Interpreting small differences in health utilities and incremental QALYS

REPORT BY THE DECISION SUPPORT UNIT

March 2026

Dan Pollard[1], Nicholas Latimer[1], Allan Wailoo[1], Emily Pulsford[1]

[1]Sheffield Centre for Health and Related Research (SCHARR), School of Medicine and Population Health, 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 Group is based at the University of Sheffield with members at York, Bristol, Exeter, Leicester, Warwick, Swansea 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 funding awarded by Wellcome through the "Effective Regulation and Evaluation of Digital Mental Health Technologies" project grant (ref: 226466/Z/22/Z). 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.

Funding/Support

This work was supported by Wellcome through the "Effective Regulation and Evaluation of Digital Mental Health Technologies" project grant (ref: 226466/Z/22/Z).

Role of the funder/sponsor

Wellcome had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of this report.

Acknowledgements

We would like to thank James Fotheringham, David Whitehurst, John Cairns, Lizzie Walker and Paul Tappenden who took part in our expert consultation. We also benefitted from discussions with several NICE staff: Lucy Beggs, Sophie Hughes, Bhash Naidoo, Dionne Bowie and Sarah Fothergill.

This report should be referenced as follows:

Pollard D, Latimer N, Wailoo A, Pulsford E. Interpreting small differences in health utilities and incremental QALYS. NICE DSU Report, 2026.

EXECUTIVE SUMMARY

Introduction

This project has arisen from the National Institute for Health and Care Excellence's (NICE's) work on digital mental health technologies (DMHTs), which are often characterised by relatively small differences in Quality Adjusted Life Years (QALYs) between the health technologies under consideration and their relevant comparators.

Small differences in QALYs (defined as differences less than 0.01) can raise challenges for decision makers and raises questions about: appropriate methods for conducting economic evaluations; how to characterise and represent uncertainty in economic evaluations; and the way results from economic evaluations should be interpreted by decision makers. Two key challenges are firstly most clinical studies are not powered to detect differences in health utilities or QALYs. This may lead to decision makers questioning whether DMHTs can be considered beneficial when small and non-statistically significant differences in point estimates of QALY differences are observed. Secondly, when incremental QALY differences are small, marginal changes in other parameters in an economic evaluation can lead to substantial changes in estimates of incremental cost-effectiveness ratios (ICERs). This volatility may be seen as a challenge both for interpretation of the results and decision-making.

We explored the overall research question of when observed mean overall QALY and/or component parts of health-related quality of life (HRQoL) differences between health interventions are small, how should these differences be interpreted?

We took multiple approaches to address this overall research question. We conducted a review of 9 case studies, all of which were economic evaluations with small QALY differences, to identify what current approaches are being taken when analysing and presenting results with small QALY differences. These case studies were selected pragmatically and spanned both model-based NICE appraisals and trial-based analyses in the academic literature. Our case study review focussed on how uncertainty in the analysis was presented and how conclusions or NICE committee decisions were influenced by the small differences in QALYs.

We then undertook a targeted literature review covering four topics of potential relevance, determined through discussion with NICE before the project started, when small QALY differences are observed:

i) a review of novel methods for or limitations of probabilistic analysis
ii) approaches for the presentation of decision uncertainty in the context of small QALY differences
iii) the relevance of the concept of Minimal Clinically Important Difference (MCID) when interpreting small QALY differences.
iv) the use of cost minimisation analysis (CMA).

From this evidence, we formed a series of draft recommendations and presented these to a group with expertise in health economics and/or decision making as members of NICE committees. Their feedback was used to inform changes that are embodied in the final set of recommendations.

Recommendations are provided in relation to a) the way in which small QALY differences should be treated analytically, b) how decision uncertainty should be conveyed, c) implications for decision makers and d) future research.

Case studies of economic evaluations with small QALY differences

Methods

We identified 9 known case studies from the academic literature and NICE appraisals known to the study team and colleagues at NICE. We described each study and the methods it used to provide context to their presentation of results. We summarised what methods were used to conduct and present the uncertainty analyses and how these were interpreted.

Results

8 out of the 9 case studies conducted a probabilistic analysis. All case studies presented the results of the evaluation with either incremental cost-effectiveness ratios (ICERs) or net benefits. 7 out of the 9 case studies presented the uncertainty in the results with either a cost-effectiveness acceptability curve (CEAC) or a cost-

effectiveness plane. Four out of the 9 case studies either had an interpretation of the small QALY differences in either the conclusion for academic papers or the published Committee discussions for NICE appraisals.

Summary

Standard methods for presenting uncertainty in economic evaluations, were used to present uncertainty in the results of our case studies. Where novel methods for presenting uncertainty were adopted in our case studies, they were either: an atypical presentation of standard techniques to describe uncertainty in economic evaluations (e.g. a bar chart of incremental net benefits); and/or, were not explicitly described as addressing uncertainty arising from the small QALY differences. Overall, standard methods are currently being used to present uncertainty in our case studies with small QALY differences.

Targeted review of literature

Methods

We conducted the targetted review of literature to identify novel literature on the following topics which were identified as high priority areas in our initial project discussions:

- novel uncertainty communication techniques for economic evaluations
- probabilistic analysis limitations and methods in context of small differences in simulated outputs
- use of MCIDs in interpreting QALY differences
- the role of CMA (if any) in interpreting cost-effectiveness analyses with small QALY differences.

We conducted three sets of searches, using backwards and forwards citation searching of 7 known papers covering areas of methodological uncertainty in economic evaluations with small QALY differences; a set of artificial intelligence assisted searches with Astra on probabilistic analysis methods; and traditional database searches. Screening was conducted according to inclusion and exclusion criteria by a single reviewer. Key points of the included studies were summarised narratively.

Results

No literature was found on the topics of probabilistic analysis limitations and methods in context of small differences in simulated outputs or the role of CMA in interpreting cost-effectiveness analyses with small QALY differences. 6 studies were found, that covered novel ways to present uncertainty in economic evaluations. None of these studies were specifically designed to present uncertainty in economic evaluations with small QALY differences. 7 studies were included that covered the use of MCIDs in interpreting QALY differences. Two studies suggested a role for MCIDs within the calculation of QALYs or in the interpretation of economic evaluations. Two studies described the MCID's of utility scores as being used to interpret epidemiological study results or in epidemiological study power calculations. Two studies were opinion pieces, outlining arguments against using MCIDs in economic evaluations. The remaining study was a highly cited methods piece on how to calculate MCIDs, that was included because it was a seed study in our search strategy.

We did not find any studies on novel techniques for conducting and interpreting probabilistic analysis or in applying CMA instead of cost-effectiveness analysis (CEA) when there are small QALY differences. This suggests that no changes in current practice should be adopted with respect to changing probabilistic analysis methods or in adopting a CMA approach after a CEA has been conducted and the QALY differences have been found to be small.

Whilst there are many novel ways of presenting uncertainty in economic evaluations that we found in our literature review, none of these were specifically designed to address the presentation and interpretation of uncertainty when an economic evaluation produces a small QALY difference. Analysts should consider if the novel techniques are useful for presenting uncertainty in their economic evaluation, and if they think they are useful, they should present uncertainty using these techniques.

Standard practice in economic evaluation is to not apply MCIDs when calculating QALYs or in interpreting CEA results. Whilst some literature suggests a role for MCID in economic evaluation, it is important to recognise the preference-based nature of

instruments used to inform estimates of HRQoL in economic evaluations that in turn inform QALY estimates. This implies that any difference is of relevance to inform decision making.

Expert commentator discussions

There were 6 key themes in our discussions with the expert commentators.

1) There was some consensus that cost and QALY differences, and what drives them, should be described in more detail when QALY differences are small, as would help understand what was driving uncertainty in the economic evaluation.
2) There was a general view that presenting many scenario analyses without rationale is not helpful. However, addressing if there are widely varying cost-effectiveness results this is useful to know. When there are small QALY differences the rationale for each scenario should be carefully thought through and presented.
3) Presenting net benefits rather than ICERs could be useful, as it avoids volatility seen in ICERs. However, this would need a very clear explanation, as net benefits may be misunderstood by Committees.
4) There was generally a view that there was not a need to address uncertainty in a different way when QALY differences are small.
5) There was some support for exploring domain-level data from HRQoL instruments to provide context for understanding small QALY differences.
6) The experts were hesitant to recommend that small QALY differences should trigger a closer inspection of the validity of the HRQoL instrument used in the analysis.

Recommendations

Our recommendations based on our interpretation of the case studies, the targeted literature review and the expert commentator discussions are:

Evidence and Analytical considerations

1) Explanation should be provided on the source and robustness of QALY differences when those differences are relatively small.

2) Small QALY differences in and of themselves do not indicate the need to conduct additional scenario analyses.

3) Explanation of the cost and QALY differences and what drives them should be provided, and Committees should consider these before discussing ICERs or net benefits. This will help Committees interpret ICERs and scenario analyses. Part of this explanation may include analysis of domain-level data from the preference-based instrument used, and comparison with expectations and data yielded from other outcome measures.

5) Committees should always consider the source of the evidence when appraising QALY differences, and this may provide particularly important context when QALY differences are small. For instance, small QALY differences may be interpreted differently, or with a different level of confidence, if they are derived from head-to-head RCT evidence, compared to indirect comparisons or observational data.

6) Decision uncertainty stems from the joint uncertainty in health and cost differences. It should be expressed in the same ways irrespective of whether QALY differences are small or not. Analysts should typically use the following methods for exploring uncertainty:

a. Probabilistic sensitivity analysis should be conducted, with all uncertain parameters appropriately determined and included in the analysis

b. To explore structural uncertainty, relevant scenario analyses should be explained and reported.

Presentation of results

7) All standard methods of presenting the results of an economic evaluation (results tables, cost-effectiveness planes, CEACs and cost-effectiveness acceptability frontiers (CEAFs)) should be presented and explained.

8) Results should be disaggregated to demonstrate whether volatile ICERs are associated with small differences in QALYs or costs. Total and incremental costs and QALYs should be presented. Where relevant, the QALYs and costs accrued in each modelled health state should be reported. The contribution to the total QALY difference associated with differences in length of life and quality of life could also be reported.

9) Consider using net benefits to present the results, as they may be easier to interpret because they are more stable than ICERs. However, net benefits should be explained clearly so that committee members who are used to interpreting ICERs as results of economic evaluations can clearly understand what is being shown and what this means.

Interpretation of results

10) The concept of MCID is not relevant to differences in preference-based instruments.

11) If the source of the differences and robustness of the QALY estimation are not apparent from the economic evaluation methods, analyses should be undertaken to identify whether small QALY differences are due to small differences in many patients, or large differences in few patients. Presenting the HRQoL or QALY data in a way that shows the distribution of relevant outcomes and highlights the impact of outliers/rare events could be useful.

12) Consider presenting HRQoL changes at the domain level for preference-based instruments, if the individual level data used to directly generate QALYs or health state utility values for a model is available. This will indicate whether changes in QALYs or health state utility values are driven by changes in specific domains of the HRQoL instrument and whether changes appear to be explainable.

Further Research

13) Research should be conducted on how to best report HRQoL changes at the domain level.

CONTENTS

TABLES

FIGURES

ABBREVIATIONS AND DEFINITIONS

ADHD	Attention deficit hyperactivity disorder
AD-SUS	Adult Service Use Schedule
AG	Assessment Group
BMJ	British Medical Journal
BNF	British National Formulary
CBT-SH	Cognitive behavioural therapy self-help
CE	Cost-effective
CEAC	Cost-effectiveness acceptability curve
CrI	Credible interval
CSLT	Computer speech and language therapy
DMHT	Digital mental health technology
DSU	Decision Support Unit
EAG	External Assessment Group
EQ-5D	EuroQol-5 Dimensions
EQ-5D-3L	EuroQol-5 Dimensions (3-level version)
EQ-5D-5L	EuroQol-5 Dimensions (5-level version)
EVPI	Expected value of perfect information
HRQoL	Health-related quality of life
IAPT	Improving Access to Psychological Therapies
ICER	Incremental cost-effectiveness ratio
iNMB	Incremental net monetary benefit
MBCT-SH	Mindfulness-based cognitive therapy self-help
MCID	Minimal Clinically Important Difference
NHB	Net Health Benefit
NHS	National Health Service
NICE	National Institute for Health and Care Excellence
NMB	Net monetary benefit

ONS	Office for National Statistics
PROMS	Patient-reported outcomes measures
PSA	Probabilistic sensitivity analysis
QALY	Quality-adjusted life year
RCT	Randomised Controlled Trial
SAVI	Sheffield Accelerated Value of Information
SD	Standard deviation
TA	Technology Appraisals

1. INTRODUCTION

This project has arisen out of NICE's work on digital mental health technologies (DMHTs), which are often characterised by relatively small quality adjusted life year (QALY) differences.

These small differences raised several issues, in particular:

- Most studies are not powered to detect small differences in health utilities or QALYs, which led to uncertainty about whether DMHTs are beneficial.
- When incremental QALY differences are small, small changes in other parameters in an economic evaluation can lead to substantial changes in estimates of incremental cost-effectiveness ratios (ICER), which can make interpretation and decision-making challenging.

In an initial investigation info these issues, the NICE team identified a new methodological guide for implementing and interpreting probabilistic sensitivity analysis (PSA) that highlights small QALY differences as an area for further consideration,[1] a commentary arguing that small QALY benefits can be badly misinterpreted,[2] and papers that explore Minimal Clinically Important Differences (MCID) and approaches to communicating and interpreting uncertainty.[3, 4] NICE wish to explore these issues in more detail, which is the purpose of this report.

The research question we address is as follows:

- When observed mean overall QALY and/or component parts of health-related quality of life (HRQoL) differences between health interventions are small, how should these differences be interpreted?

To address this question, the report is organised as follows:

- In Section 2, we define "small" differences in QALYs, explain other key relevant concepts and summarise NICE guidance around the presentation of uncertainty. We also describe two related issues, but which are not the key focus of this report: insensitive measures of HRQoL and cost-minimisation analysis.

- In Section 3, we present summaries of 9 case studies where the issue of small differences in outcomes for a cost-effectiveness analysis (CEA) was a feature. We include case studies from the published literature as well as NICE technology appraisals. We focus on analyses/appraisals of mental health topics and digital mental health technologies but also consider case studies from other settings where small QALY differences arise. We summarise how results were presented and reported in these case studies, to demonstrate currently used 'standard' methods and variation in current practice.
- In Section 4, we present a targeted literature review of papers that focus on small QALY differences in the health economics literature. Methods for conducting PSA and related analyses are well established and we do not attempt to extensively review this literature. Instead, we focus on newer literature that may point to novel claimed limitations in these methods, proposed approaches to the presentation of uncertainty which may improve on the standard approaches such as cost-effectiveness acceptability curves (CEAC) or cost-effectiveness plane scatterplots, and the relevance of the concept of MCID in relation to QALYs and/or HRQoL,
- In Section 5, we present recommendations regarding the way in which small QALY differences should be treated analytically, how uncertainty should be conveyed, and implications for decision makers.

