Quality of Life Research
https://doi.org/10.1007/s11136-024-03830-w

Valuation survey for SF-6Dv2 in Japan based on the international protocol

Takeru Shiroiwa[1] · Yosuke Yamamoto[2] · Tatsunori Murata[3] · Brendan Mulhern[4] · Jakob Bjorner[5] · John Brazier[6] · Takashi Fukuda[1] · Donna Rowen[6] · Shun-Ichi Fukuhara[7]

Accepted: 31 October 2024

Abstract

Purpose The SF-6D Classification System was recently updated (SF-6Dv2). We performed a valuation survey to construct a value set for the SF-6Dv2 in Japan.

Methods An online discrete choice experiment (DCE) with duration was used to estimate a value set for the SF-6Dv2 for Japan based on public preferences. The target sample number was 3800. Respondents were asked to complete 15 choice tasks. A conditional logit model that estimates interactions between time and each dimension was used to develop the value set.

Results The collected sample included 3933 respondents for the DCE tasks. The results of all the unconstrained models showed some inconsistencies. In particular, inconsistencies in the two most severe levels of the role limitation (RL) and vitality (VT) dimensions were observed in all models. The number of inconsistencies was smallest in a core model (n = 3) and in a model for core and common health states (n = 2). The physical functioning (PF) and pain (PA) dimensions had the greatest influence on utility at the overall level across all models. RL, VT, and social functioning (SF) had smaller overall impacts on utility. The PF weights for the two most severe levels are much lower than those in the UK and Australia. The Japanese scores tended to be lower compared with the UK SF-6Dv2 scores.

Conclusion We obtained a value set for Japan (model 5). With the development of this value set, it is now possible to calculate quality-adjusted life years for economic evaluation in Japan when the SF-6Dv2 has been used.

Keywords Utility · Quality of life · SF-6D · QALY · Health technology assessment

Background

An economic evaluation generally calculates quality-adjusted life years (QALYs) to measure the efficiency of healthcare technologies. Health technology assessment (HTA) agencies, including the National Institute for Health and Care Excellence (NICE) in the UK, ask pharmaceutical (and medical device) companies to submit cost-effectiveness data using QALYs [1]. In such countries, measurement of QALYs is important not only for academic researchers but also for pharmaceutical and medical device companies because it influences the reimbursement or pricing of pharmaceuticals and medical devices. Japan has the same situation as other countries. The Japanese government enacted a new pricing system in 2019 that uses economic evaluation to recalculate pharmaceutical or medical device prices [2]. The Japanese HTA organization, Center for Outcomes Research and Economic Evaluation for Health (C2H), requests QALY-based outcome data for cost-effectiveness analysis [3].

✉ Takeru Shiroiwa
t.shiroiwa@gmail.com

1 Center for Outcomes Research and Economic Evaluation for Health (C2H), National Institute of Public Health, 2-3-6 Minami, Wako, Saitama 351-0197, Japan

2 Department of Healthcare Epidemiology, School of Public Health in the Graduate School of Medicine, Kyoto University, Kyoto, Japan

3 Crecon Medical Assessment Co., Ltd., Tokyo, Japan

4 Centre for Health Economics Research and Evaluation, University of Technology Sydney, Sydney, NSW, Australia

5 QualityMetric, an IQVIA Company, Johnston, RI, USA

6 Sheffield Centre for Health and Related Research, University of Sheffield, Sheffield, UK

7 Section of Clinical Epidemiology, Department of Community Medicine, Graduate School of Medicine, Kyoto University, Kyoto, Japan

Published online: 27 November 2024

QALYs can be calculated by multiplying life years by quality of life, or utility, weights which are anchored on a scale of 0 (death) and 1 (full health) where values below zero reflect that the health state is considered as being worse than dead. Preference-based measures (PBMs) or preference-weighted measures are generally used to provide the utility weights for QALYs. PBMs include a set of dimensions defining health states and a value set including weights for every health state described. Value sets are derived using a preference elicitation method, and are usually country specific (for example, many PBMs have Japanese value sets [4–11], including widely used generic measures such as the EQ-5D-5L, and SF-6Dv1).

SF-6Dv1 is a generic PBM developed in the UK [12]. SF-6D consists of six domains [physical functioning (PF), role limitations (RL), pain (PA), vitality (VT), social functioning (SF), and mental health (MH)] that can be scored from the SF-36 Health Survey. A Japanese value set for the SF-6Dv1 was developed by Fukuhara and colleagues [6]. The SF-6Dv1 scores can be derived from 11 SF-36v1 or SF-36v2 items. The valuation survey of SF-6Dv1 was based on the standard gamble (SG) method where respondents trade a risk of death or severely impaired health to avoid impaired health [13, 14]. The use of SG is sometimes criticized because the respondents' risk aversion leads to relatively higher values for severe states. For example, the value for the most severe health state for the SF-6Dv1 in the UK is 0.29. Therefore, in some countries, DCE-based value sets have been published [15] using DCE with duration (i.e. a profile consisting of a health states experienced for a specified number of life years) rather than the SG method. In another Australian study, the value of the worst health states using DCE with duration was − 0.36 to − 0.44 depending on the model. Based on these and other results, the DCE method was applied to the SF-6Dv2. The SF-6Dv2 [16] health state classification system was developed to improve on the SF-6Dv1. It consists of the same six dimensions as SF-6Dv1, but with changes to the descriptors. Valuation surveys of the SF-6Dv2 have already been completed in the UK [17], Australia [18], China [19] and the US to generate country-specific value sets.

The primary objective of this study was to perform a valuation survey of the SF-6Dv2 based on an international protocol that included three DCE designs and to obtain a Japanese value set. The Japanese preference for each item in a PBMs is sometimes quantitatively and qualitatively different from that of Western countries [4–11]. Consequently, it is not appropriate to apply an existing value set developed in other countries. Actually, C2H requests the use of a value set that "reflects the preferences of the general population in Japan". In our survey, we used the DCE with duration method to elicit the value set. The DCE method has been increasingly used for valuation surveys, including cancer-specific EORTC QLU C-10D [20], and FACT-8D [21]. Second, the Japanese value set was compared with those in the UK, Australia, and China, where published SF-6Dv2 value sets exist.

