THYROID
Volume 00, Number 00, 2026
Mary Ann Liebert, A Part of Sage

DOI: 10.1177/10507256261425624

A Thyroid-Cancer-Specific Utility Index: Development and Valuation of the Thyroid Cancer Quality of Life Index

Carrie E. Cunningham,[1,2] Sam van Dijk,[1,3] Michelle M. Langer,[4] Shaidy Moronta,[1,5] Tianna Herman,[1] Reagan A. Collins,[1] Catherine Digennaro,[1,6] Andrea K. Galvan,[7] Jeena M. Varghese,[7] Miriam Lango,[8] Mohammad Jalali,[2] Sarah Fisher,[9] Elizabeth G. Grubbs,[9] Karen Donelan,[10,11] and John Shannon Swan[2]

Background: Thyroid cancer survivors experience distinctive, persistent burdens that diminish health-related quality of life (HRQoL). Utilities from patient preference-based measures are needed for quality-adjusted life-year estimation and decision-making. Generic instruments lack thyroid-specific content, limiting applicability in this population. We sought to develop a thyroid-cancer-specific utility measure, Thyroid Cancer Quality of Life Index (TCQOLI) to support clinical research, cost-effectiveness analyses, and policy applications.
Methods: We conducted a multicenter, multiphase, mixed-method, cross-sectional study. Phase 1 defined the TCQOLI domains and items using input from multidisciplinary experts and patients with thyroid cancer. Phase 2 evaluated the instrument via cognitive interviews ($n = 50$) and a mailed/phone-assisted psychometric survey in adults with thyroid cancer ($n = 163$), followed by confirmatory factor analysis (CFA) and reliability/validity analyses. Phase 3 valued health states in a separate sample ($n = 103$) using interviewer-administered visual analog scales (VAS; 0–100) and standard gamble (SG). Levels of morbidity in each health domain with VAS and SG were used for assessing preferences for three clinical marker states. We derived a weighted dead-to-full-health lower VAS anchor, estimated a VAS to SG power mapping solution to apply to the model overall, and constructed additive, multiplicative, and unweighted indices. Agreement of the indices with direct VAS was summarized by Pearson r, mean absolute error (MAE), overall standard deviation (OSD) of differences, and intraclass correlation (ICC).
Results: Ten candidate domains were finalized; because one domain, reproduction concern, had the weakest psychometrics and the lowest model weight, the primary instrument uses nine domains (a 10-domain version was also evaluated). CFA supported a general HRQoL factor plus a voice/swallow factor with acceptable composite reliability and model fit. The instrument-level ceiling effect was low (3.8%) with no floor effect. The 9-domain additive multiattribute utility theory index correlated with direct VAS ($r \approx 0.74$–0.75) and showed the best agreement. MAE/OSD was 0.045/0.095 after SG mapping and a good to excellent ICC of 0.74.
Conclusions: TCQOLI is a psychometrically robust, thyroid-cancer-specific, preference-based measure with patient-anchored valuation, suitable for health-economic evaluations. This concise index supports comparative-effectiveness research in thyroid cancer and informs resource allocation across clinical and policy settings.

Keywords: thyroid cancer, preference-based utility, patient-reported outcomes

[1]Department of Surgery, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA.
[2]MGH Institute for Technology Assessment, Harvard Medical School, Boston, Massachusetts, USA.
[3]Department of Surgical Oncology and Gastrointestinal Surgery, Academic Center for Thyroid Disease, Erasmus MC Cancer Institute, Rotterdam, The Netherlands.
[4]Department of Medical Social Sciences, Feinberg School of Medicine of Northwestern University, Chicago, Illinois, USA.
[5]Department of Surgery, Danbury Hospital, Danbury, Connecticut, USA.
[6]Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA.
[7]Endocrine Neoplasia and Hormonal Disorders, University of Texas MD Anderson Cancer Center, Houston, Texas, USA.
[8]Department of Head and Neck Surgery, University of Texas MD Anderson Cancer Center, Houston, Texas, USA.
[9]Department of Surgical Oncology, University of Texas MD Anderson Cancer Center, Houston, Texas, USA.
[10]Heller School for Social Policy and Management, Brandeis University, Waltham, Massachusetts, USA.
[11]Mongan Institute, Massachusetts General Hospital, Boston, Massachusetts, USA.

Introduction

Thyroid cancer is common, with over 700,000 survivors in the United States and a lifetime risk of approximately 1.2%.[1,2] Differentiated thyroid cancer accounts for ~90–95% of cases, and 10-year survival exceeds 95%.[3] Despite these favorable outcomes, thyroid cancer survivors often face distinctive psychological, physical, and financial burdens.[4] These concerns result in meaningful decrements in health-related quality of life (HRQoL) comparable to cancers with poorer prognoses.[5,6] The relatively young age at diagnosis and long survivorship further amplify the importance of accurately capturing HRQoL impacts.

Health systems increasingly rely on preference-based measures to quantify how patients value their HRQoL and health states.[7,8] These instruments yield quality-adjusted life years (QALYs), the standard currency in cost-effectiveness analyses, reimbursement, and resource allocation over much of the world. Widely used generic preference-based instruments (EQ-5D,[9,10] SF-6D,[11] HUI[12]) often lack content validity and have limited sensitivity with notable ceiling effects, indicating such instruments are unable to detect expected changes in quality of life in the patients with thyroid cancer.[13] This gap motivates a validated, responsive, patient-centered, disease-specific utility measure for thyroid cancer.[14]

We developed the Thyroid Cancer Quality of Life Index (TCQOLI), a compact, thyroid-cancer-specific instrument built around domains prioritized by patients and clinicians. In this study, we (1) describe the TCQOLI domains and item development; (2) establish its factor structure, reliability, and validity; and (3) conduct a valuation survey to derive a scoring algorithm that converts TCQOLI responses to utilities suitable for QALY-based applications.[12,15–17] The result is a patient-valued, thyroid-cancer-specific utility measure ready for application in clinical research, health policy, and cost-effectiveness analyses.

Methods

Study overview and reporting

We conducted a multicenter, multiphase, mixed-methods, cross-sectional study in three phases. The psychometric analysis phases include (1) TCQOLI domain and item development and (2) cognitive interviewing and psychometric evaluation, followed by (3) valuation of TCQOLI health states to utilities (Fig. 1). Institutional review boards (IRBs) at the participating centers approved all procedures, and participants gave informed consent (protocols Massachusetts General Hospital [MGH] 2002P002595, The University of Texas MD Anderson Cancer Center [MDACC] 2021-0793).

Setting and participants

This study was conducted at two tertiary cancer centers from March 2021 to September 2025. Across phases, we enrolled adults (≥21 years) with a confirmed thyroid cancer diagnosis (papillary, follicular, medullary, or anaplastic) within the prior 10 years at any point in the care pathway (active surveillance, preoperative, postoperative, routine follow-up, or long-term survivorship). We excluded non-English speakers, patients with noninvasive follicular thyroid neoplasm with papillary-like nuclear features, and those with a prior nonthyroid primary cancer diagnosed or treated within five years, with exceptions for basal cell carcinomas. Potential participants were identified in clinic and, according to the study phase (Fig. 1), were mailed either the

