NATIONAL INSTITUTE FOR HEALTH AND CARE EXCELLENCE

CHTE methods review

Health-related quality of life

Task and finish group report

July 2020

Contents

Summary of case for change and proposals

1. EQ-5D and non-reference-case measures of quality of life

Current methods

The NICE guide to the methods of technology appraisal acknowledges that EQ-5D is not always suitable for every condition. It details the evidence that should be provided to show this and notes that in these circumstances, alternative non-reference-case measures of health-related quality of life may be used. The task and finish group explored whether NICE should be more specific about when and how to use alternative measures of health-related quality of life and also whether to provide guidance on what to do when EQ-5D data are unavailable or are insufficient to populate a model.

Case for change – yes, minor

Evidence suggests the EQ-5D works well for most diseases and conditions except sensory disorders and some mental health conditions. For conditions where there is mixed evidence that EQ-5D performs well, a review of technology appraisals in these conditions shows it has been possible for committees to make recommendations based on EQ-5D.

Evidence would support specifying in the methods guide that the Health Utilities Index 3 (HUI3) is used instead of EQ-5D for hearing disorders, and to a lesser extent the Recovering Quality of Life (ReQoL) for some mental health conditions. However, we propose retaining the current broader guidance. This allows the case for the HUI3 or ReQoL to be made, but also would apply to other disease areas if evidence on the performance of EQ-5D becomes available. We suggest adding more guidance about which alternatives are preferred when it is shown that the EQ-5D is not appropriate.

A potential concern for rarer diseases is that there may be insufficient EQ-5D data to assess whether it adequately reflects changes in quality of life. Evidence other than psychometric measures could be presented and considered in these specific circumstances. But it is important to maintain the expectation that EQ-5D is used in most circumstances unless there is strong evidence that it is inappropriate.

There is currently no guidance on what to do if EQ-5D is not available in the clinical trials or the literature, and it is not possible to map from another measure to EQ-5D. This can be a problem in any appraisal where health states or events are rarely observed but is more common in appraisals for rare diseases. Previous appraisals and highly specialised technology evaluations show that vignettes are often used, but the methods of creating them and the approaches used to value them vary markedly. So, there is a case for providing more guidance about the preferred approach to measuring and valuing quality of life in these situations.

Proposals
Preferred option: Add hierarchy to methods guide (see figure 1 in section 4 of report 1: hierarchy of preferred health-related quality of life methods). This:

- Draws together the different situations in which EQ-5D is either not available or not appropriate.
- Restates some information that is already in the methods guide.
- For situations in which the EQ-5D is not available, includes new guidance on using vignettes and utility values from proxy conditions.
- For situations in which the EQ-5D is not appropriate, adds more detail than there is currently in the methods guide on alternative measures.

2. Carer quality of life

Current methods
The current methods guide states that the perspective on outcomes includes ‘all direct health effects, whether for patients, or when relevant, carers’. There is no further guidance on when and how carer health effects should be considered.

Case for change – yes, but not within this update
There is a case for providing more guidance on when and how to include carer health-related quality of life in appraisals. The Decision Support Unit found that this has been done inconsistently in previous appraisals and the general quality of the evidence was low. More explicit guidance could lead to a more consistent approach and higher-quality evidence.

The benefits from providing more explicit guidance must be weighed up against the possibility that doing so encourages more submissions to include carer effects. There are concerns that if carer health effects are considered in more appraisals, activity for which the carer benefits have not been considered may be displaced. Options for accounting for this are presented in the task and finish group report.

Providing clear guidance about when it is appropriate to include carer effects and the standard of the evidence that needs to be provided may also limit the increase in submissions and give committees clearer grounds for accepting or rejecting proposals. Specifying minimum evidence requirements could decrease uncertainty in appraisals including carer effects, which may outweigh the risk of more appraisals including it.

Proposals
A draft list of minimum evidence requirements has been produced. Many of the items included in the list require normative judgements. These should be discussed fully in a workshop with stakeholders representing patient groups, academia, committee members and industry. It has not been possible to do this in the current timescales.

It is also likely that some of the technical issues, relating to including carer health-related quality of life in economic models, need further research by academic groups.

3. Age-adjusted utility values

Current methods

The methods guide states ‘in some circumstances, adjustments to utility values, for example for age or comorbidities, may be needed’.

Case for change – yes, minor

Utility values from trials often need to be extrapolated over long time horizons. The general consensus is that when doing so, it is appropriate to adjust values to ensure that they do not exceed general population values, for example, ISPOR recommends this as best practice. Many submissions adjust utility values for this reason and appraisal committees normally prefer analyses that do.

There is more than 1 method for adjusting values, leading to inconsistencies. In the appraisals examined, most used the multiplicative method rather than the additive method. At higher utility values, the adjusted values are similar using either method, but at lower baseline values, the additive approach can result in values close to or less than zero, which does not occur with the multiplicative approach.

Exploratory analysis showed that adjusting utility values reduces health gain compared with no adjustment and, for a given disease, the impact is greater for cohorts with a younger starting age. Adjustment could have a greater impact on cost-effectiveness results at older starting ages, because the health gain is lower because of shorter life expectancy, but the examples explored did not indicate this. Overall, adjusting utility values over time does not appear to disproportionately affect older populations.

There may be some situations in which it is not appropriate to adjust utility values, so amendments to the methods guide should give the scope for those arguments to be made and considered by the committee.

Proposals

Update the methods guide to state:

- If baseline utility values are extrapolated over long time horizons, they should be adjusted to reflect decreases in quality of life seen in the general population and to ensure that they do not exceed general population values at a given age.
- If this is not considered appropriate for a particular model, supporting rationale should be provided.

- A multiplicative approach is generally preferred, and the methods used for adjusting utility values should be clearly documented.

4. Core outcome sets

Current methods

There is nothing specifically on core outcome sets.

Case for change – yes, minor

There is a concerted international effort to use core outcome sets in health technology assessment, and they are already used in other NICE programmes, including the Centre for Guidelines. The main advantages of core outcome sets are that they allow more comparative assessments between studies and outcomes are selected with patient and clinician input and subject to peer review.

The methods guide could be aligned with the Centre for Guidelines methods guide so that core outcome sets are identified and quality assessed during the scoping phase of appraisals. The task and finish group report sets out how this might be implemented.

However, a significant drawback of adopting core outcome sets is that searching and quality assessing them would increase the resource intensity of scoping for the NICE team. This additional effort may not be justified, given that a review of oncology scopes found a significant degree of overlap between outcomes that are routinely included in scopes and those in core outcome sets.

It is felt that some of the key benefits of core outcome sets could be achieved by encouraging in the methods guide that all outcomes should be relevant to patients.

Proposals

Update the methods guide to state that:

- Outcome measures in studies should be selected in consultation with people with the condition or disease, so that the study reflects what matters to them.
- A high-quality core outcome set, developed with input from people with the disease or condition, may help with outcome selection.
- Patient-reported outcomes can capture important aspects of conditions and interventions. Patient-reported outcome measures should be appropriately validated, and the methods used to collect the data should be clearly reported.

5. Children's health-related quality of life

Current methods

The methods guide says: when necessary, to consider using 'preference-based measures of health-related quality of life that have been designed for use in children'.

Case for change – yes, minor

The case for change has 4 components:

- A Decision Support Unit review of published guidance in children and young people showed wide variation in methods and poor reporting of the source of utilities.
- EQ-5D is not designed for use in children but is used in many appraisals including children and young people.
- Submissions rarely use age-appropriate measures so children and young people can assess their own health-related quality of life; this goes against NICE's patient and public involvement policy.
- Stakeholders would welcome clearer methods guidance.

However, the academic literature is not mature enough to recommend specific health-related quality of life measure(s) and value set(s).

Proposals

The proposed amendments for appraisals and evaluations that include children and young people are:

- Recommend measuring health-related quality of life using a generic measure that has been shown to have good psychometric performance in the relevant age range(s).
- Explain desirable characteristics of a measure but do not recommend specific measures.
- Recommend clear reporting on who completes the measure and how utility values for the model were calculated and selected.

6. EQ-5D-5L value set

Current methods

NICE's position statement on the use of the EQ-5D-5L value set for England says that the EQ-5D-5L questionnaire may be used to collect quality of life data but that the UK EQ-5D-3L value set should be used. NICE currently recommends the van Hout tool for mapping 5L to 3L.

Case for change – yes, minor

The Decision Support Unit has developed an alternative to the van Hout mapping tool. There is no strong reason for choosing 1 method over the other based on performance metrics alone. One advantage of the Decision Support Unit method is that it allows mapping from utility values, not just directly from questionnaire responses.

Proposals

Most of the position statement should be incorporated into the methods guide. The exception is the recommendation on mapping. The updated methods guide should recommend the Decision Support Unit rather than the van Hout mapping tool. A new data set to inform mapping is expected be available in August 2020. When the new data set is available, the NICE technical team will form recommendations on which data set to recommend and then consult the working group.

Report 1: EQ-5D and non-reference-case measures of quality of life

Background

NICE's preferred measure of health-related quality of life is the EQ-5D. However, there may be situations in which EQ-5D is not the most appropriate instrument, and some situations in which EQ-5D data are not available. The report is divided into 5 sections:

Section 1: EQ-5D not appropriate

The NICE guide to the methods of technology appraisal states: 'In some circumstances the EQ-5D may not be the most appropriate… empirical evidence on the lack of content validity for the EQ-5D should be provided… in these circumstances alternative health-related quality of life measures may be used…'.

However, although stakeholders have sometimes argued that the EQ-5D is not appropriate for a particular disease, appraisal committees have rarely been presented with the sort of evidence specified in the methods guide to support these claims.

This report will explore whether the methods guide should be more specific about when and how to use other measures of health-related quality of life instead of the EQ-5D (that is, 'non-reference-case measures').

Section 2: EQ-5D unavailable

The methods guide does not provide guidance on what to do when EQ-5D data are unavailable or available EQ-5D data are insufficient to populate an economic model.

Health-related quality of life data are usually obtained from clinical trials of the intervention or from the literature. When evaluating technologies for rarer diseases, it is often the case that such data are sparse or poor quality. However, the problem of having insufficient health-related quality of life data to populate an economic model is not unique to rare diseases. For some of the treatments assessed by the technology appraisals programme, there may be certain health states or events in models that occur infrequently in trials or are challenging to measure.

This report will consider whether the methods guide should provide more guidance on what to do in situations when the EQ-5D is not available.

Section 3: Rare diseases

The specification for the health-related quality of life task and finish group specifically mentions measuring and valuing quality of life for rare diseases. Aspects of both pieces of work on situations when the EQ-5D is not appropriate and when EQ-5D is not available may apply in rare diseases. Section 3 of the report will summarise the relevant conclusions for rare disease.

Section 4: Case for change and options
The case for changes to the methods guide will be assessed and options presented.

Section 5: Equality considerations
Equality implications will be described and considered.

Section 1: EQ-5D not appropriate

1.1 Decision Support Unit (DSU) reports and published literature on EQ-5D performance

The appropriateness of the EQ-5D for a specific condition is examined by assessing validity, responsiveness and reliability. Definitions of these criteria are summarised here and described in detail in the DSU technical support document on an introduction to the measurement and valuation of health for nice submissions (TSD8).

Validity is defined as ‘the extent to which an instrument measures what it is intended to measure’. Different criteria can be assessed including:

- Content validity: whether the instrument excludes dimensions of health that are important to patients.
- Construct validity: the extent to which the scores produced by a measure agree with other measures of the dimensions of health-related quality of life considered relevant to patients. Usually tested by assessing 1 of the following:
  - Convergent validity: whether 2 instruments of similar concept agree with each other. Can be evaluated by using correlations, to assess the degree to which the questionnaire is related to comparison measure.
  - Known groups differences: whether the instrument is able to distinguish between groups known or expected to differ in their characteristics (for example, severity of condition, or patients versus general population).

Responsiveness concentrates on the capacity of the instrument to reflect clinically significant changes in health, for example, by comparing people’s health-related quality of life before and after a successful treatment. Change is typically assessed by examining whether there are statistically significant differences in utility scores.

Reliability is the ability of a measure to reproduce the same value on 2 separate administrations when there has been no change in health. This can be over time, between methods of administration or between raters.

1.1.1. DSU reports

In 2010, the DSU examined claims that EQ-5D-3L was not appropriate in disease areas such as cancer and mental health. As part of this work, it produced a number of reports investigating the evidence and proposing solutions in addition to a number of technical support documents (TSDs). At the time, the DSU and the NICEQoL project supported by the Medical Research Council focused on the appropriateness of EQ-5D-3L in asthma, urinary incontinence, rheumatoid arthritis, visual disorders (Wailoo et al. 2010, Tosh et al. 2010, Tosh et al. 2012) and hearing disorders (Yang et al. 2010, Yang et al. 2013). There was limited evidence that EQ-5D-3L was not appropriate in asthma, urinary incontinence and rheumatoid arthritis. There were mixed results for visual disorders and the EQ-5D-3L did not perform well in studies of hearing disorders. Additionally, the DSU produced a series of 5 technical support documents on utilities (technical support documents 8 to 12) in the area of measuring and valuing health benefits for use in economic evaluation. The technical support document on alternatives to EQ-5D for generating health state utility values (TSD11) discusses the evidence required and the methods to use to show that EQ-5D is not appropriate, it discusses and assesses the alternatives to EQ-5D for generating utility values when it is proven to be not appropriate. A summary of this body of evidence and its conclusions is given below.

The DSU (Wailoo et al. 2010) aimed to assess and verify claims that EQ-5D-3L is inappropriate in specific disease areas, and identify solutions and alternatives to overcome any deficiencies. The authors investigated claims made during the Kennedy review of the value of innovation (Kennedy 2009) and specific technology appraisals (TAs). They performed systematic reviews in asthma, urinary incontinence and rheumatoid arthritis. The authors also referred to a systematic review for visual disorders (Tosh et al. 2010, Tosh et al. 2012). From the Kennedy review, the authors identified claims that EQ-5D-3L is not sensitive to detect changes in health or that not all relevant health issues are captured for evaluating quality of life in incontinence, mental health, cancer and fertility. For mental health and cancer, no or little evidence was provided by the claimants. The authors identified and discussed 5 previous NICE technology appraisals where it was claimed that the reference case would not capture all relevant health benefits. The disease areas were multiple myeloma, asthma, macular degeneration, deafness and diabetes.

In the asthma review, the authors identified 93 publications, of which 15 were included in the review. Most were cohort or cross-sectional studies and 1 randomised controlled trial. The authors concluded that EQ-5D-3L could distinguish between known groups and could reflect changes in health over time. Overall, there was no difference in validity and responsiveness between results from EQ-5D-3L and alternative generic preference-based measures (SF-6D, HUI3). However, disease-specific measures such as AQLQ were more sensitive.

In the incontinence review, the authors identified 68 publications, of which 17 were included in the review. These studies included randomised controlled trials, cohort studies and cross-sectional studies. Overall, EQ-5D-3L showed known-group validity and responsiveness and performed as well as alternative generic preference-based measures (SF-6D, AQoL [Assessment of Quality of Life] measures) and disease-specific measures (I-QoL).

In the rheumatoid arthritis review, the authors identified 1 previously published review article and 1 cross-sectional study. The published review focused on the validity and comparative performance of generic preference-based measures in rheumatoid arthritis and was not limited to EQ-5D-3L. That review included 26 publications. There were good correlations between EQ-5D-3L and alternative generic instruments (SF-6D, Health Utilities Index 2 [HUI2], HUI3) and condition-specific measures such as RAQoL and HAQ. EQ-5D was the most responsive measure when health deteriorated but less responsive than other measures when health improved.

1.1.2. Published systematic reviews

Since the last update of the methods guide in 2013, there have been a number of peer-reviewed systematic reviews that examined the appropriateness of EQ-5D-3L for assessing quality of life in different conditions. We used a targeted approach to identify the most recent relevant reviews. These are summarised below.

Payakachat et al. 2015
This review focused on the responsiveness of the EQ-5D to detect meaningful change in health status across multiple conditions. The authors conducted a systematic review of published articles reporting psychometric properties of the EQ-5D. The assessment of the EQ-5D responsiveness was based on the utility values only, whereas data collected using the EQ-VAS (visual analogue scale) were not considered. The responsiveness evidence was reported from the systematic reviews as well as additional evidence from recent literature.

Responsiveness was defined as ‘the extent to which an instrument can detect a clinically significant or practically important change over time’. Three relevant measures were developed based on this definition and included the differences in the EQ-5D health utility scores between responders and non-responders by clinical or self-reported measure and the change in EQ-5D health utility values over a period of time in which health status is expected to change.

There was a large degree of heterogeneity between studies in terms of study design, population characteristics, outcome measures and methods used to assess responsiveness. A total of 145 studies were included in the systematic review, focusing on 56 conditions within 19 categories of disease area.

EQ-5D was found to be responsive in 25 conditions: cardiac rehabilitation, heart failure, stroke, chronic opioid dependence, mental health, attention deficit hyperactivity disorder, anxiety and depression, social phobia, somatoform disorder, inflammatory arthritis, rheumatoid arthritis, asthma, chronic obstructive pulmonary disease, prostate and breast cancer, liver metastases, multiple myeloma, dementia, epilepsy, type 2 diabetes, surgery, HIV, pain and skin conditions (psoriasis, acne).

The authors concluded that EQ-5D lacked responsiveness in 4 conditions: alcohol dependency, schizophrenia, limb reconstruction and hearing impairment. In the other 27 conditions, there was limited or mixed results for the EQ-5D responsiveness.

There was some heterogeneity between studies relating to measuring responsiveness. The authors found that the EQ-5D was more responsive in severe conditions or if the change observed was large, rather than mild to moderate conditions and small changes. Additionally, the timing of follow up is important when measuring responsiveness and it should be adapted to the condition. The different country-specific value sets used to derive utilities can affect the magnitude of responsiveness. Moreover, coping with the disease may affect health change detection.

The authors recommended that in conditions where there is mixed evidence of responsiveness, the use of condition-specific measure along with the EQ-5D should be considered. They also suggested that future research using the EQ-5D with 5 response levels should focus on these conditions.

Trenaman et al. 2017, rheumatology

This review aimed to identify which patient-reported outcome measures (PROMs) are used to generate quality-adjusted life years in rheumatology and reach a consensus regarding the subdomains that might be missing from the most commonly used PROMs identified from the search (EQ-5D and SF-6D). The authors included 39 studies in their final analysis. In these studies, 5 generic preference-based measures were identified: EQ-5D, SF-6D, 15D, HUI3 and Quality of Well Being Scale (QWBS).

Of these, EQ-5D was the most commonly used (32 studies across 5 rheumatic conditions), followed by SF-6D (9 studies across 3 rheumatic conditions). A special interest group consisting of 23 participants, including methodologists (n=8), clinicians (n=13) and patients (n=2), compared these 2 instruments with the ASAS-HI, which is a disease-specific measure, to identify the subdomains that might be missing from EQ-5D and SF-6D. Participants identified energy or drive, and sleep as the 2 key subdomains that were missing from the EQ-5D. They expressed concerns regarding the focus of these measures only on health rather than other process or non-health outcomes. They were also concerned about the wording of the levels, which may not fully reflect life with rheumatic diseases.

Three potential ways of addressing these issues were proposed: 1) using an alternative generic measure that is not currently being widely used (Computerised Adaptive Tool-5 Domains), 2) the use of bolt-on subdomains, or 3) to generate a set of societal weights for an existing condition-specific PROM. No clear consensus around a preferred method was reached and research was recommended to assess the value and feasibility of these approaches.

Cooper et al. 2017, HIV

This overview of systematic reviews assessed both generic and HIV-specific health-related quality of life measures. Nine generic measures met their inclusion criteria. These were the COOP/WONCA charts, EQ-5D, FLZM Questions on Life

Satisfaction, HUI, McGill Quality of life questionnaire, SF-12, SF-20, SF-36 and WHOQOL-BREF.

The authors reported that 4 included reviews that assessed EQ-5D for use in adults with HIV provided evidence of construct and convergent validity, as well as responsiveness to treatment initiation, the development of opportunistic infections and adverse effects with small to medium effect sizes.

The authors reported that EQ-5D has frequently been used in research with people with HIV, and several authors recommended it for use in this population. However, the authors noted that its use in individuals with early, asymptomatic HIV infection may not be recommended because of problems with ceiling effects. They suggested that this applies to the EQ-5D-3L, whereas the newer EQ-5D-5L is likely to be more suitable. Of the reviews included in Cooper et al. 2017, Wu et al. (2013) recommended the use of EQ-5D alongside an HIV-specific measure (the MOS-HIV) to obtain HIV-specific quality of life alongside this utility measure.

The authors concluded that the measures supported with most psychometric evidence in the included systematic reviews were the EQ-5D, SF-36, WHOQOL-BREF and MOS-HIV.

Finch et al. 2018, multiple conditions

This is the most comprehensive and up-to-date review identified. The objective of this review of systematic reviews was to summarise the validity and responsiveness of 5 generic preference-based measures: EQ-5D, SF-6D, HUI3, AQoL and 15D (15 Dimensions), across a variety of disease areas. Where possible, the authors followed the 27-item PRISMA checklist for systematic literature reviews and meta-analyses.

The main inclusion criteria were:

- Population: adult patients (18 years or over).
- Intervention/Comparators: 1 or more of EQ-5D, SF-6D, HUI3, 15D, AQoL.
- Outcomes: responsiveness and validity, results reported at study level.
- Study type: systematic literature reviews (unless reported aggregated results, were not in English, or only available in a poster presentation).

The quality of each review was assessed using a modified version of the AMSTAR checklist for systematic reviews in which the weight of importance of each item was based on the research team's views. The questions on the comprehensiveness of the search, the presence of a quality assessment tool and the use of quality scores to formulate conclusions were considered essential and had the most weight.

A total of 30 reviews were included. The reviews differed by the number of studies included, which varied from 5 to 122. Studies included in the reviews were a mix of randomised controlled trials, cross-sectional, cohort and longitudinal studies, or other experimental or observational designs. The studies included in the reviews were not all relevant to the research question; however, more than 180 studies included in the 30 reviews were of interest and provided evidence.

Among these studies, 3 reviews were related to the DSU reports mentioned in section 1.1.1 (Davis and Wailoo 2013 on urinary incontinence, Tosh et al. 2012 on visual disorders and Yang et al. 2013 on hearing disorders).

The quality of included reviews varied, with 2 reviews assessed as excellent quality, 14 of good quality and 14 of poor quality. The main reason for poor quality assessment was that the quality of papers included in the review was not assessed.

Most of the reviews reported information on EQ-5D (29 reviews), followed by SF-6D (12 reviews), HUI3 (8 reviews), AQoL (3 reviews) and 15D (2 reviews). EQ-5D psychometric measures were reported for conditions across 16 disease classes plus aesthetic surgery and older population. The SF-6D was reported for conditions across 9 disease classes, HUI3 across 7 classes, and 15D and AQoL across 2 classes.

The type of evidence reported varied between reviews. For known-group testing, comparisons were based on disease severity, patients versus general population, different types or number of diseases and conditions, and patients with or without complications. In terms of convergent validity, most studies reported correlations with other measures. Responsiveness was mostly based on the comparison of people’s health-related quality of life before and after treatment, but also on patient groups receiving different treatments.

Findings relating to EQ-5D

The findings on the performance of EQ-5D by condition is summarised in Table ***1*** below. The main findings were in terms of validity and responsiveness. For validity, 2 criteria were tested; known groups and convergent validity.

According to the authors, EQ-5D was found to have good validity or responsiveness in type 2 diabetes, urinary incontinence, depression and anxiety, cancer, injuries, skin conditions and respiratory conditions. The EQ-5D was found to have poor validity and/or responsiveness for:

- hearing impairment
- multiple sclerosis
- personality disorders
- schizophrenia
- dementia.

Mixed results were reported in visual disorders, bipolar disorder, heart disease and HIV.

The authors reported that there was a lack of reviews for some disease areas or group of patients that were sparsely investigated: autoimmune system, haematological problems, gynaecological problems, musculoskeletal conditions, conditions related to the nose, and some reviews focused on older patients.

