MEASURING AND VALUING HEALTH-RELATED QUALITY OF LIFE WHEN SUFFICIENT EQ-5D DATA IS NOT AVAILABLE

REPORT BY THE DECISION SUPPORT UNIT

31st July 2020

Donna Rowen, John Brazier, Ruth Wong, Allan Wailoo

School of Health and Related Research, University of Sheffield

Decision Support Unit, ScHARR, University of Sheffield, Regent Court, 30 Regent Street
Sheffield, S1 4DA

Tel (+44) (0)114 222 0734
E-mail dsuadmin@sheffield.ac.uk
Website www.nicedsu.org.uk
Twitter @NICE_DSU

ABOUT THE DECISION SUPPORT UNIT

The Decision Support Unit (DSU) External Assessment Centre is based at the University of Sheffield with members at York, Bristol, Leicester and the London School of Hygiene and Tropical Medicine. The DSU is commissioned by The National Institute for Health and Care Excellence (NICE) to provide a research and training resource to support the Institute's Centre for Health Technology Evaluation Programmes. Please see our website for further information www.nicedsu.org.uk.

The production of this document was funded by the National Institute for Health and Care Excellence (NICE) through its Decision Support Unit. The views, and any errors or omissions, expressed in this document are of the authors only. NICE may take account of part or all of this document if it considers it appropriate, but it is not bound to do so.

Acknowledgements

We have no acknowledgements at this stage.

This report should be referenced as follows:

Rowen D, Brazier J, Wong R, Wailoo A. Measuring and valuing health-related quality of life when sufficient EQ-5D data is not available. NICE DSU Report. 2020.

EXECUTIVE SUMMARY

The National Institute for Health and Care Excellence (NICE) provides recommendations as part of its methods guides around measuring, valuing and sourcing information about health related quality of life (HRQoL) benefits in adults. These recommendations are for use in guidance producing programmes that require the valuation of HRQoL benefits within the Centre for Health Technology Evaluation (CHTE). This includes TA (technology appraisal), and HST (highly specialised technology) evaluations and the diagnostics appraisal program. The use of the EQ-5D is preferred across all evaluations. EQ-5D is a generic preference-based measure with a simple classification system for respondents to describe their health, and a value set elicited from members of the UK general population to generate utility values for all health states. The preferred use of EQ-5D raises questions around what should be done when EQ-5D data is unavailable, when EQ-5D is inappropriate, or when EQ-5D data is insufficient.

The NICE Methods Guide for TA provides recommendations around what should be done in situations where EQ-5D data, collected directly from patients within the relevant clinical studies of the health technology being assessed is unavailable. It states that EQ-5D utility values should be either: sourced from a literature review using a systematic search of the literature; or generated using statistical mapping applied to relevant data; or generated by conducting a separate study to collect EQ-5D data to populate the economic model.

NICE recommends that in order to determine whether EQ-5D is inappropriate empirical evidence is required demonstrating poor performance regarding content validity, construct validity or responsiveness. A NICE Technical Support Document makes recommendations around what should be done when it can be demonstrated using evidence that EQ-5D is inappropriate. Other generic or condition-specific preference-based measures, can be used where EQ-5D is demonstrably inappropriate.

However, there are situations where EQ-5D may be considered an appropriate outcome measure but available EQ-5D data are insufficient for generating values for

all health states required in the economic model of the TA or HST evaluation. For example, this could be due to rare health states or events, particularly where patient populations are small. This creates challenges for generating QALYs (quality adjusted life years) to assess the cost-effectiveness of the technology. In rare conditions, the wider literature on HRQoL benefits in the relevant patient population may also be sparse. Mapping studies may also be infeasible because no appropriate dataset exists that can be used to generate mapped estimates of utilities. In some cases, for example, where the number of patients with the condition in question are very small, the prospects of conducting a study to collect EQ-5D data in the short term are limited. However, if these approaches are not viable, there are no clear recommendations in the current NICE methods guide around preferred alternatives.

This report aims:

- To provide guidance on the alternative methods for measuring and valuing HRQoL in situations where available EQ-5D data from clinical studies are insufficient for the requirements of the economic evaluation.
- To make recommendations about the circumstances when each method may be used and suggest a hierarchy of methods where more than one option is feasible and when such recommendations are defensible.
- To help inform NICE's future considerations about recommendations for generating utilities where sufficient evidence is not available to estimate EQ-5D data.

The report focuses upon TA and HST evaluations, and the TA methods guide since this is the most detailed, but all methods and discussion are equally applicable to the diagnostics appraisal program (DAP). The methods discussed are not applicable for the medical technologies evaluation program (MTEP), which uses cost-minimisation analyses rather than cost-utility analyses.

The study objectives are:

1. To identify examples of methods and approaches used in previous TA and HST evaluations to obtain HRQoL estimates when insufficient EQ-5D data was available;
2. To critically examine each of the potential alternative methods, including points for consideration in their study development and interpretation of results;

3. To provide best practice recommendations around what should be done when sufficient EQ-5D data is not available.

The report is limited by the paucity of academic literature on this topic, and the report reflects both the limited literature that is available and the authors' opinions.

Alternative methods used in TA and HST evaluations where sufficient EQ-5D evidence is not available

The alternative methods for which examples are reported are as follows:

1) Vignettes, which are bespoke descriptions of an impaired health state. The content and format of vignettes varies widely, for example from a relatively brief outline to a detailed lengthy description, or from describing more condition-specific and symptom-specific elements to also incorporating broader domains of HRQoL. The valuation of vignettes can vary substantially, and these can be summarised under three approaches:
    a) EQ-5D estimates generated by some relevant group (e.g. clinical experts, patients, general population) completing the EQ-5D classification system for each vignette;
    b) preference elicitation techniques such as time trade-off (TTO) typically with a sample of the general population;
    c) elicitation with clinical experts, for example using Delphi panels or preference elicitation methods such as TTO;
2) "Proxy condition" utility values, involving the use of utility values from another condition as a proxy for the condition in question. This involves the use of utility values for another condition that is thought to have a similar impact on health and quality of life in terms of both the dimensions of health that are impacted and the degree of impairment.

Summary of proposed recommendations, for consideration by NICE

The following points are the key report recommendations:

- The use of EQ-5D directly administered to patients and scored using general population preferences is the preferred option to generate utility values. The

use of any other method where insufficient evidence cannot be observed remains a second-best alternative, as failure to develop a sufficient body of evidence using EQ-5D, where it would have been desirable and feasible leads to unnecessary uncertainty and incomparability to other appraisals. Evidence should be provided demonstrating why the reference case method of EQ-5D has not been used. Where evidence demonstrates that it was not possible to directly administer EQ-5D to patients, the 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 to populate the economic model.

- Where EQ-5D is demonstrated using empirical evidence to be inappropriate, it is recommended to generate evidence using either an alternative generic preference-based measure or a condition-specific preference-based measure. If this is not possible, patient valuations of their own health state can be used. Where EQ-5D is inappropriate, this should be evidenced using psychometric evidence. This could include evidence around content validity, construct validity, responsiveness and reliability. In addition, the development of the measure should be fully described, including health state descriptions and health state values. Evidence should be provided on the alternative measure, including the psychometric properties of the measure and the impact on utility values due to the use of the alternative measure instead of EQ-5D.
- Where sufficient EQ-5D evidence is not available, what is important is the appropriateness of the methods and evidence used to generate the utility values, the appropriateness of the utility values used in the economic model, and their comparability to EQ-5D. The appropriateness and selection of either vignettes or “proxy condition” utility values will vary depending on the condition and economic model.
- The reason why EQ-5D is unavailable or insufficient is important. There is a distinction between unavailable or insufficient data due to poor planning or failure to include EQ-5D in clinical studies, and unavailable or insufficient data due to characteristics of the patient population and/or states required to populate the economic model. It is understandable if appropriate attempts have been made to use EQ-5D data but this is not fully achievable due to the patient population and/or states required in the economic model. Poor planning or

failure to include EQ-5D in clinical studies where EQ-5D is appropriate is unjustifiable. In addition relevant evidence can be obtained from a range of different study types, and is not limited to clinical studies. Early consideration of the evidence requirements can avoid the need to use alternative methods to EQ-5D.

- If EQ-5D data are unavailable or insufficient clear evidence is required demonstrating why it was not possible to use self-report EQ-5D data. The evidence should demonstrate that it was not possible to: directly administer EQ-5D to patients as part of the clinical studies of the technology in question; source utility values from the literature; estimate EQ-5D utilities using mapping; or conduct a separate study to collect EQ-5D data. The use of other methods to compensate for inadequate evidence generation introduces unnecessary uncertainty into the appraisal of health technologies.
- The utility values should be generated using appropriate methods, free from bias and as comparable as they can be to the reference case method of EQ-5D. The studies underpinning the evidence should be well-conducted, transparent, and clearly described with all study advantages and limitations highlighted. The studies underpinning the evidence should clearly acknowledge any limitations to enable the ERG, NICE committee members and wider audience to interpret the evidence.

