HEDS Discussion Paper 08/12

Published paper
None.

Related material
http://www.pcpoh.bham.ac.uk/publichealth/methodology/projects/RM03_JH11_PD.shtml

ScHARR

SCHOOL OF HEALTH AND RELATED RESEARCH

ScHARR

Health Economics and Decision Science Discussion Paper Series

No. 08/12

The relative societal value of health gains to different beneficiaries: a summary

Paul Dolan[1], Richard Edlin[2], Aki Tsuchiya[3] (On behalf of the NICE Social QALY team)

[1] Imperial College, London

[2] University of Leeds

[3] University of Sheffield

Corresponding author:
Aki Tsuchiya
Health Economics and Decision Science,
School of Health and Related Research
University of Sheffield
Regent Court
30 Regent Street
Sheffield
S1 4DA
UK
Tel: 0114 222 0710
Fax: 0114 222 4095
Email: a.tsuchiya@sheffield.ac.uk

The NICE Social QALY team

Chris Armitage, University of Sheffield
Amy Hukin, University of Sheffield
John Brazier, University of Sheffield
Rachel Ibbotson, Sheffield Hallam University
Stirling Bryan, University of Birmingham
Dick Eiser, University of Sheffield
Jan Abel Olsen, University of Tromsø

Acknowledgements
This discussion paper gives an overview of the project of the same name, funded by the NCCRM and NICE (RM03/JH11). We are grateful to the Centre for Research Evaluation at Sheffield Hallam University who collected the data, and all the study participants. We must also acknowledge Allan Williams' pioneering work in this area and his influence on each of the authors.

1. Background

It is widely accepted that one of the principal objectives of government expenditure on health care is to generate health. Since health is a function of both length of life and quality of life, the quality-adjusted life-year (QALY) has been developed in an attempt to combine the value of these two into a single index number (Dolan, 2000). QALYs are increasingly being used in the evaluation of health care interventions and have been recommended by the National Institute for Health and Clinical Excellence (NICE) for use in cost-effectiveness analyses of health technologies (www.nice.org).

There are, however, many concerns about the appropriateness of using QALYs to inform resource allocation decisions. These relate to issues about the extent to which QALYs adequately capture the individual benefits and social value from health care interventions. In what follows, we will assume that QALYs are an appropriate currency in which to express individual health benefits. At the social level, the de facto standard in economic evaluations is that "a QALY is a QALY is a QALY" i.e. that a given health benefit is valued equally regardless of who gets it. However, there is now growing evidence that this assumption is not a good representation of societal preferences (Dolan et al, 2005), and yet the evidence does not allow us to estimate equity weights for QALYs. This is the purpose of the present project.

Our specific goal is to elicit preferences from members of the general public that can be fed into a social welfare function (SWF), which allows us to estimate the weight given to the health of one population group relative to another. Most of the studies in this area have adopted a social perspective, in which respondents are asked to consider allocation decisions that they personally may not be affected by. We propose to do the same. Standard economic theory is concerned with the optimising behaviour of self-interested individuals, and so many economists are rather distrustful of preferences that contain no self-interest at all. However, although self-interest exists, it does not necessarily follow that it must – or should – be the basis for social welfare evaluation (Menzel, 1999; Dolan et al, 2003).

Before we could elicit equity weights, we needed, first, to determine the relevant attributes and levels over which to elicit preferences and, second, to develop methods which facilitate the elicitation of stable preferences. The next two sections deal very briefly with these two initial phases of the project. Section 3 discusses the general design and analysis issues in generating equity weights in the form a SWF. Section 4 presents the questionnaire format and specific questions for the main and additional studies, the analyses of these questions and the respondents in our main studies. Section 5 presents the main results. Section 6 discusses some of the methodological and policy issues that these results raise.

2. What to elicit preferences over and how to elicit them

Phase 1 of the project consisted of four studies. Studies A and B used focus groups with members of the general public to identify the relevant attributes and to identify meaningful levels of these attributes, respectively. Study A was a qualitative study involving 15 members of the public in group sessions. The attributes presented for consideration were: age, social class, length of time with condition, dependents, quality of life without treatment, and whether the condition was caused by NHS

negligence. All of these attributes were considered relevant in Study A in that respondents chose to depart from simple QALY maximisation in order to take them into account. Study B asked 42 individuals (across two rounds) to trade-off attributes against QALY gains. The study varied the number of QALYs that would have to be sacrificed to target a priority group (100 QALYs across 1000 individuals versus 500 QALYs across 1000 individuals). A smaller proportion of respondents were ready to diverge from QALY maximisation where the cost of doing so was larger.

Attributes and levels for which respondents were willing to sacrifice health gains across both levels of QALY sacrifice included: quality of life without treatment (when the lower level was 40%); time with condition (where one party had the condition for one year or more); and age (15-24 year-olds relative to 65-74 year-olds; and under-5s relative to 25-34 year-olds). Finally, Study C aimed to identify the most policy relevant combinations of attributes from a postal survey of NHS staff (n=172). The modal result across all attributes was that the attribute is relevant to NHS policy "depending on what else is known", and no attribute was regarded by a majority as relevant to NHS policy "no matter what". The three attributes where the majority of respondents thought the attribute in question "would be relevant depending on what else is known" were: quality of life without treatment (40% as opposed to 70%); length of time with condition (more than one year as opposed to recent); and age (5-25 as opposed to 60-80). The next most important attribute, but not a majority view, was NHS negligence.

Studies A to C identified three attributes for taking forward into later phases of the project: age, quality of life without treatment, and responsibility. Age was clearly important but there was no real consensus about when age mattered for policy and when it did not – except in the case of children versus adults. This is the focus of our comparisons. For quality of life without treatment, a value of 25% was chosen for comparison with full health. We can be confident that 25% health meets the threshold for the general public to treat quality of life as significantly lower than full (100%) health. In terms of responsibility, NHS negligence appears to be somewhat important. Since it seems unrealistic to contrast NHS responsibility with all other causes of ill health, or with 100% patient responsibility, it was decided to present this attribute as three categories: NHS responsibility and no patient responsibility; no NHS responsibility and limited patient responsibility; and no NHS responsibility and no patient responsibility.

Study D explored preferences relating to the concentration and dispersion of benefits across beneficiaries who are equal in all relevant aspects. Whilst people may prefer to spread out health benefits to a larger number of patients than to concentrate on a smaller number if the benefit per person is large enough, they may also prefer to concentrate than to disperse if the benefits per person from dispersion are below a certain threshold. From 68 respondents surveyed in group meetings, 2.6 years was identified as the threshold or 'tipping point'. This finding is used to guide the design of trade-off exercises used in the remainder of the project i.e. we make sure that the difference between two groups in any one scenario is at least 2.6 years.

Once the attributes and levels had been decided upon, Phase 2 of the project surveyed members of the public using two different designs to determine which of them was more conducive towards eliciting stable preferences. The "resource intensive" (RI)

design involved a group discussion with fellow participants prior to an individual, face-to-face interview, whereas the "interview only" (IO) design did not involve the group discussion stage. The objective of Phase 2 was to identify the impact on peoples' preferences of the group discussion, and of the opportunity to deliberate over the issues between this group discussion and the individual interview. The stability of preferences was captured by administering a series of attitudinal questions on resource prioritisation at multiple time points in the study design. There were 56 respondents in the RI design and 232 for the IO design.

The results suggest that design appears to have had no significant effect on the willingness to prioritise different groups. Deliberation does have an effect on general prioritisation preferences, but this appears to be limited to the different stages within the RI design rather than between the two designs. However, given the sample size, and the different distributions of background characteristics across the two sub-samples of this study, the interpretation of this is not quite so straightforward. At a practical level, the IO design proved to be far more straightforward in terms of recruiting respondents and it was much less costly. Since the prior elicitation of beliefs and attitudes appear to be more important in generating stable preferences than discussion groups, we use the IO design in the main elicitation phase of the project.

3. General design and analysis issues

Four important considerations are addressed here. First, the ways in which the trade-offs are to be specified: we favour a SWF approach. Second, what precisely the trade-offs are to be over (the 'distribuendum' as Dolan and Olsen, 2001, call it): we favour weightings over lifetime health. Third, the kinds of questions that follow from these first two considerations. Fourth, the method of analysis used to parameterise the SWF.

3.1 The trade-off (social welfare) function

In economics, the SWF is typically assumed to be a function of individual utilities, which are then weighted within the function to provide a trade-off in the utilities received by different beneficiaries (Layard and Walters, 1994). Several studies have used a SWF in the area of health economics to model preferences and balance the competing demands of efficiency and equity (Dolan 1998, Dolan and Robinson, 2001, Abásolo and Tsuchiya, 2004). In health contexts, non-health outcomes for an individual are often disregarded and the focus is instead on health rather than utilities (Dolan, 1998).

Several functional forms have been suggested for the SWF, and these typically involve some form of concavity conferring value to a more even distribution of outcomes. Prominent amongst these has been the constant elasticity of substitution (CES; Dolan, 1988; Lindholm and Rosen, 1998) in which the health of two equal sized groups is assessed:

$$W = \left[\alpha v_1^{-r} + (1-\alpha) v_2^{-r}\right]^{-\frac{1}{r}}, \qquad \alpha \in [0,1],\ r \in [-1,\infty) \setminus 0$$

where: v_X is the lifetime health of Group X,
α is the weight placed on the health of Group 1,
and r reflects the overall strength of inequality aversion.

The objective of the empirical study is to identify the inequality aversion parameter (r) and the relative weight (α), so that the marginal rate of substitution (MRS) for specific combinations of health between the two groups can be calculated. The MRS in this context represents the relative social value of a marginal change in the social value of health to one group relative to the other, keeping the total level of social welfare constant. If the MRS is 1.5, that means that if the health of Group 1 deteriorates by 1 unit, then the health of Group 2 will need to improve by 1.5 units in order to maintain the current level of social welfare. This will then suggest that the marginal social value of the health of group 1 is 1.5 times that of group 2, indicating the relative values to be used in resource allocation decisions.

Through the r parameter, the CES SWF can represent a variety of different attitudes towards relative inequalities, and hence also iso-welfare contours. All these contours are *homothetic*, so that the trade-offs between different factors (the MRS) are unaffected by proportional increases in all variables. In other words, homotheticity implies that the value of inequality reduction is expected to differ according to the *relative* difference between v_1 and v_2. For $r = -1$ and $\alpha = 0.5$, the function is a simple sum of the lifetime health of the two groups and no value is given to reducing inequality. And as r rises, increasing value is given to equity. At the extreme, as r approaches infinity, only the group whose lifetime health is perceived to be worse is given any importance. The α parameter allows for the groups to be weighted differently aside from any health differences between them. The trade-off between the health of both groups (the MRS) is given by $\alpha/(1 - \alpha)$ along the 45^0 line.

In a case where there are no differences to base an unequal relative weight on, then $\alpha = 0.5$ and social welfare is a function of only one parameter (r) for inequality aversion. As we shall see more fully in Section 3.2 below, "lifetime health judgements" (which represent the social value attached to profiles of health over a lifetime) can be written as a function of two parameters, and so as few as three sets of pairs of indifference points are sufficient to find a social welfare function covering efficiency and inequality with an additional pair of indifference points necessary to find each weight for non-health characteristics. In practice, we can use more than this in order to have more confidence in the results. Of the three main attributes considered in the study, timing and severity of ill-health are interpreted as part of the definition of lifetime health judgements, and its effects are reflected in r. Condition cause/responsibility is considered a non-health characteristic and its effects are reflected in α.

