Establishing disability weights from pairwise comparisons for a US burden of disease study

JÜRGEN REHM^{1,2,3} & ULRICH FRICK^{4,5}

1 Center for Addiction and Mental Health, Toronto, Canada

2 Dalla Lana School of Public Health, University of Toronto, Toronto, Canada

3 Institut für Klinische Psychologie und Psychotherapie, Technische Universität (TU) Dresden, Germany

4 Department of Healthcare Management, Carinthia University of Applied Sciences, Feldkirchen, Austria

5 Psychiatric University Hospital, Regensburg, Germany

Key words

DALYs, valuation, pairwise comparison, ranking, disability weights

Correspondence

Jürgen Rehm, Centre for Addiction and Mental Health, 33 Russell Street, Toronto, Ontario, M5S 2S1, Canada.
Telephone (+1) 416 535 8501 ext. 6173
Fax (+1) 416 595 6033
Email: jtrehm@gmail.com

Abstract

To determine valid and reliable disability weights for a U.S. burden of disease study, a convenience sample of 68 clinical experts was recruited, including representatives from over 20 NIH institutes and Centers for Disease Control and Prevention. Experts were given various health state valuation tasks including pairwise comparison, ranking, and Person Trade Off. Materials consisted of standardized descriptions of 11 attributes per health state (Classification and Measurement System of Functional Health, CLAMES). Attributes comprised up to 5 ordinal levels of disability. All states were displayed either with or without health state labels. Health state descriptions were taken from an existing comprehensive Canadian system. Conditional Logistic (CLR) and Probit Regression (PR) were used to derive disability weights. CLR and PR converged in yielding stable regression weights to construct disability weights, with a correlation of 0.816. The overall test-retest reliability amounted to 92.5% identical decisions. No significant difference was found for the presentation of health states with or without labels. A comparison of the expert valuations from our study with a standard gamble based valuation in the general population resulted in agreement of $r = 0.61$. The chosen methodology yielded valid and reliable and disability weights. As it is based on a modularized set of attributes, this methodology will allow derivation of disability weights on the basis of existing descriptions using the CLAMES. Copyright © 2013 John Wiley & Sons, Ltd.

Introduction

Disability-adjusted life years (DALYs; Murray, 1994, 1996; Anand and Hanson, 1997), a measure of health summarizing years of life lost due to premature mortality with years of life lost due to disability, have replaced mortality as the

primary indicator to characterize population health of nations not only in the statistics of the World Health Organization (WHO; e.g. in the World Health Reports since 2000), but also in the deliberations of the World Bank (see Ruger, 2005, and various World Development Reports) and other United Nation (UN) organizations.

This has led to various iterations of the Global Burden of Disease Study (Murray and Lopez, 1996; Lopez *et al.*, 2006a; WHO, 2008) as well as to national burden of disease (BoD) studies (e.g. Mathers *et al.*, 2001; Melse *et al.*, 2000) providing updates on DALYs. One key element of BoD, and thus of DALYs, is the non-fatal disease burden, the years of life lost due to disability (YLDs; Lopez *et al.*, 2006b).

YLDs are calculated by multiplying the incidence of various disease states with average duration and their respective disability weights (DWs). A DW is a metric for the decline of health associated with a certain health state, varying between zero (perfect health) and one (death). Most BoD studies have been based on either the original set of DWs of the first BoD study (Murray, 1996) or the Dutch DWs (Melse *et al.*, 2000). This may be a problem as several authors have emphasized that the degree of disability associated with health states may be region or country specific (Badia *et al.*, 2001; Bernert *et al.*, 2009; see also Ustun *et al.*, 1999). As a consequence when starting a US BoD study, we planned for a component to establish US specific, valid and reliable DWs. The theoretical underpinnings and exact procedures are summarized elsewhere (Rehm and Frick, 2010; Frick *et al.*, 2012). Thus, the main objectives of the here described study are (1) quantitative determination of DWs for the United States, (2) determination of validity and reliability of DWs elicited by different judgmental tasks and derived from different statistical analyses.

Methods

The methodology and its underlying rationale, based on a systematic literature review, have been presented in detail elsewhere (Rehm and Frick, 2010). Therefore, we will restrict ourselves to presenting only the key information on sample and materials, but will go into more detail regarding the statistical methods.

Workshops and participants

The study population was comprised of a convenience sample ($n=68$) recruited from a range of health backgrounds, mostly clinical. The study was carried out in the form of workshops in the following institutions: National Institutes of Health (NIH) (two meetings), Bethesda, MD; Centers for Disease Control and Prevention (CDC), Atlanta, GA; University of Texas, Southwestern, Dallas, TX; Center for Addiction and Mental Health, Toronto, Canada. Care was taken to obtain representation from a variety of specialties, with over 20 of the NIHs

being present. All workshops lasted one to two days and took place in 2008. Homework was distributed two weeks before the workshops, including pairwise comparison tasks and ranking exercises (see later). Some task teams of two people had to discuss and answer questions (this technique was mainly used for complicated trade-offs – see later).

