EQ-5D-5L and SF-6D utility measures in symptomatic thyroid nodules: Acceptability and psychometric evaluation

Authors: Carlos K.H. Wong[1*] PhD (ORCID: 0000-0002-6895-6071), Brian H.H. Lang[2*] MS (ORCID: 0000-0002-9362-0086), Hill M.S. Yu[1], Cindy L.K. Lam[1] MD

[1] Department of Family Medicine and Primary Care, The University of Hong Kong

[2] Division of Endocrine Surgery, Department of Surgery, The University of Hong Kong

* Joint corresponding author

Address for Correspondence:

Dr Carlos KH Wong

Department of Family Medicine and Primary Care, The University of Hong Kong

3/F, Ap Lei Chau Clinic, 161 Ap Lei Chau Main Street, Ap Lei Chau, Hong kong

Tel.: (852) 25185688, Fax No.: (852) 28147475, Email: carlosho@hku.hk

Dr Brian HH Lang

Department of Surgery, The University of Hong Kong

Queen Mary Hospital, 102 Pokfulam Road, Pok Fu Lam, Hong kong

Tel.: (852) 22554232, Fax No.: (852) 28172291, Email: blang@hku.hk

Keywords: Quality-of-life; thyroid nodules; Psychometrics; Validity; Reliability;

Running Title: Psychometrics of EQ-5D-5L and SF-6D in thyroid nodules

Informed consent: Informed consent was obtained from all individual participants included in the study.

Abstract

Purpose: To examine the acceptability, validity and reliability of EuroQoL 5-dimension 5-level (EQ-5D-5L) and SF-6D health utility measures in patients with symptomatic benign thyroid nodules.

Methods: Data from a randomized controlled trial (ClinicalTrials.gov Identifier: NCT02398721) of 294 patients with symptomatic thyroid nodules were utilized for this psychometric evaluation of HRQOL measurement. Three HRQOL questionnaires, generic 12-item Short Form Health Survey (SF-12v2) and EQ-5D-5L, SF-6D, were interviewer-administered at baseline and 2 weeks afterwards. Responses to SF-6D were transformed to SF-6D utility scores using Hong Kong population scoring algorithm derived by Standard Gamble whereas response to EQ-5D-5L were mapped onto EQ-5D-3L response via interim mapping algorithms and then converted to EQ-5D-5L utility scores using Chinese-specific value set. Construct validity was determined by evaluating Spearman correlation between SF-12v2 scores and utility scores. Two-week test-retest reliability was assessed using intra-class correlation coefficient.

Results: No significant (>15%) floor and ceiling effects were observed for SF-6D utility scores. The SF-6D utility scores had a moderate Spearman rank correlation with the SF-12v2 domain score providing evidence for adequate construct validity. The SF-6D utility scores showed good test-retest reliability (0.794; range: 0.696-0.860). Better reliability was observed in SF-6D utility score than in EQ-5D-5L utility score.

Conclusions: While the EQ-5D-5L instrument was less reproducible, the SF-6D instrument appeared to be applicable, valid, and reliable measure in assessing the HRQOL of Chinese patients with symptomatic benign thyroid nodules. Impact of utility score selection on the effectiveness and cost-effectiveness of clinical interventions targeted to these patients needs further exploration.

Clinical trial number and registry: NCT02398721, ClinicalTrials.gov

Key points

- The SF-6D instrument appeared to be applicable, valid, and reliable measure in assessing the HRQOL of Chinese patients with symptomatic benign thyroid nodules.
- The adoption of SF-6D utility score for health economic evaluation of clinical management on patients with symptomatic benign thyroid nodule was supported.

Manuscript Text

INTRODUCTION

Thyroid nodules are common and can be discovered by clinical palpation in 5% of normal individuals and by high-resolution ultrasonography (USG) in 60% of the general population[1, 2]. Although most nodules are benign and do not grow, some nodules do become large and symptomatic [1, 2]. Some patients may experience pain in the neck, jaw or ear and cosmetic unsightliness that adversely impacts appearance and social functioning. If left untreated, some patients may even experience difficulty with breathing and swallowing and even hoarseness of voice leading to impaired physical and mental health-related quality-of-life (HRQOL).

Patients' HRQOL are not only affected by thyroid nodules themselves, but also by the subsequent treatments. There are essentially three therapeutic options for symptomatic benign thyroid nodules[3]. First is close surveillance with lifetime regular follow-ups. Although it is the least costly option, more costly treatments may follow if the disease progresses. Second is surgical resection or surgery. This is a "one-off" treatment and is associated with a high initial cost and patients may still need regular visits to check on their thyroid function afterwards. Third is non-surgical minimal invasive option which includes any form thermal ablative therapies or ethanol injection. This third option is usually a "one-off" treatment in most instances but it has the highest initial cost of the three options and regular visits are necessary. Recent longitudinal study on 40 patients with symptomatic nodule [4] reported that the thermal ablation after a two-year period was associated with physical and mental aspects of HRQOL improvements, possibly due to the reduction in compressive and cosmetic symptoms[5]. Based on a recent study[6], the non-surgical, minimally invasive option may