3.2 The trade-off (lifetime health) space

Consistent with most of the work in this area, and in keeping with the design of our studies in Phase 1, 'equity weights' refer to any conscious departure from the assumption that all QALYs should be weighted equally. There is an issue, however, about whether we start with all QALY gains (i.e. the benefits from treatment) as being equally weighted or whether we consider final outcomes, which combine starting point and gains (Dolan and Olsen, 2001). Whilst it is possible to frame the questions in either way, the final outcomes space takes account of potentially relevant additional information e.g. in relation to overall health. It might be possible to break down the final outcomes into current position and gains but this would require yet more

information for respondents to process. Since there is evidence that the greater the number of attributes presented simultaneously, the more likely individuals are to employ heuristics or shortcuts in making decisions (Payne et al, 1988; Lloyd 2003) rather than a substantive evaluation of the question.

When gains have arguably been the most salient consideration in the framing of the questions, there is some support for the notion that respondents are focussing on the final outcomes. If we consider the gains-space, then preferences should satisfy the Pareto Principle; that is, we should prefer to give benefits to one group if it does not imply a loss to the other group. However, we have found that up to 20% of respondents violate this basic principle (Dolan et al, 2002). This implies a backwards bending SWF (Abásolo and Tsuchiya, 2004) and it also provides evidence in support of a focus on final outcomes (which in these questions is more equal when the Pareto Principle is violated). Not violating the Principle may still be indicative of a focus on outcomes but without such a strong preference for reducing overall inequalities (Tsuchiya and Dolan, 2008). In any event, the use of gains is also problematic in that it requires the identification of a reference point (Kahneman and Tversky, 1979).

The choice of outcome space as opposed to the gains space is also associated with the use of lifetime health experience as the relevant distribuendum. It is possible to discuss priority and equity in health care resource allocation based only on the current health of patients, with no reference to the lifetime health of these patients. For example, one may argue that if the cost-effectiveness of treatment is the same, patients who are currently suffering in severe health should be given higher priority over those patients who are only suffering a mild health problem. However, it may be the case that this severe suffering is only for a very brief duration, whereas the mild suffering is to last much longer; and if so, it is not obvious that the former patients should always get priority over the latter patients.

In the simplest case, we could treat the social value of lifetime health as equal to the number of QALYs the person will live. The QALY gives no explicit weight to health at different life stages, as it is formed using only quality of life and duration. In contrast, societal weights may lead to different conclusions. We define a generalised QALY measure using attributes for childhood versus adult health and severe versus good health states (from Phase 1).

We use a functional form for v_i that equals:

$$v_i = \sum_{t=1}^{\infty} V(h_{it}, t),$$

where V is a weighting function based on health-related quality of life (h_{it}) of group i at time t and timing (t) that increases in health ($\frac{dV}{dh_{it}} > 0$). If v_i is multiplicatively separable into health and timing components then:

$$v_i = \sum_{t=1}^{\infty} x(h_{it})T(t), \qquad \frac{dx}{dh_{it}} > 0,\ T(t) > 0.$$

Within the study, we used a dichotomous variable for timing representing whether health is experienced up to or after 18 years of age, and quality of life at 0.25 (or 25% health) and 1.00 (100% health). A quality of life level at 0.00 (0% health, or dead) is

also used for computational reasons, although this is not valued. In contrast, the QALY takes the form $\sum h_{it}$ and thus includes no timing dimension or social weights. Expanding v_i we can write:

$$v_i = FHC \cdot y_{FHC} + SHC \cdot y_{SHC} + DC \cdot y_{DC} + FHA \cdot y_{FHA} + SHA \cdot y_{SHA} + DA \cdot y_{DA}, \quad \text{(Eq. 1)}$$

where:

- *FHC* is the value of a year in full health whilst aged < 18,
- *SHC* is the value of a year in 25% health whilst aged < 18,
- *DC* is the value of being dead whilst aged < 18,
- *FHA* is the value of a year in full health whilst aged ≥ 18,
- *SHA* is the value of a year in 25% health whilst aged ≥ 18, and
- *DA* is the value of being dead whilst dead aged ≥ 18,

and the y_{XXX} variables give the number of years spent in each health/time combination. For comparability with the QALY, *FHA* = 1 and *DA* = 0. If we can also say that health and timing are multiplicatively separable then:

$$v_i = y_{FHA} + FHC \cdot y_{FHC} + SHA \cdot y_{SHA} + (FHC \times SHA) \cdot y_{SHC}. \quad \text{(Eq.2)}$$

As an example, consider the case where the following two 'states' of affairs are regarded as equally good in terms of social welfare: In the first state, Groups 1 and 2 both experience 60 years of full health; in the second state, Group 1 experiences 65 years of full health, and Group 2 experiences 56 years of full health. Here, let us assume that $\alpha = 0.5$. Suppose, initially, that the value of health during childhood equals the value of health during adulthood (*FHC*=*FHA*=1), so that v_1=v_2=60 in the first state and $v_1 = 65$, $v_2 = 56$ in the second state. Here,

$$\left[0.5 \cdot (60)^{-r} + 0.5 \cdot (60)^{-r}\right]^{-\frac{1}{r}} = \left[0.5 \cdot (65)^{-r} + 0.5 \cdot (56)^{-r}\right]^{-\frac{1}{r}}.$$

Solving this numerically we find $r = 2.00$.

Suppose instead that the value of health during childhood is twice as high as the value of health during adulthood (*FHC* =2), so that now in the first state v_1=v_2=78 (since the first 18 years receive twice the weight of adult years) and $v_1 = 83$, $v_2 = 74$ in the second. Here,

$$\left[0.5 \cdot (78)^{-r} + 0.5 \cdot (78)^{-r}\right]^{-\frac{1}{r}} = \left[0.5 \cdot (83)^{-r} + 0.5 \cdot (74)^{-r}\right]^{-\frac{1}{r}}.$$

Solving this numerically we find $r = 2.89$. This illustrates how the parameters defining lifetime health judgements and inequality aversion are not independent from each other. Any observed level of aversion to in equality at the societal level is due to a combination of both the difference in lifetime health judgements and the significance of this difference to society.

Therefore, the inequality aversion parameter cannot be identified unless we know how big health differences are, their relative size, and the trade-offs society would make. Equally, as the preferences we observe are influenced by both inequality aversion and judgements about the value of health received, we cannot assess the size of the parameter defining FHC without knowing inequality aversion. We are, however, able to estimate these together.

In relation to the effect of non-health characteristics, consider the case where society is indifferent between the following states: In the first state, Groups 1 and 2 both experience 66 years of full health; in the second state, Group 1 experiences 60 years

of full health, and Group 2 experiences 75 years of full health. If all years of full health receive the same weighting, then this suggests $r = 2.65$. If the health of Group 1 deteriorates by six years and the health of Group 2 improves by nine years, then it follows from that SWF that the level of social welfare will remain unchanged. Suppose that after personal responsibility characteristics are added into questions the second state needs to be changed to 62 years of full health to Group 1, and 75 years of full health to Group 2 in order to achieve indifference between the two states. Society is now willing to give up only four years of Group 1's health in order to get nine years of Group 2's health. Group 1's health is therefore given a greater value than before. In other words, a nine-year improvement to the health of Group 2 will now only make up for a four-year deterioration to the health of Group 1 as opposed to six, so the health of Group 1 is now given a greater value than before.

Here, adding personal responsibility characteristics have led to increased value on the health of Group 1; the best-fit α increases to 0.614 (versus 0.50 previously) since:

$$\left[0.614 \cdot (66)^{-2.65} + 0.386 \cdot (66)^{-2.65}\right]^{-\frac{1}{2.65}} = \left[0.614 \cdot (62)^{-2.65} + 0.386 \cdot (75)^{-2.65}\right]^{-\frac{1}{2.65}}$$

The inequality aversion parameter (r) does not influence preferences where the health of both groups is equal. The relative weight α is the marginal value of an improvement to the health of Group 1 relative to an improvement to the health of Group 2 where health is equal (0.614/0.386 = 1.59).

3.3 Finding states with equal social welfare

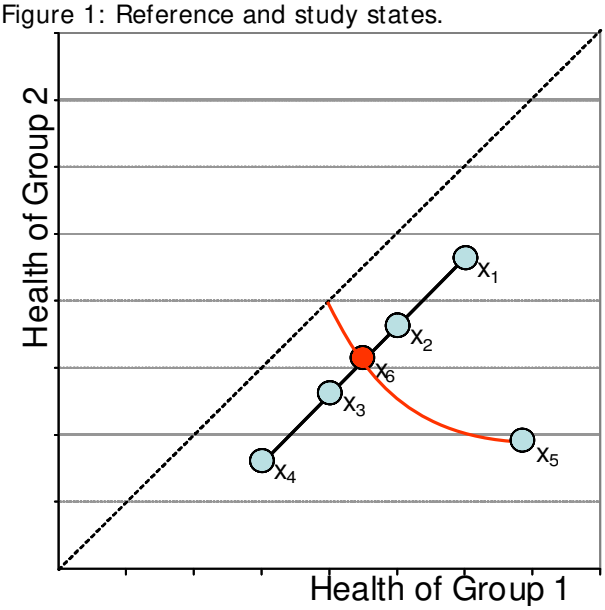
Our general approach is motivated around identifying pairs of states with the same level of social welfare attached to them, where each state represents a different combination of lifetime health to two population groups. In order to find these states, our questions could use different methods. Our preferred method is a simple pairwise task that asks respondents which of a pair of states they prefer and may also allow for indifference.

Given the use of pairwise data, we need to be able to identify pairs of equally good states from a societal point of view. We do this by asking series of choices that compare the same state ("study state") against a series of four "reference states". For each study state we have a "Choice Set" of four independent choices. Within each Choice Set, we aggregate preferences to find a sixth point that has the same social welfare as the study state. In Figure 1, where the axes represent the health of Groups 1 and 2 respectively, reference states are labelled as x_1 to x_4, the study state is labelled as x_5.

The reference states are constructed so that, when graphed, it is possible to draw a straight line through all four states. Since health is higher for both groups in x_1 than in x_4, it is reasonable to assume that individuals will prefer x_1. In the same way, x_1 is preferred to x_2, x_2 is preferred to x_3, and x_3 is preferred to x_4. This is represented as: $x_1 \text{ f } x_2 \text{ f } x_3 \text{ f } x_4$. Suppose that (as in the example above) an individual prefers the study state (x_5) over the worst two reference states (x_3 and x_4) but prefers the best two reference states (x_1 and x_2) to the study state. For such an

individual, x_1 f x_2 f x_5 f x_3 f x_4. This tells us where x_5 lies in the order of x_1 to x_4 for that individual.

Figure 1: Reference and study states.



In practice individuals make errors when making choices and might, for instance, indicate that they prefer x_5 to x_3 when the opposite is the case. This would mean that we might form the wrong preference ordering based on their individual data. In other cases, errors may prevent us from finding a preference ordering at all. For example, if an individual states that x_5 f x_2 and x_5 f x_3 then this suggests that x_3 f x_2, but we know that x_3 p x_2. In this case, we cannot form a coherent ordering between the individuals, and the assumption of *transitivity* is violated.

