Measuring Health-Related Quality of Life in Thyroid Eye Disease

Anu Sharma,1 Marius N. Stan,2 and Dan B. Rootman3,4

1Division of Endocrinology, Metabolism and Diabetes, University of Utah School of Medicine, Salt Lake City, UT 84112, USA
2Division of Endocrinology, Mayo Clinic, Rochester, MN 55905, USA and
3Stein and Doheny Eye Institutes, University of California, Los Angeles 90095, USA

Correspondence: Daniel B. Rootman, MD, MS, Division of Orbital and Ophthalmic Plastic Surgery, Stein and Doheny Eye Institutes, University of California, Los Angeles, 100 Stein Plaza #2-267, Los Angeles, CA 90095, USA. Email: Rootman@sei.ucla.edu.

Abstract

Health-related quality of life (HQOL) is a concept that aims to understand the totality of an individual’s experience of their disease state. This can include the physical, psychosocial, emotional, and psychological effects of a disease state. A complex and multifactorial concept, HQOL can be challenging to measure accurately and reliably. Thyroid eye disease (TED), as a multifaceted physically debilitating and facial disfiguring disorder, presents unique challenges and opportunities in the measurement of HQOL. Multiple distinct tools have been developed for this purpose, each has been constructed, assessed, and utilized. This discussion surveys the landscape of TED-related QOL measurement and presents challenges for the future. Clinicians and clinical researchers should implement TED-related QOL measurement as part of routine TED care and as a primary outcome in TED clinical trials. We recommend utilizing the Graves’ ophthalmopathy (GO)-QOL routine in clinical practice and as a primary outcome in TED clinical trials. If the GO-QOL is too time-consuming or mild TED, a faster alternative is the TED-QOL.

Key Words: thyroid eye disease; quality of life; measurement; patient reported outcomes

Abbreviations: CAS, clinical activity score; MCID, minimal clinically important difference; GD, Graves’ disease; GO, Graves’ ophthalmopathy; HQOL, health-related quality of life; QLS, quality of life scale; QOL, quality of life; SF, Short Form; STED-QoL, Singapore Thyroid Eye Disease Quality of Life; TED, thyroid eye disease.

What is Quality of Life in Thyroid Eye Disease

Quality of life (QOL) is an increasing focus of healthcare delivery. The intention of focusing on QOL is to better understand the effect of the disease and associated therapeutic interventions directly on the patient experience. The patient experience, however, can be notoriously difficult to quantify.

The World Health Organization through the World Health Organization Quality of Life assessment, defined QOL as an “individuals’ perception of their position in life in the context of the culture and value systems in which they live and in relation to their goals, expectations, standards and concerns.” It is a broad definition describing QOL in terms of “physical health, psychological state, level of independence, social relationships and their relationship to salient features of their environment” (1). This definition emphasized the importance of cultural context in developing QOL questionnaires and furthered the paradigm shift from physician assessed and resource focused parameters, to in-depth and comprehensive patient-centered questionnaires.

All-encompassing general health QOL assessment tools can be useful in their applicability to multiple disease states and disparate populations. These tools assess core aspects of health-related QOL (HQOL) that are important to a wide array of health states. They generally have strong methodological underpinnings and undergo extensive psychometric testing in large populations. They have the distinct advantage of allowing comparison across disease states; however, this comes at the cost of being too nonspecific to characterize granular changes in any disease state. At the other end of the spectrum are disease-specific QOL tools, that focus on 1 disease state, and, though accurate for that particular disease, may suffer from challenges in comparing different populations. Of course, this is a spectrum and many intermediate or “supercategories” of health states can be defined, balancing generalizability and specificity. For instance in thyroid eye disease (TED) scales can range from general QOL surveys such as Short Form (SF)-36 (2), through vision-related QOL such as the National Eye Institute Visual Function Questionnaire (NEI-VFQ) (3) to more specific QOL measures such as the TED-QOL (4) and GO-QOL (5).

Though each type of QOL questionnaire has been utilized in TED, the utility of a disease-specific QOL measure for TED appears self-evident to the practicing clinician. TED is a heterogeneous disease that has a multitude of variables that can negatively impact daily activities and patients’ perception of self and their surrounding environment. This dysfunction is heterogeneous as well, involving both physical disability and psychosocial challenges. Some of this heterogeneity is captured in the parameters typically used to describe TED (proptosis, diplopia, lid retraction, inflammation, corneal health, optic nerve status, and so on); however, these parameters fail to reflect the patients’ lived experience with the disease. TED encompasses a range of effects impacting psychological functioning, physical abilities, and appearance.
The longstanding challenges in measuring TED severity are compounded by the myriad classification systems (6-8) and varying primary outcome measures (9-12). Unfortunately, the disability patients experience in TED continues to be insufficiently captured. The journey to clearly understand, characterize, and measure QOL in TED has involved development, testing, and utilization of many general, partially specific, and specific TED QOL measures. Each provides important contributions to the assessment and understanding of TED-related experiences. These instruments, the knowledge they have led to, and a pathway toward a more ideal instrument are discussed through the remainder of this section.

What Are the Best Instruments to Measure QOL?

Over time, there have been 2 approaches to measurement of QOL in TED: questionnaires and semi-structured interviews. Questionnaires are standardized, have a prespecified set of answers, and can be administered without the active participation of the physician. Structured interviews are more open constructs that allow for an individualized understanding of a particular TED experience, at the expense of more time, dedicated clinician involvement, and variable generalizability.