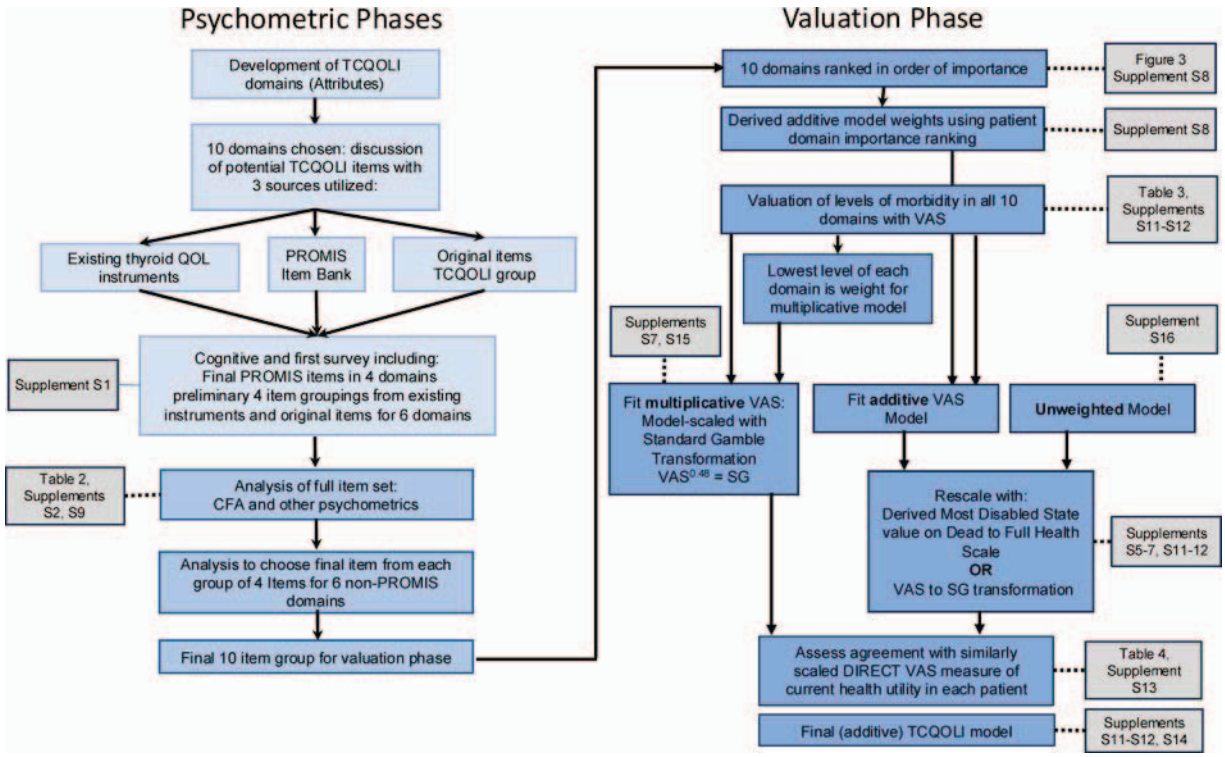

FIG. 1. Overall study flowchart for TCQOLI development. TCQOLI, Thyroid Cancer Quality of Life Index.

psychometric survey or the valuation survey. For each phase of the study, trained research assistants subsequently conducted a structured telephone interview to complete and verify responses (~20–30 minutes for the psychometric survey; ~30–40 minutes for the valuation survey). Full instruments were mailed or emailed to respondents in advance of the interview so that the person could attempt to complete the items on their own; if the instrument was not available to the person, it was resent and the interview rescheduled. Data were captured in a REDCap database with prespecified survey flows, branching logic, audit trails, and clinical/sociodemographic characteristics, which were abstracted from medical records.[18] Patients received a $20 gift card after completing the survey for remuneration. Two distinct patient cohorts contributed to this study and are reported separately: the psychometric sample (instrument statistical performance) and the valuation sample (QALY/utility estimation).

Instrument development and item selection. Multidisciplinary coinvestigators (thyroid cancer clinicians, survey scientists, and HRQoL methodologists) from both institutions met weekly over the course of months to define a domain framework that blends generic health aspects with disease-specific concerns covering relevant physical and psychological content. We explicitly included a financial toxicity domain due to the known impact and financial vulnerability of long-term survivorship.[19] Candidate items were drawn from the patient-reported outcomes measurement information system (PROMIS) item banks[20,21] (reviewed by the study psychometrician, M.M.L.), and considered from thyroid-cancer-specific or relevant psychometric survey instruments such as the EORTC QLC-THY34,[22] THYCA-QoL,[23] FACIT-COST,[24] THYCAT,[25] COH QOL-CS,[26,27] and VHI10[28] (permissions were obtained for any final items utilized). For the other domains needing content, items were developed by our group. All items used a five-level response set, higher values indicating worse HRQoL. Given the structural independence needed in preference-based (utility) indices, a single item was planned to represent each domain of health.[15] For non-PROMIS domains, we compiled four candidate items per domain and asked patients to complete all items using past-week recall concerning their own quality of life and rank each candidate on subject matter, relevance, and clarity. Participants also ranked the importance of the 10 proposed domains.

Cognitive testing and psychometric analysis

After item drafting and before the full psychometric survey, we conducted iterative think-aloud interviews with targeted probes under the supervision of our survey-scientist coauthor (K.D.). In a sample of 50 adults, interviews (run in iterative rounds) assessed item relevance, interpretation, clarity, the past-week recall period, and the consistency of the five-level response labels, continuing until thematic saturation (i.e., hearing no new concerns or ways of understanding the questions). Minimal wording/layout adjustments were anticipated to improve clarity and reduce ambiguity. The full cognitive questionnaire is provided in Supplementary Data S1.

Participants completed the psychometric survey using past-week recall during phone-assisted interviews following a mailed survey to minimize missing data. For item screening, we summarized response distributions (frequencies, means, standard deviations, missingness), computed corrected item-total and inter-item correlations, and incorporated patient rankings with clinical judgment to select the final items. Differential item functioning (DIF) by sex and education was tested with logistic regression. The goal was a single high-performing item per domain consistent with index construction.

Psychometric analyses (M.L.), including measures of interpretability, composite reliability, convergent validity, known-groups validity, and confirmatory factor analysis (CFA), are detailed in Supplementary Data S2. Psychometric analyses were performed in R (version 4.5; R Core Team, 2025) by M.L., and valuation analyses (below) were performed with MedCalc® Statistical Software version 8.11 (MedCalc Software Ltd, Ostend, Belgium; https://www.medcalc.org; 2025) by J.S.S.

Valuation (converting responses to QALY measurement/utilities)

Overview. As shown in Figure 1, valuation proceeded in a structured sequence: (i) collect visual analog scale (VAS) and standard gamble (SG) inputs, (ii) anchor the "most disabled state" (MDS) on the dead–full-health scale, (iii) estimate a VAS to SG mapping using marker states, and (iv) construct candidate scoring models and compare them with direct VAS.

Inputs (interview-administered). We obtained VAS (0–100) ratings for each domain level, a global in-the-past-week health VAS, and VAS for three clinically progressive marker states based on clinical expertise, and SG choices for the same three marker states. These values were put into the multiattribute utility models (below) and converted to disutilities (1 − utility). SG preferences were elicited using a probabilistic titration grid with iterative risk adjustment to the point of indifference.[29–31] Participant materials are shown in Supplementary Data S3, and a detailed procedural description is shown in Supplementary Data S4.

Anchoring the MDS. To place the VAS on the dead–full-health (normalized to 0–1) scale while accommodating different views about how the worst TCQOLI state relates to "dead," 147 respondents chose one of three VAS scales that best reflected their view of where MDS lies relative to dead. The scale-specific summaries were combined using a weighted mean for the overall group to obtain the study's MDS anchor. This anchor was applied as a linear rescale to the dead-to-full-health scale (see Supplementary Data S5for computational details).

VAS to SG mapping using marker states. Because VAS and SG diverge at poorer health yet share a natural zero, we fit candidate mappings (linear, power-utility, power-disutility, cubic, quadratic) to the marker-state data using regressions through the origin.[15,32] The mapping above gave an estimate of SG conversion from VAS on an MDS to full-health scale. Detailed discussion and analysis in Supplementary Data S6-S7.

Multiattribute utility theory scoring models and agreement. Patients' domain importance rankings (1–10) from Phase 1 were converted to additive weights by reverse-ranking

and normalization so weights sum to 1.0 (Supplementary Data S8). We analyzed three index forms of the model on the most-disabled-to-full-health VAS scale: additive multiattribute utility theory (MAUT),[33] multiplicative MAUT,[15,33] and unweighted.[34] Agreement with direct VAS was summarized using Pearson *r* (association), mean absolute error (MAE), overall standard deviation (OSD; a measure of population standard deviation), mean difference, and intraclass correlation (ICC; two-way mixed, single-measures, absolute-agreement), reported for the most-disabled-to-full-health, dead–full-health, and SG-mapped scales (details in Supplementary Data S11).

Results

We first summarize how TCQOLI content was defined and narrowed to one concise item set.

Domain and item selection

Ten candidate domains were finalized by consensus after iterative multidisciplinary review: recurrence concern, appearance, financial hardship, voice problems, swallowing difficulty, reproduction concern, pain interference, depression, fatigue, and cognitive problems. Four domains (fatigue, pain interference, depression, cognitive) drew items from PROMIS; the remaining domains used items adapted from thyroid-cancer-specific HRQoL instruments or developed *de novo*. Review of PROMIS calibrations prioritized items with strong discrimination.[35,36] Item-response theory-based parameters were used for selection of items by the study psychometrician (M.L.). The selected PROMIS items were pain interference ("How much did pain interfere in your day-to-day activities?"; threshold range "b" –0.2 to 2.0; slope "a" 6.5), depression ("I felt depressed"; b –0.1 to 2.3; a 4.3), fatigue ("How often did you run out of energy?"; b –1.0 to 3.2; a 3.4), and cognitive ("My thinking has been slow"; b –1.9 to –0.1; a 3.2). These items, together with semifinal candidates for non-PROMIS domains, formed the psychometric survey.

Instrument performance (psychometrics)

We next evaluated feasibility and measurement properties of the TCQOLI, starting with cognitive testing, followed by psychometric analyses.

Cognitive testing. Cognitive testing ($n = 50$) confirmed that patients understood item wording, the past-week recall, and the five-level response labels; no structural changes were required. All potential items were completed without difficulty, and the within-domain ranking task (subject matter, relevance, and clarity) was well tolerated. Given no required revisions, these observations were rolled forward into the overall psychometric sample.