2. DEFINING 'SMALL DIFFERENCES' AND RELATED CONCEPTS

In this section we outline key concepts that are crucial for the understanding of this report, and cover issues that are related to small differences, but which are not the focus of this report.

2.1. Defining small QALY differences

We define small QALY differences as differences of 0.01 or less. This was determined in discussion with NICE staff and from reviewing the distribution of incremental QALYs from a large sample of historical NICE appraisals. These QALY differences can either

mean differences from a probabilistic analysis or deterministic differences from an analysis that does not include uncertainty in analysis outputs. Techniques covered in this report may also be applicable and useful for problems where the QALY differences are bigger than the 0.01 QALY limit. Small QALY differences has been raised as an issue that is particularly of concern for DHMTs, this issue can happen in any economic evaluation, and is commonly seen in other areas such as diagnostics, screening, preventative technologies and where there are multiple health technologies.

If the mean QALY difference is small, it will often be the case that decision uncertainty is relatively high. This is because when QALY differences are small in absolute terms, estimates of cost effectiveness become highly volatile as changes both to QALYs and other parameters. This is an inevitable consequence of the ICER being calculated as a ratio statistic. This uncertainty may or may not be captured within a probabilistic analysis. For example, uncertainty in the parameter inputs to a chosen time to event parametric model should be included in the uncertainty in a probabilistic analysis but the effect of choosing alternative time to event parametric model (or even non-parametric models) would not and could only be explored using scenario analyses.

However, this may not be the case for all analyses with small QALY differences. For example, in screening studies only a small proportion of people could benefit from a new screening test (the ones who have an undetected condition at the time the test is conducted). Consequently, small QALY differences are likely. However, if there is high decision certainty around the test characteristics of the new and old tests, the decision certainty may be high in this case.

2.2. Key concepts for the understanding of uncertainty in CEA

2.2.1. Probabilistic Analysis

Probabilistic Analysis, also known as Probabilistic Sensitivity Analysis (PSA), is where samples of the joint uncertainty in costs and QALYs in the analysis is assessed. This is done as sampling every input parameter in a model from an assigned probability distribution or directly sampling costs and QALYs in a trial-based analysis. The process is repeated multiple times (usually at least 1,000 times) and the results are

then saved for each iteration. The conclusions of the economic analysis are based on the mean differences in the results across all uncertainty analyses.

If the deterministic base case and the probabilistic base case economic analyses differ, the probabilistic results are preferred in NICE appraisals.[5]

2.2.2. Deterministic Sensitivity Analysis

Deterministic sensitivity analyses all work on the principle that using the means inputs in a health economic evaluation predicts the probabilistic model well. In many, but by no means all, economic evaluations using mean parameter values in a model adequately approximates the probabilistic results. If they are not equal, then analysts may need to apply some of the types of deterministic sensitivity analyses (for example scenario analyses) in a probabilistic and not a deterministic model.

There are three types of scenario analyses typically conducted using deterministic models, these are:

2.2.2.1. Scenario Analyses

In a scenario analysis an alternative data source is used, or alternative assumption is made, and the analysis is rerun. In a trial based economic evaluation this may involve changing the HRQoL instrument used to generate utilities (for example from EQ-5D-5L to SF-6Dv2) and in a model based evaluation this can involve changing a parameter or set of parameters in the model (for example changing the preferred time to event model for predicting overall survival from using an exponential parametric model to a generalised gamma parametric model). In both cases the results of the economic analysis is presented in comparison to the base case analysis and the magnitude of the change in incremental costs, incremental QALYs and ICERs are compared. Usually, some comment is made on how impactful the change is and whether the change is an important factor to consider when interpreting the uncertainty in the ICER. Scenarios can be combined, especially in NICE appraisals where external assessment groups (EAGs) may wish to present a different set of assumptions to a company as being the most reasonable for decision making by a Committee. In this

case showing each change, the EAG wishes to make in isolation and finally in combination can be useful for representing uncertainty.

2.2.2.2. One way sensitivity analysis

This sensitivity analysis is usually only presented in model-based analysis, albeit it can be done in trial-based analyses on some inputs (for example the unit costs). Each parameter is varied one at a time within an arbitrary range decided by the analyst. Common examples are using the upper and lower bounds of the 95% confidence intervals or +/- 20% of the mean parameter value. The ICERs or incremental net benefits are then typically shown on a Tornado plot if only two options are being compared in the economic evaluation.

2.2.2.3. Threshold analyses

Sometimes alternative data sources are not known or the values that are known are subject to significant amount of bias, leaving the true value unknown. If a model based economic evaluation is being used for the economic evaluation, a threshold analysis could be considered. In this analysis, the unknown input parameter is varied across its plausible range. The model results are recorded for each tested value of the parameter. This means analysts and committees can see the parameter value changes the decision at a particular threshold ICER. The value can then be checked against a clinically plausible range for the parameter. If the value is inside the plausible range, then it may be subject to considerable debate about what decision to make.

2.2.3. Net benefit

Net benefit can either be defined in monetary terms as net monetary benefit (NMB) or net health benefit (NHB). The formulae are as follows:

$$NMB = QALYs * Threshold\ ICER - Costs$$

$$NHB = QALYs - \frac{Costs}{Threshold\ ICER}$$

The threshold ICER needs to be defined for this analysis. The NICE technology appraisal (TA) and highly specialised technologies (HST) guidance: the manual, recommends that two sets of NHB statistics are calculated, one using £20,000 per QALY gained threshold and one using £30,000 per QALY gained.[5]

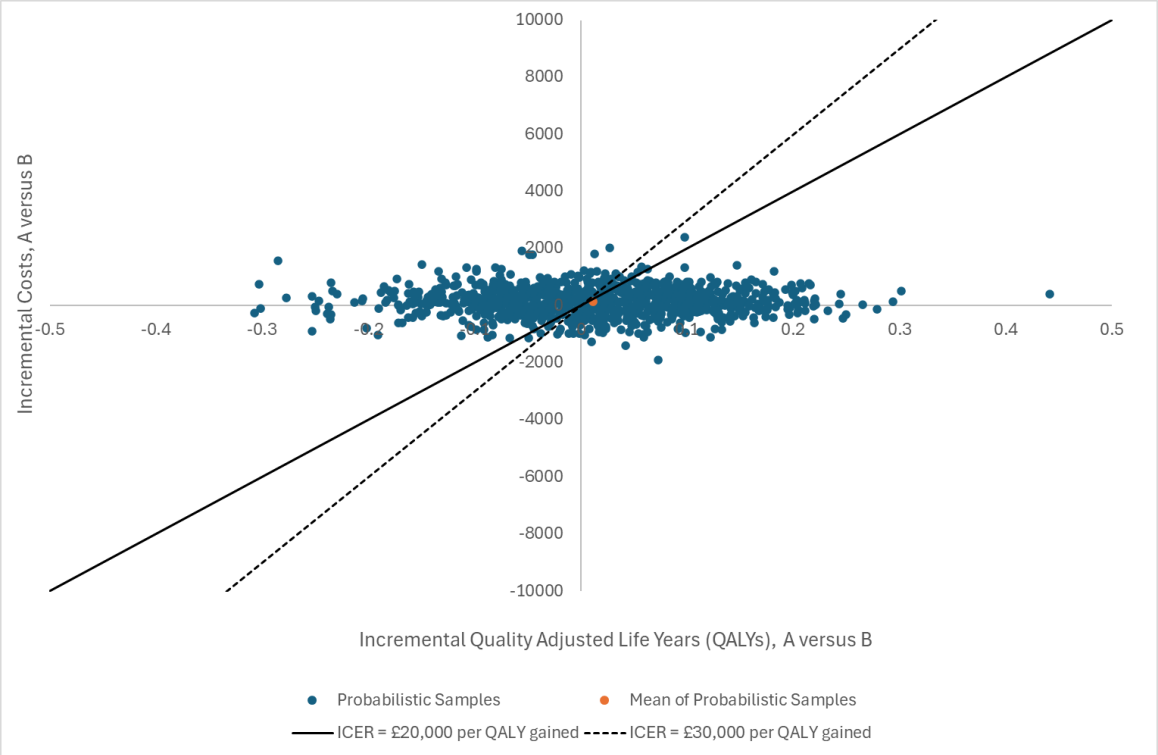
The decision rule when working with either net benefit is simple, the strategy with the biggest net benefit is economically optimal. This will be consistent across the two net benefits, as the only difference is the units. NMB is in monetary terms and NHB is in QALY terms. However, using net benefits in decision marking requires a predefined threshold ICER to exist, which in a NICE appraisal is Committee not an analyst decision.

NMB is useful for easily drawing cost-effectiveness acceptability curves (CEACs) as they can be calculated at thresholds of £0 per QALY gained whereas the NHB is more difficult to use for this purpose. Division by 0 is not mathematically possible. The NICE Methods and process Manual has a preference for presenting NHB, with NMB being an outcome that can also be presented.[5]

2.2.4. Cost-Effectiveness Plane

The cost-effectiveness plane shows the incremental costs and incremental QALYs in a scatterplot. Each probabilistic analysis iteration is drawn as a point on the scatterplot, and often the mean of the probabilistic analysis iterations is indicated with a different marker in addition to the probabilistic analysis iterations. It is common to also draw a diagonal line on the cost-effectiveness plane indicating threshold ICERs. An example Cost-Effectiveness plane, comparing two strategies A and B is shown in Figure 1.

Figure 1: An example cost-effectiveness plane.

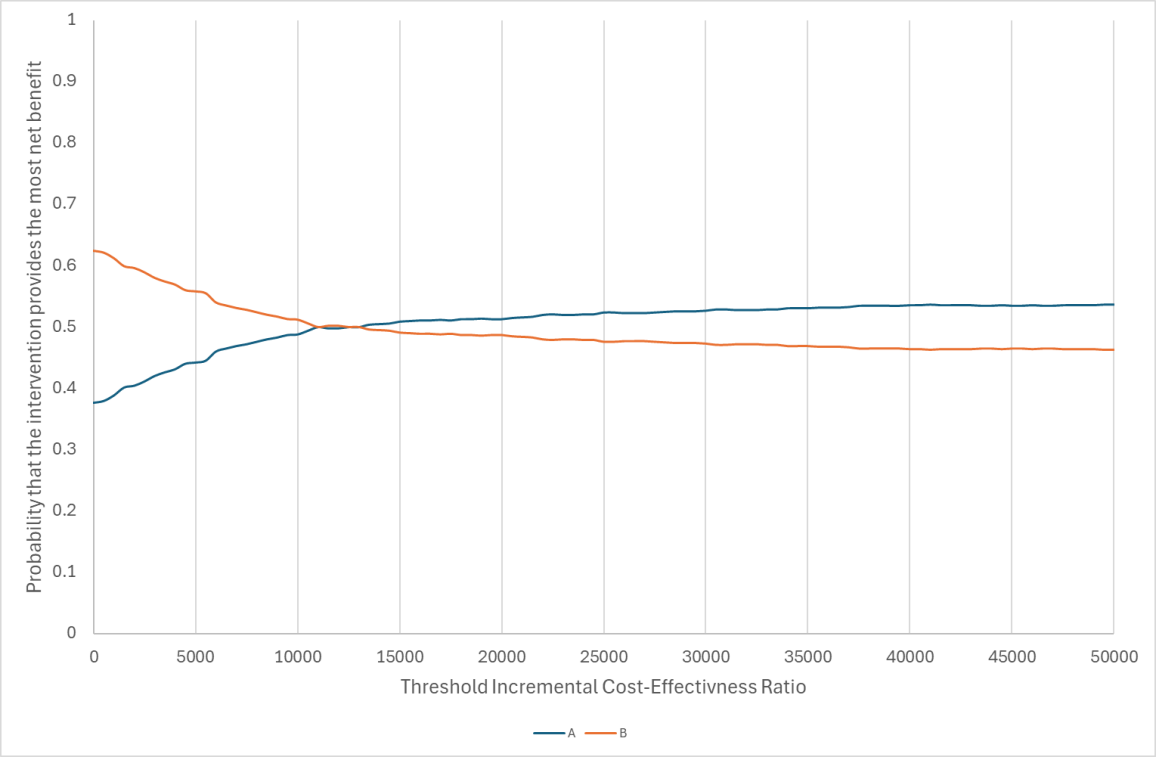


If more than two strategies are being considered in an economic evaluation, it is common for the scatterplots to significantly overlap. As such, a cost-effectiveness plane might not be shown in these circumstances.

2.2.5. Cost-Effectiveness Acceptability Curve

CEACs diagrammatically represent the probability that each intervention under consideration provides the most net benefit on the y axis and the threshold ICER for the decision maker on the x-axis.[6] This is used to supplement the cost-effectiveness plane and the central estimate of cost-effectiveness to show uncertainty in the base case probabilistic results. An example cost-effectiveness acceptability curve is shown in Figure 2. This graph shows the same underlying data as is shown in the example cost-effectiveness plane in Figure 1.