Procedures and materials

Three main methodologies were used to elicit health state valuations. (1) Pairwise comparison: two health states, each characterized by 11 attributes were displayed and the respondents had to choose the more disabling health state (see Figure 1). (2) Ranking: six health states were presented with the task to rank order them from most disabling to least disabling. (3) Two person trade-off (PTO) tasks were presented based either on a rationing scenario, or on preventive interventions. Both PTO tasks were analyzed only with regard to their initial preference statement for the purpose of this paper. The number of persons representing the point of judgmental indifference was not part of the analysis mentioned later.

We used the Classification and Measurement System of Functional Health (CLAMES; McIntosh *et al.*, 2007) that was developed by the Health Analysis and Measurement Group, Statistics Canada for defining different health states. It was adapted from three leading generic health status instruments: the Health Utilities Index Mark III (HUI3; Feeny *et al.*, 2002; Furlong *et al.*, 2001), the Medical Outcomes Study Short-Form 36 (SF-36; Ware and Sherbourne, 1992); and the European Quality of Life Five-Dimensions Index Plus (EQ-5D; Brooks and EuroQoL Group, 1996; EuroQoL Group, 1990; Rabin and de Charro, 2001). Each health state was characterized by given level of 11 attributes, six core attributes (pain or discomfort, physical functioning, emotional state, fatigue, memory and thinking, social relationships) and five other attributes (anxiety, speech, hearing, vision, use of hands and fingers). Table 1 lists all attributes and their levels.

Health state descriptions were based on a multi-year project of Health Canada aimed at producing a unified set of common characterizations using the CLAMES (Evans *et al.*, 2005; McIntosh *et al.*, 2007; Murphy *et al.*, 2005). In addition, we asked participants of the NIH and CDC workshops for five descriptions from their respective specialty area applying the CLAMES system. These descriptions could include health states composed of comorbid conditions (see Table 4 later for some examples). Overall, 379 health state descriptions were used, with 1704 different pairwise combinations

NIH Health Disability Analyzer

Read Me

NIH Study: Disability of Different Health States

In the following, you will be asked to make comparisons between several sets of two health states. Each of the health states will be described by different attributes. For each set, please choose the health state that you believe has a higher level of disability. If Health State I is more disabling, please choose 1. If you believe that Health State II is more disabling, please choose 2.

HEALTH STATE 1	HEALTH STATE 2
Pain and discomfort: Severe pain or discomfort	Pain and discomfort: Generally free of pain and discomfort
Physical functioning: Severe limitations in physical functioning	Physical functioning: Mild limitations in physical functioning
Emotional state: Very unhappy	Emotional state: Somewhat happy
Fatigue: Always feel tired and have no energy	Fatigue: Sometimes feel tired and have little energy
Memory and thinking: Very forgetful and have great difficulty when trying to think or solve day to day problems	Memory and thinking: Able to remember most things, think clearly and solve day to day problems
Social relationships: Severe limitations in the capacity to sustain social relationships	Social relationships: No limitations in the capacity to sustain social relationships
Anxiety: Moderate levels of anxiety experienced regularly	

Instruction

Indicate the more disabling health state by clicking on one of the two buttons below. Then press SUBMIT. Please note, once your response is submitted, you cannot go back to your previous question to change/resubmit.

Which one is more disabling?

13/35 ☒ Health State 1 ☐ Health State 2

start NIH Health Disabil... 100% 9:30 PM

Figure 1 Example of a pairwise comparison (computer screen).

presented. In other words, while each health state was presented at least once, not all of the possible pairwise combinations were presented. In the homework and the first phases, key diseases from the Canadian project were presented in a random order. In the last phases of the project, the newly created health state descriptions were presented paired with the Canadian descriptions. On average, each rater judged more than 200 different pairs.

Labeling: all three elicitation methods were presented either with or without health state labels. In the labeled case, health states were characterized by clinical diagnoses combined with details regarding the duration or severity of the health state (example: “diabetes type II: uncomplicated, chronic” or “first degree burn: standard”). Participants were instructed to compare the health states based on the assumption that states lasted exactly the same time period. Pairwise comparisons were either made directly or were indirectly derived from ranking or PTO.

Statistical analysis

Objective 1: determination of DWs

Two analysis techniques were used to transform respondents’ decisions into DWs: conditional logistic regression (CLR; Hosmer and Lemeshow, 1989), and probit regression (PR; McCulloch and Nelder, 1989). The main technique was CLR and the PR was used as a sensitivity analysis.

CLR could be used, as the pairwise comparison was always based on within subject decisions. Step 1 determined the beta weights for each level of the attributes, coded as dummies. If two adjacent levels of the same attribute did not significantly differ, these levels were combined in step 2. A new CLR was then calculated to estimate the regression coefficient for the combined levels. Step 3 tested the inclusion of all first-order interaction terms for the six core descriptive attributes. For inclusion, an a priori criterion of 5% improvement in model fit was set, with

Table 1 List of attributes to characterize each health state based on the CLAMES system (McIntosh *et al.*, 2007)