Whilst preferences at an individual level are "noisy", we would expect these sorts of errors to balance out at an aggregate level. For this reason, we focus on preferences from all individuals together. That is, instead of finding where x_5 falls in a preference ordering for each individual, we instead consider where it falls in an *overall* ordering. We know that one reference state (x_4) is the worst state and has the lowest value amongst the four reference states and, as we move towards the best of the states (x_1), social welfare increases. The aim of the analysis is to find a point along this progression that has the same social welfare as the study state, which can then be called the "equivalent state". The equivalent state is labelled x_6 (see Figure 1 above). The pair of the "study state" and the "equivalent state" can then be used to estimate the two key parameters of a CES social welfare function.

When we prefer one state to another, it is because the social welfare assessment we use (in that particular comparison) for the preferred state is higher than the social welfare assessment of the non-preferred state (in that particular comparison). If we had two identically-good states, then we would expect that half our sample would prefer one state and half would prefer the other state. $p(x_i, x_j)$ is defined as the proportion of the sample preferring x_i to x_j.

Thurstone (1927a, 1927b) outlined a simple method for scaling pairwise data that creates a cardinal scale, and we use the simplest (Case V) version here to scale social welfare judgements. This assumes that the social welfare assessments are independently and normally distributed with a common variance σ^2. The Thurstone score assumes that x_i is preferred to x_j when an individual perceives that x_i has the higher social welfare than x_j. If these assessments of social welfare are $W(x_i)$ and $W(x_j)$, then $p(x_i, x_j)$ represents the proportion of our sample who perceive $W(x_i)$ as larger than $W(x_j)$. Under the assumptions above, this is a function of the underlying mean difference between the two states ($\overline{W_i} - \overline{W_j}$). The Thurstone score transforms each proportion into standard normal scores and finds the average score for all of the comparisons using a state. This average score is our (unscaled) estimate of $\overline{W_i}$, where we consider all those choices using x_i. Without loss of generality, we scale these scores so that $\overline{W_2} = 1$ and $\overline{W_3} = 0$.

The Thurstone scores are based around a calculation of how often each state is preferred when it is compared to a state randomly chosen from all the possible states being compared (including itself). In this project, we consider a total of five states in each question. Here, there are 25 possible pairwise permutations, of which five see a state compared against itself (each state has a 50% chance of being preferred here). Of the remaining 20 permutations, we can infer the value of half of these from the remaining ten, since we know that one or the other must be preferred in each case. For the four reference states, the order of these states (x_1 f x_2 f x_3 f x_4) allows us to infer data since the monotonically superior state should be chosen in almost all cases. Six comparisons are provided in this way, leaving only four comparisons – x_5 versus x_1 to x_4 – to be identified in our survey work.

The score for the study state ($\overline{W_5}$) gives information about whether the study state is preferred in aggregate to each of the reference states. However, since these scaled scores are cardinally measurable social welfare values it also gives important information about how close the study state is (in social welfare terms) to these reference states.

Consider Figure 1. The lower is $\overline{W_5}$, the worse is the more unequal study state relative to the reference states, and so the less we have to move upwards from x_4 towards x_1 to find an equivalent point that has equal social welfare to the study state. The equivalent state (x_6) is defined as $x_3 + \overline{W_5}\,(x_2 - x_3)$. Where $\overline{W_5} = 0$, the equivalent state is x_3, whilst where $\overline{W_5} = 1$ the equivalent state is x_2. For $\overline{W_5} < 0$, the equivalent state is worse than x_3, and for $\overline{W_5} > 0$ it is better than x_2. It is possible that the equivalent state might be worse than the worst reference state or better than the best reference state in any comparison. The method will still allow the quantification of by how much better or worse the equivalent state (and thus the study state) is with respect to the reference states.

We can compute the scaled Thurstone scores using the aggregate preferences across all individuals, which then allow the construction of a set of equivalent states. These can then be used to estimate parameters in the SWF. Note, that whilst the values for the unscaled Thurstone scores are normally distributed, this is not the case for the scaled Thurstone scores, and hence also for the Thurstone-based equivalent state and the subsequent SWF parameters. As simple point estimates of uncertainty are of limited usefulness, we use bootstrapping to infer uncertainties. Bootstrapping assumes that the observed data are representative of the variation in the underlying population. A bootstrapping algorithm will select individuals at random and add them to a new (bootstrapped) dataset without removing them from the original dataset. This process continues until the bootstrapped dataset is of equal size to the original dataset. The analysis can be re-run on each new bootstrapped dataset and estimates obtained for equivalent states. Subsequent analyses can then also be re-run.

3.4 Parameterising the SWF

The Thurstone scores are used to infer an equivalent state (x_6) that is approximately indifferent to the study state (x_5). Given a set of n choice sets producing a pair of indifferent states (a study state plus equivalent state), define x_{i5} and x_{i6} as the ith pair of such states, where each is a vector of the time spent in each health/timing state (death whilst a child, severe ill health whilst a child, full health whilst a child, severe ill health as an adult, and full health as an adult). For simplicity, we re-define the function for lifetime health (consistent with the definition of Equation 1) as a function of the state considered and the parameters defining health for each group, l, that is:

$$v_l(x_{ij}; DC, SHC, FHC, SHA).$$

As our questions will always consider death after childhood, the value for *DC* will not be identified explicitly within the study, and so is excluded from the definition above.

Social welfare correspondingly becomes:

$$\mathrm{W}(x_{ij}; r, \alpha, SHC, FHC, SHA)$$
$$= \left[\alpha\left(v_1(x_{ij}; SHC, FHC, SHA)\right)^{-r} + (1-\alpha)\left(v_2(x_{ij}; SHC, FHC, SHA)\right)^{-r}\right]^{-\frac{1}{r}}.$$

In our estimated social welfare function, indifferent points should receive the same social welfare value (W), so that any difference between them can be interpreted as an error. The sum of squared differences errors across our data (X) is:

$$\mathrm{E}(r, \alpha, SHC, FHC, SHA \mid X)$$
$$= \sum_{i=1}^{n}\left(\mathrm{W}(x_{i6}; r, \alpha, SHC, FHC, SHA) - \mathrm{W}(x_{i5}; r, \alpha, SHC, FHC, SHA)\right)^2.$$

Within any question, a higher value for *FHC* will tend to leave the absolute differences (in spatial terms) between x_{i5} and x_{i6} reasonably constant, but will tend to decrease their relative differences – so a larger r is required, as in the example above. Unfortunately, as r rises, the effect is to compress the values found for W and so reduce the size of the differences between x_{i5} and x_{i6}. This leads to non-convergence since we can always reduce the sum of squared differences by increasing *FHC* and solving for the remaining parameters.

We therefore take a different approach in which we multiplicative separability and allow *SHC* to vary (so, use Equation 1). The value of the remaining parameters are found given a value for the inequality aversion:

$$\text{minimise } \mathrm{E}(SHC, FHC, SHA \mid r, X)$$
$$\text{with respect to } SHC, FHC, SHA.$$

Given this formulation, we define the error-minimising values of the parameters as $FHC(r)$, $SHA(r)$ and $SHC(r)$. Now also $k(r) = FHC(r) \times SHA(r) - SHC(r)$ measures the difference between the value of *SHC* against the result that could be derived with multiplicative separability. Where $k(r) = 0$, Equation 2 holds. Within our testing, there appears to be a single solution in *r* for this case and so we are able to parameterise the basic social welfare function where $\alpha = 0.5$.

Within the general form of the SWF, the parameter α is used to capture the degree to which groups are treated differently for non-health reasons. The general method in finding values for this is to compare cases where questions include a particular characteristic – for example, a difference in the cause of a condition – with others whether the question do not.

So given a set of questions, we first find sets of societally-equivalent points and these are then used to populate a social welfare function. However, this gives only a point estimate for the parameters of the social welfare function. This would be of relatively little value to decision makers (as a framework) even in an ideal world where all relevant methodological questions were answered, as it does not provide any indication of the inherent uncertainties in the analysis. To resolve this we use bootstrapping (sampling with replacement). Assuming that the data are representative of the underlying uncertainty, bootstrapping allows the construction of additional samples of the same size as the original sample. By re-running the analysis on these samples it is possible to estimate a distribution for each parameter. This distribution allows an estimate of uncertainty in the point estimate of each parameter. In this way, we can address the non-methodological uncertainties of our estimates.

4. Study design

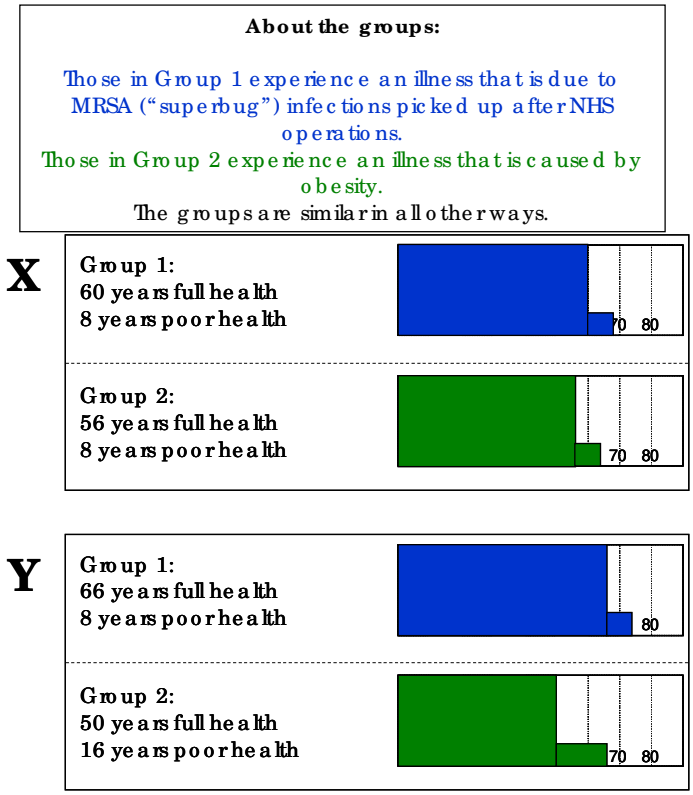
4.1 Questionnaire format

For each choice in the main preference elicitation task, respondents were asked which of two states they preferred and indifference between the states was accepted. Each choice is prefaced by a text box giving information about the choice context followed by the choice itself. As an illustration, Figure 2 displays one of the choices used in Choice Set 5.1. Responses are asked "Which scenario would you prefer NICE to bring about?" and can indicate a preference for either option or indifference ("I don't mind if it's X or Y").

The main preference elicitation task is comprised of 16 choice sets in six questions, each composed of four choices between two states at a time. Table 1 summarises these choice sets and their purposes. Choice Sets 1.1-1.4 involve choices where all life years are lived in 100% health, after which all those in the groups die. There are no differences between the groups in the type of condition experienced, the rarity of the

condition, or the numbers in each group. They are used to test the CES functional form (where the health of adults and children are valued the same). These responses can also be used to find a value for the inequality aversion (r) parameter within the SWF, and provide an opportunity to gain familiarity with the question format.

Figure 2: Example of choice context box and choice diagrams.



