NATIONAL INSTITUTE FOR HEALTH AND CLINICAL EXCELLENCE

Briefing paper for methods review workshop on structured decision making

The briefing paper is written by members of the Institute's Decision Support Unit in collaboration with Professor Nancy Devlin from the Office of Health Economics. It is intended to provide a brief summary of the issues that are proposed for discussion at a workshop to inform an update to the Institute's Guide to Methods of Technology Appraisal. It is not intended to reflect a comprehensive or systematic review of the literature. The views presented in this paper are those of the authors and do not reflect the views of the Institute.

The briefing paper is circulated to people attending that workshop. It will also be circulated to the members of the Method's Review Working Party, the group responsible for updating the guide.

For further details regarding the update of the Guide to the Methods of Technology Appraisal please visit the NICE website at http://www.nice.org.uk/aboutnice/howwework/devnicetech/technologyappraisalprocessguides/GuideToMethodsTA201112.jsp

1 Review of the 'Guide to Methods of Technology Appraisal'

The Institute is reviewing the 'Guide to the methods of technology appraisal', which underpins the technology appraisal programme.

The original Methods Guide was published in February 2001, and a revised version was published in 2007. The Methods Guide provides an overview of the principles and methods used by the Institute in assessing health technologies. It is a guide for all organisations considering submitting evidence to the technology appraisal programme and describes appraisal methodology.

The current 'Guide to methods of technology appraisal' is available from the NICE website at

http://www.nice.org.uk/aboutnice/howwework/devnicetech/technologyappraisalprocessguides/guidetothemethodsoftechnologyappraisal.jsp

The review of the Methods Guide will take place between October 2011 and April 2012. As part of the process, a number of workshops will be held to help identify those parts of the Guide that require updating. These workshops will involve a range of stakeholders, including methods experts, patient representatives, industry representatives, NHS staff and NICE technology Appraisal Committee members.

A summary of the discussion at the workshop will be provided to the Methods Review Working Party, the group responsible for preparing the draft update of the Methods Guide. Further details of the process and timelines of the review process are available from the NICE website.

The revised draft of the Methods Guide will be available for a 3-month public consultation, expected to begin in May 2012. We encourage all interested parties to take part in this consultation.

2 Background

The appraisal of health technologies by NICE can be viewed as being founded on the principle that the primary (but not only) purpose of the NHS is to improve health. Considering whether a new technology helps to achieve this objective, some measure of health improvement is required, which ought to reflect key criteria or attributes of health (e.g., length of life and various dimensions of its quality) with weights that reflect the preferences of the community served by the NHS. Since NHS resources are limited it is also important to know what additional NHS costs are required to improve health measured in this way. For this reason much of NICE methods of appraisal focus on how evidence can be used to estimate the likely improvement in health (measured by QALYs) offered by the technology and the additional NHS costs required. The combination of health benefits offered with associated NHS cost are commonly summarised as an incremental cost-effectiveness ratio (ICER). A key question is whether the health expected to

be gained from the use of the technology exceeds the health likely to be forgone elsewhere as a consequences of additional costs displacing other NHS activities. The cost-effectiveness threshold is intended to represent this aspect of opportunity cost (the additional NHS cost likely to displace one QALY elsewhere). The determination of NICE's threshold range (£20,000 to £30,000 per QALY) currently has a limited empirical basis (House of Commons Select Committee 2008; NICE 2008a). However, recent work suggests it is likely to be an appropriate order of magnitude (Martin, Rice and Smith 2008), and further research promises to strengthen the evidence base to inform the choice, albeit in the context of considerable uncertainty. What is important to recognise, however, is that the key underlying consideration in appraisal is not cost-effectiveness per se but the likely *net* health effects of a technology. A comparison of an ICER with the threshold helps inform this assessment of whether or not these *net* health effects are likely to be positive or negative.

If the objective of the NHS was *only* to improve health, and the measure of health available (QALYs) captured *all* socially valuable aspects of health, then the task of the Appraisal Committee would be restricted to exercising judgements about the scientific evidence, i.e., considering whether the evidence and analysis on which estimates of health gained and additional costs are based are judged to be reliable and reasonable. If they are, then decisions could simply be based on a comparison of ICER to the threshold, which is equivalent to asking whether the estimate of health gained exceeds the health expected to be forgone.

However, the value judgements which must be made by the Appraisal Committee must extend beyond considerations regarding the ICER for two reasons:

i. Even if the objective of the NHS was restricted to health improvement, no metric of health, no matter how sophisticated, can hope to capture all socially valuable aspects of health. For example, some types of health gain might be deemed more important and more socially valuable than others due to the characteristics of the disease (e.g., severity and

burden) or the characteristics of the recipients (e.g., children or disadvantaged populations).

ii. Although improving health might be the primary purpose of the NHS, other objectives, not directly related to health gain, might also be important (e.g., improving equity and wider social benefits).

Therefore, while cost-effectiveness (the net health effects of a technology measured by QALYs) might be a key consideration, other factors are also considered relevant and are taken into account by NICE. Indeed NICE is increasingly clear about what these factors are (NICE 2008b), and the way that it has reflected these 'social value judgements' in its decisions (Rawlins et al. 2009). NICE says that it recognises a number of criteria as relevant to its technology appraisals, and that it does so by applying 'special weightings' to these criteria when making judgements about cost effectiveness – for an overview, see Appendix 1. The way in which these factors are taken into account is set out in NICE's social value judgement document (NICE 2008b).

> *"Decisions about whether to recommend interventions should not be based on evidence of their relative costs and benefits alone. NICE must consider other factors when developing its guidance, including the need to distribute health resources in the fairest way within society as a whole." (Principle 3 – NICE 2008b p.18)*

Currently these other factors are taken into account by NICE as mitigating factors relative to the cost effectiveness threshold range of £20,000 to 30,000 per QALY gained. Specifically, the decision-making process by which the ICER and other factors are combined is described as follows:

> *"…interventions with an ICER of less than £20,000 per QALY gained are considered to be cost effective. Where advisory bodies consider that particular interventions with an ICER of less than £20,000 per QALY gained should not be provided by the NHS they should provide explicit reasons (for example that there are significant limitations to the generalisability of the evidence for effectiveness). Above a most plausible ICER of £20,000 per QALY gained, judgements about the acceptability of*

the intervention as an effective use of NHS resources will specifically take account of the following factors.

- *The degree of certainty around the ICER. In particular, advisory bodies will be more cautious about recommending a technology when they are less certain about the ICERs presented in the cost-effectiveness analysis.*

- *The presence of strong reasons indicating that the assessment of the change in the quality of life inadequately captured, and may therefore misrepresent, the health gain.*

- *When the intervention is an innovation that adds demonstrable and distinct substantial benefits that may not have been adequately captured in the measurement of health gain.*

As the ICER of an intervention increases in the £20,000 to £30,000 range, an advisory body's judgement about its acceptability as an effective use of NHS resources should make explicit reference to the relevant factors considered above. Above a most plausible ICER of £30,000 per QALY gained, advisory bodies will need to make an increasingly stronger case for supporting the intervention as an effective use of NHS resources with respect to the factors considered above." (NICE 2008b p.18-19).

Potential benefits of a more structured approach

It seems beyond dispute that factors other than net health gain measured by QALYs (i.e., cost-effectiveness) matter (Shah, Praet, Devlin et al 2011). However, it remains unclear to many outside NICE exactly how important these other considerations are, and how they are incorporated into the current deliberative approach to decision-making. The identification of these factors by NICE indicates that they must count for something, but not how much. That is, it is not clear what weight is attached to each in the decision-making process, or the trade-offs that NICE is prepared to make between QALYs gained and these other factors. Furthermore, the information provided in published NICE guidance "may not fully reflect all of the individual factors

considered by the Appraisal Committee at the time of the appraisal" (Tappenden, Brazier, Ratcliffe, et al. 2007).

Arguably, being more explicit about the factors that influence decisions, and the way these are taken into account, could serve to:

- Improve the transparency of the decision-making process and the accountability of NICE to taxpayers
- Improve the consistency of decision-making – for example, by ensuring that each of NICE's four Appraisal Committees treat these considerations in a similar manner
- Facilitate greater consistency between the way NICE decides on new technologies and the way the NHS decides how to allocate its budgets
- Provide an opportunity for NICE to engage the public in decisions about what criteria to use, and their relative importance – leading to more 'buy-in' to the difficult decisions NICE is required to make
- Sharpen the signals to industry about what aspects of innovation NICE (acting as an agent for the NHS) values and where research and development (R&D) efforts should be directed

NICE needs to consider to what extent the multiple criteria its committees need to take into account should be combined quantitatively as part of the technology appraisal process. There is a spectrum of possibilities regarding how much quantification is undertaken and it is not obvious that the optimal approach to decision making involves a highly technical solution (Devlin and Sussex 2011). Arguably, given the nature of the decisions being made by NICE, there will inevitably be a role for exercising judgement via a deliberative process (Culyer 2009). In advising NICE on the criteria which might be employed in guiding its decisions, NICE's Citizens' Council has adopted a deliberative framework to establish the strengths and weaknesses of competing criteria that might be considered (NICE 2011). The pertinent

question is therefore whether that deliberative process could be improved by the use of decision aids to structure and facilitate the consideration of multiple criteria; and to make more explicit and consistent the trade offs between criteria that are currently implicit in the deliberative process.

