Qual Life Res (2014) 23:2355–2363
DOI 10.1007/s11136-014-0670-3

Comparison between the EQ-5D-5L and the EQ-5D-3L in patients with hepatitis B

Y. X. Jia · F. Q. Cui · L. Li · D. L. Zhang · G. M Zhang · F. Z. Wang · X. H. Gong · H. Zheng · Z. H. Wu · N. Miao · X. J. Sun · L. Zhang · J. J. Lv · F. Yang

Accepted: 6 March 2014 / Published online: 14 March 2014

Abstract

Objectives The purpose of the study was to compare psychometric properties of the EQ-5D-5L (5L) and the EQ-5D-3L (3L) health outcomes assessment instruments in patients with hepatitis B in China.

Methods Patients, including hepatitis B virus carriers and those with active or inactive chronic hepatitis B, compensated cirrhosis, decompensated cirrhosis or hepatocellular carcinoma, answered a questionnaire composed of 5L, socio-demographic information, 3L, and the visual analog scale (VAS), respectively. After 1 week, a retest was conducted for inpatients. We compared acceptability, face validity, redistribution properties, convergent validity, known-group validity, discriminatory power, ceiling effect, test–retest reliability, and responsiveness of 5L and 3L.

Results A total of 369 outpatients and 276 inpatients were recruited for the first interview. Of the inpatients, 183 were used in the retest. Most patients preferred 5L–3L. The 3L–5L response pairs had an inconsistency rate of 2.4 %. Correlation with the VAS was greater with 5L than with 3L. Age, education, and comorbidity were associated with health-related quality of life (HRQoL). 5L discriminated more infectious conditions than 3L. In all dimensions, the Shannon's index from 5L was larger while in three dimensions the Shannon's evenness index from 5L was slightly larger. The ceiling effect was reduced in 5L. In patients with stable health states, no significant difference was detected in the weighted kappa between 5L and 3L, but intraclass correlation coefficient of 5L was higher than that of 3L. In patients with improved health states, HRQoL was seen as increased in both 5L and 3L, without significant difference.

Conclusions The EQ-5D-5L was more suitable than the EQ-5D-3L in the patients with hepatitis B in China.

Keywords EQ-5D · Health-related quality of life · Psychometrics · Hepatitis B

Introduction

The EQ-5D is an instrument used to measure health outcomes and health-related quality of life (HRQoL) in cost-effectiveness analysis [1–4]. The original version of EQ-5D, EQ-5D-3L (3L), has been used successfully in many countries. Its validation and psychometric properties have been explored and demonstrated in normal populations and in patients with various diseases, including hepatitis B [5–10].

Although the EQ-5D-3L has benefited from a condensed format, studies have shown that the 3L version suffers from deficiency in descriptive richness and from serious ceiling effects, especially in mild disease states [11–13]. In order to reduce these problems, a new version (EQ-5D-5L, 5L) was

Y. X. Jia · F. Q. Cui (✉) · L. Li · G. MZhang ·
F. Z. Wang · X. H. Gong · H. Zheng · Z. H. Wu ·
N. Miao · X. J. Sun
National Immunization Program, Chinese Center for Disease Control and Prevention, Nanwei Road, Xicheng District, Beijing, China
e-mail: cuifuq@126.com

D. L. Zhang
Ningbo Center for Disease Control and Prevention, Ningbo, China

L. Zhang · J. J. Lv
Shandong Center for Disease Control and Prevention, Jinan, China

F. Yang
Qingdao Center for Disease Control and Prevention, Qingdao, China

developed in 2009. According to studies utilizing the 5L version, sensitivity and discriminatory powers were improved, and ceiling effects were reduced [14–18]. Until now, there has been little use of 5L in a Chinese population, nor has there been a comparison between these two systems.

We conducted a cross-sectional study in patients with hepatitis B to compare the psychometric properties of 5L and 3L in a Chinese population and evaluated acceptability, face validity, redistribution properties, convergent validity, known-group validity, discriminatory power, ceiling effect, test–retest reliability, and responsiveness. We report the findings of our study.

Methods

Subjects and settings

We chose Jinan Hospital for Infectious Diseases and Shandong Tumor Hospital, the largest hospitals for hepatitis B infections and for tumors in Shandong Province, respectively, to recruit patients in August, 2013. All inpatients and outpatients with hepatitis B, including carriers, patients with active or inactive chronic hepatitis B (CHB), compensated cirrhosis (CC), decompensated cirrhosis (DC), or hepatocellular carcinoma (HCC), were invited to take part in the survey using consecutive sampling. We attempted to recruit all patients who were conscious and not in severe pain who registered sequentially in either of the two hospitals. For each infectious condition, at least 100 patients were recruited. Patients who were below 18 years old, or not capable of reading or speaking Chinese, or with mental disorders or dementia were excluded. Infectious conditions were confirmed by case histories, and the medical history of other diseases was recorded.