Choice Sets 2.1-2.4 again involve choices in which there are no differences between groups on rarity and condition cause/responsibility grounds. In these choices, individuals may experience a period of ill health at 25% quality of life. These questions are used to find weights given for childhood versus adult health, 25% health versus 100% health, and a value of r. The next choice sets introduce a condition cause/responsibility dimension into choices that are otherwise identical to those in Choice Sets 2.1 and 2.3. Choice Sets 3.1 and 3.2 compare a Group 1 with NHS caused illnesses with a Group 2 that has illnesses partly caused by the patient's lifestyle. Choice Sets 4.1 and 4.2 compares a Group 1 with NHS caused illnesses with a Group 2 that has illnesses not caused by the patient's lifestyle. These questions are used to find global (α) weights representing the effect of condition cause.

Choice Sets 5.1 and 5.2 repeat Sets 3.1 and 4.1 but name a more specific cause of illness rather than a general description of the type of illness. In place of NHS caused illnesses we have MRSA (Methicillin-Resistant Streptococcus Aureus) infections; non-NHS, partial patient condition causes are instead labelled "obesity"; and non-patient, non-NHS illnesses are labelled as "workplace exposure to hazards" in Phase 2. (In Phase 3 "genetic disorder" is used). These questions are used to find the effect of providing further information about illnesses. As the description of the causes are the only differences between Choice Sets 3.1 and 5.1 (and 4.1 and 5.2), we can tell

Table 1: The structure of the preference elicitation task

Question	Choice Sets	Illness	Data used to derive:
1	1.1-1.4	No	Inequality aversion (standard QALY model assumed to hold)
2	2.1-2.4	Yes	Lifetime health judgements (social value of timing of ill health) inequality aversion (where QALY model does not hold).
3 & 4	3.1-3.2, 4.1-4.2	Yes	Effect of condition cause/responsibility (abstract descriptions) (with 2.1-2.4)
5	5.1-5.2	Yes	Effect of condition cause/responsibility (labels) (with 2.1-2.4, 3.1, 4.1)
6	6.1-6.2	Yes	Effect of condition rarity (with 1.2)

Choice Sets	Description of Group 1	Description of Group 2
1.1-2.4	Groups 1 and 2 are identical except in the health that they experience	
3.1-3.2	Those in Group 1 experience an illness that is not the result of their lifestyles but is instead caused by errors within the NHS.	Those in Group 2 experience an illness that is caused by a combination of factors including poverty, genetics, pollution, and the patients' lifestyles.
5.1	Those in Group 1 experience an illness that is due to MRSA ("superbug") infections picked up after NHS operations.	Those in Group 2 experience an illness that is caused by obesity.
4.1-4.2	Those in Group 1 experience an illness that is not the result of their lifestyles but is instead caused by errors within the NHS.	Those in Group 2 experience an illness that is caused by a combination of factors including poverty, genetics and pollution, but is not caused by patients' lifestyles or by NHS error.
5.2	Those in Group 1 experience an illness that is due to MRSA ("superbug") infections picked up after NHS operations.	Those in Group 3 experience an illness that is due to workplace exposure to hazardous substances (e.g. asbestosis). Those in Group 3 experience an illness that is due to a genetic condition that affects the health of people in middle-age.
6.1 6.2	Those in Group 1 and Group 2 both experience illnesses that are caused by a combination of factors including poverty, genetics and pollution, but are not caused by patients' lifestyles or by NHS error. There are equal numbers of patients in both groups. Those in Group 3 and Group 4 both experience illnesses that are caused by a combination of factors including poverty, genetics and pollution, but are not caused by patients' lifestyles or by NHS error. The illness affecting those in Group 3 is extremely rare, and the illness affecting those in Group 4 is slightly more common. Your choice will affect an equal number of patients in Groups 3 and 4.	

whether preferences differ in the more-abstract categorical (Choice Set 3.1/4.1) and the less-abstract labelled case (Choice Set 5.1/5.2).

Choice Sets 6.1 and 6.2 are largely stand-alone and assess the effect of rarity of diseases on choices. In Choice Set 6.1, both groups suffer from an equally common condition and in Choice Set 6.2 one group suffers from a rarer condition. Respondents are instructed that the numbers affected in each group remain the same in all four cases.

An additional study considers further the issues of timing and severity to provide additional information to parameterise the SWF. It introduced illnesses at times other than at the start and end of life with illnesses affecting individuals at 10 and 30 years of age. It also considers 50% health in place of 25% health. Data were obtained in nine choice sets across three questions (A1-A3). Question A1 includes four Choice Sets (A1.1-A1.4), of which the first two are used only to give respondents some experience of the question format before the substantive data is obtained. In Question A1, all ill health occurs at the end of life. Question A2 varies the time at which ill-health occurs from the start of life (A2.1, A2.2) to 10 years of age (A2.3) and 30 years of age (A2.4). Question A3 uses a study state from the main study to allow comparison between the 25% and 50% study states.

Table 2 outlines the study and reference states used in the additional study. (Choice Sets A1.1 and A1.2 are not included in this table as they are used only to "warm up" respondents.) Note that Choice Sets A1.3-A1.4 and A2.1-A2.2 are identical to Choice Sets 2.1-2.4 with the 25% health states substituted for 50% health states with half the duration. Choice Sets A2.3 and A2.4 are identical to A1.3 with 8 years of illness moved from the end of life.

Table 2: Additional study questions

Study States	Group 1 Health	Group 2 Health
Choice Set A1.3	66 years in 100% health 4 years in 50% health	50 years in 100% health 8 years in 50% health
Choice Set A1.4	72 years in 100% health 8 years in 50% health	48 years in 100% health 8 years in 50% health
Choice Set A2.1	66 years in 100% health 8 years in 50% health	2 years in 50% health 54 years in 100% health 2 years in 50% health
Choice Set A2.2	4 years in 50% health 72 years in 100% health 4 years in 50% health	48 years in 100% health 8 years in 50% health
Choice Set A2.3	66 years in 100% health 4 years in 50% health	10 years in 100% health 8 years in 50% health 40 years in 100% health
Choice Set A2.4	66 years in 100% health 4 years in 50% health	30 years in 100% health 8 years in 50% health 20 years in 100% health
Choice Set A3.1	66 years in 100% health 8 years in 25% health	50 years in 100% health 16 years in 25% health

Table 2 (cont.):

Reference States	Group 1 Health	Group 2 Health
x_1	62 years in 100% health 8 years in 50% health	60 years in 100% health 4 years in 50% health
x_2	60 years in 100% health 4 years in 50% health	56 years in 100% health 4 years in 50% health
x_3	59 years in 100% health 2 years in 50% health	54 years in 100% health 4 years in 50% health
x_4	58 years in 100% health	52 years in 100% health 4 years in 50% health

4.2 Analysis

The homotheticity property of the CES function allows us to make predictions about societal preferences regarding different distributions of health. In Question 1, the groups are identical except in the health they receive and so α =0.50 is assumed. Given this assumption, our analysis for Question 1 (Choice Sets 1.1 to 1.4) finds the *r* values that correspond to a SWF that assumes that the standard QALY model holds. As Choice Sets 1.3 and 1.4 were derived by halving the total number of life years enjoyed in Choice Sets 1.1 and 1.2 in all states, then it also halves the number of QALYs received. If the number of QALYs describes how society judges individual lifetime health, then the CES function predicts that the same amount of inequality aversion should be exhibited in both cases. If this does not hold then either the QALY does not measure individual preferences or homotheticity does not hold (or neither holds).

When analysing Sets 1.1 to 1.4, we assess the consistency of individual choices against the predictions of the CES-SWF. We do this by comparing responses across questions to look for differences in inequality aversion in similar-sized questions (1.1 vs. 1.2; 1.3 vs. 1.4), and when homotheticity is tested (1.1 vs. 1.3; 1.2 vs. 1.4). Choice Sets 2.1 to 2.4 are used to construct both an alternative to the standard QALY when considering lifetime health from a societal perspective and to estimate the degree of inequality aversion exhibited in the data. These choices are analysed in terms of "Adult Healthy Year Equivalents" or AHYEs, v_1 and v_2, which like the QALY combine information regarding both health (quality of life) and the timing of health. In Question 2, we use $\alpha = 0.5$ as Groups 1 and 2 are identical in all respects except the health they receive.

The timing variable takes two values, and distinguishes between the health experienced prior to 18 years of age and all health experienced at and above this age. Health takes three levels, being dead, 25% and 100% health. These judgements may be consistent with conventional unweighted QALYs but they may give different weights to ill health at different ages than suggested by the QALY model. The AHYE reflects these judgements and values a profile of health using the number of years in full health as an adult that would be equivalent to it. Where the conventional unweighted QALY model adequately describes societal health judgements, the two concepts coincide.

Given the study and Thurstone-equivalent states for Choice Sets 2.1 to 2.4, we can estimate a basic form for the SWF (using AHYEs) in the case where the groups are identical in non-health respects ($\alpha = 0.5$). Earlier choice sets from Question 1 are ignored here because they do not consider periods of ill-health, and there is a concern that this may affect the responses given. (Later choice sets from Questions 3 onwards are ignored at this stage because α will not typically equal 0.5.) Of interest here is whether, and how, these estimates differ from the normal assumptions of cost-effectiveness analysis:

- Do overall preferences exhibit inequality aversion – that is, is the inequality parameter (r), typically greater than negative one?
- Is there extra weight placed on the health of children versus those of adults – that is, does $FHC = FHA = 1$?
- Is there a premium (discount) placed on the value of 25% health over the 25% that cost-effectiveness analysis assumes?

Within the general form of the SWF, the parameter α is used to capture the degree to which groups are treated differently for non-health reasons. In order to assess the effect of condition cause/responsibility, we compare cases where this information is provided, with otherwise identical cases where it is not. The states in Choice Set 2.1 also appear as 3.1, 4.1, 5.1 and 5.2. The states in Choice Set 2.3 appear as 3.2 and 4.2. In each case, the condition cause choice sets in questions 3 and 4 are compared to the baseline choice set in question 2.

The trade-offs where health differences do not exist between the groups – i.e. where only α may differ – are found by again comparing the results from Choice Sets 2.1 and 2.3 with those of Choice Sets 3.1-5.2. For Choice Sets 2.1 and 2.3, we solve to find the α providing identical social welfare values between the study and equivalent states in each case. This provides a baseline figure (α_0) correcting for any residual error in the question – which is likely to exist because the main estimates of *r*, *FHC* and *SHA* are based reducing error across Choice Sets 2.1-2.4 as a whole. For these baseline cases, the marginal rate of substitution in the absence of health differences equals $\alpha_0/(1-\alpha_0)$.

We also solve for the individual parameter values (α_1) for Choice Sets 3.1-4.2, with corresponding marginal rates of substitution ($\alpha_1/(1-\alpha_1)$). The effect of the condition cause label (in the absence of health differences) is the ratio of the two marginal rates of substitution. Similarly, by comparing Choice Sets 3.1 and 5.1, and 4.1 and 5.2, we can consider the impact of labels versus the more abstract descriptions. Condition rarity is considered in Question 6. As with the condition cause questions above, we can define Choice Set 6.1 – where there is no difference in rarity – as the baseline and Choice Set 6.2 – where the same numbers are used but Group 3 has a very rare condition – as the comparator.

Bootstrapping is used to assess the uncertainty in the parameter estimates. Microsoft Excel was used to resample the data, with a random number generator to identify rows within the database containing complete data on Choice Sets 1.1 to 6.2 in the preference elicitation task. By sampling the same number of rows as we have individuals with complete data, we can find a new dataset that reflects a similar level of heterogeneity as the original dataset. By repeating the analysis (finding equivalent

states and parameter values) multiple times, we can estimate the uncertainty in each parameter value. For the main parameter estimates, a sample of 5000 observations was used, as this was considered likely to allow convergence when estimating of parameter uncertainty. Convergence was assessed by examining how quickly parameter estimates of standard deviation reached its final value. This was then used to estimate the necessary number of iterations in subsequent analyses.