TED Structured Interviews

A primary approach to assessing HQOL in patients with TED relates to the use of structured interviews. This approach is certainly more time-consuming but can highlight impact areas that the HQOL questionnaires might not easily point out. One such investigation reported by Estocurt et al (13), highlighted important aspects of patients with TED’s adaptive response. These responses are representing defense mechanisms, and include social avoidance, frustration associated with medical encounters that seemed to offer minimal benefit, and the change in “identity” that reflected the changed appearance and functional capacity. These concepts represent important aspects of the experience, and they should be “hypothesis-generating” as they suggest areas that should be explored in more detail when trying to formulate a therapeutic plan and assessment thereof.

Thus, TED structured interviews likely represent the richest source of data regarding patients’ TED-related QOL; however, they are time-consuming and impractical in standard clinical practice. Further, such interviews are not easily comparable across sites of care, limiting their utility in the development of clinical protocols and the dissemination of research on effective therapy.

Questionnaires

The most commonly utilized general HQOL form is the Medical Outcome Short Form, commonly abbreviated to SF-36 or MOS-36 (2). The only other general health HQOL tool used in patients with TED is the sickness impact profile (14), which had very limited utilization (15). These instruments are useful for comparing the impact of TED on HQOL in relation to other diseases; however, they are less effective at describing the impact of TED and of the changes induced by its therapy.

SF-36

This tool comprises 36 questions that explore 8 areas of HQOL: physical functioning, bodily pain, role limitations due to physical health problems, role limitations due to personal or emotional problems, emotional wellbeing, social functioning, energy/fatigue, and general health perceptions (2). In 1 study in elderly people (age 65 and older) it took a median of 8 minutes to complete this questionnaire (16). Therefore, for time efficiency, 2 shorter versions (SF-12 and SF-24) have been created and used in clinical trials, including TED trials (see Table 1). In the area of thyroid autoimmunity, it has been used initially for assessing the impact of therapies for Graves’ disease (GD) (17). It has also been used for the assessment of patients with GD who developed TED (18). Patients with TED had lower HQOL than patients with GD without TED. Interestingly, patients with TED had similar SF-36 scores to patients who had inflammatory bowel disease, and lower scores than some other serious comorbidities (emphysema, congestive heart failure, and diabetes mellitus) (15). This underscores the depth of dysfunction experienced by patients with TED, which may be out of proportion to the physical measurable characteristics such as proptosis and diplopia in TED.

| Table 1. Randomized clinical trials for TED therapy utilizing GO-QOL and other QOL outcomes (alphabetical listing) |
|--------------------------------------------------|------------------|----------------|-----------------|-------------------------------|----------------|
| Intervention                                      | Comparator        | Investigators   | HQOL instrument | Clinical outcome              | HQOL outcome     |
| IV steroids high-dose                             | IV steroids moderate and low-dose | Bartalena et al | GO-QOL           | Improved CAS and composite outcome | Improved similarly in all groups |
| Mycophenolate + IV steroids                       | IV steroids      | Kahaly et al   | GO-QOL           | No difference at 12 weeks      | Improved similarly in all groups |
| Octreotide                                       | Placebo           | Wemeau et al | SF-36            | No clinical differences        | Not reported     |
| Rituximab                                        | IV steroids      | Salvi et al    | GO-QOL           | Improved CAS                   | Not reported     |
| Selenium                                         | Pentoxifylline Placebo | Marcocci et al | GO-QOL           | Improved                     | Improved        |
| Teprotumumab                                     | Placebo           | Smith et al   | GO-QOL           | Improved proptosis and CAS     | Improved        |
| Teprotumumab                                     | Placebo           | Douglas et al | GO-QOL           | Improved proptosis and CAS     | Improved        |
| Tocilizumab                                      | Placebo           | Perez-Moreiras et al (ref) | SF-36 GO-QOL | Improved proptosis and CAS     | No difference Improved |

Abbreviations: CAS, clinical activity score; GO, Graves’ ophthalmopathy; HQOL, health-related quality of life; IV, intravenous; QOL, quality of life; TED, thyroid eye disease.
Other studies have specifically investigated the association between indicators of TED severity and the SF-36. These investigations have found a correlation between TED severity and SF-36 scoring. This was clearly demonstrable for the physical component of the tool, which measured the lowest scores of the 8 subscales in cases of more severe diplopia and/or sight-threatening disease (19). Interestingly, patients with more severe proptosis had lower emotional scores than those with significant functional impairment from diplopia (20), consistent with the different impact that various TED features are expected to cause on different SF-36 domains. The variability in these results may be related to the poor correlation between the clinical elements of TED (like disease duration and severity) and the whole-body, disease agnostic HQOL.

One of the first uses of SF-36 in a TED clinical trial compared the response to intravenous vs oral steroids. This trial reported higher SF-36 improvement rates after intravenous steroids (80% reported good or excellent health) compared with oral steroids group (54% with good or excellent health). Another study where intravenous steroids were compared with external radiation did not find a similar difference (22).

Not surprisingly the utility of the SF-12/SF-36 instrument in the assessment of specific disease characteristics and the effect of treatment is limited. However, the use of this instrument establishes a comparative baseline for understanding the TED experience in relation to other well-known disease states such as inflammatory bowel disease and emphysema.

### Psychological Health Questionnaires

Given the mental burden of TED, tools used for assessing psychological health have also been applied to these patents. These again are general, and generalizable to other disease states; however, are poorly associated with clinical indicators of physical dysfunction or disease severity. They include the Life Experience Survey (exploring critical life events), Hospital Anxiety and Depression Scale, Profile of Mood States, Beck Depression Inventory (these 3 exploring affective disorders), and Ways-of-Coping Checklist, which targets coping mechanisms (19, 20, 23).

