# Weighting and valuing quality-adjusted life-years using stated preference methods: preliminary results from the Social Value of a QALY Project

Odejar, Maria and Baker, Rachel and Ryan, Mandy and Donalson, Cam and Bateman, Ian J. and Jones-Lee, M and Lancsar, Emily and Mason, Helen and Pinto Paredes, JL and Robinson, A and Shackley, P and Smith, R and Sugdem, R and Wildman, John

University of Aberdeen HERU, Glasgow Caledonian University, University of Aberdeen HERU, Glasgow Caledonian University, University of Exeter, Newcastle University Business School, Australian National University, Glasgow Caledonian University, Glasgow Caledonian University, University of East Anglia, University of Sheffield, University of Exeter, University of East Anglia, Newcastle University Medical Sciences

1 May 2010

# Weighting and valuing quality-adjusted life-years using stated preference methods: preliminary results from the Social Value of a QALY Project

R Baker,[1,2] I Bateman,[3] C Donaldson,[1,2*] M Jones-Lee,[4] E Lancsar,[1,4] G Loomes,[5] H Mason,[1] M Odejar,[6] JL Pinto Prades,[7,8] A Robinson,[9] M Ryan,[6] P Shackley,[10] R Smith,[11] R Sugden[5] and J Wildman[4] (the SVQ Research Team)

[1]Institute of Health and Society, Newcastle University, Newcastle upon Tyne, UK
[2]Yunus Centre, Glasgow Caledonian University, Glasgow, UK
[3]School of Environmental Sciences, University of East Anglia, Norwich, UK
[4]Newcastle University Business School, Newcastle upon Tyne, UK
[5]School of Economics, University of East Anglia, Norwich, UK
[6]Health Economics Research Unit, University of Aberdeen, Aberdeen, UK
[7]Department of Economics, University Pablo de Olavide, Sevilla, Spain
[8]Fundación Centro de Estudios Andaluces, Sevilla, Spain
[9]School of Medicine, Health Policy and Practice, University of East Anglia, Norwich, UK
[10]School of Health and Related Research, University of Sheffield, Sheffield, UK
[11]Department of Public Health and Policy, London School of Hygiene & Tropical Medicine, London, UK

*Corresponding author

# Abstract

To identify characteristics of beneficiaries of health care over which relative weights should be derived and to estimate relative weights to be attached to health gains according to characteristics of recipients of these gains (relativities study); and to assess the feasibility of estimating a willingness-to-pay (WTP)-based value of a quality-adjusted life-year (QALY) (valuation study), two interview-based surveys were administered – one (for the relativities study) to a nationally representative sample of the population in England and the other (for the valuation study) to a smaller convenience sample. The two surveys were administered by the National Centre for Social Research (NatCen) in respondents' homes. Discrete choice results showed that age and severity variables did not have a strong impact on respondents' choices over and above the health (QALY) gains presented. In contrast, matching showed age and severity impacts to be strong: depending on method of aggregation, gains to some groups were weighted three to four times more highly than gains to others. Results from the WTP and SG questions were combined in different ways to arrive at values of a QALY. These vary from values which are in the vicinity of the current National Institute for Health and Clinical Excellence (NICE) threshold to extremely high values.
With respect to relative weights, more research is required to explore methodological differences with respect to age and severity weighting. On valuation, there are particular issues concerning the extent to which 'noise' and 'error' in people's responses might generate extreme and unreliable figures. Methods of aggregation and measures of central tendency were issues in both weighting and valuation procedures and require further exploration.

# Introduction

Over the past 25 years, the quality-adjusted life-year (QALY) has become the dominant measure of benefit assessment in health economic evaluation.[1,2] Its use is now widespread, particularly in the various health technology assessment agencies around the globe, and most notably in the UK through the assessment procedures undertaken by the National Institute for Health and Clinical Excellence (NICE).[3] Nevertheless, almost since the QALY was introduced to the health economics literature, the importance of the context in which health gains are produced has been discussed, raising the question of whether 'a QALY is a QALY is a QALY'.[4,5] Williams has said: 'there is nothing in the QALY approach that requires QALYs to be used in a maximising context . . . more complex rules will almost certainly be needed if collective priority-setting is to reflect the views of the general public.'[6] This issue has come to the fore again recently through NICE. A prescribed task of NICE is to assess health interventions in terms of their health gains relative to their costs, and to make recommendations as to whether or not an intervention should be adopted by the rest of the National Health Service (NHS) in England.

However, it has been recognised that the Appraisal Committee at NICE will take characteristics of beneficiaries of such interventions into account in its deliberations.[7] This raises an important policy question concerning whether quantitative estimates, reflecting the relative weight to be attached to health gains derived by different beneficiaries, can be elicited from a survey of the general public and thus be used to assist the NICE process.

Furthermore, the NICE approach involves making recommendations based on evaluations of single health-care interventions, which inevitably involve judgements about whether the QALYs gained are worthwhile.[8,9] If it is thought necessary to have such benefits and costs expressed in a common metric, usually money, this raises a second policy question: what is the monetary value of a QALY? The threshold monetary value of a QALY used by NICE was queried by the House of Commons Health Select Committee in 2007,[10] thus highlighting the importance of this policy issue.

The Social Value of a QALY (SVQ) Team was contracted from October 2004 to January 2008 to undertake two studies, each based on a survey of the population in England. The first was the 'relativities study', which had the following aims:

- to identify characteristics of beneficiaries of health care over which relative weights were to be derived; and
- to estimate the relative weights to be attached to health gains according to the characteristics of the recipients of these gains.

The second was the 'valuation study', which had the following aim:

- to assess the feasibility of deriving a monetary value of a QALY.

The rest of this report is organised into six further chapters as follows. We begin with a brief review of the literature on QALY weights and survey-based approaches to assessing the monetary value of a QALY, highlighting articles of particular interest and concluding with the challenges encountered in this project. Chapters 3–5 focus on the relativities study. In Chapter 3 we describe the methods and results of exploratory and developmental work to identify attributes and the development of a diagrammatic approach to the presentation of survey questions. We adopted two methodological approaches (which nevertheless share some commonalities) to the elicitation of preference data to be used in estimating relative weights, with respondents to the relativities survey answering both matching (or person trade-off) and discrete choice questions. The design, analysis and results of the discrete choice part of the survey are described in Chapter 4, with the same aspects of the matching part of the survey being outlined in Chapter 5. The valuation survey is described in Chapter 6. Two further introductory remarks are worth making at this point. Despite the extensive analyses undertaken to date, the results are nevertheless preliminary; for example, there is considerable scope to link the two data sets from the relativities survey, which may help to resolve some of the issues raised later in the report. An initial attempt at this is presented in Chapter 7.

Until such analyses have been completed, it is fair to say that the groups within the Team [broadly, Newcastle-led and University of East Anglia (UEA)-led] have differing perspectives on the two main approaches to the relativities work; the Newcastle-led group thinks that both exercises have their merits, while the UEA-led group stands by the methods and results from the matching study. The report, therefore, concludes with further details on these differing perspectives, as well as a brief discussion and recommendations, largely for future research as opposed to current policy.

# Weighting and valuing QALYs: literature and challenges

## Estimating relative weights for QALYs

Several authors have discussed the theoretical, ethical and practical issues around distributional weights for QALYs,[11–13] and there have been a number of attempts to estimate weights.[11,14,15]

Dolan *et al.*[16] provide the most comprehensive review in this field to date. Using a 'citation pearl growing' search strategy, they identify 78 papers, dated 2001 or earlier, of which 64 include empirical data. There is growing evidence from (mainly survey-based) studies of the general population that the number of QALYs gained is likely to be traded off against other factors. Efforts to identify these other factors have indicated a wide range of possibilities, but there remain inconsistencies and contradictory findings. The list of factors identified by Dolan *et al.* (drawn from both the empirical and theoretical literature) are as follows: age, severity of illness (or starting point health state), end point/final health state, culpability/responsibility for ill health, having dependants, socioeconomic characteristics, gender, ethnicity, inequalities in health, and the concentration or dispersion of the distribution of a fixed health gain. The reviewers conclude that despite a growing body of literature, there are contradictory findings, many studies involve small samples and few attempt to estimate weights. Earlier, Schwappach[17] described two categories of factors that could influence social value: (1) characteristics of beneficiaries, and (2) characteristics of the intervention's effect on patients' health, adding the factors of prior health consumption of patients, the duration of benefit and whether the gain in health is an improvement or the prevention of a decline.

In the period since these reviews, there have been further contributions to the literature. Two of these used discrete choice data, examining attributes such as age, culpability (e.g. related to alcohol consumption), expected length of survival, time on waiting list and whether a previous transplant had been received;[18] and lifestyle, socioeconomic status, age, life expectancy, quality of life (QoL) after treatment and level of past use of health care.[19] The results indicate that several factors, in addition to health gain, influence people's choices. However, unlike the studies described in this report, these studies were either condition specific or not based on a population sample of respondents.

Projects attempting to derive QALY weights are faced with three significant challenges: identifying characteristics of beneficiaries over which weights *should* be derived; designing and presenting questions so that respondents can understand complexities and make choices; and elicitation of quantitative preference data from members of the general public to allow the estimation of QALY weights.

Designing questionnaires that respondents can engage with was a particular challenge in this study because our aim was to estimate the relative value of different types of decontextualised, generic QALY gains. Without context, however, questions can seem overly abstract to respondents. There is also evidence that different 'types' of decontextualised QALY (e.g. life-saving or QoL-enhancing QALYs) will be regarded differently,[20,21] and so the presentation of questions needed to be flexible enough to incorporate different QALY types.

## Valuing QALYs in monetary terms

The concept of willingness to pay (WTP) has existed for a long time.[22,23] However, not until the 1980s did government Transport Departments worldwide consider using the method to value lives saved from safety projects, rather than the gross output ('productivity') approach used previously.[24] Arguably, the most natural measure of the extent of a person's preference for anything is the maximum amount that they would be willing to pay for it. Under what has naturally come to be known as the 'willingness-to-pay' approach to valuation of safety,

one seeks to establish the maximum amounts that those affected would individually be willing to pay for (typically small) improvements in their own and others' safety. These amounts are then aggregated across individuals to arrive at an overall value for the safety improvement concerned, thus reflecting society's overall resource constraint.

Estimating a WTP-based monetary value of a QALY can also be viewed as a group-aggregate WTP for marginal gains in health, at least in the case of a randomly-selected sample of the public. Indeed, this argument has been used in promoting an insurance-based approach to valuing publicly-provided health care, whereby respondents are informed of the probabilities of needing care, as well as of it being successful, before providing a valuation.[25,26]

The WTP method was first applied in health to value heart attack prevention.[27] Subsequently, there were few such studies in health, probably as a result of the view that such monetary valuation was unethical. In addition, the use of WTP to inform decisions about allocation of health care, which is supposed to be on the basis of (some notion of) need, may look problematic because WTP is obviously associated with ability to pay. However, it has been shown that this need not impede the use of WTP in health economic evaluation[28] and that, indeed, QALYs suffer from the same phenomenon.[29] Since the early 1990s, the feasibility of using WTP in health economics has again been recognised,[25,30] and more studies have been undertaken.[31,32]

Thus, in health, WTP methods historically addressed decision-making dilemmas assessing relative utility of treatments at two main levels: (1) for a given group of patients (involving elicitation of values from samples of such patients), and (2) across disparate programmes funded by geographically-defined health organisations (involving elicitation from the community of WTP values for each programme at stake). Methods have been developed which work well in terms of WTP values reflecting patient preferences.[33] In the latter area, methods have been more problematic, but are improving.[34,35] As in other public sector areas, results have been mixed on how sensitive WTP responses are to the size of the good (i.e. the health change/numbers treated) on offer to respondents[36–38] and to other aspects related to 'framing' and programme information presented to respondents.[39,40] However, innovations in valuing safety improvements derived by Carthy *et al.*[41] have shown promise in overcoming these issues. Methods based on these developments, consistent with the notion of starting with 'marginal' gains (in this case, in health), are described below. For valuing a QALY, the challenge is to start with a health gain or detriment which is not so large as to hit respondents' budget constraints, but not so small that respondents are unwilling to trade it off against a risky situation in the standard gamble (SG) type question, also involved in a procedure of the sort devised by Carthy *et al.*[41]

Through the 1990s, development of national-level technology assessment agencies led to calls for monetary values of a QALY to aid decision making at a national level.[42,43] In the UK, there has been significant debate about the empirical basis of the cost-per-QALY threshold above/below which NICE would recommend rejection/adoption of a therapy by the NHS. For example, the proceedings of the 2007 House of Commons Health Select Committee criticise the current NICE threshold on the bases that it '. . . is not based on empirical research and is not directly related to the NHS budget, nor is it at the same level as that used by PCTs [primary care trusts] in providing treatments not assessed by NICE.' Following previous literature, the implication is that, once a budget is set for the NHS (such budget setting not being the responsibility of NICE), we can infer a threshold by observing the cost per QALY of treatments which are funded vis-à-vis those that are not.[44]

Two related responses to these arguments can be made: (1) it is well known that the NHS at the local level is not systematic in how it makes such decisions, at least in economics terms;[45] and (2) because the NHS is not good at curtailing existing therapies which are poor value for money, it is not really known whether the marginal cost per QALY within the rest of the NHS is indeed out of line with (i.e. lower than) that used by NICE.

Given these significant challenges to 'discovering' a threshold, an alternative is to ask members of the public about their WTP for health gains. It may be thought that asking individuals about their WTP for such health gains from their own pockets would not be relevant to the issue of establishing a threshold value contingent on the size of the health-care budget. Indeed, Culyer *et al.*[44] (p. 57) state 'Therefore, information about how much an individual or society values improvements in health (i.e. their WTP for a QALY) is not at all relevant to

the NICE remit. These values could only be used as the appropriate threshold by NICE if it were also given responsibility to set the NHS budget.'

This is an internally consistent position. However, it does not diminish the importance of trying to establish what value(s) people actually do place on QALY gains. There are two main reasons for this.

First, in a democratic society there is a case to be made for ensuring that the government's budget-setting process should, as far as possible, be informed by the preferences of members of the public. While it is reasonable to expect negotiations between the Treasury and Department of Health to take account of a number of factors, information regarding the public's WTP for health care should arguably constitute an important consideration.

Second, a theoretical argument is that, when assessing WTP questions in surveys, if respondents think of the NHS as being at full efficiency and unable to provide more services (or QALYs) without extra payments being made, then expressed WTP amounts would be a reasonable representation of a value of a QALY at the margin for the NHS and not far removed from what a budget-holder, like a PCT, might say is the value (if PCTs used QALYs and if they behaved in an economically rational and QALY-maximising fashion!).

If the present study were to suggest that eliciting a robust monetary value of a QALY is feasible, and if a suitably representative sample survey were then undertaken, it would make a significant contribution to policy with respect to thresholds.

However, suitable existing evidence is scant and of variable quality. Some estimates have been made of the value of a QALY based either on modelling approaches or on survey research.[46,47] Modelling studies have been reviewed elsewhere and values of a QALY vary greatly depending on how the data are manipulated.[48] Moreover, survey work on the value of a QALY has been limited. Typically, individuals have been asked about their WTP for health gains for which quality adjustment factors have been obtained from another sample without fully adjusting for uncertainty (i.e. by presenting scenarios involving certain gains in QoL) and, in some cases, eliciting values from patients and not from members of the general public.[47,49,50] Only one such estimate exists for a European country.[47] The research undertaken in this study, therefore, represents a significant advance in the methods in this area

# Identification and presentation of attributes for QALY weights

## Introduction

The first objective of the relativities study, 'to identify characteristics of beneficiaries of health care over which relative weights are to be derived', requires qualitative enquiry. This precedes the quantitative estimation of the relative importance placed on those characteristics, once established. While the quantitative study is reported in two subsequent sections, reflecting two methodological approaches, these approaches share a common empirical foundation based on in-depth exploratory and developmental research.

The selection of attributes for inclusion in the relativities study is crucial. The inclusion of an attribute without a robust rationale for doing so, or the omission of an important attribute, will lead to misleading conclusions. Qualitative techniques are increasingly used to establish appropriate attributes, particularly for discrete choice studies.[51] In this study we have taken a predominantly qualitative approach and supplemented conventional methods with other techniques. This exploratory phase of the project was an iterative process, involving three waves of focus groups with members of the general public, and use of a range of methods including: open-ended discussion; simple ranking procedures; experimentation with sample questions; and a more complex ranking task involving card sorting (Q methodology). Qualitative findings were interpreted alongside the results of the other methods used.

Most of this chapter is taken up by reporting on the three main methods used to identify the most important characteristics of beneficiaries, followed by a description of the development of methods used to present information to respondents using innovative diagrammatic representations, before, finally, describing the format of the questionnaires, presented using a computer-assisted personal interview (CAPI).

## Methods

Focus groups were facilitated by two or more members of the research team depending on the size of the group. Discussions were introduced and guided by the focus group leader. Other researchers were available to help distribute materials and answer questions during individual exercises. Group discussions were recorded using a digital voice recorder and transcribed verbatim. All focus groups received an introductory description of the project and the problem at hand. This took the form of a brief presentation followed by the opportunity to ask questions. Participants were then guided through two or more tasks.

During the first wave of focus groups, we adopted open-ended qualitative techniques to elicit views and to probe responses. Respondents were advised that resources to provide health services are constrained and, as such, difficult choices must be made about the types of treatments and interventions that are provided by the NHS and, by implication, those that are not. The focus group leader introduced notions of scarcity of resources, carefully, and in simple terms (see Appendix 1). Essentially respondents were asked to accept the inevitability of rationing (although that term was not used) and that with or without the views of the general public, priority setting will happen. There was positive acceptance of these facts and respondents were comfortable proceeding on that basis. They were asked to suggest what sorts of things should be taken into account when such decisions are made.

Despite interesting discussions (a summary of which is presented below), participants often had difficulty absorbing and expressing opinions about the concepts we wanted them to explore. Generally, respondents readily proposed issues such as the size of the health gain, cost, QoL and life extension, but when asked to 'go beyond'

these concepts there was difficulty (or perhaps reluctance). If the group discussion then stalled, examples of possible issues for discussion were suggested by group facilitators, but the results of this approach raised concerns about leading respondents and endowing particular attributes with validity simply by mentioning them.

In the second wave of focus groups, we included a simple ranking task both as a means of generating some crude data and to stimulate discussion. Respondents were asked to rank order a set of 10 cards printed with issues that might be considered in priority setting, such as 'quality of life of patients before treatment' or 'the social class of patients typically affected'. Participants in these focus groups were also presented with some examples of the types of questions that would be used in the quantitative study, stimulating debate about the inclusion of different attributes as well as providing valuable information on the appropriateness of different modes of presentation.

In the third wave of focus groups, we introduced Q sort techniques which are sufficiently distinct to warrant a separate section and this follows the general findings. Q sorting involves arranging a number of cards, printed with statements about the topic, according to an instruction such as 'from most agree to most disagree'. Allowing focus groups to begin with individual Q sort activities and following this with discussion enabled respondents to express their views (via the Q sort) before engaging in discussion with others. They also entail a common stimulus set (in this case 46 cards printed with different statements about the issue at hand). These results are, therefore, unencumbered by the input of the focus group leader, or by the views of others during discussion.

Focus groups were conducted with groups of between 4 and 10 respondents and organised in three waves of data collection in Newcastle upon Tyne and Norwich between March and November 2005. Participants in the Newcastle groups were recruited through a social research company (NWA Social and Market Research) based in the north-east of England and £20 was paid to each participant in recognition of their travel expenses and time. Participants in Norwich were recruited from an existing university database of members of the general public who had consented to be contacted for research. A total of 126 respondents (42 in Norwich and 84 in Newcastle) took part in focus groups.

## Qualitative findings based on open-ended discussion

As already stated, the use of open-ended qualitative techniques was only a partial success. This may reflect the fact that respondents do not necessarily have a readily articulated account of their views on such complex issues, or that they were unsure about the kinds of things we were asking them to consider. In general, respondents were far more comfortable talking about health-related characteristics than they were discussing social or personal characteristics of the beneficiaries of health care. We do not report a complete, formal qualitative analysis because of the nature of these data, our own objectives and word limits. Instead, a summary is given of the nature of the discussion on each issue, including some brief illustrative quotes, and the results of the simple ranking exercise are included.

### Age

Age was important to participants for a variety of reasons (which are also well documented in the literature). In different accounts, the young were favoured because of: their longer life expectancy; fair innings arguments; current and future contributions to society; and productivity. This was not uncontested; others argued the deservedness of older people, whose life-long contributions to the financing of the health services should be recognised:

> I would still veer towards [the] 7 year old because the 7 year old's got all its life in front of it, whereas the 70 year old has had 70 years of life.
>
> Focus group respondent, Newcastle, May 2005

Age is also a 'proxy' for a range of other characteristics. Whether or not potential patients are economically active has already been mentioned, but patients' social and familial networks were also linked to their age. The most obvious connection made was ages when patients are likely to be in their childbearing and childrearing years.

'The average age of patients at time of illness' was ranked 4th highest of 10 in the simple ranking exercise (below cards listing QoL, life expectancy and whether or not other treatments were available).

## Dependants

In early focus groups, discussion about prioritising health care for people with dependants sparked off quite significant disagreements. As well as the view that we had expected, i.e. that some respondents might attach positive weight to health gains to people with young children or other dependants, there was also strong opposition to this view which rejected the diversion of health care funds in favour of people who had made a choice to have children:

> . .. first of all I thought that might influence my decision if they had dependants . . . but then I thought about it in another way, it seems unfair if you do have dependants that just because you've got dependants its influencing your decisions . .. that you get that intervention, so it seems unfair.
>
> Focus group respondent, Newcastle, May 2005

Such positive weighting of health gain for people with dependants was also seen as discriminatory against the childless. Others pointed out that people can be good and bad 'carers' but that favouring those with dependants seems to indicate a moral worthiness to this role in exclusion of other roles in a community:

> Making moral judgements is dodgy . . . part of it seems to be about how worthy somebody is to be given resources. Because they may have dependants but their quality of care might not be great to those dependants, not all parents are good at parenting. Whereas their value in life might be that they're very good at their job, or they're very good as a friend.
>
> Focus group respondent, Newcastle, May 2005

In the simple ranking, the average rank for 'Whether or not the patients have dependants' was 7/10.

## Lifestyle

In a similar vein to the comments on dependants, the issue of whether lifestyle (or 'culpability') should be taken into account divided respondents. The process of discussion and the views of others also seemed to affect respondents' stated views. Here, issues of choice and control, addiction and social/environmental influences were mentioned. Respondents appreciated the difficulty many addicts have in quitting, as well as the fact that, for older people, information about the risks of some behaviours had not been available (in the following quotes, different focus group respondents are distinguished using letters, e.g. respondent G, respondent B):

> G Lifestyle is a much more complex thing than just yes or no choice.
> B Yes, its what people, people like smokers they're . . .
> G Stressed or . . .
> B . . . they're stressed or living in absolute poverty and have to . . . you know, that's why they've got . . .
> K And actually smoking is very difficult to give up. It's no good saying, oh well people smoke therefore it's not right, it's very difficult to give up.
>
> Focus group discussion, Newcastle, May 2005

A small number wanted to prioritise those who take care of their own health above those who smoke or drink, but most had difficulty sustaining a logical argument in the face of disagreement. Obesity seemed to generate different views than smoking and alcohol and was used as an example of the 'slippery slope' down which such discussion can descend. The culpability argument was applied to a wide range of activities, including sports injuries, for example:

> P But it's starting on the slippery slope isn't it? Where do you draw the line?
> N It doesn't harm them drinking a little bit.
> P Going on from alcohol to diet . . . because people are fat should we penalise them?
>
> Focus group discussion, Newcastle, May 2005

Discussion about liver replacement (which usually centred on the 'George Best case') resulted in more respondents wanting to incorporate lifestyle into decisions, but the dominant view was that everyone should be entitled to a first chance at treatment, regardless of past lifestyle. After that, failing to follow medical advice would be regarded negatively. Discussions usually concluded (not necessarily with consensus) that this is a problematic area and that prevention and health education are important areas for funding.

In the simple ranking task, 'Whether or not the patients live a healthy lifestyle' was ranked in the middle, at 5th on average.

### Socioeconomic status: some conflicting evidence

Findings from open discussions about the importance of socioeconomic status were often difficult to interpret, in particular because a range of issues are conflated, although the simple ranking and Q sort data would suggest that socioeconomic group should not be included as an attribute.

There are several socioeconomic issues which were mentioned. The first (and probably the main one we had anticipated) is linked to alleviation of deprivation and the prioritisation of interventions aimed at this over other interventions which may have a higher potential health gain, but which do not deal with inequalities. Individuals' ability to pay was commonly mentioned in this area and appeared to cloud the issue of inequality – the rich being able to pay being seen as a pragmatic solution rather than an issue of equity. Some respondents were adamant that socioeconomic considerations should not be taken into account, and cited the foundations of the NHS and equal treatment of all. In group discussions, the different socioeconomic issues were generally not delineated or articulated clearly. In one group, a respondent (who was also a health professional) raised the issue of inequalities. Otherwise people argued for dealing with poverty, not for giving poor people 'preferential treatment'. Several respondents were appalled at the suggestion that health care might be prejudiced against people with higher socioeconomic status, especially as they had contributed to the NHS through higher taxes.

'The social class of patients typically affected' was ranked last (10/10) on average in the simple ranking. There were, however, some concerns that social class, without further explanation, was being interpreted by some to mean discriminating *against* those in more deprived groups. (The Q sort statements made more explicit the 'direction of effect'.)

### Quality of life of beneficiaries

This health-related factor was discussed at some length in all focus groups. There were two main arguments. The first related to the relationship between length of life and QoL, the thrust of opinion focusing on the unnecessary extension of life in older people experiencing poor levels of QoL:

> You wouldn't want to live longer in a worse health state, quality of life is the important thing.
>
> Focus group participant, Norwich, March 2005

The second argument related to the 'starting point' QoL before treatment and the relationship between that starting point and the amount of QoL gained through treatment. Here, some respondents observed that an improvement in QoL for people in very poor health would be more important than an identical improvement in QoL for people in relatively good health.

The following illustrative quote refers to an example of a question in which QoL is represented on a scale from 0 to100, using percentages for ease:

> 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.
>
> Focus group participant, Newcastle, May 2005

## Results based on simple ranking

A subgroup of 19 respondents (aged 20–62, 10 female) rank ordered a set of possible attributes according to their importance and discussed their rankings. They were also invited to add any additional attributes (writing them onto blank cards provided) and incorporate those into their ranking. *Table 1* presents the average ranking of each item. This is only illustrative; respondents' rankings are not intended to be interpreted in isolation of their comments and the results of other methods.

Respondents' comments during this task revealed multiple understandings of the attributes as well as a small number of common views. Respondents were comfortable and confident talking about health-related attributes and less so when discussing non-health-related attributes. Socioeconomic status and gender were considered irrelevant to issues of prioritisation by all respondents. Whether or not patients had had a lot of health care in the past was often construed

*TABLE 1 Simple ranking exercise*

| Rank | Average rank[a] | Attributes |
|---|---|---|
| 1 | 2.9 | QoL of patients before treatment |
| 2 | 3.1 | Whether there is no other treatment available |
| 3 | 3.2 | The life expectancy of patients before treatment |
| 4 | 4.4 | The average age of patients at time of illness |
| 5 | 5.9 | Whether or not the patients live a healthy lifestyle |
| 6 | 6.1 | Whether or not the patients have had a lot of health care in the past |
| 7 | 6.3 | Whether or not the patients have dependants |
| 8 | 7.4 | Whether or not the patients are currently working |
| 9 | 9.0 | The gender of patients typically affected |
| 10 | 9.2 | The social class of patients typically affected |

a Average rank is simply the mean ranking given to the listed attributes ($n$ = 19).

as the health service having failed them and issues of 'orphan drugs' were difficult for respondents to appreciate and were not covered by the lack of any other treatment.

While the qualitative data and simple ranking data provide a good grounding in the issues in question, we felt it was insufficient for the selection of attributes and so incorporated a third method, Q methodology. This is relatively unfamiliar to most and so requires separate explanation below.

# Q methodology

## The basic features of Q methodology

Q methodology[52–54] is used to study the nature of views, opinions and beliefs. It is a useful addition to qualitative methods, especially where respondents do not necessarily have readily articulated accounts of their views on a topic.

The two main features of a Q study are the data collection method – which is based on a card sorting technique (the 'Q sort') – and a form of factor analysis which is used to analyse patterns between the card sorts to reveal a small number of underlying perspectives. There are several key terms that are used in Q studies. The 'Q sort' provides the primary data source in Q methodology. Respondents sort a set of statement cards, known as the 'Q set'. The Q set comprises a number of statements which cover the range of viewpoints and opinions on the particular topic of interest. Respondents consider each card in turn and assign it in a quick, initial sort, to one of three piles: agree, disagree or neutral. A more detailed arrangement of cards then follows, using a grid such as the one reproduced in *Figure 1*. Factors are the result of Q analysis, the aim of which is to identify shared views and meanings that exist around a topic (via correlations between the positioning of cards by respondents).

Each space in the grid indicates the positioning of a card on the continuum from –5 to +5. Two items are placed in the '+/–5' positions, four items in the '4' and '3' positions and so on. Making use of the three initial piles, respondents are asked to consider the cards in their 'agree' pile, select two cards that they 'most agree' with and place these in the +5 column. Next, selecting from the cards that they disagree with, respondents are asked to select the two cards that they 'most disagree' with, and place those cards in the –5 column. This process is repeated until all cards are placed (46 in this example), finishing at the centre of the distribution. Often the Q data (i.e. the positioning of the cards in the Q sort) are supplemented by a brief interview. In this study the Q sorts were followed by group discussion.

## Factors and factor loadings in Q methodology

In Q methodology, 'factors' are distinct accounts, each one a shared point of view, relating to the topic studied, based on the correlations between respondents' Q sorts.

