THE INCORPORATION OF HEALTH BENEFITS IN COST UTILITY ANALYSIS USING THE EQ-5D

REPORT BY THE DECISION SUPPORT UNIT

Allan Wailoo, Sarah Davis, Jonathan Tosh
School of Health and Related Research, University of Sheffield

15 November 2010

Acknowledgements:
We thank Aki Tsuchiya, John Brazier and Ben van Hout for helpful comments and advice and Jenny Dunn for formatting the report. The content of this report is the responsibility of the authors alone.

SUMMARY

Economic evaluation is an important component for decision making, underpinning much of the guidance issued by NICE. Typically these economic evaluations have taken the form of cost-utility analyses where health benefits are expressed in terms of Quality Adjusted Life Years (QALYs). One of the most widely used preference based instruments for the assessment of Health Related Quality of Life (HRQoL) that can be used to generate QALYs is the EQ-5D. This is stated to be the preferred instrument for NICE.

The EQ-5D has been the subject of criticisms levelled at NICE. This report considers these criticisms of EQ-5D. It aims to identify the types of claims that have been made about EQ-5D and identify the empirical evidence to support such claims. This is used to inform a series of case studies in different disease areas, where evidence of the performance of EQ-5D is systematically identified and reviewed.

The report identifies many claims that have been made as part of individual technology appraisals and in evidence submitted to the Kennedy review into the value of innovation. Many of these claims relate to the QALY as a measure of outcome per se or the decision rule of QALY maximisation. There were few examples of claims that EQ-5D is inappropriate as a measurement tool within these general frameworks. Where such claims were identified they were rarely supported by empirical evidence. These claims can be broadly categorised as relating to situations where a specific relevant dimension of health is not directly included in the EQ-5D instrument, such as fatigue or sensory impairment, or where the disease course is characterised by flares of unpredictable symptom severity. Several claims were made regarding inappropriateness in broad disease areas such as cancer and mental health.

We conducted case study reviews in the areas of rheumatoid arthritis, asthma and incontinence and refer to a separately funded report on visual disorders.

In general terms we found that there were several studies that suggest EQ-5D is less responsive or sensitive than disease specific outcome measures. This was the case for both preference and non preference based outcomes. Other generic preference based measures do not seem to systematically perform differently to EQ-5D. Where an alternative generic instrument includes a specific dimension of relevance to the disorder, such as the HUI3 in visual disorders, then it is more sensitive to changes than EQ-5D. There are also instances where EQ-5D may be a more appropriate instrument than some disease specific outcome measures.

These data inform assessments derived from psychometrics. Numerous cautions must be considered when making such assessments. There are no definitive tests in this situation. The data provide circumstantial evidence that must be combined with intuition and judgement in order to reach conclusions about the appropriateness or otherwise of EQ-5D or any other instrument. There is no gold standard and assessments are a question of degree. In particular, the requirement of the Institute to make consistent decisions across a broad range of diseases, patients and technologies must be considered.

The case studies also highlight the requirement to review a wide range of literature in assessing EQ-5D. Studies that contribute evidence usually are not designed with an

assessment of EQ-5D in mind. Detailed, critical examination of the studies is required to assess their relevance.

Several developments to the EQ-5D are in development including a 5 level variant and the use of “bolt-ons”. The former may help to overcome problems where sample sizes are insufficient to detect changes in the standard 3-level EQ-5D. The relevance of the latter to NICE is dependent on how a number of other considerations are resolved. For example, what is required of health state valuation methods in order to achieve consistency in decision making and what is the appropriate conceptual nature of health.

CONTENTS

Tables

ABBREVIATIONS

ACD	Appraisal Consultation Document
ACQ	Asthma Control Questionnaire
ACS	Asthma Control Scale
ADSS	Asthma Disease Severity Scale
AQLQ	Asthma Quality of Life Questionnaire
AQLQ(S)	Asthma Quality of Life Questionnaire (standardized version).
AQL-5D	Asthma Quality of Life- 5 Dimension
ASUI	Asthma Symptom Utility Index
ATAQ	Asthma Therapy Assessment Questionnaire
BFLUTS	Bristol Female Lower Urinary Tract Symptoms Questionnaire
CQLQ	Cough Quality of Life Questionnaire.
DAS28	Disease Activity Score in rheumatoid arthritis
DIS	Detrusor Instability Scores
ERG	Evidence Review Group
FAD	Final Appraisal Determination
FEV_1	Forced Expiratory Volume in the first second
FAI	Frenchay Activities Index
GETE	Global Evaluation of Treatment Effectiveness,
GINA	Global Initiative for Asthma Classification,
HRQoL	Health Related Quality of Life
HUI3	Health Utilities Index Mark 3
ICER	Incremental Cost Effectiveness Ratio
ICSQol	International Continence Society – benign prostatic hyperplasia study Quality of Life Instrument
IIQ-7	Incontinence Impact Questionnaire-short form
I-PSS	International Prostate Symptom Score
I-QOL	Incontinence specific Quality of life Questionnaire
ITT	Intention To Treat
KHQ	King's Health Questionnaire
LCQ	Leicester Cough Questionnaire.
MVH	Measurement and Valuation of Health
MTA	Multiple Technology Appraisal
NASQ	Newcastle Asthma Symptoms Questionnaire
NHP	Nottingham Health Profile
NICE	National Institute for Health and Clinical Excellence
PGI	Patient Generated Index
PP	Per Protocol.
QALY	Quality Adjusted Life Year
SF-36	Medical outcomes study 36-Item Short-Form Health Survey
SF-6D	An instrument for measuring utility in economic evaluations
SGRQ	St Georges Respiratory Questionnaire
STA	Single Technology Appraisal
SSI	Symptom Severity Index
S/UIQ	Stress and Urge Incontinence Questionnaire
TA	Technology Appraisal
TACQOL	Netherlands Organisation for Applied Scientific Research Academic Medical Centre Questionnaires for Children's Health-Related Quality of Life
TTO	Time Trade Off
UDI-6	Urogenital Distress Inventory-short form
UISS	Urinary Incontinence Severity Score
VAS	Visual Analogue Scale

1. INTRODUCTION

Decision making committees across the range of activities undertaken by the National Institute for Health and Clinical Excellence (NICE) draw on cost effectiveness analyses as part of their considerations. Typically these economic evaluations have taken the form of cost-utility studies where health benefits are expressed in terms of Quality Adjusted Life Years (QALYs). This approach facilitates comparisons to be made across different interventions, patients and disease areas and is therefore consistent with the remit of the Institute. This approach is also consistent with the majority of health economic evaluation work undertaken outside the NICE setting.

Estimation of the QALY benefits of an intervention in turn requires estimation of the changes in health status of patients, the duration of those changes and valuation of the health changes. There are several methods by which these measurement and valuation components can be obtained. However, the most widespread approach in practice is for patients to indicate their health status using a generic classification tool for which a set of corresponding values exist from the general population. Several options exist and, up until the publication of the Institute's most recent "Guide to the Methods of Technology Appraisal"[1], were each considered acceptable provided certain characteristics were met. These were that the valuations came from the UK general public using a choice based method. Whilst a preference for the use of the EQ-5D instrument was indicated in the 2004 Methods Guide, this was strengthened considerably in the 2008 Guide. The content of the Methods guide is relevant not just for Technology Appraisals, but these methods are widely implemented across the range of the Institute's activities and also inform economic evaluation methods more generally.

The widespread use of EQ-5D specifically, and cost per QALY analysis more generally, has been the subject of criticisms levelled at NICE over the last ten years. It is claimed that this approach fails to adequately capture all issues that are important to patients, carers and society. These criticisms were repeated in submissions to the recent review of how "innovation" is valued in the technology appraisal process conducted by Sir Ian Kennedy. Recommendations to NICE made as part of that study included one to conduct "research to determine whether the instruments used to calculate QALYs and capture health benefits are entirely appropriate to NICE's needs and whether they are applied properly and consistently" (Recommendation 4, Kennedy 2009[2]).

The Kennedy report was clear that appraisals should continue to be based on costs and health benefits of technologies, and the cost per QALY approach was not itself to be revised. To date no systematic consideration of the claims made regarding the inadequacy of the NICE approach has been performed. The purpose of this report is to provide such a review in order to inform future research and development proposals.

This report specifically aims to:

a) Identify the types of claims that have been made by various NICE stakeholders as to the appropriateness of EQ-5D
b) Assess whether these claims can be considered legitimately concerned with the EQ-5D instrument itself as opposed to other elements of the NICE

appraisals methods, such as broader critiques of the cost per QALY approach or the health service perspective

c) Explore the empirical evidence relating to those claims
d) Review the evidence relating to the performance of EQ-5D in a range of diseases, interventions and patient populations.
e) To identify whether any potential ways of resolving deficiencies have been proposed and what potential impact these may have for NICE

In the following section we describe the EQ-5D instrument, its development and how this relates to the Technology Appraisals Methods Guide (2008[1]). We describe the types of data that have been proposed as a means of investigating the appropriateness of EQ-5D in any specific situation. Section three reviews the types of claims that have been made relating to the deficiencies of the NICE approach to valuing health benefits. We consider claims submitted to the Kennedy Study and previous NICE Technology Appraisals. In section 4, we describe how we use this review to select a number of case studies for de novo systematic reviews. Section 5 describes these reviews in the areas of asthma, incontinence and rheumatoid arthritis. It also draws on findings from a separate MRC funded study[3] in the area of visual disorders. Section 6 discusses the findings from the report including the combined results from these reviews and the implications of proposed developments to the EQ-5D. Section 7 concludes.

2. THE EQ-5D AND ECONOMIC EVALUATION AT NICE

2.1. THE EQ-5D

The EQ-5D™ (ref EuroQol Group) is a standardised instrument intended to measure and value health outcomes across a wide range of diseases and treatments. It is therefore described as a generic rather than a condition specific instrument. It consists of two main components. First, a classification or descriptive system that covers five health domains: mobility, self-care, usual activities, pain/discomfort and anxiety/depression. Each domain has three levels: no problems, some problems, severe problems. There are therefore 243 health states that can be described in what is generally accepted as a simple approach to describing health. Second, a single valuation (EQ-5D index or tariff) is provided for each particular health state in the descriptive system. In the UK, valuations were obtained for a subset of 42 of these states, using the time-trade-off (TTO) method in interviews with a sample of 3395 members of the general public in the UK in 1993. A linear regression was then used to predict the valuations for all states except full health[4].

Those involved in the development of the EQ-5D instrument were motivated by their recognition of the need for a "global" instrument, "the descriptive content of which is neither condition–specific nor treatment-specific" (Williams, 1995, p.1[5]). The dimensions "were intended to be relevant to patients across the spectrum of health care, as well as to members of the general population." (Gudex, 1995, p.19[6]). They deliberately do not mention specific diseases, diagnoses or symptoms, although the latter is not such a clear distinction. Depression for example can be seen as a symptom or diagnosis but it was felt by the EuroQol group that dimensions such as this must also have relevance to a wide range of patients and the general population[6].

In his description of the Measurement and Valuation of Health (MVH) programme of work that culminated in the development of the EQ-5D valuation set, Williams describes a set of qualitative work that was used to inform the selection of EQ-5D as the most appropriate instrument from a series of descriptive systems appraised. Two important points are raised from this description. The first is that whilst individuals initially responded to questions about their ideas of health or ill-health with symptom based descriptions, this changed to descriptions of functional capacity, feelings and general fitness on reflection and further consideration. Second, it is apparent that the EQ-5D was selected because it was considered to cover the majority of items that individuals felt were important, and included those that were of most importance and relevance[5]. The EQ-5D developers recognised the need for compromise between a truly comprehensive instrument and one that is feasible and practical, including in relation to the aim of producing a single valuation tariff for all described states.

The EQ-5D is thus a generic instrument considered to cover the "salient" or core features of health which should always be of interest but was never intended to measure "all kinds of HRQoL in sufficient detail for all purposes." (Williams, 1993, p.6[7]). Indeed it has always been the case that recommendations from the EuroQol group included that a condition specific measure be administered alongside EQ-5D. To what extent this is useful in the context of generating QALYs for decision making informed economic evaluation is debateable. Dowie (2002[8]) for example provides a detailed critique of this approach.

2.2. THE NICE 2008 METHODS GUIDE

The current version of the NICE Methods Guide (2008[1]) states that the EQ-5D may not be appropriate in all circumstances. It goes on:

"If the EQ-5D is considered inappropriate, empirical evidence should be provided on why the properties of the EQ-5D are not suitable for the particular patient population. These properties may include the content validity, construct validity, responsiveness and reliability of EQ-5D."

The guide therefore focuses on psychometric properties. In addition, it refers specifically to the case of children but provides no further detail on where EQ-5D may not be considered appropriate.

2.3. IDENTIFYING WHERE EQ-5D IS "INAPPROPRIATE"

The EQ-5D forms part of the NICE reference case because this is felt to provide a consistent approach to assessing the value of health benefits, irrespective of the characteristics of the therapeutic area, the intervention or the patient group (NICE 2008, p. 38, 5.4.4[1]). Other generic measures appropriate for use in cost utility analysis do not provide comparable results – they are based on different health classification systems, different sample populations used to value the health states, different methods of valuation and different statistical approaches to the estimation of values, *inter alia*.

Consistency of methods is clearly an important component given NICE's remit. Indeed, the same rationale applies for using the cost per QALY approach and an ICER

threshold more generally. The importance of consistency in decision-making is referred to directly in Sir Ian Kennedy's recommendation to the Institute. It is clear that the appraisal of any instrument(s) for assessing health benefit in the NICE setting must be mindful of this facet of good decision making.

This requirement is of particular importance to Technology Appraisals with the introduction of the Single Technology Appraisal (STA) Process. In this process, it is only the manufacturer of the health technology in question that submits an economic evaluation to NICE and, at the same time, may hold all or a large part of the evidence in relation to utility values in the relevant patient group. There are obvious incentives for sponsoring manufacturers to select the approach to estimating QALYs that is most favourable to a product and NICE methods and processes should be mindful of those incentives.

However, it is not necessarily the case that consistency of methods is a means of achieving consistency of assessment. Brazier and Tsuchiya[9] present a view that argues consistency is achieved by using a common numeraire (e.g. a year in full health) and common approaches to valuation. According to this view, different descriptive systems are entirely warranted and retain consistency provided that in all situations the descriptive system reflects the issues of importance to patients.

A number of conditions must be fulfilled in order for consistency to be maintained without using the same descriptive system[10]. Most important is the requirement that the descriptive system must capture all, or at least the same amount, of relevant health issues in each situation. So, for example, if a condition specific measure were to be used in one situation with EQ-5D used in others, then it must be the case that coverage is equal in all these situations for consistency to be feasible. There are several reasons why this may not be the case. Condition specifics may fail to identify adverse events and comorbid conditions for example[9].

There may also be more fundamental reasons why it is not feasible to achieve equal coverage of descriptive systems across different conditions, which stem from the observation that "appropriateness" of any descriptive system is not a binary concept but rather is a question of degree. Even in situations where a generic instrument such as EQ-5D apparently "works" this is no proof at all that the EQ-5D really captures all or even most of a treatment effect, just that it captures some, unknown, element of it. If this is the case then it is not realistic to achieve similar coverage of descriptive systems whether using the same system universally or not.

If instead it is felt that consistency requires the same valuation and descriptive system, and it is felt that consistency is a sufficiently important part of a fair decision making process, then the implications are quite different. Information that claims to demonstrate "inappropriateness" of the preferred generic instrument (e.g. EQ-5D) should motivate the development of better, more appropriate generic instruments rather than a departure from the preferred approach in isolated cases.

In order to reach a definition of "inappropriate" one must consider what it is that is intended to be reflected in health benefit valuations. Elements of the intended conceptual properties of health state utilities can be deduced from the selection of cost utility analysis as the preferred approach to economic evaluation, other elements can

be deduced from the selection of EQ-5D itself as the preferred measure and further elements are clear in the NICE methods guide.

Firstly, health state utility values are intended to reflect the valuation, on a common scale anchored at 1 for full health for one year and zero for death, of the general public of being in a particular health state. Patients, or in some situations their proxies, should provide the "measurement" of changes in health related quality of life. This distinction between patients/measurement and general population/valuation is an important one as it has implications for how differences in valuations obtained via different classification systems may be interpreted.

Secondly, the EQ-5D classification system operates via a description of health that operates via five general dimensions of health. These are a mixture of functions (mobility, self-care and usual activities) and general symptomatic type issues (pain/discomfort, anxiety/depression), though as described above these were intended to be of wide relevance and focus on the impact of health on quality of life. This contrasts with alternative measures that are focussed to a greater extent, though not exclusively, on symptoms and disability. The Health Utilities Index Mark 3 (HUI3) is an example of such a tool. This also contrasts with approaches which focus on a capabilities approach such as those being developed in relation to older people[11] or on overall wellbeing[12].

In psychometrics, the concern is often to consider how a summary measure performs compared to some objective means of reporting the construct it is designed to measure. However, the assessment of health state utility instruments is a fundamentally different setting since there is no such gold standard against which any measure can be evaluated. It is not therefore possible for any such test to provide conclusive evidence that a health utility measure such as EQ-5D is appropriate or not. These tests can highlight differences between measures but judgements must also be made for stronger conclusions to be drawn[13;14].

Williams (1993)[7] recognised this:
"In general, establishing "validity" requires the investigator to address the question " does your measure measure what it purports to measure?". But since there is no "gold standard" for the measurement of health-related quality of life, this seems an unanswerable question. So what people fall back on instead are appeals to plausibility, for instance: testing whether the measure contains the kind of elements that we would expect such a measure to have; whether it goes up when we would expect it to go up and down when we would expect it to go down; and so on. These are all very subjective notions, and ultimately rely heavily on intuition and professional judgement" (Williams (1993) p.4[7])

"My own personal view is that searching for "validity" in this field, at this stage in the history of HRQoL measurement, is like chasing a will o' the wisp, and probably equally unproductive." (Williams (1993) p.5[7])

Dowie (2002[8]) discusses the relevance of condition specific measures (CSMs) versus generic ones (GENs) to decision making. He presents a similar dismissal of the value of comparisons that purport to demonstrate that condition specific measures are more

responsive or sensitive than generic measures (such as EQ-5D) which are "now almost endemic in the literature" (Dowie (2002) p.5[8]):
"But how, if at all, is the comparative sensitivity of a GEN and a CSM to be established? The answer is surely not by putting the changes in CSMs and GENs alongside each other and seeing which is bigger. If some action produces an 10% effect in a CSM but does not showup on a GEN (or produces, say, only a 1% effect) nothing at all follows from those two facts, other than that the former number is indeed bigger than the latter. There can be no more warrant for saying that the GEN is less sensitive or less responsive than the CSM than for saying that the distances between London and Sheffield and London and Exeter are longer when measured on a 1:25 000 map, than when measured on a 1:250 000 one." (Dowie, 2002, p.5[8])

Brazier and Deverill (1999)[14] provide a comprehensive account of the psychometric approaches to assessing instruments and relate this to the use of quality of life instruments in economic evaluation and economic theory more broadly. They argue that since measures such as EQ-5D are intended to reflect health state preferences, the true gold standard test against which they may be judged is revealed preference, that is, the decisions individuals actually make. However, unlike other economic markets, health care is characterised by features such as uncertainty that make it difficult to infer preferences from the actual choices of patients.

2.3.1. Acceptability and feasibility.

Acceptability and feasibility can together be thought of as practical considerations in the selection of quality of life instruments. Acceptability relates to the extent to which the instrument is acceptable to study subjects and is one determinant of both the response rate and the quality of the responses. Considerations of feasibility focus on the burden to researchers in administering, collecting and processing the instrument. These practical considerations may result in EQ-5D being considered inappropriate, for example, if response rates were particularly low. However, these issues are likely to have implications for study design rather than the selection of an alternative instrument altogether. For example, face-to-face administration or completion of EQ-5D by a patient proxy may provide alternative approaches that enhance acceptability and feasibility. It is well established that EQ-5D does achieve good response and completion rates.

2.3.2. Reliability

Reliability refers to the reproducibility or stability of an instrument. This can be thought of as stability of results over time in the same unchanged population (test-retest reliability), between raters or interviewers (inter-rater reliability) or between the location where the instrument is administered.

Test-retest reliability is perhaps the most relevant concept for which empirical evidence could be provided in the context of EQ-5D. However, this test relies on the assumption that a population remains unchanged between different administrations of the EQ-5D. In the absence of a gold standard it is extremely challenging to ensure and demonstrate that this assumption is valid. Other measures, both clinical and preference based, can provide evidence to support the validity of the assumption.

In addition, we must be mindful of the ultimate role of EQ-5D in economic evaluation. Our first requirement is that the estimate of the mean is unbiased. Whilst it is clear that individual responses to the EQ-5D demonstrate some degree of variability over time, provided the "error" is random this requirement will still be met. This does however, have implications for the efficiency of the estimates as larger samples will be required to reduce uncertainty around the mean which arises from this random variation.

2.3.3. Validity

The concept of validity refers to the extent to which an instrument measures what it is intended to measure. From this definition it is clear that "validity" is a question of degree. In the case of EQ-5D it is difficult to establish how such tests may be undertaken given the lack of a gold standard. In the extreme, EQ-5D itself could be argued to constitute the gold standard. Thus, in many applied investigations it is possible to provide evidence of differences between EQ-5D and other measures of outcomes rather than formal tests of validity. Brazier and Deverill (1999)[14] identify the following criteria that psychometricians use in the absence of a gold standard measure:

2.3.3.1. Content Validity

Content validity is concerned with whether the items of the instrument are appropriate for the health dimension being measured. So for a generic preference based tool such as the EQ-5D, one is concerned with whether the five items in the descriptive system together provide a reflection of the accepted conceptual model of health.

2.3.3.2. Face validity

Face validity is concerned with whether the items are sensible and appropriate for the population it is being administered to. There may be occasions where it is considered inappropriate to ask patients about the dimensions of health covered by instruments such as EQ-5D, although these few instances seem related to the ethics of asking patients to focus on symptoms or problems that may be distressing at that time, rather than founded on a basis that the conceptual model of health on which the instrument is based is itself sensible or inappropriate.

There are no formal statistical tests for face or content validity. Both are predominantly qualitative judgements. It could be argued that these "tests" cannot be applied to the consideration of whether EQ-5D, or any other instrument, is appropriate to use in the context of any single technology appraisal but must be undertaken at the system level. Both must consider "validity" with reference to the conceptual model of health that is accepted to define the "quality of life" element of QALY calculations.

Of course, it may not be practically feasible or ethically permissible to include a measure such as EQ-5D in a particular clinical trial for reasons that relate both to face validity, practicality and reliability. However, these situations do not themselves diminish the advantages of estimating EQ-5D valuations based on other methods than administering the instrument on the patients in the key studies of clinical

effectiveness. For example, proxy completion or statistical modelling between different measures could be used for estimation purposes.

2.3.3.3. Construct validity

Construct validity in general is concerned with the extent to which a scale measures or correlates with the psychological construct which it aims to measure. Thus, in the context of HRQoL instruments like EQ-5D, one would seek to assess the relationship with health state preferences. Again, the lack of a gold standard makes construct validity difficult to operationalise as a test rather than a description of differences between measures.

There are at least three main tests for construct validity as applied in psychometrics:

Firstly, there is the concept of *known-groups validity*. Here comparisons are made between scores of groups that are known to differ in the concept of interest. The instrument would be expected to detect such differences. However, since groups cannot be defined in terms of their health state utility by any gold standard, this process cannot be considered a formal "test" but a description of differences between measures that can then be used together with subjective judgements to make inferences about the degree of validity of a measure. It is possible, for example, to investigate whether patients that are known to differ in terms of a particular clinical indicator or its severity also differ in terms of EQ-5D. The usefulness of these types of descriptions depends on the extent to which the study design enables the investigator to ensure that differences in EQ-5D scores (or lack of) are due to the instruments rather than other factors which influence quality of life.

Convergent validity and *discriminant validity* are related subtypes of construct validity. Convergent validity refers to the situation where a dimension of the instrument is highly correlated with other measures that one believes should be correlated, that is, measures of the same underlying construct should converge. Discriminant validity is where measures of constructs that theoretically should not be related to each other are observed to not be related to each other, that is, dissimilar constructs can be discriminated. Walters (2009)[13] gives the example of correlations between physical functioning as measured by the MOS 36 item Short-Form Health Survey (SF-36) and the Frenchay Activities Index (FAI) as an example of the former and low correlation between physical function and pain as an example of the latter.

Brazier and Deverill (1999)[14] suggest that this type of examination should be conducted on the individual dimensions of instruments like EQ-5D rather than the overall preference based score (Brazier and Deverill (1999) p.45[14]). That is, they favour the examination of convergent validity. For example, we may wish to compare the distributions of responses to the EQ-5D's anxiety and depression item between patients with different severities of depression, perhaps defined using some disease specific measure such as Hospital Anxiety and Depression Scale (HADS). This would allow a more detailed understanding of the reasons why EQ-5D may not correspond with the comparator measure. For example, the analyst could determine whether situations where EQ-5D does not show any change when one is expected are due to insensitivity in the classification system rather than the fact that the change is not

valued by patients or is confounded by changes in other dimensions of health not considered by HADS.

