Centre de Referència en Economia Analítica

Barcelona Economics Working Paper Series

Working Paper nº 116

Measuring the Health of Populations: The Veil of Ignorance Approach

José-Luis Pinto-Prades and José-María Abellán-Perpiñán

March, 2004

MEASURING THE HEALTH OF POPULATIONS: THE VEIL OF IGNORANCE APPROACH

José-Luis Pinto-Prades*
Center for Health Economics and Department of Economics
Pompeu Fabra University

José-María Abellán-Perpiñán
Department of Applied Economics
University of Murcia

Barcelona Economics WP nº 116

*Correspondence

Department of Economics
Pompeu Fabra University
Ramon Trias Fargas 25-27
08005 Barcelona
Spain
Email:jose.pinto@econ.upf.es
Tel: 3493522638
Fax:392521746

Acknowledgements: this paper has been funded through a research grant by Pfizer Ltd.

Abstract

We report the results from two surveys designed to explore whether an application of Harsanyi's principle of choice form behind a veil of ignorance (VEI) can be used in order to measure the health of populations. This approach was tentatively recommended by Murray et al. [1, 2] as an appropriate way of constructing Summary Measures of Population Health (SMPH) for comparative purposes. The operationalization of the VEI approach used in this paper was suggested by Nord [3]. We test if VEI and person trade-off (PTO) methods generate similar *quality-of-life* weights. In addition, we compare VEI and PTO weights with individual utilities estimated by means of the conventional standard gamble (SG) and a variation of it we call Double Gamble. Finally, psychometric properties like feasibility, reliability, and consistency are examined. Our main findings are next: (1) VEI and PTO approaches generate very different weights; (2) it seems that differences between PTO and VEI are not due to the 'Rule of Rescue'; (3) the VEI resembled more a DG than a classical SG; (4) PTO, VEI, and DG exhibited good feasibility, reliability and consistency.

Key words: population health, burden of disease, preferences, veil of ignorance, DALYs.

1. Introduction

The Global Burden of Disease (GBD) Project, funded by the World Health Organization (WHO), tries to quantify the global burden of diseases, injuries, and risk factors on human population. In order to estimate the disease burden, the WHOdeveloped a summary measure of population health known as Disability-Adjusted Life-Year (DALY). The DALY estimates the health gap of a population in the form of lost years of healthy life due to either disability/morbidity or premature death.

Disability weights were originally calculated using Rating Scale (RS) valuations [4]. However, the valuation protocol was completely changed in 1995, and the RS was replaced by the Person Trade-Off (PTO) method. Apparently the reason had to do with the idea that disability weights should encapsulate not only individual utilities but also a broader set of societal values. In such a way DALYs would be a better instrument to inform resource allocation decisions [5]. As it is claimed that the PTO can encapsulate social preferences [6] it was the method finally chosen. .

In recent years, however, some of participants in the GBD Project have called the use of the PTO into question. Murray and colleagues have argued that DALYs were deviced as a Summary Measure of Population Health (SMPH) and their "dominant objective" is to measure the health of populations (the burden of disease) (cfr., Murray and Frenk [7], p.10). They separate out the issue of measurement from the potential applications of SMPH. According to them (Murray et al, [2], pp. 14-15) SMPH can be used to compare the health of populations, to monitor changes in the health of a given population… and to inform debates on priority setting. Resource allocation is just one potential use of SMPH. It is now suggested that the design of SMPH has been too closely linked with the issue of resource allocation and this has led to construct SMPH on the basis on methods like the PTO which "bring us far from the common-sense statement that one population is healthier than another" (Murray et al., [2], p. 23).

In order to compare the health of populations Murray et al. [2] tentatively propose a method based on an application of the Harsanyi's principle of choice from behind a 'veil of ignorance' [p.24]. Specifically, Murray et al. [2] proposed that "the relation 'is healthier than' can be defined such that population A is healthier than population B if and only if an individual behind a veil of ignorance would prefer to be one of the existing individuals in population A rather than an existing individual in population B, holding all non-health characteristics of the two populations to be the same" (p. 24).

Although Murray et al. have not proposed explicitly to use the veil of ignorance (VEI) approach in order to value health states it seems reasonable to think that this is a logical follow-up of their proposal. Nord ([3], p. 142) suggests a straightforward way of applying the veil of ignorance approach to the valuation of health states. We transcribe literally his proposal:

> Suppose one wants to obtain a disability weight for severe asthma, subjects can then be faced with two hypothetical cohorts, A and B, of 100 people each (Table 1):
>
> **[TABLE 1]**
>
> Each subject can then be asked: Behind a veil of ignorance, to which cohort would you rather belong? One can then change the number of asthma cases until the subject is indifferent. If the median indifference number in an appropriate sample of subjects is in fact 20, the disability weight for asthma will be 5:20 = 0.25.
>
> This operationalization may be seen either as a variant of the person trade-off technique or as a probability trade-off resembling the standard gamble technique (Menzel 1990; 1999;

Nord 1999). Whether it should be seen as the former or the latter depends on the considerations that people take into account when they respond.

Nord raises an important question, namely, if people take into consideration equity concerns then the veil of ignorance approach will not overcome the problem that Murray et al. wanted to avoid, and preferences should be seen as a mixture of individual and societal value judgement (which is called by Menzel [8] a *Self&Other* perspective and by Dolan et al. [9] *socially inclusive personal* preferences). The existence of this ambivalence (*i.e.*, mixed preferences vs personal preferences) with respect to the actual meaning of VEI values motivates our paper. This paper tries to test if the approach suggested by Murray et al in order to measure the health of populations produces results closer to individual or to social values when it is applied to the valuation of health states.

