Volume 13 • Number 2 • 2010
VALUE IN HEALTH

The Impact of Differences in EQ-5D and SF-6D Utility Scores on the Acceptability of Cost–Utility Ratios: Results across Five Trial-Based Cost–Utility Studies

Manuela Joore, PhD,[1,2] Danielle Brunenberg, MSc,[1] Patricia Nelemans, PhD,[3] Emiel Wouters, MD, PhD,[4] Petra Kuijpers, MD, PhD,[5] Adriaan Honig, MD, PhD,[6] Danielle Willems, MSc,[1] Peter de Leeuw, MD, PhD,[7] Johan Severens, PhD,[1,2] Annelies Boonen, MD, PhD[7]

[1]Department of Clinical Epidemiology and Medical Technology Assessment, Maastricht University Medical Centre, Maastricht, The Netherlands; [2]Department of Health, Organisation, Policy & Economics, Maastricht University Medical Centre, Maastricht, The Netherlands; [3]Department of Epidemiology, Maastricht University Medical Centre, Maastricht, The Netherlands; [4]Department of Pulmonology, Maastricht University Medical Centre, Maastricht, The Netherlands; [5]Department of Psychiatry, Maastricht University Medical Centre, Maastricht, The Netherlands; [6]Department of Psychiatry, Saint Lucas Andreas Hospital Amsterdam, Amsterdam, The Netherlands; [7]Department of Internal Medicine, Maastricht University Medical Centre, Maastricht, The Netherlands

ABSTRACT

Objective: This article investigates whether differences in utility scores based on the EQ-5D and the SF-6D have impact on the incremental cost–utility ratios in five distinct patient groups.

Methods: We used five empirical data sets of trial-based cost–utility studies that included patients with different disease conditions and severity (musculoskeletal disease, cardiovascular pulmonary disease, and psychological disorders) to calculate differences in quality-adjusted life-years (QALYs) based on EQ-5D and SF-6D utility scores. We compared incremental QALYs, incremental cost–utility ratios, and the probability that the incremental cost–utility ratio was acceptable within and across the data sets.

Results: We observed small differences in incremental QALYs, but large differences in the incremental cost–utility ratios and in the probability that these ratios were acceptable at a given threshold, in the majority of the presented cost–utility analyses. More specifically, in the patient groups with relatively mild health conditions the probability of acceptance of the incremental cost–utility ratio was considerably larger when using the EQ-5D to estimate utility. While in the patient groups with worse health conditions the probability of acceptance of the incremental cost–utility ratio was considerably larger when using the SF-6D to estimate utility.

Conclusions: Much of the appeal in using QALYs as measure of effectiveness in economic evaluations is in the comparability across conditions and interventions. The incomparability of the results of cost–utility analyses using different instruments to estimate a single index value for health severely undermines this aspect and reduces the credibility of the use of incremental cost–utility ratios for decision-making.

Keywords: cost–utility, EQ-5D, SF-6D, utility.

Introduction

Instruments that estimate a single index value for health are increasingly used to measure preferences for health states for the estimation of quality-adjusted life-years (QALYs) in cost–utility analyses. These measures are essentially generic health-related quality of life instruments with pre-existing preference weights that can be attached to each permutation of responses. Several widely used instruments that estimate a single index value for health are available, including the EQ-5D [1] and the SF-6D, which uses responses to 11 of the questions on the SF-36 questionnaire [2]. These measures differ in terms of scoring algorithm, and health state descriptive system, and as a result utility scores may vary according to the choice of instrument [3,4]. Indeed, for a large range of clinical conditions there is evidence for differences between the utility estimates of these two instruments for a given patient [5–14]. Moreover, evidence suggests there are differences in the level of agreement between the two instruments over the range of ill health, potentially causing differences in the estimated change in health state utility across patient groups [3].

Address correspondence to: Manuela Joore, Department of Clinical Epidemiology and Medical Technology Assessment, Maastricht University Medical Centre plus. PO Box 5800, 6202 AZ Maastricht, The Netherlands. E-mail: m.joore@mumc.nl

10.1111/j.1524-4733.2009.00669.x

The EQ-5D and SF-6D scoring algorithms are derived using different protocols using different valuation methods (time trade-off and standard gamble, respectively). The literature suggests a cross-over of standard gamble and time trade-off values: Standard gamble values are higher for more severe states, and the opposite applies for the milder states [15]. It has indeed been observed that for milder states the EQ-5D time trade-off utilities were higher than the SF-6D standard gamble utilities [4]. This may partly explain the narrower range of the SF-6D-derived utilities as compared with EQ-5D-derived utilities, which may indicate less sensitivity of the former. With regard to the descriptive system, the operationalization of health in the two instruments is not exactly the same: The "vitality" and "social functioning" dimensions of the SF-6D are not explicitly included in the EQ-5D. This would cause the SF-6D to be more sensitive in situations when impact on these dimensions of health is present. Also, the SF-6D dimensions have more levels than the EQ-5D. Furthermore, because there is evidence for floor effects in the SF-6D and ceiling effects in the EQ-5D, the instruments differ in their description of "full health" and "worse health" [3,4,7]. As a result, the EQ-5D is thought to be sensitive in patient groups with severe health state at baseline, but less sensitive in patient groups with mild health states at baseline. The reverse would apply to the SF-6D.