Recently, there have been a number of calls for decisions about resource allocation generally, and those made by NICE's Appraisal Committees in particular, to be moved along that spectrum by incorporating more quantification of other relevant criteria (Dowie 2008; NICE 2009a; Devlin and Sussex 2011). These calls have often referred to the use of multi-criteria decision analysis (MCDA) which is a set of methods of varying types which typically seek to score, weight and ultimately aggregate the various criteria relevant to a decision into an overall composite measure of benefit (Peacock, Richardson, Carter et al. 2007; Thokala 2011). MCDA approaches have been used by local NHS organisations to aid resource allocation decisions, and elsewhere in the UK public sector (for example, Department of Transport, in its evaluation of transport investment options) (Devlin and Sussex 2011),

In January 2009, NICE commissioned Professor Sir Ian Kennedy to carry out a short study of the way in which NICE values innovation when it appraises medicines (NICE 2009a). In response to the study, NICE modified its processes and documentation in order to achieve greater transparency in the way health benefits are taken into account. These changes relate to the way in which the Appraisal Committee’s deliberations are reported, but have not changed the way in which the decisions are made. However, in its submission to the Kennedy report, the Association of the British Pharmaceutical Industry called for a

> *“new structured approaches to decision-making to account for these important factors; and use of these factors should be far more transparent than currently.” The submission further suggests that “Where additional aspects of benefit and value cannot be incorporated within the QALY framework, evidence on them could be considered by NICE alongside the ICER. This will require a different decision making model capable of dealing with different sorts of evidence. Options include:*

- *the use of multi-criteria decision analysis (MCDA), where both the criteria themselves, and the weights applied to each, are explicit and transparent*

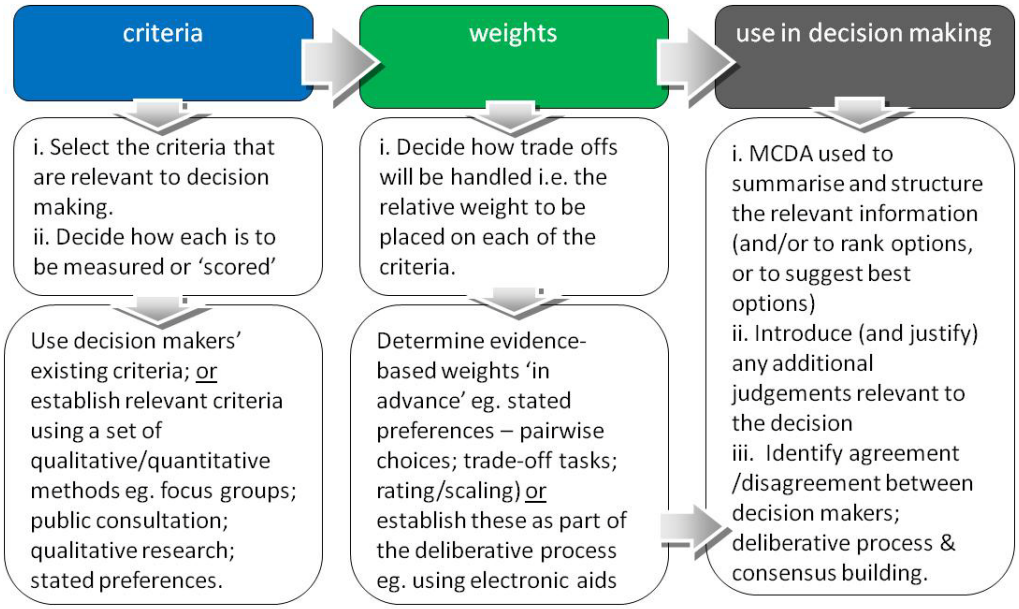
- *retention of QALYs as the principal measure of health outcome, and the ICER as the evidence on cost effectiveness, but other sorts of evidence being more formally and explicitly introduced and considered alongside these, either through MCDA or other means in a more transparent way."*

In broad terms MCDA can be regarded as a set of methods to aid decision-making, which make explicit the impact on the decision of multiple criteria that might be applied and the relative importance attached to them. This definition of MCDA encompasses a wide range of different approaches, both 'technical' and 'non-technical' in nature. Some types of MCDA involve algorithms to suggest optimal choices; others simply aim to provide some structure to a deliberative process. All aim to facilitate replicability and transparency in decision-making.

What is MCDA?

There are numerous different approaches to MCDA, which in various forms have been used in the NHS, other government departments and some HTA bodies in other countries (Devlin and Sussex 2011; Thokala 2011). All attempt to be clear about the criteria being taken into account, and the influence of multiple criteria on decisions. Beyond that, the methods and the way they are used in decision-making vary widely. An overview of the main elements is presented in Figure 1.

Figure 1 An overview of MCDA methods



Appropriately specifying MCDA requires the following questions to be addressed:

i. Which criteria should be included (see Section 3.1 below) and how can performance (that is, the extent to which a given technology achieves those criteria) be measured and scored (e.g., the criteria set out in Appendix 1 along with expected health (QALY) benefits)?

ii. What weights should be assigned to performance on each of the criteria (see Section 3.2 below)?

iii. How should the costs and opportunity costs of achieving an improvement in a composite (multi criteria) measure of benefit be considered (see Section 3.3)?

iv. Even if an appropriately specified MCDA process could be developed, unless the criteria and weights can fully reflect all aspects of social value then judgements will inevitably still need to be made. Therefore, how could the transparency of the deliberative process be improved and is there an appropriate form of MCDA that can aid rather than replace deliberative processes (see Section 3.4)?

From the outset it should be recognised that the NICE methods and process of appraisal already places it on the highly quantitative end of the spectrum of decision making that runs from the implicit and intuitive to the explicit and algorithmic. For example, decision analytic modelling is central to NICE's approach to technology appraisal, and represents an explicit, quantitative and evidence based way of transforming multiple criteria (e.g., impact on a range of clinical end-points, adverse events, resource use etc) into composite estimates of health gained (measured by QALYs) and net NHS costs. Furthermore, the QALY itself is an example of a rather sophisticated form of MCDA (see 3.2). It involves the aggregation of estimates of (changes in) life expectancy and health-related quality of life (HRQoL), where the latter is defined by different levels of performance across multiple dimensions (criteria) of health related quality of life, with a series of weights based on preferences. In NICE's Reference Case, these preferences are elicited from a sample of the general public, using stated preference techniques involving tradeoffs between length and quality of life.

The issue, therefore, is not whether NICE should use MCDA to support its decisions, but the extent to which such methods should be extended to bring together the various criteria NICE currently uses to inform its decisions or could use in the future. In other words, where on the spectrum of quantification should NICE locate its decision making approach? It is not the purpose of this briefing paper to argue for a particular location. Rather, the aim is to specify some of the key requirements that need to be adhered to if MCDA was to be more fully implemented within NICE methods, to identify some of the potential dangers of a poorly specified approach as well set out the potential benefits of a more accountable, consistent and predictable approach to making the necessary social value judgements.

3 Proposed issues for discussion

After consideration of the developments in this methodological area, the current Methods Guide and the requirements of the Institute’s Technology

Appraisal Programme, it is proposed that the following key areas are discussed at the workshop.

3.1 Which criteria might be included and how could performance be measured and scored?

Criteria as attributes of benefit

It is important to carefully determine which criteria or attributes should be included. In part, this involves careful consideration of which aspects of social value ought to be included alongside currently available measures of health benefit. Therefore, criteria should relate directly to attributes of a composite measure of social benefit.

A review of the use of MCDA in supporting resource allocation decisions elsewhere in health care (Thokala 2011), sometimes reveals a confusion about what are appropriately considered to be attributes of a measure of benefit and the necessity to consider the additional costs and opportunity costs associated with interventions that improve composite (multi-attribute) benefits (see Section 3.3).

Uncertainty and the relevance of evidence has sometimes been included as a separate and apparently independent attribute in some MCDA studies (Thokala 2011). This poses two difficulties:

i. All attributes of benefit, whether formally considered within a quantitative (MCDA) framework or a more deliberative approach, require evidence and will be estimated with uncertainty. The uncertainty associated with any composite measure of benefit and its expected consequences can inform research decisions and may also influence NICE Guidance if the type of research required cannot be conducted once a technology is approved or approval commits (opportunity) costs which cannot be recovered (Claxton, Palmer, Longworth, et al. 2011). Therefore, uncertainty and its consequences is not so much an attribute of benefit, but an important assessment to inform approval and research decisions intended to improve (multi-attribute) benefits for current and future

patient populations (the NICE Methods Guide Working Group will consider how only in research recommendations might be informed).

ii. Some examples of MCDA have included the quality and relevance of evidence as an attribute in its own right (Devlin and Sussex 2011; Thokala 2011). This risks confusing evidence about the effects of a technology on an attribute of benefit with choices about how important the attributes of benefit might be. It implies that the former can effectively be traded-off against the latter on the basis of preferences. This potential for confusing scientific and social value judgements should be avoided as it may threaten rather than enhance the transparency and accountability of the appraisal process. For example, important evidence might be disregarded on the basis of 'preference' rather than explicit consideration and reasoning with the implications fully explored so they can be scrutinised by stakeholders and ultimately held to account.