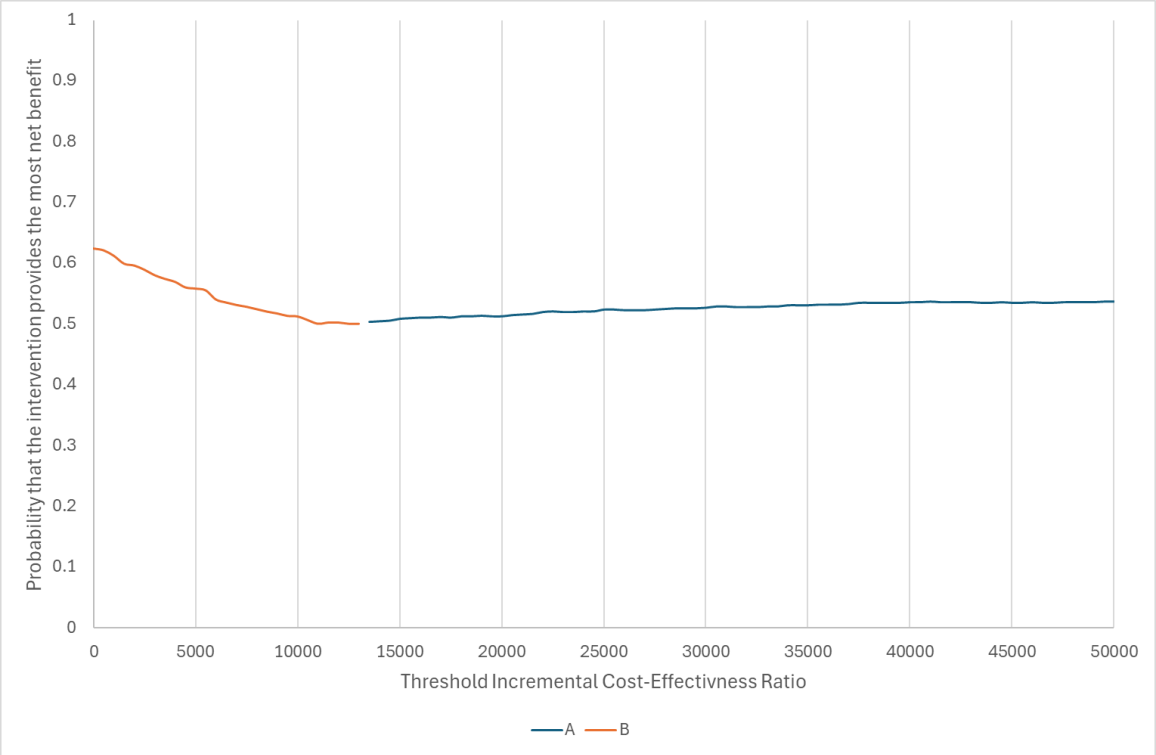
Figure 2: An example cost-effectiveness acceptability curve



2.2.6. Cost-Effectiveness Acceptability Frontier

The Cost-Effectiveness Acceptability Frontier (CEAF) is an extension of the CEAC.[7] It only shows a probability that each intervention is cost-effective if that intervention provides the most expected net benefit. This means that sometimes an intervention is show on the CEAF that has a less than 50% chance of providing the most net benefit. This is because it only provides the most net benefit in a minority of simulations, but once the magnitude of the net benefit is accounted for the mean net benefits still suggest that the strategy will provide the most (mean) net benefit, so is optimal. An example CEAF is shown in Figure 3. Again, the data in this graph is the same as in Figure 1 and Figure 2.

Figure 3: An example cost-effectiveness acceptability frontier.



2.2.7. Value of Information

Value of information analyses are conducted on probabilistic model results. There are three levels typically used in health economic evaluations. There is the expected value of perfect information (EVPI), which assesses the health economic value of being certain about the estimated costs and QALYs in the evaluation over all input parameters. There is the expected value of perfect parameter information (EVPPI), which assess the health economic value of being certain on an input parameter or subset of input parameters. There is also expected value of sample information (EVSI), but this is less useful for assessing uncertainty in economic results, as it is the value of conducting a particular study to reduce uncertainty in a subset of the inputs to the economic analysis.

2.3. NICE guidance on addressing uncertainty in CEA

The NICE technology appraisal and highly specialised technologies guidance makes 24 points on addressing uncertainty in CEA (points 4.7.1 to 4.7.24).[5] The key points raised are:

- assessment of uncertainty is important for Committees (4.7.1)
- the aim of uncertainty analyses should be to quantify the decision uncertainty (4.7.2)
- structural uncertainties should be explored in economic analysis (4.7.3 to 4.7.6)
- uncertainty in the choice of data sources should be explored in economic analyses (4.7.8 to 4.7.9)
- uncertainty in parameter precision should be explored with a probabilistic analysis, and probabilistic analyses are preferred over deterministic analyses. Deterministic analyses can be considered if shown to be appropriate for an individual decision problem, either through predicting the probabilistic model results well or a probabilistic analysis being infeasible (4.7.10 to 4.7.13)
- the cost-effectiveness plane, cost-effectiveness acceptability curve are described as typically being good ways of showing uncertainty in economic results (4.7.15)
- if the deterministic model well predicts the probabilistic model, one way sensitivity analyses can explore which individual or sets of parameters to which a decision is most sensitive. This is less appropriate if the deterministic model does not predict the probabilistic model results well. (4.7.17)
- Threshold analysis can be important for high uncertain parameters to identify at what parameter value, the decision switches at, at a given threshold cost per QALY gained. (4.7.22)

2.4. NICE guidance on assessing HRQoL instrument suitability.

The NICE technology appraisal and highly specialised technologies guidance, details that EQ5D may be insensitive to a health condition in point 4.3.10.[5] It recommends that if it is not the most appropriate measure than qualitative evidence should be generated to show that there is a lack of content validity. Additional evidence should be provided around EQ-5D performance of construct validity and responsiveness in the target patient population. This should be derived from a synthesis of the published literature.

Whilst it does not state that small QALY differences are something to consider, if the EQ-5D was insensitive and was used to estimate utilities, it would produce small QALY

differences between arms as respondents would answer the EQ-5D questionnaire similarly regardless of their health.

2.5. Small QALY differences due to insensitive HRQoL instruments

This report assumes that the QALY differences are genuinely small differences between potential health care strategies. Small differences can be found for other reasons. A common reason for this in health economic analyses would be that the chosen instrument for measuring patient self-reported HRQoL is insensitive for the condition and technology under appraisal. The correct approach to making a decision in these conditions would be to use an appropriate sensitive measure that reflects the impact of the intervention more accurately. This could be particularly important in DMHTs as it may be the case that standard instruments, such as the EQ-5D-5L, may not be sensitive to changes in HRQoL, for patients with some mental health problems.

The NICE technology appraisal and highly specialised technologies guidance details the following points on when using an instrument other than EQ-5D is appropriate.

"*4.3.10 In some circumstances the EQ-5D may not be the most appropriate measure. To make a case that the EQ-5D is inappropriate, provide qualitative empirical evidence on the lack of content validity for the EQ-5D, showing 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 (that is, it does not perform as would be expected) 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. These 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*"[5]

If the reason for the estimated differences in QALYs between competing interventions is thought to be the insensitivity of the instrument used to measure HRQoL, then evidence supporting or suggesting this should be provided. New evidence using an appropriate instrument to measure HRQoL should be generated. This report does not

cover these cases – instead, our focus is on situations where there are small QALY differences even when an appropriate and sensitive measure is used.

This was covered as being a key point of Committee discussions in HTG748 on digital therapies for chronic tic disorders and Tourette syndrome, were the Committee concluded "*evidence is lacking on the effect of digital therapies on quality of life*." There would be no analysis solution in an existing economic evaluation to address this concern. The only solution would be to generate evidence from a new valid instrument on the effects of digital therapies on quality of life for people with chronic tic disorders and Tourette syndrome.[9]

2.6. Minimal Clinically Important Difference (MCID) in QALYs

There has been much work in estimating MCIDs in many outcome measures, as well as patient reported outcome measures and utility scores. The basic principle is that at some level the effects can get to be so small, that they are no longer relevant. Whilst this works somewhat at the individual level, it can lead to problems in interpretation of studies which can only estimate population level effects.

For example, if an intervention increased the proportion of patients getting a MCID in their health from 10% with usual care to 60% with an intervention. This intuitively seems like there are a sufficient number of patients obtaining a meaningful improvement in their health that this should be interpreted as improving patient outcomes. As an additional 50% of the cohort are getting the MCID. However, if the population level effects were calculated by statistically analysing the individual level data, the difference in health between the two arms was approximately half of the MCID. If an MCID was applied to this population effect, it may be erroneously concluded by analysts that an important difference has not being shown. It is potentially misleading to use MCID to determine if a mean change in a population is important or not.

There has been much work on the MCIDs in patient reported outcomes measures (PROMS), with step-by-step guides on how to estimate these differences being published in the BMJ.[10] MCID in SF-6D utilities have been estimated by Walters and

Brazier.[11] They estimated these differences by using data from multiple studies from which SF-36 data was collected. From this SF-6D utility scores and an anchor question could be derived and the MCIDs statistically estimated. They found that the MCIDs was between 0.010 to 0.048, depending on the dataset used to estimate the MCIDs.

However, it is not apparent that this should apply to patient reported outcome measures that are used to generate utilities. Whitehurst *et al* put forward the argument that MCID is irrelevant, because the valuation exercise for PROMS that is used to calculate utilities involves an assessment of the relative importance of each domain of the PROM by those taking part in the valuation study.[12] Consequently, all differences from this are meaningful differences, even if they are small.

2.7. Cost-minimisation analysis

If it is appropriate to assume that there are no differences in health between study arms, then a cost minimisation analysis may be considered, as the decision problem then simplifies to doing the option that is expected to the cheapest over a lifetime. However, a NICE DSU report detailed that this should only happen under very limited circumstances.[13] The key recommendations relevant to post-hoc deciding whether technologies are equivalent are: it should be carefully considered whether it is plausible that two technologies are equivalent (recommendation 3); and, if determined that a technology is non-inferior to another, then it must be cheaper with no margin for increased cost (recommendation 12). We do not go into the full rationales for conducting cost-minimisation analysis in this report.

3. Case studies on economic evaluations with small QALY differences

3.1. Introduction

The purpose of reviewing 9 case studies was to identify current practice with respect to analyses and reporting when QALY differences are small. We reviewed case studies from NICE appraisals, as well as from the wider academic literature, to determine whether different techniques were used in these different settings.

3.2. Case study selection

Case studies were selected pragmatically, drawing on cases familiar to the research team and through consultation with colleagues at NICE. Reviewing up to 10 case studies was expected to be sufficient for identifying key trends in analytical techniques and reporting. Studies were selected on the following criteria: the topic area, some DMHT topics were included but also a wide range of clinical areas were addressed; type of economic evaluation method (within trial analysis or economic model); and, whether the case study was published in the academic literature or as part of a NICE appraisal. QALYs differences had to be clearly reported for a study or appraisal to be included. Studies were included with QALY differences up to 0.05, however studies if the QALY differences of less than 0.01 where preferred. For academic papers data from the main text was extracted. For NICE appraisals, the Committee discussions and the evidence assessment group reports (if available) were extracted. In cases where evidence assessment group reports were not available, the company submissions were used instead.

3.3. Case studies overview and summary of methods

Nine case studies were selected, four being NICE appraisals and five being academic papers. A summary of how each case study presented the results of their analyses is in Table 1. Four case studies explicitly addressed small QALY differences in the conclusions or the in the published Committee discussion for NICE appraisals. Three of these case studies from the academic literature, but one NICE appraisal published Committee discussions mentioned this as a consideration. An overview of how the

case studies have addressed uncertainty in the results is provided Table 1. Most case studies did a PSA and presented incremental cost effectiveness ratios (7 out of 8). Most case studies also presented a CEAC or a cost-effectiveness plane to present the PSA results (5 out 8). Also, of note is that a minority of studies presented Net Monetary Benefit or incremental Net Monetary Benefit to present the results (4 out of 8). 3 out of 8 case studies did other analyses to present uncertainty in the economic results: NICE HTG676 did a population expected value of perfect information analysis; NICE HTG729 presented the sensitivity analysis showing changes in incremental net monetary benefit rather than changes in the ICERs; NICE TA878 presented changes in the incremental net monetary benefit for each scenario analysis. 4 out of 8 case studies noted the small incremental QALYs as important in either the conclusions or Committee discussions (for NICE appraisals). In our case studies, the small QALY differences were more typically in the discussions of the academic studies than the Committee discussions of the NICE appraisals, with only one appraisal Committee explicitly including the small QALY differences issue in the written minutes of the meetings. No case studies appeared to clearly and explicitly link the small QALY differences to uncertainty in the probabilistic analysis methods. NICE HTG 729 conducted a threshold analysis on several key input parameters to the economic model, to see at what parameter values the decision changed. This was to address uncertainty in the true values of these parameters in the literature, as there were concerns about the validity of all available sources for these parameters. A detailed description of the analysis methods, results presentation and whether small QALY differences were addressed in the discussion for each case study is provided in section 3.4.

Overall, we believe that standard methods for handling uncertainty have been conducted in the case studies that we have selected, with no adjustment to these methods due to the small QALY differences being evident. The key change appears to be that more studies than we would typically expect place emphasis on net monetary benefit to present their results rather than ICERs, as just under half the studies have done this.

Table 1: A summary of the presentation of uncertainty in the small QALY gain case studies.

Study	Probabilistic Analysis	ICERs	Net Monetary Benefit	Net Health Benefit	CEAC	Cost-effectiveness plane	Other analysis	Interpretation accounts for small QALY differences
Digital Mental Health Technologies								
NICE HTG676[14]	✓		✓				✓	
NICE HTG729[15]	✓	✓	✓		✓		✓	
Richards 2020[16]	✓	✓			✓			
Other mental health interventions								
LIGHTmind[17]	✓	✓			✓	✓		✓
NICE TA367[18]	✓	✓			✓	✓		✓*
Other digital technologies								
Big Cactus[19]	✓	✓			✓	✓		✓
Other examples with small QALY differences								
HUBBLE[20]	✓	✓			✓			✓
NICE TA878[21]		✓	✓				✓	
Mandrik 2024[22]	✓	✓	✓			✓		

ICERs, incremental cost-effectiveness ratios; CEAC, cost-effectiveness acceptability curve; QALY, quality adjusted life year

*, interpretation of small QALY differences was mentioned in National Institute for Health and Care Excellence Committee decision making

3.4. Case Studies in detail

3.4.1. NICE HTG676

NICE HealthTech guidance 676 assessed digitally enabled therapies for adults with anxiety disorders.[14] Improving Access to Psychological Therapies (IAPT) was included as a comparator. The key effectiveness evidence was source from a rapid review of cohort studies, pilot studies and randomised controlled trials of the considered therapies. Many different digitally enabled therapies for adults with anxiety disorders were included in this appraisal.

A decision tree model structure was used to assess the technologies over a time horizon of 15 months. It was assumed that a proportion of the population responded fully to treatment at 4-6 months, and this proportion would remain in full response over the remaining 15-month time horizon. For the proportion of the population who did not respond they would either have their treatment escalated if IAPT was being modelled or would be referred to IAPT if they were on a digital technology after 7-9 months. Patients could then either respond or not to this second line of therapy. Again, if patients responded they remained in a response health state until the end of the 15-month time horizon. This model structure was used to model: Generalised Anxiety Disorder, Body Dysmorphic Disorder, Other Anxiety Descriptors, Post Traumatic Stress Disorder, social anxiety disorder. However, the body dysmorphic disorder scenarios will not be discussed further here as only one strategy had full reporting of the economic outcomes.