Instruments

The EQ-5D is composed of a descriptive system and a visual analog scale (VAS). The descriptive system consists of five dimensions: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. For each dimension, there are three levels in the 3L version and five levels in the 5L version. The VAS is a thermograph-like scale rated from 0 (the worst imaginable health) to 100 (the best imaginable health). Subjects were asked to choose one level in each dimension and mark the scale according to their current health states. HRQoL was derived from descriptive systems using value sets from the general population. Many countries have estimated their 3L value sets by time trade-off or VAS methods, but not with a Chinese population. Since the 5L version was relatively new, there was no current value set [1, 2].

In the 3L version, the three levels in each dimension were defined as ‘no problems,’ ‘some problems,’ and ‘unable to/extreme problems.’ If a subject chooses one level in each of the five dimensions, the combination of all five levels could define a unique health state, so the 3L descriptive system could define up to 243 (3^5) unique health states as levels ranging from 11,111 (best health) to 33,333 (worst health). In the 5L version, the levels in each dimension have been expanded to five, including ‘no problems,’ ‘slight problems,’ ‘moderate problems,’ ‘severe problems,’ and ‘unable to/extreme problems,’ so 3,125 (5^5) unique health states could be defined as levels ranging from 11,111 (best health) to 55,555 (worst health) [1, 2].

We used the simple Chinese versions of both the 3L and 5L [2]. Each consenting patient received a questionnaire composed of 5L, socio-demographic information, 3L, and the VAS, respectively. Consistent with the literature, we maintained this sequential order to reduce bias. For example, if respondents scored 3L first, there could be a tendency to not use the second and fourth levels in 5L [16]. The socio-demographic section was placed between the two systems in order to reduce the memory effects.

Since outpatients were difficult to follow-up, we conducted a retest 1 week after the first survey with inpatients to measure test–retest reliability and responsiveness of the two systems. We asked doctors to give judgments on the progressions of patients’ health states based on symptoms and laboratory tests defined as improved, stable, or exacerbated. Since the symptoms and laboratory tests of patients with hepatitis B were complex, doctors were given autonomy to make these judgments.

Analysis

HRQoL

To obtain HRQoL for 3L, we used the value set from a Japanese cohort that was estimated using a time trade-off method on 543 members of the public population [19]. Since no 5L value sets were available, the EuroQol Group recommended mapping the interim scoring of the 5L descriptive system to 3L to obtain HRQoL [20]. Being independent of the value set that is used, it can be applied to transform any 3L value set into 5L index values. Mean and standard deviation (SD) were used to describe the central and discrete tendency of HRQoL.

Acceptability and face validity

These two constructs were based on direct choices by each patient from face-to-face interview [16]. Acceptability was evaluated by two questions: ‘In your opinion, which one of the two instruments is easier to answer?’ and ‘Did you have any problems in any of the two systems?' Face validity was evaluated by another question: 'In your opinion, which one of the two systems expresses your ideas better?'

Redistribution properties

Redistribution properties were defined as proportions of 3L–5L response pairs in each dimension from the same patient [14, 21]. We used 3L-1, 3L-2, and 3L-3 to represent the subgroups in which patients chose level 1, level 2, and level 3, respectively, in 3L in each dimension. Each 3L–5L response pair was described by P(x, y) according to the levels chosen in both 3L and 5L, where x indicated the 3L level and y indicated the 5L level. In each subgroup of paired responses, the mean of VAS was calculated, except for inconsistent pairs. P(1, 1), P(1, 2), P(2, 2), P(2, 3), P(2, 4), P(3, 4), P(3, 5) were defined as consistent while the others were considered inconsistent.

In each dimension, P(1, 1) represented the best health states while P(3, 5) represented the worst. We hypothesized the VAS should be increasing as pairs going from P(3, 5) to P(1, 1). The proportion of P(1, 2) and P(2, 2) could be evidence for the validity of an extra level between level 1 and level 2 in the 3L system, the same as P(2, 4) and P(3, 4), supporting an extra level between level 2 and level 3.

Convergent validity

We compared the correlation between 3L and the VAS with 5L and the VAS. The levels of each dimension were treated as ordinal variables, and Spearman's rank correlation coefficient (r_s) was computed. Strength of correlation was interpreted as follows: absent ($r_s < 0.20$), weak ($0.20 \leq r_s < 0.35$), moderate ($0.35 \leq r_s < 0.50$), and strong ($r_s \geq 0.50$) [14]. We hypothesized a stronger correlation in 5L. Cohen's t test was used for comparison between r_s from 3L to 5L [22].