Methods

SF-6Dv2 classification system

The SF-6Dv2 is a classification system comprising six dimensions: physical functioning (PF), role limitations (RL), pain (PA), vitality (VT), social functioning (SF), and mental health (MH), with five to six severity levels (only PA has six levels) Similar to the SF-6Dv1, the SF-6Dv2 scores can be derived from SF-36v2 items. The SF-6Dv2 can also be scored from an independent six item instrument, the SF-6Dv2 Health Utility Survey (HUS) [22]. The Japanese version of SF-6Dv2 HUS was established by the research team. The Japanese team drafted the translation of SF-6Dv2 HUS to be consistent with existing Japanese SF-36 translation, back-translated into English for review by UK team. After that, cognitive debriefing was performed for 10 Japanese people. Considering and reflecting the feedback from the cognitive debriefing, the final Japanese version of SF-6Dv2 HUS was completed.

Discrete choice experiment

We used DCE with duration for valuing SF-6Dv2 health states. In the DCE survey, participants were required to imagine hypothetical health states, which consisted of health states derived from the SF-6Dv2 classification system and life years (1, 4, 7, and 10 years). Subsequently, two health states (states A and B) were presented, and the participants chose the one they preferred between the two options. In addition, we used the ternary method, in which three health states (states A, B and "immediate death") were shown to respondents, who were asked to identify what they thought was the best and the worst health state.

Survey process and design

Respondents were asked to choose their preferred profile for each choice set. A total of 15 choice sets were presented, consisting of three training tasks, two "common tasks", eight core tasks and two ternary tasks. The two common tasks were randomly selected from a set of 76 choice tasks across 38 blocks based on health states that are commonly experienced by the general population. These choice tasks used health states selected from the 200 most common health states identified in general population surveys. The choice set was identified using

the Fedorov algorithm implemented in NGene. Regarding core tasks, respondents were randomly allocated to a set of 304 core tasks across 38 blocks, which were selected among all of the health states described by the SF-6Dv2. As before, the choice set was constructed on the basis of the Federov algorithm. Two ternary tasks were randomly selected from a set of 76 choice tasks that include a third choice of immediate death. In contrast to normal DCE tasks, respondents were asked to select the best and worst health states from the three options.

These three types of tasks were presented in order. Two pairs (common), eight pairs (core), and two pairs (ternary) were randomly allocated to each participant from each of the 38 blocks. In each task, the order in which the questions were presented was randomized and the presentation positions (left or right) of the two health states were randomized to avoid a positioning effect.

The sample size of 3800 was chosen to match the power of the original UK study, which used a sample size of 3000 respondents, a set of 300 core choice tasks, and 60 ternary tasks. The UK respondents were grouped into 30 subgroups of 100 respondents that each answered 10 core choice tasks and 2 ternary tasks. In the current study, that each answered 2 common choice tasks, 8 core tasks, and 2 ternary tasks for a total of 76 common choice tasks, 304 core tasks and 76 ternary tasks.

Survey participants

An online survey was also conducted. Respondents (aged 20–79) were recruited through a Japanese web panel based on quota sampling by sex and age to represent the general population. This means that an equal number of respondents were collected from the 12 groups [age categories (20–29, 30–39, 40–49, 50–59, 60–69, 70–79) multiplied by sex categories]. If the target number of respondents was included in the survey in one group, the recruitment for the group was closed. Respondents were invited to this survey by an email and asked to click the link if they wanted to join the survey. Respondents had to provide informed consent to proceed to complete the survey. Background information on respondents was collected after 15 DCE tasks were completed. Respondents who completed all the tasks could obtain a small incentive. When the required number of responses was collected, the web page for the survey was closed.

The inclusion criteria were as follows: (a) being aged 20 years and over (definition of "adult" citizens in Japan), (b) currently living in Japan, (c) providing informed consent, (d) possessing literacy skills in Japanese, and (e) having access to a device with an internet connection. The survey was conducted in March 2022.

Statistical analysis

We calculated the number and percentage of background factors. A conditional logit model was used for the analysis of the choice tasks. The model for the estimation of coefficients was based on Bansback et al. [23] and Norman et al. [24] and included continuous duration (time) and the interaction between duration and the severity of each dimension (with the least sever level, level 1, as the baseline). Let t be the duration, and u_{ij} be the utility of profile j for individual i. In that case, u_{ij} can be formulated as follows:

$$U_{ij} = \beta_1 t_{ij} + \beta_2 x_{ij} t_{ij} + \varepsilon_i \quad (1)$$

where ε_{ij} is an error term. However, the estimated β_2 is not anchored on the 0 (death) to 1 (full health) scale. To change the latent coefficients to the disutility of each level, we can calculate the utility weight using the following equation:

$$-\hat{\beta}_2/\hat{\beta}_1 \quad (2)$$

In the immediate death profile of the ternary tasks, duration was treated as 0. We also included an interaction term (WORST) to assess the impact of the worst level of each dimension in the analysis. If the profile had one or more than one dimension at the worst level, the WORST term was defined as 1 (the "worst" model). If the estimated disutility was not logically consistent (consistency implied that "weights at the higher level in the same dimension were higher, and those at the lower level were lower"), inconsistent levels were combined and the dataset was analyzed by the same models.

We analyzed four different subsamples of the data and 9 models. Model 1 included only core task responses (eight tasks per respondent, from the total of 304 included in the design) for analysis without a worst term. Model 3 included only the core task responses, but included a worst term. Model 4 included eight core tasks and two common tasks (10 tasks per respondent). Model 6 included the eight core tasks and two ternary tasks (10 tasks per respondent). Finally, model 8 included eight core tasks, two common tasks, and two ternary tasks (12 tasks per respondent). Corresponding to each of the above models, a constrained model was applied if inconsistencies were observed (models 2, 5, 7 and 9). The only exception was model 3, where the number of inconsistencies was deemed to be too high to attempt a constrained model. The parameters were estimated using *Phreg* in SAS 9.4 and *clogit* in STATA 17. These two approaches gave the same results. We compared the models using log likelihood, number of logical inconsistencies (where as severity increases utility increases), coefficients of each level and distribution of utilities. To obtain the distribution of all utilities that can be generated by SF-6Dv2,

utilities of $5^5*6 = 18{,}750$ health states were calculated using the parameter estimates for each level of each dimension.

This study was approved by the ethics committee of the National Institute of Public Health, to which the first author belongs (NIPH-IBRA #12338).

