ORIGINAL ARTICLE

Reliability, Validity and Discriminatory Ability of Spitzer's QL-index in Dialysis Patients

T O Lim, Zaki Morad, Department of Nephrology, Hospital Kuala Lumpur, Jalan Pahang, 50586 Kuala Lumpur

Summary

Quality of life outcome (QOL) on dialysis is important. We determined the measurement properties of Spitzer's QL-index, a QOL measure, in our patients on chronic haemodialysis. The QL-index measures 5 dimensions of QOL (activity, daily activities, general health, social support and psychological outlook). 59 haemodialysis (HD) patients from 2 centres were rated by 5 raters. Inter-rater agreement for the total score was good with a mean intra-class correlation coefficient 0.66 (range 0.47-0.81). That for dimension scores however was poor (weighted kappa range 0.07-1). Systematic differences between raters were also observed. Intra-rater agreement was generally better than inter-rater agreement. Significant gradients in scores were observed by age, serum albumin, comorbid disorders, previous hospitalisation, capacity for self care HD and rehabilitation status thus providing evidence for construct validity. The distribution of total scores was skewed indicating poor discriminatory ability. Nevertheless, QL-index has acceptable measurement properties for application in dialysis patients.

Key Words: Spitzer's QL-index, Quality of life, Observer agreement, Validity, Dialysis

Introduction

Routine assessment of the outcome of dialysis treatment, as reported by most Registries [1-3] including our own [4], focus mainly on quantity of patient survival. Dialysis treatment is life long and has considerable impact on patients' lifestyle. It is therefore pertinent that assessment of the outcome of dialysis treatment should also examine patients' health related quality of life. The Malaysian Dialysis and Transplant Registry has decided to include quality of life data as part of the comprehensive data set on each patient to be reported to the Registry annually.

To develop a new instrument to measure quality of life for our application is unnecessary. There are numerous existing quality of life (QOL) measures from which to choose. Selection of an instrument for application is guided by considerations of appropriateness, practicality and measurement properties including measurement validity, reliability and discriminatory ability [5,6]. We have chosen Spitzer's Quality of life index (QL-index) for the following reasons:

1. The QL-index was developed for use in patients with cancer and other chronic diseases.[7] It has been applied to many different patient populations with chronic disease, including end stage renal failure.[8]

2. As is appropriate for a modern QOL measure, the QL-index reflects the multidimensionality of quality of life. It measures 5 dimensions of quality of life: activity level, activities of daily living, feelings of healthiness, quality of social support and psychological outlook [7]. In contrast,

early instruments like the Karnofsky scale for example, emphasises only the physical dimensions of quality of life.

3. The QL-index is brief, it takes about 2 minutes to complete. It is designed to be administered by a doctor or other health professionals [7]. In research and clinical settings, economy and ease of use would favour a patient self-administered instrument [5,6]. This however is not necessarily true in dialysis practice. Through repeated and longstanding contact, dialysis personnel are familiar and have intimate knowledge of their patients' personal lives. A dialysis staff administered instrument takes advantage of that. Further, in our dialysis setting; a staff administered instrument is particularly important. We need to avoid the difficulties associated with translation and cross cultural adaptation of English language instruments [9]. With few exceptions, all QOL measures so far developed are in English language and are intended for use in English speaking countries. To our knowledge, no Malay, Chinese or Tamil language instruments for assessing QOL have ever been developed. To translate and adapt existing English language instrument into these 3 widely used languages in Malaysia in order for the instrument to be self-completed by patient is beyond our means. Moreover, self-administered instruments require patients to be literate and sighted; these requirements cannot be met for the present by the majority of our patients.

Hence, on initial review, Spitzer's QL-index appears to be a suitable and practical QOL measure for our application. What remain to be established is its measurement properties when applied in our patient population. Reported reliability of an instrument in particular is difficult to generalise to other patient populations. We report here a study to determine the reliability, validity and discriminatory ability of QL-index in our dialysis patients.

Methods

The Spitzer's QL-index contains 5 items; each item measures a dimension of QOL. The 5 dimensions covered are activity level, activities of daily living, feeling of healthiness, social support and psychological outlook. Each dimension is scored on a scale from 0 (worst health) to 2 (best health for that dimension). The 5 scores are summed to give a total score ranging between 0 and 10.