In those situations where vignettes are used, we make the following recommendations:

- Obtain high quality appropriate and reliable evidence to inform vignette development. This could consist of, and be strengthened by the use of, multiple different types of evidence including published literature, qualitative evidence and quantitative HRQoL data.
- Vignettes should be developed to meet the requirements of the economic model structure for the TA and HST evaluations; should be formatted to enable easy reading, comprehension and understanding; should include generic aspects of HRQoL; should include all important and relevant aspects of HRQoL; be worded with minimal potential for both ambiguity and misinterpretation; reflect the typical patient experience for the disease state in question; should avoid probabilistic statements; should not include disease or treatment labels

where possible; and should not use value-laden or irrelevant phrases or content (such as ‘devastating’).

- Vignette refinement and validation should be undertaken using input from clinical experts and/or patients to ensure that the vignettes are a clear and accurate description of the disease state or adverse event that they are intended to represent in the economic model.
- The reporting of the process of vignette development needs to be full and transparent.
- Utility values for vignettes are generated using an appropriate sample of patients completing the EQ-5D for each vignette, and this is then scored using the appropriate and relevant value set for EQ-5D, provided EQ-5D is appropriate.

1. CONTENTS

TABLE OF FIGURES

ABBREVIATIONS AND DEFINITIONS

AQoL-6D	Assessment of Quality of Life- 6 Dimensions
CUA	Cost-utility analysis
DAP	Diagnostics appraisal program
DCE	Discrete Choice Experiment
EQ-5D	EuroQol- 5 Dimensions
ERG	Evidence Review Group
HRQoL	Health-related quality of life
HST	Highly Specialised Technology
HUI	Health Utilities Index
HUI2	Health Utilities Index Mark II
HUI3	Health Utilities Index Mark III
DS	Dravet Syndrome
LGS	Lennox-Gastaut syndrome
MTEP	Medical technologies evaluation program
NICE	National Institute for Health and Care Excellence
QALY	Quality-adjusted life year
ScHARR	School of Health and Related Research
SG	Standard gamble
TA	Technology Appraisal
TSD	Technical support document
TTO	Time trade-off
VAS	Visual Analogue Scale

2. INTRODUCTION

2.1. BACKGROUND

The National Institute for Health and Care Excellence (NICE) provides clear recommendations around measuring, valuing and sourcing HRQoL benefits in adults for use in TA and HST evaluations considered by NICE. The use of the EQ-5D is preferred across all evaluations.

The National Institute for Health and Care Excellence (NICE) provides recommendations as part of its methods guides around measuring, valuing and sourcing information about HRQoL benefits in adults. These recommendations are for use in guidance producing programmes that require the valuation of HRQoL benefits within the Centre for Health Technology Evaluation (CHTE). This includes TA (technology appraisal), and HST (highly specialised technology) evaluations and the diagnostics appraisal program. The use of the EQ-5D is preferred across all evaluations[1-3]. However, this raises questions around what should be done when EQ-5D data is unavailable, when EQ-5D is inappropriate, and when EQ-5D data is insufficient.

NICE provide clear recommendations around what should be done in situations where directly observed EQ-5D data from relevant studies is unavailable [1]. NICE recommends that appropriate utility values are either: sourced from a literature review using a systematic search of the literature (see for example [4]); or generated using statistical mapping applied to relevant data (see for example [5-8]); or generated by conducting a study to collect EQ-5D data to populate the economic model.

There are guidelines provided by NICE around how to determine whether EQ-5D is inappropriate, and a NICE Technical Support Document makes recommendations around what should be done when it can be demonstrated using evidence that EQ-5D is inappropriate [1,9]. Alternative measures to EQ-5D, for example other generic or condition-specific preference-based measures, can be used where EQ-5D is demonstrated as not being appropriate using evidence of the psychometric performance of EQ-5D in the relevant patient population [1,9]. For example, it could

be shown that EQ-5D does not have good construct validity in the relevant patient population, meaning that EQ-5D does not capture known group differences for the condition. In these circumstances, a measure that does have good psychometric performance in the relevant patient population can be used instead.

However, there are situations where EQ-5D may be considered an appropriate outcome measure but available EQ-5D data are insufficient for generating values for all health states required in the economic model of the TA or HST evaluation. For example, this could be due to rare health states or events, particularly where patient populations are small. This creates challenges for generating QALYs (quality adjusted life years) to assess the cost-effectiveness of the technology. In rare conditions, the wider literature on HRQoL benefits in the relevant patient population may also be sparse. Mapping studies may also be infeasible because no appropriate dataset exists that can be used to generate mapped estimates of utilities. In some cases, for example, where the number of patients with the condition in question are very small, the prospects of conducting a study to collect EQ-5D data in the short term are limited. However, if these approaches are not viable, there are no clear recommendations in the current NICE methods guide around preferred alternatives.

2.2. AIMS AND OBJECTIVES

This report aims:

- To provide guidance on the alternative methods for measuring and valuing HRQoL in situations where available EQ-5D data from clinical studies are insufficient for the requirements of the economic evaluation.
- To make recommendations about the circumstances when each method may be used and suggest a hierarchy of methods where more than one option is feasible and when such recommendations are defensible.
- To help inform NICE's future considerations about recommendations for generating utilities where sufficient evidence is not available to estimate EQ-5D data.

The report focuses upon TA and HST evaluations, and the TA methods guide since this is the most detailed, but all methods and discussion are equally applicable to the

diagnostics appraisal program (DAP). The methods discussed are not applicable for the medical technologies evaluation program (MTEP), which uses cost-minimisation analyses rather than cost-utility analyses.

The study objectives are:

1. To identify examples of methods and approaches used in previous TA and HST evaluations to obtain HRQoL estimates when insufficient EQ-5D data was available;
2. To critically examine each of the potential alternative methods, including points for consideration in their study development and interpretation of results;
3. To provide best practice recommendations around what should be done when sufficient EQ-5D data is not available.

3. ALTERNATIVE METHODS USED IN PREVIOUS TECHNOLOGY APPRAISALS AND HIGHLY SPECIALISED TECHNOLOGY EVALUATIONS TO OBTAIN HRQOL ESTIMATES WHERE SUFFICIENT EQ-5D DATA WAS NOT AVAILABLE

Examples are reported for different methods in each section below. The examples were selected with input from NICE. The chosen examples in each approach are useful for showing the various types of studies and evidence collected to inform appraisals, and are not intended to be exhaustive (see Appendix Table A1 for an overview of the included examples).

3.1. VIGNETTES

Vignettes are bespoke descriptions of an impaired health state. These descriptions are not based on or defined using a health state classification system for a preference-based measure, such as EQ-5D or SF-6D and can contain aspects of health and quality of life covering functioning, symptoms and clinical aspects of the condition. The content and format of vignettes varies widely, for example from a relatively brief outline to a detailed lengthy description to simulation of symptoms (for example, through use of contact lenses) or video demonstrations. Vignettes can vary from describing more condition-specific and symptom-specific elements to also incorporating broader domains of HRQoL. The process for the generation of vignettes can also vary in terms of the source(s) of evidence used to inform their development. This can include published literature, qualitative research with patients, qualitative research with clinical experts, social media data (though concerns may be raised with this, see Section 3.1 below) and quantitative data of patient HRQoL. Often input from both patients and clinical experts is used to refine and validate vignette descriptions.

3.1.1. Valuation of vignettes

Vignettes can be valued using a range of different techniques: clinical experts (or general population or patients) proxy-completing EQ-5D for the vignette health state; valuation with members of the general public or patients using preference elicitation techniques such as time trade-off (TTO); or eliciting values from clinical experts (for

example using a Delphi method). Importantly, the development of the vignettes can be influenced by the method that will be later used to generate the utility values. For example, it may not be appropriate to include clinical information in vignettes valued by members of the general population, but clinical information could be appropriate for inclusion in vignettes valued by clinical experts.

3.1.1.1. Direct valuation of vignettes by members of the general public or patients

Vignettes can be valued directly using an accepted preference elicitation technique (such as TTO), which generates a utility value for the impaired health state. The valuation can be undertaken by samples of the general public, patients, carers, or clinicians. The elicitation of values from members of the general population provides greatest alignment with the methods used for EQ-5D.