To our knowledge, two articles reported on the impact of the choice of instrument on the incremental cost–utility ratio. McDonough et al. [16] concluded, based on a review of the

1098-3015/10/222 222–229

literature that across studies the EQ-5 D tended to provide more favorable incremental cost–utility ratios than the SF-6 D. Grieve et al. [17] showed that in the case of a model-based study of antiviral therapy for patients with mild hepatitis using the SF-6D rather than the EQ-5D resulted in more favorable incremental cost–utility ratios. Nevertheless, of key concern is the fact whether the differences in utility estimates have an impact on the estimated impact of health-care interventions, and hence incremental cost–utility ratios, because these are used in the appraisal of a medical technology.

This article investigates whether the differences in utility scores based on the EQ-5D and the SF-6D have impact on the probability that the incremental cost–utility ratio is acceptable in five distinct patient groups. This issue is addressed using five empirical data sets of patients with different disease conditions and different health state severity. By examining the impact of the choice of instrument on whether the incremental cost–utility ratios are acceptable across patient groups this study specifically focuses on the difference in the description of "full health" and "worse health" between the two instruments by examining floor and ceiling effects. The article is structured as follows. We first introduce the data sources we used to investigate the impact of the choice of EQ-5D or SF-6D utility estimates on the incremental cost–utility ratios. Next we describe the EQ-5D and SF-6D instruments and the analyses we performed. In the results section we compare within and across studies: baseline utility scores, incremental QALYs, incremental cost–utility analyses, and the uncertainty surrounding the incremental cost–utility ratios.

Methods

Data Sources

This research concerns secondary analyses of the data from five separate studies conducted in The Netherlands, comprising in total 794 patients. The studies concerned patients with cardiovascular (hypertension), pulmonary (asthma), mental (panic disorder), and musculoskeletal (ankylosing spondylitis and osteoarthritis) disorders. All studies, except for the study with osteoarthritis patients, were cost-effectiveness analyses conducted alongside a randomized clinical trial. The osteoarthritis study was designed as a historical comparison between two matched patient groups. In all studies active treatments were compared.

The study on hypertension concerned outpatients with hypertension who were randomized to either home blood pressure management, or continuance of office blood pressure management [18]. Measurements took place at baseline, 2, 4, 6, and 12 months. The patients with mild asthma (18 years or older, GINA severity stages I to III) were randomized to a nurse-led telemonitoring intervention or care as usual [19]. In this study assessments took place at baseline, and 4, 8, and 12 months. The group of patients with active ankylosing spondylitis received a 3-week course of spa treatment in a spa-resort either in Austria or in The Netherlands, or continued usual care at home [20]. Quality of life and cost data were collected at baseline, and 7, 12, 26, and 52 weeks later. The patients with panic disorder were treated with sertraline or received usual care [21]. Measurements took place at baseline, and 12 and 24 weeks later. Finally, patients with osteoarthritis who underwent a total hip replacement either received a joint recovery program or received care or usual [22]. Quality of life and costs were assessed at baseline, and 4 and 40 weeks later.

In each study both the EuroQol and the Medical Outcomes Study 36-Item Short Form Health Survey (SF-36) were completed at baseline and at each follow-up by the patients themselves. From the data sources seven incremental cost–utility ratios could be calculated: A. home blood pressure management versus care as usual for hypertensive outpatients; B. nurse-led telemonitoring versus care as usual for asthmatics; C. spa treatment in Austria versus spa treatment in The Netherlands, and both spa treatments versus care as usual for ankylosing spondylitis (D and E); F. sertraline versus care as usual for panic disorder; and G. a joint recovery program versus care as usual after hip replacement surgery in osteoarthritis. Care as usual depends, of course, on the specific disease treated and in each study reflected current care for these patients in The Netherlands according to clinical guidelines if available.

Health-Related Quality of Life Measures

The EQ-5D instrument was developed by a European Group as a standard nondisease-specific instrument for describing and valuing quality of life [2]. It is a questionnaire with a descriptive classification system consisting of five dimensions (mobility, self-care, usual activities, pain/discomfort, anxiety/depression), each with three levels. The descriptive system allows for 243 discrete health states. These health states are assigned values, using a tariff based on the time trade-off. Several tariffs are available, but for this study the original UK population tariff was applied in all studies [23]. The EQ-5D utilities range from –0.59 to 1.00.

The SF-36 is a generic health status instrument, comprising eight scales [24]. For the SF-6D, the items of the SF-36 are converted into a six-dimensional health state classification system, with between two and six levels. This yields 18,000 different health states. The health states are assigned preference weights derived from valuations of a sample of 249 SF-6D health states using the standard gamble in a representative sample of the UK population [25]. The SF-6D utilities range from 0.29 to 1.00.

See Supporting information for the scoring algorithms used to calculate utilities from the data sets in this study at: http://www.ispor.org/Publications/value/ViHsupplementary/ViH13i2_Joore.asp

Data Analyses

For this study only patients who completed both the SF-36 and the EQ-5D at baseline were included in the analysis. Floor and ceiling effects in the baseline utility scores were investigated by computing the proportion of patients at the lowest and highest possible utility score. For each patient two outcomes in QALYs were calculated: one based on the EQ-5D utility scores and the other based on the SF-6D utility scores. The patient-level QALYs were estimated by applying the area-under-the-curve method, thus assuming linear change between the discrete follow-up points in time [26]. See Supporting information for details on the QALY calculation at: http://www.ispor.org/Publications/value/ViHsupplementary/ViH13i2_Joore.asp

The time horizon was the study duration. To obtain the incremental QALYs in each study multiple regression analysis was applied to control for differences in baseline utility between the study arms [27]. All analyses were performed using SPSS, version 16.0.

