For debate
Putting the quality into quality-adjusted life years
Adam Oliver

Summary
Over recent decades, a great deal of effort has been devoted towards developing instruments that can be used to elicit health state values. All of these instruments are conceptually very different from one another and all suffer from serious inherent biases. In this paper I outline the conceptual foundations and empirical limitations of the three principal health state value elicitation instruments. Given that the conceptual parameters internalized within an instrument influence the elicited health state values, I argue that it is necessary to attain a broad agreement on what the appropriate parameters ought to be. When the appropriate conceptual parameters have been identified we will be in a position to develop the methodology of a single standardized instrument.

Keywords: QALYs, rating scale, time trade-off, standard gamble

Introduction
In every country in the world, there is pressure on governments to provide those services that ‘society’ values but that would not be provided, or would not provided equitably, if their provision were left entirely to the private market. These services are many and varied and include, for example, education, defence, housing, law and order and, of course, health care. If there were unlimited available public sector funds, all of the demands on all public sector services would be satiated. Unfortunately, however, public sector funds are not, and can never be, unlimited, and a responsible government ought to strive towards allocating the available resources across the different services in a manner that is consistent with generating social welfare and improving social justice. Therefore, each of the services will be allocated insufficient resources to meet all the demands that are placed on them. In short, the resources available to each service – including health care – are, and always will be, scarce.

The scarcity of resources within each of the public sector services inevitably leads to the need to ‘ration’ the interventions, programmes and policies that are potentially available within those services. When the demand for a service exceeds its resource capacity, rationing, even if undertaken on a purely ad hoc basis, will occur. The introduction of some coherent, structured rationing process is synonymous with an effort to ensure that scarce resources are used in the most socially advantageous ways. To deny that rationing is inevitable is naive and, from the perspective of society, potentially very damaging.

Within health care, there are many mechanisms through which structured rationing takes place. For example, the general practitioner’s role as a gatekeeper for referrals to secondary care, waiting lists and patient copayments are all means by which to ration or prioritize health care. One such method that is becoming increasingly important at the policy-making level in a number of countries is economic evaluation, the results of which are used to help inform guidance on which health care interventions ought to be provided within a publicly provided health service.

Economic evaluation in health care is concerned with analysing an intervention for its value for money, which requires an assessment of both the costs of the intervention and the benefits that it offers. One method of economic evaluation is cost–utility analysis (CUA). The measure of health care benefit that is used in CUA represents an attempt at combining length of life with quality of life into a single index, the most well-known outcome measure of which is the quality-adjusted life year, or ‘QALY’. For example, if one year with severe angina is assigned a ‘quality weight’ of 0.5, then 20 years with severe angina will produce 10 QALYs (assuming zero discounting over life years). The merit of this approach is that, in theory, it allows us to compare the health benefits enjoyed by all patient groups in receipt of differing health care interventions.

To compare the total number of QALYs generated by each health care intervention we would need to sum the QALYs
accrued over different states of health. If this summation is to be meaningful, an important requirement for the quality weights (or ‘health state values’) is that they are ‘cardinal’ measures of value; that is, the value that an individual assigns to any particular health state should accurately reflect the individual’s relative strength of preference for that health state. For example, if an individual considers paraplegia to be half as good as having severe angina (assuming an equal length of time – for example, one year – in each health state), then the value that is elicited for the former health state ought to be half that of the latter. Over the past three decades, economists and psychologists have devoted a great deal of effort towards developing instruments that can be used to elicit cardinal health state values.

A number of elicitation instruments have been developed, all of which are conceptually very different from one another and all of which suffer from serious biases. In applied cost–utility analyses, however, the conceptual differences between, and the biases inherent within, the various health state value elicitation instruments are rarely (if ever) noted. Thus, such analyses are likely to generate biased, study-specific findings, seriously undermining their usefulness for comparing the relative effectiveness (and cost-effectiveness) of different health care interventions.