Utilities were obtained from the literature, with key study being a study by Gega *et al*.[23] Costs were obtained from commercial providers of the digital mental health technologies, and other costs were obtained from standard UK unit costs sources and for resource use the population were assumed to follow resource use pathways according to their health state.

The primary results from HTG 676 are summarised in Table 2. The approach taken was to calculate net monetary benefits and to calculate the 95% CI around these

coming from the probabilistic analyses conducted with the economic model. No deterministic results appear to have been presented.

Table 2: A summary of the primary results of the health economic evaluation in HTG 676 (adapted from External Assessment Group (EAG) report[14], Tables 19,21,22,23)

Strategy	Costs	QALYs	NMB*	NMB* 95% CI
Generalised Anxiety Disorder†				
IAPT current pathway	£494	0.81	£15,771	12707, 18615
Silver Cloud	£410	0.81	£15,811	12641, 18714
Other Anxiety Descriptors†				
IAPT current pathway	£587	0.81	£15,538	12459, 18455
Silver Cloud	£459	0.81	£15,725	12540, 18625
Post traumatic stress disorder†				
IAPT current pathway	£1,289	0.72	£13,044	9807, 16214
Spring	£496	0.75	£14,542	11362, 17444
Social Anxiety Disorder†				
IAPT current pathway	£1,433	0.73	£13,233	10060, 16141
Silver Cloud	£1,168	0.74	£13,578	10520, 16519

QALYs, quality adjusted life years; NMB, net monetary benefit; CI, confidence interval

*, estimated at £20,000 per QALY gained; †, one or more options are not presented as either costs, QALYs or NMB were confidential

One way sensitivity analyses were conducted, but most of these were redacted from the EAG report, so it is unclear how these results were presented to Committee members. The public Committee papers do present the results of one scenario analysis on the recovery rate parameter fully. In this scenario, a full set of results tables were repeated for each of the four conditions being appraised. Furthermore, a population expected value of information analysis was conducted on a population of 330,000 people per year, assuming that any new research to reduce uncertainty would be relevant for 5 years. No CEACs or cost-effectiveness planes appear to have been presented by the EAG.

The Committee discussions on the economic evidence focused on the gaps in the clinical literature in this appraisal, and the Committee considered that digital mental

health technologies in these cases were likely to cost less but may be less effective than standard care so may end up costing more in the long run. On this basis the Committee concluded the cost and resource section that "*further evidence on clinical effectiveness and resource use is needed to reduce uncertainty*" (point 3.14).[14] The Committee papers do not discuss the small incremental QALYs.

3.4.2. NICE HTG 729

NICE HealthTech guidance 729 looked at digital health technologies for assessing attention deficit hyperactivity disorder (ADHD).[15] The key options being compared are: QBtest and usual care. The guidelines more widely were assessing other digital mental health technologies for ADHD, namely: EFSim Test Web Version, Nesplora Attention Adults Aquarium. The key effectiveness evidence for the economic assessment came from the AQUA trial.

The EAG used state transition models to capture three parts of the ADHD pathways: One state transition model for ADHD diagnosis; one for the follow up pathway for people diagnosed with ADHD and one for the follow up pathway for people without ADHD. Utility values were obtained from various literature sources. Resource use was based on previous published literature and NICE Technology Appraisals (TA), with standard sources for reference costs being used to estimate total costs. Differences between treatments were modelled as different probabilities of diagnosing ADHD at 6 months post-assessment.

A more extensive set of cost-effectiveness results were presented. The base case results are replicated in Table 3. Unlike most other case studies, incremental net monetary benefits and probability of providing the most net benefit in the PSA at £20,000 and £30,000 per QALY gained are also included in the main results table. These values where the upper and lower limits of NICE's cost-effectiveness threshold at the time of the analysis.

Table 3: The base case probabilistic results in NICE HTG 729 (adapted from Table 28 in the EAG report[15])

Strategy	Mean Costs	Mean QALYs	Incr Costs	Incr QALYs	ICER	£20,000 threshold		£30,000 threshold	
						iNMB	Prob CE	iNMB	Prob CE
Standard	£6,005	6.9083	-	-					
QbTestAll	£6,243	6.9469	£238	0.0385	£6184	£533	0.922	£918	0.884

QALYs, quality adjusted life years; Incr, incremental; ICER, incremental cost-effectiveness ratio; iNMB incremental net monetary benefit; CE, cost-effective

Uncertainty in the results were explored using cost-effectiveness planes, CEACs and a table detailing how the model results changed in each of conducted scenario analysis. Four threshold analyses were conducted, changing the values of four key parameters in the model (QbTestAll sensitivity, proportion of patients with a missed diagnosis, prevalence of ADHD, proportion of patients with an unclear diagnosis). Changes in the incremental net monetary benefit were plotted against the various values of the parameters altered in each threshold analysis.

The Committee did not explicitly consider the uncertainty in the QALYs in their guidance. The key issues identified by the Committee were mostly focused on the lack of evidence in two other populations (complex prior cases and using the test to evaluate response to treatment).

3.4.3. Richards 2020

Richards *et al*. assessed the cost-effectiveness of internet delivered cognitive behavioural therapy (iCBT), compared to a wait-list control in a UK setting. Participants in the control arm received the intervention treatment after 8 weeks.[16]

Within trial health economic analysis methods were undertaken, data was collected from trial participants on self-reported resource used and EQ-5D-5L were collected. Costs were calculated by combining unit costs with self-reported resource use. QALYs were calculated by first calculating utility scores for the EQ-5D-5L using the van Hout cross walk algorithm and the EQ-5D-3L value set for the UK and secondly calculating

QALYs from the observed utility scores using the area under the curve method. The costs and QALYs were calculated over an 8-week time horizon in the base case. The base case used the complete cases in the intention to treat population. A scenario analysis was conducted where missing data were imputed using multiple imputation using chained equations.

Probabilistic analysis was undertaken by non-parametric bootstrapping. To adjust for pre-baseline costs and baseline utility score, OLS regression models were fitted within each bootstrap and the coefficient corresponding to the difference in costs and QALYs between study arms was extracted. Scenario analyses were conducted in which: missing data where imputed; cost and QALY differences where extrapolated.

Scenarios were presented on the CEAC, but not in other formats, such as in a results table or on the cost-effectiveness plane in the main text. The small differences in QALYs are not explicitly mentioned in the discussion of the paper.

3.4.4. LIGHTmind randomised controlled trial.

The LIGHTmind Randomised Control Trial (RCT) looked at two types of cognitive behavioural therapy, practitioner-supported mindfulness-based cognitive therapy self-help (MBCT-SH) versus cognitive behavioural therapy self-help (CBT-SH) for adults with mild to moderate depression.[17] Within trial analysis health economic methods were undertaken , so the sole source of clinical evidence was from the LIGHTmind RCT.

Utility scores were calculated at the individual level using participants self-reported EQ-5D-5L and mapping this to an EQ-5D-3L response using the van Hout *et al* mapping algorithm and an EQ-5D-3L value set.[24-26] QALYs were calculated using the area under the curve approach. Resource use data was collected by individual participants using Adult Service Use Schedule (AD-SUS), with unit costs from UK national sources (e.g. NHS reference costs, Unit costs of Health and Social care) to calculate total costs. The costs and QALYs were estimated over 42-week time horizon, as this was the duration of the trial.

Cost and QALY differences were estimated using regression within a bootstrapping process. Although the regression procedure (e.g. ordinary least squares, generalised linear models, generalised estimated equations) fitted within each bootstrap iteration is unclear, albeit with the control covariates provided. In the base case a complete case analysis was conducted.

The base case results were presented as in Table 4, whilst it was stated that MBCT-SH was dominant, there is no explicit detail in the results covering the small differences in QALYs. Uncertainty was shown in the main text and supplementary materials through a cost-effectiveness acceptability curve, which is reproduced in Figure 4. In spite of the small QALY differences in this analysis, these results are relatively certain with the probability of being cost-effective being over 80% up to thresholds of £50,000 per QALY gained. In the supplementary material, sensitivity analyses are presented including a sensitivity analysis using imputed data.

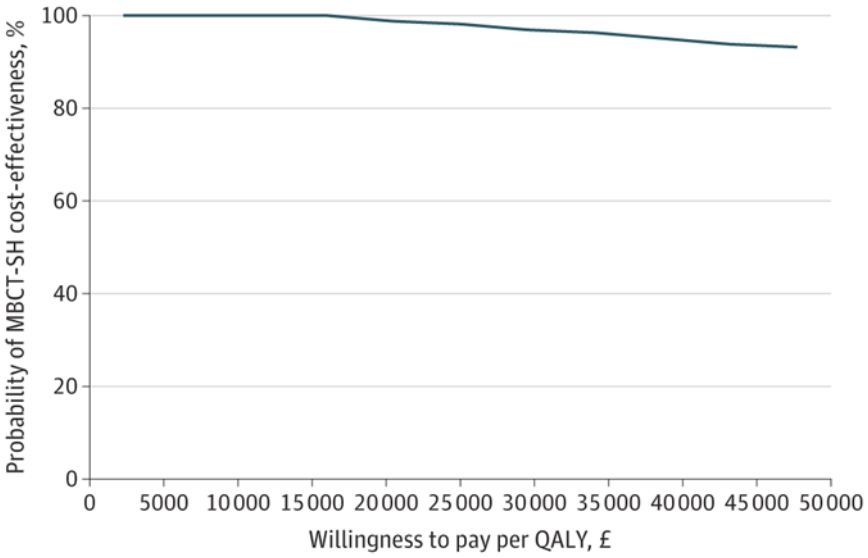
Table 4: The base case results of Strauss *et al*.[17] (adapted from Table 3 and results text)

	Mean (SD)		**Unadjusted MBCT-SH v CBT-SH**		**Adjusted* MBCT-SH v CBT-SH**	
	MBCT-SH (n=145)	CBT-SH (n=147)	Mean difference (95% CI)	P value	Mean difference (95% CI)	P value
Costs	£944 (£1,102)	£1,574 (£2,704)	-£627 (-1109 to -145)	0.01	-£526 (-£908 to -£144)	0.007
QALYs	0.65 (0.13)	0.65 (0.12)	0 (-0.03 to 0.03)	0.82	0.01 (-0.01 to 0.03)	0.48
ICER	-	-	Dominant	-	Dominant	-

SD, standard deviation; MBCT-SH, practitioner-supported mindfulness-based cognitive therapy self-help; CBT-SH, cognitive behavioural therapy self-help; CI, confidence interval; QALYs, quality adjusted life years; ICER, incremental cost-effectiveness ratio

*, adjusted by regression for baseline variable of interest, baseline utility, baseline PHQ-9, site and follow up time

Figure 4: Cost-effectiveness acceptability curve for the base case analysis in Strauss et al.[17] (reproduced under the CC-BY licence)



The authors conclude that MBCT-SH is cost-effective, because it is cheaper and has similar QALY outcomes. The rational for this interpretation that the small incremental QALY outcomes were similar and not different, is not stated.

3.4.5. NICE TA367

NICE technology appraisal 367 assessed, Vortioxetine, for treating major depressive episodes.[18] The key clinical study underlying the evidence submission was the REVIVE RCT (efficacy study) and TAK318 (tolerability study), although the health economic analysis primarily used the REVIVE RCT, which compared Vortioxetine to placebo. The analysis primarily compared the use of many treatments, Vortioxetine, Venlafaxine, Citalopram, Sertraline, Agomelatine.

The economic evaluation was conducted using a decision tree that modelled whether patients had a sustained remission, stopped treatment or switched to 3rd line therapies. For patients who did not achieve switched to a 3rd line therapy, their outcomes on the 3rd line therapy were modelled using a state transition model. The state transition models had different time horizons depending on what time patients left the model. The effectiveness estimates were obtained from an indirect treatment comparison of

the REVIVE to other RCTs for the other treatments in the decision problem to inform the probability of remission and withdrawal. Health State Utilities Values were estimated by fitting regression models to the EQ-5D data collected in the REVIVE RCT. Treatment pathways for the health states were assumed and costed according to standard UK sources (e.g. British National Formulary (BNF), NHS reference costs, etc).

A standard set of base case results were presented by the company, apart from a full incremental analysis based on the mean Total Costs and mean Total QALYs from the probabilistic analysis was not presented. A replication of these results is presented below in Table 5.

Table 5: The base case results in NICE TA367 (modified from CS[18], Table B82

Comparators	**Total Costs**	**Total QALYs**	**Incremental Costs**	**Incremental QALYs**	**ICER (£ per QALY gained)**
Deterministic					
Venlafaxine	£964	0.675	-	-	-
Vortioxetine	£971	0.694	£7	0.019	£378
Citalopram	£976	0.664	£5	-0.030	Dominated
Sertraline	£977	0.664	£0	-0.001	Dominated
Agomelatine	£1,082	0.676	£105	0.012	Dominated*

ICER, incremental cost-effectiveness ratio; QALYs, quality-adjusted life years

*, extendedly dominated by Vortioxetine. Costs for 2014

Tornado plots are presented in a pairwise fashion. Pairwise cost-effectiveness planes are presented. Pairwise and full incremental Cost-Effectiveness Acceptability curves are presented. Although the probability that Vortioxetine is cost-effective only established for each pairwise comparison. Structural sensitivity analyses are presented, by presented a new full incremental analysis table for each of the scenarios.

The Committee noted that the QALY differences were small in point 4.17 of their discussion. They noted that there were two scenario analyses conducted by the company where the estimated ICER was "*extremely unstable*".[18] The Committee

noted the points made by the ERG that these differences were driven by the short time horizon of the analysis and the data suggesting that vortioxetine was neither more or less effective than other antidepressants. The Committee stated that they factored this instability into their decision making. Ultimately, the Committee recommended the Vortioxetine.

3.4.6. Big Cactus

Palmer *et al* analysed the health economic effects of providing computerised speech and language therapy + usual care versus attention control + usual care versus usual care alone for people with post-stroke aphasia.[19] The design of the underlying clinical study was a randomised controlled trial (RCT), to assess the effects of the different study arms on co-primary outcomes of change in impairment (measured as the ability to use vocabulary of personal important) and activity (measured as assessing a 10 minute video of a conversation using the activity scale of the Therapy Outcome Measures).

The primary method for the economic evaluation was a state transition model that moved the patients between Aphasia, Good Response (6 months), Good Response (9 months), Good Response (12months and beyond) and Dead. Different transitions were modelled between the health states in each arm of the model, with transitions modelled from Big Cactus and ONS life table data. Data on utilities and costs came from Big Catus, supplemented by literature-based values where appropriate.