Two haemodialysis (HD) centres in 2 different states (out of 23 centres spread over 13 states in the Ministry of Health HD programme) were selected based on convenience for this study. Five raters participated; they comprised 2 dialysis staff from each centre and the one Nephrologist who covered both centres. All raters received prior training and instruction on how to use the instrument. While they use their individual and independent judgment in assessing patients, they were guided by detailed description of the QL-index. All raters interviewed and assessed each patient in their respective centre once on the same day. One of the dialysis staff at each centre repeated the assessment on the same patients 4 to 7 days later.

All out-patients on chronic HD at the 2 centres were included except for those on temporary HD while awaiting elective living related renal transplant and those who were hospitalised at the time of the assessment. All patients who were eligible agreed to participate when approached. All 59 chronic HD patients at both centres participated; 28 patients from one centre and 31 from the other. The sample size is within the range of those commonly employed in validation studies [7,10-12] . No formal sample size calculation was performed. 22 consecutive renal transplant patients attending the transplant clinic were used as controls. These were assessed by the same nephrologist once. To be eligible for assessment, a transplant patient must have been transplanted for at least 6 months and have a serum creatinine less than 200 umol/L.

Apart from QOL assessment, each HD patient's medical record was reviewed to abstract data on major comorbid disorders present, hospitalization in the previous 2 months, current serum albumin and haemoglobin levels. Comorbid conditions included were cardiovascular disorders (ischaemic heart disease, cerebrovascular disease, peripheral vascular disease), diabetes mellitus, malignancy, symptomatic and/or radiological renal bone disorder, and clinically evident liver cirrhosis. Data on patients' work related rehabilitation status and capacity for self-care HD at the same year were obtained from the Registry database.

Statistical analysis

Reliability assessment

Inter and intra-rater agreement was assessed for each dimension score as well as for the summated total score. For inter-rater agreement, the scores obtained by each rater were compared with those essment 4 to 7 days later. For dimension score comparison, agreement was quantified with percentage of agreement and quadratic weighted kappa (kw) which is appropriate for ordinal data 13,14. Systematic difference (bias) in dimension scores obtained by essment 4 to 7 days later. For dimension score comparison, agreement was quantified with percentage of agreement and quadratic weighted kappa (kw) which is appropriate for ordinal data [13,14]. Systematic difference (bias) in dimension scores obtained by different raters on the same subject was assessed by the test statistic for marginal homogeneity or symmetry of the marginal frequencies [14]. If bias exists these marginal frequencies will not be symmetrical. The test statistic has a chi-squared distribution with 2 degrees of freedom for a 3 x 3 table. It is computed with a program provided by Brennan and Silman [14]. Statistical significance is accepted at 5% level. For total score comparisons, assuming the score is interval scaled, agreement was quantified with intraclass correlation coefficient (ICC) calculated by repeated measure analysis of variance [13]. Systematic difference (bias) in total scores between raters was assessed by paired t test and by calculating the 95% confidence interval for the mean difference in scores between raters [14]. As there were numerous pairwise comparisons; we present only the mean and range of the values of Kw and ICC.

Validity assessment

In the absence of a gold standard measure for health related QOL as well as other previously validated QOL measure in this country for comparison; we can only examine construct validity [13,15]. In such validation, hypothesis or construct concerning differences in quality of life between groups of patients are examined by the measures being validated. We formulate the following hypotheses a priori: the QL-index total score is significantly lower (indicating poorer quality of life) in: -

1. HD patients who were older.
2. HD patients with significant comorbid disorders.
3. HD patients who had been hospitalised in the previous 2 months.
4. HD patients aged less than 55 years and who were unable to work.
5. HD patients who were incapable of self-care HD.
6. HD patients whose serum haemoglobin levels were lower.
7. HD patients whose serum albumin levels were lower.
8. HD patients compared with renal transplant patients with a serum creatinine less than 200 umol/L.

Kruskal-Wallis one way analysis of variance was used to test the above hypotheses. Statistical significance is accepted at 5% level. To the extent that the greater number of the above hypotheses are supported by the data, the greater then is the confidence we may have in the validity of QL-index.