Known-group validity

Known-group validity is a direct and convenient assessment of both construction validity and sensitivity [23]. HRQoL was used to make the comparison instead of a single level. Tests for age and education were performed using r_s, and education was recorded as three levels (1 = primary or lower, 2 = middle school, 3 = college or higher). Gender, infectious conditions, and presence of comorbidities were assessed using the Kruskal–Wallis H statistic. According to previous studies [24, 25], we expected patients that were female, young, with high education level, or without comorbidity to have higher HRQoL. Patients with mild symptoms were hypothesized to have higher HRQoL.

Discriminatory power

We used Shannon's index (H') and Shannon's (J') evenness index to compare discriminant power of 3L and 5L. Shannon's index and evenness index are measures of discriminatory power of any health measurement scale [26]. In a dimension where a certain level has a very low or very high endorsement, e.g., less than 0.05 or more than 0.95, little information is gained because one predicts the other with more than 95 % certainty. Originating from the field of information theory, the Shannon's index and Shannon's evenness index have already been used to assess many health states classifications successfully [14, 16, 27, 28]. The Shannon's index was calculated as follows:

$$H' = -\sum_{i=1}^{L} p_i \log_2 p_i$$

Where L represented the number of levels in one dimension, and p_i represented the proportion of patients who chose the ith level. The maximum of H' was marked as $H'_{\max}$, which amounted to $\log_2 C$, where C represented the number of levels in each dimension. For the 3L and 5L system, $H'_{\max}$ amounted to 1.58 and 2.32, respectively. The Shannon's evenness index was calculated as follows:

$$J' = \frac{H'}{H'_{\max}}$$

Although H' increased when C increased, J' would increase only if the new levels added were actually used. H' was an absolute measure and J' combined both the number and the efficiency of the levels that were used. If answers from patients were evenly distributed across all levels, J' would reach its zenith; if there were redundant levels, J' would decrease as well. We hypothesized H' would increase and J' would not decrease significantly in the 5L system compared with the 3L system.

Ceiling effect

Ceiling effect was assessed as the proportion of subjects reporting no problem in each dimension. With the decrease in ceiling effect, the sensitivity should increase. We expected a decrease in the ceiling effect in 5L compared with 3L.

Test–retest reliability

For patients with stable health states 7 days after the first survey according to doctors, weighted kappa was used to evaluate the test–retest reliability. We applied Fleiss's standard for the strength of agreement for kappa values, as follows: poor ($\kappa < 0.4$), fair to good ($0.4 \leq \kappa < 0.75$),

Table 1 Distribution of subjects by age, gender, resource, and disease progression in the first interview and review

Infectious condition	First interview						Review			
	No.	Mean age	Gender		Resource		No.	Disease progression		
			Male (%)	Female (%)	Outpatients (%)	Inpatients (%)		Improved	Stable	Exacerbated
HBV carriers	104	27.4	72 (69.2)	32 (30.8)	104 (100.0)	0 (0.0)	–	–	–	–
Inactive CHB	103	38.9	83 (80.6)	20 (19.4)	91 (88.3)	12 (11.7)	–	–	–	–
Active CHB	112	39.1	87 (77.7)	25 (22.3)	84 (75.0)	28 (25.0)	27	23 (85.2)	2 (7.4)	2 (7.4)
CC	114	48.5	78 (68.4)	36 (31.6)	44 (38.6)	70 (61.4)	48	24 (50.0)	21 (43.8)	3 (56.2)
DC	107	52.1	78 (72.9)	29 (27.1)	35 (32.7)	72 (67.3)	45	35 (77.8)	8 (17.8)	2 (4.4)
HCC	105	56.6	86 (81.9)	19 (18.1)	11 (10.5)	94 (89.5)	63	38 (60.3)	19 (30.2)	6 (9.5)
Total	645	43.9	484 (75.0)	161 (25.0)	369 (57.2)	276 (42.8)	183	120 (65.6)	50 (27.3)	13 (7.1)

excellent ($\kappa \geq 0.75$) [14]. At the same time, two-way random intraclass correlation coefficient (ICC) was used to compare HRQoL before and after the 7-day interval [23].

Responsiveness

For patients whose doctors reported improved health states 7 days after the first survey, we assumed an increase in HRQoL. Wilcoxon signed-rank sum test was used to compare HRQoL before and after the 7-day interval in 3L (5L) or between the two systems. There was an insufficient number of patients with deteriorated health states for statistical analysis.

All the statistical analyses were performed by R version 2.15.2, and differences were considered statistically significant for P values less than 0.05.

Results

Subjects

A total of 1,063 outpatients and 445 inpatients were invited to take part in the survey; the response rate was 34.7 and 62.0 %, respectively. Eventually 369 outpatients and 276 inpatients were included and answered the questionnaires. Seven days later, inpatient participants were invited to retake the survey, and 183 answered the same questionnaires. Judgments of their disease progressions were received from doctors (Table 1).