lead to greater improvement in HRQOL as compared to surgery but is unlikely to be cost-effective over time at its current price. Besides treatment issues, patients enrolled to conservative management or close monitoring strategy may add uncertainty, anxiety and worries about the development of thyroid cancer. Often, the cost-effectiveness analysis of health interventions requires the input of health utility score, a composite preference-weight of HRQOL theoretically ranging from zero for death to one for full health. Although there is no general recommendation for a reference case and preferable health utility measure for economic evaluations in Hong Kong, a valid and reliable utility instrument is desirable to influence decision-making and cost-effectiveness analysis of treatment options for benign thyroid nodules.

Therefore, measurement of HRQOL and utility scores are important in the subjective outcome and health economic evaluation of management for patients with symptomatic thyroid nodules. Indeed, to our knowledge, there are currently no disease-specific HRQOL instruments specifically designed for symptomatic benign nodular goitre, suggesting the use of existing generic HRQOL and utility score instruments might be acceptable. Although one previous study[6] estimated utility scores of patients with thyroid nodules through EuroQoL 5-dimension (EQ-5D) instrument, the psychometric properties of the commonly used utility instruments, EQ-5D and Short-form 6-dimension (SF-6D) and their pairwise comparisons in patients with symptomatic benign thyroid nodules were not determined. Therefore, the present study aimed to assess the acceptability, validity, and reliability of the EQ-5D and SF-6D instruments in patients with symptomatic benign thyroid nodules by doing a head-to-head comparison on psychometric properties among instruments.

METHODS AND PATIENTS

Study design and Patients

This was an interim analysis of data collected from an ongoing prospective trial (already approved by the local institutional review board and registered with www.clinicalTrials.gov (NCT02398721). Over a 9-month period, consecutive patients presenting for the first-time with a thyroid swelling were evaluated. To be eligible, first, the thyroid swelling had to be benign (i.e. Bethesda class II on fine needle aspiration cytology (FNAC) within 3 months of recruitment and a low or very-low suspicion sonographic pattern on USG). Second, the swelling (which could either be a solitary nodule or a dominant nodule in a multinodular gland) had to be causing obstructive or pressure symptoms. A swelling simply causing non-specific neck complaints or cosmetic concern was not included. Third, the index nodule had to have all three orthogonal dimensions ≥10mm on USG. Also, patients with severe existing medical co-morbidities or terminal malignancy were not eligible. Patients with previous history of thyroid surgery, indeterminate, suspicious of malignancy or malignant FNAC (i.e. Bethesda class III or above) were excluded from current analysis.

After informed consent, eligible patients were interviewed and were asked to fill in a structured HRQOL questionnaire consisting of the traditional Chinese (Hong Kong) version of SF-12 Health Survey version 2, SF-6D, EQ-5D 5-level (EQ-5D-5L), and questions on socio-demographics. As for consistency in administration modes, all patients were interviewer-administered because elderly patients with low literacy level were incompetent to self-complete questionnaires. The SF-12v2, SF-6D and EQ-5D-5L questionnaires were interviewer-administered at baseline, whereas three mentioned questionnaires and global rating of change scale were interview-administered at 2-week after baseline to assess test-retest reliability.

Study Instruments

EuroQoL 5-dimension 5-level (EQ-5D-5L)

The EQ-5D-5L is a generic preference-based measure developed by the EuroQol Group for measurement of health-related quality-of-life, providing descriptions of five dimensions of health status. The EQ-5D-5L has five domain scales (mobility, self-care, usual activities, pain and discomfort, and anxiety and depression) and five levels for each domain. Since the Chinese-specific EQ-5D-5L value set / tariff is currently not available, we applied a two-step indirect approach to estimate EQ-5D-5L scores applicable for Chinese population, as adopted in previous studies[7-9]. The first step was the application of an indirect interim mapping method[10]. The EQ-5D-5L health status was transformed to EQ-5D-3L health status according to the transition probability matrix. Finally, EQ-5D-3L health status were scored according to a recently developed Chinese-specific EQ-5D-3L value set[11], ranging from -0.149 for the worst health status (‘33333’) to 1 for the full health (‘11111’). A higher score in EQ-5D-5L indicated better HRQOL.

Short-form 12-item Health Survey (SF-12)

The Chinese (Hong Kong) SF-12 Health Survey version 2 (SF-12v2) has been validated and normed on the general Chinese population in Hong Kong. It measures eight domains of HRQOL on physical functioning (PF), role physical (RP), bodily pain (BP), general health (GH), vitality (VT), social functioning (SF), role emotional (RE) and mental health (MH) on a scale with theoretical range from 0 to 100. A higher score indicates better HRQOL. The eight domain scores were aggregated based on population-specific weights to calculate two summary scores, the physical (PCS) and mental component summary (MCS) scores.

