NATIONAL INSTITUTE FOR HEALTH AND CLINICAL EXCELLENCE

Briefing paper for methods review working party on companion diagnostics

The briefing paper is intended to provide a brief summary of the issues that are proposed for discussion by the Methods Review Working Party 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.

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 revised draft of the Methods Guide will be available for a 3-month public consultation, expected to begin in May 2011. We encourage all interested parties to take part in this consultation.

2 Background

2.1 What are companion diagnostics?

Companion diagnostics are tests that are typically developed to select patients who will benefit from specific treatments, usually pharmaceuticals, by improving the responder rates or decreasing side effects. The US FDA definition requires that the companion diagnostic provide "information that is essential for the safe and effective use of a corresponding therapeutic product". Most companion diagnostics use genetic or protein markers to identify patients who will benefit from targeted treatments. These markers to be measured by companion diagnostic tests are usually referred to in the marketing authorisation for the treatment. Examples of treatments based on specific markers appraised to date are shown in the table below.

Appraisal Title	**Marker**
Trastuzumab for the adjuvant treatment of early-stage HER2-positive breast cancer (TA107)	HER-2 (protein marker)
Bevacizumab and cetuximab for metastatic colorectal cancer (TA118)	EGFR (protein marker)
Cetuximab for the first-line treatment of metastatic colorectal cancer (TA176)	KRAS (genetic marker)
Gefitinib for the first-line treatment of locally advanced or metastatic non-small-cell lung cancer (TA192)	EGFR TK mutations (genetic marker)
Trastuzumab for the treatment of HER2-positive metastatic gastric cancer (TA208)	HER-2 (protein marker)

2.2 Regulatory requirements

The US FDA draft guidance on companion diagnostics generally requires that the companion diagnostic and the treatment be evaluated contemporaneously, although there are a number of exceptions. To date, the EMA does not explicitly deal with the evaluation of companion diagnostics. Diagnostics are regulated in accordance with the European In-Vitro Diagnostics Directive. Marketing authorisations granted for a pharmaceutical by the EMA may specify a patient sub-population requiring the testing for a

genetic or protein marker but the specific companion diagnostic to be used is not stated. In some cases the pharmaceutical SPC may indicate that only validated tests should be used.

2.3 Relevance of the topic to NICE technology appraisals

Increasingly, the marketing authorisations for new pharmaceuticals require the use of companion diagnostics. It is therefore important that within the appraisal of pharmaceuticals, adequate consideration is given to companion diagnostics. This should be balanced against the need to develop appraisals of pharmaceuticals with companion diagnostics within the normal resources and timeframes of the technology appraisals programme. NICE methods for the evaluation of companion diagnostics will develop over time and are likely to involve the technology appraisals and diagnostics assessment programmes. This review of the 2008 Technology Appraisals Methods Guide is an important opportunity to ensure adequate provision for the evaluation of pharmaceuticals requiring the use of companion diagnostic products.

The establishment of the new diagnostics assessment programme (DAP) has raised the profile of NICE with the diagnostics community and there is an expectation that NICE will evaluate companion diagnostics in conjunction with assessments of pharmaceuticals. The programme used to evaluate the diagnostic technology could be either TA or DAP depending on the question being considered. This briefing paper highlights key issues related to companion diagnostics that need consideration in the Methods Guide review.

2.4 Companion diagnostics in DAP or TA

When the marketing authorisation of a newly licensed drug includes the use of a diagnostic test to identify the eligible population the Appraisal Committee is likely to need to take the companion diagnostic into consideration when developing the guidance for the new drug. It would generally be inefficient to split the NICE processes between TA and DAP, and also this would not lead to timely guidance for the new drug. Taking account of the specific companion diagnostic used in clinical trials is also relatively straightforward as the patient outcomes observed in the trials are those from the treatment informed by that

specific companion diagnostic. Assessment of the pharmaceutical and companion diagnostic "package" can be undertaken in much the same way as for pharmaceuticals without companion diagnostics. However, in circumstances where alternative tests are available (e.g. proprietary test kits or "in-house tests" for the same marker that would fulfil the requirements of the pharmaceutical marketing authorisation), the amount of extra effort to fully evaluate these alternative options is likely to exceed the available resources and timeframe in technology appraisals.

When, after a drug is in established use, a diagnostic technology is introduced as a companion diagnostic to improve the responder rates, or decrease side effects, the diagnostic technology would typically by evaluated by either the Medical Technologies Evaluation programme (MTEP) or DAP rather than Technology Appraisals.

Companion diagnostics in Technology Appraisals

The 2008 Technology Appraisals Methods Guide refers to companion diagnostics in section 5.7.5 which reads:

"If the use of the technology is conditional on the outcome of a diagnostic test, the accuracy of the test and associated costs should be incorporated into the assessments of clinical and cost effectiveness."

In the 134 Technology Appraisals published since 2006 a specific diagnostic tool was described as part of the marketing authorisation and in the actual NICE recommendations of 47 Appraisals. Of these, the majority related to tools to assess disease severity, many included imaging, histology or other tests, and only the 5 listed in Table 1 could be referred to as true companion diagnostics.

The issues in previous appraisals around companion diagnostics were as follows:

1. Target population is a *post hoc* subgroup; The 2008 Technology Appraisals Methods Guide states: "The characteristics of patients in the

subgroup should be clearly defined and should preferably be identified on the basis of an *a priori* expectation of differential clinical or cost effectiveness due to known, biologically plausible mechanisms, social characteristics or other clearly justified factors." Often the information related to subgroups with a specific biomarker was not prospectively included in the trial.

2. Comparator data for a different population: If the data on the comparator technology are not from the clinical trial of the new pharmaceutical, then the comparator data will not usually be available for the specific target population.

3. Uncertainty over the use of the test in practice: Committee decisions were informed by clinical specialists' opinion, rather than firm evidence as to how the testing will be handled in clinical practice.

4. Test accuracy: The biggest issue relates to tests other than the specific one used in the clinical trial which may still fulfil the requirements of the marketing authorisation (e.g. alternative proprietary tests or "in-house tests" for the same marker). Often there is no evidence of the accuracy of the alternative test or its impact on the efficacy of the treatment. Tests may have serious false positive or negative rates impacting the value of testing/treatment. A second issue relates to changes over time of the knowledge base of what mutations are affected by the treatment. As more relevant mutations are discovered, the utility of any diagnostic test may change.

5. Testing increases costs for the NHS: The costs for testing all potentially eligible patients are included, but only those patients who get treated will benefit. Often the prevalence of the biomarker[1] is not known. A low prevalence of the biomarker means that more people are tested per patient identified to benefit from the new treatment which increases the cost per patient found and impacts cost effectiveness.

[1] In this paper the term "biomarker" is used in its general sense to include any biological marker that may affect the treatment. These can include nearly any lab result and is not restricted to protein markers.

The 2008 Technology Appraisals Methods Guide also includes some general coverage of diagnostics (see Appendix). Following the establishment of the DAP, standalone diagnostic technologies will not be assessed in TA and the relevance of these sections should be reviewed.

Companion diagnostics in the DAP

The DAP methods, designed for the assessment of diagnostics generally, are suitable for the assessment of multiple companion diagnostic options. They are also suitable for assessing diagnostics technologies with the potential to be used to improve the targeting or use of pharmaceuticals already used in clinical practice.

3 Proposed issues for discussion

It is expected that only the single companion diagnostic test option used in clinical trials would be fully considered in technology appraisals of pharmaceuticals since evaluating multiple diagnostic options would dramatically increase the time and resources required for the pharmaceutical evaluation. It is important, however, to acknowledge that other tests could potentially be used in clinical practice and that in using alterative tests, there is a risk that the alternative tests do not select exactly the same population as the test originally used in the clinical trials. Correspondingly different outcomes from treatment could also result. Management of this key issue within the technology appraisal of pharmaceuticals with companion diagnostics is important in ensuring optimal and cost effective use of the pharmaceuticals. This issue is avoided when the test used in clinical trials and considered within the technology appraisal is also adopted in clinical practice. In some cases, it may be possible to report the diagnostic accuracy of the test used in the clinical trials. Any alternative tests should then be validated and compared to the companion used in the trials prior to adoption. In many cases, however, diagnostic accuracy data (in this case, accuracy may mean the test's ability to predict treatment efficacy) may not be available – the only data available may be the trial outcomes resulting from treatment informed by

the test used in clinical trials. The specific test used in the clinical trial then becomes the “reference standard” with which alternatives should be compared. That is, the accuracy of alternatives tests is based on their agreement with the reference standard.

A challenge in the technology appraisal of pharmaceuticals with companion diagnostics is providing appropriate guidance and warnings on the potential use of alternative tests without detailed evaluation of the various test options. This could be as simple as a discussion within the committee considerations section or where appropriate, guidance on the diagnostics accuracy that would need to be demonstrated prior to the adoption of an alternative test. In particularly complex cases it may be appropriate to undertake a DAP assessment of the alternative companion diagnostic options following the initial technology appraisal.

A further key issue for the assessment of pharmaceuticals with companion diagnostics is how to handle the costs associated with the companion diagnostic testing. Even for pharmaceuticals that do not have companion diagnostics, the identification of patient populations for treatment often still requires significant diagnosis – and such costs are not normally included within the assessment.

Issue 1 – Should the costs of the companion diagnostic be included as part of the total costs in a technology appraisal of the treatment, and, if so, how does this impact the assessment?

Most technology appraisals start with an identified population that has been diagnosed. In this setting, the costs of the diagnostic process are not included and the diagnostic processes are generally assumed to be cost effective. The costs assessed usually begin with the treatment and include the costs of the treatment plus any further health costs influenced by the treatment or the disease in question. These can include costs of the disease and its further treatment as well as costs of dealing with the side effects stemming from the treatment or downstream treatments. The Diagnostics Assessment Programme, when assessing diagnostic tests, includes all costs stemming

from the point of the diagnostic test. The assumption is that the treatment and comparator are all cost effective.

It has been argued that one can differentiate between diagnostic processes that are carried out to diagnose a condition in general (and then choose from a number of established treatment options) and a diagnostic test that is carried out to make a decision for treatment with a specific drug. On that basis, it has been suggested that, when evaluating a treatment that has a companion diagnostic, the costs of testing should be included in the assessment. This is because the treatment cannot be initiated without the companion diagnostic and hence the cost of testing is part of the cost of treatment. However, as mentioned above, all treatments require some type of diagnosis before use, but the diagnostic costs are not generally included in appraisals of treatments.

For discussion:

1. Is it reasonable to include the diagnostic costs when looking at treatments with companion diagnostics, but not when treatments use diagnostic tests that are already is commonly in use?

2. If diagnostic costs are included in the appraisal, should it be required that separate ICERs be provided for the therapeutic with those diagnostic costs included and excluded?

3. If it is decided that costs and any direct outcomes for a companion diagnostic need to be included, what should be done when a further drug requiring the same particular companion test is subsequently appraised?

4. How should the situation be handled where the companion test is initially (but perhaps only initially) made "free" by the manufacturer?

Issue 2 – If a treatment is appraised that has been trialled with a particular companion diagnostic, what should the guidance say about the characteristics of the diagnostic test?

In most cases, information on the companion diagnostic that was used to select patients for the clinical trial(s) of the related treatment will be available. The diagnostic test will assess some marker (genetic, protein, or other) presumed to be relevant to the treatment efficacy. In some cases there will not be any other “gold standard” reference test available. However, it may be the case that the diagnostic test does not assess the marker perfectly and this may not be known. It also may not be known whether the treatment would be more effective if the test were perfect (i.e. 100% sensitive and 100% specific for the marker).

When alternative tests are available or likely to be available and used, then the question of relative accuracy becomes an issue. If there are trials of the treatment using the alternative test, then again those data would provide end outcomes directly and test accuracy, *per se*, is not an issue. If an alternative test is only compared to the test used in the trials and does not perfectly agree with that test in all cases, then there can be uncertainty about which test is more effective in maximising the benefits from the treatment.

For discussion:

1. Are there circumstances where it would be appropriate to recommend only the specific test used in the clinical trials even if this is not specified in the marketing authorisation?

2. If a true gold standard exists for the marker that has never been trialled with the treatment, under what circumstances can it be assumed that it is the appropriate marker for maximising treatment benefits? Where such a gold standard exists, should test accuracy standards (sensitivity and specificity) for alternative companion diagnostics be provided in the guidance?

3. If no such gold standard exists, should test accuracy standards relative to the companion diagnostic used in the clinical trial be provided in the guidance?

4. Alternatively, should general warnings be given on the potential consequences of using alternative companion diagnostics in the recommendations and/or committee considerations?

5. What information on the companion diagnostic used in the clinical trial(s) and the potential alternative tests should be requested as part of the manufacturer submission?

Issue 3 – Should the current sections on methods for assessing diagnostics continue to be included in the Technology Appraisals Methods Guide?

As TA will no longer appraise standalone diagnostics since those would be evaluated by MTEP or DAP, it may be appropriate to delete the current wording about diagnostics (see Appendix). A new section on companion diagnostics will probably be needed following consideration of the issues raised in this paper.

For discussion:

1. Should the current sections on diagnostics be deleted or replaced with a reference to the DAP programme manual?

4 References

FDA, Draft Guidance for Industry and Food and Drug Administration Staff - In Vitro Companion Diagnostic Devices,http://www.fda.gov/medicaldevices/deviceregulationandguidance/guidancedocuments/ucm262292.htm

NICE (December 2011) Diagnostics Assessment Programme Manual.

5 Author/s

Prepared by Hanan Bell, Nick Crabb, Elisabeth George

6 Appendix

General coverage of diagnostics in 2008 Technology Appraisals Methods Guide states

5.17 Diagnostic technologies can be used in different ways (for example, for disease identification, monitoring of disease progression and treatment, assessment of disease prognosis, or initial screening) and this should be reflected in the evidence submitted to the Institute.

5.18 Evidence for the appraisal of diagnostic technologies should normally incorporate evidence on the accuracy of the diagnostic technology. It is also important to incorporate the predicted changes in health outcomes and costs as a result of treatment decisions based on the test result.

5.1.9 The general principles guiding the assessment of the clinical and cost effectiveness of diagnostic technologies should be the same as for other technologies. However, particular consideration of the methods of analysis may be required, especially in relation to evidence synthesis. Evidence for the effectiveness of diagnostic technologies should include the costs and outcomes for people whose test results lead to an incorrect diagnosis as well as those who are correctly diagnosed.

5.1.10 As for other technologies, RCTs have the potential to capture the pathway of care involving diagnostic technologies, but their feasibility and availability may be limited. Other study designs should be assessed on the basis of their fitness for purpose, taking into consideration the aim of the study (for example, to evaluate outcomes, or to evaluate sensitivity and specificity) and the purpose of the diagnostic technology.

5.3.3 Assessments of diagnostic technologies should follow the general principles of systematic reviews as recommended here for other healthcare technologies. However, it is recognised that the specifics of, for example, the meta-analysis of studies of the sensitivity and specificity of diagnostic tests are different from reviews of the effects of therapeutic interventions. This is an area of active methodological research.

NATIONAL INSTITUTE FOR HEALTH AND CLINICAL EXCELLENCE

Briefing paper for methods review working party on choosing comparators

The briefing paper is intended to provide a brief summary of the issues that are proposed for discussion by the Methods Review Working Party 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.

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 revised versions were published in 2004 and 2008. 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 revised draft of the Methods Guide will be available for a 3-month public consultation, expected to begin in June 2012. We encourage all interested parties to take part in this consultation.

2 Background

2.1 Relevance of topic to NICE technology appraisals

Clinical and cost effectiveness are relative concepts. A technology cannot be described as “cost effective” *per se*, but is either cost effective (or not) in comparison to some other alternative. It is therefore critical that the additional costs and benefits of a new technology under appraisal by NICE are assessed relative to the appropriate comparator or comparators, to avoid a misleading view of the value of the new technology.

Whilst the choice of comparator can entirely change the assessment of cost effectiveness, there is frequently some judgement to be made about which is the appropriate comparator(s). The purpose of this paper is to highlight and discuss a) the current NICE guidance on the choice of comparator and to consider this in the light of the economic principles that underpin the use of cost effectiveness analysis, b) outline a series of issues that have arisen in past appraisals which collectively demonstrate those situations in which more detail in the Methods Guide may have been advantageous and c) present a number of issues for consideration that arise from a) and b).

2.2 What the current Methods Guide advises with respect to choosing comparators

The 2008 Methods Guide provides only broad guidance as to the selection of appropriate comparator(s). “Routine **and** best practice in the NHS” (emphasis added) is specified throughout. This wording helps to identify the set of potential comparators but does not provide any detail on which from that set should be selected as the basis for calculating the ICER, or if they are to be combined in some way, and if so, how that should be done. Furthermore, the guide does not specify whether “best practice” refers to the option that is most effective or most cost effective. Greater clarity here may help to resolve some of the challenging situations discussed below.

The following quotes exemplify the broad guidance found in the current methods guide:

"Technologies can be considered to be cost effective if their health benefits are greater than the opportunity costs measured in terms of the health benefits associated with programmes that may be displaced to fund the new technology." (Section 1.4.2.)

In relation to the scope, Section 2.2.4 of the Methods Guide states that:

- Relevant comparators are identified, with consideration given specifically to routine and best practice in the NHS (including existing NICE guidance) and to the natural history of the condition without suitable treatment.
- There will often be more than one relevant comparator technology because routine practice may vary across the NHS and because best alternative care may differ from routine NHS practice. For example, this may occur when new technologies are used inconsistently across the NHS.
- Relevant comparator technologies may also include those that do not have a marketing authorisation (or CE mark for medical devices) for the indication defined in the scope but that are used routinely for the indication in the NHS.
- Comparator technologies may include branded and non-proprietary (generic) drugs.
- Sometimes both technology and comparator form part of a treatment sequence, in which case the appraisal may need to compare alternative treatment sequences.

"Relevant comparators for the technology being appraised are those routinely used in the NHS, and therapies regarded as best practice when this differs from routine practice." (section 5. 1.1)

2.3 Guidance from the economic evaluation literature

The 2008 Methods Guide is consistent with standard economic theory to the extent that all relevant comparators are included within the set of technologies that are considered within an evaluation. Most standard texts (see for example Drummond *et al* 2005) describe how the set of potential comparators should be considered alongside the new technology of interest (we refer to this here

as the “decision set” for short). Texts then go on to outline the decision rules that should be implemented in order to identify the optimal choice from each of the comparators included within that set, that is, an incremental analysis. This detail is important because it is possible to calculate a ratio of difference in cost/difference in benefit between *every* pair of technologies in the decision set. There is the potential for such a set of pairwise comparisons to lead to confusion and they may be misleading. Some previous appraisal submissions have failed to include appropriate incremental analyses (see for example retigabine for epilepsy and trastuzumab for HER2 metastatic gastric cancer).The decision rules for incremental analysis are as follows.

- Where only two therapies are in the decision set, the relevant ratio on which to base decisions is the ratio of incremental cost to incremental benefit (the ICER).

- Where there are more than two options:

(adapted from Glick et al. 2007)

1	Rank order therapies in ascending order of either effect or cost
2	Eliminate therapies that are dominated
3	Compute ICERs for each of the remaining adjacent pairs of therapies
4	Eliminate therapies that have a smaller effect but a larger cost effectiveness ratio compared to the next highest ranked therapy (extended dominance)
5	Recalculate the ICERs for each remaining adjacent pair of therapies (steps 4 and 5 may need to be repeated)
6	Select the option with the largest ICER that is less than the maximum willingness to pay (i.e. the cost-effectiveness threshold)

These rules are consistent with the aim of identifying the technology from the decision set with the greatest measure of health benefit and a cost effectiveness ratio that does not exceed the cost effectiveness threshold. In the next section we consider the extent to which this process for identifying the optimal technology can be adopted in NICE Technology Appraisals.

2.4 NICE Technology Appraisals and the scope

In order to consider the relevance of the full incremental analysis as described above, or any other approach to defining appropriate comparators for NICE Technology Appraisals, it is necessary to consider the scope and broad aims of the programme as a whole. Clarity on the following issues will help to provide more detailed guidance, and therefore greater consistency, than that which currently exists in the 2008 guide.

- What is the relevance of current NHS practice when that is not also best practice? Should best practice be defined as the most effective alternative or should it be the most cost effective alternative?

- Should decision rules about appropriate comparators be based on consideration of the set of options that are directly the subject of the specific technology appraisal guidance i.e. the appraised technologies (in which case there is a clear difference between STA and MTA)?

- Alternatively, should the guidance that could be issued via other NICE programmes also be considered relevant in considering appropriate comparators when formulating Technology Appraisal guidance? Are any other ways in which NHS practice could be influenced, beyond those routes open to NICE, relevant when considering which comparator is appropriate?

There are several situations that have arisen in previous appraisals where there is a conflict between the technology that may be considered optimal according to the decision rules that are standard for economic evaluation and the guidance that NICE is able to publish as part of the Technology Appraisal process. The examples discussed below all demonstrate how these conflicts arise as the result of two issues:

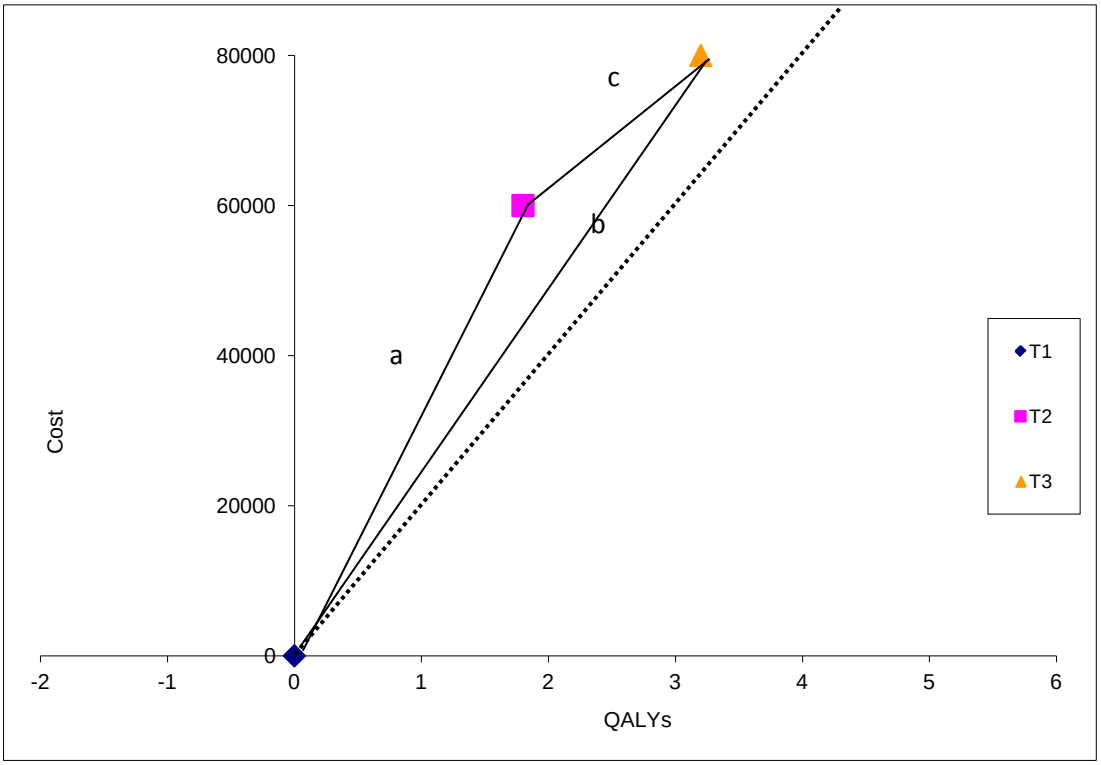
i) the fact that the guidance that NICE may issue as part of a specific Technology Appraisal does not always extend to all potential comparators.

ii) One or more of the comparator technologies in NHS use is not cost effective

Figure 1 illustrates a situation where there are three technologies that form the decision set and an assumed threshold value of £20,000 per QALY gained (represented by the dashed line). Standard decision rules would conclude that the optimal technology is T1. This is because neither options T2 nor T3 have an ICER compared to T1 that is below the threshold of £20k (the gradients of lines *a* and *b* are steeper than the dashed line). It is also the case that T2 would be excluded on the basis of extended dominance, that is, there is a combination of T1 and T3 that would cost less and generate more QALYs than T2.

It should be noted that the gradient of line *c*, the cost effectiveness ratio of T3 compared to T2, is less steep than the dashed threshold line i.e. the ICER is less than £20,000 per QALY gained. If T1 is not an option, then T3 is preferable to T2.

Figure 1: Cost effectiveness plane where a comparator is not cost effective



Given the current NICE process, there are situations whereby the existence of T1 may be deemed by some to be irrelevant. In these situations, it may be argued that "*c*" does represent the appropriate comparison for the problem at hand since this represents the differences in costs and benefits that will occur in the NHS depending on whether NICE Technology Appraisal guidance is positive or negative.

As previously mentioned, if the "best" alternative practice is defined in terms of clinical rather than cost effectiveness then this situation may arise (T2 is more effective than T1). Likewise, if the focus is on current NHS practice then it is feasible that this is T2 rather than T1. This problem may be more acute in the Single Technology Appraisal (STA) process, where the focus is T3 as the new technology, since the only guidance possible is either to recommend or not recommend T3. No recommendations will be made directly about T1 or T2 within the appraisal since these lie outside the remit. In the Multiple Technology Appraisal (MTA) process the problem will occur if neither T1 or T2 are among the technologies specified in the remit of the appraisal i.e. it is not possible to issue guidance on T1 or T2 as part of the appraisal. If T1 alone is included then the recommendation is likely to be for T1 only because T3 is not cost effective relative to it. If T2 alone is included then the appraisal might not recommend either T2 or T3 which would only leave T1 despite their being no formal NICE guidance on it.

It is therefore clear that T3 can be recommended in those situations where there is a focus on current NHS activity as the comparator, where the comparator is defined as "best practice" in terms of clinical effectiveness rather than cost effectiveness and where the optimal strategy is to be chosen only from those which NICE may directly issue guidance on within an appraisal (the appraised technologies) . This latter point requires that the broader set of activities in which NICE, or the NHS in general, may engage in order to influence NHS practice are not considered relevant to the issue of comparators in an appraisal.

There are several reasons why T2 may be current NHS practice, despite the fact that it is cost ineffective compared to T1. For example

- T2 has not been appraised by NICE. This can include the possibility that relevant comparators emerge or are only licensed after the point at which a scope is produced for a particular appraisal.
- T2 may represent off-label use for the specific indication in question. This does not currently rule it out as a comparator in either the STA or MTA process but it does mean that NICE would normally not be able to issue guidance on its use in the NHS as part of the Technology Appraisals Programme. However, it is worth noting that NICE Clinical Guidelines can make recommendations regarding off-label use (NICE Guidelines Manual p.110). This situation is most common in paediatrics, although it has also featured in a number of non-paediatric technology appraisals. This can make it difficult to ascertain whether a treatment really does represent routine practice in the NHS and therefore can increase the debate as to whether a treatment is an appropriate comparator in accordance with the strict definition given in the current Methods Guide.
- T2 may have been appraised by NICE but the technology has been adopted in the NHS contrary to NICE guidance. One previous example of this situation relates to the appraisal of natalizumab for Multiple Sclerosis (MS). Within this STA, it was accepted that current NHS treatment for these patients is beta interferon or glatiramer acetate, provided by the Department of Health supported “Risk Sharing Scheme”, which permits patients to continue to receive these treatments despite the fact that NICE did not recommend them for NHS use on the basis of their cost effectiveness.
- A similar issue is likely to arise in relation to the Cancer Drugs Fund (CDF). The CDF aims to ensure that drugs which have been deemed cost ineffective and are therefore not recommended by NICE are still made available to NHS patients in England only. It is administered regionally around Strategic Health Authority established panels and is intended to be a temporary measure until the expiry of the Pharmaceutical Price Regulation Scheme (PPRS) at the end of 2013. This temporary nature of

the scheme may distinguish these treatments from those provided through other means in the NHS.

In each of these situations, it can be argued, and indeed has been in previous appraisals (see for example Natalizumab for MS), that the “theory of the second best” becomes relevant. Essentially this accepts that efficiency within the limited set of NHS options that NICE Technology Appraisals can influence is the goal of the appraisal. Current NHS practice may be cost ineffective, but if this is not something that NICE Appraisal guidance is able to advise on, then further departures from an inefficient situation may be warranted. A broader view of the remit of the Technology Appraisals Committee, for example that includes as part of its considerations the range of NICE activities that may, at some point in the future, allow a much broader set of guidance to be issued, would lead to a different conclusion. Indeed, a view that considers a full range of NHS activities including disinvestment, implementation and research may be one which provides a rational framework for the consideration of costs, benefits and their associated uncertainties.

One important implication of this view, if accepted for all the various situations highlighted above, is that the guidance that emerges from the MTA process may sometimes be very different from that which would emerge from the STA process. The scope of an STA is limited to issuing guidance on the use of the new technology, whereas an MTA would seek to issue guidance about all technologies in the decision set in many, though not all the examples above.

In each of these situations it is also worth noting that the patient group could be perceived as having already benefitted more than other groups since a non cost effective therapy is available to them. To issue positive guidance for another new technology on the basis of a comparison to a cost ineffective alternative may be seen to exacerbate an already unfair situation.

There are practical issues that must be considered if the view is taken that “best” practice rather than current NHS practice is the relevant comparator. Particularly important is the potential for comparators to emerge after the scope for an appraisal has been finalised. Such comparators may have

gained licensing approval in the interim, or even been through the Appraisal Process. This means that at the point of Committee consideration and guidance production, the appropriate comparators may have changed, be at the point of changing, or be subject to consideration at the same point in time as the technology in the scope in preparation. Indeed, it may be reasonable to assume that such a new comparator would become future NHS practice. In such situations, there are obvious challenges to the submitting manufacturers in terms of access to data on clinical effectiveness (and other parameter values) relating to the new comparator, as well as the time constraints of the appraisal process.

In many settings standard NHS practice may be clear. However, there are situations in which standard care will vary substantially and a number of different comparators may each comprise a significant proportion of current care. There have been situations where it has been argued that the additional costs and benefits of the new technology should be calculated against some form of “average” costs and benefits associated with the mix of current approaches, sometimes referred to as a “blended comparator”. This could be seen as an appropriate approach if, as described above, the goal of a NICE appraisal is considered to be restricted to identifying whether a single new technology is efficient compared to current NHS practice as a whole. The approach also assumes that the displacement of existing practices will occur in the same proportion as in current use. It should also be clarified that where different comparators can be identified for identifiable patient groups then these should be dealt with as separate subgroups. Furthermore, there are several practical issues to be considered even if this goal is considered appropriate. These include:

How should the “average” be determined? Should weights be applied to each of the comparator technologies according to their estimated NHS use? Where should estimates of use come from?

Accepting a blended comparator approach will require at least some NHS practitioners to switch away from their current treatment approach to a new

NICE recommended approach that is relatively cost ineffective and may even be less clinically effective. At the extreme this could entail switching to a less clinically effective option.

This situation arose in relation to the appraisal of lapatinib for the treatment of women with previously treated or metastatic breast cancer. In this case, the manufacturer argued that lapatinib was cost effective compared to trastuzumab-containing treatment regimens and that these were in widespread NHS use. Lapatinib did not appear likely to be cost effective compared to other potential comparators. The manufacturer presented a "blended comparator", which was comprised of a weighted average of the costs and benefits of three treatments, including trastuzumab, where weights were estimated from market research data. The NICE Decision Support Unit (DSU) report on this appraisal argued against the concept of the blended comparator and the Appraisal Committee also adopted this view which was upheld at appeal. However, it should be noted that there are several issues specific to this appraisal that the Appraisal Committee considered pertinent and that may make it inappropriate to infer that the concept of the "blended comparator" was rejected in principle. In particular, a forthcoming NICE guideline regarding the use of trastuzumab containing therapies, that trastuzumab was being used in an unlicensed indication in this situation and the lack of evidence of the magnitude of treatment effect were considered relevant factors.

The "blended comparator" has been raised in other appraisals, including that of azacitidine for the treatment of myelodysplastic syndromes, chronic myelomonocytic leukaemia and acute myeloid leukaemia. Following appeal the committee did accept a blended comparator comprising of best supportive care, low dose chemotherapy and standard dose chemotherapy. The ACD is clear however that several appraisal specific issues led the committee to accept this comparator as the basis for decision making only in this specific instance rather than accepting this as a general decision rule. In particular, the Committee heard that the populations for each of the comparator conventional care regimens could not be clearly defined.

Further examples of potential differences between full incremental and pairwise cost effectiveness ratios

There are many technologies which could form part of a sequence of treatments for individual patients. In these situations, the fundamental principles of economic evaluation still apply. Namely, the costs and benefits of each feasible alternative sequence should be compared in an incremental fashion. This approach considers each sequence of treatments, including a sequence that excludes the new technology entirely, as if they were separate individual treatments and it is correct to do so because they are mutually exclusive: patients can only receive one sequence.

An example of an appraisal where this issue was debated was tocilizumab for the treatment of rheumatoid arthritis. Here, the manufacturer compared a number of different treatment sequences that included tocilizumab in a pairwise fashion to a sequence that excluded tocilizumab (current treatment). Since all generated ICERS were approximately equal it was argued that guidance should permit tocilizumab in any position in the sequence, including as first-line treatment. A full incremental analysis gave very different results and suggested the optimal sequence was one where tocilizumab is used as a second-line treatment within the sequence.

There may also be differences between the incremental and pairwise approaches when consideration of different patient subgroups or strategies for using a technology are considered. For example, it is often the case that separate subgroups of patients can be identified distinguished by those that are naive to currently available treatment and those that are not. For the naive group comparisons can be made between the new therapy and both current care and “do nothing”. Within the licensed indication for a new therapy it is possible to consider a range of different uses of that therapy. NICE appraisals often consider starting and stopping rules for example. As with the use of therapies in a sequence of treatments, these strategies can be considered mutually exclusive and therefore an incremental analysis may be appropriate.

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 by the Methods Guide Review Working Party.

1. What are the general principles that should govern the selection of comparators? Specifically, should NICE Technology Appraisals focus on comparing to "best" or "standard NHS" practice? If "best" should this be defined in terms of clinical or cost effectiveness?

2. Should the Appraisals committee consider only the narrow set of options that can be influenced directly by its guidance or should a broader view be taken? If the latter, what boundaries should be set e.g. the set of NICE activities as a whole, the broader NHS?

3. Should NICE appraisals ever consider the relevant ICER for a new technology to be that compared to a technology that itself is not cost effective, though in use in the NHS? If so, in which circumstances?

    - Comparator recommended for use by DoH despite NICE guidance
    - Comparator available via Cancer Drugs Fund
    - Comparator not been appraised by NICE (does it matter why not appraised? Too new, not licensed, other?)

4. In which circumstances, if any, is it appropriate to consider the comparator to be a "blend" of other options?

5. Should the identification of comparators (during the scoping stage) be focussed on providing a 'protocol' for the appraisal (i.e. a binding list of comparators to be used in the appraisal) or as more of a source of information about all possible options to be defined during the appraisal?

- If the latter approach is taken, should clear guidance on choosing options for the Committee be given?

4 References

Drummond, M.F., Sculpher, M.J., Torrance, G.W., O'Brien, B.J., Stoddart, G.L. (2005) Methods for the Economic Evaluation of health care programmes, Oxford: Oxford University Press.

Glick, H.A., Doshi, J., Sonnad, S.S., Polsky, D. (2007) Economic Evaluation in Clinical Trials, Oxford: Oxford University Press.

NICE (2009) The Guidelines Manual

5 Author/s

This briefing paper has been prepared by Allan Wailoo, Rebecca Trowman and Andrew Stevens.

NATIONAL INSTITUTE FOR HEALTH AND CLINICAL EXCELLENCE

Briefing paper for methods review working party on costs

The briefing paper is intended to provide a brief summary of the issues that are proposed for discussion by the Methods Review Working Party 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.

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 revised versions were published in 2004 and 2008. 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 revised draft of the Methods Guide will be available for a 3-month public consultation, expected to begin in June 2012. We encourage all interested parties to take part in this consultation.

2 Background

2.1 Relevance of topic to NICE technology appraisals

When appraising the cost effectiveness of a technology, it is critical to get an accurate estimate of the true costs associated with its use. Costs that must be calculated include the actual purchase costs of the intervention of interest and any comparators, but also administration costs and costs associated with living with the condition (for example monitoring, any additional treatments, potentially palliation) and so on. Without accurate costs to input into an economic model, there is a risk that the subsequent cost-effectiveness estimates (those used for decision-making) may be inaccurate and even unreliable.

Particular issues that arise when ascertaining the costs to use in an economic model include the use of Healthcare Resource Group (HRG) codes and the use of list prices. The varying use of HRG codes and of list prices in technology appraisals can lead to inconsistent cost-effectiveness estimates, which could in turn result in differing decisions and recommendations on the use of technologies. Therefore greater clarity and direction surrounding the use of HRG codes and list prices within technology appraisals would be beneficial for those that create evidence submissions and for those that have to use them for decision-making purposes.

2.2 Introduction to costs

When calculating the costs associated with a technology, there are a number of issues that must be considered. Firstly the purchase price of the technology, but also associated costs, such as the administration of the technology and additional treatments that will be given (such as pain medication) and length of stay in hospital, rehabilitation and so on. In order to do this, Healthcare Resource Groups (HRGs) are used. HRGs are standard groupings of clinically similar treatments which use common levels of healthcare resources (http://www.ic.nhs.uk/services/the-casemix-service/new-to-this-service/what-are-healthcare-resource-groups-hrgs). For example “complex neurosurgical pain procedures” is a HRG code. HRGs are used as

they are readily available, standardised, estimates of what a particular treatment for a condition will cost the NHS. They can reduce the need for local micro-costing (that is, costing of each individual component that is involved along a pathway of care in the NHS). A further benefit is that for acute care they are readily understood by people working within the NHS due to them becoming the main contracting currency.

However, there are issues when using HRGs in determining the costs to be used in an economic model. It is possible that the HRG codes used in calculating treatment costs are incomplete or incorrectly compiled and the costs are therefore underestimated. This is however becoming increasingly less likely as Reference Cost submissions have to reconcile to an organisation's annual accounts to ensure full cost recovery. Furthermore, the Audit Commission have undertaken an assurance programme since 2007 and note increasing improvements. Latest figures suggest that coding errors found were 0.03% of total Payment By Results (PbR) expenditure.

HRG codes can be considered too crude and it is possible that they do not adequately discern between treatments for a particular condition (for example if one chemotherapy takes much less time to administer than another, but it is still included in the costlier HRG). Sometimes, the HRGs are costed in such a way that they do not represent the totality of costs within one HRG. For example, there will be one cost for an admission for cardiac arrest, but if the patient has spent time in critical care then this will be captured and costed as a separate HRG. This unbundling is designed to make the costs of high cost care more visible, or to facilitate delivery of care across different organisations, but it also makes it more complex for health economists to use them accurately. Additionally, it is possible that if a new intervention provides innovative benefits (such as reduced administration time, reduced monitoring requirements and so on), then it may be unfair to use an existing HRG that is no longer reflective of the intervention of interest. It is therefore crucial that the Appraisal Committee are presented with sufficient detail to ascertain whether or not the HRG codes have been applied correctly and appropriately in a technology appraisal.

Recently the HRG4 has been published, and this major revision to HRG codes introduced new groupings, which increased from 650 to over 1,400. The new and updated groupings are intended to more accurately reflect treatment pathways in the NHS, with more refinement and consideration of disease severity and associated complications and comorbidities. Whilst the HRG4 is likely to improve granularity and accuracy of costs for individual HRGs, it may also make it more complicated for analysts to determine which HRGs are most relevant.

When HRGs are considered inappropriate for use, it may be possible to micro-cost every component of the treatment pathway using costs from other sources (such as from existing literature, from other countries, from registry data or from surveys and/or clinical opinion). In these circumstances, clear justification as to why HRGs are not used and full details of the methodology that has been used are rarely presented to the Appraisal Committee. This can mean that the exact components that contribute to the cost estimates can be unclear and without confidence in the costing estimates the robustness and reliability of the subsequent cost-effectiveness estimates can therefore be reduced.

A further issue in estimating the costs of technologies is the use of list prices or prices that are discounted when purchased in the NHS. List prices of a technology are those that are set nationally and are available in the British National Formulary which is currently updated twice per year, however, the NHS Electronic Drug Tariff (http://www.ppa.org.uk/ppa/edt_intro.htm) is updated monthly and includes costs for all drugs prescribed within primary care. However, very often the price that the NHS actually pays for a technology can be much lower than the list price due to discounts that have been negotiated with the supplier when buying in bulk. This is particularly the case for technologies that are off patent and technologies that are widely used. This is however rare for newer technologies to be discounted when first launched, unless it is part of a patient access scheme agreed with the Department of Health. This in itself gives rise to issues of transparency as the level of discount within a patient access scheme is often held as commercial-

in-confidence. In some technology appraisals the effect on the cost-effectiveness estimates in using a list price or a price with an NHS discount can be substantial. For example if a comparator is off patent and available to the NHS at a heavily discounted price but the intervention of interest is new and no discounts are available, then an analysis using list prices will result in a small cost difference between the technologies, whereas an analysis using the prices with the NHS discounts will result in a much larger difference between the two.

It is important that if prices with NHS discounts are used instead of list prices in a cost effectiveness analysis, that they are nationally available throughout the NHS and it is clear how long the discounts will apply for. The Commercial Medicine Unit (CMU) collects some information on the discounts that are available for generic (that is off-patent) drugs bought in the NHS via its Electronic Marketing Information Tool (eMIT).

There may be some situations where it is appropriate to use prices with NHS discounts rather than list prices. However, there is often limited discussion or detail as to why a price with an NHS discount rather than a list price has been used in a technology appraisal. Clear justification for this choice is rarely presented to the Appraisal Committee. Additionally, if a price with an NHS discount is used, full details of how the discounts were identified and accompanying descriptions (such as where the discounts are available and for how long) are rarely presented clearly to the Appraisal Committee.

2.3 What the current Methods Guide advises with respect to costs

The current methods guide contains a reasonable amount of detail and flexibility surrounding costing:

> *5.6.1.1 For the reference case, costs should relate to resources that are under the control of the NHS and PSS where differential effects on costs between the technologies under comparison are possible. These resources should be valued using the prices relevant to the NHS and PSS. Where the actual price paid for a resource may differ from the public list price (for*

example, pharmaceuticals, medical devices), the public list price should be used. Sensitivity analysis should assess the implications of variations from this price. Evidence should be presented to demonstrate that resource use and cost data have been identified systematically.

5.6.1.2 Given the perspective in the reference case, it is appropriate for the financial costs relevant to the NHS/PSS to be used as the basis of costing, even though these may not always reflect the full social opportunity cost of a given resource. As far as possible, estimates of unit costs and prices for particular resources should be used consistently across appraisals. A first point of reference in identifying such costs and prices should be any current official listing published by the Department of Health and/or the Welsh Assembly Government.

5.6.1.3 The methods of identification of resource use and unit cost data are not as well defined as for evidence for the identification of clinical effectiveness. Where cost data are taken from literature, the methods used to identify the sources should be defined. Where several alternative sources are available, a justification for the costs chosen should be provided. Where appropriate, sensitivity analysis should be used to assess the implications for results of using alternative data sources.

5.6.1.4 Value added tax (VAT) should be excluded from all economic evaluations but included in budget impact calculations at the appropriate rate (currently 17.5%) when the resources in question are liable for this tax.

5.6.2 Although not part of the reference case, there will be occasions where non-NHS/PSS costs will be differentially affected by the technologies under comparison. In these situations, the Institute should be made aware of the implications of taking a broader perspective on costs for the decision about cost effectiveness. When non-reference case analyses include these broader costs, explicit methods of valuation are required. In all cases, these costs should be reported separately from NHS/PSS costs.

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 by the Methods Guide Review Working Party.

- Can clearer guidance on the appropriate use of HRG codes be provided?
    - Are there any situations where HRG codes are always inappropriate?
    - Should the justification for choice of HRG codes be made more explicit?
    - Should justification for departing from HRG codes be made more explicit?

What could be the impact of providing explicit wording in the methods guide on the use of HRG codes?

- Can further guidance on the use of prices that are available at a discount, rather than list prices be provided?
    - Are there any situations where list prices are inappropriate?

What could be the consequences of specifying situations where list prices are appropriate or inappropriate in the methods guide?

- Can further guidance be given on justifying and identifying prices that are available at a discount? Should the onus be on the evidence submitter to provide certainty on the discounts that are available?

What could be the impact of providing clear direction of how discounts should be identified and then subsequently presented?

4 Authors

This briefing paper has been prepared by Rebecca Trowman, Andrew Stevens and Jenni Field.

NATIONAL INSTITUTE FOR HEALTH AND CLINICAL EXCELLENCE

Briefing paper for methods review working party on discounting

The briefing paper is intended to provide a brief summary of the issues that are proposed for discussion by the Methods Review Working Party 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.

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 revised draft of the Methods Guide will be available for a 3-month public consultation, expected to begin in June 2011. We encourage all interested parties to take part in this consultation.

2 Background

2.1 Introduction to discounting

The concept of discounting in health economics has been the source of much debate over the last two decades. This paper attempts to summarise the main principles of discounting in health technology assessment, but it is not an exhaustive review of the literature on the topic.

Discounting is an economic method which is used to assess benefits and costs that may occur in different time periods. In order to allow comparison, costs and benefits are converted to present values by applying a discount rate to the entire duration of both benefits and costs. Discounting reflects the view that people generally prefer to receive benefits or goods now, but pay for them later (time preference). Discounting also attaches declining weights to benefits and cost over time to reflect the opportunity cost (that is, the cost of paying up-front for treatment and the value of other treatment that is displaced as a result). The discount rate is generally based on values of social opportunity cost and/or social time preference (Fox-Rushby, 2005).

The mathematical implementation of discounting is relatively simple: for every year where costs are incurred and benefits received, the future value of costs and benefits are multiplied by a discount factor (DF) as follows:

$$DF(T) = \frac{1}{(1+r)^T},$$

where r is the discount rate and T is the future year to which the present value refers. The discounted present value is then obtained by adding up the reduced future values over the entire time horizon. The above equation is based on the assumption that costs and outcomes are valued periodically (e.g. every year) throughout the time horizon. A different equation is used for implementation of discounting in the case of continuous evaluations.

Figure 1 shows an example of the effect of applying a fixed rate discount to health benefits over a long time horizon. Owing to the compound effect of

discounting, the choice of a particular rate can have a substantial effect on the outcome of the economic modelling.

The impact of discounting varies depending on when costs are incurred during the time horizon and also on when health benefits are gained. The nature of the health-care intervention therefore has a bearing on the effect of discounting. For instance, discounting has the potential to have a substantial differential impact on costs and benefits in cases where costs are incurred upfront and benefits occur in the far future. This is particularly evident in public health programmes such as screening and paediatric vaccination (Severens 2004).

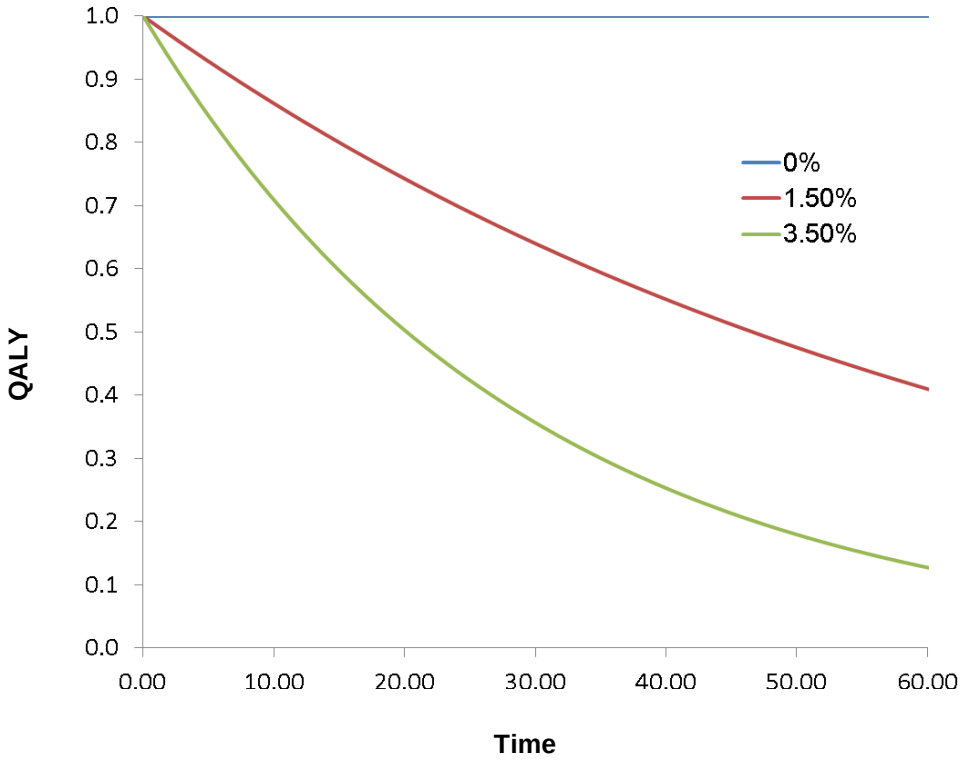


Figure 1: The impact of compound discounting. The graph shows the effect when the discount rate is varied over a time horizon of 60 years, starting at 1 QALY and the reduction in values in subsequent years. Note that discount rates are fixed throughout the period and no change in health states is assumed throughout the time horizon.