Psychometric sample. Of 401 patients screened, 26 were ineligible; 163 completed the psychometric survey (response rate 40.6%). Noncompletions comprised 20 refusals, 84 contacted but did not attempt, and 108 with no response (Fig. 2). Study staff obtained demographic and clinical information from the electronic medical record, under IRB approval. Median age was 53 years (interquartile range [IQR] 24.5); 66% were female (Table 1, Supplementary Data S10). Most participants identified as White (85.3%); 10% reported Hispanic ethnicity. Papillary thyroid cancer predominated (74.8%), followed by follicular (8.0%), medullary (8.0%), poorly differentiated/anaplastic (4.9%), and Hürthle cell (4.3%). Surgical management included total thyroidectomy (64%), lobectomy (36%), and completion thyroidectomy (13%). Approximately half had lymph node metastases at

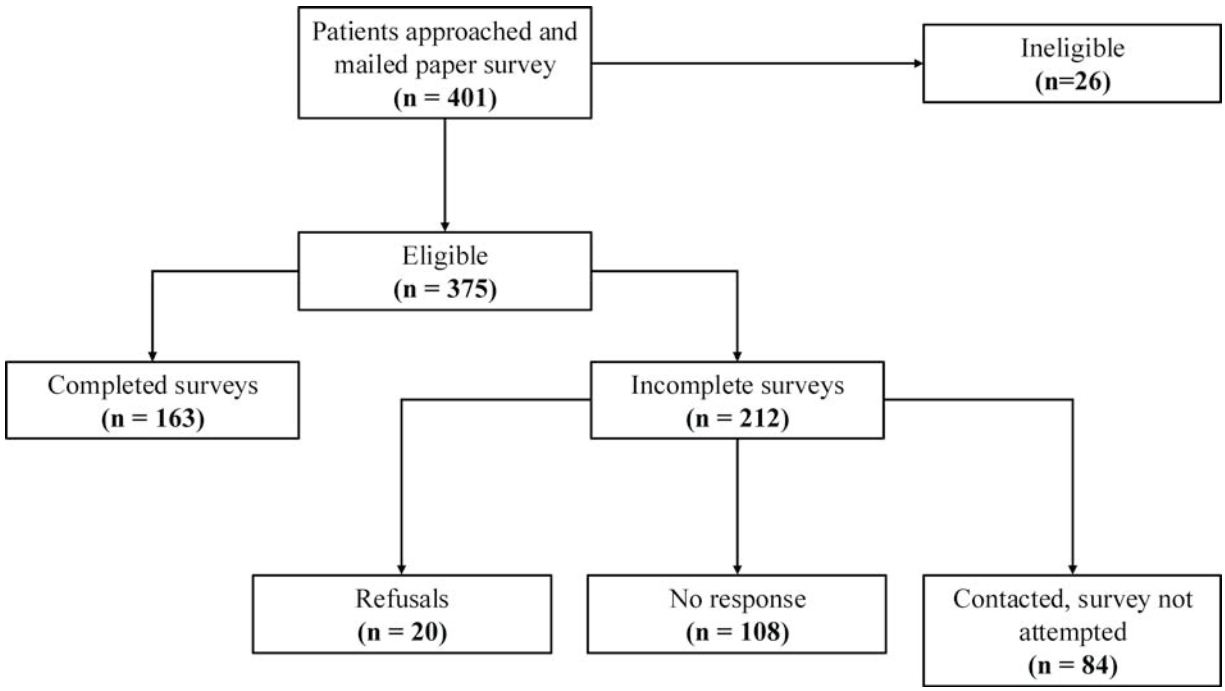

FIG. 2. Participant flow for the psychometric evaluation phase. Counts reflect pooled data across both sites.

Table 1. Sociodemographic and Clinical Characteristics of the Included Patient Cohorts

	Psychometric sample N = *163*			*Valuation sample* N = *105*		
Characteristic	*Overall*	*MGH*	*MDACC*	*Overall*	*MGH*	*MDACC*
Number of patients	163 (100)	114 (70.0)	49 (30)	105 (100)	92 (87.6)	13 (12.4)
Age, years ($n = 2$)						
Mean (SD)	51.8 (15.9)	52.4 (16.3)	50.4 (14.8)	51.8 (15.3)	50.8 (15.7)	58.7 (10.6)
Median (IQR)	53 (24.5)	54 (26.3)	51 (23.0)	52 (24)	51 (25)	59 (6)
Range	22–85	22–85	24–77	21–81	21–81	36–77
Sex at birth						
Male	55 (33.7)	42 (36.8)	13 (26.5)	29 (28)	28 (30)	1 (7.7)
Female	108 (66.3)	72 (63.2)	36 (73.5)	76 (72)	64 (70)	12 (92)
Hispanic/Latinx						
Not Hispanic/Latinx	146 (89.6)	105 (92.1)	41 (83.7)	96 (91)	84 (91)	12 (92)
Hispanic/Latinx	17 (10.4)	9 (7.9)	8 (16.3)	9 (8.6)	8 (8.7)	1 (7.7)
Race						
African American, of African descent, or Black	5 (3.1)	3 (2.6)	2 (4.1)	1 (1.0)	1 (1.1)	0 (0)
Asian or Asian American	10 (6.1)	7 (6.1)	3 (6.1)	6 (5.7)	6 (6.5)	0 (0)
Native American/Alaska Native	0 (0.0)	0 (0.0)	0 (0.0)	0 (0)	0 (0)	0 (0)
Native Hawaiian or another Pacific Islander	0 (0.0)	0 (0.0)	0 (0.0)	0 (0)	0 (0)	0 (0)
Other or more than one race	9 (5.5)	6 (5.3)	3 (6.1)	7 (6.7)	7 (7.6)	0 (0)
White or Caucasian	139 (85.3)	98 (86.0)	41 (83.7)	91 (87)	78 (85)	13 (100)
Thyroid cancer subtype ($n = 1$)						
Papillary	122 (74.8)	82 (71.9)	40 (81.6)	90 (87)	77 (91)	13 (100)
Follicular	13 (8.0)	9 (7.9)	4 (8.2)	7 (6.7)	7 (8.2)	0 (0)
Oncocytic	7 (4.3)	6 (5.3)	1 (2.0)	1 (1.0)	1 (1.2)	0 (0)
Medullary	13 (8.0)	11 (9.6)	2 (4.1)	4 (3.8)	4 (4.4)	0 (0)
Poorly differentiated/Anaplastic	8 (4.9)	6 (5.3)	2 (4.1)	2 (1.9)	2 (2.2)	0 (0)
Type of surgery						
Hemithyroidectomy	59 (36.2)	39 (34.2)	20 (40.8)	37 (35)	34 (37)	3 (23)
Total thyroidectomy	105 (64.4)	74 (64.9)	31 (63.3)	65 (62)	55 (60)	10 (77)
Completion thyroidectomy	21 (12.9)	18 (15.8)	3 (6.1)	12 (11)	12 (13)	0 (0)
Lymph node metastases at initial diagnosis						
No lymph node metastases	79 (48.5)	55 (48.2)	26 (53.1)	57 (54)	51 (55)	6 (46)
Central level VI	66 (40.5)	54 (47.4)	12 (24.5)	38 (36)	38 (41)	0 (0)
Ipsilateral level II–V	41 (25.2)	29 (25.4)	12 (24.5)	20 (19)	17 (18)	3 (23)
Contralateral level II–V	13 (8.0)	11 (9.6)	2 (4.1)	2 (1.9)	2 (2.2)	0 (0)
Distant metastases at initial diagnosis ($n = 4$)						
No	139 (85.3)	104 (91.2)	35 (71.4)	90 (89)	84 (93)	6 (55)
Yes	19 (11.7)	7 (6.1)	12 (24.5)	4 (4.0)	2 (2.2)	2 (18)
Don't know	5 (3.1)	3 (2.6)	2 (4.1)	7 (6.9)	4 (4.4)	3 (27)
Recurrence since diagnosis ($n = 3$)	34 (20.9)	27 (23.7)	7 (14.3)	8 (7.8)	8 (9.0)	0 (0)

Continuous variables are reported as median (IQR); for age, we additionally report mean (SD) and range. Categorical variables are shown as *n* (%). The number of missing observations is shown in parentheses immediately after the variable label. Percentages use the nonmissing denominator and may not total 100 due to rounding. Additional data on education level, household income, and site-specific response rate details are shown in Supplementary Data S10.

IQR, interquartile range; MDACC, The University of Texas MD Anderson Cancer Center; MGH, Massachusetts General Hospital; SD, standard deviation.

diagnosis (52%); distant metastasis was present in 11.7%, and recurrence was documented in 21%.