The authors presented the probabilistic results, with 95% credible intervals on the incremental costs and QALYs. The results are presented as a series of pairwise analyses, but for brevity this is presented here as a full incremental analysis.

Table 6: The base case results in Palmer et al (adapted from Palmer et al[19] Table 24)

Strategy	QALYs	Costs	Incremental QALYs	Incremental Costs (£)	ICER (£/QALY gained)
AC	4.1991	38.14	-	-	Dominated by UC
UC	4.1992	0	0.0001 (95% CrI -0.02 to 0.02)	-38.14 (95% CrI -34.94 to -41.50	UC dominates AC
CSLT	4.2164	732.73	0.0172 (95% CrI –0.05 to 0.10)	32.73 (95% CrI 674.23 to 798.05)	42,686

QALYs, quality adjusted life years; ICER, incremental cost-effectiveness ratio; AC, attention control; UC, usual care; CrI, credible interval; CSLT, computer speech and language therapy

All three pairwise comparisons were presented using cost-effectiveness planes and a CEAC was presented. In the main text, the scenario analysis results were described, but a summary of the results tables were provided in an appendix. The estimation of the utility scores was identified as a key area of uncertainty in this economic evaluation, as when the EQ-5D-5L responses from carers rather than patients were used, the ICER fell below decision making thresholds whereas in the base case this was above decision making thresholds.

The key methodological limitation for this study identified by the authors, was that it was unclear whether the EQ-5D-5L used in the study was appropriate, as it had only had the face validity of the instrument assessed. With aphasia, this led to uncertainty around whether responders to the questionnaire fully understood the questions asked in the EQ-5D-5L. Sensitivity analyses using proxy scores indicated that this may be an important consideration as the results were more favourable to the intervention.

Overall, no novel techniques have been used to analyse the small differences in QALYs (0.0001 difference between AC and UC). Standard cost-utility analyses have been applied. The small differences in QALYs have been treated as a feature of the analysis (albeit a highly uncertain feature due to methodological uncertainty in how to

measure HRQoL in this population), and something to factor into the results interpretation rather than requiring additional analyses.

3.4.7. HuBBLE

Alshreef *et al* calculated the cost-effectiveness of two types of surgery (rubber band ligation (RBL) versus Haemorrhoidal artery ligation (HAL)) in people with haemorrhoids.[20] The design of the underlying clinical data was a randomised controlled trial. The primary outcome of the trial was the proportion of patients with recurrent haemorrhoids at 12 months after procedure.

The primary method for economic evaluation was an analysis of trial collected EQ-5D-5L from which QALYs were calculated, trial collected resource use from which costs were calculated, and the time horizon was one year. The data were analysed by imputing missing data using multiple imputation with chained equations (predictive mean matching procedure) and between arm differences in costs and QALYs were estimated by fitting a seemingly unrelated regression to the trial data.

Table 7: The base case results in Alshreef *et al*[19] (reproduced from Table 3)

Analysis	Incremental cost [£]: HAL–RBL (95% CI); *p* value	Incremental QALYs: HAL–RBL (95% CI); *p* value	ICER £ per QALY gained	Probability that HAL is cost effective at the threshold 20,000/QALY (£30,000/QALY)
Base-case analysis	1027 (782–1272); <0.001	0.01 (−0.02 to 0.04); 0.49	104,427	0.00 (0.05)

HAL, Haemorrhoidal artery ligation; RBL, rubber band ligation; QALYs, quality adjusted life years

Several subgroup and scenario analyses were conducted. Furthermore, a long-term extrapolation analysis was conducted as a sensitivity analysis. This was the only sensitivity analysis with a QALY gain of more than 0.01. CEACs were presented for the base case and the long-term extrapolation scenario.

The small differences in QALYs were noted briefly in the conclusion. The authors caution against the long-term scenario analysis, noting that the evidence used to inform the long-term recurrence rates were of poor quality and that there may be other

interventions that should be included in the comparisons – which would need a network meta-analysis of RCTs to be conducted.

3.4.8. NICE TA878

NICE technology appraisal 878 assessed the effectiveness and cost-effectiveness of six different drugs for treating coronavirus disease 2019 (COVID-19).[21] They key clinical studies underlying the evidence submission were COVID-NMA initiative and metaEvidence initiative, both of which were at the time the appraisal was conducted living systematic reviews and meta-analyses of COVID-19 randomised controlled trials. This led to limitations in that the appraisal had to use the assumed effects from the meta-analysis without the usual assessment of how to appropriately model the treatment effects from the included studies.

The economic evaluation used an economic model. This model was a three-state partitioned survival model, that characterised COVID-19 by three mutually exclusive health states: a) discharged from hospital and alive; b) hospitalised with or without COVID-19; and c) dead. Health state occupancy was informed by Overall Survival and time to discharge, both from RECOVERY study and treatment effects. Health state utility values were from three studies Ara and Brazier for age-gender matched utilities, Rafia *et al* for-hospitalisation utilities and Evans *et al* for the effects of long covid.[27-29] Health care resource utilisation was assumed and standard UK sources for costs were attached.

The base case model did not run PSA, and instead several deterministic analyses were conducted. This was because this was a short notice appraisal and one particular calculation that would be required to conduct PSA led to the model having an infeasible computation time. Otherwise, a standard set of cost-effectiveness analyses when having multiple comparators was conducted, these are summarised in Table 8.

Table 8: The base case results for mean efficacy in the high risk population in NICE TA878[21] (replicated from the AG report, table 20)

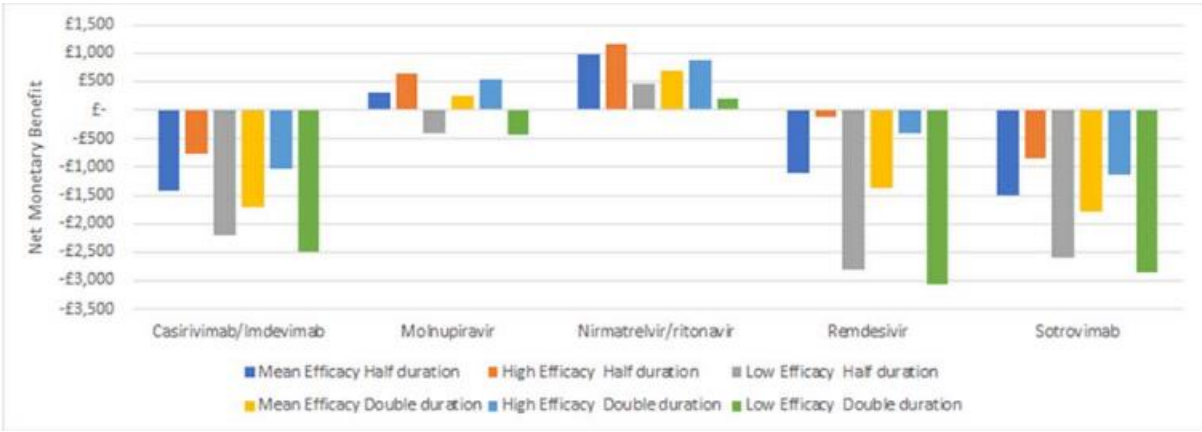
Intervention	Discounted Costs (£)	Discounted QALYs	Cost per QALY compared with SoC	NMB compared with SoC* (£)	Cost per QALY Incremental Analyses (£)
SoC	413	10.05	-		
Nirmatrelvir/ritonavir	670	10.11	£4,438	£904	£4,439
Molnupiravir	1027	10.10	£13,684	£283	Dominated
Remdesivir	1923	10.07	£88,320	-£1,169	Dominated
Casirivimab/Imdevimab	2450	10.08	£74,907	-£1,493	Dominated
Sotrovimab	2662	10.09	£65,922	-£1,567	Dominated

QALYs, quality adjusted life years; SoC, standard of care; NMB, net monetary benefit

*, assuming a threshold of £20,000 per QALY gained

Scenario analyses were explored, and there was a relatively novel way of presenting the results, as a series stacked bar charts were used to present the results of the scenario analysis. This is replicated in Figure 5. As a PSA was not conducted, more standard ways of presenting uncertainty in probabilistic results, such as a CEAC were not presented.

Figure 5: The NMB results for patients in the community with COVID-19 who are at high-risk of hospitalisation when the duration of long COVID is halved and doubled (replicated from AG report[21], Figure 22)



The small QALY gained in the analyses were not identified in the Committee discussion as a major issue.

3.4.9. Mandrik 2024

Mandrik *et al* is an academic paper assessing the cost-effectiveness and long term impact of screening for Bladder and Kidney cancer using a home urine dipstick test.[22] They used an individual level model to conduct their economic evaluation. The key clinical evidence was obtained from the Health Survey for England, the YORKSURe randomised controlled trial, and multiple clinical sources informing the sensitivity and specificity of the screening interventions tested. Utilities were sourced from literature values. Costs were calculated using a mixture of assumed diagnostic pathways, NHS reference costs and literature-based estimates of the lifetime cost of treating Kidney or Bladder cancer.

Mandrik present their results not as per person results but as per 100,000 people.[22] Furthermore, other clinical predictions are also presented in the results (e.g. number of bladder cancer cases). Otherwise, a set of deterministic results are presented in the main text as a table, a PSA is conducted, and a cost-effectiveness plane is presented in the main text. Incremental net monetary benefit compared to no screening was used to present the results of the scenario analyses.

4. Targetted literature review

4.1. Introduction

We conducted a series of targeted literature reviews to cover the current academic literature on four key topics that were identified as being potentially important to interpreting economic evaluations with small QALY differences. These topics were:

1) Novel literature on probabilistic analyses either on limitations or on how to implement probabilistic analyses when the expected outcomes were relatively small.
2) Proposed approaches to the presentation of uncertainty in economic evaluations that may improve on standard approaches (e.g. CEAC)
3) The relevance of the concept of MCID in relation to QALY or utility calculation or to the interpretation of the results of economic evaluations
4) Novel literature on the use of CMA, after a CEA result has produced a small and uncertain QALY difference.

These topics were identified as being important to review and summarise, as these topics were identified as key areas of uncertainty in the interpretation of small QALY differences by NICE.

4.2. Searches and inclusion criteria

The purpose of the literature review was to identify new literature that claims to further develop analyses and methods that are related to small QALY differences. The aim was to identify any promising techniques for analysing and reporting small QALY differences that are not being used in practice – based on the case studies reported in Section 3. We focus on papers related to probabilistic analysis and the reporting of uncertainty and also review recent papers exploring the concept of MCID in relation to QALYs and HRQoL.

4.2.1. Search details

Three separate sets of searches were conducted.

- Forward and backwards reference and citation searching of seven key seed papers was conducted, with the key seed papers being: Xie *et al*; Otten *et al* ; Geenen *et al*; Bates *et al* ; McClure *et al*.; Whitehurst *et al* and, Wang *et al*.[1-4, 10, 12, 30] These papers were selected because they were known to have been published papers or commentaries on one of the following topics: small QALY differences; interpretation of uncertainty in model results; or, MCID in utilities or QALYs. The citation searching was conducted using Scopus.
- Searching of Asta, an AI assisted search tool, to identify studies on probabilistic analysis methods.
- Traditional database searches

Full details of the search terms are provided in the Appendices.

The remaining papers were sifted by one reviewer (DP) by assessing whether the paper was on either methods of handling uncertainty in economic evaluations or directly on small benefits or an economic evaluation that had developed a method

within in it for presenting results of economic evaluations with small QALY differences. The full inclusion and exclusion criteria are given in Table 9.

Table 9: The inclusion and exclusion criteria

	Inclusion	**Exclusion**
Population	Any	None
Intervention	Any	None
Comparator	Any	None
Outcomes	Any	None
Study design	Methods paper Case study	Any other paper type
Other	Substantial text on small QALY differences Substantial text on minimum clinically important differences in QALYs Substantial text on new uncertainty analysis methods Substantial text on cost-minimisation analysis instead of cost-effectiveness analysis to address small QALY differences	Minimal text on small QALY differences, minimum clinically important differences in QALYs or uncertainty analysis

QALYs, quality adjusted life years

A narrative synthesis was conducted to synthesise the key points made in each paper. The narrative synthesis was done separately by topic area (Uncertainty communication techniques; Probabilistic analysis methods for small QALY differences; Use of MCIDs in interpreting QALY differences; Cost minimisation analysis), to make the synthesis clearer.

4.3. Search results

After the automatic removal of duplicates, 461 papers were considered for inclusion (72 astra searches, 282 from citation searching in Scopus and 107 from the database searches). In addition to the 7 seed papers of the citation searching, we found 6

additional papers that met our inclusion and exclusion criteria. The included studies are discussed by review topic area.

4.4. Overview of included studies and summary

No literature was found on the topics of probabilistic analysis limitations and methods in the context of small differences in simulated outputs or the role of CMA (if any) in interpreting cost-effectiveness analyses with small QALY differences. 6 studies were found that covered novel ways to present uncertainty in economic evaluations. None of these studies were specifically designed to present uncertainty in economic evaluations with small QALY differences. 7 studies were included on the topic of use of MCIDs in interpreting QALY differences. Two studies suggested a role for MCIDs within the calculation of QALYs or in the interpretation of economic evaluations. Two studies described the MCID's of utility scores as being used to interpret epidemiological study results or in epidemiological study power calculations. Two studies were opinion pieces, outlining arguments against using MCIDs in economic evaluations. The remaining study was a highly cited methods piece on how to calculate MCIDs, that was included because it was a seed study in our search strategy.

We did not find any studies on novel techniques for conducting and interpreting probabilistic analysis or in applying CMA instead of cost-effectiveness analysis when there are small QALY differences. This suggests that no changes in current practice here should be adopted.

Whilst there are many novel ways of presenting uncertainty in economic evaluations that we found in our literature review, none of these were specifically designed to address the presentation and interpretation of uncertainty in when an economic evaluation produces a small QALY difference. None of the novel techniques appear to be so useful in presenting and interpreting uncertainty when there are small QALY differences, that we believe they should be mandatory when there are small QALY differences. Analysts should consider if the novel techniques are useful for presenting uncertainty in their economic evaluation, and if they think they are useful, they should present uncertainty using these techniques.

Standard practice in economic evaluation is to not apply MCIDs when calculating QALYs or in interpreting QALY differences in an economic evaluation. Whilst there is some literature suggesting a role for MCID, there is also an equivalent amount of literature against applying MCIDs, with the key point in that literature being is that you can only assess QALY differences alongside cost differences in an economic evaluation. Overall, we do not believe that there is a clear case for the use of MCIDs when there are small QALY differences.