Table 1 Summary of findings for EQ-5D – Finch et al. 2018

Disease area	Review	Condition/pop	Quality of review	Performance of EQ-5D by condition (number of studies) Validity – Known groups	Performance of EQ-5D by condition (number of studies) Validity – Correlation	Performance of EQ-5D by condition (number of studies) Responsiveness
Autoimmune system	Castelino 2013	Systemic lupus erythematous	Poor	Not reported	Sparsely investigated (1)	Sparsely investigated (1)
Autoimmune system	Holloway 2014	Systemic lupus erythematous	Poor	Sparsely investigated (1)	Sparsely investigated (1)	Not reported
Cardiovascular	Dyer 2010	Heart disease	Good	Mixed results (12)	Mixed results (6)	Mixed results (33)
Ear	Yang 2013	Hearing impairment	Good	Poor validity (1)	Poor validity (2)	Poor responsiveness (4)
Endocrine, nutritional and metabolic diseases	Janssen 2011	Type 2 diabetes	Good	Good validity (21)	Good validity (9)	Good responsiveness (7)
Endocrine, nutritional and metabolic diseases	Speight 2009	Type 2 diabetes	Poor	Not reported	Not reported	Not reported
Eye	Tosh 2012	Visual impairment	Good	Mixed results (25)	Mixed results (9)	Mixed results (3)
Genitourinary system	Davis and Wailoo 2013	Urinary incontinence	Good	Good validity (5)	Good validity (9)	Good responsiveness (8)

Disease area	Review	Condition/pop	Quality of review	Performance of EQ-5D by condition (number of studies) Validity – Known groups	Performance of EQ-5D by condition (number of studies) Validity – Correlation	Performance of EQ-5D by condition (number of studies) Responsiveness
Genitourinary system	Wu 2013	HIV	Good	Good validity (1)	Not reported	Mixed results (5)
Gynaecological problems	Sanghera 2013	Menorrhagia	Poor	Sparsely investigated (1)	Sparsely investigated (2)	Sparsely investigated (1)
Haematological problems	Szende 2003	Haemophilia	Good	Good validity (2)	Sparsely investigated (1)	Not reported
Musculoskeletal system	Bansback 2008	Rheumatoid arthritis	Poor	Not reported	Sparsely investigated (1)	Not reported
Musculoskeletal system	DeVine 2011	Chronic low back pain	Poor	Not reported	Sparsely investigated (1)	Sparsely investigated (1)
Mental health	Brazier 2014	Bipolar disorder	Good	Mixed results (3)	Mixed results (5)	Not reported
Mental health	Papaioannou 2013	Personality disorder	Good	Mixed results (3)	Mixed results (2)	Good responsiveness (3)
Mental health	Papaioannou 2011	Schizophrenia	Good	Good validity (1)	Poor validity (8)	Mixed results (3)
Mental health	Peasgood 2012	Depression/ anxiety	Good	Good validity (10)	Good validity (6)	Good responsiveness (17)

Disease area	Review	Condition/pop	Quality of review	Performance of EQ-5D by condition (number of studies) Validity – Known groups	Performance of EQ-5D by condition (number of studies) Validity – Correlation	Performance of EQ-5D by condition (number of studies) Responsiveness
Neurological conditions/ Nervous system	Hounsome 2011	Dementia	Poor	Not reported	Mixed results (8)	Not reported
Neurological conditions/ Nervous system	Kuspinar and Mayo 2014	Multiple sclerosis	Excellent	Poor validity (4)	Mixed results (6)	Not reported
Neoplasm	Longworth 2014	Cancer	Good	Good validity (31)	Good validity (17)	Good responsiveness (43)
Neoplasm	Pickard 2007	Cancer	Poor	Good validity (8)	Good validity (1)	Good responsiveness (2)
Nose	Linder 2003	Acute sinusitis	Excellent	Not reported	Not reported	Sparsely investigated (1)
Others	Ching 2003	Aesthetic surgery	Poor	Not reported	Not reported	Sparsely investigated (1)
Others	Derrett 2009	Injuries	Poor	Good validity (4)	Good validity (7)	Poor responsiveness (1)
Others	Haywood 2005	Older patients	Poor	Not reported	Sparsely investigated (1)	Sparsely investigated (1)
Respiratory system	Petrillo 2011	Chronic obstructive	Poor	Mixed results (2)	Not reported	Good responsiveness (3)

Disease area	Review	Condition/pop	Quality of review	Performance of EQ-5D by condition (number of studies) Validity – Known groups	Performance of EQ-5D by condition (number of studies) Validity – Correlation	Performance of EQ-5D by condition (number of studies) Responsiveness
		pulmonary disease				
Respiratory system	Pickard 2008	Asthma/chronic obstructive pulmonary disease	Good	Good validity (11)	Good validity (8)	Mixed results (4)
Skin and subcutaneous tissues	Yang 2015	Psoriasis, acne, hand eczema, leg ulcers	Good	Good validity (9)	Good validity (7)	Good responsiveness (11)

Findings on other instruments

The SF-6D was found to have good validity or responsiveness in depression and anxiety, and the nervous and genitourinary systems. In age-related macular degeneration, it showed better performance than EQ-5D; however, evidence was limited to 1 study. SF-6D was also found to have good convergent validity in multiple sclerosis but no data were available for responsiveness.

The HUI3 was found to perform better than EQ-5D in hearing impairment, with most of the responsiveness tests showing an ability to detect changes in health status before and after treatment. HUI3 also showed good validity and/or responsiveness in cancer, disease of the eye, the ear, the nervous system, depression and anxiety. HUI3 was found to have good validity in multiple sclerosis (1 study) and in depression and anxiety.

Good psychometric properties were reported in musculoskeletal and genitourinary conditions for AQoL and genitourinary, diabetes, nutritional and metabolic conditions for 15D.

The authors concluded that most of the evidence retrieved was on EQ-5D. SF-6D and HUI3 were investigated in substantially fewer systematic reviews. The psychometric assessment also included some limitations because some studies only reported convergent validity, or reported comparisons with only 1 indicator, limiting the conclusions that can be drawn. It was also not clear whether the included reviews published after 2009 focused only on EQ-5D-3L or also included EQ-5D-5L.

The authors underlined that whenever evidence was available, it would usually support the performance of the generic measures. However, there is inconsistency in the breadth and depth of the evidence between disease areas, instruments and type of assessment, as well as a lack of head-to-head comparison between the instruments. As a result, any attempt to compare the instruments is limited.

1.1.3. Summary

Overall, the DSU reports and published systematic reviews considered covered multiple conditions including hearing, vision, mental health, cancer, rheumatology and others. The most recent overview of systematic reviews identified was that by Finch et al. 2018. It includes the studies conducted by the DSU between 2010 and 2014 in addition to other reviews on the topic. The search cut-off date for that review, however, was in 2016.

To ensure that no major reviews have been published since this cut-off date, we ran a focused search in PubMed restricting it to reviews published since 2016 using the terms ‘quality of life’, ‘psychometric properties’, ‘validity’, ‘responsiveness’ and EQ-5D. This did not identify any more recent overviews of systematic reviews. However, systematic reviews focusing on single conditions were identified. These covered HIV, hip fracture, lumbar surgery, Parkinson’s disease, peripheral arterial disease, low back pain and oncology. None of these reviews raised concerns regarding the appropriateness of EQ-5D-3L.

The findings of our targeted review of the available evidence suggest that there is evidence that EQ-5D may not be appropriate for hearing-related conditions, where HUI3 has been found to perform better on psychometric tests (Finch et al. 2018). For visual impairment, evidence suggested that EQ-5D did not perform well for age-related macular degeneration and diabetic retinopathy. Results were mixed in cataracts, whereas evidence supported its use in other eye conditions (for example, conjunctivitis, Tosh et al. 2012).

For mental health, conclusions on the appropriateness of EQ-5D varied from 1 condition to another with no problems identified for depression and anxiety, whereas evidence suggested that EQ-5D might not be appropriate for schizophrenia, personality disorders and alcohol dependency. EQ-5D may also lack validity and/or responsiveness in HIV and dementia. Mixed evidence on EQ-5D appropriateness was found in multiple sclerosis. However, no alternative measure was found to be universally more appropriate than EQ-5D for all psychometric properties of interest.

It must be noted that some of the studies included in these reviews were relatively old and the natural history of some of the conditions covered could have changed over the intervening period. However, this is a general limitation of the body of evidence available in this area. Additionally, the reviews examined did not specify for any of these conditions whether they have focused on specific levels of severity or stage. As such, the conclusions drawn here are general to these conditions.

Finally, none of the reviews identified covered the appropriateness of EQ-5D use in rare and ultra-rare diseases, so their findings cannot be generalised to these conditions. A discussion paper reporting on a systematic review of quality of life instruments for Duchenne muscular dystrophy highlighted the very limited evidence on the validity of EQ-5D and assessed that available evidence is of very low quality, flagging the need for better research in this area (Powell et al. 2019).

1.2 Alternatives when EQ-5D not appropriate

The current methods guide specifies that where evidence exists that EQ-5D is not appropriate in a disease area, alternative measures may be used. The choice of an alternative measure or method to use to generate utilities must be accompanied by a carefully detailed account of the methods used to generate these values, their validity, and how these methods affect the utility values compared with using EQ-5D.

In 2011, the DSU produced a technical support document on alternatives to EQ-5D for generating health state utility values (TSD11) examining the validity and appropriateness of the alternatives to be used when EQ-5D is considered inappropriate. The cases where EQ-5D was considered inappropriate were specified as those where relevant domains of health known to be affected by the disease are clearly absent from EQ-5D (lack of content and construct validity) or where the performance of EQ-5D on other psychometric tests assessing reliability and responsiveness have shown it performs poorly.

In these cases, alternatives were suggested with the caveat that evidence supporting the appropriateness of the chosen alternative should be provided. This should take the form of a study comparing the performance of the chosen alternative and the EQ-5D in terms of content validity, construct validity, reliability and responsiveness. However, it was stressed that 'any decision by NICE regarding the appropriateness of one measure over another is ultimately a judgement'.

The 4 alternatives and the DSU's recommendations are discussed below.

Other generic preference-based measures
Generic measures of health are developed for use across all patient groups by focusing on the core dimensions of health. The descriptive system consists of a number of domains, each of which has several response options or 'levels'. Combinations of the different domains and their levels describe several health states (for example, EQ-5D-3L has 5 domains with 3 levels, which generate 243 unique health states). Preference-based measures are those where the descriptive system is accompanied by a value set generated through valuation techniques (for example, time trade-off, standard gamble and rating scales). The values generated can be obtained from surveys of the general population or patients. NICE's reference case specifies that the values should be obtained from a general population sample.

The most commonly used generic measure is EQ-5D. Others include SF-6D, HUI3 and AQoL. Despite being generic, resultant utility values from measures can differ substantially. One of these measures may have a role if EQ-5D is found to be inappropriate. They are preferred to condition-specific measures because they can reflect the impact of adverse effects and comorbidities not assessed in condition-specific measures. Condition-specific measures also have other limitations (see below).

Condition- or disease-specific preference-based measures of health
The acceptability of a condition-specific preference-based measure depends on the content of the descriptive system as well as the valuation technique used to generate utility values. Rigorous application of qualitative methods and psychometric analysis techniques to generate the descriptive system, either novel or from an existing non-preference-based condition-specific patient-reported outcome measure, is necessary as well as the focus on health-related quality of life rather than symptoms only.

Limitations of condition-specific measures are summarised in the DSU technical support document on alternatives to EQ-5D for generating health state utility values (TSD11) as follows:

- Largely describe symptom/symptom burden rather than measure health-related quality of life.
- Some use valuation techniques that produce values rather than utilities, because they are not choice-based (for example, visual analogue scale).
- Prone to a ‘focusing effect’ during the valuation task, which results in overestimation of utility decrements.
- Some measures result in utility values not anchored onto the 1 to 0 scale (representing full health to dead).
- Inability of some measures to capture the impact of comorbidities and side effects of treatment due to being focused on the disease impact.

TSD11 recommends that for a condition-specific measure to be acceptable, its descriptive system should be based on a validated measure and the method used to derive its value set should be comparable with the methods used to derive the EQ-5D-3L value set.

Direct valuation of health state vignettes

Vignettes are used to describe the health sates associated with a disease or condition. Respondents are then asked to value these health states using the standard valuation techniques mentioned above (for example, time trade-off and standard gamble). The advantages of using vignettes, as listed in TSD11, are that they are relatively easy to construct, can be prepared with no patient data, allows valuing health states that are not possible to collect data for either practical or ethical reasons and they can be constructed to capture comorbidities or side effects of interest. However, they are still disease-specific. The validity will also depend on the rigour with which they were developed and whether they have been validated by external independent experts. Additionally, there is a limit to the number of vignettes that respondents can value, therefore the full distribution of the likely utility values would not be available. For these reasons, vignettes are considered to have a very limited role as an alternative to use when EQ-5D is not appropriate, particularly those that are not developed based on the domains included in validated health-related quality of life measures.

Direct valuation of own health

Asking patients to value their own health state is fundamentally different to generating these values by asking members of the general public. It is reported that patient valuations of physical health states are usually higher, because of adaptation and their valuation of mental health states are usually lower compared with the general public. There are also ethical and technical challenges involved in asking patients to value their own health using techniques that use life and death questions such as the time trade-off and standard gamble. This approach does have the advantage of avoiding the reported limitations associated with using descriptive systems such as poor coverage, insensitivity and lack of meaning. Hence, utility values derived using this approach should be considered very carefully on a case by case basis.

Other options or developments

EQ-5D bolt-ons

A ‘bolt-on’ is an additional sixth domain that is ‘bolted’ on to the existing 5 domains of the EQ-5D. The use of bolt-ons to address the problem of the dimensions missing from EQ-5D-3L has been considered and research is still ongoing in this area. Areas where this research has been undertaken include sleep, vision and hearing, fatigue and cognition bolt-ons (Brazier at al. 2019; Finch et al. 2019). Findings from this research suggest that not all bolt-ons have significant impact on utility values (for example, sleep bolt-on was not found to have significant effect). Brazier explained that the observed effect of these bolt-ons was not additive as expected but rather results in changes in the magnitude of the coefficients of the original 5 dimensions. This signals that there is overlap between the original 5 dimensions and the bolt-ons, which in turn means that any new additional dimension will need a new value set to be estimated. Finally, the original concern of compromising the generic nature of the measure will remain.

EQ-5D-5L

The newer, 5-level version of EQ-5D is reported to have fewer ceiling effects and be more sensitive than 3L. Thus, the newer 5L descriptive system (which is recommended in the methods guide) may have better responsiveness, and potentially convergent validity, than 3L. The 3L, however, is the most relevant version because it is the one whose value set is currently used in NICE submissions.

Moreover, the psychometric properties of EQ-5D depend on the value set used to generate the utility values in the analysis. A new EQ-5D-5L valuation study for the UK is expected to be completed by the middle of 2021. This development is covered in detail elsewhere under the quality of life task and finish group

Recovering Quality of Life (ReQoL)

ReQoL is a new patient-reported outcome measure that has been developed to assess the quality of life for people with different mental health conditions. There are 2 versions of the ReQoL: ReQoL-10 and ReQoL-20 with 10 and 20 mental health items respectively. Both versions contain 1 physical health question. They are suitable for use across all mental health populations including common mental health problems, severe and complex and psychotic disorders (except dementia and learning disabilities). Preference weights to generate quality-adjusted life years for use in cost-effectiveness studies were derived using the ‘measurement and valuation of health time trade-off protocol’ for this valuation. This is the same protocol used to value EQ-5D-3L in the UK.

The ReQoL was developed with considerable input from mental health service users, and the ReQoL-20 has been selected by the International Consortium for Health Outcome Measurement (ICHOM) as the quality of life measure in their standard set of measures for psychotic disorders.

One concern raised with condition-specific measures is that if the classification system is focused on a particular set of symptoms, the impact on quality of life can be exaggerated because these have not been placed within the context of other symptoms or more generic aspects of health. However, the developers attempted to minimise this effect by including a question about physical health.

Future research is planned to compare the relative psychometric performance of ReQoL and EQ-5D and SF-6D in trials.

Extending the quality-adjusted life year (QALY)
The University of Sheffield School of Health and Related Research's (ScHARR's) extending the QALY project aims to develop a broad measure of quality of life for use in economic evaluations across health and social care. According to the project website, there are key distinctions between existing health and quality of life measures, such as EQ-5D, and this new instrument, such as:

- Capturing the benefits of interventions in health, social care, and for carers with the aim of informing resource allocation decisions across healthcare, social care and public health.
- Including aspects of quality of life important to patients, social care users and carers and are impacted by their health condition, the care or treatment they receive or their caring role.

The project began in May 2017 and is expected to conclude by the end of 2020. The project is currently at the stage of selecting the questions to be used in the long and short forms of the measure. In 2020, the researchers will do a valuation study of the short form in England. NICE is supporting the project. When the new instrument is developed and a value set is available, NICE will assess its performance and consider whether and when it should be used to inform NICE evaluations.

1.3 Case studies from published technology appraisals

The literature summarised in the preceding sections of the report showed possible lack of appropriateness of EQ-5D-3L for hearing disorders, and mixed evidence on its appropriateness for:

- some visual disorders (such as age-related macular degeneration, diabetic retinopathy and cataracts)
- some psychological and mental health conditions (such as personality disorders, schizophrenia, bipolar disorders and alcohol dependency)

- dementia
- multiple sclerosis.

NICE has appraised interventions in some of these conditions. This section aims to summarise the committee's conclusions about the health-related quality of life measures presented in those appraisals.

Background details of the included technology appraisals are presented in appendix 1. In addition, for rare diseases, no literature was available to assess whether the EQ-5D is appropriate or not. As such, a summary of the health-related quality of life measures used in NICE's 12 published highly specialised technologies guidance and the committee's considerations were also summarised (see section 3).

A. Hearing disorders

One technology appraisal for hearing disorders was identified: cochlear implants for children and adults with severe to profound deafness (TA556), which is a multiple technology appraisal. Three cost-effectiveness models were available, 2 from participating companies and 1 from the assessment group. All models included the preference-based measure HUI3, which was accepted by the committee. It is unclear from the final appraisal document if the committee discussed EQ-5D or its inappropriateness for this appraisal.

B. Visual disorders

The literature summarised in the preceding sections of the report suggested that the EQ-5D may not be appropriate for age-related macular degeneration, diabetic retinopathy and cataracts, so technology appraisals for these conditions were searched for. There were no appraisals for diabetic retinopathy, glaucoma or cataracts. Two technology appraisals for age-related macular degeneration were identified: ranibizumab and pegaptanib for the treatment of age-related macular degeneration (TA155; multiple technology appraisal) and aflibercept solution for injection for treating wet age-related macular degeneration (TA294; single technology appraisal).

- TA155: The assessment group and 1 of the companies used utilities derived from people with age-related macular degeneration (Brown et al. 2000), whereas the other company used utilities derived from the general population using simulation contact lenses (Brazier et al. 2006). Both sets of utility values had been derived using time trade-off direct elicitation. The committee was aware that a generic preference-based measure, such as the EQ-5D or HUI3, would have been more appropriate. However, it agreed that, based on study results, HUI3 (Espallargues) may not fully capture the impact of age-related macular degeneration on people's quality of life. The committee concluded that on balance, the Brazier utility values provided the most plausible set of utility values for use in the economic models. It is unclear from the final appraisal document if the committee discussed EQ-5D's inappropriateness for this appraisal.

- TA294: The company included EQ-5D in its base-case analysis and did a scenario analysis using time trade-off derived utility values from a study by Czoski-Murray et al. (2009) that used simulation contact lenses. The evidence research group used time trade-off derived utility values (Brown et al. 2000) for the better seeing eye and company's proposed utility values for the worse-seeing eye. The final appraisal document does not refer to the committee making specific conclusions about health-related quality of life and utility values.

C. Psychological and mental health conditions

C.1 Schizophrenia and bipolar disorder

One relevant technology appraisal for schizophrenia and 1 for bipolar disorder were identified as follows:

- Schizophrenia: aripiprazole for the treatment of schizophrenia in people aged 15 to 17 years (TA213; single technology appraisal). The company collected Paediatric Quality of Life and Enjoyment and Satisfaction Questionnaire (P-QLES-Q). However, for the model they used EQ-5D data from the literature (Briggs et al. 2008). The committee did not comment on the use of EQ-5D itself but was concerned whether adult data were generalisable to young people.

- Bipolar disorder: aripiprazole for treating moderate to severe manic episodes in adolescents with bipolar I disorder (TA292; single technology appraisal). The company submission included the EQ-5D. The final appraisal document does not refer to any concerns about the health-related quality of life data used in the company's economic model.

C.2 Alcohol dependence

One relevant technology appraisal for alcohol dependency was identified: nalmefene for reducing alcohol consumption in people with alcohol dependence (TA325; single technology appraisal). The company collected SF-36 and EQ-5D in the trials. The committee agreed that the EQ-5D data were appropriate for its decision making.

D. Dementia

One relevant technology appraisal for Alzheimer's disease was identified: donepezil, galantamine, rivastigmine and memantine for the treatment of Alzheimer's disease (TA217; multiple technology appraisal). Three cost-effectiveness models were available, 2 from participating companies and 1 from the assessment group. All models included the preference-based measure EQ-5D either directly or from mapping. The committee concluded that the assumptions and inputs about utilities in the assessment group's model were appropriate.

E. Multiple sclerosis

Nine technology appraisals for multiple sclerosis were identified; 6 used the EQ-5D (TA127, TA254, TA303, TA312, TA320 and TA585), whereas a measure was not specified in 3 appraisals (TA527, TA533 and TA616). When EQ-5D was specified, the committee concluded that the approach was reasonable or did not comment specifically on it.

F. HIV

There are no technology appraisals of treatments for HIV, because these have historically not been within the remit of the programme. However, the All Wales Medicines Strategy Group has appraised such products, and some of these included cost–utility analyses using EQ-5D (for example, darunavir).

Summary

Seven technology appraisals for drugs in disease areas where EQ-5D's appropriateness may be a concern were identified. The disease areas included hearing disorders, visual disorders and psychological disorders. The committee preferred EQ-5D when EQ-5D and alternatives were presented together, which might indicate that the evidence submitted to support the use of an alternative measure, if any, may not have met the requirements specified in the methods guide. The committee accepted other preference measures when EQ-5D data were not available. In NICE's technology appraisal on cochlear implants for children and adults with severe to profound deafness, the committee accepted HUI3, which was the only presented preference measure. In an age-related macular degeneration appraisal, the committee accepted the time trade-off direct elicitation out of the 3 presented preference measures. However, it stated that it would have preferred a generic preference-based measure. It is not mentioned in the final appraisal documents for the 2 appraisals whether the committee was presented with evidence of EQ-5D's inappropriateness.

1.4 Conclusions: EQ-5D not appropriate

Based on the currently accepted methods for assessing the psychometric properties of health-related quality of life measures, the literature suggests that EQ-5D-3L performs poorly in conditions involving hearing, where HUI3 has been used in a past appraisal. For dementia, personality disorders and schizophrenia, there is mixed evidence on the appropriateness of EQ-5D-3L.

If there is published evidence showing that EQ-5D-3L may not be appropriate in a health condition, because of either a lack of necessary dimensions or poor performance on psychometric tests, then other methods of deriving utility values might be considered. NICE strongly prefers the use of another generic preference-based measure rather than condition-specific measures. Valuing vignettes or own health have limited roles and are less robust compared with methods using generic standardised descriptive systems.

For hearing conditions, evidence would support using the HUI3 instead of the EQ-5D. This alternative will have a limitation because there is no UK value set or valuation algorithm for HUI3. Additionally, using HUI3 as an alternative measure can impact the comparability of decisions made across the technology appraisal program, given that utility values derived using different measures can be very different.

For mental health conditions (except dementia and learning disabilities), if the evidence of the EQ-5D performance is mixed, then ReQoL could be considered. This measure has the advantages of being developed with input from mental health service users, it includes a physical health question to reduce the focusing effects sometimes associated with condition-specific measures, and the valuation method was comparable with EQ-5D. However, the problem remains that recommending an alternative measure to the EQ-5D could reduce comparability between appraisals. In addition, although comparative psychometric studies on the performance of ReQoL and EQ-5D are planned, they have not been carried out yet.

Previous experience shows that technology appraisal committees preferred EQ-5D if both EQ-5D and alternatives were presented. One committee has accepted HUI3 for hearing impairment when EQ-5D data were not available. A committee has also accepted time trade-off direct valuation, for age-related macular degeneration when EQ-5D data were not available but stated that it would have preferred a generic preference-based measure.

Section 2: EQ-5D not available

The methods guide does not provide guidance on what to do when EQ-5D data are unavailable or available EQ-5D data are insufficient to populate an economic model.

Health-related quality of life data are usually obtained from clinical trials of the intervention or from the literature. However, when evaluating technologies for rarer diseases, it is often the case that such data are sparse or poor quality. Because of this, the modelling of health-related quality of life can be challenging. The highly specialised technologies (HST) programme was developed to evaluate technologies for very rare diseases. In HST evaluations where data on health-related quality of life have been lacking, a number of approaches have been adopted to source health-related quality of life data.

The problem of having insufficient health-related quality of life data to populate an economic model is not unique to rare diseases. For some of the treatments assessed by the technology appraisals programme, there may be certain health states or events in models that occur infrequently in trials or are challenging to measure, leading to a paucity of data to populate economic models. Examples may include few people in trials having either very mild or severe disease, or rare but important adverse events. These situations may occur for relatively common diseases, for which high-quality EQ-5D data have been collected in the trials for most of the health states and events in the economic model.

Although rarity of a disease or particular health state is a key determinant in the decision to use the alternative methods outlined in this report, this report will not attempt to define ‘rarity’. Any changes to the methods guide in this area will highlight that it is the responsibility of the intervention company or sponsor to make the case that it is not feasible to use standard measures to measure health-related quality of life for a particular disease or health state.

2 Decision Support Unit (DSU) reports on options when EQ-5D data not available

2.1.1 Technical support document 11

The DSU’s technical support document on alternatives to EQ-5D for generating health state utility values (TSD11) notes that alternative methods for generating health-state utility values will be considered by NICE in place of EQ-5D when EQ-5D data are unavailable. Unavailability should be established from a systematic literature search. If EQ-5D data are unavailable, but there are data collected using another health-related quality of life instrument that can be mapped to EQ-5D, then this may be an acceptable alternative.

The technical support document summarises the advice in the methods guide on alternative methods to generate health-state utility values, and includes:

- providing supporting arguments and evidence for the choice of alternative method
- basing descriptions of health states being valued on validated patient-reported measures of health-related quality of life
- ensuring methods used for valuation are comparable to those used to value the EQ-5D
- comparing the impact of using alternative methods on the results of the economic evaluation compared with EQ-5D where possible.

In TSD11, the DSU reviewed the alternatives that can be used when EQ-5D data are unavailable. Alternative methods to generate utility values reviewed in the document include other generic preference-based measures, condition-specific preference-based measures, vignettes, or direct valuation of own health (such as time trade-off elicitation technique). These are reviewed in section 1.2 above.

TSD11 concludes that alternatives to EQ-5D most likely to be accepted are preference-based measures derived from validated measures of health-related quality of life, with the value set obtained from the general population, preferably using techniques similar to the protocol used to obtain the UK EQ-5D value set. Any new measures should be validated. Empirical evidence on the preference-based measure should be provided to enable the decision maker to determine the impact of using the measure in terms of the comparability, credibility, reliability and validity of the quality-adjusted life year estimates.

2.1.2 DSU report 'Measuring and valuing health-related quality of life when sufficient EQ-5D data are not available'

In 2020, the DSU published a report that examined alternative methods for measuring and valuing health-related quality of life when sufficient EQ-5D data are not available (Rowen et al. 2020). It highlighted a lack of guidance in situations where NICE's preferred methods for measuring and valuing health-related quality of life could not be followed. NICE's preference for patient-reported EQ-5D utility values can become problematic when: EQ-5D is inappropriate, EQ-5D data are unavailable or when available EQ-5D data are insufficient. Below is a summary of what the DSU considered to be inappropriate, unavailable and insufficient, in relation to EQ-5D data.

- EQ-5D inappropriate: The DSU highlighted that for EQ-5D to be considered inappropriate, empirical evidence is needed to show its poor performance in terms of content validity, construct validity or responsiveness. This is covered in section 1.1 of this report.

- EQ-5D data are unavailable: If EQ-5D data are unavailable from the clinical trials, clear evidence is needed to show why it was not possible to collect self-reported EQ-5D data. The DSU note that not having self-reported EQ-5D data available is possible if a patient population and/or health states required in the economic model prohibit its use. However, poor planning or failure to include EQ-5D in clinical studies where EQ-5D is appropriate is unjustifiable.