Item behavior and final item set. We analyzed 163 surveys (114 at Massachusetts General Hospital [MGH]; 49 at MDACC) with low missingness (two refusals on recurrence concern; six on reproduction concern). Non-PROMIS candidates were screened on distributions, missingness, patient rankings, and DIF. Corrected item to total correlations and inter-item correlations were obtained, which are measures of internal consistency of a group of items, or the degree of measurement of the same underlying psychological construct that a scale purports to measure. The selected items and their origin were voice (internally developed, "Changes in my voice have been troubling to me"), swallowing (THYCA-QoL,[23] "Have you had trouble swallowing?"), appearance (COH QOL-CS,[26,27] "Has your illness or treatment caused negative changes in your appearance?", wording clarified

with "negative" added), reproduction concern (internally developed, "Is your ability to have [more] children a concern for you?"), recurrence concern (COH QOL-CS, "To what extent are you fearful of recurrence of your cancer"), financial hardship (FACIT-COST,[24] "My illness has been a financial hardship to my family and me", plus PROMIS pain interference, depression, fatigue, and cognitive problems as above. The response sets are shown in the valuation survey. Item–total correlations ranged from 0.174 to 0.586, and inter-item correlations ranged from 0.105 to 0.341 (Table 2). The reproduction item showed poor variability and the weakest associations.

Patient-prioritized domains. Participants ranked the importance of the ten proposed domains, with recurrence concern as most important and reproduction concern as least important overall (Fig. 3). Among those aged 21–35 years ($n = 32$), reproduction concern remained second-least important (Supplementary Data S8).

Factor structure and invariance. CFA supported a two-factor solution, comprising a general HRQoL factor and a voice/swallow factor. Model misfit was minor. Pain showed the lowest factor loading. Invariance testing supported combining our two sites into one for analysis. Further details are shown in Supplementary Data S9.

Reliability and interpretability. Item-level response distributions showed ceiling effects at the best (least severe) response option ranging from 10.6% to 67.5% (mean 44.3% excluding reproduction; 47.9% including reproduction) and floor effects from 0.6% to 18.0% (mean ≈ 5%). At the full instrument level, TCQOLI total scores showed minimal range restrictions: 3.8% of respondents achieved the maximum score (ceiling) and none scored at the minimum (floor). The patient-reported global VAS showed 10.7% at the ceiling and no floor effect. Composite reliability from the final nine-item structure was acceptable: omega 0.767 (general factor) and 0.699 (voice/swallow factor); coefficient *H* 0.788 and 0.714, respectively.

Table 2. Item-Total and Inter-Item Correlations

Item for each index domain of TCQOLI	*Item-total correlation (corrected)*	*Average inter-item correlation by item*[a]
Appearance	0.505	0.293
Depression	0.489	0.288
Financial hardship	0.483	0.285
Fatigue	0.586	0.341
Pain	0.393	0.234
Recurrence concern	0.480	0.284
Thinking	0.513	0.301
Reproduction concern	0.174	0.105
Swallowing	0.500	0.292
Voice	0.439	0.266

[a]Overall average inter-item correlation was 0.269 including all items.

Item-total correlations measure how related an item is to the sum of all the other items' responses and discriminatory power (expected correlations > 0.3). Inter-item correlations assess each item's correlation to each other item (expected correlations 0.2–0.5).

Known-groups validity. We compared patients with lymph-node metastases at diagnosis to those without ($n = 50$ vs. 48). The voice/swallow factor mean was worse in the nodal group ($p = 0.012$–0.034) as expected, suggesting worse HRQoL. Model fit for this comparison was good by all usual measures of model fit, which is discussed in detail in Supplementary Data S9.

Valuation (converting to utilities)

We next converted TCQOLI responses to utilities by summarizing the valuation cohort, describing the MDS anchor choices and VAS-SG mapping, and estimating a final scoring model.

Valuation sample. In total, 250 patients were approached and mailed the valuation survey. Nine were ineligible, leaving 241 eligible patients. Of these, 92 completed the survey (response rate 38.2%). Among the 149 nonrespondents, 54 refused, 51 were contacted but did not attempt the survey, and 44 did not respond. Among participants who initiated the survey at the primary site ($n = 93$), all but one completed it (Fig. 4). We enrolled 105 patients (92 at MGH; 13 at MDACC). The median age was 52 years (IQR: 39–63); 72% were female. Most identified as White (87%); 8.6% were Hispanic/Latinx. Papillary thyroid cancer predominated (87%), followed by follicular (6.7%), oncocytic (1.0%), medullary (3.8%), and poorly differentiated (1.9%). Surgery was performed in 97%; procedures comprised total thyroidectomy in 64%, hemithyroidectomy in 24%, and hemithyroidectomy with completion thyroidectomy in 12%. At diagnosis, 46% had lymph node metastases; distant metastases were present in 4%, and recurrence occurred in 7.8%. Detailed characteristics are shown in Table 1 and Supplementary Data S10; U.S. race/ethnicity benchmarks are also provided in Supplementary Data S10.

Response quality and task performance. VAS and SG tasks were completed in interviewer-administered sessions. We excluded 30 (31.5%) participants' SG data who failed comprehension and logical checks.

Most Disabled State (MDS) anchoring. Of 147 respondents in an early phase of the study, 119 (81%) selected scales 1 and 2, and 28 (19%) selected scale 3. The 20% trimmed means were 0.2610 for scales 1 and 2, and −0.4212 for scale 3 after transformation. Weighting by scale endorsement frequencies yielded an MDS estimate of 0.13 on the dead (0)–full-health (1.0) VAS (Supplementary Data S5).

Visual Analog Scale (VAS) to Standard Gamble (SG) mapping. The power-utility form $SG = VAS^{0.48}$, which resulted from our marker state regression analyses on an MDS to full-health scale, outperformed linear, power-disutility, and published cubic and quadratic alternatives (Supplementary Data S6).

Estimating and validating utilities: decrements, weights, and agreement analyses. We estimated additive MAUT models for both the 9-domain (summarized in Table 3, detailed in Supplementary Data S11) and 10-domain TCQOLI specifications (Supplementary Data S12-S13). In each model, single-domain disutility increased across response levels (Level 2 ≈ 0.13–0.20; Level 3 ≈ 0.35–0.47; Level 4 ≈ 0.68–0.75. In the

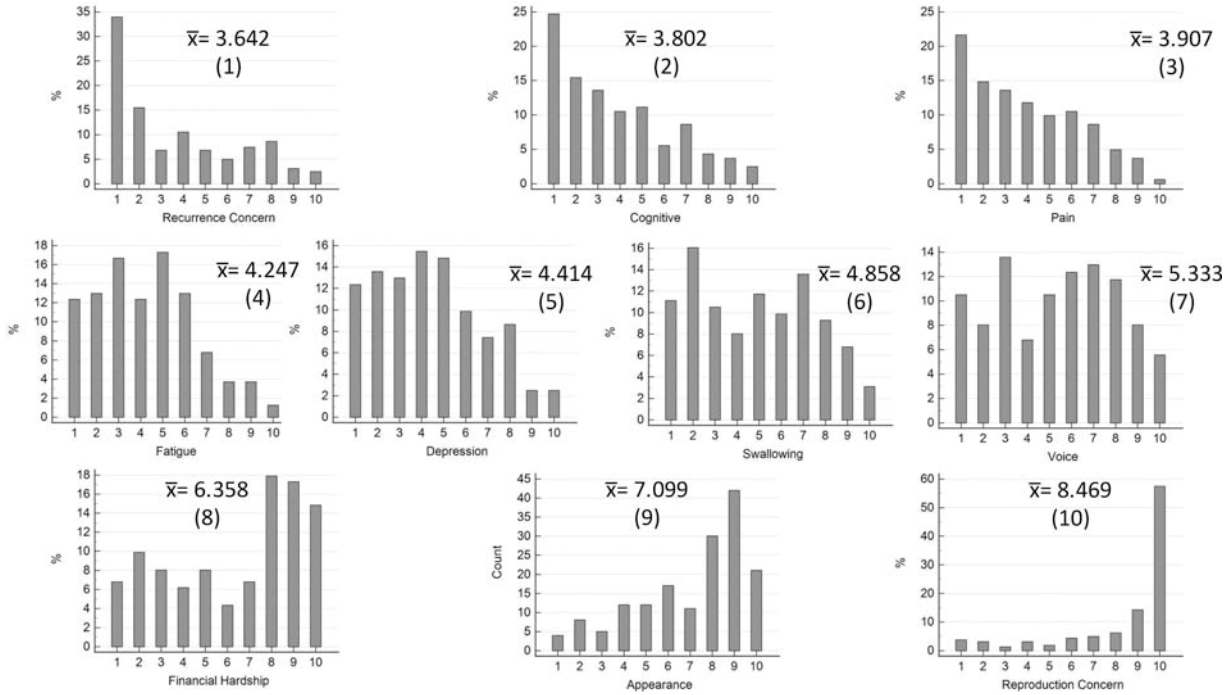

FIG. 3. Importance ranking (1–10) frequencies of TCQOLI Domains by all reporting patients ($n = 162$), with lower numbers equating with higher importance. Ranking frequencies are shown along with mean ($\overline{X}$) importance by domain and overall rank in parentheses.