We also consider whether the SWF parameters differ according to background characteristics in the main study. In each of the 12 respondent characteristic group, values for the parameters *r*, *FHC* and *SHA* (with $SHY = FHY \times SHA$) are found using preferences from Choice Sets 2.1-2.4. Uncertainty is again computed using bootstrapping, and comparisons are made the 12 respondent subgroups.

If differences are found by background characteristics, then the main analysis results may depend on the makeup of the sample itself. We therefore construct a virtual sample that is broadly representative of the general public where background differences affect preferences. In this sample, we require that the twelve background groups (gender × education × age) should be selected in the proportions they appear in the general public. This can be done using bootstrapping methods to provide both a central estimate for parameter values and uncertainty. We therefore re-run our analyses by resampling to our original sample size but require that each population subgroup provides the "correct" number of respondents when split by age, sex and education.

Within the main analysis of the project, we assume that all health below the age of 18 is "childhood" health, and all health after the age of 18 is "adult" health. Whilst the split between adult and childhood health is necessary within the analysis, the precise cut-off does not necessarily fall at 18.

Table 3 reports the background of the sample against data from the 2001 Census (or closest equivalent). In general, the sample is roughly representative in terms of age and gender (although our sample includes more 60-69 year olds). We slightly under-represent non-white ethnicities and those with disability/chronic illness, and over-represent the retired and those with higher or further education. Note, however, that the disability comparison is slightly different ("Do you consider yourself to be a disabled person" versus limiting long term illness. In any event, given that we correct for non-representative preferences in the sensitivity analysis (on age, sex and education), any non-representativeness is not a major concern.

Overall, nine interviewers were used to obtain the 559 interviews with complete data. Recruitment to the additional study took place in the two months following the main study. The additional study sampling was not designed to provide a mix of backgrounds but instead aimed to allow a quicker convenience sample which would provide indicative results only. Of the 130 interviews, 129 individuals provided complete data over the choice sets A1.3-A3.1. This sample was 47% female, 95% white, and with 37% above the age of 60. 53% were employed and 31% retired, with and 15% disabled. A larger proportion of those in the additional study owned their own homes or were mortgagees (91%).

Table 3: Background of the sample

Sample size		Main sample (%)	2001 Census (%)
Gender:	Female	55	52
Age:	18-29	20[a]	19
	30-39	16	20
	40-49	17	17
	50-59	15	16
	60-69	17	12
	70+	15	15
Ethnicity	White	95	92[b]
Employment status:	Self-employed	7	8
	Other Employed	39	52
	Retired	29	14
Education:	School only	47	78b
	HE/FE	53	22[b]
House ownership:	Owned/mortgage	71	71[c]
Disabled?	Yes	14	18[d]

[s] Includes 14 aged below 18.
[b] Ages 16-74 only
[c] 2000 data. Office of National Statistics.
[d] Limiting long-term illness.

5. Results

5.1 Inequality aversion

The Thurstone scores were found to place all the study states between x_2 and x_3 in preference terms, and this is consistent with the CES-SWF at an aggregate level. The scaled Thurstone score for the study state W(x_5) can also be used to define an "equivalent" state:

$$x_6 = \mathrm{W}(x_5)\,x_2 + (1\text{-}\mathrm{W}(x_5))\,x_3$$

Table 4 gives the study states and equivalent states that would be valued equally for each choice set. Within Choice Sets 1.1 to 1.4, the aggregate preferences suggest a trade-off between total health and reducing inequalities. In Choice Set 1.1, for example, respondents are willing to sacrifice 5.84 QALYs (70.00 – 64.16 QALYs) to the better off group in order to obtain 2.16 QALYs (58.16 – 56.00 QALYs) for the worse off group. This suggests an implicit marginal rate of substitution between the health of the worst off to the best off of 2.7; that is, the health of the worst is worth 2.7 that of the health of the best off. For the other choice sets, this figure varies from 1.42 (Choice Set 1.4) to 1.94 (Choice Set 1.3).

Table 5 gives the inequality aversion parameters, *r*, and measures of uncertainty within this for Choice Sets 1.1 to 1.4 under the assumption that the QALY describes the way that society judges lifetime health. For example, the trade-off for Choice Set 1.1 suggests an inequality aversion parameter of $r = 5.24$. The central estimates for the *r* parameters appear to differ across the four choice sets. Consider a standard case where Group 1 has a life expectancy of 70 years in full health and Group 2 has a life expectancy of 60 years in full health. The inequality aversion estimate from Choice

Set 1 would suggest that this is equivalent to case where both groups live 63.83 years in full health. Here, Group 1 loses 6.17 years and Group 2 gains 3.83 years, suggesting that across these improvements Group 2's health is worth 61% more than Group 1's health (6.17/3.83-1). For Choice Sets, 1.2, 1.3, and 1.4, the comparable figures are 21%, 38% and 13%, respectively.

Table 4: Question 1 Choice States and Equivalents

Choice Set & States		Group 1 Health	Group 2 Health	Trade-offs
1.1	Study State	70 years in 100% health	56 years in 100% health	2.70 QALYs (Group 1) per QALY (Group 2)
	Equivalent	64.16 years in 100% health	58.16 years in 100% health	
	Difference	- 5.84 QALYs	+ 2.16 QALYs	
1.2	Study State	74 years in 100% health	52 years in 100% health	3.64 QALYs (Group 1) per QALY (Group 2)
	Equivalent	63.80 years in 100% health	57.80 years in 100% health	
	Difference	- 10.2 QALYs	+ 2.80 QALYs	
1.3	Study State	35 years in 100% health	28 years in 100% health	1.94 QALYs (Group 1) per QALY (Group 2)
	Equivalent	32.36 years in 100% health	29.36 years in 100% health	
	Difference	- 2.64 QALYs	+ 1.36 QALYs	
1.4	Study State	37 years in 100% health	26 years in 100% health	2.03 QALYs (Group 1) per QALY (Group 2)
	Equivalent	32.31 years in 100% health	29.31 years in 100% health	
	Difference	-4.69 QALYs	+ 2.31 QALYs	

Results from samples of 5000 bootstrapped observations

Table 5: Inequality parameter (*r*) estimates by choice set

Choice Set	Mean	Minimum	Maximum	Std Dev	95% CI
1.1	5.24	3.64	6.96	0.50	(4.266, 6.216)
1.2	1.51	0.78	2.19	0.19	(1.139, 1.871)
1.3	3.16	1.68	5.01	0.45	(2.289, 4.041)
1.4	0.55	-0.05	6.96	0.18	(0.207, 0.895)

Estimates based on 5000 bootstrapped observations

However, the differences found between Choice Sets 1.1 and 1.3, and between Choice Sets 1.2 and 1.4 do appear to suggest a general violation of the CES-SWF in the case considered here, so that QALYs and AHYEs are expected to differ. We therefore relax the assumption that the standard QALY model is being used by individuals and in such cases we would not expect these pairs of choice sets to produce the same values for *r*, since halving standard QALYs is unlikely to halve the societal value of that health. As an illustration, suppose that childhood health is valued ten times as much as adult health. Living for 30 years (= 18+12) in full health provides 18×10+12 = 192 AHYEs. Halving this, 15 years in full health provides 15×10 = 150 AHYEs. Where timing affects preferences, we cannot test homotheticity directly.

Table 6 shows both the study states and their social welfare equivalents in Choice Sets 2.1 to 2.4. Within each of the equivalent states, ill-health always occurs at the end of life so, for instance, the equivalent state in Choice Set 2.1 involves 59.23 years of full health, followed by 4.92 years in severe health. The trade-offs defined in these states

are complex, as they involve periods in ill-health, periods in good health, health as children and health as adults. By changing the parameters in the SWF, we seek to find the solution that gives the least difference in social welfare across the pairs of states.

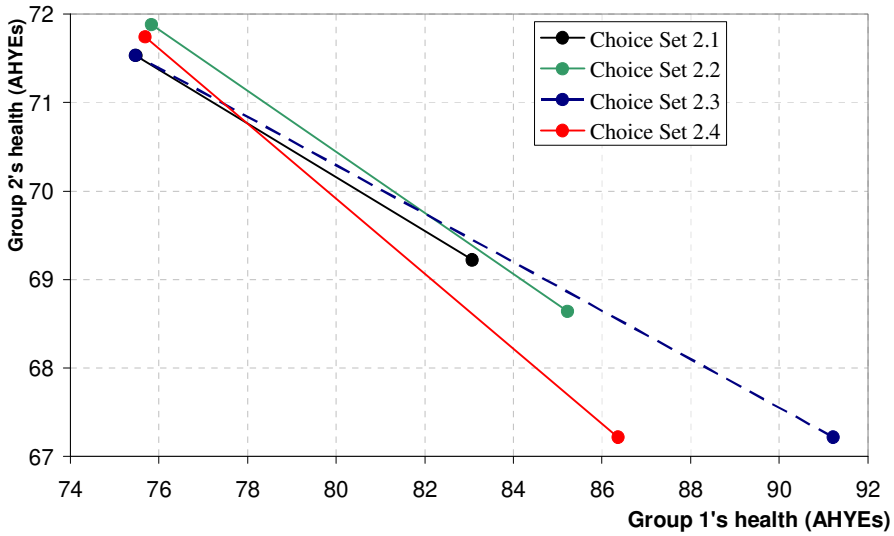
Table 6: Question 2 study states and equivalents

Choice Set & States		Group 1 Health	Group 2 Health
2.1	Study State	66 years in 100% health 8 years in 25% health	50 years in 100% health 16 years in 25% health
	Equivalent	59.23 years in 100% health 4.92 years in 25% health	54.46 years in 100% health 8 years in 25% health
2.2	Study State	66 years in 100% health 16 years in 25% health	4 years in 25% health 54 years in 100% health 4 years in 25% health
	Equivalent	59.40 years in 100% health 5.61 years in 25% health	54.81 years in 100% health 8 years in 25% health
2.3	Study State	72 years in 100% health 16 years in 25% health	48 years in 100% health 16 years in 25% health
	Equivalent	59.23 years in 100% health 4.92 years in 25% health	54.46 years in 100% health 8 years in 25% health
2.4	Study State	8 years in 25% health 72 years in 100% health 8 years in 25% health	48 years in 100% health 16 years in 25% health
	Equivalent	59.34 years in 100% health 5.34 years in 25% health	54.67 years in 100% health 8 years in 25% health

Results from samples of 5000 bootstrapped observations

Figure 3 plots the study and equivalent states together and it is clear that straight lines between these points do intersect (the two steeper lines (2.2, 2.4) are those where childhood illness may occur.) However, we can still note that there is a clear trade-off between average health and reducing inequalities in all cases, since the slope of all curves is less than one.

Figure 3: Choice Set 2 study state and equivalent points: AHYEs



The mean parameters, their estimated standard deviations, and 95% CI are presented in Table 7, along with the assumptions of a standard QALY-based CEA. The standard

CEA assumes no inequality aversion in its objective function (r = -1), whilst the SWF found here has significantly higher inequality aversion, with an inequality aversion significantly above 5. Consider a standard case where Group 1 has a life expectancy of 70 years in full health and Group 2 has a life expectancy of 60 years in full health; in terms of social welfare this is equivalent to a case in which both receive 64.76 years in full health.