Responsiveness
The responsiveness of an instrument refers to the concept of being able to reflect changes that occur in patients over time. It is therefore closely related to comparisons of known groups. In known groups comparisons the concern is with discriminating between patient groups that differ when defined in terms of some clinical or other outcome measure. Responsiveness involves comparisons across the same patients as their health status or other outcome measure changes over time. There is a clear requirement for changes in potential confounding factors to be considered in this type of analysis.

2.4. GENERAL CONSIDERATIONS

The concepts of feasibility, practicality, validity and reliability provide the general terms within which judgements about the appropriateness of EQ-5D and other preference based instruments can be developed.

Practicality and feasibility are often considered by looking at response rates, both of the overall instrument and individual items. However, there are unlikely to be many situations where alternative generic instruments will have been compared in this way. It could be argued that poor performance on these items requires alternative study designs to be considered e.g. proxy completion on behalf of the patient. It is also unlikely that studies will have compared alternative study designs in order to make this assessment.

When comparing the summary EQ-5D scores with any other clinical measure, this has descriptive value only. This cannot be considered a test of validity in the absence of a gold standard. Other contextual information is required in order to make a judgement about the differences between patients. For example, in a randomised controlled trial which demonstrates treatment efficacy in terms of a key primary clinical endpoint to compare the EQ-5D scores of the two arms could be seen as a test of known groups validity. However, the failure of EQ-5D to detect any change could be due to a) adverse events from the treatment cancelling out any benefit in the specific clinical dimension, b) other changes in health status not measured in the primary outcome of the trial, c) lack of sensitivity on the relevant components of the EQ-5D classification, d) clinical changes are not those for which patients are prepared to trade-off length of life.

The use of a Visual Analogue Scale (VAS) could be a useful outcome measure against which EQ-5D summary scores could be assessed. However, consideration needs to be given to the design of the study in which VAS is used since there are many well–documented issues with using VAS, for example spreading effects and end point biases. VAS may not reflect all aspects of HRQoL and may also differ from generic preference based measures because it reflects patient values, not those of the general population.

The suggestion in Brazier and Deverill (1999)[14] that comparisons could be made between the EQ-5D and TTO values elicited directly from patients may also suffer

similar shortcomings. Since observed differences between these measures may be due to the characteristics of the respondents, in particular the fact that the former are the general public and the latter are patients, differences in results are difficult to interpret.

Considerations of convergent validity are better examined by comparisons to the EQ-5D individual dimensions rather than the summary scores.

Lack of sensitivity in a dimension of EQ-5D does not necessarily make EQ-5D an inappropriate instrument. Lack of sensitivity may have implications for study design and, in particular sample size calculations. In many cases the variation in responses implies that a larger sample may be required to detect small changes in EQ-5D. Even where such samples are unfeasibly large, or the variation in responses is insufficient, there are alternatives available that do not require abandoning EQ-5D as the yardstick for the economic evaluation. At the very least, the existing tariff provides the feasible ranges for utility change that could be used in a sensitivity analysis. Alternatively, there is the option to statistically model the relationship between EQ-5D and a different outcome measure using external large datasets, thereby treating the EQ-5D as a continuous measure whilst simultaneously harnessing the greater statistical power that may be limited in other study designs such as clinical trials where sample sizes are based on other considerations.

A further issue for consideration when assessing empirical evidence in this area relates to the concept of statistical significance. Where changes in EQ-5D summary score are detected in the expected direction but these fail to achieve significance at traditional levels, caution should be exercised in interpretation. In many situations we may observe improvements from treatment when measured by a particular clinical outcome. Studies are invariably powered to detect precisely these differences. Many would argue that the mean EQ-5D difference should be used in the economic evaluation and the associated uncertainty represented appropriately. The ability to draw strong inferences about the appropriateness of EQ-5D as a tool in this intervention based on this kind of evidence is extremely limited.

3. A REVIEW OF CLAIMS

3.1. METHOD

The purpose of this section is to identify and classify the types of claims that have been made to NICE that relate to the inadequacy of EQ-5D in technology appraisals. We aimed to identify all such claims from two sources: i) The Kennedy Review of the Value of Innovation and ii) specific technology appraisals.

We reviewed all submissions made to the Kennedy study and were provided details of all responses to the Kennedy report that were considered potentially relevant by NICE staff. We also identified past technology appraisals by informal contacts with technology appraisals staff, appraisal committee chairs and the chair of the Institute. We cross referenced individual appraisals with the Kennedy submissions where sufficient information was provided. We obtained all relevant documentation

(assessment reports, manufacturer submissions, consultee comments, ACD and FAD documents) in order to identify the basis of each claim and any supporting evidence.

For each of the submissions made to Kennedy we extracted information about the claims which covered the precise nature of the claim, whether the claim was genuinely about features of the EQ-5D instrument or about other broader issues such as QALY maximisation or cost utility analysis per se, or the perspective adopted in technology appraisals.

For each of the technology appraisals we again identified the precise nature of the claim and considered whether it related specifically to EQ-5D rather than other elements of the NICE appraisal methods.

In all cases we critically reviewed any empirical evidence submitted or referred to in the submission, using the issues identified in section 2 as a guide.

3.2. FINDINGS

3.2.1. Submissions to the Kennedy Study

The Kennedy Study invited submissions from industry, experts and patient groups that provide the basis for this report. They were asked to comment on how NICE's current methodology may not fully capture the value of innovation in their appraisal of new technologies.

The comments that may be relevant to the general area of benefit valuation from the 50 organisations who submitted a response can be grouped into three themes:

1. Issues relating to the use of cost utility analysis and QALYs. These covered issues where it was felt the current approach is not well suited to reflecting all benefits e.g. patient safety issues or the process of care, or where the QALY maximisation approach does not reflect potential equity criteria sufficiently e.g. treatments for rare disorders.

2. Issues relating to the NHS and Personal Social Services perspective adopted in appraisal methods. For example, the impact of health technologies on productivity.

3. Issues relating to situations where EQ-5D is not considered appropriate.

Table 1 reports details of each of the claims identified.

The vast majority of the claims we identified that related to benefit valuation in NICE Technology Appraisals were not direct criticisms of the EQ-5D instrument but more general critiques of the cost per QALY approach per se or the perceived limitations of the perspective adopted by NICE. There were several instances where claims were made about the inadequacies of QALYs in reflecting benefits in specific circumstances but insufficient detail was provided to distinguish whether the claim relates to the concept of the QALY itself or is characteristic of particular measurement approaches such as EQ-5D.

A prominent theme in submissions was the perspective adopted in NICE appraisals. Within this there were two particular issues of concern. The first was that the societal benefits from technologies that permit individuals to be more economically active are not directly included. The second was that benefits to the carers of patients from improved treatments are not included. The perspective taken in NICE appraisals is determined by the remit it received from the Department of Health and the choice of appropriate perspectives for health care decisions has been the subject of a recently published review which the Department of Health commissioned[15].

Claims of inadequacy in the EQ-5D instrument were dominated by two features which are not necessarily mutually exclusive: where the instrument is considered insufficiently sensitive to detect changes in health and where the dimensions used to describe health states do not capture all relevant health issues.

Specific claims about the health benefits not captured by the QALY approach as operated by NICE included the following clinical areas: incontinence, mental health, cancer, palliative care, care of the elderly and fertility. More general characteristics of either health conditions or the patients affected that, it is claimed, make EQ-5D inappropriate are safety, route of administration, cognitive impairment, fatigue, the elderly, contraception and adverse events.

Mental health

Some of the claims regarding the inappropriateness of EQ-5D in the field of mental health focussed on issues of appropriateness, acceptability or feasibility (BIA submission, 4.11[16]). There are obvious practical problems with administering the instrument to groups that have learning difficulties, for example. No specific evidence to support these claims was provided and, as discussed above, it is questionable that these practical grounds could support the view that EQ-5D is inappropriate per se. These claims do potentially have relevance for primary study design issues such as sample size or proxy completion.

There were also claims made about the lack of sensitivity and "ceiling effects" of EQ-5D (BIA, 4.11[16]; Johnson and Johnson, p.5[17]; GSK, p.3 and p.8[18]). Little empirical evidence is provided to furnish these claims. Johnson and Johnson refer to Chisholm et al (1997)[19] to support their claim. However, this dated review paper actually notes that there was a paucity of evidence of the use of QALYs in mental health. No evidence at all is presented in relation to EQ-5D in this paper or by any of the other submissions. GSK refer to the NICE appraisal of newer drugs for the control of epilepsy but do not provide any details.

Cancer

Several bodies that submitted evidence to the Kennedy Study made claims about the inadequacy of EQ-5D in the context of cancer but, as with the vast majority of the claims made, there was no evidence provided to support the claims. Some of these claims centred on specific symptoms associated with cancer or its treatment such as fatigue (GSK, p.8[18]) or vitality (Amgen p.2[20]) that are not explicitly included as dimensions of the EQ-5D instrument. It is not clear here whether the claimed lack of sensitivity arises due to a lack of a specific dimension for these symptoms or that the number of levels within the existing five domains are insufficient. It could be argued

that the domain of usual activities in particular might be expected to capture at least part of the effect on quality of life of these types of symptoms.

Myeloma UK (part iii[21]) stated that they "tested" the EQ-5D with patients and concluded the instrument is insensitive. No further details are provided in the submission.

Fertility and contraception
Both the ABPI and the Medical Technology Group make the claim that QALYs do not reflect health benefits in relation to fertility. The latter give the example of new treatments for fibroids that may avoid the need for hysterectomy, the current standard treatment. Whilst the precise nature of these claims are difficult to disentangle, it does not appear that these are criticisms of EQ-5D per se, and may be reflections of a failure of specific applications to include the estimated health benefits of the unborn child.

The ethical judgements that must be made when considering issues around the value of the unborn child can lead to a decision not to use QALYs in the economic evaluation. For example, a recent project considered how to undertake an economic evaluation of fetal MRI scanning, which aims to provide more accurate diagnostic information to parents where traditional ultrasound has identified potential abnormalities. The decision taken here was not to include QALYs in the evaluation because this would require the estimation of the foregone health benefits associated with termination of abnormal and normal foetuses. Existing methods would allow such estimation, but the ethical acceptability of QALY maximisation as a decision rule in this setting is contentious.

3.2.2. Previous NICE appraisals

As part of the NICE response to the submissions made to the Kennedy Study, specific examples were provided of appraisals which, it was claimed, had recognised additional health benefits not captured by the reference case approach. Additional cases were identified by discussion with the technical team at NICE and Appraisal Committee chairs.

Several of the suggested topics did not reveal issues associated with concerns about EQ-5D upon detailed inspection of the appraisal documentation. In several cases there were substantial concerns about the methods that had been used to estimate health state utility values in the absence of EQ-5D or other measures having been applied within the clinical studies. For example, the requirement to estimate the relationship between clinical outcomes and EQ-5D in the appraisals of drugs for rheumatoid arthritis and psoriasis raised concerns. These relate to the statistical issues associated with such regression analysis rather than problems with EQ-5D per se.

Five appraisals were identified in total:

- Bortezomib monotherapy for relapsed multiple myeloma
- Omalizumab for severe persistent allergic asthma

- Pegaptanib and ranibizumab for the treatment of age-related macular degeneration
- Cochlear implants for severe to profound deafness in children and adults
- Long-Acting Insulin Analogue for diabetes

Case 1: Bortezomib monotherapy for relapsed multiple myeloma
The manufacturer's (Ortho Biotech) submission included an economic evaluation which used life-years (LYs) rather than QALYs as the measure of benefits. The manufacturer's reasons for not providing a QALY estimate of benefits, which are discussed in section 3.2.4 of their submission, were as follows;

- Increased survival is the single most important outcome for patients and clinicians given the nature of the condition.
- Direct evidence on EQ-5D from one of the clinical trials of bortezomib, APEX, is not usable due to poor completion rates, cross-overs and early termination of the trial.
- EQ-5D is unlikely to be an appropriate utility measure in this patient group.

The third point is supported by evidence from a focus group involving seven patients with multiple myeloma. Based on this evidence, it is claimed that the EQ-5D lacks face validity and that psychological adaptation compromises its applicability to this condition. Two domains were identified which the focus group did not feel were adequately captured by the EQ-5D. One was the experience of hospitalisation and the other was fear and anxiety regarding future disease progression.

The appraisal was an STA, and the ERG's report identified utility data from two studies[22;23]. The first was a randomised controlled trial (RCT) using EQ-5D directly in patients after treatment with either intensive chemotherapy or myeloablative treatment. The EQ-5D scores are reported at various follow-up points for each treatment group for patients in remission. For those patients not in remission an estimate was used to adjust the average EQ-5D scores for the age group (0.80) to account for multiple myeloma not in remission giving a score of 0.644. The second study mapped a cancer specific quality of life measure to 15D values (mean utility 0.789). The ERG observe that general population values for people aged 60 to 69 years range from 0.806 to 0.829[24] and therefore that health utility values in multiple myeloma "may be expected to be somewhat lower".

NICE requested cost per QALY estimates from the manufacturer and these were provided using EQ-5D utility estimates identified from the published literature. The Manufacturer's report of this additional analysis states that they conducted a literature review and identified three HRQoL studies in multiple myeloma patients all of which were reported to be in newly diagnosed symptomatic patients and two of which had also been identified by the ERG. The additional study mapped a cancer specific quality of life measure to EQ-5D values.

The study which used EQ-5D directly in a multiple myeloma population[22] was used by the manufacturer to provide cost per QALY estimates as requested by the Technology Appraisal Committee. The manufacturer states that their reasons for choosing this as the most appropriate estimate were that it reported HRQoL according

to responder rate and that it uses EQ-5D directly rather than using an indirect mapping approach. Although it should be noted that whilst the utility in responders was obtained directly from the EQ-5D, assumptions were used to calculate the utility in non-responders.

The Manufacturer's revised model uses an estimate of 0.81 for pre-progression utility and 0.644 for post-progression utility, both of which are based on EQ-5D estimates from the van Agthoven et al, 2004 study[22]. They state that their previous cost per QALY estimates used a utility of 0.81 to convert survival into QALY gains. They claim that the similarity of this value to general population values may reflect high adaptation in this population.

The Appraisal Committee's interpretation of the evidence presented by the Manufacturer and the ERG is summarised in section 4.6 of the FAD which states that the Committee did not accept the Manufacturer's view that life-years rather than QALYs were the most appropriate measure of benefit in this case as the impact of multiple myeloma and its treatment on HRQoL were likely to be important to patients and that there was evidence available to allow QALYs to be estimated. They were concerned that the utilities used in the cost per QALY calculations may not adequately capture the HRQoL impairment associated with relapsed multiple myeloma and therefore that the ICERS were likely to have been underestimated.

In summary, the only evidence presented regarding the inappropriateness of the EQ-5D for this population came from a focus group. There were EQ-5D estimates available for this population and these were used to derive cost per QALY estimates although some assumptions were required to estimate utility in non-responders.

Case 2: Omalizumab for severe persistent allergic asthma

Omalizumab (Xolair) is a humanized antibody drug for patients with moderate-to-severe or severe allergic asthma. The manufacturer (Novartis) included a cost-utility analysis as part of their submission which provided QALYs as the measure of benefits. The clinical trial data primarily used for the economic model was from the INNOVATE trial. Within the trial, participants completed the Asthma Quality of Life Questionnaire (AQLQ) which provides a disease specific quality of life outcome measure. This AQLQ data was converted to the EQ-5D using a regression function (see Tsuchiya et al 2002[25]).

This trial data was used to estimate the day-to-day asthma symptom state for the manufacturer's Markov model. However, another important component of health related quality of life in the economic model relates to the impact of exacerbations. The valuation for these states was not taken from the trials, but were instead taken from a prospective study by Lloyd and colleagues. Utilities for clinically significant, non-severe and clinically significant severe exacerbations were taken from a prospective study conducted in four UK asthma centres[26].

Subsequent discussions apparent in the appraisal documentation highlight that there were concerns with whether the trial was a more appropriate source for estimating the impact of exacerbations and whether the definition of exacerbations was equivalent

between the model and the prospective study. There was little discussion about the appropriateness of EQ-5D in this setting.

Case 3: Pegaptanib and ranibizumab for the treatment of age-related macular degeneration (AMD)

This appraisal was conducted under the MTA process. The Assessment Group's model and the manufacturer's model for pegatanib both used published estimates of utility for different visual acuity levels from a study which measured TTO directly in patients with AMD[27]. The manufacturer's model for ranibizumab used directly elicited TTO utility values measured in a general population sample who experienced simulated visual acuity states through the use of custom-made contact lenses[28]. The data from the Czoski-Murray study[28] were unpublished at the time of the appraisal and were included as confidential material within Novartis' submission. The manufacturer's submission also discussed utility values derived using the HUI-3 instrument[29].

In the first ACD, section 4.3.12 describes the Committee's discussion of the utility evidence as follows. "It considered that it may have been more appropriate to use utilities derived using a generic and validated classification system such as the EQ-5D, rather than those used in both the Assessment Group and manufacturers' models. It noted that use of the EQ-5D might result in a much smaller difference, perhaps by as much as a factor of 4, between utilities reflecting the best and worst vision states in the economic models, but nevertheless accepted the utilities used in the Assessment Group model as a guide to its decision making". After the consultation on the first ACD the Committee requested the Decision Support Unit to conduct a literature search to identify utility values and also requested additional analyses exploring the impact of using the Czoski-Murray and Espallargues estimates in the Assessment Group model and the economic model submitted by the manufacturer of pegatanib.

Novartis's comments on the first ACD included a criticism that the Committees statement regarding potential differences in utility gains when using the EQ-5D was not evidence based. Novartis' comments on the additional analysis included criticism of the HUI-3 instrument as being a crude generic description that is not sensitive enough. However, one of the key conclusions of the Espallargues study is that HUI3 is more sensitive that other preference based measures such as EQ-5D and SF-6D and that HUI3 had a higher correlation with Visual Acuity than direct patient TTO. The DSU report on the additional analysis using the Espallargues study states that it should be treated with some caution as the categories reported do not appear to discriminate well between the subgroups being considered.

In the second ACD and the subsequent FAD the Committee concluded that whilst in principle it is more appropriate to use utility values from a standardised and validated generic, the HUI-3 values may not fully capture the impact of AMD on patients quality of life and that the Czoski-Murray values provided the most plausible set of utility values for use in the economic models. The FAD refers to the fact that the Espallargues study found a small utility difference of 0.02 between people with visual acuity ranging from 6/12 to 6/24 and people with visual acuity ranging from 6/24 to 3/60 when using HUI3.

There were two appeals received on the FAD. One of the points raised during the appeal related to the size of utility loss associated with poor vision in one eye vs poor vision in both eyes and the relation this had to whether treatment should be recommended in the first eye affected or in the better seeing eye. The appellant on behalf of the manufacturer of pegatanib cited evidence from Williams (1998)[30] showing that quality of life did not vary significantly depending on whether one or both eyes were affected, and the Appraisal Committee responded that the utility loss for being blind in one eye was about 0.1 compared to a loss of 0.5 for being blind in both eyes. The estimate of 0.1 appears to be referring to a TTO exercise in patients[31] which is cited by the manufacturer of ranibizumab in their response to the additional analysis requested after the first ACD. Issues regarding the appropriateness of the instrument used to measure utility do not appear to have been raised in relation to this point. The Appeal Panel considered that the estimates used by the Appraisal Committee were reasonable in view of the range of utility analyses available to it.

A second point was raised during appeal regarding the difference in utility between patients with a visual acuity of 6/60 and patients with a visual acuity of 6/96 and the relation this had to the cut-off for treatment made in the recommendation for ranibizumab. There were difference in the utility estimates for visual acuity at the lower end of the range between the Brown and Czoski-Murray estimates and these were acknowledged, but issues regarding the appropriateness of the instrument used to measure utility do not appear to have been raised in relation to this point. The appeal panel concluded that there was no evidence of lesser efficacy or cost-effectiveness in those with poor baseline visual acuity.

In summary, the Appraisal Committee appear to have accepted the use of utility estimates derived from a direct TTO in a general population sample who were able to experience visual acuity states through the use of contact lenses. These are accepted despite there being estimates available from the Espallargues study which used both HUI3 and EQ-5D. This study showed that HUI3 was more strongly correlated with VA than EQ-5D or direct TTO. The impact of applying the HUI3 data was explored within the economic model, but it appears to have been accepted that the HUI3 data from this study was not able to adequately capture utility gain in this condition as only small differences in utility were observed for patients with a visual acuity of 6/12 to 6/24 compared to 6/24 to 3/60.

Case 4: Cochlear implants for severe to profound deafness in children and adults

This appraisal was conducted under the MTA process. The Assessment Report includes a systematic search for utility data to population the Assessment Group's economic model. Their search found that the HUI3 instrument has been widely used in this population but other generic preference based measures such as the EQ-5D have not been widely used. Data derived from the HUI-3 instrument were therefore used in the Assessment Group's economic model. The Assessment Report discusses the limitations of the HUI-3 instrument in terms of its focus on disability and the fact that the utility weights are obtained from a Canadian rather than a UK general population sample. HUI based utilities were also used in the economic models submitted by manufacturers and in many of the published models.

EQ-5D data was only reported in one of the clinical effectiveness studies which was a small (n=25) RCT comparing bilateral to unilateral cochlear implants in adults. In this study the Glasgow Health Status Inventory (GHSI) showed a significant difference in favour of bilateral implants. The HUI3 showed a small and non statistically significant difference in favour of bilateral implants whereas the EQ-5D showed a small but significant difference in favour of unilateral implants. The small but non significant improvement found using the HUI3 was used in the model. It was supported by similar results from a direct TTO valuation in normal-hearing volunteers who have familiarity with deaf adults. As there was no evidence available on the benefit of bilateral implants compared to unilateral implants in children, the Assessment Group applied the utility gain from adults to children.

In the first ACD the Appraisal Committee concluded that the gains in quality of life for pre-lingual children were likely to be higher than for those measured in adults, but that the benefits were less likely to have been underestimated for post-lingual children. They therefore recommended bilateral implants for pre-lingual but not post-lingual children. In the second ACD, the Committee stated that there was insufficient reliable evidence to quantify the benefits of bilateral implants over unilateral implants in either pre- or post-lingual children and therefore bilateral implants were not recommended for either group. Ultimately in the FAD the Appraisal Committee stated that the utility gain in children is likely to be larger than in adults but the size of the gain is highly uncertain and recommended bilateral implants for children.

The evidence regarding the size of utility gain associated with bilateral implants in children was the subject of many comments during the two rounds of consultation. Some consultees/commentators stated that there was insufficient evidence of benefit to support bilateral implants as a cost-effective intervention (Yorkshire and Humber SCG, South Central SCG). One consultation comment from ENT-UK, BCIG & BAA raised concerns regarding the ability of the HUI3 instrument to detect changes in this population given the limited states between which patients can move on the HUI3 instrument. Similar concerns were raised by Clinical Expert 1. NHS Newborn Hearing Screening Programme commented on the "remarkable degree of variability, lack of sensitivity and poor fitness for purpose for the commonly used health utility indices (SF-6D, EuroQoL and HUI3) for adults" They cited studies by Barton et al 2004[31]; Barton et al 2005[32] and Davis et al 2007[33]. LINK Centre for Deafened people commented on the "limitations in our current ability to measures with adequate sensitivity the full impact on quality of life" and the impact of this on the clinical effectiveness estimates for bilateral implants in adults.

In summary, there were difficulties in quantifying benefits in this appraisal, but these were mostly related to a lack of evidence on HRQoL in children having bilateral compared to unilateral cochlear implants. There was a paucity of EQ-5D evidence in this population but HUI3 data were available for many of the required health states. One consultee/commentator specifically criticised the sensitivity of EQ-5D and other instruments in this population and specific concerns were raised regarding the sensitivity of the HUI3 in this population.

Case 5: Long-Acting Insulin Analogue for diabetes

This was an MTA conducted in 2002. In estimating cost effectiveness it was found that a key driver of cost effectiveness was the utility value applied to the fear of hypoglycaemia. In particular, there was debate between the values used in the manufacturer model versus those in the assessment group model. The former were ten times higher than the latter.

The fear of hypoglycaemia data came from a cohort study which looked at the number of episodes in the past 3 months. Frequency of episodes was regressed against EQ-5D utility in order to estimate the decrement that should be attributed to reducing that frequency.

Other than the dispute between the assessment group and manufacturer regarding the magnitude of these estimates, there was little evidence that quality of life was considered an issue in this appraisal. The FAD highlights this as a key driver of the ICER and recommended this as a priority area for future research. It does not however demonstrate evidence of any claim that EQ-5D is not an appropriate tool to capture the health benefits of reduced frequency of hypoglycaemic events.