One example of the use of vignettes valued using this method is for short bowel syndrome [10]. A vignette study was undertaken using TTO with a UK general population sample. The study aimed to identify the utility values associated with a specific number of days per week of parenteral support. Eight vignettes were developed using a targeted literature review, interviews with patients (n=12) and clinical experts (n=4), and then health states were piloted using cognitive debriefing with members of the UK general population (n=5) [11]. The vignettes varied parenteral support from 0 to 8 days, and each vignette also described the condition, symptoms, treatments and impact on EQ-5D dimensions using bespoke wording focussed on the condition. The vignettes were valued by a convenience sample of 100 members of the UK general population using TTO and visual analogue scales (VAS). In this instance, it appears that vignettes were not used to provide evidence where there was insufficient evidence available using EQ-5D, since in the company's submission utility values were also presented using trial data with EQ-5D scored using the US tariff, and via mapping SF-36 onto EQ-5D. Vignettes showed larger utility decrements for more severe health states than the other methods.

Two TAs for Cannabidiol for different conditions (Dravet syndrome and Lennox-Gastaut syndrome) involved the use of vignettes [12-13]. The vignette study for Dravet syndrome (DS) involved the development of 23 health states for patients and 3 health

states for carers [12]. The patient health states were developed to reflect the range of seizure severity using both number of convulsive seizures per month and seizure-free days per month, with additional health states reflecting mid-points in seizure severity. The carer states were developed to reflect severity of convulsive seizures with a moderately severe health state, less severe health state and a convulsive seizure-free state. The vignettes were developed using clinical and demographic information from trial data, and the descriptions included age (11 years), number of convulsive seizures per month, number of previous and current treatments, and current health. Pilot testing was undertaken where age was defined as either 11 or 15 years of age. Pilot participants reported that their responses would not differ by age, and therefore the main study only reported age as 11 years. The patient was named (David, male) and the carer was also named (Ally, female), and participants were asked to imagine they were the patient.

The valuation of the vignettes involved an online study with epilepsy patients (DS and Lennox-Gastaut syndrome (LGS) and other epilepsy patients) and carers (of patients with DS or epilepsy) using VAS. The rationale for eliciting the values for health states from patients rather than the general population was that patients with epilepsy and their carers have a better understanding of the impact of both seizures and seizure-free days on quality of life and wellbeing. Participants were recruited via a UK patient association, and the number of participants recruited is not reported in the publicly available NICE documents. Respondents scored all health states using VAS, where 0 was defined as worst health and 100 was defined as best health. The 23 health states were used to generate values for 9 health states in the model using averages. In sensitivity analyses, the VAS values were mapped to standard gamble values using a published algorithm [14].

The vignette study for LGS used similar methodology [13], with a different patient name (Ben), the different condition specified (LGS) and included a total of 39 vignettes. Seizures were focussed on the number of days without drop seizures and the number of seizure-free days. The vignette LGS included contextual information including that Ben has a “rare form of epilepsy that is difficult to treat” and “had LGS … from a young age”. In contrast the DS vignette included information that David had a “rare form of epilepsy” and “had DS … from early infancy”. For the vignettes in both

studies, some visual information is used to differentiate between the vignettes in terms of the number of seizures and seizure-free days.

Another example of the use of vignettes is for a submission to NICE around untreated acute myeloid leukaemia, where 200 members of the UK general population (n=193 in useable data) valued vignettes using the time-trade-off elicitation technique in interviews [15]. The vignettes were developed using discussions with clinical experts, and described symptoms (hair loss, risk of infection, fatigue and tiredness) and a detailed description of the treatment including setting, administration of the treatment and details of recovery following treatment. Each participant in the valuation study valued 12 vignettes.

3.1.1.2. Values elicited for vignettes by clinical experts

Vignettes, particularly where these may be more focussed on clinical stages of disease, can be directly valued by clinical experts using a range of techniques. This could be undertaken using an accepted preference elicitation technique such as TTO with clinical experts for different clinical stages of disease. This can differ from vignettes valued by members of the general population or patients as there is no requirement that the description of the health state focusses upon the impact on health and quality of life; rather a clinical definition can be provided and valued.

An alternative technique is the use of Delphi panels to achieve consensus in utility values elicited from different experts. This is achieved through the use of multiple rounds of questionnaires distributed to a group of experts to generate utility values for a series of disease stages or health states, and to agree upon these utility values. The utility values could be elicited using TTO completed by the experts, or potentially could be directly provided on a 1-0 scale where 1 represents full health and 0 represents death. One example of the use of Delphi panels to inform utility estimates in TA and HST evaluations is through their application to develop carer utility values for short bowel syndrome using a panel of 9 experts involved in treating patients with the condition [10]. The panel valued low, mid and high parenteral support requirements on the 1-0 full health-dead scale, and whilst the methods are unclear it is suggested that the experts directly generated the values on the 1-0 scale without the use of a

preference elicitation technique such as TTO. In the company submission carer utility values were generated as the mid-point between EQ-5D utility values from a survey of carers in the UK (n=47), and those generated using the Delphi panel. This approach combines utility values using self-report EQ-5D, with clinical expert judgement, and does not meet reference case guidelines due to the inclusion of clinical expert judgement.

3.1.1.3. Vignettes valued indirectly by completing the EQ-5D

This method involves the use of clinical experts to complete the EQ-5D health state classification system (or the classification system of an alternative preference-based measure) for vignettes representing different stages of disease or common clinical states experienced by what they would consider to be the typical patient. Clearly this task could also be undertaken by patients or members of the general population. The utilities for the EQ-5D health states can then be generated using the relevant value set.

For example, this approach was used in a submission to NICE TA for treating neuronal ceroid lipofuscinosis type 2, where bespoke vignettes were developed and clinical experts were asked to complete the EQ-5D-5L for each vignette as a patient-proxy [16]. Vignettes were generated for nine health states, with separate vignettes for the two different treatments in the economic model, generating 18 vignettes in total. Vignettes contained clinical information focussed on motor and language domains, treatment (and how administered) and symptoms, and were validated by a clinical expert. Clinical experts then completed the EQ-5D-5L as a patient-proxy for each of the vignettes, which were then mapped to EQ-5D-3L and scored using the UK EQ-5D-3L value set. Whilst EQ-5D-5L data was collected in trials, the reasons provided for why these were not used were that the sample size was small and would not provide utility values for all health states in the economic model.

Another example is where this approach was used to treat inherited retinal dystrophies caused by RPE65 gene mutations, where 6 clinical experts completed the EQ-5D-5L and HUI3 as a patient-proxy for bespoke vignettes for 5 health states [17]. The vignettes were developed involving an expert advisory board (n=12), patients and

carers (n=5), and interviews with clinicians. Utility values were generated for HUI3 and EQ-5D-5L using their value set and a crosswalk to EQ-5D-3L, respectively [18]. The Evidence Review Group (ERG) raised issues around: the small number of respondents used in the study; focussing effects where the clinical experts may focus on vision loss rather than considering the overall health of the patient; and ordering effects as assessing the best health state first may have led to a capping of utilities at the upper end of the scale. Further concerns were raised as the values did not match patient experience described by the ERG's clinical advisors, in particular where for the most severe state some degree of adaptation is expected. The ERG then used published values from the literature that involved members of the public completing TTO exercises for 8 health states [19].

Another example is in the treatment of spinal muscular atrophy (SMA) [20]. The base case utility estimates were derived using mapped EQ-5D from Peds-Qol data, using an existing mapping function [21]. However, the ERG recommended that an alternative study commissioned by the company should be used in the base case analyses, due to concerns around the face validity of the mapped EQ-5D estimates and the lack of similarity of the population that the mapping function was estimated in and applied to. This study involved the construction of bespoke vignettes to describe a typical child with SMA according to their HRQoL, symptoms and physical limitations. The vignettes were developed using interviews with clinical experts (n=5) and a literature review, where the vignettes were described as case studies rather than vignettes. Clinical experts (n=5) then completed the EQ-5D-Y as a patient-proxy for each of the vignettes, which were then scored using the UK EQ-5D-3L value set [22]. However concerns were raised by the ERG around the face validity of the estimates, and estimates were also generated 'based on clinical judgement' though these were not preference-based.

Another example is a submission to NICE TA for X-linked hypophosphataemia in children and young people [23]. Scenarios which required utility values for the model were described using vignettes, and clinical experts (n=6) in the condition reported the equivalent EQ-5D-5L health state for each of these vignettes. Utility values were then generated for these EQ-5D-5L health states using the crosswalk to EQ-5D-3L [18]. Twelve vignettes were developed using both published qualitative studies and

involvement from clinical experts (n=5) via interviews, and included functioning (walking, gait, bowed legs, usual activities, stature and stockiness, speed and strength, pain, sleep, tiredness, mental health, school work, relationships, respiratory functioning, oral health, history of fractures), symptoms and clinical information. The vignettes were described as case histories, and involved four different severities (healed, mild, moderate, severe) each with 3 different age groups (1-4 years, 5-12 years, 13 years and above). Two clinicians did not report EQ-5D-5L for the most severe state as they did not treat patients with states as severe. In this study, mean values were not used; instead, the moderate state was used as an anchor with the difference in utility from the moderate state generated. Some concerns were raised by the ERG around the healed health state values, as it could be argued that the descriptions of the healed states did not accurately depict some aspects around residual deformity and fracture risk, where they were described as less severe than would be expected in practice. This raises the concern that EQ-5D utilities generated for the healed vignettes may have been overestimated.