- EQ-5D data are insufficient: The DSU also highlighted that there are situations where EQ-5D is considered appropriate for use but available EQ-5D data are insufficient to generate the health states required for an economic evaluation. This could be because the population with the condition is particularly small or a given health state is rarely observed, for example, very uncommon adverse events.

The report aimed to identify and assess the appropriateness of alternatives to NICE's recommended methods in the situations outlined above where measuring, valuing or sourcing patient-reported EQ-5D is problematic. The authors note the methods used in previous technology appraisals and HST evaluations, and outline the strengths, limitations and appropriateness of these methods in each of these situations.

The DSU notes that there may be situations where there is EQ-5D evidence available, but this is based on poor quality data with small sample sizes. It recommends that wherever possible EQ-5D evidence should be used, and other evidence used only where necessary. It highlights that sensitivity analyses can be used to explore, including non-EQ-5D evidence where EQ-5D evidence is available but is of poor quality.

Alternative methods used in technology appraisals and HST evaluations when EQ-5D data unavailable or insufficient

The report identified 2 different methods used in previous technology appraisals and HST evaluations to obtain health-related quality of life estimates where sufficient EQ-5D data were not available. The most commonly used method was the use of vignettes. The report describes vignettes as 'bespoke descriptions of impaired health states'. What was included, the format in which it was presented, and the evidence used to inform vignette descriptions varied widely. The authors explain that utility values were derived from vignettes using a variety of techniques, including using clinicians, carers or members of the public as a proxy for a patient with the condition being considered. The health states were valued by the proxy completion of the EQ-5D, the proxy completion of preference elicitation techniques like time trade-off, or through the Delphi method using a group of clinical experts to reach a consensus on plausible utility values.

Using utility values from other conditions as a proxy was the second method identified in the review of previous technology appraisal and HST guidance. The use of proxy condition utility values is described in the report as 'the use of utility values for one condition to be used as a proxy to represent utility values for another condition'. For example, if utility values are available for another condition that is deemed to have a similar impact on health-related quality of life, then those utility values were sometimes assumed representative of the condition of interest.

Best practice recommendations – when EQ-5D unavailable or insufficient
When EQ-5D data are unavailable from the clinical trials, in line with recommendations from the DSU technical support document on alternatives to EQ-5D for generating health state utility values, the DSU recommended that the next best alternatives are: undertaking a search of the literature; generating EQ-5D estimates by mapping from other sources of health-related quality of life data or conducting a study to collect data.

The authors summarise the typical reasons why EQ-5D data could be insufficient, including small population, and rare events or health states that are unlikely to be observed in a clinical study. They recommend that evidence should be provided showing that it was not possible to: source sufficient EQ-5D estimates from the literature, estimate EQ-5D utility values from using mapping, or directly administer EQ-5D to patients. If it can be shown that EQ-5D data are insufficient, the use of vignettes or proxy utility values are recommended.

Vignettes
The vignette methodology can be used to generate bespoke health-state utility values for economic evaluations. This approach involves constructing a vignette or scenario to describe each of the frequently occurring states associated with a condition and its treatment for respondents to value. They can incorporate a range of information about the impact of the condition and its treatment. This method involves either members of the public, clinicians or carers, valuing descriptions of impaired health states (vignettes). Vignettes can be valued using preference-based elicitation methods, such as time trade-off or standard gamble, or by methods to obtain a consensus view from clinicians (for example, the Delphi method).

In its review of TA and HST evaluations, the DSU found that vignettes varied in content and format. To improve the quality and limit variation between utility values generated by vignette studies, the DSU outlines best practice recommendations for the development of vignettes. In summary, when vignettes are used, high-quality, appropriate and reliable evidence is needed to inform their development and should be developed to meet the requirements of the economic model structure. Refinement and validation should be undertaken using input from clinical experts and/or patients to ensure that the vignettes clearly and accurately describe the disease state they intend to represent in the model. The process of vignette development should be fully described and transparent.

Best practice recommendations for developing vignette studies from DSU report, figure 3

1. **Obtain high-quality appropriate, reliable and informative evidence to inform vignette development. This could consist of, and be strengthened by, multiple types of evidence:**

   - Published literature, for example, reviews or original studies including qualitative studies around the health-related quality of life of patients with the condition.
   - Qualitative studies (for example, interviews or focus groups) with patients, and if relevant carers.
   - Qualitative studies (for example, interviews) with clinical experts.
   - Qualitative analysis of social media data (for example, online patient discussion forums) though care should be taken with interpretation and representativeness because patients may not be representative and formal diagnosis is not ensured.
   - Quantitative data (for example, patient-reported outcome measures of health-related quality of life in clinical trials or observational studies).

2. **Vignette development including content and format**

   - The number of vignettes and the required severity or disease state of each of these vignettes should be selected to meet the requirements of the economic model structure for the TA and HST evaluation. Considerations include the requirement that vignettes meaningfully differ, because subtle differences in descriptions may not be captured in the valuation stage, but these differences should not be exaggerated.
   - Vignettes should be presented and formatted to enable easy reading and comprehension, for example, simple language where possible if presented to members of the general public, appropriate font size, use of boldening or underlining to highlight different levels of severity.
   - Vignettes should be presented and formatted to enable the target audience to easily understand the differences between the different vignettes. For example, the aspects of health described in the vignette should always be presented in the same format and order for a given participant. This is important because it can impact on the utility values that are elicited because some participants may provide relative values for the vignettes while considering all vignettes.
   - Vignettes should include descriptions of the generic dimensions of health-related quality of life, for example, using the EQ-5D dimensions and descriptions. This can reduce focusing effects where respondents may focus on the symptoms or treatment effects described rather than considering these in a wider context of health-related quality of life.
   - Vignettes should include all important and relevant aspects of health-related quality of life to ensure accuracy and minimise bias. Important and relevant aspects should be identified using good quality evidence.

- Vignettes should be easy to understand with minimal potential for ambiguity and misinterpretation by the target audience. Clinical experts, for example, may interpret clinical stages differently in terms of their impact on health-related quality of life, so care should be taken to describe the aspects of health-related quality of life rather than clinical stages, because this is the focus of utility values.
- Each vignette should reflect the typical patient experience for the disease state in question, rather than extremes, though some vignettes may present plausible ranges, for example, 5 to 8 events per month.
- Vignette descriptions should provide clarity and certainty where possible and avoid probabilistic statements, to reduce the variability in the interpretations made by the target audience. Where there is a probability of different outcomes, separate vignettes can be valued for the different outcomes and combined using probabilities to generate the state required in the economic model.
- Carefully consider whether to include the disease label and/or the treatment in the health state. Where possible, it is recommended to avoid the condition or treatment because there is a chance that this could lead to biased estimates. If aspects of treatment are unavoidable, for example, mode of administration, these should be clearly explained to target populations who may be unfamiliar with these.
- Ensure wording is not leading or outside of the context of what should be reasonably considered, for example, avoiding descriptive phrases such as ‘devastating’, ‘debilitating’ or ‘difficult to treat’, naming the patient, or issues around burden of illness or disease history unrelated to the current state.

3. Vignette refinement, validation and interpretation

- Input from clinical experts and/or patients through interviews, focus groups or patient involvement meetings should be undertaken to ensure that the vignettes are a clear and accurate description of the health state or adverse event that they are intended to represent. Vignette descriptions before and after this stage should be presented to identify the changes, and the rationale behind the changes should be transparent and explicit.
- Before the main valuation study, it is recommended to ensure that the descriptions can be understood and are clear for the target audience. For example, the general population may need explanations of some aspects such as seizures, and this could be examined using a pilot study.

Methods for deriving utility values from vignettes

In its review of TA and HST evaluations, the DSU found that a range of methods were used to generate utility values from vignettes. The DSU recommend the following methods, in order of preference:

a) General population, clinical experts or patients complete the EQ-5D for each vignette and this is then valued using the relevant value set for EQ-5D, provided EQ-5D is appropriate.

b) Preference elicitation techniques such as time trade-off with a sample of the general population.

c) Preference elicitation techniques such as time trade-off with patients.

d) Utility values elicited directly for each vignette from clinical experts, for example, using Delphi panels or preference elicitation methods including time trade-off.

The report notes that in many of the examples from previous NICE appraisals where vignettes have been used, utility values were derived by clinicians completing EQ-5D questionnaires based on the vignette. Using the EQ-5D more closely aligns to NICE's reference case. However, the DSU concludes that a sample of the general population or patients would be preferable to clinicians completing the EQ-5D. This is because clinicians may incorporate information from their experience, not included in the vignette description when they are making their judgements. In contrast, members of the public are more likely to be unbiased since they have no experience of the health states nor any expectation they are likely to experience these health states. However, the authors note that there may be differences between the vignettes that may be difficult to interpret by people with no prior knowledge of these aspects of health or treatment. Patients on the other hand, have greater understanding of the symptoms and treatment and how these impact on health for people with the condition

Sourcing utility values from other conditions
The DSU recommends that utility values from another condition can be considered an appropriate proxy if it has the same effects on a person's quality of life as the condition of interest. They recommend that a qualitative assessment involving clinicians and patients should be done to assess the similarity in terms of the dimensions (aspects) of quality of life affected and the severity. Any differences between the proxy and condition of interest should be clearly described and acknowledged. The DSU also states that consideration should be given to the availability of proxy condition utility values that meet the NICE reference case using the EQ-5D.

2.1.3 Summary

Overall, the DSU reports recognise that there are situations in which measuring and valuing EQ-5D may be challenging or not possible. It is acknowledged that these situations are more likely to be observed when appraising or evaluating rarer diseases or rare health states.

The reasons why EQ-5D data are unavailable or insufficient are important. The DSU notes that early consideration of the evidence requirements can make it possible to use EQ-5D.

Alternative methods are available and have been used in previous TAs and HST evaluations. However, the DSU report noted that the lack of guidance on appropriate methods means that there has been variation in when and how these alternative methods were applied.

Where insufficient EQ-5D data are available, the DSU notes that the choice between generating utility values from vignettes or utility proxy condition utility values depends on the appropriateness of the proxy condition utility values and the quality of the vignette study.

2.2 Conclusions: EQ-5D not available

The use of EQ-5D completed by patients and scored using general population preferences is the preferred option to generate utility values. When evaluating technologies for rarer diseases or rare health states, it is often the case that the supporting evidence is sparse or of a poor quality. If trial data are not available, a number of approaches may be used to estimate health-state utility values.

If evidence shows that it was not possible to directly administer EQ-5D to patients, the DSU recommended options are to source EQ-5D utility values from the literature, estimate EQ-5D utilities using mapping, or conduct a study to collect EQ-5D data. This is consistent with the current methods guide.

The DSU report suggested that if none of the above options are possible, then vignettes can be considered. The task and finish group considered that it would be useful to signpost people to the DSU’s best practice guidance for developing vignettes.

The DSU outlines the various approaches that can be taken to derive utility values from vignettes. These include EQ-5D being completed by patients, clinicians or a sample of the general population, and then valued using the relevant value set, or using a technique such as time trade-off, or consensus techniques with clinicians. The DSU considered that EQ-5D completed by a sample of patients or the general population to be preferable to clinicians completing EQ-5D.

The task and finish group considered that it should be possible to recruit a large sample of the general population to complete EQ-5D for vignettes and that this may be more straightforward than some of the other options such as time trade-off. The task and finish group reflected that although patients should have a critical role in developing the content of vignettes, the vignettes should in general describe the health-related quality of life effects of the condition clearly enough for the general population to appreciate without needing specific knowledge of the condition.

However, the task and finish group recognised that in some cases there may be clinical aspects of a condition that are not well understood by members of the general population, such as seizures in epilepsy. In these circumstances, it may be more appropriate for people with the condition to complete the EQ-5D. The task and finish group did not consider this to conflict with the preference for health state valuations to be based on general population preferences. This is because the tariff used to translate the EQ-5D scores into utility values is based on public preferences. Patients mapping the vignettes to EQ-5D is similar to patients in a trial self-reporting their quality of life using the EQ-5D. The task and finish group felt it was more appropriate to ask patients than clinicians in these circumstances because clinicians may have a limited understanding of how the condition impacts on the day-to-day life of patients, and it may be easier to recruit a larger sample of patients.

The task and finish group considered that utility values from a proxy condition may have a role to play on some occasions. This is particularly the case if they were derived using reference case methods. Evidence should be provided that the proxy condition has a similar impact on health-related quality of life as the condition of interest, such as clinical expert assessment, or research with patients.

Section 3: Rare diseases

Both section 1 on when the EQ-5D is not appropriate and section 2 on when the EQ-5D is unavailable have conclusions that are relevant to rare diseases, so a summary of the conclusions is provided below.

Table 2 Summary of main conclusions relevant to rare diseases

-	**Main conclusions relevant to rare diseases**
EQ-5D not appropriate	• EQ-5D found to perform well for a wide variety of diseases. Main exception is hearing disorders. • In general EQ-5D is preferred unless companies provide evidence it is not appropriate. • Key limitation of work is that there may not be relevant published literature to assess whether the EQ-5D is appropriate or not in certain disease areas, particularly rare diseases.
EQ-5D not available	• Acknowledges that it may not be possible to collect EQ-5D data for small populations. • Suggests alternative approaches that might be considered, including using utility values from similar conditions and vignette studies.

As noted in the table above, for rare diseases, there may not be sufficient published literature to provide evidence that the EQ-5D does not perform well. However, although there may not be evidence available to show that the EQ-5D performs poorly on psychometric measures for a rare disease, a lack of content validity could be supported, by providing evidence that the EQ-5D lacks specific dimensions of health that are important to patients. However, it is important to maintain the expectation that EQ-5D is used in most circumstances unless there is strong evidence that it is inappropriate.

As part of this project, the DSU was commissioned to produce a report investigating developing methods to provide evidence in situations in which it is felt that the EQ-5D captures some aspects of health-related quality of life, but does not capture all important aspects. The DSU provided an initial report describing how this could be done, but these methods will not be sufficiently developed to include in the methods guide update. If these methods become sufficiently developed, it may be possible to use them to show which aspects of health-related quality of life the EQ-5D is not capturing for a rare disease.

It is possible that if there are insufficient EQ-5D data to assess whether EQ-5D adequately reflects changes in quality of life for a rare disease, then there may be insufficient EQ-5D data to populate the model. In which case, the recommendations from the rare health state work may apply, such as vignettes or utility values from a proxy condition. Although many rare diseases are associated with combinations of symptoms common to more prevalent diseases, there may be some aspects of certain rare disease that would not be well understood by the general public. In these situations, it may be more appropriate for patients with the condition to score the vignette using EQ-5D.

Review of highly specialised technologies (HST) guidance

As part of this work, a review of NICE's 12 published HST evaluations was carried out. Of these, 2 appraisals were in children only, 4 in adults only and 6 in both. The review shows that for these rare diseases, EQ-5D was either collected in the trials (n=2 appraisals, 17%) or available from the literature (n=4 appraisals, 33%) in half of cases. Where EQ-5D data were not collected in trials or available from the literature, other generic measures were used, such as the SF-36 or Health Utilities Index 3 (HUI3, n=3 appraisals, 25%). Other methods were sometimes used to obtain EQ-5D values, such as experts completing EQ-5D questionnaires for vignettes, which were then valued using the relevant value set (n=4 appraisals, 33%) or using mapping algorithms to derive EQ-5D values from another measure (n=1 appraisal, 8%). The evaluation documents do not record any instances of the committee being presented with data on, or having concerns about, the EQ-5D not being appropriate. The committee's concerns were focused on the limitations of the methods used to obtain the EQ-5D values or the model structure not capturing all benefits of a treatment.

Table 3 Review of HST guidance

Disease area	Sources of health-related quality of life data	Committee conclusions
HST1: Eculizumab for atypical haemolytic uraemic syndrome	EQ-5D from trial used in model	• Substantial benefits not captured in model, no specific discussion on EQ-5D appropriateness
HST2: Elosulfase for mucopolysaccharidosis type IVa	EQ-5D-5L from natural history study MOR-001 used in model	• Patient expert: EQ-5D focus is on day-to-day activities, but ability might be similar before and after treatment – critical difference is how patient feels after doing activity • Committee: health-related quality of life not robustly modelled, because of lack of evidence, not EQ-5D
HST3: Ataluren for Duchenne muscular dystrophy	• Model used HUI3 from literature • PedsQL, PODCI and activities of daily living collected in trial	• No specific discussion on appropriateness of HUI3

Disease area	Sources of health-related quality of life data	Committee conclusions
HST4: Migalastat for Fabry disease	• EQ-5D from literature used in model for complications (for example, stroke) • SF-36, Brief Pain Inventory and gastrointestinal symptoms rating scale collected in trial	• No specific discussion on appropriateness of EQ-5D
HST5: Eliglustat for Gaucher disease	• SF-36 from registry mapped to EQ-5D used in model • Utility increment for oral therapy preference derived from vignette study scored by EQ-5D, completed by a sample of general population • Brief Pain Inventory and SF-36 from the trial	• No specific discussion on appropriateness of mapped EQ-5D • Committee: oral therapy would provide health-related quality of life benefit, but questioned the extent of the benefit and the incremental value derived, which seemed high • Committee preferred Evidence Review Group's alternative utility increment
HST6: Asfotase alfa for paediatric-onset hypophosphatasia	• EQ-5D-5L completed by clinical experts for vignettes based on 6MWT test • CHAQ, PODCI and LEFS collected in trial	• Committee understood 6MWT did not capture all symptoms of condition or important domains of EQ-5D but noted that clinicians may have taken these into account when scoring vignettes • Committee: EQ-5D values in company's model reasonable estimates for 6MWT health states
HST7: Strimvelis for ADD-SCID	• EQ-5D from literature • PedsQL, Lansky performance in trial	• No specific discussion on appropriateness of EQ-5D
HST8: Burosumab for X-linked hypophosphataemia	• Vignettes scored by EQ-5D-5L completed by 6 experts • POSNA-PODCI in trial	• Committee noted vignettes scored by experts and not patients and not from trial – limitations of the data • EQ-5D-5L utility values uncertain but suitable for decision making
HST9: Inotersen for hereditary transthyretin amyloidosis	• EQ-5D from literature or registry • Norfolk QoL-DN and SF-36 collected in trial	• No discussion on appropriateness of EQ-5D but concern non-UK value set used for EQ-5D

Disease area	Sources of health-related quality of life data	Committee conclusions
HST10: Patisiran for hereditary transthyretin amyloidosis	• EQ-5D from the trial • Norfolk QoL-DN collected in trial	• Company: not all aspects of autonomic neuropathy captured by EQ-5D, added additional disutility • Evidence Review Group: EQ-5D routinely used in functional bowel disease and gastrointestinal conditions and has been used to measure autonomic neuropathy • Committee: EQ-5D appropriate but might not fully capture the impact of autonomic neuropathy
HST11: Voretigene neparvovec for inherited retinal dystrophies caused by RPE65 gene mutations	• 6 clinicians completed HUI3 and EQ-5D for each health state in model (vignette) • Company used HUI3 because it has a vision component whereas EQ-5D is believed to have poor convergent validity in visual disorders	• Committee: HUI3 values lacked face validity, acknowledged rationale for using HUI3 (includes sensory domain) but considered EQ-5D more appropriate because of potential focus on vision by clinicians on scoring the vignettes • Evidence Review Group preferred utilities from a general public time trade-off study based on valuing states defined by visual function questionnaire • Committee considered neither source sufficiently robust, concluded it would consider both
HST12: cerliponase alfa for neuronal ceroid lipofuscinosis type 2	• EQ-5D in trial but not used in model • For model, clinical experts completed EQ-5D-5L for vignettes • PedsQL Parent report for Toddlers, PedsQL-FIM, CLN2-based QoL collected in trial	• Committee concerned about robustness of the vignettes – they contained additional disease-specific elements (for example, frequency of seizures) • EQ-5D-5L scores moved in the same direction as PedsQL scores • Committee concluded that it would consider analyses based on EQ-5D values from the vignette study

Abbreviations: 6MWT, 6-minute walk test; CHAQ, Childhood Health Assessment Questionnaire; HST, highly specialised technologies guidance; HUI, Health Utilities

Index; LEFS, Lower Extremity Functional Scale; PedsQL, Paediatric Quality of Life Inventory; PODCI, Paediatric Outcomes Data Collection Instrument; POSNA, Pediatric Orthopaedic Society of North America; QoL-DN, Quality of Life-Diabetic Neuropathy.

Section 4: Case for change and options

A review of the evidence suggests that the EQ-5D works well for most diseases and conditions except for sensory disorders and some mental health conditions. For conditions where there is mixed evidence that EQ-5D performs well, a review of previous NICE technology appraisals in these disease areas shows that it has been possible for committees to make recommendations based on EQ-5D.

The evidence would support specifying in the methods guide that the Health Utilities Index 3 (HUI3) is used instead of the EQ-5D for hearing disorders. To a lesser extent, the ReQoL could be specified for some mental health conditions (except dementia and learning disabilities), although the evidence is more mixed here, and there has not yet been a study conducted comparing the psychometric properties of the ReQoL and the EQ-5D.

Specifying alternative measures for these 2 disease areas could decrease comparability between appraisals, but may increase accuracy if the EQ-5D does not capture relevant aspects of quality of life. However, the advantage of retaining the current wording of the methods guide is that it allows the case to be made for using alternatives to the EQ-5D in these disease areas and in other disease areas, should evidence emerge that the EQ-5D does not perform well.

There is a case for providing more guidance about which alternative measures of health-related quality of life are preferred when it can be shown that the EQ-5D is not appropriate. This could help to increase the likelihood that measures closer to the reference case (for example, other generic preference-based measures) are used in preference to measures departing further from the reference case (for example, vignettes).

The Decision Support Unit (DSU) is developing novel methods to provide evidence in situations when the EQ-5D captures many aspects of health-related quality of life, but does not capture all important aspects of the condition. The DSU provided an initial report describing how this could be done, but these methods are not sufficiently developed to include in this update to the methods guide.

The current methods guide states that companies should present evidence that EQ-5D is inappropriate to justify the use of alternative measures. It is proposed that this is maintained. A potential concern for rarer diseases is that there may be insufficient EQ-5D data to assess whether EQ-5D adequately reflects changes in quality of life. Although there may not be evidence available to show that the EQ-5D performs poorly on psychometric measures for a rare disease, other evidence could be presented and considered by the committee. However, it is important to maintain the expectation that EQ-5D is used in most circumstances unless there is strong evidence that it is inappropriate. There is a case for providing clearer guidance on methods for measuring health-related quality of life in small populations.

There is currently no guidance from NICE on what to do if EQ-5D is not available from the clinical trials or the literature, and if it is not possible to map from another measure to EQ-5D. This can be a problem in all appraisals where health states or events are rarely observed but is more commonly a feature of appraisals for rare diseases.

Case studies from published technology appraisal and highly specialised technology evaluations show that vignettes are often used, but the methods of creating them and the approaches used to derive utility values from them vary markedly (see table 2 and appendix 1). As such, there is a case for providing more guidance about the preferred approach to measuring and valuing quality of life in these situations.

4.1 Options

1. Do not make any changes to the methods guide in this area.

2. **Preferred option:** Add figure 1 below to methods guide. Adding hierarchy about preferred and acceptable alternative methods could lead to improved methodological consistency. This figure:

    - Draws together the different situations in which EQ-5D is either not available or not appropriate.

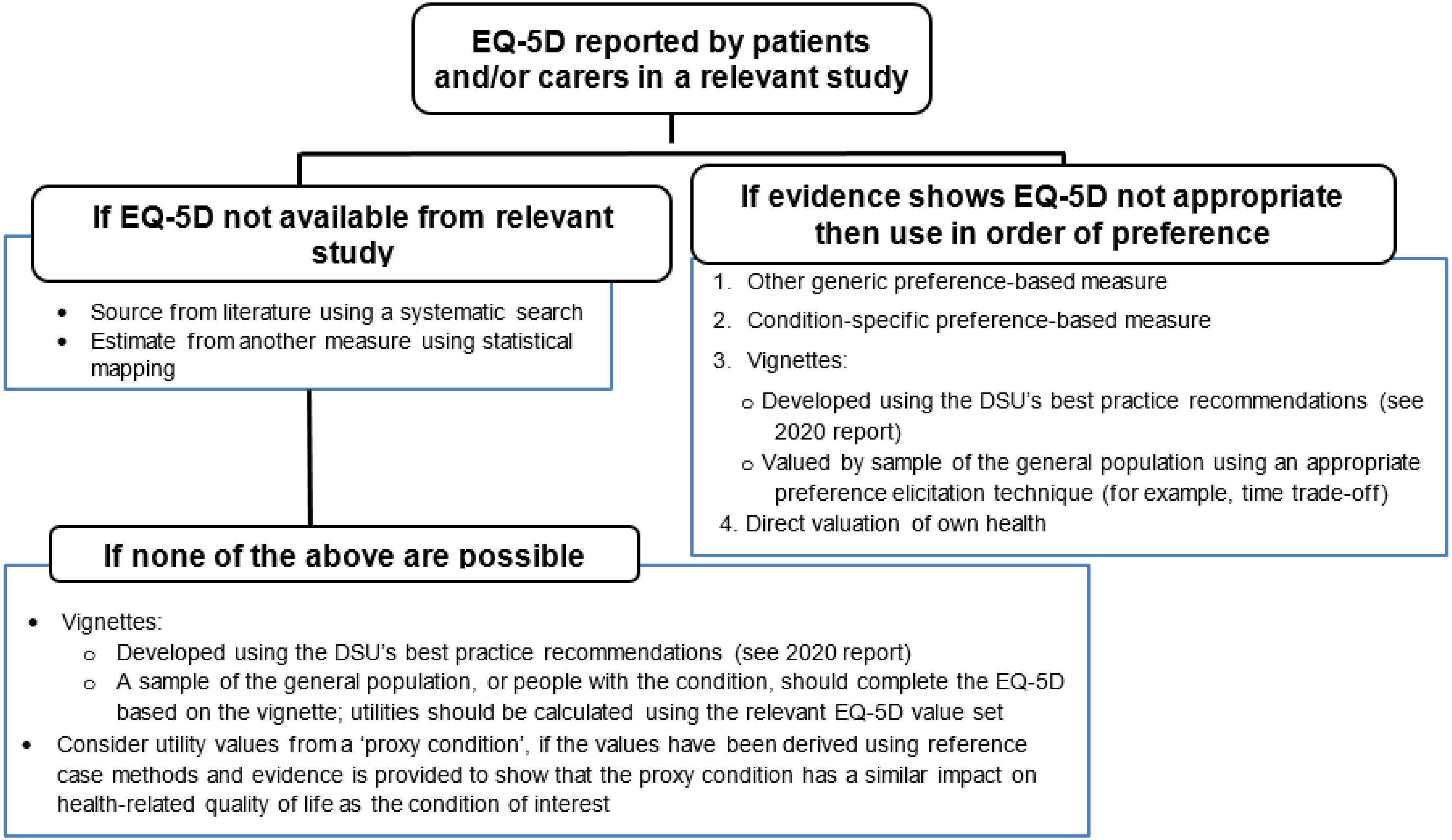
    - Restates some information that is already in the methods guide (that if EQ-5D is not available in a relevant study, source from a systematic review of the literature, or estimate using statistical mapping).