2.1.1 Choice of time preference rate

NICE currently bases its discount rate for costs and benefits on the recommended rate set by the HM Treasury for public sector investment appraisal. This rate is a social time preference rate, rather than an individual

preference rate. The social time preference is defined as the value society attaches to present, as opposed to future, consumption. The social time preference rate comprises two components (Green Book, 2003):

- the rate at which future consumption is discounted over present consumption, on the assumption that no change in per capita consumption is expected. This rate is made up of two elements: catastrophic risk (L) and pure time preference (δ),

and

- an additional component, if per capita consumption is expected to grow over time, reflecting the fact that these circumstances imply future consumption will be plentiful relative to the current position and thus have lower marginal utility. This effect is represented by the product of the annual growth in per capita consumption (g) and the elasticity of marginal utility of consumption (μ) with respect to utility.

Mathematically the social time preference (r) is represented by the following equation:

$$r = (L + \delta) + (\mu * g)$$

The current HM Treasury social time preference rate is made up of the following values: r = (0.01 + 0.005) + (1.0 * 0.02) = 3.5 per cent.

The rate of personal time preference has been studied in a UK-wide study, TEMPUS (Cairns and van der Pol, 2000). In this study the personal time preference was found to be higher than the social time preference value used by the Treasury (median discount rate ranging from 3.8% to 6.4%). The relationship between individual time preference and the social rate of discount has been debated over the years; however the TEMPUS study was not designed to address the normative question of the appropriate discount rate to use in economic evaluations. Nevertheless, Cairns and van der Pol argue that the personal preferences could be seen as an input into discussion about the appropriate rate of social discount.

2.1.2 Differential, uniform and time-varying discounting

In spite of being mathematically simple, health economists have had a long-standing debate on how discounting should be applied in the case of non-monetary units such as QALYs which measure health benefits (see for instance Cairns, 1992; Lipscomb 1996, Brouwer 2005; Drummond 2007; Gravelle 2007; Claxton 2010). The principal focus of the debate is whether health benefits should be discounted at the same rate as costs (uniform discounting) or at a lower rate (differential discounting). Furthermore, studies have shown that individuals' time preference can change during the time-horizon, which gives rise to an argument for the use of geometric discounting in long-term models, whereby discount rates are reduced as a function of time (time-varying discounting) (Severens and Milne, 2004).

Uniform discounting is based on the premise that time has an identical effect on both costs and benefits, that is, the nature of the future event is not relevant. A key argument for applying uniform discounting is that if health benefits were to be discounted at a lower value than costs, it would lead to a situation whereby successively delaying an intervention would appear to increase the cost effectiveness (lower the ICER) (Keeler and Cretin, 1983). On the other hand, arguments against uniform discounting include the assumption that the relationship between perceived value of life years and costs remain independent of time, which may not be the case (Gravelle, 2006). Furthermore, it has even been suggested that health benefits should not be discounted at all, because quality of life may already be incorporated into an individual's time preference, especially when utility is measured using the time trade off or standard gamble method (Krahn, 1993).

Generally it is argued that for consistency, uniform discounting should be applied, however in a recent publication (Nord, 2011) argues that much of the debate has focused on logical and arithmetic arguments, with little regard to societal values and empirical research, which may justify differential rates of discounts for costs and health benefits.

In a recent paper (Claxton, 2011) the authors argue that rates should be equal for health benefits and costs in situations where the cost-effectiveness

threshold is expected to remain constant. The authors also support the idea that the discount rate applied to health benefits should probably be lower than the current 3.5% recommended by the NICE methods guide.

2.2 Relevance of topic to NICE technology appraisals

Because many economic analyses considered in the technology appraisals programme have a time horizon reflecting whole of the remaining life expectancy of the cohort under consideration (in some cases several decades) discounting is required to reflect the present value of future costs and benefits.

NICE's recommendations relating to discounting have varied historically. Before the publication of the first methods guide, NICE recommended discounting of costs at 6% and health benefits at 1.5%. This reflected Department of Health policy at the time. In 2003, the Treasury updated its guidance for appraisal and evaluation in central government in a publication named the 'Green Book'. The updated guidance introduced a new rate of 3.5% which was based on social time preference. In the 2004 version of the methods guide, NICE reduced the discounting of costs to 3.5%, in line with the 'Green Book', and at the same time stipulated that costs and benefits should be discounted at an equal rate. Therefore the discount rate for benefits also changed to 3.5%.

Within the UK, the Joint Committee on Vaccination and Immunisation (JCVI) follows a decision making process similar to that of NICE. Owing to the typically long time lag between vaccination and the benefit accrued, the discount rate used for economic evaluation of vaccinations is particularly important. The JCVI analyses use a 3.5% discount rate for costs and benefits based on the Green Book, but generally present sensitivity analyses using 1.5% and 0% discount rates to inform decision making.

Discounting practices in other countries vary. As a general rule, guidelines recommend *examining* the impact of discounting in a sensitivity analysis, and several guidelines also recommend *reporting* undiscounted costs and effects.

However, it is often less clear how these sensitivity analyses are subsequently used in the decision making process and firm decision rules are often lacking.

2.3 What the methods guide currently says

The 2008 edition of the methods guide includes the following text.

> *"5.6.1 Cost-effectiveness results should reflect the present value of the stream of costs and benefits accruing over the time horizon of the analysis. For the reference case, an annual discount rate of 3.5% should be used for both costs and benefits. When results are potentially sensitive to the discount rate used, consideration should be given to sensitivity analyses that use differential rates for costs and outcomes and/or that vary the rate between 0% and 6%.*
>
> *5.6.2 The need to discount to a present value is widely accepted in economic evaluation, although the specific rate is variable across jurisdictions and over time. The Institute considers it appropriate to discount costs and health effects at the same rate. The annual rate of 3.5%, based on the recommendations of the UK Treasury for the discounting of costs, should be applied to both costs and health effects."*

Following the publication of the methods guide, the NICE board discussed how discounting should be implemented in the special case of treatments that are expected to offer curative benefits experienced over a very long time horizon. The NICE Board, having given consideration to the circumstances where it expects advisory bodies to use the sensitivity analysis on the impact of discounting of health effects, issued the following clarification in section 5.6 of the Guide to the Methods of Technology Appraisals (additions shown in bold):

> *"5.6.2 The need to discount to a present value is widely accepted in economic evaluation, although the specific rate is variable across jurisdictions and over time. The Institute considers it appropriate to* ***normally*** *discount costs and health effects at the same rate. The*

annual rate of 3.5%, based on recommendations of the UK Treasury for the discounting of costs, should be applied to both costs and health effects. ***Where the Appraisal Committee has considered it appropriate to undertake sensitivity analysis on the effects of discounting because treatment effects are both substantial in restoring health and sustained over a very long period (normally at least 30 years), the Committee should apply a rate of 1.5% for health effects and 3.5% for costs.****"*

It is important to note that the change to the text reflects a clarification of how the Committee should deal with sensitivity analyses in these particular circumstances. It does not constitute a change to the reference case.

3 Proposed issues for discussion

What is the appropriate discount rate to be applied in the reference case and should costs and health benefits be discounted at the same rate?

In the case of a very long time horizon, should discounting rates for costs and/or health benefits deviate from the standard rates, for instance through application of variable discount rates or reduced discount rates?

Should discount rates for health benefits be lower in specific circumstances, for instance when calculating health benefits for interventions that provide a cure to an otherwise terminal illness?

How should the discount rate be explored in sensitivity analyses?

How should the Committee deal with ICERs that are highly sensitive to the discount rate?

4 References

Brouwer WBF, Niessen LW, Postma MJ *et al.* (2005) Need for differential discounting of costs and health effects in cost effectiveness analyses. *British Medical Journal,* 331:446-448.

Cairns J. (1992) Discounting and health benefits: another perspective. *Health Economics*, 1:76-79.

Cairns JA, van der Pol MM. (2000) The estimation of marginal time preference in a UK-wide sample (TEMPUS) project. *Health Technology Assessment*, 4(1).

Claxton K, Paulden M, Gravelle H, *et al*. (2011) Discounting and decision making in the economic evaluation of health care technologies. *Health Economics*, 20:2-15.

Drummond M, Sculpher M, Torrance S, *et al.* (2007*) Methods for the Economic Evaluation of Health Care Programmes*. Oxford Medical Publications, Oxford.

Fox-Rushby J, Cairns J. (2005) *Economic Evaluation*. Open University Press, London.

Gravelle H, Brouwer W, Niessen L, *et al.* (2007) Discounting in economic evaluations: stepping forward towards optimal decision rules. *Health Economics,* 16:307-317.

HM Treasury. (2003) *Green Book.* (http://www.hm-treasury.gov.uk/data_greenbook_index.htm)

Keeler EB, Cretin S. (1983) Discounting of life-saving and other nonmonetary effects. *Management Science*, 29:300-306.

Krahn M, Gafni A. (1993) Discounting in the economic evaluation of health care intervention. *Medical Care*, 31:403-418.

Lipscomb J, Torrance G, Weinstein M. (1996) In *Time preference. Cost-effectiveness in Health and Medicine*. Oxford University Press, Oxford.

Nord E. (2011) Discounting future health benefits: the poverty of consistency arguments. *Health Economics,* 20:16-26.

Severens JL, Milne RJ. (2004) Discounting health outcomes in economic evaluation: the ongoing debate. *Value in Health,* 7(4):397-401.

5 Author/s

This document was prepared by Pall Jonsson and Janet Robertson. Thanks to Paul Tappenden, John Brazier and Jenny Dunn for their helpful comments.

NATIONAL INSTITUTE FOR HEALTH AND CLINICAL EXCELLENCE

Briefing paper for methods review working party on treatment sequences and downstream costs

The briefing paper is intended to provide a brief summary of the issues that are proposed for discussion by the Methods Review Working Party 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.

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 revised versions were published in 2004 and 2008. 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 revised draft of the Methods Guide will be available for a 3-month public consultation, expected to begin in June 2012. We encourage all interested parties to take part in this consultation.

2 Background

2.1 Relevance of topic to NICE technology appraisals

In some technology appraisals, a new intervention may be positioned at several points in an existing sequence of treatments. As such, the comparison in the economic evaluation can be between alternative sequences of treatments, rather than a head to head comparison between the intervention and a specific comparator treatment. Rather than **X** (new treatment) vs A vs B (comparators), the evaluation may be (**X**,A,B)[*] vs (A,**X**,B) vs (A,B,**X**), which is equivalent to evaluating **X** at 1st, 2nd or 3rd line. Due to the impact that a treatment may have on the long run costs and benefits accrued, and potentially whether a patient progresses to subsequent treatments in a sequence, a lifetime perspective is required, and therefore the economic analyses have attempted to model the possible alternative sequences.

This is particularly common in technology appraisals of chronic conditions (for example rheumatoid arthritis). In such appraisals, it has been necessary for sequences of 5 of more treatments to be modelled and compared with each other. In fact, this only represents a small proportion of the overall potential sequences of rheumatoid arthritis therapies. The selection of the alternative sequences, and the assumptions and evidence used to model the sequences can have a substantial impact on the incremental cost effectiveness of competing decision alternatives.

The current methods guide highlights that the "...*main technology of interest, its expected place in the pathway of care, the comparator(s) and the relevant patient group(s) will be defined in the scope developed by the Institute"* (5.2.6). Also *"…many technologies have impacts on costs and outcomes over a patient's lifetime. This is particularly the case with treatments for chronic diseases. In such instances, a lifetime time horizon for clinical and cost effectiveness is appropriate"* (5.2.14). It is specific when stating; *"Sometimes both technology and comparator form part of a treatment sequence, in which*

[*] For clarity, a sequence of treatments is presented within parentheses. The order in the parentheses represents their order in the sequence. A treatment in bold represents an addition into the sequence

case the appraisal may need to compare alternative treatment sequences" (2.2.4).

Therefore, the Methods Guide suggests that the modelling of sequences of therapies should be considered if the scope defines alternative possible positions of the new technology. However the Methods Guide does not at present provide specific guidance with respect to modelling sequences of treatments.

A second, related problem arises when standard NHS care given either alongside or after the treatment of interest has been given, is very costly and/or not very effective. In this situation, the effect of a significant proportion of the modelled cohort surviving long enough to receive these downstream treatments may make a new, effective intervention appear cost ineffective, purely because it increases the opportunity to receive subsequent cost ineffective treatments. This situation is challenging for the Appraisal Committee, with new interventions that do extend survival appearing cost ineffective purely because of the downstream treatments. This situation can be deconstructed into two components. Firstly, whether direct costs (related to the primary condition) and/or indirect costs (unrelated to the primary condition) should be included in the economic evaluation. In particular, whether these costs should be included if they occur in additional life-years gained as a result of the intervention. Secondly, how can the impact of the inclusion or exclusion of these costs be made transparent, to aid the Committee when these challenging situations occur?

The Methods Guide's principal comment on this issue is in its recommendation for a life-long time horizon (5.2.14). It does suggest alternative scenarios for extrapolation beyond trial data, but it does not provide specific guidance on how to compare alternative scenarios with respect to downstream treatment possibilities.

2.2 Introduction to modelling treatment sequences

There are a number of issues that surround modelling sequences of treatments within the context of a NICE Technology Appraisal.

Firstly, the order in which the treatments are given within the sequences may have an impact on the effectiveness and the costs of the technologies. If the sequence (A,B) is identical to A and B in isolation, then what is inferred is that the costs and effects are not influenced by the position in the sequence, and therefore the sequence should begin with the most cost effective treatment.

However, it is often the case that the sequence as a whole must be considered, because it is not the case that the cheapest or most cost effective in isolation necessarily comes first in the most cost effective sequence. The position in a sequence may have an influence on factors that affect cost effectiveness (e.g. shorter duration on treatment, a lower chance of response).

A second important issue that can occur is that limited possible treatment sequences are modelled and the new intervention is added to the original treatment sequence. For example, an existing treatment sequence of 3 technologies (A,B,C) exists for a condition, and a new intervention technology **X** is modelled at the beginning of the treatment sequence compared with the original treatment sequence (**X**,A,B,C). In this example, the addition of the new technology to the start of the treatment sequence raises questions. Firstly, will a treatment 'drop out' of the sequence if **X** was recommended at first line? Secondly, does (A,B,C) represent the full treatment sequence that is routinely delivered in the NHS? Is (A,B,C) more complicated, because in fact NICE guidance allows for conditional sequences (e.g. first line options (A or B), second line options (C, or if B at first line then A)).

If a complex conditional (set of) sequence(s) has emerged as standard practice in the NHS, then the question arises as to whether the appraisal of the new technology should look to identify the 'optimal' sequence of treatments and use this as the comparator. If previous NICE guidance has recommended treatment options that would no longer be cost effective in comparison with the new technology (perhaps dominated by it), then a review of all treatments using the multiple technology appraisal process would be required to update the previous guidance. An MTA review of all treatments may require a factorial set of sequences to be modelled and evaluated, which

despite representing a computationally and empirically challenging task, may allow the optimal sequence of treatments to be identified.

Hypothetical example

Existing NICE Guidance recommends:

First line: A or B or C
Second line: D or E or F

All (9) possible sequences have slightly different estimates of costs and QALYs. The sequence (A,D) has been identified as optimal, and it is more effective and less costly than (C,F).

The question is, when evaluating **X** as a new treatment, should it be recommended if it improves the optimal sequence (A,D), and offers a positive net gain to the NHS, or should it be recommended because it can improve the sequence (C,F), but the new sequence is not optimal compared to (A,D).

If the latter, this would mean that the NHS has not gained by recommending **X**, and in reality it is unlikely that **X** will see uptake in the NHS or capture any market share. **X** may represent a 'me too' product, or may have other attributes of value for specific groups of patients.

Finally, it is unlikely to be sufficient to model the new intervention at only one point of the treatment sequence if the marketing authorisation permits its use elsewhere in the sequence (for example [A,**X**,B,C] or [A,B,**X**,C] may be potential options to be modelled). In fact, a manufacturer may only present one position of their treatment (perhaps the position that would capture the greatest market share), whereas the treatments' optimal position (from an NHS perspective) may be at a later point in the sequence, which represents a less desirable position for the manufacturer.

Related to the example above, if there are a number of technologies included in the sequences, it can become increasingly challenging to know what the true treatment effects are likely to be for every technology in every position in

the sequence. For example, if treatments have been studied in clinical trials as first-line treatments, but then are placed second or third-line in a sequence, the efficacy of these treatments in the sequence may be very different from that observed in the trial. The corollary is that, as seen frequently with modelling treatment sequences, there is a danger that what is modelled has moved dramatically from the available trial evidence. It could be argued that validation of trial evidence, by the use of expert opinion or other external data, may help ensure that modelled treatment effects appropriate reflect reality. In particular, treatment effect decrements have been used in NICE appraisals, which suggest that a treatment's effectiveness 'down the line' is diminished. These decrements could potentially be informed by external data, such as registries, or expert opinion; however they would be open to potential bias. Observational studies could potentially be used to estimate treatment effects, although the limitations of this approach have been widely discussed.

2.3 Related and unrelated downstream costs

Another problem occurs when the costs and effects of cost ineffective downstream treatments are included within the calculations of cost effectiveness of a new technology. This situation is most common in appraisals of technologies that are life-extending, for example, technologies that prolong life in terminal diseases such as cancer. For example, consider a technology that extends life by approximately 3 years, the treatment (and therefore treatment costs) are incurred for a short proportion of this time, say 3 months. The rest of the increased survival is associated with additional treatment costs that are a result of living with the condition (such as monitoring, palliation and so on). In this case, downstream but related (to the original condition) costs have been included. The effect of including these downstream related treatment costs can result in very high cost effectiveness ratios for the new technology compared with standard NHS care, simply because the new technology increases survival such that more time is spent in the expensive treatment state.

The Methods Guide makes no explicit comment regarding which (related or unrelated) future health care costs should be included in the economic

analysis. Costs could be related or unrelated to the condition for which the treatment was provided, and could be specific to time that would have been lived anyway, or specific to time that has been gained as a result of the treatment being appraised. This issue has been raised in previous literature (Meltzer, 1997), and as part of a briefing paper for the last update of the Methods Guide (Miners, 2007). As Miners states; *"is it possible to establish whether a tumour that develops 10 years after radiotherapy, but in a different location to the original tumour, is related or unrelated to the index tumour or radiotherapy?"*

Future related health costs are likely to be a necessity, in that the initiation of a treatment reflects a decision about a course of action for the patients' condition, and therefore an evaluation of its cost effectiveness should include health costs attributed by that treatment on the condition. However it may lead to age-discrimination, and would be contrary to current NICE methods that prioritise treatments that offer life extension at the end of life.

Gold (1997) provide a useful taxonomy of induced costs in cost effectiveness analyses (see Table 1).

The identification of future health care treatments which are cost ineffective may provide a disinvestment opportunity for the NHS, and offer a clear representation of 'the margin', from which cost effectiveness analyses have emerged. However it may be that these apparently cost ineffective treatments have attributes for which society may potentially be willing to pay for (end of life therapy, rule of rescue).

Table 1: Gold et al. Future costs (table derived from p.47)

Category	Sub-category	Details	Considerations for NICE
Costs **related** to the intervention, incurred during years of life that would have been lived without the intervention	-	These include related diseases in the original lifespan, and adverse events.	These costs are routinely included in NICE appraisals where a life-long time horizon is required.
Costs **unrelated** to the intervention, incurred during years of life that would have been lived without the intervention		By definition these are costs that are the same irrespective of the intervention, and so will be cancelled out in the analysis.	Because these costs would cancel out in an analysis, is it not necessary for NICE to require these costs to be included.
Costs that incur in years of life **added** (or subtracted) by an intervention	Health care costs related to the primary disease	Health care costs which occur after the initial treatment, and extend into the years of life added (or subtracted) by the intervention.	Downstream treatments and activities may not be cost effective, and may be provided due to other attributes. These may 'wash out' the cost effectiveness of the initial treatment.
	Health care costs for other diseases	Costs for diseases unrelated to the intervention and occurring in added years of life.	If interventions are compared across different age groups and these costs are included, the ranking of cost effectiveness will alter from the same set but with these costs excluded.
	Non health care costs	Relates to the perspective of the overall analysis.	Should be considered alongside any alteration in the perspective of NICE's decision-making.

2.4 What the current Methods Guide advises with respect to treatment sequences and downstream costs

The Methods Guide provides the following statements regarding sequences and future health care costs:

The time horizon for estimating clinical and cost effectiveness should be sufficiently long to reflect all important differences in costs or outcomes between technologies being compared (section 5.2.13)

Many technologies have impacts on costs and outcomes over a patient's lifetime. This is particularly the case with treatments for chronic diseases. In such instances, a lifetime time horizon for clinical and cost effectiveness is appropriate (5.2.14).

Sometimes both technology and comparator form part of a treatment sequence, in which case the appraisal may need to compare alternative treatment sequences (2.2.4).

There is limited discussion of modelling treatment sequences in the 2008 Methods Guide. It is acknowledged as a possibility in Section 2.2.4, but no further guidance on when and how treatment sequences should be modelled is provided.

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 questions are discussed by the Methods Guide Review Working Party.

- Under what circumstances is it acceptable to model only individual lines of therapy, rather than treatment sequences? Should downstream treatments be assumed to incur the same cost between groups?

- Should future health care costs **related** to the primary disease which are incurred due to life extension be included in the economic analysis?

- Should future health care costs **unrelated** to the primary disease which are incurred due to life extension be included in the economic analysis?

What could be the impact of providing further direction on when the modelling of treatment sequences is appropriate?

- How can the methods guide ensure that modelling of treatment sequences is undertaken consistently across appraisals?

- Should explicit guidance on aspects of modelling treatment sequences be given?

    - When and how should sequences be identified? Which should be modelled?

    - What effectiveness estimates and model parameters can be reasonably used when a treatment is included in different places in different sequences?

    - What level of primary and sensitivity analyses should be reasonably expected?

What could be the possible consequences of including further guidance in the methods guide on exactly how downstream costs should be modelled?

- What can be done in the situation where a cost ineffective treatment is given either in combination with an intervention, or given after an intervention, such that it results in the intervention itself appearing cost-ineffective?

    - Should downstream treatments that are cost ineffective be included for the primary analysis, or just limited to a secondary analysis?

- Should a head-to-head comparison of the technologies of interest (i.e. with no downstream treatments included) be requested?

What are the potential consequences of recognising this issue in the methods guide and providing guidance on how it could be approached?

4 References

Gold, MR. Siegel, JE. Russell LB. Weinstein MC. Cost-effectiveness in health and medicine. 1996

Meltzer, D. Accounting for future costs in medical cost-effectiveness analysis. Journal of Health Economics. 1997. 16(1) p 33-64

Miners, A. Costs: Briefing paper for the update of the NICE methods guide. 2007

5 Author/s

Jon Tosh, DSU
Rebecca Trowman, NICE

November, 2011

Thanks to Paul Tappenden, Janet Robertson and Andrew Stevens for comments on a draft version of this report

NATIONAL INSTITUTE FOR HEALTH AND CLINICAL EXCELLENCE

Briefing paper for methods review workshop on equity

The briefing paper is written by members of the Institute's staff. 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 June 2011. We encourage all interested parties to take part in this consultation.

2 Background

The Technology Appraisals Methods Guide contains the following relating to equity.

> ***"****1.4.3 The Institute is committed to promoting equality, eliminating unlawful discrimination and actively considering the implications of its guidance for human rights. The Institute will take into account relevant provisions of legislation on human rights, discrimination and equality. 'NICE's equality scheme and action plan 2007–2010' describes how the Institute meets these commitments and obligations."*
>
> *"2.31 During the consultation on draft scopes in Technology Appraisals interested parties are asked for their views on an appropriate remit for the appraisal and important issues to be considered. This consultation process is important to define the relevant issues to be considered and, in*

particular, to: […] identify any equality or diversity issues that need to be taken into consideration. "

"3.4.4 The Institute considers equity in terms of how the effects of a health technology may deliver differential benefits across the population. Evidence relevant to equity considerations may also take a variety of forms and come from different sources. These may include general-population-generated utility weightings applied in health economic analyses, societal values elicited through social survey and other methods, research into technology uptake in population groups, evidence on differential treatment effects in population groups, and epidemiological evidence on risks or incidence of the condition in population groups."

"3.4.5 The Institute is committed to promoting equality and eliminating unlawful discrimination, including paying particular attention to groups protected by equalities legislation. The scoping process is designed to identify groups who are relevant to the appraisal and reflect the diversity of the population. The Institute consults on whether there are any issues relevant to equalities within the scope of the appraisal, or if there is information that could be included in the evidence presented to the Appraisal Committee to enable them to take account of equalities issues when developing guidance."

"5.10.10 The Appraisal Committee will pay particular attention to its obligations with respect to legislation on human rights, discrimination and equality when considering subgroups."

"6.1.3 When formulating its recommendations to the Institute, the Appraisal Committee has discretion to consider those factors it believes are most appropriate to each appraisal. In doing so, the Appraisal Committee has regard to the provisions of NICE's Establishment Orders and legislation on human rights, discrimination and equality. In undertaking appraisals of healthcare technologies, the Institute is expected to take into account Directions from the Secretary of State for Health […] as follows [...]

6.2.6 … [T]he, the Chair ensures that the Committee considers: […] .the relevant legislation on human rights, discrimination and equality […]

6.2.20 The Committee will take into account how its judgements have a bearing on distributive justice or legal requirements in relation to human rights, discrimination and equality. Such characteristics include, but are not confined to: age; sex/gender or sexual orientation; people's income, social class or position in life; race or ethnicity; disability; and conditions that are or may be, in whole or in part, self-inflicted or are associated with social stigma."

The purpose of this paper is to focus on equity and health from a public health perspective in order to identify a number of core considerations for discussion at the workshop.

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 Definition

Consistent terminology in the arena of equity has been found to be helpful by the World Health Organisation (WHO) which has had a long standing interest in the matter. Recently the World Health Organisation's Commission on the Social Determinants of Health (WHO, 2008) used definitions arising from the work of Whitehead (Whitehead, 1992; 2006; Whitehead and Dahlgren 2006) and Solar and Irwin (2007; 2010). The critical definitions are:

- Health equity – the absence of unfair and avoidable or remediable differences in health among social groups (Solar and Irwin, 2010:14).
- Health inequity – unfair and avoidable or remediable differences.

- Health Inequality – health differences which are not avoidable or are not the consequence of human actions and activities and are based on genetic or constitutional individual differences, age or biological sex. These are sometimes also referred to as variations (Kelly et al 2007).

It is important to note that the difference here between inequity and inequality is not used universally and many writers and commentators use the two terms as synonyms. Also the distinction between individual differences which are based on human biology and differences arising from interaction between the organism and some man made hazard externally is in reality a difficult one to draw in anything other than an analytic sense. Empirically the divide is much fuzzier than these definitions suggest. However as a way of finding some clarity the distinction is helpful.

The gist of the argument about equity and inequity is that they are not the products of nature they are the products human actions and are socially, economically or politically produced and therefore theoretically, at least, modifiable (Whitehead and Dahlgren, 2006:2; Kelly and Doohan, 2011).

The questions raised by the definitional work of WHO for the workshop are

(i) Are NICE's definitions clear?

(ii) Do they correspond to those used by WHO?

(iii) What are the bases of the definitions used in the legislation?

(iv) Are any scientific problems generated by the appeals to principles such as unfairness, social justice and human rights?

3.2 Three characteristics of health inequities

It has been argued that there are three characteristics of health inequities: patterning, causation and unfairness (Whitehead and Dahgren, 2006).

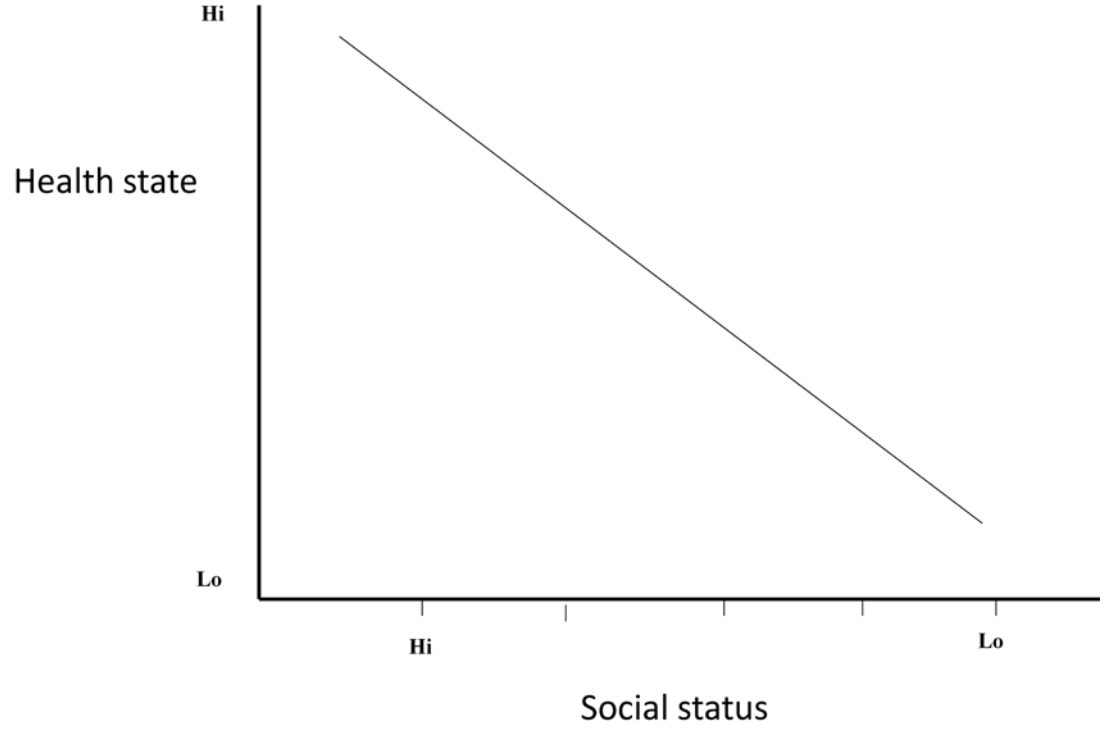
Patterning

The data reveal that health inequities are systematic or patterned. The patterning reflects various dimensions of social difference in populations -

socio economic group, gender, ethnicity, geography, age, disability and sexual orientation. The patterning occurs locally, at regional level, within countries, and between countries. This social patterning is universal in human societies, but its extent and magnitude varies between different societies (Whitehead and Dahlgren, 2006).

The pattern is conventionally referred to as the social gradient in health. The gradient describes a pattern which is formed by comparing measures of mortality and morbidity with some measure of social position. Originally, the social measure was occupation or occupation of head of household. Occupation has tended to be readily available in official statistics and has been a good proxy for a range of other aspects of life chances including education, income, housing tenure and social class (Graham and Kelly 2004).

Figure 1.The schematic health gradient



Source Kelly 2010

The difference in health experiences between the top, middle and bottom of the socioeconomic hierarchy varies considerably between countries. For example in Nordic countries there are relatively small disparities in health

across the population compared to the UK and the USA. In middle income and rapidly developing income countries the health differences may be very great with a mix of relatively good health among the well to do and extremes of low life expectancy and high infant mortality among the very poor. The policy implications will therefore vary considerably depending on the nature of the health gradient in particular societies (Kelly et al 2007).

The causes of the patterns

The second feature is that the differences are produced socially, politically or economically – they are not the products of nature or biology. The causes of these social, economic and political processes are collectively conventionally called the social determinants of health or sometimes the causes of the causes of health inequities (Kelly and Doohan, 2011).

Injustice

The third characteristic of the definition is that the differences are judged to be unfair (Whitehead and Dahlgren, 2006). In other words a further principle is invoked of or appealed to in the form of some notion of social justice or human rights.

The great majority of the data relating to health differences and the health gradient uses occupation, income or education as the measure of social difference. It is important to note that although there is a weight of evidence relating to these dimensions, the legislation under which NICE operates focuses on aspects of social difference for which the evidence base is much less robust (Meads et al 2012).

It is important also to note that in empirical and theoretical terms we know almost nothing about the interactive effects on health outcomes of the relationship between socio economic grouping, gender, ethnicity, sexual orientation, and that the research on these intersections or interactions is inconclusive (Meads et al,. 2012; Kelly 2010).

The questions this raises for the workshop include:

(i) To what degree are the patterns described in the literature on health inequities mirrored in clinical data sets?

(ii) To what degree are the questions about the causes of the pattern relevant in appraisals of new technologies?

(iii) How easy is it to operationalise questions of injustice and fairness?

3.3 Policy implications

There are conventionally three different ways in which the inequities are described in relation to policy: health disadvantage, health gaps and health gradients (Graham and Kelly, 2004). Health disadvantage simply focuses on differences, acknowledging that there are differences between distinct segments of the population, or between societies. The health gaps approach focuses on the differences between the worst off and everybody else, often assuming that those who are not the worst off enjoy uniformly good health. The health gradient approach relates to the health differences across the whole spectrum of the population, acknowledging a systematically patterned gradient in health inequities.

Conceptually, narrowing health gaps means raising the health of the poorest, fastest. It requires both improving the health of the poorest and doing so at a rate which outstrips that of the wider population. It focuses attention on the fact that overall gains in health have been at the cost of persisting and widening inequalities between socioeconomic groups and areas. It facilitates target setting. It provides clear criteria for monitoring and evaluation. An effective policy is one which achieves both an absolute and a relative improvement in the health of the poorest groups (or in their social conditions and in the prevalence of risk factors).

However, focusing on health gaps can limit the policy vision because it shifts attention away from a whole population focus. Some may object that if we single out some groups as ‘more deserving’ because they were wronged, then we are abandoning the principle that in medical contexts we ought to focus on need.

This is why the health gradient is also important. The penalties of inequities in health affect the whole social hierarchy and usually increase from the top to the bottom. Thus, if policies only address those at the bottom of the social hierarchy, inequities in health will still exist and it will also mean that the social determinants still exert their malign influence. The approach to be adopted should involve a consideration of the whole gradient in health inequities rather than only focusing on the health of the most disadvantaged. The significant caveat is that where the health gap is both large and the population numbers in the extreme circumstances are high, a process of prioritizing action by beginning with the most disadvantaged would be the immediate concern.

This approach is in line with international health policy. The founding principle of the WHO was that the enjoyment of the highest attainable standard of health is a fundamental human right, and should be within reach of all 'without distinction for race, religion, political belief, economic or social condition' (WHO, 1948). As this implies, the standards of health enjoyed by the best-off should be attainable by all. The principle is that the effects of policies to tackle health inequities must therefore extend beyond those in the poorest circumstances and the poorest health. Assuming that health and living standards for those at the top of the socioeconomic hierarchy continue to improve, an effective policy is one that meets two criteria. It is associated with (a) improvements in health (or a positive change in its underlying determinants) for all socioeconomic groups up to the highest, and (b) a rate of improvement which increases at each step down the socioeconomic ladder. In other words, a differential rate of improvement is required: greatest for the poorest groups, with the rate of gain progressively decreasing for higher socioeconomic groups. It locates the causes of health inequity, not in the disadvantaged circumstances and health-damaging behaviours of the poorest groups, but in the systematic differences in life chances, living standards and lifestyles associated with people's unequal positions in the socioeconomic hierarchy (Graham and Kelly, 2004).

Figure 2. The health gradient showing uniform improvement

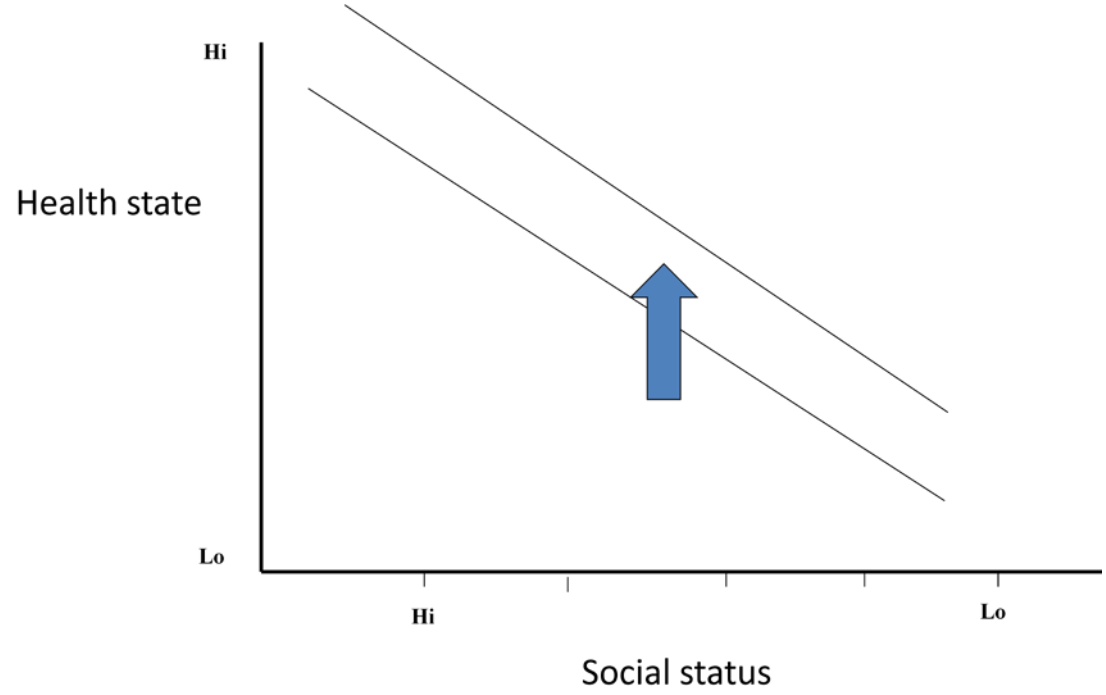


Figure 3 .The health gradient showing relative health inequalities getting worse

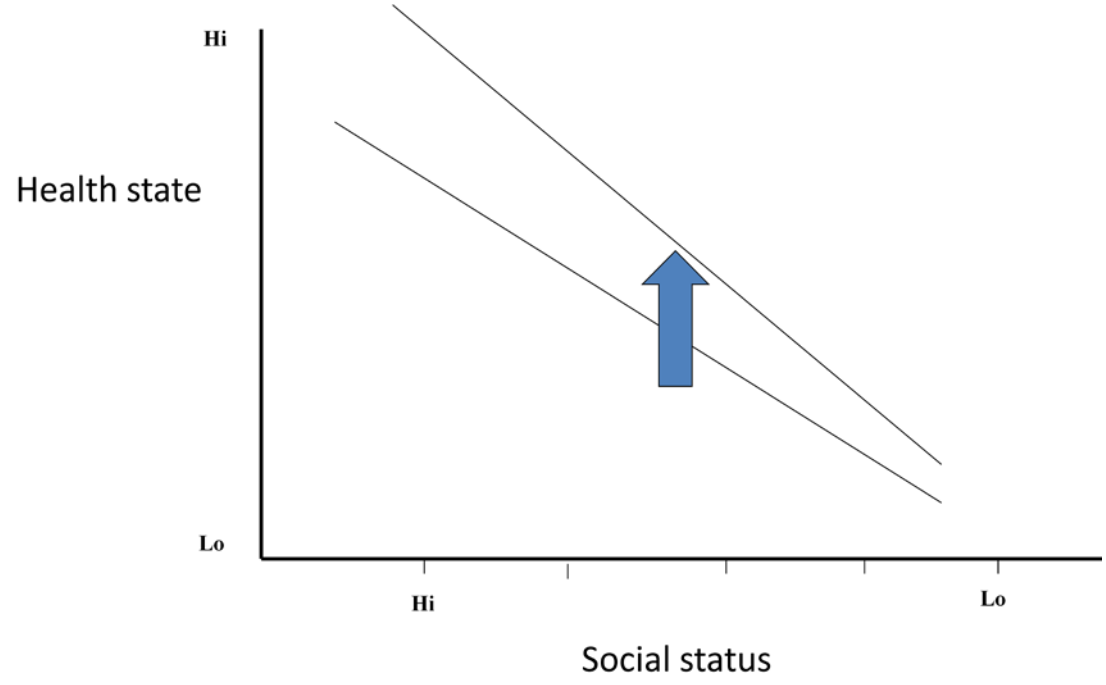
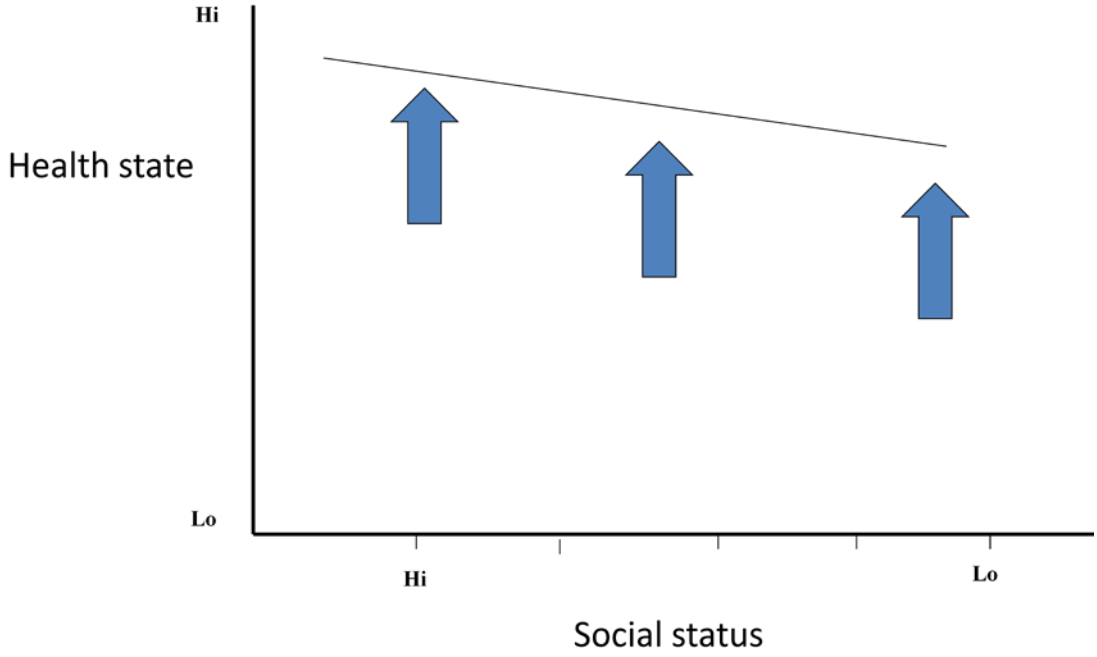


Figure 4. Shifting the health gradient through universal and targeted action



A number of questions suggest themselves here.

(i) Do we have sufficient knowledge of differential effectiveness across social groupings to be able to manipulate interventions in such a way that they would have an impact on the gradient?

(ii) What is the underlying purpose of the legislation in terms of gaps, gradients and equity, and to what degree should our attempt to work with the legislation have coherence in terms of policy goals?

Finally we need to note that in the public health while an enormous amount is known about the descriptions of inequalities especially in respect of class and income, the literature says almost nothing in practical terms about what ought to be done by policy makers or practitioners to remedy the situation. There are high level solutions which describe income equalization and greater public expenditure for example, but the evidence to support such approaches is at best equivocal, and in any event this is a domain where NICE has no responsibilities. So in effect what we have here is a very old problem philosophically speaking between our ability to describe the world as it is – empirical fact, and the vision of the world as we might like it to be – value. And

in the next section the problems of values as they impinge on these matters is discussed.

3.4 Some important philosophical and value underpinnings

The explicit value which is evident in much of the writing on equity, as we have already noted, is that health differences that exist at population level within and between societies are unfair and unjust. This is not a scientifically derived principle; it is a value position which asserts the rights to good health of the population at large. It stands in contrast particularly to the value position that argues that differences in health are a consequence (albeit an unfortunate consequence) of the beneficial effects of the maximization of individual utility in a relatively unfettered market. It is important to note that individual and collective utilities may be at odds with respect to the rights to health (Macintyre, 1984).

There is an important literature which explores the issue. Anderson (1999) for example alerts us to the fact that , WHO's efforts notwithstanding, (i) the concept of equality means a number of different things depending on the underlying political value position and the epistemological assumptions of the theory (e.g. utilitarianism or socialism); (ii) that it is an entirely rationalist concept – it is not empirically grounded; (iii) that for the most part many writers on health inequity do not explore the underpinning value positions and assert instead that things are unjust and unfair because they could be changed; (iv) that most writings on health inequalities take as their starting point the *a priori* rationalist and political position that there is something morally wrong about health inequalities; (v) that the empirical data on health inequalities is aggregated individual data and fails to explore the relational elements of inequity. i.e. that inequity reflects power, coercion and force between groups in the social world as they compete for scarce resources (vi) that the compassionate dimension is important - the unnecessary suffering and death that the inequalities involve is surely the most important reason for dealing with the question, along with the associated waste and cost to the exchequer; (viii) that most of the literature fails to address the question of causation adequately.

Pogge (2003) draws our attention to the fact that the notion of what is just or unjust is not a given, but rather has an array of different meanings. This is because justice is a relational concept, i.e. is about relations between people and is therefore a social construct arising in social interaction and judgements about it are made morally or metaphysically. Science cannot provide answers or solutions.

The question which this prompts is:

(i) Does the NICE social value judgements paper deal with these thorny problems?

3.5 Individual differences versus patterning

In much of the literature and certainly in the legislation two analytic causal levels are confused - the individual and the social. (Kelly 2010). This has some potentially important implications for the approach which might be taken in Technology Appraisals.

It is relatively straightforward to understand the causal pathway at the level of the individual. Pathology occurs in the human body, in an individual's cells and systems. The individual feels pain and suffers and the consequences of such morbidity are familiar to everyone. Medicine provides detailed explanations of the origins of such biological events in the individual. And also in many cases provides an ameliorative or curative therapy based on an understanding of the causal pathway. The origins of the pathology may be proximal, such as chance exposure to a virus or bacteria. Sometimes the originating cause is more distal in some aspect of environmental or occupational exposure to hazards like radiation or asbestos. But even in these cases of distal origins, the explanatory pathway is clear and operates at the level of the individual. This by and large is the territory of clinical medicine.

However, there is another equally important pathway that operates at the level of the social or population. And it is the outcomes of these pathways which is the focus of political and value concerns about equity.

There are clear patterns of population health as we noted above. One way of thinking about the patterns is to assume that they represent the aggregation of individual events. So the differences in mortality and morbidity at population level are simply the summation of lots of different individual disease episodes. And of course so it is. But the patterns can also be conceptualized as an analytic reality of their own. The fact is that the patterns themselves repeat themselves and reproduce generation after generation. The pattern has a quality of systemness or structure which exists above and beyond the individual events.

Two ideas illustrate this point. First, in the mid-19th century in Britain the principal causes of death were infectious disease. In the early 21st century the principal causes are diseases associated with smoking, diet, alcohol misuse and lack of exercise. Although the biological mechanism involved in the pathology then and now are quite different, the associated diseases still kill more of the relatively disadvantaged prematurely than those from more privileged backgrounds, just as was the case in the 19th century. In other words, quite different biological processes produce startlingly similar patterns.

Second, at geographical level the data also have quite remarkable permanent patterning. In 1862, William Gairdner, the first medical officer of health in Glasgow, in his treatise on air, water and cholera, drew up tables to show where the highest rates of infant and premature mortality were to be found. His list shows an eerily familiar overlap with contemporary albeit more finely-grained data. There is not an exact match but somewhere like Tower Hamlets in the East End of London was an unhealthy place in 1862 and it is today. The population has changed considerably in that time by national and ethnic origin, but the pattern of health inequality is reproduced (Kelly, 2010).

So an explanation is needed both of the individual disease outcomes *and* the patterns. The two causal pathways overlap, certainly, and the factors involved interact with each other, but there are two different things to be explained. The 19th-century pioneers in public health understood this at least intuitively. One can certainly draw the impression reading Gairdner's work or that of Duncan, the first medical officer of health in Liverpool that they tried to understand

social level causes as they described the social conditions of their cities. The great sanitation schemes of Bazelgette in London and similar efforts in continental Europe attest to an understanding of the possibility of intervening at population level and influencing the social level very effectively. Indeed, to some extent the major advances in the health of the public of the early period of public health were mostly attributable to the impact of these population level inputs.