There are several types of information of interest in the interpretation of factors. The main source of information is a 'collective' Q sort (known as a *factor array*) for each factor, which is calculated from the individual Q sorts making up that factor and based on weighted averages. In other words,

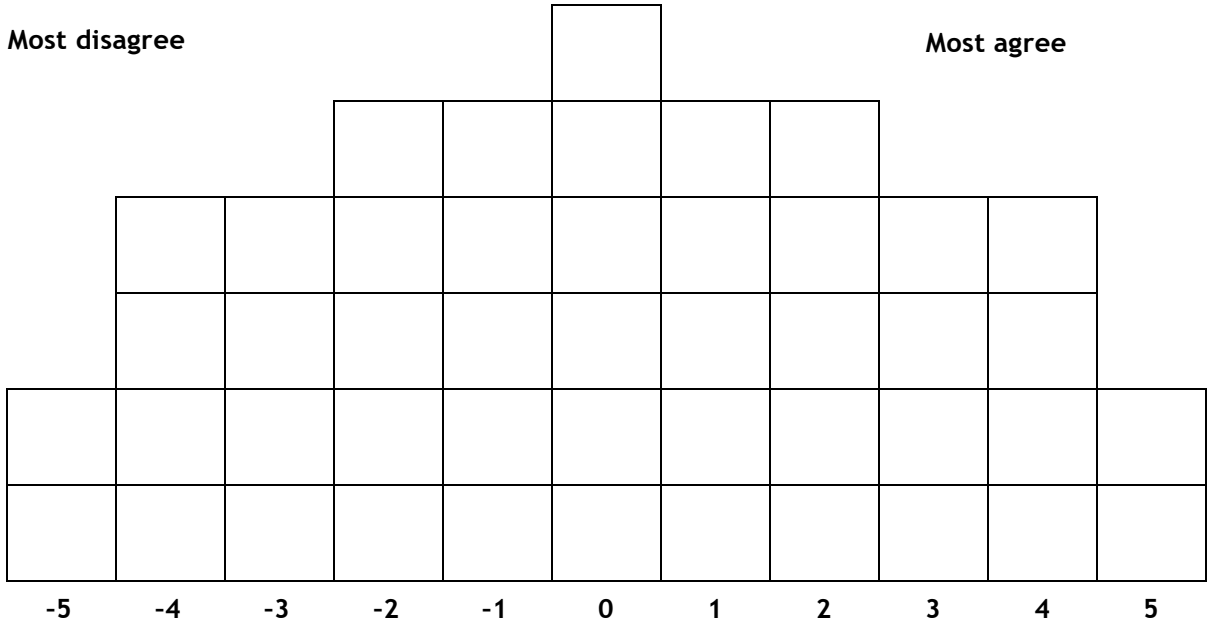


***FIGURE 1*** *Q sort response grid.*

for each factor, all of the statement cards can be placed on the Q grid, representing the relative importance of each statement to each factor. The statements placed in the +/–5 and +/–4 positions (representing strongly held views) are important in the interpretation of factors.

Information is also generated that identifies those statements which significantly distinguish factors, as well as consensus statements which are positioned similarly across factors. 'Factor loadings' give us further information about the correlation between each individual's own Q sort and each factor – see Appendix 2, which presents factor loadings for each respondent. Qualitative data collected during and after the Q sorts, including spoken and written comments, are also used to help understand the meanings contained in the Q sorts.

## The SVQ Q study

Statements for inclusion in the Q sort were taken from the first wave of focus groups, conducted in March 2005. Using the audio-recordings of the group discussions, opinions expressed by respondents which related to the topic of interest were listed. A small number of statements thought to be of interest to the research questions, but not raised in the groups, were added. Duplicate statements were then deleted, selected statements were reworded to make them clearer, and the list was reduced to a set of 46 statements through a process of discussion within the research team. Pilot Q sorts were completed by a sample of the general public ($n=26$) and a sample of Newcastle University staff members ($n = 23$). During this pilot, respondents were asked for comments about the set of statements; and in particular, to suggest anything that they felt was relevant that was not included in the statements. A revised set of statements was produced based on their comments. This final set of statements is reproduced in Appendix 3.

In the final wave of focus groups, a subgroup of 27 respondents (aged 20–84, 16 male) sorted a set of 46 cards printed with statements about the topic in question, according to their agreement with them.

## SVQ Q findings

A three-factor solution was derived using centroid factor analysis and judgemental rotation.[55] Further information on the analysis is available from the authors.

### *Factor 1: egalitarianism*

The first factor identified appears to be an egalitarian account, rejecting prioritisation of health care according to characteristics such as social class, lifestyle or whether respondents have dependants. It is an account which is concerned not primarily with outcome but with *entitlement*, and equal access to health care for everyone.

*Table 2* lists the statements of most importance to factor 1. In the top half of the table are statements that found strong agreement in factor 1 (i.e. these were statements placed at +4 and +5 positions on the Q sort grid by respondents associated with this factor). The statements in the bottom half of the table (below the emboldened line) are those that provoked strong disagreement. All the statements of importance in this factor (with the possible exception of 6) reflect an egalitarian position: no distinction should be made between age groups, socioeconomic groups or those without dependants. In keeping the statements in the bottom half of the table, the –4 and –5 positions are rejected by factor 1, supporting the interpretation. This factor refuses prioritisation on the basis of the characteristics of beneficiaries even when, as in statement 14, there is some implied gain in overall health in doing so.

### *Factor 2: health benefits*

Factor 2 reveals a somewhat different point of view. This second account puts emphasis on outcome and the size of the health benefits – as revealed through the placing of statements 29, 31, and 44 in the 'most agree' columns of the grid for this factor (*Table 3*). Interestingly, this is coupled with a rejection of any statements that make reference to socioeconomic and financial issues. All of the statements placed in the –4 and –5 positions ('most disagree') are of this nature. Preventive health care is also important in this, as it is in all three accounts.

### *Factor 3: children and experts*

A third factor shared some views with those already described, but is distinguished by a concern for children's health and a belief that health prioritisation decisions should be made by experts. Examination of the full set of statements and factor scores (see Appendix 3) reveals that factor 3 often shares views with factors 1 and 2 or occupies a space between them. However, focusing specifically on significantly distinguishing statements, we can detect the views that set this account apart – for example, statement 35 ('The decisions about which services to fund, and how to spend NHS money should be made by the experts'), placed in the +3 position for factor 3 and in –3 and 0 for factors 1 and 2 respectively), and statement 13 (*Table 4*), placed in the +4 position for factor 3 but 0 for both factors 1 and 2.

Once again, there is a notable reluctance to prioritise on the basis of other factors evidenced by the statements that are rejected.

### *Overall Q findings*

Respondents associated with all three factors thought that health care should be based on some concept of need and not on other factors, such as socioeconomic characteristics (statement 30) or lifestyle factors (statement 41). There was a rejection of socioeconomic issues and statements about lifestyle, and dependants were placed in the middle (irrelevant) or at the 'disagree' end of the scale in all three factors.

## Summary

Based on our qualitative and Q methodological enquiry, we rejected both lifestyle and socioeconomic status as attributes. Age, QoL and length of life were clearly issues of importance. The issue of whether beneficiaries of health care have dependants was slightly more difficult to resolve, but ultimately we took account of considerations of policy relevance, in consultation with representatives of NICE. It is difficult to conceive of a situation where NICE would recommend an intervention be made available only to people with dependants.

Arriving at a list of key attributes for inclusion in quantitative survey questions is a difficult process. Increasingly, qualitative methods are built into the early stages of study designs. In the context of this study, the use of open-ended qualitative methods alone proved insufficient to determine the attributes, perhaps because of the complexity of the subject matter. Respondents appeared to be led by the suggestions of the focus group facilitator, and there remained uncertainties around particular attributes (such as socioeconomic status) following analysis of the qualitative data. In this case, Q methodology provided additional structure and the opportunity to use a standardised stimulus in both the generation of data and the analysis. With respect to socioeconomic status, for example, we were able

***TABLE 2*** *Salient statements for factor 1*

| Number | Statement | Factor score |
|---|---|---|
| 11[b] | Life is equally valuable whether you are young or old. | +5 |
| 15[b] | Everybody, no matter what you are, whether you are young or old, should get the same access to and choice of treatment. | +5 |
| 41 | Health care should be based on *need*, not on social circumstances, or addiction or weight or smoking or anything else. | +4 |
| 6[b] | If someone is given treatment, like George Best, and then abuses their treatment, they should not be given repeated chances. If there are finite resources and a person has failed to take advantage of it, someone else should be given a chance. | +4 |
| 30 | Social class should make no difference whatsoever for prioritising health care. If people need treatment, they need treatment. How well off they are shouldn't come into it. | +4 |
| 25[a] | People with dependants should not be given priority over people without dependants. A human life is a human life, I think it should be irrelevant how many dependants they've got. | +4 |
| 14[b] | The age of the patient is important; if you were treating children rather than older people then you would have a longer improved life. | −4 |
| 20 | People with dependants should be prioritised over people without dependants because their treatments would benefit others as well as the patient themselves. | −4 |
| 46 | People who have already benefited from a lot of health care should take second place to people who have not used the health service as much. | −4 |
| 24 | Whether or not people are currently working should be taken into account when we prioritise health care. | −4 |
| 3 | People who live a healthy lifestyle should be prioritised because they would respond better to treatment. | −5 |
| 16[b] | You should prioritise the younger age group, because they are still able to have children. | −5 |

a Denotes a statement which distinguishes factor 1 from factors 2 and 3 ($p<0.05$).
b Marks a significance level of $p<0.01$.
Consensus statements are shaded.

to divide the broad issue into sub-issues that could be described in discrete statements. Factor analysis showed a good deal of consensus that these issues should *not* be part of health-care priority setting, and we selected age and severity as the attributes to bring forward into the quantitative analysis. Future research investigating the views of the public around complex issues should consider using Q methodology in addition to more typical qualitative methods.

## Diagrammatic questions

The presentation of the discrete choice and matching questions, including a detailed introductory explanation, was developed iteratively in focus groups. By far the most successful method was presentation of concepts of health (QoL), age, and health gains using diagrams. These diagrams were first explained by building them in small steps for respondents using an animated powerpoint presentation (reproduced in Appendix 4), which was ultimately incorporated into a CAPI. The diagrams were then presented either as choice questions or as matching questions, examples of which are shown in *Figures 2* and *3*. Choice questions (e.g. *Figure 2*) present respondents with a one-off choice between option A and option B, which differ in terms of health gains, ages of patients and levels of QoL. The two options are presented both diagramatically and descriptively in the accompanying text. Choice questions are explained in more detail in Chapter 4.

Matching questions (e.g. *Figure 3*) present respondents with a series of iterative choices where the numbers of people in Groups A and B are varied until a point of equivalence is reached. The same size of health gain is presented in Group A as in Group B for each set of iterative choices. These questions are explained in more detail in Chapter 5.

*TABLE 3* *Salient statements for factor 2*

| Number | Statement | Factor score |
|---|---|---|
| 29[a] | The quality of life of patients and their life expectancy are the most important things. The characteristics of patients like whether they are employed, or whether they have dependants, or what gender they are shouldn't matter. | +5 |
| 40 | Priority should be given to preventive health care rather than always focusing on cure once people are ill. | +5 |
| 31 | The amount of health and quality of life improvement is the most important. It's about getting the greatest benefit for the most people. | +4 |
| 30 | Social class should make no difference whatsoever for prioritising health care. If people need treatment, they need treatment. How well off they are shouldn't come into it. | +4 |
| 23 | Priority should be given to preventive health care especially education in schools about diet and lifestyle choices. | +4 |
| 44[a] | It's no good saving lives if the quality of those lives is really bad. Some treatments are keeping people alive for too long. You've got to have a decent quality of life otherwise what's the point of being alive. | +4 |
| 18[a] | There should be 'positive discrimination' towards people who are disadvantaged and in ill health because they've got a lot to contend with already. | 4 |
| 21[a] | Older people deserve to be given priority. They have been paying into the NHS all their lives, they deserve to be able to draw on those resources when they need it. | 4 |
| 24 | Whether or not people are currently working should be taken into account when we prioritise health care. | 4 |
| 9 | People who smoke and drink pay enough in extra taxes to pay for their own health care. | 4 |
| 26[a] | Poorer people should be given priority because they don't have the same opportunities to take care of their own health. | 5 |
| 28[a] | Whether or not patients can contribute financially towards the cost of the treatment should be taken into account because it would allow you to treat more people who can't afford to 'go private'. | 5 |

a Marks a significance level of $p<0.01$.
Consensus statements are shaded.

The questionnaire instrument was then piloted using cognitive interviews with 42 respondents (27 in Norwich and 15 in Newcastle) and was well received. Respondents followed the introduction and understood the questions. Their comments resulted in only minor adjustments to the wording and number of questions in each version of the questionnaire.

## Survey sample and administration

The discrete choice and matching questions were part of a longer questionnaire (incorporating attitudinal questions and sociodemographics), which was administered face to face, using a CAPI, by interviewers from NatCen. The interview began with some basic demographic and household questions, following which an animated powerpoint presentation explained the meaning of the diagrams step by step. Next the respondents answered six matching questions followed by eight discrete choice questions. There were four attitudinal questions, before some more detailed sociodemographic and health questions. On average, interviews lasted 41 minutes.

The survey was administered by NatCen to a random sample ($n = 587$) of the population in England from February to April 2007. The sample was generated by NatCen from the population of adults (aged 18 and over) living in England. Thirty addresses were selected from each of 40 postcode areas, which were stratified by Government Office Region (nine regions) and the proportion of manual/non-manual households. Within each household, only one adult was eligible for inclusion in the study. In households with more than one eligible adult present, interviewers randomly selected one interviewee. A total of 243 (41%) were male, the mean age of the whole sample being 52 years and, thus, females and older people are slightly over-represented.

***TABLE 4*** *Salient statements for factor 3*

| Number | Statement | Factor score |
|---|---|---|
| 40 | Priority should be given to preventive health care rather than always focusing on cure once people are ill. | +5 |
| 23 | Priority should be given to preventive health care especially education in schools about diet and lifestyle choices. | +5 |
| 13[a] | Age shouldn't come into it, unless you're talking about children. Children's health should be given priority over adults. | +4 |
| 41 | Health care should be based on *need*, not on social circumstances, or addiction or weight or smoking or anything else. | +4 |
| 30 | Social class should make no difference whatsoever for prioritising health care. If people need treatment, they need treatment. How well off they are shouldn't come into it. | +4 |
| 31 | The amount of health and quality of life improvement is the most important. It's about getting the greatest benefit for the most people. | +4 |
| 24 | Whether or not people are currently working should be taken into account when we prioritise health care. | −4 |
| 20 | People with dependants should be prioritised over people without dependants because their treatments would benefit others as well as the patient themselves. | −4 |
| 3 | People who live a healthy lifestyle should be prioritised because they would respond better to treatment. | −4 |
| 17 | For relatively minor conditions patients who are of working age should take priority over people who are retired. | −4 |
| 9[a] | People who smoke and drink pay enough in extra taxes to pay for their own health care. | −5 |
| 46 | People who have already benefited from a lot of health care should take second place to people who have not used the health service as much. | −5 |

a Marks a significance level of $p < 0.01$.
Consensus statements are shaded.

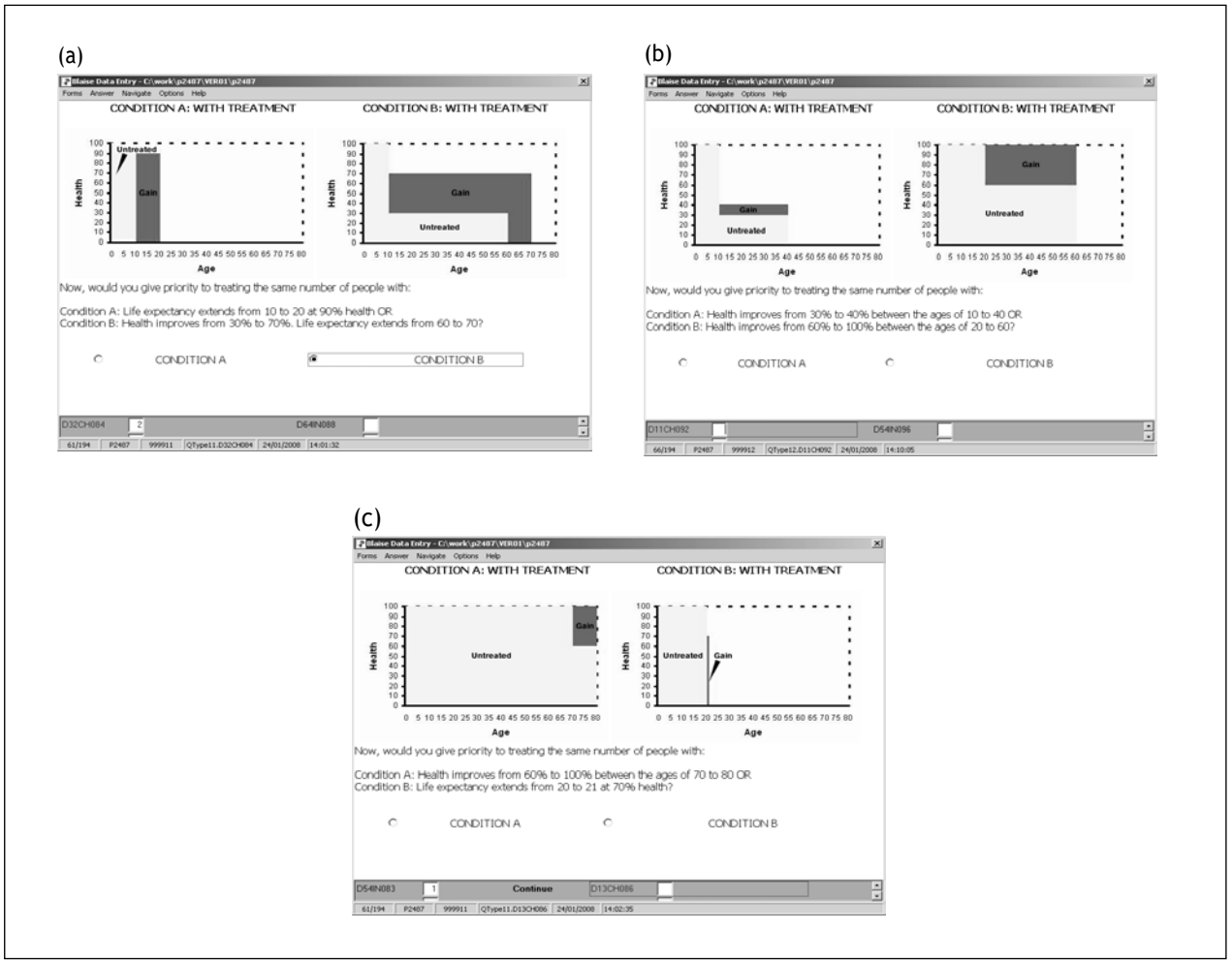


***FIGURE 2*** *Examples of choice questions. (a) Screen 1: treatment for condition A results in a gain in life-years; treatment for condition B results in a gain in both quality of life (QoL) and length of life. (b) Screen 2: treatment for either condition A or condition B results in a gain in quality of life. (c) Screen 3: treatment for condition A results in a gain in quality of life; treatment for condition B results in a gain in length of life.*

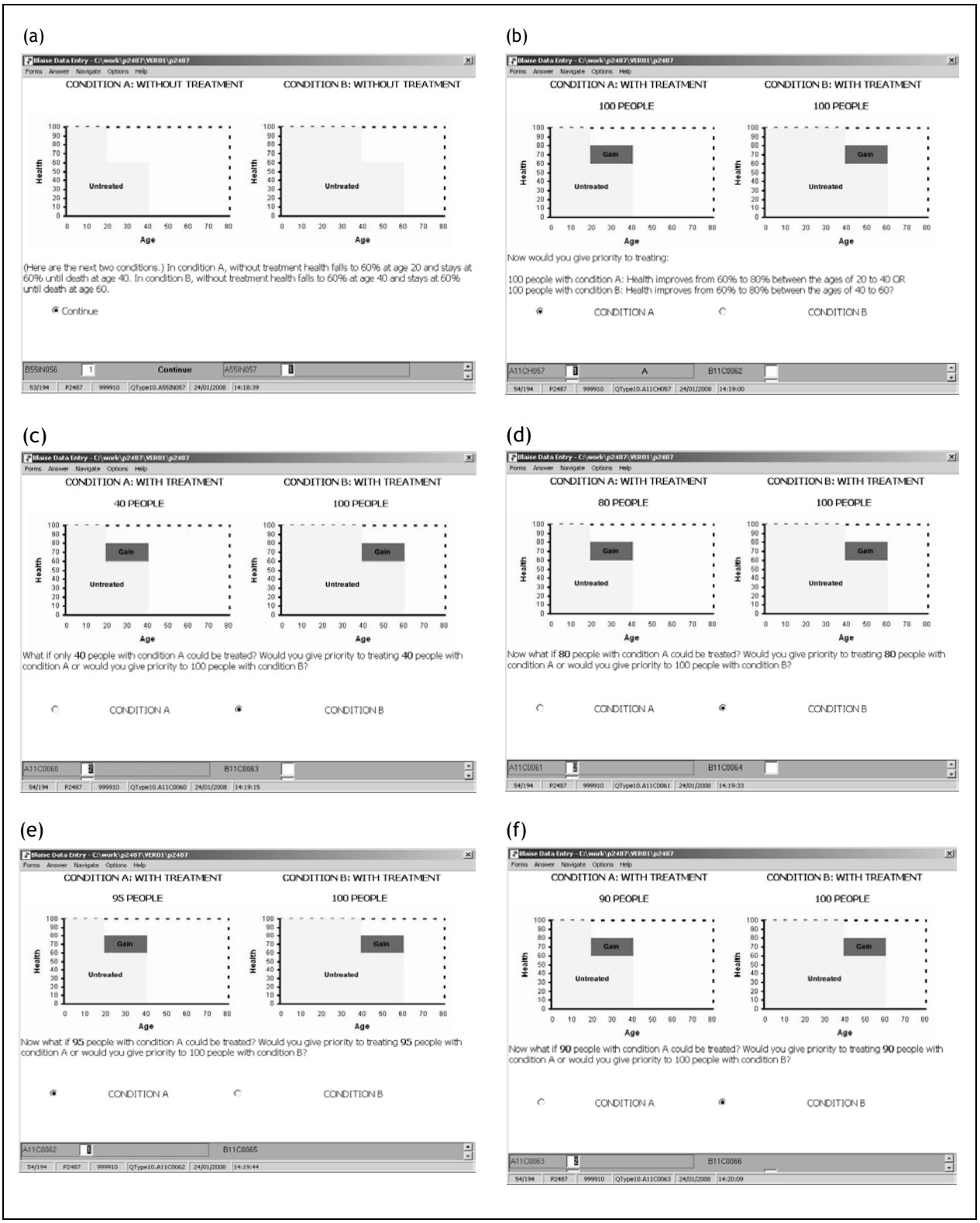


***FIGURE 3*** *Example of iterations within a single matching question based on responses shown. (a) Screen 1: untreated scenarios. (b) Screen 2: 'with treatment' scenarios showing health gains - condition A is chosen by the respondent. (c) Screen 3: when 40 people with condition A or 100 people with condition B can be treated, the respondent chooses B. (d) Screen 4: when 80 people with condition A or 100 people with condition B can be treated, the respondent chooses B. (e) Screen 5: when 95 people with condition A or 100 people with condition B can be treated, the respondent chooses A. (f) Screen 6: finally, when 90 people with condition A or 100 people with condition B can be treated, the respondent chooses B.*

# Discrete choice study

## Basic design

In a discrete choice study, respondents are presented with a series of choice sets (usually pairwise). Each scenario in the set is defined according to some predefined attributes (ours being predetermined from the research reported in Chapter 3) and levels of such attributes. The attribute levels vary across scenarios and choice sets. In each choice set, the respondent is asked which scenario they prefer or would choose. Faced with a series of such choices, respondents essentially reveal how much weight they attach to each of the attributes, the actual weights being derived through statistical analysis of the data (see below).

*Table 5* lists the attributes introduced in the preceding chapter along with the levels for each that were used in the discrete choice study.

The levels for the age variable were chosen to represent a range of stages of life: a very young child, a fully grown child and a young adult, followed by two or three further stages of adulthood into old age. The QoL if untreated variable again ranges from death through a series of better (though some still quite serious) states to 90%. The variables representing gains in life expectancy and QoL follow in large part from what was chosen for age and QoL if untreated. For example, to give a full number of life-years to a 1-year-old who would otherwise die, the gain in life expectancy would be 79, and to return someone in a health state valued at 60% back to full health, a gain of 40% would be offered. Some levels of the life expectancy and QoL gain variables also reflect the desire to have some scenarios where people might get an improvement but not to full health or maximum life-years (the maximum here being 80).

Using these variables and imposing some assumptions (such as people being in full health before the onset of illness), we constructed diagrams of the sort shown in *Figure 2*. QALY gains (shown as a dark shaded area in the diagrams) are calculated from the five attributes listed in *Table 5*.

A full factorial design using the attributes listed in *Table 5* would have resulted in (6)(7)(8)(4)(6) = 8064 possible profiles but there are many combinations of levels on these attributes that result in implausible scenarios. For example, 'age at death' must be greater than 'age at onset' of illness, and the gain in life expectancy added to the age at death if untreated cannot exceed some reasonable maximum age, which was constrained to be 80 years. The full list of constraints is listed in Appendix 5. After imposing these constraints, 6572 of the possible 8064 profiles were implausible, leaving 1492 profiles (19% of the total).

The experimental design software SAS, which allows for constraints, was initially used to select choice sets from the 1492 profiles. This resulted in a design with over 200 profiles describing 'age at onset' as 1-year-olds and only 12 profiles describing 70-year-olds. At this stage the design was altered manually to improve the balance of questions about different age groups. Clearly, such a severely-reduced set of available profiles, together with the manual alteration of the design to achieve greater

***TABLE 5*** *Attributes and levels*

| Description of attribute | Levels |
|---|---|
| Age at onset (years) | 1, 10, 20, 40, 60, 70 |
| Age at death if untreated (years) | 1, 10, 20, 40, 60, 70, 80 |
| Gain in life expectancy (years) | 0, 1, 5, 10, 20, 40, 60, 79 |
| QoL if untreated (represented as %) | 0, 30, 60, 90 |
| Gain in QoL with treatment (%) | 0, 10, 20, 40, 70, 100 |

QoL, quality of life.

balance, negatively affect the design properties that are desirable in discrete choice experiments (DCEs). Nevertheless, before administering the questionnaire, data were simulated to ensure that a model could be estimated on the basis of the amended design.

## Functional form and empirical approach

Response data in DCEs are modelled within a random utility framework of the general form:

$$U = V + \varepsilon \quad (1)$$

in which utility, $U$, is separated into parts which are explainable, in this case $V$, and unexplainable, $\varepsilon$. In this study, we are concerned with estimating $V$, which represents an underlying continuous and latent variable which is nevertheless unobservable.

If the standard QALY model is true, $V$ would simply be a function of QALYs. If individuals are concerned about other characteristics, then these will also be part of the utility function. In this simple case we assume that utility is a function of age at onset *(AO)*, age at death without treatment *(AD)*, QoL lost without treatment *(QL)* and QALYs gained from treatment (*QALY*). This gives:

$$V = f(AO, AD, QL, QALY) \quad (2)$$

Quality of life lost (*QL*) is transformed from a variable in *Table 5*, by subtracting 'quality of life if untreated' from 1. This is done in order to facilitate the log transformation required below – with some scenarios involving instant death, and thus a QoL of zero, which could not otherwise have been log transformed. Note, therefore, that when interpreting this variable, the larger the 'quality of life lost' at the onset of illness, the more 'severe' the health state.

The first two terms on the right-hand side of (2) detect age effects, *QL* detects severity and *QALY* is the health gain. If expressed as an additive function, this would mean that gains in utility could be experienced even if QALY gains were zero, i.e. age and severity would have effects on utility irrespective of whether or not QALY gains are incurred. The alternative, therefore, was to use a multiplicative form of the utility function. The QALY itself is a multiplicative function of life-years and QoL gained. By extension, in a multiplicative form of the above function, utility is derived from QALYs multiplied by the magnitudes of the other variables, ensuring that with zero QALYs gained there is a zero impact on utility. Empirically, the multiplicative models presented below consistently outperformed those based on an additive functional form, which were also investigated. Although, it may appear that a multiplicative model of the form QALYs × AGE × SEVERITY, with just one age-related variable, would make more theoretical sense, we took the more pragmatic view that this would leave too much riding on the 'age at onset' variable in terms of what respondents might be thinking about in relation to age, and so we included age at death as well. In addition, this was the best performing model empirically, which, it could be argued, is important for estimating weights.

If we assume such a multiplicative underlying model, we may use a log-linear model of the form:

$$\log(V) = \beta_1 \log(AO) + \beta_2 \log(AD) + \beta_3 \log(QL) + \beta_4 \log(QALY) \quad (3)$$

This is a standard log-linear utility function where the $\beta$s are parameters to be estimated. Given that discrete choice response data are based on choices over alternative combinations of the dependent variables, then, assuming that the $\beta$ parameters are identical across all individuals, a simple model of the following form can be estimated:

$$\Delta\log(V) = \beta_1 \left(\Delta\log(AO)\right) + \beta_2 \left(\Delta\log(AD)\right) + \beta_3 \left(\Delta\log(QL)\right) + \beta_4 \left(\Delta\log(QALY)\right) \quad (4)$$

where $\otimes$ represents the differences in levels of any given attribute reflected in the pairwise choices presented.