Formally, this is analysed in terms of AHYEs where Groups 1 and 2 receive 84.9 (1.828*18 + 1.000*52) and 74.9 (1.828*18 + 1.000*42) AHYEs. This is equivalent to a case in which both receive 79.66 AHYEs (1.828*18 + 1.000*46.76). Group 1 loses 6.12 years and Group 2 gains 3.88 years of full health as an adult; over this change, Group 2's health is valued 57% more highly than the health of Group 1. (Since both changes concern adult full health years the same trade-off is made regardless of whether we consider QALYs or AHYEs.)

Table 7: SWF parameters; standard CEA assumptions and study estimates

Choice Set	Label	CEA	Study Estimates	Standard Deviation	95% CI
Inequality aversion parameter	r	-1.00	6.32	0.29	(5.76, 6.88)
Lifetime health judgements					
Value of 100% health as an adult	FHA	1.000	1.000	-	-
Value of 100% health as a child	FHC	1.000	1.828	0.031	(1.768, 1.888)
Value of 25% health as an adult	SHA	0.250	0.268	0.012	(0.244, 0.292)
Value of 25% health to a child	SHC	0.250	0.490	0.027	(0.439, 0.542)

Estimates based on 5000 bootstrapped observations

The standard QALY values 25% health for an adult as worth 0.250 times as much as a full year in 100% to an adult: at 0.268, this weight is not significantly different in the AHYE. The standard QALY also values 25% health for a child as worth 0.250 times as much as a full year in 100% to an adult. In contrast, our results suggest that 25% health as a child is worth 0.490 AHYEs, which is significantly more than its QALY weight (since the AHYE and QALY both agree on the value of an adult full health year). Overall, the AHYEs gives 96% more weight to the first 25% health for children relative to that given by the standard QALY.

5.2 Condition cause/responsibility

Table 8 provides the study states and equivalent states up to Choice Set 4.2, along with summary information regarding the cause of illness. Each pair of study and equivalent states suggests a trade-off between the health of one group and the health of the other. For Choice Set 2.1, which is neutral with respect to the allocation of responsibilities and thus will be used as the baseline, Group 1 loses 6.77 years in 100% health and 3.08 years in 25% health, whilst Group 2 gains 4.46 years in 100% health whilst losing 8 years in 25% health. In terms of societal judgements of health, Group 1 loses 7.60 AHYEs (6.77 + 0.268×3.08) and Group 2 gains 2.31 AHYEs (4.46 – 0.268×8).

In Choice Set 3.1, Group 1 has an NHS-caused disease and Group 2 has a disease due to non-NHS causes that are partially due to the patient's lifestyle. Here, Group 1 loses 7.03 AHYEs and Group 2 gains 2.86 AHYEs. Compared to the neutral Choice Set 2.1, here in Choice Set 3.1 where condition cause/responsibility is included, the public is less willing to sacrifice health to Group 1 in order to gain more health for Group 2; a AHYE to Group 2 is worth 3.28 AHYEs to Group 1 in Choice Set 2.1, and 2.46 AHYEs to Group 1 in Choice Set 3.1. The difference between these figures is interpreted as being due to the condition cause/responsibility, with a higher value is placed on NHS-caused diseases than ones in which the patient is partially responsible.

Table 8: Equivalent states by condition cause/responsibility

Choice Set(s) and States	Group 1 Health	Group 2 Health	Trade-offs
2.1, 3.1, 4.1 Study state	66 years in 100% health 8 years in 25% health	50 years in 100% health 16 years in 25% health	
2.1 Equivalent state	(Baseline) 59.23 years in 100% health 4.92 years in 25% health	(Baseline) 54.46 years in 100% health 8 years in 25% health	3.28 AHYEs (Group 1) per AHYE (Group 2)
	-7.60 AHYEs	+ 2.26 AHYEs	
3.1 Equivalent state	(NHS causes) 59.50 years in 100% health 6.01 years in 25% health	(Non-NHS, patient) 55.01 years in 100% health 8 years in 25% health	2.46 AHYEs (NHS causes) per AHYE (Non-NHS, patient)
	-7.03 AHYEs	+ 2.86 AHYEs	
4.1 Equivalent state	(NHS causes) 59.42 years in 100% health 5.67 years in 25% health	(Non-NHS, non-patient) 54.83 years in 100% health 8 years in 25% health	2.68 AHYEs (NHS causes) per AHYE (Non-NHS, non-patient)
	-7.21 AHYEs	+ 2.69 AHYEs	
2.3, 3.2, 4.2 Study state	72 years in 100% health 16 years in 25% health	48 years in 100% health 16 years in 25% health	
2.3 Equivalent state	(Baseline) 59.23 years in 100% health 4.92 years in 25% health	(Baseline) 54.46 years in 100% health 8 years in 25% health	3.65 AHYEs (Group 1) Per AHYE (Group 2)
	-15.74 AHYEs	+ 4.31 AHYEs	
3.2 Equivalent state	(NHS causes) 59.52 years in 100% health 6.08 years in 25% health	(Non-NHS, partial patient) 55.04 years in 100% health 8 years in 25% health	3.09 AHYEs (NHS causes) per AHYE (Non-NHS, patient)
	-15.14 AHYEs	+ 4.90 AHYEs	
4.2 Equivalent state	(NHS causes) 59.45 years in 100% health 5.80 years in 25% health	(Non-NHS, non-patient) 54.90 years in 100% health 8 years in 25% health	3.22 AHYEs (NHS causes) per AHYE (Non-NHS, non-patient)
	-15.29 AHYEs	+ 4.75 AHYEs	

Results from samples of 5000 bootstrapped observations

In Choice Set 4.1, Group 1 has an NHS-caused disease and Group 2 has a disease due to non-NHS causes that are not due to the patient's lifestyle. Here, Group 1 loses 6.58 years in 100% health and 2.33 years in 25% health (7.21 AHYEs), whilst Group 2

gains 4.83 years in 100% health and loses 8 years in 25% health (a gain of 2.69 AHYEs). Compared to the neutral Choice Set 2.1, here in Choice Set 4.1 where condition cause/responsibility is included, the public is again less willing to sacrifice health to Group 1 in order to gain more health for Group 2; a AHYE to Group 2 is worth 3.28 AHYEs to Group 1 in Choice Set 2.1, and 2.68 AHYEs to Group 1 in Choice Set 4.1. This suggests that more value is placed on NHS-caused diseases than ones in which illness is not due to non-NHS, non-patient causes.

Taken together, these two findings suggest that NHS-caused diseases may attract a higher value than the other two categories, with more value placed on diseases that are not due to patient lifestyle than those that are. Comparing Choice Sets 2.3 (no condition causes), 3.2, and 4.2 suggests a similar pattern.

Table 9 shows the significance of these weights. Relative to the case of non-NHS, non-patient causes, the case where patient lifestyle is a contributing factor is given 8.3% less weight (significant) in the questions based on Choice Set 2.1 and 4% less weight (insignificant) in those based on Choice Set 2.3. The case where illnesses are caused by NHS actions is significant in both comparisons, receiving 22.6% and 14.1% higher weight.

Table 9: Weights on condition cause/responsibility

	Choice Sets 2.1, 3.1, 4.1		Choice Sets 2.3, 3.2, 4.2	
	Mean	95% CI	Mean	95% CI
Non-NHS, non-patient causes	1.000	-	1.000	-
Partial patient causes	0.917	(0.842, 0.992)	0.960	(0.919, 1.001)
NHS causes	1.226	(1.107, 1.334)	1.141	(1.076, 1.207)

Estimates based on 5000 bootstrapped observations per phase

These findings are based on relatively abstract descriptions of the cause of illness and in practice these may cover a range of causes that people would have very different reactions to. In Phase 2, we compared MRSA (NHS cause), obesity (partial patient cause) and workplace exposure to hazardous substances (non-NHS, non-patient cause). In Phase 3, we replaced this last factor with a genetic condition affecting people in middle age.

Choice Sets 5.1 and 5.2 are used to assess the effect of these labels, and are in effect "labelled" versions of the more abstract Choice Sets 3.1 and 4.1. Again, the difference between the equivalent and study states for each choice set allows the identification of a change in health for both groups that yields the same social welfare. These changes can, as above, be represented in terms of societal health judgements and an average trade-off can be found. Table 10 presents these trade-offs for both comparisons (Choice Sets 3.1 vs 5.1 and 4.1 vs 5.1). There is generally less preference to treating obesity-related conditions over MRSA than when considering NHS caused versus partially patient caused diseases (comparing Choice Sets 3.1 and 5.1). In Phase 3, there appears to be a very similar priority the labelled case as generally given to the unlabelled case. In Phase 2, there appears to be slightly more priority given to treating

workplace hazards (vs MRSA) than to treating the more generic non-patient caused diseases.

The figures here suggest that individuals would accept *some* inequality whereby those who are obese receive less lifetime health than those who suffer illness due to a genetic condition. Larger inequalities would mean that society prioritises the health of the obese group over those with genetic conditions; smaller inequalities would mean that society prioritises the group with the genetic condition. Since these effects modify the degree to which inequality is taken into account, it is inappropriate to use them in isolation of the main inequality aversion parameter. Or, in other words, the weights here are derived from the α weights alone, and not the trade-offs made in the social welfare function where health differences are considered.

Table 10: Equivalent health changes and trade-offs: label effects

Choice Set(s) and States	Group 1 Health	Group 2 Health	Trade-offs
Study State	66 years in 100% health 8 years in 25% health	50 years in 100% health 16 years in 25% health	
3.1 Equivalent state	(NHS causes) 59.50 years in 100% health 6.01 years in 25% health	(Non-NHS, patient) 55.01 years in 100% health 8 years in 25% health	2.46 AHYEs (NHS causes) per AHYE (Non-NHS, patient)
	-7.03 AHYEs	+ 2.86 AHYEs	
5.1 Equivalent state	(MRSA) 59.66 years in 100% health 6.65 years in 25% health	(Obesity) 55.33 years in 100% health 8 years in 25% health	2.11 AHYEs (MRSA) per AHYE (Obesity)
	- 6.70 AHYEs	+ 3.18 AHYE	
4.1 Equivalent state	(NHS causes) 59.42 years in 100% health 5.67 years in 25% health	(Non-NHS, non-patient) 54.83 years in 100% health 8 years in 25% health	2.68 AHYEs (NHS causes) per AHYE (Non-NHS, non-patient)
	-7.21 AHYEs	+ 2.69 AHYEs	
5.2 (Phase 2) Equivalent state	(MRSA) 59.30 years in 100% health 5.21 years in 25% health	(Workplace hazards) 54.61 years in 100% health 8 years in 25% health	3.20 AHYEs (MRSA) per AHYE (Workplace hazards)
	- 7.49 AHYE	+ 2.34 AHYE	
5.2 (Phase 3) Equivalent state	(MRSA) 59.35 years in 100% health 5.41 years in 25% health	(Genetic disorder) 54.70 years in 100% health 8 years in 25% health	2.72 AHYEs (MRSA) per AHYE (Genetic disorder)
	- 7.30 AHYE	+ 2.69 AHYE	

Results from samples of 5000 bootstrapped observations

Note, however, that if the obesity group achieves less lifetime health then these health differences would suggest (in isolation of condition-cause/responsibility weights) they should receive *greater* priority. Beyond a critical difference in health, society would prefer to treat the obesity-related conditions because the differences in lifetime health outweighs the differences in condition cause. This was the case in Choice Set 5.1, where the overall preference gave twice the weight to treating less healthy, obese group relative to the healthier group suffering from MRSA.