Attribute	Level	Description
<i>Core attributes</i>		
Pain or discomfort (PD)	1	Generally free of pain and discomfort ¹
	2	Mild pain or discomfort
	3	Moderate pain or discomfort
	4	Severe pain or discomfort
Physical functioning (PF)	1	Generally no limitations in physical functioning ¹
	2	Mild limitations in physical functioning
	3	Moderate limitations in physical functioning
	4	Severe limitations in physical functioning
Emotional state (EM)	1	Happy and interested in life ¹
	2	Somewhat happy
	3	Somewhat unhappy
	4	Very unhappy
	5	So unhappy that life is not worthwhile
Fatigue (FA)	1	Generally no feelings of tiredness, no lack of energy ¹
	2	Sometimes feel tired and have little energy
	3	Most of the time feel tired and have little energy
	4	Always feel tired and have no energy
Memory and thinking (TH)	1	Able to remember most things, think clearly and solve day-to-day problems ¹
		Able to remember most things but have some difficulty when
	2	trying to think and solve day-to-day problems
	3	Somewhat forgetful, but able to think clearly and solve day-to-day problems
	4	Very forgetful, and have great difficulty when trying to think or solve day-to-day problems
Social relationships (SR)	1	No limitations in the capacity to sustain social relationships ¹
	2	Mild limitations in the capacity to sustain social relationships
	3	Moderate limitations in the capacity to sustain social relationships
	4	Severe limitations in the capacity to sustain social relationships
	5	No capacity or unable to relate to other people socially
<i>Supplementary attributes</i>		
Anxiety (AN)	1	Generally not anxious ¹
	2	Mild levels of anxiety experienced occasionally
	3	Moderate levels of anxiety experienced regularly
	4	Severe levels of anxiety experienced most of the time
Speech (SP)	1	Able to be understood completely when speaking with strangers or friends ¹
		Able to be understood partially when speaking with strangers but able to be
	2	understood completely when speaking with people who know you well
	3	Able to be understood partially when speaking with strangers and people who know you well
	4	Unable to be understood when speaking to other people
Hearing (HE)		Able to hear what is said in a group conversation, without a hearing aid,
	1	with at least three other people ¹
		Able to hear what is said in a conversation with one other person in a quiet room, with or without a hearing aid, but require a hearing aid to hear
	2	what is said in a group conversation with at least three other people
		Able to hear what is said in a conversation with one other person in a quiet room, with or without a hearing aid, but unable to hear what is said
	3	in a group conversation with at least three other people
	4	Unable to hear what others say, even with a hearing aid

(Continues)

Table 1 (Continued)

Attribute	Level	Description
Vision (VI)	1	Able to see well enough, with or without glasses or contact lenses, to read ordinary newsprint and recognize a friend on the other side of the street ¹
		Unable to see well enough, even with glasses or contact lenses, to recognize a friend
	2	on the other side of the street but can see well enough to read ordinary print
Use of hands and fingers (HF)		Unable to see well enough, even with glasses or contact lenses, to read ordinary newsprint or to recognize a friend on the other side of the street
	1	No limitations in the use of hands and fingers ¹
		Limitations in the use of hands and fingers but do not require special tools or the help of another person
	2	Limitations in the use of hands and fingers, independent with special tools and do not require the help of another person
	3	Limitations in the use of hands and fingers, require the help of another person for some tasks
	4	Limitations in the use of hands and fingers, require the help of another person for most tasks

¹Reference category.

model fit measured by the Estrella index (Estrella, 1998). Step 4 anchored the final solution into the range between completely healthy (DW = 0) and dead (DW = 1) using two tracer diagnoses (see section entitled “Transformation into DWs”).

A random effects PR was used as sensitivity analysis. A linear additive model with the following equation was estimated:

$$DW_{ij}^* = \sum a_{dl} X_{dl} + e_{ij} + u_j$$

where DW_{ij}^* is the still unanchored disability weight for outcome state i as valued by individual j , X_{dl} is a vector of dummy variables where d represents the attribute from CLAMES and l the level of that attribute (see Table 1). The α represented the parameter weights that were estimated for all levels of an attribute relative to level 1, and e was the error term due to health states, while u was an error term due to differences among respondents.

Model estimation of CLR was performed using SAS PROC MDC (SAS Institute Inc., 2004). PR was obtained by STATA (Stata Corporation, 2008).

Objective 2: retest reliability and stability

In each of the workshops, including the homework, some judgmental comparison tasks were presented repeatedly several times. In total, 66 of the 68 persons were involved in judging at least one pairwise comparison twice. For the

most part, the repetitions were chosen at random. On average, each rater judged 18 pairs at least twice.

This allowed calculation of retest reliability coefficients, within and across elicitation methods. Correlation techniques were used to ascertain stability across elicitation methods and statistical techniques. Based on the scale level of the variables, either Pearson or Spearman correlations were used.

The impact of presenting the same materials with and without labels was tested in a hierarchical random effects regression (HRER; Raudenbush and Bryk, 2002) with percent choice of a given alternative for each pair as the dependent variable. Level 1 of the HRER is constituted by the percent choice separate by label yes/no with the number of judgments as weighting variable; Level 2 of the HRER was the pair of health states to be judged. In addition, the correlation between the percent choices with and without label was calculated. HLM 6.0 was used for the hierarchical random effects model (Raudenbush *et al.*, 2004).

The resulting DWs were compared to coefficients expressing the level of disability derived from 146 town hall meetings using standard gamble as elicitation method (McIntosh *et al.*, 2007).