Discriminatory ability

This refers to the ability of the instrument to discriminate between different levels of health related QOL. This is assessed by the frequency distribution of the total score obtained from the measure; a highly skewed distribution indicates poor ability in discerning subtle differences in level of health related QOL between subjects.

Results

59 chronic HD patients participated in this study. Their mean age was 50 years (range 14 to 79), 66% male, 36% were part or full time employed and 20% were diabetics. Their mean (SD) serum albumin was 43 (6) g/L.

Inter-rater agreement (Table I)

Only 58 patients had 3 measurements required for

Table I
Reliability of Spitzer's QL index in dialysis patients

	Score range	Inter rater agreement			Intra rater agreement		
		n	Percentage agreement mean (range)	Weighted Kappa mean (range)	n	Percentage agreement mean (range)	Weighted Kappa mean (range)
Activity	0 - 2	58	0.70 (0.46-0.87)	0.53 (0.29-0.75)	54	0.73 (0.57-0.88)	0.62 (0.58-0.66)
Activity of daily living	0 - 2	58	0.74 (0.53-0.93)	0.41 (0.15-0.68)	54	0.76 (0.61-0.92)	0.61 (0.45-0.78)
Feeling of healthiness	0 - 2	58	0.73 (0.67-0.83)	0.44 (0.35-0.52)	54	0.85 (0.71-1)	0.72 (0.43-1.00)
Social support	0 - 2	58	0.82 (0.64-1)	0.61 (0.07-1.00)	54	0.87 (0.75-1)	0.79 (0.57-1.00)
Psychological outlook	0 - 2	58	0.77 (0.61-0.86)	0.40 (0.14-0.68)	54	0.91 (0.82-1)	0.79 (0.58-1.00)
			intra-class correlation coefficient mean (range)			intra-class correlation coefficient mean (range)	
Total score	0 -10	58	0.66 (0.47-0.81)		54	0.71 (0.68-0.73)	

inter-rater assessment. Percentage agreement values for dimension scores among raters were generally high, ranging from 0.46 to a perfect 1. Weighted kappa values however were more variable; ranging from as low as 0.07 to perfect 1. All mean weighted kappa values exceeded 0.4. Systematic differences (bias) in dimension scores obtained by different raters were evident as shown in Table II. The mean intraclass correlation coefficient (ICC) for total score comparisons among raters was 0.66 (Table I); much better than the weighted kappa obtained from individual dimension score comparisons. Systematic differences (bias) in total scores obtained by raters were only evident in one centre and only between dialysis staff rater and the Nephrologist rater (Table II). The score obtained by the Nephrologist was systematically higher than both dialysis staff in that centre. For rater 1 versus Nephrologist; mean difference (d= staff rater score - Nephrologist score) in score was -1.3, 95% confidence interval of mean difference (95% CI) was -2.0 to -0.6. For rater 2 vs Nephrologist, d= -1.4 and 95% CI for d -2.0 to -0.9.

Intra-rater agreement (Table II)

Only 54 patients had repeated measurements required for intra-rater assessment. Both percentage agreement and weighted kappa were generally high; and as expected better than inter-rater agreement results. Likewise was the ICC for total score comparisons.

Validity (Table III)

The total QL-index scores obtained by the Nephrologist rater on 59 HD patients and 22 transplant patients were analysed. Table III shows the distribution of total QL-index scores by age, serum haemoglobulin (HB) level, serum albumin level, previous hospitalisation, capacity for self-care HD, modality of renal replacement therapy, rehabilitation status and for patients with comorbid disorders like diabetes, cardiovascular disorders and renal bone disease. All the hypotheses with regard to differences in QL-index scores among patient groups, except for grouping by serum Hb level, that we have stipulated in advance were supported; thus providing evidence of construct validity.

