Comparison of direct and indirect methods of estimating health state utilities for resource allocation: review and empirical analysis

David Arnold, medical student,¹ Alan Girling, senior research fellow,² Andrew Stevens, professor of public health,² Richard Lilford, professor of clinical epidemiology²

¹Cardiff University Medical School, Cardiff

²School of Health and Population Sciences, University of Birmingham, Birmingham B15 2TT

Correspondence to: R Lilford
r.j.lilford@bham.ac.uk

Cite this as: *BMJ* 2009;339:b2688
doi:10.1136/bmj.b2688

ABSTRACT

Background and objective Utilities (values representing preferences) for healthcare priority setting are typically obtained indirectly by asking patients to fill in a quality of life questionnaire and then converting the results to a utility using population values. We compared such utilities with those obtained directly from patients or the public.

Design Review of studies providing both a direct and indirect utility estimate.

Selection criteria Papers reporting comparisons of utilities obtained directly (standard gamble or time trade off) or indirectly (European quality of life 5D [EQ-5D], short form 6D [SF-6D], or health utilities index [HUI]) from the same patient.

Data sources PubMed and Tufts database of utilities.

Statistical methods Sign test for paired comparisons between direct and indirect utilities; least squares regression to describe average relations between the different methods.

Main outcome measures Mean utility scores (or median if means unavailable) for each method, and differences in mean (median) scores between direct and indirect methods.

Results We found 32 studies yielding 83 instances where direct and indirect methods could be compared for health states experienced by adults. The direct methods used were standard gamble in 57 cases and time trade off in 60 (34 used both); the indirect methods were EQ-5D (67 cases), SF-6D (13), HUI-2 (5), and HUI-3 (37). Mean utility values were 0.81 (standard gamble) and 0.77 (time trade off) for the direct methods; for the indirect methods: 0.59 (EQ-5D), 0.63 (SF-6D), 0.75 (HUI-2) and 0.68 (HUI-3).

Discussion Direct methods of estimating utilities tend to result in higher health ratings than the more widely used indirect methods, and the difference can be substantial. Use of indirect methods could have important implications for decisions about resource allocation: for example, non-lifesaving treatments are relatively more favoured in comparison with lifesaving interventions than when using direct methods.

INTRODUCTION

For resources to be allocated fairly, according to benefit gained per unit cost, between different patient groups, health economists and policy makers use a common currency of benefit. This standardisation is usually done by attaching different utilities (also known as preference weightings, health state values) to different health states.¹ Utilities are captured on a scale where 1 represents perfect health, 0 represents death, and states worse than death have negative values.

Measurement of utilities is a tricky and controversial area. Broadly, two groups of methods exist (fig 1). The first is based on mapping preferences directly onto the utility scale. This can be done by means of a trade off (standard gamble or time trade off) or visual analogue scale, which is less favoured.^{2,3} We refer to these as direct measures of utility.

The second is based on mapping preferences onto the utility scale indirectly via a generic health related quality of life questionnaire (such as European quality of life five dimensions [EQ-5D], short form six dimensions [SF-6D], and health utilities index mark 2 [HUI-2] and mark 3 [HUI-3]). Questionnaire responses are converted to utilities by means of “tariffs” or “weights.” These tariffs are available as a result of separate and previous exercises in which various possible health states have been calibrated by means of a trade off method from a sample of the general population. We refer to this group of methods as indirect methods of utility measurement.⁴⁻⁷

Indirect methods bypass the time consuming process of asking respondents to trade health states for different risks of death (or years of remaining life) each time a study is carried out; such trade offs need be done only once and the results obtained can thereafter be used to derive utilities using a simple questionnaire. The indirect methods encountered in this study are summarised in table 1.

The time trade off and standard gamble interviews must be implemented by a trained interviewer who explains the concept clearly without distressing or leading the patient. By comparison, questionnaires

Fig 1 | Direct versus indirect methods of utility elicitation See Brazier et al, 1999.²⁵

such as EQ-5D are simple, require little explanation, and stratify data into a number of different dimensions that the direct methods do not register (such as mobility and pain).