Transformation into DWs

The final step in our analysis was to transform the results of the CLR into DWs. A logistic regression by definition yields probabilities between zero and one. The endpoints

of DWs, i.e. zero and one, are clearly defined; zero means perfect health, whereas value one denotes death. In order to anchor the results from the CLR we thus had to define the value of the most disabling health state. Medical experts were asked to imagine the health state with the most severe disabling attributes and scale it on a magnitude scale, the usual thermometer scale for utility measurements (Torrance, 1987) ranging between zero and one. An average value of 0.94 resulted. In addition, we had to anchor the lower part of our DW as well. We used deafness as our anchor with a DW of 0.23 in agreement with the Dutch (Stouthard *et al.*, 1997) and Australian (Mathers *et al.*, 2001) DWs.

Results

Table 2 presents an overview of the CLR for each level of attribute. Overall, the Estrella goodness of fit (GOF) index was 0.676, which can be interpreted analogously to pseudo R^2 that 67.6% of the “variance” could be explained (Estrella, 1998). As can be seen in Table 2, for the overwhelming majority of attributes a clear rank order of judged disability could be established based on progressing severity of the descriptions. This is especially true for the six core attributes of the CLAMES: pain or discomfort, physical functioning, emotional state, fatigue, memory and thinking, social relationships.

Two non-core attributes (speech, use of hands and fingers) had to be excluded, as they led to inconsistent results, i.e. more disabling attributes were associated with lower beta coefficients for disability. The reason for this seemingly paradoxical result lies in the fact that the presented health states were not composed of random subsets of attributes, and thus not balanced. Instead, to facilitate experimental realism and meaningful judgments real health states were selected (Rehm and Frick, 2010) and their attributes of course are markedly correlated and not orthogonal. This means that many health states manifest themselves in the same cluster of attributes, often in the core attributes. However, health states with high levels of disability on uncommon attributes can by chance be linked to less overall disability, thus resulting in decreasing order of coefficients.

Fifteen two-way interaction terms (i.e. joint impact of two attributes is higher than the sum of their main effects on the logarithmic level) were tested. Statistical analysis yielded a combined improvement of pseudo R^2 , as measured by the Estrella GOF index, by 0.8% to 68.4%. This is markedly less than the criterion value of 5% increase or an additional increase of GOF by 3.4%. As a result the decision was taken to choose the more parsimonious model without interactions.

How did the results from pairwise comparison and ranking compare to the results from PTO elicitation? This question could not be answered directly in a regression model as the sample of PTO judgments was not large enough to allow either CLR or PR. Table 3 instead presents an indirect comparison between the results of the different elicitation methods. It is based on the 10 most frequently presented pairs of health states in the PTO. As can be seen, both PTO and pairwise comparisons yielded comparable preferences ($r_{\text{pairwise,PTO}} = 0.925$; $p < 0.001$).

The last column of Table 3 displays the predicted levels of preference based on the CLR. Correlations with direct pairwise comparisons (column “pairwise comparison”; $r_{\text{CLR,pairwise}} = 0.950$) and with indirectly derived from PTO (column “PTO”; $r_{\text{CLR,PTO}} = 0.869$) were high.

The correlation between statistical techniques was also high and amounted to $r_{\text{CLR,PR}} = 0.816$ ($p < 0.01$; results for PR are given in the Appendix, Table A1).

The correlation between judgments based on health state comparisons with and without labels was $r_{\text{unlabeled,labeled}} = 0.791$ ($p < 0.01$) in the CLR and 0.755 ($p < 0.01$) in the PR. There were 133 different kinds of health state pairs presented at least three times each with or without a label (average n of presentation without label: 42.1; 95% confidence interval (CI): 36.8–47.4; median: 34; with label: 15.8; 95% CI: 13.7–17.8; median: 19). A formal test of significance of differences in the proportion of which of the two health states was chosen more disabling showed no differences between labeled and unlabeled condition (random intercept model: coefficient for label yes/no: -0.00844 ; standard error (SE) = 0.00994; $t = -0.849$; $df = 264$; $p = 0.397$).

Finally, there was temporal stability in judgments: overall, 92.5% of the 2960 comparisons that were rated at least twice by the same participant resulted in the same decision. Thus, $n = 66$ out of 68 raters were exposed to (unannounced) repeated tasks.

Expert ratings of the disabling effects of each level of the nine health attributes (coefficients from Table 2) from pairwise comparisons resulted in similar effect estimates when compared to the results from the Canadian town hall meetings using standard gamble elicitation method and published by McIntosh *et al.* (2007, table 8). Here, the correlation of regression coefficients was $r_{\text{CLR,SG}} = 0.61$.

Overall, we found remarkable stability in judgments which of two health states was rated more disabling. Neither the cognitive framing of the task (pairwise comparison, ranking, PTO) nor the presentation with or without a health state label had a profound impact on the outcome. Finally, statistical methodology to estimate regression coefficients also yielded very similar results.