9-domain additive model, domains with the largest weights were recurrence concern (0.14) and cognitive issues (0.14), followed by pain (0.13), depression (0.12), fatigue (0.12), swallowing (0.11), voice issues (0.10), finance (0.08), and appearance (0.06). In the 10-domain additive MAUT model, weights were relatively similar overall, with reproduction receiving the smallest weight (0.04).

We compared the additive index, multiplicative index, and the unweighted scale against direct VAS under three scaling conditions. In the 9-domain model, agreement was similar with ICC $\sim$0.73–0.74 throughout the additive models, with MAE/OSD best after mapping to SG 0.045/0.095 (Table 4). In the 10-domain model, we observed a similar pattern (ICC $\sim$0.74–0.75) in additive models (Supplementary

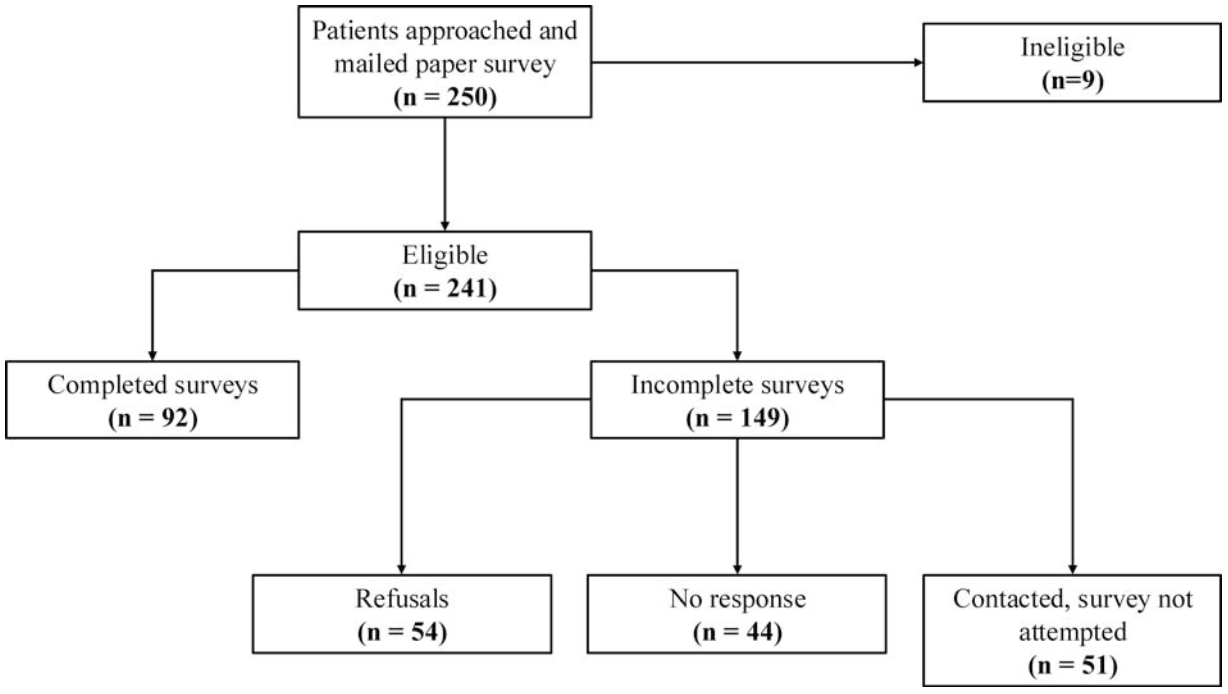

FIG. 4. Participant flow for the valuation evaluation phase. Note: This flow diagram reflects the primary site data only. The completed response data ($n = 13$) are not shown because nonresponse/refusal counts are unavailable.

Table 3. Final TCQOLI Multiattribute Disutility Function (9-Domain Model): Level-Specific Disutilities and Patient-Importance Weights

Domain	*Level 1*	*Level 2*	*Level 3*	*Level 4*	*Level 5*	*Weight*
Recurrence	0.00	0.16	0.40	0.72	1.00	0.14
Appearance	0.00	0.16	0.42	0.73	1.00	0.06
Finance	0.00	0.17	0.43	0.75	1.00	0.08
Voice	0.00	0.17	0.40	0.72	1.00	0.10
Swallowing	0.00	0.19	0.45	0.75	1.00	0.11
Pain	0.00	0.19	0.41	0.71	1.00	0.13
Depression	0.00	0.13	0.38	0.69	1.00	0.12
Fatigue	0.00	0.14	0.37	0.68	1.00	0.12
Cognitive	0.00	0.13	0.35	0.69	1.00	0.14

Normalized single-domain disutilities on the MD–FH VAS scale and patient-importance weights (sum = 1).

Additive multiattribute disutility model: $\bar{u}(x) = \sum_{j=1}^{n} wj\bar{u}j(xj)$.

$\bar{u}(x)$ = multiattribute disutility value of an overall health state x (most-disabled-to-full-health scale here).

$u(x)$ = multiattribute utility value (utility = 1 − disutility, most-disabled-to-full-health scale).

$\bar{u}_j(x_j)$ = single attribute disutility value of attribute or domain j.

$\sum$ = summation symbol, indicating addition through all attributes/domains of health as below.

w_j = attribute or domain j's weight.

$j = 1$ refers to the first domain or attribute, or starting point of additive function.

n = number of domains or attributes, stopping point of additive function.

Example Calculation.

For example, TCQOLI survey responses are as follows: recurrence (level 2), appearance (level 2), finance (level 2), voice (level 1), swallowing (level 2), pain (1), depression (1), fatigue (level 3), and cognitive (level 1).

$\bar{u}(x) = (0.14 \times 0.16) + (0.06 \times 0.16) + (0.08 \times 0.17) + (0.1 \times 0) + (0.11 \times 0.19) + (0.13 \times 0) + (0.12 \times 0) + (0.12 \times 0.37) + (0.14 \times 0)$.

$\bar{u}(x) = 0.1109$, $u(x) = 1 - 0.1109$; thus, the utility value of health state $u(x) = 0.8891$ on a 0–1.0 scale.

MD–FH, most-disabled–full-health; SG, standard gamble; VAS, visual analog scale.

Data S13). The final user-facing, publicly available survey is downloadable from Supplementary Data S11. A detailed explanation of TCQOLI scoring for both the 9- and 10-domain specifications is provided in Supplementary Data S12 and S13).

Discussion

Main findings

This study delivers a new preference-based, thyroid-cancer-specific utility measure (TCQOLI) with a strong psychometric foundation and a patient-anchored scoring function. Using PROMIS item-bank parameters for included PROMIS items and reasonable psychometrics for the other items, multidisciplinary review, and patient rankings, we selected one high-performing item per domain. The reproduction-concern item showed the weakest psychometric performance and the lowest weight in the weighted index options; therefore, our primary specification is a 9-domain index. For potential use in settings where reproduction is salient (e.g., younger women planning pregnancy), we also report a 10-domain TCQOLI. We then

Table 4. Comparison of Standard VAS to Analyzed MAUT Scales Representing Agreement Between VAS and TCQOLI: 9-Domain Model

	VAS vs. additive index[d]	*VAS vs. additive index*[e]	*VAS vs. additive index*[f]	*VAS vs. multiplicative index*[f]	*VAS vs. unweighted scale*[g]	*VAS vs. unweighted scale*[h]
MAE[a]	0.108	0.069	0.045	0.319	0.096	0.083
OSD[b]	0.163	0.141	0.095	0.421	0.184	0.160
Mean Diff[c]	0.115	0.020	0.013	0.318	0.058	0.051
Pearson correlation (r)	0.752	0.752	0.743	0.727	0.734	0.734
Pearson CI, p value	0.654–0.825, <0.0001	0.654–0.825, <0.0001	0.641–0.819, <0.0001	0.620–0.807, <0.0001	0.630–0.812, <0.0001	0.630–0.812, <0.0001
ICC, CI[i]	0.744, 0.641–0.820	0.744, 0.641–0.820	0.734, 0.630–0.813	0.180, 0.087–0.450	0.683, 0.478–0.802	0.683, 0.478–0.802

Multiattribute utility theory (MAUT)-related indices.

[a]Mean absolute error (MAE): y_i = index or predicted value, x_i = direct VAS or other compared value. n = total number of data points.

[b]Overall standard deviation: OSD = $\sqrt{[\sum(x-\mu)^2/(n-1)]}$, x is each predicted value, μ is the overall 20% trimmed mean for the variable (group), n = total number of datapoints in the group.

[c]Mean difference: mean of individual paired differences between comparators.

Comparators are on the:

[d]Most-disabled-to-full-health scale.

[e]Dead-to-full-health scale.

[f]Most-disabled-to-full-health scale standard gamble, from derived power function: $VAS^{0.48}$ = standard gamble, all where 0 = most disabled and 1.0 = full health. Multiplicative MAUT Index is similarly transformed on the most-disabled-to-full-health scale.