The key point is this. Action to deal with patterns of health inequities will in the end require actions which operate at population level in various ways from legislation to nudging, from education to screening. Moreover the broader patterns of inequalities in society themselves provide for much of the explanations of differences seen in population patterns of health. By and large medical interventions operate at the individual level and while individual interventions will clearly benefit the individuals concerned it does not follow that this will have an impact at social level (Capewell and Graham, 2010). The underlying problem with the legislation as framed is that the duties it imposes operate on individuals, but do not operate in terms of the broader social structures.

4 References

Anderson, E.S.(1999) What is the point of equality? *Ethics*; 109: 287-337.

Capewell S and Graham H (2010) Will Cardiovascular Disease Prevention Widen Health Inequalities? *PLoS Med* 7(8): e1000320. doi:10.1371/journal.pmed.1000320

Graham H, Kelly MP. *Health inequalities: concepts,frameworks and policy*. London: Health Development Agency; 2004. http://www.nice.org.uk/page.aspx?o=502453

Kelly, M.P. (2010) The axes of social differentiation and the evidence base on health equity. *Journal of the Royal Society of Medicine*, 103: 266-72

Kelly, M.P., Morgan, A., Bonnefoy, et al. (2007) *The social determinants of health: Developing an evidence base for political action,* Final Report to the World Health Organization Commission on the Social Determinants of Health, from Measurement and Evidence Knowledge Network, The hub coordinating the Measurement and Evidence Knowledge Network is run by: Universidad

del Desarrollo, Chile, and National Institute for Health and Clinical Excellence, United Kingdom
http://www.who.int/social_determinants/resources/mekn_report_10oct07.pdf

Kelly, M.P. and Doohan, E. (2012) The Social Determinants of Health, in, Merson, M.H., Black, R.E., Mills, A.J. (eds) *Global Health: Diseases, Programs, Systems and Policies, 3rd edition,* Burlington, MA: Jones and Bartlett. pp 75-113.

Macintyre, A. (1984) *After Virtue: A Study in Moral Theory*, Notre Dame, Indiana: University of Notre Dame Press.

Meads, C., Carmona, C., Kelly, M.P. (2012) Lesbian, gay and bisexual people's health in the UK : A theoretical critique and systematic review. *Diversity in Health Care,* in press

Pogge, T.W. (2003) Relational conceptions of justice: responsibilities for health outcomes, in Anand, S., Fabienne, P., Sen, A. (eds) *Health, Ethics and Equity*, Oxford: Clarendon.

Solar O, Irwin A. (2007). Towards a conceptual framework for analysis and action on the social determinants of health. WHO / Commission on Social Determinants of Health: Geneva.

Solar O, and Irwin A. (2010) A conceptual framework for action on the social determinants of health. Social Determinants of Health Discussion Paper 2 (Policy and Practice). WHO: Geneva.

Whitehead, M. (1992) Perspectives in health inequity, *International Journal of Health Services*; 22: 429-45.

Whitehead, M. (2007) A typology of actions to tackle social inequalities in health, *Journal of Epidemiology and Community Health*; 61: 473-478.

Whitehead, M and Dahlgren, G. (2006) Levelling Up (Part 1) A Discussion Paper on Concepts and Principles for Tackling Inequities in Health, Copenhagen: WHO.

WHO (1948) Constitution of the World Health Organisation, London: WHO.

WHO (2008) Closing the Gap in a Generation: Health Equity Through Action on the Social Determinants of Health, Geneva: WHO.

5 Author/s

Prepared by Professor Mike Kelly PhD FFPH Hon FRCP

Director, Centre for Public Health Excellence

National Institute for Health and Clinical Excellence

January 2012

NATIONAL INSTITUTE FOR HEALTH AND CLINICAL EXCELLENCE

Briefing paper for methods review working party on extrapolation and crossover

The briefing paper is intended to provide a brief summary of the issues that are proposed for discussion by the Methods Review Working Party 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.

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 revised versions were published in 2004 and 2008. 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 revised draft of the Methods Guide will be available for a 3-month public consultation, expected to begin in June 2012. We encourage all interested parties to take part in this consultation.

2 Background

2.1 Relevance of topic to NICE technology appraisals

When conducting appraisals of the clinical and cost effectiveness of technologies, the Appraisal Committee considers all relevant costs and consequences of treatment, often these will occur over a lifetime horizon. However, it is very rare that sufficient data on the effectiveness of a technology will be available at the time of an appraisal. Frequently within the pivotal trials of a technology the length of data follow-up is relatively short. This is often the case for chronic conditions as it is very rare for all patients to be followed up throughout the full course of their treatment and subsequent experience of the condition. It is possible that data may be available from non-randomised controlled trial (non-RCT) sources and these data may serve a role in informing estimates of the longer-term effects of a treatment. However, such data are rarely available and therefore extrapolation of the observed trial data is undertaken to estimate benefits within the unobserved period. A number of methods can be used to extrapolate available trial data; however limited alternatives are usually presented to the Appraisal Committee. Clear justification for the choice of the extrapolation model used, especially in those instances whereby different curve fits produce very different cost effectiveness estimates, is rarely provided for the Appraisal Committee. Additionally, where very short-term data are available for extrapolation it is challenging to ensure that the most appropriate method for extrapolation has been used. In these circumstances external data sources and full, detailed justifications for the method of extrapolation are rarely presented to the Appraisal Committee.

Another circumstance whereby the true treatment effect is essentially unknown is when patients in clinical trials may be switched from the placebo arm to the active treatment arm. This is particularly common when the active treatment is deemed effective early in the trial, and there is a perception that it would be unethical to retain patients on the less effective treatment. In order to control for this potential ‘diluting’ effect of the treatment crossover, a variety of statistical methods have been used and presented to the Appraisal

Committee. These techniques are being sometimes used within technology appraisals, and are generally presented without rationale or clear justification for the choice of analytical method.

2.2 Introduction to extrapolation

Often, follow-up data within clinical trials are short-term and incomplete and do not follow-up the long-term experiences of all of the participants in the trial. Frequently, important outcomes such as disease progression and overall survival are collected during the trial and for a limited period after the end of the trial but this data collection then stops (that is, the data are right censored). Particularly for chronic conditions, this means that only limited data on the number of people who progress and survive with and without treatment are available to inform the mean estimates of the clinical effectiveness and cost effectiveness of health technologies.

In instances whereby follow-up is incomplete, assumptions are regularly required to fully estimate the long-term benefits of a technology. Restricting decision making to the observed data available, especially in the presence of high levels of censoring, is likely to provide inaccurate and potentially biased estimates of the long-term effect of treatments and may ultimately lead to inaccurate estimates of the cost effectiveness of a technology.

In some instances, non-RCT evidence (or ‘real-world’ observations) are available and these can be used to estimate what would have happened if the participants in the trials had continued to be observed. It is however rare that these real-world observations are available; particularly in the case of newly licensed technologies, such long-term data simply do not exist. In these circumstances extrapolation of the observed data must be performed.

A number of methods are available for performing extrapolation. Exponential, Weibull, Gompertz, log-logistic or lognormal parametric models can be used. In addition, a number of more complex and flexible models are available such as piecewise exponential models. Some of these models allow for assumptions of proportional hazards between treatment arms, whilst others do not. The different methods have varying functional forms and the choice of

which model should be used varies according to the available data and each model has different characteristics which may make it more or less suitable for use in particular circumstances.[1;2]

The importance of extrapolation is often paramount in a technology appraisal. It is possible that the choice of the survival model can have a substantial effect on the resulting estimates of benefit (for example overall survival) which can subsequently have a dramatic effect on the mean cost effectiveness estimates. Therefore, the choice of extrapolation method is critical and should be considered a key issue for decision makers.

There are a number of techniques that can be used to determine which model is the most appropriate for extrapolating the data of interest. Firstly, a visual inspection (or 'eyeballing') the various curve fits to the observed data can be conducted. This method can be informative; however it is considered subjective and is therefore potentially inaccurate. Additionally, it is common that a number of parametric curves appear to fit the data well, and therefore visual inspection alone should be used with caution and is not considered sufficient for decision making purposes.[2]

Further, a number of statistical tests can be used to compare alternative models and their relative fit to the observed trial data. Log cumulative hazard plots or plots of residuals can be used to ascertain the nature of the observed data which in turn can inform the suitability of particular functions that can or cannot be used given the data. Once the curves have been fitted to the data, the relative 'goodness of fit' of the curves can be tested using methods such as the Akaike's Information Criterion and Bayesian Information Criterion tests. In addition, several other methods have been used in previous technology appraisals to justify the choice of curve fit to the observed data. It should be noted that patient-level data are required to conduct many of these tests; these are often not available within NICE appraisals.

The major limitation of extrapolating and the subsequent justification of extrapolation method used is that the techniques all rely on the observed data. The curves that are fitted can only be tested for goodness of fit to the

observed data (rather than the unobserved period). Thus, whilst it is possible to assess how well alternative curves fit the observed data, this does not provide any information with respect to the plausibility of the extrapolated curve beyond the observed trial follow-up period. Frequently, it is the long-term effect of a treatment on survival that has the greatest impact on estimates of the cost-effectiveness of the technology; this is particularly the case in many technology appraisals of cancer treatments. In these circumstances, the justification for selecting a particular curve fit is challenging, but the use of expert opinion, external data sources (such as historical cohort datasets or other relevant trials), and an assessment of the biological plausibility of the projected curves is recommended.[2]

Introduction to treatment switching ('crossover')

In randomised controlled trials, it is possible that participants randomised to the control group can be allowed to switch treatment group and subsequently receive the active intervention. This most commonly happens in trials of cancer treatments, whereby the participants in the control arm are switched to the intervention arm after they have experienced disease progression. This means that estimates of progression-free survival are often considered accurate, but that the switching of participants may confound the overall survival treatment effect. Within NICE appraisals in which this issue arises, methods to control for treatment switching are sometimes used to modify the estimate of treatment effect to be used in the health economic model.

In general, the use of the intention to treat principle is used to evaluate treatment effects within randomised controlled trials. This principle dictates that the treatment groups are analysed according to the treatment that patients were randomised to, regardless of the treatment that the patient actually received. However, where treatment switching has occurred and the active intervention is considered effective, then undertaking this form of analysis can lead to a 'dilution' of the overall survival treatment effect. One approach that is considered is censoring participants that crossed over from the control arm to the active intervention arm. This method is seen regularly by the Appraisal Committee, but is associated with limitations if the treatment

switching is not random and/or if a large proportion of the trial participants switched treatment and there are too few data left to conduct meaningful analyses.

Novel statistical methods for controlling for treatment switching have been presented to the NICE Appraisal Committee. For example the Rank Preserving Structural Failure Time (RPSFT) and the Inverse Probability of Censoring Weight (IPCW) models have been recently presented to the Appraisal Committee. The RPSFT method uses the randomisation of the trial in its estimation procedure in order to estimate counterfactual survival times (survival times that would have occurred if treatment crossover had not occurred). This method does not however change the level of evidence against the null hypothesis and therefore will always produce very wide confidence intervals around the point estimates, even if the point estimate is much reduced when these methods are applied. The IPCW makes a 'no unobserved confounders' assumption in order to create a 'pseudo population' consisting of control group patients that did not crossover, by making use of measurements of prognostic covariates over time. These methods are considered very complex and there are few experts who are competent in their use.

A recent review by the Decision Support Unit[3] considered these methods, among a number of other statistical techniques that can be used to control for treatment switching. It is also possible that further techniques will be developed in the future. Whilst it is rare that such statistical techniques are used, when they are the justification for the choice of method is seldom given to the Appraisal Committee.

2.3 What the current Methods Guide advises with respect to extrapolation and crossover

The current Methods Guide is detailed with regards to the need for extrapolation:

> *3.2.3 ... However, it is important to recognise that, even for the analysis of relative treatment effects, RCT data are often limited to selected*

populations and may include comparator treatments and short time spans that do not reflect routine or best NHS practice. Therefore, good-quality non-randomised studies may be needed to supplement RCT data….

Section 5 (Time Horizon) says:

5.2.14 … For a lifetime time horizon, extrapolation modelling is often necessary. When the impact of treatment beyond the results of the clinical trials is uncertain, analyses that compare several alternative scenarios reflecting different assumptions about future treatment effects should be presented (see section 5.7 on modelling). Such assumptions should include the limiting assumption of no further benefit as well as more optimistic assumptions…

Section 5 (Modelling Methods) says:

5.7.3 Modelling is often required to extrapolate costs and health benefits over an extended time horizon. Assumptions used to extrapolate treatment effects should have clinical validity, be reported transparently and be clearly justified. Alternative scenarios should be considered to compare the implications of different assumptions around extrapolation for the results. For example, for the duration of treatment effects scenarios might include when the treatment benefit in the extrapolated phase is: (i) nil; (ii) the same as during the treatment phase and continues at the same level; or (iii) diminishes in the long term.

There is currently no discussion of methods to control for treatment switching ('crossover') in the 2008 Methods Guide.

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 by the Methods Guide Review Working Party.

3.1 Extrapolation

Currently extrapolation is mentioned in reasonable detail in the methods guide. However, the consistency in submissions varies widely:

- Should further direction be given regarding the use of extrapolation?
- Should the choice of extrapolation methods and functions be explicitly specified (or at least preferences for particular methods and functions stated)?
- Should the methods guide be more explicit about distinctions between extrapolating the baseline curve and the relative treatment effect?

What are the potential consequences of explicitly specifying what methods and functions should be used?

- Should a number of alternative fits always be shown?

What could be the impact of specifying a minimum number of curve fits (for example, stating that a single curve fit is not acceptable)?

- Should the justification of choice of model be made more explicit
    - Should goodness of fit statistics be used?
    - Should external sources, such as clinical opinion, always be sought to support extrapolation?
    - How should the biological plausibility (face validity) of an extrapolation be demonstrated?
    - Can a ‘minimum’ for justifying the choice of extrapolation be defined? For example goodness of fit alone, clinical opinion alone, a mixture of this and other components?

- How should these concerns differ according to the availability of patient-level trial data?

What could be the impact of always requesting goodness of fit statistics and/or additional support for extrapolation methods? What could be the impact of defining a ‘minimum’ process of justifying extrapolation methods?

How should NICE Methods Guide draw on the information presented within the NICE DSU Technical Support Document on survival analysis when patient-level data are available?

3.2 Treatment switching (‘crossover’)

There is currently no mention of when it is appropriate to consider controlling for the effects of treatment switching and what methods should be considered. Given that treatment switching is being seen more and more commonly in clinical trials:

- Should any guidance be given with respect to when it may be appropriate to consider statistical methods to control for treatment crossover?

What could be the consequence of including such direction in the methods guide?

- If any guidance is to be given, should specific statistical methods be referred to? Note: it will be important to consider not restricting the methods guide in terms of potential methodological developments but also the likely impact of on the review groups when receiving submissions that employ the variety of available complex methods

- If statistical methods are to be used, should guidance on how the choice of method is justified be given?

What might be the impact of including explicit guidance on preferred methods that could be used? What might be the consequences of including guidance on how to justify the choice of any methods used?

4 References

1 Collett D. Modelling Survival Data in Medical Research 2009. Chapman and Hall: Florida

2 Latimer, N. NICE DSU Technical Support Document 14: Undertaking survival analysis for economic evaluations alongside clinical trials - extrapolation with patient-level data. 2011. Available from http://www.nicedsu.org.uk

3 Morden JP, Lambert PC, Latimer N, Abrams KR, Wailoo AJ. Assessing methods for dealing with treatment switching in randomised clinical trials. Available from http://www.nicedsu.org.uk/Crossover%20and%20survival%20-%20final%20DSU%20report.pdf

5 Authors

This briefing paper has been prepared by Rebecca Trowman, Nick Latimer, Andrew Stevens and Paul Tappenden.

NATIONAL INSTITUTE FOR HEALTH AND CLINICAL EXCELLENCE

Briefing paper for methods review workshop on Measuring and valuing health effects

The briefing paper is written by members of the Institute's Decision Support Unit. 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

2.1 Current NICE reference case on health effects

The NICE reference case on measuring and valuing health effects contains the following key features:

Quality Adjusted Life Years (QALYs) as the measure of health effects

The QALY combines the effects of an intervention on survival and health related quality of life (HRQL) into a single measure, by placing HRQL onto a scale where full health is one and dead is zero. This allows all health outcomes to be expressed in a common metric that allows comparisons across interventions. This has been a cornerstone of cost effectiveness methods in NICE Technology Appraisals for many years. There have been concerns that this focus excluded potential impacts on non-health outcomes

and aspects of the processes of care. This issue was partly addressed in the workshop on perspectives and is considered briefly in this review.

HRQL should be reported by patients

The Methods Guide states clearly that it expects to see the HRQL data to come from patient self-report. This reflects the evidence that carers and professionals reporting on the health of patients is often in disagreement with those of the patient, particularly for the more subjective dimensions of pain or mood. However, the Guide recognises that there may be circumstances where patients are unable to report their own health (e.g. cases of severe cognitive problems) and in this case the Methods Guide specifies close carers as the source for proxy data. This issue is not considered further in this review.

The health effects on carer givers is included

It is sometimes forgotten that the NICE Methods Guidance does allow the impact on caregiver's health to be included in the QALY calculation (unless the caregiver is employed by the NHS). There is a question of whether the measurement of external effects should extend to other family members not involved in caring and there is some debate in the literature on this subject (to which we return later).

Health effects valued using a choice-based method by the general public

The use of choice-based methods has been stipulated in the last two versions of the NICE Methods Guide, which are those preference elicitation techniques that require respondents to explicitly consider a trade-off between HRQL and some other part of their utility function, such as longevity (i.e. time trade-off (TTO)) or risk of death (i.e. standard gamble (SG)), rather than rating scales that ask for an assessment of feeling about a state. EQ-5D is the preferred instrument and it uses TTO to value health states (see below). Where other preference-based measures are used the Methods Guide requires the use of comparable valuation methods to EQ-5D, in other words TTO. This review examines this issue in as much as the EQ-5D will be re-valued in the near future probably using a different variant of TTO.

NICE decided on using general public values rather than those of patients or others due to the perspective of the decisions being taken. However, this continues to be a subject of considerable debate in the literature and policy circles, but not for this review

EQ-5D as a preferred measure of HRQL

The EQ-5D is a generic preference-based measure of health and has been validated in many conditions. The version of EQ-5D currently in use consists of 5 dimensions (mobility, ability to self-care, ability to perform usual activities, pain and discomfort, and anxiety and depression) and each dimension is described by a single 3 level item. Patients complete a 5 item one page questionnaire in order to assign them to one of the 243 states this descriptive system defines. There are a set of estimated preference-based health state values for each of the 243 states using values elicited from the UK general population by TTO. Recently the EuroQol Group has produced a 5 level version and is currently embarking on a programme of work to produce a new UK value set. An important consideration is whether these developments should be incorporated into the NICE reference case and if so, when.

NICE prefers a single measure of HRQL to be used in cost effectiveness models to promote consistency across appraisals. There is substantial evidence that other preference-based measures of health, be they generic (e.g. HUI3 or SF-6D) or condition specific, produce different values for the same patient. In order to compare across studies it is important to use the same measures. However, this raises the issues of what to do when EQ-5D data are not available and when EQ-5D is not appropriate in the patient groups.

EQ-5D is appropriate but unavailable: use of mapping

The amount and coverage (e.g. by medical condition) of EQ-5D data are increasing all the time. However, the NICE methods Guide recognised that sometimes there may not be sufficient relevant EQ-5D data and so they recommended the use of mapping (or cross-walking) methods in order to generate EQ-5D values from other measures of HRQL or even other clinical measures. This raises two important questions: when is it appropriate to use

mapping and how should mapping be undertaken? Nothing more specific was said in the last Methods Guide about when mapping should be undertaken. The Guide specifies that mapping should be based on empirical data (i.e. rather than judgement), it should have clearly described statistical properties and it should be validated. The NICE DSU Technology Support Document on Mapping (TSD 10) provides useful advice on how to undertake mapping studies. An issue to be addressed at this workshop is whether any of this advice should be incorporated into the NICE Methods Guide.

Appropriateness of EQ-5D and the alternatives

NICE recognised that there may be conditions or treatment effects that will not be adequately captured by the 5 dimensional 3 level EQ-5D. However, this was anticipated to be the exception rather than the rule. The inappropriateness of EQ-5D needs to be demonstrated with evidence on the properties of content validity, construct validity, responsiveness and reliability in the relevant patient population. Where an alternative measure is used, then the submission should give the reasons supported by evidence on these same properties.

Guidance on alternatives to EQ-5D was that these should be based on the direct valuation of a standardised and validated HRQL measure. This would seem to suggest that those states developed by experts, sometimes known as vignettes, should not be used in the reference case because they do not relate to patients reporting on their HRQL and so have little empirical basis. Another generic measure or a condition specific measure may be considered. However the NICE Methods Guide states that '…*the valuation of the descriptions should use the TTO method in a representative sample of the UK population, with 'full health' as the upper anchor, to retain methodological consistency with the methods used to value the EQ-5D*'. Of course, those submitting evidence are allowed to use other methods in any sensitivity analyses. This continues to be a contentious topic and there has been substantial research in the use of condition specific measures, so it seems right to re-consider it at this workshop.

Use of measures in children

It is recognised in HTA that there are important conceptual differences between children and adults in terms of the dimensions of HRQL as well as linguistic differences. It was recognised that there was not an obvious candidate measure for the reference case measure of HRQL in children. At the time of preparing the last Method Guide there was just the HUI2, but this was not felt to have the same status and uptake as EQ-5D to justify making it the preferred measure. The Guide asks those submitting evidence to consider the use of standardised and validated preference-based measures of HRQL, including the HUI2. Since that time a number of measures have been developed for use in children and so this review will consider this question further to see whether there is a clear candidate measure and the issue of whose values.

Use of the literature and other secondary sources

It is recognised that for populating cost effectiveness models, EQ-5D data may come from a number of different sources. Clinical trials have the attraction of internal validity, but they may not be generalisable to the populations being modelled, they may not follow-up outcomes for long enough and may not have sufficient data on key events (e.g. adverse effects). For this reason, it will often be appropriate to use other sources of data, such as observational studies, routine data sets (e.g. UK PROMS) or values published in the literature.

The NICE Methods Guide requires that estimates for the utility values for health states from published literature must be shown to have been identified and selected systematically. Where there is more than one plausible source of health state values sensitivity analyses are recommended. The ever growing published literature makes this an increasingly important source of values. This review does not propose to look at this issue any further, but readers are recommended to consult TSD 9 which provides detailed recommendations on how to conduct reviews of health state utility values (Papaioannou et al, 2011).

2.2 Relevance of health effects to the Appraisal Committee

The measurement and valuation of the health effects of technologies is a fundamental component of the assessment of the cost effectiveness of health care interventions for the NICE Technology Appraisal. The previous review of NICE methods published in 2008 provided an overview of the core issues in measuring and valuing health including what is to be measured (e.g. should it be the QALY?), how it is to be described (e.g. should it be generic or specific to the condition?), how it is to be valued (e.g. time trade-off or standard gamble), whose values (e.g. general population or patient) and how should it be aggregated (e.g. is a QALY is a QALY regardless of who gets it or should QALYs be weighted in some way) (Brazier, 2007). The purpose of this briefing paper is not to revisit these core issues in general but to address a number of specific questions that have emerged since the publication of the last review of methods where it is deemed that there needs to be firmer guidance or where there have been important developments or research that have implications for the existing reference case methods in this area. Some of these have arisen from the development of the NICE DSU Technical Support Document (TSD) series of 5 on utilities (for further information see www.nicedsu.org.uk). The TSD series provides a review of the state of the art across a number of important issues in this topic to assist those making Technology Appraisal submission to NICE, but it is not a formal part of the NICE Methods Guide.

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:

1. When is the EQ-5D not an appropriate measure of health-related quality of life?

2. What are the alternative instruments and when are they more appropriate?

3. When is mapping the preferred approach? What principles underpin good mapping analysis?

4. Should NICE adopt the new 5 level version of EQ-5D and its associated value set?

5. What preference-based measure of HRQL should be used in children?

6. Measurement and valuation of health effects on people other than the recipient of the intervention. How should ‘related individuals’ be defined, measured and aggregated?

3.1 When is the EQ-5D not an appropriate measure of health related quality of life?

The NICE reference case expressed a preference for the EQ-5D to measure HRQL in adults. It permits the use of other measures where it can be demonstrated that EQ-5D is not appropriate and provide reasons for the alternative supported by evidence. The Guide specifies the properties of reliability, content validity, construct validity, and responsiveness for assessing appropriateness.

Reliability takes two forms. One is random variation between assessments, and this has implications for sample size and precision of estimates for any given sample size. Where sample sizes are small, then this may be a cause for concern and there could be a case for using estimates from another instrument prone to less variation. However, in most cases a larger sample size will be the solution. Of more concern is unreliability from variation between methods of assessment. There is little evidence on this issue but what there is suggests there may be little difference between pencil and paper and computer completion of EQ-5D (Lloyd et al, 2011). However, there is evidence of significant differences between patient self-report and carer proxy report.

The assessment of validity is far more problematic due the lack of a gold standard measure of HRQL. While some health economists have been sceptical as to whether it is possible to assess the validity of preference-based measures, it is an important challenge facing the measurement of all psychological phenomena. The methods Guide recognised that the EQ-5D is not appropriate in all populations (Brazier and Longworth, 2011; Wailoo et al, 2010).

Content validity is concerned with whether the instrument covers all the dimensions of HRQL of importance to patients. This can be assessed through qualitative work with patients to identify ways in which their health status impacts on their physical, psychological and social functioning and wellbeing. Construct validity requires quantitative evidence on whether the measure reflects known differences between groups or converges with other relevant measures. Responsiveness is the extent to which the measure reflects changes in HRQL overtime. These criteria would preferably be assessed across the 5 dimensions of the EQ-5D as well as the overall index, though this is rarely done. Careful consideration must be given to the relevance of the variables used to test validity, which are often clinical assessments of symptoms, since these may not be important for patient's HRQL. Furthermore some of the conventional psychometric criteria may not apply to preference-based measures. In conventional psychometric analyses it is the instrument with the largest difference that is deemed best, for example as assessed using a standardized effect size (mean difference divided by standard deviation of the difference). However, bigger is not necessarily right since a highly focused measure of symptoms may achieve the highest effect size and yet not reflect the impact on overall HRQL and not be valued by the general public. The other extreme would be to argue that there must be no differences between the known groups or changes over time for the EQ-5D to be judged invalid. The truth may lie somewhere between the two and there will always remain a considerable degree of judgment in deciding on the validity of EQ-5D in any one patient group and whether another measure is more appropriate (as a measure of HRQL that matters to the general population).

An empirical literature on the validity of EQ-5D and other preference-based measures has begun to emerge over recent years. The standards of testing are often not high and are prone to the problems of interpretation highlighted above. While this is not the place to present a detailed review of the evidence, there have been a number of reviews recently conducted of the literature that give some idea of the extent of the problem. Evidence from recent reviews on construct validity and responsiveness suggests the EQ-5D is probably not appropriate for assessing the impact of hearing loss, some specific forms of visual impairment and schizophrenia (TSD 8). On the other hand it would seem that the EQ-5D is more appropriate in areas including depression and anxiety, a number of key cancers and skin conditions. However the evidence is at best patchy and often poor quality, with little evidence on content validity. In many cases, there is simply not sufficient evidence one way or the other to make definitive judgements about the suitability of EQ-5D for a given condition and there is often even less evidence on other generic or the condition specific preference-based measures.

Where alternative measures are used, those submitting evidence are required to demonstrate the likely impact on the cost effectiveness of the intervention (i.e. through sensitivity analyses). In many cases, it may not impact on the decision. However, where there is a potential impact on the decision, this still leaves the Appraisal Committee with a judgement about which should be used.

Discussion points:

- Should NICE penalise products that don't have EQ-5D data?
- How strong an evidence base is required to decide a measure is inappropriate?
- Should NICE provide more guidance on what evidence is required, how it should be reviewed and how it should be presented?
- Should NICE stipulate in advance, such as in the scoping stage, whether other measures are deemed more appropriate than EQ-5D (e.g. HUI3 in hearing loss)

3.2 What are the alternative instruments and when are they more appropriate?

The reference case argues strongly that alternatives to EQ-5D should be based on a validated patient reported outcome measures rather than vignettes based on expert opinion and they should be valued using methods comparable to those for the EQ-5D. While the range of alternative generic measures in adults has not changed, there has been a large increase in the number of condition specific preference-based measures covering diseases such as asthma, cancer, dementia, sexual functioning, Parkinson's disease, visual function, urinary incontinence, mental health and many more with 28 identified in a recent review (Brazier et al, 2011). The concern about condition specific measures comes from the lack of comparability across them and so limits their use in making decisions across programmes. This problem may arise even where the valuation methods are the same (i.e. same upper anchor, same valuation method and same source of values) due to focusing effects (whereby respondents overemphasise those specific dimensions mentioned in the state since they are not placed in the context of overall HRQL), use of disease labels (that may lead to respondents in valuation surveys bringing irrelevant prior beliefs about the condition into their responses), and problems capturing important side effects and comorbidities (that may interact with condition specific dimensions).

There has been little work comparing these new measures to existing generic measures like EQ-5D in terms of their validity. What there is suggests that the condition specific measures do not tend to produce larger differences in utility values, though there are cases of that, but rather they provide more precise estimates because they are associated with smaller standard deviations. This is important for reducing the uncertainties around specific estimates. However, the evidence on whether they are more sensitive to particular differences is mixed, with some evidence that they are better at reflecting differences at the upper end of HRQL (Brazier et al, 2011). This does not suggest that focusing effects are unimportant, but rather that comorbidities

seem to be more important. However, this experience is likely to vary between conditions and measures.

More recently researchers have begun to examine the potential for adding on extra dimensions to EQ-5D as another means of overcoming the apparent lack of sensitivity or relevance in some conditions. Research has examined 'bolt-on' dimensions for vision, hearing, sleep and cognition. This research is at an early stage, but it has the potential of improving EQ-5D in some key conditions while at the same time overcoming some of the limitations with condition specific measures.

Other alternatives continue to be used, such as vignettes and patient's own valuations (where they value their own state using TTO or SG). Should these data continue to be admissible as evidence in submissions to NICE technology Appraisals? If so, should this be agreed at the scoping?

Finally, there is a concern that important elements of patient's experience of the processes of care are excluded from outcomes measures like EQ-5D, such as regular hospital attendance, oral versus insulin medication for diabetes and dignity of care. These have been dealt with using vignettes in some submission that brings concerns about having a poor evidential basis. There is a growing literature using techniques such as DCE to combine process and outcome attributes. This is promising where patient experience is being assessed using validated patient reported measure, but it raises two concerns. One is that this extends the scope of the appraisal of benefits beyond health and hence beyond the current reference case. Secondly, even where it is decided such benefits should be taken into account they are often on a different scale to the health effects. There has been work attempting to treat the process attribute as a bolt on to the EQ-5D, but this is at a very early stage of development.

Discussion points:

- When should alternative measures be used? When EQ-5D is appropriate (instead of mapping) or only where EQ-5D is inappropriate?

- Should alternatives be presented in the main analyses where EQ-5D has been shown to be inappropriate?
- For those conditions where EQ-5D is shown to be inadequate, should NICE express a preference for an alternative measure (e.g. preference-based VFQ-25 in visual functioning).
- Does there need to be evidence demonstrating how use of an alternative has impacted on QALY estimates?
- What should be the role of other (i.e. non-reference case) alternatives such as vignettes, patient values and should the experience of process benefits be taken into account?

3.3 When is mapping the preferred approach? What principles underpin good mapping analysis?

Mapping (or cross walking) involves the development and use of an algorithm to predict EQ-5D values using data on other measures or indicators of health. The mapping algorithm should be based on statistical association and not expert judgement. The estimation of mapping functions requires an estimation sample containing the target variable (i.e. EQ-5D) and the source variable (e.g. another measure of HRQL). A statistical model is then estimated mapping the source onto the target using a range of possible specifications and estimation techniques and then it is applied. It can be used to predict EQ-5D values from data sets where it was not used, such as clinical trials or observational studies that are being used to populate an economic model. A recent review found that one quarter of submissions to the TA programme had used health state values from mapping algorithms (Tosh et al, 2011).

Mapping is usually a second best solution to using the EQ-5D in the population of interest (but there may be some exceptions to this such as where the sample in the trial is too small to obtain sufficiently precise estimates). As described below, there are known errors in mapping models that are best avoided. So an important question is when should it be the preferred approach? It should only be used when there are insufficient

relevant EQ-5D data. For some health states in a model, relevant EQ-5D values may already exist in the literature and so predictions based on mapping functions would be inferior. To ascertain the existence of relevant EQ-5D values requires a systematic search and review of existing literature. (For advice on how to do this see TSD 9). It might also be necessary to adjust published values to make them suitable for the population in the economic model, such as age or the existence of comorbidities. There are methods for making such adjustments and these are described in TSD 12 (Ara and Wailoo, 2011). The choice of using health state values from mapping, literature sources or EQ-5D data from specific trials depends on context. Mapping may be preferable to the literature where the latter does not cover the right population or misses important side effects of treatment, on the other hand literature values may be based on direct use of EQ-5D and better reflect the population of interest in the model than a pivotal trial.

Details concerning the methods of mapping are provided in TSD 10 (Longworth and Rowen, 2011). In summary the key concerns in mapping cover the estimation sample, model type, the model specification, uncertainty and validation. The characteristics of the estimation sample should be similar to the sample to which the mapping function will be applied. The choice of model should depend on the nature of the data and the expected relationships. EQ-5D data are not easy to model due to the skewed, censored and multi-modal nature of the distribution of the values. Ordinary Least Squares (OLS) regression models tend to be the most widely used, but this has theoretical limitations though often performs better than the alternatives. Attempts to improve on OLS include the Tobit, CLAD (but this provides median estimates), two part models, splining or mixture models. Some have modelled the responses to the classification rather than the EQ-5D index, which involves a two stage procedure of modelling onto the 5 dimension responses and then applying the EQ-5D value set. It is not possible to recommend any one method in all cases at this stage.

There should be clear reporting of the model and its performance. This should include statistical properties such as coefficients (e.g. size, significance),

mean absolute and root mean squared error; error reported across the EQ-5D score range and plots of observed to predicted values. Mapping functions should ideally be validated using external datasets. Mapping functions are often poorly reported in the literature with little attention given to such things. One solution is to be more prescriptive about reporting standards for mapping functions used to populate models and even require the data sets on which they are based to be made available to the Technology Assessment Groups where it has not been published.

A common finding is that mapping functions overestimate at the lower end and under-estimate at the upper end, and this can result in a reduction in the size of differences between health states based on severity or changes overtime. On the other hand mapping functions can result in less variability than the original EQ-5D. There is a tendency to ignore the uncertainties underlying the statistical model in the sensitivity analyses. NICE and other using mapping functions need to better understand the impact of using values estimated by mapping functions than observed EQ-5D values in cost effectiveness models.

Discussion points:

- When should mapping be used compared to using original EQ-5D data or literature values? When should mapping be used rather than alternatives (see above)?
- Should NICE recommend specific systematic reviews or databases of HSUVs for those submitting evidence and reduce the need for mapping?
- Does NICE need to be more prescriptive about the principles or methods used to estimate mapping functions and how they are presented? Or is the advice in TSD 10 sufficient?
- Should NICE recommend stand mapping functions or agree on one to be used at the scoping stage?

- How should the uncertainties underlying mapping functions be reflected in the cost effectiveness model?

3.4 Should the NICE Methods Guide adopt the new 5 level version of EQ-5D and its associated value set?

3.4.1 The 5 level EQ-5D

While the 3 level version EQ-5D has been shown to be valid and responsive in many conditions it has been criticised for the crudeness of having just 3 levels. With just 3 levels there are large proportion of patients at the ceiling (i.e. many respondents with health problems are allocated to state 11111) and a general insensitivity to change when the response categories involve such large steps. The EuroQol Group has been developing a five level version that retains the 5 dimensions with the descriptors adapted to a 5 level version as follows: no problem, slight problems, moderate problems, severe problems and unable to or extreme problems. The worst level of mobility has been changed from ‘confined to bed’ to ‘unable to walk’ and usual activities from performance to doing. Papers are starting to emerge using theEQ-5D- 5L and an important question is whether and how this version should be incorporated into NICE Guidance.

The argument for using the 5L for collecting data is that it would provide a more sensitive instrument. The evidence to date that this is the case is quite limited. There is evidence of a reduction in the numbers at level 1, a more even distribution across the levels and a modest increase in the correlations with related measures of health (Bas et al, 2011). There are only a couple of studies and these have been conducted by members of the EuroQol Group and not by independent researchers (Pickard et al, 2007, Janssen MF et al, 2011). There would be an intuitive case for saying 5 levels is an improvement, but the size and extent of the improvement across conditions is not known. Furthermore, there is no published evidence on the extent to which general population respondents in a valuation survey are able to distinguish between the 5 levels.

Another limitation is that no UK tariff for the 5L currently exists. There are plans to produce one in the UK (to be funded by DH), and these are discussed below. Mapping functions have been estimating for scoring the 5L from the 3L tariff. A number of methods have been explored for estimating mapping functions including OLS, non-parametric models, ordered logistic regression and item response theory (Janssen et al, 2011). These seem to achieve a similar fit with RMSEs of around 0.12. These functions suffer from familiar problems with a reduced range (since it does not predict one in many cases) and slightly flatter gradient to the predictions than would be expected from an exact fit. The implication of this error for differences between key states has not been explored (e.g. between grades of severity of different conditions).

There is a major cost to NICE in recommending an instrument that will in the end produce different values for the same patients as the EQ-5D-3L. Possible recommendations to NICE include: never adopt, adopt after further evidence and the value set become available or recommend it is used now and in the interim use the value set from the 3L. Assuming the 5L brings advantages and there will be a re-valuation of the EQ-5D in the UK in case (as planned), then to never adopt may become an untenable position. To delay recommending the use of the 5L until the next review would delay any benefits by four years. To recommend its use now and suggest the mapping function will bring some of the problems associated with mapping and indeed further statistical complications for mapping from other measures onto EQ-5D (i.e. the double mapping problem).

3.4.2 The new value set for EQ-5D

The reference case tariff of values was obtained from members of the UK general population more than 15 years ago using TTO. The version of TTO was the MVH protocol where for states better than dead respondents are asked to compare living in health state h for 10 years and x years in full health (where x<10). At the point of indifference the value of h is x/10. For states worse than dead the choice is between (a) health state h for y years followed by full health for x years, after which they will die, or (b) immediate death.

Years in the health state, y (=10-x), and years in full health (x), are varied to determine the point where the respondent is indifferent between the two options. The value of h that is consistent with the theory is –x/y (i.e. x/(10-x)) . However when using the TTO protocol where t=10 this produces values bounded at -39 for the minimum possible value for any health state, where x=9.75 (i.e. 3 months followed by full health for 9 years and 9 months). State worse than dead responses have a larger impact on the model predictions than better than dead responses. For this reason the TTO data for states worse than dead were rescaled to onto 0 to -1 using for formula –x/10 (Dolan, 1997). The values for states better than dead and the transformed values for states worse than dead are pooled and modelled using regression techniques to estimate the tariff.

This review does not address the more general concerns with using TTO (such as the assumption of constant proportional trade-off) or the use of preferences rather than patient experience to value states). An important criticism of the value set, aside from its age, is the handling of states worse than dead. This is important since a third of mean EQ-5D health state values are negative and so worse than being dead and all other states have some negative responses. It currently uses a different valuation procedure for states worse than dead and respondents may view the prospect of returning to full health following a severe health state as unrealistic. The rescaling is arbitrary and it has been argued that the values can no longer be interpreted as utility values. The values produced by the two procedures are arguably not on the same scale.

One approach to deal with the latter problem is to incorporate the correct formulae for states better and worse than dead into the econometric model via an ‘episodic random utility model (Craig et al, 2009). The main contribution of the episodic RUM model is that all TTO responses are treated identically in the model specification. Yet this does not resolve the problems outlined earlier that the TTO choice tasks are different for states valued as better than or worse than dead. This approach is not being used by the EuroQol Group.

Another proposal is to use a different TTO procedure, such as one that introduces a ‘lead time’ whereby a period in full health is added to the start of the usual TTO, meaning that states worse than dead can be valued by cutting in to the lead time (Devlin et al, 2011). The ‘lead time’ TTO task provides respondents with a choice between (a) full health for f years followed by health state h for 10 years, after which they will die, or (b) full health for f+x years, after which they will die. Years in full health, x, is varied to determine the point where the respondent is indifferent between the two options where x can be negative where the lead time is exhausted. The utility for health state h is calculated using x/10. This approach has the advantage that it does not draw attention to the fact that respondents are valuing a state as worse than dead, yet this may mean that respondents are not fully aware of what their responses indicate. The lead time can be exhausted and so respondents may have to revert to a different procedure in some cases. This method also makes a strong assumption of additive separability where the value of state h should not be affected if it is preceded by full health for period f, and may suffer from the problem of ordering effects in moving from full health to a poor health state. This new procedure and others (including a ‘lag’ time TTO) is the subject of further methodological research being undertaken by members of the EuroQol Group and elsewhere.

Finally there has been research looking into the use of ordinal methods for valuing EQ-5D states. Initially this looked at the use of rank data (Solomon, 200X), but more recently the research has begun to look at discrete choice tasks. Asking respondents to compare EQ-5D states will provide values for those states on a latent scale, but leaves the problem of how to anchor onto the full health-dead scale required for calculating QALYs. One solution to this problem is a hybrid approach, whereby some states are valued by TTO and then the DCE and TTO data combined through anchoring or mapping to produce a value set (Rowen et al, 2011)). Lastly, there is a DCE task where survival is added as a sixth dimension and in effect providing a new TTO task where the pairs of scenarios are determined by a statistical design rather than a standard elicitation procedure. This does not get away from some of the concerns with TTO, such as the assumption of constant proportional trade-off

(although this can be tested within this approach), but it avoids the need for a different task for states worse than dead. Initial testing of this approach suggests it has promise (Bansback et al, 2010) and is currently being examined by the EuroQol Group.

The EQ group has been testing the various alternatives and currently has not decided on the best approach. However, it intends to make a decision in the near future.

- Does the 5L represent a sufficient improvement for NICE to recommend it is used: 1) as a reference case or 2) to collect data for the time being, and be adopted as the reference case at a later point in time?
- Will the 5L compromise NICE's need to be consistent in decision making? How will submissions using the 5L be compared to previous 3L ones?
- Should NICE adopt the EuroQol Groups final decision regarding the method of valuation?
- If yes, when should a new tariff be adopted as the reference case by NICE?
- What should be the transition arrangements for moving from 3L to 5L??

3.5 What preference-based measure of HRQL should be used in children?

There are now three preference-based measures for children or adolescents (HUI2, AQOL-6D or AQOL-8D, and CHU-9D) and one in development (EQ-5D-Y). The HUI2 has 6 dimensions (sensation, mobility, emotion, cognition, self-care and pain) and comes with a UK SG value set elicited from adults (in addition to the Canadian used in most published studies. It was developed by experts based on a survey of parents in Canada. The AQOL-6D has six dimensions (independent living, mental health, coping, relationships, pain and senses) that were adapted from the adult instrument (ref) and there is an

Australian value set obtained using TTO elicited from adults. The AQOL-6D also has a valuation tariff from adolescents which was developed using a transformation of the adult values from a sample of states valued by adolescents. The AQOL-8D contains 2 additional dimensions to the AQoL6D and has a valuation tariff from adults. The CHU9D is the only instrument where the content of the descriptive system was developed from interviews with children about the way their health impacts on their HRQL. It was developed in children aged between 7-11, but has been used in adolescent children up to 17 years. Finally there is the EQ-5D-Y whose descriptive system has been developed from the adult EQ-5D without any alteration of the conceptual dimensions, just a change in language to make it understandable by young people. This continues to be under development and currently does not come with a value set. While these measures are starting to be used more in research, particularly in their self-reported form, there is no single measure that stands out in terms of being more widely adopted or performing notably better.

The measurement and valuation of HRQL is more complicated in children and raises important practical problems and normative issues. While self-report is being increasingly used, there are difficulties in younger age groups. There is little experience in younger children (e.g. <7), where measures of health tend to be confounded by childhood development (e.g. scores can improve simply because the child gets older). Indeed the relevance of any of these measures in the under 5 population is questionable. It is also not clear where the boundaries are between childhood, adolescence and adulthood, and how the transition between instruments should be handled when calculating QALYs or trying to make cross programme comparisons. All existing instruments use adults to value the states, but there has been interesting work in trying to elicit preferences form older children using ordinal methods that is showing promise (Ratcliffe et al, 2011), though problem of anchoring onto the full health-dead scale remains. The question of whose values presents an important normative dilemma and one that will vary by age (if for no other reason than younger children may not understand the task).

Research into measuring and valuing HRQL in children is on-going and many of these issues cannot be resolved at this workshop.

Discussion points:

- Are separate measures required for children, adolescents and adults, if so, what should be the ages of transition?
- Should NICE be encouraging self-report (at least in older children)?
- Should one instrument be preferred over the rest for certain age group?
- Are adult values acceptable or should NICE be encouraging the development of values sets based on the values of children and adolescents?
- How should the transition between instruments be dealt with in a cost effectiveness analyses of interventions with impacts across age groups and should comparison be made across programmes by age?

3.6 Measurement and valuation of health effects on people other than the recipient of the intervention. How should 'related' individuals be defined and how should the effects be measured and aggregated?

With an ageing population, the health system increasingly relies on close family and friends to provide informal care. This may impact on the caregiver's health, and the current NICE reference case allows for the incorporation of these health effects in the calculation of the overall QALY impact. Within the reference case this would normally be measured using the EQ-5D. The time of carers is not currently included within the NHS and social care perspective taken in the reference case.

There are important questions regarding who should be counted as a caregiver. It does not include professional caregivers who are already compensated for their time and effort and will be included in the staff cost in an economic evaluation. Providing informal care has been shown to impact on physical and/or psychological health, and has even been associated with a

higher risk of morality (Brouwer, 2006). This is taken into account in the current NICE reference case. However, there may be significant others who do not provide care (e.g. the children of ill parents) whose health will be affected by having members of their family who are unwell, particularly through their emotional well-being (Bobinac et al, 2010). To exclude such 'family' effects' requires the separation of family members not only into two groups, those who give care and those who do not. It also means having to net out the family effect in those who give care. To include them substantially increases the data requirements of economic models.

Another consideration is whether carer and/or family effects are already proxied by the EQ-5D. In which case, there is little need to add it into the QALY estimate since it will impact on all interventions equally. However, it is suspected that for a given EQ-5D score the impact will vary by condition, severity of condition, age (e.g. children), type of treatment (e.g. at home or in hospital), type of care being provided and the nature of the relationship. However little is understood about these relationships at present.

Finally there is the question of how to measure the impact on carers and family beyond health effects. Recent years has seen the development of quality of life scales for use with carer. However, these scales are not anchored on the full health-dead scale and even if they could it raises an important problem of how to aggregate broader measures of quality of life in carers with the health effects of the patients. If the measure for carers uses a broader notion of quality of life then why should the measure for patients be limited to HRQL?

Discussion points:

- Should the impact on significant others be broadened out to include other members of the family who are not directly involved in care?
- Should the impact on carers and significant others be limited health effects or extended to quality of life more generally?
- How should impacts on carers, significant others and parents be aggregated?

4 References

Ara R, Wailoo A. NICE DSU Technical Support Document 12: The use of health state utility values in decision models. 2011. Available from http://www.nicedsu.org.uk

Bobinac A, Van Exel NJ, Rutten FF, Brouwer WB. Caring for and caring about: disentangling the caregiver effect and the family effect. J Health Econ 2010 Jul;29(4):549-56.

Brazier J. Valuing health states for use in economic evaluation. *Pharmacoeconomics* 2007, 26(9):769-779.

Brazier J, Rowen D. NICE DSU Technical Support Document 11: Alternatives to EQ-5D for generating health state utility values. 2011. Available fromhttp://www.nicedsu.org.uk

Brazier J, Longworth L. NICE DSU Technical Support Document 8: Applying the NICE reference case to the measurement and valuation of health. 2011. Available from http://www.nicedsu.org.uk

Brazier J, Rowen D et al. Developing and testing methods for deriving preference-based measures of health from condition specific measures (and other patient based measures of outcome) Health Technology Assessment (forthcoming)

Craig BM, Busschbach JJ. (2009) The episodic random utility model unifies time trade-off and discrete choice approaches in health state valuation. Population Health Metrics. 13 no. 7: 3.