Table 2 Impact of different attributes on judgment of disability (effects of CLR)

Severity of disability	Health state attribute									
	Estimates	Pain or discomfort (PD)	Physical functioning (PF)	Emotional state (EM)	Fatigue (FA)	Memory and thinking (TH)	Social relationships (SR)	Anxiety (AN)	Hearing (HE)	Vision (VI)
Category 0 “no disability”		Reference	Reference	Reference	Reference together with category 1	Reference	Reference	Reference	Reference	Reference
Category 1 “mild”	Coefficient	0.2524	1.0307	0.3534	Combined	0.2050	0.4081	0.8577	0.9355	0.7906
	Standard error <i>p</i> - value for <i>t</i>	0.0914 0.0058	0.0907 < 0.001	0.0882 < 0.001	with category 0	0.0890 0.0213	0.0793 < 0.001	0.0956 < 0.001	0.1643 < 0.001	0.0681 < 0.001
Category 2 “moderate”	Coefficient	1.0013	1.3238	0.6786	0.1801	combined	0.7536		Combined with category 1	
	Standard error <i>p</i> - value for <i>t</i>	0.1021 < 0.001	0.1116 < 0.001	0.0849 < 0.001	0.0707 0.0108	with category 1	0.0890 < 0.001		3.0532 0.1666	Combined with category 1
Category 3 “severe”	Coefficient	Combined	2.8781	1.0467	0.733	0.7141	1.5248	Combined with category 1		
	Standard error <i>p</i> - value for <i>t</i>	with category 2	0.1285 < 0.001	0.1319 < 0.001	0.1105 < 0.001	0.0851 < 0.001	0.1204 < 0.001		< 0.001	
Category 4 “no capacity” (only EM and SR)	Coefficient			2.2762			2.0234			
	Standard error <i>p</i> - value for <i>t</i>			0.2297 < 0.001			0.3278 < 0.001			

Note: Estrella GOF measure: 0.6759.

Table 3 Stability of results based on different methods

Health states to be compared		Judgmental task				Conditional logistic regression
		Pairwise comparisons		PTO		
Health state 1 ...	Health state 2 ...	<i>n</i> of comparisons	% health state 1 more disabling	<i>n</i> of comparisons	% health state 1 more disabling	Predicted % health state 1 more disabling
3 3 3 3 1 2 2 2 0 2 3 0 2 3 1 3 3 2 1 0 2 3		46	95.3	51	94.1	100.0
2 2 2 2 0 1 2 1 0 2 0 0 0 0 0 0 2 0 3 3 0 0		38	94.7	63	80.9	80.8
3 3 4 0 4 4 3 0 0 0 0 0 0 0 0 0 2 0 3 3 0 0		38	100.0	29	96.6	99.5
3 3 3 2 4 3 1 3 0 2 3 3 3 0 2 0 2 1 0 0 0 0		37	100.0	64	87.5	93.1
3 3 3 3 3 3 2 0 0 0 0 0 3 3 4 0 4 4 3 0 0 0		37	40.5	34	5.9	43.0
1 1 1 0 3 3 2 0 0 0 0 0 0 0 2 1 1 1 1 0 0 0		37	100.0	52	98.1	93.0
3 3 3 3 3 3 2 0 0 0 0 0 1 1 1 0 3 3 2 0 0 0		34	100.0	33	96.9	98.2
3 3 4 3 2 4 3 1 1 1 3 1 1 1 1 1 3 1 4 4 1 1		30	90.0	47	78.7	91.2
3 3 4 3 2 4 3 1 1 1 3 3 2 3 3 4 3 3 1 1 1 1		25	76.0	43	69.7	47.0
1 1 1 1 1 3 1 4 4 1 1 3 2 3 3 4 3 3 1 1 1 1		25	12.0	60	25.0	7.9

Note: Numbers reflect health attributes PD-PF-EM-FA-TH-SR-AN-SP-HE-VI-HF (see Table 1).

Table 4 DALY weights for selected health states

Health state	Health state attributes											DALY weight
	PD	PF	EM	FA	TH	SR	AN	SP	HE	VI	HF	
Dental caries (acute)	1	0	0	0	0	0	0	0	0	0	0	0.001
Erectile dysfunction (chronic)	0	0	1	0	0	1	1	0	0	0	0	0.009
Alcohol dependence (mild to moderate) & Alzheimer (mild)	1	1	1	2	1	1	1	0	0	0	0	0.117
Diabetes Type 2 (uncomplicated chronic) & Angina CDS Cl. III	2	2	0	1	0	1	1	0	0	0	0	0.176
Deafness (severe hearing disorders early acquired)	0	0	0	0	0	2	0	3	3	0	0	0.230
Alcohol dependence (severe)	1	1	2	1	2	3	2	1	0	0	1	0.381
Chronic Obstructive Pulmonary Disease (COPD) (severe chronic)	2	2	2	2	1	2	2	0	0	0	0	0.486
HIV and Hepatitis C co-infection with treatment	2	2	2	3	2	2	2	0	0	0	0	0.611
Cancer – palliative end stage	2	2	3	2	1	3	2	0	0	0	2	0.723
Second degree burns (moderate)	3	3	0	3	0	3	2	0	0	0	0	0.834
Multiple sclerosis – severe chronic	3	3	3	3	1	2	2	2	0	2	3	0.908
AIDS end stage	3	3	3	3	3	3	2	0	0	0	0	0.920

Note: For the definitions of the health state attributes see Table 1.

Table 4 gives selected DWs based on the transformation procedure described earlier. As can be seen, the chosen method easily allows for determination of DWs

for comorbid health states. Of course, there is a prerequisite for a valid vector of attributes for each health state to be considered.