Characteristics of criteria

i. Criteria must be clearly defined and based on clearly articulated and generally accepted principles.

ii. To achieve the objectives of improved transparency, consistency and accountability the criteria and how performance would be measured and scored may need to be pre-specified so it can be applied consistently throughout the appraisal process.

iii. Specifying how the performance of an intervention in meeting each criterion is measured, including the type of evidence and analysis that would support any claims for improvement in the attribute, is also very important. Without it the assessment of performance may become subjective and unaccountable, undermining the very reason for taking a more quantitative approach

iv. Measures of performance might be based on the value of the attribute itself, e.g., QALYs gained or burden of disease, which would itself require careful definition with agreed and consistent measurement. However, other criteria might be categorical or qualitative (e.g. 'low',

'medium' or 'high'). Partly for this reason measures of performance are often expressed as performance scores on an ordered categorical scale (e.g., 1, 2, 3 etc). However, specifying how performance scores are related to measures of the attribute and the evidence required to support claims is important. It would also require better understanding of what constitutes 'high' or 'low' performance for each attribute (e.g. what is a high (or low) burden of disease in the NHS). Without it performance scores become subjective and might lead to lack of accountability since a judgment about the social value maybe conflated with scientific judgment about quality and interpretation of the evidence.

v. Criteria should be independent attributes of benefit. That is, they should not be alternative measures or proxies of the same underlying principle (e.g., evidence of clinical effectiveness and QALYs gained). If not there is a danger that the same attribute of benefit will be double counted when performance scores across the criteria are weighted. For the same reason the criteria should not significantly overlap and ideally should be separable and independent. Few of the criteria cited as potential candidates fully achieve this and, even those that come close (e.g., QALY gains and burden of disease), will often be related. If double counting is to be avoided the weighting of criteria would need to be much more sophisticated, providing weights of combinations of performance scores across different types of attribute (see Section 3.2).

vi. In principle, the criteria should represent a complete description of all the attributes judged to be of value and relevant to the type of decisions made in NICE appraisal. A complete description, which also meets all the requirements above, seems unlikely to be possible. Furthermore, inclusion, exclusion and measurement are likely to be contentious. Therefore, some form of deliberative process is still likely to be required (see 3.4).

How might criteria be selected?

i. A natural starting point might be the existing list of special circumstances described in NICE's social value judgements (NICE 2008b). However, it

ought to be recognised that this has been an evolving process, partly informed by the deliberative process of the NICE Citizen's Council and partly reflecting higher level concerns of the Department of health (DH) and secretary of state (SoS) (e.g., end of life (NICE 2009b)).

ii. Some of the calls for a more structured approach have also suggested adding or refining these criteria (e.g., alternative definitions of innovation (NICE 2009a)). Since relevant criteria are often disputed and the desire for completeness tends to conflict with the need to avoid double counting, consideration would need to be given to how they might be developed either though existing deliberative process of the Citizens Council or wider public consultation.

iii. However, it should also be recognised that criteria ought to reflect, or at least be consistent, with higher level objectives and policies (e.g., the SoS and DH). For example, the consultation on the Value Based Pricing (VBP) scheme, due to start in 2014, suggests that it will include criteria based on burden of illness, scale of therapeutic improvement, innovation, and wider social benefits alongside health benefits measured by QALYs (Department of Health 2010 and Claxton Sculpher and Carroll 2011). This poses a question of remit (who should ultimately be responsible for specifying the criteria), what coordination is required and when should this be done (i.e., extending MCDA prior to VBP may be premature). It is also not yet clear what analytic framework will be used to reflect these other aspects of value in VBP, i.e. some form of MCDA or applying weights when estimating costs and QALYs within existing methods of appraisal (see briefing paper on QALY weights).

3.2 How can weights be assigned to performance on each of the criteria?

Once criteria have been identified and the measurement of performance and any associated score defined, the weights to be applied to performance on each attribute need to be established.

How can weights be established?

The range of alternative approaches can be considered as falling into four broad areas:

i. Weights can be established as part of the decision making process itself, e.g., they can emerge during the process of decision making. Some MCDA approaches, such as 'decision conferencing' (Phillips 2006), help to structure those discussions, feeding back the decisions and implied weights via an iterative process. The outcome is a consensus on both the decisions themselves and the set of weights that have been applied. The advantage of this approach is that it would make the judgements that emerge from the deliberative process more explicit. The difficulty is that to achieve improved predictability and full consistency the weighting of attributes may need to be pre-specified so they can be applied consistently throughout the appraisal process, including across each of the four appraisal committees.

ii. It is also possible to conduct forms of sensitivity analysis by asking which criteria and weights would have to be deemed appropriate for each of the alternatives to be regarded as offering the most benefit. Although instructive to explore how sensitive decisions might be to the definition of criteria and specification of weights, it is unlikely that transparency and consistency would be improved in this way.

iii. Simple approaches which add up performance scores to arrive at an 'overall score' or number of 'benefit points' have been used and were proposed in submissions to the Kennedy review (e.g., Comprehensive Benefits and Value; Precision Health Economics (NICE 2009a and Thokala 2011). The problem with these rudimentary approaches is that the empirical question of performance is conflated with the question of social value. Their use would imply that each criterion was equally valuable or that the (sometimes arbitrary) scale for performance scores reflected relative social value. It would also imply that each of the attributes is valued in a separable and additive way (see below).

iv. Alternatively, weights might be pre-specified based on other evidence, gathered via related studies or processes. Sets of weights can be generated by asking selected participants to state their preferences. This draws on a set of well-established methods to uncover preferences about the importance of the various attributes (criteria) through the choices participant make between alternatives with different levels of the attributes to be valued (Ryan et al. 2008). These sorts of choice based exercises are widely used in health services research including NICE's use of QALYs where the weighting of HRQoL against the length of life uses choice-based methods (the value one attribute is expressed in terms of a willingness to forgo others). There are a number of approaches to preference elicitation which satisfy the choice-based criterion including standard gamble, time trade-off, as well as discrete choice experiments and contingent valuation methods. This logic of requiring choice-based methods of preference elicitation in NICE's current use of MCDA through QALYs would seem also to apply to the evidence required to inform the selection of weights in MCDA.

Who might provide the weights?

i. Improving transparency and consistency suggests that weights may need to be pre-specified rather than be determined by the Appraisal Committee during its deliberations. Since appropriate weights are questions of social value that are necessarily disputed, some claim for legitimacy, in terms of whose preferences are used to establish them, will be important. Therefore, adopting the view of any particular stakeholder group would seem inappropriate.

ii. Inclusive deliberative processes could be used, e.g., NICE's Citizens' Council has approached many of its topics by reflecting on the value of a given attribute on the basis of what others may need to be forgone to achieve it.

iii. NICE's current use of MCDA through HRQoL could be taken as a starting point where the preferred source of preference for weights, defining trade-offs between length of life and different attributes of

quality, is the UK general public. Although there is also a case for the use of patients' preferences for this purpose (Brazier, Akehurst, Brennan et al. 2005), few have advocated the adoption of the preferences of other stakeholders. The logic behind NICE's current use of public preferences to define weights within the QALY would seem relevant to deciding whose preferences should be used to supply the weights for a wider set of benefit attributes.

iv. Some potential criteria, however, are not directly related to the characteristics of patients or the type of health benefit, but to economic effects outside the NHS (e.g., wider social benefits). The relative weight that ought to be attached is not so much a preference but rests on estimates of the relative values of the NHS threshold and the consumption value of health (Claxton, Walker, Sculpher, et al. 2010). Which estimates of value are appropriate and which economic effects ought to be included and how they should be measured are judgments of social value. However, once these have been made, the appropriate weight (relative to health effects) is not so much a preference but a logical deduction (see Perspective briefing paper).