In a Korean study, 32.7% of patients were found to have Beck Depression Inventory scores indicating clinical depression. The authors did not find an association between disease severity and likelihood of depression (19). An American study investigating psychological stress revealed that more severe TED was associated with greater psychological stress, although the authors did not report whether the overall mean deviated from the general population (20). In a German population assessed with the Hospital Anxiety and Depression Scale, Kahaly et al found that depression and anxiety were much higher in patients with TED at 40.2% and 22% than in the general population at 5% and 8% respectively (21). These studies demonstrate a high prevalence of psychological dysfunction in patients with TED across multiple cultural contexts. These psychological indicators are as expected inconsistently associated with physical disease characteristics implying that patients may be significantly affected by signs and symptoms that physicians may classify as mild.

### Vision-specific Questionnaires

As an ophthalmic disease, a natural extension to a more specific HQOL instrument would involve utilization of previously validated vision-specific questionnaires. The NEI-VFQ, a questionnaire directed primarily to ascertain the impact of conditions such as cataract, macular degeneration, glaucoma, and other nondisfiguring diseases and developed by the National Eye Institute, has been utilized in TED populations (3). This 25-question survey covers 12 subscales, and the results are reported in a range from 0 to 100, with higher scores indicating better HQOL.

Bradley et al (24) assessed VFQ performance in a TED population, finding that some subscales did correlate with TED activity and severity. Unfortunately, significant aspects of the disease, particularly those related to appearance and impaired functioning in social settings, were poorly captured by this instrument. A Chinese translation demonstrated similar findings in a series dominated by patients with TED with dysthyroid optic neuropathy (23). Again, visual function represents only a small portion of overall disability and as such VFQ scores are, as expected, limited in the overall assessment of QOL in TED.

### TED-specific QOL Measures

As we have seen, general QOL instruments have proven to be ineffective in measuring TED-specific disease severity to a level that would be adequately granular for clinical assessment or scientific research. Therefore, the creation of a TED-specific QOL instrument became a necessity for modern medicine.

Creation of an ideal TED-specific QOL measure has proven to be a challenging task. This is not surprising given the heterogeneity of the disease in terms of physical profile and the wide ranging social and physiological effects of those physical changes. Even finding agreement on the salient physical characteristics has been challenging. A group of physicians treating patients with TED and a group of patients found common ground regarding the importance of less than 50% of various TED signs and symptoms considered (26).

### Graves’ Ophthalmopathy Quality of Life

The first disease-specific TED-related QOL questionnaire was the Graves’ ophthalmopathy (GO)-QOL, which was published in 1998 (5). The authors had previously investigated the use of general HQOL questionnaires but found them lacking in specificity to detect clinical changes in TED (15). They utilized existing vision-related questionnaires (27, 28) and performed a small survey of patients affected by TED to generate the 16 items included in the GO-QOL instrument. Questions were ranked on a 3-point Likert scale and split into 2 categories: effects of TED on visual functioning and the psychosocial effects of the physical appearance of TED (5). Patients with thyroid function abnormalities were excluded. The GO-QOL was compared with the Medical Outcomes Study Short-form General Health Survey (MOS-24) (29) and 3 subscales of the Sickness Impact Profile (14). Of note, the last 2 surveys were used in their initial study on HQOL that led to the development of GO-QOL (15). Cronbach’s alpha indicated good reliability for both the visual functioning and physical appearance categories in GO-QOL. The GO-QOL correlated well with the sickness impact profile, more so than MOS-24; however exact numbers were not reported. This suggested good construct validity. Of the 112 patients initially recruited, 70 returned the survey and 46 answered...
all questions. All participants had an assessment of TED status by the NOSPECS (8) score and the CAS (clinical activity score) (7). While the visual functioning scale correlated with these clinical TED scores, physical appearance scores did not (5). The authors attributed this lack of correlation of the GO-QOL with objective clinical scores to the varying perceptions of disease activity by patients.

The GO-QOL was developed in the Netherlands, with questions geared toward the Dutch population. It has now been translated into 19 languages (30) (11 are available for download at https://www.eugogo.eu/en/what-do-we-offer/downloads/) and studied in many different patient populations. Given that not all patients engage in biking or driving, the questionnaire has been adapted to different cultures with good construct validity in many countries (31-36). Interestingly, despite the short time to complete the GO-QOL (3.1 minutes), in 1 study there was a low completion rate (78%) compared with other TED-specific questionnaires (100% for TED-QOL and 94% for GO-quality of life scale (QLS (4)).

Both criterion validity and longitudinal validity have been documented extensively for GO-QOL. In mild to severe TED, multiple therapeutic studies across countries have shown the ability of both visual function and appearance subscale scores for GO-QOL to show improvement with therapeutic success (10, 37, 38). In a well-conducted randomized control trial investigating the effect of selenium on mild GO (10), improvement in GO-QOL scores reflected improving clinical “overall eye scores” as defined in that investigation (39). GO-QOL has also reflected improvement in clinical eye scores (as measured by the CAS) in both moderate and severe TED after treatment with steroids and orbital decompression surgeries, respectively (37).

The group responsible for the development of the GO-QOL performed a prospective cohort study on 164 patients with TED scheduled for TED treatment (radiation, decompression, eye muscle surgery, eye lid lengthening, or blepharoplasty) and established a minimal clinically important difference (MCID) (38). The MCID was noted to be 6 points for less invasive therapy (those provided for mild TED) and 10 points for more invasive therapy (those provided for moderate to severe TED). This has provided a much needed clinically useful target for clinicians and researchers to understand and assess the patient’s perception of change over time, though no subsequent validation of these values has been reported to our knowledge.