    - For situations in which the EQ-5D is not appropriate, adds more detail than is currently in the methods guide, based on the DSU technical support document on alternatives to EQ-5D for generating health state utility values (TSD11). Importantly, the figure clearly states the options in order of preference, and aligns the guidance on vignettes with that suggested for when EQ-5D is not available.

    - For situations in which the EQ-5D is not available, new guidance on using vignettes and utility values from proxy conditions.

3. Specify the disease area or conditions where EQ-5D may not be appropriate and specify an alternative generic preference base measure that should be used. The strongest evidence is for hearing and HUI3 followed by mental health conditions.

Figure 1 Hierarchy of preferred health-related quality of life methods



EQ-5D reported by patients and/or carers in a relevant study

If EQ-5D not available from relevant study

- Source from literature using a systematic search
- Estimate from another measure using statistical mapping

If evidence shows EQ-5D not appropriate then use in order of preference

1. Other generic preference-based measure
2. Condition-specific preference-based measure
3. Vignettes:
   - Developed using the DSU's best practice recommendations (see 2020 report)
   - Valued by sample of the general population using an appropriate preference elicitation technique (for example, time trade-off)
4. Direct valuation of own health

If none of the above are possible

- Vignettes:
  - Developed using the DSU's best practice recommendations (see 2020 report)
  - A sample of the general population, or people with the condition, should complete the EQ-5D based on the vignette; utilities should be calculated using the relevant EQ-5D value set
- Consider utility values from a 'proxy condition', if the values have been derived using reference case methods and evidence is provided to show that the proxy condition has a similar impact on health-related quality of life as the condition of interest

Section 5: Equality considerations

There is evidence that the EQ-5D may not appropriately capture changes in health-related quality of life for people with hearing impairments. People with a hearing impairment would be considered to have a disability under equality legislation.

Although the evidence about the potential unsuitability of EQ-5D for other conditions is less clear-cut, some of these conditions could also be considered to be disabilities.

The proposed wording of the methods guide states that the EQ-5D is the preferred measure of quality of life, but it also allows companies to present evidence that the EQ-5D is not appropriate for a condition. As such, the EQ-5D is not mandated in every circumstance. Appraisal committees must consider any evidence presented about the appropriateness of the EQ-5D for a condition and any equality implications when deciding which measures of health-related quality of life should be used for decision making.

Authors

Caroline Bregman, Dalia Dawoud, Verena Wolfram and Ross Dent on behalf of the Health-related quality of life task and finish group

References

Brown GC, Sharma S, Brown MM et al. (2000) Utility values and age-related macular degeneration. Archives of Ophthalmology 118(1): 47–51

Brazier JE, Kang K, Carlton J et al. General population survey to obtain values for age-related macular degeneration visual impairment states using lenses – Preliminary Report. 2006. Appendix 1. Lucentis (ranibizumab) therapy for the treatment of wet age-related macular degeneration. Evidence submitted to the National Institute for Clinical Excellence. Novartis Pharmaceuticals UK Ltd. August 2006 – published as Czoski-Murray C, Carlton J, Brazier J et al. (2009) Valuing condition-specific health states using simulation contact lenses. Value Health 12(5): 793–9

Brazier JE and Rowen (2011) NICE DSU technical support document 11: Alternatives to EQ-5D for generating health state utility values. Report by the Decision Support Unit

Brazier JE, Rowen D, Lloyd A et al. (2019) Future directions in valuing benefits for estimating QALYs: is time up for the EQ-5D? Value Health 22(1): 62–68

Claxton K, Longo R, Longworth L et al. (2009) The value of innovation. Report by the Decision Support Unit [accessed 2 December 2019]

Davis S, Wailoo A (2013) A review of the psychometric performance of the EQ-5D in people with urinary incontinence. Health and Quality of Life Outcomes 11:20

Finch AP, Brazier JE, Mukuria C (2018) What is the evidence for the performance of generic preference-based measures? A systematic overview of reviews. European Journal of Health Economics 19(4): 557–570

Finch AP, Brazier JE, Mukuria C (2019) Selecting Bolt-On Dimensions for the EQ-5D: Examining Their Contribution to Health-Related Quality of Life. Value Health 22(1): 50–61

Kennedy I (2009) Appraising the value of innovation and other benefits a short study for NICE [accessed 2 December 2019]

Payakachat N, Ali M, Tilford M (2015) Can The EQ-5D Detect Meaningful Change? A Systematic Review. Pharmacoeconomics 33(11): 1,137–54

Powell P, Carlton J, Woods H, Mazzone P (2019) Patient-reported outcome measures of quality of life in Duchenne muscular dystrophy (DMD): a systematic review of content and structural validity using COSMIN. Discussion Paper Series, ScHARR, University of Sheffield, UK

Tosh J, Brazier JE, Evans P et al. (2010; draft report, NICE quality of life project) A review of generic preference-based measures of health-related quality of life in visual disorders

Tosh J, Brazier JE, Evans P et al. (2012) A review of generic preference-based measures of health-related quality of life in visual disorders. Value Health 15(1): 118–127

Trenaman L, Boonen A, Guillemin F et al. (2017) OMERACT Quality-adjusted life-years (QALY) working group: Do current QALY measures capture what matters to patients? Journal of Rheumatology 44(12): 1,899–1,903

Wailoo A, Davis S, Tosh J (2010) The incorporation of health benefits in cost utility analysis using the EQ-5D. Report by the Decision Support Unit

Yang Y, Longworth L, Brazier JE (2010; draft report, NICE quality of life project) A review of generic measures of health-related quality of life in hearing disorders

Yang Y, Longworth L, Brazier JE (2013) An assessment of validity and responsiveness of generic measures of health-related quality of life in hearing impairment. Quality of Life Research 22(10): 2,813–2,828

Appendix 1 Summary of published technology appraisals in conditions identified as EQ-5D potentially unsuitable

Disease area	Technology appraisal	Technology	EQ-5D?	Other quality of life measures included	Committee conclusion	Discussion of EQ-5D appropriateness in final appraisal document
Hearing	TA566	Cochlear implant	No	HUI3	Accepted HUI3	Not discussed
Age-related macular degeneration	TA155	Ranibizumab Pegaptanib	No	Time trade-off direct elicitation Visual function questionnaire HUI3	Elicitation by time trade-off using method by Brazier most plausible	Not discussed but committee would have preferred generic measure such as EQ-5D or HUI3
Age-related macular degeneration	TA294	Aflibercept	Yes	Time trade-off direct elicitation	No specific conclusion about utility values	Not applicable as EQ-5D was accepted
Dementia or Alzheimer's disease	TA217	Donepezil Galantamine Memantine Memantine Rivastigmine	Yes	HSQ-12, Ferm's-D test QoL-AD mapped to EQ-5D	Accepted EQ-5D	Not applicable as EQ-5D was accepted
Schizophrenia and bipolar disorder	TA213	Aripiprazole	Yes	P-QLES-Q included in trial not used in assessment	Accepted EQ-5D	Not applicable as EQ-5D was accepted
Schizophrenia and bipolar disorder	TA292	Aripiprazole	Yes	No	Accepted EQ-5D	Not applicable as EQ-5D was accepted

Disease area	Technology appraisal	Technology	EQ-5D?	Other quality of life measures included	Committee conclusion	Discussion of EQ-5D appropriateness in final appraisal document
Alcohol dependency	TA325	Nalmefene	Yes	SF-36	Accepted EQ-5D	Not applicable as EQ-5D was accepted
Multiple sclerosis	TA127	Natalizumab	Yes	SF-36 and MSQLI	No specific conclusion	Not applicable as EQ-5D was accepted
Multiple sclerosis	TA254	Fingolimod	Yes	PRIMUS–QoL	Accepted EQ-5D	Not applicable as EQ-5D was accepted
Multiple sclerosis	TA312	Alemtuzumab	Yes (5L)	SF-36	Accepted EQ-5D	Not applicable as EQ-5D was accepted
Multiple sclerosis	TA303	Teriflunomide	Yes	SF-36	Accepted EQ-5D	Not applicable as EQ-5D was accepted
Multiple sclerosis	TA320	Dimethyl fumarate	Yes	Global wellbeing Visual analogue scale and SF-36	Accepted EQ-5D	Not applicable as EQ-5D was accepted
Multiple sclerosis	TA527	Beta interferons and glatiramer acetate	Measure not reported	Not reported	Not reported	Not reported
Multiple sclerosis	TA533	Ocrelizumab	Measure not reported	Not reported	Not reported	Not reported
Multiple sclerosis	TA585	Ocrelizumab	Yes	MFIS	Accepted EQ-5D	Not applicable as EQ-5D was accepted

Disease area	Technology appraisal	Technology	EQ-5D?	Other quality of life measures included	Committee conclusion	Discussion of EQ-5D appropriateness in final appraisal document
Multiple sclerosis	TA616	Cladribine	Yes	MSQoL-54	Not reported	Not reported

Abbreviations: HSQ-12, Health Status Questionnaire-12; HUI3, Health Utilities Index 3; MFIS, modified fatigue impact scale; MSQoL-54, multiple sclerosis quality of life-54; P-QLES-Q, Paediatric Quality of Life and Enjoyment and Satisfaction Questionnaire; PRIMUS–QoL, patient-reported outcome indices for multiple sclerosis-quality of life; QoL-AD, Quality of Life in Alzheimer's disease.

Report 2: Carer health-related quality of life

1. Background

The current NICE guide to the methods of technology appraisal states in section 5.1.2 that the perspective on outcomes is ‘all direct health effects, whether for patients, or when relevant, carers’. However, there is no further guidance relating to incorporating health effects for carers.

Decision Support Unit report: Modelling carer health-related quality of life in NICE technology appraisals and highly specialised technologies

The Decision Support Unit (DSU) reviewed 422 pieces of published technology appraisal (TA) and highly specialised technologies (HST) guidance and found 16 had included health effects for carers. These were:

- 8 TAs on multiple sclerosis
- 1 TA of treatments in each of the following: Alzheimer’s disease, atopic dermatitis, juvenile idiopathic arthritis and myelofibrosis
- 4 HSTs: mucopolysaccharidosis type IVa, Duchenne muscular dystrophy, adenosine deaminase deficiency-severe combined immunodeficiency and X-linked hypophosphataemia.

This list suggests that there are no clear trends or rules for when it is relevant to include health effects for carers in appraisals (for example, specific interventions, populations or disease areas). Although all multiple sclerosis appraisals have included health effects of carers, it is not clear what the particular characteristics of multiple sclerosis are that make the inclusion of carer health effects more relevant than many other disease areas.

Six of the TAs and all the HSTs incorporating health effects for carers included children, where there may be a substantial burden for carers or parents. However, there have been in total 31 appraisals that included children, so it is not the case that appraisals with children always include carer health effects.

A greater proportion of HSTs than appraisals included a quantitative assessment of direct health effects for carers (50% versus 3%). This may be related to higher frequency of paediatric populations in HSTs or because of the specific references to carer health-related quality of life in the HST methods and process guide.

Evidence for carer health effects

The standard of evidence of health-related quality of life effects for carers has generally been poor. Many of the appraisals noted that no evidence is available and have ‘borrowed’ estimates from other disease areas to populate models. Most notably, in TA127, disutilities for carers of people with multiple sclerosis were estimated from carer disutilities in the Alzheimer’s disease appraisal. These were then used in further multiple sclerosis appraisals and in the TA on juvenile idiopathic arthritis and HST2. It is unclear to what extent carer quality of life estimates are transferable between disease areas.

Approaches to modelling carer quality of life

The DSU found that several modelling approaches had been taken, such as:

- linking quality of life of carers to patient disease severity
- applying disutility to each health state to represent the carer burden
- applying different disutility for different interventions to represent the carer burden
- applying additional utility increment for the intervention and not the comparator to represent the improvement in carer quality of life
- including family quality-adjusted life year (QALY) loss when a child died.

Including a carer disutility or carer utility decrement when a patient receives a specific treatment assumes that carer quality of life is not linked to the patient's disease status, but to the treatment received. Hence, the carer would have the same quality of life benefit regardless of the size of the benefit that the patient received from treatment. This goes against the way models typically consider patient quality of life, which is linked to disease severity (influenced by treatment) rather than the treatment itself.

Modelling carer quality of life by patient disease status is more consistent with the typical approach to modelling patient quality of life. This approach may also better explain the change in quality of life for the carer: a person caring for a patient with more severe disease may have to spend more time performing caring tasks or worry more about the patient, and so the quality of life impact would be higher. Validating this requires evidence that quality of life of carers varied by the patient's disease severity. Evidence was presented that supported this in the cases where this approach was used (although sometimes from different disease areas). Modelling carer quality of life by patient disease status needs consideration to what happens to carers' quality of life when the patient dies.

Most TAs and HSTs considered the health impact on 1 carer only, but some made the case for a larger number of carers to be considered or even a wider effect on the family. Increasing the number of carers considered can have a substantial effect on the cost-effectiveness estimates.

Wider literature

There is an increasing body of literature advocating that carer health-related quality of life effects should be included in economic evaluations. Prominently, Brouwer (2019) has argued that health impacts on informal carers should be included in evaluations for reasons of efficiency and equity, even for evaluations taking a healthcare (as opposed to societal) perspective. The author argues that omitting health effects for carers is inconsistent with the goal of maximising health from a fixed budget, and risks decisions that reduce health. The author notes that if decision makers are indifferent about who receives and loses health (that is, do not concern themselves with any particular characteristics of individuals), then it is maximising health as a whole that matters.

Brouwer also highlights that there may be equality implications if carer health effects are not considered. For example, early discharge may appear cost effective, if the potential health effects for informal carers are not considered. However, evidence suggests that informal carers are more likely to be female and informal caring is more common in families from certain cultural and ethnic backgrounds, which may have equality implications.

The Brouwer paper comments on the argument that health effects for carers should not be considered because it is unclear what activity may be displaced by funding interventions where health effects for carers have been considered. The author highlights that it is also not known what activity is displaced if only patient health effects are considered, but this has not prevented healthcare systems making decisions based on average opportunity costs.

The paper notes that methods illustrating how to account for displaced activity have been published. Al-Janabi et al. (2016) sets out a framework for including family health spillover effects in economic evaluations. The framework involves adapting the conventional cost-effectiveness decision rule to include 2 multiplier effects to capture spillover effects. Each multiplier is a ratio of total health effects to health effects for patients for: 1) effects generated by funding the intervention, and 2) effects displaced from funding the intervention.

McCabe (2019) raises a note of caution, that the impact of incorporating carer quality of life might vary by socioeconomic status, if carers who can afford more respite carry a smaller burden, or carer effects may vary by household size. The paper highlights that diseases are not distributed uniformly across the socioeconomic spectrum, so changes that affect the probability of a technology being funded could potentially change the distribution of health across the population.

The author goes on to argue that evidence of the spillover burden across the spectrum of diseases and by socioeconomic status is needed to provide decision makers with insight into whether the spillover effects for identified beneficiaries are above or below the average.

On methods for measuring and valuing spillover effects, including health of carers, Prosser (2019) in an introduction to a themed issue of pharmaco-economics, notes that ‘as with any emerging field, there is a need to define methods’. The author goes on to note that valuing spillovers using conventional measures of health-related quality of life requires much additional research. An example of emerging evidence in this area is a May 2020 publication by McLoughlin et al., examining the validity and responsiveness of 5 quality of life measures in carers across 4 conditions. Prosser also notes that another crucial issue is ‘where to draw the line’ in assessing spillover effects, and questions of whose effects to include in evaluations needs additional debate.

Conclusions

Although the wider literature suggests a degree of circumspection is necessary if health effects for carers are to be considered, there is not a strong case to support removing the possibility of including health effects for carers from the methods guide.

The DSU report does provide evidence that there is a case for providing more guidance on what evidence to present to make the case for including health effects for carers and how to model those effects, given the variation in approaches in previous appraisals and the generally poor standard of evidence provided.

The potential benefit of providing further guidance that specifies robust minimum evidence requirements and standardises modelling approaches is that this could decrease the uncertainty in appraisals where the committee concludes it is appropriate to account for carer health effects. Currently, committees that are persuaded that a disease has an impact on carer health-related quality of life must often choose between poor quality, highly uncertain quantitative evidence, or take account of it qualitatively in decision making.

Providing more guidance needs to be weighed against the possibility that doing so encourages more submissions to include health effects for carers. It has been suggested that this may lead to activity being displaced for which the effects on carer quality of life have not been accounted.

Displacement and options for mitigating the impact

At the individual appraisal level, this can be mitigated by accurately capturing the health effects for carers for the intervention and all comparators.

The more difficult case is at the wider NHS level. The usual cost-effectiveness threshold range represents the rate at which, on average 1 additional QALY is produced in the NHS (£20–30,000). Therefore, when funding new interventions, if it produces additional QALYs for less than this, more health is generated than displaced. The concern is that carer effects may not have been considered when this range was defined. If carer effects are considered more widely in appraisals in the future, interventions may appear cost effective because they are being compared with a benchmark that does not take into account the carer effects associated with the portfolio of treatments currently funded by the NHS. In some cases, this may lead to more health being displaced by funding new interventions where these effects have been considered.

One possibility would be to propose that carer effects are captured widely in future appraisals but the benchmark (cost-effectiveness threshold) is also updated to take into account carer effects of currently funded treatments. That would likely result in a lower cost-effectiveness threshold range. However, it is not within the remit of the methods review to recommend changes to the cost-effectiveness threshold range.

Given this fixed threshold, an alternative approach could be to assume that on average, most conditions have some impact on carer health-related quality of life and there is an average carer benefit associated with interventions currently funded by the NHS, but that this is not captured in the threshold range. Therefore, only benefits in excess of this average should be captured. This could be done in 2 ways, both of which have drawbacks:

1. Quantitatively define the average carer benefit from current NHS activity, and subtract this from the modelled carer benefits.

2. Allow committees to determine if a condition is associated with a particularly high burden on carer health-related quality of life. In these situations, it is more likely that an intervention claiming to improve carer health-related quality of life will deliver benefits above the average expected in the NHS.

The main drawback of option 1 is that there is currently no estimate of average carer benefit deriving from the portfolio of usual NHS treatments and estimating it would be extremely challenging.

Option 2 would be simpler to implement and is perhaps closer to current committee decision making when submissions propose capturing carer benefits. However, a drawback of this approach is that in situations in which the committee concludes it is appropriate to include carer health-related quality of life, both the ‘excess’ and ‘expected’ carer benefits would be captured, if no adjustment is made.

To enable the latter approach, the first item in the proposed evidence standards asks companies to provide evidence to show that a condition is associated with a substantial impact on carer quality of life. This could be used by committees to assess whether the intervention has the potential to provide significant health-related quality-of-life benefits for carers.

Although it is possible that there will be more claims that a condition has a significant impact on carer health-related quality of life, this is only the first step. The guidance will also set out the standard of evidence needed to quantify the scale of the carer effects. This represents a significantly higher bar than currently and may discourage submissions where the case is weak and the evidence is poor. Agreed minimum evidence requirements may give committees more confidence when concluding whether the case for including carer effects is strong enough and to reject, or appropriately account for, poor quality evidence.

2. Proposed minimum evidence standards

As part of this project, a sample of key papers from the literature was examined. However, only in a minority of cases did they shed light on any of the outstanding methods issues in this area of relevance to technology appraisals. It is clear from this exercise that providing guidance on when and how to include carer health effects in appraisals will involve a degree of normative judgement.

Based on this literature and experience of previous appraisals, the NICE members of the task and finish group has drafted some potential minimum evidence standards. The purpose of these is to serve as a starting point for discussions with a wide range of stakeholders. This would ideally take the form of a workshop with participants from patient and carer organisations, academia, the life sciences industry and members of appraisal committees.

Table 1 Proposed minimum evidence standards

Item	Explanation
1. How to show that it is appropriate to include carer quality of life effects? Supporting evidence could be: • from the literature • based on clinical rationale • linked to patient EQ-5D	• There may be evidence from the literature on whether a condition is associated with a caring burden that has an impact on health-related quality of life. • A case for there being an impact on health-related quality of life of carers could be made qualitatively based on the activities carers are required to do and how this might affect their health-related quality of life. • It might be possible to link carer burden to the patients' health-related quality of life scores. For example, if patients report being unable to wash and dress themselves on EQ-5D, then this might support the case that they have informal carers.
2. Carer quality of life should be measured using the EQ-5D and collected in trials of the intervention	Using the EQ-5D has the following advantages: • can be combined with patients' quality of life • ensures comparability between appraisals.
3. Use carer utility values from literature if not collected in the trial	If utility values are not available in the trials, use published carer utility data or data from a disease with a similar severity.
4. Include health effects for primary carer only	One carer has been accepted in most technology appraisals (TAs) including carer health effects and there is likely be more robust evidence for the primary carer.
5. Do not consider family quality of life	This moves further away from direct health effects and the field is too immature to recommend including in analyses.

6. Costs relating to informal care	Healthcare costs only should be included. For example, if the impact of caring on health leads to higher NHS resource use by the carer. For example, for musculoskeletal services.
7. Use the same weighting regardless of the relationship between the carer and the patient	There should be no additional weight given to the burden of caring for someone depending on whether it is a parent caring for a child or someone caring for a spouse.
8. The effect of bereavement should not be modelled	The methods in this area are not well developed. Utility may increase on the death of the patient (as the carer is freed from caring) or decrease (for example, extreme grief with loss of a child). The impact of bereavement is unpredictable and should therefore not be included.

Supporting rationale

Item 1: When is it appropriate to include carer health-related quality of life

Background

The DSU report found there was no clear trend in the disease areas in which health effects for carers had been accepted by committees. Although more appraisals with paediatric populations incorporated carer effects, there were many more appraisals in these populations that did not consider them.

The case could be made that a wide range of diseases and conditions have effects of varying sizes on carer health-related quality of life. However, because of concerns about displacement of current activities for which the carer quality of life effects have not been included (see section 1), it is proposed that only in cases where the effects on carers are likely to substantial should they be considered.

Discussion

Whether an appraisal should include health impacts on carers should be based on clear evidence that the condition impacts on patients' ability to be independent and that the condition directly impacts on carer's health-related quality of life. This may be provided by:

- Evidence from the literature or clinical trials that shows an impact on carer quality of life.
- Evidence from the literature or clinical trials that carers spend a significant amount of time providing care and that this results in a significant impact on carer quality of life.
- Evidence from disease areas that are likely to involve similar levels of caring.
- Clinical and patient expert submissions.
- Evidence from patient-completed health-related quality of life questionnaires that show severe or extreme difficulties in relevant domains (EQ-5D preferred).

The above is not exhaustive. There is a strong preference for quantitative over qualitative evidence. Qualitative evidence may be presented to describe the impact on carer quality of life, and the appraisal committee will consider the appropriateness of including carer quality of life.

Conclusions

Evidence should be provided to show that a condition is associated with a substantial impact on carer's health-related quality of life and so it is therefore appropriate to consider how the intervention affects carers. This will help to mitigate the issue that the carer effects associated with the portfolio of standard NHS treatments is not included in the usual cost-effectiveness threshold.

Item 2: Measures to capture carer health-related quality of life

Background

Many measures have been developed that aim to capture meaningful changes in quality of life from patients' perspective. This contrasts with the limited selection of measures currently developed to capture carer quality of life.

Discussion

The NICE methods guide recommends the EQ-5D as the preferred measure of quality of life in adults and does not specify whether carer utility should be calculated by any other means. It is noted in the literature there are some concerns that the EQ-5D may not be the most appropriate measure for carers. There are bespoke instruments for measuring carer quality of life, such as the CarerQol and the Carer Experience Scale.

Emerging evidence (McLoughlin et al. 2020) in a study comparing 5 quality of life instruments for carers across 4 conditions found the EQ-5D had some validity and may be appropriate to use in health technology evaluations. The benefits of using the EQ-5D to measure carer quality of life is that it can easily be combined with patient quality of life. In addition, it allows for greater comparability across appraisals.

The methods guide does provide guidance on when alternative measures may be acceptable for use when measuring patient quality of life, and similar guidance could be applied to carer quality of life measures. That would involve submissions to NICE clearly outlining why alternative measures for measuring carer quality of life should be used, supported by evidence that the EQ-5D does not capture relevant carer domains.

Conclusion

The EQ-5D is recommended by NICE for measuring quality of life in adults. There is emerging evidence suggesting that it has some validity in measuring changes in carers' health-related quality of life. Using it for measuring changes in carer quality of life would enable carer QALYs to be included with patient QALYs more easily and encourages consistency across appraisals that include carer quality of life. Alternative measures may be considered only with supporting evidence showing that the EQ-5D does not capture the relevant domains for capturing carer quality of life. The change in carer quality of life should reflect the difference between those providing care to patients and their expected utility if they were not providing care. Where possible, attempts should be undertaken to measure how carer quality of life changes by lifetime of a condition. This may include considering health utility by disease stage for example.

Item 3: If no data collected in the trial, use carer quality of life data from literature or other disease areas

Background

Evidence on carer health-related quality of life in the relevant disease area may not be available. Data are often not collected in the clinical trials for the intervention being appraised. A key issue is whether evidence can be derived from the published literature and other disease areas thought to have a similar impact on health and quality of life in the absence of data from the relevant clinical trials.