Full details of the studies that we used as pearls in our citation searches and studies that met our inclusion and exclusion criteria are provided in Section 4.5.

4.5. Included studies in the targetted literature review in detail

4.5.1. Uncertainty communication techniques and interpretation

4.5.1.1. Xie

Xie et al present a full methods guide on conducting probabilistic analysis in health economic assessment.[1] There is a significant section in this guide on the credibility and interpretation of small QALY differences. On this particular topic, Xie et al recommend using net benefit statistics when assessing interventions with small mean incremental QALYs. They further recommend presenting results on a cost-effectiveness plane, with a confidence ellipse. They also recommend exploration of whether small QALY differences are clinically meaningful, although this is not explained in detail.

4.5.1.2. Otten

Otten *et al* conducted a review of health economic methods to more generally assess uncertainty in health economic evaluations.[4] They conducted a snowball search of 8 articles, in the Google Scholar, Web of Science and Embase databases. They identified 28 uncertainty analysis methods, albeit 16 of these methods could be categorised into broader categories of deterministic sensitivity analysis, importance analysis, value of information analysis, opportunity loss analysis. The scope of review was broad and included any uncertainty communication method. The study makes no

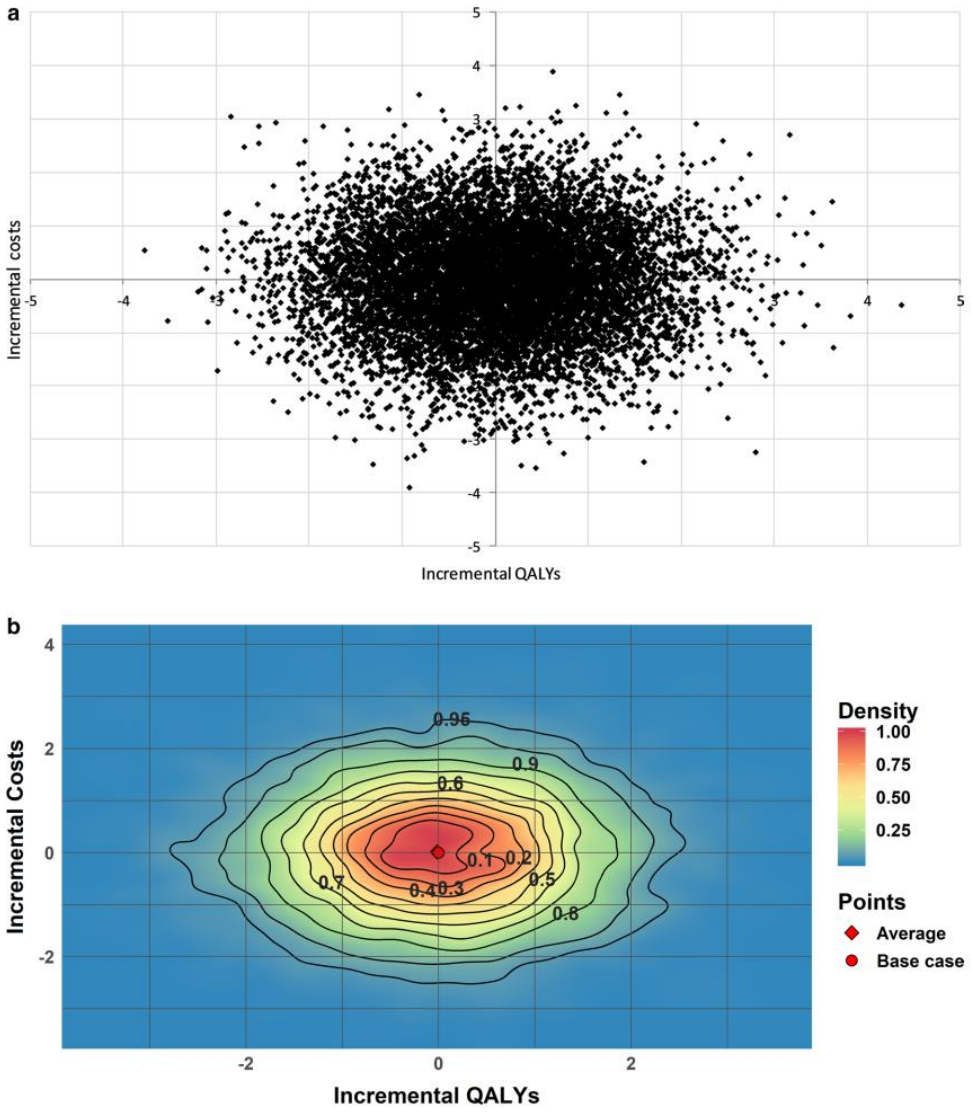
specific mention of whether these methods would be more appropriate to address small QALY differences was made.

4.5.1.3. Geenen

Geenen proposed a new method for presenting results of probabilistic analyses, the relative density plot.[30] This is designed to address two perceived limitations in the cost-effectiveness plane diagram to present PSA results. Firstly, in parts of the cost-effectiveness plane that have a high probability of occurring in the results, there is frequent overlaps in the scatter plot points which can make it difficult to assess how likely this area of the graph is. Secondly, rare scenarios are very prominent in a cost-effectiveness plane potentially can lead to those points being overinterpreted.

The case study comparing a cost-effectiveness plane to the relative density plot presented by Geenen is provided in Figure 6. Compared to the standard cost-effectiveness plane, it visually makes it clearer where the results are most likely to be.

Figure 6: A comparison of a cost-effectiveness plane and a relative density plot for an exemplar set of health economic evaluation results (replication of Greenen *et al*[30] Figure 1, used under CC-BY 4.0 licence)



4.5.1.4. Bates

Bates is a commentary piece on an applied economic analysis and primarily focuses on the volatility of economic evaluations with small QALY differences how to interpret small QALY differences in the academic literature.[2] Bates's key point is that economic outcomes are highly sensitive to there being small QALY differences. They conclude that "*There needs to be exceptional certainty in trial design and reporting of results before a low-QALY intervention can be said to be definitively cost-effective or not*"[2]

Whilst this may be a useful approach for writing a conclusion in the academic literature, it is not likely to be a valid option for NICE Committees who typically either must recommend a treatment or not. However, Committee may wish to comment that small QALY differences themselves are a key driver of uncertainty if they believe this to be the case.

4.5.1.5. Grimm

Grimm *et al* presented two new methods for analysing uncertainty in HTA, based on what was done in a scoping review of the current literature, semi structured interviews with HTA stakeholders and a stakeholder workshop going through a worked example.[31] They presented two new tools for presenting uncertainty, the assessment of risk table (ART) and the Appraisal of Risk CHart (ARCH). The ART for an exemplar analysis of pharmacological treatments, by Grimm *et al* is replicated in Table 10. The ARCH for the same analysis is presented in

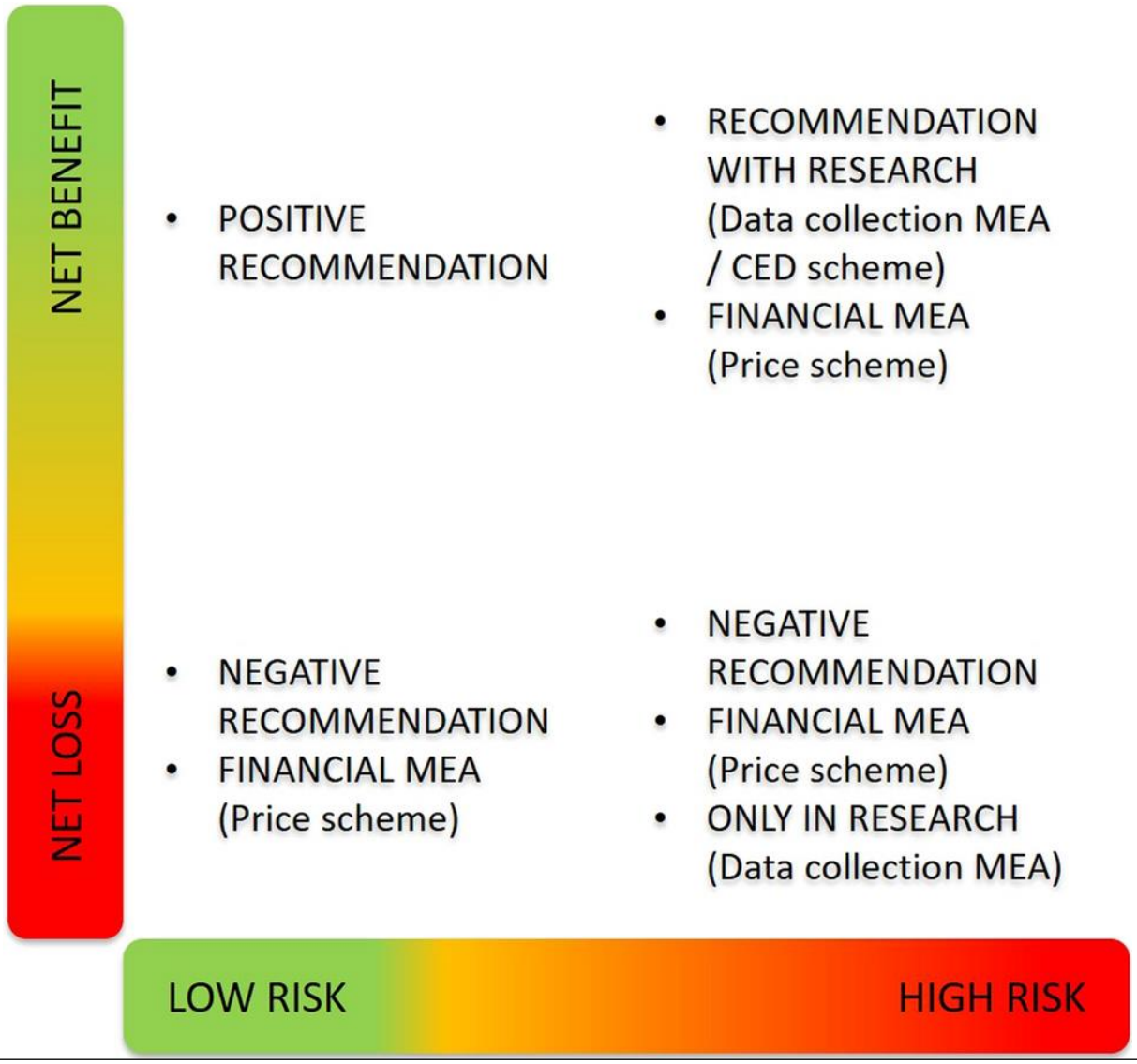
Figure 7.

Table 10: Replication of the ART for pharmacological treatments for peripheral artery disease in the Netherlands by Grimm *et al.*

1. Assessment general information					
Threshold (€/QALY)	50,000	Disease burden (€)	-	Population Size	227,499
2. Assessment outcomes					
ICER (DPI vs aspirin) (€/QALY)	26,221	Probability cost-effective	58%	Budget impact (€)	–
3. Risk and burden estimates					
EVPI (€) annual, over population	53,48,50,149	Incremental net benefit/loss DPI versus best comparator aspirin (€), over population		2,136,443,109	
4. Uncertainty information					
Uncertainties in PA	EVPPI (€)	Can be researched?	Uncertainties not in PA	Expected impact	Can be researched?
Utilities	908	Yes	Appropriate model structure unknown	Uncertain	Difficult
Costs	154	Possibly	Relative effectiveness of clopidogrel vs aspirin or DPI	Large	Possibly
			Generalisability of utility data questionable	Large	Yes

DPI dual platelet therapy with rivaroxaban+aspirin, EVPI expected value of perfect information, EVPPI expected value of perfect parameter information, PA probabilistic analysis, QALY quality-adjusted life year

Figure 7: The ARCH diagram



Neither tool was designed specifically to address small QALY differences. They were specifically designed to present uncertainty when risk management strategies are being considered.

4.5.1.6. Alarid-Escudero

Alarid-Escudero *et al*. present expected loss curves (ELC) to demonstrate uncertainty in health economic results.[32] ELCs require a probabilistic analysis to be conducted. The expected loss for each strategy is calculated by calculating the maximum net

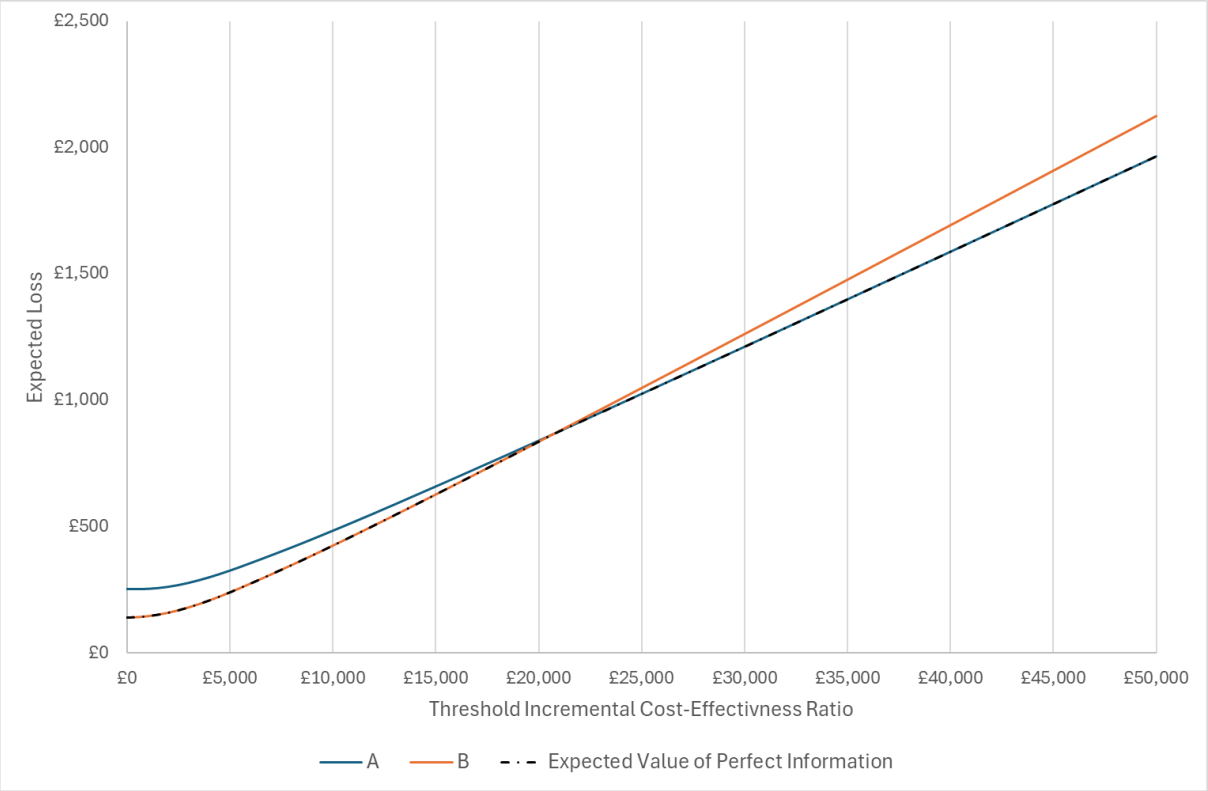
benefit in each PSA iteration. For each strategy, in each PSA iteration the expected loss is:

Max net befit for that PSA iteration – the strategies net benefit for that iteration

The expected loss is the mean loss for each strategy across all PSA runs. The expected loss of the economically optimal strategy at each threshold value is equivalent to the EVPI.