Results

The collected sample included 3933 respondents for the DCE tasks. No respondents were excluded. The mean and median total response times of the respondents to the 15 DCE questions were 13.8 min (standard deviation (SD):54.3) and 7.4 min (interquartile range (IQR) 4.7–11.2 min), respectively. The maximum time of the response was 1346.3 min. It was assumed that some of the responses to the DCE survey were interrupted. To exclude these outliers, people with response times greater than 60 min were excluded from this calculation (resulting in an N = 3861), which changed the mean response time to 9.0 min (SD:7.0) but these participants were retained for modelling purposes.

Demographic factors

Respondents' background characteristics are presented in Table 1. The median household income ranged from JPY 5 to 7 million. When compared with the average household incomes of all Japanese families of JPY 4.4 million in 2019 [25], the household income was slightly higher. According to the 2019 Labor Force Survey [26], full-time workers and part-time workers accounted for 31.6 and 13.7%, respectively. In total, 24.3% of Japanese individuals had graduated from a university or graduate school in 2017, and 61.3 and 31.6% were married and unmarried, respectively, in 2015. Overall the sample appears representative of the Japanese population for sex and age but is more educated with a higher proportion of employed individuals.

Results of the analysis for models 1, 2 and 3

Table 2 shows the coefficients of the analysis. The results of which indicated three inconsistencies for Model 1 in levels 1 (baseline)/2 and levels 4/5 in the RL dimension and levels 4/5 in the VT dimension. Model 2 is a consistent version of Model 1 where adjacent levels were combined, namely into baseline level 1 and level 2 for both RL and VT, and levels 4 and 5 for both RL and VT. Model 3 includes the interaction term, WORST, and its results showed six logical inconsistencies. The log likelihood was similar among these two models. As the fit of model 3 was not good in terms of level consistencies, a worst term was not used in subsequent analyses.

Table 1 Background factors of respondents

	Number	Percentage
Sex		
Male	1968	50.0
Female	1965	50.0
Age		
20–29	652	16.6
30–39	651	16.6
40–49	661	16.8
50–59	661	16.8
60–69	649	16.5
70–79	659	16.8
Region		
Hokkaido/Tohoku	383	9.7
Kanto	1579	40.2
Chubu	627	15.9
Kinki	759	19.3
Chugoku/Shikoku	445	11.3
Kyushu	140	3.6
Employment		
Full-time worker	1508	38.3
Part-time worker	536	13.6
Self employed	241	6.1
Housemaker	810	20.6
Retired	615	15.6
Student	158	4.0
Others	65	1.7
Education		
Elementary or Junior high school	80	2.0
High school	1145	29.1
College	768	19.5
University	1733	44.1
Postgraduate	200	5.1
Others	7	0.2
Marital status		
Unmarried	1467	37.3
Married	2170	55.2
Divorced/Bereaved	296	7.5
Household income (JPY 1mil)		
< 1	159	4.0
1 < = < 3	649	16.5
3 < = < 5	922	23.4
5 < = < 7	608	15.5
7 < = < 10	562	14.3
10 < = < 15	280	7.1
15 < = < 20	64	1.6
20 > =	43	1.1
Unknown	646	16.4

Table 2 Estimated coefficients by different models

		Model 1 (Core, Unconstrained)		Model 2 (Core, Constrained)		Model 3 (Core, Worst)		Model 4 (Core + Common, Unconstrained)		Model 5 (Core + Common, Constrained)		Model 6 (Core + Triplet, Unconstrained)	
		N = 3933											
		Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
Time		*0.348*	0.009	*0.349*	0.008	*0.322*	0.009	*0.365*	0.007	*0.364*	0.007	*0.301*	0.008
PF x Time	2	*− 0.022*	0.005	*− 0.023*	0.005	*− 0.022*	0.005	*− 0.030*	0.004	*− 0.031*	0.004	*− 0.014*	0.005
	3	*− 0.032*	0.005	*− 0.033*	0.005	*− 0.034*	0.005	*− 0.038*	0.004	*− 0.039*	0.004	*− 0.027*	0.004
	4	*− 0.120*	0.005	*− 0.119*	0.005	*− 0.122*	0.005	*− 0.120*	0.004	*− 0.119*	0.004	*− 0.091*	0.004
	5	*− 0.217*	0.005	*− 0.218*	0.005	*− 0.204*	0.005	*− 0.214*	0.005	*− 0.216*	0.005	*− 0.162*	0.004
RL x Time	2	**0.004**	0.005	0.000		**0.005**	0.005	− 0.007	0.004	− 0.006	0.004	− 0.003	0.005
	3	*− 0.035*	0.005	*− 0.038*	0.003	*− 0.034*	0.005	*− 0.045*	0.004	*− 0.045*	0.004	*− 0.023*	0.004
	4	*− 0.062*	0.005	*− 0.058*	0.005	*− 0.062*	0.005	*− 0.067*	0.004	*− 0.059*	0.004	*− 0.041*	0.004
	5	***− 0.053***	0.005	*− 0.058*	0.005	***− 0.036***	0.005	***− 0.054***	0.004			***− 0.034***	0.004
PA x Time	2	*− 0.028*	0.005	*− 0.027*	0.005	*− 0.022*	0.005	*− 0.032*	0.004	*− 0.031*	0.004	*− 0.035*	0.005
	3	*− 0.036*	0.005	*− 0.034*	0.005	*− 0.030*	0.005	*− 0.043*	0.005	*− 0.042*	0.005	*− 0.036*	0.005
	4	*− 0.082*	0.005	*− 0.082*	0.007	*− 0.082*	0.005	*− 0.088*	0.004	*− 0.089*	0.004	*− 0.063*	0.004
	5	*− 0.179*	0.005	*− 0.179*	0.004	*− 0.175*	0.005	*− 0.181*	0.005	*− 0.181*	0.004	*− 0.138*	0.004
	6	*− 0.192*	0.007	*− 0.191*	0.003	*− 0.185*	0.007	*− 0.191*	0.007	*− 0.190*	0.007	*− 0.159*	0.005
VT x Time	2	− 0.000	0.005	*0.000*		**0.004**	0.005	− 0.008	0.004	− 0.007	0.004	*− 0.017*	0.004
	3	*− 0.013*	0.005	*− 0.013*	0.005	− 0.008	0.005	*− 0.026*	0.004	*− 0.026*	0.004	**− 0.006**	0.004
	4	*− 0.041*	0.005	*− 0.036*	0.005	*− 0.034*	0.005	*− 0.048*	0.005	*− 0.042*	0.004	*− 0.037*	0.004
	5	***− 0.031***	0.005			***− 0.011***	0.005	***− 0.035***	0.005			***− 0.024***	0.004
SF x Time	2	− 0.008	0.005	− 0.007	0.005	− 0.004	0.005	*− 0.018*	0.004	*− 0.017*	0.004	− 0.003	0.004
	3	*− 0.021*	0.005	*− 0.020*	0.005	*− 0.014*	0.005	*− 0.025*	0.004	*− 0.024*	0.004	**0.000**	0.004
	4	*− 0.037*	0.005	*− 0.037*	0.005	*− 0.038*	0.005	*− 0.038*	0.004	*− 0.038*	0.004	*− 0.017*	0.004
	5	*− 0.044*	0.004	*− 0.042*	0.005	***− 0.026***	0.005	*− 0.044*	0.004	*− 0.042*	0.004	***− 0.014***	0.004
MH x Time	2	− 0.008	0.005	− 0.007	0.005	*− 0.011*	0.005	*− 0.023*	0.004	*− 0.022*	0.004	− 0.004	0.004
	3	*− 0.031*	0.005	*− 0.031*	0.005	*− 0.028*	0.005	*− 0.036*	0.004	*− 0.037*	0.004	*− 0.015*	0.004
	4	*− 0.069*	0.005	*− 0.070*	0.005	*− 0.071*	0.005	*− 0.070*	0.005	*− 0.071*	0.004	*− 0.039*	0.004
	5	*− 0.075*	0.005	*− 0.076*	0.005	***− 0.057***	0.005	*− 0.076*	0.005	*− 0.078*	0.005	*− 0.044*	0.004
WORST						*− 0.331*	0.025						
Number of observation		62,928		62,928		62,928		78,660		78,660		94,392	
Log likelihood		− 18,443.0		− 18,448.4		− 18,355.0		− 22,969.0		− 22,977.9		− 29,530.4	
Number of inconsistency		3		0		6		2		0		5	