For each study two incremental cost–utility ratios were calculated by dividing the difference in costs between the two alternatives by the difference in both QALYs. To get insight in the uncertainty around the incremental cost-effectiveness ratios nonparametric bootstrap simulations were conducted [28]. In the bootstrap simulations a sample of "costs," "QALY based on EQ-5D," and "QALY based on SF-6D" trios of equal size of the original sample was selected a thousand times with replacement. From these data 95% uncertainty intervals for the differences in

both QALYs were calculated based on the 2.5th and the 97.5th percentile. The difference in the joint distribution of the incremental results is shown in cost-effectiveness planes (CE plane). The difference in decision uncertainty is presented in cost-effectiveness acceptability curves [29]. Cost-effectiveness acceptability curves present uncertainty as the probability that an intervention has the greatest net benefit as a function of the willingness to pay (WTP) for a certain effect (in our case a QALY). The probability that one intervention is preferred over the other is represented graphically in the CE planes as the proportion of the joint density (ΔC, ΔE) to the lower right of a WTP line. A WTP line is a straight line through the origin of the CE plane that connects points with equal WTP values. This proportion can be estimated repeatedly while rotating the WTP line counter clockwise from horizontal (i.e., WTP = 0) to vertical (i.e., WTP = infinite). Hence, the shape of the cost-effectiveness acceptability curve is dependent upon the location of the joint density (ΔC, ΔE) within the CE plane.

Results

Patient Characteristics

In total 583 patients were included in the analyses in the underlying study. The characteristics of the patients included in the five data sets that were considered for this comparative study are shown in Table 1. At baseline, the ranking of the utility scores was the same for the EQ-5D and the SF-6D. The highest utility scores were observed in the hypertension data set (0.842 and 0.773, respectively), and the lowest in the osteoarthritis data set (0.339 and 0.584, respectively). The EQ-5D utility scores were higher than the SF-6D utility scores in the data sets with a relatively high utility score (hypertension, asthma) and lower in the data sets with a relatively low utility score (osteoarthritis). Floor effects were not observed. Ceiling effects were mostly found in the hypertension and asthma data sets, and more prevalent in the EQ-5D than in the SF-6D utility scores. Across the data sets, at baseline the range of EQ-5D utility scores (0.503) was 2.7 times larger than the range of SF-6D utility scores (0.189). Also, the standard deviations of the utility scores at baseline were larger for the EQ-5D utility scores.

Incremental QALYs

The incremental EQ-5D QALY varied from minus 0.011 (the equivalent of 4 days in perfect health lost; osteoarthritis) to 0.055 (the equivalent of 20 days in perfect health gained, spa treatment in Austria versus in The Netherlands for ankylosing spondylitis). For the SF-6D the incremental QALYs varied from minus 0.012 (4 days lost, asthma) to 0.031 (11 days gained, spa treatment in Austria vs. care as usual for ankylosing spondylitis). The 2.5th to 97.5th percentile confidence intervals surrounding the incremental QALYs all included zero, except for the EQ-5D incremental QALY in the asthma and Spa Austria versus care as usual data sets.

The incremental benefit from the new intervention was larger for EQ-5D QALYs than for SF-6D QALYs in the hypertension and asthma data sets. The same was found in the in ankylosing spondylitis data set for the comparison of the Spa treatment in Austria with care as usual and with the Spa treatment in The Netherlands. The largest difference in incremental QALY between the EQ-5D and SF-6D was observed in the asthma data set: 0.045 QALY difference (16 days). The confidence intervals for the EQ-5D incremental QALY were all larger than for the SF-6D incremental QALY.

Table 1 Sample size, patient characteristics, and baseline EQ-5D and SF-6D utility scores

		N		Age	Sex	EQ-5D			SF-6D			
	Data sources	Total	Included	Mean (SD)	% male	Mean (SD)	Floor	Ceiling	Mean (SD)	Floor	Ceiling	*P*-value*
Hypertension	Home BP management	216	124	56 (10)	47	0.821 (0.202)	0%	40%	0.772 (0.124)	0%	2%	0.001
	Care as usual	214	109	56 (10)	56	0.866 (0.191)	0%	54%	0.775 (0.120)	0%	1%	0.000
Asthma	Nurse-led telemonitoring	26	26	46 (15)	42	0.885 (0.128)	0%	50%	0.749 (0.134)	0%	4%	0.000
	Care as usual	27	27	46 (11)	33	0.782 (0.173)	0%	22%	0.695 (0.140)	0%	4%	0.001
Ankylosing spondylitis†	Spa The Netherlands	40	38	47 (10)	63	0.650 (0.224)	0%	3%	0.647 (0.110)	0%	0%	0.905
	Spa Austria	40	35	48 (9)	66	0.639 (0.217)	0%	0%	0.642 (0.114)	0%	0%	0.975
	Care as usual	40	37	48 (10)	84	0.723 (0.100)	0%	5%	0.649 (0.106)	0%	0%	0.000
Panic disorder	Sertraline	49	49	58 (15)	53	0.584 (0.284)	0%	9%	0.622 (0.100)	0%	0%	0.228
	Care as usual	44	44	60 (11)	55	0.601 (0.258)	0%	6%	0.611 (0.085)	0%	0%	0.750
Osteoarthritis of the hip	Joint recovery	48	46	63 (12)	35	0.366 (0.310)	0%	0%	0.603 (0.119)	0%	0%	0.000
	Care as usual	50	48	65 (13)	21	0.268 (0.343)	0%	0%	0.564 (0.117)	0%	0%	0.000