4. SUMMARY OF CLAIMS

This report has sought to clarify two sets of issues. First, on what grounds might EQ-5D be seen as an "inappropriate" choice of instrument for the calculation of QALYs in any specific context? We have presented and defined the criteria that have commonly been applied or discussed in the existing literature. Many of these concepts have been influenced by the discipline of psychometrics. It is noted however that the feasibility of performing a "test" of the EQ-5D, or any other method that seeks to reflect preferences for health states, is severely hampered, perhaps critically so, by the absence of any gold standard. It is of crucial importance that empirical evidence that claims to relate to the "inappropriateness" of EQ-5D is appropriately interpreted. In may situations, empirical data is capable only of describing differences between measures. Judgements or assumptions must be made in order to draw stronger interpretations.

It is also relevant to consider the broader decision making context of the Institute. The use of cost per QALY, a set threshold, albeit within a range, and a requirement to make decisions across diseases, patients and interventions makes consistency of decision making an important component. The attraction of the same approach to valuing health benefits is therefore obvious in this regard. It could be argued that rather than seeking to identify alternative tools to EQ-5D in specific disease areas, for example on the basis of efficiency of study design, future research should focus on alternative tools that are more globally relevant.

Our second main aim was to identify the claims made by stakeholders to the NICE Technology Appraisal process in relation to the inadequacy of EQ-5D as the tool of choice in the NICE reference case. These claims were sought in relation to submissions to the Kennedy Study into the Value of Innovation, both submissions

made to Sir Ian's report, responses to that report and the Institute's own response. We also conducted a directed review of previous NICE Technology Appraisals.

We found that very few claims were made that were directed at the EQ-5D instrument. In many cases concerns were raised about the appropriateness of either QALYs as a metric for the outcome of health interventions or the decision rule of QALY maximisation.

Empirical data to support the few claims that did relate to EQ-5D itself were sparse in both sets of documents. However, this finding does not mean that such data do not exist. The submissions made to Kennedy were not intended to constitute scientific documents nor was the issue of EQ-5D the central issue of the innovation study overall. Therefore, the lack of empirical data identified should not be considered a factor in determining which disease areas, patient groups or interventions should form the subject for future case study work.

Several common characteristics of claims can be identified:

It is claimed that EQ-5D is inappropriate in situations where the natural history of the underlying disease exhibits a relapsing-remitting feature. Examples such as asthma, diabetes, multiple sclerosis were apparent from the evidence we identified although the characteristic is a feature of numerous other conditions that have featured heavily in NICE's work programme. Two issues are claimed to raise particular difficulties: i) the feasibility of administering a quality of life instrument to patients in periods where the disorder peaks if this is a severe condition, exacerbations are short-lived, or both. ii) whether the EQ-5D instrument reflects the fear of those exacerbations at other times. The importance of "fear" as a concept was also apparent in the multiple myeloma appraisal.

Mental health and cancer were both provided as broad disease areas where unsupported claims about inadequacy of EQ-5D were made. These claims centred around feasibility and acceptability as well as issues related to the sensitivity of the instrument. Some claims also suggested that the absence of a category directly describing specific symptoms made the EQ-5D inappropriate in these fields. Similar claims are evident in relation to the previous appraisals on interventions relating to conditions with sensory impacts, i.e. hearing and sight.

There were claims made about weaknesses of the EQ-5D in relation to adverse events more generally. This also relates to some of the claims in relation to cancer, since in some situations, the differences between adverse event profiles for alternative therapies are evidence in the domains of fatigue, vitality or nausea.

These criteria provide a basis for selecting case studies where all relevant evidence may be identified, critically appraised and summarised. These should not be considered exhaustive and should not form the only basis for topic selection. In particular, the recognition of other ongoing research work would be sensible.

5. A REVIEW OF EQ-5D IN THREE DISEASE AREAS

In this section we describe three case studies where evidence of the performance of the EQ-5D has been systematically identified and critically reviewed. The case studies cover the disease areas of incontinence, asthma and rheumatoid arthritis. We also refer to a separate case study in the area of visual disorders. These areas were chosen on the basis of the review of claims made within the NICE appraisal process and the Kennedy Review. They reflect the key characteristics of those situations where EQ-5D has been claimed to be "inappropriate". Together they cover situations where there may be concern about missing dimensions within the EQ-5D descriptive system, particularly those where there are sensory impacts, and where the condition is characterised by peaks and troughs of severity of some symptoms.

We take different approaches to these reviews. In the cases of incontinence and asthma we undertake and report de novo reviews. For rheumatoid arthritis we provide a summary of previous reviews with consideration of primary research limited to those studies that have explicitly considered the relationship between a single outcome measure (the Health Assessment Questionnaire Disability Index – HAQ-DI) and EQ-5D. Finally, we provide additional discussion of a review relating to visual disorders. This review was not conducted as part of the DSU project but we have been granted access to a report undertaken by a Medical Research Council (MRC) funded project that performed this review[3].

We consider each of these reviews in turn before providing a combined discussion of the results.

5.1. ASTHMA

Asthma is a chronic disorder of the airways which results in symptoms that include recurring episodes of coughing, breathlessness, wheezing and tightness in the chest. These episodes can be variable and intermittent and commonly are triggered by exercise, smoke and allergens such as pollen, inter alia. Asthma exacerbations or attacks are acute episodes of a progressive increase in these symptoms[34]. Diagnosis and severity of asthma are sometimes based on objective tests of lung function (such as forced expiratory volume in the first second [FEV1]). Severity may also be judged according to symptoms and the amount of medication required to control the symptoms.

The objective of this review was to identify all published evidence that reported the use of EQ-5D in an asthma population in a manner that would provide empirical evidence relating to the performance of EQ-5D, as described in section 2.3 of this report.

5.1.1. Search strategy

The search strategy combined free text terms aimed at identifying papers reporting EQ-5D with free text and controlled terms (MESH and MESH-like terms) for asthma. The following databases were searched;

- BIOSIS (1969 – 15th May 2010)
- CINAHL (1982 – 15th May 2010)

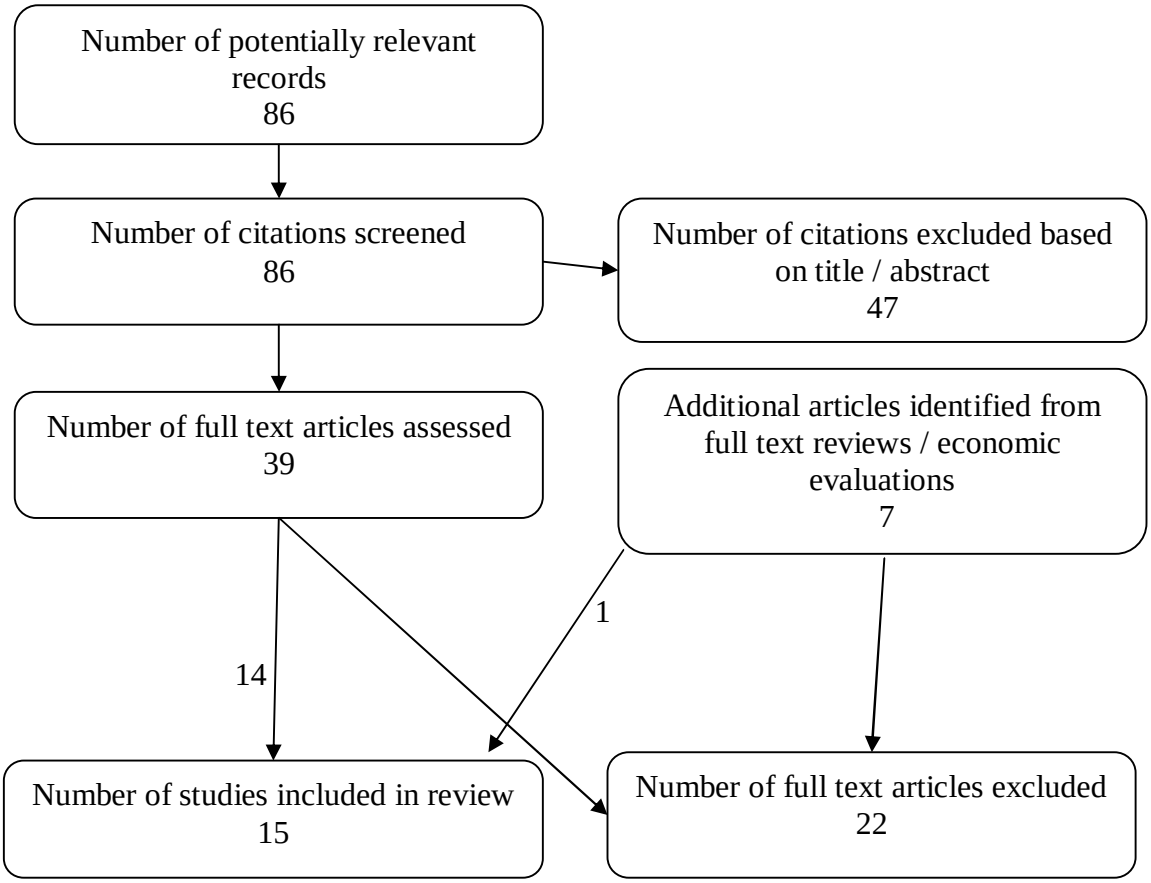
- Cochrane Library comprising the Cochrane Database of Systematic Reviews (CDSR), Cochrane Central Register of Controlled Trials (CENTRAL), NHS Economic Evaluations Database (NHS EED). (1991 – 15th May 2010)
- EMBASE (1980 – 15th May 2010)
- Euroqol website – EQ-5D references (www.euroqol.org)
- MEDLINE (In process and non-indexed 1950 – 15th May 2010)
- PsychNFO(1806 – 15th May 2010)
- Web of Science (1900 – 15th May 2010)

Included papers were those that reported EQ-5D alongside other measures of HRQoL or clinical measures in patients with asthma or in a broader population where results were reported for a subgroup of patients with asthma. Papers reporting valuations of clinical vignettes were excluded. There were no restriction relating to study design or interventions. Relevant systematic reviews and economic evaluations were ordered and their references checked for additional papers reporting primary data. Due to resource limitations, only English language studies were reviewed. Title and abstracts were sifted by two reviewers independently with discussion used to resolve any inclusion / exclusion discrepancies. Full text papers were sifted by a sole reviewer.

Data were extracted using a standardised set of forms which was similar to that used in the MRC Vision Review described in Section 5.4. Data extracted included study characteristics (country, study design, type of asthma and severity stage, treatment where relevant), participant characteristics (number, age, gender, ethnicity), outcome measures and results of responsiveness, validity and reliability assessments.

5.1.2. Results

A total of 86 citations were identified from the bibliographic searches. Of these 39 were ordered as full-text articles, although six papers (two reviews and four economic evaluations) were ordered purely to check their references for further primary studies. From these a further seven papers were identified and ordered as full-text articles.



A total of 15 papers were included in the review, the key features of which are reported in Table 1. These consisted almost entirely of cohort and cross sectional studies. Only one randomised controlled trial was identified[35] with the remainder of the papers reporting cohort or cross-sectional studies. The majority of the studies were conducted in a population with asthma although the exact criteria varied between studies, with some studies focusing on severe or difficult to treat asthma. Two studies were conducted in the general population sample who were asked about whether they had been diagnosed with a range of clinical conditions including asthma[36;37]. One study[38] recruited patients with asthma, COPD, and bronchiectasis from a general respiratory outpatients review clinic and patients with chronic cough from a specialist cough clinic. Another study recruited outpatients with various conditions, one of which was asthma[39]. These studies were included as they reported utilities for the subgroup of patients with asthma.

Table 1: Characteristics of included studies, reporting the validity and responsiveness of generic HRQoL in asthma populations

Study ref Author, Year	Country	Type of asthma (e.g severity, allergic)	Disease/treatment stage	Treatment (if any)	Study type (e.g. cross sectional, RCT, cohort)
Brusselle et al, 2009[40]	Belgium	Severe persistent allergic asthma	GINA classification of poorly controlled, severe persistent allergic asthma despite inhaled corticosteroids and long-acting beta2-agonist	Omalizumab (as add-on to usual care)	Cohort
Willems et al, 2007[35]	Netherlands	Outpatients with asthma (aged over 7)	GINA stage I to III	Telemonitoring vs usual care	RCT
Ferreira et al, 2010[41]	Portugal	Diagnosed asthma	Severity: Stage I, 17.0% Stage II or III. 76% Stage IV 7.1%		Cohort but only baseline data reported in this paper
Willems et al, 2009[39]	Netherlands	Outpatients with asthma as subset of wider sample	Asthma severity stage I to III.		Cohort
Polley et al, 2008[38]	UK (NI)	Clinically stable asthma	Recruited from respiratory outpatient review clinic		Cross-sectional
McTaggart-Cowan et al, 2008[42]	Canada	Self-reported physician diagnosed asthma	Self reported asthma control reported adequate by 87% and severity reported as mild by 38% and moderate by 33%		Cross-sectional
Chen et al, 2007[43]	US	Severe or “difficult to treat” as categorised by	GINA severity class: Mild, intermittent or persistent 7%	Continuation of current management with no specific study	Cohort

Study ref Author, Year	Country	Type of asthma (e.g severity, allergic)	Disease/treatment stage	Treatment (if any)	Study type (e.g. cross sectional, RCT, cohort)
		physician	Moderate, persistent 39% Severe, persistent 54%	intervention	
Aburuz et al, 2007[44]	UK	Adults with difficult to control asthma which was defined as one of; a) Persisting or refractory symptoms prompting specialist referral b) Minimal maintenance therapy wtih long acting beta-agonist and inhaled corticosteroids c) At least one course of systemtic corticosteroids in preceding 12mths	Recruited at regional centre (secondary care) Mean duration of illness was 20 years. Severity (ADSS score): Mild (0 to 1), 12.8% Moderate (2 to 3), 25.6% Severe (4 to 5), 45.3% Very severe (6 to 7), 16.3%		Cross-sectional
Lloyd et al, 2007[26]	UK	Moderate to severe asthma (BTS levels 4&5)	Recruited from GP and outpatients		Cohort
Szende et al, 2004[45]	Hugary	Diagnosed and previously treated asthma Inpatients and	GINA classification: Intermittent, 16% Mild, 28% Moderate, 36%		Cross-sectional

Study ref Author, Year	Country	Type of asthma (e.g severity, allergic)	Disease/treatment stage	Treatment (if any)	Study type (e.g. cross sectional, RCT, cohort)
		outpatients	Severe, 20%		
Oga et al, 2003[46]	Japan	Outpatients with asthma (American Thoracic Society definition)	All had previous medication and started at British Thorasic Society guideline step 3 or higher Mean duration of symptoms 9.9 years (sd 13.8)	Continuation of BTS guideline management	Cohort
McColl et al, 2003[47]	UK	Asthma patients in general practice setting			Cross-sectional
Garratt et al, 2000[48]	UK	Asthma patients in general practice setting			Cohort
Saarni, 2006[36]	Finland	Patients from general population self-reporting asthma diagnosis			Cross-sectional
Barton, 2008[37]	UK	Patients from general population (Age >=45) self-reporting asthma diagnosis	Not specified		Cross-sectional

The characteristics of the patients in the included studies are shown in Table 2. Four studies included less than 100 patients with asthma[38;39;44;46]. The total number of patients ranged from 20 to 4751. One paper looked exclusively at children[39] and another looked separately at adults and children[35]. One study recruited patients over the age of 12[40]. The remaining papers either looked solely at adults, or didn't specify whether children were included or excluded.

Table 2: Participant characteristics in included asthma studies

Study ref Author, Year	Number of participants	Age, mean (range)	%male/ %female	Ethnicity
Brusselle et al, 2009[40]	158 (183 screened, 160 enrolled and 158 met inclusion and had efficacy data collected)	48.16 (12-83, sd 17.18)	46.2/53.8%	94.9% Caucasian 5.1% Other
Willems et al, 2007[35]	109 (56 children, 53 adults) (of 274 potentially eligible patients approached)	Adults: control, 45.90 (sd 15.9), intervention, 45.65 (sd11.3) Children: control, 10.85 (sd 2.3) intervention, 10.57 (sd 2.1)	Adults: control, 33.3%/66.7% intervention, 42.3%/57.7% Children: control, 55.6%/44.4% intervention, 72.4%/27.6%	Not reported
Ferreira et al, 2010[41]	115	49 (sd 16.9)	29.8%/70.2%	Not reported
Willems et al, 2009[39]	56 of 86 invited to participate (asthma subgroup)	Not reported for asthma subgroup. For whole sample split by age category: 7-12 years (N=99) 10 (sd 1.5) 12-18 years (N=62), 15 (sd 1.8)	Not reported for asthma subgroup. For whole sample split by age category: 7-12 years (N=99) 59%/41% 12-18 years (N=62), 48%/52%	Not reported
Polley et al, 2008[38]	20 with asthma as subgroup of 147.	51.6 (sd 17.5)	65%/35%	Not reported
McTaggart-Cowan et al, 2008[42]	157	35.0	30%/70%	Not reported
Chen et al, 2007[43]	987	52.8 (20 to 89)	27%/73%	87% White 8% Black 3% Hispanic 1% Asian/Pacific Islander 1% Other
Aburuz et al, 2007[44]	86	42.3 (sd 15)	38.4%/61.6%	Not reported
Lloyd et al, 2007[26]	112 (had data at both time points)	Reported separately by exacerbation status: No exacerbations, 40.5	Reported separately by exacerbation	Not reported

Study ref Author, Year	Number of participants	Age, mean (range)	%male/ %female	Ethnicity
	Number by exacerbation status No exacerbations, 85 Exacerbation without hospitalisation, 22 Hospitalisation, 5	(11.6 sd?) Exacerbation without hospitalisation, 41.4 (12.0 sd?) Hospitalisation, 48.4 (11.0 sd?)	status: No exacerbations, 39.3%/60.7% Exacerbation without hospitalisation, 27.3%/72.7% Hospitalisation 40%/60%	
Szende et al, 2004[45]	228	49	34%/66%	Not reported
Oga et al, 2003[46]	54	46.8 (19 to 87) (sd 19.3)	41%/59%	Not reported
McColl et al, 2003[47]	4751 questionnaires sent out.	69.1 (sd 10.2)	57% / 43%	Not reported
Garratt et al, 2000[48]	235 (394 sent questionnaire)	Mean not reported (inclusion criteria was 18-60)	Not reported	Not reported
Saarni, 2006[36]	8028 of which 8.8% reported asthma	53 in general pop 57 in asthma group	47%/53% in general pop 38%/62% in asthma group	Not reported
Barton, 2008[37]	116 reported asthma from 1865 who returned questionnaire	Mean not reported All patients had age >=45	55%44% in general pop Not reported for asthma subgroup	98% White 2% Non-white

Table 3 shows the measures that are reported in each of the selected studies. Of the generic utility measures included we chose to include one study that included the EQ-5D child version[39]. In addition to the EQ-5D, eight studies administered the SF36 or some variant of it. Five included the SF6D utility values. Singles studies reported the HUI-3[42], the NHP[46], the 15D[36], and the TACQOL[39]. The TACQOL is a generic measure of HRQoL for use in children and their parents. Not all the studies used the UK valuation set but instead utilised those more relevant for the setting of the stud (Belgium, US, Japan). The EQ VAS was widely administered with 11 of the studies including this measure.

The main clinical measures reported were asthma severity, asthma control, exacerbation history, medication use and FEV_1. Asthma control was reported using several different validated scales (ACQ, ACS, ATAQ). Several studies reported using the GINA criteria to define either disease severity or asthma control and one study reported severity using the Asthma Disease Severity Scale (ADSS) which appears to be based on the Asthma Control Scale (ACS). It is therefore likely that measures of severity and control should be regarded as overlapping rather than mutually exclusive.

Several studies reported disease specific quality of life instruments. The most commonly reported instrument was the AQLQ which has several forms. The original AQLQ includes 32 items across four domains and includes five patient chosen items within the activity domain. There is also a standardised form called the AQLQ(S), which replaces the patient chosen activity items with the activity items most

commonly chosen by patients in the original studies. There is also a mini version of this which has 15 rather than 32 items (Mini-AQLQ).

Other respiratory specific instruments used include the St Georges Respiratory Questionnaire (SGRQ) and two instruments looking specifically at the impact of cough on quality of life (LCQ, CQLQ).

Two preference based disease specific instruments were also reported on. The AQL-5D is an instrument that has been derived from the AQLQ in much the same way that the SF-6D was derived from the SF-36[49] and it relies on a TTO valuation set[50]. The Asthma Symptom Utility Index is an instrument which has four symptom domains and one side-effect domain. Its preferences were derived using a combination of VAS and standard gamble[51].

Table 3: Measures used in the included asthma studies

	GENERIC MEASURES			OTHER MEASURES USED				
Study ref Author, Year	**Descriptive system (EQ-5D, HUI2, HUI3, SF-6D)**	**Tariff used (state tariff, or provide methods & source of valuation if no tariff used)**	**Mean value (SD, range)**	**Details of direct valuation (own health or vignettes? Whose health? Whose values? Which valuation method (TTO, VAS, SG)?)**	**Condition-specific HRQL measures used**	**Clinical measures used**	**Qualitative questions (provide details of any qualitative questions asked)**	**Missing data; completion rates of measures; patients completing study etc. (include reasons for non-completion if given)**
Brusselle et al, 2009[40]	EQ-5D	Belgian population norms.	At baseline: 0.54 (0.24, -0.16 to 1.00)	EQ-VAS	AQLQ	GINA improvement (severity) GETE Severe exacerbation (systemic corticosteroids, A&E attendance or hospitalisation) FEV_1		ITT analysis has 130 at 52 weeks PP has 105 at 52 weeks EQ-5D data only available for 51.5% of ITT and 51.4% of PP sample.
Willems et al, 2007[35]	EQ-5D* SF-6D *child version used ages 7 to 18.	UK tariff* UK tariff *child version uses adult tariff	EQ-5D at 12 mths Adults: control, 0.79 (sd 0.21) intervention, (0.90 (sd 0.11) Children: control, 0.97 (sd 0.05) intervention, 0.98 (sd 0.04) SF-6D at 12mths Adults: control, 0.74 (sd 0.14),	EQ-VAS				7/109 lost to follow-up

	GENERIC MEASURES			OTHER MEASURES USED				
Study ref Author, Year	**Descriptive system (EQ-5D, HUI2, HUI3, SF-6D)**	**Tariff used (state tariff, or provide methods & source of valuation if no tariff used)**	**Mean value (SD, range)**	**Details of direct valuation (own health or vignettes? Whose health? Whose values? Which valuation method (TTO, VAS, SG)?)**	**Condition-specific HRQL measures used**	**Clinical measures used**	**Qualitative questions (provide details of any qualitative questions asked)**	**Missing data; completion rates of measures; patients completing study etc. (include reasons for non-completion if given)**
			intervention, 0.75 (sd 0.14)					
Ferreira et al, 2010[41]	EQ-5D SF-6D	EQ-5D: UK valuation set. SF-6D: Portuguese valuation set	EQ-5D: 0.85 (0.16, 0.09 to 1.00) SF-6D: 0.86 (0.09, 0.60 to 1.00)	EQ-VAS	AQLQ(S)	Severity (stage I-IV) FEV_1 ACQ (asthma control)		
Willems et al, 2009[39]	EQ-5D (child version) [aged 7-12 proxy reported, age 12-18 self reported] TACQOL (generic non-preference based)	UK Tariff for EQ-5D		EQ-VAS			Perceived change in health status (change or no change)	47 of 56 returned 2nd questionnaire
Polley et al, 2008[38]	EQ-5D	Tariff not stated	0.63 (sd 0.38)	EQ-VAS	cough specific (not asthma specific) LCQ CQLQ	FEV_1		
McTaggart-Cowan et al, 2008[42]	EQ-5D, HUI-3, SF-6D	EQ-5D: UK valuation set. HUI-3 Canadian valuation set. SF-6D	EQ-5D: 0.85 (0.23, -0.0-6 to 1.00) HUI-3: 0.84 (0.20, 0.12 to	VAS	AQLQ(S) AQL-5D (UK TTO valuation)	ACQ (asthma control) Self reported severity and control. (Likert	None	

	GENERIC MEASURES			OTHER MEASURES USED				
Study ref Author, Year	**Descriptive system (EQ-5D, HUI2, HUI3, SF-6D)**	**Tariff used (state tariff, or provide methods & source of valuation if no tariff used)**	**Mean value (SD, range)**	**Details of direct valuation (own health or vignettes? Whose health? Whose values? Which valuation method (TTO, VAS, SG)?)**	**Condition-specific HRQL measures used**	**Clinical measures used**	**Qualitative questions (provide details of any qualitative questions asked)**	**Missing data; completion rates of measures; patients completing study etc. (include reasons for non-completion if given)**
		UK valuation set.	1.00) SF-6D: 0.79 (0.10, 0.48 to 1.00)			scale) Use of short-acting Beta-agonists FEV_1 (as % of predicted FEV_1)		
Chen et al, 2007[43]	EQ-5D	US valuation set	0.86 (sd 0.16)	EQ-VAS	Mini-AQLQ	ATAQ (asthma control) GINA (severity) FEV_1 % (lung function)	None	Excluded patients were more likely to be non-white but similar in age, sex, education, smoking, comorbidity, severity and lung function.
Aburuz et al, 2007[44]	EQ-5D	Not stated	0.47 (sd 0.33)	EQ-VAS	AQLQ	ADSS (severity) FEV_1		4/90 approached refused to participate
Lloyd et al, 2007[26]	EQ-5D	Not stated	EQ-5D after 4 weeks. No exacerbations: 0.89 (sd 00.15) Exacerbations with oral steroids: 0.57 (sd 0.36) Hospitalised: 0.33 (sd 0.39)	EQ-VAS	mAQLQ ASUI (preference based)	FEV_1 (baseline)	none	2 lost to follow-up Some HRQoL data missing from 23 participants
Szende et al, 2004[45]	EQ-5D SF-36 SF-6D	EQ-5D: UK valuation set SF-6D: UK valuation set	0.70 for whole sample. Ranged from 0.52 in poor	EQ-VAS TTO	SGRQ	Asthma control classified using GINA in treated patients	none	Not reported

	GENERIC MEASURES			OTHER MEASURES USED				
Study ref Author, Year	**Descriptive system (EQ-5D, HUI2, HUI3, SF-6D)**	**Tariff used (state tariff, or provide methods & source of valuation if no tariff used)**	**Mean value (SD, range)**	**Details of direct valuation (own health or vignettes? Whose health? Whose values? Which valuation method (TTO, VAS, SG)?)**	**Condition-specific HRQL measures used**	**Clinical measures used**	**Qualitative questions (provide details of any qualitative questions asked)**	**Missing data; completion rates of measures; patients completing study etc. (include reasons for non-completion if given)**
			control to 0.93 in good control.			FEV_1		
Oga et al, 2003[46]	EQ-5D SF-36 NHP	Japanese valuation set	Baseline: 0.808 (sd 0.187) 3mths: 0.887 (sd 0.145) 6 mths: 0.879 (sd 0.146) P<0.05 for 3mths and 6mths vs baseline.		AQLQ	FEV_1	None	Not reported
McColl et al, 2003[47]	EQ-5D SF-36	UK valuation set	Not reported		AQLQ	NASQ	None	64.4% response rate
Garratt et al, 2000[48]	EQ-5D SF-12	Not stated	Not reported		AQLQ	NASQ	None	59.6% response rate (235/394) Longitudinal data for 134.
Saarni, 2006[36]	EQ-5D 15-D	UK tariff for EQ-5D 15-D Finnish valuation set	For asthma subgroup; EQ-5D: 0.766 (SE 0.011) 15-D: 0.864 (SE 0.005)		None	None	None	
Barton, 2008[37]	EQ-5D SF-6D	UK tariff for EQ-5D Brazier algorithm for SF-6D	Not reported for asthma subgroup	EQ-VAS	None	Use of analgesics Use of any prescription medicines	None	1865 / 2770 returned questionnaire

AQLQ=Asthma Quality of Life Questionnaire, ACQ=Asthma Control Questionnaire. AQL-5D=Preference based measure based on AQLQ(S), AQLQ(S)=Asthma Quality of Life Questionnaire (standardized version). FEV_1= Forced expired volume in the first second. ASUI=Asthma symptom utility index, SGRQ=St Georges Respiratory Questionnaire. NASQ = Newcastle Asthma Symptoms Questionnarie. GINA=Gobal Initiative for asthma classification, GETE=global evaluation of treatment effectiveness, ITT=intention to treat, PP=per protocol. LCQ=Leicester cough questionnaire. CQLQ = Cough Quality of life questionnaire.