Discussion

The DSU authors conclude that it is currently unclear to what extent quality of life estimates are transferable between disease areas. However, in some appraisals, for example, TA320, TA303 and TA616 for multiple sclerosis, the committee concluded that it was appropriate to include the carer quality of life estimates. In other examples, the carer quality of life was not explicitly discussed in the guidance document but the company's model was accepted and so it is implicitly assumed that the carer quality of life estimates were considered appropriate (for example, HST2). Therefore, there has been some implicit acceptance of using carer quality of life estimates from other disease areas to populate models.

For patient quality of life, NICE recommends that if not available in the relevant clinical trials, then EQ-5D data can be sourced from the literature or estimated by mapping other quality of life measures or health-related benefits observed in the relevant clinical trial(s) to EQ-5D. When obtained from the literature, the methods of identifying the data should be systematic and transparent, and the justification for choosing a particular data set should be clearly explained. When more than 1 plausible set of EQ-5D data are available, sensitivity analyses should be carried out to show the impact of the alternative utility values. It seems logical that these recommendations could also apply to carer quality of life, that is, if not available in the relevant clinical trials, data on carer quality of life can be sourced from the literature using systematic and transparent methods with the reasons for choosing a particular data set clearly explained.

Conclusion

Data on carer health-related quality of life from the relevant clinical trial(s) or for the relevant disease are often not available. It is currently unclear to what extent quality of life estimates are transferable between disease areas but there has been implicit acceptance of this in some NICE appraisals. If evidence is not available from the relevant clinical trials and it is considered important to include carer quality of life estimates, evidence may be sourced from the wider literature or other disease areas that are thought to have a similar impact on health and quality of life. The methods for identifying the data should be systematic and transparent, and the choice of data should be clearly explained.

Item 4: When including the impact of informal caring on quality of life, how many carers should be included?

Background

Some conditions that limit patients' daily activities need 1 or more carers. However, most TAs have proposed 1 carer. There may be evidence in the literature about the number of carers on average a patient with a particular condition has. Including additional carers in the analysis would likely have the effect of reducing incremental cost-effectiveness ratios. However, the most robust evidence is likely to be for the primary carer – they are more likely to have been asked to complete quality of life data in the trials.

Increasing the number of carers may also lead to complications and discussions about whether the caring burden for each carer is the same, or whether the caring burden of 1 carer is split between multiple carers.

Conclusion

The most robust evidence is likely to be for a primary carer, so this should usually be all that is modelled. However, if a robust case can be made considering the impact on health-related quality of life for more than 1 carer, this should be conducted as an additional sensitivity analysis.

Item 5: Should family effects be included?

Background

It is thought that family effects occur in family members of the patient as a consequence of the health state of a loved one. Examples include the effects of a serious illness of a child on the wellbeing of siblings, who may not be carers.

Discussion

Although these effects may occur, it is difficult to judge how much of an impact these effects have on health in health-related quality of life. In theory, if such an effect did have an impact on health-related quality of life, then this could be captured by the EQ-5D. However, the evidence in this area is likely to be extremely uncertain.

Conclusion

There is currently no strong case for recommending family effects to be included in analyses, because the methods and evidence in this area is very sparse.

Item 6: Healthcare costs associated with informal caring

Background

As well as affecting health-related quality of life, informal caring may be associated with costs that fall on the healthcare sector. For example, GP or outpatient appointments and medication associated with the health impacts of informal caring. Other costs arising from informal care such as travel or not being able to work, do not fall within a healthcare perspective.

Discussion

If an intervention is likely to have a substantial impact on the health-related quality of life of informal carers, then these health effects could be associated with healthcare usage and this should be reflected in models.

Data on healthcare resource use of informal carers could be collected in clinical trials alongside information on the health-related quality of life effects and used to populate economic models. However, it is unlikely that this information is being routinely collected in trials.

Unless the health impacts of caring are substantial, these healthcare resource impacts are likely to be relatively small, in the broader context of the other costs associated with treatment for the disease.

An intervention that improves the quality of life of informal carers is also likely to decrease the associated healthcare costs, and as such would represent an extra benefit and potentially make these technologies appear more cost effective. However, given the likely small magnitude of these costs, this is unlikely to have as significant an impact on cost-effectiveness estimates as including carer health-related quality of life.

Conclusion

If data on healthcare costs associated with informal caring are available, these could be included in economic models. However, these are unlikely to be routinely collected in trials and may only have a small impact on cost-effectiveness estimates.

Item 7: Weighting carer effects by relationship

Background

It has been suggested that impact on carer quality of life may vary according to the carer’s relationship to the patient. For example, it might be argued that caring for a sick child would have more negative consequences on the carer’s health-related quality of life than caring for a spouse or a sibling. On the other hand, there may be differential effects according to whether the caring is for a spouse, a sibling, a parent or a friend. Consideration therefore needs to be given to whether it is appropriate to weight carer effects according to the carer’s relationship with the patient.

Discussion

There appears to be very limited evidence on how the type of relationship affects carer quality of life. Moreover, the impact of caring is highly individual and depends on personal and family circumstances. Intensity and duration of caring are known to be predictors of health effects among carers, and weighting health effects according to type of relationship could be oversimplifying, and involve subjective judgements that may not be grounded in the evidence.

It could be argued that a family member who spends long hours over a sustained period caring for an older relative with advanced dementia may have greater effects than a parent caring for a sick child over a short time period.

The NICE reference case states that an additional QALY should receive the same weight regardless of any other characteristics of the people receiving the health benefit. Therefore, an additional QALY is of equal value regardless of other characteristics of the individuals, such as their socio-demographic characteristics, their age, or their level of health.

Conclusion

Given the sparsity of data in this area, the host of individual and contextual factors involved in determining carer effects, and the current NICE reference case that an additional QALY should receive the same weight regardless of any other characteristics of the people receiving the health benefit, the conclusion of this paper is that carer utility should not be weighted according to relationship type between patient and carer.

Item 8: Should the effects of bereavement be included in models?

Background

Bereavement could have a significant impact on health-related quality of life. The DSU report notes that this has only been considered in 1 appraisal, HST7, where it was presented as a scenario analysis with additional disutility associated with bereavement after the death of a child. This relied on data from Christiansen et al. study of the cost effectiveness of the meningitis B vaccination. In this case, the additional loss of quality of life experienced by the bereaved family and network members was assumed to be equivalent to 9% of the QALYs lost by the death of the person with meningococcal disease.

The DSU stage 3 report identifies:

- a 2012 study (Hornberger et al.) in leukaemia in which utility of the patient and spouse were summed. A utility decrement of 0.60 was applied for the spouse for 1 year on death of the patient.
- Pham et al. (2014) study on end of life interventions in which QALY decrements were applied to family members from experiencing bereavement.

Discussion

Although bereavement may have a significant impact on health-related quality of life, to account for it in economic models would need the following questions to be answered:

- How should the effect of bereavement on quality of life be measured?
- How long should any effect apply for?
- Bereavement effects could potentially apply in all models where there is a mortality risk – that is, it would not just apply to appraisals where there are health-related quality of life effects for informal carers.
- Bereavement effects could apply to multiple individuals. Although there may be evidence on how many carers people with a condition have, the number of people to include bereavement effects for would be a value judgement.

Including such an effect in models would favour treatments that prevent people from dying and potentially make these interventions appear more cost effective. For some of these situations, there are other areas of the methods guide, which account for the desirability of such treatments including:

- The end of life criteria for life-extending treatments at the end of life.
- The non-reference-case discount rate, which may be applied for treatments that substantially restore people to good health, for an extended period.
- The QALY weighting in HST, for treatments that deliver substantial QALY gains. It would not be possible to achieve such gains without extending life.

Conclusion

There are no established methods for including the effects of bereavement in cost-effectiveness analyses. The minimum criteria list should not include an item relating to bereavement.

3. Conclusions and next steps

There is a case for providing more guidance around when and how to include carer health effects in appraisals. A set of minimum evidence standards would help to increase the standard of evidence and decrease the uncertainties faced by committees wishing to consider health effects for carers.

Decisions reached around minimum evidence standards involve normative judgements. These need to be explored and discussed with a wide range of stakeholders including representatives from academia, patient groups and the industry.

Some of the technical aspects of including carer quality of life in models may need further research from the DSU or other academic groups.

Authors

Zoe Charles, Ellie Donegan, Alan Moore and Ross Dent on behalf of the Health-related quality of life task and finish group

References

Al-Janabi et al. (2016) A framework for including family health spillovers in economic evaluation. Medical Decision Making 36: 176–186

Brouwer (2019) The inclusion of spillover effects in economic evaluations: not an optional extra. PharmacoEconomics 37: 451–456

McCabe (2019) Expanding the scope of costs and benefits for economic evaluations in health: some words of caution. PharmacoEconomics 37: 457–460

McLoughlin et al. (2020) Validity and responsiveness of preference-based quality-of-life measures in informal carers: a comparison of 5 measures across 4 conditions value in health (in press)

Pennington and Wong (2019) Modelling carer health-related quality of life in NICE technology appraisals and highly specialised technologies. Report by the Decision Support Unit

Prosser and Wittenberg (2019) Advances in methods and novel applications for measuring family spillover effects of illness. PharmacoEconomics 37: 447–450

Report 3: Adjusting health state utility values over time

1. Introduction

The NICE guide to the methods of technology appraisal notes that ‘In some circumstances adjustments to utility values, for example, for age or comorbidities, may be needed’. Many technology appraisals adjust utility values over time. However, there is more than 1 method for doing this, and NICE has not provided guidance on its preferred approach. In addition, age is a protected characteristic, so consideration must be given to whether adjusting utility values unfairly discriminates against people from different age groups. This report will:

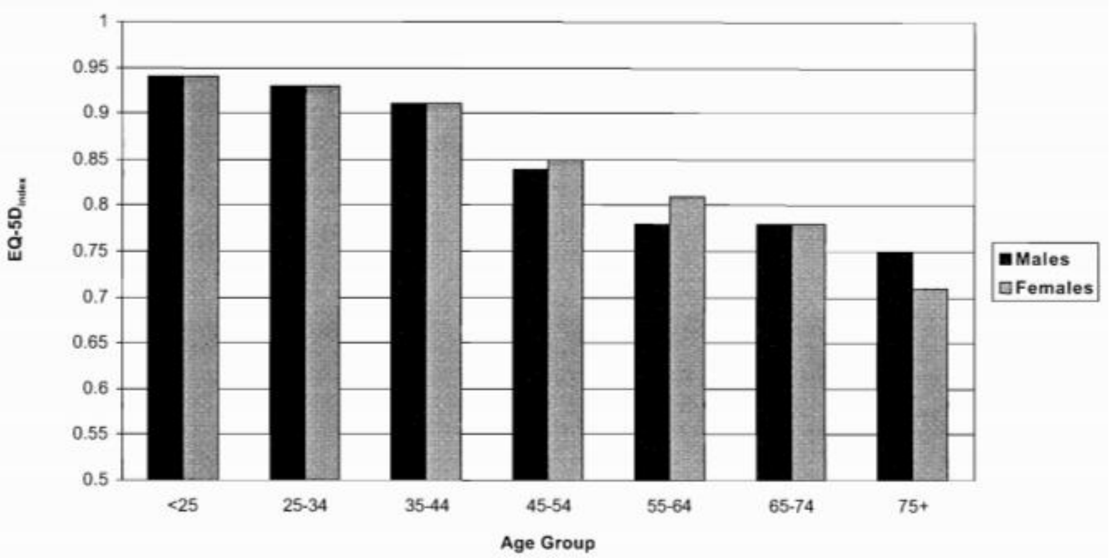
- Outline the rationale for adjusting utility values over time.
- Discuss the alternative approaches for adjusting utility values.
- Review a subset of published technology appraisal (TA) guidance to understand the approach taken to adjusting utility values and analyse the impact that doing so has on cost-effectiveness results.
- Consider the equality implications.
- Conclude with proposals for any changes to the methods guide.

2. Why adjust health state utilities over time?

Need to extrapolate over long time horizons

Many cost-effectiveness models assess quality-adjusted life years (QALYs) accrued over a lifetime horizon to capture the full benefits of treatments. However, data on health-related quality of life are usually based on observations during the trial period, or perhaps 1 value from the literature. These values must be extrapolated over the model time horizon. The simplest assumption to make is that the utility values remain constant over time. However, a decline in health-related quality of life over time has been observed as people get older (figure 1). This may be because as they get older, they accumulate comorbidities and there may be a natural decline in mental and physical functions with age.

Figure 1 UK population average EQ-5D values by age and sex (Kind et al. 1999)



Distinction between health-related quality of life and self-assessed wellbeing

It is important to note that health-related quality of life as assessed by generic preference-based measures such as the EQ-5D ask people to describe their health in terms of mobility, self-care, usual activities, pain or discomfort, and anxiety or depression. A deterioration in some of those domains would lead to lower quality of life values when valued using the general population tariff. However, older individuals themselves may have different preferences to the general population. Self-assessed wellbeing has been reported to follow a ‘U-shaped’ curve (see for example, Blanchflower, 2008), rising in older years (for example, 65 and over). However, this is measuring something different from health-related quality of life. NICE prefers EQ-5D to be valued by a representative sample of the general population both to make decisions comparable and to account for the fact that individuals in poorer health may become accustomed to their symptoms and have lower expectations of good health.

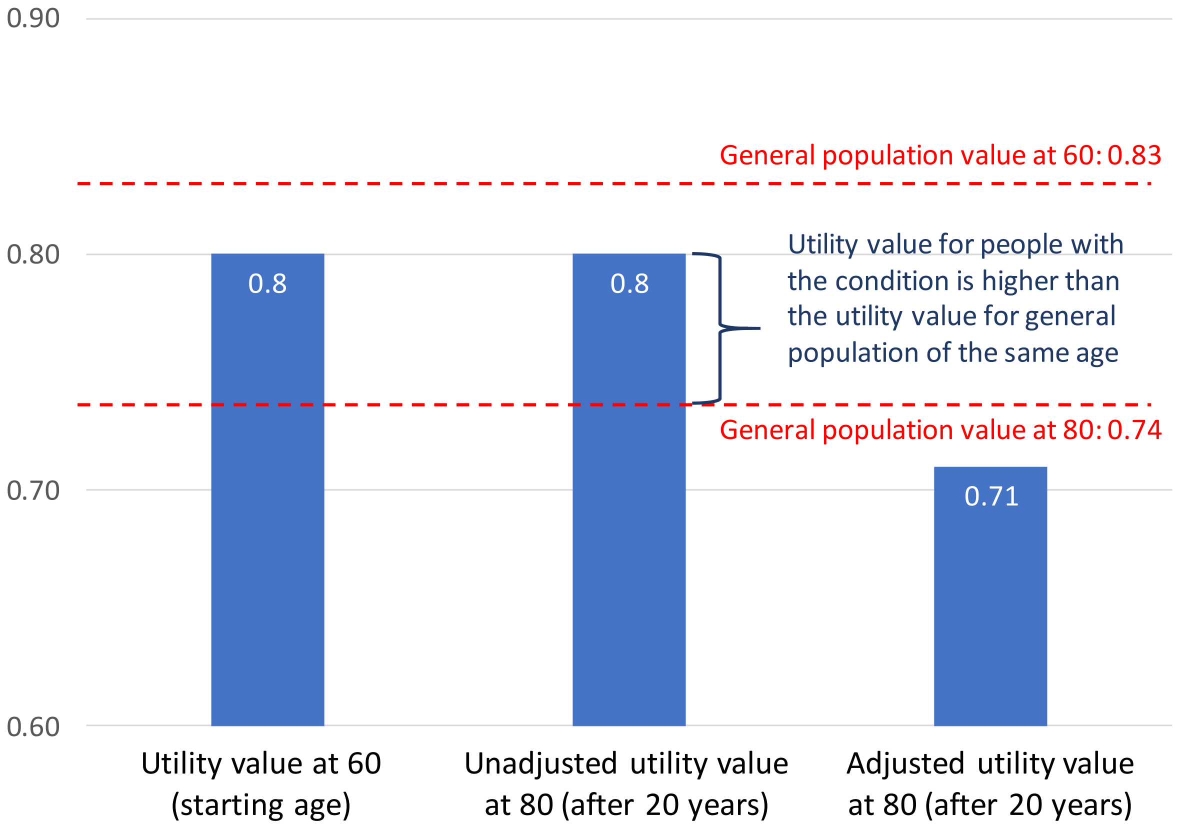
Rationale for adjusting utility values over time in cost-effectiveness analyses

If the utility data collected in the trials or from the literature are extrapolated over the remainder of the model time horizon (for example, 30 years) without adjustment, values can end up higher than what would be expected for the general population at a given age.

For example, for a particular disease, the starting age in the model may be 60 years old, to reflect the average age at diagnosis, and the initial health state may have a utility value of 0.8, based on the average at the start of the trial. This is slightly lower than the general population norm for a 60-year-old man (0.83) perhaps reflecting the impact of the disease on health-related quality of life.

Over time, many people may progress through the model to other health states, but after, for example, 20 years, a proportion of modelled patients may remain in this initial health state. Without adjustment, the utility value will still be 0.8, whereas by now those patients are 80 years old, and the population norm for 80-year-old men is 0.74. Adjusting the utility values relative to those of the general population as both progressively age (see figure below for example) means that the utility value for people with a certain disease is not higher than people of the same age without the disease.

Figure 2 Example illustrating the rationale for adjusting utility values over time in cost-effectiveness analyses



The reasons for adjusting utility values for ageing are analogous to the reasons that many cost-effectiveness models adjust survival rates so that they do not exceed general population mortality rates, which would be equally implausible.

DSU's technical support document on the use of health state utility values in decision models (TSD12) reports that at a workshop used to inform the 2008 update of the methods guide, the consensus was that 'longitudinal data from the patient group of interest would preferably be used to derive these HSUVs [health state utility values] and it was generally agreed that adjusting for the effects of age and gender should be conducted as an absolute minimum'.

In addition, a report titled Identification, review, and use of health state utilities in cost-effectiveness models: an ISPOR good practices for outcomes research task force report (Brazier et al. 2019) recommends that utilities should be adjusted to account for the natural decline with age.

Conclusions

Overall, the aim of adjusting utility values over time is to ensure that cost-effectiveness analyses reflect the difference a technology can make to someone's health-related quality of life, anchored by the average health-related quality of life that an individual in a cohort can expect to experience over the same time period. It is acknowledged that some people maintain very good health as they get older. Some of these individuals will be included in the sample used to derive the average population values. In addition, the cost-effectiveness analysis is modelling, and the appraisal committee is making decisions about, an average patient.

3. Methods for utility values over time

There is a lack of consensus about the most appropriate methods to use to adjust of utility values. In general, all methods include 2 steps:

1. Selecting the age-related distribution of utility values for the general population from published literature.

2. Using this distribution to adjust the condition-specific utility values over time, using either the additive or multiplicative method.

Age-related distribution of utilities for the general population

Ara and Brazier argue that ideally, the baseline age-related distribution of utilities in disease-free health states would be derived from people without specific condition(s) using the definitions of health states in the model (Ara and Brazier 2011). However, these data are rarely available, and the age-related distribution of utilities in the general population (irrespective of health condition) is often used as a proxy.

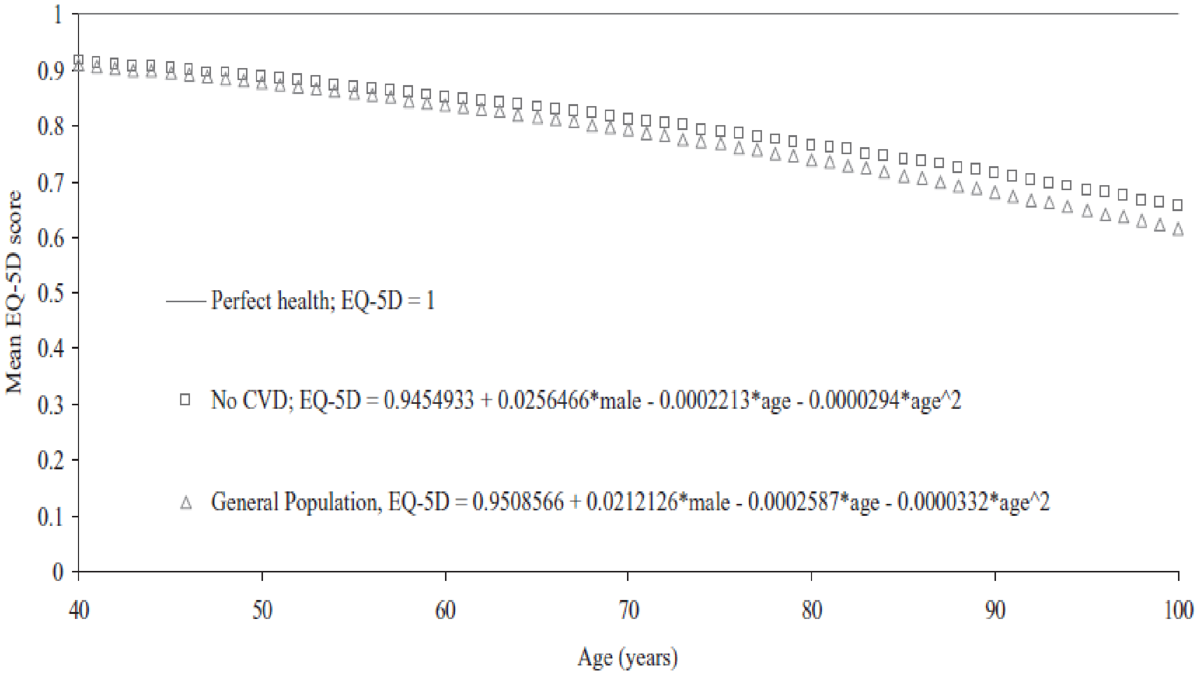
For example, Ara and Brazier (2010) showed that age-adjusted utility values from the general population are a good proxy for cardiovascular disease-free health states. Ara and Brazier (2011) also suggested that utility values from the general population could be used in place of condition-specific data (to represent the utility values associated with not having a particular health condition) in some analyses but not all. In particular, they showed that utility values from the general population may be appropriate for cohorts with multiple conditions, but less appropriate for cohorts who have just 1 health condition. In these instances, if the condition-specific data are not available, they suggested that using age-stratified mean utility values from respondents who reported they have none of the prevalent health conditions may be more appropriate.

Sources of general population data

1. *Ara and Brazier (2010)*

This study was based on data from Health Survey for England (HSE), which is conducted annually using random samples of the population living in private households in England. The 2003 and 2006 surveys included questions about history of cardiovascular disease, and a random sample of participants (aged 16 to 98 years) were asked to complete the EQ-5D questionnaire (n=26,679). The relationship between utility values, age, sex and self-reported history of cardiovascular disease was explored using linear regression and compared with a hypothetical 'perfect health' scenario (**Figure *1***). Two models were obtained, 1 for general population (using full dataset; n=26,679), and the other for individuals who reported no history of angina, heart attack or stroke (n=25,080). The former can be used to estimate the mean utility values for individuals in the general population, and the latter to estimate the mean utility values for individuals with no history of cardiovascular disease. Both equations are displayed in **Figure *1***. They describe the non-linear distribution of utilities with age.

Figure 1 Baseline utility for the event-free health state: Relationship between EQ-5D, age, sex and history of cardiovascular disease (Ara and Brazier 2010)



2. *Kind et al. (1999)*

The authors published UK population norms for EQ-5D. The data were collected in 1993 as part of the nationally representative interview survey of 3,395 people aged 18 or over living in the UK. EQ-5D population norms were stratified by a number of covariates, including sex and age (in 10-year increments; **Figure 1** in section 2 of this report). These population norms were used for adjusting the utility values in for example, TA575 or TA578. All assessments assumed a linear relationship between age and utilities to calculate the annual or per cycle utility decrement, which is a key limitation of these analyses. Also, Kind et al. (1999) grouped all people aged over 75 years into 1 age bracket (75+) so the relationship between age and utilities in this age group is unclear. These limitations were pointed out in TA575 but were not discussed in TA578.

3. *Janssen and Szende (2014)*

This publication provides population norms for a number of European countries on behalf of the EuroQol working group. UK-England population norms were based on the Health Survey for England 2010 results (computer-assisted interviews on a randomly selected sample of households in England; n=14,763). UK population norms were based on Kind et al. (1999) study. Both sets of EQ-5D population norms are presented in **Table *2***. These population norms were used for example in TA612.

Table 2 EQ-5D index population norms (country-specific time trade-off value sets) from Janssen and Szende 2014

Age group	**18–24**	**25–34**	**35–44**	**45–54**	**55–64**	**65–74**	**75+**	**Total**
UK	0.940	0.927	0.911	0.847	0.799	0.779	0.726	0.856
UK-England	0.929	0.919	0.893	0.855	0.810	0.773	0.703	0.855

Adjusting utility values over time

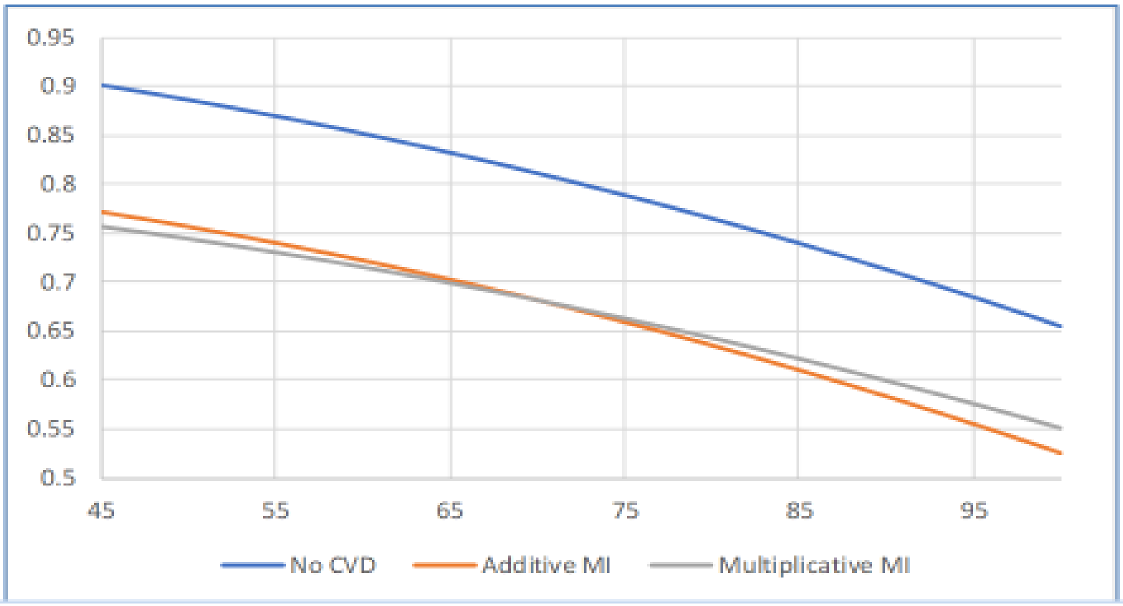
Two methods have been explored to adjust utility values over time: multiplicative and additive.