Equation (4) was estimated using a logit model (see 'simple' model under Discrete choice results), allowing for clustering of the individual standard errors to account for the fact that each individual responded to several questions. Essentially, this amounts to a conditional logit model.

A flexible functional form (referred to below as the 'powered' model) was also specified to allow for any non-linear relationships between choice and the included variables. While the standard log model allows for non-linearities, these functions are monotonic, this restriction possibly being too strong and resulting in biased estimates. Including higher order terms, in a fashion analogous to the popular translog model,[56] allows for modelling of more complex non-linearities. The use of such

flexible functional forms provides more robust estimation of the coefficients and reduces the potentially confounding problems of omitted variable bias. This also allows us to investigate the non-linearities detected in the matching data (see Chapter 5). All models were tested using the Akakie and Bayesian information criteria in order to aid model selection. Those which performed best on these criteria are reported below.

## Discrete choice results

A total of 587 respondents yielded 4696 useable responses to the discrete choice questions. The estimated models are shown in *Table 6*. We did also take a more conventional econometric approach of converting covariates into categorical variables (i.e. sets of dummy variables) within the basic multiplicative framework. However, these models did not perform as well as those reported and are more challenging for calculation of weights.

The coefficients for the simple model suggest that increasing age at onset reduces the probability of choice, as does increasing the age at death, although the former is not statistically significant. This suggests that the young are preferred to the old. The coefficient on QoL lost is also negative, suggesting that as the health state is more severe the respondents are less likely to choose that group. Finally, the impact of *QALY* is positive, as we would expect. The impact of severity, as it appears in the regression results, appears to contradict earlier literature. However, a more accurate picture of the impact of all of the variables is provided through examination of the shapes of the functional forms, which, for the simple model, are given in *Figures 4–7*. These diagrams show the shapes of the functional relationships when measured on a single scale, as represented by the 'predicted utility' axes. Here, it can be seen that that the general impact of the QoL lost variable on utility is very small, the inference being that the severity–utility relationship is essentially flat and nearly so for age.

The equivalent diagrams for the powered models are shown in *Figures 8–11*. Once again, the sign of the coefficients in the regression model inform us of the direction of impact on utility for age at onset (individuals who are aged around 10–40 years are slightly preferred to the very young and the very old), age at death (slight preferences to save those who die young and those who will die old rather than the middle aged) and severity (a preference for individuals with lower severity, with the maximum at 0.4, after which predicted utility slopes downwards showing less preference for purely life-saving interventions). However, *Figures 8–10* demonstrate, once again, the relationship between each of these variables and utility to be essentially flat. This is reinforced by the lack of statistical significance on some of the coefficients, especially for age at onset and, this time, QoL

***TABLE 6*** *Simple and powered models*

| | Simple model | | | Powered model | | |
|---|---|---|---|---|---|---|
| | **Coefficients** | **Standard error** | ***p*-value** | **Coefficients** | **Standard error** | ***p*-value** |
| log *AO* | −0.02 | 0.022 | 0.304 | −0.31 | 0.264 | 0.240 |
| (log *AO*)$^2$ | | | | 0.24 | 0.164 | 0.151 |
| (log *AO*)$^3$ | | | | −0.04 | 0.025 | 0.107 |
| log *AD* | −0.07 | 0.034 | 0.034 | 1.28 | 0.314 | 0.000 |
| (log *AD*)$^2$ | | | | −0.76 | 0.175 | 0.000 |
| (log *AD*)$^3$ | | | | 0.11 | 0.025 | 0.000 |
| log *QL* | −0.14 | 0.037 | 0.000 | −0.64 | 0.372 | 0.085 |
| (log *QL*)$^2$ | | | | −0.43 | 0.489 | 0.381 |
| (log *QL*)$^3$ | | | | −0.09 | 0.149 | 0.559 |
| log *QALY* | 0.75 | 0.033 | 0.000 | 0.45 | 0.054 | 0.000 |
| (log *QALY*)$^2$ | | | | −0.03 | 0.028 | 0.237 |
| (log *QALY*)$^3$ | | | | 0.03 | 0.007 | 0.000 |

*AD*, age at death without treatment; *AO*, age at onset; *QALY*, QALYs gained from treatment; *QL*, quality of life lost without treatment.

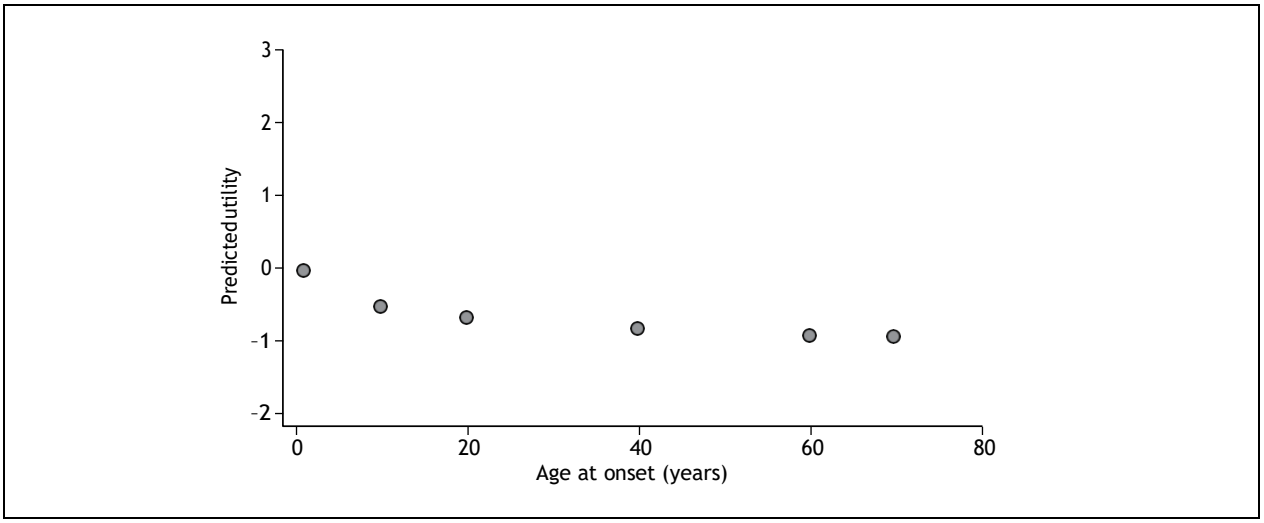


***FIGURE 4*** *Age at onset vs utility.*

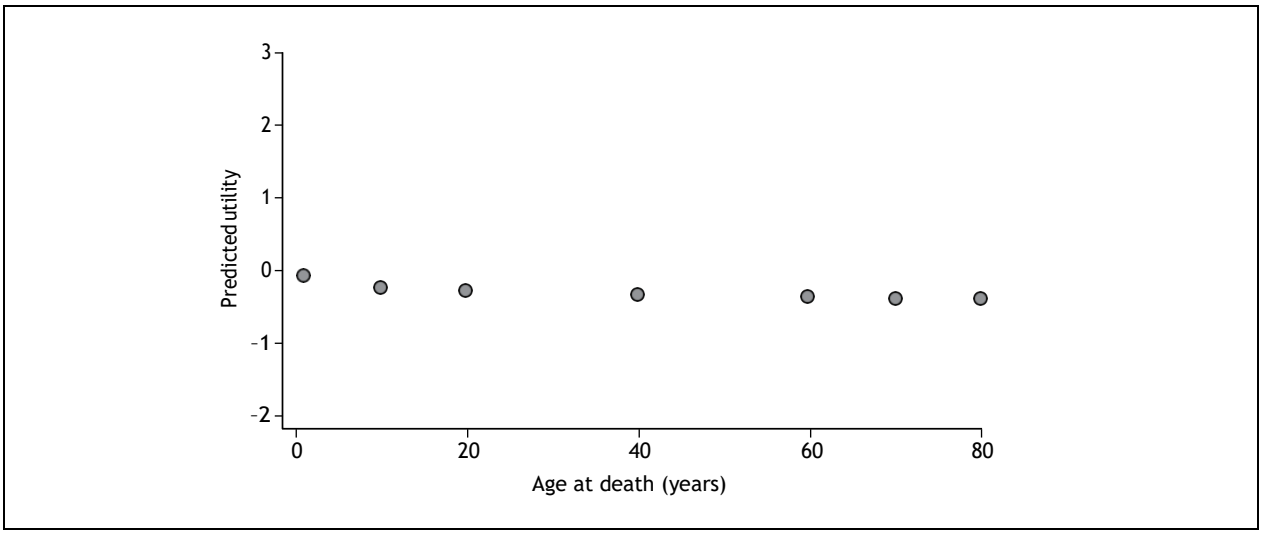


***FIGURE 5*** *Age at death vs utility.*

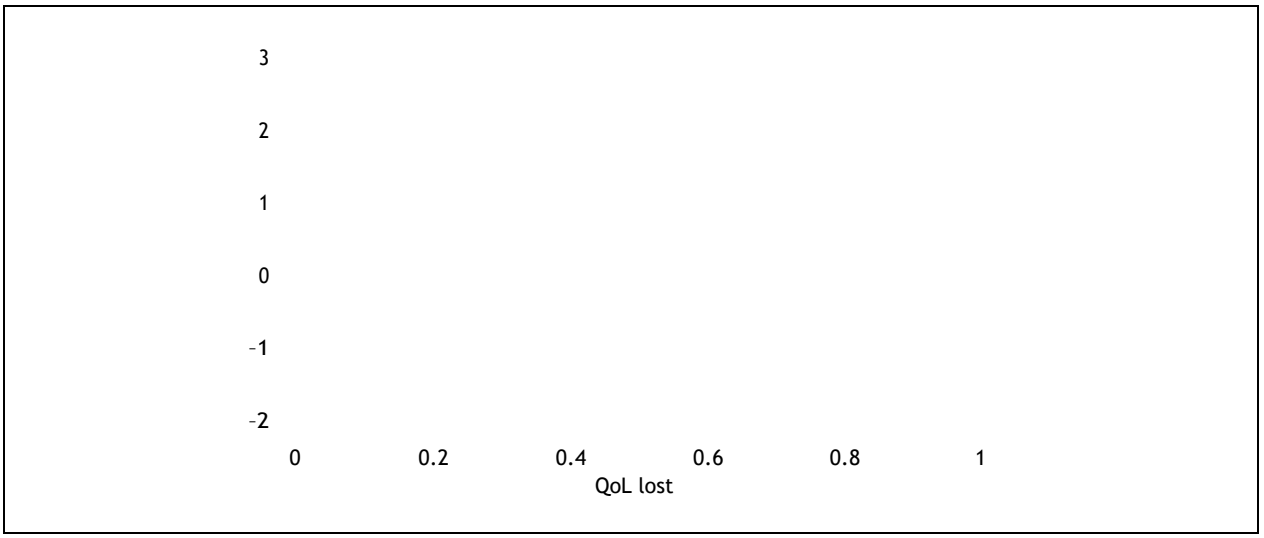


***FIGURE 6*** *Quality of life (QoL) lost vs utility.*

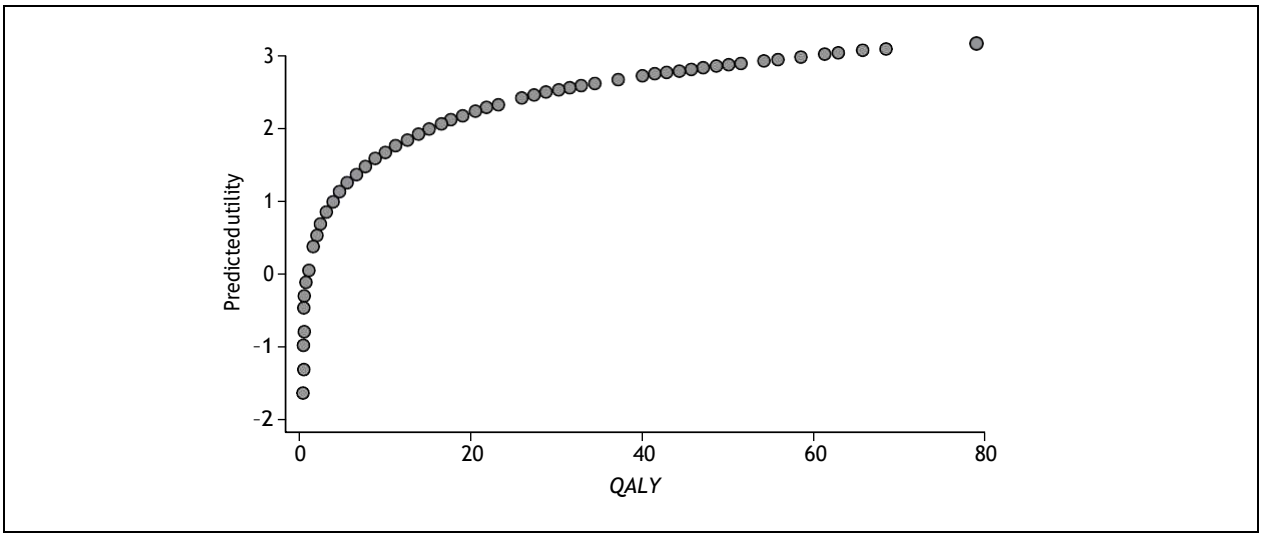


***FIGURE 7*** QALY *vs utility.*

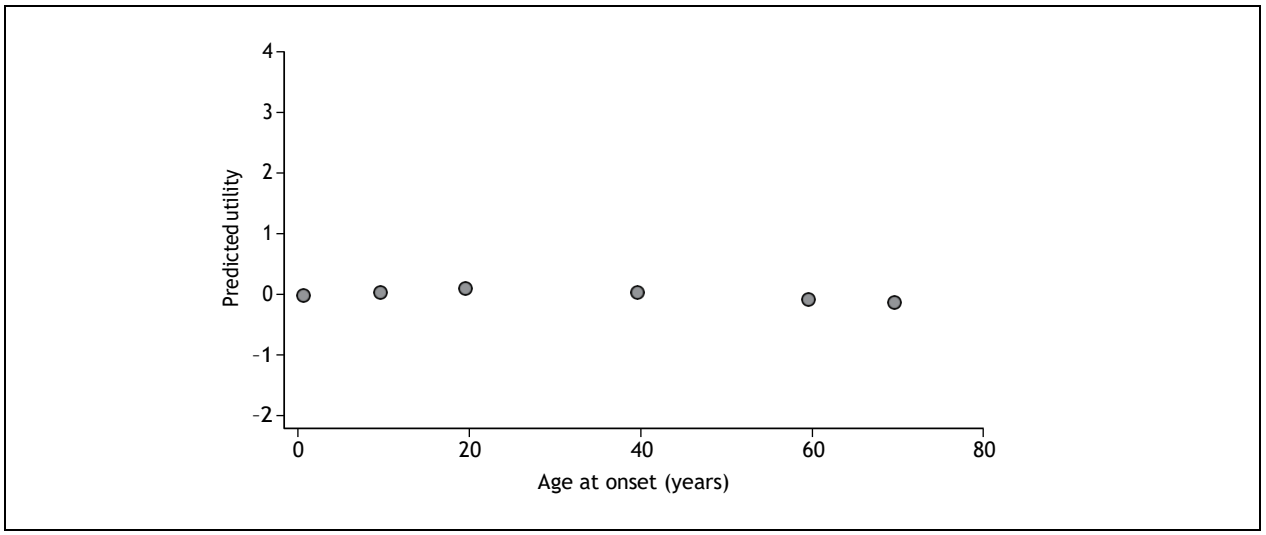


***FIGURE 8*** *Age at onset vs utility (powered functions).*

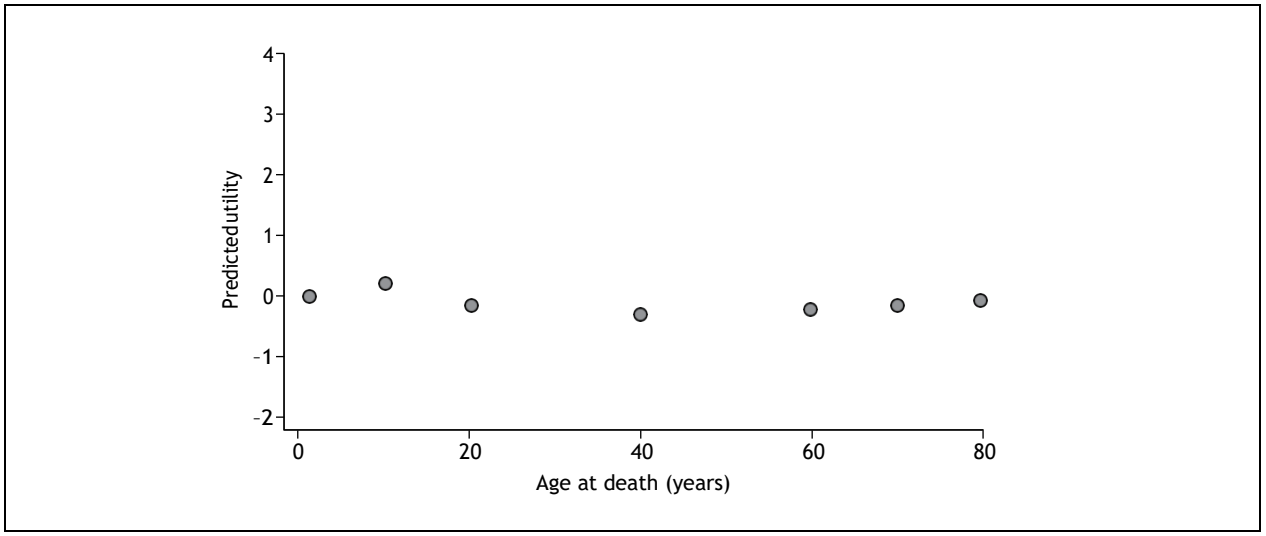


***FIGURE 9*** *Age at death vs utility (powered functions).*

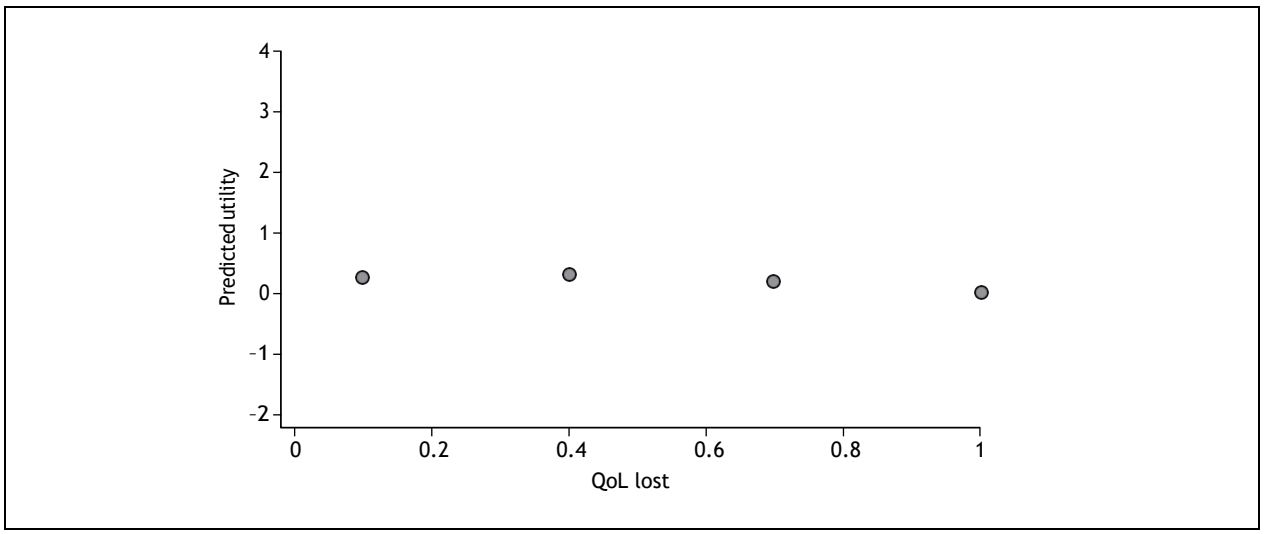


***FIGURE 10*** *Quality of life (QoL) lost vs utility (powered models).*

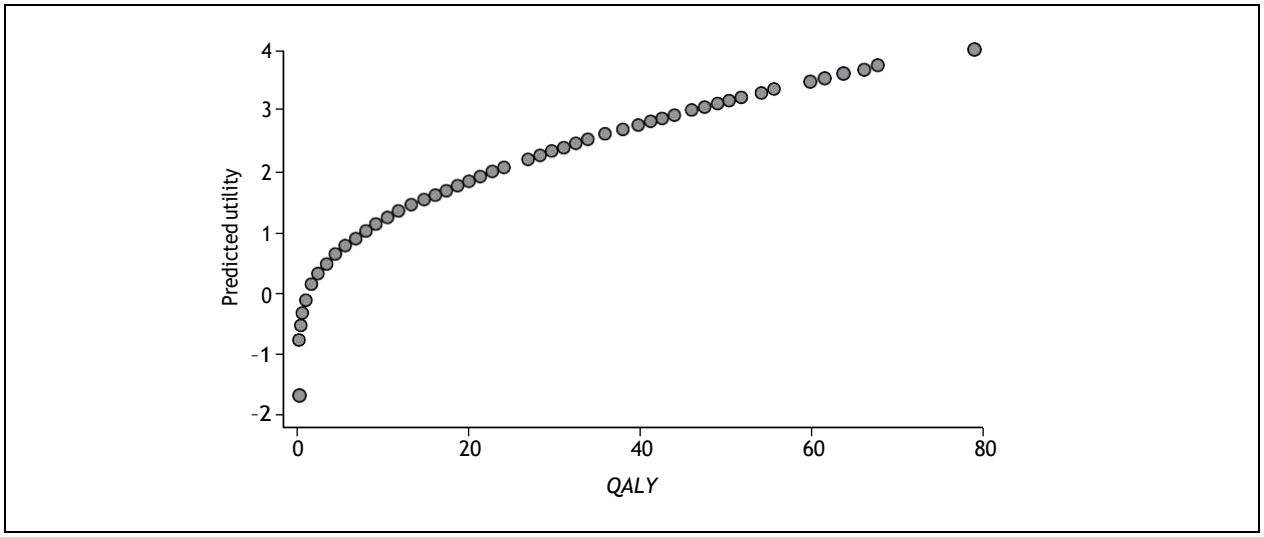


***FIGURE 11*** QALY *vs utility (powered model).*

lost. Nevertheless, weights were calculated using the point estimates from the powered regression, regardless of statistical significance. This approach is used for a number of reasons: first, the model is the best performing in diagnostic tests; second, we cannot rule out joint significance; and third, and most importantly, when calculating weights from any model, the point estimates from that model (regardless of which approach is used) are our most informed values. While it is possible to use hypothesis testing to determine whether our estimates are significantly different to zero, this does not provide better information than the point estimates. Also, because of what was revealed by the directions of the coefficients in the regression models, the results on severity are further investigated (see Further investigation of severity). *Figure 11* shows that increasing QALYs are always preferred. Concavity is still observed, but less so than in the simple model.

## Weighting QALYs using discrete choice data

Two novel ways of estimating weights from the above models have been devised, one based on predicted probabilities of choice and the other based on a compensating variation approach. As indicated above, for each, the weights presented are based on the powered model only.

## Predicted probability of choice approach

The weights are estimated using the estimated probability of choice, with a base case of $AO=40$, $AD=60$, $QL=0.7$, and $QALY=4$ (i.e. individuals fall ill at 40, lose 0.7 of their QoL, will die at 60 without treatment, and are then given four QALYs with treatment). The way in which the four QALYs have been allocated is unspecified in this model. This choice of a base case is challenging for interpretation of subsequent weights, in that it may be thought better to choose an extreme position and then measure weights for every other scenario in one direction relative to that. However, using an extreme as a base case is problematic too; the most obvious example being use of the highest age group from which it is not possible to gain any QALYs. The choice of a four-QALY gain arose to reflect a reasonable-sized gain and also to correspond with the four QALY gains which were offered to respondents in the matching study (see Chapter 5).

In order to calculate the weights we compare our base case to an alternative scenario. We then vary the number of QALYs being offered in the alternative scenario until the estimated probability of choosing the base case equals 0.5 (i.e. the individual is indifferent between the two scenarios). This is more easily demonstrated by using an example. We have our base case, $AO=40$, $AD=60$, $QL=0.7$ and $QALY=4$. We now take a scenario for which we want to find a weight, for example $AO=1$, $AD=1$ and $QL=1$ (i.e. individuals fall ill and die at age 1), this being comparison 1 in *Table 7*. We adjust the number of QALYs from treatment available to the comparison group until the probability of choosing the base case equals 0.5. This probability to calculated as:

$$\Pr_{BASECASE} = \frac{\exp\left(\log(u_{BASECASE})\right)}{\exp\left(\log(u_{BASECASE})\right) + \exp\left(\log(u_{COMPARISON})\right)}$$

where $\exp\left(\log(u_i)\right)$ is the predicted utility for the ith choice from the powered model regression results in *Table 6*. In this case the probability of choosing the base case is 0.5 when the comparison group is offered 4.1 QALYs. The weight itself is found by taking the ratio of the QALYs offered in both cases, so:

$$weight = \frac{QALY_{COMPARISON}}{QALY_{BASECASE}}$$

which in this case gives 4.1/4 = 1.025. This demonstrates that the base case is slightly preferred to the alternative, with one QALY to the base case being equal to 1.025 QALYs to the comparison. This process is repeated for each comparison scenario to generate weights. More generally, we try to illustrate this in *Figure 12*. Weights closer to 0 show a stronger preference for the comparison, weights equal to 1 show indifference between the two groups and weights greater than 1 show preference for the base case.

The weights from this procedure are given in *Table 7*. (In this table, the variables given earlier as percentages are now presented on a 0–1 scale in line with the more common representation in the QALY literature. Percentages were used earlier because that is how QoL was presented to survey respondents.) While these represent weights arising from assessing the respective comparator group against the base case, it is possible to generate weights for comparing different comparison groups. For example, if we wished to compare scenarios 5 and 8, we could do this indirectly by comparing the weights given in *Table 7*. The weight for comparison 5 is 0.79 and the weight for comparison 8 is 0.725 – from this we can conclude that scenario 8 is preferred to scenario 5. Alternatively, we could take the ratio of these weights, to generate a new weight for scenario 8 of 0.92 (0.725/0.79), which demonstrates that 0.92

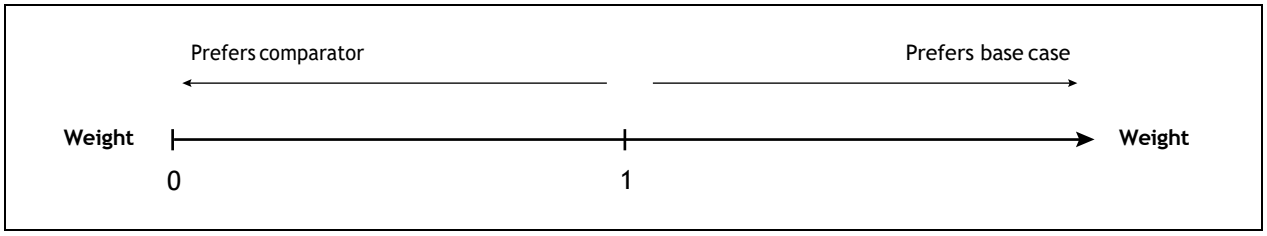


***FIGURE 12*** *Preferences and weights in the discrete choice study.*

of a QALY to scenario 8 is worth one QALY to scenario 5.

The weights presented in *Table 7* are sometimes difficult to interpret because a number of factors vary at once when comparing scenarios. We can see that most weights are less than 1, suggesting that the comparison scenarios used here are generally preferred to the base case. The only exceptions are comparisons 1, 9, 11, 14 and 15 (and comparison 12 which is the same as the base case). Apart from comparison 1, these scenarios are towards the top end of age at death and are quite severe.

For those scenarios which are purely life saving, where age at onset equals age at death (comparisons 1, 6, 11 and 14), we see that there is a preference for those with an age at onset and age at death of 10 (comparison 6). The weight for this scenario is 0.62. The weights for the older groups (comparisons 11 and 14) are greater than 1, indicating a preference for the base case. The weight for the youngest group (comparison 1) is also greater than 1, suggesting that individuals do not choose to treat those who are very young. These results suggest a non-linear relationship between age and weights, with the most preferred being the younger, but not youngest, groups.

For severity we can see that, for the same age at onset and age at death, as QoL lost decreases, moving from 0.7 to 0.1, the weight moves towards 0. The less severe comparisons (where QoL lost equals 0.1) have a weight of less than 1, showing that they are the preferred groups. This can be seen for comparisons 2 and 3, 4 and 5, 7 and 8, 9 and 10, 12 and 13, and 15 and 16.

It is difficult to make generalisations about age at onset and age at death using *Table 7*, where both variables often vary at the same time. An alternative approach to presenting the weights is to hold one of the factors constant and allow the others to vary. This is the approach used in *Tables 8* and *9*.

In *Table 8*, QoL lost is fixed at 0.7. This clearly shows that as age at onset increases the weights move towards 0 up to age 20 but then move towards 1. Weights for age at death move towards 1 up to age 40, but then fall again. This shows that there is a clear preference for treating people

***TABLE 7*** *Weights based on powered model*

| Base | Age at onset 40 | Age at death 60 | QoL lost 0.7 | QALYs gained 4 | Weight |
|---|---|---|---|---|---|
| ***Comparison*** | | | | | |
| 1 | 1 | 1 | 1 | 4.1 | 1.025 |
| 2 | 1 | 10 | 0.7 | 1.85 | 0.463 |
| 3 | 1 | 10 | 0.1 | 1.52 | 0.38 |
| 4 | 1 | 20 | 0.7 | 3.8 | 0.95 |
| 5 | 1 | 20 | 0.1 | 3.16 | 0.79 |
| 6 | 10 | 10 | 1 | 2.48 | 0.62 |
| 7 | 10 | 20 | 0.7 | 3.5 | 0.875 |
| 8 | 10 | 20 | 0.1 | 2.9 | 0.725 |
| 9 | 10 | 40 | 0.7 | 4.6 | 1.15 |
| 10 | 10 | 40 | 0.1 | 3.95 | 0.9875 |
| 11 | 40 | 40 | 1 | 6.15 | 1.54 |
| **12** | **40** | **60** | **0.7** | **4** | **1** |
| 13 | 40 | 60 | 0.1 | 3.38 | 0.845 |
| 14 | 70 | 70 | 1 | 6.45 | 1.61 |
| 15 | 70 | 80 | 0.7 | 4.2 | 1.05 |
| 16 | 70 | 80 | 0.1 | 3.55 | 0.89 |

QoL, quality of life.