Short-form 6-dimension (SF-6D)

In this study, the SF-6D utility score was directly converted from raw responses from SF-6D instrument. Specifically, estimation of SF-6D utility score is possible with the available item

responses of either SF-6D or SF-12v2[12] instrument. Previous research show that SF-6D utility score measured by SF-6D instrument was more sensitive and preferable than that derived from SF-12v2 instrument[13]. The Hong Kong SF-6D value set[14, 15] was derived by standard gamble valuation method[12]. The theoretical range of SF-6D utility score ranged from 1 for full health to 0.315 for the worse possible health state according to Chinese Hong Kong population-specific scoring algorithm [14, 15]. The SF-6D utility score serves as preference weighting input to quality-adjusted life-year outcomes in economic evaluation.

Global Rating of Change Scale

The global rating of change scale is a single item "Compared to the first visit, how would you rate your overall health now?" with 7-point Likert scale rating from 'extremely worse' (rating of -3) to 'extremely better' (rating of +3) health condition compared to previous assessment at baseline. It has widely used as an external criterion of change in heath condition[16, 17]. Patients self-rated '0' as indication of 'stable' in health condition change were selected for test-retest reliability assessment.

Statistical Analysis

Descriptive statistics including mean, standard deviation (SD) and percentage of floor and ceiling of each subscale and each summary scale were calculated.

The acceptability of instruments[18] was assessed using the proportion of missing values, and proportion of patients giving highest possible and lowest possible responses, denoted as ceiling effect and floor effect. At least 15% of subjects achieved the lowest or highest possible score was considered as the presence of floor or ceiling effect, respectively. Based on patients with stable in health condition over a 2-week duration, 2-week test-retest reliability was assessed by intra-class correlation (ICC) coefficient using a value ≥ 0.7 to

indicate adequate reproducibility. A weighted Kappa[19] of <0.2 was interpreted as poor agreement of individual domain responses between baseline and at 2-week assessments, 0.21-0.4 as fair, 0.41-0.6 as moderate, 0.61-0.8 as good and >0.8 as very good.

The construct validity[20] of the SF-6D and EQ-5D-5L utility score was assessed using correlation test against the SF-12v2 subscale scores holding similar constructs. Concerning the possible violation of normality assumption for HRQOL and health utility scores, Spearman's rank correlation test was used. We hypothesized that the utility scores were moderately correlated with SF-12v2 summary scores, since those scores were composite scores of different important HRQOL domain scores. Data analyses were conducted using SPSS Windows 23.0 (IBM SPSS Inc., Chicago, IL, USA). P-value < 0.05 was statistically significant.

Results

Baseline patient characteristics

A total of 314 patients were enrolled and of these, 20 patients (6.5%) with previous history of thyroid surgery were excluded. Therefore, a total of 294 patients was included for analysis. Table 1 shows the characteristics of patients included in this study. The male:female ratio was 1:6.4. The mean (± SD) age at presentation was 56.59 ± 11.67 years. The mean body mass index was 23.33 ± 3.29kg/m^2. All of them were euthyroid on presentation. The mean serum TSH level was 1.33 ± 0.99mIU/L (normal: 0.05 – 4.2mIU/L). The majority (n=225, 71.7%) had multiple benign-looking nodules in their gland (i.e. multinodular goitre). The mean size of the dominant (largest) nodule was 2.4 ± 2.3cm (range: 1.6 – 3.8).

Figure 1 plots the response distributions related to all dimensions of both the EQ-5D-5L and SF-6D instruments. Table 2 entails the descriptive statistics of health utility scores and detailed response distributions of EQ-5D-5L and SF-6D domains. The highest proportion of 'no problems' response referred to social functioning domain in SF-6D and self-care domain in EQ-5D-5L. The mean EQ-5D-5L and SF-6D utility scores were 0.901 ± 0.113 and 0.773 ± 0.139, respectively, where distribution of EQ-5D-5L was heavily left-skewed. Ceiling effect was observed for EQ-5D-5L. No significant (>15%) floor and ceiling effects were observed for SF-6D utility score. Table 3 depicts the two-week test-retest reliability among HRQOL and utility instruments. For two-week test-retest assessment, 130 patients completed the full set of HRQOL questionnaires. The mean follow-up duration was 17.1 days, with range from 9 days to 51 days. Based on an external criterion indicating the change in health status compared to baseline assessment, 104 patients perceived stable in health condition over the 2-week duration, where the retest time frame is commonly adopted for test-retest reliability[21]. The SF-6D utility scores showed good test-retest reliability (0.794; range: 0.696-0.860) but ICC of EQ-5D-5L score was less than 0.7.

Weighted kappa, indicating agreement between two assessments, was interpreted as fair-moderate (0.223-0.445) for 6 dimensions of SF-6D, and poor-moderate (0.149-0.344) for 5 dimensions of EQ-5D-5L. Better reliability was observed in SF-6D utility score than in EQ-5D utility score.