2. Elicitation techniques

In order to conduct this test the VEI approach has to be compared with the PTO and with a method that only encapsulates individual values. To that end, we conducted a survey amongst the general population (N=300) asking VEI questions using the operationalization of Murray et al.'s proposal suggested by Nord [3]. Respondents were also interviewed using the conventional PTO method and a variant of the SG proposed by McCord and de Neufville [10] that they called the "lottery equivalent" (LE) method. In this method individuals are asked to set the indifference between two gambles by varying the probability of one of them. In order to emphasize the common root of SG and LE (*i.e.*, in both cases indifference is reached by varying probability), in this paper we will refer to LE as Double Gamble (DG). In summary, the main hypothesis of this paper will be tested by comparing VEI, DG and PTO approaches.

We think that the traditional SG is not a proper method in order to examine if the VEI approach is closer to individual or social values. The SG may incorporate some biases that can be a confunding factor in the interpretation of results. If subjects interpret VEI questions as expected utility maximizers, the operationalization suggested by Nord [3] can be seen as a comparison between two risky lotteries with no sure outcome in any of the two. By contrast, in the SG there is one degenerate lottery (i.e., the probability of a particular outcome is 1). It is commonly assumed that the SG is affected by the so-called 'certainty effect' [11]. This elicitation bias implies that the degenerate lottery in the SG is overweighed, leading to high utilities [12]. Some authors have proposed that the certainty effect can be avoided by the DG method [13]. The second objective of the paper is to study the potential influence of the certainty effect to explain the difference between VEI and other valuation techniques.

Finally, it could be hypothesized that PTO and VEI produce different values because the PTO may be affected by the so-called 'Rule of Rescue' (RoR) which would not influence VEI. The term RoR was coined by Jonsen [14] to describe the imperative to rescue identifiable individuals facing avoidable death. In the PTO there are no identifiable individuals but it commonly compares life saving treatments with treatments that mainly improve quality of life. The response to a PTO question may be influenced not only by health considerations but also by the special consideration that live saving treatments may have [15] [16].

Hence, we hypothesize that if the standard frame of the PTO is used (life saving vs quality of life improvements), the RoR would lead to an overweighting of the 'lives saved' option. It is possible that VEI and PTO values be different because the imperative of avoiding deaths is present in the PTO but not in VEI. To test this hypothesis is the third objective of this paper.

We test this hypothesis by removing death as the bottom endpoint in the scale on which quality weights were estimated for two health states. Death was replaced by an intermediate health state worse than the health state to be evaluated. Quality weights were then estimated on a scale 0-1, where 0 was the weight attached to the worse intermediate health state and 1 was the weight attached to full health as usual.

In sum, preferences in our study were elicited using four methods, namely: DG, PTO, VEI, and the classical SG. The Visual Analogue Scale (VAS) was also included at the beginning of the surveys to familiarize subjects with the health states to be valued. Framing of questions is shown in Appendix A.

We used a multi-step procedure in order to reach indifferences with all methods. Appendix B depicts the application of this procedure for the specific case of the DG method. As it can be seen there, the two first questions (Choices 1 and 2) allow us to discriminate between those respondents who regard the health state as better-than or worse-than death. Next, we present a table with multiple choices, and for each choice the chances of one of the two gambles are varied, until the indifference is reached. The estimation of quality weights at the indifference point for each method is outlined in Appendix C.

3. Design and methods

HEALTH STATES

We calculated quality weights for five health states (henceforth they are denoted as Q, R, X, Y, and Z) representative of the loss of quality of life after stroke (*see Table 2.*)

[TABLE 2]

The instrument used to describe health states was the *Modified Rankin Scale* ([17], [18]). This instrument yields an ordinal ranking of severity, according to which the five health states can be logically ordered from better to worse as follows: Q $\succ$ R $\succ$ X $\succ$ Y $\succ$ Z, where the preference relation $\succ$ means "preferred to".

STUDY DESIGN

Two stratified random samples (stratified on age and gender) were drawn from the general population from the Barcelona area using random-digit-dialing to recruit people willing to participate in a face-to-face interview in their home. A team of 6 trained interviewers conducted the interviews over a 2 month period. Respondents were not compensated for participating. In order to test the reliability of the methods, 50 subjects were again interviewed two weeks later. 12 Euros were offered as monetary compensation for accepting to be interviewed for a second time. About 50% rejected the payment.

Table 3 summarizes our study design. One survey (henceforth, survey 1) included questions with non-traditional methods (*i.e.*, DG, PTO, and VEI), while the other survey (survey 2) included conventional SG questions. Health states and questions were randomized in the two surveys, with the exception of VAS questions that were always presented at the beginning of the questionnaire. Respondents in survey 2 only valued three health states (R, X, and Y), so comparisons of the non-traditional methods with the SG are limited to these three health states.

In order not to overburden the respondents we splitted the total sample of survey 1 into two sub-samples of equal size. Each sub-sample completed all the tasks, *i.e.*, the methods administered, for three of the five health states under evaluation. With this design we expected to conduct interviews not too long. In addition, this design allows performing between-subject tests of consistency. It has the advantage of making more difficult that respondents can perceive the ordinal logic of questions, avoiding strategic behavioural forcing consistency.

[TABLE 3]

When death was not present in questions (denoted ✓*), quality weights were then estimated in two different ways. First, as we explained in Section 2, we used a chained procedure. In this way, table 2 shows that the weight of health state R was estimated both with (sub-sample 1) and without (sub-sample 2) chaining. Hence, for the sake of clarity, from now on we will identify with R2 the case in which the weight of R was estimated in a standard way (using death as the worst reference health state) and with R1 the case in which the weight of R was chained to death throughout health state X. Equally, health state Q was chained to death by mean of the anchor health state R2.

Second, as we also explained in Section 2, the influence of the RoR was tested by removing death of the scale on which quality weights were computed. In consequence, we also estimated weights for Q and R1 by setting values for R2 and X, respectively, equal to 0.