*TABLE 8 Weights by age at onset and age at death*

| | Age at onset | | | | | |
|---|---|---|---|---|---|---|
| **Age at death** | **1** | **10** | **20** | **40** | **60** | **70** |
| **1** | | | | | | |
| **10** | 0.47 | | | | | |
| **20** | 0.95 | 0.87 | | | | |
| **40** | 1.25 | 1.16 | 1.04 | | | |
| **60** | 1.09 | 1.00 | 0.90 | 1.00 | | |
| **70** | 0.95 | 0.88 | 0.78 | 0.86 | 1.09 | |
| 80 | 0.80 | 0.74 | 0.65 | 0.73 | 0.93 | 1.05 |

who fall ill at ages between 10 and 40 rather than the very old or the very young, and a preference for treating those who will die either very young or very old. In this, and similar tables to follow, blank cells are simply representative of unfeasible comparisons.

In *Table 9*, age at death is fixed at 60 and the other attributes are allowed to vary. Again we can see that the weights for age at onset are quadratic, starting close to 1 for an age at onset of 1, moving towards 0 up to age at onset of 20 before moving towards 1 as age at onset increases further. For severity we can see a preference for treating those who are closer to the middle of the severity range. The weights are nearest 0 for a severity of 0.4 and closer to 1 for severity scores of 0.1 (least severe) and 1 (most severe).

## Compensating variation approach

Another approach to calculating the relative weights attached to different types of QALYs, or beneficiaries of QALYs, is to use the Hicksian compensating variation approach to welfare measurement. The method for calculating the compensating variation using discrete data is due to Small and Rosen[57] and was introduced to the health economics literature in the context of DCEs by Lancsar and Savage.[58]

In general, the compensating variation (CV) is calculated by valuing in monetary terms the change in expected utility due to a policy change (e.g. change in price or quality of a good/service, or, in our study, a change in health state) as the change in income required to return the individual to their initial level of utility, that is, to compensate them for the change.

The CV for discrete choice data takes the following form:

$$CV = -\frac{1}{\lambda}\left[\ln\sum_{j=1}^{J} e^{V_j^0} - \ln\sum_{j=1}^{J} e^{V_j^1}\right] \quad (5)$$

where $V_j^0$ and $V_j^1$ are the value of the indirect utility function for each choice option $j$ before and after the policy change respectively; $J$ is the number of options in the choice set; and $\lambda$ is the marginal utility of income, or its proxy.

While a monetary value is the most convenient to turn the change in expected utility into a common metric, in fact any quantitative metric could be used as the numeraire. In the current study, instead

*TABLE 9 Weights by age at onset and severity*

| | Age at onset | | | | |
|---|---|---|---|---|---|
| **Severity** | **1** | **10** | **20** | **40** | **60** |
| **0.1** | 0.925 | 0.850 | 0.750 | 0.840 | |
| **0.4** | 0.875 | 0.805 | 0.713 | 0.800 | |
| **0.7** | 1.088 | 1.000 | 0.900 | 1.000 | |
| 1 | | | | | 1.625 |

of using income it is possible to calculate the CV for a move from one health state to another using the marginal utility of QALY gains as the numeraire. That is, we calculate the number of additional QALYs required to be given or taken away in the new health state to equalise the utility derived from the base and new health states.

Like the probability-of-choice approach described previously, the CV approach can be used to value changes in, and create weights for, whole scenarios (i.e. entire health states). In addition, the CV can also be calculated to value and derive weights for individual characteristics such as age at onset, age at death and severity. To calculate both types of weights, we value changes from an initial health state, namely the same reference case as used in the calculation of the probability weights: $AO=40$, $AD=60$, $QL=0.7$ and $QALY=4$.

From this initial health state we can calculate the CV, measured in numbers of QALYs, associated with a move to a new health state, described by new levels on the above attributes. This can involve an entire scenario or the valuation of each attribute one at a time. We use these CV measures to calculate weights for entire health states and for individual characteristics that describe the health states, by taking the ratio of the total QALYs required in the new health state to equalise utility and the original number of QALYs in the base model. This welfare theoretic approach also allows investigation of the strength of preference, as indicated by the magnitude of the CV.

As with the probability-of-choice approach, the CV and relative weights reported in this section used the conditional logit results for the powered model reported in *Table 6*. The CV and relative weights for the entire health states described in *Table 7* are reported in *Table 10*. In each of these, QALY gains are held constant at 4.

Before discussing the results in *Table 10*, it is important to note that 'QALY gain with treatment' represents the number of QALYs in the health-state scenarios presented to respondents in the survey. This is not the same as the 'QALY gain' in the probability-of-choice tables, which is the result of the calculation of the number of QALYs required to

***TABLE 10*** *Weights based on compensating variation approach (version 1)*

| Base | Age at onset 40 | Age at death 60 | QoL lost 0.7 | QALY gain with treatment 4 | CV | Weight per scenario |
|---|---|---|---|---|---|---|
| ***Comparison*** | | | | | | |
| 1 | 1 | 1 | 1 | 4 | 0.0224 | 1.0056 |
| 2 | 1 | 10 | 0.7 | 4 | −0.5788 | 0.8553 |
| 3 | 1 | 10 | 0.1 | 4 | −0.7280 | 0.8180 |
| 4 | 1 | 20 | 0.7 | 4 | −0.0435 | 0.9891 |
| 5 | 1 | 20 | 0.1 | 4 | −0.1787 | 0.9553 |
| 6 | 10 | 10 | 1 | 4 | −0.3650 | 0.9087 |
| 7 | 10 | 20 | 0.7 | 4 | −0.1097 | 0.9726 |
| 8 | 10 | 20 | 0.1 | 4 | −0.2468 | 0.9383 |
| 9 | 10 | 40 | 0.7 | 4 | 0.1190 | 1.0297 |
| 10 | 10 | 40 | 0.1 | 4 | −0.0112 | 0.9972 |
| 11 | 40 | 40 | 1 | 4 | 0.3591 | 1.0898 |
| **12** | **40** | **60** | **0.7** | **4** | **0** | **1** |
| 13 | 40 | 60 | 0.1 | 4 | −0.1339 | 0.9665 |
| 14 | 70 | 70 | 1 | 4 | 0.4050 | 1.1013 |
| 15 | 70 | 80 | 0.7 | 4 | 0.0390 | 1.0098 |
| 16 | 70 | 80 | 0.1 | 4 | −0.0937 | 0.9766 |

CV, compensating variation; QoL, quality of life.

equalise the probability of choosing the base case and alternative health states.

Both the CV and weights are interpreted relative to the reference case. The same reference case is included as comparison 12, for which the CV is 0, as would be expected since the health state has not changed, and the weight is 1. In *Figure 13*, relative to this base case, a negative CV indicates that the new state is preferred to the base case and that individuals are 'willing to pay' to secure the new health state – see for example, comparison 2. The consequent weight of 0.855 indicates that a QALY given in the base state is worth only 0.855 of a QALY given in the new state. A positive CV indicates that individuals require compensation (in terms of being given additional QALYs) in the new health state – see for example, comparison 1. A weight of greater than 1 indicates preference for the base state.

For the four comparison health states that involve instant death without treatment (comparisons 1, 6, 11 and 14) there is a preference for giving the QALYs to the base case, with the exception of comparison 6 which favours the alternative. This is in accordance with the underlying utility values derived from each health state. That is, the utility derived from the base case is greater than the utility derived from the alternatives with the exception of comparison 6 (a 10-year-old who dies instantly without treatment), which yields a higher utility than the base case.

Still looking at the magnitude of the CV relative to the base case, respondents were willing to pay the largest amount (in terms of QALYs) for comparison 3, namely a health state in which the individual becomes unwell at age 1, without treatment QoL drops by 90% and they die at age 10, whereas with treatment they gain four QALYs. The least preferred beneficiary of four QALYs is someone who becomes unwell at age 70 and without treatment dies instantly (comparison 14).

Welfare measures and relative weights can also be calculated for the same complete health states described in *Table 10*, but where QALYs, in addition to the other attributes, are allowed to change between the initial and new health states. These results are presented in *Table 11*.

Again, a weight greater (less) than 1 indicates a preference for the base (alternative) health state. Looking at these results in aggregate, the number of QALYs gained is driving the preference between the base and comparison health states. Without exception, for QALY gains in the new health state of 10, 15 and 30, the new health state is always preferred to the base (which only contains four QALYs). For all comparison health states involving only one QALY gain, the base case (containing four QALYs) is always preferred. For comparison health states with two QALY gains, the base case is generally preferred to the comparison case, except when those two QALYs are given to 1-year-olds whose QoL drops by 0.1 and who die at age 10, or 1-year-olds whose QoL drops by 0.7 and who die at age 20.

The CV and relative weights associated with the individual attributes (age at onset, age at death and severity) are reported in *Table 12*. These calculations used the same initial base health state of age at onset of 40, age at death of 60, QoL lost of 0.7 and four QALYs. The CV and weights per attribute level were calculated by changing one attribute at a time in the new health state, holding all else constant.

Age at onset of 20 and 40 appear to be the most important, which could relate to considerations of the ages at which individuals are most productive to society (and also more likely to have

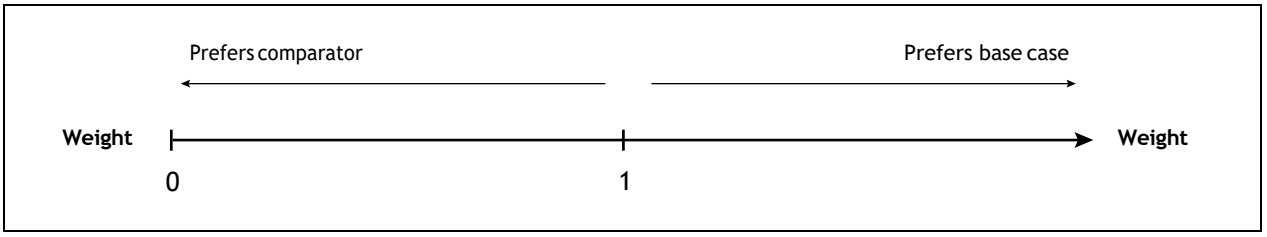


***FIGURE 13*** *Preferences and weights in the discrete choice study.*

***TABLE 11*** *Weights based on compensating variation approach (version II)*

| Base | Age at onset 40 | Age at death 60 | QoL lost 0.7 | QALY gain with treatment 4 | CV | Weight per scenario |
|---|---|---|---|---|---|---|
| *Comparison* | | | | | | |
| 1 | 1 | 1 | 1 | 1 | 0.8568 | 1.2142 |
| 2 | 1 | 10 | 0.7 | 1 | 0.3853 | 1.0963 |
| 3 | 1 | 10 | 0.1 | 1 | 0.2642 | 1.0660 |
| 4 | 1 | 20 | 0.7 | 1 | 0.8065 | 1.2016 |
| 5 | 1 | 20 | 0.1 | 1 | 0.7022 | 1.1756 |
| 6 | 10 | 10 | 1 | 1 | 0.5561 | 1.1390 |
| 7 | 10 | 20 | 0.7 | 1 | 0.7557 | 1.1889 |
| 8 | 10 | 20 | 0.1 | 1 | 0.6491 | 1.1623 |
| 9 | 10 | 40 | 0.7 | 1 | 0.9299 | 1.2325 |
| 10 | 10 | 40 | 0.1 | 1 | 0.8312 | 1.2078 |
| 11 | 40 | 40 | 1 | 1 | 1.1082 | 1.2771 |
| 12 | 40 | 60 | 0.7 | 1 | 0.8398 | 1.2100 |
| 13 | 40 | 60 | 0.1 | 1 | 0.7370 | 1.1842 |
| 14 | 70 | 70 | 1 | 1 | 1.1417 | 1.2854 |
| 15 | 70 | 80 | 0.7 | 1 | 0.8695 | 1.2174 |
| 16 | 70 | 80 | 0.1 | 1 | 0.7680 | 1.1920 |
| 17 | 1 | 1 | 1 | 2 | 0.4895 | 1.1224 |
| 18 | 1 | 10 | 0.7 | 2 | −0.0460 | 0.9885 |
| 19 | 1 | 10 | 0.1 | 2 | −0.1813 | 0.9547 |
| 20 | 1 | 20 | 0.7 | 2 | 0.4316 | 1.1079 |
| 21 | 1 | 20 | 0.1 | 2 | 0.3122 | 1.0780 |
| 22 | 10 | 10 | 1 | 2 | 0.1462 | 1.0366 |
| 23 | 10 | 20 | 0.7 | 2 | 0.3733 | 1.0933 |
| 24 | 10 | 20 | 0.1 | 2 | 0.2517 | 1.0629 |
| 25 | 10 | 40 | 0.7 | 2 | 0.5741 | 1.1435 |
| 26 | 10 | 40 | 0.1 | 2 | 0.4601 | 1.1150 |
| 27 | 40 | 40 | 1 | 2 | 0.7821 | 1.1955 |
| 28 | 40 | 60 | 0.7 | 2 | 0.4699 | 1.1175 |
| 29 | 40 | 60 | 0.1 | 2 | 0.3519 | 1.0880 |
| 30 | 70 | 70 | 1 | 2 | 0.8215 | 1.2054 |
| 31 | 70 | 80 | 0.7 | 2 | 0.5041 | 1.1260 |
| 32 | 70 | 80 | 0.1 | 2 | 0.3874 | 1.0969 |
| 33 | 1 | 1 | 1 | 10 | −0.9594 | 0.7602 |
| 34 | 1 | 10 | 0.7 | 10 | −1.6568 | 0.5858 |
| 35 | 1 | 10 | 0.1 | 10 | −1.8256 | 0.5436 |
| 36 | 1 | 20 | 0.7 | 10 | −1.0374 | 0.7407 |
| 37 | 1 | 20 | 0.1 | 10 | −1.1961 | 0.7010 |
| 38 | 10 | 10 | 1 | 10 | −1.4122 | 0.6469 |

*TABLE 11* *Weights based on compensating variation approach (version II) (continued)*

| Base | Age at onset 40 | Age at death 60 | QoL lost 0.7 | QALY gain with treatment 4 | CV | Weight per scenario |
|---|---|---|---|---|---|---|
| 39 | 10 | 20 | 0.7 | 10 | −1.1153 | 0.7212 |
| 40 | 10 | 20 | 0.1 | 10 | −1.2754 | 0.6811 |
| 41 | 10 | 40 | 0.7 | 10 | −0.8442 | 0.7890 |
| 42 | 10 | 40 | 0.1 | 10 | −0.9992 | 0.7502 |
| 43 | 40 | 40 | 1 | 10 | −0.5535 | 0.8616 |
| 44 | 40 | 60 | 0.7 | 10 | −0.9859 | 0.7535 |
| 45 | 40 | 60 | 0.1 | 10 | −1.1437 | 0.7141 |
| 46 | 70 | 70 | 1 | 10 | −0.4971 | 0.8757 |
| 47 | 1 | 1 | 1 | 15 | −1.6162 | 0.5960 |
| 48 | 1 | 10 | 0.7 | 15 | −2.3566 | 0.4109 |
| 49 | 1 | 10 | 0.1 | 15 | −2.5338 | 0.3666 |
| 50 | 1 | 20 | 0.7 | 15 | −1.6997 | 0.5751 |
| 51 | 1 | 20 | 0.1 | 15 | −1.8691 | 0.5327 |
| 52 | 10 | 10 | 1 | 15 | −2.0985 | 0.4754 |
| 53 | 10 | 20 | 0.7 | 15 | −1.7829 | 0.5543 |
| 54 | 10 | 20 | 0.1 | 15 | −1.9535 | 0.5116 |
| 55 | 10 | 40 | 0.7 | 15 | −1.4924 | 0.6269 |
| 56 | 10 | 40 | 0.1 | 15 | −1.6588 | 0.5853 |
| 57 | 40 | 40 | 1 | 15 | −1.1777 | 0.7056 |
| 58 | 40 | 60 | 0.7 | 15 | −1.6446 | 0.5889 |
| 59 | 40 | 60 | 0.1 | 15 | −1.8132 | 0.5467 |
| 60 | 1 | 1 | 1 | 30 | −3.2013 | 0.1997 |
| 61 | 1 | 10 | 0.7 | 30 | −4.0036 | −0.0009 |
| 62 | 1 | 10 | 0.1 | 30 | −4.1926 | −0.0482 |
| 63 | 1 | 20 | 0.7 | 30 | −3.2930 | 0.1768 |
| 64 | 1 | 20 | 0.1 | 30 | −3.4780 | 0.1305 |
| 65 | 10 | 10 | 1 | 30 | −3.7265 | 0.0684 |
| 66 | 10 | 20 | 0.7 | 30 | −3.3840 | 0.1540 |
| 67 | 10 | 20 | 0.1 | 30 | −3.5696 | 0.1076 |
| 68 | 10 | 40 | 0.7 | 30 | −3.0648 | 0.2338 |
| 69 | 10 | 40 | 0.1 | 30 | −3.2481 | 0.1880 |
| 70 | 40 | 40 | 1 | 30 | −2.7139 | 0.3215 |
| 71 | 40 | 60 | 0.7 | 30 | −3.2325 | 0.1919 |
| 72 | 40 | 60 | 0.1 | 30 | −3.4171 | 0.1457 |

CV, compensating variation; QoL, quality of life.

*TABLE 12 Compensating variations per attribute*

| Attribute | Level | CV | Weight |
|---|---|---|---|
| Age at onset | 1 | 0.0688 | 1.0172 |
| | 10 | 0.0043 | 1.0011 |
| | 20 | −0.0862 | 0.9784 |
| | 40 | 0 | 1 |
| | 60 | 0.1741 | 1.0435 |
| | 70 | 0.2675 | 1.0669 |
| Age at death | 1 | −0.3173 | 0.9207 |
| | 10 | −0.6568 | 0.8358 |
| | 20 | −0.1141 | 0.9715 |
| | 40 | 0.1148 | 1.0287 |
| | 60 | 0 | 1 |
| | 70 | −0.1108 | 0.9723 |
| | 80 | −0.2438 | 0.9390 |
| Severity (QoL loss) | 1 | 0.2517 | 1.0629 |
| | 0.7 | 0 | 1 |
| | 0.4 | −0.1730 | 0.9568 |
| | 0.1 | −0.1339 | 0.9665 |

CV, compensating variation; QoL, quality of life.

dependants), followed by the very young, with least importance given to the eldest two age groups of 60- and 70-year-olds. Higher preference is given to those who die young (aged 10, 1 and 20).

When interpreting the severity weights, we first note that the base case of QoL lost = 0.7 is one of the middle severity levels, so movements from this base case can represent an improvement in QoL (a loss of only 0.4 or 0.1 rather than 0.7 QoL) for which respondents were willing to pay or a detriment (QoL lost = 1) for which respondents required compensation. As movements from the base are in both directions, absolute values are used. Relative to the base case, again, more severe states are less preferred, as would be expected given the results reported above. In particular, moving from the base to the worst level of severity had the largest impact in terms of the CV; that is, respondents would have to be compensated a relatively large amount in absolute value for a health state that involved instant death. Willingness to pay (in terms of QALYs) to improve the initial level of severity – that is, to lose only 10% or 40% of QoL rather than the base case of 70% was similar across these two levels, producing similar weights (0.9567 and 0.9665 respectively).

## Further investigation of severity

In the above regression analyses, the severity variable (QoL lost) appeared to behave in the opposite direction to how it had been weighted in previous studies. Although the relationship is close to being flat, this different result is nevertheless reflected to some extent in the regressions and in the weights presented in the previous section. This issue was, therefore, thought worthy of further investigation. In particular, the question to address is why QALYs gained by those who would otherwise die instantly (i.e. whose QoL lost would be 1) were not weighted higher than other types of QALY gain.

To investigate this further, we isolated choice sets in which one scenario involved QoL lost being equal to 1 (i.e. age at onset = age at death), referred to here as the 'life-saving' option. By selecting choices with just one life-saving option, by definition, the comparator scenario within any such choice would be any of all other types (e.g. QoL enhancing only and L-shaped gains which enhance QoL, but also add years to life). This amounted to 1253 choice sets in total.

On 446 occasions, the life-saving option was chosen. On these occasions, the mean QALY gain being offered in the life-saving option was 12.8 (median = 7) as opposed to a mean QALY gain of 10.7 (median = 7.8) in the alternative (which was not chosen). The life-saving option, by definition, presents QoL lost as 1, and the mean QoL lost presented in the rejected scenario was 0.54. On these occasions it seems that, on average, QALY gains are larger and beneficiaries are more severe in life-saving options, so it is not surprising that this was chosen.

On the 809 occasions when life saving was not chosen, the mean QALY gain being offered in the life-saving option was 5.43 (median=4) as opposed to 13.4 (median = 7.9) in the alternative; the alternative having a mean QoL lost of 0.53. On these occasions it seems that, on average, QALY gains are much larger when beneficiaries are in a less severe state. This seems to have driven the choices made, making severity appear less important than it might be if choices were offered in which the scenarios were closer together in terms of numbers of QALYs gained. This may explain why results of matching-type studies, which essentially set QALYs equal and highlight differences in equity variables of interest (such as severity), tend to reveal more enhanced weights for such variables with beneficiaries in more severe states being more highly valued. An improved design might have meant that the interaction between QALYs and severity could have been allowed for more comfortably, which may have shown that the shape of the relationship between severity and utility was positive when treatment provides large QALY gains – a preference shown for the most severe – and negative when treatment provides small QALY gains – a preference for the least severe. That is, when small amounts of QALYs are being offered to someone who is already in a very bad state (or who will die imminently) it might seem reasonable not to place additional value or weight on that more severe state. Indeed, this is one interpretation of the results from one other study we are aware of, which provides a result that runs counter to the usual view of severity;[59] however, our data are not able to show this one way or the other. Other studies on severity, based on matching-style approaches, are discussed in Chapter 5, Concluding remarks.

## Concluding remarks

The discrete choice part of the relativities study has exposed the reader to three things: the challenges of deriving relative weights for QALYs using discrete choice methods; two novel approaches to eliciting relative QALY weights from a discrete choice experiment; and some initial empirical weights.

Overall, the estimated weights are similar across different regression models and different weighting procedures. There are obviously differences between the simple model and that including higher-order polynomials due to the imposed restriction of monotonicity in the former. Comparing the CV weights with those from the probability-of-choice approach, we can see that the range of weights is smaller (0.8–1.1) for the former, and that they are generally nearer to 1, suggesting no difference between the base case and the comparison scenario. More generally, this would indicate that age and severity have little impact on QALY gains. The range for the probability-of-choice approach is 0.38–1.61. As the weights are based on the same regression model, it is unsurprising that the same scenarios result in the lowest and highest weights in both methods.

The narrow range of weights, especially using the CV approach, and the lack of statistical significance of age at onset and severity in the regression models, raise a number of issues. It might be that, when presenting information on such variables within a context in which health gains also vary, the impacts of age and severity are diminished. Indeed, this could be seen as an advantage of a discrete choice approach in which equity attributes are embedded within scenarios in which health gains (or QALYs) are also allowed to vary. However, it could also be the case that this result is driven by the compromises made necessary in the experimental design.

There are some clear patterns. For example, as the age at onset increases, the weights move towards 0 (until age 20) and then move towards 1. This shows a preference for those who fall ill between the ages of 10 and 40, with 20 being the most preferred. As age at death increases, the weights move towards 0 (until age 10) and then move towards 1 (up to age 40), before moving back towards 0. This shows a preference for individuals who will die at 1, 10 and 80 and less of a preference for those between 20 and 70. As severity increases, there is a preference for those who are slightly ill, a QoL loss of 0.4 and 0.1, compared with those who lose 0.7 and 1. Of course, the advantage with these approaches is that it is possible that age effects outweigh severity effects, or vice versa, meaning that generalisations are harder to make unless we can invoke ceteris

paribus assumptions. In other words, different weights on QALYs are required for a variety of contexts.

With respect to more methodological issues, we have demonstrated two novel approaches to deriving weights from our empirical data. First, what issues arise from people's perceived similarities or differences in the weights? At a conceptual level, the CV and probability-of-choice approaches to deriving relative weights are similar. However, instead of equalising the probability of choice between the two health states, the CV method equalises the utility associated with the two health states – a subtle but important difference. Second, does this difference matter? One possibility, of course, is that using the CV approach overcomes one limitation of the probability approach in that the CV can be used to calculate weights per attribute in addition to weight for entire health-state scenarios.

The controversy of these results lies in the small impact of the equity variables of interest (age and severity). However, this should not necessarily be taken as an indication that weighting of QALYs is not desirable. In the short term, the results need to be reinforced by calculation of confidence intervals around the weights that have been presented. This is a significant task; hence such data have not been presented in this report. In addition, despite the controversial nature of this result, as mentioned above, it may have arisen due to one potential advantage of a discrete choice approach, whereby variations in equity attributes are considered alongside variations in health gains. Our main longer-term recommendation would, therefore, be for more research, as replication of such a result would be crucial prior to any subsequent policy recommendations. This research should also pursue alternative experimental design strategies in order to address the issue of implausible scenarios while still maintaining desirable design properties, so ensuring we can estimate the effects of interest with improved efficiency. Further important potential challenges, such as the nature of the (multiplicative) functional form used in the statistical analysis and the transformation of variables from the original forms in which they were presented in *Table 6,* are discussed further in Chapter 5, Concluding remarks.

# Matching study

## Basic approach

We report here the 'matching' relativities study which, along with the DCE study (see Chapter 4), set out to determine whether the UK public wish to attach more weight to some QALYs than to others. To recap, a QALY is 1 year in full health and years spent in less than full health are 'quality adjusted' in order to take account of the morbidity associated with disability, disease or illness. As QALYs combine morbidity (QoL) and mortality (length of life) on a single scale, they allow comparisons to be made across interventions with different types of health outcomes (e.g. QoL enhancing versus life extending). In the standard 'unweighted' QALY model, all QALYs are of equal value. For example, the age of recipients does not matter, as long as the QALY gain is the same. Likewise, the standard model assumes that equal QALY gains are of equal value regardless of how severely ill the patients are prior to treatment. The aim of the matching – and DCE studies – is then to address the question of whether a QALY is a QALY is a QALY. Or put another way, is the 'standard' model correct?

While each respondent in the survey answered a set of matching and discrete choice questions, we concentrate here on the results of the matching study, the main aims of which were:

- to estimate the relative weights to be attached to a QALY according to the age of recipients
- to estimate the relative weights to be attached to a QALY according to the severity of illness of recipients.

The secondary aims of the matching study were to assess the impact of certain other attributes considered to be of interest to respondents and which featured in the design such as gain in life expectancy and gain in QoL.

## Methods

### Matching questions

Briefly, 'matching' questions ask people to state the number of outcomes of one kind they consider to be 'just as good as' a specified number of outcomes of another kind. A typical matching question in this setting would be something along the following lines:

Consider two groups of 100 people: 100 of type A and 100 of type B (these types being described and differentiated by age and severity of illness prior to treatment). With treatment, each of these types could experience some given health benefit. If there were only enough resources to treat one group, which would you prefer it to be?

Suppose the respondent chooses to give priority to Group A. The number of beneficiaries in Group B is held fixed at 100 while the number of A-types is reduced to the point where the respondent finds it hard to prioritise between that number (X) of A-types and 100 B-types. By systematically varying the age and health of the people described in Groups A and B in a series of questions, it is possible to investigate the relative weights attached to health gains in people of different ages and levels of health.

### The QALY grid approach

Recall that the animated powerpoint introduction to the CAPI used a graph with QoL (from 0% to 100%) on the vertical axis and Age (from 0 to 80) on the horizontal axis. Clearly, all questions posed in the relativities studies – both matching and DCE – must involve health gains that fall within this defined space. Further, all questions must be 'logical' in that there must be some positive gain in terms of either QoL or length of life, or both. While it was necessary to incorporate such considerations explicitly into the DCE design (the underlying principle of which is to allow all attributes to vary independently), 'illogical' scenarios were ruled out a priori in the matching design. There remain, however, an enormous number of potential pairings that can be made, and the challenge was to come up with the most parsimonious design that allowed all the research questions to be addressed. We were keen that the design be capable of testing the standard 'unweighted' QALY model, i.e. where all relative

weights would be equal to one, as well as being able to detect systematic departures from this baseline assumption with respect to the age of patients and their 'untreated' QoL. To that end we developed the 'QALY grid' approach, summarised briefly below.

Consider again the graph with QoL (from 0% to 100%) on the vertical axis and Age (from 0 to 80) on the horizontal axis. By partitioning the vertical into five 20% ranges and dividing the horizontal into four 20-year intervals, we get 20 'cells' (numbered 1–20 in *Table 13*).