Table 2 (continued)

		Model 7 (Core + Triplet, Constrained)		Model 8 (Core + Common + Triplet, Unconstrained)		Model 9 (Core + Common + Triplet, Constrained)	
		Estimate	SE	Estimate	SE	Estimate	SE
Time		*0.299*	0.008	*0.331*	0.007	*0.331*	0.007
PF x Time	2	*− 0.016*	0.005	*− 0.025*	0.004	*− 0.026*	0.004
	3	*− 0.029*	0.004	*− 0.035*	0.004	*− 0.036*	0.004
	4	*− 0.091*	0.004	*− 0.095*	0.004	*− 0.096*	0.004
	5	*− 0.162*	0.004	*− 0.166*	0.004	*− 0.166*	0.004
RL x Time	2	− 0.002	0.005	*− 0.012*	0.004	*− 0.011*	0.004
	3	*− 0.022*	0.004	*− 0.031*	0.004	*− 0.031*	0.004
	4	*− 0.036*	0.004	*− 0.047*	0.004	*− 0.040*	0.003
	5			***− 0.038***	0.004		
PA x Time	2	*− 0.031*	0.005	*− 0.033*	0.004	*− 0.031*	0.004
	3	*− 0.033*	0.005	*− 0.042*	0.004	*− 0.041*	0.004
	4	*− 0.061*	0.004	*− 0.069*	0.004	*0.069*	0.004
	5	*− 0.136*	0.004	*− 0.141*	0.004	*− 0.140*	0.004
	6	*− 0.156*	0.005	*− 0.161*	0.005	*− 0.159*	0.005
VT x Time	2	*− 0.010*	0.004	*− 0.021*	0.004	*− 0.018*	0.004
	3			***− 0.017***	0.004		
	4	*− 0.030*	0.004	*− 0.044*	0.004	*− 0.036*	0.004
	5			***− 0.030***	0.004		
SF x Time	2	− 0.002	0.004	*− 0.016*	0.004	*− 0.011*	0.003
	3			**− 0.006**	0.004		
	4	*− 0.015*	0.003	*− 0.021*	0.004	*− 0.020*	0.003
	5			***− 0.019***	0.004		
MH x Time	2	− 0.003	0.004	*− 0.020*	0.004	*− 0.020*	0.004
	3	*− 0.016*	0.004	*− 0.020*	0.004	*− 0.021*	0.004
	4	*− 0.039*	0.004	*− 0.044*	0.004	*− 0.043*	0.004
	5	*− 0.046*	0.004	*− 0.049*	0.004	*− 0.051*	0.004
WORST							
Number of observation		86,526		110,124		110,124	
Log likelihood		− 27,480.8		− 34,073.8		− 35,702.8	
Number of inconsistency		0		5		0	

Bold inconsistent coefficient, *Italic* $P < 0.05$; *PF* physical functioning, *RL* role limitations, *PA* pain, *VT* vitality, *SF* social functioning, *MH* mental health

Results of the analysis for models 4 and 5

Model 4 included data from eight core tasks and two common tasks. Table 2 shows the coefficients analyzed using models 4 and 5. The results showed only two inconsistencies (level 4/5 of the RL and VT domains). These levels are constrained in model 5.

Results of the analysis for models 6, 7, 8, and 9

Models 6 and 7 used data from eight core tasks and two ternary tasks. Models 8 and 9 used data from eight core tasks, two common tasks, and two ternary tasks. Models 6 and 8 have the same logically inconsistent levels; levels 4/5 of RL, and levels 2/3 and levels 4/5 of the VT and SF dimensions. These logically inconsistent and adjacent levels are constrained to generate consistent models in models 7 and 9.