METHODS

It is well established in the utility measurement literature that negative valuations may induce heavily skewed valuation data (Patrick et al., 1994) distorting means largely. Hence, we checked the extent to which our data were influenced by negative values. *Kolmogorov-Smirnov* goodness-of-fit tests with *Lilliefors* modification rejected normality. In addition, descriptive statistics revealed that a few outliers had a large effect on means. Consequently, we opted for analyze data by using medians and nonparametric statistics. Other analyses were also performed (*e.g.*, trimmed means) but they are not reported here. Results from other analyses are available on request to the authors.

Comparisons between non-traditional methods (*i.e.* DG, PTO, and VEI) were performed by using the *Friedman ranks sum test*. In case Friedman tests were significant we then should perform multiple pairwise comparisons. Two procedures were used in order to control Type I error in multiple comparisons : (i) The so-called *Nemenyi's Critical Difference* procedure.; and (ii) series of *Wilcoxon signed ranks tests* with *Bonferroni-Holm* correction. The application of two different procedures served the purpose to test the robustness of our findings[4].

We will only report the results of the two tests separately if they yield different conclusions.

3

4 See Nemenyi (1961) and Holm (1974) for a description of the abovementioned procedures.

,

The *Mann-Whitney* U test was used to test for differences between independent samples. In this way we compared SG weights (survey 2) with VEI and DG weights (survey 1) respectively. Bonferroni-Holm correction for multiple significance testing was considered if appropriate.

We also assessed various psychometric properties (*i.e.*, feasibility, reliability and consistency) of the VEI approach in relation to PTO and DG methods. These properties have been largely studied for traditional methods like the SG [27], [28], [29], [30] but there is little evidence for the PTO [31] and it is unknown for the DG.

Test-retest reliability was estimated by mean of *Spearman's rank coefficients*. Feasibility was examined by computing the percentage of complete responses. Finally, logical consistency, *i.e.*, the extent to which the health states used in the study was rank-ordered as $Q \succ R \succ X \succ Y \succ Z$ within the different methods, was tested by performing all possible pairwise comparisons between health states values obtained by the same method, *e.g.*, U(Q) vs U(Y) within the DG method. These comparisons were performed using the same non-parametric methods described above.

4. Results

RESPONDENTS

Survey 1 was administered to 300 respondents. There was no statistically significant difference between sub-sample 1 (N_1=150) and sub-sample 2 (N_2=150) in any sociodemographic characteristic (*i.e.*, age, gender, education level, and employment status). Mean age was around 48 years in sub-sample 1 and 47 years in sub-sample 2. Proportions of women were slightly higher than men in both sub-samples. Most people had elementary or secondary educational degree, and the majority of them were employed at the time of the interview. Interviews lasted approximately 50 minutes on average.

Survey 2 was administered to 100 respondents. Both surveys were alike demographically, but there was a difference in employment status (Chi-square, $P<0.05$). The percentage of employed was 88% in survey 2, while it did not exceed 73% in survey 1.

PSYCHOMETRIC PROPERTIES

All subjects were able to fully complete all tasks. Thus, the completion rate among the respondents was 100%. It seems that the VEI approach is highly feasible when it is conducted in a *face-to-face* interview.

Test-retest reliability was also satisfactory. Spearman rank correlation coefficients were larger than 0.9 for all methods considered ($P<0.01$ in all comparisons). The highest correlation was found for the DG method, reaching the unity, whereas in the case of VEI was 0.943. The lowest coefficient was for the PTO (0.928).

We also tested the logical consistency of the three methods. Comparisons between all pairs of health states within each of the sub-samples were significant ($P < 0.001$) and they went in the direction we expected (*i.e.*, $R1 \succ X \succ Z$ for sub-sample 1, and $Q \succ R2 \succ Y$ for sub-sample 2). Once we found that consistency was satisfied within each sub-sample and method, we then test whether such a consistency also hold between the two sub-samples. Again, all the comparisons were significant (Mann-Whitney U test, $0.001 < P < 0.05$) with the exception of R1 and R2. In summary, we found that the rank-order of health states followed the logical order that we expected, *i.e.*, U(Q) > U(R) > U(X) > U(Y) > U(Z).

COMPARISONS OF METHODS

Table 4 displays the median (and the interquartile range) for each health state across the four methods tested (DG, PTO, VEI and SG) plus the VAS.

[TABLE 4]

As it can be seen, respondents were ordinally consistent in the initial VAS task (Wilcoxon test, P < 0.001). We found significant differences between PTO, VEI and DG (Friedman test, P < 0.0001). We then examined the differences between each pair of methods.

[TABLE 5]

The picture that seems to emerge from inspection of table 5 is:

First, the PTO clearly differs from the rest of the methods, since it generates significant higher values. In consequence, it seems that preferences elicited by VEI and PTO methods are different.

Second, DG and VEI valuations were quite similar. It seems that respondents behave as expected utility maximizers to evaluate VEI questions. For health state R1, the difference between DG and VEI is significant at the 5% level by the Wilcoxon test (P = 0.021) but it is not significant by the Nemenyi test.

Third, the SG produced higher utilities than VEI and DG (Mann-Whitney test, P < 0.001 for states R2 and X, and P < 0.01 for state Y). This result seems to support our hypothesis of the influence of the certainty effect. It also suggests that in order to test the relation between VEI and methods based on self-interest the SG does not seem the appropriate comparator.

As we explained in previous sections, we also computed quality weights for health states Q and R1 by setting equal to 0 the value of health states used as bottom endpoints in elicitations (states R2 and X respectively). We observe that even though death is removed from weights calculations, significant differences between VEI and PTO persist. Median value for Q was 0.51 for VEI and 0.63 for PTO ($P < 0.05$). In the case of state R1 the median was 0.5 for VEI and 0.81 for the PTO ($P < 0.05$). The tendency observed in the rest of health states, namely, that PTO weights are higher than VEI ones, clearly remains in spite of the absence of death. It seems that the difference between the PTO and the VEI is not only due to the rule of rescue (as we defined it).