The SF-6D utility score had a moderate-strong Spearman rank correlation with SF-12v2 domain and summary scores (0.461-0.630) that conceptually measures the similar construct providing evidence for adequate construct validity (Table 4). Only fair-to-moderate Spearman rank correlations between EQ-5D-5L score (0.310-0.552), and SF-12v2 domain scores were observed. Particularly Spearman correlations between EQ-5D-5L score and SF-12v2 summary scores were fair-to-moderate (0.257-0.457).

Discussion

This psychometric validation study is the first report to evaluate acceptability, reliability, and validity of HRQOL and utility instruments in patients with symptomatic benign thyroid nodules. In evaluating acceptability, about half of patients self-reported the EQ-5D-5L health profile of '11111', suggesting an observed ceiling effect for the EQ-5D-5L score. As such, EQ-5D-5L score may reflect the better health status condition of benign thyroid nodules and the lacking in room for improvement in health utility score due to clinical interventions. There was no observed floor and ceiling effects for the SF-6D score, demonstrating substantial acceptability of two instruments. The reliability of instruments are essential psychometric property of the HRQOL and utility assessment. It refers to the ability to reproduce consistent responses within a short period of time. The test-retest reliability of SF-6D score was satisfactory, whereas EQ-5D-5L score had an inadequate reproducibility. The response test-retest reliability measured by weighted kappa was fair to moderate for SF-6D

dimensions but poor to moderate for EQ-5D-5L dimensions, in line with test-retest reliability of their utility scores. The reliability of SF-6D instrument was superior to that of EQ-5D-5L instrument for symptomatic benign thyroid nodules. To ascertain the construct validity of HRQOL and utility instrument, the generic SF-12v2 instrument was used as an anchor assessing whether dimension of SF-12v2 holds good correlation with dimensions in other instruments holding similar construct. Correlation between SF-12v2 dimension scores and utility scores reflected the significant strength of SF-6D over EQ-5D-5L. All subscale and summary scores of SF-12v2 were more correlated with SF-6D score than EQ-5D-5L score, in part due to the fact that items in SF-6D were extracted from SF-12 instrument. Moreover, SF-6D score had moderate correlation to the physical (r=0.630) and mental (r=0.461) composite summary scores of SF-12v2 but weak correlation (r=0.257) between the EQ-5D-5L score and mental composite summary score of SF-12v2 was observed.

One possible explanation for the differences in psychometric properties between SF-6D and EQ-5D-5L utility scores was in part due to the differential scoring algorithms available for the use in Hong Kong general and patient populations. Performance of acceptability and psychometric properties may be influenced by the scoring algorithms used for conversion. Unlike direct approach for calculation of SF-6D score, EQ-5D-5L utility score was derived via two-step indirect approach, interim mapping algorithm plus EQ-5D-5L value set derived by mainland China, which neither of them was not established based on Hong Kong Chinese population. Moderate correlation (r=0.556) between two utility scores reflected the reality that concepts in one utility score were not perfectly captured by another utility score, and thus echoed importance of utility instrument selection. Impact of utility score selection, referring to SF-6D score against EQ-5D-5L score, on quality-adjusted life-years, incremental cost-effectiveness ratio and decision making of clinical interventions for thyroid nodule needs further exploration.

Limitations

Notably, four instruments used in current study may not be the best available HRQOL and utility instruments designed for benign thyroid nodules. A recent systematic review[22] appraising quality of 14 standardized instruments recommended the use of ThyPRO instrument, initially developed by Danish research group, for HRQOL assessment in patients with benign thyroid diseases. The cross-cultural validity of ThyPRO have been assessed at multiple clinical sites at multiple countries[23]. However, the Chinese version, irrespective of simplified or traditional version, has not been validated and was available at the time of study. Future works on psychometric properties of ThyPRO instrument or other existing HRQOL instruments specific to symptomatic thyroid nodules are warranted to enable the subjective evaluation of clinical interventions over time as well as integration of HRQOL measurement into routine care of thyroid nodules. Secondly, unlike psychometric testing on instruments used for moderate and severe disease, HRQOL scores in benign thyroid nodule were not hypothesized to have significant correlations with any clinical characteristics such as size of nodule, number of nodules and treatment modalities. Therefore, no analysis was conducted to assess correlations between HRQOL and clinical characteristics. Thirdly, the SF-12v2 scores were in principle more correlated with SF-6D scores than other scores, given the fact that SF-6D is the abbreviated form of SF-12v2. There was a possibility of favoring SF-6D in evaluation of construct validity. In current study, The SF-6D score was not calculated from seven items responses from SF-12v2 instrument, i.e. SF-12 derived SF-6D score. Both the SF-6D and SF-12v2 instruments were separately administered to each patient, gathering different set of responses from two instruments. The SF-12 was assumed to be ‘gold standard’ of HRQOL measurement testing against other HRQOL and utility scores. Nevertheless, this is presumably the best ‘gold standard’ among four instruments because this generic instrument has been validated in Hong Kong general population[24]. Since

instruments were not administered in a randomized order, there was also the possibility of context effect (or 'order effect') on HRQOL and utility measurement. Fourth, study was the secondary analyses of data collected within a RCT design. Since the RCT was not designed for purpose of psychometric evaluation, this paper lacks important information for a comprehensive acceptability investigation such as time for instrument completion, and patients' perceived views and feedback on each instrument. Finally, the patients were sampled from endocrinology surgical outpatient clinic of one hospital in Hong Kong and so our findings might be less generalizable in non-Chinese populations.