How can the weights be used?

i. Once appropriate weights have been assigned they need to be combined with measures of performance on each attribute. The most obvious approach is simple linear aggregation, i.e. each score on each criterion is multiplied by the weight for that criterion and these weighted scores are then summed to determine an overall score for that option, which may be compared to the scores for other options under consideration. This is a simple and very common approach in MCDA. However, there are serious drawbacks. It implies that attributes are valued in an additive and separable way, so the value of an improvement in one is independent of the level of that attribute and also of the levels of all the other attributes (i.e. the value of the combination of levels of attributes is simply the sum of its parts). In other contexts (e.g., HRQoL) this strong assumption generally doesn’t hold and would not be

regarded as acceptable. This problem is likely to be particularly acute when criteria inevitably overlap to some extent or are related in some way. Therefore, the need for completeness in specifying criteria combined with simple linear aggregation might mean that the alternative with the highest score might not necessarily offer the greatest social value and lead to decision based on MCDA that are widely regarded as unacceptable.

ii. This problem is widely recognised when constructing measure of HRQoL. For example, NICE's preferred measure (EQ-5D) comprises 5 dimensions (criteria) of quality of life each with 3 levels (performance scores). However, the tariff for EQ-5D (the weights for different possible combinations of levels of each attribute of quality) are not simply based on 5 weights (one for each criteria) or 15 weights reflecting every level in each dimensions (one for each performance level within each criteria) but a weight for each of the 243 possible combinations which define the possible health states. This is a considerable task, entirely comparable to the problem of weighing criteria in MCDA, which requires a large and representative sample of respondents (nevertheless some assumptions are still required). Measures of HRQoL have gone much further than most examples of MDCA in estimating weights (although some have used multi-attribute utility theory). Therefore adopting MCDA with weights that impose much stronger assumptions than are acceptable in current QALY measures are likely to be widely criticised especially when approval is restricted or withheld based on poor performance on some attributes. Relaxing these assumptions to provide a more complete tariff of weights for the possible combinations of levels of performance across all criteria would require a considerable valuation task but would not avoid all assumptions even if undertaken.

iii. Some approaches to MCDA seek to establish the dominance or extended dominance of options, by drawing on various ways of establishing weights and combining scores across criteria (e.g., strong dominance, outranking and data envelopment analysis). However, those

measures of dominance that are unaffected by assumptions of separability and additivity (e.g., an alternative is better on all criteria; or better on some criteria and no worse on others), is unlikely to have discriminatory power in most circumstances. Furthermore, the additional cost associated with an alternative also needs to be considered (see Section 3.3) even if it strongly dominates others in the multi-attribute benefit it offers.

3.3 How should the costs and opportunity costs of achieving an improvement in a composite measure of benefit be considered?

The criteria included in MCDA should relate directly to attributes of a composite measure of benefit. However, some of the recent calls for extending the use of MCDA for HTA bodies like NICE seem to have confused attributes of a measure of benefit and the necessity to consider the additional costs and opportunity costs associated with interventions that improve composite (multi-attribute) benefits by including cost-effectiveness (summarised as an ICER) as a criterion. Interestingly, where MCDA has actually been used to inform investment decisions in health care the attributes of benefit have been scored and weighted first and then the composite benefits of the options have been compared to their costs, sometimes summarised as a cost-benefit score (Wilson, Sussex, Macleod, et al. 2007; Epsom and St Helier University Hospitals Trust, 2009).

Weighting ICERs?

As outlined in Section 2 it is not cost-effectiveness (the ICER) per se that is an attribute of benefit but an assessment of the health benefits (in QALYs) and likely net health effects (also in QALYs) offered by the intervention. Of course, an ICER is related to both, although both require knowledge of the value of the denominator (not just the ratio) and the latter also requires knowledge of the numerator and an estimate of the threshold. Therefore, including an ICER as criteria to be weighted in MCDA poses a number of problems:

i. Since an ICER is derived from estimates of health effects and resource use it will not be mutually exclusive and will overlap considerably with others related to health effects and cost (e.g., evidence of clinical effect).

ii. Although ICERs are related to health gains offered, any weight assigned to an ICER implies different weights assigned to health benefits (because the ICER is a ratio). Without knowledge of the denominator and numerator in the ICER it is not possible to know the implied weight that is being assigned to the health (QALY) gains. Therefore, deriving weights that show how health gains should be traded against other aspects of social value cannot be achieved by asking respondents to weight ICERs. It is for this reason that implementation and evaluation of end of life criteria focuses on the weights that might be attached to QALY gains at the end of life rather than weights applied directly to ICERs or the threshold (NICE 2009b; Shah, Tsuchiya, and Wailoo 2011). Once weights for health gains (and other attributes) have been derived it is possible to solve for the implied equivalent weight attached to the ICER (or the threshold to be applied) for the particular intervention. However, this will differ depending on weights associated with other attributes, the numerator and denominator in the ICER and what other aspects of value are forgone due to additional costs (see below).

iii. In many NICE appraisals, including Single Technology Appraisal, there is more than one alternative to the technology being considered. In these circumstances, there are a number of ICERs that summarise the trade-off between QALYs gained and NHS cost. Weighting ICERs in MCDA, poses the question of which ICER to weight - with dangers of weighting inappropriate comparisons (comparators which are dominated or extendedly dominated).

Opportunity costs and the threshold

If attributes directly related to social benefits are specified and appropriate weights derived then the application of MCDA would generate an estimate of the additional composite (multi-attribute) benefit offered by each intervention, along with estimates of their additional cost, i.e., in the same way that current methods provide quantitative estimates of additional cost and QALY gains. Any decision will turn on whether the composite benefits gained are likely to

exceed the same composite benefits forgone due to the additional costs. It will require comparison with a threshold that not only reflects the QALYs forgone but also the other attributes associated with displaced NHS activities.

i. Current research to estimate the QALY threshold for the NHS is based on estimating how changes in expenditure and outcome are allocated across disease areas (groups of ICD codes) so can indicate the types of QALYs most likely to be forgone. Therefore, in principle, at least, any weights attached to the different types of health gained (e.g., burden of disease or other criteria that can be linked directly or indirectly to ICD code) can also be attached to the types of health forgone, providing an estimate of a weighted QALY threshold or a composite cost-benefit threshold. An ICER with a denominator of composite benefits could then be compared to a threshold for the same composite benefits.

ii. This is very important because if additional criteria are only applied to the benefits offered but are not reflected in opportunity costs, then decisions lead to more social value forgone than is gained; defeating the purpose of extending the use of MCDA because it may reduce rather than improve the definition of social value embodied in the section of criteria and weights.

iii. This also has an important implication which did not seem to be recognised in some of the submissions to Kennedy review (NICE 2009a) – given that budgets are fixed, incorporating other criteria (if done appropriately) will inevitably mean that some technologies, that would have been regarded as cost-effective based only on a QALY ICER, will be rejected or access restricted because they perform relatively poorly on some attributes compared to their comparators and/or what is likely to be forgone elsewhere in the NHS.

In some circumstances this problem of estimating a threshold that reflects the other attributes and their value that are likely to be forgone can be avoided.

i. If the circumstances described in Appendix 1 are indeed special, in the sense that they are very uncommon (in other NHS activities) then taking

them into account without suitable adjustment to the threshold might be reasonable on the basis that health and heath care associated with these characteristics are very unlikely to be forgone. This may be reasonable when special circumstances are narrowly defined as exceptions (even then it is an empirical question). However, extending the criteria to attributes which are more common or associated with all health effects (e.g., burden of illness) will require these aspects of value to be reflected in the threshold. Adding criteria to the benefits side which are not possible to incorporate in the opportunity cost side would seem self defeating – leading to decisions which reduce rather improve social value.

ii. If approval (investment) of a new technology could be considered alongside the current NHS activities which could be curtailed to accommodate the additional NHS costs, then all investment and matching disinvestment options could be evaluated using the same criteria and weights. Some applications of MCDA are undertaken in this way, e.g., its use in Programme Budgeting and Marginal Analysis. There are many examples of these sorts of approaches to decision making being used by Primary Care Trusts. Similarly, if the context is making an investment decision when the resources available to the decision maker have already been allocated specifically for that purpose, only the attributes of each of the options available within that budget constraint need be considered. In the longer term, there may be scope to develop a set of criteria and weights for use across the NHS. However, at present there is no mechanism for reconciling local and national priorities or for NICE to consider the specific disinvestments which would be required to accommodate a new technology. Therefore, the impact on the threshold of extending the use of MCDA cannot be avoided unless other criteria are restricted to exceptional and special circumstances.

3.4 How could the transparency of the deliberative process be improved?

The current deliberative process in NICE appraisal recognises that current measures of health gain (QALYs) cannot reflect all aspects of social value associated with the decisions that NICE must make. However, it also recognises that questions of social value are complex, nuanced and quite naturally disputed.

Moving to an entirely algorithmic process, where the only judgments required are ones of scientific rather than social value, would avoid deliberation. However, it would require criteria and weights to fully reflect all aspects of social value in a way that was regarded as legitimate and carry some broad consensus. The discussion in Sections 3.1, 3.2 and 3.3 suggested that this is unlikely to be possible. For example, the criteria would need to represent a complete description of all the attributes judged to be of value. This seems unlikely, not least because views about social value (the purpose of the NHS) quite legitimately differ and are disputed. Even if some broad consensus was possible about which attributes should be included, which weights should be applied and which assumptions are reasonable when doing so, are also not self evident. Therefore, extending the use of MCDA seems unlikely to avoid deliberation. Nor would it avoid disputes about social values and their relative weights when technologies are rejected or their use restricted and especially when some technologies, which would have been acceptable based on health gain alone, are unacceptable once other criteria are applied.

If a complete and legitimate description of social value is not possible then maybe the most important question is not whether extending quantitative use of MCDA can overcome some of the difficulties or substitute for deliberation, but how an unavoidably deliberative process can be improved in two respects: i) how the considerations are undertaken; and ii) how the reasoning and impact on decisions can be reported to improve transparency and accountability. This chimes well with the findings of the Kennedy review:

> *"Because I have concluded that those benefits which I say should be taken account of should (be – sic) incorporated into NICE's estimation of*

> *health gains as against health losses, the appraisal system should make it clear how this is to be done…But it must do so in a way that does not perpetuate the unfortunate idea, which could currently be entertained, that there is a methodology based on ICER/QALY and then there is some set of afterthoughts. If indeed social judgements, values or benefits do form part of NICE's appraisal as NICE claims and it is a "deliberative process", then they should overtly be identified as part of that deliberative approach…" (Kennedy Review 2009 p. 29-30 – emphasis added)*

The principles of MCDA may help to identify ways in which deliberation can be undertaken in a more structured and transparent way throughout the appraisal process, i.e., aiding rather than substituting for deliberative decision making.