It is well known that different methods of utility estimation yield systematically different values.⁸⁻¹⁴ Standard gamble, time trade off, and visual analogue scale have all been compared across studies in a systematic way.¹⁵⁻²¹ However, the relation between direct and indirect utility measures has not been systematically documented across studies, despite the fact that differences between them have been identified in particular examples. Such examples have contributed to a widely held impression among health economists that direct methods tend to yield higher utilities (reflecting better reported health) for given health states than do indirect methods, irrespective of the type of direct or indirect method used (for example, time trade off versus standard gamble or EQ-5D versus SF-6D). The aims of the current study were therefore: to systematically confirm or refute the impression that direct methods tend to yield higher utility values than indirect methods; to

quantify the magnitude of any such differences; and to describe the relation between direct and indirect measures.

Data were collected by searching systematically for studies in which the same respondents completed a direct trade off and a generic quality of life questionnaire leading to an indirect utility assessment. In view of the acknowledged differences among the various direct and indirect methods, we examined our results overall and by subgroups based on the specific methods used.

Previous studies have shown that health state valuations obtained directly from affected populations are often higher (that is, health is rated as better) than those obtained from unaffected populations asked to make hypothetical judgments.^{22,23} For this reason we supplemented holistic comparisons of direct and indirect methods with an analysis stratified by the type of population.

METHODS

Literature search

We obtained studies from two sources. As the primary method, the search strings in table 2 were used to look for papers in the PubMed databases (covering all diseases) that mentioned a direct and an indirect method in the title or abstract. We then looked at each article to identify those in which the same group of patients had contributed direct utilities and had completed a health related quality of life questionnaire, the results of which had been translated into indirect utilities by means of pre-existing tariff values.

To verify the primary method, we also reviewed the Tufts database of utilities²⁴ for any studies that compared the utilities within the same group of respondents.

As a further check that our search strategy had not missed important studies, we searched PubMed (using the strings shown in table 3) for studies mentioning all

Table 1 | Indirect methods of utility assessment contributing to this study

Utility measurement	Description of questionnaire*	Tariff weighting obtained	Range of possible utility scores	Comment
European quality of life five dimensions (EQ-5D)	Five dimensions (mobility; self-care; usual activities; pain/discomfort; anxiety/depression), each of which is assigned one of three levels, yielding 243 possible combinations, each describing a particular health state	Sample of European general population (n=3395); time trade off valuation; hypothetical scenarios	-0.59 to 1.00	Preceded by EQ-6D; mostly used in continental Europe and the UK
Short form six dimensions (SF-6D)	Six dimensions (physical functioning; role limitations; social functioning; pain; mental health; vitality), with four to six levels, which define 18 000 health states.	Sample of UK general population (n=611); standard gamble valuation; hypothetical scenarios	0.30 to 1.00	Derived from the longer SF-36; applicable to SF-12 and SF-36
Health utilities index mark 2 (HUI-2)	Seven dimensions (sensory; mobility; emotion; cognitive; self-care; pain; fertility), with three to five levels, which define 24 000 health states	Parents of unaffected schoolchildren in Ontario, Canada (n=203); visual analogue scale transformed into standard gamble†; hypothetical scenarios	-0.03 to 1.00	HUI-2 originally used to capture quality of life in children; mostly used in Canada
Health utilities index mark 3 (HUI-3)	Eight dimensions (vision; hearing; speech; ambulation; dexterity; emotion; cognition; pain), with five to six levels, which define 972 000 health states	Representative sample of adults in Ontario, Canada (n=504); visual analogue scale transformed into standard gamble†; hypothetical scenarios	-0.36 to 1.00	Closely related adaptation of HUI-2, with a more detailed descriptive system; mostly used in Canada

*Dimensions used in indirect methods of assessment are sometimes referred to as multi-attribute utility scales (MAUSs). For further information see Brazier.²⁵

†Transformation carried out using conversion factors established in earlier studies.

Table 2 Search strings used for studies where direct and indirect utilities were compared in the same population

Search	Direct terms used	Indirect terms used	Total number of hits	Hits that fulfilled criteria
1	Standard gamble or SG	EuroQol or EQ-5D or EQ5D	58	15
2	Time trade off or TTO	EuroQol or EQ-5D or EQ5D	91	18
3	Standard gamble or SG	Short Form or SF-6D or SF6D	88	3
4	Time trade off or TTO	Short Form or SF-6D or SF6D	52	5
5	Standard gamble or SG	Health utilities index or HUI	116	9
6	Time trade off or TTO	Health utilities index or HUI	46	8

utility/preference ratings in four conditions for which utility measures are commonly made: asthma, diabetes, rheumatoid arthritis, and stroke. We then scrutinised studies to find out whether they compared direct and indirect methods in the same group of people.