Devlin, N., Tsuchiya, A., Buckingham, K.J., Tilling, C. A Uniform Time Trade Off Method for States Better and Worse than Dead: Feasibility Study of the 'Lead Time' Approach. *Health Economics* 2011; Forthcoming.

Dolan, P. 1997. Modeling valuations for EuroQol health states. *Medical Care* 1095-1108

Janssen MF, Birnie E, Haagsma JA et al. Comparing the standard EQ-5D three level system wuth a five level version. Value in Health (in press).

Janssen MF et al. Values sets for the EQ-5D-5L. EuroQoL Paper.

Lloyd, A.J., Kind, P., Thompson, T., Leese, B., Nixon, A., Quadri, N. Paper to web: equivalence testing of EQ-5D report to the Department of Health. 2011.

Tosh JC, Longworth LJ, George E. Utility values in National Institute for Health and Clinical Excellence (NICE) Technology Appraisals. *Value in Health* 2011;14(1):102-9.

Longworth L, Rowen D. DSU Technical Support Document 10:The use of mapping mehtods to estimate health state utlity values. Available from http://www.nicedsu.org.uk

McCabe CJ, Stevens KJ, Brazier JE. Utility scores for the Health Utilities Index Mark 2: an empirical assessment of alternative mapping functions. Med Care 2005 Jun;43(6):627-35.

National Institute of Health and Clinical Excellence (NICE). Guide to the methods of technology appraisal. London: NICE; 2008.

Papaioannou D, Brazier J, Paisley S. (2011) NICE DSU Technical Support Document 9:The identification, review and synthesis of health state utility values from the literature. Available from http://www.nicedsu.org.uk

Pickard AS, De leon MCV, Kohlmann T et al. Psychometric comparison of the standard EQ-5D to a 5 level version in cacner patients. Med care 2007; 45(3):259-263.

Salomon, J.A. 2003. Reconsidering the use of rankings in the valuation of health states: a model for estimating cardinal values from ordinal data. *Popul.Health Metr.*, 1, (1) 12 available from: PM:14687419

Stevens K. Developing a descriptive system for a new preference-based measure of health-related quality of life for children. Qual Life Res 2009 Oct;18(8):1105-13.

Wailoo A, Davis S, Tosh. The incorporation of health benefits in CUA using the EQ-5D. NICE DRU Report, 2010. http://www.nicedsu.org.uk/PDFs%20of%20reports/DSU%20EQ5D%20final%20report%20-%20submitted.pdf

5 Author/s

Prepared John Brazier (Health Economics and Decision Science, School of Health and Related Research, University of Sheffield) on behalf of the Institute's Decision Support Unit, October 2011.

6 Acknowledgements

The author is grateful for comments on earlier drafts from Meindert Boysen, Carole Longson, Louise Longworth, Donna Rowen, Andrew Stevens, Paul Tappenden and Allan Wailoo.

NATIONAL INSTITUTE FOR HEALTH AND CLINICAL EXCELLENCE

Report to the Methods Review Working Party

Key issues arising from workshop on measuring and valuing health effects

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

There was no indication during the workshop discussions that it is necessary to deviate from the EQ-5D as the preferred utility measure in the reference case. A few participants thought that there should be no exceptions to the use of the EQ-5D, but the majority agreed that there are circumstances where other generic measures or mapping of condition-specific measures could be used. The following topics were addressed:

- **When is EQ-5D not appropriate?** Participants felt that the current Methods Guide does not adequately describe when the EQ-5D is not appropriate, and what criteria determine which preferred measure should

be presented instead. Where alternative measures are presented, there was a preference for these to be presented alongside the EQ-5D wherever possible, and also for any alternative measure to use preference-based valuation methods, ideally the time-trade-off method. It was generally agreed that the EQ-5D was not appropriate in evaluations of treatments for children, or treatment for hearing, vision or mental health conditions, and that more guidance should be included in the Methods Guide on this.

- **How to identify the most appropriate measures (EQ5D or other):** Participants did not conclude how utility values should be identified. It was stated that ideally these should be identified in the same way as clinical data, but that this was not considered to be practical or do-able. There was no consensus on how systematic/exhaustive searches for utility values should be. However, there was overall agreement that searches should be transparent, explicit and reproducible.

  It was generally agreed that trial-based utility values should not be used in isolation, and that evidence providers should attempt to validate these values with values identified in the literature, explore any differences, and address issues related to extrapolation and generalisability.

  Participants agreed that no established quality criteria currently exist to choose most appropriate utility values. Participants agreed that synthesising or pooling utility values from different sources was not appropriate, and that the most important criterion for selecting a utility value is the relevance to the health state being modelled. It was generally agreed that utility values that are selected are fully justified, and that any uncertainty should be explored in sensitivity analysis.

- **Mapping**: Participants agreed that the mapping function chosen should be fully described, its choice justified, and it should be adequately demonstrated how well the function fits the data. Sensitivity analyses should be presented if there are several algorithms in the literature but no clear preference over which algorithm should be used, and to explore any uncertainties in the mapping algorithm.

During the workshop discussions it was frequently suggested that the utility values for a particular appraisal could be determined during the scoping stage between the evidence provider and NICE. Furthermore, it was suggested that NICE could set up a publicly available archive of previously used utility values, exceptions to the EQ-5D accepted by the Appraisal Committee, and what alternatives, including specific mapping algorithms, have been accepted by the Appraisal Committee in which circumstances.

The following 3 additional topics were also addressed:

- **EQ-5D 5L:** Participants agreed that NICE could not currently use the EQ-5D-5L as reference case because it is yet not fully evaluated. However, participants felt that NICE needs to make statements about the usefulness of measuring quality of life in parallel with both the 3L and the 5L version. This was because participants were concerned about any problems with the transition from the 3L to the 5L version bearing in mind the time it would take to collect EQ-5D data before a NICE appraisal.

- **Adjustments to utility values:** Participants agreed that in some circumstances adjustment, for example for age or co-morbidities, to utility values could be required in order to provide unbiased estimates of health effects. However, it was stated that the appropriate method by which to undertake such adjustment, and the impact of it, will vary by the context and the technology. Also, because of the current ongoing methodological debate about adjustments, participants felt that the Methods Guide should not be too prescriptive in respect of utility adjustments, but expressed the importance of a pragmatic approach tailored to the data used in each appraisal.

- **Quality of life benefits for people other than patients and carers:** Although the reference case stipulates 'all health effects on individuals' the text in section 5.2.7 mentions *'patients or, when relevant, other people (principally carers)'*. There was no consensus at the workshop about the appropriateness of this wording. Some participants thought that only health effects on patients should be taken into account, some thought that not

only carers but also significant others (e.g. children, grandparents, parents, foster parents, brothers, sisters, companions, dependents), and some thought that section 5.2.7 was appropriate, as carers are performing a different role to that of significant others, i.e. one which in some cases would have to be provided by the NHS or Social services. Participants expressed concern that including significant others would favour conditions affecting people of child bearing age who are more likely to have dependents. Participants, however, agreed that if any health benefits beyond patients are included, it has to be based on empirical data, rather than claims only. Participants also felt that there were methodological problems to address with aligning tools to measure carer quality of life with the EQ-5D. As far as non-health quality of life benefits are concerned, participants agreed that this depends on the final decision on the perspective and needs to be consistent across all individuals whose health effects are included.

There was much discussion about of the **DSU's technical support documents** (TSDs) and the need to clarify the role of these in relation to the Methods Guide, whether they should be referenced or the technical advice embedded in the Methods Guide. Some Participants raised the concern that these documents have not been consulted upon. The specific TSDs mentioned were

- TSD 9 ('The identification, review and synthesis of health state utility values from the literature')
- TSD 11 ('Alternatives to EQ-5D for generating health state utility values')
- TSD 12 ('The use of health state utility values in decision models')

2 Questions posed to the workshop participants

1. Should the EQ-5D be the reference case measure across all patient populations (e.g. adults, children) and diseases? Are there any exceptions in which an alternative measure (e.g. the SF-6D or the HUI-3) should be used? If such exceptions arise, which alternative measures should be recommended? How should NICE's Appraisal Committee deal with appraisals whereby the available reference case utility values are not considered plausible?

2. How, and when, should NICE adopt the new EQ-5D-5L tariff? Should NICE follow the lead of the Euroqol group or should it set its own agenda?

3. When, if ever, should mapping (cross-walking) be preferred over the direct valuation of health (e.g. using the EQ-5D)? Can a consistent set of criteria be set out to define such circumstances? What are these criteria? Should such analyses be presented as the base case analysis or as a secondary analysis? How can the Appraisal Committee ensure that the mapping is reasonable?

4. How should potentially relevant utility values be identified? Is a systematic review of utility studies necessary? How should appropriate utility values be selected? How should values be synthesised across studies such that the uncertainty is adequately reflected)?

5. Should models reflect changing utility over time (for example, between disease progression and death), and if so how? Should utility values for health states be adjusted for age and/or sex when incorporated into economic models?

6. How should health effects on people other than the recipient of the intervention (e.g. parents, carers) be defined, measured and valued within technology appraisals?

3 Summary of the workshop discussions

3.1 The EQ-5D as the reference case

The current Methods Guide states in section 5.4.1 that '*The EQ-5D is the preferred measure of HRQL in adults. The methods to elicit EQ-5D utility values should be fully described. When EQ-5D data are not available or are inappropriate for the condition or effects of treatment, the valuation methods should be fully described and comparable to those used for the EQ-5D.'* It also states a preference for the *EQ-5D* in section 5.4.4. Section 5.4.9 and 5.4.10 explain in more detail what is required if the EQ-5D is not considered appropriate (see full text in Appendix 1).

3.1.1 *Should the EQ-5D be the reference case measure across all patient populations (e.g. adults, children) and diseases?*

In general, participants felt that the EQ-5D is an appropriate reference case because it is a standardized tool, based on rigorous research, the most widely used measure, and because it works in most cases. However, views differed on the extent to which in exceptional cases other measures can be used instead or alongside the EQ-5D. Some participants were of the view that no exceptions should be permitted because of the need for comparability and consistency. However, the majority of participants agreed that exceptions to the reference case would be appropriate if there is well substantiated evidence that the EQ-5D is not suitable for the particular patient population or disease.

3.1.2 *What could be exceptional circumstances?*

In general, it was felt that the Methods Guide needs to be very explicit about which exceptions are allowed and to guide the manufacturer on how to demonstrate which situations may be exceptions.

Some participants felt that it would only be appropriate to use another measure if evidence from literature (academic publications, rather than unpublished analyses presented by the manufacturer) demonstrates that the

EQ-5D should not be used in specific circumstances. Some participants suggested that the EQ-5D should always be used and if a manufacturer wanted to present disease specific or other measures, these should only be included in sensitivity analysis. This is because the EQ-5D was considered to be a reference standard – and that it is not possible to assess alternative utility values if there is nothing to compare them with.

Some participants thought that the EQ-5D is generally not appropriate for children and that the Methods Guide should provide more guidance. Participants were aware of the children's version of EQ-5D descriptions (EQ-5DY) but that this was not valid for children under 5 years of age. Participants felt that there is the need for research on whether valuations obtained from adults are applicable to children. Some participants questioned whether children value health differently to adults, but then generally agreed that it is appropriate for the valuation to be carried out by adults, as these are the tax-paying general population.

Participants stated that two important areas of health-related quality of life which are not captured with the EQ-5D were hearing and vision because loss of these two senses may matter to the patient in more ways than affecting mobility, usual activity, and anxiety and depression. It was suggested that the HUI may be better in these circumstances, but by allowing other measures could give advantages to specific diseases (if the instruments are chosen that detect the largest differences). It would therefore be important to use EQ-5D alongside disease specific questionnaire.

Other specific examples where participants felt the EQ-5D might not work were:

- Mental health – potential problems with patients self-reporting
- Co-morbidities - EQ5D not considered valid as no able pick up differences
- Quality of life in diagnostics

- Ultra orphan diseases – lack of data, epidemiology very poor.

3.1.3 *Which alternative measures should be recommended?*

Participants agreed that disease specific measures should be used when EQ-5D does not capture all relevant dimensions, but again, evidence that this is so would need to be provided.

Participants also felt that if other measures are presented alongside the EQ-5D then the valuation method of this alternative descriptive system needed to be comparable to how it is done with the EQ-5D, i.e. using TTO. Also, any direct valuation of health states could also be used using TTO, but in this case the disease is known and some diseases carry more emotional weight than others, which may change the way patients rank or value health. With generic descriptions this does not happen.

Participants noted that the EQ5D valuation maybe out of date (1996) with societal preferences having moved on. Also, some participants felt that it was important to mention that EQ-5D is not aimed to measure functioning unlike some disease specific measure and that it this needs to be borne in mind to avoid confusing the sensitivity of the EQ-5D in detecting functional impairments and in detecting changes in HRQoL.

Participants were aware that for vision and hearing ‘bolt ons’ to the EQ-5D may be an option but that more research is needed on this methodology and the consequences of its use.

Some participants suggested a list on NICE’s website with the exceptions accepted by the Appraisal Committee – this list could be updated as more evidence becomes available. Others suggested a discussion about alternatives to be included at the scoping stage.

Participants felt that the DSU’s Technical Support Document (TSD) 11 (‘Alternatives to EQ-5D for generating health state utility values’) provides a good basis for this and should be referenced in the methods guide.

3.1.4 *How should NICE's Appraisal Committee deal with appraisals whereby the available reference case utility values are not considered plausible?*

If the available reference case utility values are not considered plausible, participants suggested that the Committee need to use common sense in considering utility values obtained, taking into account the accumulated experience of previous appraisals. Also, exploring the importance of the utility values for the cost effectiveness through sensitivity analysis was considered appropriate.

Participants felt that if significant difference between values obtained from alternative or condition-specific measures and from EQ-5D were found that more explorations of the reasons are needed. This may involve requesting more information from the manufacturer. It may also be necessary for the Committee use judgment and deliberation, but that this needs to be explained well.

3.2 How, and when, should NICE adopt the new EQ-5D-5L tariff? Should NICE follow the lead of the EuroQol group or should it set its own agenda?

Participants agreed thought that NICE could not recommend the EQ-5D-5L as the reference case at this moment in time, but expressed the view that NICE could not ignore the EQ-5D-5L, and that some guidance was required. Specifically, some groups thought that if NICE had specific views on how the valuation of the EQ-5D-5L should be undertaken, it should request such research proactively. Concerns were raised that NICE's view on the EQ-5D-5L could have implications for the continued development of the EQ-5D-5L. Many groups confirmed that a signal of support from NICE could aid its development internationally, and support the development of more evidence.

Participants suggested that before NICE can consider the incorporation of the *EQ-5D-5L* as the reference case, evidence needs to be available

- about the sensitivity of the EQ-5D-5L,
- that people can appropriately differentiate between the levels,
- to map from EQ-5D-5L to the EQ-5D-3L values, and
- that addresses the issue that the 55555 state is unlikely to be equal to the 33333 due to changes in the method of valuing states worse than dead.

Many groups discussed the time lags between the EQ-5D-5L validation by the EuroQol group, generation of data and the adoption by NICE, and agreed that industry needs guidance now for products that may be appraised in several years. There were concerns that 'parallel running' of EQ-5D-3L and EQ-5D-5L in trials could lead to gaming. However some groups thought that data produced in parallel would be informative.

Participants agreed that any EQ-5D-5L generated should be presented in a sensitivity/secondary analysis, with EQ-5D-3L always being presented as base case.

There was some discussion about the valuation methods, which EuroQol may use, but there was no clear consensus about this. Some groups voiced concerns about the TTO method, and others thought that discrete choice evaluation was a promising methodology for this new valuation.

3.3 Mapping

3.3.1 *When, if ever, should mapping (cross-walking) be preferred over the direct valuation of health (e.g. using the EQ-5D)? Should such analyses be presented as the base case analysis or as a secondary analysis?*

The current Methods Guide states in section 5.4.6 that '…. *When EQ-5D data are not available, methods can be used to estimate EQ-5D utility data by mapping (also known as 'cross-walking') EQ-5D utility data from other HRQL measures included in the relevant clinical trial(s). This can be done if an adequate mapping function can be demonstrated and validated. Mapping*

should be based on empirical data and the statistical properties of the mapping function should be clearly described.'

In general, workshop participants considered that mapping should only be undertaken in exceptional circumstances such as when:

- EQ-5D is not suitable
- The study from which the health-related quality of life measures were derived is poorly designed or has a very small population (e.g. for rare diseases)
- A specific quality of life measure is collected in the trial at time points which cannot reasonably inform the model

Some participants noted that it would be preferable to use an existing mapping algorithm (if it has been appropriately validated and is still up to date). However, participants also expressed concern that often manufacturers use algorithms which have been previously considered and accepted by Committee, even if they are now largely out of date, or contain an error. In light of this, participants agreed that the mapping function chosen should be fully described, its choice justified, and it should be adequately demonstrated how well the function fits the data. Any other available mapping functions should also be included as secondary analyses. Uncertainty around the mapping function used should also be clearly described and tested in sensitivity analyses.

Some participants debated whether mapping to the utility values or to the dimensions (response mapping) is more appropriate. It was noted that response mapping is not widely undertaken in the UK at present; however it does have advantages in being able to include a comparison of patients across diverse instruments and in having flexibility in the degree of precision desired.

3.3.2 *Can a consistent set of criteria be set out to define such circumstances? What are these criteria? How can the Appraisal Committee ensure that the mapping is reasonable?*

Participants considered that if mapping is required, there are likely to be significant variations in the methodology used unless explicit instructions are provided to the manufacturer. Participants stated that the current Methods Guide does not adequately describe when the EQ-5D is not an appropriate measure and what criteria determine which preferred measure should be presented instead. They suggested that a hyperlink to the DSU's TSD 12 ('The use of health state utility values in decision models') should be embedded in the methods guide, and a summary of key points from the TSD should also be presented, to help readers understand how to determine which method to use if EQ-5D is not available/not appropriate.

Some participants suggested that the modelling developers should advise NICE during the scoping process whether utility values were directly collected in the key trials, or whether a mapping function will be used. NICE should then advise on the most appropriate approach to derive utility values. However, concern was expressed that some Assessment Teams do not have mapping specialists in their teams and therefore it would be challenging for them to determine the most appropriate mapping function to use.

Participants suggested that advisory meetings could be held to determine which mapping functions are most appropriate for each therapeutic area. Additionally, a systematic review of all published mapping functions should be undertaken. One participant cited work from Danny Frybach (using a US database of underlying health state measures) and suggested that NICE should review it and use it to inform which mapping algorithms should be considered if EQ-5D is not available.

Participants also highlighted the benefit of producing an archive (or publically available database) which contains all previous mapping functions used in technology appraisals, alongside a list of criteria to determine the most appropriate function to use for each therapeutic area. This would also ensure

a systematic collection of the strengths and weaknesses of previous mapping methods used, and assist the Committee with the decision about the appropriateness of the mapping function used in an appraisal.

Participants concluded that more explicit instructions should be included in the methods guide and in the NICE submission template to help justify the choice of mapping function and adequately explore and describe any uncertainties. In particular, it is important that the sample used to derive the health-related quality of life measures adequately matches the patient population under consideration in the appraisal (ideally all health states should come from the source data) and cover all disease states. In addition, any mapping algorithm should be externally validated (preferably on a separate patient sample) and results from model fit tests be provided.

3.4 How should potentially relevant utility values be identified?

The current Methods Guide states in section 5.4.11 that '*When health-related utility values have been obtained from the literature, the methods of identification of the data should be systematic and transparent. The justification for choosing a particular data set should be clearly explained. Health-related utility data that do not meet the criteria for the reference case should be accompanied by a carefully detailed account of the methods used to generate the data and a consideration of how these methods may affect the values. When more than one plausible set of h health-related utility data are available, a sensitivity analysis should be undertaken.'*

Participants discussed a number of different ways to identify utility values:

- **Searches**: There was no overall consensus on how systematic/exhaustive searches should be. Participants felt that given time constraints involved in the appraisal process, systematic literature search methods may not always be feasible. However, there was overall agreement that searches should be transparent, explicit and reproducible e.g. for the ERG/Assessment Group to re-run the search.

Participants could not find a consensus about how prescriptive NICE should be in defining a systematic search strategy and a selection process and whether evidence providers should be offered more explicit guidance on how systematic reviews should be conducted e.g. from Decision Support Unit technical support documents.

- **Clinical trials**: Often the manufacturer will decide a priori to collect utility scores within the clinical trial that is used as part of their submission. There was general agreement that such data would probably be the most useful in capturing HRQoL impact for the relevant population in the submission. However, there were issues about extrapolating utility values measured from a relatively short time period (within a clinical trial) over a longer term horizon, as required in an economic model. It was generally agreed that trial-based utility values should not be used in isolation and that evidence providers should attempt to validate these values with utility values identified in the literature and explore any significant differences. There was also concern on the generalisability of the patient population in the clinical trial to the UK NHS setting.

- **Electronic Databases**: Participancts suggested that existing databases e.g. PROMS may be used. However, there were concerns about how comprehensive such databases were in terms of the disease areas covered. Two concerns were raised against using PROMs databases: (1) that patient heterogeneity was masked due to the large patient numbers, and (2) that the narrow confidence intervals were not representing the true uncertainty around the estimates.

- **Expert elicitation:** Participants generally agreed that this should be done only if utility values are not identified in a literature search or collected within a trial.

- **Previously published Technology Appraisals**: As there are a growing number of appraisals which include utility values as part of the

economic model, one participant suggested that a database of utility values used in published appraisals should be set up by NICE.

- **Previously published systematic reviews:** Participants agreed that if there has been a recent, well-conducted systematic literature review already conducted either in a previous appraisal or journal article, it would be reasonable to identify relevant utility values from these sources. This would avoid duplication of effort especially given time constraints involved in the appraisal process. However, it was agreed that older reviews (e.g. > 5 years?) may be out-of-date and that using poorer quality reviews may lead to replication of errors. Therefore, some quality assessment of previously published reviews may be necessary.

3.4.1 *Quality criteria for the selection of utility values*

Participants agreed that no established quality criteria currently exists (e.g. quality checklists) in order to choose most appropriate utility values. There was some discussion of what such quality criteria should be, e.g. sample size, missing data, country, type of instrument, generalisability to the patient population in the UK NHS and internal validity.

However, relevance of the utility values to the health states that are being modelled was seen as the most important issue when selecting utility values. There was however, overall agreement that cherry-picking utility values (selection bias), with no explicit or transparent method of identification, should be avoided. It was generally agreed that utility values that are selected are fully justified and that any uncertainty should be explored (in sensitivity analysis)

3.4.2 *How should values be synthesized across studies such that the uncertainty is adequately reflected?*

Participants discussed if formal methods of data synthesis including pooled utility values should be used when there are sufficient number of homogenous utility values available (e.g. from same patient population and using the same

instrument). Some form of meta-regression may also be useful to explore any causes of variation between utility values. However, it was agreed that this will not always be feasible if the degree of heterogeneity is too high. It was suggested by some workshop participants that identifying the most relevant single utility value is more crucial than attempting to pool or synthesise multiple utility values when a large number of potentially relevant values are identified in the literature.

Participants acknowledged that ongoing MRC-funded research is in progress to explore evidence synthesis methods applied to utility values but does not report until the end of 2012.

It was suggested that it was more important to explore any uncertainty in the utility difference between relevant health states rather than the uncertainty in any baseline utility values.

Participants generally agreed that current methods to deal with uncertainty around utility values may be appropriate e.g. using alternative utility values if available. If alternative utility values are unavailable then threshold analyses may be appropriate i.e. varying the utility values between plausible ranges to explore how the ICERs are affected.

There was reference to the DSU's TSD 9 ('The identification, review and synthesis of health state utility values from the literature') but participants were uncertain on its role as guidance documents in relation to the NICE methods guide.

3.5 Should models reflect changing utility over time (for example, between disease progression and death), and if so how? Should utility values for health states be adjusted for age and/or sex when incorporated into economic models?

The current Methods Guide does not make reference in section 5.4 to any adjustments that may be necessary when incorporating utility values into economic models.

This topic was added after the development of the briefing paper for this workshop and therefore the briefing paper did not cover the topic of potential adjustments to utility values. This meant that the responses from the workshop participants to this question may not have been as focused as otherwise possible.

Participants acknowledged that there is ongoing debate in the health economics community about appropriateness of adjusting utility values. The overall consensus was that a pragmatic approach should be taken and the Methods Guide should incorporate flexibility to allow the most appropriate approach tailored to the individual appraisal.

Participants focussed their discussions on adjustments for varying utility over time, age and sex. Generally, the majority of participants expressed the view that it was appropriate to vary utility over time in the modelling of health benefits and that this is already the current practice within NICE appraisals. However, some participants felt that such an adjustment was rare and should not normally happen in NICE appraisals.

It was commonly accepted that age adjustment is in fact a proxy to adjust for the average increased comorbidity and decline of function (e.g. with respect to eyesight or hearing), which generally occurs as people age. However, some delegates explained that often it is difficult to separate the effect of disease progression from disease-unrelated comorbidities, and that the impact of this can vary according to the disease area. Also, participants discussed that multiple related comorbidities may have a lesser impact on quality of life than two completely unrelated comorbidities.

Participants expressed the opinion that ideally models should contain data from a representative mix of people of all ages and disease severity as appropriate. Where this is available, directly observed utility values should be used instead of age-adjusted or comorbidity-adjusted values, but it was accepted that in most cases such data are not available. Most participants expressed the view that in this situation, the utility values should be adjusted for age in order to describe plausible health gains. Moreover, some participants expressed the view that face validity of models could be compromised if adjustment is not performed, particularly when modelling over an extensive period of time.

Some participants noted that the Methods Guide currently stipulates that future costs that are considered to be unrelated to the condition or technology of interest should be excluded from the evaluation. These participants therefore questioned whether the adjustment of utility values for future unrelated comorbidities might imply a different perspective for health effects compared to costs. However, other participants suggested that the inclusion of changes in health unrelated to the condition or technology of interest was required in order to estimate overall health gains from the technology of interest and that this did not constitute an inconsistency in terms of perspective.

Some participants were concerned that adjustment for age could be interpreted as indirect age discrimination. Other participants expressed the view that age-adjustment was not discriminatory because it captured the average increase in comorbidity as people get older, and that adjusting was the correct methodological approach to capture benefit over time.

Participants also discussed adjustment for sex and stated that the difference in life expectancy between men and women could affect accumulated life years, and subsequently any discounting of QALYs could therefore have a differential impact on women and men. Furthermore there could be a difference in the natural history of illness between men and women and this could also cause a differential effect. To mitigate this, participants felt that gender-specific utility data would ideally be used, however conceded that data

on both condition-specific and health-specific utility values were unlikely to be available for most appraisals.

In summary, participants from more than one group expressed the importance of a pragmatic approach to this problem and that in some circumstances adjustment could be required in order to provide unbiased estimates of health effects. Participants felt that the Methods Guide should not be too prescriptive in respect of utility adjustment. It was stated that the appropriate method by which to undertake utility adjustment, and the impact of any adjustment, will vary by the context and the technology. There was also discussion about the unresolved issue of the implementation of various methods, including methods that are multiplicative, additive, mixed and non-linear. Therefore, participants felt that the best option would be to take an approach that does not restrict the appraisal to a particular method of adjustment. Many participants stated the need for presentation of ICERs including both adjusted and unadjusted utility values in the economic model. This would illustrate the impact of the adjustment on the estimated health effects, and would be valuable in cases where the Appraisal Committee deemed that the method of adjustment used was inappropriate.

Some delegates suggested that the Methods Guide could make reference to the DSU's TSD 12 ('The use of health state utility values in decision models') that provides more information on appropriate methods. However, other participants noted that the TSDs are not currently put out for public consultation, unlike the Methods Guide.

3.6 Should the impact on significant others be broadened out to include other members of the family who are not directly involved in care

The current Methods Guide states that '*all health effects on individuals' should be taken into account* (table 5.1; page 30) and section 5.2.7 specifies that '*the perspective on outcomes should be all direct health effects, whether for patients or, when relevant, other people (principally carers).*'

Participants at the workshop generally felt that it was appropriate to consider benefits both to carers and to significant others, but had a number of concerns which may be difficult to overcome.

Participants preferred that benefits to significant others should be included as part of the deliberative process, rather than including these benefits formally in the economic model, and that if it is done it should be based on empirical evidence. This was because of the limitations to current methodology for including benefits to carers and significant others in economic evaluations and difficulties in data collection. In general, participants were unclear about the likely effect on cost effectiveness of including such data and whether the benefits of including such data would be outweighed by the effort to collect it.

Participants raised concerns about where the boundary of significant others would be drawn (e.g. children, grandparents, parents, foster parents, siblings, co-habitants, dependents), and that without clear boundaries this could lead to great inconsistencies if some evidence provides include more ‘beneficiaries’ than others.

Participants considered that including health benefits to significant others raised equalities issues in that such an approach could favour some conditions over others such as those affecting people of child bearing age who are more likely to have a number of dependents. It was noted by one PCT attendee that factors such as family status were not considered in individual funding requests for equalities reasons.

These concerns were felt to be particularly important in the context of the introduction of value based pricing where the price of a product could be influenced by these factors, thereby potentially incentivising evidence providers to identify and incorporate such benefits. There was one suggestion that it could be left for the manufacturers/sponsors to submit such evidence if they wished. Other participants, however, felt that with the introduction of value based pricing there was a need for consistency and more specific guidance from NICE.

Some participants suggested that if benefits to carers and other significant others were included, this would affect opportunity cost and would affect the cost-effectiveness threshold.

Participants noted that if evidence was available for health effects on other people but the patient that could be included in an economic evaluation then this may be most appropriately submitted as a sensitivity analysis to the reference case.

There was no consensus about the appropriateness of the wording in the current methods guide. Specific points raised were:

- the current reference case allowing the inclusion of carer benefits in economic modelling is not appropriate, and that the economic modelling should focus on the benefits to the patient.
- the current methods guide is inconsistent and that if NICE accepts carer benefits then for consistency, it should be all significant others and not limited to those of carers.
- the current methods guide is appropriate, as carers are performing a different role to that of the non-caring wider family, one which in some cases would have to be provided by the NHS or Social services if it didn't exist.

3.6.1 *Extend benefits to non-health-related quality of life*

Participants also discussed if only health effects or non-health-related quality of life effects should be taken into account, and noted that this question was related to the decision about the most appropriate perspective, and whether the perspective should remain that of the NHS or be extended to a wide societal perspective. Participants agreed that if non-health-related quality of life was included for carers and significant others, then this should also be done for patients. Likewise if only health-related benefits were considered for patients, then this should also be done for carers and significant others.

3.6.2 *Aggregation of impacts on carers and significant others*

Participants also discussed how any impacts on carers and significant others should be aggregated and agreed that there would be challenges in collecting and analysing data from carers and significant others. For example to enable collection of data from carers or significant others in a trial, such people would need to formally consent to, and be enrolled in, the trial. This would add additional administration and cost to trials.

If only health benefits were included, some concern was raised about whether the EQ-5D would be sufficiently sensitive to identify the impact on carers and significant others. Further it was noted that currently care-specific measures include a more general focus on quality of life rather than health related quality of life and are not anchored to 0 and 1 in the same way as the EQ-5D. Participants considered that it would not be possible to aggregate different measures for example EQ-5D for patients and CarerQOL for carers. Participants also questioned whether it is appropriate to assume an equal weight for benefits to carers and significant others compared to benefits to patients. For all these reasons, before data for carers or significant others were to be formally considered, there is a need for methodological research for this to be done appropriately.

Participants felt that these methodological issues meant that currently it would not be appropriate for NICE to make quantification of benefits to carers and significant others a requirement. Instead, the consensus was that that the Appraisal Committee should deliberate these benefits.

4 Key issues for consideration by Working party

1. Should the Methods Guide be more descriptive than in the current sections 5.4.9 and 10 about the circumstances in which utility measures other than EQ-5D are acceptable?

2. If so,

    a) What are these circumstances?

b) What supporting evidence needs to be provided?

3. Should decisions about alternative utility measures be made at the scoping stage of an appraisal?

4. Should decisions about alternative utility measures be made be based on ‘case law’ developed through previous appraisals?

5. Should the EQ-5D always be presented alongside any alternative measures?

6. Should NICE encourage the parallel use of EQ-5D 5L?

7. If so which data should be used for decision making in an appraisal where both sets of data are available?

8. Should more information on mapping than in the current section 5.4.6 be included in the Methods Guide?

9. Should the Methods Guide be more explicit about when utility values need adjusting, for example for age, and if so, how to carry out such adjustments?

10. Should in the definition of who benefits (‘all health effects on individuals’) be changed to be more specific about who those individuals are?

11. Should there be a difference between carers and significant others?

12. Should the only health benefits for people other than patients be included?

13. Should the DSU’s TSDs be included/ embedded/ referenced in the Methods Guide?

5 Authors

Prepared by Elisabeth George on the basis of workshop feedback from Moni Choudhury, Matthew Dyer, Rita Faria, Zoe Garrett, Susan Griffin, Martin

Hoyle, Pall Jonsson, Brendan Mulhern, Fiona Rinaldi, Marta Soares, Jon Tosh and Helen Tucker, whose contributions are gratefully acknowledged.

January 2012

6 Appendix 1

Extract from the Current Methods Guide

Section 5.4 Measuring and valuing health effects

5.4.1 For cost-effectiveness analysis, the value of health effects should be expressed in terms of QALYs for the appropriate time horizon. For the reference case, the measurement of changes in HRQL should be reported directly from patients and the value of changes in patients' HRQL (that is, utilities) should be based on public preferences using a choice-based method. The EQ-5D is the preferred measure of HRQL in adults. The methods to elicit EQ-5D utility values should be fully described. When EQ-5D data are not available or are inappropriate for the condition or effects of treatment, the valuation methods should be fully described and comparable to those used for the EQ-5D. Data collected using condition-specific, preference-based measures may be presented in separate analyses. The use of utility estimates from published literature must be supported by evidence that demonstrates that they have been identified and selected systematically.

5.4.2 The QALY is a measure of a person's length of life weighted by a valuation of their HRQL over that period. The HRQL 'weighting' usually comprises two elements: the description of changes in HRQL itself and a valuation of that description of HRQL. Information on changes in HRQL as a result of treatment should be reported directly by patients (and directly by carers when the impact of treatment on the carer's health is also important). The valuation of changes in HRQL reported by patients should be based on public preferences elicited using a choice-based method in a representative sample of the UK population.

5.4.3 When it is not possible to obtain information on changes in patients' HRQL directly from patients, then data should be obtained from their carer (not from healthcare professionals). The valuation of changes in

HRQL measured in patients (or carers) should be based on a valuation of public preferences from a representative sample of the UK population.

5.4.4 To quantify the effects of technologies on HRQL, the EQ-5D (a standardised and validated generic instrument) is preferred. Different classification systems produce different utility values; therefore, results from the use of different systems cannot always be compared. Given the comparative nature of the Institute's work and the need for consistency across appraisals, a single classification system, the EQ-5D, is preferred for the measurement and valuation of HRQL.

5.4.5 The EQ-5D is a widely used measure of HRQL and has been validated in many different patient populations. The EQ-5D comprises five dimensions of health: mobility, ability to self-care, ability to undertake usual activities, pain and discomfort, and anxiety and depression. The system has been designed so that people can describe their own HRQL using a standardised descriptive system. A set of preference values elicited from a large UK population study using a choice-based method of valuation (the time trade-off method) is available for the EQ-5D classification system. This set of values can be applied to people's self-reported descriptions of their HRQL to generate health-related utility values.

5.4.6 Data using the EQ-5D instrument may not always be available. When EQ-5D data are not available, methods can be used to estimate EQ-5D utility data by mapping (also known as 'cross-walking') EQ-5D utility data from other HRQL measures included in the relevant clinical trial(s). This can be done if an adequate mapping function can be demonstrated and validated. Mapping should be based on empirical data and the statistical properties of the mapping function should be clearly described.

5.4.7 When EQ-5D utility data are not available, direct valuations of descriptions of health states based on standardised and validated HRQL measures included in the relevant clinical trial(s) may be submitted. In these cases, the valuation of descriptions should use the time trade-off

method in a representative sample of the UK population, with ‘full health’ as the upper anchor, to retain methodological consistency with the methods used to value the EQ-5D.

5.4.8 Data that have been collected directly in relevant clinical trials using condition-specific, preference-based measures should be presented in a separate economic analysis.

5.4.9 The EQ-5D may not be an appropriate measure of health-related utility in all circumstances. If the EQ-5D is considered inappropriate, empirical evidence should be provided on why the properties of the EQ-5D are not suitable for the particular patient population. These properties may include the content validity, construct validity, responsiveness and reliability of EQ-5D. When an alternative measure is preferred, those submitting analysis should provide reasons, supported by empirical data on the properties of the instrument used. They should also indicate any evidence that will help the Committee understand to what extent their choice of instrument has impacted on the valuation of the QALYs gained. If direct valuations of descriptions of health states based on HRQL measures other than the EQ-5D are used, the valuation methods must be comparable to those used for the EQ-5D (see section 5.4.5).

5.4.10 It is recognised that the current version of the EQ-5D has not been designed for use in children. When necessary, consideration should be given to alternative standardised and validated preference-based measures of HRQL, such as the Health Utility Index 2 (HUI 2), that have been designed specifically for use in children.

5.4.11 When health-related utility values have been obtained from the literature, the methods of identification of the data should be systematic and transparent. The justification for choosing a particular data set should be clearly explained. Health-related utility data that do not meet the criteria for the reference case should be accompanied by a carefully detailed account of the methods used to generate the data and a consideration of how these methods may affect the values. When more

than one plausible set of h health-related utility data are available, a sensitivity analysis should be undertaken.

NATIONAL INSTITUTE FOR HEALTH AND CLINICAL EXCELLENCE

Briefing paper for methods review working party on mixed treatment comparisons

The briefing paper is intended to provide a brief summary of the issues that are proposed for discussion by the Methods Review Working Party 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.

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 revised versions were published in 2004 and 2008. 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 revised draft of the Methods Guide will be available for a 3-month public consultation, expected to begin in xxx. We encourage all interested parties to take part in this consultation.

2 Background

2.1 Relevance of topic to NICE technology appraisals

The quantity and nature of clinical evidence submitted for technology appraisals varies considerably. Commonly there may be one or two directly relevant head-to-head trials which compare an intervention of interest with a comparator of interest, but evidence to draw comparisons across the full range of treatment options specified as comparators in the scope is lacking. In such situations, it is also common for there to exist a number of indirectly relevant trials in which the intervention(s) of interest, or the comparator(s) of interest, are compared with other treatments which may or may not be within the appraisal scope. The use of mixed treatment comparisons to synthesise such evidence is becoming increasingly used for NICE technology appraisals. This may be the consequence of a number of factors including a lack of direct head-to-head trials of all relevant decision alternatives, increased awareness of indirect methods, developing methodology as well as the direction of the 2008 Methods Guide. Where such approaches are employed, it is essential that the scope and methods of evidence synthesis are appropriate, robust and transparent for NICE's Appraisal Committees.

As with any pooling of studies, it is crucial that there can be confidence that the trial populations and methods are comparable and that decisions about trial inclusion into the network are both unbiased and transparent. However, it is very rare for manufacturers' submissions to present a full critical appraisal of the mixed treatment comparison which includes full details of how the mixed treatment comparison has been constructed and full details of the trials and participants included in the mixed treatment comparison. In addition, the network of trials can often be very large, which from a practical viewpoint, can result in problems for the Evidence Review Groups and Assessment Groups in systematically reviewing and appraising mixed treatment comparisons. In addition, manufacturers' submissions rarely present a full examination of the effects of individual trials on the results of the mixed treatment comparison

and sensitivity analyses exploring the inclusion and exclusion of key trials are rarely submitted.

A critical appraisal checklist has recently been developed as part of the DSU series of Technical Support Documents on evidence synthesis methods (see Appendix 1).[1-7] This checklist[7] covers a number of pertinent synthesis issues including the scope of the analysis, the search strategy used to identify relevant trials for inclusion in the analysis, the definition of the interventions, the choice of outcome measure(s), the presentation of data, statistical methods employed, software considerations, issues surrounding inconsistency, and the use of the analysis within economic decision models.

2.2 Introduction to mixed treatment comparisons

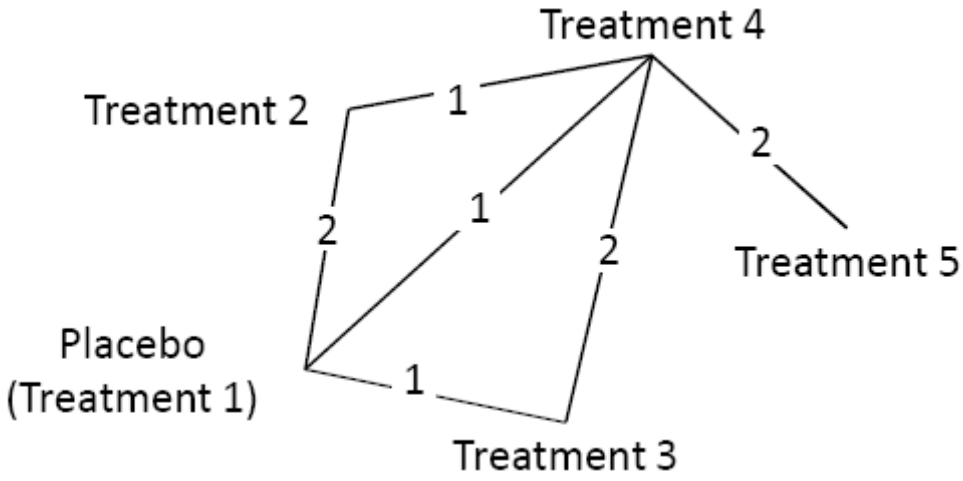
Frequently, and particularly in the case of newly licensed technologies, there are very few head-to-head randomised controlled trials that directly compare the intervention of interest (that is, the new technology) with the comparator of interest (that is, routine and standard practice in the NHS). It is therefore common in practice to see indirect comparisons or mixed treatment comparisons conducted in order to provide sufficient evidence on which to ascertain the relative effectiveness of the new technology compared with the comparator(s) of interest.

In order for a robust mixed treatment comparison to be possible, a number of conditions must be satisfied. Firstly, the trial populations must be truly comparable at baseline and secondly there must be comparable treatment circumstances. For example, if the trial populations or trial methodologies differ greatly, then it may be inappropriate to pool these studies. In addition, the whole trial network needs to be constructed in an unbiased manner (that is, the same inclusion and exclusion criteria are applied to all of the trials considered). These conditions are consistent with those that would be expected if conducting a more conventional head-to-head or ‘classical’ piecewise meta-analysis.

In order to explain the concept of indirect and mixed treatment comparisons, consider a scenario in which there are three technologies of interest, A, B and

C. Define the effect in a trial which compares A to B by d_{AB}. This is a *direct* estimate of AB. An *indirect* estimate of AB can also be obtained if there are AC an BC trials since $d_{BC} - d_{AC} = d_{AB}$. A mixed treatment comparison analysis (also known as network meta-analysis or mixed treatment meta-analysis) allows the synthesis of AB, AC, BC and ABC (i.e. three-arm) trials and estimates each pairwise treatment effect from both the direct and indirect evidence *without breaking randomisation*. Mixed treatment comparisons, are essentially extension of a traditional meta-analysis; these comparisons synthesise data from a series of trials allowing different comparisons to be made among the technologies of interest. Mixed treatment comparisons require a connected network; that is, for each treatment, there is a chain of pairwise comparisons that connects it to every other treatment. The construction of network diagrams can clearly describe the different possible evidence structures (see Figure 1). Within this form of network diagram, each edge represents a treatment; connecting lines indicate pairs of treatments which have been directly compared in randomised trials. The numbers on the lines indicate the numbers of trials making that comparison.

Figure 1 Example network diagram[1]



The methodology can be particularly useful when no, or little, direct head-to-head evidence exists on comparisons of interest. Also, when conducting a series of pairwise meta-analyses, it is difficult or impossible to rank all technology options in terms of effectiveness or cost-effectiveness. In contrast,

this is straightforward within a mixed treatment comparison and allows the estimation of the probability each technology is optimal across individual or multiple clinical endpoints.

It is important to recognise that mixed treatment comparison networks can become relatively large and complicated. This happens when there are a number of relevant interventions and comparators. The benefit of combining all of the direct and indirect evidence is that a decision is being made on all of the available evidence. However, there are a number of issues with mixed treatment comparisons that frequently arise within NICE technology appraisals. Firstly, the size of the network, that is, the number of trials and additional comparators included within the mixed treatment network, can become very large. In this situation, undertaking a comprehensive review and appraisal of the mixed treatment comparison can become cumbersome and time consuming. This can cause problems for Evidence Review Groups and Assessment Groups, especially in terms of checking that all the relevant studies have been included, that no inappropriate studies have been included and that the results of the analysis are both robust and reliable.[1]

In practice, previous mixed treatment comparisons from similar appraisals are often used as a basis for the network, into which additional trials are added. This may mitigate the intensity of the checking activity required if this ‘base network’ is considered reliable. However, checking of the original network, and any amendments made to it, must always be conducted. In addition, conducting an appraisal of a mixed treatment comparison can also include checking with experts in the field and comparator manufacturers, relevant stakeholders and conducting additional systematic reviews. From a practical point of view, these activities can be very resource intensive especially when the scope of network is large and complex and if the mixed treatment comparison has been submitted later on in the appraisal process, for example in response to an Appraisal Consultation Document.

A second issue is that the presentation of the mixed treatment comparisons usually does not facilitate understanding of the individual trials and of the trial participants and characteristics. Often, the descriptions of the individual trials

are limited and key differences between trials are not exposed. This means that it is often difficult to assess the face validity of the results from a mixed treatment comparison. The same criticism can also however be made with respect to classical piecewise meta-analysis which can also contain a large number of trials. As with any pooling of studies (such as in a conventional meta-analysis or indirect comparison), it is essential that the studies included are comparable in terms of design, participants, and other key factors. However, in mixed treatment comparisons, especially those with large networks, there are more trials and more decisions being made when constructing the network and therefore an increased possibility for trials that are not completely comparable to enter the network. In instances where the network (and hence individual trials) is poorly described, it can also be difficult to exactly ascertain how comparable the trials are and what the effect of this may be.

One example of when a mixed treatment comparison can be strongly affected by trial inclusion is when trials with non-comparable control groups are included in a network. For example, consider a Technology A that in trials appears slightly better than a placebo, but that the placebo arm in that trial also performed relatively well. Technology B in trials (the comparator to Technology A) appears to be much more effective than placebo, but the placebo arm in that trial has performed relatively poorly. In this situation, the relative effectiveness of Technology A compared with placebo is small and the relative effectiveness of Technology B and placebo is large. If these were combined in a mixed treatment comparison together with a number of other trials within the evidence network, it is possible that the results of the mixed treatment comparison would be misleading and the reasons for this inconsistency would be difficult to tease out if the network and the individual trials are poorly presented. In this situation one can, for example, question whether the scale of measurement (log ORs) is correct, or one can adjust for baseline risk if one believes that this has an impact of the *relative* effects.[5]

A further point, however, applies to pairwise comparisons or to cases where there is just one trial in the evidence base. Suppose that we only had the

Technology A vs placebo trial, and the target population was in fact the one that appears in the Technology B vs placebo trial, or vice versa. In both cases we would completely misjudge the efficacy of the active treatment. Or consider we had three A vs placebo trials, all with different baseline efficacies. Whilst these are clearly difficult situations for the interpretation of evidence, it would be a mistake to blame indirect comparisons as the root cause of the problem.[3]

A third issue is that often mixed treatment comparisons are presented as the reference case and little, or no, exploration of the suitability of the mixed treatment comparison is presented. In particular, the results of some mixed treatment comparison networks may be heavily influenced by one or two key trials and the inclusion and exclusion of these trials and the subsequent effect of this on the overall result is rarely presented clearly as sensitivity or scenario analyses to the Appraisal Committee. Particularly in the cases described above, whereby the inclusion of some trials may be open to question, it is important that the effect of these studies on the overall results is clearly presented.

In summary, there remains a need to undertake coherent analyses and further clarity about when this should include a mixed treatment comparison could be beneficial. In addition, there is an outstanding need for further direction on the reporting standards and appropriate sensitivity and scenario analyses that should be conducted when undertaking a mixed treatment comparison.

The decision support unit (DSU) have been commissioned to write a number of technical support documents (TSDs) that address many oif the points raised in this briefing paper. In particular TSD7 is a checklist for reviewers of mixed treatment comparisons and many consider that this would be of great value going forward and could have a role within the methods guide itself.