3.2. "PROXY CONDITION" UTILITY VALUES

"Proxy condition" utility values involve the use of utility values for one condition to be used as a proxy to represent utility values for another condition. For example, this involves the use of utility values for another condition where the condition is thought to have a similar impact on health and quality of life in terms of both the dimensions of health that are impacted and the degree of impairment.

One example is the use of utility values elicited for age-related macular degeneration (AMD) as a proxy for utility values in the treatment of limbal stem cell deficiency after eye burns [24]. The utility values were generated in a study that elicited utility values from members of the general population using TTO where symptoms of AMD were simulated using contact lenses [25]. The utility values used in the economic model focussed on the relationship that was reported in the study between TTO utility values and visual acuity.

Another example is the use of utility values of patients with ROS1-positive advanced non-small cell lung cancer as a proxy for utility values of patients with ALK-positive

advanced non-small cell lung cancer [26]. The “proxy condition” utility values were sourced from trial data using EQ-5D-3L, and it was argued that the two conditions have similarities in their patient populations and their characteristics.

The use of “proxy condition” utility values is arguably not uncommon, as mapping functions may have been developed in one patient group and used to generate utility values in another. For example, a mapping function estimated from a cancer-specific measure to EQ-5D for a common cancer could be applied to estimate utility values for a rare cancer where it may be more challenging to generate a dataset suitable for estimating mapping functions.

4. BEST PRACTICE RECOMMENDATIONS

4.1. How to determine whether there is insufficient EQ-5D evidence available

As outlined above, the recommended method for generating utility values for use in TA and HST evaluations is EQ-5D self-reported by patients in a relevant study. The use of a relevant study where EQ-5D is directly administered to patients is preferred wherever possible, or using proxy-complete by carers (including parent-reporting) if this is infeasible. Relevant studies can include the clinical trial of the treatment, but can also include other studies such as observational studies. Wherever possible EQ-5D evidence should be provided. The use of utility estimates in economic models that are not generated using EQ-5D causes uncertainty in the estimates and reduces the comparability of the results to all other TA and HST evaluations.

However, EQ-5D evidence may not always be available, EQ-5D may be inappropriate for a certain condition, or EQ-5D evidence may be insufficient to generate all utility values required in the economic model. Figure 1 outlines the different recommendations for generating utility values depending on whether EQ-5D data is unavailable, inappropriate, or where there is insufficient data available.

4.1.1. Where EQ-5D data is unavailable

If it is not possible to directly administer EQ-5D to patients, the next best alternative is any one of the following: undertake a literature review using a systematic search; generate EQ-5D estimates via mapping using available evidence; or conduct a study to collect EQ-5D data. Conducting a study to collect EQ-5D data can be used alongside existing data, for example, if existing trial data has small sample size. Existing studies that include EQ-5D data should be searched for potential usage, and this can include clinical audits, registries, and observational studies. All other methods are less preferred alternatives and are not justified on the grounds of insufficient planning or time to submission of the TA or HST evaluation. It is therefore recommended that organisations who will be preparing an evaluation for submission to NICE should seek advice as early as possible from experts in health economic outcomes and clinicians around the best methods for measuring, valuing and sourcing utility values for their economic model. Wherever possible, the sourcing and use of

EQ-5D data is recommended, and any other methods outlined in this report should be used only in the limited circumstances defined in sections 3.1.2 and 3.1.3.

4.1.2. Where EQ-5D is inappropriate

NICE recommendations about what should be done where EQ-5D is inappropriate are:

"In some circumstances the EQ-5D may not be the most appropriate. To make a case that the EQ-5D is inappropriate, qualitative empirical evidence on the lack of content validity for the EQ-5D should be provided, demonstrating that key dimensions of health are missing. This should be supported by evidence that shows that EQ-5D performs poorly on tests of construct validity and responsiveness in a particular patient population. This evidence should be derived from a synthesis of peer-reviewed literature. In these circumstances, alternative health-related quality of life measures may be used and must be accompanied by a carefully detailed account of the methods used to generate the data, their validity, and how these methods affect the utility values." ([1], p45).

In order for it to be concluded that EQ-5D is inappropriate for a condition (or population) psychometric evidence demonstrating this is required. It cannot be simply claimed that EQ-5D is inappropriate, rather this needs to be demonstrated using evidence on the psychometric performance of EQ-5D. Psychometric evidence could include evidence around content validity, construct validity, responsiveness and reliability.

Where an alternative preference-based measure of HRQoL is used, this could be a generic preference-based measure or a condition-specific preference-based measure. Recommendations on alternative measures to EQ-5D are also detailed in a NICE Technical Support Document (TSD) [9]. The TSD makes recommendations around the details that should be provided around the use of any alternative preference-based measure that is used to generate utility estimates [9]. The TSD outlines that the development of the measure should be fully described, including health state descriptions (how health states are derived, typically based on an existing measure)

and health state values (valuation methods and comparability of these methods to those used to value EQ-5D) [9]. The TSD also recommends that empirical evidence should be provided on the alternative measure, including the psychometric properties of the measure such as content validity, construct validity, responsiveness and reliability [9]. It is also recommended that the impact on utility values due to the use of the alternative measure instead of EQ-5D are detailed [9].

Selection of the alternative measure should be informed both by the appropriateness and relevance of the measure for patients with the condition, the rigour and quality of the development of the measure, and whether the methods used to value the measure are in accordance with NICE recommendations (i.e. public preferences elicited using a choice-based method). It is important that the measure has been developed using good scientific practice, for many reasons, since this can impact on the accuracy of the utility estimates. For example, how easy the questions are to understand and complete by patients is important since answers should be accurate. Measures developed using good scientific practice are also likely to include important and relevant dimensions of HRQoL with appropriate response options. The valuation methodology used to generate utility values should also be considered, since different preference elicitation techniques can generate different values, and values elicited from patients can differ to values elicited from members of the general population. For comparability to EQ-5D, it is recommended that utility values are elicited from a representative sample of the general population using a choice-based technique. However, use of a measure that may be deemed of poor quality or relevance should not be justified on the basis that it has been valued using a choice-based method. The selection of either a generic preference-based measure or a condition-specific preference-based measure is likely to be context specific, depending both on the condition and the measures that are available (see [27] for a summary of available condition-specific preference-based measures). NICE stipulate that if other measures are used they should be accompanied “by a carefully detailed account of the methods used to generate the data, their validity, and how these methods affect the utility values” p45, [1]. It is recommended that the measure, its development, and where possible the impact on the utility values should be transparent, and clearly described. It is also recommended that any limitations of the measure or its usage are clearly

acknowledged to enable the ERG, NICE committee members and wider audience to interpret the evidence.

The NICE TSD outlining alternative measures to EQ-5D discusses other methods that generate utilities, including vignettes and patient own health state valuations [9]. Vignettes are not recommended for use when EQ-5D is unavailable due to the limitation that they are study-specific and are rarely based on self-report data from patients. Vignettes are recommended for use only when EQ-5D data is insufficient, and these recommendations are detailed below in section 3.2. Section 3.2 also provides a more detailed discussion of the limitations of vignettes.

Studies collecting patient valuations of their own health state typically ask patients to complete a time trade-off, visual analogue scale or standard gamble exercise for their own health today. The health of the patient today is not defined for the valuation exercise using an existing measure, and often patients do not self-report their health status using a preference-based measure. One advantage of this is that the lack of description of own health avoids any problems of poor coverage, insensitivity or irrelevance that are often used as criticisms for generic preference-based measures. However, there are ethical and practical considerations around the administration of valuation exercises that require a patient who may be in state of very poor health to consider whether they would rather live in their current health state or die. In addition, patient valuations differ in important ways to valuations from the general public (see [28] for an overview). There is no consensus across all health conditions around whether patient valuations are higher or lower that general public valuations, though evidence suggests that patient values are generally higher for physical health conditions and lower for mental health conditions. Patient values can incorporate adaptation, because patients have adapted to their ill health. Patient values may also be impacted by a response shift, where patients may, for example, have changed their expectations of full health to instead reflect full health for their age, or value their health relative to the health of other patients also in poor health.