An MCID of 6 or more has therefore been used in larger trials as secondary endpoints. In the teprotumumab for the treatment of active thyroid eye disease study (40), participants in the teprotumumab arm reported a mean change in GO-QOL of 13.79 vs 4.43 points in the placebo arm. Given that the criterion and longitudinal validities of the GO-QOL have been established, the interpretation of a change of 6 or more, and in this case more than 10, reflects a clinically significant quality of life improvement for invasive therapy in moderate to severe TED. Reported GO-QOL improvements with other medical therapies for TED have been statistically significant but not as robust. Intravenous steroids have shown improvements from 6 to 12.8 points depending on the dose of steroid used (41, 42). There was a 6 or greater point improvement in GO-QOL in 58% of patients receiving rituximab at 72 weeks whereas 77% report improvements at 52 weeks in visual functioning (11); 47% of patients reported ≥8 points improvement in GO-QOL with tocilizumab compared with 33% with placebo (43).

Over the years, the impact of various therapies in GO-QOL has been tested in several studies (see Table 1). Interestingly, clinical outcomes and GO-QOL outcomes are not consistently associated. For instance, in the comparison trial of different intravenous steroid regimens (42), improved clinical benefit was noted at the highest dose (7.5 g); however, this was also associated with a greater number of side-effects and thus no benefit in GO-QOL could be identified (though all groups had an overall improvement). In contrast, the teprotumumab trials (12, 40) and the selenium trial (10) did identify GO-QOL benefit toward the active intervention arms compared with placebo (as well as pentoxifylline in the selenium trial).

The impact of surgery on GO-QOL has been studied in 1 large EUGOGO study with 140 patients and multiple surgical approaches (44). The results revealed a benefit in the appearance score of around 20 points; however, the functionality score decreased in many cases and the surgical approaches associated with higher incidence of diplopia were most likely to be associated with a decline. Of interest, a study limiting the surgical approach to the lateral orbital wall and orbital fat resection demonstrated a positive impact on both appearance and functional scores which almost doubled in value (9).

It is no surprise that the biggest improvement in GO-QOL was reported in the functionality scale on 2 patients treated for dysthyroid optic neuropathy. This may be related to the lower incidence of diplopia in lateral wall decompression, although not specifically tested. These studies again point to the balance of multiple dimensions in the assessment of TED-related QOL, each of which can be affected in positive or negative ways by a single intervention.

Graves’ Ophthalmopathy Quality of Life Scale
While GO-QOL had been extensively utilized in Europe, its validity in the United States was poorly established (45, 46). In addition to this, the poor correlation with clinical activity of the disease was listed by others as part of the impetus for developing more precise instruments (47). Yeatts (47) used existing general QOL instruments including the SF-12 (48), Dermatology-Specific Quality of Life questionnaire (49), NEI-VFQ (3), and visual function questions specific to TED to create a 105-item form. This questionnaire was mailed to 325 patients with TED who had previously visited the oculoplastic clinic at Wake Forest University in North Carolina. There was a 62% response rate. Additionally, the questionnaire was administered to 53 consecutive new patients who presented to the clinic with TED. Thirty-three people with no history of thyroid disorders also completed the questionnaire and served as controls. Using scaling and other statistical techniques, the 105 questions were paired down to 9 that had the highest discriminative ability for TED (47).

The GO-QLS demonstrated good correlation with disease severity and was better correlated with TED severity than the more general NEI-VFQ (1 of the scales from which it was derived). The GO-QLS also demonstrated good construct and content validity. The GO-QLS was noted to correlate with GO-QOL and TED-QOL surveys with good internal consistency; however, 19% of patients felt the GO-QLS was confusing on the first attempt to respond and 12% on the
second attempt (4). More data are required to determine its reliability, criterion validity, generalizability to distinct populations (4), and longitudinal validity of this instrument.

Importantly, content validity of GO-QOLs was based on item development performed in more general populations of visually and otherwise impaired patients (SF-12, Dermatology-Specific Quality of Life questionnaire, NEI-VFQ), and again did not involve patients affected by TED in the process of item generation. Further, testing for comprehension and ease of use was an important aspect of this research that had been previously unreported in the GO-QOL scale evaluations. This proved to be problematic in retrospect, given 1 out of 5 of patients found it confusing. The GO-QLS has not gained wide popularity. It has been rarely tested in clinical trials with limited data to support its clinical utility.

**Thyroid Eye Disease Quality of Life**

The TED-QOL lands at the opposite end of the understandability spectrum consisting of only 3 questions (4) assessed on a visual analog scale, borrowed from the pain assessment literature. The questions focus on overall QOL, ability to perform daily activities, and physical appearance. The advantage of this approach is ease of use and cross-cultural comprehension; however, it does sacrifice depth and diversity in the domains assessed. The TED-QOL was assessed on 100 consecutive patients who were referred to an ophthalmology clinic at the University of British Columbia, each completing the scale twice (4). The GO-QOL and GO-QLS were also administered to the same patients simultaneously to determine construct validity and internal consistency. Disease severity was assessed using the vision, inflammation, strabismus appearance (VISA) classification (6). This study had 100% completion rate for all 3 questionnaires at baseline and 58% on follow-up. TED-QOL correlated well with both GO-QOL and GO-QLS on both visual function scores and physical appearance subscales (Pearson correlations 0.62-0.75). All 3 questionnaires also showed good test–retest reliability. The GO-QOL and GO-QLS had better correlations between visual functioning questions and all 4 components of the VISA classification whereas the TED-QOL correlated with only 2 components of the VISA classification (inflammation and diplopia). With a completion time of 1.6 minutes, the TED-QOL is practical and patient friendly (4).