Improving a person's health by any one cell, i.e. by 20% per year for 20 years, would, under the 'standard' model, give four QALYs. But are all QALYs weighted equally, irrespective of where they are located in the grid? For example, suppose that type A people were those whose QoL would, if untreated, be 60% at (and after) age 20 and treatment could raise that QoL to 80% for the next 20 years, whereas type B people would be at 80% from the age of 60 but could receive treatment that would restore them to 100% for the subsequent 20 years. Both types would gain four 'standard' QALYs; but do respondents give both treatments equal priority? If not, which group of 100 do they give priority to? And if the number of beneficiaries in that group is reduced, at what value of $X < 100$ do the two treatments receive equal priority? For example, if a respondent finds it hard to prioritise when the same total expenditure of resources could treat either 40 people of type A or 100 people of type B, that would suggest that such a respondent gives two-and-a-half times as much weight to a QALY in cell 14 as in cell 20.

If we have data on respondents' relative values for the 20 cells, and the value of a 'reference' cell is set equal to 1 (suppose we call that $v_1$), we can derive a set of relative weights ($v_2$–$v_{20}$) for the remaining 19 cells in the grid. Our basic aim was to estimate these relative values. If each cell is paired with every adjacent cell, we can estimate ratios such as $v_1/v_2$ and $v_1/v_5$. This requires 31 pairings. An example of a question respondents were actually presented with in the matching study is reproduced in *Figure 14*.

While this is a parsimonious design, extrapolating from these 31 'single cell' questions requires a number of strong assumptions about the nature of respondents' preferences; assumptions we did not feel able to make without first testing their validity.

So another 41 questions in the overall design are intended to embody a number of such tests. Some of these questions look at blocks of cells: for example, each row P–T is paired with every adjacent row, and each column with every adjacent column. One important difference between the 'whole column' and 'single cell' pairings is that the former *necessarily* involve 'life-extending' QALYs (as they offer an additional 20 years of life expectancy). For example, pairing cells 10 and 11 asks respondents to consider an additional four QALYs (moving from 40% to 60% health for 20 years) either between the ages of 20 and 40 or between the ages of 40 and 60. In contrast, pairing columns B and C asks respondents to consider extending life (at 100% health) for 20 years at the age of either 20 or 40. If respondents were to have preferences over 'life-extending', rather than 'life-enhancing' QALYs, we might expect the results of the questions involving the whole column to be somewhat different than implied by the 'sum of the parts', i.e. those pairings of the single cells that make up the columns.

The design also allows us to test for consistency via chaining with equal numbers of steps, e.g.

***TABLE 13** The QALY grid*

| QoL | A | B | C | D | |
|---|---|---|---|---|---|
| 80–100% | 17 | 18 | 19 | 20 | **P** |
| 60–80% | 13 | 14 | 15 | 16 | **Q** |
| 40–60% | 9 | 10 | 11 | 12 | **R** |
| 20–40% | 5 | 6 | 7 | 8 | **S** |
| 0–20% | 1 | 2 | 3 | 4 | **T** |
| Age | 0–20 | 20–40 | 40–60 | 60–80 | |

QoL, quality of life

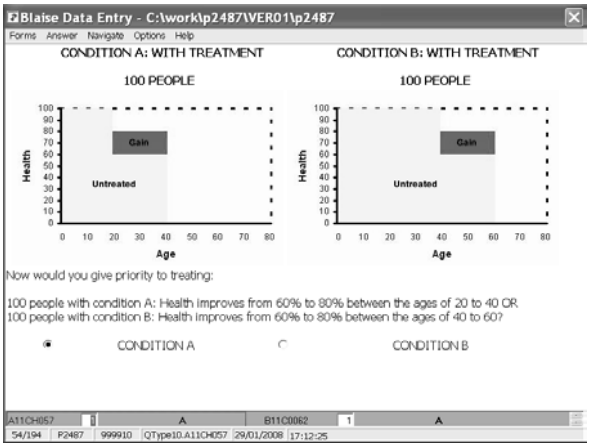


***FIGURE 14** Matching question from CAPI.*

$(v_1/v_5)*(v_5/v_9)$ should equal $(v_1/v_9)$. To test for chaining with different numbers of steps, we have built in some *non-adjacent pairings*. This has been done for *single cells* (e.g. compare the pairs {1, 2}, {2, 3} and {1, 3}), for *whole columns*, e.g. compare {A, B}, {B, C} and {A, C} and for *whole rows*, e.g. compare {R, S}, {S, T} and {R, T}.

Finally, there are the tests for independence of the profile in which a cell appears. Consider the different ways in which cell 14 may appear in a scenario. In *Figure 15*, 14S shows the patient to be in good health until aged 20 when – without treatment – their QoL falls from 100% to 60% and they die at the age of 40.

This depicts the 'standard' representation that applies to all other single cell comparisons used here. In 14X the patient is at 80% of good health until the age of 20, when their QoL falls to 60%, and again they die at the age of 40. In 14Z, the patient is at 80% of good health until the age of 20, when their QoL falls to 60%, but now they live until the age of 80. Finally, in 14Y the patient is in good health from birth and lives to the age of 80. Each profile test compares the 'standard' representation with one of the three other representations of that cell (i.e. 14S versus 14X, 14S versus 14Z, 14S versus 14Y). Three different sets of profile tests were carried out, involving cells 7, 11 and 14. 'Fair innings' type arguments would suggest a tendency, all else being equal, to favour those patients whose total lifetime QALYs were lower.

For each of the three cells used in the profile tests, a 'fair innings' type argument would predict that *more* weight would be given to S than to Y (as the total lifetime QALYs is less in S than in Y), but S would be given *less* weight than X (as the total lifetime QALYs is greater in S than in X). Although not shown in detail here, S would be given less weight than Z in the cases of cells 7 and 11, but S would be given more weight than Z in the case of cell 14.

The 72 questions were spread evenly across 12 versions of the matching questionnaire, each respondent answering six matching questions (along with eight DCEs).

## Aggregating results

Suppose the following represents data from five respondents who were indifferent between treating X patients in Group A and Y patients in Group B (with either X or Y = 100):

| | **Group A** X | **Group B** Y |
|---|---|---|
| Resp 1 | 100 | 50 |
| Resp 2 | 100 | 90 |
| Resp 3 | 100 | 50 |
| Resp 4 | 50 | 100 |
| Resp 5 | 90 | 100 |

This pattern indicates that respondents 1 to 3 preferred to give priority to Group B at the first matching iteration, i.e. when there were 100 in each group, and subsequently set the number treated in that group (Y) at some number lower

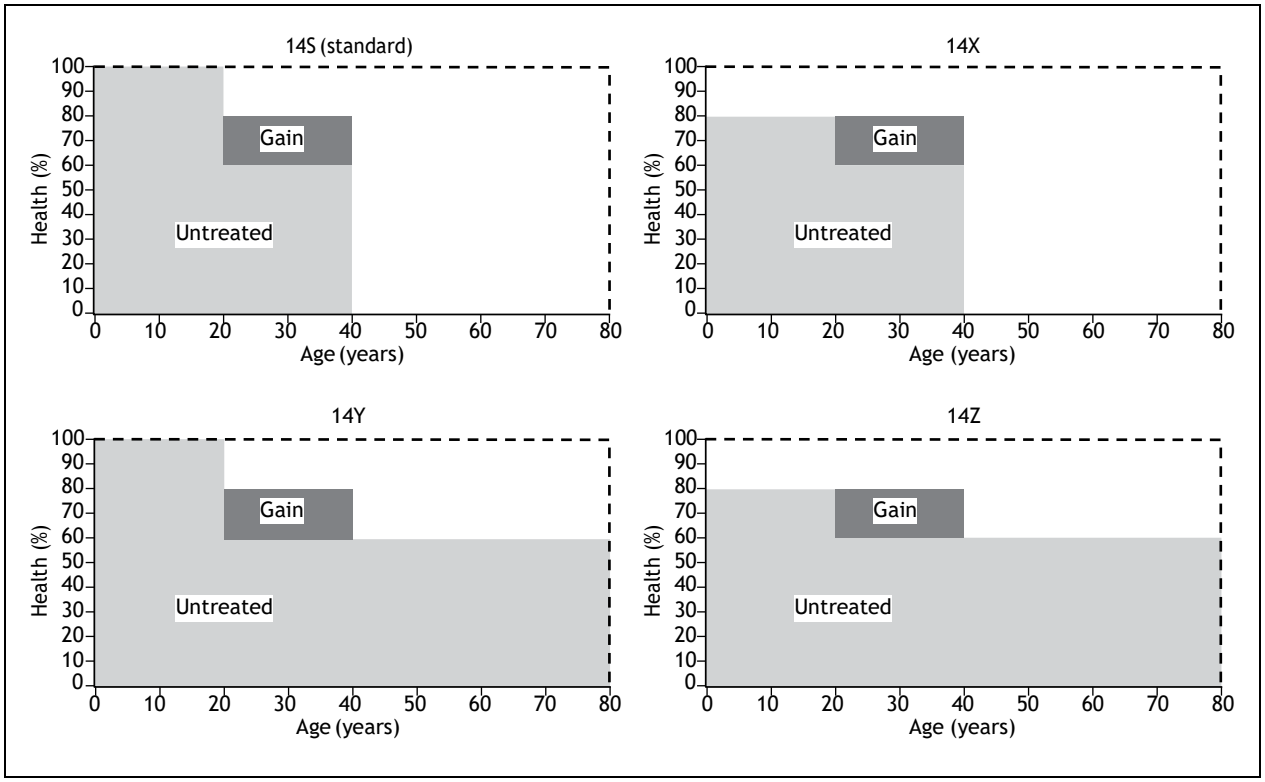


***FIGURE 15*** *The profile tests (cell 14).*

than 100. In contrast, respondents 4 and 5 preferred to give priority to Group A initially, and subsequently set the number treated in that group (X) at some number lower than 100. Clearly, respondents 3 and 4 have identical, but opposite, preferences. Yet, the ratio X/Y is 2:1 for respondent 3 and 0.5:1 for respondent 4 (and vice versa for Y/X) resulting in a very different effect on the arithmetic mean of individual ratios. As the decision to take the ratio X/Y or Y/X is arbitrary, this asymmetry is an undesirable property of an aggregation method.

There are two ways in which we shall aggregate the results of the matching questions in order to overcome problems due to lack of symmetry. The first of these – the ratio of means method – is the method used previously in work commissioned for the Health and Safety Executive (HSE), which set out to investigate whether premia ought to be attached to the prevention of certain types of deaths.[60]

### *The ratio of means method*

In this method the most favoured group attracts a value of 1 and the less favoured a value equal to the number of patients treated in the more preferred group divided by the number treated in the less preferred group. For example, Group B was respondent 1's 'most preferred' group and they set Y at 50. Group B then attracts a value of 1 and Group A a value equal to Y/X = 0.5. Following this principle for all five respondents yields the following values:

| | **Group A** | **Group B** |
|---|---|---|
| Resp 1 | 0.5 | 1 |
| Resp 2 | 0.9 | 1 |
| Resp 3 | 0.5 | 1 |
| Resp 4 | 1 | 0.5 |
| Resp 5 | 1 | 0.9 |

Means of each column are then taken and the ratio of those means are calculated. In this case, the mean value for Group A is 0.78 while the mean value for Group B is 0.88. The implied weight of Group B relative to Group A is then 0.88/0.78 = 1.13. Alternatively, the implied weight of Group A relative to Group B is 1/1.13 = 0.88, i.e. the measure is symmetrical.

### *Aggregating ratios derived from individual responses*

Alternatively, the ratios X/Y may be computed for each individual respondent and then the median can be identified. The ratios are as follows:

| | X/Y |
|---|---|
| Resp 1 | 2:1 |
| Resp 2 | 1.11:1 |
| Resp 3 | 2:1 |
| Resp 4 | 0.5:1 |
| Resp 5 | 0.9:1 |

Taking medians of the individual ratios results in a weight of Group A relative to Group B of approximately 1.11. As with the ratio of means method above, this measure is symmetrical, i.e. the median of the Y/X ratios is 1/1.11 = 0.9.

In this particular example, both measures produce very similar measures of central tendency. However, as we shall see below, this may not always be the case.

# Results

## Responses to first matching question

Before going on to look at the weights derived from the aggregate matching data, we begin by looking at the number of respondents choosing to give priority to one group or the other in the first matching iteration, i.e. when the number of patients treated was 100 in both groups. These data are the most straightforward to interpret and are arguably the simplest way of illustrating the *general patterns* to emerge from the relativities study.

With regard to the general pattern relating to age, a total of 269 respondents answered a matching question involving an equivalent gain going either to 40- to 60-year-olds or to 60- to 80-year-olds. Of these 269, 173 (64%) preferred to give priority to the younger patients. Similarly, a total of 294 respondents answered a matching question involving an equivalent gain going either to 20- to 40-year-olds or to 40- to 60-year-olds. Of the 294, 182 (62%) preferred to give priority to the younger patients. In contrast, a total of 295 respondents answered a matching question involving an equivalent gain going either to 0- to 20-year-olds or to 20- to 40-year-olds. Of the 295, only 107 (36%) preferred to give priority to the younger patients. This suggested that the pattern relating to age is not a simple one and that 'non-linearities' may exist with respect to age weights.

This pattern is borne out if we look only at responses to the whole column questions (detailed above) which involved gains of 20 QALYs that were necessarily life extending. Of the 51 respondents who were asked to compare column A (20 years in full health for a newborn) with column B (20 years in full health for a 20-year-old), 17 (33.3%) preferred to treat the younger patients. In contrast, of the 54 respondents who compared column C (20 years in full health for 40-year-olds) with column D (20 years in full health for 60-year-olds), 36 (66.7%) preferred to treat the younger patients.

With regard to the general pattern relating to severity, a total of 294 respondents answered a matching question involving a move either from 60% to 80% health or from 80% to 100% health (for a given age group). Of the 294, 170 (58%) preferred to give priority to patients in the more severe health state. In contrast, of the 280 respondents who answered a matching question involving a move either from 0% to 20% health or from 20% to 40% health (for a given age group), only 105 (38%) preferred to give priority to those patients in the more severe health state. Again, observing these initial choices suggested that non-linearities may exist with respect to severity weights. The pattern is borne out by observing responses to the 'whole row' questions. Of the 44 respondents who were asked to compare row T (a move from 0% health to 20% health between birth and age 80) and row S (a move from 20% health to 40% health between birth and aged 80), 17 (38.6%) preferred to treat those patients who were worse off initially. In contrast, of the 44 respondents who were asked to compare row Q and row P, 32 (71.1%) preferred to treat those patients who were worse off initially.

As it seemed plausible that the desire to prioritise one age group over another may be related to the age of respondents, we also broke these data down by age of respondent. The sample was divided into three broad age groups: the under 40-year-olds, the 40- to 60-year-olds and the over 60-year-olds. These categories were chosen as they best coincided with those used in the QALY grid (as respondents had to be 18 or over to participate in the survey, there are too few respondents to have a separate category for 0- to 20-year-olds). In *Tables 14* and *15*, the numbers in the cells relate to those

*TABLE 14* *Preferences for treating 0- to 20-year-olds over 20- to 40-year-olds by respondent age*

| Age of respondent | Preference given to 0- to 20-year-olds | Preference given to 20- to 40-year-olds | *n* |
|---|---|---|---|
| Under 40 | 48 (49%) | 51 (51%) | 99 |
| 40–60 | 35 (34%) | 67 (66%) | 102 |
| Over 60 | 24 (26%) | 70 (74%) | 94 |
| Total | 107 | 188 | 295 |
| $\chi^2(2) = 11.79$; $p = 0.003$. | | | |

*TABLE 15* *Preferences for treating 40- to 60-year-olds over 60- to 80-year-olds by respondent age*

| Age of respondent | Preference given to 40- to 60-year-olds | Preference given to 60- to 80-year-olds | *n* |
|---|---|---|---|
| Under 40 | 60 (68%) | 21 (32%) | 81 |
| 40– 60 | 65 (61%) | 30 (39%) | 95 |
| Over 60 | 48 (49%) | 45 (51%) | 93 |
| Total | 173 | 96 | 269 |
| $\chi^2(2) = 7.65$; $p = 0.022$. | | | |

respondents preferring to treat one group over the other at the initial iteration in the matching question, i.e. when there were 100 patients in both groups. The numbers in brackets relate to the percentage of respondents in each age group preferring to treat that group over the other. In each case, differences across age groups have been tested using chi-squared tests with two degrees of freedom.

Thus, it does appear as if responses aresignificantly different according to age of respondent. The general pattern to emerge from the data (including some not shown here) is that older respondents seem more inclined than younger respondents to give preference to older patients, but this does not appear to be related just to a self-interested desire to prioritise their own age group (as they are also more likely to give preference to 20- to 40-year-olds than to 0- to 20-year-olds).

As it also seemed plausible that a desire to prioritise in terms of severity may also be related to respondent age, we broke down the data relating to severity in this way. In *Tables 16* and *17* the numbers in the cells represent respondents preferring to treat one group over the other, now defined by severity of health state (and holding age of patients constant).

Again, it does appear as if responses are significantly different according to age of respondent. In particular, *Table 17* shows that older respondents are more likely to favour that group in better health initially when the gains take place towards the top of the severity spectrum. This may reflect the fact that older respondents place greater weight on returning patients to full health than their younger counterparts.

While observing that these first choices is a useful means of illustrating the *general* patterns to emerge from the data, we go on to look at the aggregate results from the iterative matching procedures that followed.

## The relative weights

Recall that our 'QALY grid' partitions the vertical axis of *Figure 14* into five 20% ranges and the horizontal axis into four 20-year intervals, giving 20 'cells', each worth four QALYs under the 'standard' model. We then set out to determine whether all QALYs were weighted equally, irrespective of where they are located in the grid. We begin by presenting the results of the 31 single cell comparisons and the ratio of means aggregation method. The implications of using alternative aggregation methods are outlined below.

***TABLE 16*** *Preferences given to gain from 0% to 20% health over gain from 20% to 40% health (for given age of patient) by age of respondent*

| Age of respondent | Preference given to gain from 0% to 20% | Preference given to gain from 20% to 40% | *n* |
|---|---|---|---|
| Under 40 | 38 (46%) | 45 (54%) | 83 |
| 40–60 | 28 (28%) | 73 (72%) | 101 |
| Over 60 | 39 (41%) | 57 (59%) | 96 |
| Total | 105 | 175 | 280 |

$\chi^2(2) = 7.30$; $p = 0.026$.

***TABLE 17*** *Preferences given to gain from 60% to 80% health over gain from 80% to 100% health (for given age of patient) by age of respondent*

| Age of respondent | Preference given to gain from 60% to 80% | Preference given to gain from 80% to 100% | *n* |
|---|---|---|---|
| Under 40 | 43 (68%) | 20 (32%) | 63 |
| 40–60 | 75 (61%) | 49 (39%) | 124 |
| Over 60 | 52 (49%) | 55 (51%) | 107 |
| Total | 170 | 124 | 294 |

$\chi^2(2) = 7.30$; $p = 0.026$.

***TABLE 18*** *Horizontal (age) weights: ratio of means method (cell numbers are shown in square brackets)*

| QoL | | | | |
|---|---|---|---|---|
| 80–100% | [*17*] 0.837 | [*18*] 1.559 | [*19*] 1.834 | [*20*] N/A |
| 60–80% | [*13*] 0.781 | [*14*] 1.131 | [*15*] 1.564 | [*16*] N/A |
| 40–60% | [*9*] 0.676 | [*10*] 1.039 | [*11*] 1.383 | [*12*] N/A |
| 20–40% | [*5*] 0.716 | [*6*] 1.298 | [*7*] 1.433 | [*8*] N/A |
| 0–20% | [*1*] 0.867 | [*2*] 1.228 | [*3*] 1.687 | [*4*] N/A |
| Age | 0–20 | 20–40 | 40–60 | 60–80 |

N/A, not applicable; QoL, quality of life.

The numbers in each cell of *Table 18* signify the weight attached to that cell relative to the cell to the right, as derived via the method set out under 'Aggregating results'. Values greater than 1 indicate that more weight is attached to that cell than to the one to the right. Values less than 1 indicate that less weight is attached to that cell than to the one to the right. For example, a value of 1.564 in cell 15 indicates that a gain (in this case from 60% to 80% health) accruing to 40- to 60-year-olds is valued at 1.564 times the equivalent gain accruing to 60- to 80-year-olds. In contrast, the value of 0.676 in cell 9 indicates that a gain (in this case from 40% to 60% health) accruing to 0- to 20-year-olds is worth 0.676 times the equivalent gain accruing to 20- to 40-year-olds.

Following this through for all other adjacent horizontal cell comparisons, the general pattern to emerge with respect to age (holding severity constant) is as follows:

- Less weight is given to treating 0- to 20-year-olds than 20- to 40-year-olds.
- More weight is given to treating 20- to 40-year-olds than 40- to 60-year-olds.
- More weight is given to treating 40- to 60-year-olds than 60- to 80-year-olds.

***TABLE 19*** *Vertical (severity) weights: ratio of means method (cell numbers are shown in square brackets)*

| **QoL** | | | | |
|---|---|---|---|---|
| 80–100% | [*17*] N/A | [*18*] N/A | [*19*] N/A | [*20*] N/A |
| 60–80% | [*13*] 1.064 | [*14*] 1.369 | [*15*] 1.203 | [*16*] 1.618 |
| 40–60% | [*9*] 0.792 | [*10*] 1.144 | [*11*] 1.179 | [*12*] 1.247 |
| 20–40% | [*5*] 0.955 | [*6*] 1.185 | [*7*] 1.235 | [*8*] 1.247 |
| 0–20% | [*1*] 0.786 | [*2*] 0.759 | [*3*] 0.857 | [*4*] 0.801 |
| Age | 0–20 | 20–40 | 40–60 | 60–80 |

N/A, not applicable; QoL, quality of life.

Though not shown in detail here, this broad pattern is borne out in the 'whole column' comparisons, indicating that this pattern holds for life-extending QALYs involving larger gains too.

Turning now to the vertical – or severity – weights, the numbers in each cell of *Table 19* signify the weight attached to that cell relative to the cell above it. Values greater than 1 indicate that more weight is attached to that cell than to the one above it and vice versa. For example, a value of 1.618 in cell 16 indicates that, for a given patient age group, a gain from 60% health to 80% health is valued at 1.618 times that from 80% to 100% health. In contrast, a value of 0.801 in cell 4 indicates that a gain from 0% to 20% health is valued at 0.801 times that from 20% to 40% health (again keeping age of patient constant).

The general pattern to emerge with respect to severity (holding age constant) is as follows:

- Less weight is given to a move from 0% to 20% than from 20% to 40%.
- More weight is given to a move from 20% to 40% than from 40% to 60%.
- More weight is given to a move from 40% to 60% than from 60% to 80%.
- More weight is given to a move from 60% to 80% than from 80% to 100%.

(The exception to this pattern is the 0–20 age range, where a move from 60% to 80% is weighted most highly and where the weights drop away progressively above and below that cell.)

Given the broad regularity of the patterns for both age and severity – in both cases, the general shape is a (somewhat off-centre) inverted U –one simple way of combining the two into a single set of weights for the 20 cells is as follows:

1. Identify the average pattern showing how weights vary just with age, independent of severity.
2. Likewise, identify the average pattern showing how weights vary just with severity, independent of age.
3. Compute the weight for any cell as the cross-product of the appropriate age and severity weights, normalising overall so that the most highly weighted cell is indexed at 1.

***TABLE 20*** *Horizontal (age) weights inferred from* Table 18 *(cell numbers are shown in square brackets)*

| **QoL** | | | | |
|---|---|---|---|---|
| 80–100% | [*17*] 0.837 | [*18*] 1.000 | [*19*] 0.641 | [*20*] 0.350 |
| 60–80% | [*13*] 0.781 | [*14*] 1.000 | [*15*] 0.884 | [*16*] 0.565 |
| 40–60% | [*9*] 0.676 | [*10*] 1.000 | [*11*] 0.962 | [*12*] 0.697 |
| 20–40% | [*5*] 0.716 | [*6*] 1.000 | [*7*] 0.770 | [*8*] 0.538 |
| 0–20% | [*1*] 0.867 | [*2*] 1.000 | [*3*] 0.814 | [*4*] 0.483 |
| **Age** | 0–20 | 20–40 | 40–60 | 60–80 |
| Average for each age range | 0.775 | 1.000 | 0.814 | 0.527 |

QoL, quality of life.

One way of achieving that result is as follows. It is clear from *Table 18* that at every level of severity the 20–40 age range is most highly weighted, so assign those cells an 'age weight' of 1 and use the ratios reported in *Table 18* to infer relative weights for the other cells on each row. The result is reported in *Table 20*. The average weights for each age range are shown in the bottom row.

Likewise, noting that the highest weight on the severity axis for three of the four age ranges is accorded to a move from 20% to 40%, we can assign weights of 1 to those cells and infer relative weights for the other cells in each column (*Table 21*). Then the average for each severity level can be calculated; this is shown in the last column.

Taking the cross-product of the average weights for age and severity gives *Table 22*.

*Table 22* gives the estimated weight of each cell relative to the cell with the highest value – cell 6, where a gain from 20% to 40% health accrues to 20- to 40-year-olds. The general pattern is that the

***TABLE 21*** *Vertical (severity) weights inferred from* Table 19 *(cell numbers are shown in square brackets)*

| QoL | | | | | Average for each severity |
|---|---|---|---|---|---|
| 80–100% | [*17*] 1.243 | [*18*] 0.539 | [*19*] 0.571 | [*20*] 0.397 | 0.688 |
| 60–80% | [*13*] 1.322 | [*14*] 0.738 | [*15*] 0.687 | [*16*] 0.643 | 0.848 |
| 40–60% | [*9*] 1.047 | [*10*] 0.844 | [*11*] 0.810 | [*12*] 0.802 | 0.876 |
| 20–40% | [*5*] 1.000 | [*6*] 1.000 | [*7*] 1.000 | [*8*] 1.000 | 1.000 |
| 0–20% | [*1*] 0.786 | [*2*] 0.759 | [*3*] 0.857 | [*4*] 0.801 | 0.791 |
| Age | 0–20 | 20–40 | 40–60 | 60–80 | |

QoL, quality of life.

***TABLE 22*** *Implied relative weights: ratio of means method (cell numbers are shown in square brackets)*

| QoL | | | | |
|---|---|---|---|---|
| 80–100% | [*17*] 0.533 | [*18*] 0.688 | [*19*] 0.560 | [*20*] 0.362 |
| 60–80% | [*13*] 0.658 | [*14*] 0.848 | [*15*] 0.690 | [*16*] 0.446 |
| 40–60% | [*9*] 0.679 | [*10*] 0.876 | [*11*] 0.713 | [*12*] 0.461 |
| 20–40% | [*5*] 0.775 | [*6*] 1.000 | [*7*] 0.844 | [*8*] 0.527 |
| 0–20% | [*1*] 0.613 | [*2*] 0.791 | [*3*] 0.644 | [*4*] 0.417 |
| Age | 0–20 | 20–40 | 40–60 | 60–80 |

QoL, quality of life.

***TABLE 23*** *Implied relative weights: median of individual ratios method (cell numbers are shown in square brackets)*

| QoL | | | | |
|---|---|---|---|---|
| 80–100% | [*17*] 0.47 | [*18*] 0.72 | [*19*] 0.54 | [*20*] 0.24 |
| 60–80% | [*13*] 0.55 | [*14*] 0.85 | [*15*] 0.64 | [*16*] 0.29 |
| 40–60% | [*9*] 0.60 | [*10*] 0.92 | [*11*] 0.69 | [*12*] 0.31 |
| 20–40% | [*5*] 0.65 | [*6*] 1.00 | [*7*] 0.75 | [*8*] 0.34 |
| 0–20% | [*1*] 0.49 | [*2*] 0.75 | [*3*] 0.56 | [*4*] 0.26 |
| Age | 0–20 | 20–40 | 40–60 | 60–80 |

QoL, quality of life.

valuation function peaks at cell 6, but then 'falls away' in all directions. As the top right-hand corner cell combines the lowest average weight assigned to severity with the lowest average weight assigned to age, it carries the lowest weight of any cell; on this basis, the differential between the highest and lowest valued cells is roughly 2.75:1.

As noted under 'Aggregating results', the ratio of means method is one method of aggregating the matching data. However, another set of weights can be produced on the basis of medians. The method of construction is essentially the same: the median ratios between rows and between columns allow us to fix the weight of cell 6 at 1, and set the weights for all other cells in the same row and column as cell 6, on the basis of those medians; the weights in all other cells are then computed as the appropriate cross-products. The results are given in *Table 23*.

Not surprisingly, weights based on means do not exactly coincide with the median-based measures. The averaging process appears to make the slope away from the 'peak' weight less steep in nearly all directions, so that the median-based weights are somewhat lower in most cases: while the most extreme comparison from *Table 22* is 2.75:1, the corresponding ratio in *Table 23* is approximately 4:1. However, although the specific weights may vary according to the measure of central tendency chosen and the particular method of aggregation, the general pattern of weights dropping as we move in any direction away from cell 6 is robust to these differences.

## Chaining tests

As outlined in the methods sections, a number of non-adjacent pairings were carried out in order to test for chaining with different numbers of steps with the basic idea being that $(v_1/v_5)*(v_5/v_9)$ should roughly equal $(v_1/v_9)$. In *Table 24*, the 'direct' column reports actual responses to the non-adjacent pairings, i.e. $(v_1/v_9)$ in the example used here, while the values in the 'chained' column are derived by multiplying together the weights of the component parts of the 'chain', i.e. $(v_1/v_5)*(v_5/v_9)$ in the example used here. Chaining tests were also carried out for whole columns, e.g. compare {A, B}, {B, C} and {A, C} and for whole rows, e.g. compare {R, Q}, {Q, P} and {R, P}. In each case the values relate to the aggregate matching results and make use of the ratio of means aggregation method.

For each of the non-adjacent pairings in the chaining tests we can see that the value implied by 'chaining' is more extreme than that derived directly. For example, the second row of *Table 24* shows that $(v_{11}/v_{15})*(v_{15}/v_{19}) = 1.42$ while $(v_{11}/v_{19}) = 1.18$. Likewise, row 4 shows that $(v_{18}/v_{19})*(v_{19}/v_{20}) = 2.85$ while $(v_{18}/v_{20}) = 2.12$.