HRQoL and the VAS

In 5L, HBV carriers and DC had the highest and lowest respective means for both HRQoL and VAS (0.813 vs. 0.661). In 3L, HBV carriers and inactive CHB patients had the highest mean HRQoL (0.817), while DC had the lowest (0.665). HBV carriers had the highest mean of VAS (88.0), while DC had the lowest (58.4) (Table 2).

Table 2 Mean and SD of HRQoL and VAS in the first interview

Infectious condition	HRQoL (mean ± SD)		VAS (mean ± SD)
	5L	3L	
HBV carriers	0.813 ± 0.035	0.817 ± 0.044	88.0 ± 10.1
Inactive CHB	0.808 ± 0.043	0.817 ± 0.051	82.3 ± 11.1
Active CHB	0.771 ± 0.060	0.773 ± 0.087	78.4 ± 15.8
CC	0.763 ± 0.056	0.759 ± 0.084	72.0 ± 15.5
DC	0.661 ± 0.107	0.665 ± 0.133	58.4 ± 15.7
HCC	0.699 ± 0.119	0.684 ± 0.152	71.1 ± 17.9

Acceptability and face validity

There were 63.6 % of patients who deemed 5L to be easier to answer than 3L; 13.5 % chose 3L; 20.3 % indicated there was no apparent difference between them; and 2.6 % did not answer the questions. A total of 7.2 % of patients reported some problems in filling in 3L, while 2.5 % complained 5L was tedious. A total of 68.2 % of patients claimed 5L could express their true thoughts; 8.2 % chose 3L; 18.2 % indicated there was no apparent difference; and 5.4 % did not answer the questions.

Redistribution properties

A total of 77 (2.4 %) pairs were inconsistent. Usual activities had the most inconsistent pairs ($n = 36$, 5.6 %) while self-care had the least ($n = 5$, 0.8 %).

Table 3 shows the mean VAS of 3L–5L pairs. With most of the pairs, the VAS decreased as going from $P(1, 1)$ to $P(3, 5)$. In self-care, the VAS of $P(2, 4)$ equaled the VAS of $P(2, 2)$, while $P(3, 5)$ was higher than the whole 3L-2

Table 3 Redistribution properties from 3L to 5L: consistent responses

Dimension	3L level	5L level	Subgroup	N	Proportion (%)	VAS mean
Mobility	1	1	P(1,1)	459	90.9	80.7
		2	P(1,2)	46	9.1	70.2
	2	2	P(2,2)	87	69.0	63.6
		3	P(2,3)	33	26.2	49.4
		4	P(2,4)	6	4.8	36.9
	3	4	P(3,4)	0	–	–
		5	P(3,5)	5	100	29.8
Self-care	1	1	P(1,1)	587	98.2	76.7
		2	P(1,2)	11	1.8	65.0
	2	2	P(2,2)	31	72.1	50.3
		3	P(2,3)	9	20.9	45.8
		4	P(2,4)	3	7.0	50.0
	3	4	P(3,4)	1	25.0	–
		5	P(3,5)	3	75.0	53.0
Usual activities	1	1	P(1,1)	380	91.6	82.3
		2	P(1,2)	35	8.4	76.0
	2	2	P(2,2)	67	40.4	73.8
		3	P(2,3)	74	44.6	65.0
		4	P(2,4)	25	15.0	58.2
	3	4	P(3,4)	4	14.3	32.0
		5	P(3,5)	24	85.7	48.4
Pain/discomfort	1	1	P(1,1)	301	71.5	80.8
		2	P(1,2)	120	28.5	74.5
	2	2	P(2,2)	152	72.7	69.0
		3	P(2,3)	44	21.1	56.7
		4	P(2,4)	13	6.2	44.6
	3	4	P(3,4)	3	100.0	35.0
		5	P(3,5)	0	–	–
Anxiety/depression	1	1	P(1,1)	266	67.0	80.3
		2	P(1,2)	131	33.0	74.4
	2	2	P(2,2)	164	74.2	70.3
		3	P(2,3)	44	19.9	65.2
		4	P(2,4)	13	5.9	55.2
	3	4	P(3,4)	6	50.0	61.3
		5	P(3,5)	6	50.0	50.8

Table 4 Correlation coefficient between 3L, 5L and VAS

Dimension	r_s between 5L and VAS	r_s between 3L and VAS	t	P value
Mobility	−0.554	−0.510	−4.32	$P < 0.001$
Self-care	−0.321	−0.334	1.16	$P = 0.246$
Usual activities	−0.561	−0.503	−6.37	$P < 0.001$
Pain/discomfort	−0.450	−0.401	−2.55	$P = 0.011$
Anxiety/depression	−0.366	−0.317	−2.14	$P = 0.032$

subgroup. In usual activities, the VAS of P(3, 5) was higher than that of P(3, 4). In anxiety/depression, the VAS of P(3, 4) was higher than the VAS of P(2, 4).