This overall picture that the PTO seems to produce values that are different from VEI and DG values is reinforced if we analyse preferences for health states in relation to death for health states Y an Z (in the rest of health states the number of people that considered those states as worse than death was quite small).

[TABLE 6]

[TABLE 7]

Tables 6 and 7 show that the percentage of people that considered a health state as better or worse than death was quite similar for the VEI and DG methods. However, there was a tendency (about 30% of respondents) to consider a health state as better than death with the PTO but as worse than death with VEI or DG. This means that a large number of people said that for themselves it was worse to live the rest of their lives in situations Y or Z than dying. However, if they had to choose between a group of patients in situation Y or Z and another group of patients about to die, they would prioritise the group about to die, This indeed seems to confirm that for an important number of subjects the saving of life is given special status and the RoR is playing a role in PTO.

In summary, the PTO values seem to be quite different from VEI and DG values. One of the reasons seems to be the RoR but it does not seem to be only reason.

5. Conclusion

The veil of ignorance approach proposed by Murray et al. [1,2], using the operationalization suggested by Nord [3], seems closer to individual utilities than to social (PTO) values. This finding suggests, as indeed Murray et al. proposed, that it is possible to construct SMPH based on the Harsanyi's idea of the choice from behind a veil of ignorance. A direct implication of this finding for the GBD Project would be that DALYs, as far as we want to distance the measurement of population health from allocation considerations, could be based on VEI utilities rather than on PTO values. In general, the VEI approach could be useful not only to construct health gap-based measures like DALYs, but also to produce health expectancy-based measures like QALYs.

This does not mean, however, that VEI weights can be seen as conventional standard gamble utilities. Our data suggest that the VEI is not affected by the 'certainty effect', but it is closer to a 'double gamble' method originally devised to avoid the probability distortion caused by the classical standard gamble technique.

The close resemblance between VEI and DG values suggests that both approaches indeed elicit personal preferences rather than social or mixed preferences in health states valuations. However, we cannot conclude that both methods are interchangeable for constructing SMPH. The advantage of the VEI approach (at least as Murray et al. [2] argue) is that it agrees with common-sense definitions of population health. Murray et al. [2; p. 24-27] define several common-sense criteria to compare the health of populations. These criteria encompass more dimensions apart from reductions or improvements in quality of life. Comparisons between health of two populations at any instant in time can be characterized by different attributes like mortality rates, prevalence of a disease or remission for some intermediate health state to ideal health. In all these examples the veil of ignorance approach would coincide with a common-sense notion of population health. For example, if in society A infant mortality is lower than in society B, we all would agree (common-sense) that health in A is better than in B. A SMPH should show that health in A is higher than in B. A SMPH based on the veil of ignorance approach would show that health in A is better than in B. The reason is that any person, behind a veil of ignorance, would prefer to be a member of the population with the lower mortality rate. In this way the VEI approach can be used as a basis for constructing SMPH consonant with *any* of the dimensions described by Murray et al. It is not apparent that the DG method can be used on a similar basis.

It also seems that our findings about the relation between VEI and PTO cannot be entirely due to the effect of the 'Rule of Rescue', at least as this rule has been understood along this paper. Differences between both methods remain when death is not present. Of course, the RoR could be understood in more general terms as a tendency to help those who are worse-off. Our results for the PTO could then be interpreted as being influenced by the RoR in all cases.

Other papers have used the veil of ignorance perspective in order to analyze several health care issues (see Nord [32] p. 43-47 for a review and comment). However there is only one (Richardson and Nord [15]) with a similar perspective to ours.

Richardson and Nord used a "personalized" version of the PTO (that they called PTO-self) that incorporates the veil of ignorance perspective. The PTO-self questions explicitly invited respondents to have in mind their own interests when they made their choices. The effect of the change of perspective (from impersonal to personal) resulted in higher values for the PTO-self than for the PTO. This result seems to contradict ours. We believe that there are, at least, two explanations for this discrepancy.

The first explanation is that both framings seem to be quite different. Simply, there may be various possible operationalisations of the 'veil of ignorance' approach, such as Menzel [8] assumes to differentiate between *PTO-Self* and *PTO-Self&Other* questions. Indeed, this is also emphasized by Nord [3] when he suggests that "to avoid ambiguities and difficulties of interpretation, the veil of ignorance approach therefore needs some further specification" (p. 143). However, the present study seems to show that personal preferences may be captured by the VEI approach even without asking individuals to think only on their own self-interest

The second explanation is that the results obtained by Richardson and Nord are in some respect surprising. It is true that they find that the PTO-self produces higher values than the PTO however it is also true that in the same paper the Time Trade-Off utilities are also higher than the PTO. However, PTO values are commonly found higher than individual based utilities [33][34][35]. Then one possible explanation of the result that PTO-self is higher than PTO is that, for some unknown reason, in their paper PTO values are too low.

We think that the change in preferences in front of death is an interesting result because it illustrates quite clearly a potential criticism to the PTO, namely, "paternalism". We understand by paternalism to impose preferences for others that we personally do not hold, "this is bad for me but good for you". If I have to distribute resources only for myself then I would prefer resources to be addressed to treatments for problems Y or Z but if resource allocation has to influence the welfare of others then I prioritise life saving treatments. This is similar to saying that if I have to buy private health insurance I would prefer a policy insurance that gives treatment for Y or Z over a policy insurance that offers live-saving treatments. However, if I have to vote how to distribute public resources for others I would vote for live-saving treatments. We think that an interesting topic for future research would be to make people aware of the potential violations of individual preferences that would occur if resources were distributed according to their own PTO preferences. It is not clear to us if people would respond changing their individual preferences or changing PTO values.