2.3 What the current Methods Guide advises with respect to mixed treatment comparisons

During the last review of the methods guide, the subject of mixed treatment comparisons was a central discussion point. As a result, the methods guide

includes a number of paragraphs (5.3.13 to 5.3.22) on mixed treatment comparisons. The methods guide states the following:

5.3.13 Data from head-to-head RCTs should be presented in the reference-case analysis, if available. When head-to-head RCTs exist, evidence from mixed treatment comparison analyses may be presented if it is considered to add information that is not available from the head-to-head comparison. This mixed treatment comparison must be fully described and presented as additional to the reference-case analysis (a 'mixed treatment comparison' includes trials that compare the interventions head-to-head and indirectly). When multiple technologies are being appraised that have not been compared within a single RCT, data from a series of pairwise head-to-head RCTs should be presented. Consideration should also be given to presenting a combined analysis using a mixed treatment comparison framework if it is considered to add information that is not available from the head-to-head comparison. If data from head-to-head RCTs are not available, indirect treatment comparison methods should be used (an 'indirect comparison' is a synthesis of data from a network of trials). The principles of good practice for standard meta-analyses should also be followed in mixed and indirect treatment comparisons.

5.3.14 The Institute has a preference for data from head-to-head RCTs and these should be presented in the reference-case analysis when available.

5.3.15 An 'indirect comparison' refers to the synthesis of data from trials in which the technologies of interest have not been compared in head-to-head trials, but have been compared indirectly using data from a network of trials that compare the technologies with other interventions. A 'mixed treatment comparison' refers to an analysis that includes trials that compare the interventions of interest head-to-head and trials that compare them indirectly. The principles of good practice for systematic reviews and meta-analyses should be

carefully followed when conducting mixed and indirect treatment comparisons. The rationale for the identification and selection of the RCTs should be explained, including the rationale for the selection of treatment comparisons that have been included. A clear description of the methods of synthesis is required. The methods and results of the individual trials should be documented. If there is doubt about the relevance of a particular trial, sensitivity analysis should also be presented in which these trials are excluded. The heterogeneity between results of pairwise comparisons and inconsistencies between the direct and indirect evidence on the technologies should be reported.

5.3.16 *There may be circumstances in which data from head-to-head RCTs are less than ideal (for example, the sample size may be small or there may be concerns about the external validity). In such cases additional evidence from mixed treatment comparisons can be considered. In these cases, mixed treatment comparisons should be presented separately from the reference-case analysis and a rationale for their inclusion provided. Again, the principles of good practice apply.*

5.3.17 *When multiple technologies are being appraised, data from RCTs (when available) that compare each of the technologies head-to-head should be presented in a series of pairwise comparisons. Consideration may be given to presenting an additional analysis using a mixed treatment comparison framework. In these situations, the Appraisal Committee will consider the results of both analyses with particular reference to the methods of synthesis and the appropriateness of the inclusion or exclusion of studies.*

5.3.18 *There may be situations when data from head-to-head RCTs of the technologies (and/or comparators) are not available. In these circumstances, indirect treatment comparison analyses should be considered.*

5.3.19 When evidence is combined using indirect or mixed treatment comparison frameworks, trial randomisation must be preserved. A comparison of the results from single treatment arms from different randomised trials is not acceptable unless the data are treated as observational and appropriate steps taken to adjust for possible bias and increased uncertainty.

5.3.20 Analyses using indirect or mixed treatment comparison frameworks may include comparator interventions (including placebo) that have not been defined in the scope of the appraisal if they are relevant to the development of the network of evidence. The rationale for the inclusion and exclusion of comparator interventions should be clearly reported. Again, the principles of good practice apply.

5.3.21 Evidence from a mixed treatment comparison may be presented in a variety of ways. The network of evidence may be presented in tabular form. It may also be presented diagrammatically as long as the direct and indirect treatment comparisons are clearly identified and the number of trials in each comparison is stated.

5.3.22 When sufficient relevant and valid data are not available for including in meta-analyses of head-to-head trials, or mixed or indirect comparisons, the analysis may have to be restricted to a qualitative overview that critically appraises individual studies and presents their results. In these circumstances, the Appraisal Committee will be particularly cautious when reviewing the results of analysis.

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 by the Methods Guide Review Working Party.

Currently indirect and mixed treatment comparisons are described in great detail in the methods guide. However, the consistency in submissions varies widely:

- Should further direction be given of the use of mixed treatment comparisons?
    - Is the current content in the methods guide regarding mixed treatment comparisons excessive?

What would be the impact of reducing the level of content on mixed treatment comparisons in the next methods guide?

- Should components of 'best practice' in conducting mixed treatment comparisons be more clearly specified?
    - As a minimum, should a full list of all trials included in the mixed treatment comparison, with baseline participant characteristics and key outcomes be provided?
    - Can any guidance on the size of networks by provided?

How should the technical support documents created by the decision support unit be incorporated into the Methods Guide?

- Should TSD 7 (checklist for reviewers) be recommended as a standard reference within the methods guide?

What are the potential consequences of requiring a full list of all trials (with participant characteristics and key outcomes)? Should the methods guide state how the information should be presented?

- Should guidance be provided on checking the face validity of a mixed treatment network (for example, contacting experts in the field, checking other appraisals in the same disease area)?

What would be the impact of providing instruction within the next methods guide on checking face validity?

- Should sensitivity and scenario analyses involving the mixed treatment comparison networks always be requested?

What would be the impact of always requesting sensitivity analyses? Should sensitivity analyses only be requested if there are inconsistencies within the mixed treatment comparison?

- Should potential inconsistencies within a mixed treatment comparison network always be formally explored?

What are the consequences of requesting formal exploration of inconsistencies within the method guide? Could specific methodology be referred to if this was included in the methods guide?

4 References

1. Dias, S., Welton, N.J., Sutton, A.J., Ades, A.E. (2011) NICE DSU Technical Support Document 1: Introduction to evidence synthesis for decision making. Available from http://www.nicedsu.org.uk
2. Dias, S., Welton, N.J., Sutton, A.J. & Ades, A.E. NICE DSU Technical Support Document 2: A Generalised Linear Modelling Framework for Pairwise and Network Meta-Analysis of Randomised Controlled Trials. 2011; last updated August 2011. Available from http://www.nicedsu.org.uk
3. Dias, S., Sutton, A.J., Welton, N.J., Ades, A.E. (2011) NICE DSU Technical Support Document 3: Heterogeneity: subgroups, meta-regression, bias and bias-adjustment. Available from http://www.nicedsu.org.uk
4. Dias, S., Welton, N.J., Sutton, A.J., Caldwell, D.M., Guobing, L. & Ades, A.E. NICE DSU Technical Support Document 4: Inconsistency in Networks of Evidence Based on Randomised Controlled Trials. 2011. Available from http://www.nicedsu.org.uk
5. Dias, S., Welton, N.J., Sutton, A.J., Ades, A.E. (2011) NICE DSU Technical Support Document 5: Evidence synthesis in the baseline natural history model. Available from http://www.nicedsu.org.uk

6. Dias, S., Sutton, A.J., Welton, N.J. & Ades, A.E. NICE DSU Technical Support Document 6: Embedding evidence synthesis in probabilistic cost-effectiveness analysis: software choices. 2011. Available from http://www.nicedsu.org.uk

7. Ades T, Caldwell TM, Reken S, Welton NJ, Sutton AJ, Dias S. NICE DSU Technical Support Document 7: Evidence synthesis of treatment efficacy in decision making: a reviewer's checklist. Available from http://www.nicedsu.org.uk

5 Author/s

This document was prepared by Rebecca Trowman, Andrew Stevens and Paul Tappenden. Thanks for Tony Ades and Alex Sutton for their helpful comments.

Appendix 1 Checklist Table. Abbreviations: Y/N/na Yes, No, Not Applicable; SA Sensitivity Analysis.

		Y/N/na	Comments, SA needed ?
A. DEFINITION OF THE DECISION PROBLEM ***A1. Target population for decision***			
A1.1	*Has the target patient population for decision been clearly defined?*		
A2. Comparators			
A2.1	*Decision Comparator Set: Have all the appropriate treatments in the decision been identified?*		
A2.2	*Synthesis Comparator Set: Are there additional treatments in the Synthesis Comparator Set, which are not in the Decision Comparator Set?*		
A3 Trial inclusion / exclusion			
A3.1	*Is the search strategy technically adequate?*		
A3.2	*Have all trials involving at least two of the treatments in the Synthesis Comparator Set been included?*		
A3.3	*Have all trials reporting relevant outcomes been included?*		
A3.4	*Have additional trials been included?*		
A4 Treatment Definition			
A4.1	*Are all the treatment options restricted to specific doses and co-treatments, or have different doses and co-treatments been "lumped" together?*		
A4.2	*Is a dose-response model fitted, or are the sub-components of the treatment modelled?*		
A5 Trial outcomes and scale of measurement chosen for the synthesis			
A5.1	*Where alternative outcomes are available, has the choice of outcome measure used in the synthesis been justified?*		
A5.2	*Have the assumptions behind the choice of scale been justified?*		
A6 Patient population: trials with patients outside the target population			
A6.1	*Do some trials include patients outside the target population?*		
A6.2	*What assumptions are made about the impact, or lack of impact this may have on the relative treatment effects?*		
A6.3	*Has an adjustment been made to account for these differences? If so, comment on the adequacy of the evidence presented in support of this adjustment, and on the need for a sensitivity analysis.*		
A7 Patient population: heterogeneity within the target population			
A7.1	*Has there been a review of the literature concerning potential modifiers of treatment effect?*		
A7.2	*Are there apparent or potential differences between trials in their patient populations, albeit within the*		

	target population?		
A8 Risk of Bias			
A8.1	*Is there a discussion of the biases to which these trials, or this ensemble of trials, are vulnerable?*		
A9. Presentation of the data			
A9.1	*Is there a clear table or diagram showing which data have been included in the base-case analysis?*		
A9.2	*Is there a clear table or diagram showing which data have been excluded and why?*		
B. METHODS OF ANALYSIS AND PRESENTATION OF RESULTS ***B1 Meta-analytic methods***			
B1.1	*Is the statistical model clearly described?*		
B1.2	*Has the software implementation been documented?*		
B2. Heterogeneity in the relative treatment effects			
B2.1	*Have numerical estimates been provided of the degree of heterogeneity in the relative treatment effects?*		
B2.2	*Has a justification been given for choice of random or fixed effect models? Should sensitivity analyses be considered?*		
B2.3	*Has there been adequate response to heterogeneity?*		
B2.4	*Does the extent of unexplained variation in relative treatment effects threaten the robustness of conclusions?*		
B3 Baseline model for trial outcomes			
B3.1	*Are baseline effects and relative effects estimated in the same model? If so, has this been justified?*		
B3.2	*Has the choice of studies to inform the baseline model been explained?*		
B3.3	*Has the statistical heterogeneity between baseline arms been discussed?*		
B4 Presentation of results of analyses of trial data			
B4.1	*Are the relative treatment effects (relative to a placebo or "standard" comparator) tabulated, alongside measures of between-study heterogeneity if a RE model is used?*		
B4.2	*Are the absolute effects on each treatment, as they are used in the CEA, reported?*		
B5 Synthesis in other parts of the natural history model			
B5.1	*Is the choice of data sources to inform the other parameters in the natural history model adequately described and justified?*		
B5.2	*In the natural history model, can all the differences between treatments be explained by their differences on randomised trial outcomes?*		

C. ISSUES SPECIFIC TO NETWORK SYNTHESIS ***C1 Adequacy of information on model specification and software implementation***			
C1.1	*Is the statistical model described, or was a citation for the statistical model given?*		
C1.2	*Is the source of the computer code used in the synthesis cited?*		
C1.3	*Is programming code for the synthesis provided?*		
C2. Multi-arm trials			
C2.1	*If there are multi-arm trials, have the correlations between the relative treatment effects been taken into account?*		
C3 Connected and disconnected networks			
C3.1	*Is the network of evidence based on randomised trials connected?*		
C4 Inconsistency			
C4.1	*How many inconsistencies could there be in the network?*		
C4. 2	*Are there any* a priori *reasons for concern that inconsistency might exist, due to systematic clinical differences between the patients in AB, AC, etc trials?*		
C4.3	*Have adequate checks for inconsistency been made?*		
C4.4	*If inconsistency was detected, what adjustments were made to the analysis, and how was this justified?*		
D EMBEDDING THE SYNTHESIS IN A PROBABILISTIC COST EFFECTIVENESS ANALYSIS ***D1. Uncertainty Propagation***			
D1.1	*Has the uncertainty in parameter estimates been propagated through the model?*		
D2 Correlations			
D2.1	*Are there correlations between parameters? If so, have the correlations been propagated through the model?*		

NATIONAL INSTITUTE FOR HEALTH AND CLINICAL EXCELLENCE

Briefing paper for methods review working party on uncertainty and only in research recommendations

The briefing paper is intended to provide a brief summary of the issues that are proposed for discussion by the Methods Review Working Party 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.

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 revised versions were published in 2004 and 2008. 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 revised draft of the Methods Guide will be available for a 3-month public consultation, expected to begin in June 2012. We encourage all interested parties to take part in this consultation.

2 Background

2.1 Relevance of topic to NICE technology appraisals

The NICE technology appraisals programme makes recommendations about health technologies close to regulatory approval through the single technology assessment (STA) process. Inevitably these decisions are made when there may be substantial uncertainty about their clinical effectiveness and cost-effectiveness. In these circumstances the acquisition of further evidence could lead to better decisions in the future. The decision to recommend a technology for use in the NHS could have an impact on the prospects of acquiring further evidence because the incentives on researchers, including those from marketing authorisation holders, are diminished once the technology has been recommended. Therefore, it has been suggested that the decision to recommend a technology should account for the potential costs to future NHS patients in terms of the value of evidence that may be forgone by early adoption.

2.2 What the current Methods Guide advises with respect to 'only in research' recommendations

Sections 6.2.11 and 6.2.12 set out the discussion of 'only in research' recommendations in the 2008 Methods Guide in terms of factors to be consider by the Appraisal Committee, but no detailed criteria or thresholds for making such decisions are provided. The concept of 'approval with research' does not feature in the 2008 Methods Guide.

6.2.11 When evidence of effectiveness is either absent or weak, the Appraisal Committee may recommend that particular interventions are used within the NHS only in the context of research. Factors that will be considered before issuing such recommendations include the following.

- *The intervention should have a reasonable prospect of providing benefits to patients in a cost-effective way.*

- *The research can realistically be set up, is already planned, or is already recruiting patients.*
- *There is a real prospect that the research will inform future NICE guidance.*
- *The broad balance of the benefits and costs of conducting the research are favourable.*

6.2.12 Recommendations on the use of technologies only in the context of research will not include consideration of which organisation (public or private) will fund the research.

2.3 Relevant methodological research

The MRC and NIHR methodology programme recently funded the Universities of York and Brunel to undertake research to help inform when NICE should recommend the use of health technologies only in the context of an appropriately designed programme of evidence development (Claxton K., Palmer, Longworth L., et al. 2011).

This paper categorised previous NICE technology appraisal guidance with a research element as either 'only in research' (OIR) recommendations (interpreted for the purposes of this briefing as meaning that the technology is recommended to be used **only** in the context of research, the nature of which is specified in the guidance) or 'approval with research (AWR) recommendations (that is, the technology is recommended alongside a further recommendation for research or data collection).

The executive summary of the CHE publication of this research (HTA monograph forthcoming) is appended to this document (Appendix A).

3 Proposed issues for discussion

In consideration of the developments in this area resulting from the MRC project, the current Methods Guide and the requirements of the Institute's

Technology Appraisal Programme, it is proposed that the following key areas are discussed by the working party.

3.1 Uncertainty about clinical effectiveness

Currently the recommendations in the methods guide focus on situations where *"evidence of effectiveness is either absent or weak".*

- Should this focus of on the estimate of effectiveness remain, or should other aspects of uncertainty in the estimates of cost effectiveness be considered.

3.2 Key principles and assessments needed for OIR recommendations

Should the methods guide recommend a more formal method of assessing the need for further research in the conduct of technology appraisals? The MRC researchers suggest the use of checklists as an aid to these judgments (see Appendix B).

- Are the checklists outlined in section 3 of the CHE research paper in Appendix B useful for committee decision making
- What additional information and analysis – over and above that already conducted in the course of an appraisal – might be required to allow the committee to be more systematic in its exploration of the value of undertaking further research
- How can research commissioners be involved when the Appraisal Committee are considering AWR/OIR recommendations ?

3.3 The concept of approval with research

The methods guide does not currently include the concept of AWR.

- Is the concept of AWR, or a similar concept, useful in circumstances where the committee is considering use of the technolgy in the context of research

4 References

Claxton K., Palmer.S, Longworth L., et al. Uncertainty, evidence and irrecoverable costs: Informing approval, pricing and research decisions for health technologies? University of York; CHE Research Paper 69; 2011.

Claxton K., Palmer S., Longworth L., et al.. Informing a decision framework for when NICE should recommend the use of health technologies only in the context of an appropriately designed programme of evidence development. Health Technology Assessment forthcoming.

5 Author/s

Prepared by Bhash Naidoo and Janet Robertson, on behalf of NICE Technology appraisal Programme

November 2011

Appendix A

Executive summary of CHE Research Paper 69[1]

The general issue of balancing the value of evidence about the performance of a technology and the value of access to a technology can be seen as central to a number of policy questions. This research was commissioned to inform when NICE should approve health technologies only in research (OIR) or with research (AWR). It has implications for policy (e.g., NICE guidance and drug pricing), the process of appraisal (e.g., greater involvement of research commissioners) and methods of appraisal (e.g., should additional information, evidence and analysis be required). However, establishing the key principles of what assessments are required and how they might be informed has much wider relevance beyond NICE and the UK NHS (e.g., informing the questions posed by coverage with evidence development initiatives).

Key principles and assessment needed

The key principles and assessments needed fall into four broad areas: i) expected cost-effectiveness and population net health effects (including benefits, harms and NHS costs); ii) the need for evidence and whether the type of research required can be conducted once a technology is approved for widespread use; iii) whether there are sources of uncertainty which cannot be resolved by research but only over time; and iv) whether there are significant (opportunity) costs which will be committed and cannot be recovered once the technology is approved.

Decisions (NICE Guidance) will depend on the combined effect of all these assessments because they influence whether the benefits of research are likely to exceed the costs and whether any benefits of early approval are greater than withholding approval until additional research is conducted or

[1] Claxton K., Palmer.S, Longworth L., et al. Uncertainty, evidence and irrecoverable costs: Informing approval, pricing and research decisions for health technologies? University of York; CHE Research Paper 69; 2011. Available from URL: http://www.york.ac.uk/che/publications/in-house/. Other related documents available from URL: http://www.york.ac.uk/che/research/teams/teehta/workshops/only-in-research-workshop/

other sources of uncertainty are resolved. The sequence of assessment and judgments required is represented as an algorithm, which can be summarised as a simple seven point checklist.

Each sequence of assessment and decision, leads to different categories of guidance (e.g., Approve, AWR, OIR or Reject) for technologies with differing characteristics, indications and target populations. Different 'types' of apparently similar guidance can be identified. This illustrates how the same category of guidance might be arrived at in different ways, helping to identify the particular combination of considerations which might underpin decisions.

The principles suggest that restricting approval to OIR, or making it conditional on research through AWR, has wider application than is reflected in previous NICE guidance. For example, OIR may be appropriate when a technology is expected to be cost-effective. Even when research is possible with approval, OIR or even Reject maybe appropriate if there are significant irrecoverable costs. Therefore, the full range of categories of guidance ought to be considered for technologies, which on the balance of existing evidence and current prices, are expected to be cost-effective. It is only approval that can be ruled out if a technology is not expected to be cost-effective, i.e., cost-effectiveness is a necessary but not sufficient condition for approval and lack of cost-effectiveness is neither necessary nor sufficient for rejection.

Distinguishing principles (what assessment are needed) from methods of analysis (how they might be informed) allows potentially wide application of principles embodied in the algorithm and associated checklist, whist recognising that how the assessment might be made is likely to differ in different contexts.

Implications for value based pricing

Any change in the effective price of the technology, either through patient access schemes (which offer some form of discount that reduces NHS costs), or direct price changes (possibly negotiated though a value based pricing scheme) will affect the key assessments, leading to different categories of guidance. The price at which a technology is just expected to be cost-

effective is commonly regarded as its value based price. This describes the threshold price below which Approve rather than Reject would be appropriate if OIR or AWR are not available as policy options. However, if they are available there are often a number of relevant price thresholds. Once uncertainty, the need for evidence and the impact of irrecoverable costs are recognised, the threshold price that would lead to Approval rather than OIR will always be lower than a single value based price based on expected cost-effectiveness alone, i.e., disregarding uncertainty in costs and effects.

Even if price negotiation becomes possible alongside NICE appraisal, it will be important to retain OIR and AWR as available categories of guidance for two reasons: i) there is no guarantee that manufacturers will always agree to the lower price below which Approval rather than OIR or AWR would be appropriate; and ii) there may be many circumstances when no effective price reduction which would make Approval appropriate, e.g., Reject or OIR guidance may be appropriate even if the effective price of a technology was zero if there is substantial uncertainty about its effectiveness and/or potential for harms.

Incentives for evaluative research

An explicit assessment of OIR and AWR provides clear signals and an incentive to ensure the type of evidence, requiring research that cannot be conducted once approved for NHS use, is available and is sufficient at launch (e.g., relative effectiveness and subtle but important differences in side effect profiles). Therefore, a predictable OIR and AWR policy signals what type of evidence is likely to be most important at an early stage. It offers manufacturers a choice, to either: i) accept OIR Guidance at a higher price but restricted volume; ii) reduce the effective price to achieve Approval, or AWR where that is possible; or iii) conduct the evaluative research at an earlier stage so that additional evidence is available at launch.

How the NHS and manufacturers are likely to share the value of evidence might inform whether manufacturers should be expected to conduct the research specified in AWR or OIR guidance or contribute to the costs of publicly funded research which may ultimately benefit their product. The

success of AWR when manufacturers are asked to conduct the research will depend on whether appropriate contractual arrangements can be established, i.e., those that can be monitored and enforced with credible penalties to ensure agreed research is conducted and in the way intended. At present, NICE does not have a credible mechanism since removing approval of a technology simply because recommended research had not been conducted was not considered an ethical or credible threat.

The assessments required can be used to consider the value to the NHS of: i) being able to conduct research while a technology is approved (value of AWR); ii) making evidence that is needed by the NHS available at launch; and iii) being able to acquire evidence more quickly. This can inform investments in better data collection, registries or information systems that might make AWR possible. The value to the NHS of having access to the evidence needed at launch can inform a range of policies, such as early advice, public investment in transitional and evaluative research earlier in the development process or other incentives for research and development. Understanding the relationship between the time taken for research to report and the value of the evidence to future populations can help to inform: i) investments which might make research findings more quickly available; ii) the trade-off implicit in the choice of alternative research designs; and iii) those areas where if research is to be undertaken there must be confidence that it can report quickly.

The value of early evidence at launch and AWR can also be considered from the perspective of the manufacturer and inform whether they or the NHS might be expected to conduct the research needed. In principle, AWR and OIR research could be publicly funded rather than undertaken by manufacturers if the costs of research could be recovered directly from manufacturers or indirectly through other price discounts. Since the costs of public research are likely to be substantially lower than for manufacturers this might be mutually beneficial in many circumstances; providing appropriate support to innovation, while allowing wider access to the data generated and more transparency and accountability in the conduct of the research.

How should the assessment be undertaken?

The order of the assessments in the checklist relate to the sequence of decision nodes that fully describe the algorithm in Appendix A. This order of considerations means that all 7 assessments do not necessarily need to be made when an earlier judgement can lead directly to guidance. Therefore, one model for an efficient process of assessment would be to consider points 1-5 routinely. The Appraisal Committee would then be in a position to either rule out OIR or AWR and issue guidance in the usual way or indicate in the appraisal consultation document (ACD) that OIR or AWR was provisionally recommended subject to advice from a research advisory committee and subsequent analysis to support an assessment of points 6 and 7 of the checklist prior to FAD. This model would avoid unnecessary analysis and incorporate the judgments of the research community without necessarily delaying appraisal.

Some assessment of: i) the type of research needed to address the key uncertainties; ii) whether this will be regarded as ethical and can be undertaken while the technology is approved for use; iii) whether it is likely to be a priority for public funding and be commissioned; and iv) when it is likely to report is required. Although the NICE appraisal process may be well suited to identifying the need for evidence, these other critical assessments (the type of research and its priority) are not necessarily ones for which NICE and its advisory committees, as currently constituted, have particular expertise. Informed judgements and better decisions might be possible through greater involvement of the research community. For example, a research advisory committee could be constituted which could consider provisional OIR or AWR guidance (at ACD), making recommendations about the type of research needed, its ethics, feasibility and likely priority during the consultation period before final appraisal and guidance. It might also make recommendations about whether research should be publicly funded or undertaken by the manufacturer with appropriate contractual arrangements (which may require the involvement of DH at some stage).

What additional information and analysis might be required?

In the assessments, cost-effectiveness was presented as net health effects per patient treated and for the population of patients over time. This provides information in a way that is directly relevant to the assessments that need to be made using information generally already available during appraisal.

An early indication of potential importance of irrecoverable costs can be based on their scale relative to expected net health effects, the point at which any initial losses are expected to be compensated by later gains, whether treatment decisions are reversible and what opportunities to improve health might be forgone by a delay to initiating treatment.

The question of whether further research might be worthwhile requires some assessment of: i) how uncertain a decision based on expected cost-effectiveness might be; and ii) what the consequences, in terms of population NHE, are likely to be if an incorrect decision is made. This can be made in a series of steps each presenting what is already available within current methods of appraisal but in ways that can more directly inform the assessment required. How the consequences of uncertainty between as well as within scenarios can be presented and interpreted is also explored.

An assessment of the type of evidence needed requires judgements about: i) how important particular types of parameters (inputs to the economic model) are to estimates of cost and QALY; ii) what values these parameters would have to take to change a decision based on expected cost-effectiveness; iii) how likely is it that parameters might take such values and iv) what would be the consequences if they did, i.e., what might be gained in terms of population NHE if the uncertainty in the values of these parameters could be immediately resolved? The methods of analysis presented in Section 3 take these steps in turn; presenting what is available within current appraisal but in ways that more directly inform the assessment required. It is only when assessing the consequences of uncertainty associated with particular parameters that additional analysis is required to provide quantitative estimates.

The current appraisal process generally already provides the information and much of the analysis required to complete all the quantitative assessment reported in Section 3. However, the information required to assess whether other sources of uncertainty will resolve over time requires information that is not commonly sort as part of NICE appraisal. NICE many need to consider how access to this type of information can be provided or whether it should extract this type of information itself at an earlier stage of appraisal.

Any additional analysis to support a more explicit consideration of OIR and AWR would need to be included in manufacturers' submissions and be reviewed by the ERG. Although the additional analysis itself is limited (most is already required but sometimes presented in different forms), more explicit consideration of OIR and AWR and their link to price would make the critique of how uncertainty and its consequences has been characterised more important. An assessment of whether the point estimate of cost–effectiveness is reasonable is inevitably a more limited task than also assessing whether the uncertainty surrounding that assessment is credible. Any additional burden on ERGs (and manufacturers) might be eased with clear guidance on the details of how analysis should be conducted and presented, what common assumptions are deemed reasonable and provision of additional information by the Institute as well as only considering points 6 and 7 on the checklist after ACD and following advice from a research advisory committee.

Appendix B

The following checklists and algorithm are reproduced from CHE Research paper 69[2]

Checklist for OIR and AWR (technologies expected to be cost effective)

Point	Assessment	Judgement Yes	No
1	Is it cost-effective?	Yes	
2	Are there significant irrecoverable costs?		
3	Does more research seem worthwhile?		
4	Is the research possible with approval?		
5	Will other sources of uncertainty resolve over time?		
6	Are the benefits of research greater than the costs?		
7	Are the benefits of approval greater than the costs?		

Checklist for OIR and AWR (technologies not expected to be cost effective)

Point	Assessment	Judgement Yes	No
1	Is it cost-effective?		No
2	Are there significant irrecoverable costs?		
3	Does more research seem worthwhile?		
4	Is the research possible without approval?		
5	Will other sources of uncertainty resolve over time?		
6	Are the benefits of research greater than the costs?		
7	Are the benefits of approval greater than the costs?		

[2] Claxton K., Palmer.S, Longworth L., et al. Uncertainty, evidence and irrecoverable costs: Informing approval, pricing and research decisions for health technologies? University of York; CHE Research Paper 69; 2011. Available from URL: http://www.york.ac.uk/che/publications/in-house/

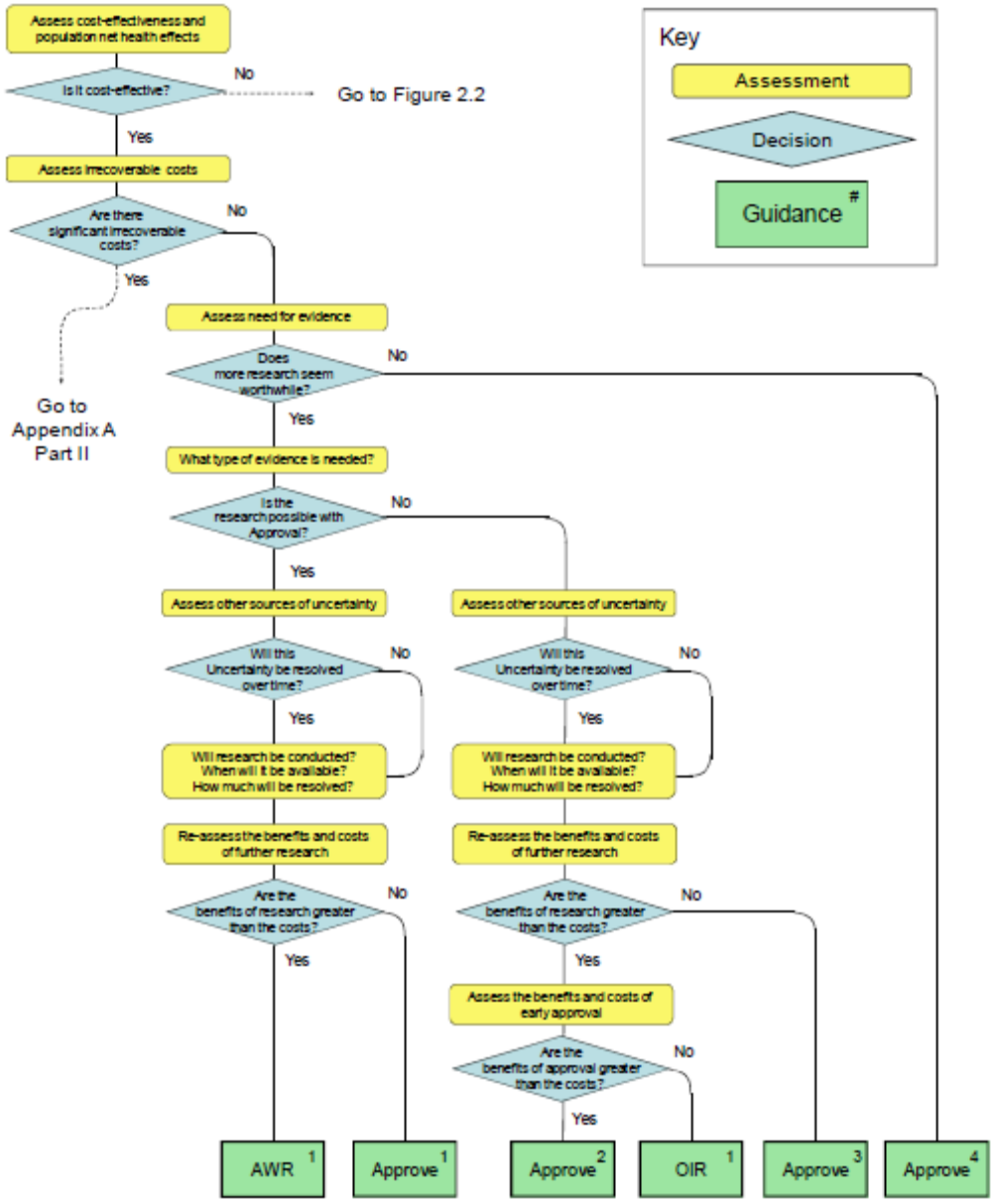


Figure 2.1 **Technologies expected to be cost-effective**

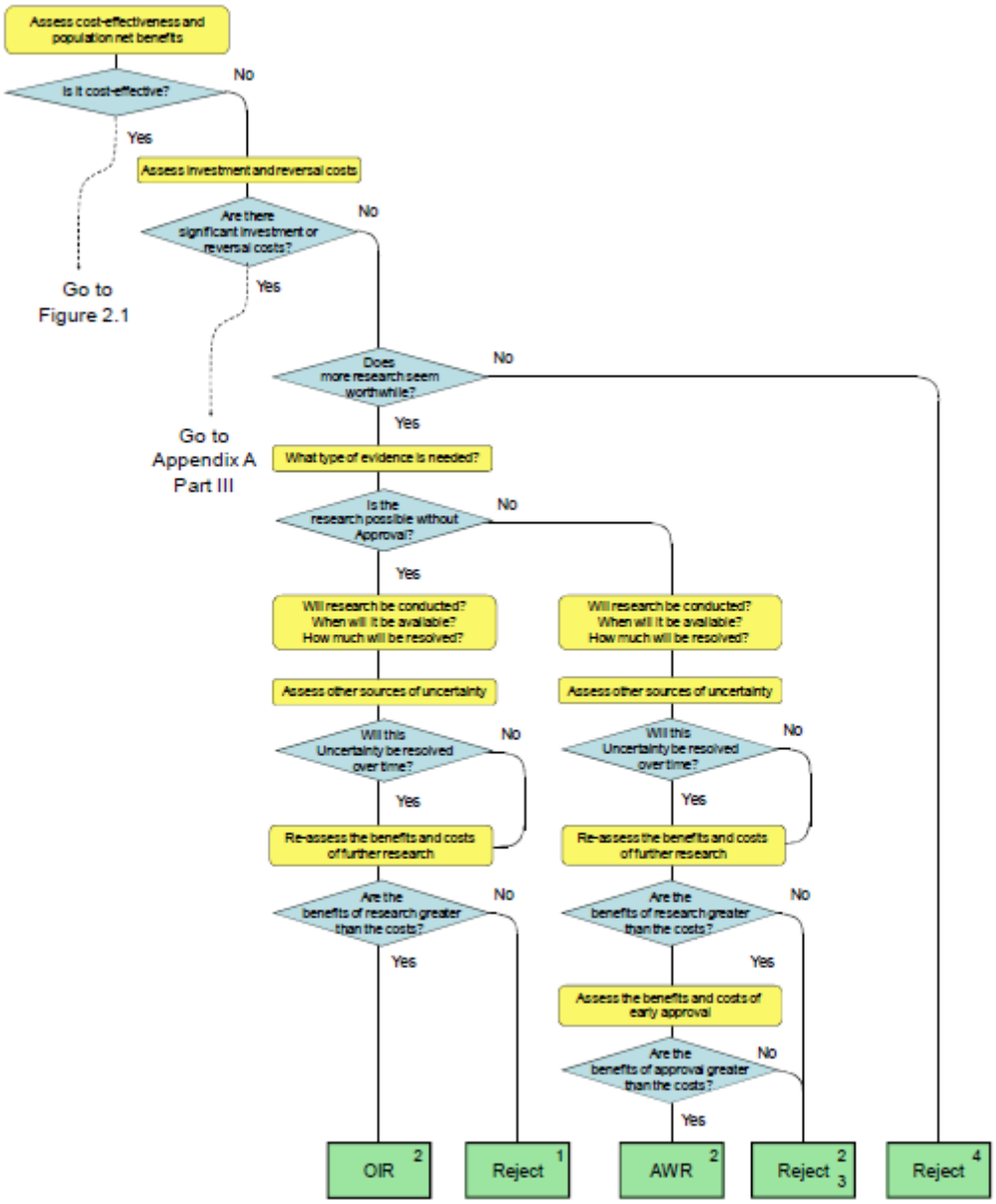


Figure 2.2 **Technologies not expected to be cost-effective**

NATIONAL INSTITUTE FOR HEALTH AND CLINICAL EXCELLENCE

Briefing paper for methods review workshop on patient evidence 1: making the most of patient-based evidence and patient and public involvement

The briefing paper is written by Dr Sophie Staniszewska in collaboration with members of the Institute's Technology Appraisals team. 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/technologyappraisallprocessguides/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 2011. We encourage all interested parties to take part in this consultation.

2 Background

The current 'methods guide' states the following (see section 4.3):

> *"Submissions are invited from all patient/carer groups involved in the appraisal. Patient evidence can include the views, assessments and evaluations of: individual patients, individual carers, groups (such as groups of patients, carers or voluntary organisations that represent patients). Patient evidence refers to any information originating from patients and/or carers that may inform the appraisal of a technology. […]*

There are two principal reasons for presenting patient evidence. Patients and carers are a unique source of expert information about the personal impact of a disease and its treatment, which can help set the correct scope for the assessment of the evidence and enable the realistic interpretation of the clinical and economic data as the appraisal progresses. Patient evidence can identify limitations in the published research literature; in particular, the failure to capture the true concerns of individual patients related to HRQL over and above measurements using standardised instruments (such as questionnaires) developed using psychometric techniques.

For the purpose of informing its technology appraisals, the Institute is looking for a concise and balanced overview that reflects the range of patient and carer perspectives. Two groups of experts – clinical specialists and patient experts – are selected by the Committee Chair from nominations provided by (non-manufacturer) consultees and commentators. Clinical specialists and patient experts provide written evidence and attend the Committee meeting to help in the discussion of the technology being appraised."

Section 4.5 gives further guidance to people attending committee meetings as experts:

"The experts attending the Committee meeting are asked to submit, in advance, a brief written personal view of the current management of the condition, the (expected) role of the technology and its use in the NHS, as well as to provide oral commentary during the meeting. The purpose of the oral commentary provided by the experts is to explore the evidence that is provided in the written submissions from consultees. During the open part of the meeting, clinical specialists and patient experts are encouraged to interact fully in the debate with the Committee, including responding to and posing questions. The clinical specialists and patient experts are asked to withdraw from the meeting before the Committee discusses the content of the guidance.

Views expressed orally by the experts at the Committee meeting can usefully inform the debate in a variety of ways, including the following.
- Identifying important variations in clinical practice in both the management of the condition in general and specifically in the current use of the technology. […] - Giving personal perspectives on the use of the technology and the difficulties encountered, including the important benefits to patients and the range and significance of adverse effects as perceived by patients. - Providing views on the nature of any rules, informal or formal, for starting and stopping use of the technology. This might include the requirement for additional analysis: to identify appropriate subgroups of patients for treatment with the technology, to assess response to treatment and the potential for discontinuation.
- Responding to queries that arise from: the lead team presentation (the lead team being two Committee members who make a brief presentation to introduce the topic of the appraisal), issues raised by the Chair and other Committee members, issues raised by other experts.

A lead team, selected from the Committee members at the start of each STA, helps the NICE technical lead prepare a summary of the evidence, known as the premeeting briefing. One of the lay representatives on the Committee is also selected to advise the lead team when developing the premeeting briefing. At the Appraisal Committee meeting, the lead team makes a brief presentation, based on the premeeting briefing, to introduce the STA topic.

The ‘lay lead’ role was designed to further develop the role of the 12 lay members on the Technology Appraisals Committees. When starting this lay lead process, two main areas of potential impact were proposed: increasing lay member involvement with the work of the committee, and increased visibility of patient/carer evidence. The three lay members per committee take it in turns to be the lay lead, with one of them being assigned to every topic. They advise the lead team about the key patient, carer and public issues and evidence within the committee topic documentation. This helps ensure that these issues and evidence are explicitly referred to in the presentations given at the start of the committee meeting.

NICE Patient Experience Guidance

The NICE Patient Experiences Guidance will be published in 2011. A scoping study of patient experiences was carried out as part of this work, to identify key generic dimension of patient experience that apply to all patients (Staniszewska et al 2011, in review). This scoping study, which was included in an appendix in the NICE Consultation on this Guidance, may provide a helpful context for discussions about the dimensions of experience that can be considered in technology appraisal.

3 Proposed issues for discussion

From the description in the current methods guidance it is clear that NICE Technology Appraisal Committees consider a variety of patient-based evidence.

This workshop will focus on exploring whether current processes of technology appraisal are maximising the potential for using patient-based evidence and the potential for patient and public involvement in the identification, synthesis and interpretation of patient-based evidence. This paper provides some context for this discussion and considers the concept of patient-based evidence and the levels of patient and public involvement.

3.1 The concept of patient-based evidence

The conceptual framework drawn on to inform this discussion includes clinical evidence, economic evidence and patient-based evidence (Staniszewska et al 2010, Rycroft-Malone 2004, Doll 1974). Patient-based evidence includes qualitative and quantitative forms of evidence, such as studies that have used qualitative methods to explore patient experiences, surveys that have attempted to measure different dimensions of patient experiences. Patient-based evidence can also include patient-reported outcomes (PROs) with measures patients' assessments of their health status and well-being (Staniszewska 2010).

Compared to clinical and economic forms of evidence, patient-based evidence is less well defined conceptually and methodologically. This makes it more difficult to integrate automatically with the clinical and economic forms of data, as there are few agreed frameworks to facilitate this process, although some research has started to examine the possibilities (McInnes et al 2011). In the absence of ready-made frameworks and methods, the role and contribution of patient-based evidence needs to be carefully considered within Technology Appraisal to ensure the benefits of this form of evidence are maximised.

The synthesis of qualitative data will be considered more fully in Ruth Garside's presentation. There are also issues around the synthesis of experiences data with quantitative experiences data, or other forms of patient-based evidence, such as patient-reported outcome measures. In addition, the syntheses of qualitative data with data from quantitative systematic reviews that identify interventions to enhance some aspect of patient experience also needs to be considered.

3.2 Patient and public involvement

Patient experts can be nominated by a range of organisation which has been identified as having a close interest in the technology under appraisal. As well as nominating one or more experts to attend the committee meeting, patient organisations are also invited to make written statements or submissions. The patient expert who attends the meeting presents their own opinion, which may differ from the views presented by the nominating organisation.

Patient experts provide evidence, sometimes through a formal presentation, that contributes to discussions about the appropriateness, relevance and acceptability of a particular technology. The patient experts may have different philosophical underpinnings and may vary in the forms of knowledge and evidence they contribute to the process. Some initial unravelling of philosophical perspective and nature of evidence that patient experts may provide is given below to stimulate discussion about the key questions:

- **Philosophical underpinning**: The philosophical underpinnings that guide a patient expert in relation to level of involvement may influence

the way in which they provide evidence and their expectations of the process. For example, consultative forms of involvement might involve patient experts being asked for a view, but they may expect less or no involvement in the discussion or synthesis of evidence or the decision on a recommendation. Some patient experts may favour more collaborative roles where they are inherently involved in contributing to the synthesis of different forms of evidence and in the formation of a recommendation. Some patient experts may also be familiar with the concept of user-led research, where users of service or user-researchers lead a project. The way patient experts from this background may provide evidence may differ from those more used to collaborative or consultative forms of involvement.

- **Experiential knowledge or perspective**: The patient expert may be someone who has experiential knowledge based on their own experiences. In this way they offer a perspective, which can generate valuable discussion. This issue of representation in this context is really a red herring as the focus should be on their perspective, as with other experts. Alternatively experiential knowledge may be drawn from the experiences of a broader constituency of people who have come together in some form, for example, as a patient organisation and may represent the range of views.

- **Research-based knowledge:** The patient expert may be someone with a broader knowledge and evidence base about experiences with a particular technology. Their analysis and synthesis of research-based knowledge may be undertaken with a different 'lens,' appraising aspects of experience according to different criteria in the context of a technology. The knowledge or evidence they are aware of may come from research, such as a meta-ethnography or may be more diverse and can include grey literature.

- Research-based knowledge can include **methods critiques**, for example, whether assumptions made in economic modelling have validity. For example, that people can make a choice between

interventions when they have not experienced a condition. Another example is patient-reported outcomes measures (PROMS) where concerns have emerged about the extent to which PROMS capture outcomes of importance to patients (Haywood et al, 2011 Staniszewska et al 2011).

4 Questions for discussion

1. How can we maximise the potential for identifying and incorporating evidence from patients and carers in technology appraisals, within current processes?

2. Is the methods guide clear on the level and nature of involvement we expect from patient experts in technology appraisal?

3. Does the methods guide give clear guidance on the nature and type of evidence and knowledge we expect patient experts and patient organisations to contribute?

4. How could the guidance on nature and types of evidence and knowledge be improved?

5. What role could patient experts and patient organisations have in evaluating the adequacy of PROMS data, in relation to content validity?

5 References

Doll R. Surveillance and monitoring. International Journal of Epidemiology, 1974; 3: 305–314.

Haywood K, Staniszewska S, Chapman, S. (2011) Quality and acceptability of patient reported outcome measures used in Chronic Fatigue Syndrome/Myalgic Encephalomyelitis (CFS/ME): a structured review. Quality of Life Research, May 18. [Epub ahead of print]

McInnes E Seers K & Tutton L(2011) 'Older people's views in relation to risk of falling and need for intervention: a meta-ethnography.' *Journal Of Advanced Nursing* Early view - first published online: 1 JUN 2011 | (0309-2402)

Rycroft-Malone J, Seers K, Titchen A et al (2004). What counts as evidence in evidence-based practice. Journal of Advanced Nursing, 47 (1):81-90.

Staniszewska S, Haywood K, Brett J, Tutton (2011). Patient and public involvement in PROMS : Evolution not revolution. The Patient Patient-Centered Outcomes Research). In press.

Staniszewska S , Crow S, Badenoch D, Edwards C, Savage J, Norman W (2010). The PRIME Project: Developing a Patient Evidence-Base. Health Expectations, 13 (3): 312-322

6 Author/s

Prepared by Dr. Sophie Staniszewka, Senior Research Fellow, Lead for Patient Experiences and Public Involvement Programme, Royal College of Nursing Research Institute, University of Warwick

October 2011

NATIONAL INSTITUTE FOR HEALTH AND CLINICAL EXCELLENCE

Briefing paper for methods review workshop on patient evidence 2: patient evidence, qualitative research and synthesis

The briefing paper is written by Dr Ruth Garside in collaboration with members of the Institute's Technology Appraisals team. 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/technologyappraisallprocessguides/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 2011. We encourage all interested parties to take part in this consultation.

2 Background

The Technology Appraisals uses a variety of types of evidence to arrive at its recommendations. Section 3.4 of the guide includes the following text:

> *In addition to evidence on treatment effect and cost effectiveness, the appraisal of health technologies requires consideration of a range of other issues. A variety of types of evidence generated from a range of sources, of both quantitative and qualitative origin, is relevant to these areas. […]*

Information on whether a health technology is considered to be an acceptable or appropriate technology (compared with alternative technologies) by patients, carers or healthcare professionals is useful. Individuals or groups may prefer particular health technologies, for example, because of the frequency or nature of adverse events or the route or frequency of administration. The health impact of most of these factors (for example, adverse events) is expected to be reflected in the estimation of HRQL. In addition, individuals or groups may be concerned about the ethics of using a particular technology. These are relevant considerations for an appraisal because they influence judgements on the usefulness of technologies, inform the nature of choice between alternative technologies and provide important evidence on the extent to which these considerations have been adequately captured in measurements of HRQL. Evidence relevant to these considerations can come in various forms, be based on quantitative or qualitative measurements, and originate from a range of sources that have different methodological strengths and weaknesses. Such evidence includes literature reviews, adverse effect/adherence/continuation data collected in research studies, patient surveys (for example, of adverse effects or preferences) and summarised testimonies from clinical specialists and patients.

Thus, in addition to seeking the views and experiences of patients through their direct involvement in Committee meetings and in consultations on documents, other types of evidence on the experience of patients can also contribute to the evidence base for a Technology appraisal.

Whilst there is a laudable aim to ensure that the patient voice is heard in the appraisal process, the current methods guide conflates a number of issues that make the *purpose* of doing this and the *methods* for it unclear. For example, within a technology appraisal, it seems to conflate *de novo* patient/public involvement strategies and understandings about the kind of pertinent research evidence that might already exist. Potentially, this leads to

neither goal (meaningful patient involvement with the process and use of existing patient centred research) to being effectively reached.

There is a need to distinguish clearly between the following types of evidence:

1. Quotes and written submissions from people who have a disease or condition (or their family or carers) about their experience of this, and/or its treatment. This could be called **Qualitative Evidence** – that is, evidence in the form of text/words (such as that provided through the contributions of patient experts at committee meetings) which has not be subject to formal research methodology in order to collect or analyse it.