*Paired samples *t* test.
†Comparisons: 1) spa treatment in Austria versus spa treatment in The Netherlands; 2) spa treatment in Austria versus care as usual; and 3) spa treatment in The Netherlands versus care as usual.
BP, blood pressure.

Incremental Cost–Utility Analyses

The point estimates of the incremental cost–utility ratios indicated dominance in the comparison of the two spa treatments for ankylosing spondylitis, irrespective of the utility instrument used. The cost–utility ratios in the hypertension, asthma, and ankylosing spondylitis (spa treatment in Austria vs. usual care) data sets were more acceptable when based on EQ-5D utility. The ratios in the panic disorder, osteoarthritis, and ankylosing spondylitis (spa treatment in The Netherlands vs. usual care) data sets were lower when based on SF-6D utility. The incremental costs, QALYs, and cost–utility ratios are presented in Table 2, ranked from the data set with the highest baseline utility score (A: hypertension) to the data set with the lowest baseline utility score (G: osteoarthritis).

On the CE planes (Fig. 1) it is shown that the uncertainty surrounding the point estimates of the incremental cost–utility ratios in all seven comparisons was larger if the QALY is based on EQ-5D utility. The difference in the probability of acceptance of cost-effectiveness between the EQ-5D and SF-6D is shown in the acceptability curves (Fig. 2: The comparisons are ranked from higher [A] to lower [G] baseline utility score). In Figure 1A, B, F and G the joint density (ΔC, ΔE) is, partly, in the southwest quadrant, indicating less costly and less effective. Therefore, if the WTP increases, more of the joint distribution will fall above the WTP line, and the probability of acceptance decreases. Therefore the cost-effectiveness acceptability curves in Figure 2A, B, F and G are falling. For ceiling ratios between €0 and €80,000 per QALY the smallest differences between EQ-5D- and SF-6D-derived cost per QALY were observed in the ankylosing spondylitis data. The largest differences were observed in the asthma and panic disorder data (Fig. 2B and F). At a ceiling ratio of €40.000 per QALY the probability of acceptance of nurse-led telemonitoring for asthma was 0.55 larger when using EQ-5D utility estimates to calculate QALYs (Fig. 2B). In the panic disorder data set it is the other way around. At €40,000 per QALY the probability of accepting sertraline was 0.45 larger when using SF-6D utility estimates to calculate QALYs (Fig. 2F). By "reading" Figure 2 from left to right and from top to bottom, it is shown that first (higher baseline utility scores; A, B) the probability that the intervention is cost-effective is higher based on the EQ-5D, while later (lower baseline utility scores; F, G) the probability is higher when based on SF-6D utility scores.

Discussion and Conclusions

We investigated the impact of the differences in utility scores based on the EQ-5D and the SF-6D on the probability that incremental cost–utility ratios were acceptable, in five distinct patient groups. Our main findings are the following. First, EQ-5D utility scores were higher than SF-6D utility scores for disease states with higher baseline utility scores and lower for states with lower baseline utility scores. This is in line with previous evidence: Healthier individuals tend to have higher mean scores on the EQ-5D, and less healthy individuals tend to have higher scores on the SF-6D [30]. Also, the considerable ceiling effects observed in the EQ-5D utility scores in the data sets with relatively mild conditions (hypertension, asthma) are in line with the literature [7]. Floor effects were not observed. Second, the observed differences in incremental QALY were very small, and mainly occurred in the trials with baseline utility scores at either end of the health spectrum. The EQ-5D provided more favorable incremental cost–utility ratios for data sets with higher baseline utility and the opposite in the data sets with lower baseline utility. In light of the observed ceiling effects, which are expected to decrease sensitivity to change, in mainly the data sets

Table 2 Incremental costs, QALYs, and cost–utility ratios based on the EQ-5D and SF-6D utility estimates