The TED-QOL has been translated into Korean (50) and Chinese (51). In the Korean population (50), TED-QOL was administered to 90 patients with GD and TED, 14% of whom had active TED. Face validity was rated as good by patients in a post hoc manner. In an almost duplicate of the Korean study, the TED-QOL (referred to as TAO-QOL in this study) was also investigated in 182 Chinese patients with GD and TED, 15.3% with active TED (51). Both studies noted good content, convergent and discriminant validities of the TED-QOL (50, 51). There were moderate correlations with clinical disease activity and physical appearance as measured by the CAS, NOSPECS, Gorman diplopia score, and VISA grading.

Change in TED-QOL score before and after treatment was studied by the original developers of this instrument (52). Fifty-six patients with TED requiring surgery completed the TED-QOL and GO-QOL before and after rehabilitative surgery (orbital decompression, strabismus repair, and lid procedures). Both TED-QOL and GO-QOL visual functioning scores differed significantly with surgery type, with a larger difference noted in the TED-QOL. There was no difference in appearance scores for either questionnaire after surgery. The authors concluded there was greater sensitivity with the TED-QOL compared to the GO-QOL. The TED-QOL was deemed more responsive across all surgery types whereas the GO-QOL only showed visual functioning responsibility with strabismus surgery. The GO-QOL did have significant responsiveness for all types of surgery on the physical appearance scale. The TED-QOL therefore is more likely to detect a clinical change in TED within a subject compared with the GO-QOL.

These findings emphasize the tradeoffs in tool implementation, as the TED-QOL performs well across populations in terms of ease of understanding and use. It however lacks depth and breadth in measurement of disease-related QOL, correlating inconsistently with physical measures of the disease.

**Singapore Thyroid Eye Disease Quality of Life**

The Singapore Thyroid Eye Disease Quality of Life (STED-QoL) is the first questionnaire developed focusing on the culture, epidemiology and clinical characteristics of an Asian population (53). Item development was performed utilizing qualitative methodologies, unlike instruments developed earlier. Focus groups involving Singaporean patients with TED of Chinese descent were conducted until idea saturation was achieved. Thematic analysis was utilized to develop a 12-item questionnaire, which was administered to 59 patients with mild, moderate, and severe TED to assess criterion validity. After noting a suboptimal fit to the Rasch model (particularly precision and disordered thresholds), the 12-item scale was paired down to the final 10 item STED-QoL (53). While the STED-QoL had a 100% response rate with strong internal consistency, construct validity, and criterion validity, there are no data on its replicability in other populations, even within Singapore.

The development of this instrument was focused on patient experience and demonstrates excellent face validity for this reason. Additionally, the mixed-model approach produced robust psychometric results. It has however not been widely tested on populations over time, or after treatment, to establish longitudinal or discriminant validity. The development in a small, Chinese-speaking population also presents some challenges for generalizability to English-speaking patients. At this point the STED-QoL is the most recently developed of the TED-specific HQOLs and the least tested in practice. It is noteworthy however, that there is significant overlap between this scale and the others (Table 2).

**Opportunities and Challenges**

The current quandary regarding appropriate use of TED-specific QOL instruments relates to each representing a significant development in the assessment of TED. However, each has important shortfalls. The dire need for improvements presents opportunity for the TED community. Each of the scales have been poorly associated with the physical markers of disease severity as outlined previously. It is further known that TED decreases QOL to levels similar to or worse than those of classically disabling diseases such as inflammatory bowel disease and emphysema (15), which may be out of proportion to the severity of TED-related physical changes defined by clinicians (54). Further, changes in appearance and improved function with surgery have proven to be inconsistently associated with QOL improvement (44, 55).
Additionally, the levels of psychiatric disability in patients affected by TED are much higher than the general population (19, 20, 23). Unfortunately, psychological functioning is not specifically addressed by any measure of TED-related QOL. Finally, disability in terms of economic and social productivity, has been examined in a European population (56); however, it remains poorly understood in TED. Although peripherally assessed in some of the current measures, the reliability of these data compared with other indices of economic and social productivity are not established. Taken together, measurable physical manifestations of TED and their effects on patient experience may be poorly understood.

Two possible explanations for this incongruence between physical manifestations of TED and apparent negative impact on QOL can be proposed: either current measures of physical change insufficiently describe the effects of the disease or the QOL measures currently employed poorly measure physical disease domains, or both. An effective TED-specific QOL measure should be capable of unwrapping this discrepancy. At least in a partial manner, the lack of consensus on the validity and reliability of a single TED QOL measure may push clinicians and scientists toward the use of the more “objectively reliable” physical characteristic measurement as a default. Of course, these measures are more comfortably quantified by primarily clinical observers. The ease of quantifying physical measures may have significant influence opposing the adoption of a more patient-centered approach as a primary outcome in TED assessment, despite ample evidence that the social dysfunction associated with TED typically outweighs the physical changes.

A robust, valid, and reliable patient-centered QOL measure would go a long way toward enacting a shift to more thorough assessment of TED disease states and the influence of clinical management on it. To date, only the GO-QOL and TED-QOL (52) have been correlated with clinical outcome, with the GO-QOL utilized the most in clinical trials (10, 40, 42, 43, 45, 57).

An ideal TED-related QOL measure would have excellent face validity, parsimoniously broad, be understandable and intuitive to patients and clinicians, generalizable to multiple populations of patients (culturally and by disease severity), and include psychometric characteristics. Each is explored below.

**Face Validity**

The core of any measure is fundamentally face validity, otherwise referred to as content validity. Face validity represents the extent to which an instrument measures the concept of interest (58). In a measurement tool, this is reflected by the degree to which the scores are representative of the patient experience.