This question is quite problematic and it is related to the issue of perspective in preference elicitation that has been discussed in other papers [9] [36]. What this paper shows is that the kind of considerations that people seem to take into account when responding to PTO questions are very different from the considerations they use under the VEI approach.

This study has also found that acceptability, reliability and consistency are good enough for the different methods used in this paper. This is relevant since, to the best of our knowledge, such features have not been examined before neither for the DG nor for the VEI using a representative sample of the general population, and, in the case of the PTO, the revision of the empirical evidence available is considered “inconclusive”, “relatively unknown”, or even, “virtually non-existent” ([15]).

One potential limitation of our study is that the results can be too dependent on the framing we have used. This is suggested by the opposite results obtained by Richardson and Nord [15]. We did not include different framings so it is important to test to what point our results depend on the specific framing we used. More specifically, we have elicited quality weights in a curative context, it could be that in a preventive context where the potential influence of the Rule of Rescue is reduced the discrepancy between VEI and PTO could also be reduced. In fact, in the DALY protocol of the European project ([37]), the PTO was framed in a preventive context.

In sum, our results show that if it was judged "necessary to distance the development of summary measures from the complex values that must be considered in the allocation of scarce resources" (Murray et al. [2]: p. 23) it seems that the veil of ignorance approach suggested by Murray and operationalised as Nord had proposed can indeed achieve this goal.

References

[1] Murray, C., Salomon, J., and Mathers, C. A critical examination of summary measures of population health. Bulletin of the World Health Organization 2000; 78: 981-994.

[2] Murray, C., Salomon, J., and Mathers, C. A critical examination of summary measures of population health.In CJL Murray, JA Salomon, CD Mathers and AD Lopez (eds.) *Summary Measures of Population Health: Concepts, Ethics, Measurement and Applications*, WHO, 2002.

[3] Nord, E., My goodness – and yours: A history, and some possible futures, of DALY meanings and valuation procedures; p. In CJL Murray, JA Salomon, CD Mathers and AD López (eds.) *Summary Measures of Population Health: Concepts, Ethics, Measurement and Applications*, WHO, 2002.

[4] Murray, C., Quantifying the burden of disease: the technical basis for disability adjusted life years. *Bulletin of the World Health Organisation*, 1994; 74: 429-445.

[5] Murray, C. Rethinking DALYs. In Murray, C. And Lopez, A. (eds.) *The Global Burden of Disease.* World Health Organisation/Harvard University Press, 1996.

[6] Nord, E.The person-trade-off approach to valuing health care programs. *Medical Decision Making* 1995; 15: 201-208.

[7] Murray, C. and Frenk, J. Summary measures of population health in the context of the WHO framework for health system performance assessment. In CJL Murray, JA Salomon, CD Mathers and AD López (eds.) *Summary Measures of Population Health: Concepts, Ethics, Measurement and Applications*, WHO, 2002.

[8] Menzel, P. How should what economist call 'social values' be measured. *Journal of Ethics*, 1999; 3: 249-273.

[9] Dolan, P., Olsen, J.A., Menzel, P. and Richardson, J. An inquiry into the different perspectives that can be used when eliciting preferences in health. Health Economics 2003; 12: 545-551.

[10] McCord, M. and De Neufville, R. "Lottery equivalents": reduction of the certainty effect problem in utility assessment. *Management Science*, 1985; 32: 56-60.

[11] Kahneman, D. and Tversky, A. A Prospect theory: an analysis of decision under risk. *Econometrica*, 1979; 50: 277-323.

[12] Hersey, J.C. and Schoemaker, P. H. J. Probability versus certainty equivalence methods in utility measurement. *Management Science*, 1985; 31: 1213-1231.

[13] Wakker, P. and Sttigelbout, A. Explaining distortions in utility elicitation through the rank-dependent model for risky choices. *Medical Decision Making*, 1995; 15: 180-186.

[14] Jonsen, A. R. Bentham in a box: technology assessment and health care allocation. *Law, Medicine and Health Care*, 1986; 14: 172-174.

[15]Richardson, J. and Nord, E. The importance of perspectives in the measurement of quality-adjusted life years. *Medical Decision Making*, 1997; 17: 33-41.

[16] Richardson, J. and McKie, J. The Rule of Rescue. CHPE Working Paper 112, 2000.

[17] Rankin, J. Cerebral vascular accidents in patients over the age of 60. Scottish Medical Journal, 1957; 2: 200-15.

[18] Bonita, R. and Beaglehole, R. Modification of Rankin Scale: Recovery of motor function after stroke. Stroke, 1988; 19: 1497-1500.

[19] Conover, W.J. Practical Nonparametric Statistics , third edition, Wiley, 1999.