Conclusions

The SF-6D instrument demonstrated satisfactory acceptability and psychometric properties in patients with symptomatic benign thyroid nodule, whilst the EQ-5D-5L instrument was less reproducible than SF-6D instrument at the test-retest assessment. These findings supported the use of the SF-6D instrument to evaluate the HRQOL in routine care as well as the adoption of its utility score in the health economic evaluation of patients with symptomatic benign thyroid nodule.

Compliance with Ethical Standards

Financial support for this study was provided by Health and Medical Research Fund (HMRF#12132941) of Food and Health Bureau, HKSAR. The funding agreement ensured the authors independence in designing the study, interpreting the data, writing, and publishing the report.

Conflict of Interest: Carlos KH Wong, Brian HH Lang, Hill MS Yu, and Cindy LK Lam declare that he / she has no conflict of interest.

Ethical approval: All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent: Informed consent was obtained from all individual participants included in the study.

Author contributions: CKHW wrote the manuscript, researched data, contributed to statistical analysis and interpretation of results. BHHL contributed to study design, acquisition of data and reviewed/edited the manuscript. HMSY contributed to acquisition of data and reviewed/edited the manuscript. CLKL reviewed/edited the manuscript.

References

1. Gharib H, Papini E, Garber JR, Duick DS, Harrell RM, Hegedüs L et al. American Association of Clinical Endocrinologists, American College of Endocrinology, and Associazione Medici Endocrinologi Medical Guidelines for Clinical Practice for the Diagnosis and Management of Thyroid Nodules – 2016 Update. Endocrine Practice. 2016;22(5):622-39. doi:10.4158/EP161208.GL.

2. Haugen BR, Alexander EK, Bible KC, Doherty GM, Mandel SJ, Nikiforov YE et al. 2015 American Thyroid Association Management Guidelines for Adult Patients with Thyroid Nodules and Differentiated Thyroid Cancer: The American Thyroid Association Guidelines Task Force on Thyroid Nodules and Differentiated Thyroid Cancer. Thyroid. 2016;26(1):1-133. doi:10.1089/thy.2015.0020.

3. Che Y, Jin S, Shi C, Wang L, Zhang X, Li Y et al. Treatment of Benign Thyroid Nodules: Comparison of Surgery with Radiofrequency Ablation. American Journal of Neuroradiology. 2015;36(7):1321-5. doi:10.3174/ajnr.A4276.

4. Valcavi R, Tsamatropoulos P. Health-related quality of life after percutaneous radiofrequency ablation of cold, solid, benign thyroid nodules: a 2-year follow-up study in 40 patients. Endocrine practice. 2015;21(8):887-96. doi:10.4158/ep15676.or.

5. Spiezia S, Garberoglio R, Milone F, Ramundo V, Caiazzo C, Assanti AP et al. Thyroid nodules and related symptoms are stably controlled two years after radiofrequency thermal ablation. Thyroid. 2009;19(3):219-25. doi:10.1089/thy.2008.0202.

6. Yue W-W, Li X-L, Xu H-X, Lu F, Sun L-P, Guo L-H et al. Quality of Life and Cost-Effectiveness of Radiofrequency Ablation versus Open Surgery for Benign Thyroid Nodules: a retrospective cohort study. Scientific Reports. 2016;6:37838. doi:10.1038/srep37838.

7. Pan C-W, Sun H-P, Zhou H-J, Ma Q, Xu Y, Luo N et al. Valuing Health-Related Quality of Life in Type 2 Diabetes Patients in China. Medical Decision Making. 2016;36(2):234-41. doi:10.1177/0272989x15606903.

8. Pan C-W, Sun H-P, Wang X, Ma Q, Xu Y, Luo N et al. The EQ-5D-5L index score is more discriminative than the EQ-5D-3L index score in diabetes patients. Quality of Life Research. 2015;24(7):1767-74. doi:10.1007/s11136-014-0902-6.

9. Cheung PHW, Wong CKH, Samartzis D, Luk KKD, Lam CLK, Cheung KMC et al. Psychometric validation of the EuroQoL 5-Dimension 5-Level (EQ-5D-5L) in Chinese patients with adolescent idiopathic scoliosis. Scoliosis and Spinal Disorders. 2016;11(1):19. doi:10.1186/s13013-016-0083-x.