[g]Non-MAUT 0–1 unweighted scale, with max–min normalization.

[h]Non-MAUT 0–1 unweighted scale, with max–min normalization then rescaled with 0.13 VAS value for the most-disabled state on dead-to-healthy scale.

[i]Intraclass correlation coefficient (ICC): 2-way mixed model, for single measures and absolute agreement.

CI, confidence interval; ICC, intraclass correlation coefficient; MAE, mean absolute error; OSD, overall standard deviation (SD of paired differences, index – VAS); SG, standard gamble; VAS, visual analog scale.

converted TCQOLI responses into utilities on a 0–1 scale elicited directly from patients. TCQOLI thus provides utilities suitable for cost-effectiveness and QALY-based analyses, moving the instrument from HRQoL measurement to decision-analytic applicability.

Psychometric properties

Missing data were modest with phone-assisted administration, and item response patterns supported selecting one high-performing item per domain. Items behaved as expected and were not redundant, showing good internal consistency; the reproduction-concern item was the exception, showing limited variability and the weakest associations. Item total correlations show how related each item is to the sum of all the other items' responses in the scale and also each item's discriminatory power for the construct of interest (can the item discriminate people with high versus low total scores). Such correlations are expected to be at least 0.3. Inter-item correlations assess each item's correlation to each other item. Items are expected to show correlations of at least 0.2–0.5, with greater than 0.5 suggesting redundancy. As is shown in Table 2, all items perform adequately with the exception of the reproduction concern item. This result likely reflects the cohort's age (mean $\sim$52 years), which matches the average age at thyroid cancer diagnosis in the United States, when reproduction is typically less important.[37] Preliminarily, in the younger subgroup, reproduction was ranked second least important. Clinically, fertility is usually preserved after thyroid cancer treatment, though pregnancy may be delayed, which may further attenuate concern.[38]

CFA supported a parsimonious two-factor structure (a general HRQoL factor plus a voice/swallow factor) with fit. The pain item loaded lower than the other indicators, likely because pain is less prominent for many thyroid cancer survivors.[5,39] We retained the pain domain to preserve coverage of generic HRQoL constructs relevant for QALY estimation, and pain may be more salient in aggressive subtypes.[40] In our use of scalar invariance measurement for known groups validity, a significant difference in swallowing and voice troubles was seen in the patients with lymph node metastasis versus those without. Composite reliability, or how consistently items within each factor measure the same concept, was acceptable.[41] Together, these findings indicate that the TCQOLI classification is psychometrically coherent and suitable for subsequent valuation and scoring.

At the instrument level, the TCQOLI showed a low ceiling effect of 3.8% and no floor effect, both well below the conventional 15% threshold for concern.[42] This ceiling is lower than ranges reported for widely used generic utility indices, including PROPr and EQ-5D-5L.[43,44] By contrast, item-level ceilings were higher, which is expected in survivorship cohorts with generally good health. These distributional properties support good interpretability and adequate headroom to detect differences and meaningful changes in thyroid-specific symptoms.

The result of strong correlation of the current health VAS with the summated items and, in turn, the additive model suggests content and construct validity of the TCQOLI health classification system. Rescaling, whether it is linear as was done with 0.13 or nonlinear, with the 0.48 power transformation of VAS to SG, appears to increase agreement.

Valuation interpretation

TCQOLI Pearson correlations were above 0.70, and the additive MAUT model achieved good–excellent absolute agreement (ICC $\sim$0.73–0.75 across the 9- and 10-domain models),[45] superior to multiplicative MAUT (low ICC). For context, head-to-head comparisons of established utility indices often show only fair agreement (ICCs $\sim$0.3–0.5), for example, SF-6D versus HUI3 (ICC = 0.41)[46] and PROPr versus EQ-5D-3L (ICC = 0.27).[47] Likewise, individual-level SG–HUI agreement has been reported around 0.45–0.57 in rheumatoid arthritis and was described as moderate to strong;[48] therefore, our ICCs in the $\sim$0.7 range represent comparatively strong agreement. Details of the multiplicative MAUT and unweighted models are shown in Supplementary Data S14-S15.

Clinical implications and relation to existing measures

Beyond generic measures, such as PROPr,[49] which maps PROMIS domains to a preference-based index, cancer-specific utility systems offer closer alignment with oncology constructs of interest in trials and ongoing care. Condition-specific approaches have emerged in other cancers, including the FACT-LUI for lung cancer[50] and the PORPUS-U for prostate cancer,[51] each deriving utilities from disease-targeted content. In thyroid cancer, the European cancer-specific EORTC QLU-C10D has demonstrated good validity, minimal ceiling effects, and higher statistical efficiency than EQ-5D-5L in most head-to-head comparisons.[52] These findings highlight how instrument choice shifts utility levels and sensitivity to change. Our TCQOLI complements this landscape by using thyroid-cancer-specific content prioritized by patients and clinicians, yielding utilities tailored to thyroid cancer decision contexts. Unlike generic preference-based measures and tumor-agnostic cancer utilities, the TCQOLI is designed to capture thyroid-specific burdens while remaining compact enough for utility valuation and routine use. For decision modeling, we recommend future sensitivity analyses that compare TCQOLI-based utilities with the QLU-C10D to quantify downstream effects on QALYs and cost-effectiveness conclusions.

Preference weighting clarifies which domains and severity levels mostly decrease patient's quality of life, supporting shared decision-making and informing policy and reimbursement where incremental cost per QALY is considered. Importantly, utilities depend on who values the health states. A cross-sectional, time-trade-off study of low-risk thyroid cancer states showed that general-population volunteers assigned consistently lower utilities than thyroid cancer survivors.[53] This indicates that the choice of source population can shift QALY estimates for thyroid cancer. Because values were derived from patients, the TCQOLI aligns with patient-centered care and may better reflect experienced burden than population-based tariffs in this context.

TCQOLI utilities can also be integrated with the thyroid cancer policy model (TCPM), a state-transition microsimulation that follows individual patients through thyroid cancer natural history, treatments, and survival.[54,55] TCQOLI weights can be assigned to TCPM health states (e.g., active surveillance; postlobectomy or total thyroidectomy with/without complications; adjuvant therapies; biochemical or structural

recurrence; and survivorship) to compute state-specific and lifetime QALYs. Using disease-specific TCQOLI utilities within the TCPM strengthens cost-utility analyses and scenario testing (e.g., surveillance intensity, operative strategies, systemic therapy choices), thereby improving the clinical and policy relevance of the model outputs.

Strengths and limitations

This multiphase, multicenter study combined rigorous, iterative instrument development with patient-centered valuation. Items were selected with patient input and cognitive interviews, then supported by CFA across sites, yielding a compact classification that captures thyroid-specific concerns without sacrificing usability (one item per domain as required for utility indexes). Utilities were elicited directly from patients and calibrated into a transparent, additive scoring function suitable for clinical research and cost-effectiveness applications. To balance the theoretical appeal of choice-based SG with respondent burden and feasibility, we paired SG with VAS; this approach is consistent with arguments that VAS can validly capture health-state preferences.[32,56] Finally, multicenter enrollment across disease types and care pathways supports generalizability, and we provide ready-to-use calculators for both the 9- and 10-domain specifications to facilitate transparent implementation.

Several limitations warrant consideration. Smaller sample sizes were seen due to pandemic-related delays, falling short of the conventional $\sim$200 participants often cited for CFA. Nevertheless, the sample was adequate for our planned analyses, and prior work shows smaller samples can yield stable solutions.[57–59] Our case-to-item ratio ($\sim$16:1 for 10 items) exceeds typical rules of thumb ($\approx$5–20:1),[60] supporting the suitability of the dataset for these psychometric procedures. In prior studies of this type, we have achieved higher rates of response. Previously, we were able to place clinical research coordinators in clinics to have a so-called "warm handoff" between the clinical team and the research team. This was not possible during the pandemic and even after the pandemic when virtual visits became commonplace. Second, valuation was cross-sectional; responsiveness and longitudinal validity of utilities were not assessed. Third, the English-language eligibility requirement likely contributed to the underrepresentation of Hispanic/Latinx participants (9–10% in our samples vs. $\sim$20% in the U.S. population[61,62]), thereby reducing the precision and representativeness of preference estimates for this group. Fourth, these types of questions cover complex medical issues. To minimize the cognitive burden, we provide the instrument to the person and conduct these interviews by phone rather than as self-administered paper or online surveys. We included basic measures of literacy and numeracy in these surveys, and our past research has shown that people with a range of literacy or numeracy challenges are less likely to participate in the research but can do so with assistance. Although comprehension checks and interview support were used, cognitive burden inherent to SG tasks could introduce noise for some respondents.[63] Thirty participants failed logical checks and were excluded from SG-based mapping analyses. These exclusions, while necessary for data quality, further reduced the SG sample size and may bias results toward more numerate respondents. Finally, the expression of utilities on a dead–full-health SG scale relied on an indirect, two-step approach (MDS anchoring and VAS $\rightarrow$ SG power mapping). While this strategy is pragmatic and benchmarked for plausibility, it introduces model dependence and additional uncertainty that warrants external validation.