To maintain strong face validity, the items and domains within an instrument should be developed by patients, based on freeform discussion of their experiences in a nonjudgmental environment devoid of constraints imposed by treatment relationships.

Various methodologies have been employed for item development involving a range of methods and conceptual frameworks. With regard to TED, the STED-QoL may be the best example of such a patient-centered item and domain generation approach. The use of focus groups to achieve idea saturation placed the responsibility of defining the critical issues

### Table 2. Similar items on 4 TED-specific quality of life measures

| GO-QOL(47) | GO-QOL(5) | TED-QOL(4) | STED-QoL(53) |
|------------|-----------|------------|--------------|
| **Daily activities** | **Noticing objects or activities** | **How is your eye disease currently affecting your ability to carry out daily activities?** | **Are your daily activities affected by your vision or walking down the stairs?** |
| **Driving** | **How is your eye disease currently affecting your ability to drive?** | **Are your daily activities affected because of TED?** | **Do you change your appearance to camouflage your appearance because of TED?** |
| **Subjective personal assessment of appearance** | **How is your eye disease currently affecting your appearance?** | **Do you feel confident to drive because of TED?** | **Do you feel that your appearance has changed because of your thyroid eye disease?** |
| **Self confidence** | **How is your eye disease currently affecting your self-confidence?** | **Do you feel that your appearance is still the same because of TED?** | **Do you feel that your appearance has changed because of your thyroid eye disease?** |
| **Photography** | **Do you feel confident to drive because of TED?** | **Do you feel that your self-esteem is affected because of TED?** | **Do you feel that you appear less often on photos than before you had thyroid eye disease?** |
| **Masking changes** | **Do you feel that you appear less often on photos than before you had thyroid eye disease?** | **Do you change your appearance to camouflage your appearance because of TED?** | **Do you feel that your appearance has changed because of your thyroid eye disease?** |

**Abbreviations:** GO; Graves’ ophthalmopathy; QLS, quality of life scale; TED-QoL, Singapore Thyroid Eye Disease Quality of Life; TED, thyroid eye disease.
in TED-related QoL on the patients (53). The use of semi-structured interviews influences the agenda for discussion, and focus groups in general can be subjected to groupthink and other social context biases. These challenges are balanced against alternatives such as open-ended interviews, which can lack iterative thought generation and community consensus. Both approaches have advantages.

The STED-QoL appears the most robust instrument to date in terms of content validity of all available, and, like all its comparators, could be improved. Some improvement could be made by involving individual, more open-ended interviews and through the addition of an iterative feedback process in terms of item understandability and ease of use after operationalization, as was performed in developing the GO-QoL (47).

Parsimoniously Broad

Striking a balance between maintaining enough depth in assessment and avoiding duplication and response fatigue is critical in deploying an instrument on a large scale. On 1 extreme, the TED-QoL includes only 3 items, likely insufficient to discriminate reliably across a diverse heterogenous population (52). At the other end of this spectrum is the original 105 item GO-QoL (47), which is far too long for practical use and carried significant redundancy, as evidenced by quantitative item reduction to a total of 9 (47). In between are the 16-item Go-QoL (5) and the 10-item STED-QoL (53). These last 2 likely lie in the range of appropriate parsimony, with the caveat that the very limited use of STED-QoL will require a lot more studying before further comments can be made about its clinical use.

Practical

The administration of a widely utilized tool would ideally would be simple and straightforward, requiring minimal input from trained survey administrators. Self-administered questionnaires have several advantages including expediency, economic efficiency, relative reliability, and can be electronically distributed. For self-administration the QoL tools should be easily understandable by patients and avoid confusing language. In this regard, the TED-QoL and GO-QoL perform reasonably well; however, the GO-QoL was less comprehensible (4). It is notable that this was an informal assessment and was not performed during item development. Although some patients were consulted regarding GO-QoL comprehensibility, these results are not extensively reported (5). During TED-QoL development, subjects were queried regarding comprehensibility and, as it performed well, no changes were made based on this (4). In the future, an iterative process of item operationalization involving comprehensibility assessment by patients may lead to a more reliable instrument.

Generalizable

It is vital that a measurement tool performs accurately across a range of disease states and social-cultural situations. As such the results can be appropriate for clinical assessment in diverse practices and research findings can be appropriately applied by clinicians broadly. In both development and validation, utilizing diverse populations representative of the range of severities for TED and the social-cultural contexts in which they experience disease are critical in the determination of generalizability. In this regard, general health questionnaires (eg, SF-36 and NEI-VFQ) are likely to have an advantage, compared with the TED-specific ones.

TED-specific instruments have proven to be limited in the inclusion of either a range of severity or diversity in populations during development. From this group though GO-QoL is likely the most robust. Having been translated into many languages and trialed in a wide range of populations, it has the potential for broad based utility. Studies have shown similar scoring across different jurisdictions, although these have not been directly compared (33, 35, 36). The TED-QoL has also been assessed across cultural circumstances (in a more limited manner) and produced similar results (50).

The limited sociocultural inclusion is consistent in all available scales, being mostly developed in a single population. This is most notable in the GO-QoL, which includes bicycle riding as a typical activity of daily living. This is of course highly variable across populations and modified versions have eliminated or mitigated this item (31). For example, GO-QoL now includes check boxes for “never learned to ride a bike” or “no driver’s license,” with the calculated score adjusted if these boxes were checked. There is room in this space for the inclusion of broader severity and sociocultural inclusion. Each of these instruments were developed in homogenous populations (eg, Asian, White European) and with a bias toward moderate to severe inflammatory phenotype patients (4, 38).