Trial		Comparison*	ΔCosts	ΔQALY					Incremental cost–utility ratio											
				EQ-5D			SF-6D		EQ-5D					SF-6D						
			Mean	Mean	Percentiles		Mean	Percentiles	ICUR	Distribution (%)				ICUR	Distribution (%)					
			€		2.5th	97.5th		2.5th	97.5th	€/QALY (95% CI)	NE	NW	SW	SE	€/QALY (95% CI)	NE	NW	SW	SE
Hypertension	A	Home BP management vs. CAU	45	0.014	−0.025	0.054	−0.004	−0.028	0.022	3,156 (inferior–833)	36	17	6	41	Inferior (inferior–2,045)	17	36	27	20	
Asthma	B	Nurse-led telemonitoring vs. CAU	1163	0.033	0.004	0.071	−0.012	−0.070	0.030	35,180 (291,583–16,380)	96	4	0	1	Inferior (inferior–38,767)	17	83	0	0	
Ankylosing spondylitis	C	Spa Austria vs. The Netherlands	−209	0.044	−0.008	0.101	0.010	−0.024	0.043	Dominant (26,125–dominant)	37	2	3	58	Dominant (8,708–dominant)	29	10	18	43	
	D	Spa Austria vs. CAU	1269	0.055	0.006	0.106	0.031	−0.003	0.065	23,198 (211,500–11,972)	99	1	0	0	40,879 (inferior–19,523)	96	4	0	0	
	E	Spa The Netherlands vs. CAU	1478	0.011	−0.046	0.064	0.021	−0.008	0.054	131,903 (inferior–23,093)	62	37	0	0	68,920 (inferior–27,370)	90	10	0	0	
Panic disorder	F	Sertraline vs. CAU	−1089	−0.014	−0.041	0.012	0.003	−0.007	0.013	77,785 (26,780–dominant)	4	17	70	9	Dominant (155,571–dominant)	17	5	26	53	
Osteoarthritis of the hip	G	Joint recovery program vs. CAU	−821	−0.011	−0.107	0.081	0.009	−0.030	0.049	73,265 (7,673–dominant)	4	15	43	38	Dominant (27,367–dominant)	11	8	22	59	

*The comparisons are ranked from higher (A) to lower (G) baseline utility score.
CAU, care as usual; CI, confidence interval; ICUR, incremental cost–utility ratio; NE, northeast; NW, northwest (inferior); QALY, quality-adjusted life-year; SE, southeast (dominant); SW, southwest.