The methodologies used to obtain “tariffs” (weights to convert the results of health related quality of life questionnaires to utilities) for the four indirect methods included among the comparative studies identified by our search have already been reviewed.²⁵ The methods used were time trade off in the case of EQ-5D, standard gamble in the case of SF-6D, and a visual analogue scale that was transformed into a standard gamble (using a conversion factor obtained in other studies) in the case of HUI.

Data extraction

Each eligible study was scrutinised by two authors (DA and AG) and the following information was extracted: title, authors, date, disease topic, the methods of elicitation of direct utilities (standard gamble, time trade off, or both); the generic questionnaires used (EQ-5D, SF-6D, HUI-2, HUI-3). Mean utility values for each method, with sample sizes, were extracted for the disease states in each included study. Medians were extracted when means were not reported. Standard deviations were recorded if they were available or could be deduced by working back from cited confidence intervals. Each study was classified according to whether the respondents were actual patients with direct experience of a certain condition (we refer to these as “current patients”) or were asked to imagine the experience of the condition (“hypothetical patients”).

Statistical methods

Many of the studies contained data on more than one health or disease state. Where the same state had been

assessed by separate groups of respondents (for example, baseline assessments in the two arms of a randomised controlled trial) these groups retained their identity in the analysis. In studies where the same group of respondents had provided utilities for more than one state, these states were combined (within a study) by pooling the data to ensure independence between different groups contributing to the analysis. This process entailed calculating “average medians” for studies that did not cite means. The end result was that each of the independent groups contributed a single mean (or median) observation on each of several utility methods. Differences between the utility observations obtained in this way cannot be relied on to follow a constant distributional form, even if they could be attributed to chance alone. Thus the sign test was used for paired comparisons of direct and indirect methods.

Average relations between direct and indirect utilities across the independent groups were fitted using least squares regression lines, with the constraint that all methods assign the value 1 to perfect health. In each case the scores from the direct method were regressed on the scores from the indirect method, to test the hypothesis that direct utilities could be predicted by applying a linear correction to the indirect utilities. The predictive value of the fitted lines was assessed using an approximate method. Residuals about the fitted line were calculated and standardised with respect to variation between subjects (in both utilities). The method used only those studies that contributed means and standard deviations to the data set. The latter were converted to standard errors. Where a group had been formed (as described above) by pooling values over different health states, average standard errors were used—guaranteed to be greater than the true standard errors in the presence of (unknown) correlations between states. For most studies it was not possible to estimate residual standard deviations accurately because the correlations between direct and indirect utilities were not available. Instead, a conservative approach was employed, leading to underestimates of the absolute magnitude of the standardised residuals. For a fitted line $y = a + bx$ the standardised residual for the point (x', y') with standard errors (s_x, s_y) was estimated as $(y' - a - bx') / (s_y + |b|s_x)$. The denominator of this expression is guaranteed to be greater than the true residual standard deviation. We applied χ^2 tests to the sums of squares of the estimated standardised residuals. The approximations described here mean that these tests are conservative—they understate the

Table 3 Verification search strings used and hits returned

Disorder	Search strings	All Hits	Contained empirical utility data	Comparisons within same population	Relevant hits not found with original search
Asthma	Utilit* Asthma Health	126	15	1	0
Diabetes	Utilit* Diabetes Health	437	58	1	0
Rheumatoid arthritis	Utilit* Rheumatoid Arthritis Health	156	32	5	0
Stroke	Utilit* Stroke Health	295	22	0	0

Fig 2 | Bland-Altman plot comparing direct and indirect utilities for average health states within independent groups of participants. Vertical axis represents half the difference in utilities; horizontal axis is the average of the two utilities. Broken lines enclose the feasible region for utilities constrained to lie between 0 and 1. The single point outside this region represents a state with a negative EQ-5D estimate.

discrepancy between the reported utility values and the fitted lines. Small P values indicate that such discrepancies cannot reasonably be attributed to variation between the respondents contributing to the same study.

RESULTS

The main search for studies that contained direct and indirect methods of comparison returned 32 studies. The Tufts database search did not yield any studies comparing direct and indirect utilities in the same population that had not been retrieved by the original search. The same was true for the disease specific search (table 3). Four of the 32 studies were about health states experienced by children. In three of these studies, parents had stood as proxies to provide preference values for their own children. The other study asked young children to rate their own health, but the investigators acknowledged methodological difficulties that could have compromised the results. The four studies of children were excluded from further consideration.