Significant age gradients were found while no significant difference was found between men and

Table II
Systematic difference (bias) in QL-index scores between raters

Dimension	Centre 1 (n=28 patients)			Centre 2 (n=30 patients)		
	rater 1 vs rater 2	rater 1 vs nephrologist	rater 2 vs nephrologist	rater 3 vs rater 4	rater 3 vs nephrologist	rater 4 vs nephrologist
Activity	ns	*	*	*	ns	ns
Activities of daily living	ns	*	ns	*	*	ns
Feeling of healthiness	ns	ns	ns	ns	*	ns
Social support	ns	ns	*	ns	ns	ns
Psychological outlook	*	ns	*	ns	*	ns
Total score	ns	*	**	ns	ns	ns

ns: No significant systematic difference (bias) in score (p>0.05)
** : Significant (P<0.05) difference in score between raters; based on test statistic for marginal homogeneity.*
***: Significant (P<0.05) difference in score between raters; based on pair sample t test.*

women in their QL scores. Patients who had been hospitalised in the previous 2 months, diabetics and those with cardiovascular disorders and renal bone disorders had poorer quality of life. Transplant patients scored better than HD patients. Those HD patients who were capable of self-care and those who could return to part or full time employment had higher scores as expected. Interestingly a significant gradient in QL scores was found for serum albumin level. A lesser gradient was also seen for serum haemoglobin level though this was not statistically significant.

Discriminatory ability (Figure 1)

The frequency distribution of total QL-index scores is shown in Figure 1. This was highly skewed; the median score was 9.

Discussion

In the study, all raters found the instrument easy to understand and to use. Each assessment in this study took less than 5 minutes to complete; supporting the developer's claim that the instrument is easy to administer and that it has found ready acceptance by physicians [7].

The evidence for the construct validity of QL-index is good. The hypothesised differences in scores among various patient groups were mostly observed. The discriminatory ability of QL-index in our HD patient population was poor. The scores were highly skewed to the right indicating poor ability in discerning subtler differences in quality of life among well outpatient HD patients. This is expected. QL-index is a brief instrument comprising only 5 items and few response categories and yet broad in scope. As such, it is not expected to and it cannot discriminate adequately the quality of life of well people, according to the instrument developer [7]. Many outpatient HD patients are relatively well; at least in comparison with other

Table III
Mean total scores of QL-index in relation to sociodemographic and clinical variables

Variable	n	Mean Total QL-Index	p value index
Age (years):			
=<40	13	9.5	0.01
40-60	32	8.7	
>60	14	7.6	
Sex:			
Males	39	8.7	0.27
Females	20	8.4	
Sr. Haemoglobin (mg/dL)			
<7.5	23	8.2	0.49
7.5-10	25	8.8	
>10	11	9.1	
Sr. Albumin (g/L)			
<40	12	6.7	0.002
40-45	27	8.9	
>45	20	9.5	
Hospitalised in previous 2 months:			
Yes	14	6.6	0.0001
No	45	9.3	
Capable of self-care:			
Yes	32	9.5	0.0001
No	27	7.6	
Diabetic:			
Present	12	7.3	0.0007
Absent	47	9.0	
Cardiovascular disorder:			
Present	14	7.6	0.002
Absent	45	9.0	
Renal bone disorder:			
Present	3	6.6	0.02
Absent	56	8.7	
Modality of renal replacement therapy:			
HD	59	8.6	0.009
Transplant	22	9.6	
Rehabilitation status of patient age <55:			
Full or part time employed	21	9.7	0.0001
Unable to work	6	5.7	

