Multi-method approach to valuing health states: problems with meaning

Erik Norda,*, Paul Menzelb and Jeff Richardsonc
a Norwegian Institute of Public Health, Oslo, Norway
b Pacific Lutheran University, Tacoma, USA
c Center for Health Program Evaluation, Melbourne, Australia

Summary

In an earlier article in Health Economics, Salomon and Murray argue that by applying maximum likelihood techniques to predetermined functional forms and to a data set where a number of health states are valued by means of four standard valuation techniques, underlying ‘pure’ valuations of health may be teased out, together with estimates of parametric relationships between these ‘pure’ valuations and valuations based on the four standard techniques. We argue below that ‘pure’ valuations of health are ordinal rather than cardinal and that the ‘pure’ values that result from the multi-method approach give a false impression of being cardinal. They are therefore not usable as weights for life years. In the unlikely event that the authors should be able to demonstrate cardinality in ‘pure’ valuations of health, it must be possible to have subjects express these valuations directly, in which case there seems to be no need for the indirect multi-method approach. Copyright © 2006 John Wiley & Sons, Ltd.

Keywords health state values; utility; DALY; QALY; multi-method approach

Introduction

Health state valuation techniques like the standard gamble, time trade-off, person trade-off and the rating scale typically yield quite different values for the same states. The reason is that with each technique, subjects’ valuations are affected not only by their feelings about and judgements of a health state per se, but also by characteristics of the technique. For instance, the standard gamble, time trade-off and person trade-off invite considerations of risk, time preference and fair distribution respectively, while the rating scale has end point anchoring effects and does not ask for trade-offs at all. Salomon and Murray [1] argue that by applying maximum likelihood techniques to predetermined functional forms and to a data set where a number of health states are valued by means of all four valuation techniques, underlying ‘pure’ valuations of health may be teased out, together with estimates of parametric relationships between these ‘pure’ valuations and valuations based on the four standard techniques.

The authors state that in developing their approach, they were influenced by our arguments concerning the need for preference weights that are clearly meaningful and which pass the test of ‘reflective equilibrium’. While we admire the technical ingenuity of their solution, we fail to see that it satisfies the requirements we have suggested.

*Correspondence to: Norwegian Institute of Public Health, P.O. Box 4404, Nydalen, N-0403 Oslo, Norway.
E-mail: erik.nord@fhi.no
Unjustified move from ordinal to cardinal level

We agree with Salomon and Murray that when subjects are asked to value health states, it is reasonable to assume that they first make some kind of ‘raw’ judgements of the relative badness (undesirability, unhealthiness) of the states in question. We also accept that these prior raw judgements are free of some of the specific mental operations involved in the standard gamble, the time trade-off, the person trade-off and the visual analogue scale. But we need to consider the scaling level of the raw judgements. Broadly speaking, there are two possibilities. One is that the raw judgements are vague, based largely on feeling and have an ordinal property in relation to the value of other health states. This only allows statements such as ‘I think A is worse than B’ or ‘I think A is much worse than B’. The other possibility is that the raw judgements consist of precise, cognitive operations that result in judgements at a cardinal level. This allows statements such as ‘I think A reduces well-being twice as much as B’. Without discussing which of these two possibilities is likely to be the case Salomon and Murray produce estimates of the raw judgements that they claim have cardinal properties. (They must be claiming this, since the numbers are supposed to be used as weights for life years.) But the claim can only be true if the initial judgements made by the subjects were at a cardinal level. (The statistical treatment of the data cannot raise the scaling level of these judgements from ordinal to cardinal.)

We therefore think Salomon and Murray, before proceeding further, need to discuss the likelihood of the initial raw judgements being cardinal. If they are not, we think their approach constructs a variable that does not exist in the real world.