The ELC then graphically plots these expected losses on the Y axis against the maximum acceptable ICER to the decision maker on the X axis. The graphs in this paper cannot be replicated here, as the paper is not open source, but they can be seen in the original paper. An expected loss curve using the same data used to demonstrate the cost-effectiveness plane, the CEAC and the CEAF in Section 2 is presented in Figure 8.

Figure 8: An example expected loss curve



The authors state that the ELCs are address three limitations of CEACs and CEAFs, not small QALY differences in economic evaluations. Firstly, CEACs and CEAFs do not show the differences in net benefit, so you cannot judge how much better the optimal strategy is than the others or the net benefit of any non-optimal strategies. Secondly, two strategies could have similar net benefits, but there is large difference in the probabilities of being cost-effective. Thirdly, value of information must be shown on another Figure. No mention is made of small QALY differences, or if ELCs are better able to demonstrate uncertainty in these situations.

4.5.1.7. Eckerman

Eckerman *et al*.[33] advocate for presenting health economic results on the cost-disutility plane. This involves recalculating the x axis of the standard cost-effectiveness plane, so that it presents increasing disutility as the graph goes from left to right, instead of increasing utility. They also advocate for expected loss curves, as in Alarid-Escudero *et al*.[32] We do not demonstrate a cost-disutility plane here. The paper does not identify that either of these methods were designed to present uncertainty when there are small QALY differences.

4.5.2. Use of Minimal Clinically Important Differences (MCID) in interpreting QALY differences

4.5.2.1. Xie

Xie et al also discuss MCIDs in QALYs and how to interpret this in the context of small population level incremental QALY differences.[1] They recommend that if QALYs do not meet a MCIDs then other factors may become more important in decision making. This is because if the differences fall below certain minimum levels, then other factors may drive policy decisions. They suggest that the other factors should be considered are costs, equity, patient preferences and accessibility.

4.5.2.2. McClure

McClure *et al* present a method of applying MCIDs to HRQoL data collected within trials.[3] They estimate a change in utility that is meaningful at the individual level from an external data source, utility scores that are under the meaningful utility change from baseline are set to the baseline level (effectively ignoring the change for any individual whose difference in utility is below the MCIDs). QALYs are then calculated using normal methods.

Given our concerns about applying a MCIDs approach to interpreting or estimating utilities outlined in the introduction (see Section 2.6), we would urge caution in using this to help interpret small QALY differences, as it has no basis in economic theory and ignores the value that has been placed on these health conditions. Furthermore, it appears that this would be difficult to apply to model based analyses, unless health state utility values were calculated from individual participant level data.

4.5.2.3. Whitehurst

A full description of the points that Whitehurst *et al*. is made in Section 2.6.[12] In brief, this is a commentary paper on why MCIDs are not an important concept to consider when changes in utility or QALYs. The key reason for this is that every difference is important in a utility score, as the changes in level of the PROM have been valued using people's preferences. Any change in utilities or QALYs alone is potentially worthwhile, although they can cost too much to be economically viable.

4.5.2.4. Wang

Wang *et al* was only included as a seed for literature on MCIDs in PROMs.[10] As it is a step by step guide on how to establish MCID in PROMS generally, it does not address why MCIDs are an important for interpreting utility values or why this is an important concept for health technology assessments.

4.5.2.5. Cheng

Cheng *et al* present a systematic review of EQ-5D (either index scores or VAS) MCIDs studies, that used anchor based methods.[34] Most of their paper is not pertinent to our review as focuses on what the MCID is for the EQ-5D and how it varies across different studies. However, in their introduction, they note three key uses for MCIDs 1) interpreting change over time or between group differences in EQ-5D scores; 2) estimate sample sizes for new studies; and 3) preform responder analysis. They do not note a use within an economic evaluation.

They found that the MCIDs in EQ-5D varied by baseline EQ-5D score and treatment type. Within the discussion, they do not explicitly note any use the MCIDs directly in QALY calculations or in economic evaluations.

4.5.2.6. Hays

Hayes *et al* is a paper on how MCID should be interpreted.[35] They make two key points that could be relevant to interpreting small QALY differences. Firstly, MCID thresholds are not useful for decision making, as they ignore the costs required to achieve any changes in HRQoL. Secondly, MCIDs are usually derived from mean changes in HRQoL. However, estimating a MCID in this way ignores all of the individual response patterns to that HRQoL instrument.

4.5.2.7. Henson

Henson *et al* was a systematic review of MCID in the context of health utility estimates.[36] They found: many studies on estimating MCID for the EQ-5D-3L, EQ-5D-

5L and SF-6Dv1 instruments, but little data on other instruments; MCID varied greatly across the instruments; the method to estimate MCIDs greatly affected the estimated MCID values; many studies did not report on key items for an MCID study such concurrent validity between the instrument of interest and the anchor instrument or the precision in the estimate of the MCID.

They recommend that MCID make 11 recommendations for estimating MCIDs based on the literature. They don't state how MCIDs for utility instruments should be accounted for in economic evaluations and the paper instead states the having valid and reliable estimates of MCID "*may encourage the inclusion of MAUIs in a wider range of studies, including clinical trials*".

4.5.2.8. Summary

Only two papers advocate for the use of MCID in HTA, Xie *et al* and McClure, the rest are neutral or against the inclusions of MCID in HTA and only advocate for the use MCID interpreting clinical results or are against the use of MCID in HTA. The systematic reviews on MCID of utility instruments both found that the MCID varied greatly. We believe that MCID should not be directly applied in HTA.

5. Expert review comments

We presented our analysis and draft recommendations to a group of expert commentators who consisted of academics, NICE Committee chairs, ex and current appraisal Committee members and senior EAG members who produce independent reports on economic evaluations submitted to NICE. The purpose of these discussions was not to generate research evidence but instead to evaluate the issues and ideas raised from the case studies and review for theory, practice and policy.

The experts' discussions focused on six key themes.

First, there was some consensus that cost and QALY differences, and what drives them, should be described in more detail when QALY differences are small. Committee decision-making would benefit from considering what is driving

uncertainty, before appraising ICERs. Understanding whether there is particular uncertainty in estimates of costs or QALYs would help the Committee interpret ICERs. Similarly, understanding whether uncertain small QALY differences are driven by estimates of quality of life or length of life could be important. The experts stressed that a full appraisal of the drivers of small and/or uncertain QALY differences would allow better interpretation of subsequently discussed ICERs.

Second, the value of scenario analyses was discussed. There was a general view that presenting a large number of scenario analyses without rationale is not helpful. Committees find this particularly challenging when scenarios produce cost-effectiveness results that range from dominated to dominating, a frequent occurrence when there are small QALY differences. However, it was highlighted than when relevant scenarios do result in widely varying cost-effectiveness results, this is important for the Committee to know. Hence, scenario analyses can be important in the context of small QALY differences, but the rationale for each scenario should be carefully thought through and presented.

Third, issues around presenting results using net benefits rather than ICERs was discussed. The experts generally agreed that presenting net benefit is useful and avoids some of the volatility seen in ICERs. However, there were also views that Committees did not always fully understand results reported as net benefits, which may lead to incorrect interpretation. This may indicate a training need for NICE Committee members. This discussion also touched upon the use of cost minimisation analysis (CMA). It was highlighted that when QALY differences were small, Committees may instead focus on costs, to determine how confident they could be about cost increases or decreases. The experts were concerned about this, as the fact that QALY differences are estimated to be small and/or uncertain does not mean that they should be ignored. CMA approaches were not supported.

Fourth, the experts discussed whether there was any need for different analytical approaches when estimated QALY differences were small. There was a general view that there is not a need to address uncertainty in a different way when QALY differences are small, standard methods for presenting uncertainty in economic

evaluations remained appropriate. As discussed in the first and final points, this could be supplemented with additional analyses that provided context on the drivers of the small QALY differences. The experts did not believe that small QALY differences should lead a Committee to consider CMA, or MCIDs. With respect to MCIDs, the experts agreed that differences in utilities or QALYs at either an individual or a cohort level were important, whatever their size.

Fifth, there was some support for exploring domain-level data from HRQoL instruments to provide context for understanding small QALY differences. For instance, this could demonstrate whether differences are driven by consistent improvements in specific domains, or whether they appear to be driven by more random changes across domains. This could help a Committee understand whether a small QALY difference was 'credible or not, which could in turn help the interpretation of cost-effectiveness results. Associated with this, there was some support for conducting 'responder analysis', where data are analysed to determine whether small QALY differences were driven by large differences in a small proportion of patients, or by small differences across a large number of patients. There was some concern as to how a Committee should interpret this information, but some experts did feel that this would help decision-making.

Finally, the experts were hesitant to recommend that small QALY differences should trigger a closer inspection of the validity of the HRQoL instrument used in the analysis. It was suggested that the generic instruments used for economic evaluation will never be perfect for any specific condition and could be under-estimating benefits even when larger QALY differences are estimated. In general, the validity of the instrument should be considered for any appraisal, irrespective of the size of the QALY difference.

6. Recommendations

6.1. Evidence and Analytical considerations

1) Explanation should be provided on the source and robustness of QALY differences when those differences are relatively small.

2) Small QALY differences in and of themselves do not indicate the need to conduct additional scenario analyses.

3) Explanation of the cost and QALY differences and what drives them should be provided, and Committees should consider these before discussing ICERs or net benefits. This will help Committees interpret ICERs and scenario analyses. Part of this explanation may include analysis of domain-level data from the preference-based instrument used, and comparison with expectations and data yielded from other outcome measures.

5) Committees should always consider the source of the evidence when appraising QALY differences, and this may provide particularly important context when QALY differences are small. For instance, small QALY differences may be interpreted differently, or with a different level of confidence, if they are derived from head-to-head RCT evidence, compared to indirect comparisons or observational data.

6) Decision uncertainty stems from the joint uncertainty in health and cost differences. It should be expressed in the same ways irrespective of whether QALY differences are small or not. Analysts should typically use the following methods for exploring uncertainty:

a. Probabilistic sensitivity analysis should be conducted, with all uncertain parameters appropriately determined and included in the analysis

b. To explore structural uncertainty, relevant scenario analyses should be explained and reported.

6.2. Presentation of results

7) All standard methods of presenting the results of an economic evaluation (results tables, cost-effectiveness planes, CEACs and CEAFs) should be presented and explained.

8) Results should be disaggregated to demonstrate whether volatile ICERs are associated with small differences in QALYs or costs. Total and incremental costs and QALYs should be presented. Where relevant, the QALYs and costs accrued in each modelled health state should be reported. The contribution to the total QALY difference associated with differences in length of life and quality of life could also be reported.

9) Consider using net benefits to present the results, as they may be easier to interpret because they are more stable than ICERs. However, net benefits should be explained clearly so that committee members who are used to interpreting ICERs as

results of economic evaluations can clearly understand what is being shown and what this means.

6.3. Interpretation of results

10) The concept of MCID is not relevant to differences in preference-based instruments.

11) If the source of the differences and robustness of the QALY estimation are not apparent from the economic evaluation methods, analyses should be undertaken to identify whether small QALY differences are due to small differences in many patients, or large differences in few patients. Presenting the HRQoL or QALY data in a way that shows the distribution of relevant outcomes and highlights the impact of outliers/rare events could be useful.

12) Consider presenting HRQoL changes at the domain level for preference-based instruments, if the individual level data used to directly generate QALYs or health state utility values for a model is available. This will indicate whether changes in QALYs or health state utility values are driven by changes in specific domains of the HRQoL instrument and whether changes appear to be explainable.

6.4. Further Research

13) Research should be conducted on how to best report HRQoL changes at the domain level.