Discussion

DWs were computed on the basis of pairwise comparisons of health state descriptions based on a standardized set of attributes. Experts, mostly clinical, undertook the valuation task and attested the feasibility and meaningfulness of the approach chosen. Feasibility seemed to apply to all forms of operationalization for the pairwise comparisons, i. e. direct comparison of two health states, rank ordering exercise of sets of six health states, and PTO. Results clearly converged into reliable and stable choices, thereby contradicting the results of (Krabbe *et al.*, 1997), who found pairwise comparisons not reliable in student volunteers who were not necessarily trained in medical conditions.

Different forms of statistical analysis revealed a high degree of consistency of judgments within and between experts and methods. Moreover, the statistical analysis techniques, based on different assumptions, led to similar results.

The overall feasibility, consistency, and reliability of judgments do not necessarily mean that these judgments are valid or meaningful to be used as DWs in BoD studies. There were strong arguments following the original determination of DWs in the inaugural Global Burden of Disease Study (for a description see Murray, 1996) that DWs should not be based on experts but on the general population (Boyd *et al.*, 1990; Wiseman *et al.*, 2003).

First, some of these arguments may be more valid for utility studies in the quality-adjusted life year (QALY) tradition (Baker *et al.*, 2010) than for determining DWs. The theoretical underpinning for DWs, i.e. the concept of “disability”, is different from measuring quality of life (Broome, 2004; Rehm and Frick, 2010; for calculatory differences see also Airolidi and Morton, 2009). DWs simply want to measure decrements of health, a concept easily understood by experts and in the general population. They do not necessarily carry all the assumptions and implications of utility measurement to determine resource allocation.

However, comparisons of some of the specific degrees of disabilities associated with certain health states require clinical expertise, if labels are required. This was one of the main reasons for our choice of participants (for details see Rehm and Frick, 2010).

Second, irrespective of the theoretical position taken in this discussion, we thought it would be of merit to compare

the expert judgment with the perspective of the general population as derived from large series of town hall meetings. The correlation between expert judgments and general population valuations displayed significant overlap but a substantial amount of unexplained variation. Such a measure did not compare equals with equals. We asked our participants to identify the more disabling health condition, whereas in the town hall meetings the point of equivalence between choices was elicited. Additionally, the methodology differed: the more technical Standard Gamble (town hall meetings) versus the methodological mix in our study. Consistent with the theoretical model of (Stiggelbout and de Vogel-Voogt, 2008) we postulate that the different framings resulted in different appraisal processes. Further research should explore the most pronounced discrepancies in the contribution of different attributes to the overall judgment.

The epidemiology of the non-fatal part of BoD is methodologically less developed compared to the epidemiology of mortality. While DALYs have been integrated into routine reporting systems of UN agencies, this does neither mean that their calculation is standardized (e.g. with respect to age weighting and discounting, see Murray *et al.*, 2002) nor that methodological guidelines for the development of the underlying key concepts such as DWs have been agreed on (Murray *et al.*, 2002). The methodology for the present article has been based on a systematic research of the literature where all choices for single steps have been explained in detail (Rehm and Frick, 2010). This does not mean that other scientists in the field will necessarily accept our choices. However, we would hope that differing opinions would be substantiated by empirical exemplifications, and thus that the field as whole will grow based on empirical evidence rather than a priori settings.

Acknowledgment

Source of Financial Support: Contract No. HHSN267200700041C of the National Institute of Alcohol Abuse and Alcoholism (NIAAA), Bethesda, MD.

Declaration of interest statement

The authors have no competing interests.

References

Airolidi M., Morton A. (2009) Adjusting life for quality or disability: stylistic difference or substantial dispute? <i>Health Economics</i>, 18, 1237–1247. DOI: 10.1002/hec.1424	Anand S., Hanson K. (1997) Disability-adjusted life years: a critical review. <i>Journal of Health Economics</i>, 16, 685–702. DOI: 10.1016/S0167-6296(97)00005-2	Badia X., Roset M., Herdman M., Kind P. (2001) A comparison of United Kingdom and Spanish general population time trade-off values for EQ-5D health states. <i>Medical</i>