When EQ-5D is considered inappropriate, the use of an alternative generic or condition-specific preference-based measure is a recommended option over the patients' valuation of own health. Studies that involve patients valuing their own health

face challenges around patient recruitment, ethical acceptability, comparability and practicality of undertaking such a study. In addition, the values generated from these studies are elicited from patients, and may incorporate factors such as adaptation to the poor health state that arguably should not be included. For these reasons, the use of alternative generic or condition-specific measures is the preferred option when EQ-5D is inappropriate.

4.1.3. Where EQ-5D data is insufficient

The cases where EQ-5D data is insufficient are typically due to small populations, rare events, or health states that may occur in real life but not in clinical studies. This means that sufficient EQ-5D data for all states required in the model is not available. The situations where EQ-5D may be insufficient can be summarised as follows (this is adapted from [29] and their description of when to use vignettes):

- Where the condition is rare, and this means that sufficient EQ-5D data for all states required in the economic model cannot be observed.
- Where the population assessed is children and adolescents and potentially rare, and this means that there are small sample sizes in relevant studies and that sufficient EQ-5D data for all states required in the economic model cannot be observed.
- Where the condition is common but there are health states, or adverse events, that are not commonly observed (i.e. rare events), and this means that sufficient EQ-5D data for all states required in the economic model cannot be observed.
- Where some states required for the economic model occur years later and outside the time frame of the clinical studies and are not commonly experienced in study datasets (i.e. rare events), and this means that sufficient EQ-5D data for all states required in the economic model cannot be observed.
- Where some states required for the economic model are not observed in studies due to the study design, and this means that sufficient EQ-5D data for all states required in the economic model cannot be observed.
- Where health states are temporary, for example a flare or exacerbation, and it is not possible to ask respondents to complete EQ-5D during this time, and this

means that sufficient EQ-5D data for all states required in the economic model cannot be observed.

In order to use other methods on the grounds that there is insufficient evidence available:

- A clear explanation should be provided detailing why the use of alternative methods has been undertaken;
- Evidence should be provided that demonstrates that it was not possible to source sufficient EQ-5D estimates either by: directly administering EQ-5D to patients; sourcing utility values from the literature; or estimating EQ-5D utilities using mapping.

It is not sufficient to argue that the clinical trial(s) did not have a large enough sample(s), since the data collection of utility values should be considered in advance of the TA and HST submission to NICE. Furthermore EQ-5D evidence can be generated using a variety of sources and not solely from trial data. This can include clinical audits, registries and observational studies, and include proxy-reported EQ-5D data. For rare events EQ-5D data may be available for another condition, for example, utility decrements for febrile neutropenia observed in one cancer may be representative of utility decrements for febrile neutropenia in other cancers.

There may be situations where there is EQ-5D evidence available, but this is based on poor quality data with small sample size and all states required in the economic model are not observed. This means that there is insufficient EQ-5D evidence available for all states. Wherever possible EQ-5D evidence should be used, and other evidence used only where necessary. Sensitivity analyses can be used to explore the inclusion of non-EQ-5D evidence where EQ-5D evidence is available but it is insufficient.

Sections 3.2 and 3.3 outline best practice recommendations and important considerations for the two approaches that are recommended for use when insufficient EQ-5D data is available. The two approaches are vignettes, and “proxy condition” utility values. Section 3.4 outlines how to choose whether vignettes or

“proxy condition" utility values are most appropriate for a given TA or HST evaluation.

Figure 1 Steps to identify appropriate methods to generate utility values

Recommended method: EQ-5D self-reported by patients in a relevant study

- **If EQ-5D data is unavailable, EQ-5D can be:**
  - Sourced from a literature review using a systematic search of the literature;
  - Or estimated using statistical mapping;
  - Or generated by conducting a study to collect EQ-5D data to populate the economic model.

- **If EQ-5D is considered to be inappropriate (demonstrated with empirical evidence) then use (in order of preference):**
  - a) Other generic or condition-specific preference-based measures;
  - b) Valuation of own health, for example using time trade-off to value patient own health.

- **If insufficient EQ-5D data is available either from relevant studies, literature or using mapping, utility data can be generated using either:**
  - Vignettes, valued using (in order of preference):
    - a) Indirectly, by clinical experts, patients or general population completing the EQ-5D for each vignette and this is then scored using the appropriate and relevant value set for EQ-5D, provided EQ-5D is appropriate, patient samples are recommended;
    - b) Directly using preference elicitation techniques such as time trade-off with a sample of the general population or patients;
    - c) Utility values elicited directly for each vignette from clinical experts, for example using Delphi panels or preference elicitation methods including time trade-off.
  - Or “Proxy condition” utility values, if relevant.

4.2. VIGNETTE RECOMMENDATIONS

4.2.1. Recommendations for vignette development

The development of vignettes can vary widely, and this may be expected due to the heterogeneity of conditions and circumstances where vignettes are generated. Figure 2 reports recommendations of best practice for vignette development, adapted from recommendations reported in [29].

Figure 2: Proposed recommendations of best practice for vignette development

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

- Published literature, for example reviews or original studies including qualitative studies around the HRQoL 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 since patients may not be representative and formal diagnosis is not ensured.
- Quantitative data (for example patient-reported outcome measures of HRQoL in clinical trials or observational studies).

Vignette development including content and format

- The number of vignettes and the required severity/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, as 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 e.g. simple language where possible if presented to members of the general public, appropriate font size, use of boldening/underlining to highlight different levels of severity.

- Vignettes should be presented and formatted to enable ease of understanding of the target audience of 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 since it can impact on the utility values that are elicited as some participants may provide relative values for the vignettes whilst considering all vignettes.
- Vignettes should include descriptions of the generic dimensions of HRQoL, for example using the EQ-5D dimensions and descriptions. This can reduce focussing effects where respondents may focus on the symptoms or treatment effects described rather than considering these in a wider context of HRQoL.
- Vignettes should include all important and relevant aspects of HRQoL 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 HRQoL, so care should be taken to describe the aspects of HRQoL rather than clinical stages, since 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.

Vignette refinement, validation and interpretation

Input from clinical experts and/or patients via 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.

Prior to the main valuation study it is recommended to ensure that the descriptions are able to 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.

Adapted from [29].

Vignettes are typically used to meet specific requirements of events or health states required in an economic model for TA and HST evaluation. Therefore the requirements of the economic model can be used to determine the number of vignettes and the required severity/disease state for each vignettes. The vignettes should meaningfully differ from each other, as subtle differences in descriptions may not be captured in the valuation stage. However these differences should not be exaggerated since this will not accurately reflect the event or state.

Where possible, the use of multiple sources of evidence should be encouraged since this enables the vignettes to be grounded in a wider amount of evidence. For example, there may be patient-reported outcome data on HRQoL available in a trial or observational study on functioning, usual activities and symptoms that can be used as evidence to inform the vignette development. Qualitative evidence, either from existing published sources or from social media such as online patient discussion forums, can also be used to obtain descriptions of patients HRQoL. However, it should be noted that there are limitations with these techniques where the qualitative evidence may not be representative of the wider population. Patient involvement during the vignette development can be one way of ensuring that the vignettes focus on the aspects that are relevant for patients, since this may differ to what is reported by clinical experts.
Vignettes can vary widely in: their length (from brief to very detailed); number of aspects covered (from a small number through to a large number); and their focus (from specific symptoms, to include side effects, to the inclusion of generic aspects of functioning and usual activities, to the inclusion of treatment specific and process attributes associated with the treatment). It is important that the respondent completing the valuation exercises can understand the health state that is described. Brief vignette

descriptions may not contain sufficient detail, whereas very detailed descriptions may contain too much detail. If the vignette description is too brief there may be ambiguity and potential for misinterpretation, as respondents in the valuation exercise may imagine aspects of health that are not included in the description. In contrast, long descriptions may encourage respondents to focus on a subset of aspects and to not pay attention to all information provided in the vignette description.

Limitations of vignettes stem from the fact that they are not (usually) based on patient-reporting of their own health and are not standardised, meaning that they are potentially more open to bias and inaccuracy. The validity and accuracy of the vignettes and the utilities they are associated with depends on their development, and the following limitations can apply:

- Vignettes may not include all aspects of health that are deemed important to patients.
- Vignettes may not accurately depict all aspects of health.
- Vignettes may focus on some aspects of health and downplay others. In the valuation stage this may be leading for the target audience and open to bias.
- Vignettes may focus on symptoms and treatments and exclude functioning and generic aspects. In the valuation stage this may cause a focussing effect which leads to lower utility values.
- Vignettes may include disease labels and this can cause bias in the valuations (see for example [30]).
- Vignette descriptions may encourage greater differentiation between health states across treatments and greater exaggeration of side effects than is experienced by patients.
- Vignettes focus on the ‘typical patient’ for each disease or treatment stage, which does not reflect the variation that is typically observed in patients. Whilst it is possible to report a range across some aspects which may be more believable and more representative, for example episodes experienced per month, this can make it more difficult to value and may be more open to interpretation by people valuing the states.
- Vignettes containing a large amount of information can be difficult to value, and respondents may focus on a subset of the information provided. In contrast,

vignettes containing little or limited information can be open to interpretation and misinterpretation if respondents imagine the aspects of health that are not described. However, this argument can be applied to preference-based measures as well as vignettes.