***TABLE 24*** *Results of 'chaining' tests*

| | Direct | Chained |
|---|---|---|
| 9 vs 17 | 0.92 | 0.84 |
| 11 vs 19 | 1.18 | 1.42 |
| 10 vs 12 | 1.38 | 1.44 |
| 18 vs 20 | 2.12 | 2.85 |
| 6 vs 8 | 1.83 | 1.86 |
| A vs C | 0.93 | 0.71 |
| R vs P | 1.91 | 2.46 |

## Profile tests

Finally, we turn to the tests for independence of the profile in which a cell appears. Recall that *Figure 15* illustrated the alternative ways in which we might have presented the QALY gains embodied in the cells in the QALY grid. Profile tests were carried out on three different cells: 7, 11 and 14. We have argued that a 'fair innings' type argument – in which the total lifetime QALYs was relevant – would always predict that *more* weight would be given to the 'standard' profile (S) than to profile Y (as the total lifetime QALYs is less in S than in Y), and *less* weight to profile S than profile X. Finally, profile S would be given less weight than Z in the cases of cells 7 and 11, but S would be given more weight than Z in the case of cell 14.

The results are given in *Table 25* where values greater than 1 indicate that the gain was valued more highly when embedded in the profile S than in the alternative and vice versa. For example, the first column of *Table 25* gives the results of comparing the profile S with profile X (see *Figure 15*). As the values are each greater than 1, there appears, if anything, to be a tendency to favour

***TABLE 25*** *Results of the 'profile' tests*

| | S vs X | S vs Y | S vs Z |
|---|---|---|---|
| Cell 7 | 1.33 | 0.74 | 1.24 |
| Cell 11 | 1.22 | 1.36 | 1.08 |
| Cell 14 | 1.06 | 0.55 | 0.78 |

those with higher lifetime health, which is the opposite of what a 'fair innings' argument would suggest. Column 2 shows that for two of the three cells (7 and 14) the standard profile was valued less highly than profile Y, which had lower lifetime health – a finding which does fit with a 'fair innings' type argument – but that the opposite was true in the case of cell 11. The results of the S versus Z comparison also appear to contradict the 'fair innings' argument.

To sum up, it is difficult to detect any clear pattern in responses to the profile tests conducted here, but, if anything, the data appear to contradict the 'fair innings' argument, which is puzzling.

## Concluding remarks

We reported here the results of a computer-assisted 'matching', or person trade-off, valuation study carried out in a sample of 587 members of the public in the UK. The results showed that, when asked questions of the type posed here, respondents *do* differentiate between 'types' of QALYs and are willing to trade off (sometimes considerable) numbers of patients treated in order to prioritise according to age and severity of illness. The matching data show a general tendency to give more weight to younger patients and to those in poorer health, but the pattern is not a simple one. In particular, less weight is given to the youngest patients and to those in poorest health, reversing the pattern found elsewhere in the data. As these patterns may be observed in the simple choice data from the first matching iteration (when there were 100 patients in each group), they are robust to differences in methods used to aggregate the data.

The qualitative findings reported in Chapter 3 may offer some explanation for the pattern of results uncovered here. Recall that age was found to be important to focus group participants for a variety of reasons including life expectancy, which would explain the general tendency to favour the young over the old. Recall also, however, that age was found to be used as a 'proxy' for a range of other characteristics including economic activity and the existence of dependants. One possible explanation for the tendency to favour 20- to 40-year-olds over younger patients, therefore, is that such considerations outweigh the impact of life expectancy around those ages. If we were to represent the 20 discrete choice weights with the 20 cells of the QALY grid, the peak (most preferred cell) would be cell 14. This is consistent with the most preferred age group from the matching data (20–40), but implies that less-severe states are preferred (60% to 80%), although the magnitude of the weights is obviously very different.

Returning specifically to the matching data, it is also reported in Chapter 3 that respondents expressed the view that an improvement in QoL for people in poor health would be more important than an identical improvement in QoL for people in relatively good health. Again, this would seem to be consistent with the general tendency to favour those in more severe health states initially. Some concern was also expressed that prolonging life in a really bad health state was not desirable. One possible explanation for why respondents favoured a move from 20% to 40% health over one from 0% to 20% is that 20% was still considered to be not worth living. Indeed, one of the respondents quoted in Chapter 3 almost makes this point explicitly: 'whereas 20% is pretty close to death.' This raises issues about how the QoL axis – which necessarily relied on a 'visual analogue scale (VAS)-like' representation in order to put the concept across to respondents – was being interpreted, particularly towards the lower end.

It is also possible that the tendency for the weights to 'dip' towards the top right-hand corner of the grid may be linked to how respondents were interpreting the QoL scale. While the diagrammatic representation of QoL over the patient's lifetime made it clear that the QoL related to the relevant stage of the patient's life, such that 100% health at the age of 70 meant in good health *for a 70-year-old*, it may be that certain respondents were considering QoL as an *absolute* measure. If this were the case, then it might have been considered that 80% health, for example, was the best that old people could reasonably expect to achieve and little was to be achieved by trying to improve their health beyond this. If there were a certain degree of this going on when respondents were answering the matching questions, this raises the possibility that the weights estimated here are perhaps penalising the 'healthy old' rather too heavily.

While the general pattern of weights estimated for the QALY grid is robust to the different methods of aggregating the data, clearly the magnitude of the weights differs markedly. The ratio of means method of aggregating the data is the more conservative of the two methods deployed here and has been used to derive relative weights elsewhere in the economy.[60] It is important to acknowledge, however, that theoretical arguments exist in favour

of other methods of aggregating the data. It is not our intention to go into these issues further here as decisions regarding the appropriate magnitude of any weights used for policy purposes will clearly be guided by additional considerations, such as acceptability and practicality.

We conclude this section by comparing how the magnitude of the weights estimated here compares with those reported in previous studies. The review paper by Dolan and colleagues[16] cites a number of studies that have set out to estimate age-weights for life-years gained that have come up with a range of values. For example, Johannesson and Johanesson[61] report weights for life-years gained at ages 30, 50 and 70 years of 1.0, 0.22 and 0.1 respectively, indicating a 10:1 differential between the most preferred and least preferred age groups. Adjusting for QoL affected these estimates somewhat, with three QALYs gained for a 50-year-old or nine QALYs gained for a 70-year-old judged equivalent to one QALY gained for a 30-year-old. Similarly, Nord and colleagues[11] derived weights for life-years gained at ages 10, 20, 60 and 80 of 1.1, 1.0, 0.4 and 0.1 respectively, again indicating a more than 10:1 differential between the most preferred and least preferred age groups. This pattern was very similar when life-improving, rather than life- extending, interventions were considered. Finally, Busschbach and colleagues[62] compared QoL improvements at ages 5, 10, 35, 60 and 70 and estimated weights for the utility of health at these ages to be 0.2, 1.5, 1.0, 0.7 and 0.7 respectively, indicating a 7.5:1 differential between the most preferred and least preferred age groups. It is interesting to note that these weights are more extreme than even the less conservative of our two methods of aggregating the data.

In addition, a number of studies have looked at the impact of severity of health state and found it to be an important factor, although the results are mixed. Nord and colleagues[11] and Ubel and colleagues[63] asked subjects to compare improvements on a disability scale (where lower numbers indicate better functioning) with approximately equal distances between the levels. Respondents were asked to indicate how many patients moving from level 5 to level 1 was equivalent to moving a smaller number of patients from level 6 to level 4. Both studies reported a marked preference for treating the more severely ill patients. In contrast, a recent study by Dolan and Tsuchiya[64] found that respondents gave consistently higher priority to patients with better prospects without treatment, indicating that the more severely ill attract lower priority than the better off, a finding which remains unexplained and is perhaps more in line with the result on severity in Chapter 4.

# Exploring the feasibility of eliciting a monetary value of a QALY

## Basic strategy

Other Government departments/agencies, e.g. the Department for Transport (DfT), the Department for Environment, Food and Rural Affairs (DEFRA) and the HSE, have drawn on the stated preferences of representative cross-sections of the population to assign monetary values to the prevention of injury, illness and premature death. If it were possible to obtain a comparable preference-based monetary value for a QALY, this would offer the prospect of being able to apply cost–benefit analysis to an even broader range of public health and safety policies in a more coherent and consistent manner – and in a way which better reflects the values of the people who are paying for, and benefiting from, those policies.

So the objective of this part of the study was to explore the feasibility of eliciting such a value and, in the process, raise and address the theoretical and practical issues involved in trying to do so.

As a starting point, consider the way that preference-based monetary values have been estimated for use in road safety appraisal. Typically, surveys have sought to elicit the maximum amounts that respondents would individually be willing to pay for (usually small) reductions in their own (and possibly others') risk of being killed or injured in road accidents. These amounts are then aggregated across individuals to arrive at an overall value for the safety improvement concerned. The resultant figure thus aims to show what the safety improvement is 'worth' to the affected group, relative to the alternative ways in which that group might have spent its money.

For example, suppose that the members of a representative sample of the population were each asked for their WTP to reduce their risk of death on the roads during the next year by 1 in 100,000. If the mean response were £12, the implication would be that 100,000 members of the population would, between them, be willing to pay a total of £1.2M to achieve an overall reduction in risk that would, on average, reduce the number of fatalities on the roads by 1. Hence, the preference-based component of the value of preventing a (statistical) fatality (VPF) would be set at £1.2M. [While this is the largest element in the overall VPF, it is not the only one: it is usually supplemented by figures which represent the loss of net consumption entailed by premature death and some estimate of the costs (medical, police, etc.) of dealing with the consequences of a typical fatal accident.]

A similar approach could be envisaged for estimating a societal WTP-based monetary value of a QALY. The simplest case would be if people had a clear idea of what a QALY is, and could then be asked for their WTP to reduce the risk of suffering the loss of 1 QALY by 1 in 1000 (for example). If a representative sample stated a mean WTP of, say, £18 for such a risk reduction, the implication would be that 1000 members of the population would, between them, be willing to pay a total of £18,000 for some intervention that would, on average, prevent the loss of one QALY.

Other closely-related variants are possible. For example, each respondent could be asked to imagine that they face a 1 in 1000 risk of an illness which, if untreated (or treated conventionally), would entail the loss of one QALY. They could then be offered an 'insurance policy' which would ensure that they received treatment (or treatment over and above the conventional provision) that would restore them to their current/normal health and avoid that QALY loss. Again, if the average WTP for this insurance were £18, the aggregate value of the treatment that generates a one-QALY benefit would be £18,000.

That is the basic principle. In the remainder of this part of the report, we describe how we explored different ways of trying to put that principle into practice, and we discuss the issues raised both by the process and by the responses elicited.

## Overview of the questionnaire

In practice, very few members of the public know what a QALY is. One possibility might have been

to begin any interview by trying to explain the concept in a general way (as was attempted in the 'relativities' part of the project) and then ask individuals to think about how they would envisage personal QALY gains or losses and how much they would be willing to pay to achieve such gains or avoid such losses.

We do not rule this out as one possible way of proceeding; indeed, in the 'relativities' study there appeared to be a reasonable degree of receptiveness for the initial broad explanation of what QALYs are. However, such an explanation would have required quite a lot of time – especially if supplemented by some attempt to get respondents to relate the concept to their own present and future lives – and our focus was upon testing some of the assumptions underlying QALYs and their possible monetisation, rather than accepting them as uncontested. So instead of asking people to value (chances of) QALY gains/losses directly, we adopted the strategy of asking people to value avoidance of, or reductions in the risk of, illness states described quite naturalistically; and we then tried to elicit in a standard way some QoL index measure of those states. Putting the two together would – if certain assumptions hold – enable us to assign money values to QALYs; but separating the two tasks would also allow us to examine more closely whether the necessary assumptions do in fact hold.

To be more specific, the questionnaire centred around two illness states – one involving recurrent stomach/bowel problems, the other involving recurrent episodes of head pain. For each state, there were three possible durations: 3 months, 12 months and 'the rest of your life'. And for various combinations of these, there were questions about how much the respondent would be willing to pay to prevent the certainty of the illness and how much they would be willing to pay to eliminate some risk of the illness. Having obtained these monetary responses, the questionnaire went on to elicit information about trade-offs the respondent would make in terms of the relative probabilities of different health states, including some measure of the QoL index assigned to the illness state. In theory, this trade-off information makes it possible to compute for each individual the QALY losses entailed by the various illness descriptions; and this information, combined with the individual's WTP to avoid/reduce the risk of the various illness scenarios, should (if the underlying assumptions hold) enable us to take readings of that individual's monetary value of a QALY.

The questionnaire can be viewed as divided into four sections. Part A asked questions about each individual's current self-assessed state of health and then introduced them to the particular illness descriptions that constituted the focus of subsequent questions. Part B introduced the general idea of WTP for health benefits before eliciting a series of WTP values for different scenarios. These scenarios were varied between four different subsamples in order to provide broader coverage and allow certain tests of conformity (or otherwise) with standard assumptions. Respondents were allocated at random to one of the four subsamples, and the fact that Part A and the first question in Part B were common to all respondents allowed us to check for the comparability of those different subsamples. Part C consisted of four questions eliciting various risk trade-offs designed both to test the standard assumptions invoked for this part of the QALY computation and to provide a QoL/health-state index measure for the illness state under consideration. The final part of the questionnaire consisted of a series of questions collecting socioeconomic/demographic information. This allowed further checks of comparability between subsamples.

## Implementation and results

There were four versions of the questionnaire, one for each of the four subsamples. A copy of Version 2, together with the relevant supporting material, is included in Appendix 6. The ways in which the other versions differed from Version 2 will be explained in the text that follows.

But first, the sample. In order to make the most of the project budget, sampling for this part of the project involved revisiting many of those who had previously participated in the 'relativities' part of the study and then enlarging the sample opportunistically from that base. Thus, this was a 'convenience' sample; and while it covered the full range of educational level, income and social class, it was *not* intended to be representative of the population as a whole. What is important for present purposes is that within this sample, respondents were allocated at random between the four versions of the questionnaire: our objective was to examine within-subject consistency and make between-subsample comparisons. If it is judged that the survey instruments pass the tests conducted on this basis, it will require a larger and more representative sample to provide estimates suitable for public policy.

The four versions were as follows. Versions 1 and 2 focused on scenarios about stomach illness while Versions 3 and 4 were centred on scenarios about head pain. As indicated earlier, within each version there were three durations of illness: 3 months followed by return to respondent's current state of health; 12 months followed by return to respondent's current state of health; and a chronic condition where the illness lasted for the rest of the respondent's life. Each scenario was described on a separate card: the six cards are reproduced as part of the supporting materials in Appendix 7.

The key way in which Version 1 differed from Version 2, and likewise the way in which Version 3 differed from Version 4, related to the questions involving the risk of an illness: Versions 1 and 3 asked about WTP to eliminate 10% risks of the three durations of illness, while Versions 2 and 4 asked about eliminating 5% risks.

The total number of questionnaires was 409. However, after a check, six of these were excluded on the grounds that they contained very few answers to the main questions and offered little or no possibility of contributing usefully to the analysis.

The distribution of the 403 between the four versions is shown *Table 26*.

***TABLE 26*** *The distribution of versions*

| Version | Illness | % | Total |
|---|---|---|---|
| 1 | Stomach pain; chance 10% | 26 | 105 |
| 2 | Stomach pain; chance 5% | 27 | 108 |
| 3 | Head pain; chance 10% | 25 | 99 |
| 4 | Head pain; chance 5% | 23 | 91 |
| Total | | | 403 |

We now turn to the main results from the key parts of the questionnaire, starting with Part A.

## Part A: current health status and ranking of illness descriptions

At the beginning of the questionnaire, respondents were asked to complete a modified (actually, expanded) form of EuroQol 5 dimensions (EQ-5D) (five levels for each dimension rather than the usual three levels) – we shall refer to our version as EQ-5D+. Then they were asked to rate their own health on a visual analogue 'thermometer' scale, where 100 was labelled 'as good as it could be for someone of your age' and 0 was 'as bad as being dead'.

These questions were included more for purposes of 'warming up' respondents – getting them to think about what we mean by health, and giving them some tasks which involved them from the outset. They also allow us to check that there were no obvious health disparities between the respondents across the four versions – and indeed there were no significant differences between the distributions of responses to these questions across the subsamples. (Additional checks on the distributions of age and gender also showed no significant differences between the four subsamples in those respects.)

*Table 27* gives an overall summary of the responses to the EQ-5D+ questions.

To check whether there was broad correspondence between EQ-5D+ responses and the VAS, we grouped EQ-5D+ responses into four categories by adding up the rank on each dimension. So someone who reported themselves as problem free, i.e. as 11111, got a sum score of 5. Someone who reported just one 2 got a score of 6. Someone who

***TABLE 27*** *The distribution of answers for current state of health*

| | No problems | Occasional minor problems | Frequent but minor problems | Quite a lot of difficulties | Very severe difficulties |
|---|---|---|---|---|---|
| Mobility | 309 (77%) | 45 (11%) | 23 (6%) | 21 (5%) | 5 (1%) |
| Self-care | 376 (93%) | 14 (3%) | 7 (2%) | 4 (1%) | 2 (0.5%) |
| Usual activities | 317 (79%) | 44 (11%) | 20 (5%) | 18 (4%) | 4 (1%) |
| Pain and discomfort | 204 (51%) | 125 (31%) | 35 (9%) | 32 (8%) | 7 (2%) |
| Anxiety and depression | 300 (75%) | 76 (19%) | 16 (4%) | 8 (2%) | 2 (1%) |

reported themselves as 21232 was scored as 10; and so on. *Table 28* gives the main VAS statistics for each of four EQ-5D+ categories.

Comparisons across the groups using parametric and non-parametric tests [analysis of variance (ANOVA) and Kruskal–Wallis] reject the null hypothesis in favour of the alternative that there are significant differences: that is, VAS scores are correlated with EQ-5D+ responses and VAS scores are as we should expect. In addition, both appear to correlate with age – the older people are, the worse their self-reported health.

The questionnaire next asked each respondent to consider the various illness descriptions reproduced in Appendix 7 – 3 months of stomach illness, 12 months of stomach illness, the stomach illness for the rest of life, 3 months of head pain episodes, 12 months of head pain episodes, the head pain episodes for the rest of life, and sudden painless death – and to rank these seven scenarios from least bad down to worst. The main point of this exercise was to introduce people to the descriptions and to give them a task intended to get them to read them fairly carefully.

Piloting had shown that it was quite demanding simply to give these seven cards to respondents and ask them to rank them. So the actual procedure involved giving them four cards to start with – the 3-month and 12-month durations of both the stomach and the head pain symptoms – and ask them first to rank those. Once these had been ranked, each respondent was given the two 'lifetime' durations and was asked to add those to the ranking. Finally, they were asked to locate the 'sudden painless death' card by putting it above any of the other descriptions the respondent considered worse than death, or by putting it at the bottom of the ranking if the respondent considered death to be worse than all of the six illness scenarios.

There is no single obviously 'right' ranking over all seven cards, but at the very least we might expect that within 'stomach' and within 'head' we should find the lifetime duration to be ranked worst and the 3-month duration to be ranked least bad. However, 21% of respondents failed to satisfy that expectation for at least one body area. This is arguably rather disappointing.

Subsequent analysis shows that the 'aberrant' 21% did not give systematically different answers to the sets of questions that followed. Our conjecture is that although there was evidence of some confusion when asked to process these descriptions on first presentation, the later questions, which focused on subsets of no more than three at a time, did not pose the same difficulties. Still, one clear message is that if a large-scale representative survey were to be undertaken in the future, and if this were to involve a significant ranking task (especially near the beginning), extra time and care should be given to allow respondents to process the task. However, in the present feasibility study, the ranking exercise was intended mainly as a warm-up/familiarisation task, and respondents' particular rankings play no role in the subsequent analysis.

## Part B: willingness-to-pay questions

Having completed the illness ranking exercise, respondents were introduced to the idea of WTP to prevent/reduce illness with a 'practice' question, common to all versions. This question asked about WTP to reduce the duration of a sore throat while on holiday from 3 days to 1 day. The overall mean WTP was just over £40, with a median of £20. There were no differences between the distributions of responses across the four subsamples – again, some reassurance that the assignment of respondents to version was such as to allow comparability across subsamples.

***TABLE 28*** *Main statistics for VAS values between EQ-5D groups*

| | Sum of ranks = 5 (11111) | Sum of ranks = 6 | Sum of ranks = 7/8/9 | Sum of ranks = 10+ |
|---|---|---|---|---|
| *n* | 169 | 87 | 86 | 61 |
| Mean | 90.3 | 83.9 | 79.67 | 63.42 |
| Median | 94 | 85 | 80 | 70 |
| Standard deviation | 12.52 | 13.10 | 12.83 | 16.56 |

Respondents then embarked on a series of five WTP questions. These are labelled B4, B5, B6, B7 and B8 in the questionnaire, and this report will stick with those labels. B4 asked respondents what they would pay for 'a simple, safe and painless cure that would avoid' the (otherwise) certainty of the 3-month illness (stomach in Versions 1 and 2, head in Versions 3 and 4); B5 asked for WTP to avoid the certainty of the 12-month illness; B6 asked for WTP to eliminate either a 10% (Versions 1 and 3) or a 5% (Versions 2 and 4) risk of the 3-month illness; B7 asked the same for the 12-month illness; and B8 asked for WTP to eliminate the 10%/5% risk of the respondent suffering the illness for the rest of their life.

The precise wording of these questions can be seen in the copy of Version 2 in Appendix 6. The essential features were as follows. First, respondents were asked to suppose that their income would be unaffected and to focus just on how the illness would affect their health, with health being explicitly framed with reference to the EQ-5D+ dimensions they had considered in the first question in Part A.

In each of both B4 and B5, respondents were first asked whether avoiding the health effects of the illness would be worth at least something, even if it were only a few pence. If they responded that it would not even be worth paying a few pence, they were asked to say why they felt that way, and their response was noted by the interviewer. However – and especially for B4 and B5 (the questions expressed in terms of certainty) – the overwhelming majority said that they *were* willing to pay something. These respondents were then given a small pack of cards, each of which had a different sum of money printed on it. (The amounts were: £1, £5, £25, £50, £100, £250, £500, £1000, £2000, £3000, £5000, £10,000, £20,000, £50,000, £100,000 and £1M.) Once the pack had been shuffled, they were asked to sort the cards into (up to) three piles: those amounts they definitely *would* pay for an instant cure that would avoid all aspects of the illness; those amounts they definitely would *not* be prepared to pay; and any amounts about which they were unsure (initially, at least) whether or not they would pay.

After initially sorting all of the money amounts, they were asked to think about any amounts they had placed in the 'unsure' pile and were invited to relocate them (if they wished – although there was no pressure to do so) into either of the other two piles. The interviewer then recorded the highest amount in the 'definitely would pay' pile and the lowest amount in the 'definitely would not pay' pile, reminded the respondent of the highest amount in the 'definitely would pay' pile and offered them the opportunity to nominate any higher amount they felt they would be prepared to pay. If the respondent did nominate a higher amount, that amount was recorded and was used in the subsequent analysis; if the respondent did not nominate a higher amount, the largest amount in the 'definitely would pay' pile was used in the analysis.

Exactly the same procedure – and as far as possible, the same wording – was used for B6, B7 and B8: the only substantive difference was that respondents were now being asked to suppose that a test had shown there was a 5% (alternatively, 10%) chance of developing the illness in question and were asked for their WTP to eliminate that risk.

### *Avoiding/preventing the certainty of illness*

It should be noted that for questions B4 (3-month certainty) and B5 (12-month certainty), there were very few people giving zero responses. For the 12-month stomach (S) condition in B5, there were only two, while only seven others gave a WTP below £25; for head pain (H), the figures were one zero response and four others below £25. This contrasts with many WTP surveys where there is often a significant proportion of 'protest' zero responses given by people who feel that the good in question ought to be provided without specific charge and/or that money values are inappropriate. In this sample there appeared to be no such reaction to the avoidance of an otherwise certain illness: on the contrary, there was general recognition that the benefits offered in B4 and B5 were significant, and there was a widespread preparedness to indicate that with positive WTP responses.

Besides establishing whether there was or was not some general willingness to state a monetary value to avoid certain illnesses, questions B4 and B5 were included for two other reasons.

The first objective was to see whether spending some period(s) in one health state was regarded as significantly worse than spending the same period(s) in the other health state, as reflected by some significant difference between WTP to avoid a given duration in each state. (The intention was

to compare this judgement with the health-state indices elicited via later questions.)

The second objective was to examine the relationship between WTP and the duration of the illness.

For this latter purpose, the head pain cases are the most clear-cut. As can be seen from the illness description cards V and A in Appendix 7, the *only* difference between the two is the duration: V lasts for 3 months and A lasts for 12 months. But how might this difference be reflected in WTP amounts?

There are various possibilities, not all working in the same direction. First, even though the periods of time are relatively short and reasonably close to the present, it might be that the future is discounted, so that the second, third and fourth lots of 3 months that constitute the last 9 months of the 12-month period are each given progressively lower weights, with the result that experiencing the health state for 12 months is regarded as less than four times as bad as experiencing it for 3 months. Moreover, there are 'adaptation' arguments that suggest that people may get used to being ill and adapt their behaviour accordingly, so that the loss of welfare after adaptation is less than it is initially – with the result that the welfare loss over 12 months is less than four times the loss over 3 months.

In addition, and working in the same direction, it may be that paying four times as much represents more than four times the 'sacrifice' as financial budget limitations bite increasingly hard – with the result that WTP to avoid the 12-month illness is substantially less than WTP to avoid the 3-month illness.

Against that, there is the possibility that a longer period of illness becomes increasingly difficult to accommodate and/or tolerate. Two somewhat different things might be at work here. First, there may simply be a sense of being ground down by continuing ill health. Second, while it may be possible to rearrange one's life and work to some extent to cope with shorter periods of illness – putting off a holiday, for example, or postponing some tasks for a while – it may be increasingly difficult to work around a longer period of illness.

It was not obvious a priori which tendencies would be predominant, and it was not intended that this study design would be able to unscramble all of the various forces and factors. The more limited aims were these: first, to see whether the pattern of money values corresponds with the simplest QALY model, where QALY losses are supposed to be strictly proportional to the time spent in an illness state, or whether there is some other pattern; and second, whether the pattern of money values corresponds with the trade-offs made in Part C.

The same aims apply to the stomach illness scenarios. These were slightly more complex: although the recurrent bouts of stomach discomfort and sickness were of the same intensity and frequency for both the 3-month and the 12-month scenarios (and in that respect the latter might be regarded as four times the former), they began with illnesses of different severities, i.e. 3–4 days of stomach pains, diarrhoea and vomiting at the start of the 3-month illness compared with 7 days of *severe* stomach pains, diarrhoea, vomiting and fever at the beginning of the 12-month illness. However, even if the latter cannot be known to represent exactly four times as much loss of well-being as the former, we should still expect to see clear differentiation between the two, and the money values and Part C trade-offs can still be compared.

We shall come to those Part C trade-offs later. But first, we consider the issue of whether the stomach condition (S) generated significantly different WTP responses than the head pain condition (H). *Table 29* gives the summary statistics.

The presence of outliers increases standard errors and makes it more difficult to reject via *t*-tests the nulls of equal means between S and H, although these nulls *are* just rejected at the 5% level (2-tailed) for both 3 months and 12 months. The non-parametric (Mann–Whitney) tests reject the null (easily) at the 1% level in both cases. On this evidence, members of our sample seemed to regard the H state as worse than the S state.

What about sensitivity to duration? The 12-month to 3-month ratio of means is 2.305:1 for S and 2.174:1 for H – both well short of the 4:1 ratio of duration. The ratios of medians are closer, i.e. 3.33:1 for S and 4:1 for H. So there is *some* sensitivity, but it is as if there is either some discounting of the longer duration or some effect of budget constraints (or some combination of both).

If we take the *difference* between the B5 and B4 responses for each individual, we find a similar overall pattern for both S and H. For S, 11 go the wrong way (i.e. WTP strictly more to prevent

***TABLE 29*** *Summary statistics for B4 and B5*

| | | 3 months | | 12 months | |
|---|---|---|---|---|---|
| | | **Stomach** | **Head** | **Stomach** | **Head** |
| *n* | Valid | 212 | 189 | 213 | 190 |
| | Missing | 1 | 1 | 0 | 0 |
| Mean WTP (£) | | 810.27 | 1495.88 | 1867.37 | 3252.35 |
| Median WTP (£) | | 150.00 | 250.00 | 500.00 | 1000.00 |
| Standard deviation | | 1856.70 | 4658.29 | 4769.52 | 8254.29 |

3 months than 12 months), 45 give exactly the same answer to both questions (i.e. difference=0), and 156 give a higher WTP to prevent 12 months. For H, the corresponding breakdown is 10:32:147. Were we to focus exclusively on the 303 people who gave a strictly higher response to B5 than to B4, the ratio of means for S would be 2235.83:687.21 which is 3.25:1, and for H the ratio would be 3870.44:1086.78, i.e. 3.56:1. Thus, even taking the subset of responses which satisfy the (minimum) requirement that the respondent places a strictly higher value on the avoidance of the 12-month illness than on the avoidance of the 3-month illness, the means still suggest a less-than-proportional relationship between WTP and duration.

With such relatively short durations as 3 and 12 months, it might seem implausible that there is substantial discounting on time preference grounds alone – in which case, the prime suspect might be the impact of budget constraints. If this were the main cause, we might expect the effect to be reduced by moving away from certainty and asking questions about reducing risks of adverse outcomes. This brings us to questions B6, B7 and B8.

### *Eliminating the risk of illness*

*Tables 30–32* report the key statistics for the questions which asked respondents for their WTP to eliminate either a 10% risk or a 5% risk of suffering the specified illness (B6 = 3-month risk, B7 = 12-month risk and B8 = lifelong risk).