5.3 Rarity

Choice Sets 6.1 and 6.2 are used to assess the effect of condition rarity. Table 11 outlines the equivalent health changes (the difference between the study states and equivalent states) for these two choice sets. In both choice sets, society appears to be more willing to prioritise the health of Group 2, as the less healthy group. In Choice Set 5.1, where there are no differences in rarity, a single AHYE to Group 2 is worth as much as 4.66 AHYEs to Group 1, whilst in Choice Set 6.1 where rarity is introduced, this falls to 4.00 AHYEs. This suggests that society may be more willing to prioritise groups with more rare conditions over less rare conditions. These figures relate to a question in which there is a health difference between the groups.

Where there are no health differences between the groups, then If the rare condition is given a weight of 1.00 (versus another "rare" condition, $\alpha = 0.5$, MRS = $\frac{\alpha}{1-\alpha} = \frac{0.5}{1-0.5} = 1$) then the very rare condition is given a weight of 1.19 ($\alpha = 0.543$, MRS = $\frac{\alpha}{1-\alpha} = \frac{0.543}{1-0.543} = 1.19$) versus the rare condition). The difference between these two weights is not statistically significant (95% CI = 0.785-1.597).

Table 11: Equivalent health changes and trade-offs: condition rarity

	Group 1 Health	Group 2 Health	Trade-offs
6.1	(Rare) +2.01 years in 100% health -20 years in 25% health	(Rare) +5.01 years in 100% health -16 years in 25% health	4.66 AHYEs (G1) per AHYE (G2)
	- 3.35 AHYE	+ 0.72 AHYE	
6.2	(Very rare) -6.43 years in 100% health -1.72 years in 25% health	(Rare) +5.14 years in 100% health -8 years in 25% health	4.00 AHYEs (G1, more rare) per AHYE (G2, less rare)
	- 3.26 AHYE	+ 0.81 AHYE	

5.4 Background characteristics

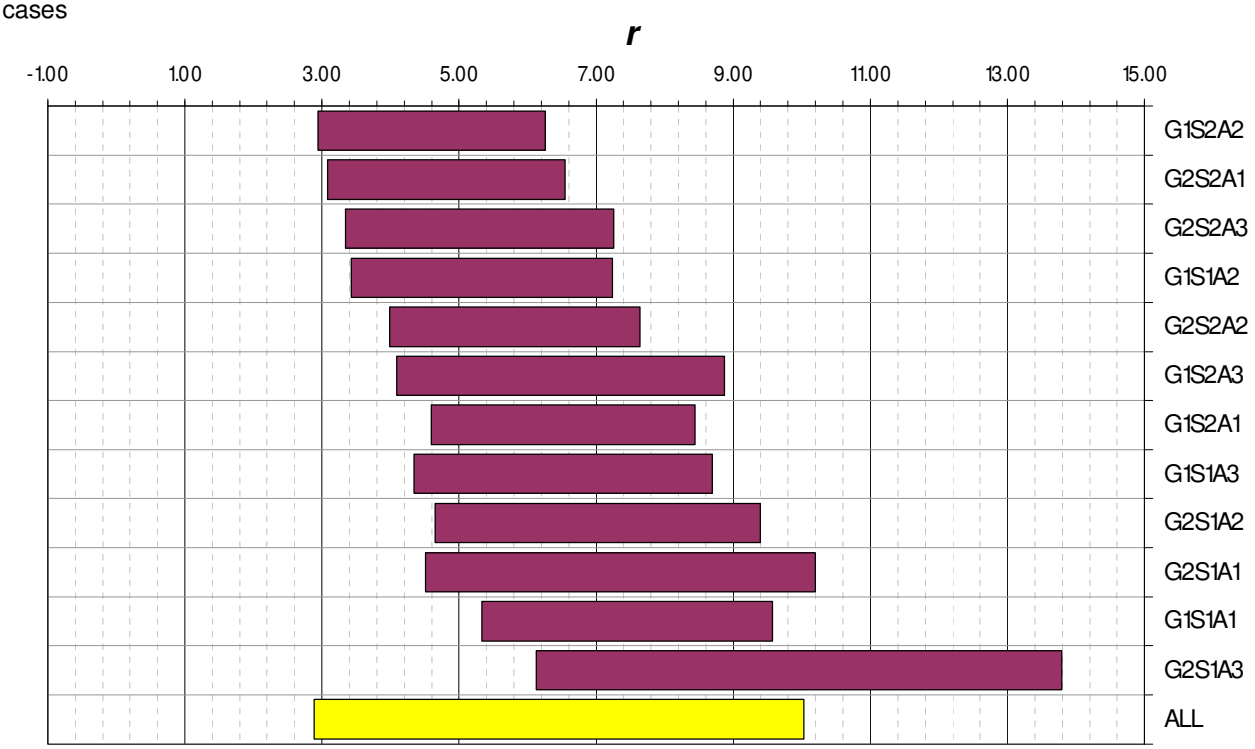
The effect of personal characteristics was assessed by comparing the values for the parameters α, *FHC* and *SHA* (with $SHC = FHC \times SHA$). The analysis compared values for both the 12 population subgroups and between those varying over a series of other background characteristics (health, disability, insurance, health and political beliefs). The bootstrapping used 200 resamples of the data for the latter comparisons, and 5000 for the former due to some instances where estimates in the population subgroups did not lead to convergence. The tables below should be interpreted with reference to Table 10 above, presenting the main baseline results with all respondents pooled.

The 12 population subgroups subdivide the main sample by age, gender and education. To simplify the analysis, we use the following notation: "GxSyAz" represents these groups; for x = 1 we have males, x = 2 females, y = 1 school educated, y = 2 HE/FE educated, z = 1 under 40s, z = 2 aged 40-59, z = 3 over 60s. Our estimates allowed all inequality aversion and indifference to inequality. Convergence issues were identified with bootstrapped resamples, and in 10 of the 12 groups, all 5000 cases converged within the bootstrapping. The other groups were female, school educated groups, with under 0.1% of cases in under 40 year olds and 0.4% of cases for over 60 year olds failing to converge.

We also found a series of cases in which $r = -1$, which represents the case where individuals are indifferent to inequality; lower values are not possible without the preferences being inequality seeking. In these cases, we also find negative values below -0.250 (and above –0.40) for health at 25% in both adult and child cases. The number of these cases varies, with 8 out of the 12 groups showing at least one case, and more than 1% of cases in the school educated, female group aged 40-59 (1.2%), and the HE/FE educated groups who were female and aged under 40 (13.9%), male and aged 40-59 (2.5%), and female and aged over 60 (4.6%). We proceed by considering only these convergent, inequality neutral/inequality averse bootstrapped cases.

Figure 4 shows the 95% confidence intervals for the *r* parameters within these convergent, inequality averse bootstrap cases (as before, based on Choice Sets 2.1-2.4). Here, we find five significant differences in the sensitivity to inequality: the male, HE/FE educated, 40-59 group (G1S2A2) has the lowest average inequality aversion of the twelve subgroups (r= 4.60), and is significantly different from the two groups with the highest estimated aversion – the male, school educated, under 40 group (r = 7.45, G1S1A1) and the female, school educated, 60+ group (r = 9.96, G2S1A3). This latter group has significantly higher inequality aversion than four other groups – the two male, 40-59 groups (r = 4.60 and 5.33; G1S2A2 and G1S1A2); and the female, HE/FE, under 40 and 60+ groups(r = 6.52, r = 6.48; G2S2A3 and G2S2A1). As a rule, the more highly educated groups (those with "S2") appear to display less inequality aversion than the other groups.

Figure 4 Inequality aversion parameter - 95% CI for converging, inequality averse cases



Respondents who were female, school educated, and aged between 40 and 59 give the highest weight to childhood health (*FHC*= 2.084, G2S1A2). This value is significantly higher than both women aged under 40 group (*FHC* = 1.770 and 1.709, G2S1A1,

G2S2A2), men who are HE/FE educated and aged between 40 and 59 (*FHC* = 1.608, G1S2A2), and HE/FE educated women aged over 60 (*FHC* = 1.661, G2S2A3). All groups gave values for childhood full health significantly above the adult value (*FHA* = 1.000 for all groups).

It appears that the value given to severe-ill health as an adult (*SHA*) generally increases with age. The under 40s groups, on average, give a value of 0.213 to the first 25% of health, compared to 0.251 from those aged 40-59. Those aged over 60 give a much higher average value, at 0.334. HE/FE educated women aged over 60s provide the highest value, (*SHA* = 0.431, G2S2A3), and this is significantly higher than all eight of the groups aged under 59 (male/female × schooling dimensions × two age groups). The lowest value is given by the female, HE/FE educated, under 40 group (*SHA* = 0.152, G2S2A1) and this group is significant against three of the four over 60s groups (G1S1A3, G1S2A3, G2S1A3).

As the preferences of the 12 population subgroups appears to differ, then the representativeness of the results becomes an issue. The subgroups within the existing sample of 559 individuals with full-data can be re-sampled in proportion to their population frequencies. These population frequencies are given in Table 12, and vary from a minimum of 10 (male, HE/FE educated, 60+) to a maximum of 89 (female, school educated, 18-40 year olds). A representative sample is typically younger, and less educated than our sample (which was selected for its spread rather than its representativeness).

Table 12: Inferred representative sample (n = 559)

Gender	Education	Age		
		< 40	40-59	60+
Male	School only	85	66	52
	HE/FE	28	21	10
Female	School only	89	74	74
	HE/FE	29	20	11

The analysis selects the number required from each population group from the main sample and analysis proceeds as normal. Whilst our sample is unrepresentative in the pattern of respondents, the results are very similar to our earlier analyses. On none of the parameters does the difference between the original and representative samples approach significance. In general, the level of inequality aversion appears to be slightly higher (r = 6.75, 95% CI 6.19-7.32), as is the value of 100% health to a child (FHC = 1.875, 95%CI 1.817-1.933), the value of 25% health as an adult (SHA = 0.272, 95% CI 0.248-295), and 25% health as a child (SHC = 0.509, 95% CI 0.458-0.561). The condition cause/responsibility weights are also similar.

5.5 The additional study

The best-fit solution here suggests that 50% health is weighted as equivalent to 0.615 of an adult healthy year – this suggests that the health between death and 50% health is worth 60% more ($\frac{0.615}{0.5} / \frac{0.385}{0.5} = 1.60$) than the health between 50% health and full health. The uncertainty in this figure is found through bootstrapping (n=200), and it appears that the value given to 50% health is significantly greater than 0.50 (95% CI, 0.533-0.698). The additional study questions also investigate the value of full health

years at different time periods. In the main study, ill health occurs either in the first 8 years or at the end of life, and the value of a full health year is split between the first 18 years (youth) and subsequent life (adult). The main study suggests that (lifetime equity aside) each year of full health prior to 18 years of age is worth 1.828 times as much as a year spent as an adult, with this latter figure serving as a numéraire (*FHA* = 1.000).