This left 28 studies for analysis covering a wide range of diseases (including rheumatic, neuropsychiatric, vascular, respiratory, gastrointestinal, renal, and infectious diseases) and encompassing data from 4688 respondents. Some studies reported utilities for several different states of health or disease. Altogether there were 83 instances in which direct and indirect methods could be compared. Sixty-eight of the 83 comparisons (from 25 studies) were based on current patients, and 15 (from three studies) were based on hypothetical patients (one group from the general public, two groups from the healthcare community).

The full list of papers, including those involving children, is given in appendix 1, with summary data from each paper in appendix 2 (see Web Extra material). A breakdown of the clinical topics investigated (and the number of disease states per topic) is given in table 4.

In some instances more than one method was used to determine utilities. For direct utilities, standard gamble alone was used in 23 of the 83 comparisons, time trade off alone in 26, and both in 34. The most popular indirect methods were EQ-5D (n=67) and HUI-3 (n=37), one or both of which figured in all but four of the 83 comparisons. Table 5 shows a breakdown.

The utility values reported by the individual studies are averages over samples of respondents, with sample sizes ranging from three to 1011 (median 62). Means and standard deviations of utility values were available for 26 studies (covering 72 of the 83 available comparisons), with the remainder reporting medians only. Table 6 summarises the distribution of the reported mean or median utility values, presented by type of respondent.

More detailed analyses were undertaken for current and hypothetical comparisons that reported EQ-5D, HUI-3, or both. Figure 2 shows the differences between direct and indirect methods for independent groups of participants.

In figure 2, all but nine of 86 points lie above the horizontal line, reflecting the tendency for direct

Table 4 | Clinical conditions contributing data to the study

Population and topic	Number of papers	Number of disease states within papers
Current patients		
Peripheral vascular disease	4	18
Hepatitis C	1	7
End stage renal failure (on dialysis)	1	4
Chronic renal disease	1	6
Gastrointestinal reflux	2	2
Age related macular degeneration	1	1
Osteoarthritis of hip	1	3
Arthritis (general)	1	1
Rheumatoid arthritis	3	9
Survivors of extreme prematurity	1	2
Motor neurone disease	1	4
Gout	1	1
Systemic sclerosis	1	1
Severe mental disease	1	1
Depression	1	1
Uncomplicated diabetes	1	1
Benign prostatic hyperplasia	1	1
Chronic obstructive pulmonary disease	1	1
Oesophageal cancer	1	4
Hypothetical patients		
Supportive care, non specific	1	6
Breast cancer	1	4
Rheumatoid arthritis	1	5
Total	28	83

Table 5 | Breakdown of indirect methods used

Numbers of indirect methods	EQ-5D (N=67)	SF-6D (N=13)	HUI-2 (N=5)	HUI-3 (N=37)	Numbers of comparisons
One (N=48)	✓	—	—	—	37
	—	✓	—	—	4
	—	—	—	✓	7
Two (N=31)	✓	—	—	✓	24
	✓	✓	—	—	5
	—	—	✓	✓	2
Three (N=4)	✓	✓	—	✓	1
	—	✓	✓	✓	3
Total	—	—	—	—	83

utilities to exceed indirect utilities. However, the spread of the differences diminishes for states approaching perfect health, a consequence of the fact that utility estimates cannot be greater than 1. The direct methods (time trade off, standard gamble) produced significantly higher values than the indirect methods (EQ-5D, HUI-3) in every case where a statistical comparison was feasible (table 7). By contrast, the differences in utility values between each of the direct methods and between each of the indirect methods were not statistically significant.

The discrepancy between individual direct and indirect measures is reflected in figure 3. These plots are constructed using information taken directly from the studies, without aggregation into independent groups. If direct and indirect methods gave the same results, then the points would be distributed equally above and below the 45° line in each panel. The great majority of points in all panels, however, fell above this line. In each panel, the broken line represents the predicted direct utility score from a regression on the indirect score, as computed from the “current patient” comparisons. Table 7 shows the slopes for these lines and those based on hypothetical comparisons. The lines represent average relations only, with statistically significant ($P<0.05$) departures from the line in all but one instance, which was based on very low sample numbers. This finding suggests that the variation between participants within studies was not sufficient to account for the discrepancies between the plotted points and the fitted lines. At best, these lines characterise “average” relations between direct and indirect methods across a collection of different health states; it

cannot be assumed that they will produce accurate conversions from one type of utility to another.