The first issue is whether responses were sensitive to the size of risk reduction. If such questions are to provide a sound basis for public policy, we should ideally like to see the values for eliminating the 10% risks being close to double the values for eliminating the corresponding 5% risks; and at the very least, we should expect the former to be significantly higher than the latter.

However, *Tables 30–32* show that there was very limited sensitivity to the size of the risk being eliminated. Out of the six 10% versus 5% between-sample comparisons – three for S and three for H – none of the *t*-tests showed a statistically significant difference between means. Indeed, and rather worryingly, the mean values for eliminating the 5% risk of H were in fact all *higher* than the corresponding means for eliminating the 10% risk. To some extent, this can be explained in terms of a couple of high outliers in the 5% H subsample (note the much larger standard deviations). However, there are other worrying signs – not least that the medians for the H subsamples, in particular, show insufficient sensitivity (and in B8, actually go the wrong way); and out of the six relevant Mann–Whitney tests, only one, for the 12-month stomach illness, showed a difference that was statistically significant at the 5% level.

We can also compare on a within-subject basis the difference between what a respondent said they were willing to pay to prevent the certainty of an illness and what they said they were willing to pay to eliminate a 10% or 5% risk of the same illness. Generally, about 80% of the sample were willing to pay strictly more to prevent the certainty than to eliminate the risk, which would seem to be a minimum requirement. However, that still leaves about 12% who gave the same WTP, and about 8% who said they were willing to pay *more* to eliminate the risk than to prevent the certainty.

What about within-subject sensitivity to duration, given the baseline risk? For each of the four versions, we can examine at the individual level the difference between the B8 (lifelong) response and the B7 (12-month) response, and also between the B7 and B6 (3-month) responses.

Running *t*-tests of whether these differences are significantly greater than zero gives a clear picture for the B8 versus B7 differences – all reject the

***TABLE 30*** *B6 – eliminating risk of the 3-month illness*

| | Stomach | | Head | |
|---|---|---|---|---|
| | **10%** | **5%** | **10%** | **5%** |
| *n* | 104 | 106 | 98 | 90 |
| Mean WTP (£) | 375.83 | 230.44 | 403.36 | 477.34 |
| Median WTP (£) | 25.00 | 25.00 | 50.00 | 50.00 |
| Standard deviation | 1133.79 | 1030.80 | 1283.01 | 2221.59 |

***TABLE 31*** *B7 – eliminating risk of the 12-month illness*

| | Stomach | | Head | |
|---|---|---|---|---|
| | **10%** | **5%** | **10%** | **5%** |
| *n* | 105 | 106 | 99 | 90 |
| Mean WTP (£) | 914.44 | 451.70 | 623.87 | 877.29 |
| Median WTP (£) | 100.00 | 50.00 | 150.00 | 100.00 |
| Standard deviation | 3364.82 | 1666.58 | 1460.66 | 2337.15 |

***TABLE 32*** *B8 – eliminating risk of the lifelong illness*

| | Stomach | | Head | | |
|---|---|---|---|---|---|
| | **10%** | **5%** | **10%** | **5%** | **5%[a]** |
| *n* | 104 | 108 | 98 | 86 | 86 |
| Mean WTP (£) | 3235.39 | 1883.42 | 3008.94 | 14,249.64 | 3203.13 |
| Median WTP (£) | 325.00 | 250.00 | 600.00 | 750.00 | 750.00 |
| Standard deviation | 9565.18 | 4145.36 | 6471.16 | 107,630.34 | 6637.52 |

a Replaces one observation of £1M with £50,000.

null at the 1% level. However, the means and medians for B8 are mostly only three to five times the values for B7, whereas average remaining life expectancy is much more than 3–5 years: the mean and median ages of the sample were 51.37 and 52 respectively, so that average remaining life expectancy would be in the region of 30 years.

The differences are predominantly in the right direction for the B7 versus B6 comparison, but here there are many more zero differences – 160 across the sample as a whole, many but not all the result of zero WTP to eliminate these risks – so the null is not rejected for Version 4 and is on the borderline for Version 1. The ratios of the means are 2.43:1, 1.96:1, 1.55:1 and 1.84:1 for the four versions – all well short of the 4:1 ratio of duration, and overall lower than the corresponding ratios for B4 and B5, despite the fact that budget considerations might have been expected to have exerted greater constraints on the B5:B4 ratios. The ratios of medians are somewhat higher, at 4:1, 2:1, 3:1 and 2:1, but mostly still fall short of the ratio of duration and are also more muted than in the B5:B4 comparisons. This may suggest that it is not simply an issue of budget constraints but that the use of probabilities introduces further 'noise' and diminished sensitivity into people's patterns of response.

A possible side effect of the reduced sensitivity to questions involving these risks was that the difference between S and H appeared to be somewhat attenuated. There are three comparisons for the 10% baseline, and another three for the 5% baseline: none of the six *t*-tests showed a statistically significant difference, and only one of

the Mann–Whitney tests came close ($p = 0.056$ in the case of the 5% risk of the 3-month illness).

Thus, despite the theoretical advantages of using risk reduction questions to elicit values, the evidence from this feasibility study seems to suggest that those questions do not, in practice, show the kind of sensitivity, either to the size of the risk reduction or to the duration of the illness, that would be desirable as a basis for a robust value of a QALY.

If probabilities seem to pose additional problems for WTP questions, what are the implications for questions that use probabilities to express trade-offs between health states and durations? The study explored this issue in Part C of the questionnaire, to which we now turn.

## Part C: standard gamble questions

There were four questions in this part of the questionnaire. The basic structure was the same for all four. In each case, respondents were asked to consider two alternatives. On the left of a showcard (see the supplementary material in Appendix 8) was displayed some prospect that would be faced with certainty – for example, in the first of these questions, that prospect was the certainty of 3 months of either S or H. On the right of the showcard was an uncertain prospect with two possible outcomes: the good outcome, which was in all cases the respondent's current health, with no ill effects; and the bad outcome, which varied from one question to another, but which was always worse than the certain prospect on the left – for example, in the first of these questions it was 12 months of either S or H.

So the uncertain alternative offered some chance of being in a better state than the certain prospect, but also entailed some chance of being worse off. What each respondent was asked to identify were the probabilities of being better or worse off that would make them feel that there was nothing to choose between the two alternatives. In all four questions, these probabilities were elicited by an iterative process. The full details can be found in the body of the questionnaire in Appendix 6, but the essential idea was as follows.

Initially, the respondent was asked to choose between the certain prospect (of 3 months in S, say) and a treatment whose outcome was uncertain but which offered a 90% chance of success (which meant continuing in current health, with no ill effects) and a 10% chance of failure (resulting in S lasting for 12 months rather than 3 months). The next stage of the question depended on the answer to that first choice. If they had preferred the certain prospect to the 90%:10% alternative, the chances were changed to make the uncertain alternative more attractive – this time offering a 99% chance of success alongside a 1% chance of failure – and the respondent was asked to choose afresh. On the other hand, if they had preferred the uncertain alternative, the chances were changed to 50%:50% to make that prospect less attractive, and the respondent was then asked to choose on this basis.

By altering the chances on the right hand side in response to each choice, it was possible to home in on a pair of probabilities which balanced the two alternatives in terms of the respondent's preferences.

As indicated above, the first question of this kind – labelled C and administered in conjunction with the relevant showcard – involved balancing the certainty of a 3-month illness against a risky treatment offering instant recovery to current health if successful but suffering for 12 months if this treatment failed.

The next question – labelled D in the questionnaire – involved the certainty of 12 months' illness versus a risky treatment offering instant recovery to current health if successful but suffering for the rest of their lifetime if this treatment failed.

Question E juxtaposed the certainty of the lifelong condition versus the prospect of current health if the uncertain treatment succeeded but (painless immediate) death if it failed.

Finally, Question F was the 'complement' of D. That is, the certain prospect involved 1 year in current health, which would then be followed for sure by the rest of their life in either S or H; while the risky prospect offered current health if it succeeded, but if it failed it entailed S or H starting immediately (i.e. this meant losing the first year in current health).

Question C was intended to explore the relationship between the subjective loss entailed by the 3-month illness compared with the loss of well-being associated with suffering the condition for four times as long. As no money is involved

in these types of question, the focus is upon how much worse the 12-month duration is compared with the 3-month period of illness. If (as a simple QALY calculation supposes) the 12-month illness involves four times as much QALY loss as the 3-month illness, a respondent behaving broadly according to the QALY model should feel that the two alternatives are evenly balanced when the risk of failure is 0.25: that is, the 0.75 chance of avoiding the 3 months of illness would be regarded as exactly balancing a 0.25 chance of suffering an extra 9 months.

As *Table 33* shows, a relatively small proportion (overall, about 15%) of each subsample gave that answer. The majority (about 60% overall) were willing to accept only a smaller risk. Neither parametric nor non-parametric tests suggest any difference between the distributions of responses for S and H. On average, the 12-month illness appears to be regarded as five or six times as bad as the 3-month duration; while the interpretation based on the medians is that most people regard the loss of well-being (or to use decision theoretic terminology, utility) from the 12-month illness as being between seven and 14 times the loss involved in a 3-month illness.

If we took these ratios as a reliable reflection of individual attitudes to duration, it would suggest that considerations of the difficulty of adapting to anything more than a fairly short illness and/or the 'intolerability' of continuing ill health outweigh any 'time discounting'. These results contrast with the relativities suggested by the responses to B4 and B5, where the ratios of means were only slightly above 2:1 and the ratios of medians were between 3:1 and 4:1. So if the SG responses do reflect preferences, the implication is that budget constraint effects are really quite strong. However, an alternative possibility is that framing these questions in terms of 'chances' and requiring respondents to think probabilistically and confront risk and uncertainty may prompt excessively cautious responses, exerting a downward influence on the chances of failure they would accept, and seeming to inflate the magnitude of the loss entailed by 12 months as compared with 3 months. As will be seen shortly, there is *some* evidence from this study which is consistent with this latter interpretation; but the evidence is decidedly mixed, as is shown by the next question.

Question D involved risking the illness for the rest of their life in order to avoid the certainty of 12 months of the illness. As the lifelong illness involves more years of health loss for those with the greatest remaining life expectancy, one might suppose that the acceptable risk of failure should be lower for younger people. *Table 34* divides the full sample into three roughly equal age groups and then reports the mean and median risks of failure for each age group for both S and H.

However, for the most part there are no statistically significant differences between age groups within S or within H, whether judged by *t*-test or by Mann–Whitney test. The only exception is the comparison between the middle and older age groups for S: but there, the difference is in the *opposite* direction to the expected one – that is, older respondents are willing to take less risk, even though they would suffer for fewer years.

The lack of – or even, perverse – sensitivity of response to age is a discouraging result. But here, by contrast with Question C, the means appear to *underweight* the magnitude of loss: given that average remaining life expectancy is in the region

***TABLE 33*** *The statistics for Standard gamble C*

| | Stomach | Head |
|---|---|---|
| *n* | 202 | 179 |
| Failure risk < 0.25 | 130 | 102 |
| Failure risk = 0.25 | 29 | 28 |
| Failure risk > 0.25 | 43 | 49 |
| Mean | 0.177 | 0.207 |
| Median | 0.075 | 0.15 |
| Standard deviation | 0.19 | 0.21 |

***TABLE 34*** *The statistics for Standard gamble D*

| | Mean risk of failure | Median risk of failure |
|---|---|---|
| ***S*** | | |
| Age ≤ 43 (*n* = 61) | 0.101 | 0.055 |
| Age > 43 but < 59 (*n* = 66) | 0.137 | 0.060 |
| Age ≥ 59 (*n* = 73) | 0.084 | 0.025 |
| All S | 0.108 | 0.055 |
| ***H*** | | |
| Age ≤ 43 (*n* = 62) | 0.117 | 0.045 |
| Age > 43 but < 59 (*n* = 56) | 0.123 | 0.030 |
| Age ≥ 59 (*n* = 60) | 0.150 | 0.050 |
| All H | 0.129 | 0.035 |

of 30 years, overall means of 0.108 (for S) and 0.129 (for H) imply that suffering the conditions for the rest of life would only be between 8 and 10 times as bad as suffering for 1 year – a result that seems implausible unless we assume respondents to be discounting the future very heavily and/or to be anticipating considerable adaptation. Another possibility is that the arithmetic mean is not a particularly good measure of central tendency in these cases: the medians of 0.055 and 0.035 – implying that the lifetime illness is between 18 and 28 times as bad as the 12-month illness – may constitute a better reflection of the 'typical' respondent.

Before considering Question E, it may be useful to look at the results from Question F. As mentioned above, Question F is the 'complement' of Question D. For both questions, the possible outcomes of the risky treatment are either to continue in current health or else to suffer lifelong illness from now on. In Question D, the certainty is 12-months' illness followed by a return to current health for the rest of life, so that the possible gain from the risky treatment is to avoid the illness for the coming year, while the possible loss is to drop from current health for the rest of life *after* this year to illness for that remaining lifetime.

In Question F, the certainty is that the respondent will remain in current health this year, but the onset of illness cannot be delayed beyond that and so the respondent will then be ill for the rest of their life. Thus, the potential gain from the risky treatment would be to avoid illness for all life after next year and instead spend those years in the health currently expected (i.e. the mirror image of the potential loss in Question D), while the possible loss in Question F is that instead of spending the next year in current health, the effects of the illness start now (i.e. the mirror image of the potential gain in Question D). On standard assumptions, whatever chance of failure makes the respondent feel that the alternatives are finely balanced in Question D should be the chance of success that would make things finely balanced in Question F. Put another way, for each respondent the sum of the two risks of failure should come to 1. *Table 35* reports the vital statistics for Question F and for the D + F sums.

The null hypothesis that D + F = 1 is rejected at the 0.1% level for both S and H. To give some rough idea of the distributions, consider *Table 36*, where the observations are grouped in four ranges.

It is not clear exactly what this result signifies. If a similar exercise has been conducted in previous work, we are not aware of it; and so interpretation must be cautious and speculative. However, one possible interpretation is that many people are averse to the possible bad outcome of a risky prospect but do not know exactly how averse they are, and therefore in one or both questions they respond with extra caution, thus producing at least one and possibly two understatements of the downside risk they would accept.

While this might fit with the tentative suggestion that respondents exhibited excessive caution in Question C, it does not fit so neatly with the responses in Question D, which might be seen as reflecting a propensity by a significant proportion of the sample to take too much risk when the downside involves suffering the illness for the rest of their life.

On the other hand, it does seem likely that the risks people were willing to take in Question F were very conservative. An 'average' respondent with 30 years' remaining life expectancy is facing, on the left-hand side, the certainty of spending the last 29 of those years in either S or H; but the uncertain treatment on the right-hand side offers some chance of avoiding that and instead spending all of the rest of their life in the health they currently expect – the downside risk being that if the uncertain treatment fails they lose

***TABLE 35*** *The statistics for Standard gamble F and for the sums of D and F*

| | **Stomach** | | **Head** | |
|---|---|---|---|---|
| | **F** | **D + F** | **F** | **D + F** |
| *n* | 202 | 195 | 180 | 172 |
| Mean | 0.474 | 0.583 | 0.520 | 0.653 |
| Median | 0.550 | 0.575 | 0.550 | 0.676 |
| Standard deviation | 0.33 | 0.39 | 0.33 | 0.41 |

***TABLE 36*** *Indicative distributions for D + F*

| | Stomach | Head |
|---|---|---|
| $D + F \leq 0.5$ | 87 | 63 |
| $0.5 < D + F < 0.9$ | 60 | 51 |
| $0.9 \leq D + F \leq 1.10$ | 35 | 41 |
| $D + F > 1.10$ | 13 | 17 |

the coming year of current health. The median response – wanting a 45% chance of avoiding the chronic illness in order to be prepared to take a 55% chance of suffering the illness immediately – appears at first pass either to place a very high value on the next year relative to the rest of life (not reflected in their Question D responses) or to reflect misunderstanding/confusion. In any event, it may prompt us to treat with caution other data emerging from SG questions.

This brings us to Question E, which is a vital component in any attempt to infer a monetary value of a QALY from the present survey.

The idea of Question E was to get each individual's assessment of the health state (S or H) in the form of a 'health-state index' number. The underlying assumption here is that current health is indexed at 1 and death at 0 and that an individual's index number for either S or H is independent of the number of years spent in that state relative to spending those years in current health or else being dead for that time. So the loss of health (relative to 1) is assumed to be revealed by the risk of death that would make the individual indifferent between the certainty of S/H for the rest of their life and the uncertain cure-or-kill treatment. The results for this are portrayed in *Table 37*.

Although the mean for H was higher than for S, the medians were the same, and neither a *t*-test nor a Mann–Whitney test registered any significant difference. The picture here, then, is more similar to the one emerging from the risk reduction WTP questions B6–B8 than from the B4 and B5 questions asking about WTP to avoid the certainty of the illnesses, which suggested that H was regarded as significantly worse than S.

***TABLE 37*** *The statistics for Standard gamble E*

| | Stomach | Head |
|---|---|---|
| *n* | 196 | 177 |
| Mean | 0.104 | 0.144 |
| Median | 0.025 | 0.025 |
| Standard deviation | 0.17 | 0.25 |

## Using the willingness-to-pay and standard gamble data to derive a monetary value of a QALY

In order to get an estimate of each individual's monetary value of a QALY, our procedure involves converting an illness description into a QALY loss by using the individual's response to Question E, and then combining that with their money value for avoiding that loss, taken from the WTP questions.

There are various ways we could combine individuals' responses to different questions, but perhaps the simplest to explain is to take the 12-month illness, compute the QALY loss involved in that, and combine that with WTP to avoid that loss as stated in response to B5. It will be seen that the issues raised by this route are so fundamental that it is really rather superfluous to pursue the other variants, none of which circumvent those fundamental problems. So we focus on this route, which works as follows.

Consider someone who is just willing to accept a 5% risk of death in Question E. This is taken to signify that the illness state is indexed at 0.95 on the scale that runs from 1 (normal health) to 0 (dead, or as bad as being dead). In other words, each year spent in that state is rated at 0.95 of a normally healthy year and thus represents a loss of 0.05 (i.e. one-twentieth) of a QALY. If that same individual states in Question B5 that they are willing to pay £800 to avoid that loss, then 20 people like this respondent would, collectively, be prepared to pay £16,000 to avoid health losses that add up to one QALY (i.e. $20 \times 0.05$). Thus, this individual's responses amount to saying that for a population consisting of similar people, the money value of a QALY is judged to be £16,000. By calculating the 'value of a QALY' implied by each individual's responses combined in this way, we can derive mean and median values for this sample.

The potential problem, of course, is that if someone says they are only willing to take a 1 in 100,000 risk (or less) of death in Question E, their WTP to avoid the 12-month illness is multiplied

by 100,000, potentially giving an astronomical figure for the value of a QALY. (Someone who says they are only willing to take a 1 in 100,000 risk of death to avoid the chronic illness state is taken to be indexing that health state at 0.99999, i.e. a year spent in that state is taken to amount to the loss of 0.00001 of a QALY. So if 100,000 such people were each willing to pay, say, £300 to avoid the 12-month illness, they would between them be paying £30M and their combined benefit would add up to just one QALY.) Sure enough, we find that 115 respondents in total – 59 in the S subsample and 56 in the H subsample – give responses to Questions E and to B5 which, in combination, imply values of a QALY of more than £1M. With some of these combinations generating values of thousands of millions of pounds, the mean values for a QALY are £3 × 108 from the S subsample and £7 × 108 from the H subsample. The medians are somewhat more terrestrial: £26,666.67 and £57,142.86 respectively. On the grounds that there are no statistically significant differences between the two subsample distributions, we might pool them to obtain an overall mean of £5 × 108 and an overall median of £40,000.

It has been widely accepted that when such values are being used to guide public policy, it is the mean figure which should be used as the best indicator of social welfare. However, clearly, a mean value for a QALY of £5 × 108, or even a figure one-thousandth as big as that (i.e. £500,000), would be totally anomalous in a world where the VPF is about £1.5M and where this, in the context of road accident fatalities, represents preventing a death which on average entails the loss of about 40 years of life expectancy.

Nevertheless, applying the procedure in the way described does generate the kinds of extremely skewed distributions reported, and that does produce means which are pulled up by some extremely high individual figures.

This raises the question of whether there is something that could be done – some other way of managing the procedure or analysing the data, perhaps – that might give figures more compatible with those in use in other areas of public health and safety policy?

As a start to answering this question, recall what is involved in a study to elicit the value of preventing a (statistical) fatality or the value of preventing a particular (statistical) injury. Typically, respondents are presented with a description of a specified physical outcome – death, or a particular injury description – and are asked for their WTP to reduce their risk of experiencing this outcome by some given amount. We then effectively sum the responses over a representative population big enough that their individual risk reductions add up to preventing the (statistical) expectation of one death or one injury of the particular type.

In the course of this procedure, no attempt is made at the individual level to ascertain the number of QALYs each individual perceives they would lose in the event of being killed or injured. In principle, this *could* be done. Were things to be done that way, it might very well be that the combination of some individuals giving high WTP with responses by them implying very small individual QALY scores would produce a similar highly-skewed distribution and upward pull on means. But as things stand, it has not been the usual practice. So, in the light of the way in which things actually have been done, what might be the implication for the present study?

Take the case of a non-fatal road injury, which is a closer analogue than a fatality to the illness descriptions used in the present study. Call this road injury 'J'. Typically, respondents have been given descriptions of what J involves, have been asked to think how such an injury would impact upon their life, and have then either been asked for their WTP to prevent/reduce the risk of such an injury or expressed some relativity between the injury and death. In this latter case, the value of preventing each J is then 'pegged' against the VPF according to the average relativity expressed between J and death. But however it is arrived at, for the sake of example, let us suppose that the process generates an average money value for preventing injury J of £60,000.

If one were to want to go further and infer a value of a QALY from the established value of preventing injury J, the most obvious way of doing so would be to undertake a survey eliciting from a cross-section of the population their judgement of the QALY loss they would suffer if they sustained injury J. Again, for the sake of example, suppose that responses to such a survey ranged quite widely from one individual to another but that the average judged QALY loss turned out to be 2.5 QALYs.

On this basis, a public body undertaking a road safety scheme costing £600,000 which is expected

to prevent 10 cases of injury J could (just) justify that expenditure on cost–benefit grounds (assuming all other projects yielding an excess of benefit over cost were also undertaken). And although this body could not say in advance which particular individuals would benefit from the scheme by avoiding injuries they would otherwise have suffered, it could assert that on average the benefit would amount to preventing a loss of 25 QALYs, translating to an average value of £24,000 per QALY.

We can mimic that kind of calculation on the basis of the data collected in the feasibility study if it is processed in the following way.

First, we have reported in *Table 29* the WTP-based mean values for avoiding the 12-month durations of S and H: respectively, £1867.37 and £3252.35. Let us round these figures to £1870 and £3250. Had our sample been a large, representative cross-section of the population, it could have been argued that health-care resources allocated to preventing/curing such illnesses could (just) be justified if they cost those amounts for every 12-month episode prevented/cured.

Then we might ask what QALY gains would result from such expenditure. We cannot know which particular individuals will benefit, but on the basis of our sample members' responses to Question E as summarised in *Table 37*, together with the standard QALY procedure for combining health-state indices with periods of time spent in a health state, we could say that, on average, each avoided case of 12-month S would avoid a loss of 0.104 of a QALY, while each avoided case of 12-month H would generate an average benefit of 0.144 of a QALY.

Putting the relevant figures together, we would have an expenditure of £1870 on 12-month S, bringing an average QALY benefit of 0.104, which translates to a money value of about £17,980 per QALY, while each £3250 spent preventing 12-month H would 'save' 0.144 QALYs, implying a value per QALY of approximately £22,570.

Processed this way, the data obtained via two different health-state descriptions produce figures reasonably close to each other on either side of £20,000 per QALY. But why are these figures so very different from the astronomical means based on the value-per-QALY figures generated by combining Question E and B5 responses at the level of each individual? And which – if either – is the correct basis for deriving an estimate?

The answer to the first of these questions is fairly straightforward and can be illustrated by an example (which may also be helpful when trying to answer the second question).

Consider a subsample of 10 people. In order to keep things simple, suppose each of them states the same WTP value to avoid 12 months of H – let us say £250 (which works out at about £10 for each fortnightly episode avoided). Suppose also that in response to Question E, 8 of these 10 would take a 1% risk of death in order to get a 99% chance of avoiding H for the rest of their lives, while one person would take a 2% risk and the remaining member of the subsample would only take a 0.1% chance. Averaging those responses gives an average QALY loss of 0.0101 per 12-month illness which, combined with the average WTP of £250 to avoid it, generates a value per QALY of just over £24,750.

Contrast this with the figure given by first combining each individual's WTP with their Question E response and *then* averaging. For eight respondents, the implied value of a QALY is £250 ÷ 0.01, which gives £25,000. For the respondent prepared to take a 2% risk, the figure is £250 ÷ 0.02, which gives £12,500. And for the respondent who will only accept a risk of 1 in 1000, the calculation is £250 ÷ 0.001, which produces £250,000. On this basis, the subsample mean is £46,250, i.e. almost double the figure arrived at by the other method.

The reason for the difference is this. Under the first method of computation, the two responses which differed from 0.01 were *added* to each other and to the eight 0.01s; and as they diverged from 0.01 in different directions but by almost the same amount, (+0.01 in one case, –0.009 in the other), this adding and averaging more or less cancelled them out, so that the subsample average diverged from 0.01 by just 0.0001. The net effect of this was to nudge the computed value of a QALY down to a little below £25,000.

However, under the second computation method, the (reciprocals of the) probabilities operate *multiplicatively* on the WTP responses before any averaging occurs. Thus the difference between 0.01 and 0.02 acts to halve the implied value of a QALY for that individual from £25,000 to £12,500 – a money difference of –£12,500. At the same time, the slightly smaller difference in the opposite direction between 0.01 and 0.001 acts to multiply by 10 the implied value – producing £250,000 as compared with £25,000 – a money difference

of +£225,000. When added and averaged at *this* stage, the two 'outliers' nowhere near cancel each other out: on the contrary, one individual's implied value comes to more than all of the other nine put together, and this has the effect of almost doubling the mean.

So that is why the two methods of computation produce very different results; and the divergences in the example, striking though they may be, are dwarfed by the divergences in the actual sample, where still smaller probabilities had the effect of multiplying WTP responses by tens and hundreds of thousands, and even by millions.

So which is the *correct* method to use? In theory, if our analysis is based on the premise of each member of the population valuing the same QALY gain, it is the second method, i.e. the method used initially in the analysis of the survey data which generates stratospheric means. But, as with any theory, the validity and usefulness of the results depend crucially on the extent to which the underlying assumptions are valid. If we subscribe to the conventional precepts of welfare economics and if we could be confident that individuals have values and preferences which conform with standard assumptions and that their responses to our questions reveal those values with total accuracy and precision, the second method would be the appropriate one to use. In the light of the example which shows how one person's value can outweigh the values of nine others, this may seem a surprising conclusion; but if all 10 really were reporting their true values accurately, and if the guiding principles entail giving each person and their values equal weight, then taking the mean of that distribution, however skewed, is the appropriate thing to do.

But what if certain assumptions do not hold perfectly? In particular, what if people are not always able to report their values with total accuracy, but are liable to give responses which contain elements of 'noise', error and/or bias? We are still some way from having very good models of the noise, error, bias and imprecision in human judgement; but we know enough to appreciate (a) that such things exist and (b) that they may not always be neutral or 'white' in their effects. So although we cannot say categorically how to model these factors, let us consider how they *might* impact upon the results of the present feasibility study.

Consider again the example set out a few paragraphs earlier. Suppose that all 10 respondents not only have the same 'true' WTP to avoid the illness, but also would have their QoL diminished by the illness to exactly the same degree – that is, by 1%. If they each reported their values and judgements with complete accuracy, we should infer a mean value of a QALY of £25,000.

However, suppose that they do not all process probability judgements with unerring accuracy. To keep things simple, suppose that just two of them give erroneous responses – one reporting a willingness to accept a 2% risk, the other setting the 'break-even' risk of failure at 0.1%. Of course, this is only an illustration: we do not know enough about judgemental error to say exactly which and how many deviations from the underlying 'true' values are likely to occur. The point is, however, that if those two responses involve (seemingly small) errors, they can throw the estimates off in different ways and to very different degrees, depending on how they are processed. In the case of the first (and theoretically vulnerable) method of taking the arithmetic mean WTP separately from the arithmetic mean of the failure risks and then deriving a value of a QALY on the basis of the combination of these two means, the errors act to produce an estimate which is roughly £250 (and thus about 1%) below the true value; whereas in the case of the second (and theoretically preferable) method, the errors act multiplicatively before any averaging process occurs, and thereby produce an estimate which is £21,250 (and thus 85%) above the true value.

No general claim is being made here about the relative performance of the two methods: this is just a simple and convenient, albeit stylised, example. But what we do know is that people generally find probabilities quite challenging to manipulate, and their responses are liable to deviate a good deal, and in ways we do not well understand, from what is assumed by standard decision theories. (Recall the evidence relating to the sums of responses to Questions D and F, as reported in *Tables 35* and *36*.) So a method of calculation which is especially liable to magnify 'errors' by including them multiplicatively – and which we know in this case is liable to produce phenomenally large and implausible values – seems hard to justify as a basis for public policy.

Unfortunately, there is no way of knowing with any great confidence whether any serious distortions occur as the result of using the other method outlined here. It may be, as in the example, that this method has a tendency to understate the

'true' value. But equally, it is possible to produce examples where that is not the case; and even if it were the case in this study, we know too little about the nature of the imprecision/error in people's responses to say by how much any estimates diverge from the 'truth'. Perhaps the most that can be said is that this method, flawed though it may be, seems broadly compatible with the ways in which other values used by government departments have been derived; and to the extent that there is a desire to have a monetary value for a QALY established on a similar footing to the values used in DfT, DEFRA, etc., there appears to be some argument for this method of estimation, while keeping all the caveats in mind.