In this paper I will outline the conceptual foundations and empirical limitations of the three principal health state value elicitation instruments: the rating scale, the time trade-off and the standard gamble. I will then discuss some steps that need to be taken before we can have any confidence in using cost–utility analyses to inform decisions at the policy-making level.

The health state value elicitation instruments

The rating scale

The rating scale was developed by psychologists and requires the respondent to place health state descriptions on a scale similar to that illustrated in the Figure. The best and worst conceivable health states (e.g. full health and immediate death) are respectively marked at the top and bottom of the scale. The top and bottom of the scale are typically assigned values of one and zero (or 100 and 0). When the rating scale is used, the respondent should be informed that the spaces they leave between the points on the scale that mark their valuations of the intermediate health states ought to reflect their relative strength of preference for each health state. For example, if a respondent considers severe angina to be half as good as full health, and paraplegia to be half as good as severe angina, then they should respectively mark severe angina and paraplegia at 0.5 and 0.25 on the scale (assuming an equal length of time in each health state). These ratings consequently represent the rating scale values.

Unfortunately, the rating scale is subject to context and end aversion biases. Context bias refers to the observation that the rating scale value for any particular health state depends on the number of better or worse health states that the respondent is also asked to consider. That is, if a health state is presented together with many better health states its value tends to be depressed, and if it is presented with many worse health states, its value tends to be enhanced. End-aversion bias refers to the observation that people show a reluctance to use the parts of the scale that lie close to the ends, and thus bunch their ratings towards the centre.

Notwithstanding these serious empirically observed biasing effects, the rating scale is conceptually limited in that it does not internalize any notion of choice. It merely involves placing health states on a line. In real-world settings, when the individual (whether they are a policy-maker, a doctor or a patient) assesses the potential outcomes that health care interventions offer, they are invariably required to choose (if only between treatment and no treatment). The very act of choosing may influence the values that the individual places on health outcomes because when faced with choices they are implicitly required to think about the trade-offs (i.e. the opportunity costs) of realizing one outcome rather than another. A theoretically sound value elicitation instrument will allow for the possibility that ‘choice’ influences ‘value’.

The time trade-off

The time trade-off was developed in the early 1970s and, unlike the rating scale, does internalize the notion of choice. Application of the time trade-off involves presenting the respondent with two options. In the first option the respondent is asked to imagine that they will be in an intermediate health state (for example, paraplegia) for a fixed period of time (for example, 5 years) and then they will die. In the second option the respondent is asked to imagine that they will be in full health for a shorter period of time (for example, 4 years) and then they will die. The respondent is asked to choose between the two options.

Figure The rating scale.
If, in the context of the above example, the respondent prefers the 4 years in full health, the exercise is repeated with less time in full health on offer in the second option. The exercise is continually repeated, varying the time available in full health, until we reach a point where the respondent is indifferent between the two options. At this point, the time trade-off value for paraplegia is calculated by dividing the time in full health in the second option by the time with paraplegia in the first option. For example, if the respondent is indifferent between 5 years with paraplegia and 1 year in full health, then that respondent’s time trade-off value for paraplegia is \(1/5 = 0.2\).

An empirical problem with the time trade-off is that the preference scores elicited with this method will be biased if the respondent discounts future life years; i.e. if the respondent values more distant years of life differently from more proximate years.\(^{12-14}\) It is well documented that people generally discount life years – both positive and negative discount rates have been observed.\(^{15-17}\) Positive discount rates are most commonly reported, and imply that the respondent will place a higher value on a more immediate year of life than on a more distant year of life; e.g. the respondent will value the next year more than a year 10 years from now. That is, they have a positive rate of ‘time preference’.