- Vignette development is not standardised and this means that the development may be both less transparent and more open to researcher influence than directly administering a preference-based measure to patients (or their carers or proxies) in a trial or observational study.

The use of clinical experts is crucial in ensuring that the descriptions are appropriate for the exact states required in the economic model. Clinical experts should be selected on the basis that they have relevant experience. Nurses may have more experience than medical clinical doctors on the day-to-day HRQoL impact of patients, particularly where this may involve treatment undertaken over a long time period with large nurse involvement, for example chemotherapy or stoma management.

One problematic issue with the use of vignettes is around the quality of the study, and this can vary widely across studies used to generate utility values for economic models. Matza et al [29] state that “vignette utility studies must be well designed, carefully conducted, clearly reported, and interpreted with appropriate caution” [29]. Where vignette studies are well conducted, transparently reported, well motivated and there is no more appropriate technique, they may be an acceptable and useful approach for generating utilities. This is evidenced by the fact that vignettes are acceptable methods for estimating utilities for some international agencies [31,32].

Recommendations around transparent reporting of vignette development are: names and conflicts of interest of experts involved in the development of vignettes; clear and informative reporting of methods used, to include the specific input from clinical experts and patients during each stage of the vignette development; the role of each source of evidence in developing the vignettes; sample sizes at each stage and whether the sample composition differed across the different stages of the vignette development; vignette descriptions before and after validation to identify any changes, and the rationale behind any changes made to the descriptions.

Comparability between EQ-5D dimensions and the vignette descriptions is important. Vignette descriptions will arguably be more comparable if they include generic aspects of HRQoL related to function and usual activities. For example, vignettes could cover the same dimensions as the EQ-5D, with additional aspects of health also included. The inclusion of treatment effects and administration, treatment labels, disease labels, symptoms, burden of disease and disease history differs from EQ-5D (where none of these are included) and hence reduces comparability. Therefore, the inclusion of any additional aspects within a vignette should be considered carefully, and wherever possible excluded. These should be included only where necessary and should be appropriate for the economic model.

4.2.2. Recommendations for valuation of vignettes

4.2.2.1. Vignettes valued indirectly by completing EQ-5D

This is where participants proxy-complete the EQ-5D for a vignette. The rationale for taking this approach to value the vignettes rather than eliciting values from members of the general population was summarised in the HTA for X-linked hypophosphataemia "…it is difficult to determine the accuracy of the vignettes themselves and the general public rating the states may not fully recognise the relevance of aspects of the disease burden or may over emphasise the impact of certain issues." p226, [23]. Therefore, the rationale underlying this approach is firstly that it may reduce some of the problems raised with vignettes, and secondly that the measure and value set used to score the health states can meet the NICE reference case through the use of EQ-5D. Whilst the examples presented in Section 2.1.2.3 above involved the use of clinical experts to complete the EQ-5D for the vignettes, this could be patients or members of the general population if the vignette descriptions are worded to ensure they are understandable. One method that can be used to generate the utility values involves asking participants to report the EQ-5D for each vignette and then take the mean and standard deviation of the utility values that are generated. An alternative method is to ask participants to state the range of levels expected for each vignette to generate a range of utility values, since the patient group may be

heterogeneous (for example this approach was used in [23], which can be viewed as an attempt to reflect the variability in outcomes and HRQoL experienced by patients). One advantage of this is that it will generate a range of utilities that could be used to inform sensitivity analyses of the economic model. However, it would not be expected that this approach would produce the same distribution and natural variability in health states as would be observed if different patients self-report their own health. Furthermore, this approach may be too complex in some cases to be undertaken by members of the general population or patients.

One challenge is that the vignettes cannot fully contain EQ-5D information already, hence they must be focussed more upon symptoms and aspects related to treatment, meaning that any interpretation of how the vignette impacts on the generic dimensions covered by EQ-5D will be subjective. In addition the dimensions of the EQ-5D may not reflect all the aspects of health described in the vignette.

Where clinical experts report the EQ-5D, they will be reporting from their experience how they have observed that this impacts on the dimensions included in the EQ-5D. Their experience may not be accurate, and they may have limited understanding of how the condition impacts on the day-to-day life of patients across different disease stages. Whilst clinicians are likely to observe patients with the condition, clinicians can only outwardly observe what it is like for a patient, and may not be able to accurately report their generic functioning and usual activities, particularly where these are around how the patient is feeling. This is why patients are typically asked to self-report their own health in trials, since clinical experts are poor proxies for patients. In addition, clinical experts may not observe all disease stages particularly where the patient population is very small, meaning that they may not be able to proxy-report EQ-5D for all of the vignettes across the disease stages (se for example [23]). Clinical experts are also not independent of the condition, and may bring with them their own preconceptions and opinions. This can provide greater understanding of the different vignettes, but can also lead to values that may incorporate bias or that may not be based on the vignette descriptions. As discussed above, nurses may have more understanding of the day-to-day impact of the condition than medical clinical doctors. In contrast, whilst members of the general population are not expected to have any experience of the health states, they are independent and unbiased.

The main argument in favour of asking patients to complete EQ-5D for the vignettes is that patients have greater understanding of the symptoms and treatment and how these impact on health for people with the condition. In addition vignettes can contain symptoms that are specific to the patient population, such as seizures, that are not easily understood by members of the general population. This approach is in line with NICE recommendations, since ultimately the health states are defined using the EQ-5D classification system and the values generated are EQ-5D utilities from a recommended value set. However, crucially the utility estimates will differ from utility values generated from patient self-report of EQ-5D of their own health, since they do not have the same distribution and variability of health state utility values that would be reported by patients since the utility values are based on vignettes not own health.

4.2.2.2. Direct valuation of vignettes by members of the general public or patients

Utility values for vignettes can be generated using preference elicitation techniques, for example using time-trade-off. As discussed above, if valuation of vignettes is undertaken then the method that is closest to the reference case is a choice-based method, for example TTO, completed by a representative sample of the general population. Values elicited from patients can differ to values elicited by members of the general population for many reasons (see [28] for an overview). In order to achieve comparability to EQ-5D, the elicitation of general population values is recommended rather than the elicitation of patient values.

One disadvantage of the valuation of vignettes using TTO completed by members of the general population is that the general population may overestimate the importance of condition-specific symptoms, where the symptoms are presented but the impact on more generic domains of health is not presented. For example, where only condition-specific symptoms are presented the impact on utility values may be exaggerated due to focusing effects on the symptoms that are presented, rather than considering other important aspects of health that may not be impacted. As recommended above, ensuring that vignette descriptions cover generic domains may minimise this, as the symptoms are presented within the wider context of all important health domains.

Another potential disadvantage of this approach is that the general population may be affected by labelling effects, as the labelling of the condition or potentially the treatment can impact on values (see for example [30]). For this reason, it is recommended that condition and treatment labelling is avoided in the vignettes.

4.2.2.3. Values elicited for vignettes by clinical experts

The elicitation of utility values for vignettes from clinicians is problematic since this will generate the preferences of the clinician, which should not be expected to be representative of the utility value experienced by the patient. Clinicians are unrepresentative of the general population for many reasons, including that their health and sociodemographic profile is non-representative and their views are likely to be impacted by their profession. First, techniques such as TTO incorporate time preference and participants trade between length of life and quality of life. Clinicians may be more or less willing than members of the general population to sacrifice between length and quality of life. Second, the elicitation exercises will be undertaken from a different perspective, since the clinical expert has a patient-doctor relationship with patients experiencing these health states, and this could impact on their preferences. Third, it is likely that the clinical expert is imagining an “other” patient in the health state they are doing this on behalf of, regardless of the perspective they are asked to take in the exercise, and previous research has found that valuing health states experienced by someone else, i.e. an “other” perspective, impacts on values [34]. Eliciting values from nurses or carers may be a better option than medical clinical doctors, since they may have a better understanding of how the condition impacts on the day-to-day life of patients across different disease stages.

Wherever feasible, eliciting utility values from the general population is a more preferred option, in accordance with NICE recommendations [1]. However, if general population values are infeasible, for example because there are clinical aspects of the condition that may not be understood by members of the general population such as seizures, then patients would be a better option than clinical experts. For rare conditions, this could potentially be achieved through the use of patient charities if other recruitment strategies are infeasible. The use of patients from similar conditions

(for example different types of epilepsy if the type of epilepsy is rare), and the use of Patient Involvement groups may be feasible in situations where it is not possible to recruit patients as research participants.