Anchored results

Figure 1 shows the utility weights for value sets, estimated from the coefficients of the constrained models using Eq. (2). The PF and PA dimensions were the most influential dimensions of utility in all models at the overall level (i.e. the utility decrement associated with level 5). In contrast, the RL, VT, and SF dimensions had smaller coefficients than the other dimensions. The differences of weights between PF/PA domains and RL/VT/SF are large. The calculated value of the worst state from the potential value sets ranged from of − 0.782 (model 2), − 0.722 (model 5), − 0.488 (model 7) and − 0.426 (model 9). The worst state is − 0.574 in the UK and − 0.685 in Australia (based on the WORST constrained model). The second-highest score [121111] was calculated to be 1.000 (model 2), 0.980 (model 5), 0.993 (model 7) and 0.967 (model 9). Figure 2 shows the distribution of utility of all health states described by SF-6Dv2 (the distribution of results from models 2, 5, 7 and 9 compared with the result from the preferred WORST model of SF-6Dv2 in the UK and the EQ-5D-5L in Japan). Scores by models including ternary tasks (model 7 and model 9) were higher than models 2 and 5 which do not include ternary tasks. In comparison with model 2 and model 5, the distribution of the UK SF-6Dv2 model moved to the right.

Figure 3 compares the UK and Japanese SF-6Dv2 scores (based on model 5), by calculating SF6Dv2 scores of all the

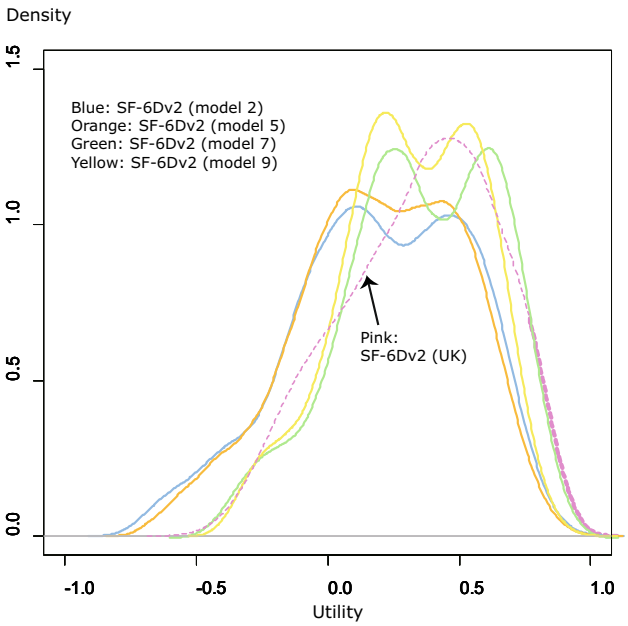

Fig. 2 Distribution of Japanese SF-6Dv2

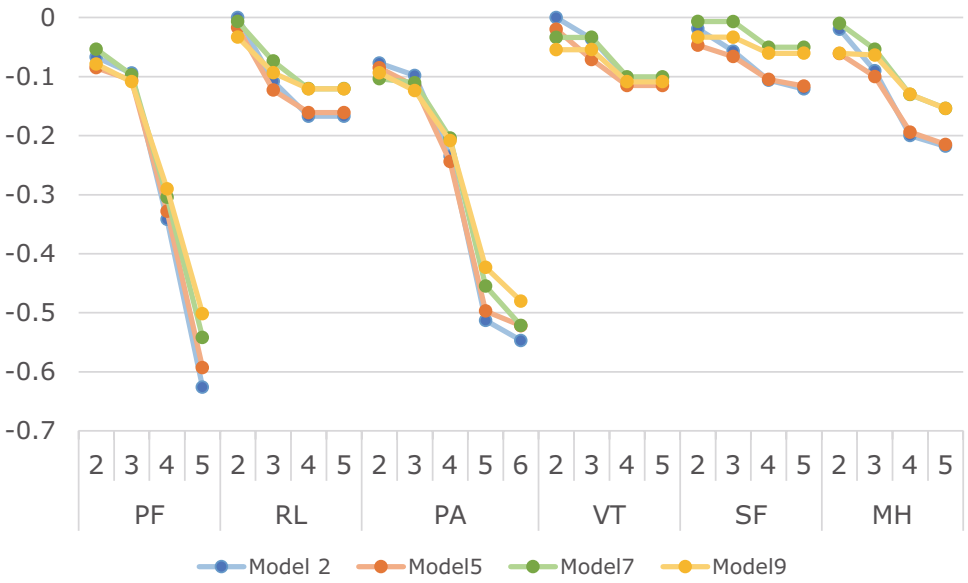

Fig. 1 Coefficients in each constrained model. *PF* physical functioning, *RL* role limitations, *PA* pain, *VT* vitality, *SF* social functioning, *MH* mental health

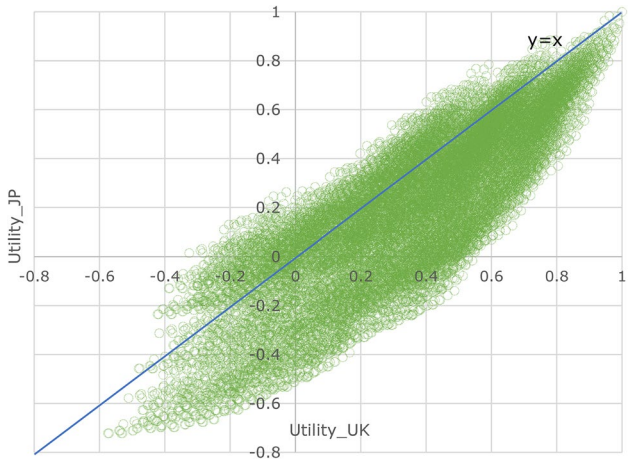

Fig. 3 Comparison of Japanese and the UK SF-6Dv2 scores

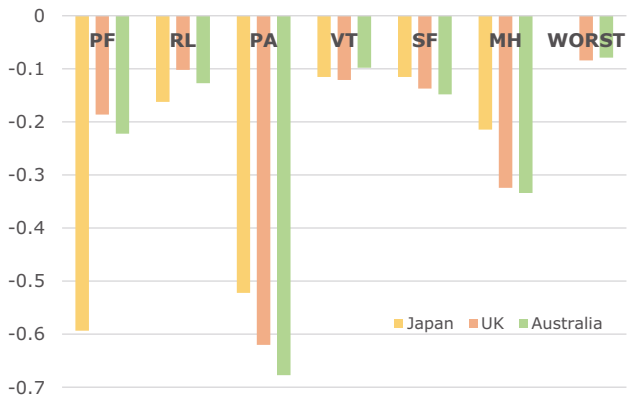

Fig. 4 Comparison of the worst level between Japan, the UK and Australia. *PF* physical functioning, *RL* role limitations, *PA* pain, *VT* vitality, *SF* social functioning, *MH* mental health

health states by both value sets. Out of all possible SF-6Dv2 health states, the Japanese utility values were lower than those of the UK in 77.9% of cases. Figure 4 shows the utility decrements of the worst level in each domain among the three countries (The UK, Australia and Japan). The Japanese utility decrements for Level 5 in PF are significantly larger than those in the UK and Australia.