Psychometric Performance

There is an extensive science of psychometric assessment, and details of this are beyond the scope of the present discussion. Both validity and reliability can be assessed in a quantitative fashion, and psychometric performance characteristics will influence the quality of expected results.

Content validity is the most critical of performance characteristics; however, both construct and discriminant validity characteristics can confirm the overall validity of the instrument after item development and iterative refinement. Construct validity describes the way in which instrument performance conforms to a priori hypotheses. This is most commonly tested by comparing the measure with other measures. In this regard the GO-QoL (59) has been assessed and found to correlate well with some aspects of commonly used QoL measures such as the Medical Outcomes Study Short-form General Health Survey (29) and the Sickness Impact Profile (14). As the items on the GO-QoL were derived from general QoL questionnaires, these too correlated well with other measures. However, when these measures are tested for association with specific physical characteristics of TED, the performance is less than ideal. As noted previously this is an area of interest for future research.

Discriminant validity describes the ability of an instrument to distinguish between qualitatively different disease states. Each of the TED-specific instruments does appear to demonstrate reasonable discriminant ability having been tested across disease severity states (5, 22, 31, 33, 36, 47, 52, 53, 60) and over the course of treatment (9, 10, 18, 37, 40, 41, 43-45).

Reliability can be defined as a measure’s ability to produce consistent, reproducible estimates of the disease state across a range of settings. The most reliable instruments will demonstrate both internal consistency (often measured with
Chronbach alpha) and test–retest reliability. Administration on large diverse populations, at multiple time points and quantitative psychometric analysis is required to understand these performance characteristics.

Test–retest reliability was assessed in a cohort of 100 patents by the group validating the TED-QOL (4). They found good performance for the TED-QOL, GO-QOL, and GO-QLS (4, 38). Another study on the GO-QOL alone also demonstrated similar internal consistency and test–retest reliability for both the visual and appearance subscales (59). Further studies on the STED-QoL may be useful to confirm these aspects of the scale performance.

Concluding Remarks

It is clear from both clinical experience and extensive research that TED has a significant impact on QOL. This impact is apparent out of proportion to (currently) measurable physical manifestations and spans across visual, social, vocational, and psychologic domains. Characterization of the disease state and response to therapy should involve a complete assessment of individual disease impact and thus TED-specific QOL measures have a critical role to play in the management of TED. Significant progress has been made over the years and TED-specific QOL measures have been developed. Each has some advantages and some disadvantages, and each demonstrates challenges in item construction and subsequent psychometric performance. There is considerable opportunity in this space to enhance the breadth and precision of TED-related QOL assessment and the ideal measure may not be currently available. Widespread adoption to QOL assessment in basic TED care may follow the promotion of such a scale. The GO-QOL was the most extensively studied, the TED-QOL is the easiest to complete, the GO-QLS correlated best with baseline disease severity, and the STED-QOL incorporated the patients’ experience the most in its creation. Until an ideal, universal scale has been developed, the GO-QOL would ideally be utilized in clinical research and academic settings given its robust reliability across populations. What is most ideal is to be consistent within a clinical practice to track QOL over time and assess the efficacy of interventions.

Disclosures

A.S., M.S., and D.R. have acted as consultants for Horizon Pharmaceuticals.

Financial support

None.

Data Availability Statement

Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

Reference

1. WHO. The World Health Organization quality of life assessment (WHOQOL); position paper from the World Health Organization. Soc Sci Med. 1995;41(10):1403-1409.
22. Kulig G, Andrysiak-Mamos E, Sowińska-Przepiera E, et al. Quality of life assessment in patients with Graves’ disease and progressive infiltrative ophthalmopathy during combined treatment with methylprednisolone and orbital radiotherapy. Endokrynol Pol. 2009;60(3):158-165.

23. Kahaly GJ, Petrak F, Hardt J, Pitz S, Egle UT. Psychosocial morbidity of Graves’ orbitopathy. Clin Endocrinol (Oxf). 2005;63(4):395-402.

24. Bradley EA, Sloan JA, Novotny PJ, Garrity JA, Woog JJ, West SK. Evaluation of the National Eye Institute visual function questionnaire in Graves’ ophthalmopathy. Ophthalmology. 2006;113(8):1450-1454.

25. Du Y, Ye H, Li K, et al. Vision-related quality of life tends to be more severely impaired in patients with dysthyroid optic neuropathy. Curr Eye Res. 2014;39(5):532-536.

26. Watt T, Hagedus L, Rasmussen AK, et al. Which domains of thyroid-related quality of life are most relevant? Patients and clinicians provide complementary perspectives. Thyroid. 2007;17(7):647-654.

27. Steinberg EP, Tielsch JM, Schein OD, et al. The VF-14. An index of functional impairment in patients with cataract. Arch Ophthalmol. 1994;112(5):630-638.

28. Mangione CM, Phillips RS, Seddon JM, et al. Development of the ‘Activities of Daily Vision Scale’. A measure of visual functional status. Med Care. 1992;30(12):1111-1126.

29. Stewart AL, Hays RD, Ware JE Jr. The MOS short-form general health survey. Reliability and validity in a patient population. Med Care. 1988;26(7):724-735.

30. Bartalena L, Wiersinga WM. Proposal for standardization of primary and secondary outcomes in patients with active, moderate-to-severe Graves’ orbitopathy. Eur Thyroid J. 2020;9(Suppl 1):3-16.

31. Park JJ, Sullivan TJ, Mortimer RH, Wagenaar M, Perry-Keene DA. Development of the ‘Activities of Daily Vision Scale’. A measure of visual functional status. J Endocrinol (Oxf). 2001;185(4):G43-G67.

