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.