Discussion

In this study, we used data from a large sample of respondents from the general Japanese population to estimate a value set for the SF-6Dv2 based on an international protocol using DCE. The value set obtained using Model 5 can now be used for cost-effectiveness analyses in Japan. According to the Japanese value set of the SF-6Dv1, the score of the worst state was 0.392, which was much larger than that of other PBMs, including the EQ-5D-5L. Although the valuation methods differed between the two studies (standard gamble in SF-6Dv1 and DCE in SF-6Dv2), the worst score of SF-6Dv2 was − 0.722 (Model 5). The problem of measurement range improved.

The results calculated by all unconstrained models revealed some logical inconsistencies, where as health state severity increases utility increases. Inconsistencies in the RL and VT dimensions for levels 4 and 5 were observed in all the unconstrained models, which suggests that the Japanese respondents did not distinguish between Levels 4 and 5 of the RL and VT dimensions. The preference weights of Levels 4 and 5 in the PF dimension and those of Levels 5 and 6 in the PA dimension are considerably larger in terms of their impact on utilities than the other weights. They have considerable influence on the range of the Japanese value set. Especially, compared with the UK and Australian weights, it is noteworthy that the Japanese utility decrements for levels 4 and 5 in PF are quite large (level 4: − 0.327 (Japan, model 5), − 0.092 (the UK) [17], − 0.138 (Australia) [18] and level 5: − 0.593 (Japan), − 0.186 (the UK) [17] and − 0.222 (Australia) [18]), although the UK and Australia uses the WORST model in which the weight of the worst is − 0.084 (the UK) and − 0.079 (Australia) and this is not included here with the exception of model 3. However, the coefficients of the PA and MH dimension are higher than those for the UK and Australian weights. For Japan, in contrast with UK and Australia, the lowest weight of the PF dimension in model 5 is lower than that of the PA dimension. The Chinese data showed a similar tendency in that the utility decrements of PF and PA were small, but particularly so for the PA dimension. In the case of the Japanese value set of the EQ-5D-5L [4], the utility decrement of the worst level of mobility (Mo) was the largest (− 0.243), although those of pain/discomfort (Pd) and anxiety/depression (Ad) were − 0.191 and − 0.196, respectively. Mo was the most influential item on utility, but Pd was comparable to Ad. In contrast, Devlin et al. [27] indicated that the decrease in Pd was the largest (− 0.335), and that of Ad was the second largest (− 0.289). The coefficient of Mo is − 0.274. These findings may partly result from cultural differences between other countries and Japan, where physical independence is more valued and partly from the characteristics of the SF-6Dv2.

The minimum scores obtained by all the models were lower than that of the Japanese EQ-5D-5L (− 0.025). Although the scores of the Japanese EQ-5D-5L are much higher than those of the EQ-5D-5L in almost all other countries, the scores of the Japanese SF-6Dv2 are low compared to the UK scores. The reason may be that the valuation method of SF-6Dv2 is DCE with duration, and that of EQ-5D-5L is time trade-off (TTO) where Japanese people tend to avoid choosing immediate death. Moreover, DCE with duration trades expected life years, and

the trading of death is not explicit. In contrast the ternary tasks do include a direct trade of death and impaired health. When we included data from ternary tasks in our analyses, the worst possible scores did increase (− 0.488 and − 0.426). These results support the hypothesis that Japanese people tend to avoid choosing immediate death.

The new Japanese guidelines for economic evaluation in 2024 recommend using EQ-5D-5L ("8.2.1 The Japanese version of the EQ-5D-5L is recommended as the initial choice for the PBM.") [3]. However, the guidelines also accept the use of other generic PBMs including SF-6Dv2 as the second choice (Data collected using a generic Japanese PBM with a Japanese value set other than the EQ-5D-5L). Therefore, developing a Japanese value set for SF-6Dv2 is important because increasing the number of PBMs with Japanese value sets is helpful for academia and decision-makers. A PBM can lack sensitivity or responsiveness when measuring the utility of certain conditions or diseases. Different PBMs reflect different aspects of health states in terms of utility. However, it is also essential to consider comparability among PBMs, especially for decision-makers.

Considering the number of logical inconsistencies in the coefficients, models 1 and 4 had a few inconsistencies in the coefficients, but these were remodeled with ordering imposed to produce consistent versions, Models 2 and 5. Model 4 showed only two inconsistencies: levels 4 and 5 of the RL and VT dimensions. Model 1 showed an additional inconsistency in the RL dimension between level 1 (baseline) and level 2. The absence of this inconsistency in model 4 may be due to the inclusion of data from the common design, which provides more statistical power for the estimation of utility decrements for mild heath problems. In the UK and Australia, the WORST model (Model 3 in our report) was preferred, but in Japan, model 3 had a high number of inconsistencies. For these reasons, we recommend the constrained version of model 4 (i.e. model 5) to be used for scoring the Japanese SF-6Dv2.

With the development of the value set of the Japanese SF-6Dv2, it is now possible to calculate QALYs for economic evaluation using SF-6Dv2. The value set is based on the results of an online survey completed by 3933 members of the Japanese public, and web-survey was well-controlled. One limitation of this study was the sampling method. This web survey, and recruiting from an existing web panel, does not allow respondents to be chosen randomly across Japan. In addition, this survey was performed during the latter stages of the outbreak of COVID-19. The influence of the COVID-19 outbreak, which could have changed the preferences for health states, is unknown. Compared with the numerically large weights for PF and PA dimensions, the weights were numerically smaller for other dimensions, especially RL, VT and SF.