Five studies[26;41-43;45] in total provide information that allow the mean EQ-5D to be compared across asthma groups which were considered to differ in terms of severity, control, exacerbation history, medication use or lung function. Results are displayed in Table 4. Two studies[41;42] reported EQ-5D outcomes for more than one set of known groups giving a total of eight known group comparisons. In six of the know group comparisons, the mean EQ-5D score differed across all of the groups in the direction expected. In only one comparison were these differences reported as not statistically significant. In the other two known group comparisons, the EQ-5D didn't vary consistently across the groups in the direction expected and differences between the groups were not statistically significant.

Two studies reported known group validity by asthma severity. In one study[41] EQ-5D varied in the expected direction across the groups with the difference between groups statistically significant at p=0.1. In this study the SF-6D did not vary as expected across the groups but the groups varied significantly (p=0.01). In the second study[42] EQ-5D did not vary as expected across all the four groups and differences between groups were not statistically significant at p=0.09. In this study there were similar problems of apparent inconsistency with SF-6D and HUI-3, whilst the AQL-5D (condition specific & preference based measure) did reflect differences in asthma severity. The EQ-VAS performed well in one study but not the other whilst the condition specific measures of AQLQ(s) and ACQ performed well across both studies.

Three studies reported differences between groups defined by the degree of asthma control[42;43;45]. The EQ-5D utility scores changed as expected across all of the groups in all three studies and differences between the groups were significant in one of the studies, non significant in another and not reported in the third. The study which found no significant difference between the groups based on self-reported asthma control also reported that there was a significant difference between groups stratified by ACS, although the utility scores didn't vary consistently with ACS (not included in Table 4). The SF-6D results were similar whilst the HUI-3 exhibited one apparent inconsistency in the single study in which it was included[42]. AQL-5D and EQ-VAS also achieved consistent and statistically significant results by group. The study not reporting statistical significance for difference between groups[45], also had consistent changes in the TTO and the SF-36 physical component but not the SF-36 mental component. The condition specific measures generally performed well.

One study reported differences between three groups defined by exacerbation history (No exacerbations, exacerbations with steroid use, hospitalised) and all the included measures discriminated between the groups well (EQ-5D, Mini-AQLQ, Mini-AQLQ domains, ASUI, EQ-VAS)[26]. Another study defined groups in terms of beta-agonist canisters used in the past year[42]. The majority of the included measures performed well (EQ-5D, HUI-3, AQL-5D, EQ-VAS, AQLQ(S), ACQ) with the exception of SF-6D and VAS.

One study reported differences between four groups defined by FEV_1[41]. None of the included measures, including EQ-5D, discriminated between the groups.

Table 4: Results of "known groups" comparisons in asthma studies

Study ref Author, Year	*Groups defined as*	*Instrument*	*Direction of change consistent across groups?*	*Direction change consistent with clinical expectation*	*Difference between groups statistically significant?*
Ferreira et al 2010 [41]	Asthma severity stage: I II III IV	EQ-5D SF-6D EQ-VAS AQLQ(S) ACQ	Yes No (stage IV > III) Yes Yes Yes	Yes No (stage IV > III) Yes Yes Yes	Yes at p=0.1 Yes at p=0.01 Yes at p=0.01 Yes at p=0.001 Yes at p=0.001
McTaggart-Cowan et al, 2008 [42]	Self reported asthma severity: Very mild Mild Moderate Severe	EQ-5D HUI-3 SF-6D AQL-5D EQ-VAS VAS AQLQ(S) ACQ	No (Mild highest) No (Mild highest) No (Mild=very mild) Yes No (Mild highest) Yes Yes Yes	No (Mild highest) No (Mild highest) No (Mild=very mild) Yes No (Mild highest) Yes Yes Yes	No at p=0.09 No at p=0.09 No at p=0.09 Yes at p=0.001 Yes at p=0.001 Yes at p=0.001 Yes at p=0.001 Yes at p=0.001
McTaggart-Cowan et al, 2008 [42]	Self reported asthma control: Very well Well Adequate Not	EQ-5D HUI-3 SF-6D AQL-5D EQ-VAS VAS AQLQ(S) ACQ	Yes No (well lowest) Yes Yes Yes Yes Yes Yes	Yes No (well lowest) Yes Yes Yes Yes Yes Yes	No at p=0.09 No at p=0.09 No at p=0.09 Yes at p=0.001 Yes at p=0.001 Yes at p=0.001 Yes at p=0.001 Yes at p=0.001
Chen et al, 2007[43]	ATAQ (asthma control): 0 problems 1 problem 2 problems 3 problems	EQ-5D AQLQ overall AQLQ domains EQ-VAS (at follow-up)	Yes Yes Yes Yes	Yes Yes Yes Yes	Yes at p=0.001 Yes at p=0.001 Yes at p=0.001 Yes at p=0.001

	4 problems (at baseline)				
Szende et al, 2004[45]	Asthma control; Good Mild reduced Moderated reduced poor	EQ-5D EQ-5D domains SF-6D EQ-VAS SF-36 (mental) SF-36 (physical) TTO SGRG	Yes Yes (except for self care) Yes Yes No (poor better than mod) Yes Yes Yes	Yes Yes (except for self care) Yes Yes No (poor better than mod) Yes Yes Yes	Not reported Not reported Not reported Not reported Not reported Not reported Not reported Not reported
Lloyd et al, 2007[26]	Exacerbation experience: No exacerbation Exacerbation with steroids use Hospitalised	EQ-5D Mini-AQLQ Mini-AQLQ domains ASUI EQ-VAS	Yes Yes Yes Yes Yes	Yes Yes Yes Yes Yes	Yes at p=0.001 Yes at p=0.001 Yes at p=0.001 Yes at p=0.001 Yes at p=0.001
McTaggart-Cowan et al, 2008[42]	β-agonists use (canisters in past year): 4 or less 5-12 12 or more	EQ-5D HUI-3 SF-6D AQL-5D EQ-VAS VAS AQLQ(S) ACQ	Yes Yes No (5-12 highest) Yes Yes No Yes Yes	Yes Yes No (5-12 highest) Yes Yes No (5-12 > 4 or less) Yes Yes	Yes at p=0.05 Yes at p=0.05 No at p=0.09 Yes at p=0.001 Yes at p=0.001 Yes at p<0.05 Yes at p=0.001 Yes at p=0.001
Ferreira et al, 2010[41]	FEV_1: <50% 50 to 74% 75 to 99% 100%	EQ-5D SF-6D EQ-VAS AQLQ(S) ACQ	No Yes No No No	No No No No No	No at p=0.1 No at p=0.1 No at p=0.1 No at p=0.1 Yes at p=0.01

The relationship between EQ-5D and a series of asthma specific outcome measures and generic outcomes measures was assessed via correlation or regression in ten studies in total. The results from these studies are displayed in Table 5.

Correlations with AQLQ were reported in five studies[41;42;44;47;48]. EQ-5D had moderate to strong correlations for overall AQLQ scores (four studies) ranging from 0.41 to 0.57[41;42;44;48]. These were all in the expected direction and were either statistically significant or there was a failure to report significance (one study[42]). Results were similar for the individual AQLQ domains, although the range of correlation coefficients was greater (0.31 to 0.71). Only two studies reported the correlation between AQLQ and SF-6D and the coefficients for SF-6D were marginally higher than for EQ-5D (0.61 vs 0.53 and 0.43 vs 0.41)[41;42]. One study reported correlation between AQLQ and HUI-3 and the coefficients were marginally smaller than for EQ-5D (0.40 vs 0.41)[42]. AQL-5D was highly correlated with AQLQ which was to be expected given the design of this instrument[42].

Correlations between the Asthma Control Questionnaire (ACQ) and EQ-5D were reported in two studies[41;42] and were 0.37 and -0.51 respectively. A negative relationship is as expected since the ACQ scores from 0 (totally controlled asthma) to 6 (totally uncontrolled), but in one study[42] the ACQ results were rescaled to ensure a positive correlation with the other measures. SF-6D and HUI-3 were reported in one study and had similar results with slightly smaller coefficients. AQL-5D was highly correlated with ACQ (coefficient 0.82 when ACQ scale reversed)[42].

There was evidence of moderate to strong correlation between EQ-5D and a range of other clinical measures (CQLQ, LCQ, SGRQ, NASQ). The exception was FEV_1 which was not well correlated with any preference based outcome except TTO.

EQ-VAS had moderate correlations with the EQ-5D, SF-6D, HUI-3 and AQL-5D although the SF-6D appeared to correlate less with the EQ-VAS than the EQ-5D. The EQ-5D had moderate correlations with the SF-36 instrument. SF-12 also had moderate correlation with the EQ-5D but the coefficients were lower than for SF-36.

In the two studies that considered the relationship between EQ-5D and other outcome measures using statistical regression modelling, one study found that ATAQ at baseline was a significant predictor of EQ-5D at follow-up but not change in ATAQ from baseline or severity or FEV1 at baseline[43]. The other study found a statistically significant utility decrement associated with a diagnosis of asthma for 15-D but not EQ-5D[36].

Table 5: The relationship between EQ-5D and other measures in asthma

Study ref Author, Year	*Clinical/non pref based measure*	*Correlation with EQ-5D*	*Correlation with SF-6D*	*Correlation with HUI-3*	*Correlation with AQL-5D**	*Correlation with TTO*	*regression*	*Other*
Ferreira et al, 2010[41]	EQ-VAS AQLQ(S) AQLQ domains ACQ	0.48 0.53 0.31 to 0.56 -0.51 All in expected direction All p<0.001	0.43 0.61 0.38 to 0.64 -0.49 All in expected direction All p<0.001					
Polley et al, 2008[38]	CQLQ total LCQ total FEV_1	-0.68 (p=0.002) 0.66 (p=0.002) 0.06 (p>0.1) All expected direction						
McTaggart-Cowan et al, 2008[42]	EQ-VAS AQLQ(S) ACQ (direction reversed) FEV_1 AQL-5D* *disease specific preference based	0.59 0.41 0.37 0.14 0.41 All in expected direction P values not given	0.51 0.43 0.34 0.15 0.43 All in expected direction P values not given	0.58 0.40 0.32 0.06 0.39 All in expected direction P values not given	0.60 0.91 0.82 0.26 - All in expected direction P values not given			Plot of utility across ACQ scores: AQL-5D shows consistent gradient across ACQ range EQ-5D, HUI-3, SF-6D show changing gradients across ACQ range

Study ref Author, Year	*Clinical/non pref based measure*	*Correlation with EQ-5D*	*Correlation with SF-6D*	*Correlation with HUI-3*	*Correlation with AQL-5D**	*Correlation with TTO*	*regression*	*Other*
Chen et al, 2007[43]	Mini-AQLQ	0.42 Both expected direction and p<0.0001					Multivariate regression on EQ-5D at follow-up (adjusted for age, race, education, smoking, comorbid, severity and FEV_1): Independent significant (p<0.05) predictors were ATAQ (at baseline), and some general characteristics such as comorbidities, but not severity or FEV_1 or change in ATAQ from baseline to follow-up.	
Aburuz et al, 2007[44]	AQLQ overall AQLQ domains	0.57 0.40 to 0.55 All p<0.0001 All in expected direction.						
Szende et al, 2004[45]	SGRQ FEV_1 SF-36 (physical) SF-36 (mental) EQ-VAS	-0.68 0.21 0.62 0.59 0.55	Not reported Not reported Not reported Not reported 0.48			-0.36 0.36 0.34 0.25 All in expected		

Study ref Author, Year	*Clinical/non pref based measure*	*Correlation with EQ-5D*	*Correlation with SF-6D*	*Correlation with HUI-3*	*Correlation with AQL-5D**	*Correlation with TTO*	*regression*	*Other*
		All in expected direction P not reported				direction P not reported		
McColl et al, 2003[47]	Results when generic (EQ-5D / SF-36) instrument used first: NASQ AQLQ domains SF-36 domains Result when condition specific used first: NASQ AQLQ domains SF-36 domains	-0.52 0.42 to 0.59 0.51 to 0.69 -0.52 0.42 to 0.58 0.51 to 0.71 All in expected direction P not reported						
Garratt et al, 2003[48]	NASQ AQLQ overall AQLQ domains SF-12 (PCS) SF-12 (MCS)	0.45* 0.56 0.44 to 0.60 0.49 0.37 All p<0.01 *expected to						

Study ref Author, Year	*Clinical/non pref based measure*	*Correlation with EQ-5D*	*Correlation with SF-6D*	*Correlation with HUI-3*	*Correlation with AQL-5D**	*Correlation with TTO*	*regression*	*Other*
		be negative, but no comment made so may be reporting error.						
Saarni, 2006[36]							Utility loss associated with asthma based on multivariate regression controlling for sociodemographic variables and other chronic conditions: 15D: -0.019 (SE 0.005) $p<0.01$ EQ-5D: -0.008 (SE 0.008). $p>0.01$ Statistically significant HRQoL decrement for 15-D but not EQ-5D	

Table 6 shows results from studies that provide information on the responsiveness of EQ-5D.

Two cohort studies reported significant improvements over time in both EQ-5D and the disease specific AQLQ (ref id 3, 24). Agreement was also seen in these two studies between EQ-5D and several generic non-preference based measures (EQ-VAS, SF-36, NHP). In a third study which had an RCT design there was agreement with EQ-VAS in children but not in adults (ref id 4). For adults, SF-36 data were also reported in this RCT and agreement was seen between EQ-5D and some (6/8) domains of the SF-36 and between EQ-5D and SF-6D. When considering comparisons between trial arms, the QALY gains estimated for adults in this RCT were statistically significant and positive for EQ-5D and statistically non-significant and negative for SF-6D. However, these small and/or non-significant QALY differences may reflect the fact that no significant differences were seen between the treatment and intervention groups over time in either utility measure or the clinical outcomes reported in the main study publication[52].

One cohort study reported responsiveness as the proportion achieving a pre-specified response criteria[40]. A substantial proportion were regarded to have responded across all outcomes. 57% achieved an EQ-5D utility gain of 0.074 or more. Whilst 85% achieved an AQLQ gain of >0.5.

Two cohort studies looked at changes from baseline by response category. One cohort study looked at changes from baseline for those experiencing no exacerbations, exacerbations requiring medication and exacerbations requiring hospitalisation[26]. The change in EQ-5D was significantly different across these three groups, but there were no corresponding significant differences in disease specific quality of life (mini-AQLQ) or disease specific utility (ASUI). The second study looked for a linear relationship in change from baseline across five response categories and found a significant linear relationship for EQ-5D and for disease specific measures (NASQ, AQLQ)[48]. For the SF-12, differing results were found for the two component scores in that PCS had a significant relationship whilst MCS did not. The disease specific measures produced larger F-statistics than then generic measures.

Two studies reported standardised response means for different instruments[46;48]. The standardised response means were lower for EQ-5D than for disease specific measures (NASQ, AQLQ) but were comparable to other generic instruments (SF-36, NHP, SF-12).

Table 6: EQ-5D responsiveness in asthma

Study ref Author, Year	Discrete comparisons			Continuous comparison			
	Proportion achieving change in clinical outcome	*Proportion achieving change in EQ-5D*	*Agreement with direction??*	*Δ clinical measure(s) or other preference based utility*	*Δ EQ-5D*	*Agreement with direction??*	*Agreement with stat sig*
Brusselle et al, 2009[40]	ITT results. Improved GINA: 31.0%, Good or excellent GETE: 72.3%, AQLQ gain >0.5: 84.4%, Free of severe exacerbations: 65.6%, PP	EQ-5D gain >0.074: 56.7%,	Yes. Improvements for all outcomes.	Reduction in severe exacerbations per annum: 1.78 ITT and 1.79 for PP. (no p values or sd presented) Mean AQLQ gain 1.79 (sd1.13, p<0.01) EQ-VAS mean gain: 14.22 (95%CI 9.11 to 19.34) p<0.001 FEV_1 mean improvement: 12.23,	EQ-5D mean gain: 0.142 (95%CI 0.086 to 0.199), p<0.001 for ITT	Yes	Yes for AQLQ and EQ-VAS. Statistical significance not reported for reduction in severe exacerbations.

Study ref Author, Year	**Discrete comparisons**			**Continuous comparison**			
	Proportion achieving change in clinical outcome	*Proportion achieving change in EQ-5D*	*Agreement with direction??*	*Δ clinical measure(s) or other preference based utility*	*Δ EQ-5D*	*Agreement with direction??*	*Agreement with stat sig*
	analysis: Similar results to ITT but marginally better for all outcomes						
Willems et al, 2007[35]				Significant ($p<0.05$) variations over time within 6/8 SF-36 domains and SF-6D (adults only). No significant ($p>0.05$) variation in time for EQ-VAS in adults. Significant ($p<0.05$) variation in time for EQ-VAS in children. No differences between treatment	Significant ($p<0.05$) variation in time for EQ-5D (adults and children). No differences	Results reported at multiple time points and direction of change not consistent over time.	Yes for some but not all.

Study ref Author, Year	Discrete comparisons			Continuous comparison			
	Proportion achieving change in clinical outcome	*Proportion achieving change in EQ-5D*	*Agreement with direction??*	*Δ clinical measure(s) or other preference based utility*	*Δ EQ-5D*	*Agreement with direction??*	*Agreement with stat sig*
				groups in time for SF-36 (adults only) or EQ-VAS (both adults and children) No differences between groups in time for SF-6D (adults only) QALY gains using SF-6D in adults: -0.01 (-0.07 to 0.03)	between groups in time for EQ-5D (adults and children) and SF-6D (adults only) QALY gains for intervention vs control were; Using EQ-5D in adults, 0.03 (0.00 to 0.07) Using EQ-5D in children, 0.01 (0.00 to 0.02)		
Lloyd et al, 2007[26]				Mean change from baseline reported for; a) No exacerbation b) Exacerbation with steroid use c) Hospitalised	 0.02 -0.10 -0.20	 Yes Yes Yes	 p=0.007 for differences between the groups

Study ref Author, Year	Discrete comparisons			Continuous comparison			
	Proportion achieving change in clinical outcome	*Proportion achieving change in EQ-5D*	*Agreement with direction??*	*Δ clinical measure(s) or other preference based utility*	*Δ EQ-5D*	*Agreement with direction??*	*Agreement with stat sig*
				No significant difference in any of the mini-AQLQ domains or ASUI			
Oga et al, 2003[46]				Significant ($p<0.05$) improvements from baseline in AQLQ, FEV_1, SF-36 (all domains) and NHP (3 of 6 domains) for at least one follow-up point. Effect sizes and standardised response means: AQLQ had high responsiveness (over 0.8) in all but one domain (environment) SF-36 was low to	Mean scores were; Baseline 0.808, 3mths 0.887 6 mths 0.879. Both 3mths and 6mths were sig ($p<0.05$) different from baseline Effect sizes and standardised response means: EQ-5D: 0.32 to 0.41 Correlations between changes in EQ-5D and changes in AQLQ were	Yes, generally better scores after baseline for all measures.	Yes, generally although some SF-36 and NHP domains were non-sig at some time points.

Study ref Author, Year	Discrete comparisons			Continuous comparison			
	Proportion achieving change in clinical outcome	*Proportion achieving change in EQ-5D*	*Agreement with direction??*	*Δ clinical measure(s) or other preference based utility*	*Δ EQ-5D*	*Agreement with direction??*	*Agreement with stat sig*
				high (0.28 to 0.95) NHP was low to moderate (0.20 to 0.61)	statistically significant and ranged from 0.37 to 0.45.		
Garratt et al, 2000[48]				Mean changes by self-reported transition e.g much better: Significant linear relationship for NASQ (F=24.90), AQLQ (F=17.79 overall and 4.44 to 25.94 across domains), SF-12 PCS (F=-4.96) but not SF-12 MCS Standardised response means: NASQ -0.82 AQLQ overall	Mean changes by self-reported transition e.g much better: Significant linear relationship for EQ-5D (F=4.07) Standardised response means: 0.29	Yes	Yes

Study ref Author, Year	**Discrete comparisons**			**Continuous comparison**			
	Proportion achieving change in clinical outcome	*Proportion achieving change in EQ-5D*	*Agreement with direction??*	*Δ clinical measure(s) or other preference based utility*	*Δ EQ-5D*	*Agreement with direction??*	*Agreement with stat sig*
				0.55 AQLQ domains 0.32 to 0.77 SF-12 PCS 0.35 SF-12 MCS 0.03			

Only one study reported on re-test reliability for EQ-5D in asthma (ref id 7). The intraclass correlation coefficients for EQ-5D utility were 0.39 (95% CI 0.05 to 0.64) for asthma patients aged 7-12 (n=32) and 0.19 (95%CI -0.44 to 0.69) for asthma patients aged 12-18 (n=11)

Summary of asthma findings
EQ-5D demonstrated validity in the majority of known group comparisons. None of the alternative generic preference based measures (SF-6D, HUI-3) performed consistently better than EQ-5D, however disease specific measures such as AQLQ did show a greater degree of responsiveness than the generic measures. In those studies where there was a significant improvement in clinical or disease specific measures, the EQ-5D was generally found to be responsive.