32. Kashkouli MB, Heidari I, Pakdel F, et al. Assessing quality of life in Taiwanese patients with Graves’ ophthalmopathy. J Ophthalmol. 2009;48(1):75-78.

33. Choi YJ, Lim HT, Lee SJ, Lee SY, Yoon JS. Assessing Graves’ ophthalmopathy-specific quality of life in Korean patients. Eye 2012;26(4):544-551.

34. Delampady K, Reddy SVB, Yadav SB. Assessing the quality of life in Indian Graves’ orbitopathy patients and validation of Hindi version of GO-QOL questionnaire. Indian J Ophthalmol. 2020;68(8):1617-1621.

35. Ponto KA, Hommel G, Pitz S, Elflein H, Pfeiffer N, Kahaly GJ. Quality of life in a German Graves orbitopathy population. Am J Ophthalmol. 2011;152(3):483-490.e1.

36. Sawicka-Gutaj N, Bednarczuk T, Daroszewski J, et al. Quality of life assessment in patients with Graves’ orbitopathy. J Ophthalmol. 2015;66(4):362-366.

37. Kashkouli MB, Heidari I, Pakdel F, et al. Assessing quality of life in Taiwanese patients with Graves’ ophthalmopathy. J Formos Med Assoc. 2015;114(11):1047-1054.

38. Kashkouli MB, Heidari I, Pakdel F, Jam S, Honarbaksh Y, Miramandehi B. Change in quality of life after medical and surgical treatment of Graves’ ophthalmopathy. Middle East Afr J Ophthalmol. 2011;18(1):42-47.

39. Terwee CB, Dekker FW, Mourits MP, et al. Interpretation and validity of changes in scores on the Graves’ ophthalmopathy quality of life questionnaire (GO-QOL) after different treatments. Clin Endocrinol (Oxf). 2001;54(3):391-398.

40. Bartalena L, Kahaly GJ, Baldeschi L, et al. The 2021 European Group on Graves’ orbitopathy (EUGOGO) clinical practice guidelines for the medical management of Graves’ orbitopathy. Eur J Endocrinol. 2021;185(4):G43-G67.

41. Douglas RS, Kahaly GJ, Patel A, et al. Teprotumumab for the treatment of active thyroid eye disease. N Engl J Med. 2020;382(4):341-352.

42. Barthela L, Krasas GE, Wiersinga W, et al.; European Group of Graves’ Orbitopathy. Efficacy and safety of three different cumulative doses of intravenous methylprednisolone for moderate to severe and active Graves’ orbitopathy. J Clin Endocrinol Metab. 2012;97(12):4454-4463.

43. Perez-Moreiras JV, Gomez-Reino JJ, Maneiro JR, et al. Efficacy of tocolizumab in patients with moderate-to-severe corticosteroid-resistant Graves orbitopathy: a randomized clinical trial. Am J Ophthalmol. 2018;195:181-190.

44. Mourits MP, Bijl H, Altea MA, et al; European Group on Graves. Outcome of orbital decompression for disfiguring proptosis in patients with Graves’ orbitopathy using various surgical procedures. Br J Ophthalmol. 2009;93(11):1518-1523.

45. Stan MN, Garrity JA, Carranza Leon BG, Prabin T, Bradley EA, Bahn RS. Randomized controlled trial of rituximab in patients with Graves’ orbitopathy. J Clin Endocrinol Metab. 2015;102(2):432-441.

46. Lee THB, Sundar G. Quality of life in thyroid eye disease: a systematic review. Ophthalmic Plast Reconstr Surg. 2020;36(2):118-126.

47. Yeatts RP. Quality of life in patients with Graves ophthalmopathy. Trans Am Ophthalmol Soc. 2005;103:368-411.

48. Ware J Jr, Kosinski M, Keller SD. A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity. Med Care. 1996;34(3):220-233.

49. Anderson RT, Rajagopalan R. Development and validation of a quality of life instrument for cutaneous diseases. J Am Acad Dermatol. 1997;37(1):41-50.

50. Son BJ, Lee SY, Yoon JS. Evaluation of thyroid eye disease: quality-of-life questionnaire (TOD-QOL) in Korean patients. Can J Ophthalmol. 2014;49(2):167-173.

51. Han I, Wang H, Li X, Li J, Lin G, Yan J. Assessment of quality of life in Chinese patients with thyroid associated orbitopathy. Biomed Res (Tokyo) 2017;28(4):1774-1780.

52. Fayers T, Fayers PM, Dolman PJ. Sensitivity and responsiveness of the patient-reported TOD-QOL to rehabilitative surgery in thyroid eye disease. Orbit. 2016;35(6):328-334.

53. Wong MHY, Fenwick E, Aw AT, Lamoureux EL, Seah LL. Development and validation of the Singapore thyroid eye disease quality of life questionnaire. Trans Vis Sci Technol. 2018;7(5):14.

54. Terwee CB, Dekker FW, Bonsel GJ, et al. Facial disfigurement: is it in the eye of the beholder? A study in patients with Graves’ ophthalmopathy. Clin Endocrinol (Oxf). 2003;58(2):192-198.

55. Bahmani-Kashkouli M, Pakdel F, Astrakhi A, et al. Quality of life in patients with thyroid eye disease. J Ophthalmic Vis Res. 2009;4(3):164-168.

56. Ponto KA, Merkasdsal F, Hommel G, Pitz S, Pfeiffer N, Kahaly GJ. Public health relevance of Graves’ orbitopathy. Eur J Ophthalmol. 2018;28(4):1774-1780.