- Decision Making*, **21**(1), 7–16. DOI: 10.1177/0272989X0102100102
- Baker R., Bateman I., Donaldson C., Jones-Lee M., Lancsar E., Loomes G., Mason H., Odejar M., Pinto Prades J.L., Robinson A., Ryan M., Shackley P., Smith R., Sugden R., Wildman J. (2010) Weighting and valuing quality-adjusted life-years using stated preference methods: preliminary results from the Social Value of a QALY Project. *Health Technology Assessment*, **14**(27), 1–162. DOI: 10.3310/hta14270
- Bernert S., Fernandez A., Haro J.M., Konig H.H., Alonso J., Vilagut G., Sevilla-Dedieu C., de Graaf R., Matschinger H., Heider D., Angermeyer M.C., The ESEMeD/MHEDEA 2000 investigators. (2009) Comparison of different valuation methods for population health status measured by the EQ-5D in three European countries. *Value in Health*, **12**(5), 750–758. DOI: 10.1111/j.1524-4733.2009.00509.x
- Boyd N., Sutherland H., Karen Z., Heasman D., Cummings B. (1990) Whose utilities for decision analysis? *Medical Decision Making*, **10**, 58–67. DOI: 10.1177/0272989X9001000109
- Brooks R., EuroQoL Group. (1996) EuroQoL: the current state of play. *Health Policy*, **37**(1), 53–72. DOI: 10.1016/0168-8510(96)00822-6
- Broome J. (2004) *Weighing Lives*, Oxford, Oxford University Press.
- Estrella A. (1998) A new measure of fit for equations with dichotomous dependent variables. *Journal of Business and Economic Statistics*, **16**(2), 198–205.
- EuroQoL Group. (1990) EuroQoL – a new facility for the measurement of health-related quality of life. *Health Policy*, **16**(3), 199–208. DOI: 10.1016/0168-8510(90)90421-9
- Evans W.K., Connor Gorber S., Spence R.T., Will B.F. (2005) *Health State Descriptions for Canadians*: Cancers, Ottawa, Statistics Canada.
- Feeny D., Furlong W., Torrance G.W., Goldsmith C.H., Zhu Z., DePauw S., Denton M., Boyle M. (2002) Multi-attribute and single-attribute utility functions for the health utilities index mark 3 system. *Medical Care*, **40**(2), 113–128.
- Frick U., Irving H., Rehm J. (2012) Social relationships as a major determinant in the valuation of health states. *Quality of Life Research*, **21**(2), 209–213.
- Furlong W.J., Feeny D.H., Torrance G.W., Barr R.D. (2001) The Health Utilities Index (HUI) system for assessing health-related quality of life in clinical studies. *Annals of Medicine*, **33**(5), 375–384. DOI: 10.1186/1477-7525-1-54
- Hosmer D., Lemeshow S. (1989) *Applied Logistic Regression*, New York, John Wiley & Sons.
- Krabbe P.F., Essink-Bot M.L., Bonsel G.J. (1997) The comparability and reliability of five health-state valuation methods. *Social Science & Medicine*, **45**(11), 1641–1652. DOI: 10.1016/S0277-9536(97)00099-3
- Lopez A.D., Mathers C.D., Ezzati M., Jamison D.T., Murray C.J.L. (2006a) Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data. *Lancet*, **367**(9524), 1747–1757. DOI: 10.1016/S0140-6736(06)68770-9
- Lopez A.D., Mathers C.D., Ezzati M., Jamison D.T., Murray C.J.L. (2006b) *Global Burden of Disease and Risk Factors*, New York, Oxford University Press.
- Mathers C.D., Vos E.T., Stevenson C.E., Begg S.J. (2001) The burden of disease and injury in Australia. *Bulletin of the World Health Organization*, **79**(11), 1076–1084.
- McCulloch P., Nelder J. (1989) *Generalized Linear Models*, London, Chapman & Hall.
- McIntosh C.N., Connor Gorber S., Bernier J., Berthelot J.M. (2007) Eliciting Canadian population preferences for health states using the Classification and Measurement System of Functional Health (CLAMES). *Chronic Diseases in Canada*, **28**(1–2), 29–41.
- Melse J.M., Essink-Bot M.L., Kramers P.G., Hoeymans N. (2000) A national burden of disease calculation: Dutch disability-adjusted life-years. Dutch Burden of Disease Group. *American Journal of Public Health*, **90**(8), 1241–1247. DOI: org/10.2105/AJPH.90.8.1241
- Murphy K., Connor Gorber S., O'Dwyer A. (2005) *Health State Descriptions for Diabetes*, Ottawa, Statistics Canada.
- Murray C.J.L. (1994) Quantifying the burden of disease: the technical basis for disability-adjusted life years. *Bulletin of the World Health Organization*, **72**(3), 429–445.
- Murray C.J.L. (1996) Rethinking DALYs. In Murray C.J.L., Lopez A. (eds) *The Global Burden of Disease: A Comprehensive Assessment of Mortality and Disability from Diseases, Injuries, and Risk Factors in 1990 and Projected to 2020*, pp. 1–98, Boston, MA, Harvard School of Public Health.
- Murray C.J.L., Lopez A.D. (1996) *The Global Burden of Disease*, Cambridge, MA, Harvard School of Public Health (on behalf of the WHO and World Bank).
- Murray C.J.L., Salomon J., Mathers C., Lopez A. (2002) *Summary Measures of Population Health: Concepts, Ethics, Measurement and Applications*, Geneva, WHO.
- Rabin R., de Charro F. (2001) EQ-5D: a measure of health status from the EuroQoL group. *Annals of Medicine*, **33**, 337–343. DOI: 10.3109/07853890109002087
- Raudenbush S., Bryk A., Cheong Y.F., Congdon R. (2004) *HLM6: Hierarchical Linear and Nonlinear Modeling*, Lincolnwood, IL, Scientific Software International Inc.
- Raudenbush S.W., Bryk A.S. (2002) *Hierarchical Linear Models. Applications and Data Analysis Methods*, London, Sage Publications.
- Rehm J., Frick U. (2010) Valuation of health states in the US study to establish disability weights: lessons from the literature. *International Journal of Methods in Psychiatric Research*, **19**(1), 18–33. DOI: 10.1002/mpr.300
- Ruger J.P. (2005) The changing role of the World Bank in global health. *American Journal of Public Health*, **95**(1), 60–70. DOI: 10.2105/AJPH.2004.042002
- SAS Institute Inc. (2004) *SAS/ETS® 9.1 User's Guide*, Cary, NC, SAS Institute Inc.
- Stata Corporation. (2008) *Stata Statistical Software. Release 10.1*, College Station, TX, Stata Corporation.
- Stiggelbout A., de Vogel-Voogt E. (2008) Health state utilities: a framework for studying the gap between the imagined and the real. *Value in Health*, **11**(1), 76–87. DOI: 10.1111/j.1524-4733.2007.00216.x
- Stouthard M., Essink-Bot M.L., Bonsel G.J. (1997) *Disability Weights for Diseases in the Netherlands*, Rotterdam, Erasmus University Rotterdam.
- Torrance G.M. (1987) Utility approach to measuring health-related quality of life. *Journal of Chronic Diseases*, **40**(6), 593–603.
- Ustun T.B., Rehm J., Chatterji S., Saxena S., Trotter R., Room R., Bickenbach J. (1999) Multiple-informant ranking of the disabling effects of different health conditions in 14 countries. *Lancet*, **354**, 111–115.
- Ware J., Sherbourne C. (1992) The MOS 36-item short form health survey (SF-36). I. Conceptual framework and item selection. *Medical Care*, **30**, 473–483.
- World Health Organization (WHO). (2008) *The Global Burden of Disease: 2004 Update*, Geneva, WHO.
- Wiseman V., Mooney G., Berry G., Tang K. (2003) Involving the general public in priority setting: experiences from Australia. *Social Science & Medicine*, **56**, 1001–1012. DOI: 10.1016/S0277-9536(02)00091-6