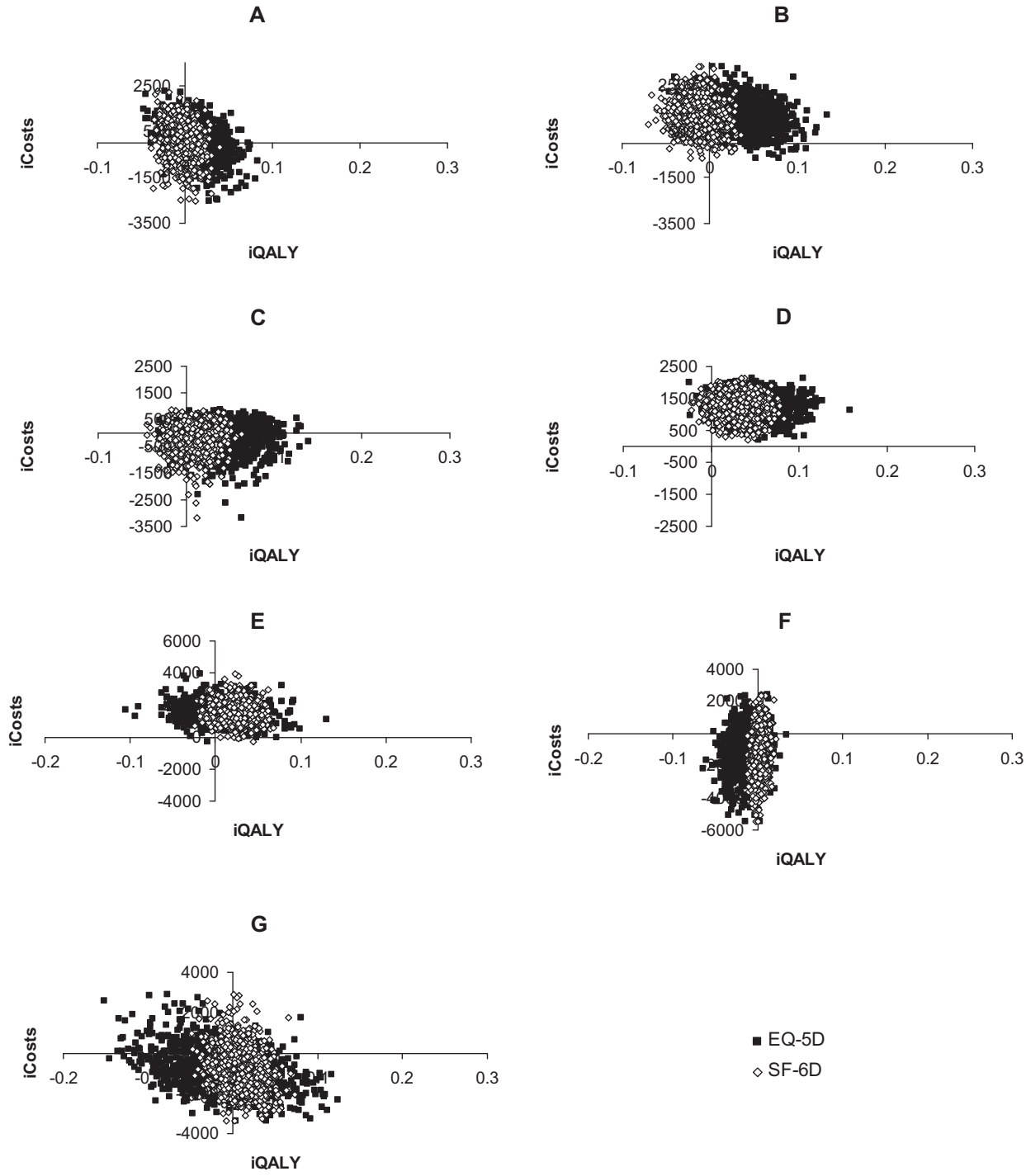

Figure 1 Distribution of incremental cost–utility results based on EQ-5D and SF-6D quality-adjusted life-year (QALY) estimates for each comparison ranked from higher (A) to lower (G) baseline utility score (€/QALY). (A) Hypertension trial; (B) asthma trial; (C) ankylosing spondylitis trial (spa Austria versus spa The Netherlands); (D) ankylosing spondylitis trial (spa Austria vs. care as usual); (E) ankylosing spondylitis trial (spa The Netherlands vs. care as usual); (F) panic disorder trial; (G) osteoarthritis of the hip trial.

with higher baseline utility, this finding seems somewhat counterintuitive. Third, the uncertainty surrounding the incremental cost–utility ratios was larger when using EQ-5D utility scores in each data set. This is not surprising, taking into account the considerably larger standard deviation of EQ-5D utility scores. Fourth, the probability that the incremental cost–utility ratios were acceptable was considerably larger when using EQ-5D utility scores in the data sets with higher baseline utility scores; although the opposite was found in the data sets with lower baseline utility scores.

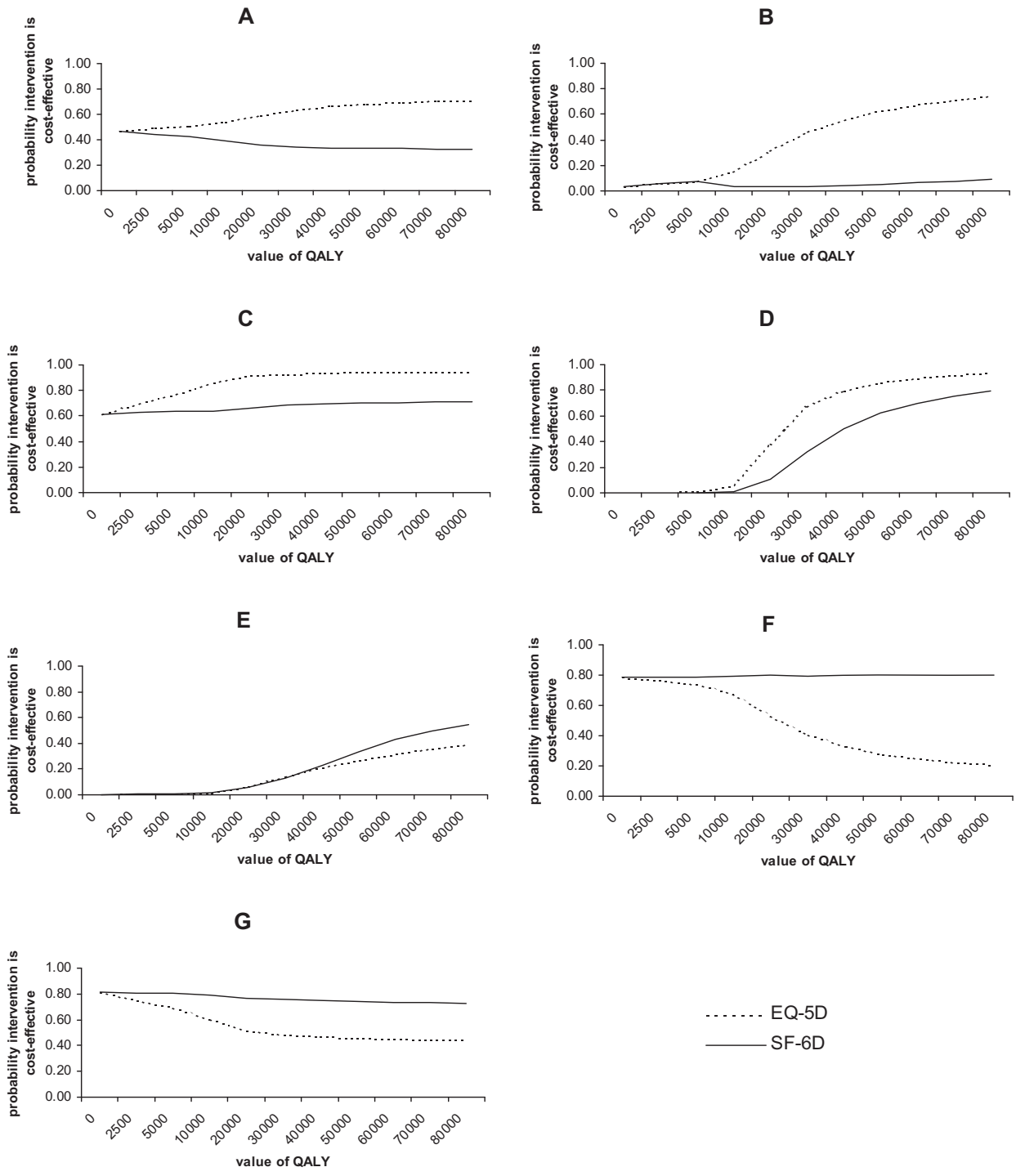

Figure 2 Cost-effectiveness acceptability curves based on EQ-5D and SF-6D quality-adjusted life-year (QALY) estimates for each comparison, ranked from higher (A) to lower (G) baseline utility score. (A) Hypertension trial; (B) asthma trial; (C) ankylosing spondylitis trial (spa Austria vs. spa The Netherlands); (D) ankylosing spondylitis trial (spa Austria vs. care as usual); (E) ankylosing spondylitis trial (spa The Netherlands vs. care as usual); (F) panic disorder trial; (G) osteoarthritis of the hip trial.