To demonstrate how a positive rate of time preference biases the time trade-off values, let us assume that the first option offered to the respondent is 10 years with severe angina followed by death. Further we assume that the respondent specifies that they are indifferent between 5 years in full health (i.e. the second option) and the first option. Their time trade-off value for severe angina is thus 0.5. If the respondent has a zero rate of time preference, we would also expect them to be indifferent between 20 years with severe angina and 10 years in full health. However, if the respondent has a positive rate of time preference, their marginal utility of an extra 10 years with severe angina (i.e. the difference between 10 and 20 years) will be less than their marginal utility of an extra 5 years in full health (i.e. the difference between 5 and 10 years). This is because their marginal utility of more distant years diminishes disproportionately more than their marginal utility of more proximate years, and implies that the respondent would require less than 10 years in full health to be indifferent to 20 years with severe angina. When 20 years with severe angina is offered as the first option, the respondent’s time trade-off value for severe angina would therefore be less than 0.5, the implication being that when positive rates of time preference are observed, time trade-off values are downwardly biased. In short, the time trade-off values of health states are biased by a combination of the lengths of time incorporated in the method and the respondent’s time preference rate, and thus many different values for a single health state could be elicited from a single respondent, none of which would represent the respondent’s ‘true’ value.

Moreover, on a conceptual level, health care decisions invariably do not involve choice only: they also involve a degree of risk. If we take a medication, there is always a chance that it will do us more harm than good. Similarly, if we undergo an operation, there is always a chance that we will never again leave the operating theatre alive. Although the time trade-off method internalizes the notion of choice, it does not internalize attitudes towards risk; i.e. the respondent is asked to think about the choice between two options, but both, if they were to occur, occur with certainty. However, individuals may enjoy risk, or, more probably, they may have a fundamental dislike of risk, which in turn may influence the value they place on health states. For example, let us assume that an individual is involved in an accident that leaves them with a limp. The individual is told that they can have an operation that may return them to full mobility, but that there is a risk inherent in the operation and they could end up in a wheelchair as a direct consequence of surgery. It seems plausible that the risk will influence the value that the individual places on their current health state (i.e. their limp) and that they would want this influence to be taken into account in any assessment of the relative benefits of surgery versus no surgery. If this is the case, then excluding risk attitude from the elicitation instrument would remove a potentially important influence on the valuation of health care outcomes in real-world circumstances.

The standard gamble

The standard gamble is implied from the assumptions of the dominant economic theory of risk (expected utility theory), and for this reason, is considered by some health economists to be the gold standard for eliciting cardinal health state values.\(^{18,19}\)

As with the time trade-off, application of the standard gamble involves presenting the respondent with two options. Once again using paraplegia as an example of an intermediate health state, the respondent is offered the hypothetical choice between paraplegia for their remaining life expectancy and a risky treatment option that offers a probability, \(p\) (for example, \(p = 0.1\)), of living in full health for their remaining life expectancy otherwise immediate death (hence, in this case, it is assumed that if the respondent undergoes treatment, there is a 0.9 chance that the treatment will immediately kill them). If the respondent prefers living with paraplegia over a treatment option that offers only a 0.1 chance of success, the probability, \(p\), is increased until we reach a point where the respondent is indifferent between remaining paralysed and undergoing the treatment. The standard gamble value for paraplegia is given by the probability at the point of indifference. For example, if the respondent is indifferent between remaining paralysed and a treatment that offers a 0.6 chance of returning to full health otherwise immediate death, the respondent’s standard gamble value for paraplegia is 0.6. The formula used to calculate this standard gamble value is given in the Appendix.

Because the treatment option in the standard gamble is risky, this instrument will internalize the respondent’s attitude towards risk. To demonstrate how this can influence the value elicited for an outcome, let us imagine that a respondent is indifferent between 10 years with severe angina and 5 years of...
healthy life, both lived for certain (i.e. a time trade-off scenario with no risk). The value of severe angina, calculated by the time trade-off method, is 0.5. Now let us assume that under a standard gamble scenario, the same respondent is presented with the choice between severe angina for their remaining life expectancy and a treatment that offers a chance of full health for their remaining life expectancy otherwise immediate death. The respondent is asked for the probability of treatment success that they would require for them to be indifferent between the certainty of severe angina and the treatment. In the evaluation of most intermediate health states, we would expect the indifference probability to be rather high; indeed, considerably higher than 0.5. This is because most respondents will be highly averse to the risk inherent in the treatment (which, in this case, is a risk of immediate death). Thus, for risk-averse respondents, the standard gamble is likely to generate higher values than the time trade-off, and this has indeed been empirically observed.20