## Concluding remarks

Speaking of caveats, it is important to reiterate that even if the data were unproblematic and the calculation of the monetary value of a QALY were straightforward, no value generated by *this* feasibility study could be regarded as a sound basis for policy because the sample used here was *not* representative of the population. The question is, then, whether the present study gives grounds for believing that (at least some of) the techniques explored here could be used in conjunction with a large and genuinely representative sample to produce a value robust enough to be used to guide policy, or whether the evidence suggests that no sufficiently robust value is likely (ever) to emerge.

There were a number of encouraging features. Interviewers generally found that respondents were engaged and interested, and very few interviews were aborted or resulted in large numbers of missing values. And even though many people in the UK are still somewhat resistant to the idea of paying for health care over and above tax contributions, the very small number of zero responses to B4 and especially B5 suggests that the 'protest' element was small and that people were willing to 'play the game'. Moreover, although the responses to B4 and B5 did not display the ideal 4:1 ratio that full sensitivity to duration might have elicited, there was a degree of sensitivity that *could* arguably have been compatible with the effect of budget constraints.

However, there are also a number of discouraging features. Although by no means new or unexpected, there was serious between-sample insensitivity to the size of the risk being eliminated in questions B6–B8. And although responses to these questions should have been less affected by budget constraints and should therefore have shown greater sensitivity to duration, they did, if anything, exhibit rather less sensitivity to the 3-month:12-month difference; and also seemed to greatly underweight the 'remaining lifetime' duration relative to the 12-month scenario, with mean responses to B8 never more than 5 times higher than the corresponding means for B7, even though average remaining life expectancy was in the region of 30 years. It appeared that the use of probabilistic questions added complexity and dulled sensitivity.

Thus it may come as no surprise that some of the Part C questions, which used probabilities as their main 'currency', were also problematic. Although responses to Question D showed greater sensitivity to duration than was exhibited in the WTP questions, the mean response still appeared to greatly underweight remaining life expectancy relative to the next year. The summation of responses to Questions D and F showed a very substantial and seemingly systematic departure from the behaviour that would be consistent with standard assumptions.

It is hard to say whether the responses to Question E are 'reasonable' or not. Certainly, the distributions are heavily skewed: for both S and H the measure of skewness is greater than 2. Between a quarter and a third of respondents gave a 'failure risk' greater than the mean, while more than a third were unwilling to accept a risk of failure of 1 in 1000. Focusing on this latter statistic, does it really seem plausible that a condition involving significant disruption of at least some activities for between 8 hours and 3 days every couple of weeks entails a loss of welfare of less than 0.1%? Or is it that eliciting a response by means of a question involving the risk of immediate death induces ultraconservative responses from a significant minority of respondents? This is an open question. But if such responses are combined at the individual level with average WTP responses, they imply huge values of a QALY of the kind that generate the implausibly high means reported in this study.

On the other hand, if we compute means for B5 and for Question E separately and then use the ratio of these means to estimate the value of a QALY, the two figures obtained via different illness state descriptions are reasonably similar and not obviously outlandish.

Overall, then, this feasibility study sounds many notes of caution and points to a number of issues – particularly concerning the way respondents process probabilities and the extent to which their answers are sensitive to key dimensions – that would require further investigation before investing in a large representative national survey. Such questions are potentially amenable to investigation using qualitative methods and entailing considerable cognitive testing during piloting, and it would be unwise to embark on a large-scale study to generate policy values without first undertaking extensive (and probably expensive) preparatory research of this kind.

# Conclusions

The main contributions of the research described in this report have been the development and initial application of two novel approaches to eliciting weights for QALYs along with one for estimating a monetary value of a QALY.

Detailed caveats have been listed within each of the preceding three chapters and, so, only broad research and policy conclusions will be offered here.

## Weighting QALYs

The main results of the discrete choice and matching approaches show very different sets of weights for age and severity. In broad terms, the discrete choice results might be taken as suggesting that it is not worth weighting QALYs at all. By contrast, significant weights can be inferred from matching data, with gains for some groups being weighted up to 2.75 times higher than for others when using the more conservative aggregation method, the ratios going up to 4:1 using the less conservative method. Despite such differences in magnitude, discrete choice and matching weights did show similar patterns across age (but not severity) ranges.

There are two perspectives to take on the differences in results. First, it could be said that it would be premature to propose any particular set of QALY weights at this point in time: before that point is reached, there is scope for both further reconciliation and replication. Second, it might be argued that there is no scope for reconciliation and we need to choose between the results in light of the caveats of each.

### Reconciliation and replication

With respect to the reconciliation of our findings it could, of course, be argued that it is not surprising that the discrete choice and matching methods led to different results. One factor has already been pointed out. The matching study involved holding health gains constant between the two options in any pairwise choice, and varying the age and severity attributes, while the discrete choice questions presented a pair of scenarios with different health gains as well as levels of age and severity. It might be, therefore, that the matching procedure highlights age and severity while the discrete choice method dampens down their impact when the size of health gain is also varied. Indeed, if respondents have lexicographic preferences, whereby health gain matters above all else, it could be argued that the results of the discrete choice and matching approaches are entirely consistent. If this were the case, then the policy implications would require very careful thinking through and would pose major challenges to matching-based approaches in which choices presented hold QALY gains constant. Nevertheless, continuing to focus on the differences, there are additional limitations in the design of the discrete choice study which indicate caution around interpretation of results.

Other potentially important differences between the methods are listed in more detail in Appendix 9. Despite these, given that each respondent was asked a set of discrete choice questions and a set of matching questions, further detailed analysis may shed light on reasons for the differences.

Decision heuristics provides another possible avenue to reconciliation. First we have mentioned, in Chapter 5, that the lowest weight attached to the 60- to 80-year-olds in full health may have resulted from respondents thinking that 80% health is good enough for someone in such an age group. The lack of graphical representation of numbers of people in the matching study may have detracted respondents from the 'brutal' nature of trade-offs between persons, which, if dealt with otherwise, may have led to smaller trade-offs. With respect to the discrete choice approach, the simplest decision heuristic would be to compare the size of the areas representing the health gain, i.e. to maximize health and ignore age and severity. If each of these arguments had some validity, the 'true' result would be somewhere between the two.

It may be significant that the discrete choice and matching studies use different functional forms for age and severity weights.

In the discrete choice study, utility is modelled as a function of 'age at onset' (*AO*), 'age at death if untreated' (*AD*), 'quality of life lost without treatment' (*QL*), and 'QALYs gained from treatment' (*QALY*). Thus, the discrete choice functional form measures the health gain from an intervention by the size of the dark shaded area in *Figures 2* and *3*, i.e. it measures the health gain as the total number of QALYs. The functional form then treats this measure of 'total QALY gain' as one variable. The other three variables describe the lifetime health profile without treatment (the light shaded area). So, the weights produced by the discrete choice study are based on the properties of the light shaded area and the total area of the dark shaded area.

In contrast, the matching study focuses on the properties of the 'dark shaded area' of health gains. Recall that this study uses a 'QALY grid' of 20 cells, defined and numbered as in *Figure 16*. Each cell is assumed to have a subjective weight, which is applied to any health gains that occur in that cell. The subjective value of an intervention is modelled as the sum of the weighted health gains that are generated.

For example, consider the following case. In the absence of treatment, the lifetime health profile is given by $AO=10$, $AD=20$ and $QL=0.7$ (i.e. the patient is in full health to age 10, drops to 30% health from ages 10 to 20, and then dies). With treatment, the patient is maintained at 50% health from age 10 to age 30, and then dies. The health profile without treatment is represented by the light shaded area in *Figure 16*; the health gain from treatment is represented by the dark shaded area. The discrete choice study treats the effect of treatment as an undifferentiated gain of seven QALYs (i.e. an increase of 20 percentage points for 10 years, plus an increase of 50 percentage points for 10 years). The weight given to these QALYs is determined by the values of *AO*, *AD* and *QL*. In contrast, the matching study treats the effect of treatment as the creation of two QALYs in cell 2, one QALY in cell 5, two QALYs in cell 6 and one QALY in each of cells 9 and 10; each of these QALYs is given the weight of the cell in which it is located. Denoting the weight of each cell i by $w_i$, and the QALY gain in each cell i by $q_i$, a simple additive functional form can be specified as follows:

$$U = w_1 q_1 + \ldots + w_{20} q_{20}. \qquad (6)$$

It is possible that the difference between the functional forms used in the two studies is responsible for the differences in their results. We tried to test this explanation by estimating equation (6) using the discrete choice data. One might expect that if the differences between the

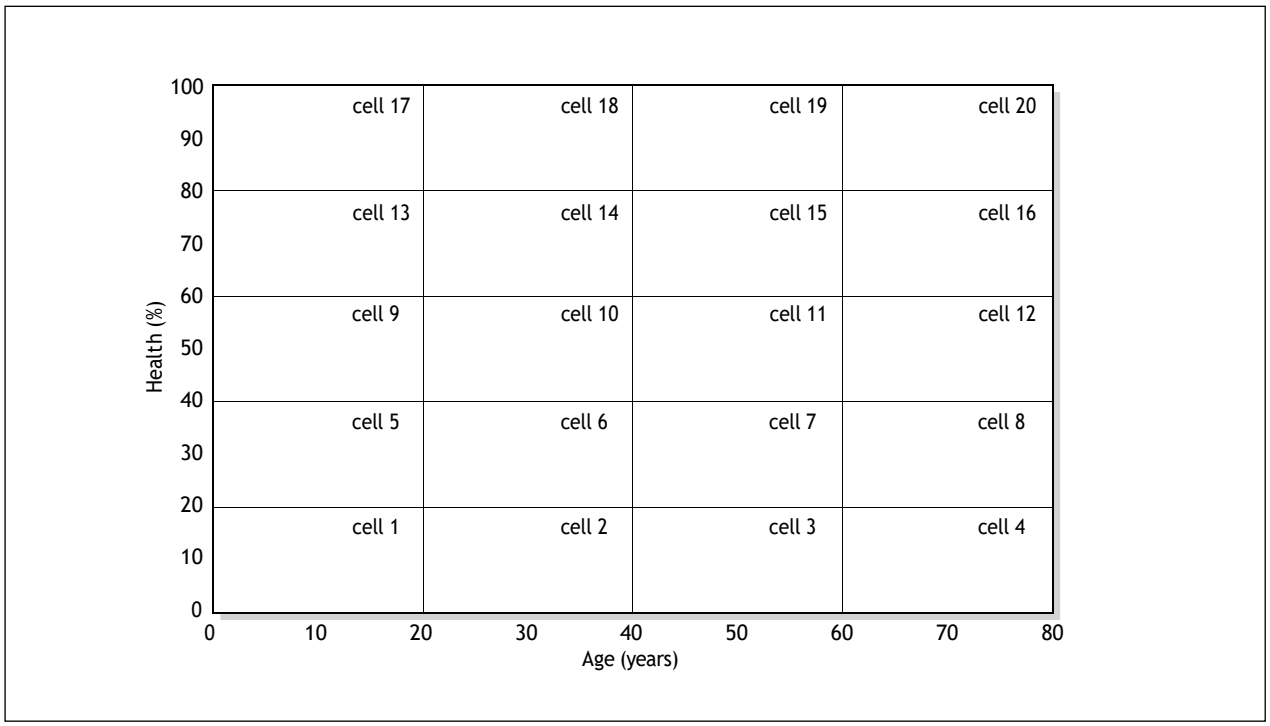


***FIGURE 16*** *A health gain in the QALY grid.*

results of the two studies were primarily due to the different functional forms used, this estimation exercise would generate weights similar to those found in the matching study. Alternatively, if it generated weights with some other systematic pattern, that might point towards some other explanation of the differences. (For example, if the weights were approximately equal in all cells, that would give further support to the conclusion that, when answering to discrete choice questions, respondents do not weight QALYs according to age or severity.) In fact, the estimated weights were highly variable, with no apparent pattern. The Bayesian information criterion clearly favoured the original model (described in Chapter 4, Predicted probability of choice approach) rather than equation (6). The lack of structure in these estimates is probably the result of econometric problems caused by multicollinearity in the data. The basic problem is that, in a typical discrete choice question, the health gains created by each intervention occur in cells that are adjacent to one another, creating strong positive correlation between the $q_i$ values for adjacent cells. This reflects the fact that the questions used in the discrete choice study had been designed to estimate a very different model.

Procedurally, too, the exercises are very different (the results of the matching questions are derived from numbers being treated, which is always constant in the discrete choice) and discrete choice is perhaps a more complicated task (there are more things varying at any one moment, such as the health gain). Other such differences are outlined below. As pointed out in Chapter 1, these further differences represent concerns that each group within the Team, largely Newcastle-led and UEA-led, have with the piece of relativities work on which they did *not* lead. The Newcastle-led group think that both exercises have their merits, whereas the UEA-led group stands by the methods and results from the matching study.

Replication is also important. Aspects of each of these approaches are new and their applications to deriving relative weights for QALYs have generated significant challenges. Further survey work is required to inform the debate on weighting of QALYs with more confidence. With respect to discrete choice methods, as we have said, we would recommend pursuit of alternative experimental design strategies in order to address the issue of implausible scenarios while still maintaining desirable design properties, so ensuring we can estimate the effects of interest with improved efficiency.

## Fundamental differences between the approaches

It is possible, of course, to choose between the approaches in light of the caveats of each. Here, we outline those caveats without making such a choice. The discrete choice approach is new with respect to its application to deriving QALY relativities. This novelty could be viewed as advantageous. However, the approach raises some serious questions that would threaten the validity of the results. The main ones are:

- The issues raised by a compromised design by standards usually applied to discrete choice experimentation and whether the more pragmatic econometric approach overcomes these to any degree.
- The theoretical meaning that can be attached to the multiplicative functional form that underlies the analysis of the discrete choice data. More specifically, the meaning of multiplying QALYs by two age variables as well as by severity, and the challenges of multicollinearity such a model engenders. In defence of this model, however, it is not clear what form such a function should take and the use of squared and cubic terms might reasonably have been expected to pick up any non-linearities implied by the functional form arising from the matching data. Also, as pointed out in Chapter 4, although, it may appear that a multiplicative model of the form QALYs×AGE×SEVERITY, with just one age-related variable, would make more theoretical sense, we took the more pragmatic view that this would leave too much riding on the 'age at onset' variable in terms of explaining what respondents might be thinking about in relation to age, and so we included age at death as well. In addition, this was the best performing model empirically, which, it could be argued, is important for estimating weights.
- Whether the transformations of original variables to a QALY composite and QoL lost represent too much of a distortion from the variables as presented to respondents in the survey.
- The challenges of coping with severity as described in Chapter 4, Further investigation of severity.

The matching approach builds on a method used previously in the literature. The results are not out of line with such earlier studies, the pattern of weights being consistent across aggregation methods. Nevertheless, in addition to QALY gains being held constant in the matching questions, there are some issues with the matching that remain to be resolved. These are that:

- There is an underlying assumption that numbers of beneficiaries presented in matching scenarios can be multiplied by the number of QALYs (implicitly) presented, and it is not clear if respondents were thinking in such a manner.
- While the 'whole column' and 'whole row' tests looked at the issue of larger QALY gains, it is unclear whether we can generalise from weights obtained by offering four QALY gains to gains of fewer than four QALYs and it is acknowledged that many interventions yield a much smaller benefit than four QALYs.
- Although a general pattern has been detected, the appropriate magnitude of the weights remains to be established. This is not a problem of the matching methodology per se, but rather an inevitable part of any elicitation exercise which is used to guide policy.

In summary, the extent to which either approach yields results that are entirely consistent with social preferences is uncertain. This work has indicated some important patterns in the data. It could be argued that the old and the healthy should receive lower weights. The challenges are with the actual numbers to adopt and so further validation and testing would be required before implementation in policy.

## The feasibility of estimating a monetary value of a QALY

It was never the intention in the valuation study to produce a monetary value of a QALY for use in policy deliberations. The purpose was simply to assess the feasibility of estimating such a value.

There is scope to use the current data set to further explore different ways of combining WTP and SG data to arrive at a value of a QALY, and to examine the sensitivity of results to different assumptions about noise in the data – and especially in the probability responses generated by the SG questions. Further consideration needs to be given to aggregation issues: for example, the weight to attach to means versus medians, and whether to use means of individual WTP/SG combinations or combinations of mean WTP and mean SG values.

The feasibility study has demonstrated that there are considerable challenges involved in trying to elicit a robust monetary value of a QALY. However, the low number of protest responses demonstrates that respondents are comfortable with these types of question. Any future national sample survey should be preceded by further extensive qualitative research and cognitive testing to resolve the main questions identified in the present study.

## Implications for research and practice

### Implications for practice

Given the methodological nature of the research reported, implications for practice are limited, but twofold:

- On relativities, it could be said that it would be premature to propose any particular set of QALY weights at this point in time: before that point is reached, there is scope for both further reconciliation and replication. However, it might equally be argued that there is no scope for reconciliation and we need to choose between the results in light of the caveats of each.
- On valuation, it was never the intention to conduct a representative survey using a definitive method. The main recommendation, therefore, is that any future national sample survey should be preceded by further extensive qualitative research and cognitive testing to resolve the main questions identified in the present study.

### Research recommendations

The research recommendations arsing from the study are that:

- The findings from the relativities study indicate that more work is required in the short term to reconcile the results obtained, although fundamental differences between the methods and results reported may challenge such reconciliation.
- Longer term, and still with respect to relativities, further methodological research should attempt to account for some of the

deficiencies of the methods (especially the particular discrete choice approach used in SVQ).

- Building on the results of the innovative methods that have been devised in this study to derive relative weights, further replication of these results is required to address this important policy issue.
- With respect to valuation, shorter-term work is required around the issues of aggregation, combining WTP and SG values and the appropriateness of different measures of central tendency.
- Longer term, more qualitative and cognitive research is required around two valuation issues in particular: first, the problem of identifying health states to present to respondents which are 'minor enough' for people to be able to express their WTP but not so minor that respondents will accept only minuscule risks of death when responding to SG type questions; and, related to the first, the extent to which 'noise' and 'error' in people's responses might generate extreme and unreliable figures.

# Acknowledgements

The authors would like to acknowledge the support of the (then) National Coordinating Centre for Research Methodology of the Department of Health, as well as the National Centre for Social Research, who conducted the two surveys reported on in this monograph.

## Contribution of authors

Rachel Baker contributed to the design of the study, analysis and interpretation of data, and drafting and editing of the report. She led the Q methodology and qualitative analysis and drafted Chapter 3. Ian Bateman contributed to the conception and design of the study, analysis and interpretation of data, revision of the report and approval of the final version. Cam Donaldson (lead author and principal investigator) contributed to the conception and design of the study, analysis and interpretation of data, and drafting and editing of the report. Michael Jones-Lee, Jose Luis Pinto Prades, Mandy Ryan, Phil Shackley, Richard Smith and Robert Sugden contributed to the conception and design of the study, analysis and interpretation of data, and revision and editing of the report. Emily Lancsar and John Wildman contributed to the design of the study, analysis and interpretation of discrete choice data, and drafting and revision of the report. Graham Loomes contributed to the conception and design of the study, analysis and interpretation of data, and drafting and editing of the report. He led the valuation study and drafted Chapter 6. Helen Mason contributed to the design of the study, analysis and interpretation of data, and revision and editing of the report. Maria Odejar contributed to the analysis and interpretation of discrete choice data. Angela Robinson contributed to the conception and design of the study, analysis and interpretation of data, and drafting and editing of the report. She led the analysis of person trade-off relativities study data and drafted Chapter 5.

# References


1. Torrance GW. Measurement of health state utilities for economic appraisal. *J Health Econ* 1986;**5**:1–30.

2. Williams A. Economics of coronary artery bypass grafting. *Br Med J* 1985;**291**:326–9.

3. National Institute for Health and Clinical Excellence. *Guide to the methods of technology appraisal*. London: NICE; 2007.

4. Donaldson C, Atkinson A, Bond J, Wright I. Should QALYs be programme-specific? *J Health Econ* 1988;**7**:239–57.

5. Weinstein M. A QALY is a QALY is a QALY – or is it? *J Health Econ* 1988;**7**:489–91.

6. Williams A. QALYs and ethics: a health economist's perspective. *Soc Sci Med* 1996;**43**:1795–804.

7. Rawlins M, Culyer A. National Institute for Clinical Excellence and its value judgements. *BMJ* 2004;**329**:224–7.

8. Devlin N, Parkin D. Does NICE have a cost-effectiveness threshold and what other factors influence their decisions? A binary choice analysis. *Health Econ* 2004;**13**:437–52.

9. Loomes G. Valuing life years and QALYs: transferability and convertibility of values across the UK public sector. In: Towse A, Pritchard C, Devlin N, editors. *Cost effectiveness thresholds: economic and ethical issues*. London: King's Fund and Office of Health Economics; 2002.

10. House of Commons Health Select Committee. *National Institute for Health and Clinical Excellence, first report session 2007–08, volume 1*. London: The Stationery Office; 2007.

11. Nord E, Street A, Richardson J, Kuhse H, Singer P. The significance of age and duration of effect in social evaluation of health care. *Health Care Anal* 1996;**4**:103–11.

12. Nord E. Towards cost–value analysis in health care? *Health Care Anal* 1999;**7**:167–75.

13. Dolan P, Olsen JA. *Distributing health care: economics and ethical issues*. Oxford: Oxford University Press; 2002.

14. Cropper ML, Aydede SK, Portney PR. Preferences for life saving programs: how the public discounts time and age. *J Risk Uncert* 1994;**8**:243–65.

15. Johannesson M, Johansson P-O. The economics of ageing: on the attitude of Swedish people to the distribution of health care resources between the young and the old. *Health Policy* 1996;**37**:153–61.

16. Dolan P, Shaw R, Tsuchiya A, Williams A. QALY maximisation and people's preferences: a methodological review of the literature. *Health Econ* 2005;**14**:197–208.

17. Schwappach DL. Resource allocation, social values and the QALY: a review of the debate and empirical evidence. *Health Expect* 2002;**5**:210–22.

18. Ratcliffe J. Public preferences for the allocation of donor liver grafts for transplantation. *Health Econ* 2000;**9**:137–48.

19. Schwappach DL. Does it matter who you are or what you gain? An experimental study of preferences for resource allocation. *Health Econ* 2003;**12**:255–67.

20. Johri M, Dmaschroder IJ, Zikmund-Fisher BJ, Ubel PA. The importance of age in allocating health care resources: does intervention-type matter? *Health Econ* 2005;**10**:461–74.

21. Mason H. *Monetary valuation of health outcomes for use in national policy formulation*. Newcastle upon Tyne: PhD; 2007.

22. Dupuit J. On the measurement of utility of public works. [orig. 1844]. *Int Econ Papers* 1952;**2**:83–110.

23. Davis R. Recreation planning as an economic problem. *Nat Resource J* 1963;**3**:239–49.

24. Jones-Lee M. *The economics of safety and physical risk*. Oxford: Blackwell; 1989.

25. Gafni A. Willingness to pay as a measure of benefits: relevant questions in the context of public decision making about health care programmes. *Med Care* 1991;**29**:1246–52.

26. O'Brien B, Gafni A. When do the 'dollars' make sense? Toward a conceptual framework for contingent valuation studies in health care. *Med Decis Making* 1996;**16**:288–99.

27. Acton JP. *Evaluating public programmes to save lives: the case of heart attacks*. Report No.: R950RC. Santa Monica: RAND Corporation; 1976.

28. Donaldson C. Valuing the benefits of publicly-provided health care: does 'ability to pay' preclude the use of 'willingness to pay'? *Soc Sci Med* 1999;**49**:551–63.

29. Donaldson C, Birch S, Gafni A. The pervasiveness of the 'distribution problem' in economic evaluation in health care. *Health Econ* 2002;**11**:55–70.

30. Johannesson M, Jonsson B, Borgquist L. Willingness to pay for anti-hypertensive therapy – results of a Swedish pilot study. *J Health Econ* 1991;**10**:461–74.

31. Olsen JA, Smith R. Theory versus practice: a review of 'willingness to pay' in health and health care. *Health Econ* 2001;**10**:39–52.

32. Smith RD. Construction of the contingent valuation market in health care: a critical assessment. *Health Econ* 2003;**12**:609–28.

33. Donaldson C. Eliciting patients' values by use of 'willingness to pay': letting the theory drive the method. *Health Expect* 2001;**4**:180–8.

34. Olsen JA, Donaldson C, Pereira J. The insensitivity of 'willingness to pay' to the size of the good: new evidence for health care. *J Econ Psychol* 2004;**25**:445–60.

35. Olsen JA, Donaldson C, Shackley P, Group E. Implicit versus explicit ranking: on inferring ordinal preferences for health care programmes based on differences in willingness-to-pay. *J Health Econ* 2005;**24**:990–6.

36. Olsen JA, Kidholm K, Donaldson C, Shackley P. Willingness to pay for public health care: a comparison of two approaches. *Health Policy* 2004;**70**:217–28.

37. Yeung RYT, Smith RD, McGhee SM. Willingness to pay and size of health benefit: an integrated model to test for 'sensitivity to scale'. *Health Econ* 2003;**12**:791–6.

38. Smith RD. Sensitivity to scale in contingent valuation: the importance of the budget constraint. *J Health Econ* 2005;**24**:515–29.

39. Protière C, Donaldson C, Luchini S, Moatti JP, Shackley P. The impact of information on non-health attributes on willingness to pay for multiple health care programmes. *Soc Sci Med* 2004;**58**: 1257–69.

40. van Exel NJA, Brouwer WBF, van den Berg B, Koopmanschap MA. With a little help from an anchor: evidence of starting point bias in contingent valuation of informal caregiver time inputs. *J Socio-Econ* 2006;**35**:836–53.

41. Carthy T, Chilton S, Covey J, Hopkins L, Jones-Lee M, Loomes G, *et al*. On the contingent valuation of safety and the safety of contingent valuation: Part 2 – the CV/SG 'chained' approach. *J Risk Uncert* 1999;**17**:187–213.

42. Johannesson M. The relationship between cost-effectiveness analysis and cost-benefit analysis. *Soc Sci Med* 1995;**41**:483–9.

43. Garber AM, Phelps CE. Economic foundations of cost-effectiveness analysis. *J Health Econ* 1997;**16**:1–31.

44. Culyer A, McCabe C, Briggs A, Claxton K, Buxton M, Akehurst R, *et al*. Searching for a threshold, not setting one: the role of the National Institute for Health and Clinical Excellence. *J Health Serv Res Policy* 2007;**12**:56–8.

45. Bate A, Murtagh M, Donaldson C. Managing to manage scarce resources in the English NHS. What can economics teach? *Health Policy* 2007;**84**:249–61.

46. Hirth RA, Chernew ME, Miller E, Fendrick M, Weissert WG. Willingness to pay for a quality-adjusted life year: in search of a standard. *Med Decis Making* 2000;**20**:332–42.

47. Gyrd-Hansen D. Willingness to pay for a QALY. *Health Econ* 2003;**12**:1049–60.

48. Mason H, Marshall A, Donaldson C, Jones-Lee M. *Estimating a willingness to pay based value of a QALY from existing UK values of prevented fatalities and serious injuries*. Birmingham: National Coordinating Centre for Research Methodology; 2005.

49. Byrne MM, O'Malley K, Suarez-Almazor ME. Willingness to pay per quality adjusted life year in a study of knee osteoarthritis. *Med Decis Making* 2005;**25**:655–66.

50. King JT Jr, Tsevat J, Lave JR, Roberts M. Willingness to pay for a quality adjusted life year: implications for societal health care resource allocation. *Med Decis Making* 2005;**25**:667–77.

51. Coast J, Horrocks S. Developing attributes and levels for discrete choice experiments using qualitative methods. *J Health Serv Res Policy* 2007;**12**:25–30.

52. Stephenson W. *The study of behavior: Q-technique and its methodology*. Chicago, IL: University of Chicago Press; 1953.

53. Brown SR. *Political subjectivity: applications of Q methodology in political science*. London: Yale University Press; 1980.

54. Baker R, Thompson C, Mannion R. Q methodology in health economics. *J Health Serv Res Policy* 2006;**11**:38–45.

55. Brown SR. A primer on Q methodology. *Operant Subj* 1993;**16**:91–138.

56. Greene WH. *Econometric analysis*. 6th edition. New Jersey: Prentice-Hall; 2007.

57. Small K, Rosen H. Applied welfare economics with discrete choice models. *Econometrica* 1981;**49**: 105–30.

58. Lancsar E, Savage E. Deriving welfare measures from discrete choice experiments: inconsistency between current methods and random utility and welfare theory. *Health Econ* 2004;**13**:901–7.

59. Dolan P, Green C. Using the person trade-off approach to examine differences between individual and social values. *Health Econ* 1998;**7**:307–12.

60. Chilton S, Covey J, Hopkins L, Jones-Lee M, Loomes G, Pidgeon N, *et al*. Public perceptions of risk and preference-based values of safety. *J Risk Uncert* 2002;**25**:211–32.

61. Johannesson M, Johansson P-O. Is the valuation of a QALY gained independent of age? Some empirical evidence. *J Health Econ* 1997;**16**:589–99.

62. Busschbach JJV, Hessing DJ, de Charro FT. The utility of health at different stages in life: a quantitative approach. *Soc Sci Med* 1993;**37**:153–8.

63. Ubel PA, Spranca MD, DeKay ML, Hershey JC, Asch DA. Public preferences for prevention versus cure: what if an ounce of prevention is worth only an ounce of cure? *Med Decis Making* 1998;**18**:141–8.

64. Dolan P, Tsuchiya A. Health priorities and public preferences: the relative importance of past health experience and future health prospects. *J Health Econ* 2005;**24**:703–14.