For example, Appendix 2 illustrates a sort of simple recording template suggested by Devlin and Sussex 2011 that could be used. This could be seen as building on and extending the table that is currently provided at the end of the 'considerations' section of ACDs, FADs and Guidance. This would address some concerns about the lack of transparency in the importance attached to these 'other criteria', i.e. those not captured in the ICER, while preserving the character of the NICE deliberative process.

What are the options?

NICE *already* uses multiple criteria in its decision making: both quantitatively, through its use of decision analytic modelling and measures of HRQoL; and qualitatively, through its use of a deliberative process. The proposed introduction of value based pricing suggests that future decision making about new health care technologies is likely to be based on weighting the types of QALYs gained and forgone.

The question of what constitutes social value is inevitably complex, nuanced and disputed. There is no obvious broad consensus nor is this question one with a 'correct' empirical answer. For this reason deliberation is unavoidable. The crucial question is what form of quantitative analysis would provide the best (secure, accountable and evidence based) starting point for deliberation and decision?

The options for NICE range from:

- Taking health improvement as the primary purpose of the NHS, for which there might be some general broad consensus, and QALYs as the best currently available metric of health improvement, i.e., taking cost per QALY gained as the start point for deliberation, with some discretion in some limited circumstances (e.g., the metric of health improvement was shown not to capture important aspect of health). The primary role of the Appraisal Committee would be to make scientific value judgements about the evidence and analysis rather than social value judgements, i.e., representing early NICE appraisal prior to 2008.

- Take cost per QALY as the start point but incorporate other aspect of social value through deliberation (reported textually in the considerations section of Guidance), but indicate how considerations might influence decisions through application of the threshold, i.e. representing the current approach post 2008.

- The use of MCDA alongside and as a supplement to existing deliberative process, serving to structure those discussions; to feed back to decision makers the weights implicit in their decisions. The current approach to the cost effectiveness threshold range might potentially be maintained, but with more explicit reporting of the way that other criteria influenced a decision to accept a technology with an ICER within or above that range.

- The use of MCDA to identify, score and weight (for example, using weights derived from stated preferences exercises with the general public) multiple criteria, to determine some aggregate incremental benefit score, to be weighed up against incremental cost. Opportunity cost would therefore need to be considered in commensurate terms (e.g. as a ‘cost per benefit points’ threshold), so the cost effectiveness threshold would need to be re-assessed on that basis.

Appendix 1. Special weightings applied by NICE in making judgements about cost effectiveness.

NICE takes a number of factors into account – and these are "given special weighting when making judgements about cost effectiveness" (Rawlins et al. 2009). The factors noted by NICE, with the examples provided by Rawlins et al. (2009) of specific decisions where these factors were taken into account, are:

1. **Severity of the underlying illness**

More generous consideration is given to the acceptability of an ICER in serious conditions, reflecting society's priorities.

Taken into account in decisions about: Riluzole (for MND); Trastuzumab (advanced breast cancer); Imatinib (for chronic myeloid leukaemia); Imatinib (for gastrointestinal stromal tumour); Pemetrexed (for malignant mesothelioma); Omalizumab (for severe asthma); Sunitinib (for advanced renal cancer); and Lenalidomide (for multiple myeloma).

2. **End of life treatments**

The public places special value on treatments that prolong life at the end of life, providing that life is of reasonable quality.

Taken into account in decisions about: Riluzole (for MND); Imatinib (for gastrointestinal stromal tumour); Pemetrexed (for malignant mesothelioma); Sunitinib (for advanced renal cancer); and Lenalidomide (for multiple myeloma).

3. **Stakeholder persuasion**

Insights provided by stakeholders e.g. on the adequacy of the measures used in clinical trials in reflecting symptoms and quality of life.

Taken into account in decisions about: Riluzole (for MND); Ranibizumab (age related macular degeneration); Omalizumab (for severe asthma); Sunitinib (for advanced renal cancer); Somatropin (growth hormone deficiency); and Chronic subcutaneous insulin infusion (childhood type 1 diabetes).

4. **Significant innovation**

Some products may produce demonstrable and distinct benefits of a substantive nature, and which are not adequately captured in the quality of life measures.

Taken into account in decisions about: Trastuzumab (advanced breast cancer); Imatinib (chronic myloid leukaemia; Imatinib (for gastrointestinal stromal tumour); Ranibizumab (age related macular degeneration); Omalizumab (for severe asthma); Sunitinib (for

advanced renal cancer); Somatropin (growth hormone deficiency); and Lenalidomide (for multiple myeloma).

5. Disadvantaged populations

Special priority is given to improving the health of the most disadvantaged members of the population e.g. poorer people and ethnic minorities.

Taken into account in decisions about: Pemetrexed (for malignant mesothelioma).

6. Children

Given methodological challenges in assessing quality of life in children, society would prefer to give 'the benefit of the doubt'.

Taken into account in decisions about: Somatropin (growth hormone deficiency); and Chronic subcutaneous insulin infusion (childhood type 1 diabetes).

Source: Devlin and Sussex (2011), based on Rawlins et al (2009).

Appendix 2. A template for explicit and transparent consideration of social value judgements in NICE's deliberative process.

	To be considered at scoping:	To be considered at the appraisal committee:	
SVJ criteria	Relevant to this technology?	Record of committee's deliberations on each SVJ deemed relevant at scoping: key points considered (free text)	Summary of the committee's view of the importance of this SVJ in considering this technology: (1 = very important to 5 = not important)
End of life	Yes ❑ No ❑		
Severity	Yes ❑ No ❑		
Children	Yes ❑ No ❑		
Social disadvantage	Yes ❑ No ❑		
Small patient numbers	Yes ❑ No ❑		
Lack of alternative treatments	Yes ❑ No ❑		
Aspects of innovation not taken into account in the ICER	Yes ❑ No ❑		
(other_____)	Yes ❑ No ❑		
(other_____)	Yes ❑ No ❑		
(other_____)	Yes ❑ No ❑		
Record of the overall (combined) impact of SVJs on the decision about this technology with respect to the cost effectiveness threshold range: Most plausible ICER for this technology £_____ Implicit weight applied to QALYs gained from combined SVJs at £20k threshold*:_____ Implicit weight applied to QALYs gained from combined SVJs at £30k threshold*: _____ Summary of the overall influence of SVJs in the deliberative process for this technology: *"As the ICER of an intervention increases in the £20,000 to £30,000 range, an advisory body's judgement about its acceptability as an effective use of NHS resources should make explicit reference to the relevant factors… Above a most plausible ICER of £30,000 per QALY gained, advisory bodies will need to make an increasingly stronger case for supporting the intervention as an effective use of NHS resources…"(NICE 2008, p.19).			

Note: **The criteria shown in this template are illustrative only. This template is reproduced with permission from Devlin and Sussex (2011).**

4 References

Brazier, J., Akehurst, R., Brennan, A., et al. (2005). Should patients have a greater role in valuing health states? *Applied Health Economics and Health Policy* 4: 201-208.

Claxton K., Walker S., Sculpher MJ. And Palmer S. (2010). Appropriate perspectives for heath care decisions. Centre for Health Economics, University of York. CHE Research Paper 54.

Claxton, K., Sculpher, M. and Carroll, S. (2011). Value-based pricing for pharmaceuticals: Its role, specification and prospects in a newly devolved NHS. CHE Research Paper 60. http://www.york.ac.uk/media/che/documents/papers/researchpapers/CHERP60.pdf. York: Centre for Health Economics, University of York.

Claxton K., Palmer S., Longworth L. et al. 2011 Uncertainty and decision: when should health technologies be approved only in or with research? University of York; CHE Research Paper 69.

Culyer, A. J. (2009). Deliberative processes in decisions about health care technologies: combining different types of evidence, values, algorithms and people. London: Office of Health Economics.

Department of Health (2010). A New Value-Based Approach to the Pricing of Branded Medicines - a Consultation. London: Department of Health.

Devlin, N. J.,Sussex, J. (2011). Incorporating Multiple Criteria in HTA. Methods and Processes. London: Office of Health Economics.

Dowie, J. (2008). The future of HTA is MCDA. The future of Health Technology Assessment lies in the use of Multi-Criteria Decision Analysis. http://knol.google.com/k/the-future-of-hta-is-mcda#. Knol - A Unit of Knowledge.

House of Commons Health Committee (2008) National Institute of Health and Clinical Excellence. First Report of Session 2007-8. http://www.publications.parliament.uk/pa/cm200708/cmselect/cmhealth/27/27.pdf

Martin, S., Rice, N. and Smith, P. C. (2008). Does health care spending improve health outcomes? Evidence from English programme budgeting data. *Journal of Health Economics* 27: 826–842.

National Institute for Health and Clinical Excellence (NICE) (2008a). Guide to the Methods of Technology Appraisal. London: NICE.