2. Research which analyses group or individual interviews or written texts with patients (or their family or carers) about a particular topic in order to produce an analytic account of the nature of living with a condition (and/or its treatment) based the experience of a number of such people. This is **Qualitative Research Evidence** – that is, evidence that has been collected and analysed using one of a number of recognised approaches to this type of research.

3. As for other forms of evidence used in technology appraisals, systematic review and synthesis procedures can be applied to existing qualitative research evidence in order to produce a coherent understanding of the body of work about living with a particular condition, and/or its treatment. This is a **Synthesis of Qualitative Research Evidence**. A range of approaches have been described for this and systematic reviews and syntheses of qualitative research are already in use by the CPHE at NICE to inform the production of public health guidance.

Patient involvement is being considered in more detail by Dr Staniszewska, so this paper focuses on patient evidence which is sourced from research. It will focus on qualitative research, which has the potential to reveal the patient experience, although quantitative surveys and questionnaires may also be a source of relevant information.

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 Using patient evidence to inform technology appraisals

3.1.1 Qualitative evidence from submissions

Section 4.3.4 of the methods guide states the following:

> *For the purpose of informing its technology appraisals, the Institute is looking for a concise and balanced overview that reflects the range of patient and carer perspectives, including majority views and potentially important views that may be held by only a few patients. The Institute is interested in capturing a range of patient and carer views on, and experiences of, living with the condition, and the impact of a technology on a patient's symptoms and physical, social, psychological and emotional state. It is also interested in what it might be like living without the technology being appraised.* ***Patient evidence is most useful when presented as a synthesis of information, balancing positive and negative views, rather than as a series of individual testimonials.***

The highlighted in bold preference above for "synthesis of information….rather than as a series of individual testimonials" seems to imply that the most useful form of patient evidence is that derived from qualitative research, whether individual reports or an evidence synthesis (2 or 3 above), although currently it collects qualitative evidence (1 above) from groups and individuals. Patient groups collate the concerns and testimonies of their members, although there are no current guidelines for how, and from whom this is done. It would, in theory, be possible for NICE, or another group, to formally analyse this

submitted, textual information in order to identify the key concerns raised by those who have provided submissions or taken part in the consultations.

3.1.2 Qualitative research evidence

Alternatively, qualitative research could be used to obtain patient evidence. Such research could include that undertaken with patients, their families and/or carers which explores areas such as:

- the impact of having a condition of disease,
- the experience of being within the healthcare system for treatment of that condition,
- the experience of undergoing specific treatments for that condition.

If new research were to be undertaken, the guidance should expand on who should undertake this research, and the methods for identifying, sampling and recruiting participants. It would also need to guide the researchers as to how and by whom, areas for investigation should be identified; for example, through reflection on the quality of life tools currently used, recognition of particular issues in comparing treatments such as balance of adverse events etc. or by allowing patients themselves to prioritise what they discuss by using more unstructured interview methods. Preferred methods of data collection and analysis might also be mentioned.

3.1.3 Syntheses of qualitative research evidence

Where existing research is to be considered as providing patient evidence, it is likely that systematic review and synthesis will provide the most useful framework to understand what is known in the literature as a whole about the experience of a condition and its treatment. Aspects of intervention design, acceptability, implementation and context, are unlikely to be illuminated by the results of quantitative research, and may also be found in qualitative research.

Section 5.3 of the methods guide gives guidance on the review and synthesis of evidence on clinical effectiveness, principally focusing on evidence from randomised controlled trials. There is no corresponding guidance on using

existing qualitative research, including methods of identification, quality appraisal or synthesis.

There are a number of approaches to such review and synthesis, which synthesise qualitative research alone, or with quantitative research including narrative synthesis, meta-ethnography and meta-synthesis (Britten, et al., 2002; EPPI-Centre, 2007; Jensen, et al., 1996; Mays, et al., 2005; Petticrew, et al., 2006; Popay, et al., 2006). The nature of the evidence identified may dictate the most appropriate synthesis methods. In addition, some aspects of the systematic review, such as the most appropriate way to identify qualitative research (Shaw, et al., 2004), and methods of appraising qualitative research, remain contentious (Dixon-Woods, et al., 2004; Wallace, et al., 2004). Despite this, there is increasing acceptance of the methods of synthesis and appreciation of its utility, including in a policy making context (Centre for Public Health Excellence, 2009). For example, recent syntheses have explored the experience of heavy menstrual bleeding (Garside, et al., 2008); strategies employed by patients to manage their psychotropic medicine taking (Britten, et al., 2010); and beliefs about skin cancer and tanning, in the context of providing information to prevent skin cancer (Garside, et al., 2009).

3.1.4 Questions for discussion

3.1.4.1 Qualitative evidence from submissions

Should existing submissions be treated as qualitative evidence which needs to be formally analysed? If so, by whom? Using what methods?

What guidance should be given about the nature and quality of submissions?

3.1.4.2 Qualitative research evidence

Should the Technology Appraisals methods guide give guidance on the use of new qualitative research?

To what extent should the submission of new qualitative research evidence be encouraged in the Technology Appraisals methods guide? From whom and with whom should such research be undertaken? How would the scope and methods of enquiry be determined?

What guidance should be given to optimise the methodological quality of qualitative research evidence used in the Technology Appraisals programme?

3.1.4.3 Syntheses of qualitative research evidence

Should the Technology Appraisals methods guide give guidance on the use of syntheses of qualitative research? By whom should these be undertaken?

How could these syntheses be incorporated into the overall clinical and cost effectiveness evidence base?

4 References

Britten N, Campbell R., Pope C. et al. (2002) Using meta-ethnography to synthesise qualitative research: a worked example. *J Health Serv Res Policy* 7 (4): 209-215.

Britten N, Riley R., and Morgan M. (2010) Resisting psychotropic medicines: a synthesis of qualitative studies of medicine taking. *Advances in Psychiatric Treatment* 16: 207-218.

Centre for Public Health Excellence(2009) *Methods for the Development of NICE Public Health Guidance (2nd edition).* 2 edition. London: National Institute for Health and Clinical Excellence.

Dixon-Woods M, Shaw R.L., Agarwal S. et al. (2004) The problem of appraising qualitative research. *Qual Saf Health Care* 13: 233-225.

EPPI-Centre(2007) *EPPI-Centre methods for conducting systematic reviews.*London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Garside R, Britten N., and Stein K. (2008) The experience of heavy menstrual bleeding: A systematic review and meta-ethnography of qualitative studies. *Journal of Advanced Nursing* 63 (6): 550-562.

Garside R, Pearson M., and Moxham T. (2009) What influences the uptake of information to prevent skin cancer? A systematic review and synthesis of qualitative research. *Health Education Research*.

Jensen LA,Allen M.N. (1996) Meta-synthesis of qualitative findings. *Qual Health Res* 6 (4): 553-560.

Mays N, Pope C., and Popay J. (2005) Systematically reviewing qualitative and quantitative evidence to inform management and policy making in the health field. *J Health Serv Res Policy* 10 (S1): 6-20.

Petticrew M, Roberts H(2006) *Systematic Reviews in the Social Sciences: A practical guide.*Oxford: Blackwell Publishing.

Popay J, Roberts H, Sowden A, et al.(2006) *Guidance on the conduct of narrative synthesis in systematic reviews.*London: ESRC Methods Programme.

Shaw RL, Booth A, Sutton AJ, et al. (2004) Finding qualitative research: an evaluation of serach strategies. *BioMed Central Medical Research Methodology* 4 (5).

Wallace A, Croucher K., Quilagars D. et al. (2004) Meeting the challenge: developing systematic reviewing in social policy. *Policy and Politics* 32 (4): 455-470.

5 Author/s

Prepared by Dr Ruth Garside

Senior Research Fellow, PenTAG, Peninsula Medical School, University of Exeter.

November 2011

NATIONAL INSTITUTE FOR HEALTH AND CLINICAL EXCELLENCE

Report to the Methods Review Working Party

Key issues arising from workshop on patient evidence

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 patient evidence. 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

- In the current methods guide, there is no clear definition of patient evidence and the myriad ways it might be obtained and used (patient attendance at Committee meetings, written statements, patient organisation submissions, qualitative research, synthesis of qualitative research, patient involvement in consultations). Importantly, there is blurring and confusion between patient involvement, and qualitative research in the methods guide.

- Some of the discussion on maximising the potential for identifying and incorporating patient evidence in technology appraisals focused on the NICE processes rather than methodology, particularly in relation to earlier involvement. There was discussion of the need for further

research looking at previous submissions so that the qualities of a good submission could be identified more clearly than is currently the case.

- There was much discussion of the integration of patient evidence into the decision making of the Committee. Many felt that evidence from patients had a low prominence in technology appraisals because it could not easily be integrated into the economic analysis that usually forms the basis of the decision. There was some discussion of how the economic modelling could incorporate patient evidence.

- The role of patient experts as critics of the assumptions made in economic analyses and of the extent to which both models and patient reported outcome measures capture outcomes that are of importance to patients was discussed. The methods guide currently mentions this role, but could perhaps give more guidance to patient experts and patient organisations on how they might best fulfil it.

- Some delegates felt that the technical language in the methods guide made it inaccessible to patient experts.

- The current methods guide expresses a preference for *"a synthesis of information […] rather than as a series of individual testimonials"*. This implies a preference for evidence derived from qualitative research. Workshop delegates were not supportive of this implied preference and generally felt that too much analysis could result in a loss of the richness of the language in the direct testimony of patients. There was little support for subjecting the written patient statements received in the current processes to formal analysis.

- Some, but by no means all, patient organisations have the capacity to conduct primary qualitative research in support of a submission and the timeframes of a NICE appraisal do not facilitate this. Appraisal Committee members among delegates generally agreed that Committee didn't necessarily prefer this type of submission to the more informal reporting of patient experiences, meaning pertinent

submissions need not be out of reach of organisations that do not have this capacity.

- There was agreement that review and synthesis of existing qualitative research could usefully contribute to technology appraisals but unless this becomes a requirement of the NICE methods, it was unclear whether it would be useful to provide guidance on how this should be done in the methods guide.

- There was agreement that patient evidence is an important form of evidence alongside clinical and economic evidence but that it is complex and needs to be teased out into its component parts. There needs clearer definition of what we mean by patient involvement and patient evidence.

2 Questions posed to the workshop participants

1. *What more should NICE do to maximise the potential for identifying and incorporating evidence from patients and carers in technology appraisals, within current processes? What is unique about the contribution that patient evidence makes within the context of technology appraisals?*

2. *Does the methods guide give clear guidance on the nature and type of evidence and knowledge we expect patient experts and patient organisations to contribute? Is it clear what level of involvement is required at different stages in the process? How could the guidance on nature and types of evidence and knowledge be improved?*

3. *What role could patient experts and patient organisations have in evaluating the adequacy of patient-reported outcome measures (PROMs) data, in relation to content validity? What guidance could be included in the methods guide to help patient experts and patient organisations contribute in this way?*

4. *Should the submissions NICE currently receives be treated as qualitative evidence which needs to be formally analysed? If so, by whom? Using what methods? What guidance should be given about the nature and quality of submissions?*

5. *To what extent should the submission of new qualitative research evidence be encouraged in the Technology Appraisals methods guide? From whom and with whom should such research be undertaken? How would the scope and methods of enquiry be determined? What guidance should be given to optimise the methodological quality of qualitative research evidence used in the Technology Appraisals programme?*

6. *Should the Technology Appraisals methods guide give guidance on the use of syntheses of qualitative research? By whom should these be undertaken? How could these syntheses be incorporated into the overall clinical and cost effectiveness evidence base?*

3 Summary of the workshop discussions

This workshop involved two presentations, each of which focused on different aspects of evidence from carers and how the technology appraisals programme might make best use of it. The first presentation and questions 1–3 above focused on the nature of patient evidence, its contribution to NICE decision making and the role of patient experts and patient organisations in supplying this evidence. The second presentation and questions 4–6 focused on clarifying what patient evidence is, and ensuring that patient involvement and qualitative research are used to their best advantage in the NICE process.

In addition to considering how the methods guide might be improved, the participants also considered how NICE might maximise the usefulness of patient submissions by providing further support and education for patient organisations submitting to the NICE technology appraisals programme. There was also consideration given to the value of research reviewing

previous submissions in order to learn from successful strategies. Workshop attendees were generally supportive of such research.

Similarly, issues around the technology appraisals process were raised relating to the stages at which patients were involved. There was an emphasis on earlier involvement if there was to be a move from a consultative approach to a more collaborative approach. The possibility of improving patient submissions by supplying an enhanced template was also considered. At the end of the process, some delegates considered that a post-appraisal debriefing for patient organisations could be useful (as is currently offered to manufacturer consultees).

Comments relating to these issues have been noted but will not be covered in detail in this paper, which will focus on the methods guide.

3.1 The contribution of patient evidence in technology appraisals

This issue was discussed in the context of the current reference case in which EQ-5D is the preferred measure of health-related quality of life in adults. The Delegates felt that direct patient involvement in the form of patient testimony and written submissions offers valuable insight into the impact of conditions and interventions on individuals' daily lives which cannot be captured in a health-related quality of life measure such as EQ-5D.

Delegates noted that concepts surrounding personal and social acceptability of interventions are best captured using directly reported patient testimony. An example was given of an intervention which was taken orally when the existing comparator intervention was given intravenously. Although the benefits of the two interventions in terms of QALYs generated might be similar, patients would favour the new oral treatment if it improved their everyday experience. The possible consequences of these preferences on adherence may also be assessed from patient evidence, particularly from accounts from patients with experience of a particular condition, rather than from data that relies on theoretical assumptions.

One delegate expressed the view that the unique contributions from patients involved in the NICE process may be divided into information that should be

(but is not) captured by the QALY [through the imperfections of the methodology], and information which cannot or should not be captured by the QALY.

3.2 Increasing the prominence of patient evidence

A view expressed by some delegates, including lay contributors to previous submissions, was that evidence contributed through patient involvement was given low prominence and visibility in technology appraisals in comparison to health economic data and modelling results. It was felt that if patient organisations were more confident that patient involvement is an integral part of the decision process they would be more enthusiastic about contributing and submitting. Other delegates suggested that some patients and carers may feel that their contribution is in some way less valuable than that from clinical and economic experts.

It was suggested that having improved written qualitative and patient evidence in the topic's evidence-base would reduce the pressure that patient experts might feel to adequately represent all those issues. However, by and large the solutions to these problems discussed by the delegates were related to improvements in the process by which patients are involved in appraisals and improvements in the support offered by NICE rather than methodological issues that could be covered by the methods guide.

3.3 The clarity of the current methods guide

The majority of delegates felt that the methods guide does not give clear guidance on the nature and type of evidence expected from patient experts and patient organisations.

There was repeated comment that technical language in the current guide may be inaccessible for patients and patient organisations. This might be related to the lack of clarity about the different types and sources of patient evidence that are potentially available and useable. While some delegates felt that more information on the types of evidence required should be included in the methods guide, others raised concerns that favouring one type of evidence over another may discourage organisations with an 'unfavoured'

evidence type from submitting at all. This was linked with concern from some patients that smaller, less financially secure patients and patient organisations could be disadvantaged if a demand for more robust evidence was explicitly favoured.

Delegates from patient organisations were clear in their desire for more guidance about what type of evidence is useful to NICE. There was repeated suggestion that examples of cases where patient evidence has had an impact on a decision in the past would be useful [this relates to the non-methods guide issues mentioned above].

3.4 The level of involvement of patient organisations

Issues relating to the level of involvement were discussed in the context of the briefing paper and presentation outlining three levels of involvement: patient-led, collaborative and consultative. The majority of delegates felt that it is not clear from the methods guide what level of involvement is required at each stage of the process. There was consensus that getting patients and patient organisations involved in the early stages of the appraisal process (that is, during scoping) is important. There was repeated suggestion across tables that the process would benefit from patient participation in the development of the economic analysis in some way to ensure that the model reflected the experience of patients, for example in terms of the health states included. This perhaps indicates a desire for more involvement; a more collaborative rather than consultative approach.

3.5 The role of patient experts and organisations in evaluating the adequacy of PROMs data

Again, this question was discussed in the context of the current methods reference case that indicated a preference for the EQ-5D. Consultees discussed the adequacy of EQ-5D in relation to capturing the impact of health technologies on patients. They also discussed the role of patient organisations and patient experts in evaluating the adequacy of EQ-5D for their particular patient populations.

It was acknowledged that current methods do not necessarily capture all the aspect of quality of life that are important, and that patient experts and patient organisations have a role in identifying those missing parts and bringing them to the attention of the Committee. There seemed to be a general consensus that EQ-5D had a number of weaknesses; for example certain domains might be missing like vision and hearing, it might include non-responsive dimensions and levels and it might be insufficiently sensitive to measure some important changes in quality of life.

One group noted that EQ-5D or any other PROMs tools are only attempting to capture health-related quality of life and will not capture experiences and impacts due to the processes involved in health care delivery. These can sometimes be more important in determining the most appropriate treatment.

One of the groups also raised the issue of ability and capability of patient organisations, in terms of resources and skill mix, in evaluating PROMs instruments and in conducting appropriate research to inform NICE decision making. They felt it was more for researchers to develop tools and measures that are as comprehensive as possible rather than to rely on directly reported patient evidence to fill the gaps.

Some delegates raised the issue of how the additional information from patients could be incorporated into decision making; would it be considered robust enough for cost-effectiveness analysis? There were concerns about how much weighting would be given to this additional information, what would be the recognised way of presenting it. It was unclear how non-preference-based PROMs are dealt with in the decision making process.

It was generally agreed that the methods guide should clarify what is expected from patient experts and patient organisations because it takes a lot of time and effort on their part to undertake this activity. There were suggestions that the methods guide should emphasise EQ-5D’s common deficiencies as well as specific deficiencies for particular patient populations.

Some delegates emphasised the use of lay and user friendly language within the methods guide for ease of understanding.

3.6 What guidance should be given to patient organisation about the nature and quality of submission?

Delegates were concerned that not all organisations have the capacity and resources to conduct extensive research in preparing their submissions. Also small organisations would have less capacity to produce submissions of as high quality as those from larger patient organisations. This could mean that their 'voice would not be as loud' as that of better funded organisations.

Appraisal Committee members among delegates generally agreed that Committee didn't necessarily want patient organisations to produce large amounts of material, meaning pertinent submissions need not be out of reach of smaller organisations. Appropriate guidance could benefit both patient organisations (who write the submissions) and committee members (who read them).

Appraisal Committee delegates identified broadly two important purposes for patient expert and patient organisation submissions and statements:

1) To provide the experiential context of the clinical decision;

2) To highlight aspects of the experience of either having the condition or taking the treatment for the condition that may not be clear or appropriately represented within the quantitative clinical or economic metrics (for example alopecia as an adverse effect of certain cancer therapy might be ignored in health economic modelling but it could be very important to some patients).

The experiential context broadly means getting a better sense of what it feels like to have the condition, and what it feels like to have the treatment for the condition. Even though this information wouldn't necessarily lead to the Committee making different recommendations, many Committee members considered this important to have as it helped to humanise the decision-making process and make the implications and importance of their decisions clearer. The second purpose was also considered important, especially where the quantitative evidence was lacking or ambiguous and the decision was near the margins.

Some participants suggested producing guidance for patient organisation for the submission like ‘Hints and Tips for Patient Experts’ produced by NICE for patient experts participating in Appraisal Committee meetings.

Some participants cautioned about the risk of being too prescriptive. This could suppress individual patients concerns from being highlighted. Getting the balance right between informing people about how to produce a submission that has the right sort of information from Committee’s perspective, without being so prescriptive that it discourages engagement was considered very important.

3.7 Submissions as qualitative evidence for analysis

At the moment, main themes are identified in patients’ submission and presented at the Committee meeting, normally by the lead team (including the lay lead) in their presentations. In addition committee members are supplied with and expected to read the original submissions. Some noted that the current method by which patient evidence submissions are presented to the committee with the specific input of a lay lead could be seen as analogous to the content analysis methodology of qualitative research although this might be done in very variable ways as it cannot be assumed that the lay leads have experience of qualitative analysis. Since no structured tools are used for the data collection, a framework analysis of patient submission is not possible. The participants agreed that quotations from patients in submissions and written statements add life to the discussion and humanise the complex clinical and statistical data presented at the meeting.

The participants agreed that main purpose of patients submission is to bring insight into the condition and were concerned that an over formal analysis of the patient submission will take the life experience out of the discussion and potentially ‘dehumanises’ the evidence, getting further away from the patient experience (although a good qualitative analysis should retain the patient voice). The other factors which can potentially discourage a formal content analysis of patient submissions were the associated time and resource cost, particularly in the light of the tight timelines of a technology appraisal. It might be the case that sometimes there would be negligible added value, especially

when the cost-effectiveness evidence strongly indicates that a technology is cost-effective (unless of course there is a view that the technology is not acceptable to patients).

3.8 The submission of new and existing qualitative research evidence

There was a lack of agreement over the extent to which the methods guide should encourage the submission of new qualitative research. Some delegates believed that such evidence would add little, because the main driver of the decision is the cost-effectiveness evidence. Unless the qualitative research evidence informs this analysis then it may not be useful. Others believed that qualitative research evidence could be useful in assessing the acceptability, appropriateness, effectiveness and utility of a technology from the patient perspective and would provide vital context for considering cost-effectiveness data.

Undertaking primary research was seen as unrealistic most of the time, but it was important that there was the opportunity to present existing qualitative research. There was a range of views as to whether it was feasible to undertake some level of (rapid) review of existing evidence during the appraisal process. However it was not established where the burden of finding this evidence and should fall. Some felt that any level of additional work would not be possible if the burden was placed on patient groups. The possibility of evidence review groups (ERGS) or Assessment Groups a rapid review was also considered, but tight deadlines and variation in the level of expertise available would place constraints on this option too (see also section 3.9 below).

There was general agreement that if qualitative research evidence was required, then any guidance regarding this evidence should not be overly prescriptive. Delegates felt that a certain minimum standard for reporting qualitative research would be required to ensure the evidence is useful to the committee. It was commented by some that poor evidence may actually harm the case being presented.

3.9 Systematic review and synthesis of qualitative research

Most tables agreed that syntheses of existing qualitative research could usefully contribute to technology appraisals. If there is useful qualitative evidence already 'out there' then there ought to be a means by which it can be incorporated into the appraisal process. One table queried whether NICE would require syntheses of qualitative research evidence for all appraisals. If so, they felt that guidance would be useful. Otherwise, it was unclear whether it would be useful to provide guidance in the methods guide on the use of syntheses of qualitative research.

One commentator expressed the opinion that relying on syntheses of qualitative research would be yet another step removed from the patient experience. As has been noted previously, there is some power in the language that patients use to describe their conditions, and because qualitative research evidence is already one step removed from the individual patient view, further synthesis would lose the context within which the information was obtained.

Again the question of who is best placed to provide this evidence arose. Most tables agreed that there were four main options: the manufacturer, patient organisations, the ERG in single technology appraisals (STAs) and the assessment group in multiple technology appraisals (MTAs) and other independent academic organisations. The delegates identified potential difficulties with each of these groups. For example, there would be an inherent assumed bias in manufacturer provided the synthesis of qualitative research. For patient groups, the problem would be one of funding and resources, as well as a lack of early enough involvement in the appraisal process. STAs present a challenge with respect to time, the involvement of ERGs is limited to eight weeks, during which it would be difficult to undertake this additional work. It might, be feasible for Assessment Groups to conduct this additional research if they were properly resourced to do so. Some of the panellists commented that only a few of the assessment groups would have the capacity and expertise necessary to undertake these syntheses.

Finally, independent academic organisations were suggested as an alternative. It was also suggested that these could potentially be commissioned by the manufacturer. One delegate expressed a concern that if the syntheses were provided or funded by the manufacturer, then they may be regarded more sceptically than if they were provided by a patient group or other source without a commercial interest.

3.10 Incorporating qualitative evidence into the overall clinical and cost effectiveness evidence base?

One table suggested that most of the additional information that could be derived from qualitative research and syntheses of qualitative research should already be captured in the QALY and the only reason that it wasn't is because EQ-5D is deficient. If EQ-5D could be improved, then it would remove the need for formally incorporating qualitative evidence.

Another suggestion was to give each element of an appraisal a fixed weight of importance. For example, cost effectiveness 40%, clinical effectiveness 40% and patient evidence 20%. It was acknowledged that this method would lend itself to being unscientifically applied and may lead to inconsistent results.

Another table suggested that qualitative research could be used alongside utilities used in the economic models, acknowledging that general population values and those from patients are usually different. It was also suggested that qualitative research could be used to assist committees in deciding what the range of acceptable incremental cost effectiveness ratio might be for a given topic. Other delegates indicated that the committee already performs this function adequately without this additional input.

4 Key issues for consideration by Working party

1. Is the current information in the methods guide on the purpose of patient evidence adequate and complete?

2. Could more helpful guidance be given on the role of patients in critiquing the clinical and economic evidence?

3. As highlighted in briefing paper 2, the current methods guide does not clearly distinguish between qualitative evidence, qualitative research evidence and syntheses of qualitative research evidence. Discussions at the workshop suggest that Appraisal Committee members do not necessarily value formal research more highly than directly reported patient testimony as they fulfil different roles. Given this:

    a) To what extent should the methods guide require or encourage the submission of primary qualitative research?

    b) Does the methods guide need to expand on the methods of qualitative research, such as the methods for identifying, sampling and recruiting participants?

4. Qualitative evidence that already exists in the literature is frequently overlooked in current appraisals. To what extent should the methods guide encourage systematic review and synthesis of existing evidence?

5 Authors

Prepared by Janet Robertson on the basis of workshop feedback from Lizzie Amis, Richard Diaz, Anwar Jilani, Bindweep Kaur, Anju Keetharuth, Heidi Livingstone, Jon Minton, Paul Richards, Donna Rowen, Will Sullivan, Jon Tosh and Helen Tucker whose contributions are gratefully acknowledged.

February 2012

NATIONAL INSTITUTE FOR HEALTH AND CLINICAL EXCELLENCE

Briefing paper for methods review workshop on perspective

The briefing paper is written by members of the Institute's Decision Support Unit. 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 2011. We encourage all interested parties to take part in this consultation.

2 Background

2.1 The current position in the NICE Methods Guide

The current Methods Guide states that

"… the perspective on outcomes should be all direct health effects, whether for patients or, when relevant, other people (principally carers). The perspective adopted on costs should be that of the NHS and PSS. Technologies for which a substantial proportion of the costs (or cost savings) are expected to be incurred outside of the NHS and PSS, or which are associated with significant non-resource effects other than health, should be identified during the scoping stage of an appraisal. In these exceptional circumstances, information on costs to other government bodies, when these are not reflected in HRQL measures, may be reported separately from the reference-case analysis. The intention to include such data will normally be agreed with the Department of Health before finalisation of the remit."[1] *(Section 5.2.7)*

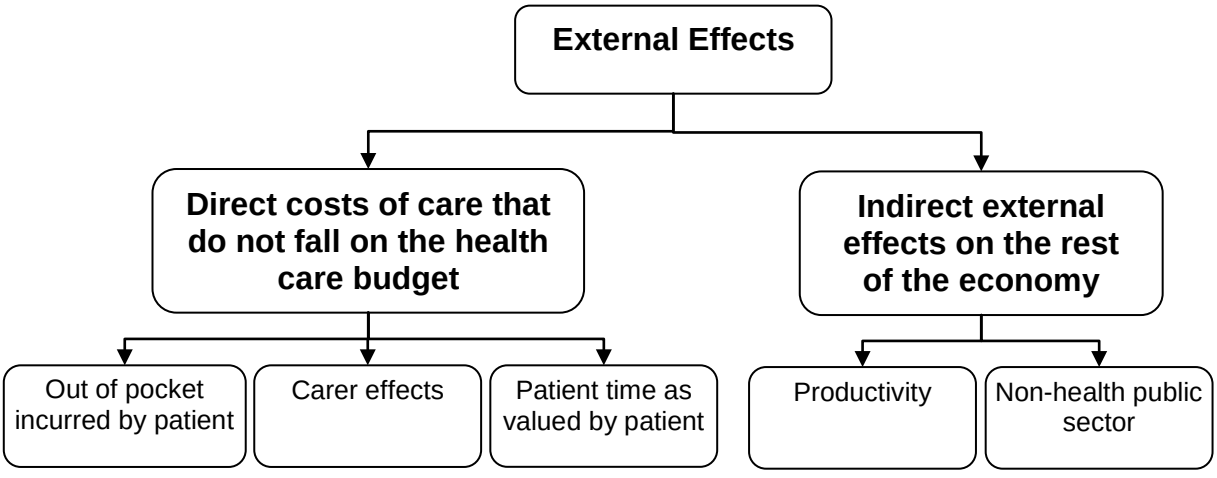
Hence the current Reference Case uses cost effectiveness analysis to compare the health benefits expected to be gained by using a technology with

the health that is likely to be forgone due to additional costs falling on the health care budget and displacing other activities that improve health. Except in the exceptional circumstances referred to above (and then only costs falling on other parts of the public sector), this approach assumes that any effects outside the health sector are small or not socially valuable compared to the effects within the health sector. Effects outside the NHS come in two general forms (see Figure 1).

2.2 Direct costs of care that do not fall on the health care budget

Some of the direct costs of care are borne by patients, such as out of pocket costs as well as their time in accessing care. It may also include the direct financial consequences of ill health (and earlier recovery) for patients and families if these are not fully captured in measures of health related quality of life (HRQoL). It can cover the time and resources devoted to caring for patients outside the health care system. These costs may be direct costs to the patient if formal (marketed) care is purchased. More often informal (non marketed) care is provided but the opportunity cost of this activity (what society loses) still needs to be valued. An effective health technology may reduce these costs (for example, a quicker recovery) or increase them (for example, prolong survival in a chronic disease state).

Figure 1. Categorising the different types of external effects



2.3 Indirect external effects on the rest of the economy

The indirect external effect on the wider economy also needs to be valued. These are effects external to patients, their families or informal carers but are valued by the rest of society. For example, returning a patient to active participation in the labour market may add to production in the economy. This will be a net benefit to society if the value of the additional production exceeds the individual's additional consumption over their remaining life expectancy. An effective health technology may provide external benefits by reducing mortality in economically active groups whose production is likely to exceed their consumption.

Table 1 summarises the approaches used to measure and value the different elements of direct and indirect external effects. It also indicates the choices and issues that exist in this regard. The table also gives examples of the types of appraisal where the different forms of external effect may be relevant.

2.4 Alternative perspectives

What alternative perspectives could NICE adopt in its decision making? The economic evaluation methods literature describes and often advocates a 'societal' perspective; that is, considering all the costs and benefits of the options being compared. It may also be considered that there are some 'middle ways'; for example, to consider costs falling only on the public sector. The problem for policy is that, in the face of budgets set by a 'higher authority' (that is, government) including the NHS budget, it is not clear how or whether a broader perspective can be implemented and reflected in NICE decisions – particularly if transfers between sectors are not regarded as a feasible option. There is also the fundamental difficultly of specifying how the trade-offs between health, consumption and other social objectives, as well as the valuation of market and non market activities, ought to be done.

Table 1. Definitions of the different components of external effects and a summary of measurement and valuation methods and issues

	Definition	Measurement	Valuation	Issues
External care effects				
Out of pocket expenses incurred by the patient or family.	Any out of pocket health care costs not covered by the NHS and falling on the patient or family. These could include: transportation costs, home improvements, additional private health care. An example of where this has been potentially relevant in appraisals: the cost of nursing homes falling on the individual patient (for example, interventions for Alzheimer's).	Monitoring of any costs incurred by the patient or family due to the patient's illness but not covered by the NHS. Can be collected prospectively using questionnaires (for example, in trials).	Based on costs recorded by patients with relevant inflation adjustment as necessary.	• A clear definition of what constitutes care costs is required; for example, do home improvements necessary to maintain a suitable quality of life come under the umbrella of care costs borne by the patient? • Possible scope for work to estimate a set of standard mean costs relating to different NHS activities (for example, GP visit, out-patient visit) or health states (for example, cost per period for given EQ5D state).

	Definition	Measurement	Valuation	Issues
Carer effects	Carer effects include any costs or benefits to the carer (formal or informal) that are not accounted for in the health budget. These are likely to be split into three sections: out of pocket, time and health effects. Example of where this has been potentially relevant in appraisals is drug therapies for Alzheimer's disease.	i) Out of pocket - as for patient/family above but falling on a carer. ii) Time: possible use of structured interviews or detailed diaries to record time inputs to care. iii) Health: possible use of EQ5D which could incorporate influence of care-giving on health. Alternative use of specific instrument for example, CarerQOL but would need to link to QALYs.	i) Out of pocket - as for patient/family above. ii) Time: for example, net market wage as a reflection of opportunity costs, reservation wage, net wage for formal carer. iii) Time can also be valued using preference elicitation methods such a conjoint analysis and contingent valuation. iv) Health can be valued in terms of QALYs which are, in principle, additive to patients' QALYs.	• Health effects on the carer are already covered in the NICE Reference Case. • If carers gain some benefits (for example, reassurance) from providing care themselves rather than employing others, then market rates may over value the true opportunity costs. • Similarly, the net wage might not represent the marginal value of a patient's leisure time as choice of working hours is often restricted. • Possibility of double counting if QALYs are used to capture health effects and additionally time is valued in monetary terms based on market rates for formal carers as the market price of a carer will include a health premium. • Potential problems in measuring the time spent providing care due to possible joint production by the carer. This occurs if the carer can undertake other activities while at the same time caring for the patient.

	Definition	Measurement	Valuation	Issues
Patient time	Patient time will incorporate all time implications of receiving health care to the patient. This includes the time taken to find or receive care. There may also be benefits to patient time if, for example, surgery reduces the time spent receiving a medical treatment. The time effects can result in forgone work or leisure time. Forgone work time has similar issues to productivity costs (see below). An example of where this may be relevant in appraisals is the development of a new product which reduces the time patient needs to spend in hospital or clinic (for example, an oral rather than intravenous medication).	Time spent identifying and consuming health care could be collected prospectively (for example, in trials).	Valuation will depend on whether leisure or work time is being forgone • Forgone leisure time likely to be considered captured in the QALY (reflected in HRQoL though, for example, the EQ5D). • The value of forgone work time due to consuming care is not expected to be captured in the QALY, but may be captured in monetary terms using similar methods as for productivity (see below).	• Cost of forgone work time may not fall on the patient, depending on nature of employment. Hence may be costs falling outside patient (for example, employer, wider economy). • Any lost work time falling on the patient could be valued at net wage.

	Definition	Measurement	Valuation	Issues
Productivity impacts on patient	Ill-health (morbidity and mortality) impacts on attendance at work and productivity whilst at work. The effect of forgone work time can fall on the patient (through reduced income and consumption) and on the wider economy. An example of where this may be relevant in appraisals is the use of a new procedure which reduces the duration of convalescence and allows patients to get back to usual activities earlier (for example, laparoscopic surgery vs. open surgery). Also, in principle, any intervention which reduces mortality risk.	Various standardised measures exist to measure changes in productivity due to morbidity, and these can be used prospectively in trials or surveys.	• The effects of ill-health on leisure time can be assumed to be captured in the QALY. • Effects on patient of lost productivity due to mortality can be assumed to be captured in the QALY (through its life years component). • Effects on patient (in terms of reduced consumption) of reduced productivity due to morbidity may be captured through the QALY. This will depend on whether responses to valuation questions (for example, for EQ5D health states) reflect possible loss of income. If not, then such effects would be captured as part of monetary valuation methods (see below).	• The main issue is whether morbidity effects on consumption can be reflected in the QALY. Recent reviews suggest that this effect is minimal with the EQ5D, in which case monetary valuation as part of the wider productivity effects would be appropriate.

	Definition	Measurement	Valuation	Issues
Non-care effects imposed on the wider economy				
Productivity impacts on those other than the patient (that is, employer, wider economy).	That proportion of the productivity effects from ill-health (mortality and morbidity) and time away from work affecting the wider economy (that is, other than the effect on the patient's consumption). In principle unpaid production should also be included. This could be a reduction in childcare and voluntary work. As above.	As for the productivity effects on patients.	There are three main alternative means of valuation (covering both morbidity and mortality). i) Human capital method whereby productivity is valued at the gross wage on the assumption that this (marginal cost) equals the value of lost production (marginal revenue product) when markets are in equilibrium. ii) Friction cost method[2] which adjusts the human capital method for various factors. Most importantly, the existence of involuntary unemployment reflects the fact that market equilibrium cannot be assumed, so productivity costs are only incurred during the period it takes to replace an ill or dead or sick worker with someone from the pool of unemployed. iii) US Panel Approach[3] whereby the effect of productivity loss on the patient is (by design) captured through the QALY. Only the productivity effect on the wider economy is reflected in financial terms. These estimates will also include the proportion of value accruing to patients from which they benefit in terms of consumption. If this has been captured separately through the QALY (see above), it needs to be netted off the value of the wider effect. This is essentially the US Panel approach.	• Each method for measurement and valuation incurs its own set of issues. • Implicit equity concerns of valuing productivity if it only relates to those in paid employment. • There are similarities in the issues with how patient time costs are valued. • Are the effects of reduced productivity (mortality and morbidity) on the patient's family adequately reflected?

	Definition	Measurement	Valuation	Issues
Non-health public sector	Covers the effects (resource costs and consequences of value in the sector) on other, non-health care, parts of the public sector. It presents the implications, beyond productivity, on the wider public sector. Examples of where this may be relevant in appraisals include the possible impact on the education and criminal justice systems of parent training programmes in the management of children with conduct disorders, the impact on criminal justice of interventions to reduce opioid dependence and the effect in education costs and outcomes of cochlear implants.	Similar issues of measurement to those relating to NHS resource use. Also need to capture non-resource consequences. Can measure prospectively in trials and other studies.	For costing, it may be possible to use of standardised unit costs, micro-costing exercises etc.; similar issues to costing in health. Also need to reflect the opportunity cost of costs falling on the budget for those other sectors. These are equivalent to the cost effectiveness threshold used by NICE. This also permits any cost or savings in these other sectors to be valued in terms of their valued outcomes.	• Few other parts of the public sector have developed a generic measure of effect such as the QALY in health care. • Similar lack of quantified cost effectiveness thresholds in other sectors. • If cost and outcome data and cost effectiveness threshold estimates are available across sectors, compensation tests can be used to assess whether interventions with costs and/or outcomes falling in different sectors are worth undertaking. This assumes some scope to adjust budgets over time.

A recent review of current UK policy and of policies adopted elsewhere reveals considerable variation in the type of perspective claimed, a lack of clarity on what constitutes a broad societal perspective and little or no consideration of the impact of fixed budgets.[4] The justification for type of policies adopted is also somewhat limited, commonly resting on literature which ignores the implications of fixed budget constraints. This lack of clarity and ambiguous terminology is also reflected in the results of an extensive review of the cost perspective adopted in published cost-effectiveness literature, with many studies claiming to take a societal perspective when in fact their analysis is restricted to the health care system.[5]

A series of challenges, therefore, presents itself to NICE is considering the appropriate perspective to adopt for technology appraisal:

- If a wider perspective is to be incorporated into the Methods Guide, should this include the full range of external effects (both direct and indirect)? Or is it possible to 'pick off' particular elements of non-NHS costs?
- What are the implications of the fact that a wider perspective would not increase the NHS budget but could effectively result in the transfer of some NHS resources to patients, their families, other parts of the public sector or the wider economy?
- There are not only external effects (costs or benefits) from new technologies, this will also be true of services that are displaced by budget re-allocations resulting from recommending new (more costly) technologies. How are these external effects from displacement to be factored in?
- Consideration of the impact of a wider perspective on other social objectives. For example, the implications would need to be assessed of including productivity costs (net of individual consumption) for older retired patients compared with younger patients active in the labour market.
- To implement a wider perspective appropriately would potentially add complexity to the analyses presented to NICE. This would be an added

challenge to critical review by assessment groups/evidence review groups and NICE.

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 Reflecting the relative value of external effects

3.1.1 Summary of the issue

Using cost-effectiveness analysis to inform NICE decisions compares the benefits expected to be gained in the health sector using QALYs to the health that is likely to be forgone due to additional costs falling on the health care budget (represented by the cost effectiveness threshold). This is the case when the incremental cost effectiveness ratio (the additional cost falling on the NHS budget, Δc_h divided by the additional health, Δh) is less than the cost effectiveness threshold, k:

$$\frac{\Delta c_h}{\Delta h} < k \qquad (1)$$

This is entirely equivalent to establishing whether the health gain from the new technology (Δh) is greater than the health forgone due to the increased cost falling on the budget (Δc_h divided by the cost effectiveness threshold):

$$\Delta h - \frac{\Delta c_h}{k} > 0 \qquad (2)$$

It is also equivalent to establishing whether the monetary value of the health gain ((∆h multiplied by the cost effectiveness threshold) is greater than the costs falling on the NHS budget:

$$k.\Delta h - \Delta c_h > 0 \quad (3)$$

As described above, this is a reasonable approach when no relevant or important effects (direct or indirect) lie outside the health sector. When external effects are considered relevant and important, it is not clear how these should be factored into the analysis and how they should impact on NICE decisions. NICE has no direct responsibility for setting the NHS budget but is charged with making decisions which use NHS resources efficiently.

When a new technology is considered cost effective in terms of health gain and NHS costs, but also generates benefits outside the health sector (for example, reducing informal care costs), a NICE decision to recommend that technology is consistent with efficiency more widely. However, there may be situations where there are clear trade-offs in the value of a technology in different sectors. For example, a new intervention for Alzheimer's disease may not be considered cost-effective in terms of health gain and NHS costs (that is, its funding would reduce net population health), but may generate net benefits outside the health sector through significant reductions in informal care costs. How should these two effects (within and outside the health sector) be traded-off?

One approach is to ignore effects outside the health sector. This may be difficult to sustain when such effects are relevant and important. Another is effectively to add these two types of effect together. That is, to express the external effects in monetary terms, add these to the costs falling on the NHS budget, relate the total net cost to the additional health gain using an ICER and compare with NICE's cost effectiveness threshold. This is inappropriate as the threshold represents opportunity costs in terms of health forgone when additional costs fall on the NHS budget but, with such an approach, not all the costs fall on the NHS budget. A third approach is to ignore the NHS budget constraint entirely and to compare an ICER made up of NHS and external

costs with some sort of 'societal willingness to pay' (that is, a value society puts on health gain expressed in terms of forgone consumption, '*v*'). However, when an NHS budget constraint actually exists, the NHS cost-effectiveness threshold ('*k*') is always relevant and cannot be ignored as it represents what is forgone in terms of health when additional costs fall on the budget.

A more feasible way of dealing with the challenge of external effects is to reflect *both* the consumption value of health, *v*, and the cost effectiveness threshold, k. We do this by expressing all the costs and benefits falling outside the health sector in terms of their positive or negative effects on society's ability to consume goods and services generally, in other words a net consumption cost, Δc_c. Now the allocation decision described in (3) can be generalised to comparing the consumption value of the health expected to be gained to the consumption value of health forgone and other net effects on consumption. The social consumption value of health, v, represents the amount of consumption that is equivalent to 1 unit of health. Within this framework, the technology should be accepted if the net consumption value is positive:

$$v.\left[\Delta h - \frac{\Delta c_h}{k}\right] - \Delta c_c > 0 \qquad (4)$$

The health expected to be gained is valued at v rather than k. But since all costs that fall on the health care budget are also health forgone these must also valued at v (the first term). Therefore, if there are no external effects ($\Delta c_c = 0$) a decision based on (3) or (4) will be the same irrespective of the value of v. When there are no external effects, maximising health or maximising the consumption value of health leads to the same decision: the value of v and whether or not $v > k$ is irrelevant, what matters for the decision is the value of k.

When there are external effects ($\Delta c_c \neq 0$) the decision can be described as a comparison of the consumption value of the net health gained in the health sector (the first term) with the net consumption costs falling on the wider

economy (the second term). If the former exceeds the latter then the technology should be adopted. Alternatively and equivalently the allocation decision in (4) can be expressed in terms of health:

$$\left[\Delta h - \frac{\Delta c_h}{k}\right] - \frac{\Delta c_c}{v} > 0 \qquad (5)$$

Now the decision can be described as a comparison of the net health gained in the health sector (1st term) with the health equivalent of the net consumption costs falling on the wider economy (2nd term). If the former exceed the latter then the technology should be adopted. If $k = £20,000$ but $v = £60,000$ then costs which fall outside the health sector get one third of the weight $\left(\frac{k}{v}\right)$ of costs that fall directly on the NHS budget. This can be clearly seen when (5) is rearranged to express the decision as a comparison of an ICER, which includes both Δc_h and Δc_c, with the threshold:

$$\frac{\Delta c_h + \frac{k}{v}\Delta c_c}{\Delta h} < k \qquad (6)$$

Therefore, assuming that $v > k$, this decision rule could be interpreted as taking external effects 'into account' but not giving it the same weight as NHS costs. Although not undertaken analytically, this could be seen as equivalent to NICE's position in the 2004 Methods Guide and, in exceptional costs, in the 2008 Methods Guide. The approach assumes that either budget transfers between sectors are possible or, if not, then Δc_h must be marginal with respect to the budget, that is, incurring these additional costs will not change the cost-effectiveness threshold.

3.1.2 Discussion points

- Does the approach of weighting external costs by $\left(\frac{k}{v}\right)$ seem a practical means of dealing with the challenge of reflecting external costs in the economic analysis informing NICE decisions?

- Where would *v* come from? Would it be based on some form of survey of public preferences using, for example, a contingent or conjoint valuation method? Or should it be based on the judgement of the policy makers or members of the Appraisal Committees?
- There are likely to be some implications of reflecting effects outside the health which may be considered inappropriate (in other words, *v* is unlikely to capture everything of social value). For example, any indirect effects of new technologies on productivity (for example, through reduced time away from paid labour) are, when considered net of individuals' consumption, likely to be greater in the young than the old and, in the latter case, may well be negative. How should these implications be dealt with?
- Is the assumption that transfers of budget (for example, between the NHS and Education) tenable? If not, is it reasonable to assume that the effects of decisions regarding new technologies will be marginal on the NHS budget? How should any non-marginal effects be dealt with?

3.2 Reflecting forgone external benefits

3.2.1 Summary of the issue

The cost effectiveness threshold (*k*) represents the health forgone when additional costs are imposed on the health care budget as a result of a new technology being recommended. These opportunity costs are incurred because the NHS budget is fixed, so the only way the local NHS can fund a new intervention is by displacing (doing less of or removing) a service entirely. If the types of external effects discussed here are to be factored into cost effectiveness analysis and more formally reflected in NICE decision-making, then the implications of displacement of services for external effects needs to be considered as well as the implications of displacement for health. In other words, when a new technology is recommended, the changes in local services that ensue as a result of the need to free up funding for the new intervention will not only effect patients' health; they may also have an impact on the direct and indirect external effects which are the focus of this briefing paper. In which case two thresholds are effectively required – the standard

one reflecting forgone health and the second relating to forgone external benefits.

It will, therefore, be necessary to provide a value for this second threshold. Providing an empirical estimate of the cost-effectiveness threshold in terms of forgone health is itself challenging but recent[6] and ongoing[7] research has conceptualised this in terms of estimating the average health effect (in terms of quality-adjusted life-years - QALYs) of a small reduction in the overall budget in the NHS. A similar concept would be relevant to external effects: what would be the impact on direct and indirect external effects of a small reduction in the NHS budget as induced by the recommendation of a new technology? The routine data sources being used to quantify the threshold in terms of forgone health (including programme budgeting data and national mortality data) would not provide any empirical estimates of the change in the types of external effect of interest.

One possible approach is to estimate a relationship between a change in health (in terms of mortality and, if possible, HRQoL) and external effects. If such a stable relationship could be estimated then, for any health forgone as a result of additional expenditure on a new technology expressed through the 'standard' cost effectiveness threshold, it would be possible also to derive an estimate of external benefits forgone. There is a literature on the relationship between health and productivity[8] which could be the basis of such estimation, but it is unlikely that the full range of external effects (including patients' costs and informal care costs) would exhibit a stable relationship with health as they are likely to vary across clinical areas.

3.2.2 Discussion points

- Should consideration of external effects be symmetrical with respect to the net external benefits of the new technology and the net external benefits forgone as a result of displaced services due to increased expenditure on the new intervention?
- What are the alternative ways of estimating the opportunity cost of displaced services in terms of external benefits? How can routine data sources be used for this purpose?

- Is there likely to be a stable relationship between health and external benefits which can be estimated and used for this purpose? Is there an existing literature on this? What data sources exist to estimate it? Would it be consistent across clinical areas or would it need to differentiate by, for example, ICD classification.

If the external benefits associated with displaced services can be approximated by a relationship with the health displaced, should this also be used in considering the external effects of new technologies? Does it suggest that external effects can be given less weight because more effective technologies (in terms of health gain) are the ones which are likely to be recommended anyway?