In all dimensions of the 3L-1 subgroup, there were higher proportions for P(1, 1), which ranged from 67.0 % (anxiety/depression) to 98.2 %, than for P(1, 2). In the 3L-2 subgroup, P(2, 1) with a proportion range from 69.0 % (mobility) to 74.2 % (anxiety/depression) dominated over P(2, 3) and P(2, 4) except in usual activities, where P(2, 3) had the largest proportion of 44.6 %. P(2, 4) had the smallest proportions in all dimensions. The most even distribution in the 3L-2 subgroup was in usual activities (67/74/25). There were not many pairs observed in the 3L-3 subgroup in mobility, pain/discomfort, and anxiety/depression, P(3, 5) had a larger proportion than P(3, 4), while for self-care and usual activities, P(3, 4) dominated the subgroup (Table 3).

Convergent validity

Table 4 shows the r_s between 5L (3L) and the VAS in each dimension. The correlations between mobility and the VAS, usual activities and the VAS were high in both of 3L and 5L while others were moderate. The r_s between 5L and the VAS was statistically higher than 3L in all dimensions except self-care.

Known-group validity

For 5L and 3L, the r_s between age and HRQoL was −0.367 and −0.320 in both 5L and 3L, respectively. This indicates a negative correlation between age and HRQoL, as expected, but the difference between r_s from 3L to 5L was not statistically significant ($t = -0.928$, $P = 0.354$). The r_s was 0.223 and 0.199 between education and HRQoL in 5L and 3L, respectively, showing the same direction as expected; however, the difference between r_s from 3L to 5L was also not statistically significant ($t = 1.75$, $P = 0.081$).

There was no significant difference between male and female ($P = 0.278$ and 0.290, respectively) in either 5L or 3L. Patients with other diseases showed lower HRQoL in both 5L and 3L (both $P < 0.001$). In 5L, there were significant differences of HRQoL between different infectious conditions except two pairs: inactive CHB and HBV carriers ($P = 0.546$) and active CHB and CC ($P = 0.355$). In 3L, HRQoL of three pairs—inactive CHB and HBV carriers ($P = 0.673$), active CHB and CC ($P = 0.168$), DC and HCC ($P = 0.245$)—were found to have no significant difference.

Discriminatory power

In all dimensions, H' from 5L was larger than that of 3L. The 5L J' value was slightly larger than 3L J' in mobility,

Table 5 Shannon's index and evenness index

Dimension	5L			3L		
	No.	*H'*	*J'*	No.	*H'*	*J'*
Mobility	645	1.201	0.518	645	0.808	0.511
Self-care	645	0.569	0.245	645	0.430	0.272
Usual activities	645	1.692	0.729	645	1.101	0.697
Pain/discomfort	645	1.445	0.623	645	0.982	0.622
Anxiety/depression	645	1.546	0.666	645	1.063	0.673

usual activities, and pain/discomfort, but in self-care and anxiety/depression, *J'* from 3L was larger than *J'* from 5L (Table 5).

Ceiling effect

There were relatively severe ceiling effects in both of the systems in all dimensions. Self-care showed the highest percentage (90.7 and 92.1 % in 5L and 3L, respectively), while anxiety/depression showed the lowest (42.6 and 62.3 % in 5L and 3L, respectively). A reduction in the proportion of patients who chose the first level of a dimension can be seen when comparing 3L and 5L percentages: Anxiety/depression showed the largest reduction (19.7 %), followed by pain/discomfort (17.7 %), mobility (7.3 %), usual activities (6.8 %), and self-care (1.4 %). Patients who chose level 1 in all dimensions decreased to 21.6 % in 5L by 16.7 % (Table 6). All reduces were statistically significant.

Test–retest reliability

The κ of 5L was higher than of 3L in mobility, usual activities, and anxiety/depression but without statistical evidence (Table 7). ICC of 3L in CC patients was higher than in 5L. In DC and HCC patients, ICC was lower than 5L, but none of these differences were statistically significant. In all of the patients with stable health states, ICC of 5L was higher than 3L with statistical evidence (Table 8). There were only 2 active CHB patients with stable health states, so ICC was not computed.