10. van Hout B, Janssen MF, Feng Y-S, Kohlmann T, Busschbach J, Golicki D et al. Interim Scoring for the EQ-5D-5L: Mapping the EQ-5D-5L to EQ-5D-3L Value Sets. Value in Health. 2012;15(5):708-15. doi:10.1016/j.jval.2012.02.008.

11. Liu GG, Wu H, Li M, Gao C, Luo N. Chinese Time Trade-Off Values for EQ-5D Health States. Value in Health. 2014;17(5):597-604. doi:10.1016/j.jval.2014.05.007.
12. Brazier JE, Roberts J. The estimation of a preference-based measure of health from the SF-12. Medical care. 2004;42(9):851-9.
13. Wong CKH, Lam ETP, Lam CLK. Comparison of direct-measured and derived short form six dimensions (SF-6D) health preference values among chronic hepatitis B patients. Quality of Life Research. 2013;22(10):2973-81. doi:10.1007/s11136-013-0403-z.
14. Lam CLK, Brazier J, McGhee SM. Valuation of the SF-6D Health States Is Feasible, Acceptable, Reliable, and Valid in a Chinese Population. Value in Health. 2008;11(2):295-303. doi:10.1111/j.1524-4733.2007.00233.x.
15. McGhee SM, Brazier J, Lam CL, Wong LC, Chau J, Cheung A et al. Quality-adjusted life years: population-specific measurement of the quality component. Hong Kong medical journal. 2011;17 (Suppl 6):17-21.
16. Guyatt G, Walter S, Norman G. Measuring change over time: Assessing the usefulness of evaluative instruments. J Chronic Dis. 1987;40(2):171-8. doi:10.1016/0021-9681(87)90069-5.
17. Guyatt GH, Deyo RA, Charlson M, Levine MN, Mitchell A. Responsiveness and validity in health status measurement: A clarification. Journal of Clinical Epidemiology. 1989;42(5):403-8. doi:10.1016/0895-4356(89)90128-5.
18. Fitzpatrick R, Davey C, Buxton M, Jones D. Evaluating patient-based outcome measures for use in clinical trials: a review. Health Technology Assessment. 1998;2(14):1-74. doi:10.3310/hta2140.
19. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychological bulletin. 1968;70(4):213-20.
20. Terwee CB, Bot SDM, de Boer MR, van der Windt DAWM, Knol DL, Dekker J et al. Quality criteria were proposed for measurement properties of health status questionnaires. Journal of Clinical Epidemiology. 2007;60(1):34-42. doi:10.1016/j.jclinepi.2006.03.012.
21. Marx RG, Menezes A, Horovitz L, Jones EC, Warren RF. A comparison of two time intervals for test-retest reliability of health status instruments. Journal of Clinical Epidemiology. 2003;56(8):730-5. doi:10.1016/S0895-4356(03)00084-2.
22. Wong CKH, Lang BHH, Lam CLK. A systematic review of quality of thyroid-specific health-related quality-of-life instruments recommends ThyPRO for patients with benign thyroid diseases. Journal of Clinical Epidemiology. 2016;78:63-72. doi:10.1016/j.jclinepi.2016.03.006.
23. Watt T, Barbesino G, Bjorner JB, Bonnema SJ, Bukvic B, Drummond R et al. Cross-cultural validity of the thyroid-specific quality-of-life patient-reported outcome measure, ThyPRO. Quality of Life Research. 2015;24(3):769-80. doi:10.1007/s11136-014-0798-1.
24. Lam CLK, Tse EYY, Gandek B. Is the Standard SF-12 Health Survey Valid and

Equivalent for a Chinese Population? Quality of Life Research. 2005;14(2):539-47.

Figure Legend

Figure 1. Response distribution of domains in SF-6D instrument (upper) and EQ-5D-5L instrument (lower).

Tables

Table 1. Baseline Characteristics of Patients with symptomatic thyroid nodules

Characteristics	Total (N=294)		Characteristics	Total (N=294)	
	N	%		N	%
Gender			The largest nodule		
Male	40	13.61	Length (cm, Mean±SD)	1.52	0.37
Female	254	86.39	Width (cm, Mean±SD)	1.26	0.42
Age (years, Mean±SD)	56.59	11.67	Height (cm, Mean±SD)	1.21	0.42
Education			Volume (mL, Mean±SD)	1.47	1.17
No formal schooling	10	3.42	Position		
Primary	54	18.49	Upper	25	9.26
Secondary	149	51.03	Middle	107	39.63
Tertiary or above	79	27.05	Lower	138	51.11
Marital Status			Side		
Married	211	72.01	Right	145	50.88
Single	36	12.29	Left	123	43.16
Widow(er)	24	8.19	Isthnmus	17	5.96
Separated or divorced	22	7.51	Number of nodules		
Currently Working			1	37	13.03
Yes	180	61.43	2	41	14.44
No	113	38.57	3	47	16.55
Unknown			4	54	19.01
Monthly income			5	29	10.21
≤HKD20,000	139	47.44	≥6-<10	28	9.86
>HKD20,000	154	52.56	≥10	48	16.90
BMI (kg/m^2, Mean±SD)	23.33	3.29	Treatment for thyroid disease		
serum TSH (mIU/L, Mean±SD)	1.33	0.99	RAI	3	1.02
FT4 levels (Mean±SD)	18.09	10.43	Radiation	7	2.39
			Operation	38	12.93
			Thyroid-related medication	3	1.02