DISCUSSION

Principal findings

Our findings show that the versatile and convenient indirect methods of utility measurement yield different results from those obtained by the direct methods. The indirect methods yielded systematically lower utility values than direct methods for a wide range of diseases (we give two examples in the box). The differences in utility values between the direct and indirect groups of methods were sometimes substantial, and tended to be greater than the differences between the individual methods making up the group.

Limitations

As with all systematic reviews, the results are constrained by the published literature which, in this case, gave access to 83 comparisons across 28 studies. The studies covered many different diseases and several different methods of utility elicitation allowing a general trend to be identified—that indirect methods tend to yield lower utilities than direct methods. However, the heterogeneity of methods and disease states in the review precludes any generalisable numerical summary of effect size that would apply to a different spectrum of patients or methods.

Explaining the findings

This study was not designed to elucidate the reasons for the differences in results between indirect and direct methods. Nevertheless, there are several possible explanations that could contribute to an understanding of the results.

One potential explanation is that the generic questionnaires used to obtain the indirect utilities impose constraints: respondents are forced to encapsulate their potentially complex condition within five to eight categories, depending on the questionnaire used. The questionnaires do not allow respondents to report, for example, potentially positive aspects of their situations that would boost utility values. There is a risk, therefore, that people contributing tariffs for the various states form a different impression to that which those completing the questionnaire would have had.

Table 6 | Distribution of preference values across all disease states assessed by included studies

	Current patients*				Hypothetical patients				Overall mean (SD)
	Studies	States	Range	Mean (SD)	Studies	States	Range	Mean (SD)	
Time trade off	19	50	0.52, 0.99	0.83 (0.10)	2	10	−0.17, 0.77	0.45 (0.27)	0.77 (0.20)
Standard gamble	16	46	0.45, 1.00	0.83 (0.12)	2	11	0.47, 0.90	0.70 (0.14)	0.81 (0.14)
EQ-5D	19	52	−0.01, 0.93	0.65 (0.15)	3	15	−0.52, 0.81	0.35 (0.34)	0.59 (0.24)
SF-6D	6	13	0.55, 0.69	0.63 (0.04)	—	—	—	—	0.63 (0.04)
HUI-2	2	5	0.55, 0.95	0.75 (0.17)	—	—	—	—	0.75 (0.17)
HUI-3	10	37	0.34, 0.89	0.68 (0.11)	—	—	—	—	0.68 (0.11)

Each disease state is associated with just one study. The statistics in each row (mean, SD, range) refer to the sample of disease states.

*The majority of studies reported mean utility values across samples of patients. Two studies of current patients reported only median utilities. These have been included in the table nevertheless.

Table 7 | Pairwise comparisons between and within direct and indirect methods

Comparison	Number of states	Mean difference (SD)	Number of independent groups	Sign test	Regression slope (SE)	Predictive P value
Current patients						
Time trade off – EQ-5D	40	0.17 (0.13)	26	P<0.001	0.54 (0.07)	P<0.001
Standard gamble – EQ-5D	39	0.20 (0.15)	27	P<0.001	0.50 (0.05)	P<0.001
Time trade off – HUI-3	24	0.19 (0.09)	13	P=0.023	0.48 (0.06)	P<0.003
Standard gamble – HUI-3	29	0.15 (0.11)	16	P=0.004	0.65 (0.08)	P<0.002
Standard gamble – Time trade off	28	0.02 (0.06)	16	P=0.804	—	—
HUI-3 – EQ-5D	25	0.03 (0.13)	14	P=0.424	—	—
Hypothetical patients						
Time trade off – EQ-5D	10	0.13 (0.12)	2	NA	0.81 (0.01)	P<0.994
Standard gamble – EQ-5D	11	0.15 (0.11)	2	NA	0.50 (0.06)	P<0.042
Standard gamble – Time trade off	6	0.22 (0.05)	1	NA	—	—

NA=not applicable. Second column shows numbers of states contributing to mean and SD of the difference. Sign tests and regression analyses are based on aggregate health states obtained from averaging data within independent groups of participants. Regression slopes refer to least squares lines for predicting direct utility from indirect utility, constrained to pass through the point (1,1). Predictive P-values assess the goodness-of-fit of the lines, taking variation within studies into account.