- Multiplicative method assumes constant relative decrement of disease health states on utilities; this method has a larger effect at higher baseline utility values.
- Additive method assumes constant absolute utility decrement of disease health states on utilities; it has a larger effect at lower baseline utility values.

Their effects on utility values over time is presented in

Figure ***2***.

Figure 2 Comparison of multiplicative and additive methods for adjusting utility values over time



Ara and Brazier (2010) describe the multiplicative method for adjusting utility values over time. Their first example shows how to calculate total QALYs accrued from avoiding a single event (angina at the age of 50) over a 50-year time horizon. By comparing the mean EQ-5D for angina (at the mean age of 69 years), and the mean EQ-5D for general population (or general population with no history of cardiovascular disease) at the same age, it's possible to calculate a multiplier, which can then be applied to calculate the expected mean utility values for angina at the age of 50, and subsequent years in the model. The steps for carrying out such a calculation are set out in the example below:

1. Utility value for men with angina at the age of 69 years is 0.61

2. Average utility for men in the general population at the age of 69 years is 0.80

The **multiplier** for men with angina is therefore 0.61/0.80 = **0.77**

To calculate the utility value for men with angina at a particular age, the utility value for the general population at a particular age is multiplied by the multiplier. In this example, at age 80, the utility value for men in the general population is 0.74 and multiplying by 0.77 from above gives an adjusted utility value of **0.57**.

In contrast the **additive approach** works out a **utility decrement** by subtracting the utility of people with angina from the value for people in the general population at the same age. In this example, 0.80 – 0.61 = **0.19**.

To calculate the utility value of men with angina at a particular age the utility decrement is subtracted from the average utility value for the general population. In this example, average general population utility at age 80 is 0.74 and subtracting the decrement from above of 0.19 gives and adjusted utility value of **0.55**.

A limitation specific to the additive approach is that under certain conditions, a very low baseline utility value could result in such a large utility decrement that over the model time horizon, utility values could be close to zero or negative. This would not happen under the multiplicative approach.

A limitation of the both the multiplicative and additive methods is that they assume that the disutility multiplier or decrement related to a particular condition or event (for example, heart attack) is constant over time (regardless of time from the event), which may not always be a valid assumption (Brazier et al. 2019). For example, the impact of heart attack on utility values was shown to be reducing with age, the impact of rheumatoid arthritis was constant with age, and the impact of skin complaints increased with age (Brazier et al. 2019).

Figure 5 Mean EQ-5D scores and average health-related quality of life decrements (relative and absolute) for myocardial infarction

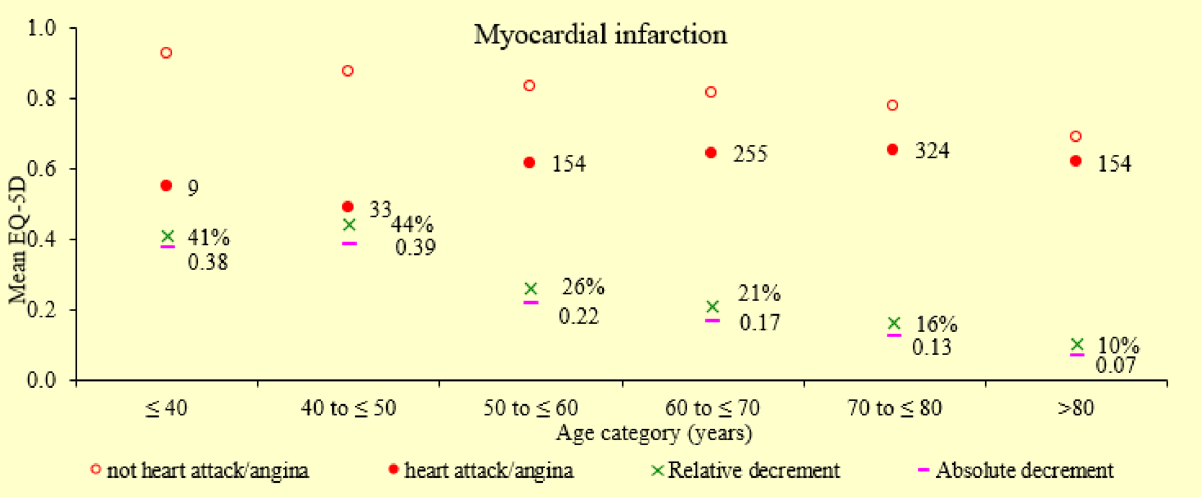


Figure 6 Mean EQ-5D scores and average health-related quality of life decrements (relative and absolute) for rheumatism

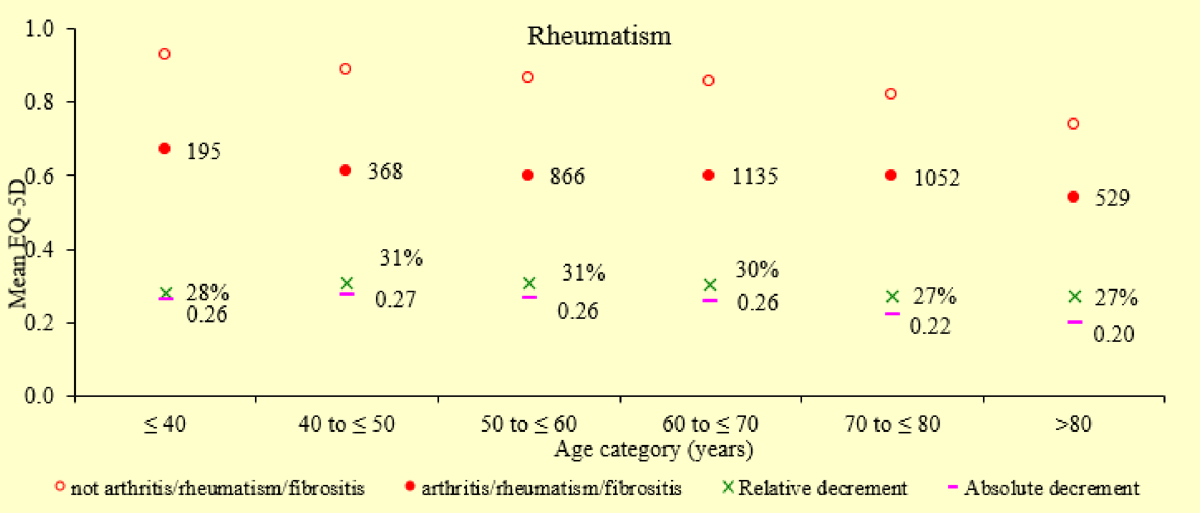
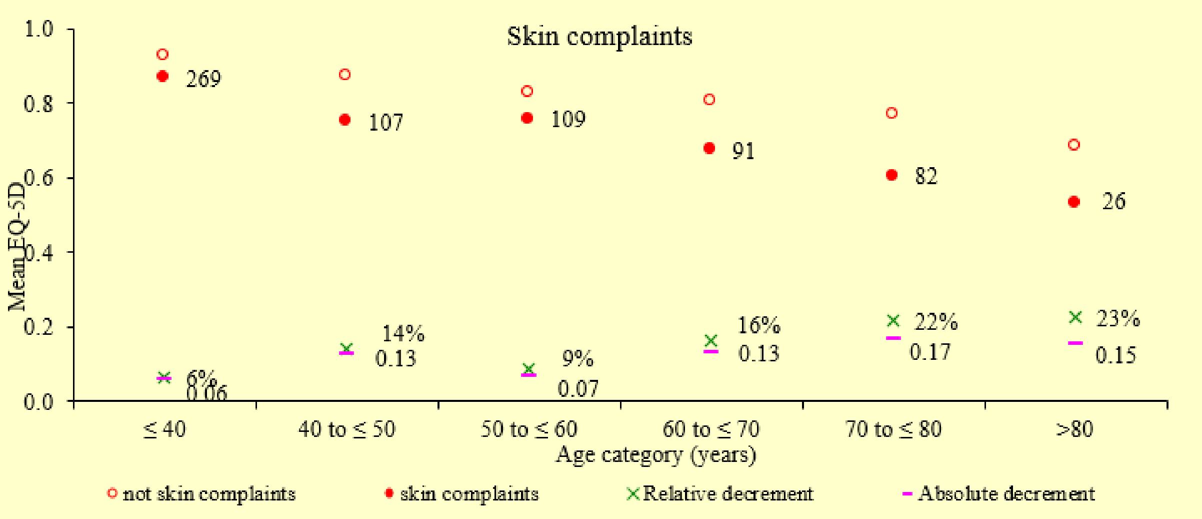


Figure 7 Mean EQ-5D scores and average health-related quality of life decrements (relative and absolute) for skin complaints



Note: The number of cases is shown next to data points for respondents who have the condition.

Other considerations

Utility values may decline over time in chronic progressive conditions. For example, multiple sclerosis is usually modelled as progression through health states defined by expanded disability scale score. As disability accumulates, each successive health state is associated with a lower utility value. Adjusting for age on top of an observed decline in utility values may double count the effects of age as older patients are generally more likely to be in severe health states than younger patients in progressive conditions (Ara et al. 2017).

In addition, the EQ-5D measures health in 5 domains. Some individuals may start off at the lowest level of a domain, for example, because they cannot walk. The aim of adjusting utility values is to account for accumulation of comorbidity and progressive loss of function such as the ability to walk over time. However, if individuals have already lost these functions then their utility over time may follow a different functional form – that is, it may not decline as much or at all. In these situations, it may not be appropriate to adjust the utility values, but the utility values are also less likely to exceed the general population values at a given age, the primary motivator for adjusting.

In TA523, the ERG adjusted the utility values using the starting age in the model as an anchor. However, the utility values were sourced from studies that included a generally older population. In response to consultation, the company acknowledged the need to adjust the utility values over time but proposed an alternative approach. It proposed using the mean age of the patients in the different utility studies as an anchor, which they considered to be more appropriate and reflective of the true utility values rather than the starting age in the model. The committee agreed this approach was appropriate, but it had a limited effect on the cost-effectiveness results.

Key questions

- How to avoid double counting the impact of age in chronic conditions?
- How to ensure distribution of utility values from general population is appropriate to use (for example, documenting the searches to show no disease-specific utility values found?)
- How to ensure assumption underlying multiplicative and additive methods (constant utility multiplier or decrement) for adjusting utility values is valid in a particular condition?

4. Impact of adjusting utility values on cost-effectiveness estimates

Case studies

A review of selected TAs was done to assess the impact of adjusting utility values on cost-effectiveness estimates. The case studies were identified by searching the NICE website using ‘age adjusted’ term for guidance published in the past year. The search was not designed to be comprehensive but rather to identify the most recent examples of guidance where age-adjustment of utilities was discussed by the committee. The search, run on 28 November 2019, identified 38 hits. Of these, 24 applied age-adjustment (either by the company or by the evidence review group), whereas the remaining 14 did not use or mention the use of age-adjusted utilities. Six of 24 provided details that allowed the impact on the cost-effectiveness estimates be assessed, and are summarised in **Table 2**.

Only 1 TA was identified, TA575, where the company adjusted utility values in its base case, and the evidence review group and committee thought that the adjustment introduced unnecessary complexity and preferred to use unadjusted utility values (**Table 2**; rationale: no effect on mortality so no need for adjustment). There were a number of TAs where the company did not adjust the utility values in their base-case, but the evidence review group and the committee thought it was appropriate to do so, to reflect the natural decline in health-related quality of life over time. The main concern was that without adjustment, utility values assigned to some health states may eventually exceed utility estimates from general population for the same age group, when a lifetime horizon is considered.

None of the reviewed guidance used age-related distribution of utilities derived from people without the specific condition(s). Distribution of utility values in the general population from Ara and Brazier (2010) and Kind et al. (1999) were most frequently used. TA612 followed age-based decline in utilities from Janssen and Szende 2014. The appendix with full details of case studies is not included because it contains confidential information, but a summary of the results is in table 2 below.

Both multiplicative and additive approaches have been used in previous TA guidance, although multiplicative approach appears to be more frequently used. The selection of a particular approach was rarely justified, and was not stated altogether in some guidance.

Adjusting utility values over time increased the cost-effectiveness estimates by 2.5% to 9% compared with unadjusted values (**Table 2** and appendix 1). This was because the unadjusted utility values overestimated the benefit of the intervention in later years.

Table 2 Summary of case studies

ID	Clinical area	Population/ mean age	Life extending?	Utility values adjusted by company?	Utility values adjusted by evidence review group (ERG)?	Committee conclusion	Method	Effect on incremental cost-effectiveness ratio (ICER) of adjusting the utility values
TA523	Acute myeloid leukaemia	45 years	Yes	No	Yes	Appropriate	Multiplicative; Ara and Brazier 2010	~7% increase
TA575	Psoriasis	46 years	No	Yes	No (age-adjustment removed)	Not appropriate	Multiplicative; Kind et al. 1999	~5.5% increase
HST12	Neuronal ceroid lipofuscinosis type 2 (CLN2)	Children (~5 years)	Yes	No	Yes	Appropriate	Unclear	~2.5% increase
TA612	Breast cancer	51 years	Yes	No	Yes	Appropriate (decision by the technical team)	Unclear; Janssen and Szende 2014	~9% increase
TA578	Lung cancer	63 years	Yes	No	Yes	Appropriate (decision by the technical team)	Additive; Kind et al. 1999	~5% increase
TA595	Lung cancer	62 years	Yes	No	Yes	Appropriate	Unclear; Ara and Brazier 2010	4.4% to 5.2% (depending on the comparator) increase

5. Equality considerations

Age is 1 of 9 protected characteristics covered in the Equality Act 2010. Within the public sector, which NICE falls under, the Equality Duty must be followed.

As a public body, NICE is required to have ‘due regard’ with respect to eliminating unlawful discrimination, and to ‘remove or minimise disadvantages suffered by people due to their protected characteristics’.

Two previously published TAs were chosen to explore whether adjusting utility values has a bigger effect when the starting age of the cohort in a model is older. The results, presented in table 3 below, are mixed.

Both models show that adjusting the utility values for ageing reduces total QALYs for the intervention and comparator. The first model shows that the absolute difference in incremental QALYs is greater for younger cohorts. However, for the second model, the trend is less clear.

Adjusting the utility values also increases the incremental cost-effectiveness ratio (ICER) by a larger absolute amount at a younger starting age in model 1, but this is not always the case in model 2.

The exact effects will depend on the relative contributions of survival and quality of life to outcomes, baseline utility values and how these differ between the different health states in the model.

Table 3 Impact on QALYs and ICERs from adjusting utility values for different starting ages, in model 1 (cardiovascular disease) and model 2 (leukaemia)

Starting age	45 (model 1)	60 (model 1)	75 (model 1)	45 (model 2)	53 (model 2)	60 (model 2)
Intervention quality-adjusted life years (QALYs)	-0.69	-0.34	-0.13	-0.15	-0.18	-0.16
Comparator QALYs	-0.51	-0.23	-0.09	-0.11	-0.13	-0.14
Incremental QALYs	-0.18	-0.11	-0.04	-0.04	-0.05	-0.03
ICER	+£4.0k	+£2.0k	+£1.5k	+£1.3k	+£1.9k	+£1.7k
% change in incremental cost-effectiveness ratio (ICER)	+11.5%	+8.0%	+5.0%	+4.5%	+5.4%	+3.9%

The analysis shows that overall, adjusting utility values is likely to reduce incremental QALY gain regardless of starting age. This is because the unadjusted utility values overestimate quality of life at older ages, but over a lifetime horizon, younger cohorts will also reach these older ages. This reduction in QALY gain could have a greater impact on ICERs for populations at older starting ages, because incremental QALYs tend to be lower because of lower life expectancy. However, this is not clearly the case in the 2 examples above. Overall, the evidence does not indicate that older populations are disproportionately affected by adjusting utility values over time.

The arguments for adjusting utility values to reflect changes in those of the general population over time are analogous to those for adjusting survival rates in models so that they do not exceed background mortality (that is, ensuring people with the disease in question are not predicted to live longer than people without the disease). As such, it is a proportionate means of achieving the legitimate aim of allocating scarce healthcare resources efficiently.

The methods for adjusting utility values over time serve as an imperfect proxy for the optimal solution, which would be to have longitudinal data on health-related quality of life for people with the disease having the treatments of interest over time.

6. Other considerations

Future unrelated costs

If it is appropriate to consider future unrelated comorbidity (effectively what adjusting utility values over time does), then it may also be appropriate to consider future unrelated care costs. This is being explored by another task and finish group.

Other HTA bodies

A review of other international health technology assessment bodies did not find any instances in which the respective methods guides refer to adjusting utility values over time.

7. Case for change and proposals

Case for change

- Ideally there would be longitudinal data on health-related quality of life, but usually there are only data for the duration of the trial.
- There is general consensus in the academic community that when extrapolating this data over long time horizons, it is appropriate to adjust values to reflect the decline in quality of life seen in the general population over time and to ensure that values do not exceed those of the general population at a given age. For example, ISPOR recommends this as best practice.
- Most companies adjust utility values over time in their initial submissions; when they do not, the ERG and the committee normally request that they do.

- Adjustment is being done inconsistently between appraisals. The multiplicative approach is used more often than the additive approach. At higher baseline utility values, the adjusted values are similar using either method, but at lower baseline utility values, the adjusted values using the additive method can end up significantly lower than when using the multiplicative method. In specific circumstances, the additive approach can lead to utility values close to zero, or negative, which does not occur with the multiplicative approach.
- Exploratory modelling shows that adjusting utility values over time does reduce the health gain compared with no adjustment; this is because unadjusted values overestimate health-related quality of life at older ages. However, over a lifetime horizon, younger cohorts will also reach these older ages, so for a given disease, adjustment can have a greater impact on QALY gain at younger starting ages.
- Adjustment could have a greater impact on ICERs at older starting ages, because the QALY gain is lower because of shorter life expectancy, but the examples explored did not indicate this.
- Overall, adjusting utility values over time is a proportionate means of achieving the legitimate aim of allocating scarce healthcare resources efficiently.
- There may be some situations in which it may be inappropriate to adjust utility values, so any amendments to the methods guide should still give the scope for those arguments to be made and considered by the committee.

Proposal

Update methods guide to state:
If baseline utility values are extrapolated over long time horizons, they should be adjusted over time to reflect decreases in quality of life seen in the general population and to ensure they do not exceed general population values at a given age.

- If this is not considered appropriate for a particular model, a supporting rationale should be provided.
- A multiplicative approach is generally preferred, and the methods used for adjusting utility values should be clearly documented.

Authors

Amy Crossley, Ewa Rupniewska and Ross Dent on behalf of the Health-related quality of life task and finish group

References

Ara R and Brazier JE (2010) Populating an economic model with health state utility values: moving toward better practice. Value in Health. 13(5): 509–518

Ara R and Brazier JE (2011) Using health state utility values from the general population to approximate baselines in decision analytic models when condition-specific data are not available. Value in Health. 14(4); 539–545

Ara R, Brazier J and Zouraq IA (2017) The use of health state utility values in decision models. Pharmacoeconomics. 35(1): 77–88

Blanchflower and Oswald, Is wellbeing U-shaped over the life cycle? Social science & medicine, 2008. 66(8): 1733–1749

Brazier J, Ara R, Azzabi I, et al. (2019) Identification, review, and use of health state utilities in cost-effectiveness models: an ISPOR good practices for outcomes research task force report. Value in Health. 22(3): 267–275

Janssen B and Szende A (2014) Population norms for the EQ-5D. In: Szende A, Janssen B, Cabases J, editors. Self-reported population health: an international perspective based on EQ-5D. Dordrecht: Springer Netherlands, 2014: 19–30

Kind P, Hardman G and Macran S (1999) UK population norms for EQ-5D (No. 172chedp)

Report 4: Core outcome sets

1. Introduction

The evidence requirements of NICE's guidance producing programmes send an important signal to researchers and trialists designing clinical studies for market access. Specifying outcomes of interest for guidance production may also have an impact on wider clinical practice.

NICE's involvement in core outcome set development and implementation could:

- improve the relevance and consistency of outcome selection and measures and
- assist decision making by ensuring the most appropriate evidence is submitted to guidance producing committees.

A preference for core outcome sets could also influence evidence generation in managed access schemes, facilitate real world evidence collection for benchmarking within healthcare settings or longitudinal monitoring of health outcomes over time.

This report will outline:

- outcome definitions and types
- core outcome sets definition
- NICE's current experience with core outcome sets
- options for using core outcome sets within technology appraisals.

2. Outcome definitions and types

There are many interrelated terms and definitions used in relation to outcomes and/or endpoints.

- **Clinical outcomes:** these may be a clinical event (for example, cardiovascular event) a composite of several events, a measure of clinical status (symptoms and function), or health-related quality of life. They can be reported by a patient (patient-reported outcomes, PROs), clinician or a carer. Clinical outcomes can be surrogate (intermediate endpoints) or final, and cover both benefits and harms.
- **Patient-relevant outcomes**: defined in most cases as an outcome that measures mortality, morbidity and/or health-related quality of life. These are not necessarily selected by patients or with patient input.
- **Patient-important outcomes:** cover the same measures as patient-relevant outcomes but are defined as outcomes that patients value directly.
- **Patient-centred outcomes:** outcomes that are meaningful, valuable and helpful to patients and their families. It involves putting patients, and their families and carers, at the heart of deciding which goals are most valuable for them.

- **Patient-reported outcomes:** cover a range of measurement types, including symptom measures (such as pain measured using a Likert scale) complex measures (such as activities of daily living or function), multidimensional measures (such as health-related quality of life) and satisfaction with treatment (patient-reported experience measures, PREMs). The key component is that the outcome is directly reported by the patient.

International perspective on clinical outcomes

The European Medicines Agency, US Food and Drugs Administration and many international health technology assessment bodies have criteria relating to the selection and use of clinical outcomes that relate to the type of outcome (such as final, surrogate, intermediate, composite, validated and PRO), statistical significance, definition of a meaningful clinical change and measurement within clinical trials.

Current guide to methods of technology appraisal

NICE's preference on clinical outcomes is 'for long term or final patient relevant outcomes that reflect how patients feel, function or survive'.

NICE's guide to the methods of technology appraisal states that 'the scope identifies principal measures of health outcome(s) that will be relevant for the estimation of clinical effectiveness. That is, they measure health benefits and adverse effects that are important to patients and/or their carers'.

The only area where the methods guide states a clearly defined preference for outcomes is for the health-related quality of life data used within cost-effectiveness analysis. This should be generated by collecting EQ-5D data within a clinical trial. However, there is no guidance on patient-reported outcomes (symptoms and function) or disease-specific health-related quality of life instruments that may be used to assess clinical effectiveness.

3. Core outcome sets

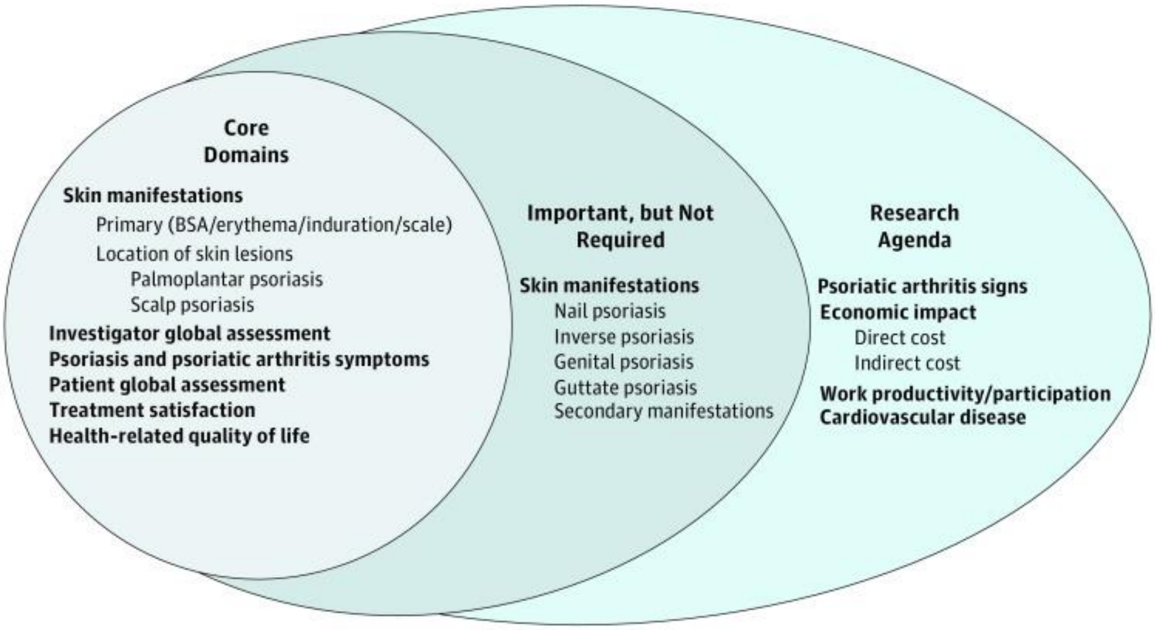
One of the main issues facing decision makers on outcome reporting is substantial variability in the outcomes that are selected and measured across healthcare settings, and even within disease areas. Such variation can make synthesising research studies difficult or impossible. This presents a particular challenge for NICE's guidance producing programmes, where it is necessary to assess the body of evidence associated with a particular decision. Inconsistency in outcome selection also highlights a potentially larger issue: that clinical research and policy decisions may not address the outcomes that matter most to patients, clinicians and healthcare commissioners. One potential solution to the issue of outcome variability is the use of core outcome sets.

Core outcome sets are an agreed minimum set of outcomes that should be measured and reported in all clinical studies for a specific disease or condition. They are increasingly being produced for clinical practice and the collection of data in real world settings. A good quality core outcome set is developed with the input of relevant stakeholders, including patients, researchers, clinicians and other healthcare decision makers. Patient engagement is central to core outcome set development, which is based on the concept of a consensus-based approach to determining which outcomes are most relevant and meaningful to patients and other decision makers.