Table 7 Kappa in each dimension in patients with stable health states

Dimension	Kappa(95 % CI)	
	3L	5L
Mobility	0.908(0.786–1.000)	0.978(0.929–1.000)
Self-care	0.928(0.802–1.000)	0.732(0.618–0.846)
Usual activities	0.744(0.567–0.921)	0.839(0.688–1.000)
Pain/discomfort	0.867(0.721–1.000)	0.790(0.620–0.960)
Anxiety/depression	0.822(0.668–0.976)	0.857(0.711–1.000)

Table 8 ICC in patients with stable health states

Infectious conditions	ICC	
	5L	3L
CC	0.840 (0.824–0.856)	0.874 (0.825–0.923)
DC	0.665 (0.653–0.677)	0.659 (0.627–0.691)
HCC	0.982 (0.974–0.990)	0.921 (0.896–0.946)
Total	0.925 (0.903–0.947)	0.828 (0.757–0.899)

Responsiveness

In patients with improved health states, HRQoL increased in all infectious conditions for both 3L and 5L. In 5L, HRQoL increased between 0.029 (active CHB and CC) and 0.073 (HCC), while in 3L, HRQoL increased between 0.025 (CC) and 0.076 (HCC) (Table 9). There was no significant difference in the increases in HRQoL between 3L and 5L.

Discussion

The purpose of our study was to compare 5L and 3L according to validity, sensitivity, reliability, and responsiveness. We believe that it is the first time to make such a comparison between the two systems and the first time to use the new 5L version of EQ-5D in China.

Table 6 Distribution of patients who chose the first level in each dimension

Dimension	5L		3L		Reduction in 5L compared with 3L		
	No.	Percent (%)	No.	Percent (%)	No.	Percent (%)	*P* value*
Mobility	461	71.5	508	78.8	47	7.3	<0.001
Self-care	585	90.7	594	92.1	9	1.4	<0.001
Usual activities	388	60.2	432	67.0	44	6.8	<0.001
Pain/discomfort	309	47.9	423	65.6	114	17.7	<0.001
Anxiety/depression	275	42.6	402	62.3	127	19.7	<0.001
All dimensions	139	21.6	247	38.3	108	16.7	<0.001

* McNemar's test

Table 9 Results of Wilcoxon signed-rank order tests for patients with improved health states

Infectious condition	No.	HRQoL (first interview)		HRQoL (review)		HRQoL (increases)		*P*		
		5L	3L	5L	3L	5L	3L	5L	3L	5L vs. 3L
Active CHB	23	0.763	0.750	0.792	0.776	0.029	0.026	0.002	0.039	0.700
CC	24	0.740	0.738	0.769	0.763	0.029	0.025	0.002	0.031	0.477
DC	35	0.658	0.666	0.717	0.718	0.059	0.052	0.009	0.004	0.171
HCC	38	0.673	0.646	0.746	0.722	0.073	0.076	<0.001	<0.001	0.435
Total	120	0.699	0.690	0.751	0.739	0.052	0.049	<0.001	<0.001	0.089

Although 3L had a more condensed format, most patients found 5L to be easier to answer and more useful in expressing their true opinion, supporting similar findings as another study [16]. The results of 3L–5L redistribution properties show that the extra levels added provide evidence to support this preference. Since very few people deemed the 5L version to be tedious, the levels added were not an extra burden for subjects.

Similar to other studies [14, 15, 18], our distribution data, especially in the 3L-1 and 3L-2 subgroups, were skewed. In the 3L-1 subgroup, most patients showed level 1 in 5L since they displayed no symptoms in at least one dimension no matter which system they used. In the 3L-2 subgroup, most patients chose level 2 in 5L in all dimensions except in usual activities. However, in other studies [16, 21], the distribution was more even. This may be due to disease characters that most patients with chronic HBV have tended to perceive as less severe, and as a result, very few chose $P(2, 4)$ or the 3L-3 subgroup. $P(2, 4)$ and $P(3, 4)$ consisted of only a small proportion, giving little evidence to an extra level between level 2 and level 3 in 3L. On the other hand, $P(1, 2)$ and $P(2, 2)$ consisted of 6.5 % (self-care) to 45.7 % (anxiety/depression) of all pairs in each dimension, showing strong evidence for an extra level between level 1 and level 2 in 3L.

The inconsistent pairs constituted a small proportion of the responses. This small proportion is similarly found in another study [16] and even lower than two others [14, 15], indicating that patients could consistently use the two systems. The trend of the VAS increasing as the pairs moved from $P(3, 5)$ to $P(1, 1)$ supports the validation of the 3L–5L extension. There was deviation in the redistribution in some severe subgroups—$P(2, 4)$ in self-care, $P(3, 5)$ in usual activities, and $P(3, 5)$ in anxiety/depression. This may because in the VAS, 0 was described as ‘worst imaginable health state,’ but in each dimension, there was no word such as ‘worst.’ This may lead to diverse rating methods from patients.

The correlation between 5L and the VAS was stronger than 3L, indicating better convergent validity. Both the two systems showed good known-group validity except in gender since men were much more prevalent than women in our survey, and symptoms could often cover the effects of gender. The 5L version being able to separate more infectious conditions than the 3L version allowed for better validation and sensitivity.