4.2.2.4. Valuation recommendations

Utility values for vignettes can be generated using a range of different methodologies. For the valuation there is a decision both around which preference elicitation technique to use and whose preferences should be elicited. As reported above, the approaches used can vary. NICE recommend implicitly that a choice-based method is preferred, and the use of TTO provides comparability to the preferences elicited for the value set for EQ-5D [1]. For NICE, in common with most international agencies [33], the use of general population preferences is recommended for valuation in their reference case, and therefore general preference values are recommended.

The methods recommended for use to value vignettes are (in order of preference): a) Indirect valuation by asking patients to complete the EQ-5D for each vignette and this is then scored using the appropriate and relevant value set for EQ-5D, provided EQ-5D is appropriate; b) Direct valuation by preference elicitation techniques such as time trade-off with a sample of the general population or patients; c) Utility values elicited directly for each vignette from clinical experts, for example using Delphi panels or preference elicitation methods including time trade-off.

4.3. "PROXY CONDITION" UTILITY VALUES

This raises the question of when it can be concluded that utility values for one condition can be deemed to be an appropriate proxy ("proxy condition" utility values) for another condition. We recommend that it is appropriate to use "proxy condition" utility values where the impact of the disease is the same, both in terms of the dimensions that are impacted and the degree in which they are impacted, for the two different conditions. The use of "proxy condition" utility values should also take into consideration whether there are appropriate utility values available for the "proxy condition" that meet the NICE reference case using the EQ-5D. However, EQ-5D data may not be available for the "proxy condition" if the "proxy condition" is also rare.

The population, disease impact including both symptoms and functioning, and where relevant, adverse events of the “proxy condition” must be representative for the condition of interest in the economic model. It should be clearly described and acknowledged if there are differences between the “proxy condition” and the condition of interest. For example, there may be differences across disease stages, events, adverse events, clinical indicators, or how the condition impacts on the HRQoL of the patient. There may be instances where there are important differences, but “proxy condition” utility values remain the most indicative utility values since they provide a more accurate description of HRQoL than other options available.

We recommend that a qualitative assessment is undertaken involving both patients and clinical experts to determine whether the “proxy condition” and the “proxy condition” utility values are considered appropriate and representative for the condition of interest. An assessment of appropriateness could also be informed by a targeted literature review examining how each condition impacts on the HRQoL of the patients. The extent to which “proxy condition” utility values are appropriate for the condition of interest may vary across different health states within the model. It is also important to take into account how the utility values are used in the economic model. For example, utility decrements for adverse events may be highly appropriate whereas utility values for other states may be inappropriate. One example is febrile neutropenia in cancer, where utility decrements for this adverse event may be representative of utility decrements for febrile neutropenia in other cancers. However, when the adverse event is combined with the impact of the cancer in the economic model, this may generate different utility values for health states in the “proxy condition” and the condition of interest.

The degree of appropriateness of the “proxy condition” utility values for the condition of interest will determine where the use of “proxy condition” utility values lies in the hierarchy of evidence presented in figure 1. This will then inform decisions around whether the “proxy condition” utility values are appropriate for use in the economic model. It is not expected that appropriateness is a binary issue of appropriate/inappropriate, but rather that there will be a range of degrees of appropriateness.

4.4. CHOICE OF VIGNETTES OR "PROXY CONDITION" UTILITY VALUES

The appropriateness of use of either vignettes or "proxy condition" utility values differs on a case-by-case basis. In addition the degree of appropriateness is not a binary appropriate/inappropriate judgement but a matter of degree of appropriateness. A well-conducted vignette study, for example, may be recommended over "proxy condition" utility values where evidence indicates that the "proxy condition" is unrepresentative of the condition of interest and how the condition impacts on the HRQoL of the patient. The factors that should be used to select whether vignettes or "proxy condition" utility values are used include:

- The appropriateness of the "proxy condition" utility values for representing the utility values of health states for the condition of interest (see previous section).
- The quality of the vignette study, including whether it follows the best practice recommendations outlined above and whether the study is well conducted, transparently reported and well motivated.
- The comparability of the vignette study or "proxy condition" utility values to the NICE reference case of EQ-5D, where the more comparable utility values are recommended for use.

5. CONCLUSIONS

The report has outlined examples of methods used to generate utility values where insufficient EQ-5D data is available to populate the economic model in TA and HST evaluations. The report makes recommendations around how to generate HRQoL evidence where EQ-5D is unavailable, where EQ-5D is inappropriate, or where EQ-5D data is insufficient. There is a limited academic literature providing recommendations and discussions of the advantages and limitations of each of the different approaches identified in Section 2. Therefore, this report reflects both the limited literature available and the authors' opinions; this is a limitation of the report.

5.1. SUMMARY OF PROPOSED RECOMMENDATIONS, FOR CONSIDERATION BY NICE

The following points are the key report recommendations:

- The use of EQ-5D directly administered to patients and scored using general population preferences is the preferred option to generate utility values. The use of any other method where insufficient evidence cannot be observed remains a second-best alternative, as failure to develop a sufficient body of evidence using EQ-5D, where it would have been desirable and feasible leads to unnecessary uncertainty and incomparability to other appraisals. Evidence should be provided demonstrating why the reference case method of EQ-5D has not been used. Where evidence demonstrates that it was not possible to directly administer EQ-5D to patients, the 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 to populate the economic model.
- Where EQ-5D is demonstrated using empirical evidence to be inappropriate, it is recommended to generate evidence using either an alternative generic preference-based measure or a condition-specific preference-based measure. If this is not possible, patient valuations of their own health state can be used. Where EQ-5D is inappropriate, this should be evidenced using psychometric evidence. This could include evidence around content validity, construct validity, responsiveness and reliability. In addition, the development of the measure should be fully described, including health state descriptions and

health state values. Evidence should be provided on the alternative measure, including the psychometric properties of the measure and the impact on utility values due to the use of the alternative measure instead of EQ-5D.

- Where sufficient EQ-5D evidence is not available, what is important is the appropriateness of the methods and evidence used to generate the utility values, the appropriateness of the utility values used in the economic model, and their comparability to EQ-5D. The appropriateness and selection of either vignettes or “proxy condition” utility values will vary depending on the condition and economic model.
- The reason why EQ-5D is unavailable or insufficient is important. There is a distinction between unavailable or insufficient data due to poor planning or failure to include EQ-5D in clinical studies, and unavailable or insufficient data due to characteristics of the patient population and/or states required to populate the economic model. It is understandable if appropriate attempts have been made to use EQ-5D data but this is not fully achievable due to the patient population and/or states required in the economic model. Poor planning or failure to include EQ-5D in clinical studies where EQ-5D is appropriate is unjustifiable. In addition relevant evidence can be obtained from a range of different study types, and is not limited to clinical studies. Early consideration of the evidence requirements can avoid the need to use alternative methods to EQ-5D.
- If EQ-5D data are unavailable or insufficient clear evidence is required demonstrating why it was not possible to use self-report EQ-5D data. The evidence should demonstrate that it was not possible to: directly administer EQ-5D to patients as part of the clinical studies of the technology in question; source utility values from the literature; estimate EQ-5D utilities using mapping; or conduct a separate study to collect EQ-5D data. The use of other methods to compensate for inadequate evidence generation introduces unnecessary uncertainty into the appraisal of health technologies.
- The utility values should be generated using appropriate methods, free from bias and as comparable as they can be to the reference case method of EQ-5D. The studies underpinning the evidence should be well-conducted, transparent, and clearly described with all study advantages and limitations

highlighted. The studies underpinning the evidence should clearly acknowledge any limitations to enable the ERG, NICE committee members and wider audience to interpret the evidence.

In those situations where vignettes are used, we make the following recommendations:

- Obtain high quality appropriate and reliable evidence to inform vignette development. This could consist of, and be strengthened by the use of, multiple different types of evidence including published literature, qualitative evidence and quantitative HRQoL data.
- Vignettes should be developed to meet the requirements of the economic model structure for the TA and HST evaluations; should be formatted to enable easy reading, comprehension and understanding; should include generic aspects of HRQoL; should include all important and relevant aspects of HRQoL; be worded with minimal potential for both ambiguity and misinterpretation; reflect the typical patient experience for the disease state in question; should avoid probabilistic statements; should not include disease or treatment labels where possible; and should not use value-laden or irrelevant phrases or content (such as ‘devastating’).
- Vignette refinement and validation should be undertaken using input from clinical experts and/or patients to ensure that the vignettes are a clear and accurate description of the disease state or adverse event that they are intended to represent in the economic model.
- The reporting of the process of vignette development needs to be full and transparent.
- Utility values for vignettes are generated using an appropriate sample of patients completing the EQ-5D for each vignette, and this is then scored using the appropriate and relevant value set for EQ-5D, provided EQ-5D is appropriate.