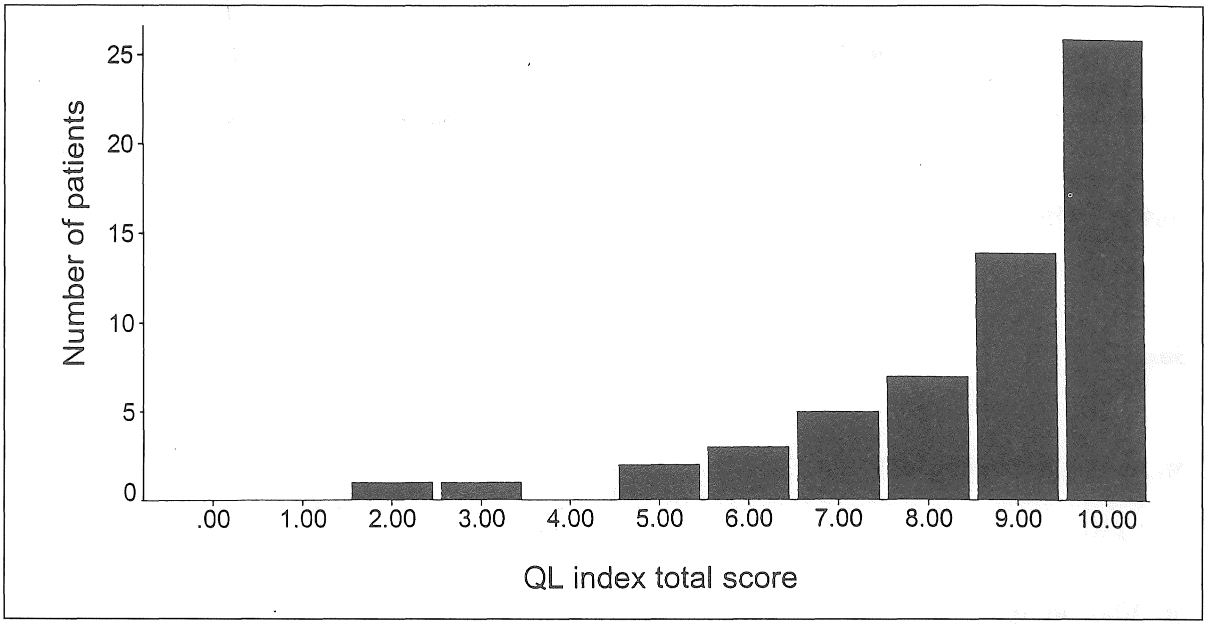

Fig. 1: Frequency distribution of QL index total scores

patients with chronic disease, for example those with cancer or symptomatic cardiovascular disease, or sick in-patients. This however should not detract from its utility for our application. Our intended application is to determine the quality of life outcome of chronic HD patients for the purpose of programme evaluation; in other words, it is intended for application in a group of patients. The mean score of a group of patients is of interest and not that of an individual patient's. It is certainly not intended for application to individual patients as, for example in clinical decision making.

The reliability of QL-index, as found in this study, is a matter of greater concern. For an instrument intended to be applied by a large number of dialysis staff raters on thousands of patients all over the country, reliability of the instrument is of course crucially important. Reliability is expressed as values of quadratic weighted Kappa (Kw) and percentage agreement for dimension score and intra-class correlation coefficient (ICC) for total score. Kw is identical to ICC [16] and may be similarly interpreted. Interpretation of these descriptive statistics however is not straightforward. Fleiss [17] suggested the following interpretation: values less than 0.4 represent poor agreement, 0.4 to 0.75 represent fair to good agreement and values more than 0.75 indicate excellent agreement. Though arbitrary, this is reasonable for QOL measures intended for application in groups of patients as in research or treatment evaluation. For such purposes, conclusions are drawn from mean score averaged across many individuals, and the sample size will serve to reduce the error of measurement in comparison of group differences. For application in individual patients for clinical decision making, a value more than 0.85 is recommended [18].

Thus, as found in this study, intra-rater agreement is acceptable; the lowest Kw was 0.43 and mean Kw values were greater than 0.6. More reassuring, the inter rater agreement for total score is good; mean ICC value was 0.66 (range 0.47 to 0.81). The QL-index total score as a single summary measure of quality of life is likely to be of greatest use to us. As for inter-rater agreement for individual dimension scores, even though all the mean Kw values were 0.4 or higher; many individual values were disappointingly poor. Within rater variation (intra-rater agreement) as discussed above is acceptably small. Thus much of the unreliability in dimension score is attributable to between rater variation. Indeed, significant systematic differences in dimension scores were found among several rater-pairs in this study. This deserves further attention. Variation can be reduced by limiting assessment to a single dialysis staff rater in each centre and devoting greater and better training on the use of the instrument to that selected rater.