Core outcome sets are widely endorsed by a number of agencies that NICE engages with, including the European Medicines Agency, the National Institute for Health Research, the European Federation of Pharmaceutical Industries and Associations and Cochrane (see appendix 2 for more details).

The figure shows an example of a core outcome set for psoriasis.

Figure 1 Core outcome set for psoriasis



Source: Onion model of core domains for psoriasis clinical trials

The COMET database is an output of an internationally funded initiative (European Commission, Medical Research Council, and the National Institute for Health Research). The database is regularly updated, free to use and has accessible searchable functionality, providing links to all freely available articles.

The tools below can be used to assess the quality, validity and rigour in its development of a particular core outcome set:

- Core outcome sets-STAP (core outcome set-standardised protocol items) statement for the content of a core outcome sets development study protocol.
- COMET handbook on the development of core outcome sets.
- Core outcome set-STAD (core outcome sets-standards for development) identifies minimum standards for the design of a core outcome sets study agreed on by an international group.
- Core outcome sets-STAR (core outcome sets-standards for reporting) statement consists of a checklist of 18 items considered essential for transparent and complete reporting in all core outcome sets studies.

4. Core outcome sets at NICE

Centre for Guidelines

Core outcome sets have been indicated for use in scoping in clinical guidelines since 2012. The unified updated manual expanded the remit for core outcome sets to include public health and social care. The Centre for Guidelines currently endorses the use of good quality published core outcome sets in guideline development (for selecting outcomes within the PICO [population, intervention, control and outcomes] for each systematic review) as stated in NICE's guideline manual:

'Core outcome sets should be used if suitable based on quality and validity; one source is the COMET database. The core outcome set standards for development (core outcome sets-STAD) and core outcome set standards for reporting (core outcome sets-STAR) should be used to assess the suitability of identified core outcome sets.'

In addition, the guideline surveillance team is piloting the use of core outcome sets to define outcomes to prioritise within their decision making for updates and reviews.

Evidence resources directorate

Core outcome sets are suggested for the development of NICE's evidence standards framework for digital health technologies, which specifically states:

'The outcome measures reported should reflect best practice for reporting improvements in the specific condition, using validated outcome measures such as those in the COMET core outcome set.'

Centre for Health Technology Evaluation

Core outcome sets are one of the resources searched for by the information services team to inform the development of the scope. There are limited suggestions on how to systematically and transparently identify, select or measure outcomes. There is no direct reference to the use of core outcome sets within the technology appraisal methods guide.

Health and Social Care Directorate

Core outcome sets are used within the quality indicators programme for identifying outcomes to align with those selected by Centre for Guidelines whenever possible.

Scientific Advice

NICE Scientific Advice does not systematically use core outcome sets, they will use all sources of available information to address questions relating to outcome selection. This has on occasions included core outcome set measurement instruments (COSMIN).

5. Case for change

Core outcome sets offer the following potential benefits:

- Patients' perspective and patient-important outcomes identified; greater assurance that outcomes important to patients are informing decisions.
- Patient-reported outcome measures are often included in core outcome sets; these are important for patient-centred approaches.
- Increased transparency from peer-reviewed selection and identification of outcomes.
- Greater ability to make meaningful comparative assessments of evidence and enables real world evidence to inform reviews.
- Increased consistency of outcome selection across NICE guidance programmes.
- Broad adoption creating greater certainty for trialists, researchers and companies.
- Enhancing NICE's reputation as innovative and a leader in the area of evidence synthesis and use.
- Linked products resulting from guidance such as shared quality standards, shared decision tools, and a range of advice products could also use the same core outcome sets within a disease or health area to encourage consistency and a connected approach to guidance.

6. Options for using core outcome sets in technology appraisals

Scoping

Core outcome sets could be used (alongside the methods currently used) to identify patient-relevant outcomes for including in PICO within the scope of an appraisal. This could easily align with the methodology used in the Centre for Guidelines and the use of the COMET database.

To assess the quality, validity and robustness of a core outcome set, a number of tools are freely available. Feedback from Centre for Guidelines indicates that these are minimally resource intensive and do not need a high level of technical skill to complete.

For the 'Outcomes' section of the PICO, to ensure that the COMET database is searched for core outcome sets, the detail on search terms and availability and use of core outcome sets (within a disease area, that is, oncology, liver disease) and specified population should be documented.

It may not be appropriate to use an identified core outcome set, but the rationale for this decision should be clearly documented. If individual outcomes are removed as indicated by stakeholder consultation and/or clinical input (this may include feedback from scoping consultation and clinical input on the relevance of an outcome), then a rationale should also be indicated. Lack of measurement within a trial is not a sufficient rationale to exclude from a scope.

The scope should ideally contain all the outcomes that NICE requires for its decision making. Outcomes identified in a core outcome set could be highlighted by the use of formatting and the core outcome set referenced within the scope.

Possible amendment to methods guide

Wording could align with that already used within NICE guideline manual:

‘Core outcome sets should be used if suitable based on quality and validity; one source is the COMET database. The Core Outcome Set Standards for Development (core outcome sets-STAD) and Core Outcome Set Standards for Reporting (core outcome sets-STAR) should be used to assess the suitability of identified core outcome sets.’

Appraisal

Outcomes not provided as requested within the scope should be detailed in the company submission, as is current practice. Core outcome set outcomes could be differentiated by formatting.

Outcomes not provided within a submission should ideally be noted by the evidence review group within their report and noted within the technical report, appraisal consultation decision and the final appraisal document.

No changes to the methods guide would be required for this.

Likely impact of potential changes

Including core outcome sets in scopes is likely to have a minor impact on the outcomes that are included. Most core outcome sets will include the key outcomes relating to length of life and quality of life, which are needed for deriving quality-adjusted life years.

A review of technology appraisal oncology scopes suggested that most of the time, the outcomes selected overlapped completely with those in core outcome sets. In the small number of exceptions to this, the core outcome sets included 2 additional outcomes to those specified in the scope.

Best practice guides for producing core outcomes sets (see for example a practical toolkit for the identification, selection and measurement of outcomes including in real-world settings) recommend around 8 outcomes. This is similar to the number of outcomes usually included in scopes.

If outcomes specified in previous scopes in the same disease area are not part of the core outcome set, these can still be included in the scope, to ensure consistency with previous appraisals.

Core outcomes sets may not be available for some diseases or conditions. In these situations, outcomes to be included in the scope would be selected in the same way as they currently are.

The assessment of the quality and suitability of core outcome sets will increase the time it takes the NICE technical team to produce scopes. In addition, where multiple core outcome sets exist, or where core outcome sets contain a large number of outcomes, the technical team will need to make decisions about what to include in the scope. However, through consultation on the scope, stakeholders would get the opportunity to comment on these decisions.

7. Conclusion

There is a concerted international effort to use core outcome sets in health technology assessments, and they are already used in other NICE guidance producing programmes, including the Centre for Guidelines. The main advantages of core outcome sets are that they facilitate more comparative assessments between studies and outcomes are selected with patient and clinician input and subject to peer review.

The methods guide could be aligned with the Centre for Guidelines methods guide so that core outcome sets are identified and quality assessed during the scoping phase of appraisals. The task and finish group report sets out how this might be implemented including potential wording for the methods guide.

However, a significant drawback of adopting core outcome sets in scoping of appraisals is that searching and quality assessing them would increase the resource intensity of scoping for the NICE team. This additional effort may not be justified, given that a review of oncology scopes found a significant degree of overlap between outcomes that are routinely included in scopes and those in core outcome sets.

It is felt that some of the key benefits of core outcome sets could be achieved by encouraging in the methods guide that all outcomes should be relevant to patients. Therefore, it is suggested that the methods guide is updated to state that:

- Outcome measures in studies should be selected in consultation with people with the condition or disease, so that the study reflects what matters to them.
- A high-quality core outcome set, developed with input from people with the disease or condition, may help with outcome selection.
- Patient-reported outcomes can capture important aspects of conditions and interventions. Patient-reported outcome measures should be appropriately validated, and the methods used to collect the data should be clearly reported.

In the future, as part of ongoing transformation work, there could be the opportunity for NICE to maintain a library of quality-assured core outcome sets that can be used across all programmes. This would alleviate concerns about resource constraints in the technology appraisals programme, and ensure a consistent approach between guidance producing teams.

Authors

Katy Harrison and Ross Dent on behalf of the Health-related quality of life task and finish group

References

- COMET database
- International Consortium for Health Outcomes Measurement (ICHOM)
- Outcome Measures in Rheumatology (OMERACT)
- Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials (IMMPACT)
- European Medicines Agency (EMA) Patient registries initiative 2018
- Green Park Collaborative, Core Outcome Set Initiatives, Center for Medical Technology Policy (CMTP)

Appendix 1 Current and previous core outcome sets activity at NICE

Conferences

- Centre for Guidelines have conducted at Guideline International [GIN] Conference training session linked with COMET initiative on the development of core outcome sets 2017, 2018
- Health Technology Assessment international 2018 Vancouver-Panel session-chairing session on standardising outcome selection and measurement – is health technology assessment leverage a key
- Oral presentation by Science Policy and Research (SP&R) on core outcome sets and health technology assessment outcome selection at COMET 2018 meeting
- Posters at GIN (n=5; 2017, 2018), 2 from guidelines or quality indicators, 2 from SP&R and ISPOR (2018), COMET international meeting 2019
- ICHOM 2019 presentation – real-world evidence (RWE) and patient-centred outcomes (SP&R)

Development of core outcome sets

NICE's input into core outcome sets:

- The coreHEM initiative and the publication of a core outcome sets for clinical trials of gene therapy in haemophilia. Core outcome set for gene therapy in haemophilia: results of the coreHEM multistakeholder project. Haemophilia. 2018; 00:1–6 (CHTE committee member)
- CoreNASH (SP&R)
- CoreSCD (NSA)
- IMI HARMONY – NICE is providing and coordinating health technology assessment and methodological assistance input in 7 disease-specific and 1 topic-specific core outcome sets in the area of haematological malignancies

Engagement with key initiatives

- COMET – links and joint publications both with SP&R and Centre for Guidelines and joint PhD student 2016–2019
- ICHOM active partner on IMI EDHEN with SP&R
- Green Park Collaborative – Inclusion of NICE and its perspective on White paper developed by Green Park Collaborative (a major initiative of the Center for Medical Technology Policy): A multi-pronged strategy to improve the relevance, usefulness, and comparability of outcomes in clinical research

Methods

- Big Data for Better Outcomes projects

  - DO-IT products and outputs including:
    - A practical toolkit for the identification, selection and measurement of outcomes including in real-world settings (lead developer, NICE)
    - Public webinar on 15 May, attended by around 100 people from across sectors, representing the public-private nature of the project

Appendix 2 Organisations and initiatives actively promote the use of core outcome sets

- European Medicines Agency (EMA) patient registries initiative (5)
- EMA – certain disease-specific CHMP [Committee for Medicinal Products for Human Use] guidelines
- European Federation of Pharmaceutical Industries and Associations
- The Centre for Medical Technology Policy are leading a project working with post-regulatory decision makers to promote uptake of core outcome sets and developing core outcome sets – NICE is an active participant in many of these outputs (6)
- SBU – Swedish Agency for Health Technology Assessment and Assessment of Social Services
- Promoting the adoption of core outcome sets in clinical research. Arthritis Research UK (ARUK)
- Health Research Board (HRB)
- The IDEAL Collaboration
- SPIRIT (Standard protocol items: recommendations for interventional trials)
- Deutsche Forschungsgemeinschaft (DFG) German Research Foundation – clinical trials research proposals
- COMET Initiative – an international multidisciplinary network with Medical Research Council and EU funding, which aims to raise awareness of current problems with outcomes in clinical trials, encourage the development of core outcome sets, and provide resources to enable the development of core outcome sets (1)
- ICHOM develops ‘standard sets’ of outcomes for routine or real-world settings across a range of disease areas (standard sets also include minimum datasets, which refer to other characteristics like age or health behaviours; 2)
- Core outcome sets developers in specific disease areas include OMERACT for rheumatoid arthritis and IMMPACT for pain, which refer to ‘core domain sets’ or ‘core outcome domains’ respectively (3, 4).

Report 5: Measuring and valuing children's health-related quality of life

Background

Health-related quality of life instruments each have 2 parts:

- the questionnaire used to measure people's health status
- the value set allowing responses to be converted into utility scores, based on the public's preferences for different health states.

A briefing on the key terminology and concepts is in appendix 1.

NICE provides clear guidance on how to measure and value health-related quality of life in adults, but is less clear on preferred methods for children and young people. NICE's guide to the methods of technology appraisal states:

'5.3.11 When necessary, consideration should be given to alternative standardised and validated preference-based measures of health-related quality of life that have been designed specifically for use in children. The standard version of the EQ-5D has not been designed for use in children. An alternative version for children aged 7–12 years is available, but a validated UK valuation set is not yet available.'

A Decision Support Unit (DSU) review of past technology appraisals and highly specialised technology evaluations found 31 whose population included children and young people (Hill et al. 2019). The key finding was that the health-related quality of life of children and young people is rarely measured directly using an age-appropriate instrument. Only 7 evaluations (23%) used a paediatric questionnaire to measure the health-related quality of life of a child with the condition and used this to inform the utility value of at least 1 health state in the model. Most evaluations (27/31, 87%) used the EQ-5D scored using the UK adult value set for at least 1 health state. The population completing the adult EQ-5D was often not clearly reported, so we do not know if it was adults who had the condition, adults acting as a proxy, or children and young people themselves.

NICE's patient and public involvement policy says we produce 'guidance and standards on topics covering children and young people's health and wellbeing, which have been informed and influenced by their views and experiences'.

This paper does not address the question of whether committees should alter their standard approach to decision making because a treatment is for children or young people rather than adults; that question is addressed by the Modifiers workstream.

Case for change

The case for change has 4 components:

- The DSU review of published guidance in children and young people showed wide variation in methods and poor reporting of the source of utilities.
- The DSU review found widespread use of the adult EQ-5D. But there isn't evidence that the adult EQ-5D performs well psychometrically in paediatric populations (Noyes and Edwards 2011) and EuroQol does not recommend its use in children under the age of 12 (van Reenen et al. 2014).
- The DSU review shows that submissions rarely use age-appropriate measures that allow children and young people to assess their own health-related quality of life, which goes against our patient and public involvement policy.
- Academics, industry and NICE's scientific advice team would welcome clearer guidance on how to measure and value the health-related quality of life of children and young people.

Unfortunately, the academic literature is not mature enough to enable NICE to recommend specific health-related quality of life measure(s) and value set(s) for children and young people. Instead, we propose a light-touch clarification of methods guidance as part of the methods update, alongside support for longer-term research (summarised in appendix 2).

Proposals for the methods guide

1. For appraisals and evaluations whose population includes children and young people (that is, people under 18 years old), the methods guide should include the following recommendations.

   a. Measure the health-related quality of life of children and young people using a generic measure that has been shown to have good psychometric performance in the relevant age range(s). Not all paediatric health-related quality of life instruments have a UK value set, and there are methodological challenges when developing value sets for children and young people. Nonetheless, generic measures give valuable descriptive information about the impact of the condition and intervention on children and young people's health-related quality of life. If data from a paediatric health-related quality of life instrument are used to generate utility values, explain how this was done. If there is evidence that generic measures are unsuitable for the condition or intervention, refer to section X.

   b. NICE does not recommend specific measure(s) of health-related quality of life in children and young people. The choice of measure should be informed by evidence of psychometric performance, evidence that it is valid in the age range(s) being studied, and the quality and availability of value set(s). A report by the DSU summarises the psychometric performance of several preference-based measures (Rowen et al. 2020).

c. Report whether measure(s) of health-related quality of life were completed by adults with the condition, children and young people themselves, or proxies (for example, parents, carers or clinicians answering on behalf of a child). For self- and proxy reporting, report the age of the children and young people. If multiple data sources are available, report which data were used in the economic model and the rationale behind this choice.

2. The glossary should state: Psychometric performance refers to how well a questionnaire measures what it intends to measure. Aspects of psychometric performance include validity, reliability, responsiveness, acceptability and feasibility.

3. The methods guide already states that proxy reporting should be by carers rather than professionals; we propose that this statement is not altered.

Risk assessment

The proposals are low risk. They affect a small number of appraisals and evaluations. Current methods are variable and poorly reported.

The current methods guide could be interpreted as an implied endorsement of EQ-5D-Y. Our proposal is that the new methods guide would not refer to EQ-5D-Y (or any other measure) – this may be interpreted as a step back from endorsing EQ-5D-Y. To manage that risk, we propose liaising directly with EuroQol before consultation to explain our proposals and give them the chance to comment.

Alternative options

NICE’s future research will largely focus on 4 paediatric measures: CHU9D, EQ-5D-Y, Health Utilities Index 2 (HUI2) and HUI3 (summarised in table 1). These were chosen because they are intended to be used with children and young people, are generic, have some evidence of acceptable psychometric performance, have value sets or these are in development, are short enough to be used routinely in trials, and are widely used. The methods guide could recommend that companies choose 1 of these 4 measures. However, the evidence base does not yet support including a list of measures in the methods guide. Before being confident in such a recommendation, we need further research to:

a. Examine the content validity of these measures (that is, the extent to which they comprehensively cover the different dimensions of health and are sufficiently sensitive to changes; Brazier et al. 2017).

b. Compare their psychometric performance in large head-to-head studies.

c. Understand the impact of different methodological approaches used to develop value sets (for example, whether children or adults perform valuation tasks) so normative judgements can be made with an appreciation of their likely consequences. Move towards defining best-practice methods for valuation.

d. Assess the quality of the available value set(s) and assess how the value set affects the psychometric performance of the measure.

e. Understand the differences that arise between adult and paediatric utility values for the same condition, and how these differences should be interpreted during decision making.

Moreover, this is an active field of research and further measures may become available in the next few years (such as a shorter version of the Paediatric Quality of Life Inventory [PedsQL] that is suitable for valuation). By including a list of measures in the methods guide now, we could discourage research into measures not on the list. Ongoing research and recommendations for future research are discussed in more detail in appendix 2.

NICE could state that we do not recommend applying the EQ-5D-3L value set to data gathered using the EQ-5D-Y questionnaire. However, this recommendation is clearly stated on the EuroQol website. It would be unusual to use the methods guide to list what not to do, and this practice does not appear to be widespread. The DSU review found that 2 out of 31 past technology appraisals or highly specialised technology evaluations applied the EQ-5D-3L value set to EQ-5D-Y; the EQ-5D-Y data were generated by mapping from a disease-specific questionnaire (Hill et al. 2019).

NICE could provide guidance about the age ranges for which we prefer proxy- rather than self-report. However, the recommended age ranges vary depending on the measure (table 1). Moreover, a child’s ability to self-report their health-related quality of life depends on their developmental age, not just chronological age. An ISPOR taskforce recommends that self-report is preferred for children aged 12 and above, but notes that age cut-offs should be based on the measure and tested in the target population (Matza et al. 2013).

NICE could commission a DSU technical support document on how to measure and value children’s health-related quality of life. We think this would be more useful once some of the research outlined in appendix 2 is completed.

Authors

Rosie Lovett, Sophie Cooper and Alan Lamb on behalf of the Health-related quality of life task and finish group

References

Brazier J, Ratcliffe J, Salomon JA, Tsuchiya A. (2017) Measuring and valuing health benefits for economic evaluation: second edition. Oxford: Oxford University Press

Hill H, Rowen D, Pennington D, Wong R, Wailoo A. A review of the methods used to estimate and model utility values in NICE technology appraisals for paediatric populations. NICE DSU report 2019

Matza LS, Patrick DL, Riley AW (2013). Pediatric patient-reported outcome instruments for research to support medical product labelling: report of the ISPOR PRO good research practices for the assessment of children and adolescents task force. Value in Health. 16, 461–479

Noyes J & Edwards RT (2011). EQ-5D for the assessment of health-related quality of life and resource allocation in children: a systematic methodological review. Value in Health. 14, 1117–1129

Rowen D, Keetharuth A, Poku E, Wong R, Pennington B, Wailoo A. A review of the psychometric performance of child and adolescent preference-based measures used to generate utility values for children. NICE DSU report 2020

van Reenen M, Janssen M, Oppe M, et al. (2014). EQ-5D-Y user guide version 1.0.

Table 1 Generic measures of health-related quality of life for children

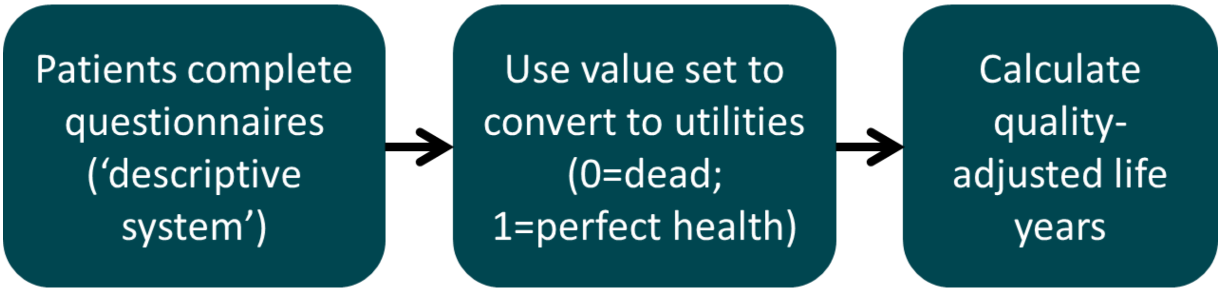
	Questionnaire (‘descriptive system’): Age range	Questionnaire (‘descriptive system’): Self or proxy report	Questionnaire (‘descriptive system’): Dimensions	UK value set: Study population	UK value set: Perspective	UK value set: Valuation method
CHU9D	Designed for 7-11, used for 4–18	4–6 years: proxy 7–18: self	9 (worry; sadness; pain; tiredness; annoyance; school; sleep; daily routine; activities)	UK adult general population	Self (adult)	Standard gamble (a new UK value set is in development using discrete choice experiment; adult perspective)
EQ-5D-Y	4–15	4–7: proxy 8–11: self 12-15: self, recommend EQ-5D-Y but can use adult EQ-5D	5 (mobility; looking after myself; doing usual activities; having pain or discomfort; feeling worried, sad or unhappy)	Not available (protocol in development)	Not available (protocol in development)	Not available (protocol in development)
HUI2	≥5	5–7: proxy 8–11: proxy (unless interviewer-administered) ≥12: self	6 (sensation; mobility; emotion; cognition; self-care; pain)	UK adult general population	Child aged 10	Standard gamble and visual analogue scale
HUI3	≥5	5–7: proxy 8–11: proxy (unless interviewer-administered) ≥12: self	8 (vision; hearing; speech; ambulation; dexterity; emotion; cognition; pain)	Canadian adult general population	Self (adult)	Standard gamble and visual analogue scale

Sources: Rowen et al. 2020 NICE DSU report, EQ-5D-Y user guide, HUI webpage.

Appendix 1: Key concepts and terminology

NICE uses cost–utility analyses to assess the value for money of interventions. These analyses require one to estimate not only how long people live, but also their **health-related quality of life**. Typically, this is done by asking patients to complete questionnaires about their health. A value set is then used to convert questionnaire responses into **utility scores**, on a scale where zero is equivalent to dead and one is perfect health. By multiplying utility scores and length of life, one calculates **quality-adjusted life years**.

Figure 1 Process of deriving quality-adjusted life years from questionnaires



Patients complete questionnaires (‘descriptive system’) → Use value set to convert to utilities (0=dead; 1=perfect health) → Calculate quality-adjusted life years

A **value set** is created by asking a large sample of people to express their preference for different hypothetical health states. NICE prefers valuation to be based on **public preferences** from a representative sample of the English population. This means that the utility scores that inform our recommendations represent the views of the public about which impairments to health matter the most. NICE does not define what constitutes a representative sample of the public, but in practice this has been interpreted as adults over the age of 18.

There are lots of ways of obtaining **preferences** (for example, time trade-off, standard gamble, discrete-choice experiment, visual analogue scale). The different methods each have advantages and disadvantages, and they give different results. NICE recommends a choice-based method, which excludes the visual analogue approach.

There are many generic preference-based measures of health-related quality of life (**generic** means they can be used for a wide range of diseases and conditions; **preference-based** means there is a value set so the measure can generate utility values). Each measure covers different aspects of health (known as **domains** or **dimensions**) and has been valued using different methods. As a result, different measures often give different utility scores for the same condition. Thus, for consistency, NICE prefers to use the EQ-5D for most evaluations for adults.

Rowen et al. (2020) note that, ideally, generic measures should show good **psychometric performance**, measured by:

- **Validity**: does the instrument measure what it claims to measure? Assessed by **known-group validity** (the ability to differentiate between groups of different disease severity); **convergent validity** (the strength of association with other measures of health-related quality of life or disease severity); **content validity** (does it comprehensively cover the different dimensions of health and is it sufficiently sensitive to changes; Brazier et al. 2017).
- **Reliability**: does the measure give consistent results over time when there has been no change in health? Assessed by whether the measure gives the same value on 2 separate administrations, this can be over time (**test-retest reliability**), between methods of administration (**inter-modal reliability**) or between raters, that is, self- versus parent report (**inter-rater reliability**).
- **Responsiveness**: does the measure capture change over time when change is expected, for example, before and after treatment?
- **Acceptability and feasibility**: are people willing to complete it, do they understand the questions, and are there high levels of missing data?

Appendix 2: Current state of knowledge and plans for future research

Current state of knowledge

Ideally, NICE's recommendations about how to measure health-related quality of life for children and young people would be informed by:

- Evidence of psychometric performance (that is, reliability, validity, responsiveness to change, feasibility of use and acceptability to users).
- A critical assessment of available value sets, to ensure they are relevant to the UK, generated using methods acceptable to NICE, and of acceptable quality.
- A thorough understanding of how the utility values generated by paediatric measures compare with those from adults.

The current state of knowledge in each of these areas is summarised below.

Evidence of psychometric performance

The Decision Support Unit (DSU) has reviewed the psychometric evidence for paediatric generic measures of health-related quality of life (Rowen et al. 2020). Broadly, for the 4 measures listed in table 1, there is evidence of known-group validity, convergent validity, acceptability and feasibility (albeit this is only for the dimensions of EQ-5D-Y because there is no value set, and the evidence is somewhat mixed for Health Utilities Index 3 [HUI3]). There are few studies of responsiveness and reliability and the evidence regarding these aspects is inconclusive. Overall, the evidence is hard to synthesise because the studies used different methods and most examined 1 measure in isolation (rather than comparing 2 or more measures in a head-to-head comparison). The DSU review did not examine content validity.