Addressing our main question: Even though the differences in incremental QALYs were rather small, the choice of instrument had considerable impact on the probability that incremental cost–utility ratios are acceptable. Moreover, across the patient groups and comparisons we included in this study, we found that the cost–utility ratios were more acceptable when using EQ-5D in relatively mild health conditions and using SF-6D in relatively serious health conditions. This result does not confirm the find-

ings of McDonough et al. [16], who concluded, based on a review of the literature, that across studies the EQ-5D tended to provide more favorable incremental cost–utility ratios. This result also is not in line with the expectation that the ceiling effects in the EQ-5D would lead to more favorable EQ-5D results in severe conditions.

The data sets we used reflect a considerable severity range and very different areas in health: musculoskeletal disease, cardiovascular pulmonary disease, and psychological disorders. Nevertheless, it certainly was a convenience sample of studies as available in our department. The systematic difference of the choice for EQ-5D or SF-6D on whether incremental cost–utility ratios are acceptable we observed can be a result of the specific sample of studies we used. For instance, we did not observe floor effects in the SF-6D, although others did [3,31–32]. This may be the reason why in the data set with a relatively serious condition (osteoarthritis), somewhat counter intuitively, the SF-6D utility scores translated into a larger probability that the intervention was cost-effective. If floor effects would be present in the SF-6D baseline utility scores, this result could be reversed. In addition, in the data sets we used only small differences in QALYs were observed between the interventions. Although this is a rather common finding when comparing a new intervention with the best available alternative, it would favor the instrument that overall is more sensitive to change. Taking into account the differences in health state description and scoring algorithm between the EQ-5D and SF-6D it is not clear which instrument would overall be more sensitive to change. It is expected that this will differ between patient groups and interventions.

The findings of this study suggest that besides the differences in the definition of worse and full health, other sources of differences in change in utility score as measured with EQ-5D and the SF-6D play a role. Therefore, it is of great importance that we further improve our understanding of the impact of the choice of the utility instrument on the probability that an incremental cost–utility ratio is acceptable. If feasible, we like to recommend the use of more than one utility instrument in trial-based economic evaluations to obtain as much comparative data as possible. In addition, we strongly encourage researchers to publish any available cost-effectiveness data in which two or more instruments are used to estimate a single index value for health. More explicitly, because in this area the burden of evidence arises from a series of analyses and not a single study, repeating our analyses for other conditions and interventions with a different type and magnitude of effect seems worthwhile. In addition, other sources of differences between the instruments need to be investigated. For instance the differences in domain structure may have impact on utility change. Furthermore, differences in utility change may arise from differences in the interval properties of the utility scales of the EQ-5D and SF-6D.

In conclusion, we observed small differences in incremental QALYs, but remarkably large differences in the probability that the incremental cost–utility ratio is acceptable in the majority of the presented cost–utility analyses. More specifically, in the patient groups with relatively mild health conditions the probability of acceptance of the incremental cost–utility ratio was considerably larger when using the EQ-5D to estimate utility. While in the patient groups with relatively serious health conditions the probability of acceptance of the incremental cost–utility ratio was considerably larger when using the SF-6D to estimate utility. A systematic difference in the probability of accepting the cost–utility of interventions as a result of the choice of utility instrument would seriously bias the comparability of the results of economic evaluations. This is problematic, because much of the appeal in using QALYs as measure of effectiveness in economic evaluations is in the comparability across conditions and interventions. The incomparability of the results of cost–utility analyses using different utility instruments reduces the credibility of the use of incremental cost–utility ratios for decision-making.

An earlier version of this work was presented at the International Health Economics Association Conference, July 2007, Copenhagen, Denmark. Two anonymous reviewers are kindly acknowledged for their valuable comments.

Source of financial support: The data in this study are taken from studies funded by The Netherlands Organisation of Health Research and Development, the Dutch Health Care Insurance Board, the Land Salzburg, the Gasteiner Tal Tourismusgesellschaft, the Kurzentrum Thermentempel, the Gasteiner Heilstollen from Austria, Zorgvoorzieningen Nederlands NV, IZA Zorgverzekering, Dick van Toll Assurantieen BV, and Yakult BV Netherlands.