Appendix

Table A1 Results from the random effects probit regression models (adjusted for rater and task)

	Health states								
	Labeled			Unlabeled			Combined		
	Coefficient	SE	p Value	Coefficient	SE	p Value	Coefficient	SE	p Value
<i>Pain or discomfort</i>									
Level 2 ¹	0.101	0.091	0.269	0.040	0.046	0.388	−0.167	0.037	<0.001
Levels 3 and 4 ¹	0.621	0.118	<0.001	0.467	0.048	<0.001	0.312	0.042	<0.001
<i>Physical functioning</i>									
Level 2 ²	0.311	0.095	0.001	0.262	0.041	<0.001	0.231	0.036	<0.001
Level 3 ²	−0.108	0.113	0.340	0.306	0.052	<0.001	0.189	0.045	<0.001
Level 4 ²	0.487	0.125	<0.001	0.763	0.056	<0.001	0.618	0.049	<0.001
<i>Emotional state</i>									
Level 2 ³	0.079	0.093	0.399	0.229	0.045	<0.001	0.181	0.038	<0.001
Level 3 ³	0.230	0.097	0.018	0.228	0.043	<0.001	0.114	0.037	0.002
Level 4 ³	0.086	0.184	0.641	0.204	0.059	<0.001	0.191	0.055	<0.001
Level 5 ³	0.017	0.373	0.963	0.741	0.101	<0.001	0.658	0.095	<0.001
<i>Fatigue</i>									
Level 3 ⁴	0.053	0.066	0.426	0.075	0.036	0.039	0.147	0.030	<0.001
Level 4 ⁴	0.307	0.113	0.007	0.234	0.057	<0.001	0.348	0.048	<0.001
<i>Memory and thinking</i>									
Levels 2 and 3 ⁵	−0.272	0.099	0.006	0.026	0.044	0.559	−0.101	0.038	0.008
Level 4 ⁵	−0.282	0.085	0.001	0.126	0.043	0.004	0.069	0.038	0.066
<i>Social relationship</i>									
Level 2 ⁶	0.345	0.079	<0.001	0.330	0.040	<0.001	0.288	0.034	<0.001
Level 3 ⁶	0.735	0.094	<0.001	0.575	0.042	<0.001	0.537	0.038	<0.001
Level 4 ⁶	1.338	0.148	<0.001	0.787	0.057	<0.001	0.795	0.052	<0.001
Level 5 ⁶	1.397	0.419	<0.001	1.185	0.157	<0.001	1.092	0.138	<0.001
<i>Anxiety</i>									
Levels 2, 3 and 4 ⁷	0.543	0.097	<0.001	0.279	0.044	<0.001	0.226	0.037	<0.001
<i>Hearing</i>									
Levels 2 and 3 ⁸	0.689	0.126	<0.001	0.184	0.094	0.050	0.432	0.072	<0.001
Level 4 ⁸	0.439	0.157	0.005	0.979	0.076	<0.001	0.518	0.063	<0.001
<i>Vision</i>									
Levels 2, 3 and 4 ⁹	0.557	0.074	<0.001	0.360	0.033	<0.001	0.385	0.029	<0.001

Note: SE, standard error.

¹Reference category is pain or discomfort (level 1).²Reference category is physical functioning (level 1).³Reference category is emotional state (level 1).⁴Reference category is fatigue (levels 1 and 2).⁵Reference category is memory and thinking (level 1).⁶Reference category is social relationships (level 1).⁷Reference category is anxiety (levels 1).⁸Reference category is hearing (level 1).⁹Reference category is vision (level 1).