Valuation

There is not yet a consensus around best-practice methods for valuing paediatric instruments. The key methods choices include:

- Population: should instruments be valued by adults or young people (to understand their perspective about which aspects of health matter the most)?
- Perspective: should the valuation task ask people to think about their own health, or that of someone else (perhaps a child, and if so what age)?
- Valuation method: options include time trade-off, standard gamble, discrete-choice experiment with or without duration, visual analogue scale and best-worst scaling. Note that some techniques such as time trade-off involve thinking about early death, and therefore are not typically used with children and young people. Thus, many valuation studies with children and young people also need

further research with adults, to ‘anchor’ the young people’s preferences onto a scale where zero equals death.

Crucially, the choice of population, perspective and method has an impact on the resulting utility values. But the reasons for these differences are, in some cases, unclear and are the subject of ongoing research. Some of these choices are also based on social values rather than science – especially the choice of population.

Among the measures in table 1, only CHU9D and HUI2 have UK value sets. HUI3 has a Canadian value set. The value sets differ in their population, perspective and valuation method (table 1). EQ-5D-Y does not have a UK value set. The EuroQol group has developed a preliminary valuation protocol, involving a sample of adults taking the perspective of a 10-year-old hypothetical child and using the time trade-off technique and discrete-choice experiments (Stolk 2019). For reference, the adult EQ-5D-3L value set for England was created using the time trade-off technique in a sample of adults who took the perspective of their own health (Dolan, 1997).

Consistency with adult values

At an internal workshop in November 2019, we discussed how to address the potential for utility values to differ between adults and children for the same health state. Some NICE staff would prefer paediatric measures to produce utility values that are consistent with those for adults. Others felt that one would not necessarily expect consistency, because diseases and treatments may impact differently on adults and children. Moreover, the wording of descriptive systems is different for children and the valuation methods may differ.

The potential inconsistencies between adult and child utilities, and their impact, will need to be better understood before paediatric measure(s) and value set(s) can be recommended by NICE.

Ongoing and planned research

We are liaising with academics and the DSU to pursue further research into the psychometric performance of the measures listed in table 1. The priority research questions are:

- Examine the content validity of these measures (that is, the extent to which they comprehensively cover the different dimensions of health and are sufficiently sensitive to changes; Brazier et al. 2017). This work should also examine the content validity of the adult EQ-5D for children aged 12 and above.
- Compare psychometric performance in large head-to-head studies with a focus on responsiveness and test-retest reliability.

The EuroQol group is funding several research projects that examine how the choice of population, perspective and valuation method impact on valuation results – and, crucially, why these differences arise.

The Australian National Health and Medical Research Council (NHMRC) recently issued a call for research applications on ‘tools to value health change in paediatric populations’. The call was developed with input from the Pharmaceutical Benefits Advisory Committee (PBAC). NICE has been invited to join the advisory group for 2 applications under this call. These are substantial research projects that, if funded, will address many of the outstanding research and policy issues.

We are in touch with international health technology assessment agencies to explore whether it is feasible to work together to produce consistent methods guidance in this area.

We are also exploring whether it would be helpful to gain input from the public regarding the social values underlying some of the methodological choices.

References

Dolan P (1997) Modelling Valuations for EuroQol Health States. Medical Care 35(11) 1095–1108

Stolk E (2019). Valuing health in children – where are we now, and what further work is needed? Presentation at ISPOR Europe, Copenhagen

Rowen D, Keetharuth A, Poku E, Wong R, Pennington B, Wailoo A. A review of the psychometric performance of child and adolescent preference-based measures used to generate utility values for children. NICE DSU report 2020

Report 6: Mapping between EQ-5D 3L and 5L

Overview of proposed changes to methods

There are 2 versions of the EQ-5D questionnaire for measuring health-related quality of life, the 3L version and 5L version.

The NICE position statement on use of the EQ-5D-5L value set for England (updated October 2019) states that the 5L version of the questionnaire may be used to collect quality of life data but that the English value set for EQ-5D-5L should not be used. NICE instead recommends that the UK 3L value set is used.

We propose only 1 change to the methods recommended in the position statement. This relates to what companies should do if they have data or utility values collected using the 5L questionnaire. The current position statement recommends that data collected using the 5L questionnaire are mapped to the 3L value set using a mapping function developed by van Hout et al. (2012). Other methods for mapping between EQ-5D 3L and 5L have been developed by the Decision Support Unit (DSU; Hernández Alava et al. 2017). These have not previously been recommended by NICE, because we were unwilling to depart from the 2013 methods guide without public consultation. This paper proposes that the updated methods guide should recommend the DSU rather than the van Hout mapping tool.

The DSU mapping tool can be informed by different data sets. The most recent data set was collected by Hernández Alava and colleagues, hereafter referred to as the "EEPRU data set."[1] The EEPRU data set has several advantages, notably a wider coverage of health states because of its larger sample size (EEPRU, 2020). We propose that the updated methods guide should recommend the DSU mapping tool informed by the EEPRU data set.

We also propose stating that mapping from a disease specific measure to EQ-5D-3L is considered to be a departure from the reference case.

We propose that the new recommendations on mapping, and the remaining unchanged recommendations from the position statement, are incorporated into the updated methods guide for ease of reference. The position statement will then be removed from the NICE website. The relevant sections of the NICE guide to the methods of technology appraisal (2013) are presented in appendix 1 for information.

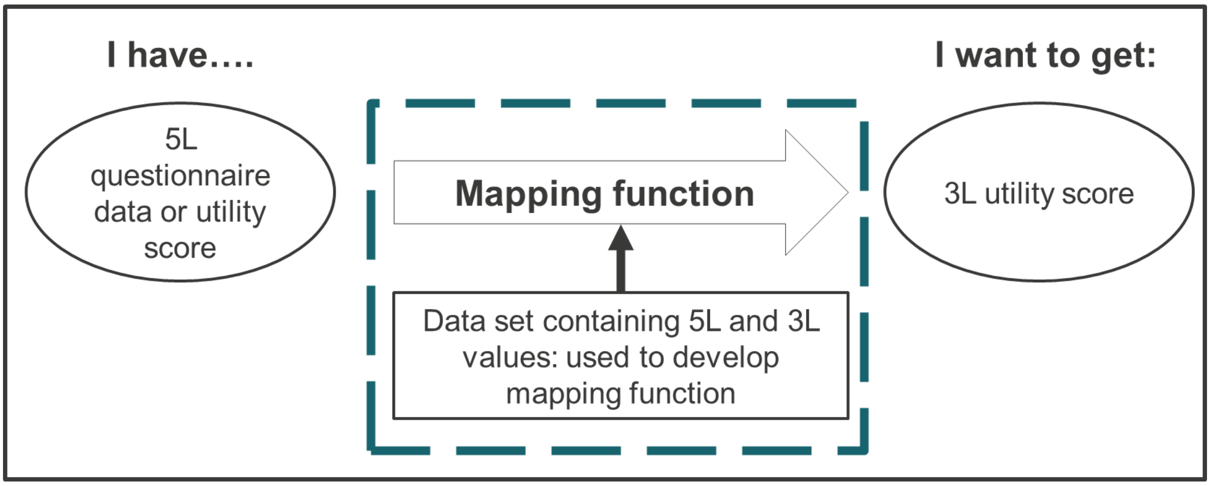
The technical issue

NICE recommends the collection of quality of life data using the EQ-5D-5L but recommends using the UK EQ-5D-3L value set. This means companies who have

[1] EEPRU is the Policy Research Unit in Economic Evaluation of Health & Social Care Interventions. Some members of EEPRU are also part of the DSU.

only collected EQ-5D-5L response data have to 'map' these to the equivalent EQ-5D-3L responses then convert to utilities using the EQ-5D-3L value set (see figure 1). The mapping tool is developed using a data set containing responses from people who have completed both the 3L and 5L questionnaires.

Figure 1 Overview of approach to mapping


I have….
5L questionnaire data or utility score
Mapping function
Data set containing 5L and 3L values: used to develop mapping function
I want to get:
3L utility score

There are several different methods of mapping available and several data sets that may be suitable for the development of mapping tools. The 3L utility score derived from the mapping process will vary depending on which are chosen, and this may affect the results of cost-effectiveness analyses, in some cases giving incremental cost-effectiveness ratios that differ by thousands of pounds (Pennington et al. 2018). Thus, it is important to recommend a single approach in the reference case in order to avoid gaming (that is, choosing whichever mapping approach or data set is most favourable to the technology being assessed). Decisions need to be made on the preferred:

- data set used to develop the mapping tool
- statistical method used to perform the mapping.

Considerations for choosing data set to inform mapping

2 data sets are available to inform mapping. The EuroQol group data set (n=3,691) includes 8 patient groups plus a healthy population and includes patients from the UK and several other European countries. The EEPRU data set (n=49,999) includes members of the public from the UK and did not collect data on any specific patient groups. The van Hout mapping tool uses the EuroQol group data set. The DSU tool can use either data set, we focus on the EEPRU data set because it was collected with a view to addressing some limitations of the EuroQoL group data set.

The key differences between the EEPRU and EuroQol data set are:

- the EEPRU data set has a larger sample size (49,999 compared with 3,691)
- the EEPRU data set covers more of the possible EQ-5D health states than the EuroQol data set (90% vs 51% for EQ-5D-3L and 43% vs 11% for EQ-5D-5L); this difference also holds for health states which were defined as ‘poor.’
- the number of responses is higher in the EEPRU data set across both 3L and 5L, in all domains and all response levels.
- the order in which the 3L and 5L questionnaires was given was randomised in the EEPRU data set, whereas in the EuroQol data set the 5L questionnaire was always given first. Randomisation of the questionnaires reduces the risk of ordering effects having an influence on the results (Hernández Alava and Pudney, in submission).
- Mapping assumes that responses to 3L and 5L are not affected by doing both questionnaires in the same survey. This assumption is more likely to be valid if the 2 questionnaires are separated by other questions rather than being given one after the other. The degree of separation of the 3L and 5L questionnaires was larger in the EEPRU data set.

One potential limitation of the EEPRU data set is that it was collected during the COVID-19 pandemic which may have affected the observed data in an unknown way. However, it seems unlikely this would affect the relationship between 5L and 3L responses. This assumption is supported by the conclusions that the mapping approaches using the EEPRU and EuroQol data set give broadly similar results.

Considerations for choosing mapping methods

Van Hout et al. (2012) and the DSU report (Hernández Alava et al. 2017) describe factors to be considered when choosing mapping methods, including:

- theoretical background
- statistical fit (how closely the values predicted by the mapping model match the values observed in the data used to develop the model)
- predictive power (how accurately the mapping model predicts values for data outside the sample used to develop the model)
- model parsimony (using the simplest appropriate modelling approach to fit the data)
- the data set used to develop the mapping method
- the functionality of the mapping method (can it do what users need?).

The 2017 DSU report compares the 2 approaches using these metrics, and a recent EEPRU report updates that comparison by including the DSU tool informed by the

EEPRU data set (EEPRU, 2020). An unpublished report sent by van Hout to NICE also provides supporting information.

The available methods to map 5L to 3L differ in terms of their statistical approach to modelling, data sources used to inform the modelling and functionality. A comparison of the mapping methods is presented in appendix 2.

Recommendations for updates to the methods guide

We recommend that:

- The new data set collected by EEPRU should be used to inform the mapping.
- The DSU method should be used to map EQ-5D-5L data or utility values to the UK EQ-5D-3L value set.

Rationale for recommendations to methods working group

Rationale for choice of data set

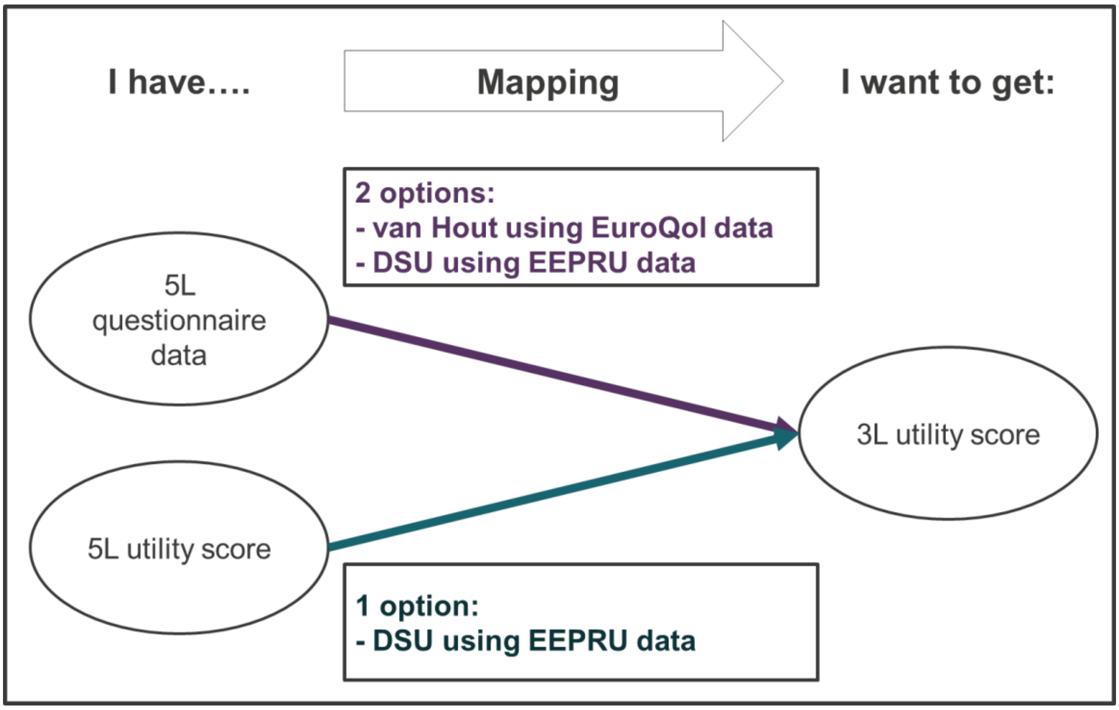
The 2 existing data sets (from the EuroQol group and EEPRU) are both suitable for deriving a mapping function. NICE technical staff note that the EEPRU study is larger, covers a greater proportion of health states and is designed to limit potential bias from the order in which the questionnaires are administered and their degree of separation. Therefore, the new EEPRU data set is preferred.

Rationale for choice of mapping method

The performance of the van Hout method (using the EuroQol data set) and DSU method (using the EEPRU data set) are similar in terms of statistical fit and predictive power. NICE technical staff advise that there is no strong reason to prefer one method over the other based on these metrics alone. The van Hout method may be preferred by some analysts as it is a simpler modelling approach. On the other hand, the DSU has argued that its method makes fewer assumptions, and thus is more data-driven, than van Hout. Of note, the van Hout approach assumes that responses at levels 1, 3 and 5 of the 5L response scale always lead to responses at levels 1, 2 and 3 of the 3L response scale. Analysis of the EuroQol and EEPRU data sets alongside 2 further out of sample data sets shows that this assumption doesn't always hold. The EEPRU report also shows that, when moving from the fourth to fifth response level on the 5L questionnaire, the van Hout mapping sometimes results in a very large utility decrement, much greater than is seen in the large EEPRU data set (for details see pages 25–26 and table 5 of the EEPRU report).

A key distinction between the van Hout and DSU methods is that the former requires individual 5L *questionnaire response data* to derive a utility value whereas the latter can map from a 5L *utility value* (see figure 2 below). While the approach of mapping directly from patient response data is preferred, these data may not always be available. While situations where this functionality is needed to map from 5L to 3L are likely to be rare, the need for mapping from literature values will increase in the future when mapping in the reverse direction (from 3L to 5L) will be required. Only the DSU method provides a way of directly mapping from literature values.

Figure 2 Mapping options from questionnaire data and utility scores



Taking all these points into consideration it is proposed that NICE recommends mapping EQ-5D-5L data or utility values to the UK EQ-5D-3L value set with the DSU method informed by the EEPRU data set.

Mapping from a disease-specific measure to EQ-5D-3L

Alternative methods for generating UK EQ-5D-3L values may be available. For example, a company may choose to collect data using a disease-specific outcome measure and map that to EQ-5D-3L instead of collecting EQ-5D-5L. A report by the DSU (Hernández Alava et al. 2019) concludes that if there is no convincing evidence of superior performance of a disease-specific mapping method then ‘generic’ methods should be used (that is, collect EQ-5D-5L and map to EQ-5D-3L). The DSU report includes an assessment of a disease-specific mapping tool, and thus provides a guide to the kind of evidence required. NICE technical staff note that using a ‘generic’ mapping method in the reference case is aligned with NICE’s general preference for using a generic measure of health-related quality of life. We consider that the issue of disease-specific mapping to EQ-5D-3L is too technical to discuss in the methods guide; it is covered by the DSU report. Therefore, we propose adding a single sentence to the methods guide: ‘Mapping from a disease-specific measure to EQ-5D-3L is considered to be a departure from the reference case; see the DSU report’. The current methods guide already contains recommendations for mapping when EQ-5D data are not available.

Quality assurance

The EEPRU report has been peer-reviewed and a manuscript detailing the work on the data set is in preparation and will be submitted to a peer-reviewed journal.

The DSU mapping method has been published in a peer-reviewed journal and the DSU report was peer-reviewed; one of the latter reviewers examined the model code.

The EEPRU report shows that the DSU method using the EEPRU data set performs similarly to the van Hout method using the EuroQol data set across 2 out of sample data sets. The similar performance of these approaches provides reassurance about the robustness of the mapping code and underlying data.

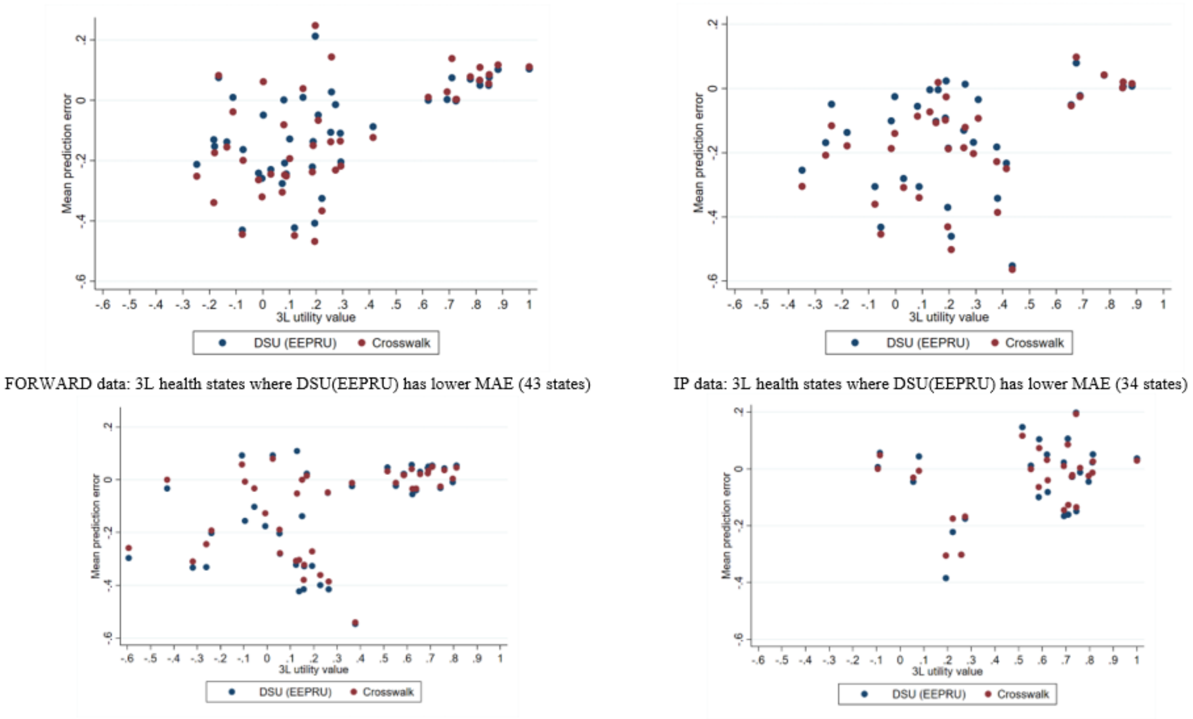
The NICE team concludes that the mapping method and data set have undergone appropriate quality assurance.

Impact of proposed changes

A comprehensive impact assessment is challenging because for each case study one would require access to both the EQ-5D-5L patient-level questionnaire responses and the economic model. This would be logistically challenging as it would require permission from data owners to access the response data and economic model. Additionally, not all economic models use utility data from an associated clinical trial to inform the health states in the model. This limits the number of potential case studies, making it challenging to obtain a fully representative set of case studies. Thus, commissioning a large impact assessment would be expensive, time-consuming, and might result in only a few case studies meaning that the results may not be generalisable to the range of health states seen in NICE appraisals. As a pragmatic approach we asked the DSU to explore the impact of a change in mapping approach conceptually.

For any set of 5L data, different mapping methods and mapping data sets will give somewhat different 3L utility scores. Figure 3 (below) shows that, overall, the van Hout and DSU models give fairly similar results. Figure 3 is based on 2 data sets where people completed both 3L and 5L, neither of these data sets had been used to inform the mapping models. For all 3L health states observed in the data, the DSU calculated the difference between the directly calculated 3L utility and the 3L utility calculated by mapping; an error of 0 means the 2 were the same. In about half the cases, the van Hout method gave smaller errors and in about half the DSU method had smaller errors. There was no section of the 3L utility scale where one mapping had clearly lower errors than the other.

Figure 3: Mean prediction errors in 3L utility scores in out of sample data



FORWARD data: 3L health states where DSU(EEPRU) has lower MAE (43 states)

IP data: 3L health states where DSU(EEPRU) has lower MAE (34 states)

FORWARD data: 3L health states where Crosswalk has lower MAE (40 states)

IP data: 3L health states where Crosswalk has lower MAE (27 states)

On average, the differences between the van Hout and DSU mapped 3L utility values tend to be quite small (for details see page 24 and figure 7 of the EEPRU report). However, there are some scenarios where the van Hout and DSU mapped 3L utility values are substantially different, for example very poor health states where data are sparse (see table 5 of the EEPRU report). Thus, there will be some economic models where the choice of mapping makes a difference to the resulting utility values and cost-effectiveness results.

In order to prevent ‘gaming’ we propose that routine sensitivity analyses using alternative methods would not be recommended in the methods guide; the EEPRU data set and DSU mapping method will be the reference case. In analogy to deviating from the reference case to use an alternative HRQoL measure to EQ-5D, the use of a non-reference case mapping method should be justified using empirical evidence that the reference-case mapping method is not appropriate.

Future considerations

A second UK valuation for EQ-5D-5L is expected to begin in 2021. Once complete, NICE will review its methods guidance on EQ-5D valuation and mapping.

Authors

Rosie Lovett and Alan Lamb on behalf of the Health-related quality of life task and finish group

References

van Hout B, Janssen MF, Peng Y-S, et al. (2012) Interim scoring for the EQ-5D-5L: mapping the EQ-5D-5L to EQ-5D-3L value sets. Value in Health 15(5): 708–715

Hernández Alava M, Wailoo A & Pudney S (2017) Report by the Decision Support Unit: Methods for mapping between the EQ-5D-5L and the 3L for technology appraisal

Hernández Alava, M. Pudney, S, & Wailoo, A. (2020) Report by the Policy Research Unit in Economic Evaluation of Health and Care Interventions (EEPRU). Estimating the relationship between EQ-5D-5L and EQ-5D-3L: results from an English Population Study.

Pennington B, Hernández Alava M, Pudney S & Wailoo A (2018) Report by the Decision Support Unit: Comparing the EQ-5D-3L and 5L versions. What are the implications for model-based cost effectiveness estimates?

Hernández Alava M & Pudney, S. (in submission) Mapping between EQ-5D-3L and EQ-5D-5L. A survey experiment on the validity of multi-instrument data.

Hernández Alava M, Chrysanthou G & Wailoo A (2019) Report by the Decision Support Unit: Disease specific versus generic mapping methods: How to link outcomes to EQ-5D

Appendix 1: Current wording of methods guide

5.3 Measuring and valuing health effects

5.3.6 The EQ-5D is a standardised and validated generic instrument that is widely used and has been validated in many patient populations. The EQ-5D comprises 5 dimensions of health: mobility, ability to self-care, ability to undertake usual activities, pain and discomfort, and anxiety and depression. For each of these dimensions it has 3 levels of severity (no problems, some problems, severe problems). The system has been designed so that people can describe their own health-related quality of life using a standardised descriptive system. A set of preference values elicited from a large UK population study using a choice-based method of valuation (the time trade-off method) is available for the EQ-5D health state descriptions. This set of values should be applied to measurements of health-related quality of life to generate health-related utility values.

[…]

5.3.12 A new version of the EQ-5D, the EQ-5D-5L, has been developed in which there are 5 levels of severity (no problem, slight problems, moderate problems, severe problems and unable to or extreme problems) for each of the 5 dimensions of health (see section 5.3.6). The EQ-5D-5L may be used for reference-case analyses. The descriptive system for the EQ-5D-5L has been validated, but no valuation set to derive utilities currently exists. Until an acceptable valuation set for the EQ-5D-5L is available, the validated mapping function to derive utility values for the EQ-5D-5L from the existing EQ-5D (-3L) may be used (available from www.euroqol.org).

In August 2017, NICE issued a position statement on the use of the EQ-5D-5L valuation set. Companies and academic groups should refer to this statement.

Appendix 2: Comparison of mapping approaches

Table 1 Comparison of mapping approaches of van Hout and Decision Support Unit

-	van Hout	DSU
Method for mapping 5L tow 3L	Non parametric calculations based on frequencies obtained when cross-tabulating responses to 3L and 5L questionnaires were used to generate transition probabilities for being in each 3L health state based on response to 5L.	Ordinal regressions with a flexible residual distribution specified as a copula mixture
Data source	EuroQol group	EEPRU
Type of tool	Excel tool available on EuroQol website	Stata command available, see details on DSU website. Excel tool has been developed and will be made available on DSU website.
Possible to map directly from utility scores?	No. Response data to questionnaire required	Yes. It is recommended that this is done only when response data to questionnaire are not available.