Secondly, it is likely that the respondents who contribute trade off values for indirect utility elicitation differ systematically from the patients that participate in direct methods. The general population used to obtain tariffs for indirect utility estimation is spread across a wider age range than patient populations typically used in direct estimation. Young people in good health and people with diseases may have very different perspectives on the relative merits of remaining in a given health state and making a trade off involving death.

Those participating in elicitation of indirect tariffs are always asked to make a hypothetical choice whereas direct utilities are usually based on the experience of people who actually have the condition, who are aware not only of the distress of living with a condition, but also of mitigating factors. They may derive benefits from family, friends, work, and recreation on a daily basis and tend, over time, to adapt to their condition.²⁶ Members of the general public contributing tariffs for health states are at a considerable emotional distance—they lack personal experience of the disease, and the descriptions given in generic questionnaires are rather bland and decontextualised.²⁵

However, the difference between direct and indirect utilities seems to remain even when values are elicited from an (admittedly small) sample of hypothetical patients. This may be because the verbal “pictures” of disease used were somewhat richer when applying direct methods to hypothetical patients than when eliciting tariffs for the calibration of indirect methods.²⁵

Finally, modifications to the methodology for deriving indirect utilities from quality of life questionnaires have recently been suggested.²⁷ If implemented these would tend to increase the valuations attached to mild health states and eliminate some of the differences observed here.

Direct or indirect?

Simply observing a systematic and substantial difference between methods, as we have done, does not resolve the decision of which method to choose. There is, however, no universally accepted theoretical basis for choosing direct or indirect methods.¹ Some people think that utilities should be derived from patients (ideally at stages throughout their illness) on the grounds that only patients really know what the condition is like. Others think that the citizen’s perspective is more relevant. Various reasons have been given for this opinion—it is argued that citizens are the natural (and democratic) locus for a decision about use of society’s resources, that citizens can be asked to factor in societal objectives (such as the pursuit of equity), and that the necessarily hypothetical choices that citizens make conform to “the axioms of utility theory”, whereby decisions are made prospectively under uncertainty. However, a preference for direct or indirect utilities does not necessarily result from these considerations: direct utilities derived hypothetically from citizens and indirect utilities calibrated through a survey of the general population can both provide the advantages of the citizen’s perspective.

Given the absence of a gold standard for the elicitation of utilities, it may be informative to consider the

Examples in which indirect methods produce considerably lower utilities than direct methods

For the condition of intermittent claudication an indirect (EQ-5D) utility of 0.57 was elicited. This value implies that a person would forgo over 40% of their remaining life or run slightly over a 40% risk of death to avoid the condition. Yet the same respondents gave direct utilities of 0.82 (time trade off) and 0.91 (standard gamble) for the condition. So, when asked directly, they would only accept a much lower risk (of around 10%) to avoid the condition.

For patients with chronic obstructive pulmonary disease, respondents gave a utility of 0.91 for the direct method (time trade off), accepting a 9% reduction in lifespan to avoid the condition. The EQ-5D utility from the same patients was 0.73, suggesting a willingness to forgo a quarter of their remaining lifespan. Readers may wish to consider whether a preventative treatment with a mortality of 25% would really be acceptable for most patients destined to develop this disease.

In both these cases, the utilities were taken from current patients.

Fig 3 | Direct utilities against indirect utilities. Plotted points are means (if available) or medians from health-states within 28 studies. Vertical and horizontal lines represent standard errors cited (or deduced) within the studies. Broken lines are regressions of direct on indirect utilities from current patient comparisons. In top left panel, one (hypothetical) point lies off the scale, with EQ-5D=-0.52, time trade off=-0.17.

practical implications of using one method or the other.

Implications for resource allocation

Had we found little or no difference between direct and indirect methods, then the latter could be recommended on grounds of efficiency since, once the tariffs have been set, indirect utilities are far simpler to elicit.