Note:

SD=standard deviation; BMI=body mass index; RAI=radioactive ablation iodine; TSH=thyroid-stimulating hormone; FT4=free thyroxine; FNAC=fine-needle aspiration cytology

Table 2. Descriptive statistics of health utility scores and response distributions related to all dimensions of EQ-5D-5L and SF-6D instruments at baseline

Utility score	Mean	Standard deviation	Observed Range	Theoretical Range	Floor (%)	Ceiling (%)
SF-6D score	0.773	0.139	0.41-1.00	0.32-1.00	0.0	1.3
EQ-5D-5L score	0.901	0.113	0.51-1.00	-0.15-1.00	0.0	47.1
	Response Distribution					
SF-6D Dimension (%)	1	2	3	4	5	6
Physical functioning	24.5	32.7	28.6	12.2	1.7	0.3
Role functioning	65.0	12.9	7.1	15.0		
Social functioning	72.8	17.0	6.8	2.7	0.7	
Pain	29.6	24.5	22.1	13.9	6.8	3.1
Mental health	24.5	48.0	21.1	5.1	1.4	
Vitality	14.6	45.2	27.9	9.9	2.4	
EQ-5D-5L Dimension (%)	1 (No problems)	2 (Slight problems)	3 (Moderate problems)	4 (Severe problems)	5 (Unable to)	
Mobility	87.4	9.5	3.1	0.0	0.0	
Self-care	97.6	2.0	0.0	0.0	0.3	
Usual activities	91.5	6.8	1.4	0.3	0.0	
Pain/discomfort	50.3	34.7	13.9	0.3	0.7	
Depression/anxiety	72.8	19.0	7.5	0.7	0.0	

Table 3. Two-week test-retest reliability of health utility scores

	Intra-class correlation (n=104)		
Utility score	Estimate	95% CI	
SF-6D score	0.794	0.696	0.860
EQ-5D-5L score	0.575	0.374	0.712
	Weighted Kappa		
	Estimate	95% CI	
SF-6D			
Physical Functioning	0.278	0.159	0.396
Role Limitation	0.342	0.215	0.469
Social Functioning	0.340	0.178	0.501
Pain	0.223	0.112	0.334
Mental Health	0.445	0.319	0.570
Vitality	0.404	0.289	0.519
EQ-5D-5L			
Mobility	0.149	-0.007	0.306
Self-care	0.241	0.098	0.384
Usual activities	0.160	-0.033	0.352
Pain/discomfort	0.344	0.193	0.495
Depression/anxiety	0.231	0.072	0.390