Future Work

Future work should include external validation in broader, multilingual, and international cohorts to assess transferability and measurement invariance. Conducting this work in academic medical centers and comprehensive cancer centers may mean that recruited patients have more advanced illness, and validation will be needed to assess generalizability. In addition, head-to-head comparisons with the other instruments will allow direct assessment of downstream consequences of utility differences per instrument for surveillance strategies, surgical approaches, and related management decisions. We plan implementation studies that integrate TCQOLI utilities into clinical pathways and economic models. Embedding the index in registries or electronic health record workflows and linking it to decision models (e.g., the TCPM) will enable estimation of lifetime QALYs and support guideline-relevant choices on diagnostic strategies, treatment, and surveillance intensity. These pragmatic evaluations will test whether TCQOLI-based utilities change model outputs relative to generic instruments and whether they improve patient-centered decision-making.

Conclusion

The TCQOLI is a thyroid-cancer-specific, preference-based measure with strong psychometric properties and patient-anchored valuation. It shows initial evidence of validity and may be considered for health economic evaluations. It converts thyroid-cancer-specific outcomes into utilities that enable QALY estimation to support clearer resource allocation across clinical and policy settings.

Authors' Contributions

C.E.C.: Conceptualization (lead), writing—original draft (colead), reviewing, and editing. S.v.D.: Patient and data accrual (interviewing patients), writing—original draft (colead), reviewing, and editing. M.M.L.: Methodology and formal analysis. S.M.: Patient and data accrual (interviewing patients), data management and analysis, reviewing, and editing. T.H.: Patient and data accrual (interviewing patients), data management and analysis, reviewing, and editing. R.A.C.: Patient and data accrual (interviewing patients), data management and analysis, reviewing, and editing. C.D.: Patient and data accrual (interviewing patients), and data management and analysis. A.K.G.: Patient and data accrual (interviewing patients), and data management and analysis. J.M.V.: Conceptualization, reviewing, and editing. M.L.: Conceptualization, reviewing, and editing. M.J.: Methodology, formal analysis, reviewing, and editing. S.F.: Conceptualization, patient accrual (interviewing patients), data management (MDACC), reviewing, and editing. E.G.G.: Site lead (MDACC), conceptualization, patient accrual (interviewing patients), data management (MDACC), reviewing, and editing. K.D.: Conceptualization, patient accrual (training research fellow in interviewing patients), data management,

reviewing, and editing. J.S.S.: Conceptualization, writing—original draft (colead), methodology, formal analysis, reviewing, and editing. All authors reviewed the final version of the article and approved of the content.

Data Availability Statement

Data extracted from included studies and other supporting materials are available from the corresponding author on reasonable request.

Author Disclosure Statement

No competing financial interests exist.

Funding Information

C.E.C., M.M.L., S.M., T.H., R.A.C., C.D., A.K.G., and J.S.S.: Supported by the American Cancer Society RSG-20-024-01-CPHPS; J.M.V., M.L., M.J., and K.D.: no funding information to declare; and S.F. and E.G.G.: no conflicts of interest to disclose.

Supplementary Material

Supplementary Data

References

1. Boucai L, Zafereo M, Cabanillas ME. Thyroid cancer: A review. JAMA 2024;331(5):425–435; doi: 10.1001/jama.2023.26348
2. Siegel RL, Miller KD, Wagle NS, et al. Cancer statistics, 2023. CA Cancer J Clin 2023;73(1):17–48; doi: 10.3322/caac.21763
3. Surveillance Research Program NCI. Seer*Explorer: An interactive website for SEER cancer statistics. National Cancer Institute. Available from: https://seer.cancer.gov/statistics-network/explorer/ [Last accessed: September 25, 2025].
4. Roth EM, Lubitz CC, Swan JS, et al. Patient-reported quality-of-life outcome measures in the thyroid cancer population. Thyroid 2020;30(10):1414–1431; doi: 10.1089/thy.2020.0038
5. Applewhite MK, James BC, Kaplan SP, et al. Quality of life in thyroid cancer is similar to that of other cancers with worse survival. World J Surg 2016;40(3):551–561; doi: 10.1007/s00268-015-3300-5
6. McIntyre C, Jacques T, Palazzo F, et al. Quality of life in differentiated thyroid cancer. Int J Surg 2018;50:133–136; doi: 10.1016/j.ijsu.2017.12.014
7. Bakker C, van der Linden S. Health related utility measurement: An introduction. J Rheumatol 1995;22(6):1197–1199.
8. Sanders GD, Neumann PJ, Basu A, et al. Recommendations for conduct, methodological practices, and reporting of cost-effectiveness analyses: Second panel on cost-effectiveness in health and medicine. JAMA 2016;316(10):1093–1103; doi: 10.1001/jama.2016.12195
9. Herdman M, Gudex C, Lloyd A, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Qual Life Res 2011;20(10):1727–1736; doi: 10.1007/s11136-011-9903-x
10. Rabin R, de Charro F. EQ-SD: A measure of health status from the EuroQol group. Ann Med 2001;33(5):337–343; doi: 10.3109/07853890109002087
11. Brazier J, Usherwood T, Harper R, et al. Deriving a preference-based single index from the UK SF-36 health survey. J Clin Epidemiol 1998;51(11):1115–1128; doi: 10.1016/S0895-4356(98)00103-6
12. Feeny D, Furlong W, Boyle M, et al. Multi-attribute health status classification systems. Pharmacoeconomics 1995;7(6):490–502; doi: 10.2165/00019053-199507060-00004
13. Lubitz CC, De Gregorio L, Fingeret AL, et al. Measurement and variation in estimation of quality of life effects of patients undergoing treatment for papillary thyroid carcinoma. Thyroid 2017;27(2):197–206; doi: 10.1089/thy.2016.0260
14. Houten R, Fleeman N, Kotas E, et al. A systematic review of health state utility values for thyroid cancer. Qual Life Res 2021;30(3):675–702; doi: 10.1007/s11136-020-02676-2
15. Feeny D, Furlong W, Torrance GW, et al. Multiattribute and single-attribute utility functions for the health utilities index mark 3 system. Med Care 2002;40(2):113–128; doi: 10.1097/00005650-200202000-00006
16. Lamu AN, Gamst-Klaussen T, Olsen JA. Preference weighting of health state values: What difference does it make, and why? Value Health 2017;20(3):451–457; doi: 10.1016/j.jval.2016.10.002
17. Prieto L, Sacristán JA. What is the value of social values? The uselessness of assessing health-related quality of life through preference measures. BMC Med Res Methodol 2004;4:10; doi: 10.1186/1471-2288-4-10
18. Harris PA, Taylor R, Thielke R, et al. Research electronic data capture (REDCap)–A metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009;42(2):377–381; doi: 10.1016/j.jbi.2008.08.010
19. Bogdanovski AK, Sturgeon C, James BC. Financial toxicity in thyroid cancer survivors. Curr Opin Endocrinol Diabetes Obes 2023;30(5):238–243; doi: 10.1097/med.0000000000000826
20. Cella D, Riley W, Stone A, et al.; PROMIS Cooperative Group. The patient-reported outcomes measurement information system (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. J Clin Epidemiol 2010;63(11):1179–1194; doi: 10.1016/j.jclinepi.2010.04.011
21. Reeve BB, Hays RD, Bjorner JB, et al.; PROMIS Cooperative Group. Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the patient-reported outcomes measurement information system (PROMIS). Med Care 2007;45(5 Suppl 1):S22–S31; doi: 10.1097/01.mlr.0000250483.85507.04
22. Singer S, Al-Ibraheem A, Pinto M, et al. International phase IV field study for the reliability and validity of the European organisation for research and treatment of cancer thyroid cancer module EORTC QLQ-THY34. Thyroid 2023;33(9):1078–1089; doi: 10.1089/thy.2023.0221
23. Husson O, Haak HR, Mols F, et al. Development of a disease-specific health-related quality of life questionnaire (THYCA-QoL) for thyroid cancer survivors. Acta Oncol 2013;52(2):447–454; doi: 10.3109/0284186x.2012.718445
24. de Souza JA, Yap BJ, Wroblewski K, et al. Measuring financial toxicity as a clinically relevant patient-reported outcome: The validation of the COmprehensive score for financial toxicity (COST). Cancer 2017;123(3):476–484; doi: 10.1002/cncr.30369
25. Aschebrook-Kilfoy B, Ferguson BA, Angelos P, et al. Development of the ThyCAT: A clinically useful computerized adaptive test to assess quality of life in thyroid cancer