WHAT IS KNOWN

Health state utilities play a crucial role in the allocation of healthcare resources

Utilities may be obtained directly (usually from patients) or, more often, indirectly, by using a quality of life questionnaire, the results of which are converted to utilities using “weights” (tariffs) obtained from the general public

Different direct and indirect methods yield different utility values

WHAT THIS PAPER ADDS

Indirect methods as a group produce consistently lower utilities (worse recorded health) than the direct group of methods

This difference may be larger than the differences between methods within each group

Reliance on indirect methods will result in less resources being allocated to life saving treatments than if direct methods were used

Conversion of indirect utilities to direct utilities is only partly successful

However, it seems that indirect methods give consistently lower levels of utility than direct methods. This means that there is more headroom for utility gain with indirect methods. The utility of death, however, is fixed at zero. Thus, in comparison with direct methods, indirect methods will favour the allocation of healthcare resources away from interventions that prevent or delay death in favour of those that alleviate non-fatal conditions. It could be argued that the popularity of indirect methods for informing rationing decisions simply expresses a legitimate societal attitude in favour of non-lethal over lethal conditions. However, it could equally be argued that the public would rather give more weight to delaying death. In that case indirect methods might risk undervaluing both personal and societal preferences. In England, the National Institute for Health and Clinical Excellence (NICE) has noted that “society places great value on extending the lives of people with life-threatening conditions” and has issued new guidance for “life-extending treatments that are above the usual range of cost effectiveness,”²⁸ in the wake of decisions that denied life extending treatments.

We are not the first investigators to suspect that many of the utilities used in health resource decisions seem too low, particularly at the upper end of the scale.²⁹ If this is so, then the data presented here show that the tendency applies especially to indirectly elicited utilities (see box).

Implications for decision makers

Those who prefer direct methods, but who wish to exploit the convenience of indirect methods, might propose using a simple correction to map indirect utilities onto the putatively more valid direct utilities. Our results show that linear adjustments of indirect utilities can achieve only a partial conversion to direct utility scores. They cannot be used to accurately predict direct utilities for particular health states. However, such an adjustment may be better than nothing, because the adjusted utilities are well calibrated on average. Further development in the science of utility estimation could make a contribution by identifying the characteristics of disease states for which the average adjustment is inappropriate.

This paper adds weight to the recommendation to be cautious when using utilities of any type. In the construction of health economic models it may be prudent to extend the range of uncertainty beyond the confines of statistical confidence limits and conduct a sensitivity analysis. In that case, the results presented here suggest that indirect utilities might best be used to populate only the lower limits of such an analysis.

Because direct and indirect methods can lead to such noticeable differences in elicited utilities, priority setting institutions should avoid using a mixture of methods for different decisions, otherwise a motivated choice of method might be used to distort the outcome in a preferred direction. NICE has certainly had to deal with inconsistent use of methods.³⁰ It may be that a general counsel against the mixing of different methods is the most important implication of this work.

We thank Stirling Bryan and Tracey Roberts for helpful discussion. We also thank the referees Steven McPhail and Cindy Lam for their detailed and perceptive comments.

Contributors: RJL conceived the idea for the paper, DA performed the literature reviews and data extraction, and AG repeated the data extraction and carried out statistical analyses. All authors prepared the manuscript and RL is guarantor.

Funding: This work was supported through the MATCH programme (EPSRC Grant GR/S29874/01) although the views expressed are entirely the authors' own.

Competing interests: None declared.

Ethical approval: Not required.