REFERENCES

1. Xie X, Schaink AK, Gajic-Veljanoski O, Yeung MW, Wang M, Li C*, et al.* A methodological guide for implementing and interpreting results of probabilistic analysis. *Expert Rev Pharmacoecon Outcomes Res* 2025;25:123–35.
2. Bates A. Trials with small QALY benefits can be badly misinterpreted. *BJOG* 2019;126:1051.
3. McClure NS, Paulden M, Ohinmaa A, Johnson JA. Modifying the quality-adjusted life year calculation to account for meaningful change in health-related quality of life: insights from a pragmatic clinical trial. *Eur J Health Econ* 2021;22:1441–51.
4. Otten TM, Grimm SE, Ramaekers B, Joore MA. Comprehensive Review of Methods to Assess Uncertainty in Health Economic Evaluations. *Pharmacoeconomics* 2023;41:619–32.
5. National Institute for Health and Care Excellence. NICE technology appraisal and highly specialised technologies guidance: the manual 2026.
6. van Hout BA, Al MJ, Gordon GS, Rutten FF. Costs, effects and C/E-ratios alongside a clinical trial. *Health Econ* 1994;3:309–19.
7. Fenwick E, Claxton K, Sculpher M. Representing uncertainty: the role of cost-effectiveness acceptability curves. *Health Econ* 2001;10:779–87.
8. Papaioannou D, Brazier J, Parry G. How valid and responsive are generic health status measures, such as EQ-5D and SF-36, in schizophrenia? A systematic review. *Value Health* 2011;14:907–20.
9. National Institute for Health and Care Excellence. Digital therapy for chronic tic disorders and Tourette syndrome: early value assessment. HealthTech guidance HTG748; 2025.
10. Wang Y, Devji T, Carrasco-Labra A, King MT, Terluin B, Terwee CB*, et al.* A step-by-step approach for selecting an optimal minimal important difference. *BMJ* 2023;381:e073822.
11. Walters SJ, Brazier JE. What is the relationship between the minimally important difference and health state utility values? The case of the SF-6D. *Health Qual Life Outcomes* 2003;1:4.
12. Whitehurst DGT, Briggs A, Lloyd AJ, Abangma G, Parkin D. An Overdue Denunciation of the Minimal Important Difference When Applied to Health State Values. *Value Health* 2026; 10.1016/j.jval.2026.01.003.
13. Wailoo A, Dixon S. The use of cost minimisation analysis for the appraisal of health technologies: report by the NICE Decision Support Unit; 2019.
14. National Institute for Health and Care Excellence. Digitally enabled therapies for adults with anxiety disorders: early value assessment. HealthTech guidance HTG676; 2023.
15. National Institute for Health and Care Excellence. Digital technologies for assessing attention deficit hyperactivity disorder (ADHD) HealthTech guidance HTG729; 2024.
16. Richards D, Enrique A, Eilert N, Franklin M, Palacios J, Duffy D*, et al.* A pragmatic randomized waitlist-controlled effectiveness and cost-effectiveness trial of digital interventions for depression and anxiety. *NPJ Digit Med* 2020;3:85.
17. Strauss C, Bibby-Jones AM, Jones F, Byford S, Heslin M, Parry G*, et al.* Clinical Effectiveness and Cost-Effectiveness of Supported Mindfulness-Based

Cognitive Therapy Self-help Compared With Supported Cognitive Behavioral Therapy Self-help for Adults Experiencing Depression: The Low-Intensity Guided Help Through Mindfulness (LIGHTMind) Randomized Clinical Trial. *JAMA Psychiatry* 2023;80:415–24.
18. National Institute for Health and Care Excellence. Vortioxetine for treating major depressive episodes. Technology appraisal guidance TA367; 2024.
19. Palmer R, Dimairo M, Cooper C, Enderby P, Brady M, Bowen A*, et al.* Self-managed, computerised speech and language therapy for patients with chronic aphasia post-stroke compared with usual care or attention control (Big CACTUS): a multicentre, single-blinded, randomised controlled trial. *Lancet Neurol* 2019;18:821–33.
20. Alshreef A, Wailoo AJ, Brown SR, Tiernan JP, Watson AJM, Biggs K*, et al.* Cost-Effectiveness of Haemorrhoidal Artery Ligation versus Rubber Band Ligation for the Treatment of Grade II-III Haemorrhoids: Analysis Using Evidence from the HubBLe Trial. *Pharmacoecon Open* 2017;1:175–84.
21. National Institute for Health and Care Excellence. Nirmatrelvir plus ritonavir, sotrovimab and tocilizumab for treating COVID-19. echnology appraisal guidance. TA878 2023.
22. Mandrik O, Thomas C, Akpan E, Catto JWF, Chilcott J. Home Urine Dipstick Screening for Bladder and Kidney Cancer in High-Risk Populations in England: A Microsimulation Study of Long-Term Impact and Cost-Effectiveness. *Pharmacoeconomics* 2025;43:441–52.
23. Gega L, Jankovic D, Saramago P, Marshall D, Dawson S, Brabyn S*, et al.* Digital interventions in mental health: evidence syntheses and economic modelling. *Health Technol Assess* 2022;26:1–182.
24. van Hout B, Janssen MF, Feng YS, Kohlmann T, Busschbach J, Golicki D*, et al.* Interim scoring for the EQ-5D-5L: mapping the EQ-5D-5L to EQ-5D-3L value sets. *Value Health* 2012;15:708–15.
25. Dolan P. Modeling valuations for EuroQol health states. *Med Care* 1997;35:1095–108.
26. Herdman M, Gudex C, Lloyd A, Janssen M, Kind P, Parkin D*, et al.* Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Qual Life Res* 2011;20:1727–36.
27. Ara R, Brazier JE. Populating an economic model with health state utility values: moving toward better practice. *Value Health* 2010;13:509–18.
28. Rafia R, Martyn-St James M, Harnan S, Metry A, Hamilton J, Wailoo A. A Cost-Effectiveness Analysis of Remdesivir for the Treatment of Hospitalized Patients With COVID-19 in England and Wales. *Value Health* 2022;25:761–9.
29. Evans RA, McAuley H, Harrison EM, Shikotra A, Singapuri A, Sereno M*, et al.* Physical, cognitive, and mental health impacts of COVID-19 after hospitalisation (PHOSP-COVID): a UK multicentre, prospective cohort study. *Lancet Respir Med* 2021;9:1275–87.
30. Geenen JW, Vreman RA, Boersma C, Klungel OH, Hovels AM, Ham R. Increasing the information provided by probabilistic sensitivity analysis: The relative density plot. *Cost Eff Resour Alloc* 2020;18:54.
31. Grimm S, Pouwels X, Ramaekers B, Wijnen BFM, Otten T, Grutters JPC*, et al.* State of the ART? Two New Tools for Risk Communication in Health Technology Assessments. *PharmacoEconomics* 2021;39:1185–96.
32. Alarid-Escudero F, Enns EA, Kuntz KM, Michaud TL, Jalal H. "Time Traveling Is Just Too Dangerous" but Some Methods Are Worth Revisiting: The

Advantages of Expected Loss Curves Over Cost-Effectiveness Acceptability Curves and Frontier. *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research* 2019;22:611–8.
33. Eckermann S, Briggs A, Willan AR. Health technology assessment in the cost-disutility plane. *Medical decision making : an international journal of the Society for Medical Decision Making* 2008;28:172–81.
34. Cheng LJ, Chen LA, Cheng JY, Herdman M, Luo N. Systematic review reveals that EQ-5D minimally important differences vary with treatment type and may decrease with increasing baseline score. *Journal of Clinical Epidemiology* 2024;174.
35. Hays RD, Woolley JM. The concept of clinically meaningful difference in health-related quality-of-life research. How meaningful is it? *Pharmacoeconomics* 2000;18:419–23.
36. Henson GJ, van der Mei I, Taylor BV, Scuffham P, Chen G, Campbell JA. A systematic review of minimum important changes for generic multi-attribute utility instruments and recommendations for their estimation. *Eur J Health Econ* 2025;26:1383–99.

Appendices

Search methods

Initial scoping of the literature before and after finalising the project brief suggested that the existing methodological literature on interpreting small differences in health utilities and incremental QALYs is limited. A mixed approach to search methods for the literature review was therefore adopted to maximise retrieval of relevant evidence, using an iterative approach that included focused systematic database searching, supplemented by citation-based and targeted methods based on known relevant academic papers.

No date or language limits were applied to the searches.

Phase 1 - Citation searches

Scoping generated an initial set of five relevant academic papers, which were used as seed papers for backwards citation searching using citationchaser (https://estech.shinyapps.io/citationchaser/) and forwards citation searching using the Scopus database. The table below summarises the seed articles and the number of references retrieved through these methods.

Seed paper	References retrieved from bibliography via citationchaser (backwards citation search)	References citing this paper retrieved via Scopus (forwards citation search)
Xie et al., 2025	41	1
Otten et al., 2023	74	13
Geenen et al., 2020	23	1
Bates et al., 2019 (blog/commentary)	0	0

McClure et al., 2021	36	4
Total	**174**	**19**

The searches were conducted in February 2026. After deduplication in EndNote, 166 references from the backwards citation search, and 18 references from the forwards citation search were initially screened at title/abstract level for inclusion in the review.

These searches were further supplemented by Scopus database citation searches of Whitehurst et al 2025 and Wang et al 2023, to further supplement coverage of the application of minimum clinically important differences to quality adjust life years and/or utilities. This returned a further 100 papers to sift.

Phase 2 - Database searches

A systematic search of relevant bibliographic databases was conducted in February 2026. The results are summarised in the table below and the full search strategies follow.

Strategy overview

Database	**Platform**	**Search date**	**No. of results**
Medline	Ovid	25.02.2026	28
Embase	Ovid	25.02.2026	49
EconLit	Ovid	25.02.2026	44
Results before deduplication			**121**
Results after deduplication			**109**

Results were deduplicated in EndNote, and the remaining references were screened for inclusion in the review.

Database search strategies

Ovid MEDLINE(R) ALL <1946 to February 24, 2026>

1	(probabilistic sensitivity analys* or PSA* or probabilistic analys*).tw.	65983
2	(methodolog* or approach* or technique* or instrument*).ti,ab.	5320208
3	Quality-Adjusted Life Years/	18982
4	("small QALY difference*" or "small QALY gain*" or "small QALY benefit*").tw.	15
5	((present* or interpret* or communicat* or representat*) adj2 uncertaint*).ti,ab.	2045
6	3 or 4 or 5	21020
7	Technology Assessment, Biomedical/	11891
8	("health technology assessment" or hta).tw.	9384
9	"health economics".tw.	4456
10	7 or 8 or 9	22027
11	1 and 2 and 6 and 10	28

Embase <1974 to 2026 Week 08>

1	(probabilistic sensitivity analys* or PSA* or probabilistic analys*).tw.	125569
2	(methodolog* or approach* or technique* or instrument*).ti,ab.	6706024
3	quality adjusted life year/	43654
4	("small QALY difference*" or "small QALY gain*" or "small QALY benefit*").tw.	31
5	((present* or interpret* or communicat* or representat*) adj2 uncertaint*).ti,ab.	2567
6	3 or 4 or 5	46188

7	biomedical technology assessment/	20590
8	("health technology assessment" or hta).tw.	17826
9	"health economics".tw.	7399
10	7 or 8 or 9	37628
11	1 and 2 and 6 and 10	49

Econlit <1886 to February 19, 2026>

1	(probabilistic sensitivity analys* or PSA* or probabilistic analys*).tw.	419
2	(methodolog* or approach* or technique* or instrument*).ti,ab.	320678
3	1 or 2	320996
4	("small QALY difference*" or "small QALY gain*" or "small QALY benefit*").tw.	1
5	((present* or interpret* or communicat* or representat*) adj2 uncertaint*).ti,ab.	367
6	(quality adjusted life year* or QALY*).ti,ab.	822
7	4 or 5 or 6	1188
8	("health technology assessment" or hta).tw.	239
9	"health economics".tw.	1151
10	8 or 9	1377
11	3 and 7 and 10	44

Supplementary database search

These database searches were supplemented by a search of Asta (asta.allen.ai), an AI-assisted search platform. This was searched on 24th February 2026 using the following query:

‘What probabilistic sensitivity analysis methods are used to communicate uncertainty or interpret small differences in health utilities and incremental QALYs in cost effectiveness analysis for health technology assessment?’

This yielded 73 references, none of which the Asta platform deemed a ‘perfect match’, and so all 73 results were screened at title/abstract level initially for inclusion in the review.

Phase 3 - Targeted searches using Lens.org

Lens.org was chosen for targeted searching because of its capabilities to search the full text of articles and its broad range of sources. Some of the relevant keywords for the topic (such as ‘small QALY gain’) may be present in the full text of a relevant article, but not in the title, abstract and indexing, making the articles harder to retrieve via conventional keyword searching on bibliographic databases.

A list of academic journals likely to contain relevant methodological literature was generated by topic experts. Each of these was searched in combination with keywords for selected relevant phases in the full text of articles, via Lens.org on 25th February 2026. The results from these searches were deduplicated and the remaining results screened at title and abstract level for inclusion in the review.

The individual searches performed during this phase are summarised below, with individual queries listed after.

keyword(s)/ journal	**Value in Health**	**PharmacoEconomics**	**Medical Decision Making**	**Health Economics**
probabilistic sensitivity analysis AND methods	17 results	29 results	3 results	11 results

small QALY gain	22 results	18 results	1 results	5 results
small QALY benefit	0 results	2 results	0 results	1 results

Total before deduplication = 109

Total after deduplication = 90

Scholarly Works (17) = Source Title: ("Value in Health") AND ((Title: (probabilistic AND (sensitivity AND analysis)) OR (Abstract: (probabilistic AND (sensitivity AND analysis)) OR (Keyword: (probabilistic AND (sensitivity AND analysis)) OR Field of Study: (probabilistic AND (sensitivity AND analysis))))) AND (Title: (methodology) OR (Abstract: (methodology) OR (Keyword: (methodology) OR Field of Study: (methodology))))))

Filters: No filters applied

Scholarly Works (29) = Source Title: PharmacoEconomics AND ((Title: (probabilistic AND (sensitivity AND analysis)) OR (Abstract: (probabilistic AND (sensitivity AND analysis)) OR (Keyword: (probabilistic AND (sensitivity AND analysis)) OR Field of Study: (probabilistic AND (sensitivity AND analysis))))) AND (Title: (methodology) OR (Abstract: (methodology) OR (Keyword: (methodology) OR Field of Study: (methodology))))))

Filters: No filters applied

Scholarly Works (3) = Source Title: ("Medical Decision Making") AND ((Title: (probabilistic AND (sensitivity AND analysis)) OR (Abstract: (probabilistic AND (sensitivity AND analysis)) OR (Keyword: (probabilistic AND (sensitivity AND analysis)) OR Field of Study: (probabilistic AND (sensitivity AND analysis))))) AND (Title: (methodology) OR (Abstract: (methodology) OR (Keyword: (methodology) OR Field of Study: (methodology))))))

Filters: No filters applied

Scholarly Works (11) = Source Title: ("Health Economics") AND ((Title: (probabilistic AND (sensitivity AND analysis)) OR (Abstract: (probabilistic AND (sensitivity AND analysis)) OR (Keyword: (probabilistic AND (sensitivity AND analysis)) OR Field of Study: (probabilistic AND (sensitivity AND analysis))))) AND (Title: (methodology) OR (Abstract: (methodology) OR (Keyword: (methodology) OR Field of Study: (methodology)))))

Filters: No filters applied

Scholarly Works (22) = Source Title: ("Value in Health") AND Full Text: ("small QALY gain")

Scholarly Works (18) = Source Title: PharmacoEconomics AND Full Text: ("small QALY gain")

Filters: No filters applied

Scholarly Works (1) = Source Title: ("Medical Decision Making") AND Full Text: ("small QALY gain")

Filters: No filters applied

Scholarly Works (5) = Source Title: ("Health Economics") AND Full Text: ("small QALY gain")

Filters: No filters applied

Scholarly Works (2) = Source Title: PharmacoEconomics AND Full Text: ("small QALY benefit")

Filters: No filters applied

Scholarly Works (1) = Source Title: ("Health Economics") AND Full Text: ("small QALY benefit")

Filters: No filters applied