survivors. Surgery 2018;163(1):137–142; doi: 10.1016/j.surg.2017.09.009
26. Ferrell BR, Dow KH, Grant M. Measurement of the quality of life in cancer survivors. Qual Life Res 1995;4(6):523–531; doi: 10.1007/bf00634747
27. Ferrell BR, Dow KH, Leigh S, et al. Quality of life in long-term cancer survivors. Oncol Nurs Forum 1995;22(6):915–922.
28. Rosen CA, Lee AS, Osborne J, et al. Development and validation of the voice handicap index-10. Laryngoscope 2004;114(9):1549–1556; doi: 10.1097/00005537-200409000-00009
29. Brazier J, Czoski-Murray C, Roberts J, et al. Estimation of a preference-based index from a condition-specific measure: The king's health questionnaire. Med Decis Making 2008;28(1):113–126; doi: 10.1177/0272989x07301820
30. Littenberg B, Partilo S, Licata A, et al. Paper standard gamble: The reliability of a paper questionnaire to assess utility. Med Decis Making 2003;23(6):480–488; doi: 10.1177/0272989x03259817
31. Ross PL, Littenberg B, Fearn P, et al. Paper standard gamble: A paper-based measure of standard gamble utility for current health. Int J Technol Assess Health Care 2003;19(1):135–147; doi: 10.1017/s0266462303000138
32. Torrance GW, Feeny D, Furlong W. Visual analog scales: Do they have a role in the measurement of preferences for health states? Med Decis Making 2001;21(4):329–334; doi: 10.1177/0272989x0102100408
33. Keeny RL, Raiffa H. Decisions with Multiple Objectives: Preferences and value Tradeoffs. Cambridge University Press; 1993.
34. Tomlinson G, Bremner KE, Ritvo P, et al. Development and validation of a utility weighting function for the patient-oriented prostate utility scale (PORPUS). Med Decis Making 2012;32(1):11–30; doi: 10.1177/0272989X11407203
35. Nguyen TH, Han HR, Kim MT, et al. An introduction to item response theory for patient-reported outcome measurement. Patient 2014;7(1):23–35; doi: 10.1007/s40271-013-0041-0
36. Stover AM, McLeod LD, Langer MM, et al. State of the psychometric methods: Patient-reported outcome measure development and refinement using item response theory. J Patient Rep Outcomes 2019;3(1):50; doi: 10.1186/s41687-019-0130-5
37. Society AC. Key statistics for thyroid cancer. 2025. Available from: https://www.cancer.org/cancer/types/thyroid-cancer/about/key-statistics.html [Last accessed: October 10, 2025].
38. Hirsch D, Yackobovitch-Gavan M, Lazar L. Infertility and pregnancy rates in female thyroid cancer survivors: A retrospective cohort study using health care administrative data from Israel. Thyroid 2023;33(4):456–463; doi: 10.1089/thy.2022.0501
39. Goswami S, Mongelli M, Peipert BJ, et al. Benchmarking health-related quality of life in thyroid cancer versus other cancers and United States normative data. Surgery 2018;164(5):986–992; doi: 10.1016/j.surg.2018.06.042
40. Pavlidis ET, Galanis IN, Pavlidis TE. Update on current diagnosis and management of anaplastic thyroid carcinoma. World J Clin Oncol 2023;14(12):570–583; doi: 10.5306/wjco.v14.i12.570
41. Field-testing: Item reduction and data structure. In: Measurement in Medicine: A Practical Guide. (de Vet HCW, Terwee CB, Mokkink LB, Knol DL., eds.) Cambridge University Press; 2011; pp. 65–95. Practical Guides to Biostatistics and Epidemiology.
42. McHorney CA, Tarlov AR. Individual-patient monitoring in clinical practice: Are available health status surveys adequate? Qual Life Res 1995;4(4):293–307; doi: 10.1007/bf01593882
43. Emrani Z, Akbari Sari A, Zeraati H, et al. Health-related quality of life measured using the EQ-5D-5 L: Population norms for the capital of Iran. Health Qual Life Outcomes 2020;18(1):108; doi: 10.1186/s12955-020-01365-5
44. Klapproth CP, Sidey-Gibbons CJ, Valderas JM, et al. Comparison of the PROMIS preference score (PROPr) and EQ-5D-5L index value in general population samples in the United Kingdom, France, and Germany. Value Health 2022;25(5):824–834; doi: 10.1016/j.jval.2021.10.012
45. Cicchetti DV. Multiple comparison methods: Establishing guidelines for their valid application in neuropsychological research. J Clin Exp Neuropsychol 1994;16(1):155–161; doi: 10.1080/01688639408402625
46. Abel H, Kephart G, Packer T, et al. Discordance in utility measurement in persons with neurological conditions: A comparison of the SF-6D and the HUI3. Value Health 2017;20(8):1157–1165; doi: 10.1016/j.jval.2017.04.008
47. Klapproth CP, Fischer F, Rose M. Scale agreement, ceiling and floor effects, construct validity, and relative efficiency of the PROPr and EQ-5D-3L in low back pain patients. Health Qual Life Outcomes 2023;21(1):107; doi: 10.1186/s12955-023-02188-w
48. Rashidi AA, Anis AH, Marra CA. Do visual analogue scale (VAS) derived standard gamble (SG) utilities agree with health utilities index utilities? A comparison of patient and community preferences for health status in rheumatoid arthritis patients. Health Qual Life Outcomes 2006;4:25; doi: 10.1186/1477-7525-4-25
49. Dewitt B, Feeny D, Fischhoff B, et al. Estimation of a preference-based summary score for the patient-reported outcomes measurement information system: The PROMIS(®)-preference (PROPr) scoring system. Med Decis Making 2018;38(6):683–698; doi: 10.1177/0272989x18776637
50. Swan JS, Lennes IT, Stump NN, et al. A patient-centered utility index for non-small cell lung cancer in the United States. MDM Policy Pract 2018;3(2):2381468318801565; doi: 10.1177/2381468318801565
51. Bremner KE, Mitsakakis N, Wilson L, et al. Predicting utility scores for prostate cancer: Mapping the prostate cancer index to the patient-oriented prostate utility scale (PORPUS). Prostate Cancer Prostatic Dis 2014;17(1):47–56; doi: 10.1038/pcan.2013.44
52. Pilz MJ, Seyringer S, Singer S, et al.; EORTC Quality of Life Group. The cancer-specific health economic measure QLU-C10D is valid and responsive for assessing health utility in patients with thyroid cancer. Thyroid 2024;34(11):1356–1370; doi: 10.1089/thy.2024.0396
53. Carlisle K, Kowalski R, Park AN, et al. Differences in health state valuation for small, low-risk thyroid cancer between general population and cancer survivors: A cross-sectional analysis. Qual Life Res 2025;34(10):2891–2900; doi: 10.1007/s11136-025-04033-7
54. Lubitz C, Ali A, Zhan T, et al. The thyroid cancer policy model: A mathematical simulation model of papillary thyroid carcinoma in the U.S. population. PLoS One 2017;12(5):e0177068; doi: 10.1371/journal.pone.0177068
55. White C, Weinstein MC, Fingeret AL, et al. Is less more? A microsimulation model comparing cost-effectiveness of the revised American thyroid association's 2015 to 2009 guidelines for the management of patients with thyroid nodules

and differentiated thyroid cancer. Ann Surg 2020;271(4): 765–773; doi: 10.1097/sla.0000000000003074
56. Parkin D, Devlin N. Is there a case for using visual analogue scale valuations in cost-utility analysis? Health Econ 2006; 15(7):653–664; doi: 10.1002/hec.1086
57. Barrett P. Structural equation modelling: Adjudging model fit. Pers Individ Dif 2007;42(5):815–824; doi: 10.1016/j .paid.2006.09.018
58. Gagne P, Hancock GR. Measurement model quality, sample size, and solution propriety in confirmatory factor models. Multivariate Behav Res 2006;41(1):65–83; doi: 10.1207/ s15327906mbr4101_5
59. Marsh HW, Hau KT, Balla JR, et al. Is more ever too much? The number of indicators per factor in confirmatory factor analysis. Multivariate Behav Res 1998;33(2):181–220; doi: 10.1207/s15327906mbr3302_1
60. Kline RB. Principles and Practice of Structural Equation Modeling. Guilford Publications; 2023.
61. United Nations. Revision of world population prospects. Available from: https://population.un.org/wpp/ [Last accessed: September 25, 2025].
62. U.S. Census Bureau. Data from: American community survey (ACS) 1-year estimates, table S0201: Selected population profile in the United States. 2024. Available from: https://data.census.gov/table/ACSSPP1Y2024.S0201
63. Lobo E, Nanda L, Akhouri SS, et al. Describing the development of a health state valuation protocol to obtain community-derived disability weights. Front Public Health 2019;7:276; doi: 10.3389/fpubh.2019.00276

Address correspondence to:
Carrie E. Cunningham, MD, MPH
Department of Surgery
Massachusetts General Brigham
Harvard Medical School
15 Fruit Street
Boston, MA 02114
USA

E-mail: ccunningham@mgh.harvard.edu