. *(Show the response card and ask respondent to say or point to the answer)* **In the last week, how much have you enjoyed watching television?**

a lot **quite a bit** **a little** **not at all**

Follow up with a prompt question: **Why is that?** *or* **Tell me a bit more about that.**

DOI: 10.3310/hta17050

For all of the questions I'm going to ask you, I want you to think about the last week.

First I'm going to ask about <u>your feelings</u>. In the last week, have you felt…….

1. **cheerful**? **	☐ a lot	☐ quite a bit	☐ a little	☐ not at all
2. **worried or anxious?**	☐ a lot	☐ quite a bit	☐ a little	☐ not at all
3. **that you are enjoying life? **	☐ a lot	☐ quite a bit	☐ a little	☐ not at all
4. **frustrated?**	☐ a lot	☐ quite a bit	☐ a little	☐ not at all
5. **confident? **	☐ a lot	☐ quite a bit	☐ a little	☐ not at all
6. **full of energy? **	☐ a lot	☐ quite a bit	☐ a little	☐ not at all
7. **sad?**	☐ a lot	☐ quite a bit	☐ a little	☐ not at all
8. **lonely?**	☐ a lot	☐ quite a bit	☐ a little	☐ not at all
9. **distressed?**	☐ a lot	☐ quite a bit	☐ a little	☐ not at all
10. **lively? **	☐ a lot	☐ quite a bit	☐ a little	☐ not at all
11. **irritable?**	☐ a lot	☐ quite a bit	☐ a little	☐ not at all
12. **fed-up?**	☐ a lot	☐ quite a bit	☐ a little	☐ not at all
13. **that there are things that you wanted to do but couldn't?**	☐ a lot	☐ quite a bit	☐ a little	☐ not at all

Next, I'm going to ask you about <u>your memory</u>. In the last week, how worried have you been about…….

14. **forgetting things that happened recently?**	☐ a lot	☐ quite a bit	☐ a little	☐ not at all
15. **forgetting who people are?**	☐ a lot	☐ quite a bit	☐ a little	☐ not at all
16. **forgetting what day it is?**	☐ a lot	☐ quite a bit	☐ a little	☐ not at all
17. **your thoughts being muddled?**	☐ a lot	☐ quite a bit	☐ a little	☐ not at all

18. **difficulty making decisions?**	☐ a lot	☐ quite a bit	☐ a little	☐ not at all
19. **poor concentration?**	☐ a lot	☐ quite a bit	☐ a little	☐ not at all

Now, I'm going to ask you about your everyday life. In the last week, how worried have you been about…….

20. **not having enough company?**	☐ a lot	☐ quite a bit	☐ a little	☐ not at all
21. **how you get on with people close to you?**	☐ a lot	☐ quite a bit	☐ a little	☐ not at all
22. **getting the affection that you want?**	☐ a lot	☐ quite a bit	☐ a little	☐ not at all
23. **people not listening to you?**	☐ a lot	☐ quite a bit	☐ a little	☐ not at all
24. **making yourself understood?**	☐ a lot	☐ quite a bit	☐ a little	☐ not at all
25. **getting help when you need it?**	☐ a lot	☐ quite a bit	☐ a little	☐ not at all
26. **getting to the toilet in time?**	☐ a lot	☐ quite a bit	☐ a little	☐ not at all
27. **how you feel in yourself?**	☐ a lot	☐ quite a bit	☐ a little	☐ not at all
28. **your health overall?**	☐ a lot	☐ quite a bit	☐ a little	☐ not at all

We've already talked about lots of things: your feelings, memory and everyday life. Thinking about all of these things in the last rate…….

29. **your quality of life overall? ****	☐ very good	☐ good	☐ fair	☐ poor

****** items that need to be reversed before scoring

DOI: 10.3310/hta17050

DEMQOL-Proxy

Instructions: *Read each of the following questions (in bold) verbatim and show the respondent the response card.*

I would like to ask you about _____ *(your relative's)* **life, as you are the person who knows him/her best. There are no right or wrong answers. Just give the answer that best describes how _____** *(your relative)* **has felt in the last week. If possible try and give the answer that you think _____** *(your relative)* **would give. Don't worry if some questions appear not to apply to _____** *(your relative).* **We have to ask the same questions of everybody.**

Before we start we'll do a practise question; that's one that doesn't count. *(Show the response card and ask respondent to say or point to the answer)*. **In the last week how much has _____** *(your relative)* **enjoyed watching television?**

a lot **quite a bit** **a little** **not at all**

Follow up with a prompt question: **Why is that?** *or* **Tell me a bit more about that.**

For all of the questions I'm going to ask you, I want you to think about the last week.