[20] Bland, J. M. and Altman, D. G. Multiple significance tests: the Bonferroni method. British Medical Journal, 1995; 310: 170.
[21] Hollander, M. and Wolfe, D. Nonparametric Statistical Methods, second edition, Wiley, 1999.
[22] Sprent, P. and Smeeton, N. C. Applied Nonparametric Statistical Methods, Chapman & Hall/CRC, 2000.
[23] Howell, D. C. Statistical Methods for Psychology, fifth edition, Duxbury Press, 2002.
[24] Holm, S. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, 1979; 6: 65-70.
[25] Aickin, M. and Gensler, H. Adjusting for multiple testing when reporting research results: the Bonferroni vs Holm methods. American Journal of Public Health, 1996; 86: 726-28.
[26] Jaccard, J. and Wan, C. K. LISREL approaches to interaction effects in multiple regression. Sage Publications, 1996.
[27] Torrance, G. W. Measurement of health-state utilities for economic appraisal: A review. *Journal of Health Economics*, 1986; 5: 1-30.
[28] Froberg, D. G. and Kane, R. L. Methodology for measuring health-state preferences – II: Scaling methods. *Journal of Clinical Epidemiology* 1989; 42: 459-471.
[29] Dolan, P. et al. Valuing health states: a comparison of methods. *Journal of Health Economics*, 1996; 15: 209-31.
[30] Bleichrodt, H. and Johannesson, M. Standard gamble, time tradeoff, and rating scale: experimental results on the ranking properties of QALYs. *Journal of Health Economics*, 1997; 16: 155-175.
[31] Green, C. On the societal value of health care: what do we know about the person trade-off technique. *Health Economics* 2000; 10: 233-243.
[32] Nord, E. Cost-Value Analysis in Health Care. Making Sense out of QALYs. Cambridge University Press: Cambridge, UK, 1999.
[33] Nord, E. The validity of a visual analogue scale in determining social utility weights for health states. *International Journal of Health Planning and Management*, 1991; 6: 234-242.
[34] Ubel, P. A., Loewenstein, G., Scanlon, D., and Kamlet, M. Individual utilities are inconsistent with rationing choices. *Medical Decision Making*, 1996; 16: 108-116.
[35] Pinto, J. L. Is the person trade-off a valid method for allocating health care resources? *Health Economics*, 1997; 6: 71:81.
[36] Erik Nord, Menzel, P. RichardsonJ. The value of life: individual preferences and social choice. A comment to Magnus Johannesson *Health Economics*, 2003; 12: 873-877
[36] Van der Maas, P.J. Disability weights for diseases in Europe. Project Program. Mimeo, Erasmus University, Rotterdam, 1997.

Appendix A: Framing of the elicitation methods

FRAMING OF THE DOUBLE GAMBLE QUESTION

When somebody has a stroke it is very important to provide medication as soon as possible to remove the blood clot. Otherwise, there can be permanent injuries in the brain that may create important chronic health problems. However, the best treatment (or better said the best dose) is not clear. In general, there are two ways of dealing with this problem. The doctor can give to the patient two different doses, namely, high or low. The problem of the high dose is that in some patients can lead to death. The low dose cannot cause death but in some patients can be ineffective and then the stroke can cause brain damage and patients can have health problems the rest of their live.

The success and failure probabilities can be different for each dose. Now we are going to change the success and failure probabilities for each dose and we would like you to tell us which dose you think it would be better for you.

Assume that you have a stroke. If you take the high dose and it fails you have a 25% probability of dying, but if it is successful you have a 75% probability of restoring your health status to good health. If you take the low dose and it fails you have a 25% probability of remaining in health state X for the rest of your life, but if it is successful you have a 75% probability of restoring your health status to good health.

FRAMING OF THE PERSON TRADE-OFF QUESTION

Now we are going to ask you to assume that you are the head of a hospital and you have received an increase in your budget. You can decide to spend this money in two alternative programs A or B, which will benefit people of your age.

Program A is a new treatment that will totally cure people that right now are in health state X. Without the program they will remain in state X for the rest of their lives. Program B is a new medicine that will totally cure people that right now are about to die. Without the program they will die.

The number of people that you can benefit in the two programs is different with the same budget. We are going to change the number of potential beneficiaries in both programs and we are going to ask you to decide if you would spend the money in program A or B in each case. You do not have the possibility of splitting the budget between both programs.

FRAMING OF THE VEIL OF IGNORANCE QUESTION

We are going to ask you to imagine next situation. Imagine that there are two small societies of 200 people each. They are all like you. Both societies are the same in everything except in the health of the inhabitants. In society A there are 150 people with good health and 50 have a terminal illness that will cause their death in a few days. In society B there are 150 people in good health and 50 with a chronic health state like X.

With this information, would you prefer to be one of the current members of society A or one of the current members of society B? If you choose to be one of the current members of society A you can be one of the 150 people with good health or one of the 50 people with the terminal illness. If you choose to be one of the current members of society B you can be one of the 150 people in good health or one of the 50 people that are in health state X.

FRAMING OF THE STANDARD GAMBLE QUESTION

Assume that you have a stroke. If you take the high dose and it fails you have a 50% probability of dying, but if it is successful you have a 50% probability of restoring your health status to good health. If you take the low dose you remain with certainty in health state X for the rest of your life.

Appendix B: Multi-step procedure used to elicit indifferences

Step 1

Respondents were first told that there exist two treatments after stroke, a high dose treatment and a low dose treatment. Respondents were then shown a card describing one health state, say state X, and asked to choose between the two following alternatives:

High dose		Low dose		Decision		
Success Probability	Probability of death	Success Probability	Probability of state X	High Dose	Indifferent	Low Dose
75%	25%	75%	25%			

After the choice was explained, respondents were asked whether they preferred the high dose, the low dose, or whether they were indifferent between the two treatments. In case a respondent preferred either the high dose or the low dose, next choice was displayed:

Step 2

High dose		Low dose		Decision		
Success probability	Probability of death	Success probability	Probability of state X	High dose	Indifferent	Low dose
1%	99%	75%	25%			

Now when a respondent preferred the low dose the interviewer determined through a sequence of choices (Step 3) probabilities of success and failure until indifference was reached:

Step 3

High dose		Low dose		Decision		
Success probability	Probability of death	Success probability	Probability of state X	High dose	Indifferent	Low dose
65%	35%	75%	25%	Go	Stop	Stop
15%	85%	75%	25%	Stop	Stop	Go
55%	45%	75%	25%	Go	Stop	Stop
25%	75%	75%	25%	Stop	Stop	Go
45%	55%	75%	25%	Go	Stop	Stop
35%	65%	75%	25%	Stop	Stop	Stop

Finally, suppose that the respondent chose the high dose when probability of success was 15%. Hence, the indifference probability would be lower than 15%.