Finally, our statistical model makes the following three assumptions: (a) linear time preference [28], (b) independence from irrelevant alternatives (IIA), and (c) a multiplicative utility function (health state × duration) [29]. According to Jonker et al. [28], the assumption of linear time preference (without discounting) is not valid; the estimated discount rate is larger than that normally used by HTA agencies, and the hyperbolic discount function is better fitted than the exponential one. However, we did not consider these time preferences in the survey. For example, a mixed logit model can ease this assumption; however, we used only a fixed model. Jonker et al. [29] showed that many respondents' choices were based on the additive utility function that does not differ from the multiplicative utility function, which is the model assumption of Bansback et al. [23]. If the respondents violated this assumption, the estimated value sets were biased; however, we analyzed the DCE data based on the multiplicative assumption. If reflecting non-linear time preference, the absolute value of the utility coefficients of the SF-6D becomes smaller [28]; in contrast, considering only respondents with a multiplicative utility function, those values become larger [29]. We do not empirically predict which influences on utility are severe; however, our estimates of utility decrements may have been affected by these factors.

Some aspects of the Japanese SF-6Dv2 have not yet been clarified because experiences with SF-6Dv2 use have not accumulated to a sufficient degree. For example, the relationship between the SF-6Dv2 and other PBMs is unknown. Moreover, the population norms [30, 31] of SF-6Dv2 may help interpret obtained data for both the general population and specific patient groups. Further studies may be required to address these issues. Nevertheless, the present study contributes to promoting and enabling economic evaluations in Japan.

Acknowledgements If projects are implemented solely in Japan, the Japanese version of the SF-6Dv2 is available royalty-free for non-funded academic users, after registration by sending your email to os@qualitest.jp. The following URL can be used for registration of international projects; https://www.qualitymetric.com/health-surveys/sf-6dv2-license-request.

Author contributions Conceptualization: [Takeru Shiroiwa, Yosuke Yamamoto, Tatsunori Murata, Brendan Mulhern, Jakob Bjorner, John Brazier, Takashi Fukuda, Donna Rowen, Shun-Ichi Fukuhara]; Methodology: [Takeru Shiroiwa, Tatsunori Murata, Brendan Mulhern, Jakob Bjorner, John Brazier]; Formal analysis and investigation: [Takeru Shiroiwa, Tatsunori Murata]; Writing—original draft preparation: [Takeru Shiroiwa]; Writing—review and editing: [Takeru Shiroiwa, Yosuke Yamamoto, Tatsunori Murata, Brendan Mulhern, Jakob Bjorner, John Brazier, Takashi Fukuda, Donna Rowen, Shun-Ichi Fukuhara],; Funding acquisition: [Takeru Shiroiwa]; Resources: [Takeru Shiroiwa]; Supervision: [John Brazier, Takashi Fukuda, Shun-Ichi Fukuhara].

Funding This study was funded by National Institute of Public Health in Japan.

Availability of data and materials The datasets generated and/or analyzed during the current study are not publicly available due to the lack of consent from participants, but are available from the corresponding author upon reasonable request.

Code availability Yes.

Declarations

Conflicts of interest The authors have no conflicts of interests to declare that are relevant to the content of this article.

Ethical approval This study was performed in line with the principles of the Declaration of Helsinki. The ethics committee of the National Institute of Public Health, to which the first author belongs (NIPH-IBRA #12338) approved.

Consent to participate Informed consent was obtained from all individual participants included in the study.

Consent for publication Participants signed informed consent regarding publishing their anonymous data.

References

1. National Institute of Health and Care Excellence (2022). NICE health technology evaluations: the manual. Retrieved November 14, 2024, from https://www.nice.org.uk/process/pmg36
2. Shiroiwa, T. (2020). Cost-effectiveness evaluation for pricing medicines and devices: A new value-based price adjustment system in Japan. *International Journal of Technology Assessment in Health Care, 36*(3), 270–276. https://doi.org/10.1017/s0266462320000264
3. Center for Outcomes Research and Economic Evaluation for Health (C2H) (2024). Guideline for preparing cost-effectiveness evaluation to the central social insurance medical council. Retrieved November 14, 2024, from https://c2h.niph.go.jp/tools/guideline/guideline_en_2024.pdf
4. Shiroiwa, T., Fukuda, T., Ikeda, S., Igarashi, A., Noto, S., Saito, S., & Shimozuma, K. (2016). Japanese population norms for preference-based measures: EQ-5D-3L, EQ-5D-5L, and SF-6D. *Quality of Life Research, 25*(3), 707–719. https://doi.org/10.1007/s11136-015-1108-2
5. Noto, S., Shiroiwa, T., Kobayashi, M., Murata, T., Ikeda, S., & Fukuda, T. (2020). Development of a multiplicative, multi-attribute utility function and eight single-attribute utility functions for the Health Utilities Index Mark 3 in Japan. *Journal of Patient-Reported Outcomes, 4*(1), 23. https://doi.org/10.1186/s41687-020-00188-8
6. Brazier, J. E., Fukuhara, S., Roberts, J., Kharroubi, S., Yamamoto, Y., Ikeda, S., Doherty, J., & Kurokawa, K. (2009). Estimating a preference-based index from the Japanese SF-36. *Journal of Clinical Epidemiology, 62*(12), 1323–1331. https://doi.org/10.1016/j.jclinepi.2009.01.022
7. Shiroiwa, T., Ikeda, S., Noto, S., Fukuda, T., & Stolk, E. (2021). Valuation survey of EQ-5D-Y based on the international common protocol: Development of a value set in Japan. *Medical Decision Making, 41*(5), 597–606. https://doi.org/10.1177/0272989x211001859
8. Shiroiwa, T., Moriyama, Y., Nakamura-Thomas, H., Morikawa, M., Fukuda, T., Batchelder, L., Saloniki, E. C., & Malley, J. (2020). Development of Japanese utility weights for the Adult Social Care Outcomes Toolkit (ASCOT) SCT4. *Quality of Life Research, 29*(1), 253–263. https://doi.org/10.1007/s11136-019-02287-6
9. Shiroiwa, T., Nakamura-Thomas, H., Yamaguchi, M., Morikawa, M., Moriyama, Y., Fukuda, T., Allan, S., & Malley, J. (2022). Japanese preference weights of the Adult Social Care Outcomes Toolkit for Carers (ASCOT-Carer). *Quality of Life Research, 31*(7), 2143–2151. https://doi.org/10.1007/s11136-021-03076-w
10. Shiroiwa, T., King, M. T., Norman, R., Müller, F., Campbell, R., Kemmler, G., Murata, T., Shimozuma, K., & Fukuda, T. (2024). Japanese value set for the EORTC QLU-C10D: A multi-attribute utility instrument based on the EORTC QLQ-C30 cancer-specific quality-of-life questionnaire. *Quality of Life Research*. https://doi.org/10.1007/s11136-024-03655-7
11. Tsuchiya, A., Ikeda, S., Ikegami, N., Nishimura, S., Sakai, I., Fukuda, T., Hamashima, C., Hisashige, A., & Tamura, M. (2002). Estimating an EQ-5D population value set: The case of Japan. *Health Economics, 11*(4), 341–353. https://doi.org/10.1002/hec.673
12. Brazier, J., Usherwood, T., Harper, R., & Thomas, K. (1998). Deriving a preference-based single index from the UK SF-36 Health Survey. *Journal of Clinical Epidemiology, 51*(11), 1115–1128. https://doi.org/10.1016/s0895-4356(98)00103-6
13. Brazier, J., Roberts, J., & Deverill, M. (2002). The estimation of a preference-based measure of health from the SF-36. *Journal of Health Economics, 21*(2), 271–292. https://doi.org/10.1016/s0167-6296(01)00130-8
14. Kharroubi, S., Brazier, J. E., & O'Hagan, A. (2007). Modelling covariates for the SF-6D standard gamble health state preference data using a nonparametric Bayesian method. *Social Science and Medicine, 64*(6), 1242–1252. https://doi.org/10.1016/j.socscimed.2006.10.040
15. Norman, R., Viney, R., Brazier, J., Burgess, L., Cronin, P., King, M., Ratcliffe, J., & Street, D. (2014). Valuing SF-6D health states using a discrete choice experiment. *Medical Decision Making., 34*(6), 773–786. https://doi.org/10.1177/0272989x13503499
16. Brazier, J. E., Mulhern, B. J., Bjorner, J. B., Gandek, B., Rowen, D., Alonso, J., Vilagut, G., & Ware, J. E. (2020). Developing a new version of the SF-6D health state classification system from the SF-36v2: SF-6Dv2. *Medical Care, 58*(6), 557–565. https://doi.org/10.1097/mlr.0000000000001325
17. Mulhern, B. J., Bansback, N., Norman, R., & Brazier, J. (2020). Valuing the SF-6Dv2 classification system in the United Kingdom using a discrete-choice experiment with duration. *Medical Care, 58*(6), 566–573. https://doi.org/10.1097/mlr.0000000000001324
18. Mulhern, B., Norman, R., & Brazier, J. (2021). Valuing SF-6Dv2 in Australia using an international protocol. *PharmacoEconomics, 39*(10), 1151–1162. https://doi.org/10.1007/s40273-021-01043-4
19. Xie, S., Wu, J., & Chen, G. (2022). Discrete choice experiment with duration versus time trade-off: A comparison of test-retest