The Shannon’s index increased from 3L to 5L, but no apparent improvement of the Shannon’s evenness index in 5L was found, indicating the use of the extra levels added was not adequate. This might be attributed to the skewed redistribution of the 3L–5L response pairs, resulting in the inadequate use of some levels. The ceiling effect decreased in all dimensions in 5L, as shown in the redistribution of the proportion of $P(1, 2)$, and spread from 1.8 % (self-care) to 33.0 % (anxiety/depression) in the 3L-2 subgroup.

In patients with stable health states, ICC was higher in 5L than 3L, but the improvement of the κ was not significant. While ICC was assessed by HRQoL as a sum of all five dimensions, κ was obtained from each dimension. As a result, κ was more sensitive to changes in a single dimension, while what ICC measured was more comprehensive. This result indicated that although the stability of each dimension was not improved in 5L, the reliability for the whole descriptive system was much improved.

For patients with improved health states, both of the two systems could detect increases in HRQoL, indicating a good responsiveness. We did not detect any significant difference with our sample size, and therefore, we could not determine which system was more effective.

In our study, we compared the two versions of EQ-5D in a wide disease spectrum of hepatitis B infection. Since the conditions varied greatly by infectious patterns, from no symptoms to being restricted in bed, we could observe the two versions from different perspectives. Most studies have evaluated the 5L version to be better than the 3L version [14–17, 21, 28]. We did observe improvement in the sensitivity and convergent validity, and we acquired evidence for the extra levels added from redistribution properties; however, our evidence was not as apparent as previous studies. This may be because our limited sample size could not detect small improvements. Additionally, during the field survey, some patients reported confusion about the dimensions and levels, for example, the third dimension, usual activities, which was described as being able to work, study, do housework, and entertainment. For some patients with early HCC, work was forbidden so the patient could rest, but the patient could do housework, confusing the patient. For patients who needed infusions regularly, they may have chosen level 1 in mobility since they could walk without the help when there was no infusion but they also could have chosen level 5 since they could not move during infusion. At least 13 patients suggested the levels should be described more concretely, rather than just changing adjectives. These problems arose due to different interpretations on the descriptive systems. In the future, studies with larger sample sizes and more types of patients may be warranted, while some revisions in EQ-5D may be needed for better applicability in a Chinese population.

There were several limitations in our study. First, there was no value set for 5L and 3L for any Chinese population. We used a Japanese value set, which may cause bias. Second, some patients who were highly anxious or depressed refused participation in the survey, and patients who were seriously ill, for example, those who were in a coma or felt grave pain or discomfort at the time of survey were excluded. Third, judgments of patients' progressions were made by doctors which may have involved subjective factors. Fourth, the interval between the first survey and retake of the survey was determined by the mean time of inpatients staying in hospitals. This interval may have been short enough to have carryover effects, and there may have been some residual memory that could have influenced the retest. Besides, the socio-demographic section was placed between the two systems in order to reduce the memory effects, but this could not eliminate this carryover effect.

In conclusion, 5L was more suitable than 3L for use on patients with hepatitis B in China. Its acceptability, convergent validity, and reliability are improved while its ceiling effects are reduced compared with 3L. Evidence for the levels added in 5L was obtained from redistribution properties. We recommend the 5L version for future studies.

Acknowledgments The use of two versions of EQ-5D was approved by EuroQol Group. This study was sponsored by National Health and Family Planning Commission and Minister of Science and Technology (No.2012ZX10002001), and we would like to express the special thanks to Rodewald Lance, who provided comments and suggestions on the manuscript.