6. REFERENCES

[1] National Institute for Health and Care Excellence. Guide to the methods of technology appraisal. 2013. Available at: https://www.nice.org.uk/

[2] NICE Interim Process and Methods of the Highly Specialised Technologies Programme Updated to reflect 2017 changes. April 2017, Available at: https://www.nice.org.uk/

[3] NICE. Diagnostics Assessment Programme manual. December 2011. Available at: https://www.nice.org.uk/

[4] Brazier J, Ara R, Azzabi I, Busschbach J, Chevrou-Severac H, Crawford B, Cruz L, Karnon J, Lloyd A, Paisley S, Pickard AS. 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, 2019; 22:267-275.

[5] Wailoo AJ, Hernandez-Alava M, Manca A, Mejia A, Ray J, Crawford B et al. Mapping to Estimate Health-State Utility from Non-Preference-Based Outcome Measures: An ISPOR Good Practices for Outcomes Research Task Force Report. Value in Health 2017; 20:18-2.

[6] Longworth L, Rowen D. The use of mapping methods to estimate health state utility values. NICE DSU Technical Support Document 10. 2011.

[7] Longworth L & Rowen D (2013) Mapping to obtain EQ-5D utility values for use in NICE health technology assessments.. *Value Health*, 16(1), 202-210.

[8] Ara R, Rowen DL & Mukuria (2017) The Use Of Mapping To Estimate Health State Utility Values. PharmacoEconomics, 35(Suppl 1), 57-66.

[9] Brazier JE, Rowen D.NICE Decision Support Unit Technical Support Document 11: Alternatives to EQ-5D for generating health state utility values. 2011. Available at: ⟨http://www.nicedsu.org.uk⟩.

[10] National Institute for Health and Care Excellence. Short bowel syndrome - teduglutide [ID885] 2019 [Available from: https://www.nice.org.uk/guidance/indevelopment/gid-ta10048/documents].

[11] Ballinger R, Macey J, Lloyd A, Brazier J, Ablett J, Burden S, Lal S. Measurement of utilities associated with Parenteral Support Requirement in Patients with Short Bowel Syndrome and Intestinal Failure. Clinical Therapeutics, 2018; 40:1878-1893.

[12] National Institute for Health and Care Excellence. Cannabidiol for adjuvant treatment of seizures associated with Dravet syndrome [ID1211] 2019 [Available from: https://www.nice.org.uk/guidance/indevelopment/gid-ta10274].

[13] National Institute for Health and Care Excellence. Cannabidiol for adjuvant treatment of seizures associated with Lennox-Gastaut syndrome [ID1308] 2019 [Available from: https://www.nice.org.uk/guidance/indevelopment/gid-ta10410].

[14] Torrance GW, Feeny D, Furlong W. Visual analog scales: do they have a role in the measurement of preferences for health states? Med Decis Making. 2001;21:329–334.

[15] National Institute for Health and Care Excellence. Liposomal cytarabine–daunorubicin for untreated acute myeloid leukaemia [TA552] 2018 [Available from: https://www.nice.org.uk/Guidance/TA552].

[16] National Institute for Health and Care Excellence. Cerliponase alfa for treating neuronal ceroid lipofuscinosis type 2 [ID943] 2019 [Available from: https://www.nice.org.uk/guidance/indevelopment/gid-hst10008].

[17] National Institute for Health and Care Excellence. Voretigene neparvovec for treating inherited retinal dystrophies caused by RPE65 gene mutations [HST11] 2019 [Available from: https://www.nice.org.uk/guidance/hst11].

[18] van Hout B, Janssen MF, Feng Y-S, Kohlmann T, Busschbach J, Golicki D, Lloyd A, Scalone L, Kind P & Pickard AS (2012) Interim scoring for the EQ-5D-5L: mapping the EQ-5D-5L to EQ-5D-3L value sets.. Value Health, 15(5), 708-715.

[19] Rentz AM, Kowalski JW, Walt JG, Hays RD, Brazier JE, Yu R, et al. Development of a preference-based index from the National Eye Institute Visual Function Questionnaire-25. JAMA Ophthalmol. 2014;132(3):310-8.

[20] National Institute for Health and Care Excellence. Nusinersen for treating spinal muscular atrophy [TA588] 2019 [Available from: https://www.nice.org.uk/guidance/ta588].

[21] Khan KA, Petrou S, Rivero-Arias O, Walters SJ, Boyle SE. Mapping EQ-5D utility scores from the PedsQLTM generic core scales. Pharmacoeconomics. 2014;32(7):693–706

[22] Dolan R. Modelling valuations for Euroqol health states. Medical Care. 1997;35(11):1095-108

[23] National Institute for Health and Care Excellence. Burosumab for treating X-linked hypophosphataemia in children and young people [HST8] 2018 [Available from: https://www.nice.org.uk/guidance/hst8].

[24] National Institute for Health and Care Excellence. Holoclar for treating limbal stem cell deficiency after eye burns [TA467] 2017 [Available from: https://www.nice.org.uk/guidance/ta467].

[25] Czoski-Murray C, Carlton J, Brazier J, Young T, Papo NL & Kang HK (2009) Valuing Condition-Specific Health States Using Simulation Contact Lenses. Value in Health(5), 793-799.

[26] National Institute for Health and Care Excellence. Crizotinib for treating ROS1-positive advanced non-small-cell lung cancer [TA529] 2018 [Available from: https://www.nice.org.uk/guidance/ta529].

[27] Rowen DL, Brazier J, Ara R & Azzabi Zouraq I (2017) The Role Of Condition-Specific Preference-Based Measures In Health Technology Assessment. PharmacoEconomics, 35(Suppl 1), 33-41.

[28] Brazier, J., Ratcliffe, J., Salomon, J., Tsuchiya, A.: Measuring and valuing health benefits for economic evaluation, 2nd edn. Oxford University Press, Oxford (2016)

[29] Matza LS, Stewart KD, Lloyd A, Rowen D, Brazier JE. Vignette-Based Utilities: Usefulness, Limitations, and Methodological Recommendations. Draft paper, 2020.

[30] Rowen D, Brazier J, Tsuchiya A, Young T, Ibbotson R. It's all in the name, or is it? The impact of labeling on health state values. *Medical decision making.* Jan-Feb 2012;32(1):31-40.

[31] Pharmaceutical Benefits Advisory Committee (PBAC). Guidelines for preparing a submission to the Pharmaceutical Benefits Advisory Committee. Version 5.0. September 2016; Available at: https://pbac.pbs.gov.au/content/information/files/pbac-guidelines-version-5.pdf, Canberra, Australia. 12 August 2019.

[32] Scottish Medicines Consortium (SMC). Guidance to submitting companies for completion of New Product Assessment Form (NPAF). June 2019; Available at: https://www.scottishmedicines.org.uk/media/4527/20190626-guidance-on-npaf.pdf.

[33] Rowen DL, Azzabi Zouraq I, Chevrou-Severac H & van Hout B (2017) International regulations and recommendations for utility data for Health Technology Assessment. PharmacoEconomics, 35(Suppl 1), 11-19

[34] Mulhern B, Bansback N, Brazier J, Buckingham K, Cairns J, Devlin N, Dolan P, Hole AR, Kavetsos G, Longworth L , Rowen D et al (2014) Preparatory study for the revaluation of the EQ-5D tariff: methodology report.. Health Technol Assess, 18(12), vii-191.

Appendix

Table A1: Selected examples of NICE TAs and HST evaluations

TA/HST	ID number	Technology	Reference
TA	ID1211	Cannabidiol for adjuvant treatment of seizures associated with Dravet syndrome	[12]
TA	ID1308	Cannabidiol for adjuvant treatment of seizures associated with Lennox-Gastaut syndrome	[13]
TA	ID885	Short bowel syndrome - teduglutide	[10]
TA	TA552/ID1225	Liposomal cytarabine–daunorubicin for untreated acute myeloid leukaemia	[15]
TA	TA588/ID1069	Nusinersen for treating spinal muscular atrophy	[20]
TA	TA467	Holoclar for treating limbal stem cell deficiency after eye burns	[24]
TA	TA529	Crizotinib for treating ROS1-positive advanced non-small-cell lung cancer	[26]
HST	ID943	Cerliponase alfa for treating neuronal ceroid lipofuscinosis type 2	[16]
HST	ID1054	Voretigene neparvovec for treating inherited retinal dystrophies caused by RPE65 gene mutations	[17]
HST	HST8	Burosumab for treating X-linked hypophosphataemia in children and young people	[23]