reliability of health utility elicitation approaches in SF-6Dv2 valuation. *Quality of Life Research*. https://doi.org/10.1007/s11136-022-03159-2
20. King, M. T., Viney, R., Simon Pickard, A., Rowen, D., Aaronson, N. K., Brazier, J. E., Cella, D., Costa, D. S., Fayers, P. M., Kemmler, G., & McTaggart-Cowen, H. (2018). Australian utility weights for the EORTC QLU-C10D, a multi-attribute utility instrument derived from the cancer-specific quality of life questionnaire, EORTC QLQ-C30. *PharmacoEconomics, 36*(2), 225–238. https://doi.org/10.1007/s40273-017-0582-5
21. King, M. T., Norman, R., Mercieca-Bebber, R., Costa, D. S. J., McTaggart-Cowan, H., Peacock, S., & Cella, D. (2021). The functional assessment of cancer therapy eight dimension (FACT-8D), a multi-attribute utility instrument derived from the cancer-specific FACT-general (FACT-G) quality of life questionnaire: Development and Australian value set. *Value Health, 24*(6), 862–873. https://doi.org/10.1016/j.jval.2021.01.007
22. Broderick, L., Bjorner, J. B., Lauher-Charest, M., White, M. K., Kosinski, M., Mulhern, B., & Brazier, J. (2022). Development of the SF-6Dv2 health utility survey: comprehensibility and patient preference. *Journal of Patient-Reported Outcomes, 6*(1), 47. https://doi.org/10.1186/s41687-022-00455-w
23. Bansback, N., Brazier, J., Tsuchiya, A., & Anis, A. (2012). Using a discrete choice experiment to estimate health state utility values. *Journal of Health Economics, 31*(1), 306–318. https://doi.org/10.1016/j.jhealeco.2011.11.004
24. Norman, R., Mulhern, B., & Viney, R. (2016). The impact of different DCE-based approaches when anchoring utility scores. *PharmacoEconomics, 34*(8), 805–814. https://doi.org/10.1007/s40273-016-0399-7
25. Ministry of Health Labour and Welfare. (2020). Comprehensive Survey of Living Conditions.
26. Statistics Bureau of Japan. (2020). Labour Force Survey.
27. Devlin, N. J., Shah, K. K., Feng, Y., Mulhern, B., & van Hout, B. (2018). Valuing health-related quality of life: An EQ-5D-5L value set for England. *Health Economics, 27*(1), 7–22. https://doi.org/10.1002/hec.3564
28. Jonker, M. F., Donkers, B., de Bekker-Grob, E. W., & Stolk, E. A. (2018). Advocating a paradigm shift in health-state valuations: The estimation of time-preference corrected QALY tariffs. *Value Health, 21*(8), 993–1001. https://doi.org/10.1016/j.jval.2018.01.016
29. Jonker, M. F., & Norman, R. (2022). Not all respondents use a multiplicative utility function in choice experiments for health state valuations, which should be reflected in the elicitation format (or statistical analysis). *Health Economics, 31*(2), 431–439. https://doi.org/10.1002/hec.4457
30. Shiroiwa, T., Ikeda, S., Noto, S., Igarashi, A., Fukuda, T., Saito, S., & Shimozuma, K. (2016). Comparison of value set based on DCE and/or TTO data: scoring for EQ-5D-5L health states in Japan. *Value Health, 19*(5), 648–654. https://doi.org/10.1016/j.jval.2016.03.1834
31. Shiroiwa, T., Noto, S., & Fukuda, T. (2021). Japanese population norms of EQ-5D-5L and health utilities index mark 3: Disutility catalog by disease and symptom in community settings. *Value Health, 24*(8), 1193–1202. https://doi.org/10.1016/j.jval.2021.03.010