Our own view is that raw judgements of health states in terms of ‘badness’ or ‘undesirability’ or ‘unhealthiness’ with few exceptions are at an ordinal level of scaling. Our belief is, in part, based on introspection which, in this case, seems quite helpful. However, we also observe everyday language in others. People use words like ‘bad’, ‘not so bad’, ‘very bad’, etc. to describe degrees of badness of health. They never support these clearly ordinal, verbal descriptions with numbers. This contrasts with cases where there is a clear underlying cardinal property. People use ordinal phrases such as ‘near’ and ‘far’; ‘light’ and ‘heavy’. But they also use the cardinal terms ‘kilometres’ and ‘kilograms’ which indicates that they are aware of and understand the underlying cardinal property.

This latter issue is not new. Rosser and Kind [2] asked subjects to make pairwise comparisons of a number of health states using the question ‘How many times more ill is a person in state X as compared with state Y?’ To help the subjects cope with this unusual question, the authors asked the subjects to think in terms of equivalence of numbers of people treated for X and Y, respectively. In other words, a person trade-off theme was introduced in order to make the initial question meaningful at a cardinal level. Similarly, in two studies of subjects who valued health states on the EuroQol visual analogue scale, a large majority of the subjects could not, when asked directly ex post, give any cardinal interpretation of their responses [3,4].

At this point one might object that if one accepts that responses to the SG, TTO and PTO have cardinal properties, and the SG, TTO and PTO are supposed to reflect judgements of the badness of health states, then one must also be assuming that the underlying (raw) judgements of health are at a cardinal level. But in our view this would be a false inference, for a number of reasons.

First, for sure it is possible to obtain responses to the SG, TTO and PTO irrespective of the measurement level of the prior (raw) judgements. Given this, one cannot infer from the fact that SG, TTO and PTO responses have cardinal meaning (which they clearly have, but each in a different way) that the prior raw judgements must also have cardinal meaning.

Second, the SG, TTO and PTO all force choices between competing scenarios described in quantitative terms. By forcing these concrete choices, more cognitive effort is demanded from the subjects compared to when they make raw internal judgements. So it is not surprising if the scaling level of the subjects’ responses is raised when they move from raw judgements to explicit trade-offs.

Third, a cardinal property can be derived from an underlying ordinal property as illustrated in the relationship between ordinal utility measurement of the value of products and the cardinal measurement of the willingness to pay for these same products. Similarly, ‘pleasurable feelings’ have no independent meaningful, and cardinal property. Only the expression of these feelings has the property. For example, if the pleasurable feelings related to ownership of a new car a
A cardinal metric could be obtained using willingness to pay but this would not necessarily imply a cardinal scale relating to feelings.

Fourth, and following from this, a cardinal metric depends, not just upon the underlying property being measured but upon the measurement scale and its peculiarities. In the example above, use of the willingness to pay 'shoehorns' 'pleasurable feelings' into a metric. However, the willingness to pay also depends upon the availability and relative price of other items which might be purchased and the resulting trade-off between obtaining pleasurable feelings and these other items. But these other considerations introduce something new. They are irrelevant for the initial feelings of pleasure. Similarly, the expression of a preference for health which can be easily understood draws upon other elements which may be irrelevant for the underlying feeling. Salomon and Murray are aware of this and claim they standardise for these extraneous elements. But our point here is that these elements which allow the SG, TTO and PTO to have a cardinal property destroy the implied logical link between measurement of the SG, TTO and PTO at the cardinal level and the existence of a cardinal property in the underlying health state.

Fifth, the SG, TTO and PTO all introduce well understood metrics (probability, time, persons) that offer themselves as meaningful vehicles for making cardinal judgements. Thus, there is clear meaning to statements like ‘a probability of 0.4 is double the probability of 0.2’; ‘6 years (in full health) is double the duration of 3 years’; ‘80 people cured is double 40 persons cured’. A similar meaningful metric does not seem to be present in the prior phase of raw judgement.