5.2. URINARY INCONTINENCE

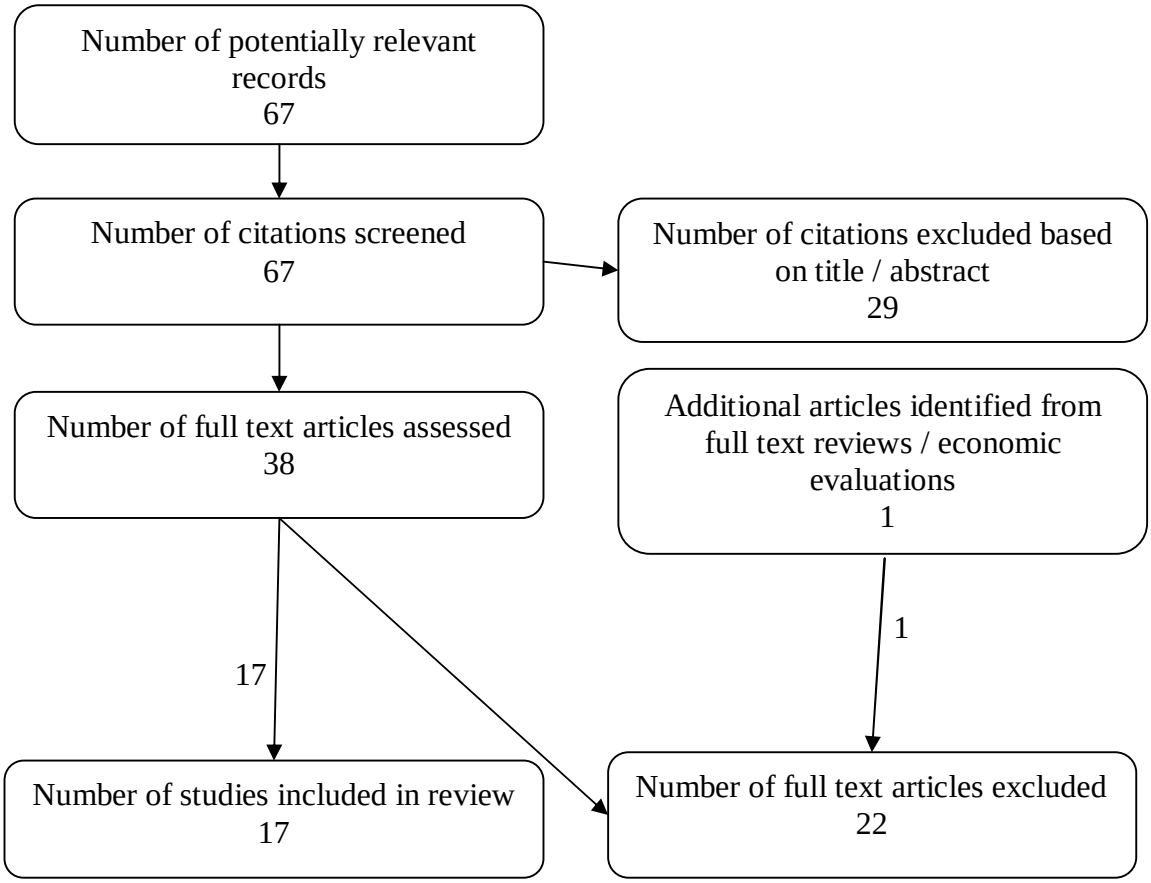
Urinary incontinence has been defined by the incontinence society as "the complaint of any involuntary urinary leakage"[53]. Urinary incontinence can cause embarrassment and can impact on daily activities and quality of life. It can lead to depression, loss of confidence, loss of self-esteem and can carry considerable social and economic costs. Urinary incontinence is often categorised as either stress, urge or mixed. Stress incontinence is associated with effort, exertion, sneezing or coughing, whilst urge incontinence is when leakage is accompanied or immediately preceded by urgency. The term mixed incontinence is used when features of both stress and urge incontinence are present. The objective of this review was to identify all published evidence that reported the use of EQ-5D in people with urinary incontinence in a manner that would provide empirical evidence relating to the performance of EQ-5D, as described in section 2.3 of this report.

5.2.1. Search strategy

The search strategy combined free text terms aimed at identifying papers reporting EQ-5D with free text and controlled terms (MESH and MESH-like terms) for urinary incontinence. The databases searched and date ranges employed were the same as for the asthma review (see 5.1.1) as were the general inclusion / exclusion criteria and the method of sifting papers. Data were extracted using the same set of standardised forms used in the asthma review with data fields relating to asthma (e.g asthma type) replaced with data fields relating to urinary incontinence (e.g incontinence type).

5.2.2. Results

A total of 67 citations were identified from the bibliographic searches. Of these 38 were ordered as full-text articles, although nine papers (four reviews and five economic evaluations) were ordered purely to check their references for further primary studies. From these a further one paper was identified.



A total of 17 papers were included in the review, the key features of which are reported in Table 7 Characteristics of included studies, reporting the validity and responsiveness of generic HRQoL in people with incontinence. Four of the studies identified were RCTs, four were cohort studies and nine were cross-sectional studies . The majority of the studies were conducted in a population with incontinence. Two studies were conducted in the general population sample who were asked about whether they had a range of clinical conditions including incontinence[36;54]. These studies were included as they reported utilities for the subgroup of patients with incontinence. One study identified patients from an academic urology unit inpatient database and examined over active bladder symptoms including incontinence[55]. One study was in men with uncomplicated urinary tract symptoms associated with benign prostatic enlargement[56] and a second study was conducted in outpatients attending urology department with urinary symptoms (not specifically incontinence) and possible benign prostatic obstruction. This study also recruited a general practice sample which was not selected for incontinence[57]. These studies were included as urinary incontinence can be experienced in patients with benign prostatic hyperplasia. Two papers reported different analyses from the PURE study[58;59]. One paper reporting EQ-5D values from a study[39] had a second associated paper[60] which was excluded as it didn't report EQ-5D values, however the EQ VAS values reported in this secondary paper are included in the results table under the primary paper.

Table 7 Characteristics of included studies, reporting the validity and responsiveness of generic HRQoL in people with incontinence

Study ref Author, Year	Country	Type of incontinence (e.g stress, urge)	Disease/treatment stage	Treatment (if any)	Study type (e.g. cross sectional, RCT, cohort)
Ternent et al, 2009[61]	UK	Stress incontinence	No details on how stress urinary incontinence defined	No details	Cross sectional (self-selected sample)
Ismail et al, 2009[62]	UK	Urodynamic stress incontinence	Urodynamic tests complied with definitions of the International Continence Society. Duration of symptoms, media=60mths (IRQ 33-150)	Magnetic energy stimulation of pelvic floor muscles No comparator arm.	Cohort
Rinne et al, 2008[63]	Finland	Stress urinary incontinence	Indication for surgical treatment. Positive cough stress test DIS score<7 Median duration of symptoms TVT: 8yrs (range 1-30) TVT-O: 9yrs (range 1-30) *p=0.004	a) Tension free vaginal tape (TVT) b) TVT obturator (TVT-O)	RCT of TVT vs TVT-O
Haywood et al, 2008[64]	UK	Women with stress and/or urge incontinence referred for physiotherapy from primary or secondary care.	Duration of symptoms 6.20 years (sd 7.07)	Physiotherapy	Cohort (RCT with combined across arms)

Study ref Author, Year	Country	Type of incontinence (e.g stress, urge)	Disease/treatment stage	Treatment (if any)	Study type (e.g. cross sectional, RCT, cohort)
Monz et al, 2007[58]	15 European Countries (UK and Ireland subgroup)	Urinary incontinence of any type. Subgroups by Stress (29%), Urge (13%), Mixed (58%)	Women seeking treatment for urinary incontinence. 23.4% receiving drug therapy, 23.9% receiving conservative treatment and 14.3% had previous surgery		Cross-sectional data from cohort study
Kobelt et al, 2006[65]	France, Germany, Italy, Sweden, UK	Stress urinary incontinence	No details on how stress urinary incontinence defined	NASHA/Dx gel	Cohort (comparison with TVT not based on comparative trial data and limited to costs only)
Dumville et al, 2006[66]	UK	Stress incontinence	Proven stress urinary incontinence requiring surgery	Laparoscopic vs open colposuspension	RCT
Currie et al , 2006[55]	UK	Subgroups with stress and non-stress incontinence identified.	Patient with incontinence identified from sample who had been treated by urology department.		Cross-sectional
Monz et al, 2005[59]	15 European countries	Women seeking treatment for urinary incontinence	Involuntary leakage of urine in past 12 months. S/UIQ used to define type as stress, urge or mixed	none	Cross-sectional data from a cohort study
Manca et al, 2003[67] (clinical outcomes from Ward 2002)	UK	Stress incontinence	Women selected for surgical management of stress incontinence	Tension free vaginal tape vs colposuspension	RCT
Kobelt, 1997[68]	Sweeden	Patients with mixed	Mean micturitions per		Cross-sectional

Study ref Author, Year	Country	Type of incontinence (e.g stress, urge)	Disease/treatment stage	Treatment (if any)	Study type (e.g. cross sectional, RCT, cohort)
		or urge incontinence who had previously received therapy from a urotherapist.	day 9.04 (sd 3.35) Mean leaks per day 3.47 (sd 3.17)		
Hawthorne, 2009[54]	Australia	General population sample with data on presence of urinary incontinence	Urinary incontinence classified as; None (N=2113) Slight/mild (N=714) Moderate (N=119) Severe (N=61)		Cross-sectional
Tincello et al, 2010[69]	Germany, UK, Sweden & Ireland	Women seeking treatment for stress urinary incontinence	Stress incontinence symptoms with or without urge symptoms (defined by clinical opinion) S/UIQ used to define type as stress, urge or mixed	36.1% receiving conservative management at baseline. 18.0% receiving drug therapy at baseline.	Cross-sectional (baseline data from cohort study)
Saarni, 2006[36]	Finland	Patients from general population self-reporting urinary incontinence			Cross-sectional
Noble et al, 2002[56]	UK	Men with uncomplicated urinary tract symptoms associated with benign prostatic		Laser therapy vs Transurethral prostrate resection Vs conservative management	RCT

Study ref Author, Year	Country	Type of incontinence (e.g stress, urge)	Disease/treatment stage	Treatment (if any)	Study type (e.g. cross sectional, RCT, cohort)
		enlargement			
Mihaylova et al, 2010[70]	Multicountry (Germany, UK & Sweden)	Stress urinary incontinence 40% had pure stress incontinence with the rest reporting both stress and urge incontinence	Treatment initiation or treatment switch at time of enrolment. Those receiving surgical management or other pharma management excluded.	Duloxetine vs conservative management vs duloxetine plus conservative managements vs no treatment	cohort
Donovan et al, 1997[57]	12 countries	Outpatients attending urology department with symptoms (not specifically incontinence) and possible benign prostatic obstruction. GP sample (not selected for condition)			Cross-sectional

DIS=Detrusor instability scores, S/UIQ= Stress and Urge Incontinence Questionnaire

The characteristics of the patients in the included studies are shown in Table 8 Participant characteristics. One study enrolled less than 100 patients[62]. The total number of patients ranged from 48 to 9487. Two papers looked exclusively at males[56;57], four had a mixed population of males and females[36;54;55;68] and the remainder looked exclusively at females.

Table 8 Participant characteristics in included incontinence studies

Study ref Author, Year	Number of participants	Age, mean (range)	%male/ %female	Ethnicity
Ternent et al, 2009[61]	105 (of 188 approached)	56.90 (28-89)	0/100%	Not reported
Ismail et al, 2009[62]	48	51 (sd 13)	0/100%	Not reported
Rinne et al, 2008[63]	267 (of 273 randomised)	TVT: 53 (sd 10) TVT-O: 54 (sd10)	0/100%	Not reported
Haywood et al, 2008[64]	174	52.50 (sd 11.75)	0/100%	Not reported
Monz et al, 2007[58]	9487	60.7 (sd 13.5)	0/100%	Not reported
Kobelt et al, 2006[65]	82 of 139 enrolled	56 years (sd 13)	0/100%	Not reported
Dumville et al, 2006[66]	291		0/100%	
Currie et al , 2006[55]	609 (from 2193 sent survey)	Men: 67.2 (sd 14.9) Women: (59.9 (sd 15.8)	67.7%/32 .3%	White: 96% Mixed: 1% Asian or Asian British: 1% Chinese or other: 0% Missing data: 2%
Monz et al, 2005[59]	9487		0/100%	
Manca et al, 2003[67]	344	Tension free vaginal tape: 50 (IQR 24 -56) Colposuspension: 50 (IQR 45 to 59)	0/100%	Not reported
Kobelt, 1997[68]	461 (541 sent questionnaire)	61.1 (sd 14.0)	5%/95%	Not reported
Hawthorne, 2009[54]	3015	45 (sd19)	49%/51 %	Not reported
Tincello et al, 2010[69]	3739 of 3762 enrolled	58.0 (20.0 to 99.0)	0/100%	Not reported
Saarni, 2006[36]	8028 of which 13.0% reported urinary incontinence	53 in general pop 64 in incontinence group	47%/53 % in general pop 23%/77 % in incontine nce group	Not reported
Noble et al, 2002[56]	340	Not reported	100%/0 %	Not reported
Mihaylova et	1510	56 (sd 13)	0/100%	Not reported

Study ref Author, Year	Number of participants	Age, mean (range)	%male/ %female	Ethnicity
al, 2010[70]				
Donovan et al, 1997[57]	1271 outpatient sample 423 GP sample (UK) (Not all had incontinence)	Not reported	100%/0 %	Not reported

Table 9 shows the measures reported in each of the included studies. In addition to the EQ-5D, five studies administered the SF36 or some variant of it[55;57;66-68]. One included SF-6D, AQoL, AQoL-8, and HUI-3[54] and one reported the 15-D[36]. Several papers reported using the UK valuation set for the EQ-5D and none reported using an alternative valuation set, although it was common for this information not to be reported. Only two studies reported the EQ VAS[58;68].

The main clinical measures reported were severity, or grade of incontinence, type of incontinence (stress / urge/ mixed), frequency of leakage episodes and pad usage or pad tests to determine volume of leakage. Some studies reported on cough stress tests or cystometry results. In the benign prostatic hyperplasia populations maximum flow rate and post void residual volume were used as measures of treatment effectiveness.

Various symptom scoring and incontinence specific quality of life tools were also used (KHQ, UISS, I-QOL, IIQ-7, SSI). Some studies included tools which were designed for use in patients with overactive bladder rather than incontinence (UDI-6, BLUTS). Some studies included scales designed to measure the impact of lower urinary tract symptoms in men (ICSQoL, IPSS). One study reported a questionnaire that assesses the likelihood of destrusor instability which may be associated with stress incontinence, based on patient history (DIS). One study reported quality of life using a patient generated index (PGI) which is an individualised health related quality of life measures.

Table 9 Measures used in the included incontinence studies

	GENERIC MEASURES			OTHER MEASURES USED				
Study ref Author, Year	**Descriptive system (EQ-5D, HUI2, HUI3, SF-6D)**	**Tariff used (state tariff, or provide methods & source of valuation if no tariff used)**	**Mean value (SD, range)**	**Details of direct valuation (own health or vignettes? Whose health? Whose values? Which valuation method (TTO, VAS, SG)?)**	**Condition-specific HRQL measures used**	**Clinical measures used**	**Qualitative questions (provide details of any qualitative questions asked)**	**Missing data; completion rates of measures; patients completing study etc. (include reasons for non-completion if given)**
Ternent et al, 2009[61]	EQ-5D	Not stated	0.598 (0.339, -0.17 to 1)	None	KHQ, PGI	None	None	73/105 completed PGI correctly 101/105 completed EQ-5D
Ismail et al, 2009[62]	EQ-5D	Not stated	Median:0.812 (IQR 0.656 to 0.919) at baseline No sig difference at end of treatment or follow-up	None	KHQ	1hr pad test Leakage episodes Pad usage	None	31/48 completed all treatments 27 attended 3mth follow-up
Rinne et al, 2008[63]	EQ-5D	Not stated	Baseline: a) 0.8	None	UISS DIS VAS	Cough stress test 24-hr pad	Satisfaction with operation.	5 excluded post randomisation 4 refused, 1

	GENERIC MEASURES			OTHER MEASURES USED				
Study ref Author, Year	**Descriptive system (EQ-5D, HUI2, HUI3, SF-6D)**	**Tariff used (state tariff, or provide methods & source of valuation if no tariff used)**	**Mean value (SD, range)**	**Details of direct valuation (own health or vignettes? Whose health? Whose values? Which valuation method (TTO, VAS, SG)?)**	**Condition-specific HRQL measures used**	**Clinical measures used**	**Qualitative questions (provide details of any qualitative questions asked)**	**Missing data; completion rates of measures; patients completing study etc. (include reasons for non-completion if given)**
			8 5 b) 0. 8 7 6 1 year: a) 0. 9 0 1 b) 0. 9 3 3 No SD's or p values provided		IIQ-7 UDI-6			cancelled, 1 switched TVT-O to TVT. 2 lost to follow-up
Haywood et al, 2008[64]	EQ-5D	States general population utility weights.	0.81 (0.24 sd)	None	I-QoL (index and individual domains)	SSI Incontinence episodes per week at baseline	Subjective treatment benefit assessed by patient.	85% had 6 week and 79% had 5mth follow-up data.

	GENERIC MEASURES			OTHER MEASURES USED				
Study ref Author, Year	**Descriptive system (EQ-5D, HUI2, HUI3, SF-6D)**	**Tariff used (state tariff, or provide methods & source of valuation if no tariff used)**	**Mean value (SD, range)**	**Details of direct valuation (own health or vignettes? Whose health? Whose values? Which valuation method (TTO, VAS, SG)?)**	**Condition-specific HRQL measures used**	**Clinical measures used**	**Qualitative questions (provide details of any qualitative questions asked)**	**Missing data; completion rates of measures; patients completing study etc. (include reasons for non-completion if given)**
						(none, a few days, half the week, most days, every day)		
Monz et al, 2007[58]	EQ-5D	Not stated	Not reported	EQ-VAS	I-QOL	UI severity (Sandvik Index) UI subtype (S/UIQ)	Bother (4 point scale)	Some patients excluded when data not available e.g EQ-5D analysis based on 6978 patients.
Kobelt et al, 2006[65]	EQ-5D	Reference suggests UK tariff used.	Baseline: 0.820 (sd 0.18) 3mths: 0.868 (sd 0.14) 12mths: 0.834 (sd 0.19)	None	None	Incontinence grade Median number of episodes per day		139 enrolled 105 trial completers 82 providing EQ-5D data
Dumville	EQ-5D	UK Tariff	Baseline:	None	None	Objective	Subjective cure*	5 withdrew after

	GENERIC MEASURES			OTHER MEASURES USED				
Study ref Author, Year	**Descriptive system (EQ-5D, HUI2, HUI3, SF-6D)**	**Tariff used (state tariff, or provide methods & source of valuation if no tariff used)**	**Mean value (SD, range)**	**Details of direct valuation (own health or vignettes? Whose health? Whose values? Which valuation method (TTO, VAS, SG)?)**	**Condition-specific HRQL measures used**	**Clinical measures used**	**Qualitative questions (provide details of any qualitative questions asked)**	**Missing data; completion rates of measures; patients completing study etc. (include reasons for non-completion if given)**
et al, 2006[66]	SF-36 (reported in related clinical paper)*		Laparoscopic, 0.827 Open 0.824 24mths: Laparoscopic, 0.844 Open, 0.825			cure* (negative 1 hr pad test)	(perfectly happy / pleased) to spend rest of life with current urinary symptoms	randomisation leaving 286
Currie et al , 2006[55]	EQ-5D SF-36	Not stated	Stress incontinence: 0.57 (sd 0.331) Incontinence other than stress incontinence: 0.625 (sd 0.317)	None	None	None	None	
Monz et al, 2005[59]	EQ-5D	Not stated	Median across all enrolled	None	I-QOL	Sandvik index (severity based on	Bothersomeness and limitations of daily activities	Maximum excluded from analysis is 5.3%

	GENERIC MEASURES			OTHER MEASURES USED				
Study ref Author, Year	**Descriptive system (EQ-5D, HUI2, HUI3, SF-6D)**	**Tariff used (state tariff, or provide methods & source of valuation if no tariff used)**	**Mean value (SD, range)**	**Details of direct valuation (own health or vignettes? Whose health? Whose values? Which valuation method (TTO, VAS, SG)?)**	**Condition-specific HRQL measures used**	**Clinical measures used**	**Qualitative questions (provide details of any qualitative questions asked)**	**Missing data; completion rates of measures; patients completing study etc. (include reasons for non-completion if given)**
			0.85. Mean UK:0.73 Median UK:0.81 (sd0.30), n=1070 Median by type: Stress: 0.85 Mixed: 0.81 Urge: 0.85			frequency and leakage amount)		for any country.
Manca et al, 2003[67]	EQ-5D SF-36	UK tariff	Tension free vaginal tape Baseline: 0.778 6 weeks: 0.788 6mths:			Objective cure (based on negative pad test and negative cystometry) Subjective cure (based on		Economic analysis uses 214 who had all complete EQ-5D data and data on theatre length of stay

	GENERIC MEASURES			OTHER MEASURES USED				
Study ref Author, Year	**Descriptive system (EQ-5D, HUI2, HUI3, SF-6D)**	**Tariff used (state tariff, or provide methods & source of valuation if no tariff used)**	**Mean value (SD, range)**	**Details of direct valuation (own health or vignettes? Whose health? Whose values? Which valuation method (TTO, VAS, SG)?)**	**Condition-specific HRQL measures used**	**Clinical measures used**	**Qualitative questions (provide details of any qualitative questions asked)**	**Missing data; completion rates of measures; patients completing study etc. (include reasons for non-completion if given)**
			0.806 Colposuspension Baseline: 0.785 6 weeks: 0.754 6mths: 0.794			BFLUTS)		
Kobelt, 1997[68]	EQ-5D SF-36	UK Tariff.		EQ-VAS* *reported in associated paper by Johannesson 1997		Frequency of micturitions and involuntary urine loss (combined measure)		461 responded of 541 sent questionnaire
Hawthorne, 2009[54]	EQ-5D SF-6D AQoL AQoL-8 (derived from	EQ-5D: UK tariff SF-6D: Not stated but Brazier 2002 referenced. AQoL & AQoL-	Reported by continence status. See validity					

	GENERIC MEASURES			OTHER MEASURES USED				
Study ref Author, Year	**Descriptive system (EQ-5D, HUI2, HUI3, SF-6D)**	**Tariff used (state tariff, or provide methods & source of valuation if no tariff used)**	**Mean value (SD, range)**	**Details of direct valuation (own health or vignettes? Whose health? Whose values? Which valuation method (TTO, VAS, SG)?)**	**Condition-specific HRQL measures used**	**Clinical measures used**	**Qualitative questions (provide details of any qualitative questions asked)**	**Missing data; completion rates of measures; patients completing study etc. (include reasons for non-completion if given)**
	AQoL) HUI-3 (deciles)	8: community TTO	results below.					
Tincello et al, 2010[69]	EQ-5D	UK tariff	Mean for UK (n=553), 0.73 (sd 0.29) (varied significantly by country)	None	None	Episodes per week	None	23 (<1%) excluded from analysis
Saarni, 2006[36]	EQ-5D 15-D	UK tariff for EQ-5D 15-D Finnish valuation set	For incontinence subgroup; EQ-5D: 0.693 (SE 0.010) 15-D: 0.835 (SE 0.004)		None	None	None	
Noble et	EQ-5D	Not stated	Mean at	None	I-PSS which	Maximum		

	GENERIC MEASURES			OTHER MEASURES USED				
Study ref Author, Year	**Descriptive system (EQ-5D, HUI2, HUI3, SF-6D)**	**Tariff used (state tariff, or provide methods & source of valuation if no tariff used)**	**Mean value (SD, range)**	**Details of direct valuation (own health or vignettes? Whose health? Whose values? Which valuation method (TTO, VAS, SG)?)**	**Condition-specific HRQL measures used**	**Clinical measures used**	**Qualitative questions (provide details of any qualitative questions asked)**	**Missing data; completion rates of measures; patients completing study etc. (include reasons for non-completion if given)**
al, 2002[56]			71/2 mths follow-up: Laser, 0.790 Resection, 0.816 Conservative, 0.772		includes a quality of life score.	flow rate Post void residual urine Number of successful procedures (based on I=PSS and maximum urinary flow)		
Mihaylova et al, 2010[70]	EQ-5D	UK tariff	Baseline EQ-5D: 0.80 (sd 0.23)			Number of leaks during 7 days		11.9% missed observation 3 13.7% missed observation 4
Donovan et al, 1997[57]	EQ-5D (UK, Denmark and Netherland only, N=359) SF-36 (UK only, N=205)	Not reported	0.79		ICSQol (ICSmale)			Some tools were only applied in certain countries. See under descriptive system for numbers.