3.3 Measuring and valuing external effects

3.3.1 Summary of the issue

Table 1 above summarises the methods available to measure and value different aspects of external effects. It also describes some of the issues and challenges relating to measurement and valuation. If NICE's perspective is broadened then its methods guidance would presumably have to define the measurement and valuation approaches the Institute would prefer to be used. There are three issues in particular which need careful consideration. The first is the valuation of carer time. Here it may be possible to use the market net wage rate since, in undistorted markets, this should reveal an individual’s marginal valuation of their time. However, this is likely to overestimate opportunity costs of most carer time. In addition, if carers gain some benefit (for example, reassurance) from providing care themselves rather than employing others, then market rates may over-value the true opportunity costs. Similarly, net wage might not represent the marginal value of a patient’s leisure time as choice of working hours is often restricted, and proposed values range from zero to the overtime wage rate. Others suggest it should depend on what time is being sacrificed to reflect the value of the different types of activities that are forgone. There are also methods which elicit carers' valuation of their own time such as conjoint analysis.

A second key methodological challenge relates to the valuation of productivity effects, both to the wider economy and to the patient directly. For example, returning a patient to active participation in the labour market will, in many circumstances, add to production in the economy. This will be a net benefit to the rest of society if the value of the additional production exceeds the individual's additional consumption over their remaining life expectancy. How to value improvements in productivity due to reduced mortality or earlier return to participation in the labour market due to improved HRQoL is a matter of debate. There are two main approaches supported in the literature: the human capital approach and the friction cost method. The human capital approach assumes that any productivity gained or lost will extend over time and should be valued based on the gross earnings of employment. Gross wages are often recommended on the basis that the gross wage in an undistorted competitive market will be equal to the social (market) value of the production (the marginal revenue product). However, some key assumptions are required: that the labour and associated product markets are competitive and undistorted and that there is no involuntary unemployment due to structural problems in sectors of the economy. Therefore, the gross wage will overestimate the value of productivity if there is unemployment in the relevant sector or if there are distortions in labour and product markets.

Others have proposed a friction cost approach to valuing productivity losses from ill health, which is based on the amount of production lost during the time it takes employers to restore the initial production levels.[9] The total friction cost will include the lost production (over a more limited time frame than human capital estimates) as well as the direct costs employers must incur to restore these initial production levels (for example, recruitment costs, training costs etc). The use of the friction cost *approach* results in much lower estimates of the value of production losses from ill health than those from the human capital approach.[2]

With respect to the patient, an important question is whether the consumption enjoyed by an individual as a consequence of improved length or quality of life is captured in estimates of HRQoL. If, when valuing health states,

respondents take account of the impact that the health state would have on their ability to work and consume, then the financial effects on the patient will already be accounted for in estimates of QALYs gained. In these circumstances adding in the additional consumption enjoyed by the patients through a human capital or friction cost approach would double-count these benefits. This is the position taken by the US panel - a multi-disciplinary group who considered best practice for economic evaluation for the US Public Health Service in 1996.[3] As described in Table 1, the US Panel argued that the value of productivity gains *to the individual patient* can and should be captured within the QALY through the values ascribed to health states by (in NICE's Reference Case) a sample of the public.

It should be noted that NICE's preferred measure of HRQL, the EQ5D, includes in its description of health states the ability to perform 'usual social role' which will include participation in the labour market and its financial implications. When valuing this health state, would individuals consider the impact of moderate or major limitations on this role on their ability to generate income in the labour market and hence enjoy consumption? Other measures of HRQL do not include social role specifically in their health sate descriptions, so they might be less likely to capture these effects in their health state valuations. The current evidence suggests consumption or income effects are not currently captured within measures of HRQL,[10-11] although this work cannot be described as definitive. In these circumstances the additional consumption enjoyed by the patient would need to be included as a benefit and set against any indirect external costs (consumption net of production). It should be clear that adding consumption as a benefit to the patient and also as a cost to the rest of society will cancel, leaving just the external value of any production as a positive benefit.

The third key issue regarding measurement and valuation relates to costs and consequences of new technologies falling in other parts of the public sector (for example, criminal justice or education). Each of these sectors has some form of budget constraint relating to its activities. As a result, there are opportunity costs (in terms of outcomes of value in those sectors which are

forgone). Therefore, additional costs falling on a non-health sector as a result of a new technology being recommended in the NHS would result in sector-specific outcomes forgone; cost savings in another sector would be of value because it would free-up resources to generate improved outcomes; and any positive or negative non-resource effect of value in these sectors (for example, changes in educational outcomes in the education sector) could be expressed in monetary terms by reflecting the budgetary cost that would need to be incurred to generate those effects.

To formally quantify these effects on other parts of the public sector methods of measurement and valuation would be needed that are comparable to those used in the NHS. These would include estimates of the resource and non-resource consequences of new technologies recommended in the NHS on other parts of the public sector and the use of standardised unit costs to value resource use in monetary terms. More of a challenge would be the need to agree which outcomes are important in each sector and the relative value between them, ideally expressed as some composite measure of outcome such as the QALY in health. More challenging still would be the need to express the opportunity cost in other sectors in terms of these outcomes using a cost effectiveness threshold similar to the one used by NICE. If these metrics were to be available it would be possible to determine whether an intervention was cost effective from a public sector perspective.[12] For example, consider an intervention which reduces opioid dependency but is not considered cost effective from the NHS perspective. Also assume it generates cost savings and improved outcomes in the criminal justice sector. If the latter are sufficient for criminal justice to compensate the NHS sufficiently to make the intervention cost effective in the NHS, whilst still leaving a net benefit in criminal justice sector, and there is some budget flexibility to allow this, then a broader public sector perspective could be implemented.

3.3.2 Discussion points

- If a broader perspective is to be used by NICE, what are the methods of measurement and valuation which will need to be defined by NICE?

- If carer's time costs are to be included in any broadening of the perspective, would a single method of valuation need to be prescribed? if so, what would it be?
- If productivity costs are considered relevant to NICE decision making, would specific methods of valuation need to be defined? If so, what would they be?
- If the costs and consequences of new medical technologies falling on other parts of the public sector are to be formally considered in a broader perspective, what methods of measurement and valuation can NICE define?

3.4 Making a broader perspective work in decision making

3.4.1 Summary of the issue

There are clearly a number of challenges to be faced if a decision is made to broaden the perspective taken by NICE in technology appraisal. Some of these are amenable to resolution through careful judgement - for example, the most appropriate means of valuing productivity effects. Some could be addressed through further research - for example, deriving composite measures of outcomes and estimating cost effectiveness thresholds for other parts of the public sector, and estimating the external effects associated with displaced health services resulting from the funding of new technologies from a finite NHS budget. However, some of the challenges are potentially intractable. The first of these is the implications of non-marginal effects of the cost of new technologies on the NHS budget such that the NHS cost effectiveness threshold and $\left(\frac{k}{v}\right)$ changes.

The second is the fact that formally defining the trade-offs in the costs and consequences of new technologies between the NHS, other parts of the public sector and the wider economy using, for example, the approach described in Section 3.1, without fully specifying a social welfare function may lead to prescriptions which conflict with other legitimate objectives of social

policy and principles of the NHS. This is particularly the case when wider considerations will inevitably lead some technologies, which would have been accepted as cost-effective from the perspective of the health care system, to be rejected. These will tend to be technologies in older populations or which offer life extension in chronic diseases where a return to productive activity is not possible. Such decisions might be very difficult to sustain if they rest on measurement and valuation of consumption benefits which are not widely accepted or if they conflict with other objectives of social policy or widely held social value judgements.

It is important to recognise that consideration of external effects would only reallocate existing NHS resources, not add to them. It would change the mix and relative priority of particular technologies; at the margin it would prioritise less effective technologies which offer net consumption benefits over more effective technologies which impose net consumption costs. Therefore, in the short run, it would reduce the overall health gains from the NHS budget. This would be more pronounced with value-based pricing as, if price flexibility is achieved, the price would be set to a level which effectively internalises external effects onto the NHS budget for a greater proportion of newly licensed pharmaceuticals. This may be desirable if all the social objectives and arguments that are relevant are encapsulated in the framework used (that is, it reflects a fully defined social welfare function), but this is unlikely to be the case. There may be sense, then, in avoiding a formalised and analytical approach to incorporating wider costs and consequences in NICE decisions. Instead a deliberative approach to handling these issues can provide a means of balancing the complex network of social objectives and constraints which can almost certainly not be defined mathematically. It is possible to characterise the 2004 (and possibly 2008) methods guidance in these terms - a Reference Case made up of NHS costs and health effects, but wider impacts taken into account in a deliberative process. As noted in Section 3.1, this can also be seen as consistent with the analytical approach of including non-NHS costs and benefits, but down-weighting them by $\left(\frac{k}{v}\right)$.

The various challenges of formalising a broader perspective - both technical and in terms of social values - may also suggest the current NICE policy of only considering non-NHS effects in circumstances where they are likely to have a major impact, is appropriate. It is quite possible to retain this position and to consider the implications of wider costs and benefits using a deliberative framework. However, there may be a need for NICE to be more specific about when these circumstances exist and the methods to be used when they do. For example, is there a need for more clarity about the fact that the external effects of a new technology can be relevant and important when there are benefits (for example, reduction in carer time) or costs (for example, costs imposed on another sector such as education)? If the existing wider perspective (public sector outside the NHS) is used when these are significant or in general, how will value in other sectors be assessed? Will it be possible to initiate some form of budget transfer if appropriate?

3.4.2 Discussion points

How can an analytical approach to reflecting the external costs and benefits of health technologies avoid decisions conflicting with other social objectives?

When reflecting broader costs and benefits, what needs to be defined terms in the Methods Guide regarding how this process would work?

What further information should be provided in the Methods Guide on the criteria used to define circumstances for incorporating wider perspectives?

4 References

1. National Institute for Health and Clinical Excellence (NICE). *Guide to the Methods of Technology Appraisal*. London: NICE, 2008.
2. Koopmanschap MA, Rutten FFH, van Ineveld BM, van Roijen L. The friction cost method of measuring the indirect costs of disease. *Journal of Health Economics* 1995;14:123-262.
3. Gold MR, Siegel JE, Russell LB, Weinstein MC. *Cost-Effectiveness in Health and Medicine*. New York: Oxford University Press, 1996.
4. Claxton K, Walker S, Palmer S, Sculpher M. Appropriate Perspectives for Health Care Decisions. CHE Research Paper 54

(http://www.york.ac.uk/inst/che/pdf/rp54.pdf). York: Centre for Health Economics, University of York, 2010.

5. Neumann P. Costing and perspective in published cost-effectiveness analysis. *Medical Care* 2009;47:S28-S32.
6. Martin S, Rice N, Smith PC. The link between health spending and health outcomes for the new English primary care trusts. London: The Health Foundation, 2009.
7. Centre for Health Economics University of York. Methods for estimation of the NICE cost-effectiveness threshold. http://www.york.ac.uk/che/research/teams/teehta/projects/methodological-research/ 2011.
8. Epstein D, Jimenez Rubio D, Smith PC, Suhrcke M. An economic framework for analysing the social determinants of health and health inequalities. Centre for Health Economics Research Paper No. 52 York: Centre for Health Economics, University of York, 2009.
9. Koopmanschap M, van Ineveld B. Towards a new approach for estimating indirect costs of disease. *Social Science and Medicine* 1992;34(9):1005-10.
10. Brouwer WBF, Koopmanschap MA, Rutten FFH. Productivity costs in cost-effectiveness analysis: numerator or denominator: a further discussion. *Health Economics* 1997;6:511-14.
11. Brouwer WBF, Koopmanschap MA, Rutten FFH. Productivity costs measurement through quality of life? A response to the recommendation of the Washington panel. *Health Economics* 1997;6:253-59.
12. Claxton K, Sculpher M, Culyer AJ. Mark versus Luke? Appropriate methods for the evaluation of public health interventions. Centre for Health Economics (CHE) Research Paper 31. York: CHE, University of York, 2007.

5 Author/s

Prepared by Mark Sculpher (Centre for Health Economics, University of York, on behalf of the Institute's Decision Support Unit, September 2011.

6 Acknowledgements

The author is grateful for comments on earlier drafts from Meindert Boysen, Carole Longson, Janet Robertson, Andrew Stevens, Paul Tappenden and Allan Wailoo.

NATIONAL INSTITUTE FOR HEALTH AND CLINICAL EXCELLENCE

Report to the Methods Review Working Party

Key issues arising from workshop on Perspective

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 Introduction

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

This report describes the key responses to these questions under a number of broad headings to assist consideration at the Working Party. Key issues for consideration by the Working Party are proposed at the end of the report.

2 Definition and inclusion of external effects

Q1: What are the relevant external effects of a new technology which could, in principle, be considered in an appraisal? Under what circumstances would these constitute costs versus benefits?

Q2: Are there any circumstances where it would be permissible for population health to be forgone (through greater costs falling on the NHS) in order to realise external benefits?

Q3: If external effects should be considered in appraisals, is it necessary for all effects identified in 1. above to be assessed? If not, which elements of external effects should be selected for consideration? Does this depend on the characteristics of the technology or patient population? What ethical principles are relevant to these considerations?

Groups generally agreed that the categories described in section 2.2, figure 1 of the briefing paper are reflective of the relevant external effects which could potentially be considered in an appraisal. Other external effects not specifically included in the briefing paper and that might be appropriate for consideration were patient experience, patient choice and care in the community (that is, current NHS objectives). However, it was noted that these external effects may already be captured in the HRQoL measures such as the EQ-5D, or could be considered within the overall decision framework.

Participants broadly agreed that, in principle, it would be reasonable for population health to be forgone in order to realise external benefits but that this should only be considered in exceptional circumstances. One representative suggested incorporating decision rules such those used for the supplementary advice on appraising life extending treatments at the end of life criteria for each of the external effects in the Methods Guide to describe the exceptional circumstances in which they each could be considered.

Groups recognised that the Guide to Methods will need to list the external effects that can be considered for inclusion in a technology appraisal, as it

does now for costs incurred outside of the NHS and PSS. This would promote the transparency, standardisation and comparability of the appraisal process across topics. Even if, overall a deliberative approach is taken when considering the impact of external effects, it was felt that a description of the preferred external effects to be considered and the evidence required should be specified as clearly as possible in the methods guide. Some described this as a ‘sub-reference case’.

Participants generally considered that selection of specific external effects will need to depend on the topic under appraisal, and that the scoping phase of an appraisal would be the best time and place to explore and, perhaps, agree them. It was suggested that consultation with a broad panel of stakeholders including ministries of education, transport, defence, justice etc. would be necessary for inclusion of costs to the public sector not directly related to health.

The majority of the participants felt that out-of-pocket expenses should be considered only in exceptional circumstances and should not be part of the routine technology appraisal process. Those who agreed for inclusion of ‘carer effects’ acknowledged that not much research have been done in this area and suggested that a conservative approach of minimum wages for working age people could be a starting point. They suggested that reassurance benefit could be assumed to be levelled out by the loss of leisure time (if not estimated at a higher wage).

Productivity was one of the components that participants found most challenging to consider for inclusion in the broadening of the perspective. The likely ethical consequences of inclusion of the impact of a new technology on the productivity of patients, and possibly their carers, was explored and participants disagreed about the appropriateness of inclusion. Participants did agree that including ‘productivity’ benefits is likely to disadvantage older cohorts of people compared with a younger cohort. It was suggested that possible equality issues should be highlighted with estimation of loss or gain of productivity rather than not doing it at all. A pragmatic approach was also suggested to consider productivity only in the appraisals where there is very

significant potential of getting working people back to work early (for example laparoscopic surgery versus laparotomy). However, participants also cautioned that only including these effects when it is likely to provide positive effects for the technology will invite criticism of being too selective.

Many of the participants felt that health related quality of life measurement already captures patients' time during illness and treatment phase and additional consideration would lead to double counting; more so if loss of productivity is also considered.

Most participants agreed that effect on non-health public sector should be considered only if there is a general willingness and formal agreement/understanding across Government sectors about budget reallocation. Some participants were sceptical about the feasibility of such arrangements. Those who favoured its inclusion suggested that it should be considered in all appraisals with simultaneous negotiations with the government about budget reallocations.

3 Opportunity cost and displacement

Q4: If external effects are considered, how should any trade-off with health be quantified? (that is, what should the 'exchange rate' be between health and external effects?) If the consumption value of health is relevant to this, how should it be estimated?

Q5: Should the external effects of displaced activities (as a result of a technology imposing additional costs on the NHS budget) be formally considered? If so, how should this be quantified?

The briefing paper discussed three methods for determining the 'exchange rate' between health and external effects; each of which was discussed at the workshop by participants.

- The first method involved expressing the external effects in monetary terms, adding these to the costs falling on the NHS budget, relating the total net cost to the additional health gain using an ICER and comparing

with NICE's cost effectiveness threshold (‘k’). In general, participants were not in favour of this ‘lumping’ approach as they acknowledged that these external costs would not fall under the NHS budget. However, some participants supported this approach stating that: ‘In NICE Committee experience so far, external effects have not had a large impact on appraisals, therefore this simple lumping approach may be practical, assuming that external effects are marginal’, and that ‘PCT’s have been considering carer costs as a proxy for external costs within NHS budgets and the lumping approach would be in line with this’.

- The second method involved comparing an ICER made up of NHS and external costs with some sort of 'societal willingness to pay' (that is, a value society puts on health gain expressed in terms of forgone consumption, 'v'). Participants indicated this was not a good decision rule as the NHS faces a budget constraint and therefore ‘k’ cannot be ignored. Moreover, issues around the estimation of ‘v’ would apply to this method as well.

- The third method involved reflecting both the consumption value of health, v, and the cost effectiveness threshold, k and comparing the net health gained in the health sector with the health equivalent of the net consumption costs falling on the wider economy. This amounts to weighting external costs by k/v and, assuming that v>k, this decision rule could be interpreted as taking external effects into account but not giving it the same weight as NHS costs. Most participants agreed that while the k/v weighting approach was reasonable, v was a difficult concept and open communication around it was very important. Some participants were concerned that quantification would result in a lack of flexibility and judgement that Appraisal Committee’s are established for. On the other hand, it was considered that a deliberative approach could potentially result in a lack of transparency and that it was very important to set out everything clearly. Participants suggested that by way of a scenario analyses, a range of k/v values could be presented to the Committee for a deliberative discussion. Most participants agreed that while the

measure of evaluation should be formalised, the decision should be deliberative. However, some participants felt that before any question on quantifying the trade-off could be addressed it would have to be assumed that all external effects can be measured in monetary terms (for example, crime) and this made them cautious about commenting.

One participant said that the Department of Health is already using a value for 'v' for cost-benefit analyses; set at £60,000 per QALY gained, and based on evidence adapted from the Department of Transports 'value of a life' work. Alternatively, participants stated that values can be informed by trade-offs of individual preferences expressed in hypothetical choices using contingent valuation or discrete choice experiments; the work by Donaldson on the social value of a QALY was specifically referred to.

It was generally felt that if external effects are to be considered for a technology which is being appraised then the external effects of displaced activities should also be considered in order to ensure a consistent approach. Some expressed the view that even if aspects of a broader perspective will be only considered on a case by case basis for individual technology appraisals, as proposed above, all aspects of an agreed broader perspective should be taken into account in quantification of the threshold of cost effectiveness. A number of attendees rejected the notion that external effects for displaced activities should be considered, on the grounds that it is impossible to know what they are and so to accurately measure them.

Attendees raised a number of issues around the feasibility of considering external effects of displaced activities. For example, there is currently a lack of data about what is disinvested following the introduction of a new technology, which would present challenges to researchers who were attempting to establish the external effects of these disinvestments.

One participant suggested that the 2004 Guide to the Methods of Technology Appraisal appeared to have provided for consideration of a broader perspective in establishing the range of cost-effectiveness ratios that reflect the opportunity cost of accepting a new technology as an effective use of NHS

resources. In the 2004 Guide, 'wider societal costs and benefits' is included in the shortlist of factors likely to inform a judgement about the acceptability of the technology as an effective use of NHS resources; noting though that the 2004 Guide also indicates that this is only expected 'where appropriate'.

It was generally felt that the methodology for assessing the external effects of displaced activities is not yet in place. It was considered by some that only if and when NICE stipulates that external effects for displaced activities should be formally considered, would more research be conducted in this area, and so would relevant methodology be developed. The following suggestions around how the external effects of displaced activities might be quantified were made:

- It was noted that in effect, at present, NICE values external effects with a value of zero for both the new technology and for any displaced technologies. It was felt by some that the external effect could as well be positive as negative and that, in fact a value of zero may well be reasonable. Other attendees expressed this same issue using different language: it was felt by a substantial number of attendees that the external costs and benefits of adopting a new technology may cancel out the external cost and benefits of any displaced activities. In this regard, there was a leaning towards the status quo.

- Some attendees thought that if the cost and/or effects of the displaced activities are substantial enough, then they will be accounted for using NICE's existing methodology.

- There was a suggestion that data could be gathered on actual displacement seen. A study looking at displaced NHS activities has already been conducted which adopts this type of approach, although one of the authors conceded that this was a challenging study in itself and raised questions over the feasibility of such a study in a wider context, looking at displaced activities both within and outside the NHS.

- It was noted that currently PCTs are specifically asked about the likely activities that might be displaced if a technology is adopted, during the process of an appraisal. Such an approach could be an option to explore in order to obtain data on displaced activities, albeit within and not outside of the NHS. Nonetheless, such an approach in itself could be far from reliable or consistent due to regional variations and uncertainties around what is being displaced.

- A suggestion was proposed that a study could be commissioned which sets out to investigate the external effects of displaced activities, possibly by disease area. The results from this study could then be applied to any technology being appraised in the specific disease area.

- Another approach which was suggested was to develop a regression equation which predicts the costs and effects of displaced activities. In such a study, some people thought that it would be necessary to use a wide range of possible variables, so as to ensure that all possible effects were captured. Other people thought that it might be possible to identify key effects to include, so to limit the number of variables in the equation.

4 Measurement and valuation

Q6: How should the various elements of external effect be measured and valued? To what extent should the NICE Methods Guide be prescriptive about these methods?

There was consensus among the participants that the measurement and valuation of external effects is a significant challenge. Many participants suggested that a conservative approach needs to be adopted. It was suggested that in practice it is inconceivable that the full integration of external effects could be included in one step, so it will be necessary to proceed in stages, perhaps based on crude assumptions at the beginning. These methods for measurement and valuation could be improved as research established more robust methodology.

Some participants felt out-of-pocket patient expenses are hard to quantify consistently and should not be considered. If they are to be considered, the methods guide should be prescriptive about what expenses could be included.

The question of whether the financial effects of ill-health on the patient is included in the QALY and the problem of double counting was discussed. It was generally felt that EQ-5D does not measure productivity or consumption of an individual very well and a more accurate picture could be obtained by use of a well-being measure which could capture broader dimensions in more detail. The difficult of converting this type of well-being measure into QALYs or monetary terms was mentioned. Some participants felt that an additional problem with well-being was the potential number of attributes and there was some discussion regarding the use of a multi-criteria approach to deal with this.

There was some discussion about the best method of valuing carers' time based on for example net market wage or the minimum wage. One group agreed that the minimum wage was probably the most practical approach as although it is conservative in value it also balanced the positive effects associated with caring but that this should be reviewed as methodology develops.

If productivity gain or loss is to be considered most felt that the "frictional cost" method of measuring it was the preferred (pragmatic but not perfect) method. Frictional cost method was suggested to be challenging from ethical point of view as it values people from their earning power. One group suggested relating the method of productivity evaluation to the indication: human capital method might be suited to long-term chronic illness and a friction-based approach to short term illness.

Many participants felt that measuring and valuing external effects on the non-health public sector was important for some technologies. It was recognised that it would be very difficult to work out parameters such as k and v for different departments. However some participants felt that using information

from the Green Book (H.M. Treasury), current methodology from the Dept. Of Health and other national health evaluation agencies, the external effects on non-health public sectors could be quantified and this could be improved as methodology develops in the future.

5 Decision making

Q7: If external effects should be considered in appraisals, should this be undertaken formally as part of the economic analysis? Or should they be considered as part of the Appraisal Committee's more general deliberation? What should the Methods Guide specify regarding any deliberative approach?

When considering the question of the extent to which external effects should form part of the formal analysis, delegates expressed a range of views encompassing both ends of the spectrum of opinion.

Those in favour of incorporating external effects into the formal analysis as part of the NICE reference case gave the following reasons:

- If an expanded perspective is to be considered then it should be done to the same standard of evidence as analyses according to the current reference case.
- Even though it accepted that methods are not fully developed and high quality evidence may not be available for all the additional components of an expanded analysis, at least all the assumptions would be explicit.
- It would be difficult for committees to be consistent in their decisions without a formal analysis. In order to have such consistency, the preferred methods of analysis incorporating external effects should be specified in the reference case.
- External effects are unlikely to be significant for every technology appraisal, but one would need to undertake a formal analysis to know that for certain.

- The deliberative process does not work – the outcome depends too much on who is at the table.

Those thought that external effects should be considered only as part of the deliberative process gave the following reasons:

- The status quo already allows for the Appraisal Committee to depart from the reference case and it has occurred very infrequently. There is nothing to stop manufacturers presenting external effects in their submissions if these are thought to be significant.

- Formal analysis adds in more complexity with much more uncertainty. Cost effectiveness analysis is as complex as necessary now.

- The mathematical approach may not produce better decisions. Formal analysis makes the decision more transparent but does not necessarily make it right.

- For a lot of technologies external effects will not impact on the analysis and the extra effort (and cost) put into the analysis would be wasted.

- Evidence for quantifying external effects for the purposes of inclusion in the analysis will be weaker than the evidence used for current reference-case analyses.

- Formal analysis may introduce opportunities for gaming by manufacturers.

Some delegates suggested a compromise between the two extremes of a formal analysis including external effects becoming the new reference case and consideration of the additional factors only in a deliberative approach.

1. Only consider inclusion of external effects in the formal analysis in cases where it is appropriate. The need for an expanded-perspective analysis could be identified at the scoping stage. A ‘sub-reference case’ could be specified for these analyses which could allow a quantification of external effects to be made, but not include them in

the baseline ICER. The main advantage of this approach was thought to be its efficiency in that the more resource-intensive approach was only used where necessary. Disadvantages included a lack of consistency between appraisals, and the lack of good information on which to make the decision at the scoping stage (it would be a 'guess' as to whether an expanded-perspective analysis was needed).

2. Those who thought that a deliberative approach was preferred nevertheless thought that deliberation should be informed and systematic. Therefore information should be sought on the external effects that are likely to impact on the decision. It was suggested that it might be useful to develop some standard ways of quantifying external effects while not necessarily including them in the analysis. This could possibly take the form of a tariff including predefined valuations for things like carer time, absence from work, travel time and costs borne by patients.

3. Some thought that limited expansion of the formal analysis was warranted, perhaps taking a 'government' perspective rather than the current reference case which is limited to NHS and personal social services. Remaining external effects would continue to be considered as part of the deliberative process.

6 Rapporteurs

Meindert Boysen, Janet Robertson and Andrew Stevens

On the basis of feedback from Anju Keetharuth, Anwar Jilani, Bernice Dillon, Claire McKenna, Clara Mukuria, Helen Starkie, Janet Robertson, Jon Tosh, Raisa Sidhu, Richard Diaz, Sarah Willis and Tess Peasgood; whose contribution is gratefully acknowledged.

7 Key issues for consideration by the Working Party

Assuming that the Directions to NICE allow for consideration of a broader perspective than only the NHS and PPS in appraising the cost effectiveness of technologies, the following key issues are to be considered:

1. Is it right to include effects outside of the health sector? That is, is it desirable to sacrifice health for non health savings and outcomes?

2. If so, does the wording used in 5.2.7 of the Guide to the Methods for Technology Appraisals allow for appropriate consideration of a broader perspective than that of the NHS in economic evaluations?

3. And if so,

    a) should the exceptional circumstances be described in more detail and/or expanded upon? And how?

    b) should the requirement for agreement with the Department of Health before inclusion be removed?

    c) should a description of all possible external effects be included? And how?

    d) should some be excluded from consideration? And why?

    e) should consideration be given to the external effects of services likely to be displaced? And how?

    f) should a wider range of stakeholders be consulted? And who should it include?

4. Once measured, and in the context of decision-making, should the external effects form part of deliberation or part of formal analyses?

5. If part of formal analyses, should;

a) the net addition of effects outside of the health sector (expressed in monetary terms) to the costs falling on the NHS and subsequently related to the additional health gain, be compared with the current threshold range used by NICE? Or,

b) the net addition of effects outside of the health sector (expressed in monetary terms) to the costs falling on the NHS and subsequently related to the additional health gain, be compared with the some sort of 'societal willingness to pay'?

c) consideration be given to reflecting both the consumption value of health ('v') and the cost effectiveness threshold ('k') by expressing all the costs and benefits falling outside the health sector in terms of their positive or negative effects on society's ability to consume goods and services generally?

How should the Guide to Methods reflect on issues of measurement of (each of) the external effects? What could be the role of technical support documents and/or evidence submission template(s).

NATIONAL INSTITUTE FOR HEALTH AND CLINICAL EXCELLENCE

Briefing paper for methods review workshop on QALY weighting

The briefing paper is written by members of the Institute's Decision Support Unit. 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

2.1 What is QALY weighting?

The Quality Adjusted Life Year (QALY) is a unit of health outcome that combines longevity and quality of life into the common metric of a year in full health. It achieves this by assigning a value to health states experienced by patients, using a scale anchored at one for full health and zero for states regarded equivalent to being dead. Negative values are assigned for states considered worse than being dead.

The QALY is the unit of outcome used in reference case cost effectiveness analyses for NICE. The additional cost per QALY gained generated from a new technology compared to best or existing NHS practice is estimated and compared against a threshold value. This allows appraisals to be conducted in a consistent manner across different disease areas with the intention that the value of benefits generated by any technology recommended by NICE is

equal to or exceeds the value of those technologies that are displaced in the NHS as a result (where the latter are reflected in the cost effectiveness threshold).

In principle, it is feasible to assign different weights to health benefits generated in different situations, whether those benefits are expressed in terms of QALYs or some other outcome measure. One may wish to assign different weights to QALYs in order to reflect societal preferences relating to issues of efficiency or equity that do not coincide with the view that "a QALY is a QALY is a QALY", the underlying view embodied by the reference case (NICE 2008a). QALY weighting can therefore be defined as any approach that incorporates into the formal assessment of cost effectiveness, weights to the benefits that are not unitary in all situations. It is this potential role for QALY weighting within the current analytical framework operated by NICE that is the focus for this paper. For issues concerning QALY weights within other analytical frameworks see the Briefing Paper on Structured Decision Making.

This paper first sets out the approach adopted in the 2008 NICE Methods Guide and Supplementary Advice issued to the Appraisals Committees in January 2009. The characteristics of patients and technologies that have been suggested as those that should attract differential weights in the existing literature is then presented. Methods for estimating weights are then described and the results from studies that have applied those methods summarised. The paper also suggests methods that could be considered in order to ensure that any adoption of weights to benefits for new technologies are also reflected in the assessment of health services displaced as a result of new guidance.

2.2 The current position in the NICE Methods Guide

The 2008 Methods Guide (NICE, 2008a) states that

> *"In the reference case, an additional QALY should receive the same weight regardless of any other characteristics of the people receiving the health benefit." (Section 5.12)*

"The estimation of QALYs, as defined in the reference case, implies a particular position regarding the comparison of health gained between individuals. Therefore, an additional QALY is of equal value regardless of other characteristics of the individuals, such as their socio-demographic details, or their pre- or post-treatment level of health. There are several unresolved methodological issues concerning how and in what circumstances to apply additional weights to QALY calculations. Until such issues are resolved, the use of differential QALY weights is not recommended as part of the reference case." (section 5.12.2)

Thus, the reference case makes a clear direction regarding the equity position of a QALY is a QALY for the analysis. However, the Methods Guide does allow other factors to be considered outside of the reference case analysis, in the appraisal of the evidence (Section 6). Section 6.1.3 highlights the need for the Appraisal Committee to take into account NICE's directions from the Secretary of State for Health including:

"The degree of clinical need of patients with the condition or disease under consideration.

The potential for long-term benefits to the NHS of innovation".

"The Appraisal Committee takes into account advice from the Institute on the appropriate approach to making scientific and social value judgements. Advice on social value judgements is informed by the work of the Citizen's Council." (Section 6.1.4)

Furthermore, Supplementary Advice issued to the Appraisals Committees in January 2009 explicitly departed from the unweighted QALY approach within the reference case framework. For treatments that extend life in patients with a short life expectancy *inter alia,* the Supplementary Advice states that the Committee will consider:

"The magnitude of the additional weight that would need to be assigned to the QALY benefits in this patient group for the cost-effectiveness of the technology to fall within the current threshold range." (Section 2.2.2)

Therefore the current approach adopted by NICE could be characterised as a hybrid approach that has parallels with cost consequences analysis. In most situations, additional weights for benefits are considered as part of the deliberative framework adopted by the NICE Appraisal Committees. Whilst there is some guidance as to the situations where such social value judgements may be appropriate, it could be argued that this lacks both transparency and consistency. For those observers outside the NICE decision making process the factors that are used to determine whether a particular characteristic will be deemed relevant to the appraisal of a specific technology, and the weights that are to be applied if it is relevant, are not always clear and may not be consistent across appraisals. In the case of end of life, the circumstances in which the additional weights are to be applied is explicit, but the weights to be applied are not.

2.3 Other information of relevance to the current approach

As highlighted in the Methods Guide, there are several other areas in which NICE provides more information on the types of judgments that are deemed to be of potential relevance to its decision making committees. NICE's Social Value Judgements (2008b) states:

> *"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)*

The document goes on to provide some detail on what those other factors should and should not be. Those that are listed as relevant factors are those mandated by the Secretary of State and appear in the Methods Guide. Those that are ruled out are "rarity", "rule of rescue", "race", "age", "Behaviour-dependent conditions" and "Socioeconomic status", *inter alia*. Only where these features influence clinical effectiveness in these subgroups or "or other reasons relating to fairness for society as a whole", can differential decisions

be made (Principle 7). Whether these same judgments are applicable for weighting QALYs as well as for considering sub-groups of patients is unclear.

Rawlins *et al.* (2010) outline six sets of circumstances where special weightings have been applied to cost effectiveness considerations by the Institute's various advisory bodies. Two of these seem clearly to be situations in which additional weight has been given to benefits because of some perception of social value (severity and end of life). It is also stated that greater priority is given to disadvantaged populations, "particularly poorer people and ethnic minorities" though this would seem to conflict with the statements in the Social Value Judgements Document. The other three situations, labelled "stakeholder persuasion", "innovation" and "Children", are all justified as being relevant because they can provide reasons to doubt that all individual level costs and benefits have been adequately captured. In the case of children it is also stated that there may be an element of additional social value.

Therefore, whilst a greater degree of transparency is emerging from these documents, there are also elements of contradiction between them. To some extent this may be inevitable because the decisions made by NICE committees are live processes with deliberate flexibility built-in. But this does also highlight a genuine concern for some stakeholders, that it is not possible to know with certainty *a priori* which specific considerations other than costs, quality and length of life will be considered relevant or to what extent.

2.4 Value Based Pricing

A new system of Value Based pricing (VBP) is due to be introduced by the Department of Health to replace the Pharmaceutical Price Regulation Scheme (PPRS) which expires at the end of 2013. Whilst the full details of the Government's proposals are as yet unknown, and it is unclear precisely how the NICE appraisals programme will feature in this new process, there are some details known about the "other factors" that may be considered within VBP.

What role considerations about VBP ought to play in consideration of the current NICE Methods Guide and which order processes and methods for NICE and the Department of Health (DH) ought to be defined is debateable. However, there presumably needs to be some degree of alignment between the two organisations, either with both considering the same aspects of "value", or with one considering only those elements relating to the unweighted cost per QALY gained.

In the VBP consultation document (Department of Health, 2010) there is a clear commitment to applying different weightings to reflect "burden of illness", "therapeutic innovation and improvement", and other unnamed wider societal benefits. Work to estimate weights that may be used in VBP is currently being undertaken by the Department of Health's Policy Research Unit in Economic Evaluation based at the Universities of Sheffield and York. This includes both studies to estimate the weights that could be applied for these specific factors and studies that empirically estimate the threshold, including with the incorporation of those same weights for services displaced (see Section 3.4).

3 Proposed issues for discussion

Having considered the current guidance provided in the Methods Guide and the Supplementary Advice relating to end of life technologies, as well as the published literature in this area and the broader 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 should attract non unitary weights?

3.1.1 **Summary of the issue**

An important issue in determining which characteristics of diseases or patients should attract non unitary weights for health benefits concerns whose preferences should be taken into account.

Since the concern here is with incorporating elements of social, as opposed to individual, values for health benefits, many have advocated that the relevant characteristics should be identified by the general public. There is a large literature that has attempted to provide empirical evidence of the views of the general public, whether using random or convenience samples. It is not the purpose of this paper to conduct a detailed review of that literature but to provide an indication of the types of issues for which there is some empirical evidence, drawing on reviews by Sassi *et al*. (2001), Schwappach (2002), Olsen *et al.* (2003), Dolan *et al.* (2005) and Stafinski *et al.* (2011).

It should also be recognised that alternative views are held. Under the current NICE approach there are a range of sources for decisions about the relevance of potential weighting criteria, as outlined in Section 2. Those directed by the Secretary of State are broad in nature whilst more specific judgements are the responsibility of the NICE Board drawing on the Citizen's Council and reflected in a Social Value Judgements document. The Appraisals Committees themselves are also expected to apply their own judgements to issues of social as well as technical value as part of the decision making process. Most would argue that majority public support for the inclusion of

some criteria in determining health care resources in the absence of considerations of the ethical foundations would be insufficient (Daniels 1998). Furthermore, these types of decisions are not those which members of the general public typically have to make. It is a challenge to design experiments that are capable of yielding meaningful responses but often also require large sample sizes.

NICE itself commissioned two large studies (co-funded with the Department of Health) that estimated the weights for various factors (Dolan *et al.* 2008, Donaldson *et al.* 2008), the choice of which was informed by existing literature, a range of qualitative research with members of the general public and surveys of NHS staff.

The Dolan *et al.* study found that members of the general public chose to diverge from QALY maximisation to some extent on the basis of age, social class, length of time with the condition, dependents, quality of life without treatment, and whether the condition was caused by NHS negligence. NHS staff indicated in survey data that they were much less willing to diverge from QALY maximisation. The Dolan *et al.* study went on to estimate weights based on the age of recipients, quality of life without treatment and responsibility for illness. They also included rarity at the request of the Institute.

The Donaldson *et al.* study included various exercises to identify potentially relevant criteria, one of which was a ranking exercise. Here it was found that the most important factors were quality of life prior to treatment, where there is no other treatment available, life expectancy before treatment, age of patients and whether the patients live a healthy lifestyle. The lowest ranked were social class, gender, whether patients are working, whether they have dependents and past consumption of healthcare. The weighting element of their study selected age (at onset and at death) and severity of illness (with and without treatment) as issues to be considered.

The literature as a whole is large and variable in terms of the key characteristics of the studies. Most are based on samples from Western,

industrialised countries but many are small in size (n<100) and made up of convenience samples of students or other groups of workers. For almost every potential characteristic that has been discussed, there are conflicting findings between studies. These differences in samples, and additional variation in design issues, need to be considered when assessing the evidence.

Age

As with the two NICE sponsored studies, age of patients is one of the most commonly considered characteristics. In part this seems to have been motivated by the prominence of the concept of the fair innings. Williams (1997) argued in favour of the fair innings concept whereby lifetime health, whether measured as life years or lifetime QALYs, should be equalised. It is based on the feeling that everyone is entitled to some “normal” span of life (e.g. three score years and ten) and anyone failing achieve this has been “cheated”.

Most studies do find that respondents are willing to apply different weights to patients differentiated by age and that health gains to the old are valued less. There is some disagreement between studies as to whether the magnitude of weights peaks in childhood or at middle age, and not all studies find respondents willing to differentiate at all (e.g. Anand and Wailoo, 2000). The Dolan *et al.* weighting study concentrated on the weights for children versus adults as broad groups whereas the Donaldson *et al.* study considered age in 20 year blocks.

In those studies that do find a willingness to prioritise the young, it is unclear to what extent respondents might be motivated by the contributions to productivity or other efficiency related factors associated with different ages and, if this is a motivation, to what extent it would be appropriate for NICE to reflect such weights given the perspective currently employed in the reference case.

Where weights have been estimated some studies suggest approximately a value of 10:1 for the values of health benefits in the most preferred (usually

childhood) to the least preferred (usually old age) (see Dolan *et al.* 2005). Values were lower in the Donaldson study. However, these empirical findings contrast with the view of the NICE Citizen's Council who considered that age should not be valued more highly in some age groups than others.

Initial severity

The severity of patient health prior to receiving treatment is an issue that has been widely considered in the literature to date. The general hypothesis motivating these studies is that there may be greater social value from treating those in severely impaired health compared to those in less severe conditions, in addition to the valuations of treatment benefits at the individual level across the spectrum of disease. The topic has been discussed by the Citizen's Council in 2008, as well as playing a prominent role both NICE funded QALY weighting studies. A review by Shah (2009) provides an overview of findings from the published literature which comprised 21 empirical studies.

Most of these studies identify support for greater weight to be applied to the health gains of those in more severely impaired health states compared to those in better health, though many of these studies have extremely small convenience samples. Again, this does not have universal support across studies.

An important issue for the existing literature is that there is often the requirement to ask respondents to consider changes in quality of life which must be described in terms of some scale with interval properties. Shah highlights that if respondents do not accept or understand the assumed properties then their responses, that are assumed to reflect preferences for treating those in severely impaired health states, may in fact be reflections of their individual valuations of changes in health states that we already assume are reflected in the QALY measure. Some studies that have investigated this specific issue also support the possibility that respondents are not providing social valuations for severity as assumed. One example of a respondent that seems to follow such a line of thinking regarding his issue is cited in Donaldson *et al.*'s preliminary qualitative work:

“I went for (choice) ‘A’ because I thought that a jump from 20% to 40% would make a huge difference, a bigger difference than from 70% to 90%. I can imagine 70% being a healthy state that you could quite easily live and not have to take too many treatments and that kind of thing, whereas 20% is pretty close to death.” (p.12)

The Dolan *et al.* (2008) study found some evidence of preferences for different weights for QALYs by severity, but the greatest weight was for those in moderately severe ill health, rather than the greatest severity group. This was also found to some extent in the Donaldson *et al.* study, although the results are sensitive to method. In particular, the relationship of starting severity to the size of the health gain from treatment (or final endpoint) must be considered.

Size of the health gain and final endpoint

Schwappach (2002) highlights several studies that indicate a general reluctance for individuals to allocate resources to those situations where patients remain in a severely impaired health state after treatment, even though there may be substantial health gains from the treatment and this was also a feature of the Donaldson *et al.* study. Dolan and Cookson (2000) report qualitative evidence that supports this finding.

Responsibility for ill health

There are a large number of studies that consider the role of responsibility for disease. Dolan *et al.* (2008) included this in their weighting study based on findings in the qualitative work, choosing to focus on ill health caused by the NHS versus that caused by the individual patient. In the published literature, many examples focus on ill health due to smoking or drinking and in general there is evidence that the public attach a lower priority where these factors are assumed to cause or contribute to the requirement for treatment. Results do however tend to vary according to the precise setting, as might be expected given the subjective nature of the concept of responsibility for ill health. In addition, there seems to be some evidence that those that do not agree responsibility is a relevant criteria disagree strongly (Schwappach, 2002).

End of life

The NICE Decision Support Unit (DSU) has recently undertaken research examining attitudes of the general population to treating patients with short life expectancies (Shah *et al.* forthcoming). Preliminary findings indicate that there is support from the general public for treating patients with short life expectancies though this is not an overwhelming majority. Furthermore, there appears to be a greater concern for quality of life improvement than survival gains in these patients.

A study that aims to estimate the weights for end of life technologies is currently in progress and will report in March 2012.

Other issues

There are a range of other issues that have been discussed in the existing literature. The issue of productivity or other social role, such as caring for young children, has been widely considered in empirical studies. Clearly, responses here may be closely aligned to those regarding age and the relevance of the current NICE perspective to this issue was highlighted earlier in this section. In general, there is little support from existing studies for differential weights explicitly based on social role (though exceptions are noted in Stafinski *et al.* (2011) and Dolan *et al.* (2008)) and less for productivity. A large number of studies have considered the relevance of socio-economic disadvantage, which in some circumstances is the compensation of lower productivity groups. Few have identified majority public support for this approach, though some based on non UK samples have found relatively large minorities supporting the view. A notable exception is the Dolan *et al.* (2008) NICE study. In addition to survey results, they found that many participants in focus group studies were willing to prioritise those in lower socio-economic groups and often argued that those in higher social classes could purchase private health care. More limited evidence exists relating to the relevance of the amount of previous healthcare consumed, time spent waiting for treatment, other issues of “merit” such as priority for war veterans, and rarity. Some of these issues are not of obvious relevance to the

types of decisions faced within NICE technology appraisals and UK evidence is concentrated around organ transplantation for others.

3.1.2 **Discussion points**

- Who should decide which criteria are relevant? (Appraisal Committee members on a case by case basis, the Institute drawing on its Citizen's Council, the general public?)
- What account, if any, should be taken of the current published plans around Value based pricing?
- If the criteria should come from existing studies of the general public, are there any particular features these studies ought to have? (setting for sample, size, sampling method)
- Which, if any, criteria should be considered relevant?

3.2 How should weights be calculated?

3.2.1 **Summary of the issue**

There are several methods available for estimating the relative weights that could be applied to candidate criteria. It is to be expected that different methods will provide different estimates, as is recognised in the health valuation literature. However, the reasons for differences are less well understood in this setting because there is a smaller literature and there are few instances where investigators have conducted studies using sufficiently consistent approaches to allow comparisons of methods to be made.

Within the two NICE funded QALY projects, three methods were adopted for the estimation of weights. All three general analytical frameworks have some degree of pedigree in the previous literature, though nearly all required methodological adaptation and development in these NICE funded studies.

The Donaldson *et al.* study considered both Discrete Choice Experiments (DCE) and a "matching" or Person Trade-Off approach. Since these were the same respondents addressing issues around some of the same criteria (age

and severity), the study is able to make more informed comparisons than is often the case.

DCE is an approach whereby respondents are presented with a series of pair-wise choices. Both of the two scenarios presented in each pair are described in terms of a number of candidate characteristics (in this case age at onset, age at death, gain in life expectancy, quality of life if untreated and gain in quality of life if treated), which are themselves described as being at one from a set of levels. Respondents are assumed to choose which of the pair they would prefer to treat based on the levels of each of the characteristics. This reveals information about the relative value of each of the characteristics and levels. By sampling an appropriate number of respondents making sufficient pairwise choices, across an appropriate subset from the set of all feasible combinations of levels, the investigator can estimate the required weights based on multivariate regression analysis of the data.

There are several issues to consider in this type of design, perhaps the most significant of which are the methods and specification of the statistical analysis and the methods used to estimate the weights from the statistical analysis. Donaldson *et al.* present two different methods for performing the latter (the “predicted probability of choice approach” and the “compensating variation (CV) approach”). As the report highlights, there is therefore uncertainty in the results related to the choice of method with the weights obtained via the CV approach generally closer to one than for the probability of choice approach.

The “matching” or Person Trade-Off approach asks respondents to consider different potential characteristics at different levels in a different format to the DCE. Respondents are asked to assess whether they prefer to treat group A or Group B where groups are initially equal in size but differ in terms of age and severity of illness prior to treatment. The size of one of the groups is then altered to find a point at which the respondent is indifferent between them. The choices provide information about how individual respondents value the differences in levels of each characteristic and, with an appropriate combination of respondents and choices, it is possible to estimate the relative

weight of one set of health benefits compared to another. The complexities of this analysis, and the assumptions underpinning the analysis are described in detail in Donaldson *et al.* As with the DCE, there are different methods of analysis that can be employed.

General findings in the Donaldson study were that the matching approach results in estimated weights that are substantially larger than those obtained via DCE methods. Whilst in the DCE the general finding was that most weights are not significantly different from unity, in the matching study there were up to four-fold differences in the value of some health benefits compared to others. There is a range of possible explanations for this outlined in the study report, including the possibility that the findings are not contradictory because of differences in the nature of the characteristics that were varied.