- 1 Gold MR, Siegel JE, Russell LB, Weinstein MC. Cost-effectiveness in health and medicine. Oxford: Oxford University Press, 1996.
- 2 Stiggelbout AM, Eijkemans MJ, Kiebert GM, Kievit J, Leer JW, De Haes HJ. The 'utility' of the visual analog scale in medical decision making and technology assessment. Is it an alternative to the time trade-off? *Int J Technol Assess Health Care* 1996;12:291-8.
- 3 Torrance GW, Feeny D, Furlong W. Visual analog scales: do they have a role in the measurement of preferences for health states? *Med Decis Making* 2001;21:329-34.
- 4 Torrance GW. Measurement of health state utilities for economic appraisal. *J Health Econ* 1986;5:1-30.
- 5 Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. *J Health Econ* 2002;21:271-92.
- 6 Horsman J, Furlong W, Feeny D, Torrance G. The Health Utilities Index (HUI): concepts, measurement properties and applications. *Health Qual Life Outcomes* 2003;1:54.
- 7 EuroQol group. EuroQol—a new facility for the measurement of health-related quality of life. *Health Policy* 1990;16:199-208.
- 8 Walters SJ, Brazier JE. Comparison of the minimally important difference for two health state utility measures: EQ-5D and SF-6D. *Qual Life Res* 2005;14:1523-32.
- 9 Boonen A, van der HD, Landewe R, van TA, Mielants H, Dougados M, et al. How do the EQ-5D, SF-6D and the well-being rating scale compare in patients with ankylosing spondylitis? *Ann Rheum Dis* 2007 Jun;66:771-7.
- 10 Macran S, Weatherly H, Kind P. Measuring population health: a comparison of three generic health status measures. *Med Care* 2003;41:218-31.
- 11 Grutters JP, Joore MA, van der HF, Verschuure H, Dreschler WA, Anteunis LJ. Choosing between measures: comparison of EQ-5D, HUI2 and HUI3 in persons with hearing complaints. *Qual Life Res* 2007;16:1439-49.
- 12 van Stel HF, Buskens E. Comparison of the SF-6D and the EQ-5D in patients with coronary heart disease. *Health Qual Life Outcomes* 2006;4:20.
- 13 Brazier J, Roberts J, Tsuchiya A, Busschbach J. A comparison of the EQ-5D and SF-6D across seven patient groups. *Health Econ* 2004;13:873-84.
- 14 Bryan S, Longworth L. Measuring health-related utility: why the disparity between EQ-5D and SF-6D? *Eur J Health Econ* 2005;6:253-60.
- 15 Robinson A, Dolan P, Williams A. Valuing health status using VAS and TTO: what lies behind the numbers? *Soc Sci Med* 1997;45:1289-97.
- 16 Lamers LM, Stalmeier PF, Krabbe PF, Busschbach JJ. Inconsistencies in TTO and VAS values for EQ-5D health states. *Med Decis Making* 2006;26:173-81.
- 17 Badia X, Monserrat S, Roset M, Herdman M. Feasibility, validity and test-retest reliability of scaling methods for health states: the visual analogue scale and the time trade-off. *Qual Life Res* 1999;8:303-10.
- 18 Bleichrodt H, Johannesson M. Standard gamble, time trade-off and rating scale: experimental results on the ranking properties of QALYs. *J Health Econ* 1997;16:155-75.
- 19 Puhans MA, Schunemann HJ, Wong E, Griffith L, Guyatt GH. The standard gamble showed better construct validity than the time trade-off. *J Clin Epidemiol* 2007;60:1029-33.
- 20 Martin AJ, Glasziou PP, Simes RJ, Lumley T. A comparison of standard gamble, time trade-off, and adjusted time trade-off scores. *Int J Technol Assess Health Care* 2000;16:137-47.
- 21 Bleichrodt H. A new explanation for the difference between time trade-off utilities and standard gamble utilities. *Health Econ* 2002;11:447-56.
- 22 Murphy R, Sackley CM, Miller P, Harwood RH. Effect of experience of severe stroke on subjective valuations of quality of life after stroke. *J Neurol Neurosurg Psychiatry* 2001;70:679-81.
- 23 Post PN, Stiggelbout AM, Wakker PP. The utility of health states after stroke: a systematic review of the literature. *Stroke* 2001;32:1425-9.
- 24 Tufts Medical Center. Center for the evaluation of value and risk in health. 2008. <https://research.tufts-nemc.org/cear/search/search.aspx>.
- 25 Brazier J, Deverill M, Green C, Harper R, Booth A. A review of the use of health status measures in economics evaluation. *Health Technol Assess* 1999;3:1-164.
- 26 Schwartz CE, Andresen EM, Nosek MA, Krahn GL. Response shift theory: important implications for measuring quality of life in people with disability. *Arch Phys Med Rehabil* 2007;88:529-36.
- 27 Craig B, Busschbach J. The episodic random utility model unifies time trade-off and discrete choice approaches in health state valuation. *Population Health Metrics* 2009;7:3.
- 28 National Institute for Health and Clinical Excellence. The NICE bulletin for parliamentarians. Issue 55. November 2008: www.nice.org.uk.
- 29 Nord E. Cost-value analysis in health care. Cambridge: Cambridge University Press, 1999.
- 30 Stein K, Fry A, Round A, Milne R, Brazier J. What value health? A review of health state values used in early technology assessments for NICE. *Appl Health Econ Health Policy* 2005;4:219-28.

Accepted: 18 February 2009