KHQ=King's Health Questionnaire, PGI = patient generated index, UISS=urinary incontinence severity score, DIS= Detrusor instability scores, VAS=visual analogue scale, IIQ-7=incontinence impact questionnaire-short form, UDI-6=urogenital distress inventory-short form, I-QOL=Incontinence specific Quality of life Questionnaire, ICSQol=International Continence Society – Benign Prostatic Hyperplasia study Quality of Life Instrument. I-PSS = International prostate symptom score. SSI= symptom severity index, S/UIQ=Stress and Urge incontinence questionnaire. BFLUTS=Bristol female lower urinary tract symptoms questionnaire

Table 10 provides a summary of those studies that compared the mean EQ-5D between groups defined in terms of incontinence severity, frequency or subtype.

Two studies defined groups by the frequency of incontinence episodes[64;69]. In one study, three groups were defined and the mean EQ-5D consistently reflected differences between groups and the differences were statistically significant[69]. In the second study, five groups were defined[64]. The mean EQ-5D was equal for two of the groups and the differences between all the five groups was not statistically significant. In the same study, the condition specific measures of SSI and I-QoL discriminated well between the groups.

Two studies reported known group validity by severity group. The definition of severity was not well described in Hawthorne 2008[54], but in Monz 2005[59] a validated severity index was used which was based on combined scores for frequency and leakage amount. EQ-5D varied between severity groups as expected in both studies and had statistically significant differences between severity groups in one study[54], whilst the other did not report whether differences were statistically significant[59]. Other preference based measures (SF-6D, AQoL & AQoL-8) generic measures (EQ-VAS) and disease specific measures (I-QoL) were found to perform equally well.

Three studies compared groups defined by incontinence type with two studies distinguishing between stress, urge and mixed incontinence[59;69] and the other study grouping patients as general incontinence, stress incontinence or none[55]. It was unclear what differences were clinically expected between the stress, urge and mixed groups. However, two studies reported greater EQ-5D scores for stress incontinence than for urge and greater utilities for urge than for mixed[59;69]. These differences were statistically significant in one study and the other did not report statistical significance. EQ-VAS had differences across the group that were consistent with the differences for EQ-5D except for when severity was reported as slight. Mean I-QoL score performed similarly to EQ-5D although the differences between the groups were not consistent for individual I-QoL domains.

In the third study EQ-5D scores were lower for general incontinence than for none as clinically expected, but statistical significance was not reported[55]. SF-36 performed equally well in distinguishing between general / stress/ none.

Table 10: Results of "known groups" comparisons in incontinence studies

Study ref Author, Year	*Groups defined as*	*Instrument*	*Direction of change consistent across groups?*	*Direction change consistent with clinical expectation*	*Difference between groups statistically significant?*
Haywood et al, 2008[64]	Number of episodes at baseline: Not at all A few days Half the week Most days Every day	 SSI I-QoL index I-QoL domains EQ-5D	 Yes Yes Yes except for I-QoL PIS Yes but same mean for two least severe groups	 Yes Yes Yes except for I-QoL PIS Yes but same mean for two least severe groups	 Yes Yes Yes No
Tincello et al, 2010[69]	Episode frequency: <=7 per week 7 to 12 per week >=14 per week	EQ-5D	Yes	Yes	Yes
Monz et al, 2005[59]	Severity (reported for each subtype) Slight Moderate Severe Very severe	 EQ-5D EQ-VAS Mean I-QoL I-QoL domains	 Yes Yes Yes Yes	 Yes Yes Yes Yes	 Not reported Not reported Not reported Not reported
Hawthorne, 2009[54]	Continence status: a) None b) Slight/mild c) Moderate d) Severe	 EQ-5D SF-6D AQoL AQoL-8	 Yes Yes Yes Yes	 Yes Yes Yes Yes	 Yes Yes Yes Yes
Currie et al, 2006[55]	Type of incontinence:	EQ-5D	General lower than none and	Yes in that general is lower than	Not reported

	General Stress None	SF-36	stress lower than non-stress As for EQ-5D	none. As for EQ-5D	As for EQ-5D
Monz et al, 2005[59]	Subtype (reported for each severity category): Stress Urge Mixed	EQ-5D EQ-VAS Mean I-QoL I-QoL domains	Stress>urge>mixed As for EQ-5D (except when severity slight) As for EQ-5D Not consistent pattern across all domains	Unclear what clinical expectation is of difference in type	Not reported Not reported Not reported Not reported
Szende et al, 2004[45]	UI subtype: Mixed Pure stress Pure urge	EQ-5D	Stress>urge>mixed	Unclear what clinical expectation is of difference in type	Yes

The degree of association between EQ-5D and a range of incontinence measures is reported in Table 11.

Correlations of EQ-5D against a variety of disease specific instruments (KHQ, PGI, I-QoL, ICS-QoL, SSI) and clinical measures (incontinence grade and number of micturitions / leakages) are provided in Table 11. Significant correlations were seen for many of the disease specific instruments, although the size of correlation varied. SSI was found to have a minimal (-0.09) and non-significant correlation with EQ-5D. The correlation with the individual ICS-QoL items varied from -0.04 and non-significant to -0.25 and significant. Correlation with the I-QoL index and domains were small to moderate (0.28 to 0.37). Small but significant correlations were found with incontinence grade (0.13) and number of micturitions and leakages (-0.20) for combined outcome.

Two studies used regression techniques to assess the impact of clinical measures on EQ-5D scores. Severity, subtype of incontinence (e.g stress / urge) and number of episodes were found to be significant predictors[58;69]. One study found that presence of incontinence was a significant predictor of EQ-5D in urology patients[55]. A second study found that incontinence was a significant predictor of both EQ-5D and 15D in a general population sample and the size of utility loss was similar between these two instruments[36].

Table 11: The relationship between EQ-5D and other measures in incontinence

Study ref Author, Year	*Clinical/non pref based measure*	*Correlation with EQ-5D*	*Regression*
Ternent et al, 2009[61]	PGI mean score KHQ domains	Positive Negative (values not given) Both expected direction and significant	
Haywood et al, 2008[64]	SSI I-QoL index I-QoL domains	-0.09 0.37 0.28 to 0.37 SSI was in expected direction but not significant at p=0.05. I-QoL index and domains were significant at p<0.001and in expected direction	
Monz et al, 2007[58]			Ordinal logistic regression used to determine relationship between EQ-5D index scores as categorical variable (1, <0.727, & >=0.727 to <1) and severity and type of stress urinary incontinence and other potentially confounding factors Severity and subtype were significant (p<0.0001) predictors of EQ-5D category.
Kobelt et al, 2006[65]	Incontinence grade	Grade 0 or 1 had utility 0.133 higher than more severe grades (p<0.05)	
Currie et al, 2006[55]			Multivariate regression used to assess impact of incontinence on mean EQ-5D and SF-36 domains (controlling for sex, age, BMI) Incontinence reduced EQ-5D by 0.107 (SEM 0.026)

Study ref Author, Year	*Clinical/non pref based measure*	*Correlation with EQ-5D*	*Regression*
			p<0.001. Also significantly (p<0.05) reduced SF-36 general health perception score and all 8 domains.
Kobelt, 1997[68]	Number of leakages and micturitions (combined outcome)	-0.20 P<0.001	
Szende et al, 2004[45]			Multivariate logistic regressions on odds of having utility=1: Number of episodes and type (pure stress vs other) were significant predictors of utility <1 (p<0.0001) Multivariate linear regression on predictors of EQ-5D index score: Similar results to logistic regression (no further details)
Saarni et al, 2006[36]			Utility loss associated with incontinence based on multivariate regression controlling for sociodemographic variables and other chronic conditions: 15D: -0.029 (SE 0.003) p<0.01 EQ-5D: -0.029 (SE 0.006). p<0.01
Mihaylova et al, 2010 [70]			Multivariate regression model for QALY at 1 year based on EQ-5D utility. Duloxetine alone (p<0.01) and duloxetine plus conservative treatment (p<0.05) were significant predictors of QALY but conservative alone was not (reference group was no treatment)

Study ref Author, Year	*Clinical/non pref based measure*	*Correlation with EQ-5D*	*Regression*
Donovanet al 1997[57]	ICS QoL items - Need to change clothes - Reduce drink intake -Interference with life -Time with symptoms -Satisfaction with rest of life with current LUTS	 -0.16, p<0.01 -0.10 (p<0.1) -0.25, (p<0.001) -0.04, (p>0.1) -0.22, p<0.001 All expected directions	

Table 12 reports results from studies that provide details on the responsiveness of EQ-5D in incontinence.

Five studies reported change in EQ-5D from baseline and compared this to changes in diseases specific or clinical measures[56;62;63;65;67]. Generally there was agreement between changes in EQ-5D and changes in clinical or disease specific measures with four studies reporting improvements in both[56;63;65;67] although two studies did not report whether the EQ-5D changes were statistically significant[56;67]. In one study there was no significant change in either EQ-5D or clinical outcomes[62].

One study reported changes from baseline for patients whose continence-specific health improved[64]. In this subgroup significant changes from baseline were seen in SSI and I-QoL, but not EQ-5D at six weeks. However, by five months when greater changes from baseline were seen for SSI and I-QoL, the EQ-5D changes were also found to be larger and statistically significant. This study also reported mean scores for responders and non-responders with response being based on patient perceived benefit. There were significant differences between responders and non-responders in two of the I-QoL domains at six weeks, but differences in SSI, I-QoL index and EQ-5D were non-significant. However, by five months EQ-5D differences were found to be significant although only one I-QoL domain remained significantly different between responders and non-responders.

Five studies reported whether the difference between treatment groups was significant for both EQ-5D and for other measures (clinical, disease specific, generic HRQoL)[56;63;66;67;70]. In three studies there were no significant differences in EQ-5D between treatment groups and this agreed with the other trial outcomes[63;66;67]. In one of these studies some significant differences were found in some domains of the SF-36 but not in the other clinical outcomes (objective and subjective cure rates)[67]. One study found differences in EQ-5D scores between the treatment arms that were consistent with the clinical outcomes, but the statistical significance of the EQ-5D differences was not reported[56]. In another study six comparisons were made between the four treatment options (three active and one no treatment)[70]. For the three comparisons of active treatment against no treatment, all three active treatments were more clinically effective than no treatment but only two had significantly better EQ-5D scores. For the three comparisons between the active treatment arms, no significant differences were seen in the clinical effectiveness, but there were significant differences in the EQ-5D scores for two comparisons.

One study reported standardised response means for different instruments[64]. The standardised response means were lower for EQ-5D than for disease specific measures (SSI and I-QoL)

Table 12: EQ-5D responsiveness in asthma

Study ref Author, Year		**Continuous comparison**			
	Comparison	*Δ clinical measure(s) or other preference based utility*	*Δ EQ-5D*	*Agreement with direction??*	*Agreement with stat sig*
Ismail et al, 2009[62]	Change over time	No significant change on any measure (KHQ,1hr pad test, pad use, leakage episodes)	No significant change	NA	Yes
Rinnie et al, 2008[63]	Change over time	24hr pad test significantly improved in both arms All condition specific measures (UISS, DIS, VAS, IIQ-7, UDI-6) significantly improved in both treatment groups EQ-VAS significantly improved in both treatment groups	Significant improvement in both arms	Yes	Yes
	Difference between treatment arms	No sig difference in objective cure, leakage, complication rate, UISS, DIS, VAS, IIQ-7, UDI-6.	No sig difference in EQ-5D	Agreement with some clinical outcomes and not others.	Yes
Haywood et	Comparison of means				

Study ref Author, Year		**Continuous comparison**			
	Comparison	*Δ clinical measure(s) or other preference based utility*	*Δ EQ-5D*	*Agreement with direction??*	*Agreement with stat sig*
al 2008[64]	for responders and non-responders	6 week data: SSI and I-QoL index had difference in expected direction but not statistically significant (at p=0.01). Two of the I-QoL domains had significant difference. 5 mth data: As for 6 weeks except only one of the I-QoL domains had significant (p<0.01) difference.	6 week data: EQ-5D had difference in expected direction but not statistically significant (at p=0.01). 5mth data: EQ-5D had difference in expected direction and statistically significant (p=0.01).	6 weeks: Yes 5mths: Yes	6 weeks: Not consistent with all 5mths: Not consistent with all.
	Mean change scores for patients reporting improvement	6 week data: Expected direction and significant (at p=0.05) for SSI, I-QoL index, I-QoL domains 5mth data: As for 6 weeks but larger changes.	6 week data: Expected direction but p>0.05 5mth data: Expected direction and p<0.05.	6 week data: Yes 5mth data: Yes	6 week data: No 5mth data: Yes
	MSRM for patients reporting improvement	6 week data: SSI, 0.70 I-QoL index, 1.01	6 week data: 0.07	6 week data: Yes	6 week data: No

Study ref Author, Year		**Continuous comparison**			
	Comparison	*Δ clinical measure(s) or other preference based utility*	*Δ EQ-5D*	*Agreement with direction??*	*Agreement with stat sig*
		I-Qol domains, 0.40 to 0.94 5mth data: SSI, 0.67 I-QoL index, 1.17 I-Qol domains, 0.80 to 1.25	 5mth data: 0.26	 5mth data: Yes	 5mth data: Yes
Kobelt et al, 2006[65]	Median incontinence episodes per day for clinical outcome but change from baseline for EQ-5D	(All patients): 3.0 at baseline, 0.7 at 3mths and 0.9 at 12 mths (p<0.0001 and p<0.001 for differences)	All patients: 3mths: 0.048 (p<0.001) 6mths: 0.014 (not significant) 12mths: "gain remained evident" Patients with utility<1 at baseline: 3mths: 0.099 (p<0.01) 6mths: 0.065 (p<0.001) 12mths: "significant improvements"	All patients 3mths: Yes 12mths: Yes Patients with utility <1 at baseline: As for all patients	All patients 3mths: Yes 12mths: Yes Patients with utility <1 at baseline: As for all patients
Dumville et al, 2006[66]	Difference between treatment arms:	Objective and subjective cure rates and SF-36 scores showed no significant difference	QALY gain based on EQ-5D utility scores showed no significant difference (CrI crossed 0)	No change in either clinical, generic HRQoL or utility	Yes
Manca et al, 2003[67]	Differences from baseline to 6mths	Pad weight decreased significantly for both groups. Significant reduction in leakage episodes in both groups (P<0.0001)	Utility increased in both arms (significance not reported)	Yes	Not reported

Study ref Author, Year		**Continuous comparison**			
	Comparison	*Δ clinical measure(s) or other preference based utility*	*Δ EQ-5D*	*Agreement with direction??*	*Agreement with stat sig*
		Significant reduction in 21/30 symptoms (BFUTS) in both groups (P<0.0001)			
	Differences between trial arms:	No significant difference in objective or subjective cure rate between trial arms SF-36 scores had significantly smaller improvement/ greater decline lower for colposuspension group vs TVT in four domains at 6 weeks and four domains (three same and one different) at 6 mths.	QALY difference between arms based on EQ-5D scores non significant at p=0.05	Agreement with clinical outcomes but didn't detect differences between arms in some SF-36 domains	Yes for clinical outcomes, no for some SF-36 domains
Noble et al, 2002 [56]	Change from baseline:	Improvements in I-PSS, maximum urine flow, and residual volume were significant (p=0.05) for laser and resection but not conservative. Improvements in I-PSS QoL were significant for all three interventions.	Means increased for laser and resection but not conservative. (p values not reported)	Yes	Not reported

Study ref Author, Year		**Continuous comparison**			
	Comparison	*Δ clinical measure(s) or other preference based utility*	*Δ EQ-5D*	*Agreement with direction??*	*Agreement with stat sig*
	Differences between trial arms:	Resection vs conservative and laser vs conservative showed significant difference in all four outcomes. Laser vs resection showed significant difference in only one outcome which was in favour of resection (maximum flow)	Gains were greater for resection than laser therapy (p values not reported)	Yes	Not reported
Mihaylova et al, 2010[70]	Comparison between active treatment arms and no treatment:	Number of leaks avoided per week was significantly ($p<0.01$) better for Duloxetine alone, conservative alone and duloxetine plus conservative (all relative to no treatment).	QALY gains based on EQ-5D utility were significant for Duloxetine alone ($p<0.01$) and duloxetine plus conservative treatment ($p<0.05$) but conservative alone was not significant and was negative (all compared to no treatment)	Yes for two of three comparisons against no treatment	Yes for two of three comparisons against no treatment

Study ref Author, Year		**Continuous comparison**			
	Comparison	*Δ clinical measure(s) or other preference based utility*	*Δ EQ-5D*	*Agreement with direction??*	*Agreement with stat sig*
	Comparison between the three active treatment arms:	No significant reduction in number of leaks for 3 comparisons between active treatment arms.	Significant ($p<0.05$) QALY gains for 2 of 3 comparisons between active treatment arms.	Yes for 2 of 3 comparisons between active treatment arms.	No for 2 of 3 comparisons between active treatment arms.

MSRM=modified standardised response mean

Key findings on re-test reliability

One study[64] reported the intraclass correlation coefficient (ICC) for patients reporting no benefits from treatment during a clinical trial (data from both trial arms combined). The test-retest correlation for EQ-5D was 0.83 (n=50).

Summary of findings for incontinence
There is no strong evidence to suggest that EQ-5D is not an appropriate outcome measure for use in economic evaluation in this patient group. In most situations EQ-5D performs well when assessed by known groups validity or responsiveness. In most of the tests performed, EQ-5D was consistent with clinical or disease specific outcome measures, including in achieving statistical significance. However, there were situations where statistical significance was not achieved. This is not surprising since the correlation between EQ-5D and these instruments tended to be moderate at best. This may indicate that no single disease specific measure provides a full description of health related quality of life when used in isolation. It is not necessarily an indication that EQ-5D is insufficiently sensitive.

5.3. RHEUMATOID ARTHRITIS

RA is a chronic, systemic auto-immune inflammatory condition that affects approximately 0.8% of the total adult UK population. The disease progresses with wide variability, causing painful swelling and damaging cartilage and bone around the joints, particularly those in the hands, feet and wrists. Since it is a systemic condition, other parts of the body can be affected including the eyes, lungs and heart. It is characterised by "flares" of inflammation: pain, stiffness and fatigue which can come and go with unpredictable frequency and duration.

In recent years, treatment advances in the form of biologic drugs have been made. These treatments are expensive and are therefore natural candidates for economic evaluations. Indeed, at the time of writing NICE had conducted technology appraisals of nine such drug treatments. In many cases some aspects of the use of EQ-5D have proved controversial. In particular, the relationship between key clinical outcomes (typically the Health Assessment Questionnaire – HAQ) and EQ-5D.

Search Strategy
Our initial searches were devised to identify all studies that reported the use of EQ-5D in an adult RA population or a mixed population that was reported in a manner that separated RA from other conditions.
A total of 241 studies were identified as potentially relevant from initial searches. RA is an extremely active area of clinical research and there was evidence of a large number of studies of relevance to this review. We did however, identify a recent review article that considered the issues of validity and responsiveness of generic utility measures in RA[71]. We therefore provide a narrative account of the relevant features of this study and supplement the findings with studies published since their review was conducted.

In addition, we provide a review of studies that have statistically modelled the relationship between EQ-5D and the HAQ, since this is the issue relating to EQ-5D

that is of most relevance to the economic evaluation of technologies in this disease area.

Included studies

Harrison et al. (2008[71])

The aim of this paper was to review evidence on the validity and comparative performance of generic utility scales in RA. It is therefore a slightly different aim to this report which is focussed on the EQ-5D. They reviewed literature published up until mid 2006 and included 26 papers in total. Whilst it is not entirely clear from the paper, the relatively low numbers of papers may be due to the application of more stringent inclusion/exclusion criteria than have been applied in our reviews of incontinence and asthma. In particular, it seems that only studies that specifically focussed on validation or methodological studies of utilities, or economic evaluations that reported the source of utilities were included. In our reviews we have included all studies that report |EQ-5D in a manner that permits relevant comparisons to be constructed. However, as was apparent with our initial searches in RA, this is a substantial task in RA and it is questionable whether such a review would really contribute more than the focussed approach taken by Harrison et al.

The review focussed on three broad concepts. First is "feasibility" which considers practical issues associated with a measure such as time to complete and response rates, as described in section 2.3.1 above. Since there is little controversy regarding EQ-5D in this respect we do not consider the findings in this area further. The second set of concerns is labelled "truth" and refers to content and construct validity. Finally "discrimination" is the term used to describe reliability and responsiveness. Within the section on responsiveness, there is a focus on the minimum important differences (MID) for alternative scales and different measures of effect sizes. As has been referred to throughout this report, this type of assessment of a measure is of particular value in calculating required sample sizes for studies but may be of more limited relevance in assessing whether the measure is "inappropriate".

Linde et al.

Published after the Harrison review, this cross sectional study compares the validity, reliability and responsiveness of generic and disease specific measures in 200 Danish patients with RA.

Construct validity

Correlations

The degree of correlation between EQ-5D and other generic measures (SF-6D, HUI-2 and HUI-3) has been explored in three studies[72-74], with strong correlations reported (range from 0.59 to 0.70). In severe RA there was evidence that the relationship between EQ-5D and SF-6D was less strong.

The EQ-5D also had strong correlation with the RAQoL, as did all the other generic measures[75]. This same study demonstrated significant and strong correlations with pain, global assessments and the HAQ. It also correlated well with self-reported disease severity and as strongly as the disease specific measures of RAQoL and HAQ.

In Linde et al (2008)[76] the correlation between EQ-5D and HAQ was -0.79 and strong correlations were also reported with all the other measures included in the study which were a range of disease specific and generic measures.

Known groups
Other studies included in the review report the ability of EQ-5D to distinguish between patients defined using a range of approaches relating to health status, disability, social support and employment status[74;77-81].

Linde et al (2008)[76] showed that EQ-5D distinguished statistically significant differences between groups defined in terms of RA activity, whether patients were in receipt of disability pensions or not and between Low and moderate DAS28 (a measure of disease activity) groups. The difference between groups with and without bone erosions was consistent but not statistically significant. The difference between high and moderate DAS28 was zero, although the former group contained just 12 patients. The results for EQ-5D were comparable to those for disease specific measures such as RAQoL and HAQ, as well as the 15D.

Responsiveness
When considering changes in the same patients over time, the EQ-5D demonstrated consistent responsiveness and better performance than the other generic measures when health deteriorated[72;82]. In one study EQ-5D was also the most responsive in detecting improvements in health (ref 41). However, in other studies the most responsive instrument varied according to the time of follow up and the definition of change[82;83].

Linde et al (2008)[76] consider responsiveness from a subgroup of 96 patients in their cohort. Patients were considered to have either improved, not changed or deteriorated based on the change in the patient reported changes in RA at 6 months. The EQ-5D showed consistent results that were statistically significant from baseline for those that had improved, and were not for the other two groups i.e. as expected for the no change group.

Reliability
Linde et al (2008)[76] report that in 87 patients that reported no change from baseline at two weeks, the intraclass correlation coefficient (ICC) was 0.79 (95% CI 0.68 to 0.87). This was lower than for some measures, such as HAQ (0.97) and RAQoL (0.96) but higher for mental and social component scores of the SF36. The mean change was not significantly different from zero.

The Harrison review reports a lower reliability for EQ-5D compared to SF-6D and HUI3 based on the ICC in four studies which compared the same patients over varying follow up periods[72;74;82;83] although the ICC figures were relatively high (0.46 to 0.66). The follow up periods ranged from one week to one year. Other studies reported high ICCs for the EQ-5D over two weeks (0.78) and three months (0.73)[78]. Test-retest scores reported in Conner-Spady et al (2003)[72] found high stability for EQ-5D and other generics at three months and one year in patients that reported no change in their health.

Summary of findings in RA

There is little evidence to suggest that EQ-5D is not an appropriate measure for use in RA. There is widespread reporting of high correlations between EQ-5D and both other generic instruments and condition specific measures such as HAQ. Indeed, widespread concerns within the NICE appraisals of various RA interventions have focussed on the issue of the nature of the relationship between EQ-5D and HAQ with the implied criticism that EQ-5D is inadequate. In fact, it is likely HAQ that is deficient as an outcome measure in RA since its focus is functional impairment and does not include pain. Hernández Alava et al (2010)[84] have demonstrated the very strong relationship when using both HAQ and pain to estimate EQ-5D. Indeed, the full HAQ instrument includes both the HAQ disability index ***and*** pain measured on a VAS.

Evidence for construct validity may also be seen in the comparisons of known groups. EQ-5D was only found to fail to distinguish between two of the groups in one study where the sample size was extremely small. This serves to highlight the inherent dangers in summarising results at the study level.

It is interesting to note that whilst EQ-5D was responsive to the extent that changes over time were consistent with what was expected by the authors, this was the most responsive measure of the generics when health deteriorated, but was less responsive in some studies when health improved.

5.4. MRC REVIEW OF VISUAL DISORDERS

The review conducted by the MRC project team into visual disorders[3] was broader in scope than is the case for this report. Whilst our focus is on the EQ-5D, the MRC team searched for studies that included any preference based measure. Nevertheless, 28 of the 32 studies did report the EQ-5D. Indeed, the evidence relating to alternative generic preference based measures was extremely limited in terms of the numbers of studies identified. There were just two studies that included the SF6D[29;41] and these both included the EQ-5D. There were six that included the HUI-3 (refs) but only two of these also included the EQ-5D[26;29].