References

1 EuroQol Group. EuroQol: a new facility for the measurement of health related quality of life. Health Policy 1990;16:199–208.
2 Brazier J, Usherwood T, Harper R, Thomas K. Deriving a preference-based single index from the UK SF-36 Health Survey. J Clin Epidemiol 1998;51:115–28.
3 Brazier J, Roberts J, Tsuchiya A, Busschbach J. A comparison of the EQ-5D and SF-6D across seven patient groups. Health Econ 2004;13:873–84.
4 Tsuchiya A, Brazier J, Roberts J. Comparison of valuation methods used to generate the EQ-5D and the SF-6D value sets. J Health Econ 2006;25:334–46.
5 Marra CA, Esdaile JM, Guh D, et al. A comparison of four indirect methods of assessing utility values in rheumatoid arthritis. Med Care 2004;42:1125–31.
6 Barton GR, Bankart J, Davis AC. A comparison of the quality of life of hearing-impaired people as estimated by three different utility measures. Int J Audiol 2005;44:157–63.
7 Bharmal M, Thomas J 3rd. Comparing the EQ-5D and the SF-6D descriptive systems to assess their ceiling effects in the US general population. Value Health 2006;9:262–71.
8 McDonough CM, Grove MR, Tosteson TD, et al. Comparison of EQ-5D, HUI, and SF-36-derived societal health state values among Spine Patient Outcomes Research Trial (SPORT) participants. Qual Life Res 2005;14:1321–32.
9 Stavem K, Froland SS, Hellum KB. Comparison of preference-based utilities of the 15D, EQ-5D and SF-6D in patients with HIV/AIDS. Qual Life Res 2005;14:971–80.
10 Van Stel HF, Buskens E. Comparison of the SF-6D and the EQ-5D in patients with coronary hearth disease. Health Qual Life Outcomes 2006;4:20–8.
11 Bryan S, Longworth L. Measuring health related utility: why the disparity between EQ-5D and SF-6D? Eur J Health Econ 2005;6: 253–60.
12 Marra CA, Woolcott JC, Kopec JA, et al. A comparison of generic, indirect utility measures (the HUI2, HUI3, SF-6D, and the EQ-5D) and disease-specific instruments (the RAQol and the HAQ) in rheumatoid arthritis. Soc Sci Med 2005;60:1571–82.
13 Szende A, Svensson K, Stahl E, et al. Psychometric and utility-based measures of health status of asthmatic patients with different disease control level. Pharmacoeconomics 2004;22:537–47.
14 Longworth L, Bryan S. An empirical comparison of the EQ-5D and SF-6D in liver transplant patients. Health Econ 2003;12: 1061–7.
15 Dolan P, Gudex C, Kind P, Williams A. Valuing health states: a comparison of methods. J Health Econ 1996;15:209–31.
16 McDonough CM, Tosteson ANA. Measuring preferences for cost-utility analysis. How choice of method may influence decision making. Pharmacoeconomics 2007;25:93–106.
17 Grieve R, Grishchenko M, Cairns J. SF-6D versus EQ-5D: reasons for differences in utility scores and impact on reported cost-utility. Eur J Health Econ 2009;10:15–23.

18 Verberk WJ, Kroon AA, Kessels AG, et al. Home versus office blood pressure measurements: reduction of Unnecessary treatment Study: rationale and study design of the HOMERUS trial. Blood Press 2003;12:326–33.
19 Willems DCM, Joore MA, Hendriks JJE, et al. Cost-effectiveness of a nurse-led telemonitoring intervention based on peak expiratory flow measures in asthmatics: results of a randomised controlled trial. Cost Eff Resour Alloc 2007;5:10.
20 Van Tubergen A, Boonen A, Landewé R, et al. Cost-effectiveness of combined spa exercise therapy in ankylosing spondylitis. Arthritis Rheum 2002;47:459–67.
21 Honig A, Dirksen C, Kuijpers P, Brunenberg D. Panic attacks and Chest Pain: Effectiveness of Medication [in Dutch]. November 2002, report. Maastricht: University Hospital Maastricht, 2002.
22 Brunenberg D, van Steyn M, Sluimer J, et al. Joint recovery programme versus usual care: an economic evaluation of a clinical pathway for joint replacement surgery. Med Care 2005;43:1018–26.
23 Dolan P. Modelling valuations for EuroQol health states. Med Care 1997;35:1095–108.
24 Ware J, Snow K, Kosinski M, Gandek B. SF-36 Health Survey Manual and Interpretation Guide. Boston, MA: New England Medical Center, The Health Institute, 1993.
25 Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. J Health Econ 2002;21:271–92.
26 Matthews JNS, Altman D, Campbell MJ. Analysis of serial measurements in medical research. Br Med J 1990;300:230–5.
27 Manca A, Hawkins N, Sculpher MJ. Estimating mean QALYs in trial-based cost-effectiveness analysis: the importance of controlling for baseline utility. Health Econ 2005;14:487–96.
28 Briggs AH, Wonderling DE, Mooney CZ. Pulling cost-effectiveness analysis up by its bootstraps: a non-parametric approach to confidence interval estimation. Health Econ 1997;6:327–40.
29 Hout van BA, Al MJ, Gordon GS, Rutten FF. Costs, effects and C/E-ratios alongside a clinical trial. Health Econ 1994;3:309–19.
30 Barton GR, Sach TH, Avery AJ, et al. A comparison of the performance of the EQ-5D and SF-6D for individuals aged ≥45 years. Health Econ 2008;17:815–32.
31 Ferreira PL, Ferreira LN, Pereira LN. How consistent are health utility values? Qual Life Res 2008;17:1031–42.
32 Davison SN, Jhangri GS, Feeny DH. Comparing the Health Utilities Index Mark 3 (HUI3) with the Short Form-36 Preference-Based SF-6D in chronic kidney disease. Value Health 2009;12:340–5.