The same multi-step procedure was used for the remaining methods, *i.e.*, PTO, SG, and VEI..

Appendix C: Estimation of quality weights

I. METHODS USED IN SURVEY 1

- **In the *VAS* method**, the respondent is asked to place health states to be evaluated plus death on a line with 100 (best imaginable health state) and 0 (worst imaginable health state) as endpoints, such that the distance between the placements correspond to differences in preferences. If death is judged to be the worst state and placed at the 0 endpoint of the scale, the weight for each of the other health states is the scale value of its placement divided by 100. If death is not judged to be the worst state and it is placed at some intermediate point on the line (say d), the weights for the remaining states are obtained as

$$(i - d)/(1 - d) \qquad (1)$$

where i is the scale placement of the health state to be evaluated.

WITHOUT CHAINING

- **In the *DG* method**, the high dose alternative is a gamble with a chance *p* of good health and a chance *(1-p)* of dead. The low dose alternative is a gamble with a chance *q* of good health and a chance *(1-q)* of the health state to be evaluated (*say I*). The respondent is then asked to choose one of the two doses. At the indifference point, the weight for health state I is:

$$W(I) = (p - q)/(1 - q) \qquad (2)$$

by assuming the usual conventions W(good health) = 1 and W(dead) = 0.

- **In the *PTO* method**, Program A is a new treatment that cures a certain number *r* of patients suffering the health state to be evaluated. Program B is a new medicine that cure a certain number *s* of patients who are about to die. The respondent is then asked to choose one of the two programs. At the indifference point, the weight for health state I is:

$$W(I) = 1 - (s/r) \qquad (3)$$

- **In the *VEI* method**, Society A has *h* members in good health and *t* members who are about to die. Society B has *v* members in good health and *w* members in the state to be evaluated. The respondent is then asked to choose one of the two societies. At the indifference point, the weight for health state I is:

$$W(I) = 1 - (t/w) \qquad (4)$$

WITH CHAINING

- **In the *chained DG* method**, the respondent is asked to choose between a gamble with a chance *p'* of good health and a chance *(1-p')* of state J, and a gamble with a chance *q'* of good health and a chance *(1-q')* of the health state to be evaluated *(state I)*. At the indifference point, the weight for health state I is:

$$W(I) = [(p' - q') + (1 - p') \times W(J)]/(1 - q') \qquad (5)$$

where W(J) is estimated according to formula (1).

- **In the *chained PTO* method**, the respondent is asked to choose between a program that cures people suffering the anchor state J and a program that cures people in the state to be evaluated. We have then

$$W(I) = 1 - (s'/r') \times [1 - W(J)] \qquad (6)$$

where W(J) is estimated according to formula (2).

- **In the *chained VEI* method**, the respondent is asked to choose between a society with *v* members in good health and *w* members in the state to be evaluated, and a society with *h* members in good health and *t* members in the anchor state J. Thus we have

$$W(I) = 1 - (t'/w') \times [1- W(J)] \qquad (7)$$

where W(J) is estimated according to formula (3).

II. METHODS USED IN SURVEY 2 (STANDARD GAMBLE)

- If the health state to be evaluated (I) is regarded by the respondent as better than death, then the SG method offers the two following alternatives: a gamble (high dose alternative) with a chance *p* of good health and a chance *(1-p)* of dead, and health state I with certainty (low dose alternative). The respondent is then asked to choose one of the two doses. At the indifference point, the weight for health state I is:

$$W(I) = p \qquad (8)$$

- If the health state to be evaluated (I) is regarded by the respondent as worse than death, then the SG method offers the two following alternatives: a gamble (low dose alternative) with a chance q of good health and a chance *(1-q)* of health state I, and immediate death with certainty (no dose). The respondent is then asked to choose one of the two doses. At the indifference point, the weight for health state I is:

$$W(I) = - q/(1 - q) \qquad (9)$$

TABLES AND FIGURES

Table 1 Two hypothetical cohorts for valuing a health state

Cohort	Healthy	Asthma at 40 with detailed description	Fatal disease at 40	Sum
A	80	20	0	100
B	90	0	5	100

Table 2· Description of the health states

Health state Q
The person has symptoms as a result of a health problem; symptoms can include numbness, minor problems with movement, or some difficulty with reading or writing. The symptoms do not interfere with the person's usual activities to any appreciable extent, but they may affect the person's enjoyment of aspects of their daily life.
Health state R
As a consequence of the health problem the person is restricted in participation in a major aspect of life that they engaged in previously. They may be unable to work or look after children if these were major roles before; they may have restricted social and leisure activities; or they may have experienced significant disruption of close relationships. They can look after their own affairs (preparing meals, household chores, shopping in the neighbourhood, looking after their financial situation…) and can attend to their bodily needs (such as washing, going to the toilet, and eating) without problems.
Health state X
As a consequence of their health problem the person is unable to live independently. They will be unable to travel alone or shop without help if they did these things previously; and they will be unable to look after themselves at home for some reason (for example they may not be able to prepare a meal, do household chores, or look after money). They can attend to their bodily needs (such as washing, going to the toilet, and eating) without problems.
Health state Y
As a consequence of their health problem the person needs assistance with some basic activities of daily living or needs help from another person with walking. Basic activities of daily living include attending to bodily needs such as washing, going to the toilet, and eating.
Health state Z
As a consequence of the health problem the person is bedridden, and unable to move from bed without assistance. They may be incontinent, and will need someone available at all times to look after them.