Nevertheless, it should be pointed out that another reason can partially explain the low Kappa values obtained. It is known that Kappa has some undesirable properties; in particular, it is affected by the prevalence of the quality of life state being measured [14,19,20]. With high prevalence, as in the case when scores are highly skewed (vast majority of patients scored 2 on a particular dimension), Kappa tended to be low. Feinstein and Cichetti [20] referred to this as the paradox of "high agreement but low Kappa". This has affected several of the comparisons in this study. An extreme example is the inter-rater agreement for social support; percentage agreement was 0.64 and yet the Kappa was only 0.07. At present, this problem cannot be resolved other than for 2 x 2 tables [19,21].

In conclusion, the Spitzer's QL-index has acceptable measurement properties for application in our dialysis programme. We have confidence in its validity. The reliability of its summated score as a summary measure of quality of life is reassuring. Reliability of individual dimension scores needs to be further improved by careful selection of rater and better training. Ongoing monitoring of instrument reliability is essential. The process of validating and assessing any measuring tool for research is not a one-off exercise but should be continuous.

References

1. Disney APS (ed). ANZDATA report. Woodville, South Australia: Australian and New Zealand Dialysis and Transplant Registry, 1994.

2. USRDS 1995 Annual Data Report. Bethesda, MD: URDS, 1995.

3. Tufveson G, Geerlings W, Brunner FP, et al: Combined report on regular dialysis and transplantation in Europe XIX, 1988. Nephrol Dial Transplant 1989; 4 (suppl 4):5-29.

4. Lim TO, ed. First report of the Malaysian Dialysis and Transplant Registry, 1993. Kuala Lumpur, Malaysia.

5. Fitzpatrick R, Fletcher A, Gore S, Jones D, Spiegelhalter D, Cox D. Quality of life measures in health care, I: Applications and issues in assessment. Br Med J 1992; 305: 1074-7.

6. Bergner M. Health status measures: An overview and guide for selection. Ann Rev Public Health 1987; 8: 191-210.

7. Spitzer WD, Dobson AJ, Hall J, Chesterman E, Levi J, Shepherd R et al. Measuring the quality of life of cancer patients: A concise QL-index for use by physicians. J Chron Dis 1981; 34: 585-97.

8. M'Clellan WM, Anson C, Birkeli K, Tuttle E. Functional status and quality of life: predictors of early mortality among patients entering treatment for end stage renal disease. J.Clin Epidemiol 1991; 44: 83 - 89.

9. Guillemin F, Bombardier C, Beaton D. Cross-cultural adaptation of health-related quality of life measures: literature review and proposed guidelines. J Clin Epidemiol 1993; 46: 1417 -32.

10. Blake RL, Vandirer TA. The reliability and validity of a ten item measure of functional status. J Fam Prac 1986; 23: 455-9.

11. Jelles F, Van Bennekern CAM, Lankhorst GJ, Sibbel CJP et al. Inter and intra rater agreement of the rehabilitation activities profile. J Clin Epidemiol 1995; 48: 407-16.

12. Laupacis A, Muirhead N, Keown P, Wang C. A disease specific questionnaire for assessing quality of life on haemodialysis. Nephron 1992; 60: 302-6.

13. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. Oxford: Oxford University Press 1989.

14. Brennan P, Silman A. Statistical methods for assessing observer variability in clinical measures. Br Med J 1992; 304: 1491-4.

15. McDowell I, Newell C. Measuring Health: a guide to rating scales and questionnaires. New York, Oxford University Press 1987.

16. Fleiss JL, Cohen J. The equivalence of weighted Kappa and the intra-class correlation coefficient as measures of reliability. Educational and Psychological Measurements 1973; 33: 613-9.

17. Fleiss JL. Statistical methods for rates and proportions. New York; Wiley 1981.

18. Weiner EA, Stewart BJ. Assessing individuals. Little Brown, Boston 1984.

19. Byrt T, Bishop J, Carlin JB. Bias, prevalence and Kappa. J Clin Epidemiol 1993; 46: 423-9.

20. Feinstein AR, Cicchetti DV. High agreement but low Kappa I. The problems of two paradoxes. J Clin Epidemiol 1990; 43: 543-9.

21. Cicchetti DV, Feinstein AR. High agreement but low Kappa II. Resolving the paradoxes. J Clin Epidemiol 1990; 43:551-8.