NICE. (2008b) Social value judgements: principles for the development of NICE Guidance. www.nice.org.uk/aboutnice/howwework/socialvaluejudgements/socialvaluejudgements.jsp [Accessed 6 August 2010].

National Institute for Health and Clinical Excellence (2009a). Kennedy Study of Valuing Innovation: Submissions. http://www.nice.org.uk/aboutnice/howwework/researchanddevelopment/KennedyStudyOfValuingInnovationSubmissions.jsp (accessed 11/5/11). London: NICE.

National Institute for Health and Clinical Excellence (NICE) (2009). Appraising Life Extending End-of-Life Treatments. London: NICE.

National Institute for Health and Clinical Excellence (NICE) (2011) Citizens' Council Fact Sheet. http://www.nice.org.uk/newsroom/factsheets/citizenscouncil.jsp (accessed 11/5/11). London: NICE.

Peacock, S., Richardson, J., Carter, R., et al. (2007). Priority setting in health care using multi-attribute utility theory and programme budgeting and marginal analysis. Social Science and Medicine 64: 897-910.

Phillips, LD. (2006) Decision conferencing. Chapter 19 in: Operational Research working papers, LSEOR 06.85. Operational Research Group,

Department of Management, London School of Economics and Political Science, London, UK. http://eprints.lse.ac.uk/22712/ [Accessed August 10 2010]

Rawlins, M., Barnett, D. and Stevens, A. (2010), Pharmacoeconomics: NICE's approach to decision-making. British Journal of Clinical Pharmacology, 70: 346–349.

Ryan M, Gerard K, Amaya-Amaya M. (eds.) (2008) Using discrete choice experiments to value health and health care. Dordrecht: Springer.

Shah K, Tsuchiya A, Wailoo A. (2011) Valuing health at the end of life: report on pilot study. DSU report for NICE.

Shah K, Praet C, Devlin N, Sussex J, Parkin D, Appleby J. (2011) Is the aim of the health care system to maximise QALYs? OHE Research Paper 11/03. London: Office of Health Economics.

Tappenden P, Brazier J, Ratcliffe J, Chilcott J. (2007) A stated preference binary choice experiment to explore NICE decision-making. *Pharmacoeconomics* 25(8):685-693.

Thokala, P. (2011). Multiple Criteria Decision Analysis for Health Technology Assessment. Report from NICE Decision Support Unit. Sheffield: ScHARR, Sheffield.

Wilson E, Sussex J, Macleod C, Fordham R. (2007) Prioritizing health technologies in a Primary Care Trust. *Journal of Health Services Research and Policy* 12(2):80-85.

5 Authors

Prepared by Karl Claxton and Nancy Devlin on behalf of the Institute's Decision Support Unit.

Centre for Health Economics, University of York and Office of Health Economics

October 2011

Report to the Methods Review Working Party

Key issues arising from workshop on structured decision making

This report is written by members of the Institute's team of analysts. It is intended to highlight key issues arising from discussions at the workshop on structured decision making. It is not intended to provide a detailed account of all comments expressed at the workshop. The report has been written independently of the people who attended the workshop.

The report is circulated to the members of the Method's Review Working Party, the group responsible for updating the guide. For further details regarding the update of the Guide to the Methods of Technology Appraisal please visit the NICE website at http://www.nice.org.uk/aboutnice/howwework/devnicetech/technologyappraisalprocessguides/GuideToMethodsTA201112.jsp.

1 Summary

- Participants at the workshop addressed a number of questions raised by the briefing paper in five groups facilitated by representatives of the NICE Decision Support Unit.

- The majority of participants agreed that once the Committee has decided what the most plausible ICER is, the decision making process should remain deliberative and flexible, rather than moving towards a fully quantitative (or algorithmic) approach. However, it was noted that the adoption of additional criteria may require more structure to the deliberative process and to the appraisal documents.

- Overall, participants agreed that incorporating a more analytical decision-making process would be associated with practical difficulties with regard to development and interpretation, and the uncertainty around the inputs. Consequently they did not feel confident that moving towards a more quantitative approach would lead to better decisions.

- Participants generated many ideas as to how a set of additional criteria should be derived, without settling on any particular way of doing so, or on how criteria should be measured.

- Some participants at the workshop stated a preference for a pure cost per QALY approach, without consideration of any other criteria. Other participants proposed the following additional criteria (with no particular ranking): Disease severity, level of innovation, unmet need, affordability, rarity of disease, burden of illness, and equity and equality. However, many of these were disputed within the groups.

- The vast majority of participants did not recommend that NICE should attempt to assign weights to any additional criteria, suggesting that flexible deliberation is important rather than stringent rules. However, participants agreed that, if used, any weights and scoring systems must be transparent, rational, defensible and established through a choice-based framework which would require an extensive evaluation and consultation exercise. It was suggested that health-related criteria should ideally be weighted and incorporated into a revised QALY measure. Consequently, participants suggested that it may be more worthwhile to think about extending the measure of health used in appraisals and that improving the current EQ-5D measure would be more feasible and useful, or it may even help to make the list of additional criteria as short as possible. All other, non-health-related criteria would then be left to a deliberative process and only applied in exceptional circumstances.

- Participants generally agreed that it would be impossible without extensive research to define the opportunity cost resulting from an increased number of criteria. Participants felt strongly that if the NHS is expected to pay for additional benefits which are currently not included, then the baseline threshold would need to be reduced. Participants concluded that this problem could be overcome if additional criteria were adopted in only exceptional circumstances, when there would not be an effect on overall opportunity cost.

- Many but not all participants agreed that more transparency and consistency is needed on how additional criteria are discussed during Committee meetings and reported in the guidance documents. It was suggested that this could be captured through a checklist of criteria which should be referred to consistently in the submission template, the Committee discussions and the documents issued by NICE, but it was also stated that this could constrain a desirable level of Committee flexibility.

2 Questions posed to the workshop participants

1. *Would you support or resist a move to a(n even) more structured decision making framework at NICE? For what reason? Do you think that such a move would change the outcomes of technology appraisal decisions?*

2. *How could NICE derive a meaningful, legitimate and usable set of criteria to be considered in addition to quality of life gain and life extension? Which benefit criteria do you think should be considered; and rank these (including quality of life gain and life extension) as first order, second order, and third order?*

3. *Should (and say why) NICE broadly prefer:*

    i. *A pure cost per QALY calculation – defining QALY as broadly as at present (that is, all health benefits to the patient and other beneficiaries)*

    ii. *A cost per QALY calculation with additional flexibility to allow factors other than quality of life gain and life extension (for example, innovation and equity) to change judgements that might otherwise be above (or below) the cost-effectiveness threshold. [The present method].*

    iii. *Some adaptation with rules for additional criteria (possibly requiring the threshold to be reduced)*

iv. *A full MCDA quantifying benefit criteria such that non-QALY benefits can be factored into the decision.*

4. *How could a legitimate set of weights and scoring systems for criteria be derived to achieve a composite measure of benefit?*

5. *How should costs and opportunity costs of achieving a composite measure be considered (as not only health but the other attributes (criteria) will also be forgone following resource reallocation)?*

6. *How could the consistency and transparency of NICE's ACs' deliberative processes (whether the current or future) be improved?*

3 Summary of the workshop discussions

The workshop discussions addressed three distinct topics:

1. What type of structured decision making should be adopted?

2. Should additional criteria and weights be taken into consideration, and if so how should they be established, and how would that impact on the opportunity cost

3. Transparency and consistency in decision making

3.1 The degree of structure in the decision making

Workshop participants expressed the view that the current approach to decision making in Technology Appraisals is well respected. However, it was felt that it is not always clear to the public how the final decisions have been reached by the Committee and that it is important that the public is reassured that the current approach is appropriate. Participants suggested that more structure around the reporting of the Committees' deliberations (in the appraisal documents and on the NICE website) may overcome this issue (see Section 3).

Participants were asked where on the spectrum between full quantification and deliberation they would like NICE to locate its decision making approach.

The majority of participants stated a preference for an approach similar to the present method of using a cost per QALY calculation with additional flexibility to allow factors other than quality of life gain and extension to be part of the decision making. There was consensus that the deliberative process in Appraisal Committee meetings is fundamental to the appraisal process. It provides an opportunity for Committee members to express their views and to develop consensus decisions. In addition, participants noted that it has generally been accepted that quantitative approaches, such as a fully algorithmic MCDA, would not remove the need for deliberation and value judgements. It was also emphasised that deliberation of the decision making criteria should be undertaken with the same thoroughness for both positive and negative recommendations.

Detailed views on the discussed options are as follows:

3.1.1 *The option of staying with the current approach*

A large number of participants stated the current approach, meaning a cost per QALY calculation with additional flexibility during Committee deliberation, to allow factors other than quality of life gain and life extension, as their choice of approach. Participants highlighted that the current framework and criteria used by the Appraisal Committee to arrive at value judgements about new technologies is adequate, and that discussing additional factors during the deliberation process provides the necessary flexibility in the absence of hard evidence. Participants felt that deliberating the importance of such factors (for example innovation, disease severity and disease burden of illness) in an unstructured manner allows for context-specific discussions; therefore, it is entirely reasonable to expect some degree of inconsistency across appraisals.