References

1. EuroQol Group. UserGuide_EQ-5D-5L. http://www.euroqol.org/about-eq-5d/publications/user-guide.html.
2. http://www.euroqol.org/.
3. Brazier, J., Deverill, M., & Green, C. (1999). A review of the use of health status measures in economic evaluation. *Journal of Health Services Research and Policy, 4*(3), 174–184.
4. Revicki, D. A., & Lenderking, W. R. (2012). Methods and issues associated with the use of quality-adjusted life-years. *Expert Review of Pharmacoeconomics and Outcomes Research, 12*(1), 105–114.
5. Zhang, F., Fu, S. Y., Wang, M., & Du, J. (2009). Correlation empirical research on the quality of life of Chinese patients with acute ischemic stoke measured by EQ-5D. *China Journal of Pharmaceutical Economics, 6*, 5–12.
6. Wu, Y. Q., Liu, K., Tang, X., Cao, Y., Wang, J. W., Li, N., et al. (2012). Empirical research of measuring elderly health utility in the outskirts of Beijing by using European quality of life 5-Dimensions. *Journal of Peking University (Health Science), 44*(3), 397–402.
7. Zhou, L., Jin, H., Deng, Z. H., Cheng, X. W., Li, Z., Hu, S. X., et al. (2012). Measuring quality of life of hospitalized patients with severe acute respiratory infection by EQ-5D instrument: An exploratory study. *Practical Preventive Medicine, 19*(5), 644–647.
8. Luo, Y. J., Fang, Y., & Ding, X. S. (2009). Quality of life of patients with type 2 diabetes measured by EQ-5D and its influential factors analysis. *China Journal of Pharmaceutical Economics, 2*, 12–18.
9. He, M. M., & Wu, M. (2009). Using EQ-5D to preliminarily analyze the relation between living mode and health condition in a district in Beijing. *Chinese Journal of Gerontology, 29*(4), 478–481.
10. Wang, H. M., Patrick, D. L., Edwards, T. C., Skalicky, A. M., Zeng, H. Y., & Gu, W. W. (2012). Validation of the EQ-5D in a general population sample in urban China. *Quality of Life Research, 21*(1), 155–160.
11. Kopec, J. A., & Willison, K. D. (2003). A comparative review of four preference-weighted measures of health-related quality of life. *Journal of Clinical Epidemiology, 56*(4), 317–325.
12. Brazier, J., Deverill, M., & Green, C. (1999). A review of the use of health status measures in economic evaluation. *Journal of Health Services Research and Policy, 4*(3), 174–184.
13. Kopec, J., & Willison, K. (2003). A comparative review of four preference-weighted measures of health-related quality of life. *Journal of Clinical Epidemiology, 56*(4), 317–325.
14. Scalone, L., Ciampichini, R., Fagiuoli, S., Gardini, I., Fusco, F., Gaeta, L., et al. (2012). Comparing the performance of the standard EQ-5D 3L with the new version EQ-5D 5L in patients with chronic hepatic diseases. *Quality of Life Research, 22*(7), 1707–1716.
15. Kim, S. H., Kim, H. J., Lee, S. I., et al. (2012). Comparing the psychometric properties of the EQ-5D-3L and EQ-5D-5L in cancer patients in Korea. *Quality of Life Research, 21*(6), 1065–1073.
16. Janssen, M. F., Birnie, E., Haagsma, J. A., & Bonsel, G. J. (2008). Comparing the standard EQ-5D three-level system with a five-level version. *Value in Health, 11*(2), 275–284.
17. Pickard, A. S., De Leon, M. C., Kohlmann, T., Cella, D., & Rosenbloom, S. (2007). Psychometric comparison of the standard EQ-5D to a 5 level version in cancer patients. *Medical Care, 45*(3), 259–263.
18. Kim, T. H., Jo, M. W., Lee, S. I., Kim, S. H., & Chung, S. M. (2012). Psychometric properties of the EQ-5D-5L in the general population of South Korea. *Quality of Life Research, 22*(8), 2245–2253.
19. Tsuchiya, A., Ikeda, S., Ikegami, N., Nishimura, S., Sakai, I., Fukuda, T., et al. (2002). Estimating an EQ-5D population value set: The case of Japan. *Health Economics, 11*(4), 341–353.
20. van Hout, B., Janssen, M. F., Feng, Y. S., Kohlmann, T., Busschbach, J., Golicki, D., et al. (2012). Interim scoring for the EQ-5D-5L: mapping the EQ-5D-5L to EQ-5D-3L value sets. *Value in Health, 15*(5), 708–715.

21. Janssen, M. F., Pickard, A. S., Golicki, D., Gudex, C., Niewada, M., Scalone, L., et al. (2013). Measurement properties of the EQ-5D-5L compared to the EQ-5D-3L across eight patient groups: A multi-country study. *Quality of Life Research, 22*(7), 1717–1727.
22. Cohen, J. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (Vol. 2). Hillsdale, NJ: Lawrence Erlbaum Associates.
23. Fayers, P. M., & Machin, D. (2000). *Quality of Life-Assessment, Analysis, Interpretation*. Chichester: Wiley.
24. Sun, S., Chen, J., Johannesson, M., Kind, P., Xu, L., Zhang, Y., et al. (2011). Population health status in China: EQ-5D results, by age, sex and socio-economic status, from the National Health Services Survey 2008. *Quality of Life Research, 20*(3), 309–320.
25. Wang, H., Kindig, D. A., & Mullahy, J. (2005). Variation in Chinese population health related quality of life: Results from a EuroQol study in Beijing, China. *Quality of Life Research, 14*(1), 119–132.
26. Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal, 27*, 379–423.
27. Janssen, M. F., Birnie, E., & Bonsel, G. J. (2007). Evaluation the discriminatory power of EQ-5D, HUI2, and HUI3 in a US general population survey using Shannon's indices. *Quality of Life Research, 16*(5), 895–904.
28. Pickard, A. S., Kohlmann, T., Janssen, M. F., Bonsel, G., Rosenbloom, S., & Cella, D. (2007). Evaluating equivalency between response systems: Application of the Rasch model to a 3-level and 5-level EQ-5D. *Medical Care, 45*(9), 812–819.