Altogether we maintain that there is a prima facie case against Salomon’s and Murray’s implicit assumption that raw judgements of health states are at a cardinal level. Until Salomon and Murray have addressed this issue and brought evidence to support their counter-intuitive assumption, it is difficult to perceive their multi-method approach as meaningful.

If meaningful, why not ask directly?

But even if they should be able to bring the evidence we request, we question the usefulness of their approach. The reason is that raw judgements can only be shown to have cardinal meaning by having subjects express them verbally in understandable, cardinal terms. As noted above, we doubt that this is going to happen. But in the event that such verbal expression turned out to happen and be convincing, it must follow that one should be able to elicit the cardinal judgements in direct questions, instead of estimating them by means of a complicated, indirect multi-method procedure. For instance, assume that states A and B score 0.6 and 0.8, respectively, on the estimated latent variable. Salomon and Murray may wish to attribute the following meaning to these numbers: ‘The subjects think going from perfect health to A is twice as large a change in health as going from perfect health to B.’ (Salomon, personal communication.) The only way to verify this interpretation is to ask the subjects if they agree with it. This would be a test of reflective equilibrium. But why would Salomon and Murray then not ask the subjects the following question to begin with: ‘How many times larger is a change in health from perfect health to A compared to a change from perfect health to B?’ If this magnitude estimation question is meaningful and covers what the latent variable is about, then there is no reason to beat around the bush. The multi-method approach would then be an awkward proxy for a direct magnitude estimation question, and the interesting statistical analysis would only be to examine the bivariate relationship between responses to the magnitude estimation question and responses to each of SG, TTO, PTO and VAS (for purposes of prediction when in some situations one elicitation technique has been used and one is interested in knowing what the responses to other techniques might have been).

Concluding remarks

There are further problems with Salomon’s and Murry’s paper. First, they present intraclass coefficients of more than 0.98 between the predicted responses and the mean observed responses, without discussing the fact that 16 variables were used to explain variation across 48 observations. In such circumstances, presumably a number of quite different models would perform at more or less the same level. No criterion is offered for choosing between them. Second, for time trade-off, a negative time preference is estimated.
Many will presumably take this as a possible sign of weakness in the approach itself and would expect a careful, critical discussion of this counter-intuitive finding rather than an extremely brief ad hoc defence in terms of selected references. Third, the authors find that ‘the respondents have preferences consistent with strong distributional concerns’ (pp. 284–285) and they show that values that take these concerns into account, are very different from ‘pure health’ values (Figure 2). It would have been useful if the authors had commented on the serious problems this raises for the use of ‘pure health’ values – if they were possible – in aiding fair societal health policy decisions.

But these are side points. Our main purpose is, as earlier, to defend meaning in health state valuations [5–7]. Unless policy makers and others fully comprehend and appreciate the meaning of the scales which measure quality of life then their ability to make appropriate judgements and trade-offs will be impaired. We are of course happy with the authors’ interest in our call for ‘measurement methods that enable meaningful statements about strength of preference’. But in our view, it is precisely this ideal that their proposed procedure does not live up to. We think judgements of value can obtain cardinal meaning when expressed as trade-offs in real, meaningful contexts. We fail to see how choice free perceptions and judgements of an abstract phenomenon like health can be understood as cardinal and treated as such in weighting life years. We thus primarily believe that Salomon and Murray are postulating a type of cardinal strength of preference that does not exist in the real world [8]. Alternatively, should it exist, it must be possible to ask directly about it, in which case there seems to be no need for a complicated multi-method approach.

Our recommendation for valuing health states is therefore to stick to meaningful trade-off techniques such as the standard gamble, the time trade-off and the person trade-off. The former two may be useful in establishing the undesirability (disutility) of different health problems from individuals’ personal perspective (ex ante or ex post). The latter is useful for eliciting judgements of fairness in resource allocation across patient groups. We have elsewhere described how these techniques may be combined in a transparent and meaningful two stage procedure for evaluating health programs [9].

References