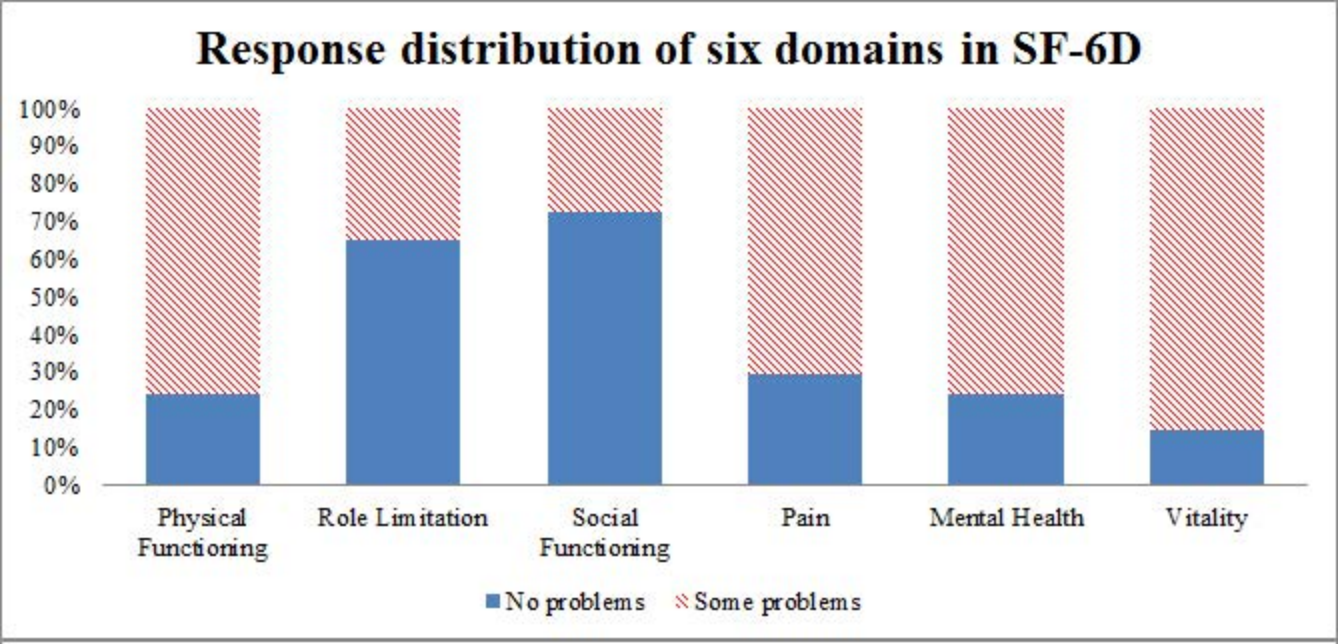
Note: CI = confidence interval

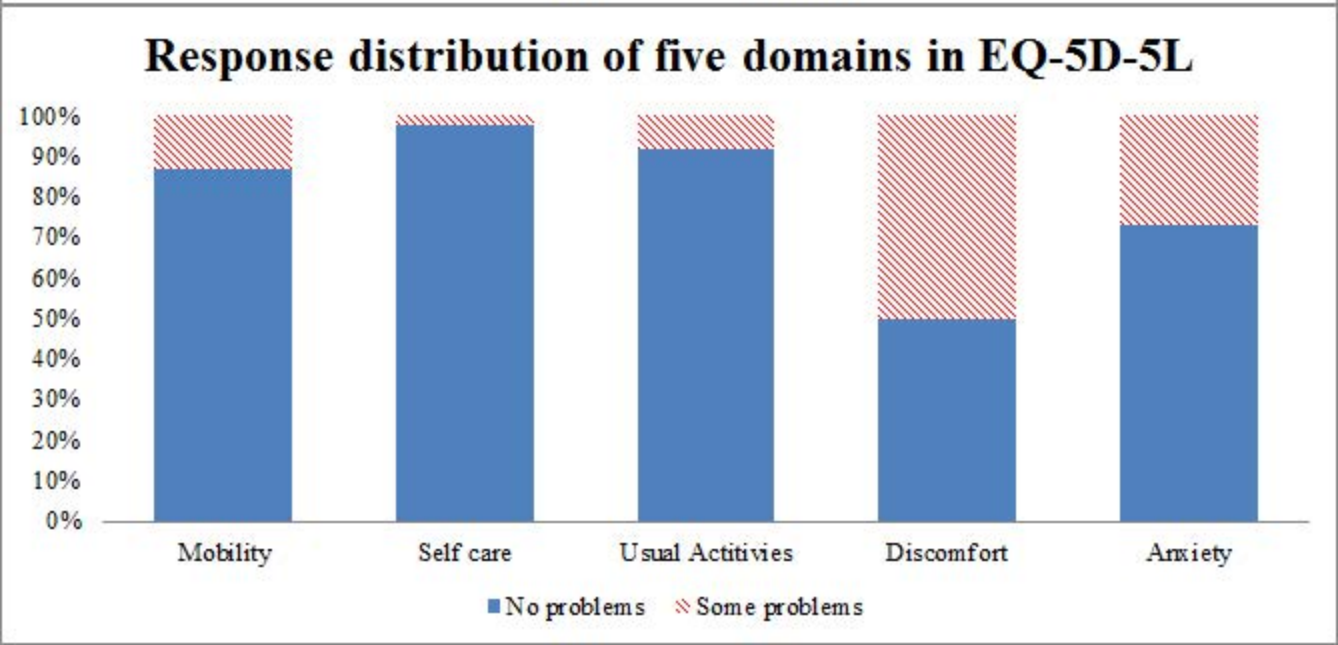
Table 4. Spearman Correlation Coefficients between SF-12v2 subscale and summary scores and the health utility scores

	SF-12v2 Subscale and Summary Scores									
Utility score	PF	RP	BP	GH	VT	SF	RE	MH	PCS	MCS
SF-6D score	0.581	0.589	0.610	0.495	0.537	0.610	0.501	0.517	0.630	0.461
EQ-5D-5L score	0.372	0.381	0.552	0.321	0.390	0.360	0.310	0.332	0.457	0.257

Note:

PF=physical functioning; RP=role physical; BP=bodily pain; GH=general health; VT=vitality; SF=social functioning; RE=role emotional; MH=mental health; PCS=Physical Composite Summary; MCS=Mental Composite Summary;


Response distribution of six domains in SF-6D
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Physical Functioning
Role Limitation
Social Functioning
Pain
Mental Health
Vitality
No problems
Some problems


Response distribution of five domains in EQ-5D-5L
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Mobility
Self care
Usual Actitivies
Discomfort
Anxiety
No problems
Some problems