A third, quite different method was adopted in the Dolan *et al.* study. The approach asked respondents to make choices between pairs of scenarios where each scenario consists of two equal sized groups of people. Those groups are described in terms of life expectancy, age, severity of health condition, responsibility for ill health and rarity. These results are used to estimate two parameters of the Social Welfare Function that represent the degree of inequality aversion between groups and the strength of weight placed on the health of one group relative to the other. Together these two parameters allow the estimation of the relative value of a change in the health of one group compared to a change in another group. The choices are analysed in terms of “Adult Healthy Year Equivalents” (AHYEs), an approach which values a profile of health using the number of years in full health as an adult that would be equivalent to it. However, the calculations required to achieve such an estimation appear particularly complex and rely on a series of analytical decisions such as the functional form of the SWF, the method of scaling of pairwise choices to a cardinal scale, and the calculation method.

The work being undertaken by both the DSU funded study into weights for patients with short life expectancies and the DH sponsored work looking at weights that might inform VBP are using DCE methods.

For all methods it is important to recognise that there may be interdependencies between different characteristics such that there is no fixed weight for any particular one. Rather, the weights are dependent on the context. For example, in relation to age, Donaldson *et al.* identify a general tendency for younger patients to be favoured over older patients, except for the very young where the pattern is reversed. However, the magnitude of the age weight is simultaneously dependent on the initial severity of the condition.

3.2.2 **Discussion points**

- Is it appropriate for NICE to specify a particular analytical approach for estimating QALY weights? If so, which should it be? If not, is it appropriate to specify some of the features that should be present in a well designed study e.g. how many characteristics should be considered, how they should be specified, how should they be presented to participants, sampling issues?

3.3 How should non unitary weights be applied to the assessment of a new technology?

3.3.1 **Summary of the issue**

If there are factors for which it is deemed relevant to apply non unitary weights then there has been a tendency to think that a relatively simple mechanism could be applied in order to reflect those weights in the cost effectiveness ratio of the technology under appraisal. However, this may not be the case (Wailoo *et al.* 2009).

Certainly, it is not appropriate to adjust the threshold in order to reflect additional weight to the new technology since the threshold is intended to reflect the value of NHS activities displaced (see Section 2). In many cases this will not be purely a presentational matter but could lead to erroneous conclusions i.e. the estimate of the cost per weighted QALY gained is not guaranteed to be free of bias. Even where this is simply a matter of presentation, any adjustment to the threshold would need to be made on the threshold that itself is already adjusted to reflect the weights relevant to NHS

services that are displaced (see Section 3.4). It is therefore recommended that weights are applied to the benefits of the new technology and this is the approach that has been reflected in the End of Life Supplementary Guidance.

Whilst this might be purely a presentational matter in some cases, in many others the differences are important. For example, where technologies are deemed to meet the current EoL criteria the Appraisals Committees currently consider the magnitude of the weight that would need to be applied to the incremental QALYs gained in order to make the technology cost-effective compared to the standard threshold. However, one interpretation of the societal preferences that the EoL supplementary advice claims to reflect is that the preference is for health gains that are generated by the extension of life, not quality of life improvements. Indeed, treatments that improve quality of life but have little or no survival benefit are explicitly excluded from the supplementary guidance. However, most technologies for which the supplementary advice is relevant generate QALYs both from life extension and from quality of life improvement prior to disease progression. In this situation, it can be argued that the “end of life weight” should be applied only to part of the incremental QALY gain. Of course, other interpretations of the End of Life Guidance are perfectly feasible, but the point is to highlight how simplistic approaches to QALY weights may often need to be avoided. To apply a uniform weight to the entirety of the QALYs gained in many cases implies that the technology itself is the characteristic that is the source of social value rather than the nature of the health gains and the recipients.

There are several other of the candidate characteristics where a simple approach to QALY weighting, that is, applying additional weight or weights to all of the incremental QALY gains, may not be an appropriate reflection of societal preferences. These include situations where individual patient characteristics change over the relevant period of evaluation of costs and benefits, and those where there is heterogeneity within the licensed population. Two examples illustrate.

When considering the incorporation of weights for “age”, attention must be given to the precise valuation tasks and definitions given to respondents in the

weighting study. If “age” is intended to reflect “baseline age” followed by a stream of health benefits over time, then no additional adjustment may be necessary. However, if the weights are intended to refer to the age of the patient at the time when the health benefit is received, then their incorporation may be less straightforward. Many decision models simulate hypothetical patients over long time horizons, particularly where disease is chronic and treatments may be disease modifying. Clearly, not all QALYs accrued should receive the same weight in these situations where patients receive benefits at different ages. In this situation, there is a requirement for a breakdown of the total QALYs generated according to the ages of patients in order that appropriate weights can be applied.

The magnitude of the treatment gain is another potential criteria whereby the simple approach may lead to misleading estimates. Consider the situation in which there is a greater weight established for treatments that provide large QALY gains compared to those that provide smaller gains. If the weight is applied to the expected incremental QALYs then this ignores the distribution of those gains. In those situations where the distribution of health gains is not symmetrical then the simplistic approach will yield a biased result. Similarly, one could imagine two different technologies that generate identical mean QALY gains but one has a much more dispersed distribution than the other. Whilst the simple approach to weighting QALYs would treat the gains from both technologies identically, this would not necessarily reflect the societal preferences reflected in the weights appropriately. This could be the case even if the distributions are both symmetrical because there is no guarantee that the weights themselves are symmetrical. The issue is analogous to the rationale for using Probabilistic Sensitivity Analysis (PSA) to obtain an unbiased estimate of the expected costs and effects as recognised in Section 5.8.4 of the Methods Guide. The mean weighted QALY gain is not necessarily the same as the mean QALY gain times the mean weight. This is also analogous to some of the other parameter values typically incorporated into cost effectiveness analyses where reflecting variability is important. An example is when we wish to reflect the costs for drugs sold by vials where vials cannot be shared with weight based dosing. The mean number of vials

required is not the same as the number of vials for the patient of mean weight (see for example the Multiple Technology Appraisal of appraisal of infliximab and adalimumab for the treatment of Crohn's disease).

In the case of "magnitude of gain", the distribution of benefits is highly likely to be skewed since typically therapies fail entirely for a significant proportion of patients but may lead to extremely large benefits for small groups of patients.

The additional complexity of the calculations required to accurately estimate the expected weighted QALY gain depends on the number of levels the weights are to be applied to (e.g. are age weights simply for children vs adults or are they more continuous?) and the characteristics of the patients in the decision problem. The same factors determine how inaccurate the simple approach will be. The obvious solution is that weighted QALYs are applied directly in the decision models used to calculate costs and benefits. However, at the extreme there could be a requirement for more complex types of decision models, particularly individual sampling models. In some situations, relatively simple cohort models designed to reflect the key drivers of costs and effects will not be capable of reflecting appropriately the weights.

As highlighted in the previous section, the weights estimated in some studies (e.g. the Donaldson *et al.* study) make it clear that there is no single "weight" for a characteristics or levels within a characteristics, rather the relevant weight is dependent on the context. This further reduces the set of circumstances in which a simple adjustment to the final estimated incremental QALYs will be feasible.

In all cases, weights are estimates from finite samples that are subject to parameter uncertainty as with other inputs to the estimation of cost effectiveness. This uncertainty should also be reflected using methods described in the existing methods guide.

3.3.2 **Discussion points**

- Should explicit weights be used and incorporated into the calculation of the Incremental Cost Effectiveness Ratio (ICER) or should a deliberative process be used?

- If part of formal analysis, do weights need to be incorporated as part of the CE model or is it acceptable to make an adjustment to the total estimated incremental QALYs gained?

- Should subgroups that align to the factors that attract differential QALY weights be considered?

3.4 How should non unitary weights be applied to the assessment of NHS services displaced?

3.4.1 **Summary of the issue**

The fundamental aims of the Technology Appraisals programme and the budget constraint the NHS faces remain whether some health benefits are considered of greater social value or not. Most candidate criteria for weighting QALYs focus on aspects of the recipients, the nature of the disease or the size of the benefits. None are specific to particular technologies *per se*, with the exception of some suggested definitions of innovation, and therefore it is likely that these same criteria are of some relevance to the assessment of forgone benefits when existing NHS services are displaced as a result of positive guidance for new technologies.(see also Briefing paper on Structured Decision Making)

The threshold is designed to reflect the value of those displaced activities and QALY weights should be reflected in the calculation of the threshold in the same manner as is proposed for NICE appraised technologies. Failure to do so creates the false impression that society has a preference for new technology *per se*. This is a definition of “innovation” that some have sought to promote. The real aim must be to establish whether the weighted benefits gained exceed the weighted benefits forgone from those NHS activities displaced due to increased costs. However, whilst the principle that QALY weights potentially apply to all NHS activities is self evident, the practice of adjusting the threshold is not necessarily straightforward.

Currently, a threshold range is operated by NICE and reflected in the 2008 Methods Guide. In broad terms, technologies with a credible ICER below

£20,000 can expect to be approved whilst above that level other factors become important. Above £30,000 the case needs to be increasingly strong. However, the current threshold range is not based on empirical estimates of what is displaced but has emerged over time. Note that a change in approach that explicitly incorporates many of the “other factors” into the analysis, implies that the circumstances in which the lower bound of the threshold range can be exceeded but the technology still achieve positive guidance must be diminished.

If the circumstances in which weights are applied to the benefits of NICE appraised technologies are infrequent or marginal, then the requirement to simultaneously reflect the same weights in the threshold value reduces. The precise definition of “marginal” is an empirical question but it seems reasonable to assume that the current end of life criteria would meet this definition, particularly given the requirement for small patient populations. However, many of the candidate criteria are common and likely to be relevant to all technologies, both those appraised and displaced, to some degree. For example, burden of disease, magnitude of the health gain and age of the patients will each have widespread relevance indicating they will need to be routinely reflected both in the benefits of the new technology and of those displaced.

In this situation, there may be a requirement for fairly radical departures from the current approach. It is also likely that all alternative approaches will necessarily be somewhat crude. One possibility would be to match disinvestment decisions to approvals of new technologies. The proposed disinvestment would be evaluated with a similar degree of rigour, including the incorporation of any QALY weights, in order to establish that there would be an expected gain in net health for the NHS as a result. This would have parallels to the Programme Budgeting and Marginal Analysis (PBMA) type approach often undertaken at a local level (see Structured Decision Making briefing paper).

Alternatively, a formal empirical estimation of the threshold can be performed. Current work being undertaken at the University of York is approaching this

task by estimating how changes in expenditure at a system level result in changes in expenditure, and subsequently changes in outcomes measured as life years and QALYs, across disease areas (classified by ICD codes). In principle, this type of analysis can use the same weights as are used for the assessment of the costs and benefits of new technologies. However, in practice this will be a challenge. The analyses are subject to precisely the same challenges as highlighted in Section 3.3. However, the option of overcoming these challenges by incorporating the weights directly into the cost effectiveness model is not available here. The calculations are necessarily much cruder than those undertaken for the assessment of the new technology.

3.4.2 **Discussion points**

- Should NICE routinely reflect QALY weights by adjusting the threshold for all technologies, in principle?
- If so, is there an acceptable and feasible method by which this can be done?

4 References

Anand, P. & Wailoo, A. (2000) "Utilities versus Rights to Publicly Provided Goods: Arguments and Evidence from Health Care Rationing", Economica, vol. 67, no. 268, pp. 543-577.

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

Daniels, N. (1998) "Distributive justice and the use of summary measures of population health status" in Summarizing Population Health: Directions for the development and application of population metrics, eds. M.J. Field & M.R. Gold, National Academy Press, Washington, D.C., pp. 58-71.

Dolan, P. & Cookson, R. (2000) "A qualitative study of the extent to which health gain matters when choosing between groups of patients", Health Policy, Vol. 51, no. 1, pp. 19-30.

Dolan, P., Shaw, R., Tsuchiya, A. & Williams, A. (2005) "QALY maximisation and people's preferences: a methodological review of the literature", Health Economics, vol. 14, no. 2, pp. 197-208.

Dolan, P., Edlin, R., Tsuchiya, A. *et al.* (2008) The Relative Societal Value of Health Gains to Different Beneficiaries,

Donaldson, C., *et al.* (2008) Weighting and Valuing Quality Adjusted Life Years: Preliminary Results from the Social Value of a QALY Project.

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 16th November 2011].

Olsen, J.A., Richardson, J., Dolan, P. & Menzel, P. (2003) "The moral relevance of personal characteristics in setting health care priorities", Social Science & Medicine, vol. 57, no. 7, pp. 1163-1172.

Sassi, F., Archard, L. & LeGrand, J. (2001) "Equity and the economic evaluation of healthcare", Health Technology Assessment, vol. 5, no. 3.

Schwappach, D.L. (2002) "Resource allocation, social values and the QALY: a review of the debate and empirical evidence", Health Expectations, vol. 5, no. 3, pp. 210-222.

Shah, K. (2009) Severity of illness and priority setting in healthcare: A review of the literature, Health Policy, 93(2-3), 77-84.

Shah, K., Tsuchiya, A., and Wailoo, A. (forthcoming) Valuing health at the end of life: an empirical study of public preferences, NICE Decision Support Unit

Stafinski, T., Menon, D., Marshall, D. & Caulfield, T. (2011) "Societal Values in the Allocation of Healthcare Resources: Is it All About the Health Gain?" The Patient, vol. 4, no. 4, pp. 207-225.

Williams, A. (1997) "Intergenerational equity: an exploration of the 'fair innings' argument", Health Economics, vol. 6, no. 2, pp. 117-132.

5 Author/s

Prepared by Allan Wailoo and Aki Tsuchiya, Health Economics and Decision Science, ScHARR, University of Sheffield, on behalf of the Institute's Decision Support Unit, September 2011.

6 Acknowledgements

The authors are grateful to Chris Skedgel and Anju Keetharuth for sharing their review work. We also received helpful comments on earlier drafts from Paul Tappenden, Mark Sculpher, Karl Claxton, Meindert Boysen and Janet Robertson.

NATIONAL INSTITUTE FOR HEALTH AND CLINICAL EXCELLENCE

Report to the Methods Review Working Party

Key issues arising from workshop on QALY weighting

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 workshop discussions addressed four distinct topics:

- Appropriateness of QALY weighting
- Identifying the type of criteria which should be weighted
- Determining weights for specific criteria and incorporating them into the appraisal process
- Opportunity cost issues arising from QALY weighting

Most workshop participants considered that QALY weighting should only be conducted for exceptional cases (5-10% of appraisals), and that it would

serve to achieve greater transparency around decision-making and improved consistency between appraisals. Participants suggested that NICE should look to the Government's Value Based Pricing proposal and the NHS Operating Framework when determining which criteria would be most appropriate for QALY weighting. It was also suggested that the ongoing work by the University of Sheffield to derive QALY weights for these criteria should also be considered.

There was little support among participants for making explicit changes to the current threshold in order to allow for formal consideration of QALY weighting in the cost effectiveness framework.

2 Questions posed to the workshop participants

1. Do you think that QALY weighting is reasonable for NICE to do? Should QALYs be weighted at all in NICE appraisals? If not, why?
2. Which criteria are those that should be considered most appropriate for QALY weighting? (For example, end of life, severity, children, size of benefit, unmet need, improvement and innovation, rarity, others). How should these criteria be defined?
3. Who should decide on the QALY weights? How should they be elicited?
4. How should QALY weights be incorporated into the assessment of new technologies? Should they be incorporated analytically into the economic analysis or should they be dealt with as part of the deliberative process?
5. If QALY weighting is formally incorporated into the cost-effectiveness framework, how should NICE deal with the opportunity cost issues arising from QALY weighting? Should NICE:
    I. Adjust the threshold?
    II. Apply the weights only in very rare situations such that the threshold would be largely unaffected?

III. Apply weights symmetrically within appraisals such that there are both positive and negative weights?

IV. Other?

How should NICE consider the issue of opportunity cost if QALY weights are reflected in a less formal manner?

3 Summary of the workshop discussions

3.1 Appropriateness of QALY weighting

Workshop participants acknowledged that the current deliberative process, which allows for implicit weighting, is generally sufficient and allows for flexibility in decision-making. However, some participants expressed the view that deliberation can lead to inconsistent decision making between appraisals and that QALY weighting may help achieve greater consistency and transparency.

Overall most workshop participants considered that QALY weighting should only be conducted for exceptional cases (5-10% of appraisals). In such instances, there should be a strong argument to justify any deviations from the base case (that is, any deviation from a QALY weighting of 1). Some participants also expressed the view that weights should also be applied to QALYs of displaced activity in the NHS.

3.2 Identifying the type of criteria which should be weighted

There was a general consensus that the purpose of QALY weighting is to incorporate elements about the technology that are not currently included in the QALY. Any criteria chosen for additional weighting should avoid double counting what is already included in the QALY. Some participants were also concerned that most criteria are correlated and therefore cannot be considered separately.

The participants considered that there should be some kind of ethical basis or justification behind any criteria which receive additional QALY weighting, and that criteria should be very stringently defined so that, in practice, additional

QALY weighting would only be applied in exceptional cases rather than in the majority of technology appraisals.

Participants noted it would be easier to weight some criteria (for example certain populations such as children) more than others (for example innovation). It was also noted that if additional QALY weighting were accepted for Technology Appraisals, this would inevitably have an impact on all of the other guidance producing programmes.

Which criteria should be weighted?

1. **Severity of disease**. Participants noted that severe diseases (such as cancer) are likely to be preferentially weighted if QALY weighting is introduced. There was uncertainty surrounding how severity of disease should be defined.

2. **Children**. It was noted that the public might attach more weight to children because they have a longer life expectancy; however this represents double counting as lifetime gains are already included in the QALY. Participants also noted however that even children who have a short life expectancy are still likely to have more weight attached to them by the public.

3. **Age.** Participants discussed the 'fair innings' argument that a lower QALY weighting could be applied to elderly people, although it was queried as to whether QALY weighing should also be applied to middle aged adults. Delegates noted that the majority of serious illnesses arise in people older than 55 years so they considered whether a lower weight could be applied to people over 55 years of age. Overall, there was limited support for a reduced QALY weighting in the middle aged or elderly adults.

4. **Size of benefit**. Most participants did not consider it appropriate to weight size of benefit as it is already included in the QALY gain. It was noted however the Government's Value Based Pricing proposal will consider magnitude of therapeutic improvement, and therefore participants considered that it would be important for NICE to ensure

that the Committee considers criteria which are consistent with the Government's proposal.

5. **Personal responsibility**. Most delegates did not consider it appropriate to negatively weight conditions which are associated with lifestyle choices, for example smoking, alcohol or drug misuse or obesity. There was recognition that it is difficult to prove a causal link between some activities and diseases.

6. **Rarity.** None of the participants considered it appropriate to give additional weight to orphan or ultra orphan diseases. Although it was noted that the objective of weighting rare disease was to incentivise drug development, participants did not think it appropriate to do this by QALY weighting.

7. **Innovation.** None of the participants considered it appropriate to weight innovation because they considered that a 'step change' in the management of a disease should already be captured (to some extent) in the QALY calculation. The participants noted that some of the benefits of innovation may be captured in other criteria such as unmet need. They also noted that often the innovative nature of a technology has no impact on the patient beyond what is already measured in the QALY. Although participants did not consider it appropriate to weight innovation, they did consider it useful for innovation to be taken into consideration in the Committee's deliberations.

8. **Unmet need.** None of the participants considered it appropriate to weight unmet need. It was noted however the Government's Value Based Pricing proposal will consider magnitude of therapeutic improvement, and therefore participants considered that it would be important for NICE to ensure that the Committee considers criteria which are consistent with the Government's proposal.

9. **End of life.** Participants noted that although extension to life is already weighted through the end-of-life criteria, it is quality of life which is often more important to patients. They considered this to be particularly important when the impact of recommending a technology could mean

that the provision of palliative care is displaced. There were several participants who expressed their dissatisfaction with the way in which the current end-of-life criteria were added to the Methods Guide as a supplement. Some participants suggested that the existing QALY weighting for end-of-life treatments should be removed from the methods guide. Instead, it should be replaced by a more evidence-based approach (such as weighting based on disease severity).

10. **Patient preference and the process of care**. Two of the five groups thought that patient preference should be weighted as it is not reflected in the QALY gains. This could include changes to the delivery of care (to reduce anxiety to the patient, or treatment which fits in better with family life), or a weight which is attached to a certain type of treatment (for example less invasive treatments compared with standard practice). Participants thought that although these issues are captured during the Committee's deliberations, they could be weighted to ensure that they are consistently addressed in all appraisals. This would also bring the Committee's approach in line with the NHS operating framework which places a lot of emphasis on patient preference and process of care are.

How should criteria be defined?

Participants expressed confusion about how the criteria should be defined. However, they were unanimous that once criteria are selected they should be clearly defined along with their trigger points.

The groups considered how QALY weighting would be applied to severity of disease. One suggestion was that an audit of all technology appraisals conducted to date could be undertaken to rank diseases by severity. An arbitrary cut off could then be applied so that the top 5-10% of diseases were considered to meet the criteria for severity. Another suggestion was that a study should be commissioned to elicit societal preferences for weighting different severities of disease. There was also a suggestion that people with health states which are considered to be 'worse than death' were a special case which required special consideration for QALY weighting. Some

participants considered that the criteria for disease severity should be binary (yes or no) rather than a continuous scale in which different diseases could have a different weighting applied based on differing severities.

When discussing how additional QALY weighting would be applied to children, several issues were raised. For example, how should children be defined and should there be a different weighting applied to different ages of children? Some participants were concerned that a child aged 17 years might receive a different QALY weighting to that which would be applied when they turn 18 and become an adult. Participants concluded that if children were to have a different QALY weighting to adults, then it would be important to first commission a study to evaluate the societal preference of QALY weighting in children.

3.3 Determining weights for specific criteria and incorporating them into the appraisal process

Who should decide on the QALY weights and how should they be elicited?

Some participants considered that the decision over whether or not there should be QALY weighting was not for them to answer. Many suggested that it should be a political decision (by the Health Minister for example) which is then left to NICE to implement as appropriate. Some participants questioned whether the Committee was qualified to make the social judgements that may be required should QALY weighting be implemented.

Many participants acknowledged that studies to estimate the weights for specific criteria are already being undertaken (by the University of Sheffield) to support the government's Value Based Pricing proposal. It was suggested that this research could also be used to inform the QALY weights which could be applied by the Committee when assessing new technologies.

Four approaches to determining QALY weights arose from the discussions:

1. **Using population-level preferences**– this view emphasised the need to reflect the views of society based on a random sample of the general population. However, a number of weaknesses with this

approach were identified including lack of consistency in the results depending on how questions are asked, and which survey tool is used. To minimise systematic biases, one recommendation was to conduct higher quality surveys to yield more reliable responses.

There was consensus that the most appropriate method to elicit population preferences was not clear and that all the different methods currently available have limitations. However, the Discrete Choice Experiments (DCE) approach was considered to be the least worst option despite issues around consistency and bias. One view expressed by participants was to accept the imperfections of the DCE, in the same way that the EQ-5D is used to estimate quality of life despite its limitations.

Of note, participants recognised that some societal preferences may be undesirable for NICE to adopt, for example, if they were considered to promote inequities.

2. **Using a small group of expert people (not manufacturers)** – it was considered that this approach would result in more consistent judgements but that any expert elicitation would need to be justified, transparent and subject to consultation and negotiation. The question of whether a small group of people had the authority to determine QALY weights was also raised by participants.

   There was no clear opinion regarding how the QALY weights would be elicited using this approach.

3. **Political decision** – this view emphasised that weights should be handed down to NICE from politicians, namely the Secretary of State, because participants considered that the Secretary of State is the only person with the political mandate to make the decision. This view expressed the need for politicians to be explicit when they prioritise one group over another.

4. **Mixed approach** – because of the limitations associated with using a single approach, a common view was to have a mixed approach consisting of:
    - public preference followed by expert adjustment; or
    - a political decision informed by general population preferences

How should QALY weights be incorporated into the appraisal process?

Participants noted that the feasibility of incorporating QALY weights into the appraisal process would be dependent on obtaining reliable, validated QALY weight estimates, the number and complexity of criteria for consideration; and whether the criteria are binary or continuous variables.

Citing DH-EEPRU's work on burden of illness with respect to the Value Based Pricing consultation, participants noted severity as the most likely (measurable) criteria that could be included as a QALY weight for future technology assessments. Due to the nature of a decision analytic model, a severity QALY weight may be a relatively simple addition to the model. Participants noted that for criteria such as unmet need and innovation, obtaining a reliable QALY weight may be difficult; therefore, these may be better suited as context-specific discussions through the deliberative process.

Participants cited concerns over a potential increase in the complexity of the analytical models if QALY weights are formally incorporated into them, and the potential shift towards patient-level modelling, which would impose resource issues to ERGs. For example, if the criteria are continuous variables (for example, for age) the QALY weight would, hypothetically, be applied according to the distribution of simulated patients entering the model, to estimate an unbiased estimate of the mean weighted QALY gain. This would increase the level of work and complexity of the model, and some participants voiced concerns that it may complicate interpretation and reduce transparency. Many participants cited it would be preferable if the QALY weights were binary (e.g. child versus adult), as it would be more feasible to incorporate them into the economic analysis and present to the Committee as

a secondary or sensitivity analysis (not the primary analysis). Participants considered that if the QALY weights were small (i.e. close to 1), then there would be insufficient benefit gained from incorporating formal QALY weights into the economic analysis when considering the additional burden for sponsors, NICE, review groups and the Committee.

Participants noted that particular methodological issues may arise when trying to incorporate weights into a decision analytic model (survival analysis/uncertainty); however, several participants considered that these should not be seen as a reason to avoid formally incorporating QALY weights.

Participants generally felt that QALY-weighted analyses should be presented to the Committee for deliberation, where the Committee would comment on the validity of the QALY weights, its impact on the technology, and consider the other 'non-modelled' criteria (ex. innovation, unmet need, etc...). Participants on the whole agreed that the deliberative process was required to ensure scientific accuracy, including that the weights have been appropriately modelled, along with considering other non-modelled criteria.

Several groups noted that there is a preference for a simple process and a need for more transparency into how additional criteria would be discussed and reported by the Committee. Some participants suggested that incorporating QALY-weights would add to the complexity of the process, potentially reduce transparency and possibly contribute to inconsistency in how it would be assessed by the Committees. It was not clear if people felt strongly whether a deliberative consideration of all QALY weights would be more or less transparent than an analytical consideration of all QALY weights. One participant commented that if QALY weights are to be included, then this should be done correctly, via the analytic model (and solve any methods issues), rather than doing it simply and incorrectly. One participant cited that previous NICE appraisals set a precedent for future QALY-weightings and unspecified weights can be inferred through case law. Participants cited the example of end-of-life criteria, but noted that this criteria and weight has weak theoretical underpinnings.

Participants encouraged NICE to ensure that the QALY weight applied to end-of-life treatments is explicitly described in the NICE methods guide.

Opportunity cost issues arising from QALY weighting

….if QALY weights are formally incorporated into the process

There was general agreement among participants that if QALY weighting is formally incorporated into the cost-effectiveness framework, then it should also be reflected either formally or informally in the opportunity cost of displaced technologies.

There was also agreement among participants that the best way of dealing with the opportunity cost issues arising from QALY weighting was to apply weights symmetrically such that both QALY weights above and below 1 can be incorporated. It was agreed that, within the context of a fixed NHS budget, this was the optimal approach; that is, by allowing for higher QALY weights this will implicitly mean that more technologies will need to be displaced from the NHS budget to account for the higher 'value' being placed on the health benefits (QALY weighting > 1) for specific patient groups and so therefore, some technologies will need to be downgraded for other patient groups (QALY weighting < 1). By only applying positive QALY weights to specific technologies, the opportunity costs may exceed the benefits (QALYs) gained for a given NHS budget.

There was discussion among participants about some of the challenges of applying QALY weights symmetrically, in particular in situations where QALY weights have not already been applied to existing technologies which are to be displaced within the NHS. In addition, some participants acknowledged that the relative cost-effectiveness of many technologies, including those that should be displaced, are unknown and that some cost-effectiveness evaluations of technologies that are to be displaced (including any additional QALY weighting for these technologies) may need to be undertaken. This may involve considerable time costs. There was also general acknowledgement that NICE does not currently have a formal system in place to evaluate which technologies (currently funded in the NHS) should be displaced/disinvested when more cost-effective technologies are introduced. It

was also noted that if NICE provided more explicit advice on technologies which should be disinvested, this would be helpful for PCT-level decision-making.

Several concerns were raised about how QALY weighting will be implemented. Specifically, participants noted that negative QALY weights will need to be applied to some patient groups (e.g. healthier patients, less severe disease/illness) which will be politically unpopular. There were also concerns raised about the technical difficulties involved in ensuring that the positive and negative weights are equally offset across all technologies that are appraised (hat is, that it must sum to zero).

Overall, there was little support among participants for making explicit changes to the current threshold in order to allow for formal consideration of QALY weighting into the cost effectiveness framework. There were some participants who argued that if the current cost-effectiveness threshold was adjusted downwards or upwards to reflect the current NHS budget (that is, to reflect the opportunity cost of technologies displaced by new, more cost effective technologies), this may reduce the need for QALY weighting. For example, if the threshold was adjusted downwards then it may be possible to only apply positive QALY weights where necessary. However, this would require full knowledge of the costs and QALYs from all technologies funded within the NHS, which currently does not exist. There was general awareness that the current cost-effectiveness threshold has no formal empirical basis and that there is ongoing research (Claxton et al., York University) that will attempt to derive a formal, empirical estimate of the threshold.

…. if QALY weights are reflected in a less formal manner?

Some participants argued that if QALY weights are applied less frequently (that is, only in special cases) then the issue of opportunity cost and applying simultaneous negative weights may be less important, as the overall impact may be negligible and, as a consequence, a formal adjustment to the cost-effectiveness threshold would not be necessary. Subsequent to this argument, it was noted that ‘end of life’ criteria, which involves QALY weighting are applied frequently in many appraisals for advanced cancers.

There was concern among some participants that, if done on a deliberative, case-by-case basis, then judgements on QALY weights made by each individual committee may lack transparency and consistency. There was some acknowledgement of a trade-off between a more formal, transparent approach (that could be applied on a uniform basis across all committees) and a more flexible, ad-hoc approach (which would allow individual committees to estimate their own QALY weights). The general consensus was in favour of the former approach.

4 Key issues for consideration by Working party

1. Can the Methods Guide describe how the QALY weights will be applied to additional criteria and what influence they should have on decision making?
2. Will it be possible to include in the Methods Guide an explicit list of criteria together with their respective weights?
3. What are the benefits of formally including QALY weighting into the appraisal process through an algorithmic approach rather than just through deliberation?
4. How should a decision be made about which criteria should have QALY weights applied to them?

5 Authors

Prepared by Fiona Rinaldi on the basis of workshop feedback from

Helen Starkie, Claire McKenna, Jeshika Singh, Eleanor Donegan, Jon Minton, Zoe Charles, Jon Tosh, Kumar Perampaladas, Manuel Gomes, Matthew Dyer, whose contributions are gratefully acknowledged.

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*

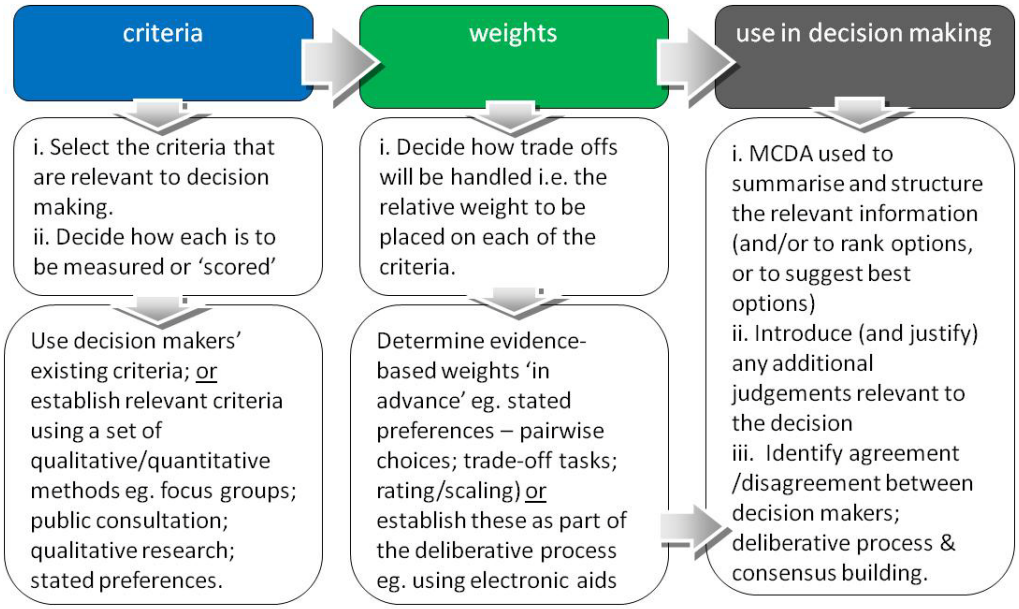
- *retention of QALYs as the principal measure of health outcome, and the ICER as the evidence on cost effectiveness, but other sorts of evidence being more formally and explicitly introduced and considered alongside these, either through MCDA or other means in a more transparent way."*

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

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*

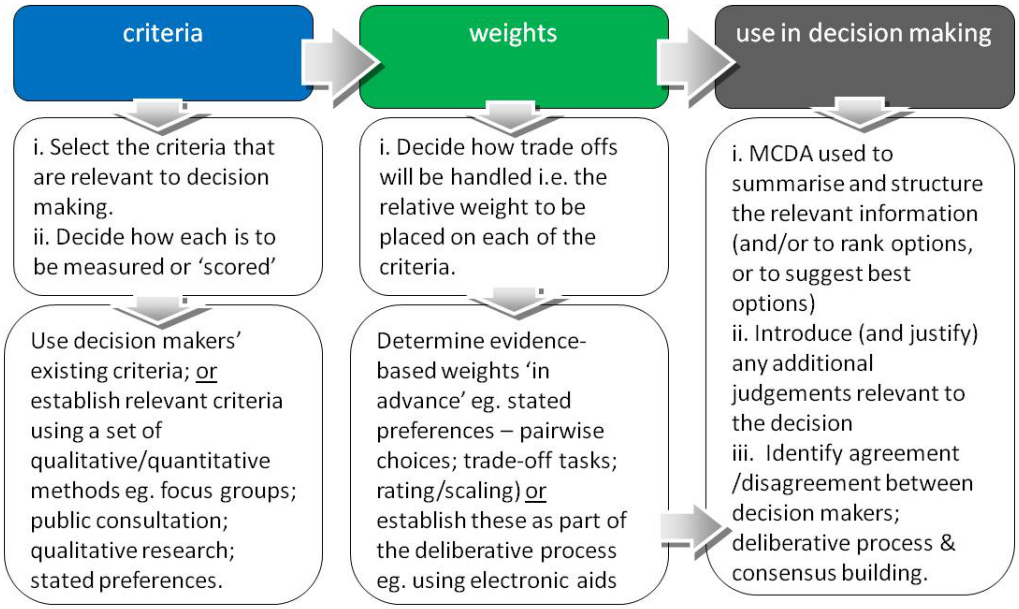
- *retention of QALYs as the principal measure of health outcome, and the ICER as the evidence on cost effectiveness, but other sorts of evidence being more formally and explicitly introduced and considered alongside these, either through MCDA or other means in a more transparent way."*

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

NATIONAL INSTITUTE FOR HEALTH AND CLINICAL EXCELLENCE

Briefing paper for methods review working party on surrogate outcomes

The briefing paper is intended to provide a brief summary of the issues that are proposed for discussion by the Methods Review Working Party 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.

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 revised versions were published in 2004 and 2008. 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 revised draft of the Methods Guide will be available for a 3-month public consultation, expected to begin in June 2012. We encourage all interested parties to take part in this consultation.

2 Background

2.1 Relevance of topic to NICE technology appraisals

The choice of outcome(s) is a key factor in any technology appraisal. In assessing the clinical and cost-effectiveness of technologies, the principal health outcome(s) should be clinically relevant, i.e. measures of health benefits and adverse effects that are important for patients and/or their carers. A clinically important (or 'final') outcome would typically include survival and/or health-related quality of life (HRQoL) that can be directly translated into quality-adjusted life years (QALYs). However, the evidence available at the time of appraisal for some (new) technologies may be solely (or largely) based on effect on surrogate outcomes (or intermediate outcomes), rather than final outcomes. In the absence of the final outcome, a surrogate outcome is defined as an outcome that is intended to both *substitute* for and *predict* the final outcome (Elston and Taylor, 2009; PBAC, 2008).

Surrogate outcomes are used as they may occur faster (than final outcomes) or may be easier to assess, thereby shortening the duration of clinical trials. In the context of In the context of health technology assessment (HTA), a surrogate outcome can include a 'biomarker' (e.g. LDL cholesterol or glycated haemoglobin [HbA_{1C}] as substitute for and predictor of future cardiovascular mortality or future major diabetic complications respectively) and also an intermediate measure of health outcome (e.g. progression-free survival as a substitute for and predictor of overall survival in cancer).

Thus, a key question for a technology appraisal, where the clinical effectiveness evidence base is principally based on a surrogate outcome, is how accurately that evidence can be used to predict the final outcomes? Or, in other words, what is the level of uncertainty associated with using a proposed surrogate outcome(s) to assess the clinical effectiveness and cost-effectiveness of a technology?

2.2 Introduction to surrogate outcomes

The use of surrogate outcomes in health policy has been controversial. Their use, at least in some applications, has led to erroneous or even harmful conclusions (Fleming and DeMets, 1996; Gotzsche *et al*, 1996).

There are a number of specific issues surrounding the use of surrogate outcomes in HTA, the first being the appropriate definition to use within this context, i.e. what meets the definition of surrogate outcome in the context of a technology appraisal? According to the US National Institutes of Health Biomarkers Definitions Working Group, a surrogate outcome is a biomarker intended to substitute for a clinical endpoint, which is ”a characteristic or variable that reflects how a patient feels or functions, or how long a patient survives” (Biomarkers Definitions Working Group, 2001). For example, the biomarkers of HbA_{1c} and LDL-cholesterol have been accepted in licensing as surrogate outcomes for risk of diabetes complications and cardiovascular disease respectively. However, a broader surrogate outcome definition is needed in the context of HTA and reimbursement that includes not only biomarkers but also what might be regarded as intermediate measures of health outcome. A common example seen in NICE appraisals is the use of the intermediate outcome of progression (or disease-free) survival to predict overall mortality (the final outcome) in cancer (Sargent *et al*., 2005; Bowater *et al*, 2008). Bone mineral density is often used as the surrogate in licensing decisions for osteoporotic treatments. However, in the context of a cost-effectiveness analysis, hip fracture risk (an intermediate outcome) may also be regarded as surrogate outcome in that it is used to substitute (and predict) for the principal health benefits related to the treatment, namely survival and HRQoL (Stevenson *et al*, 1995). Clarification at the scoping stage of an appraisal as to which outcomes are surrogate is important to inform future technical and methodological discussions for that appraisal.

A second issue is the assessment of the validity of the surrogate outcome, i.e. in a technology appraisal what evidence should be used to assess whether a proposed outcome can reasonably accepted as a surrogate outcome (or not)? A large literature has been written about the validation of surrogate outcomes,

particularly in terms of statistical approaches. In brief, three broad validity criteria have been proposed (Bucher *et al*, 1999; Lassere, 2008; Elston and Taylor, 2009):

(1) biological reasoning – is there evidence of biological plausibility of relationship between surrogate and final outcome (from pathophysiological studies and/or understanding of the disease process)?

(2) epidemiological evidence – is there evidence demonstrating a consistent association between surrogate outcome and final outcome (from epidemiological/observational studies)?

(3) trial-based evidence – is there evidence demonstrating treatment effects on the surrogate correspond to effects on final outcome (from clinical trials)? Trial-based evidence is usually not available for the specific technology in question so instead this evidence is sourced from another technology within the same class or a different technology class.

Several statistical methods have been proposed to assess these criteria, particularly for trial-based evidence (for review see Weir and Walley, 2006).

In order to appropriately assess the validity of proposed surrogate outcome in the context of a technology appraisal, a recent HTA review of surrogate outcomes has proposed that a systematic review of the evidence for each of these three criteria is needed (Elston and Taylor, 2009).

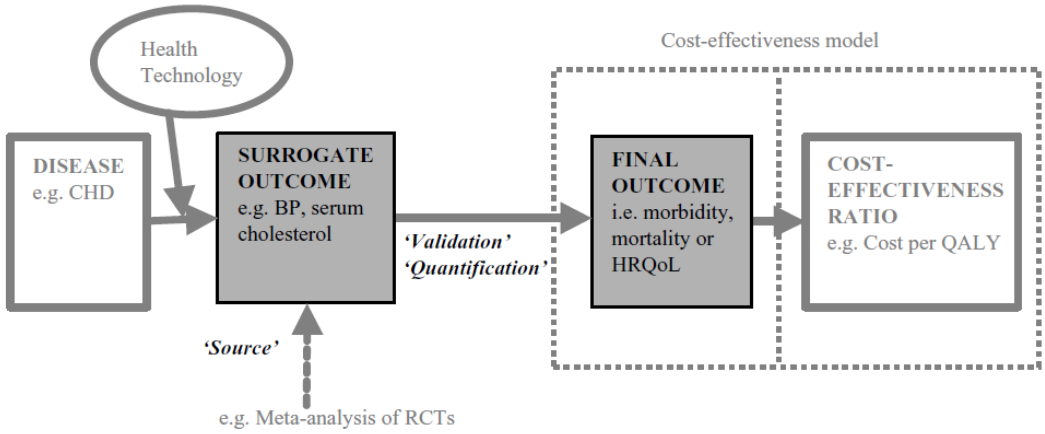
In a technology appraisal it might be expected that for an outcome to be deemed a 'valid' surrogate, it should fulfil each of the above three criteria. However, as there is currently no consensus in the HTA community on the minimum level of evidence for validation of surrogate outcomes, it could more conservatively be argued that these criteria instead need to be considered on a case-by-case basis.

The final issue relates to the prediction and *quantification* of the surrogate-final outcome relationship and how this is captured in the cost-effectiveness analysis, i.e. in a technology appraisal, how is the treatment effect on

surrogate outcome used to predict the final outcome and, thus, assess the incremental cost per QALY? As outlined above, various statistical approaches have been used to validate surrogate outcomes. In doing so, these methods effectively quantify the relationship between the treatment effect on surrogate and final outcome. For example, regression-based methods can use trial level data (meta-regression) or individual patient data from a single trial or combination of both (e.g. Johnson *et al*, 2009; Molenberghs *et al*., 2002).

Economic modelling typically involves extrapolating the clinical effectiveness evidence in order to estimate QALYs, e.g. extrapolation of trial-based observed mortality or attributing utility values to cardiovascular events observed in the trial. In doing so, such models are used to set out the potential relationship(s) between surrogate/intermediate and final endpoints (this is part of what makes them models). As such the role of surrogates is relevant to any NICE appraisal. However, in appraisals where the clinical effectiveness evidence is based only (or principally) on a surrogate outcome there is an additional element of uncertainty specifically associated with the prediction of the (unobserved) final outcome (typically survival or HRQoL) (see Figure 1). There may or may not be evidence to support such relationships. The impact of this uncertainty on cost-effectiveness needs to be fully explored, such as through the extensive use of sensitivity analyses (Elston and Taylor, 2009).

Figure 1. Schematic representation of the use of a surrogate in an HTA cost-effectiveness model (from Elston & Taylor, 2009)



2.3 What the current Methods Guide advises with respect to extrapolation and crossover

There is limited discussion on the use of surrogate outcomes in the current Methods Guide.

In the 'Suppliers of evidence, commentary and analysis' section, the methods guide says:

> *4.4.3 The written submissions […] include evidence that relates to some or all of the following. […] The identification of appropriate outcome measures and the appropriate use of surrogate outcome measures.*

In the 'Modelling methods' section, it states:

> *5.7.2 Situations when modelling is likely to be required include those where […] intermediate outcomes measures are used rather than effect on HRQoL and survival*

Furthermore, the definition of 'intermediate outcome' is given in the Glossary:

'Intermediate outcome: Outcomes that are related to the outcome of interest but may be more easily assessed within the context of a clinical study; for example, blood pressure reduction is related to the risk of a stroke.'

The methods guide also adds a 'process' consideration:

> *2.2.6 As far as possible, principal measures of health outcome are identified in the scope. For the valid analysis of clinical effectiveness, the principal outcome(s) will be clinically relevant; that is, they measure health benefits and adverse effects that are important to patients and/or their carers.*

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 by the Methods Guide Review Working Party.

- Which definition of surrogate outcome is most suitable in the technology appraisal context?
    - Should NICE's definition of surrogate outcomes be limited to biomarkers or should they include a wider category of intermediate health outcomes (e.g. fracture rate, progression free survival)?
    - Should the scoping exercise be used to clarify if the clinical effectiveness evidence in support of a technology appraisal is likely to be based principally on a surrogate outcome?

What are the potential consequences of a revision of the classical definition of surrogate outcomes in the HTA context?

- Should the methods guide require a review of the evidence to support the use of a surrogate outcome in place of a final outcome during the appraisal?
    - Does this review of evidence have to be systematic?
    - Should there be a minimum level of evidence for an outcome to be accepted as a surrogate and thereby inform the estimation of a technology's clinical effectiveness and cost-effectiveness? Should specific statistical approaches to surrogate validation be recommended/prescribed?

What could be the impact of always requesting a synthesis of evidence for the use of the surrogate outcomes in the technology appraisal process? What could be the impact of specifying a minimum level of evidence needed?

- *Should there be an explicit quantification of the uncertainty related to the use of surroga*te outcomes in the cost-effectiveness analysis?

- How should this uncertainty be estimated and presented?

What could be the impact of always requesting an explicit quantification of the uncertainty around the relationship between the surrogate and the final outcomes?

4 References

Bucher, HC, Guyatt, GH, Cook, DJ, Holbrook, A, McAlister, FA. (1999) Users' guides to the medical literature: XIX. Applying clinical trial results. A. How to use an article measuring the effect of an intervention on surrogate end points. Evidence-Based Medicine Working Group. *JAMA*, 282, 771-8.

Biomarkers Definitions Working Group (2001) Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin Pharmacol Ther,* 69, 89-95.

Bowater RJ, Bridge LJ, Lilford RJ. (2008) The relationship between progression-free and post-progression survival in treating four types of metastatic cancer. *Cancer Letters*, 272,48-53.

Elston J, Taylor RS (2009) Use of surrogate outcomes in cost-effectiveness models: a review of United Kingdom health technology assessment reports*. Int J Technol Assess Health Care*, 25, 6-13.

Fleming TR, DeMets DL. (1996) Surrogate end points in clinical trials: Are we being misled? *Ann Intern Med*, 125:605-613.

Gotzsche PC, Liberati A, Torri V, Rossetti L. (1996) Beware of surrogate outcome measures. *Int J Technol Assess Health Care*, 12:238-246.

Johnson K, Freemantle N, Anthony D, Lassere M (2009) LDL-cholesterol differences that predicts benefit in statin trials as determined by the surrogate threshold*. J Clin Epidemiol*, 63, 328-336.

Lassere MN. (2008) The Biomarker-Surrogacy Evaluation Schema: a review of the biomarker-surrogate literature and a proposal for a criterion-based,

quantitative, multidimensional hierarchical levels of evidence schema for evaluating the status of biomarkers as surrogate endpoints. *Stat Methods Med* Res, 17, 303-40.

Molenberghs G, Buyse M., Geys H, Renard D, Burzykowski T, Alonso A. (2002) Statistical challenges in the evaluation of surrogate endpoints in randomized trials. *Controlled Clinical Trials,* 23, 607-625

Pharmaceutical Advisory Benefits Committee (PBAC). *A framework for evaluating proposed surrogate measures and their use in submissions to PBAC*. December 2008 Available at: http://www.pbs.gov.au/industry/useful-resources/PBAC_feedback_files/STFOWG%20paper%20FINAL.pdf

Sargent DJ, Wieand HS, Haller DG *et al*. (2005) Disease-free survival versus overall survival as a primary end point for adjuvant colon cancer studies: individual patient data from 20,898 patients on 18 randomized trials. *J Clin Onc.* 23:8664-8670

Stevenson M, Jones ML, De Nigris E, Brewer N, Davis S, Oakley J. (2005) A systematic review and economic evaluation of alendronate, etidronate, risedronate, raloxifene and teriparatide for the prevention and treatment of postmenopausal osteoporosis. *Health Technol Assess* 9:1-160.

Weir, CJ, Walley, RJ. (2006) Statistical evaluation of biomarkers as surrogate endpoints: a literature review. *Stat Med,* 25, 183-203.

5 Authors

This briefing paper has been prepared by Oriana Ciani and Rod Taylor of the Peninsula Technology Assessment Group (PenTAG), Peninsula College of Medicine & Dentistry, University of Exeter. Thanks to Paul Tappenden for commenting on the document.

11th November 2011