First I'm going to ask you about _____ *(your relative's)* **feelings. In the last week, would you say that** _____ *(your relative)* **has felt…….**

1. **cheerful? ****	☐ a lot	☐ quite a bit	☐ a little	☐ not at all
2. **worried or anxious?**	☐ a lot	☐ quite a bit	☐ a little	☐ not at all
3. **frustrated?**	☐ a lot	☐ quite a bit	☐ a little	☐ not at all
4. **full of energy? ****	☐ a lot	☐ quite a bit	☐ a little	☐ not at all
5. **sad?**	☐ a lot	☐ quite a bit	☐ a little	☐ not at all
6. **content? ****	☐ a lot	☐ quite a bit	☐ a little	☐ not at all
7. **distressed?**	☐ a lot	☐ quite a bit	☐ a little	☐ not at all
8. **lively? ****	☐ a lot	☐ quite a bit	☐ a little	☐ not at all
9. **irritable?**	☐ a lot	☐ quite a bit	☐ a little	☐ not at all
10. **fed-up**	☐ a lot	☐ quite a bit	☐ a little	☐ not at all
11. **that he/she has things to look forward to? ****	☐ a lot	☐ quite a bit	☐ a little	☐ not at all

Next, I'm going to ask you about _____ *(your relative's)* **memory. In the last week, how worried would you say** _____ *(your relative)* **has been about……**

12. **his/her memory in general?**	☐ a lot	☐ quite a bit	☐ a little	☐ not at all
13. **forgetting things that happened a long time ago?**	☐ a lot	☐ quite a bit	☐ a little	☐ not at all

14. **forgetting things that happened recently?**	☐ a lot	☐ quite a bit	☐ a little	☐ not at all
15. **forgetting people's names?**	☐ a lot	☐ quite a bit	☐ a little	☐ not at all
16. **forgetting where he/she is?**	☐ a lot	☐ quite a bit	☐ a little	☐ not at all
17. **forgetting what day it is?**	☐ a lot	☐ quite a bit	☐ a little	☐ not at all
18. **his/her thoughts being muddled?**	☐ a lot	☐ quite a bit	☐ a little	☐ not at all
19. **difficulty making decisions?**	☐ a lot	☐ quite a bit	☐ a little	☐ not at all
20. **making him/herself understood?**	☐ a lot	☐ quite a bit	☐ a little	☐ not at all

Now, I'm going to ask about _____ *(your relative's)* everyday life. In the last week, how worried would you say _____ *(your relative)* has been about…….

21. **keeping him/herself clean (eg washing and bathing)?**	☐ a lot	☐ quite a bit	☐ a little	☐ not at all
22. **keeping him/herself looking nice?**	☐ a lot	☐ quite a bit	☐ a little	☐ not at all
23. **getting what he/she wants from the shops?**	☐ a lot	☐ quite a bit	☐ a little	☐ not at all
24. **using money to pay for things?**	☐ a lot	☐ quite a bit	☐ a little	☐ not at all
25. **looking after his/her finances?**	☐ a lot	☐ quite a bit	☐ a little	☐ not at all
26. **things taking longer than they used to?**	☐ a lot	☐ quite a bit	☐ a little	☐ not at all
27. **getting in touch with people?**	☐ a lot	☐ quite a bit	☐ a little	☐ not at all
28. **not having enough company?**	☐ a lot	☐ quite a bit	☐ a little	☐ not at all

29. **not being able to help other people?**	☐ a lot	☐ quite a bit	☐ a little	☐ not at all
30. **not playing a useful part in things?**	☐ a lot	☐ quite a bit	☐ a little	☐ not at all
31. **his/her physical health?**	☐ a lot	☐ quite a bit	☐ a little	☐ not at all

We've already talked about lots of things: _____ *(your relative's)* **feelings, memory and everyday life. Thinking about all of these things in the last week, how would you say** _____ *(your relative)* **would rate……..**

32. his/her quality of life overall? **	☐ very good	☐ good	☐ fair	☐ poor

** items that need to be reversed before scoring

EME
HS&DR
HTA
PGfAR
PHR

Part of the NIHR Journals Library

Published by the NIHR Journals Library