Table 3 · Design of the study: tasks by sample, sub-sample, and health state

	Survey 1								Survey 2	
	VAS		DG		PTO		VEI		SG	VAS
Health state	Sub. 1	Sub. 2	Sub. 1	Sub. 2	Sub. 1	Sub. 2	Sub. 1	Sub. 2		
Q	✓	✓		✓*		✓*		✓*		✓
R	✓	✓	✓*	✓	✓*	✓	✓*	✓	✓	✓
X	✓	✓	✓		✓		✓		✓	✓
Y	✓	✓		✓		✓		✓	✓	✓
Z	✓	✓	✓		✓		✓			✓

✓ denotes the task administered
* Death not present

Table 4 • Valuations for each health state: medians (IQR)

Health state	VAS	DG	PTO	VEI	SG
Q	0.68	0.78	0.99	0.76	n,a,
	(0.25)	(0.3)	(0.05)	(0.35)	
R2	0.47	0.48	0.95	0.50	0.80
	(0.25)	(0.36)	(0.22)	(0.41)	(0.49)
R1	0.39	0.61	0.92	0.67	n,a,
	(0.26)	(0.49)	(1.71)	(1.16)	
X	0.20	0.26	0.77	0.29	0.47
	(0.21)	(0.66)	(0.51)	(0.66)	(0.8)
Y	0.07	-0.04	0.49	-0.15	0.00
	(0.16)	(1.02)	(1.12)	(1.35)	(0.01)
Z	-0.02	-0.72	-0.10	-0.70	n,a,
	(0.11)	(2.2)	(1.2)	(1.2)	

Note: IQR stands for *inter-quartile range*.

Table 5 • Test for differences amongst the different methods: differences between medians (Wilcoxon's z-values and Mann-Whitney's z-values)

Health state	DG vs PTO	DG vs VEI	PTO vs VEI	SG vs DG	SG vs PTO	SG vs VEI
Q	-0.22***	0.02	0.25***	n,a,	n,a,	n,a,
	(-9.350)	(-0.857)	(-9.607)			
R2	-0.47***	-0.02	0.45***	0.32***	-0.15***	0.3***
	(-9.834)	(-0.549)	(-9.276)	(-6.894)	(-5.345)	(-5.920)
R1	-0.31***	-0.06*	0.25***	n,a,	n,a,	n,a,
	(-8.718)	(-2.317)	(-7.496)			
X	-0.51***	-0.03	0.48***	0.21***	-0.3***	0.18***
	(-9.849)	(-1.857)	(-9.175)	(-4.904)	(-4.786)	(-3.750)
Y	-0.53***	0.11	0.64***	0.04**	-0.49***	0.15**
	(-5.349)	(-0.902)	(-6.663)	(-2.863)	(-4.395)	(-3.049)
Z	-0.62***	0.02	0.6***	n,a,	n,a,	n,a,
	(-5.704)	(-1.430)	(-5.464)			

Note: Significance level alpha after sequential Bonferroni-Holm correction.
*** denotes statistical significance at alpha = 0.001; ** alpha = 0.01; * alpha = 0.05

Table 6 • Percentage of respondents (n=150) regarding state Y as better, indifferent, or worse than death

	Better	Indifferent	Worse
DG	35.3	7.3	57.4
PTO	62.7	8.0	29.3
VEI	33.3	8.7	58.0

Table 7 • Percentage of respondents (n=150) regarding state Z as better, indifferent, or worse than death

	Better	Indifferent	Worse
DG	12.7	3.3	84.0
PTO	40.0	4.7	55.3
VEI	14.0	2.7	83.3

Figure 1: The Utility Function for Years in Good Health

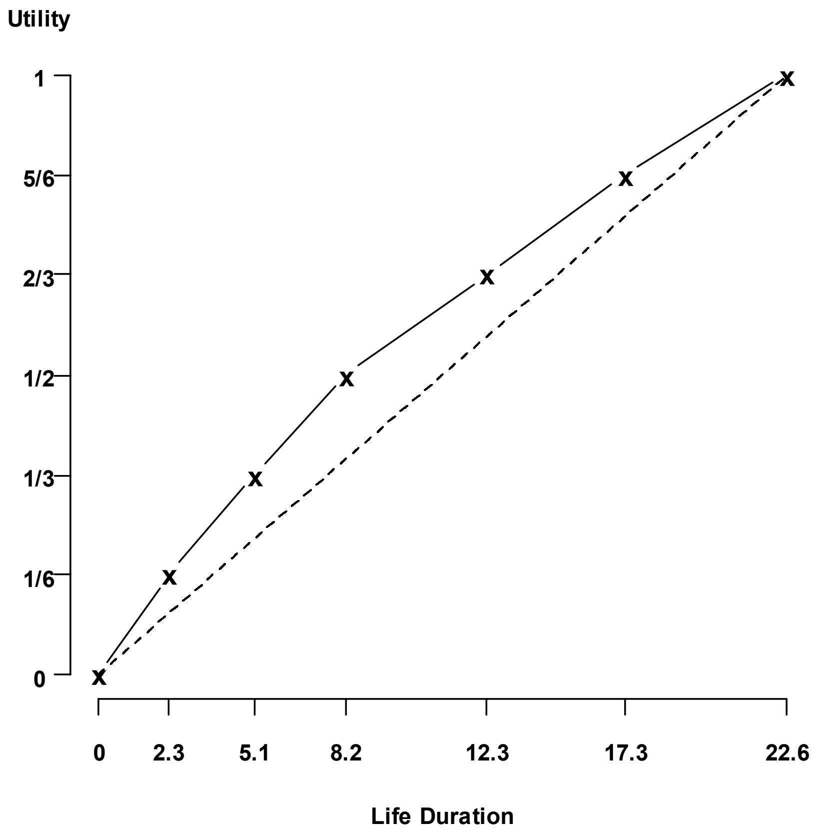

Figure 2: Utility Functions Elicited in the Second Experiment