Choice Set A2.3 considers ill-health between years 10 and 18; after 18 years of age, it is assumed that the parameters for adults apply. Between 10 and 18 years, the value that minimises squared utility errors over Choice Sets A1.3-A2.3 assumes that a year of full health whilst a child is worth 1.073 times as much a year of full health as an adult (95% CI: 0.928-1.218). That is, it appears that health after the age of 10 is viewed very much like health to an adult. The analysis varying the "adult" age cut-off to 10 years above found that this made very little difference to the values of the variables defining lifetime health (the AHYE). Therefore, whilst the level for inequality aversion may change as a result of changing the cut-off, the major results of the project are unaffected.

Choice Set A2.4 considers ill-health between years 30 and 38, and was used to construct a weighting for the early period of adulthood (18 up to 40 years). For purposes of comparison, the numéraire period for health (*FHA* = 1.000) in which a year in full health equals 1 AHYE is now the period from 40 years onwards. Between 18 and 40 years, the value that minimises errors over Choice Sets A1.3-A2.2 plus 2.4 assumes that a year in full health is worth 0.989 AHYEs (95%CI: 0.888-1.091), with 50% health worth 0.609 AHYEs (95%CI: 0.499-0.720). Again, this suggests that those aged 30-38 (or by extension 18-40) are not treated differently from "normal" adults. There does not appear to be a premium placed on helping adults during the period where productivity-based estimates would suggest special emphasis is placed on health. Both estimates suggest that where lifetime health differences over an entire lifetime are accounted for (using the inequality aversion parameter), the value of health appears to be relatively constant except for a period quite close to the beginning of life, and possibly within the first 10 years.

The final choice set from the additional study used both 25% health and 50% health in the same question. The number of QALYs received in Choice Sets A1.3 and A3.1 are identical with the only difference being that Choice Set 3.1 uses 25% health for a longer period rather than 50% health for a shorter period. (The questions in Choice Set A3.1 are identical to Choice Set 2.1 in the main study.) The weight for SHA was selected here in order to solve for equality in utilities between the Thurstone-equivalent and study state in Choice Set A3.1. The best-fit figure here suggests that the first 25% of health is worth 0.231 of a full health life year (95% CI 0.212-0.249).

6. Discussion

This project builds on the existing literature on the social value of a QALY. There is a growing literature exploring whether or not publicly funded health care systems should treat all QALYs as having the same social value, and whether there is empirical support for such policies (see Dolan et al, 2005 for a review). The purpose of this project was to develop this work further and to consider whether it is possible

to generate a set of equity weights for QALYs. This section is intended to provide a general discussion of the results from our studies.

Following other research within health economics, we use a social welfare function (SWF) that allows trades-off between the lifetime health of different individuals. Since society may give different priorities to the health received at different ages and to different levels of severity, we consider social valuations over lifetime health and in relation to what we refer to as 'adult healthy year equivalents' (AHYEs) rather than in relation to 'standard' QALYs. In keeping with the conventional SWF framework, we focus on the valuation of outcomes rather than gains. We believe this focus on outcomes is preferable given its generalisability, as well as empirical concerns over the formation of reference points and violations of the Pareto principle (that is, from a given reference point, preferring to do less for both groups).

To our knowledge, Nord et al (1999) was the only previous paper to have advocated social judgements over social valuations of health. There, social valuations were limited to the somewhat unrealistic case that society will value every year of life lived to all chronically ill or disabled individuals equally so long as the individual concerned would prefer this state to death. This assumption was made to address the conflation of social values with the measurement of individual health benefits in person trade-off data. We believe our methods and analyses are unique in estimating both the social valuation of lifetime health and the value given to more equal lifetime health together.

The main findings of the study are summarised in Table 13. The timing of ill health seems to have a substantial impact. A greater weight was given to health experienced as a child in the main study (between ages 0-18 versus 18+), and the additional study suggesting that this might only apply within the first 10 years of life. We re-ran our main analysis to compare health in the first 10 years versus other timings (i.e. a 10 year cut-off for "childhood") and found no significant differences in our SWF parameters. This supports the use of a higher cost-per-QALY threshold for treatments that involve young patients. However, those interpreting the results of this study should refrain from extrapolating values in ranges that lie beyond the values used in the elicitation task. For example, in Choice Sets 2.1-2.4, where the impact of the timing of ill health is explored, none of the states involve a person dying before the age of 18.

In relation to severity, 25% health was valued as worth 0.268 of full health, so that there was a slight premium in the first 25% of health (but not significantly more than the 0.25 weight given by the standard QALY). 50% health was valued as worth 0.615 of full health, so that again there was a premium given to lower health states versus higher health states. Comparing 25% health and 50% health in the additional study, 25% health was valued as worth 0.231 of full health, with the health between 25% and 50% health therefore worth 0.384 (0.615-0.231). This suggests a slight premium for the health between 25% and 50% health. An indirect comparison across the main and additional studies suggests that the first 25% is worth 0.268, the next 25% is worth 0.347 (0.615-0.268) and the final 50% is worth 0.385.

Table 13: Summary of results from Phases 2 and 3
Baseline threshold of £20,000 for one year in full health as an adult (i.e. one AHYE)

Issue	Study	Results	Threshold per QALY change**
Timing of health	M	Significantly higher weight is given to health under 18 as compared to over 18.	£36,560 for age < 18, £20,000 for age > 18*
	A	There is no significant difference in the weight given to 10-18 years versus 18+ years, and no difference in the weight given to 18-40 years versus 40+ years.	£36,560 age < 10*, £21,460 age 10-18, £19,780 age 18-40 £20,000 age > 40*
Severity	M	The first 25% of health is given more weight than predicted by the QALY but not significantly so	£21,440 (4 years in 25% health) £19.520 (1.33 years from 25% to 100%)
	A	The first 50% of health is given significantly more weight than predicted by the QALY, with a premium between 25% and 50%.	£18,480 (4 years in 25% health) £30,720 (4 years from 25% to 50%) £15,400 (2 years from 50% to 100%)
	I	Premium on 25%-50% health remains but is reduced when considering 50% estimate from additional study with 25% estimate in main study	£21,440 (4 years in 25% health) £26,320 (4 years from 25% to 50%) £15,400 (2 years from 50% to 100%)

Issue	Results	Threshold for health change (per AHYE)
Inequality aversion	Significant difference from the inequality neutral case (r = -1), where health is summed across individuals i.e. willingness to prioritise more equal outcomes.	Not applicable. However, thresholds can be calculated for differences between groups (e.g. social classes).
Condition cause	Significantly higher weight to NHS causes and significantly lower weight to partial patient lifestyle causes versus non-patient, non-NHS causes	£18,340 per AHYE (partial patient cause) £24,520 per AHYE (NHS cause) £20,000 per AHYE (other causes)

Main (M), Additional (A), Indirect from main and additional studies (I)
* Fixed by assumption.
** Timing changes assume moving an individual from 0% to 100% health, severity changes assume an improvement to adult health.

There are, of course, some limitations to our research. The project has considered a very complex topic which is likely to have been challenging for many participants. Our study may have benefited from some qualitative study within Phases 2 and 3, and specifically a formal "think aloud" approach during piloting as compared to the informal feedback we received from interviewers. We would also have gained much in Phase 2 from recording the discussion groups but transcribing and analysing such data, as past experience has shown, is a particularly time-consuming activity.

More generally, the questionnaire in Phases 2 and 3 required a balance between breadth of approach and triangulation. The issues raised by the study are complex and the main questionnaire required a lengthy book of prompts (74 pages) and interview script (53 pages). In an ideal world, our main SWF would be based on a larger number of choice sets and would consider different amounts of lifetime health. We

would ideally have liked more levels for our attributes, particularly in the severity dimension. However, this was not practical given the size of the questionnaire and our targeted sample size. Overall, we believe that our design balances respondent fatigue and comprehensiveness of our results as well as we could be expected of the project.

So far as has been possible within time and resource constraints, we have addressed potential weaknesses in the methods within a series of sensitivity analyses (not all of which are reported here) and additional studies. Throughout, the results sections of this paper discuss the outcomes in terms of statistically significant differences in the key parameters. However, at a practical level, what actually matters is not necessarily whether the results are robust in that way, but whether the results would make a difference in terms of the incremental cost- effectiveness of an intervention. This is a complex issue that is beyond the aims of this project, and thus this paper does not address this.

We did not consider an alternative form for the SWF but the CES form is widely used and largely accepted by economists. The results that generate the SWF should always be placed in the appropriate context; that is, in the context of macro level decisions. We look forward to all of our results contributing to on-going debate and health policy in relation to the relative societal value of health gains to different beneficiaries.

REFERENCES

Abásolo I., Tsuchiya A (2004). Exploring Social Welfare Functions and violation of monotonicity: An example from inequalities in health. *Journal of Health Economics* 23(2): 313-329.

Dolan P (1998). The measurement of individual utility and social welfare. *Journal of Health Economics* 17(1):39-52.

Dolan P (2000). "The measurement of health-related quality of life for use in resource allocation decisions in health care" in AJ Culyer and JP Newhouse (eds), *Handbook of Health Economics* pp. 1723-1760.

Dolan P, Green C (1998). Using the person trade-off approach to examine differences between individual and social values. *Health Economics* 7(4):307-12.

Dolan P, Olsen JA (2001). Equity in health: the importance of different health streams. *Journal of Health Economics* 20:823-834.

Dolan P, Olsen JA, Menzel P, Richardson J (2003), An inquiry into the different perspectives that can be used to elicit preferences to inform priority-setting, *Health Economics*, 12, 7, 545-551.

Dolan P, Robinson A (2001). The measurement of preferences over the distribution of benefits: The importance of the reference point. *European Economic Review* 45(9):1697-1709.

Dolan P, Shaw R, Tsuchiya A, Williams A (2005). QALY maximisation and people's preferences: A methodological review of the literature. *Health Economics* 14(2):197-208.

Dolan P, Tsuchiya A, Smith P, Shaw R, Williams A (2002). *Determining the parameters in a social welfare function using stated preference data: An application to health.* Sheffield Health Economics Group, University of Sheffield, Discussion Paper Series 02/2.

Kahneman D, Tversky A (1979). Prospect theory: An analysis of decision under risk. *Econometrica* 47(2):263-291.

Layard R, Walters AA (1994). Allowing for income distribution. In *Cost-Benefit Analysis* (Layard R, Glaister S. eds) Cambridge University Press (reprinted from Layard R, Walters AA, 1978, *Microeconomic Theory*, McGraw-Hill).

Lindholm L, Rosen M. (1998). On the measurement of the nation's equity adjusted health. *Health Economics* 7(7):621-8.

Lloyd A (2003). Threats to the estimation of benefit: Are preference elicitation methods accurate? *Health Economics* 12(5):393-402

Menzel P (1999). How should what economists call 'social values' be measured? *The Journal of Ethics* 3:249-273.

Nord E, Pinto JL, Richardson J, Menzel P, Ubel P (1999). Incorporating societal concerns for fairness in numerical valuations of health programmes. *Journal of Health Economics* 8, 25–39.

Payne J, Bettman J, Johnson E (1988). Adaptive Strategy Selection in Decision Making. *Journal of Experimental Psychology* 14:534-52.

Thurstone LL (1927a). A law of comparative judgment. *Psychological Review*, 34, 273-286.

Thurstone LL (1927b). The method of paired comparisons for social values. *Journal of Abnormal and Social Psychology,* 21, 384-400.

Tsuchiya A, Dolan P (2008) Equality of what in health? Distinguishing between outcome egalitarianism and gain egalitarianism, *Health Economics*, forthcoming