Even if we assume that an important conceptual requirement of a health state value elicitation instrument is the internalization of risk attitudes (which, admittedly, is a far from universally accepted assumption21), the standard gamble possesses serious internal biases. At least part of the source of these biases is that individuals significantly and systematically violate the assumptions of expected utility theory, the very assumptions from which the standard gamble is implied.22 The most controversial of these assumptions is known as the ‘independence axiom’. In the context of the standard gamble, the independence axiom implies that the extent to which the respondent values any particular health outcome will not be influenced by changing the health outcomes against which that outcome is being compared. For example, if we were to compare severe angina against a treatment that could return the respondent to full health but could also cause immediate death, the extent to which the respondent intrinsically values severe angina should not be influenced by changing either the success or the failure outcomes of treatment. If the extent to which the respondent values severe angina is influenced by altering the other outcomes, we would potentially be able to elicit multiple standard gamble values for that health state, none of which would necessarily reflect the respondent’s ‘true’ underlying value. Indeed, several studies have demonstrated that the standard gamble values for particular health states are significantly and systematically influenced by altering the outcomes in the treatment option,23–25 seriously undermining the validity of this method.

The next steps

The three most commonly used instruments for eliciting health state values (which are used as the basis for calculating QALYs) – the rating scale, the time trade-off and the standard gamble – are all subject to serious inherent biases. On empirical grounds alone, it is not yet appropriate to use any of these instruments to gather data for informing the decision-maker as to which health care interventions ought to be publicly funded (and which ought not). Far more methodological development of these elicitation instruments is required before they can be confidently used for practical purposes.

It is possible that the empirical biases could, in time, be controlled for or removed. However, a more pressing issue is to attain broad agreement on the appropriate conceptual parameters to internalize within the elicitation instrument. Is it appropriate to avoid consideration of choice and risk (as in the rating scale), to consider choice only (as in the time trade-off) or to consider both choice and risk (as in the standard gamble)? Going beyond the three instruments discussed in this paper, others argue that we ought also to internalize consideration of the number of people who will be targeted by a particular health care intervention, and the severity of their illnesses.26 There are likely to be numerous other potential considerations, some of which have probably not yet been thought of, but given that the conceptual parameters of an instrument will influence the elicited health state values, it is crucial that a broadly accepted conceptual framework is agreed upon. Research involving economists, psychologists, medical practitioners, public health specialists, philosophers, policy-makers, patients, the general public and a host of other stakeholders is required to reach this necessary agreement.

Once we have decided upon the appropriate conceptual framework, we will be in a position to develop our elicitation instrument and remove, as far as possible, any biases that are inherent within it. If CUA is to have a role in rationing health care, decision-making bodies throughout the world have an important role to play in encouraging the development of a standardized value elicitation instrument that is free of (or at least is less restricted by) the serious biases that fundamentally undermine the validity of the existing, prevailing instruments. Ultimately, the standardization and recommendation of a single instrument is necessary if we are to realize our aim of accurately comparing the relative benefits accruing to different patient groups across different health care interventions.

References

Appendix

In the hypothetical example given in the main text the respondent is assumed to be indifferent between remaining paralysed and a treatment that offers a 0.6 chance of returning to full health otherwise immediate death. The standard gamble value for paraplegia is therefore given by

\[ v(\text{paraplegia}) = 0.6v(\text{full health}) + 0.4v(\text{immediate death}) \]

where \( v(\cdot) \) is the standard gamble value function. Full health and immediate death are usually assumed to be the best and worst conceivable health states, respectively, and therefore \( v(\text{full health}) \) is set equal to one and \( v(\text{immediate death}) \) is set equal to zero. Hence

\[ v(\text{paraplegia}) = 0.6. \]