We do not reproduce the results here but provide a discussion of the findings in relation to EQ-5D, informed by the issues raised in section 2 of this report and the remit of DSU report, to inform decision making across a range of treatments and patients groups at NICE.

First, it is extremely difficult to draw firm conclusions about the relative performance of the different generic measures since the degree of evidence that exists in relation to EQ-5D is vastly more than that which exists for the other measures. Only rarely have they been compared within the same studies.

Second, the results of the review are aggregated in Table 7 of the report at the study level. As highlighted in the reviews above, this can sometimes lead to a misleading picture where there are multiple groups being compared, as is often the case in "known groups" or responsiveness comparisons.

The MRC review also highlights the fact that many of the comparisons are hampered by the failure to control for potential confounders. This point is particularly important given the nature of the patient populations being considered in most of the studies. Patients tend to be elderly with a substantial proportion over 80 years, have a high incidence of comorbidities. In these predominantly non randomised studies failure to control for these characteristics is a fundamental limitation.

An example from one study included in the review[85], that forms part of the evidence base on EQ-5D and NV-AMD illustrates these issues. Soubrane et al (2007)[85] yields apparently inconsistent results when comparing groups of patients defined in terms of VA severity in the better seeing eye. However, the limitations are so substantial as to make the findings in relation to EQ-5D of questionable value. First, no comparison can be made between the control group and any of the AMD patients, without controlling for relevant confounders. The patient group had a mean age 14 years higher than the control group as well as statistically significant imbalances in a range of visual and non visual comorbidities. These imbalances may also be present within the NV-AMD groups but this is not reported. Given the frequency of comorbidities such as cancer (8.2%), diabetes (10%) and arthritis (11%) such imbalances are likely and potentially critical determinants of HRQoL. Indeed, it is interesting to note that the authors conducted subgroup analyses that compared the NV-AMD patients as a whole group with control subjects after controlling for age and comorbidities, separately, and found consistent and statistically significant differences in mean EQ-5D scores. Furthermore, the paper does not report the numbers of patients in each of these NV-AMD subgroups, making it impossible to judge whether the apparent inconsistencies are due to chance.

Two other general and related points arise from the reviews, in particular the MRC review of visual disorders. First, the EQ-5D tends to yield "consistent" results but these are frequently not statistically significant. This is the case for known groups comparisons, the various approaches to assessing convergent validity, and the assessment of responsiveness. This may illustrate that the series of studies included in the review are generally insufficiently powered to detect changes in EQ-5D, resulting in findings that are generally consistent, frequently statistically insignificant and sometimes inconsistent. Of course, an alternative interpretation is that the EQ-5D is insufficiently sensitive. Second, the HUI-3 that includes a specific visual dimension on 6 levels is more strongly correlated with measures of visual impairment than EQ-5D. This is hardly surprising. To draw any conclusions about the appropriateness of either instrument requires a recognition that they are assessing different concepts. The EQ-5D asks respondents to indicate the degree to which their visual disorder impacts on five domains. The tariff is then derived from the general population valuation of that impact. The HUI3 vision domain however is based on asking respondents to indicate the degree of visual impairment they experience as opposed to the impact of that impairment on their lives. The general population tariff values are therefore based on how they would perceive the visual impairment to impact on their lives. In relation to this issue, it is interesting to note that in Espallargues et al (2005)[29] the strength of the relationship between the degree of visual impairment and the HUI3 is stronger than the relationship between visual impairment and own valued health using both the TTO and the VAS. As the MRC report states, there is no reason to infer from these findings that the HUI-3 is the gold standard in visual disorders.

6. DISCUSSION

The purpose of this report is to investigate the role of EQ-5D as the preferred instrument for valuing health states for use in economic evaluation across the range of NICE's activities. We have reviewed claims made to the Institute and conducted de novo reviews in order to investigate the strength of these claims. In this section we provide an overview of the findings and consider the implications for the Institute and future research and the potential impact of planned future developments around the EQ-5D instrument.

Summary of claims made to NICE

In this report we have reviewed all submissions made to the Kennedy Review of the Value of Innovation[2] in relation to the claimed inadequacy of the EQ-5D instrument. We supplemented this with a review of similar claims made within the NICE Technology appraisals process.

Very few claims were made in relation to the method of valuing health benefits within a cost utility framework and even fewer related to the EQ-5D itself. Most claims we identified related either to the appropriateness of the QALY as a measure of outcome per se or the decision rule of QALY maximisation. There were also few examples of technology appraisals where the EQ-5D's appropriateness was considered controversial. In all cases, supporting empirical evidence was either sparse or non existent.

We found that claims covered several very broad disease area such as mental health and cancer. More specific issues were raised about situations where the EQ-5D does not include a dimension that directly reflects a symptom of claimed importance, such as fatigue or sensory impairment. Some of these issues are of relevance to adverse events from treatments rather than specific diseases. There were also claims made about the perceived inadequacy of EQ-5D where the disease course waxes and wanes with flares of symptom severity that are unpredictable in nature.

These unsupported claims were used to inform our choice of case studies for detailed review.

Summary of empirical evidence from the case studies

There were several studies where EQ-5D was found to be less responsive or sensitive than disease specific outcome measures. This was the case both with disease specific preference based measures (e.g AQL-5D) and disease specific non-preference based measures (e.g AQLQ). In studies which included other generic preference based measures, these were not found to perform consistently better than EQ-5D, although there were not many studies which examined these comparisons included in the reviews. The exception to this was that in the vision case study, it was found that the HUI-3 which includes a specific vision dimension, was more sensitive to changes in vision than the EQ-5D. Conversely there is also evidence that EQ-5D may discriminate between patients better than some standard disease specific outcome measures precisely because it includes domains of relevance to patients that are missing from the disease specific measure (e.g HAQ in RA).

The review identified few studies that considered the individual domains of EQ-5D rather than the summary score. As described in section 2, considering the individual domains may be more informative.

Limitations to assessment of appropriateness of EQ-5D

In section 2.3 we described the types of comparisons that are typically used to support claims of the adequacy/inadequacy of any instrument. These are founded on tests from psychometrics and relate substantially to the concepts of validity, reliability and responsiveness.

When the instrument in question intends to measure health utility, as EQ-5D does, these comparisons are not tests. They can highlight differences between EQ-5D and other instruments such as other generic instruments, disease specific outcomes or clinical measures but since there is no gold standard it cannot be established conclusively which measure is "right". Intuition and judgement are required to draw any stronger conclusions.

Even when comparing EQ-5D with other generic, preference based measures, one must be aware of the conceptual differences between measures. This is well highlighted by the case of the HUI3, which appears to be a measure that is more closely correlated to changes in vision, for example, than EQ-5D. The instruments do not measure the same thing. The HUI3 asks patients to indicate their health state by, inter alia, describing their visual impairment on a seven point scale. These symptoms are then valued by the general public. The EQ-5D on the other hand asks patients to indicate the impact their visual disorder has on five domains that indicate the impact on their life. This description is then valued by the general public. That results differ between the two is hardly surprising. One might expect for example that the EQ-5D reflects a degree of adaptation on the part of patients that is absent from the HUI3. Which approach is "correct" requires a judgement about the conceptual basis of health which is part of a broader issue that relates to the roles of patients vis-a-vis the general public in health state valuations. We would venture though that a judgement cannot be reached by considering specific diseases or treatments in isolation.

There is an additional factor that is relevant in the context of decision making at NICE and that is the requirement for consistency. It is debateable whether consistency requires the same instrument to be used in all assessments (a point which is discussed in section 2.3) but whichever view is taken, this is an additional caveat that is often not relevant to the authors of studies who focus on "appropriateness" within a single disease area rather than across the entire health care system.

Limitations to the studies included in the reviews can further dilute the conclusions that may be drawn.

Where groups are defined in terms of some clinical measure, the distinctions between groups may reasonably not translate to differences in health utilities. An example of typical study limitations comes from the incontinence review. Haywood et al (2008)[64]found that EQ-5D was not able to fully discriminate between 5 groups. The groups were defined in terms of the number of episodes as " not at all", "a few days", "half the week", "most days" and "every day". The differences between the groups are therefore relatively small, not necessarily mutually exclusive, and it is

questionable whether there would be significant differences in the preferences of patients in some of the groups.

Furthermore, the reporting of the extent to which an instrument is consistent with groups defined in another way needs to consider how many groups are being considered. Often there are multiple groups being compared and the instrument may provide consistent results across many of them. Results that are summarised at the study level do not reflect the number of comparisons made within each study and may therefore provide a misleading view of the evidence. P-values typically relate to the null hypothesis that the mean value is equal in all the subgroups under consideration. This itself may be ambiguous because it does not consider how many of the individual pairs of comparisons are statistically significant and does not discriminate between situations where the observations are all consistent i.e. statistical significance provides support for the validity of the instrument, versus those where one or more observations appear to be inconsistent i.e. statistical significance may or may not provide support for the validity of the instrument.

Issues for consideration

The implications of some of the findings relate to sample sizes for estimating the effects of treatments rather than the appropriateness of EQ-5D per se, although the two issues are not mutually exclusive. The consideration of inadequate sample sizes within clinical trials powered on some other outcome measure to EQ-5D needs to be considered at an early stage. The NICE Scientific Advice programme has a role to play here. There are several practical steps that may be considered where single trials are likely to be underpowered for EQ-5D. For example, observations may be maximised by the inclusion of EQ-5D in all trials. Another option is to harness the additional statistical power from external datasets, such as those from observational studies, by estimating the relationship between EQ-5D and outcome measures used in the trial.

NICE must consider what is required in order to achieve consistency. The current approach indicates that the use of the same descriptive system and valuation is required, hence the preference for the EQ-5D stipulated in the methods guide. Brazier and Tsuchiya (2010)[9] outline the view that the same descriptive system is not required to achieve consistency, only the same valuation method. However, this is founded on the premise that equal coverage is achieved across diseases using different descriptive systems which is unlikely to be realistic. This is not a defence of EQ-5D itself however, but may have quite different implications for the development of future instruments or refinement of existing ones.

It is clear that the construction of a case for departure from the NICE reference case requires a systematic and detailed review of the entirety of the literature. This was a substantial and time consuming task in each of the reviews we undertook and may not be feasible in many of NICE's evaluations. In particular, this would be a substantial addition to the STA process. Future developments of the NICE Methods guide may need to reflect on the instructions to provide empirical evidence of EQ-5D being inappropriate, subject to any other changes, particularly as our review of previous technology appraisals yielded little evidence that such empirical support has been provided.

Future developments of EQ-5D

A moratorium on modification of the EQ-5D was put in place in 1993 and has largely held until the present time[86]. This has allowed the production of a large catalogue of datasets from around the world that are comparable.

There are two developments in particular that have been approved by the EuroQol group that may have relevance to the issues raised in this report.

First, the EuroQol group have recently introduced a five level version of the EQ-5D, the EQ-5D-5L. The dimensions of this instrument remain as the EQ-5D-3L, but expands the range of options. In English, the new labels are "no problems", "slight problems", "moderate problems", "severe problems" and "extreme problems/unable to perform". However, there is no current value set for the EQ-5D-5L. Studies are underway to derive these value sets but will take time to report and disseminate findings. In the interim there are attempts to estimate the values for EQ-5D-5L using data from patients that have simultaneously completed both the three and five level variants.

The expectation amongst its developers is that the five level version of EQ-5D will enhance responsiveness and sensitivity. This will have the impact of reducing the required sample size to detect small changes in health compared to the three level version. How this compares to alternative approaches for addressing inadequate sample sizes, and whether it will eradicate the need to employ these approaches, remains to be seen.

Second, EuroQol have approved the development of "bolt-ons/dimension extensions". These instruments will permit the addition of extra dimensions to the standard EQ-5D instrument in order to directly capture other issues of importance to patients. How precisely these bolt-ons are approached remains to be seen. There are both philosophical and challenging statistical considerations that need to be resolved in order for this route to provide an approach that is consistent with evaluations that use the standard five domain instrument. What is the nature of the additional domains to be "bolted on"? Will they retain the philosophy of the existing domains of EQ-5D of being non symptom specific, focussing on general functionings? Is the philosophy of the conceptualisation of health related quality of life consistent? Can comparability be achieved given that domains are not independent? As highlighted in the results of the reviews conducted in this report and for the MRC, it is simply not the case that the EQ-5D entirely misses a set of concerns or symptoms that can be easily and simply rectified by the addition of the "missing" domain.

It is interesting to note here that fatigue is one of the symptoms that was claimed by stakeholders to the Kennedy review to not be captured by the standard EQ-5D instrument. Williams (1995)[5] in his description of the development of the valuation set describes that this was the domain that was the strongest candidate for inclusion based on the qualitative and survey work they undertook. However, an experimental 6 domain version of the EuroQol Questionnaire was tested in a pilot. The results showed that it has such a small impact on valuations that it was not pursued further.

Other alternative developments need to be guided after consideration of the issues raised in this report about the conceptual nature of health and what is required of health state valuation methods in order to achieve consistency in decision making. One direction for development is in the area of condition–specific preference based approaches, although many of the same issues that are relevant to EQ-5D bolt ons are also applicable here.

7. CONCLUSIONS

This study has identified the types of claims that have been made about the EQ-5D in informing the types of decisions made across the range of NICE activities. These reviews informed case study reviews which highlight the difficulties associated with drawing conclusions about the appropriateness of otherwise of any preference based measure of HRQoL. Empirical evidence tends to be based on psychometric tests which provide only circumstantial rather than direct evidence.

However, the case studies also demonstrate that it is important to include all relevant evidence and for each included study to be considered in great detail in order to understand the strength of the evidence for any particular claim. This is because the studies that generate relevant evidence are often not designed specifically with this aim in mind. Evidence may come from studies that include patients with the same broad condition but are not precisely matched with those in question. Patients may also cover a range of severities not relevant to the intervention under assessment. Furthermore, the design of studies may be such that apparent tests of appropriateness of EQ-5D are undermined. As with reviews of clinical evidence, the strengths and weaknesses of studies must be fully appraised.

8. REFERENCES

(1) National Institute for Health and Clinical Excellence (NICE). Guide to the methods of technology appraisal. 2008.

(2) Kennedy I. Appraising the value of innovation and other benefits: a short study for NICE. 2009.

(3) Tosh J, Brazier J, Evans P, Longworth L. A review of generic preference-based measures of health-related quality of life in visual disorders (DRAFT): A report for the NICEQoL project. 2010.

(4) Dolan P, University of York.Centre for Health Economics, University of York.Health Economics Consortium, University of York.NHS Centre for Reviews and Dissemination. A social tariff for EuroQol: results from a UK general population survey. Centre for Health Economics York; 1995.

(5) Williams A. The Measurement and Valuation of Health: A Chronicle, University of York, Centre for Health Economics Discussion Paper 136. *Centre for Health Economics Discussion Paper 136* 1995.

(6) Gudex C. The descriptive system of the EuroQol Instrument. *EQ-5D concepts and methods: A developmental history* 2005;19-27.

(7) Williams A. The EuroQol instrument. Springer; 1993. 1-17.

(8) Dowie J. Decision validity should determine whether a generic or condition-specific HRQOL measure is used in health care decisions. *Health Economics* 2002; 11(1):1-8.

(9) Brazier J, Tsuchiya A. Preference based condition specific measures of health: what happens to cross programme comparability? *Health Economics* 2010; 19(2):125-129.

(10) Brazier J. Briefing paper for methods review workshop on key issues in utility measurement (NICE). 2007.

(11) Coast J, Flynn TN, Natarajan L, Sproston K, Lewis J, Louviere JJ et al. Valuing the ICECAP capability index for older people. *Social Science & Medicine* 2008; 67(5):874-882.

(12) Dolan P, Lee H, King D, Metcalfe R. Valuing health directly. *British Medical Journal* 2009; 339(jul20 3):b2577.

(13) Walters SJ. Quality of life outcomes in clinical trials and health-care evaluation: a practical guide to analysis and interpretation. Wiley; 2009.

(14) Brazier J, Deverill M. A checklist for judging preference-based measures of health related quality of life: learning from psychometrics. *Health Economics* 1999; 8(1):41-51.

(15) Claxton K, Walker S, Palmer S, Sculpher M. Appropriate perspectives for health care decisions. *Working Papers* 2010.

(16) Kennedy Review on Value: BIA submission. 2009.

(17) Kennedy Review: Johnson and Johnson. 2009.

(18) NICE short study on valuing innovation: A submission of comments from GlaxoSmithKline. 2009.

(19) Chisholm D, Healey A, Knapp M. QALYs and mental health care. *Social Psychiatry and Psychiatric Epidemiology* 1997; 32(2):68-75.

(20) Amgen response to the Kennedy review of how NICE values innovation. 2009.

(21) Kennedy study on valuing innovation (April 2009): Myeloma UK written evidence submission. 2009.

(22) Van Agthoven M, Segeren CM, Buijt I, Uyl-de Groot CA, Van Der Holt B, Lokhorst HM et al. A cost-utility analysis comparing intensive chemotherapy alone to intensive chemotherapy followed by myeloablative chemotherapy with autologous stem-cell rescue in newly diagnosed patients with stage II/III multiple myeloma:: a prospective randomised phase III study. *European Journal of Cancer* 2004; 40(8):1159-1169.

(23) Gulbrandsen N, Wisl°ff F, Nord E, Lenhoff S, Hjorth M, Westin J. Costûutility analysis of high-dose melphalan with autologous blood stem cell support vs. melphalan plus prednisone in patients younger than 60 years with multiple myeloma. *European journal of haematology* 2001; 66(5):328-336.

(24) Kind P, Dolan P, Gudex C, Williams A. Variations in population health status: results from a United Kingdom national questionnaire survey. *Bmj* 1998; 316(7133):736.

(25) Tsuchiya A, Brazier J, McColl E, Parkin D. Deriving preference-based single indices from non-preference based condition-specific instruments: Converting AQLQ into EQ5D indices. 2002.

(26) Lloyd A, Price D, Brown R. The impact of asthma exacerbations on health-related quality of life in moderate to severe asthma patients in the UK. *Primary care respiratory journal: journal of the General Practice Airways Group* 2007; 16(1):22.

(27) Brown GC, Sharma S, Brown MM, Kistler J. Utility values and age-related macular degeneration. *Archives of Ophthalmology* 2000; 118(1):47.

(28) Czoski Murray C, Carlton J, Brazier J, Young T, Papo NL, Kang HK. Valuing Condition Specific Health States Using Simulation Contact Lenses. *Value in Health* 2009; 12(5):793-799.

(29) Espallargues M, Czoski-Murray CJ, Bansback NJ, Carlton J, Lewis GM, Hughes LA et al. The impact of age-related macular degeneration on health status utility values. *Investigative ophthalmology & visual science* 2005; 46(11):4016.

(30) Williams RA, Brody BL, Thomas RG, Kaplan RM, Brown SI. The psychosocial impact of macular degeneration. *Archives of Ophthalmology* 1998; 116(4):514.

(31) Barton GR, Bankart J, Davis AC, Summerfield QA. Comparing utility scores before and after hearing-aid provision: results according to the EQ-5D, HUI3 and SF-6D. *Applied health economics and health policy* 2004; 3(2):103-105.

(32) Barton GR, Bankart J, Davis AC. A comparison of the quality of life of hearing-impaired people as estimated by three different utility measures. *International Journal of Audiology* 2005; 44(3):157-163.

(33) Davis A, Smith P, Ferguson M, Stephens D, Gianopoulos I. Acceptability, benefit and costs of early screening for hearing disability: a study of potential screening tests and models. *Health Technol Assess* 2007; 11(42):1-294.

(34) National Institute for Health and Clinical Excellence (NICE). Inhaled corticosteroids for the treatment of adults and children aged 12 years and over, Technology Appraisal Guidance 138. 2008.

(35) Willems DC, Joore MA, Hendriks JJ, Wouters EF, Severens JL, Willems DC et al. Cost-effectiveness of a nurse-led telemonitoring intervention based on peak expiratory flow measurements in asthmatics: results of a randomised controlled trial. *Cost Effectiveness & Resource Allocation* 2007; 5:10.

(36) Saarni SIH. The impact of 29 chronic conditions on health-related quality of life: A general population survey in Finland using 15D and EQ-5D. *Quality of Life Research* 2006; 15(8):Nov.

(37) Barton GRS. A comparison of the performance of the EQ-5D and SF-6D for individuals aged [greater-than or equal to] 45 years. *Health Economics* 2008; 17(7):July.

(38) Polley L, Yaman N, Heaney L, Cardwell C, Murtagh E, Ramsey J et al. Impact of cough across different chronic respiratory diseases: comparison of two cough-specific health-related quality of life questionnaires. *Chest* 2008; 134(2):295-302.

(39) Willems DC, Joore MA, Nieman FH, Severens JL, Wouters EF, Hendriks JJ et al. Using EQ-5D in children with asthma, rheumatic disorders, diabetes, and speech/language and/or hearing disorders. *International Journal of Technology Assessment in Health Care* 2009; 25(3):391-399.

(40) Brusselle G, Michils A, Louis R, Dupont L, Van de Maele B, Delobbe A et al. "Real-life" effectiveness of omalizumab in patients with severe persistent allergic asthma: The PERSIST study. *Respiratory Medicine* 2009; 103(11):1633-1642.

(41) Ferreira LN, Brito U, Ferreira PL, Ferreira LN, Brito U, Ferreira PL. Quality of life in asthma patients. *Revista Portuguesa de Pneumologia* 2010; 16(1):23-55.

(42) McTaggart-Cowan HM, Marra CA, Yang Y, Brazier JE, Kopec JA, FitzGerald JM et al. The validity of generic and condition-specific preference-based instruments: the ability to discriminate asthma control status. *Quality of Life Research* 2008; 17(3):453-462.

(43) Chen H, Gould MK, Blanc PD, Miller DP, Kamath TV, Lee JH et al. Asthma control, severity, and quality of life: quantifying the effect of uncontrolled disease. *Journal of Allergy & Clinical Immunology* 2007; 120(2):396-402.

(44) Aburuz S, Gamble J, Heaney LG, Aburuz S, Gamble J, Heaney LG. Assessment of impairment in health-related quality of life in patients with

difficult asthma: psychometric performance of the Asthma Quality of Life Questionnaire. *Respirology* 2007; 12(2):227-233.

(45) Szende A, Svensson K, Stahl E, Meszaros A, Berta GY, Szende A et al. Psychometric and utility-based measures of health status of asthmatic patients with different disease control level. *Pharmacoeconomics* 2004; 22(8):537-547.

(46) Oga T, Nishimura K, Tsukino M, Sato S, Hajiro T, Mishima M et al. A comparison of the responsiveness of different generic health status measures in patients with asthma. *Quality of Life Research* 2003; 12(5):555-563.

(47) McColl E, Eccles MP, Rousseau NS, Steen IN, Parkin DW, Grimshaw JM et al. From the generic to the condition-specific?: Instrument order effects in Quality of Life Assessment. *MED CARE* 2003; 41(7):777-790.

(48) Garratt AM, Hutchinson A, Russell I, Garratt AM, Hutchinson A, Russell I. Patient-assessed measures of health outcome in asthma: a comparison of four approaches. *Respiratory Medicine* 2000; 94(6):597-606.

(49) Young T, Yang Y, Brazier J, Tsuchiya A. The use of Rasch analysis as a tool in the construction of a preference based measure: the case of AQLQ. 2007.

(50) Yang Y, Tsuchiya A, Brazier JE, Young TA. Estimating a preference-based single index from the Asthma Quality of Life Questionnaire (AQLQ). 2007.

(51) Revicki DA, Leidy NK, Brennan-Diemer F, Sorensen S, Togias A. Integrating patient preferences into health outcomes assessment. *Chest* 1998; 114(4):998.

(52) Willems DCM, Joore MA, Hendriks JJE, Nieman FHM, Severens JL, Wouters EFM. The effectiveness of nurse-led telemonitoring of asthma: results of a randomized controlled trial. *Journal of Evaluation in Clinical Practice* 2008; 14(4):600-609.

(53) Abrams P, Cardozo L, Fall M, Griffiths D, Rosier P, Ulmsten U et al. The standardisation of terminology of lower urinary tract function: report from the Standardisation Sub committee of the International Continence Society. *Neurourology and urodynamics* 2002; 21(2):167-178.

(54) Hawthorne G. Assessing utility where short measures are required: Development of the short assessment of quality of life-8 (AQoL-8) Instrument. *Value in Health* 2009; 12(6):September.

(55) Currie CJ, McEwan P, Poole CD, Odeyemi IA, Datta SN, Morgan CL et al. The impact of the overactive bladder on health-related utility and quality of life. *BJU International* 2006; 97(6):1267-1272.

(56) Noble SM, Coast J, Brookes S, Neal DE, Abrams P, Peters TJ et al. Transurethral prostate resection, noncontact laser therapy or conservative management in men with symptoms of benign prostatic enlargement: an economic evaluation (Structured abstract). *Journal of Urology* 2002; 168:2476-2482.