3.1.2 *The option of the current approach with additional criteria requiring more structure*

A large number of participants also stated a preference for the current approach but with adaptations to allow for additional criteria and that this would require more structure to the deliberative process. These participants felt that extending the approach to explicitly include additional criteria is a necessary step in order to address the inconsistency in how these criteria are

viewed and interpreted across the four Appraisal Committees. One possible solution which some of the participants suggested was that health-related criteria could be weighted and incorporated into a revised QALY measure since these factors are frequently considered within appraisals. All other non-health-related criteria could be left to a deliberative process and are more likely to be seen only in exceptional circumstances. A few workshop participants suggested that a checklist should be presented to the Committee as a reminder of all of additional criteria which need to be considered before reaching a final conclusion, although some participants were concerned about a possibly stifling influence of a checklist approach on the deliberative process.

3.1.3 *The option of a fully quantitative approach, such as a fully algorithmic MCDA*

Participants were made aware that there have been advances in decision theory and use of MCDA in other areas of public sector decision making. Very few participants considered that a fully algorithmic MCDA could and should be incorporated into the appraisal process. Those that supported the fully quantitative approach suggested that a fuller quantification of a wider set of criteria provides the potential benefit that more appropriate decisions will be taken and that it would provide transparency and consistency across appraisals. However, other participants explained that such approaches, where used for public sector or health care decision making, have been poorly defined. For example, cost effectiveness was included as a benefit criterion or as criterion additional to effectiveness; uncertainty in, or quality and relevance of, the evidence was included as criteria in their own right, and there was double counting in selecting attributes, potential overlap between the criteria, and the issue about separability between criteria. It was cautioned that such potential aggregation of scientific and social value judgments could threaten rather than improve the transparency and accountability of the appraisal process.

The majority of participants felt that a fully algorithmic MCDA approach removes the element of discussion regarding the additional criteria and that it

would not be possible to appropriately specify all social judgements necessary. Also it was felt that aggregating criteria would be less transparent than the current approach of trying to disaggregate and discuss them separately.

Overall, participants agreed that moving towards a more quantitative decision-making process (beyond the ICER calculation) would be associated with practical difficulties with regard to development and interpretation, particularly considering the resource necessary to establish all necessary inputs (see Section 2), and the uncertainty around the inputs. Consequently most participants they did not feel confident that a more quantitative approach would lead to better decisions, and that there are not enough advantages associated with a MCDA approach and since the weighting for many criteria is likely to be small.

Participants were asked about their expectations whether or not adopting a MCDA approach would change the outcomes of technology appraisal decisions. Although this was a purely hypothetical question, the general expectation was that a more structured approach would not substantially change the outcomes of technology appraisals. There was a concern that it could lead to more incorrect than correct decisions, based on the uncertainties that would be associated with the assumptions feeding into the MCDA. Therefore, most participants agreed that undertaking a full MCDA would not add value to the appraisal process.

3.2 Establishing additional criteria, attributes, weights and scoring systems and the effect on the opportunity cost

Participants differentiated between two questions:

- which criteria to be considered,
- the weight that each criterion should have on the final decision.

Some participants noted that, at present, Committees only systematically consider clinical and cost-effectiveness, equality issues and the supplementary advice on end of life medicines. All other issues (such as

innovation, unmet clinical need etc) are not always considered for each appraisal. This raised a possible concern about inconsistency in decision making between the Committees.

Some participants questioned whether NICE should be generating a checklist list of additional criteria, other than quantity and quality of life, for its Committees to consider. A considerable number of participants at the workshop stated a preference for a pure cost per QALY approach, without consideration of any other criteria. These participants thought that quality of life gain and life extension were the only criteria that should be considered relevant to NICE and also went on to highlight that this approach would be transparent and consistent since there would be no subjective deliberative process. They claimed that this was the original approach taken in the early days of technology appraisal decision making, but others argued that additional factors have always been taken into account, and indeed the 2004 and 2008 versions of the Methods Guide reflect this.

Participants however agreed that if the QALY measure could be improved to be more sensitive and to cover all aspects of health, additional criteria would need to be considered less often.

Another issue with adding more criteria identified during the workshop was that the Committee already has limited time to consider all of the current criteria to be taken into account during an appraisal. Therefore some participants thought that adding more criteria would further complicate the process.

3.2.1 *How to derive a meaningful, legitimate and usable set of additional criteria*

- Participants generated many ideas as to how a set of criteria should be derived, without settling on any particular way of doing so. These included
    - existing relevant literature
    - criteria used in published NICE appraisals: It was suggested that an audit of all previous decisions should be undertaken to identify which

additional factors are most commonly considered by the Committee, and whether they are considered in a consistent manner across the Committees. It was further proposed that if there were more than two instances in which a criterion was deemed important, these could go into the list of criteria for consideration.

- a process similar to that used for the Kennedy report (stakeholder and academic submissions and independent evaluation)
- NICE's current stakeholder community
- Focus groups including: Suggestions for the composition of such focus groups varied widely from members of the general population, experts in the field, a similar make-up to the participants at the methods review workshops, representatives from NICE and from the Department of Health, Appraisal Committee members, Chair/vice chairs of the Appraisal Committee, Health Economists, to members of parliament.
- The public: It was generally felt that it would be best to canvas the public with a set list of criteria as otherwise too wide a range of opinions would be generated. It was thought that the public would tend to agree that all items on the list should be included, without fully understanding the implications. Therefore, participants thought that asking the public might produce useful information, but that this would not be a useful exercise for generating a final list of criteria.

3.2.2 *Which criteria?*

- If additional criteria were to be included, participants considered that they should only be taken into account in rare and exceptional circumstances. In these rare and exceptional circumstances, some participants thought that a clear list of criteria should be included in the Methods Guide to ensure consistency across appraisals. Some participants thought that only criteria linked to health should be considered.

- In addition to quantity and quality of life, the following criteria were proposed by some, but also disputed by others:
    - Disease severity: Many participants stated that baseline disease severity should be formally considered, by weighting QALYs for severity. This was because a given QALY gain for someone with very low baseline quality of life (e.g. motor neurone disease) was considered by many generally more valuable than the same QALY gain for a person with much better health (e.g. mild psoriasis).
    - Level of innovation: There were conflicting views as to whether innovation should be explicitly included and considered or not. If so, it was mentioned that the lack of innovation should also have an influence (such that for 'me too' drugs a penalty should be applied so that correct signals are sent to the industry about the value that the NHS places on innovative treatments). Participants considered that, at present, there is no agreed definition of 'innovation' and therefore it would be difficult to consistently value the innovativeness of a technology. Participants therefore considered that the impact of the level of innovation on the Committee's decision should be left to the deliberative process.
    - Unmet need was raised as a possible criterion to include, meaning that no alternative treatment options are available.
    - A few participants thought that affordability should be included as a criterion.
    - Several participants were not in favour of the existing End of life criteria.
    - Rarity of disease, burden of illness, and equality/equity could be explicit criteria, but most participants thought that the best method for dealing with these was through a deliberative process.

None of the groups of participants felt able to rank the criteria, or how criteria should be measured. Despite this, it is safe to say from the workshop discussions, that quantity and quality of life would be considered as so-called 'first order' attributes. Participants requested that explicit definitions must be provided for any criteria to be included. Participants also requested that as part of the NICE Method's Guide review it would be useful to revisit the End of Life criteria as well as differential discounting and to provide a strong scientific basis to underpin them.

3.2.3 *Weights and scoring systems for criteria*

Participants did not recommend that NICE should attempt to assign weights to any additional criteria, suggesting that flexible deliberation is important rather than stringent rules. However, participants agreed that, if used, weights and scoring systems must be transparent, rational, and defensible. Most importantly, the weights or scoring systems should be established through a choice-based framework where individuals show how they value one criterion in terms of their willingness to forgo one or more others. Participants agreed that it is only when faced with choice that people reveal how valuable something is to them. The importance of trade-off with health was emphasised as it was felt that the main objective of the NHS is to produce health and all other weights should be derived from how much health would be given up.

The majority of participants indicated that deriving weights and scoring criteria appropriately would involve a massive evaluation and consultation exercise and this added complexity may not result in any additional benefit to the decision making process. It was highlighted that the NICE process is very transparent and explicit and the disadvantages of incorporating a fully algorithmic MCDA approach including weights may outweigh any advantages. The concerns expressed were as follows:

- Criteria other than length of life and quality of life have previously not been the most important influence and formalising these additional criteria would make them disproportionately influential. The importance of balance and flexibility was emphasised and some participants considered that there may be a danger of making the decision making process too rigid.

- Some participants expressed the view that it would be impossible, because of a lack of evidence, to include a comprehensive evidence-based set of weights and scoring systems for a meaningful MCDA approach.

- There is often interdependence between attributes and it is rarely appropriate to assume that this relationship is additive. It was noted that the small body of research available in the literature will not translate into the context in which NICE has to make decisions as it would have been conducted in a much more closed setting.

- Issues around potential double-counting, adjustment of the threshold (see Section 3.2.4), how questions would be framed and thereafter integrated back into the QALY were discussed and it was stressed that, if incorporated, the technical detail around how this would be done would become very important.

- Some participants highlighted the fact that the ICER is often very uncertain anyway, and expressed concerns about attaching weights for additional criteria that are also associated with even greater uncertainty, and that this would not lead to improved decision making.

Participants were made aware by a workshop participant of a process called decision conferencing, whereby weights would be established as part of the decision making process, that is, weights would emerge from Committee precedent. However, this approach did not receive any support due to the danger of generating inconsistent decision making across appraisals. Also, most participants considered that the socio-economic profile of the Appraisal Committee would not be wide enough to develop a fully societal valuation. It was noted that existing NICE structures such as the Citizen's Council would be better placed for such an approach.

With respect to preference-based approaches, participants expressed concerns around the legitimacy of values obtained from the public as it was felt that unless the attributes related directly to health states the public may not be the best source, for example with respect to innovation. Moreover,

there may be potential for bias, for example in cases where it is thought that diseases are lifestyle-related or self-inflicted.

Consequently, participants suggested that it may be more worthwhile to think about extending the measure of health used in appraisals and that improving the current EQ-5D measure would be more feasible and more useful than assigning weights to criteria, or it may even help to make the list of additional criteria as short as possible. This should also cover elements of quality of life around convenience, satisfaction and wellbeing that are not fully reflected in the current methods. Others, however, thought that there are alternative methods of measuring HRQoL to pick up these differences, which are already permissible within the NICE reference case. If it is established that there are important quality of life differences that cannot be captured using the EQ-5D, a case can be put forward for use of for example the SF-6D, or using vignettes by employing the standard gamble or time trade off methods to get alternative estimates of health states, without needing to add any formal criteria.

Participants also expressed the expectation that the developments around Value Based Pricing would potentially inform the weighting of criteria such as disease burden, severity and innovation and should be fed into the decision-making process of Appraisal Committees.

3.2.4 *Costs and opportunity cost*

There was confusion amongst participants as to what the questions posed meant and participants found it challenging to answer this question before knowing what criteria could be included. Participants were aware that cost and opportunity costs were associated with an increased number of criteria in terms of resource needed to incorporate any additional criteria appropriately in the decision making and in terms of any increased uncertainty in the results. For the discussions, participants agreed to focus on 'opportunity cost' in terms of what would be displaced in the NHS and that an alternative phrasing of the question was how to decide the threshold.

The discussions ranged widely and touched in general on the difficulty of establishing the opportunity cost (or the currently used threshold range) due to the lack of knowledge about disinvestment decisions in the NHS.

Participants agreed that including more criteria outside of health would make it necessary to adjust the threshold, but for this it would be necessary to have more information about the cost effectiveness of what would be forgone, which is currently unknown. It was acknowledged that if the threshold were not changed in line with changes to the criteria for benefit the health service will have to be prepared to give up more services.

Part of the discussions therefore focussed on disinvestment decisions. One of the biggest problems was seen in that cost savings from NICE decisions are often made over the long term whereas the investments in the new technology are required immediately. Furthermore, savings can be made by small changes in referral criteria which can be difficult to identify. Therefore, it is challenging to observe what disinvestment decisions follow from NICE appraisals.

Participants were aware of the ongoing work on establishing the opportunity cost through the work on ICD codes and programme budgeting. However, the view was that quantifying opportunity costs is very challenging and it will only ever result in rough estimates. One suggestion was to randomly sample a set of NHS services and value them. This could be used to estimate the average opportunity cost of disinvesting in current NHS services.

Input from NHS commissioning showed that at the local level real life decisions already aim to displace the least cost-effective intervention. However, this was often not possible. Others thought that services get displaced that do not attract powerful lobbies, but that such services are often very cost effective. It was also stated that investment/ disinvestment decisions are often contained within departments (disease areas), but that this is not always possible.

One idea suggested by participants to avoid an impact of additional criteria on opportunity cost put forward was to try and balance weights such that overall

there is a zero sum. It was suggested to allocate positive and negative scoring to the additional criteria to ensure that criteria are applied in each appraisal in a balanced way. This may ensure that there is a zero sum gain over time. However, not all participants agreed that such approach is feasible and therefore, the idea of balancing the weights was not agreed by everyone.

Overall, most participants agreed that if additional criteria were taken into account in the decision making such as currently done with the End of Life criteria, this needs to be done in a way that has a 'symmetrical effect on the threshold', meaning that if the NHS is expected to pay for additional benefits which are currently not included, then the baseline threshold needs to be reduced. However, participants agreed that it would be impossible to define the opportunity cost resulting from an increased number of criteria. Participants agreed that this could be overcome if additional criteria were only adopted in exceptional circumstances, when there would not be an effect on overall opportunity cost.

3.3 Consistency and transparency

Participants expressed the view that NICE is by far the most transparent decision maker world-wide, particularly after the changes introduced in recent years, e.g. more detailed considerations, summary tables and public meetings. However, some participants agreed that there is still a need to explain the Committees' conclusions better, and for more consistency into how additional criteria are discussed and interpreted during Committee meetings and reported in the guidance documents. Participants also suggested that more explicit criteria would be useful for people using the guidance and for pharmaceutical companies, the latter of which would benefit from more predictability for pricing decisions. It was suggested that this could be captured through a checklist of criteria which should be referred to consistently in the submission template, the Committee discussions and the documents issued by NICE. In addition, participants suggested that it would help consistency having one or two Committee members being linked permanently to all Committees to ensure that the criteria is interpreted and viewed in the same way across all Committees.

Updating the Methods Guide was recognised as an important opportunity to further improve the transparency and consistency of the Appraisal Committees' deliberative process. It was suggested that the Methods Guide update could provide a more explicit, detailed framework of any additional criteria which may be considered in the appraisal, what evidence is required to justify the consideration of a criterion and also how this information will be assessed by the Appraisal Committees. This should then help inform manufacturers' expectations about NICE appraisals and also improve the predictability of outcomes for stakeholders. However, there was consensus amongst participants that because of lack of evidence it would be challenging to include an explicit list of the relevant criteria and respective weights, together with details of a revised threshold.

Suggestions to improve transparency of Appraisal Committee meetings included keeping as much of the discussion as possible in public and ensuring that all members participate to facilitate a collective decision making process. Suggestions for improving consistency in addition to the above mentioned checklist of the criteria was regular internal audit of the Appraisal Committees' deliberative process encompassing what criteria were considered and how they were dealt with. The audit could be done with standard clinical governance tools and could be used to demonstrate existing consistency between Committees and also to provide a source of shared learning to improve future consistency.

There was agreement that appropriate reporting of the appraisal outcome is an essential component of a transparent process, and that the considerations section in NICE guidance already provides an opportunity for the Appraisal Committee to give a detailed rationale for its decision. There was discussion about improving the structure of the guidance documents to make it easier for stakeholders and the public to understand the Committees' deliberation. The inclusion of a table as suggested in Appendix 2 of the briefing paper was agreed to be useful, but would need to be given more thought so that it expresses the final selection of criteria appropriately and removes double-counting. Also it was suggested that by having a list of other criteria alongside

the ICER, there may be a ‘crowding out’ effect due to the number of other criteria irrespective of their intended relative impact on the decision. It was mentioned that the current summary table could be extended or possibly using other formats such as graphical displays used by other agencies.

Overall, the broad consensus amongst the groups was that if there were additional criteria to be considered alongside cost-effectiveness, a more structured but still deliberative process is required to ensure that it is demonstrated how (not only that) all relevant criteria have been considered and to ensure transparency and consistency of the Appraisal Committees’ decisions and NICE’s accountability.

4 Key issues for consideration by Working party

1. Should the Appraisal Committees’ consideration of criteria other than clinical and cost effectiveness move towards the quantitative end of the decision making spectrum and away from the deliberative end?

2. If so,

    a) What would be the benefits of moving towards a more quantitative approach?

    b) Could a fully algorithmic MDCA approach be adopted?

3. Should additional criteria be explicitly included in the deliberative decision making process?

4. If so,

    a) How should such additional criteria be selected and by whom?

    b) Should the Methods Guide describe how additional criteria will be taken into account and what influence they should have on the decision making?

c) If so, should an explicit list of relevant criteria together with their respective weights be included in the Methods Guide?

5. Should the Methods Guide be configured such that the current set of 'additional supplementary advice' can be integrated and be part of one coherent approach?

6. If additional criteria are formally included, should the impact of considering these additional criteria on the displacement of treatment and services elsewhere in the NHS be explored?

7. If so,

   a) Should the current threshold range be amended to reflect an increased number of criteria?

   b) Should such impact be kept minimal by using additional criteria only in exceptional circumstances?

8. Should formal mechanisms put in place to ensure consistency between appraisals and Committees in the consideration of criteria other than clinical and cost effectiveness? If so how could this be done without affecting Committee independence?

9. Should the structure of ACDs and FADs be changed to more clearly explain the Appraisal Committees' deliberations?

5 Authors

Prepared by Elisabeth George on the basis of workshop feedback from Bernice Dillon, Claire McKenna, Eldon Spackman, Fiona Rinaldi, Hazel Squires, Helen Starkie, Jon Tosh, Kumar Perampaladas, Moni Choudhury, Penny Watson, Raisa Sidhu, Simon Walker, whose contributions are gratefully acknowledged.

November 2012