(57) Donovan JL, Kay HE, Peters TJ, Abrams P, Coast J, Matos-Ferreira A et al. Using the ICSOoL to measure the impact of lower urinary tract symptoms on quality of life: evidence from the ICS-'BPH' Study. International Continence Society--Benign Prostatic Hyperplasia. *Br J Urol* 1997; 80(5):712-721.

(58) Monz B, Chartier-Kastler E, Hampel C, Samsioe G, Hunskaar S, Espuna-Pons M et al. Patient characteristics associated with quality of life in European women seeking treatment for urinary incontinence: results from PURE. *European Urology* 2007; 51(4):1073-1081.

(59) Monz B, Pons ME, Hampel C, Hunskaar S, Quail D, Samsioe G et al. Patient-reported impact of urinary incontinence--results from treatment seeking women in 14 European countries. *Maturitas* 2005; 52 Suppl 2:S24-S34.

(60) Johannesson M, O'CONOR RM, Kobelt-Nguyen G, Mattiasson A. Willingness to pay for reduced incontinence symptoms. *British journal of urology* 1997; 80(4):557-562.

(61) Ternent L, Vale L, Buckley B, Glazener C, Ternent L, Vale L et al. Measuring outcomes of importance to women with stress urinary incontinence. *BJOG* 2009; 116(5):719-725.

(62) Ismail SI, Forward G, Bastin L, Wareham K, Emery SJ, Lucas M et al. Extracorporeal magnetic energy stimulation of pelvic floor muscles for urodynamic stress incontinence of urine in women. *Journal of Obstetrics & Gynaecology* 2009; 29(1):35-39.

(63) Rinne K, Laurikainen E, Kivela A, Aukee P, Takala T, Valpas A et al. A randomized trial comparing TVT with TVT-O: 12-month results. *International Urogynecology Journal* 2008; 19(8):1049-1054.

(64) Haywood KL, Garratt AM, Lall R, Smith JF, Lamb SE, Haywood KL et al. EuroQol EQ-5D and condition-specific measures of health outcome in women with urinary incontinence: reliability, validity and responsiveness. *Quality of Life Research* 2008; 17(3):475-483.

(65) Kobelt G, Fianu-Jonasson A, Kobelt G, Fianu-Jonasson A. Treatment of stress urinary incontinence with non-animal stabilised hyaluronic acid/dextranomer (NASHA/Dx) gel : An analysis of utility and cost. *Clinical Drug Investigation* 2006; 26(10):583-591.

(66) Dumville JC, Manca A, Kitchener HC, Smith AR, Nelson L, Torgerson DJ et al. Cost-effectiveness analysis of open colposuspension versus laparoscopic colposuspension in the treatment of urodynamic stress incontinence. *BJOG* 2006; 113(9):1014-1022.

(67) Manca A, Sculpher MJ, Ward K, Hilton P, Manca A, Sculpher MJ et al. A cost-utility analysis of tension-free vaginal tape versus colposuspension for primary urodynamic stress incontinence. *BJOG* 2003; 110(3):255-262.

(68) Kobelt G, Kobelt G. Economic considerations and outcome measurement in urge incontinence. *Urology* 1997; 50(6A Suppl):100-107.

(69) Tincello DS. Patient characteristics impacting health state index scores, measured by the EQ-5D of females with stress urinary incontinence symptoms. *Value in Health* 2010; 13(1):January-February.

(70) Mihaylova B, Pitman R, Tincello D, van d, V, Tunn R, Timlin L et al. Cost-Effectiveness of Duloxetine: The Stress Urinary Incontinence Treatment (SUIT) Study. *Value Health* 2010.

(71) Harrison MJ, Davies LM, Bansback NJ, Ingram M, Anis AH, Symmons DPM. The validity and responsiveness of generic utility measures in rheumatoid arthritis: a review. *The Journal of Rheumatology* 2008; 35(4):592.

(72) Conner-Spady B, Suarez-Almazor ME. Variation in the estimation of quality-adjusted life-years by different preference-based instruments. *MED CARE* 2003; 41(7):791-801.

(73) Marra CA, Esdaile JM, Guh D, Kopec JA, Brazier JE, Koehler BE et al. A comparison of four indirect methods of assessing utility values in rheumatoid arthritis. *MED CARE* 2004; 42(11):1125-1131.

(74) Luo N, Chew LH, Fong KY, Koh DR, Ng SC, Yoon KH et al. A comparison of the EuroQol-5D and the Health Utilities Index mark 3 in patients with rheumatic disease. *The Journal of Rheumatology* 2003; 30(10):2268.

(75) 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. *Social Science & Medicine* 2005; 60(7):1571-1582.

(76) Linde L, S°rensen J, Ostergaard M, H°rslev-Petersen K, Hetland ML. Health-related quality of life: validity, reliability, and responsiveness of SF-36, 15D, EQ-5D [corrected] RAQoL, and HAQ in patients with rheumatoid arthritis. *The Journal of Rheumatology* 2008; 35(8):1528.

(77) Hawthorne G, Buchbinder R, Defina J, Monash University, University of Melbourne. Functional Status and Health-related Quality of Life Assessment in Patients with Rheumatoid Arthritis. Centre for Health Program Evaluation; 2000.

(78) Hurst NP, Kind P, Ruta D, Hunter M, Stubbings A, Hurst NP et al. Measuring health-related quality of life in rheumatoid arthritis: validity, responsiveness and reliability of EuroQol (EQ-5D). *British Journal of Rheumatology* 1997; 36(5):551-559.

(79) Kobelt G, Eberhardt K, Jonsson L, Jonsson B, Kobelt G, Eberhardt K et al. Economic consequences of the progression of rheumatoid arthritis in Sweden. *Arthritis & Rheumatism* 1999; 42(2):347-356.

(80) Marra CA, Lynd LD, Esdaile JM, Kopec J, Anis AH. The impact of low family income on self-reported health outcomes in patients with rheumatoid arthritis within a publicly funded health-care environment. *Rheumatology* 2004; 43(11):1390.

(81) Harrison MJ, Tricker KJ, Davies L, Hassell A, Dawes P, Scott DL et al. The relationship between social deprivation, disease outcome measures, and response to treatment in patients with stable, long-standing rheumatoid arthritis. *Journal of Rheumatology* 2005; 32(12):2330-2336.

(82) Marra CA, Rashidi AA, Guh D, Kopec JA, Abrahamowicz M, Esdaile JM et al. Are indirect utility measures reliable and responsive in rheumatoid arthritis patients? *Quality of Life Research* 2005; 14(5):1333-1344.

(83) Russell AS, Conner-Spady B, Mintz A, Maksymowych WP, Russell AS, Conner-Spady B et al. The responsiveness of generic health status measures as assessed in patients with rheumatoid arthritis receiving infliximab. *Journal of Rheumatology* 2003; 30(5):941-947.

(84) Alava H, Wailoo AJ, Ara R. Tails from the Peak District: adjusted censored mixture models of EQ-5D health state utility values. 2010.

(85) Soubrane G, Cruess A, Lotery A, Pauleikhoff D, Mones J, Xu X et al. Burden and health care resource utilization in neovascular age-related macular degeneration: findings of a multicountry study. *Archives of Ophthalmology* 2007; 125(9):1249.

(86) Kind P, Brooks R, Rabin R. EQ-5D concepts and methods: a developmental history. Kluwer Academic Pub; 2005.

Table 13: Summary of claims as part of submissions to Kennedy Study

Consultee	Insufficient Perspective	Use of QALYs /cost-utility analysis	EQ-5D
Deltex	Returning to work earlier not valued by NICE i.e productivity costs excluded	Doesn't capture impact of faster recovery, fewer complications and earlier discharge on mental "well-being"	
ABHI	Perspective should be broadened to include: • Societal benefits. - Return to work. - Savings in other govt departments. • Carer benefits.	Factors to be considered in addition to QALYs: • Improved efficiency in delivery of care. • Support for wider NHS priorities such as "18 weeks", "duty to innovate", and "closer to home". • Patient experience. If cost-utility analysis is done too early, then this can ignore the benefits of a potentially increasing evidence base and improved effectiveness from iterative improvement to medical devices during clinical use.	
ABPI	Perspective should be broadened to include societal benefits: • Positive economic externalities such as workforce participation, reduced sick leave, and productivity. • Reduction in benefits payments/public spending. • Positive externalities from supply and production of health care (e.g employment). • Social capital (e.g better health may result in more engaged social role outside of employment). • Positive externalities for other patients. • Positive externalities in health technology (e.g R&D incentives). Perspective should be broadened to include carer benefits • Positive externalities for carer's mental and physical health.	Societal preference to treat based on unmet need, disease severity (Mason et al 2008, Dolan et al 2008, NICE Citizens Council) Disease rarity means that the expense of development is spread over a few patients giving a higher price per patient Health benefits not captured: • Creating an additional life. • Patient safety. • Option value (service availability when required). • Reductions in health inequalities (caring externalities). Non-health benefits not captured: • Improve non-health QoL (utility/happiness/life satisfaction/subjective wellbeing). • Process-of-care utility. • Desired reduction in income and social inequalities. Patient centred attributes to be considered alongside QALY gains: • Dosage. • Treatment site. • Adherence. • Independence.	
AdvaMed	Perspective should be broadened to include: • Productivity costs/employment.	Factors in to be considered in addition to QALYs: • Personal benefits – such as better management of incontinence. • Patient care attributes – management of treatment.	
Amgen	Perspective should be broadened to include: • Societal benefits eg. productivity costs/employment. • Benefits to carers' and families mental and physical health. • Community and collective wellbeing and productivity. • Productivity gains. • Reduced social service / unemployment benefit.	Factors to be considered in addition to QALYs: • Aspects of HRQoL (energy, vitality, consistency of wellness etc). • Aspects of non-HRQoL (wellbeing, satisfaction, happiness, hope). • Increased utility due to improvements in care experience (convenience, empowerment, dignity, earnings). • Health inequalities. • Spillover effects of novel techs to further research. • Spillover improvements in service redesign.	

Consultee	Insufficient Perspective	Use of QALYs /cost-utility analysis	EQ-5D
		Differential value associated with particular characteristics of the recipients of QALY gains (e.g End of life reform capturing value of life-prolonging interventions at the end of life)	
Arthritis Care	Perspective should be broadened to include: • Positive economic impact to individual and society.	Factors to be considered in addition to QALYs: • Increased convenience for service-user. • Increase dignity. • Reduction in social inequality. • Non-HRQoL (happiness and wellbeing).	
AstraZeneca	Perspective should be broadened to include: • Productivity gains for both patients and their carers.	Factors to be considered in addition to QALYs: • Patient convenience in terms of delivery mechanism.	EQ-5D valuations are out of date. Suggesting re-running EQ-5D valuations. EQ-5D (and SF-36) not sensitive to impact of convenience on quality of life.
Baker Donaldson		Valuing societal preference using "Social Value of a QALY" exercise. QALY weights elicited for age and severity of illness.	
Beating Bowel Cancer		QALY approach is biased against providing treatments which improve quality of life in those with short life-expectancy.	
BIA	Perspective should be broadened to include: • Productivity benefits. • Reduced burden on carers. • Reduced burden on social services. • Out of pocket costs for patients. • Greater societal involvement.	Factors to be considered in addition to QALYs: • Disease severity. • Rarity. • Unmet need. • Orphan drugs. Factors not captured by QALYs: • Clinical outcomes important to patient in addition to QALYs e.g maintaining independence. • Patient convenience, compliance. • Patient preferences for route of delivery. • Health equity. • Long-term outcomes such as antibiotic resistance, transmission of infectious diseases. • Improvements in non-health related quality of life e.g subjective well-being. Problems with QALY approach: • QALY doesn't distinguish between different sorts of life extension e.g. end of life. Possible solutions: • Weights could be used to account for different social values for QALY depending on characteristics of recipient.	Limitation of EQ-5D for measuring QoL. Difficult to use in some groups e.g learning disabilities, mental health. Dimensions not captured e.g cognitive function. Ceiling effects (No empirical studies cited).
BIVDA			
BMS	Perspective should be broadened to include: • Societal costs and benefits • Returning to work	Factors to be considered in addition to QALY: • Severity of disease • Bias against rarer conditions	

Consultee	Insufficient Perspective	Use of QALYs /cost-utility analysis	EQ-5D
	• Reduced burden on carers.	Problems with QALY: • QALY doesn't capture value at end of life.	
CRUK		Problems with QALY: • More weight needed at end of life	Question the appropriateness of EQ-5D, especially relating to cancer. Encourage EuroQoL and NICE to update EQ-5D.
EMIG	Perspective should be broadened to include: • Societal benefits e.g Productivity gains	Factors to be considered in addition to QALYs: • Unmet need • Patients' convenience • Attributes specific to certain therapeutic areas not valued	
Genzyme		Factors to be considered in addition to QALYs: • Unmet need	
GIG		Factors to be considered in addition to QALYs: • Rare disease • Unmet need • Patient's circumstances • Disease severity • Intervention impact	
GSK	Perspective should be broadened to include: • Social care. • Welfare benefits.	Questions weight given to cost per QALY in decisions making Cost-utility approach is biased against recommending life-extending drugs in patients with high annual costs of additional life-years. Alternative thresholds for different diseases. States that QALYs are a crude, arbitrary, population based measure that don't reflect the needs of individual patients. Non –health benefits not captured by QALYs: • educational attainment associated with lower cognitive side-effects of epilepsy drugs Factors to be considered in addition to QALYs: • Severity and unmet need • Benefits important to patients e.g safety profile, route and setting of administration • Non-HRQoL improvements • Public preferences • Government priorities	Insensitive, does not account for patient treatment preferences and safety profile. Impact on cognitive function of newer epilepsy treatment not captured. Cancer fatigue not captured. Doesn't capture non-health related quality of life measure such as happiness, well-being.
Hep C Trust	Perspective should be broadened to include: • Societal benefits of reduced transitions.	Cost utility analysis should capture health benefits beyond the treated individual for infectious diseases.	
Hooman Fenwick			
IDIS			
Isabel Health Care			
Johnson and Johnson	Perspective should be broadened to include: • Productivity/societal benefits	Cost-utility may not capture secondary benefits of drug for uses other than first licensed indication.	Insensitivity of the EQ-5D. Does not capture

Consultee	Insufficient Perspective	Use of QALYs /cost-utility analysis	EQ-5D
	• Non-health care costs • Carers and family • Patient personal costs	States that flaw of QALY approach are well documented (references 2-8). Gives examples of conditions where QALY not appropriate (contraception, acute pain, deafness, antibiotics, mental health (ref11), palliative care (ref12), acute conditions (ref13), and elderly (ref14). Equity concerns regarding bias of QALY towards curative interventions and against interventions in those with permanent disability (ref10). Patient experience, convenience, preference e.g route of administration in NSCLC and shorter recovery times and less scaring for laparoscopic surgery. System benefits e.g waiting times	small fluctuations in HRQoL. Limited depth due to number of domains and levels. Doesn't capture subjective well-being .
Karl Claxton			
Lilly UK	Perspective should be broadened to include: • Social impact (costs and benefits). • Carers. • Economic contribution.	State that QALY limitations are well documented and cites Nord ViH 2009. Focuses on best option "on average" and doesn't value providing range of options to meet varying patient needs. Doesn't consider affordability.	
Medical Technology Group	Perspective should be broadened to include non NHS costs including: • Employment. • State benefits. • Access to education due to mode of administration.	Maintaining fertility may be a benefit that is broader than health.	
Mind	Perspective should be broadened to include costs and benefits to wider society not just NHS.	Cost-utility approach shouldn't be allowed to reduce choice of options available where there is variation in individual need e.g side-effects vary by patient. Difficulties in generating evidence base for non-Pharma interventions and risk to future research and service provision of negative recommendations.	Utility measurement should be based on service user valuations of efficacy and acceptability and long-term side-effects.
MSD		Doesn't allow for the fact that clinicians need access to a range of agents within a Pharma class as individuals will respond differently	
Myeloma UK			
Novartis	Perspective should be broadened to include: • Costs of private social care, loss of income. • Carer costs and loss of earnings. • Publicly funded social care. • Tax revenue.	Current approach doesn't allow for revision to cost-utility as evidence is gathered on longer term outcomes Suggests considering the following patient factors in addition to quality and length of life: quality of death, side-effect profile, length of hospital stay, number of outpatient appointments, frequency and type of monitoring, mode of admission, mobility, independence, psychological impact and well-being, interactions with clinicians. Suggests including additional hospital factors e.g clinical outcomes and side-effects, resource use, ease of admin and storage, place of admin, patient support and counselling.	
PHE	Perspective should be broadened to include; • Financial and non-pecuniary costs to care givers and families. • Labour productivity.	Public preferences outside of QALY e.g prevention of drug resistance, reassurance provided by diagnosis QALYs discriminate against older patients with chronic disability when evaluating life-saving interventions QALYs may not reflect public preference which may weight factors such as age and disease severity and unmet need. Cost-utility analysis may not capture benefits of enhanced compliance (5 studies cited on page 5) and decreased patient inconvenience (1 citation on page 5).	Valuation of improvement depends on current HRQoL (disease severity) and life-expectancy. Proposes using EQ-5D or SF-6D to determine disease severity but

Consultee	Insufficient Perspective	Use of QALYs /cost-utility analysis	EQ-5D
		Decision needs to include care giver utility.	says these may be insensitive in some cases and disease specific instruments could be used alongside generic measures. States that EQ-5D and SF-6D exclude elements of well-being.
RCN	Perspective should be broadened to include indirect costs to the patient and society A broader definition of "health" should be considered. • Personal and social care, • Accommodation, • Finance • Education, • Employment, • Leisure, • Transport and access Carer costs (financial and emotional) and loss of patient independence should be considered.	NICE should provide guidance on what existing technology the new technology replaces (e.g disinvestment). Cost-utility approach doesn't allow decisions to be individualised Cost-utility approach doesn't capture benefits of knowledge gained from recommending use with evidence gathering Current cost effectiveness data positively discriminates against patients with long term conditions. A recent Health Service Journal Report (5th February, 2009) highlighted the fact that NICE values some lives more than others. NICE places considerable value on 'treatments which offer the possibility of extending life when we are close to death'. Preference for route of delivery which may link to employment or emotional support.	Quality of life cannot be easily reduced to measurement on a quantitative scale. Small gain over short period may be perceived by patient as having considerable value.
Reform	Wider costs and benefits to society e.g • Employment • Welfare • Social care	QALYs apply greater "weight" to survival than quality of life therefore discriminating against long-term conditions. Current evaluation of quality of life is too narrow. Lacks transparency and understanding by patients and clinicians.	
Roche		Cost-utility is biased against life-extending treatment for patients with high cost of future years e.g statins in patients with ESRD or adding biologics to treatment of patients receiving chemotherapy Patient preference regarding administration method (e.g IV vs oral, daily vs weekly) not captured by QALYS.	
Schering plough		Difficulty capturing benefits of interventions which improve efficiency of service rather than patient outcomes (e.g anaesthetic agents) or mitigate future risk such as antibiotic resistance.	
SHA	Employment related benefits should be nicluded	Arbitrary nature of QALY calculations. QALY is an artificial construct. Factors that should be considered in addition to QALYs: • Urgency. • Severity. • Availability of alternatives.	Significant value attached to mobility but none to continence (NICE citizen council).
Wyeth	Narrow NHS and PSS perspective doesn't capture societal benefits of biologics e.g return to work, reduced state benefits	Claims that QALY can't be used to value innovation Proposes Multiple Criteria Decision Analysis in addition to cost-utility.	
NICE's reply to ABPI letter			
ABPI letter regarding NICE's submission	Would like NICE to consider carer benefits	Would like NICE to consider social value judgements and process of care issues e.g • Convenience • Dignity • Independence	Welcomes NICE's proposals to discuss limitations of EQ-5D with EuroQol

Consultee	Insufficient Perspective	Use of QALYs /cost-utility analysis	EQ-5D
NICE's reply to SHCA letter			
SHCA letter regarding NICE's submission			
NHS confederation	Restriction to NHS and PSS costs has been criticised	NICE doesn't take into account affordability and ability to decommission less cost-effective services. QALYs not designed for purpose NICE uses them for.	
DSU report	Difficulty arises in determining how much health should be forgone by others in trade for non-health benefits such as labour force benefits.	Argues that factors such as convenience of administration and reduce burden on families and carers are currently included through reduced administration costs and improved compliance.	Industry submission's primary concern is that the EQ-5D is too simplistic to capture the full range of benefits to patients. These concerns are currently given consideration during decision.
NICE	Costs and other significant non-health effects outside of NHS and PSS can be considered by agreement with the DoH.	A negative decision based on the ICER for a single treatment may prevent benefits being captured from future treatments which can only be given if progression is preventing using current treatment.	States that NICE has in the past taken into account stakeholder's cases that benefits are not adequately captured. Examples given are ECT, omalizumab for asthma natalizumab for MS. tates that NICE will discuss EQ-5D with EuroQol and DoH.
NRAS Minutes	Against restriction to NHS and PSS perspective.	QALY doesn't take account of how the patient values extension of life or quality of life.	Suitability of EQ-5D should be looked at by: • Patients in NICE (PIN). • Difficulty mapping clinical scores to EQ-5D. • EQ-5D not sensitive enough. • Based on population rather than patient preferences. • Utility measurement should be incorporated into patient access scheme agreements.

Consultee	Insufficient Perspective	Use of QALYs /cost-utility analysis	EQ-5D
NRAS submission	Change NICE's remit to include wider societal costs and work related disability Mentions costs for patients and carers related to lost employment, lost pension contributions and cost to society for state benefits and social care.	Concern regarding QALY concept (cites paper by Oxford University National Perinatal Epidemiology Study Unit, "valuing a QALY: a review of current controversies" which is an editorial on measuring the social value of QALYs).	EQ-5D doesn't reflect preferences of the individual but the general public's.
Patients involved in NICE	Against current restriction to NHS and PSS costs. Would like to see wider non-health benefits included such as employment, reduction in state benefits, reduced cost of privately funded social care.		
Shona Adams (3 documents)			
ABPI	Carer benefits: improving the carer's daily life, productivity and/or daily activities is an important aspect of some medicines. Although these can be measured, they often fall outside conventional measures of health related quality of life.	QALY maximisation should not be sole objective. Additional factors over and above value for money identified from NICE Citizens Council and NICE's Social Value Judgements include: • Treatments tackling health inequalities. • Severe ill health. • Extension of life for people with terminal illness. • Limited alternative treatments (unmet need). • Value of use in children. • Impact of disease rarity and small patient numbers on uncertainty in evidence.	Welcomes recognition of the insensitivity of EQ-5D to capture all relevant health benefits and opportunity to raise this at scoping stage. HRQoL measurement is area of academic debate and development including development of EQ-5D-5L. There are many adverse events, such as alopecia, fatigue, vomiting, dizziness or visual disturbances, that have a major impact on patients but whose reduction is unlikely to be detected by the EQ-5D.
Actelion Pharmaceuticals UK Ltd		Questions the ability of QALY based analysis to capture benefits of orphan drugs specifically. Broader assessment of value including disease severity, unmet need, innovation and overall budget impact.	
AstraZeneca UK Ltd			States that there are limitations of QALY measurement via EQ-5D (no further details) NICE should be actively involved in research to establish society's willingness to pay for benefits - such as the active involvement with the EuroQOL group research.
EMIG	Appraisal Committee should consider societal gains and patient convenience.	QALY is unreliable and should not be sole measures of value.	

Consultee	Insufficient Perspective	Use of QALYs /cost-utility analysis	EQ-5D
	Should include productivity gains and indirect effect of individual's health on wider economy and society.	Appraisal Committee should consider patient convenience.	
GSK		ICER does not capture all benefits important to patients. Concerns regarding application of NICE appraisal to drugs unlikely to achieve cost-effectiveness but which have a low budget impact e.g orphan and ultra-orphan. Need to value innovation in areas with unmet clinical need.	
Myeloma UK		Welcomes proposals to identify health related benefits that are relevant to the specific condition. Myeloma patient's preferences are unlikely to be similar to other group's preferences due to their experience of myeloma. These disease specific preferences are unlikely to be captured in the standard ICER/QALY methodology.	Patients with myeloma and their families discussed the EQ-5D and concluded that the EQ-5D is insensitive and likely to substantially underestimate the quality of life gains of myeloma treatment for myeloma patients. Further research on EQ-5D and other instruments is supported.
Pfizer Ltd		Welcome the specification of health care benefits currently not be captured by QALY Additional factors identified by NICE Citizens Council should be taken into account: • Tackling health inequalities. • Illness severity. • Extension of life in terminal illness. • Unmet need or long period without new treatments. • Use in children. • Rare diseases.	EQ-5D fails to capture much that is of value to patients and their carers and there is no robust alternative.

All the above submissions are available on the NICE website at
http://www.nice.org.uk/aboutnice/howwework/researchanddevelopment/KennedyStudyOfValuingInnovationSubmissions.jsp